CellMinerCDB: NCATS is a Web-Based Portal Integrating Public Cancer Cell Line Databases for Pharmacogenomic Explorations

William C. Reinhold[1,*], **Kelli Wilson**[2,*], **Fathi Elloumi**[1], **Katie R. Bradwell**[3], **Michele Ceribelli**[2], **Sudhir Varma**[1,4], **Yanghsin Wang**[1,5], **Damien Duveau**[2], **Nikhil Menon**[2], **Jane Trepel**[1], **Xiaohu Zhang**[2], **Carleen Klumpp-Thomas**[2], **Samuel Micheal**[2], **Paul Shinn**[2,#], **Augustin Luna**[6], **Craig Thomas**[2], **Yves Pommier**[1]

[1]Developmental Therapeutics Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20892, USA

[2]National Center for Advancing Translational Sciences, NIH Bethesda, MD 20892, USA

[3]Palantir Technologies, Denver, CO 80202, USA

[4]HiThru Analytics LLC, Princeton, NJ 08540, USA

[5]ICF International Inc., Fairfax, VA 22031, USA

[6]cBio Center, Dana-Farber Cancer Institute and Department of Cell Biology, Harvard Medical School, Boston, MA 02215, USA.

Abstract

Major advances have been made in the field of precision medicine for treating cancer. However, many open questions remain that need to be answered to realize the goal of matching every cancer patient to the most efficacious therapy. To facilitate these efforts, we have developed CellMinerCDB: NCATS (https://discover.nci.nih.gov/rsconnect/cellminercdb_ncats/), which makes available activity information for 2,675 drugs and compounds, including multiple non-oncology drugs and 1,866 drugs and compounds unique to the National Center for Advancing Translational Sciences (NCATS). CellMinerCDB: NCATS comprises 183 cancer cell lines with 72 unique to NCATS including some from previously understudied tissues of origin. Multiple forms of data from different institutes are integrated, including single and combination drug activity, DNA copy number, methylation and mutation, transcriptome, protein levels, histone acetylation and methylation, metabolites, CRISPR and miscellaneous signatures. Curation of cell lines and drug names enables cross-database (CDB) analyses. Comparison of the datasets is made possible by the overlap between cell lines and drugs across databases. Multiple univariate and multivariate analysis tools are built-in, including linear regression and LASSO. Examples have been presented here for the clinical topoisomerase I (TOP1) inhibitors topotecan and irinotecan/SN-38. This web-application provides both substantial new data and significant pharmacogenomic integration allowing exploration of interrelationships.

Corresponding author: **Name:** William C. Reinhold, **Address:** National Institutes of Health, 9000 Rockville Pike, Building 37, room 5041, Bethesda, MD 20892, USA, **Phone:** 240-760-7339, wcr@mail.nih.gov.

[#]deceased

[*]Co-first authors.

The authors declare no potential conflicts of interest.

Introduction

The approach to molecular biology and pharmacology, commonly referred as Precision Medicine has been significantly changed over the last ~25 years by the introduction of omics data and the conceptual shift to the use of computer analyses of large datasets with a combination of statistics, machine learning, omics visualizations and integration of multiple disparate forms of data. Starting with the pioneering work of the Developmental Therapeutics Program (DTP) at the National Cancer Institute (NCI) (1), many projects have been and are contributing sizable blocks of data, prominently including (but not limited to) the large (~1,000 cell line) panels of the Cancer Cell Line Encyclopedia (CCLE) from the Broad/Novartis, the Genomics of Drug Sensitivity in Cancer (GDSC) from Sanger and Massachusetts General Hospital and the Cancer Therapeutics Response Portal (CTRP) from the Broad Institute.

The Genomics and Pharmacology Facility (GPF) has pioneered omics data acquisition and integration since the mid 1990's (1–9). Its efforts have led to the CellMiner and CellMinerCDB web-application (2–7,9,10) allowing pharmacogenomic database access and integrative analyses across all public cancer cell line genomics and drug response databases (2).

NCATS has established an automated compound screening platform for large compound libraries using quantitative high-throughput (qHTS) format across multiple different disease models since 2008 (11–13). For cancer cell line viability screening, NCATS created the Mechanism Interrogation PlatEs (MIPE) compound library comprising approved and investigational chemotherapeutic agents, as well as common medications for non-cancer indications. An additional design feature of the MIPE library is compound mechanistic redundancy allowing analyses across multiple compounds reported to hit the same target. Compound screening data using the MIPE library has demonstrated value for multiple cancer types, such as diffuse intrinsic pontine glioma (DIPG), Hodgkin's lymphoma, Ewing sarcoma, small cell lung cancer, glioblastoma and others (9,14–17). Published and unpublished MIPE library compound screening data have been aggregated into a unified dataset called the NCATS-NCI Cytoxicity Dataset shared internally with the NCI through the Palantir Foundry platform. A subset of this unified dataset is now being made public through CellMinerCDB.

Here we introduce the public databases and web-portal of CellMinerCDB: NCATS (https://discover.nci.nih.gov/rsconnect/cellminercdb_ncats/). CellMinerCDB: NCATS enables individual users to access and explore the large NCATS drug response database, with an emphasis on pharmacology and its relationships to molecular genomics. CellMinerCDB: NCATS is integrated with 33 datasets from multiple projects from DTP, GPF, CCLE, GDSC, CTRP, the NCI DTP Small Cell Lung Cancer Project (NCI SCLC), NCI60-DTP Almanac, MD Anderson and the Project Achilles from the DepMap portal (see "Supplementary Materials and Methods" for a full listing) (4,5,7,18–28). The omics analyses include single and two-drug activities, DNA copy number, methylation and sequencing, whole genome transcriptome, microRNA and selected protein expression, metabolite levels, and clustered regularly interspaced short palindromic repeats (CRISPR) knockouts, allowing explorations of the relationships between those data and pharmacological responses. Functionalities of the new CellMinerCDB: NCATS web-application are introduced and discussed here with multiple examples validating the database. Details about general functionalities of the CellMinerCDB (https://discover.nci.nih.gov/rsconnect/cellminercdb/) platforms have been reviewed recently (2) and a 10-minute tutorial is on YouTube (see Figure 1A).

CellMinerCDB: NCATS is a public web-application hosted in the Genomics and Pharmacology Facility of the Developmental Therapeutics Branch of the NCI Center for Cancer Research, and of the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH).

Materials and Methods

The NCATS screening data contained within the CellMinerCDB: NCATS web-application utilize RSTUDIO-2022.12.0-353 and were generated as previously described (29). Cells were treated with compounds for 48 hours in 1536 well plates and assessed for viability using CellTiter Glo (Promega, Madison, WI, USA). Data were normalized to plate controls of DMSO treated cells as 100% viability and no cells at 0% viability. A four-parameter curve fit was used to generate an IC_{50} and Area Under the Curve (AUC). Z-score AUC (across cell lines) was calculated by subtracting the mean AUC and dividing by the standard deviation of each drug across all cell lines screened.

All compounds were matched using SMILES and InChIKey to external databases to pull clinical status. NCATS Inxight, DrugBank, and CHEMBL were used as references for compound structure matching and global clinical status (30–32). Structure matching was done within the Palantir Foundry platform (Palantir Technologies, Denver CO) utilizing RDKit: Open-source cheminformatics (2021_09_4 (Q3 2021) Release); and NCATSFind Resolver. NCATS cell lines were annotated internally using Cellosaurus for disease and tissue type and matched to the other cell line sets (33). The NCATS web-application is an R/shiny app hosted on an NCI server.

Information sources for the cell lines and drugs include the NCI Thesaurus, PubChem and the scientific literature. The large amount of data coming from the included omics efforts and the platforms used to develop them has been previously described. Compound and cell line name variation across the different institutions cell line sets were resolved internally. An example is a single compound with the names 122958 (NCI-60), ATRA (GDSC), tretinoin (CTRP), and isotretinoin (NCATS). Another example is a single cell line with the names CO:COLO 205 (NCI-60), COLO 205 (CCLE), COLO-205 (GDSC), COLO205 (MD Anderson). All datasets have instances of missing data for specific cell lines, drugs or genes.

Univariate Analysis and Multivariate Analysis shown throughout were done using CellMinerCDB: NCATS functionalities or using data downloaded directly from CellMinerCDB: NCATS. The web-application generated scatterplots, tables and heatmap shown were generated using the selections described in the input boxes and figure legends. Drug versus drug activity comparisons not generated by the web-application were done by

Pearson correlation using R version 3.6.3. Bar charts were generated using GraphPad Prism version 7.0. Violin plots were generated using ggplot version 3.3.5.

Bimodal drug activity density distributions were identified using a combination of a Gaussian Mixed Model-based (*norm1mix* package (version 1.3), a kurtosis test and visual inspection. Both these calculations and the density plots were done using The R Project for Statistical Computing.

Prediction of NCATS IC50 activity using CCLE microarray transcript expression by both univariate and multivariate analysis used Pearson's correlation between drug response and gene expression of the target. The multivariate models use stepwise forward regression. Each model was initiated with a target for a given drug; multiple targets generated multiple models. Possible regression features included genes from Onco500 (34). A maximum of 10 features were added to each model and then pruned. For each iteration step, the feature with the lowest partial correlation p-value after removing the effects of already included features was added using rcellminer 2.9.1 (35). A 10-fold cross-validated predicted response was calculated at each step using rcellminerElasticNet 0.1.1. Models were pruned by examining the statistical difference in the correlation of predicted versus observed response with each added feature using cocor 1.1–3. CCLE microarray expression data from CellMinerCDB was used (2).

Data Availability

The data analyzed in this study were obtained from multiple sources. Within the application, the source of each data set is accessible within the Metadata tab, both within the "Select here to learn more about…" link and from the "Download Footnotes" tab. A description of all data sources used in CellMinerCDB: NCATS is provided in the "Supplementary Materials and Methods".

Results

The CellMinerCDB: NCATS web-application

The CellMinerCDB: NCATS publicly accessible web-application was created to both access the NCATS drug response data and enrich and expand its usefulness by integrating multiple other forms and sources of genomics, proteomics, and metabolomics data from the other public cancer cell line datasets using the CellMinerCDB platform (2).

A screenshot of the site, banner, and tabs for the CellMinerCDB: NCATS web-application is presented in Figure 1A. CellMinerCDB: NCATS allows drug comparisons and emphasizes cross-database (CDB) analyses with the other public cancer cell line databases. The *Univariate Analyses* tab allows generation of on-the-fly bivariate scatter plots and correlation analyses from a single input to compare all profiles within selected data sets. The *Multivariate Analyses* tab allows the exploration of multivariate models predictive of an observed profile. Analyzing selected tissues of origin is an option for both univariate and multivariate analyses. The *Metadata* tab allows the download of datasets of interest for further processing and archiving. The *Search IDs* tab provides the identifiers within each cell line set by data type. The *Help* tab provides explanations and descriptions of the various functionalities within the web-application. Additionally, the *Video Tutorial* tab provides a description and explanation of the CellMinerCDB functionalities. Thus, CellMinerCDB: NCATS provides new data, multiple functionalities, and data integration, allowing users to mine independently the NCATS data without having to seek support from bioinformatics teams.

The NCATS input data

CellMinerCDB: NCATS comprises 2,675 drugs and compounds tested in 183 cell lines; 2,667 of which have mechanism of action designations. The dataset was created as described in Methods and Figure 1B. The output is fully compatible and integrated with CellMinerCDB (2). An asset of CellMinerCDB: NCATS is the unique compounds and cancer cell lines included (Figure 1C and D).

NCATS contains two drug sensitivity metrics, Z-Area Under the Curve (AUC) and IC50 values. These boast a large range of screening concentrations, routinely using 11 concentrations between 0.79 nanomolar and 47 micromolar, which is an asset of NCATS drug testing (12). The drugs include 952 (36%) clinically approved, 790 (30%) that have entered clinical trials, and 908 drugs (34%) that are preclinical (Figure 1C, left). Notably, 1,877 (70%) drugs and compounds are unique to NCATS (Figure 1C, right). They have been annotated with their commonly accepted mechanisms of action. A feature of the NCATS dataset is the inclusion of 518 approved non-oncology drugs not found in the other public databases (Supplemental Table 1). Those include 103 anti-infectives (anti-bacterial, mycobacterial, viral, or fungal) for systemic use, 86 cardiovascular or nervous system drugs, 72 alimentary tract and metabolism compounds.

The 183 NCATS cell lines distribution by tissue of origin is detailed in Supplemental Table 2. They include 72 (38%) unique cancer cell lines absent in other public cancer cell line databases (Figure 1D and Supplemental Table 3). Figure 1D shows several of the rare disease subtypes including diffuse intrinsic pontine gliomas (DIPG), renal Birt-Hogg Dubé syndrome, hereditary leiomyomatosis and TFE3 fusion cancer cell lines. details the. Thus, CellMinerCDB: NCATS provides the user with substantial new drug and cell line data.

Cell line and drug overlaps of NCATS with other cancer cell line datasets

The cell lines overlaps for CellMinerCDB: NCATS as well as all other cell line sets are listed in Figure 2A. As in our other CellMinerCDB websites (https://discover.nci.nih.gov/), cell lines are matched with common tissue of origin terms based on the OncoTree ontology levels developed by the Memorial Sloan Kettering Cancer Center and Dana-Farber Cancer Institute, primarily version 1.1 as described previously (2). Additional information such as patient gender or age from which the cell line originated are also included. Comparison between drug responses in cell lines is made possible by the overlap of cell lines across databases (Figure 2A).

The drug and compound activity overlap between the multiple cell line sets is presented in Figure 2B. Information on each cell line set activity measurements are accessible in the “Data Type” input box, Metadata “Units” description or footnotes, or the provided urls. An asset for the user is that CellMinerCDB: NCATS automatically matches cell line and drug data across any cell line sets queried, which allows their comparison for identical, related by mechanism of action, or disparate drugs.

Omics data available for cross-comparisons in CellMinerCDB: NCATS

Figure 2C summarizes by cell line set and measurement type the profiles available in CellMinerCDB-NCATS, including 31,617 drug (and compound) activities, 261,848 molecular measurements and 18,119 miscellaneous signatures. All 28 included datasets are available for download from the Metadata tab (see Fig. 1A). Our curation and standardization of these datasets minimizes the task of name matching.

The data types available for exploration based on the databases with overlapping cell lines include single-drug activities, two-drug combination activities, gene copy number, methylation and mutation levels, transcript expression, protein expression, metabolite levels, the DepMap Achilles (Achilles) CRISPR genetic dependencies, and miscellaneous molecular signatures. Those miscellaneous phenotypic signatures include the antigen presenting machinery ("APM"), epithelial mesenchymal status ("EMT"), replication stress ("RepStress"), genomic instability ("HRD_LOH", "HRD-SUM", "NtAI", "LST") and neuroendocrine status ("NE"). The metadata phenotypic signatures are accessible in the *Univariate Analyses | Data Type | mda: Miscellaneous phenotypic data*. The number of data explorations one might pursue, depending on one's interest, easily jumps into the billions. The NCATS drug data can be compared to genomics data for the same cell lines in other datasets allowing one to relate the drug responses to omics features using CellMinerCDB: NCATS. The following examples illustrate the basic use of CellMinerCDB: NCATS.

Drug comparisons

The overlaps between cell lines and drugs across the "*Cell Line Sets*" facilitate multiple forms of drug comparisons. Figure 3A shows a *Univariate Analyses/Plot Data* output for two structurally related TOP1 inhibitors commonly used in clinical oncology (36), topotecan (x-axis) versus SN-38 (y-axis), the active metabolite of irinotecan. Both are measured by NCATS and displayed using CellMinerCDB-NCATS. The highly significant correlation between the two drugs ($p=9.1\times10^{-52}$) demonstrates internal assay consistency.

Similarly, Figure 3B shows a *Univariate Analyses/Compare Patterns* comparing the NCATS ALK inhibitor TAE-684 to other NCATS ALK inhibitors (by entering ALK inhibitor in the output MOA column). Of the 12 ALK inhibitors in the NCATS database, 10 show significant correlations demonstrating assay and mechanism of action reproducibility across cell lines within the NCATS drug response database.

Comparison of NCATS with GDSC and CTRP drug activities in Figure 3C and D, respectively shows the top 15 correlated compounds for each. Four protein kinase inhibitors are common between these two (shown as red bars): linifanib, sorafenib, AZD-7762 and tivozanib. Figure 3E is a *Univariate Analyses/Plot Data* analysis of one of these comparisons, AZD-7762 as measured by both NCATS (x-axis) and CTRP (y-axis), yielding a p-value of 1.1×10^{-10}. These observations demonstrate ways of comparing drug activities across databases to determine consistency across common cell line sets.

Compared globally, the average Pearson's correlation for NCATS versus either GDSC or CTRP across all compounds using Z-AUC or IC50 is 0.4. Violin plots (Figure 3F) visualize significant correlations between NCATS and 102/265 compounds (38.4%) for CTRP and 71/212 compounds (33.5%) for GDSC. The NCATS versus the PRISM drug data are not included in this analysis as none had the minimum 16 cell lines with overlap. The Figure 3 examples are only a small sampling of the types of informative comparisons one might do.

Exploration of NCATS drug responses with omics or CRISPR data

The integration of the NCATS drug responses with a wide range of molecular, phenotypic and signature data from the other omics databases (CCLE, GDSC, NCI) allows correlation queries for overlapping cell lines. We next present a small group of these as illustrations with outputs and screenshots from CellMinerCDB: NCATS.

Figure 4A validates SN-38 activity (in NCATS) versus *SLFN11* gene transcript expression (in GDSC) using CellMinerCDB: NCATS *Univariate Analyses*/*Plot data*. The scatter plot confirms the expected significant correlation between these causally linked parameters (36). Figure 4B presents additional examples between NCATS and GDSC; all showing significant correlation between a drug's activity and the transcript levels of that drug target.

A second form of omics data comparison is given in Figure 4C, comparing activity of the mTOR inhibitor VS-5584 from NCATS and MTOR DNA copy number from CCLE demonstrating significant correlation. CellMinerCDB: NCATS also shows that MTOR DNA copy number is significantly correlated to its transcript level (r=0.49, p=1.6E-61), providing the logical link between the drug activity and DNA copy number. Figure 4D provides additional examples of significant correlations between drug activities and DNA copy numbers; all linked through having the same gene both as drug target and molecular measurement. All have significant correlations between gene DNA copy number and transcript levels.

Figures 4E and F exemplify the possibility of testing NCATS drug activity versus genetic inactivation of the drug target. Figure 4E compares the growth inhibitory activity of vemurafenib (a BRAF inhibitor) to cell survival with *BRAF* CRISPR knockdown (as measured by Project Achilles). The resultant scatter plot demonstrates significant correlation between the two. Figure 4F lists other examples showing significant correlations between drug activities and CRISPR knockdown; in each case linked through having the same gene both as the drug and CRISPR target. As for the drugs in Figure 3, Figure 4 provides only a small sampling of the types of informative comparisons one might do.

To compare the predictive value of different genomics parameters, the NCATS approved and clinical trial drugs IC50 activities were each compared to the different genomics evaluations of their gene targets (transcript expression, gene copy number, methylation, mutations and CRISPR) across the other nine platforms in CellMinerCDB: NCATS resulting in 1,100 drug versus gene pairings (see Supplemental Table 4). The percent significant correlations by platform were: i) 5.3% for the CCLE DNA copy number, ii) 8.8% for the GDSC methylation, iii) 6.5% for the CCLE mutation, iv) 5.1% for GDSC mutation, v) 11.8% for the CCLE transcript microarray, vi) 10.8% for the GDSC microarray, vii) 12.8% for the

CCLE RNA sequencing, viii) 9.2% for the CCLE protein and ix) 6.2% for the Achilles CRISPR. These results demonstrate the value of RNA sequencing and proteomic analyses for predicting drug activity.

Although determination of protein levels remains limited in clinical samples, we found that both protein expression and gene expression of the proapoptotic factor BAX in CCLE are significantly correlated with the IC50 activity of SN-38 in NCATS (Supplemental Figure 1; p-values = 0.0013 and 0.0026 for 88 and 95 common cell lines, respectively). Thus, based on the analysis of drugs tested in NCATS, we conclude that RNA-seq is currently the most practical predictor of drug response.

Multivariate and miscellaneous phenotypic signature (Mda) analyses using CellMinerCDB: NCATS

Presuming that multiple factors are involved in drug response (36), we present two approaches for clinical TOP1 inhibitors (topotecan and SN-38, the active metabolite of irinotecan) using CellMinerCDB: NCATS.

The first utilizes the prior knowledge that the cytotoxicity of TOP1 inhibitors are dependent on SLFN11, apoptosis and transcription (36). Combining transcript expression of SLFN11 (Figure 5A), BPTF (Figure 5B) and HMGN1 (Figure 5C) shows how the predictive value of SLFN11 can be strengthened by using the multivariate analysis tool of NCATS:CDB (Figure 5D and E).

The second multivariate analysis available in NCATS:CDB (and the other CellMinerCDB websites) uses previously described multi-gene expression signatures, which can be retrieved using the "*mda*" tab in the "*Data Type*" pull-down menu at the left of the website (Figure 6). Together, these examples demonstrate the increased power of aggregating multiple genomic parameters to predict drug activity.

Drug activity distributions and additional multivariate analysis

Figure 7 presents another form of exploration generated from the NCATS drug database: drug activity distributions with consideration of tissues of origin. Bimodal drug distributions were identified, demonstrating both sensitive and resistant cancer cell line responses. Enrichment for specific tissues of origin in the activity peaks demonstrates novel prospective therapeutic indications. Multivariate analyses using CCLE transcriptomics visualize multivariate molecular predictors. The first example given is for filanesib, with it's bimodal activity distribution visualized in Figure 7A and the significant prediction of that activity by *KIF11*, *MYBBP1A* and *TNFRSF10D* ($p=1.2\times10^{-7}$) in Figure 7B. The second example given is for epothilone, with it's bimodal activity distribution visualized in Figure 7C and the significant prediction of that activity by *TUBB6*, *ABCG1*, *GSK3G* and *MLH1* ($p=1.2\times10^{-7}$) in Figure 7D. Diverse mechanisms of action drugs reveal enhanced activities for bladder, blood (leukemia), bone (sarcoma), bowel, brain and lymphatic cancer cells in Figure 7E.

Supplemental Table 5 presents an example of a more systematic pharmacological prediction approach of NCATS IC50 drug activity distributions using CCLE microarray transcript levels. Included are 63 significant gene-drug combinations in which the genes are known targets for those drugs. In the case of ABT-737 (a BH3 mimetic and BCL2 gene family inhibitor), the generated multivariate model includes two known targets: BCL2L2 and BCL2 (as given by NCATS annotation).

Discussion

Making the NCATS drug activities publicly available is a significant addition to the omics arena. CellMinerCDB: NCATS gathers the NCATS drug response database and integrates it with nine other genomic and proteomic projects (see Fig. 1). The NCATS 2,675 drugs and compounds is second only to the large NCI/DTP activity screening in number (2) (Figs. 1–2). Its high proportion of novel drugs, large number of non-oncology drugs and inclusion of many novel cell lines, including rare tumors add significantly to the omics cancer cell line field.

Our curation of both the cell line and drug names enables integration with our previous CellMiner databases (2,3,9). It also resolves differences, making data retrieval and comparisons available with an intuitive web application. This combined with the molecular, metabolic, phenotypic and signature data from NCI, CCLE, GDSC and other databases adds a myriad of informative molecular parameters for the purposes of exploration, discovery, prediction and verification of either previously known or novel relationships.

We find that the activity of drugs with similar mechanisms of action is in general internally consistent within NCATS and across the other drug databases (CCLE, CTRIP, GDSC, NCI) as shown in Figure 3. Activity variability for overlapping drugs between institutes is recognized and presumably comes from a combination of the type of robotics and biological techniques employed (37). NCATS uses 1536 well plates, with compounds added immediately after cell plating and 48-hour drug exposure. CTRP and GDSC use 384 well plates, with compounds added 24 hours after cells plating and 72-hour drug incubation. All three projects use CellTiter-Glo. It is unsurprising that drug activity assays done under different conditions might give different results. However, our analyses shows that multiple drugs and compounds perform similarly regardless of differences in assay parameters. Thus, our recommendation for pharmacogenomics exploration with CellMinerCDB: NCATS is to first perform inter-database analyses with drugs present in at least two platforms and prioritize drugs with consistent cytotoxicity response across databases.

CellMinerCDB: NCATS comprises two main analysis tools): "*Univariate Analyses*" and "*Multivariate Analyses*" (see Fig. 1A. The pharmacogenomics analyses shown in Figures 3–7, all generated within the CellMinerCDB: NCATS web-application, provide examples of the many types of analysis possible. With 14.7 billion drug activity versus gene molecular or phenotypic (CRISPR) measurements, practically, one is limited only by the number of questions and knowledge one has. This number does not include the many intergene molecular and interdrug activity comparisons one might do.

Figures 3A, 4A, 5A–E and Supplementary Figure 1 provide pharmacogenomic and proteomic explorations for SN-38, as prior work has causally related SLFN11 expression to the activity of TOP1 inhibitors (6,38–40). The additional transcript examples in Figure 4B, and DNA copy number examples in Figure 4C and D link various NCATS drugs to their molecular targets. The ability to perform gene knockdown (CRISPR) comparisons reflect how a gene knockdown measured in Project Achilles relates to response to drugs measured in NCATS. None of the 33 drug-target examples listed are FDA-approved biomarkers for their respective drugs; so each of them provides possible incentive for their development and use. One might easily expand this type of analysis to non-target, but biologically relevant genes based on domain knowledge.

When using “Univariate Analyses”, we find the transcript data are stronger predictors of pharmacological response than the other genomic data (gene mutations, copy number variation, or methylation) available in the cancer cell lines (Supplemental Table 4). Currently DNA mutation is a predominant biomarker used for drug prediction. Although we see the expected predictive value of BRAF mutations with the activity of vemurafenib and dabrafenib (Supplemental Figure 2), mutations only predict the activity of a relatively small subset of drugs routinely used in oncology. In addition to having reliable gene coverage and being implemented clinically RNA-seq data are advantageous for the construction of multi-gene signatures. The cell line superiority for the prediction of pharmacological response is likely to translate clinically over time, leading to its gaining dominance for that purpose.

Because pharmacological response is a product of multiple molecular factors, drug activity prediction or exploration is expected to be improved and tested using the “Multivariate Analyses” tools of CellMinerCDB: NCATS. Figures 5 provide examples of how building multigene analyses can be explored. This approach requires an understanding of the pathways and targets that determine drug response. Taking the example of SN-38 (the active metabolite of irinotecan) and topotecan (36), Figure 5 shows how “Multivariate Analyses” can be generated. CellMinerCDB also provides preexisting gene signatures. Figure 6 uses a precomputed multi-gene signature, the 18-transcript replication stress (RepStress) signature (29). Increased level of this stress parameter is significantly correlated with topotecan and SN-38 response, providing proof-of-principle and a testable preclinical model for RepStress as predictive for patient response to TOP1 inhibitors. Having precomputed signatures avoids looking up the reference, finding the genes involved, determining and then applying the algorithm for the cell line set of interest.

Downloading the data of CellMinerCDB: NCATS reveals drug activity distribution enrichments for some tissue of origins within the cancer cell line panels. All the cancer types enriched indicate prospective novel applications for those drugs, presumably with responsive subsets. Non-oncology drugs might also be studied. An example from Figure 7E is disulfiram, a drug used to discourage alcohol intake. Response to this drug is bimodal across the NCATS cancer cell lines, with improved activity in bone (sarcoma) cell lines. This result expands our prior work on the discovery of acetalax, another non-cancer drug, with activity in triple-negative breast cancer cell lines (3).

In summary, the wealth of information in the CellMinerCDB: NCATS web-application, albeit with its own limitations, allows basic and clinician researchers to explore pharmacogenomic relationships in either univariate or multivariate fashion. One may consider drug response in the context of multiple forms or combinations of outputs that easily run into the billions. The web-application facilitates the user's ability to explore those relationships and explore potential pharmacogenomic parameters applicable to clinical studies.

Limitations of the data come in multiple forms requiring multiple solutions. Missing data might be addressed by simply carrying out the salient form of analysis to fill those gaps.

More complete analysis of variability between platforms might be done by adding overlapping cell lines, drugs, or assays of interest. Algorithmic approaches that better consider the limitations and proper interpretation of datasets can improve results at that level, including the expansion of multivariate analysis functionality and approach selection. Recognitions of signatures predictive of pharmacological response should yield improved success in that area. It should be noted that the relationships found do not constitute proof of causality. The continued exploration and definition of how best to integrate cancer cell lines omics data with that from patients and to integrate clinical data into the omics format remain fields in their infancy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Our studies are supported by the Center for Cancer Research, the Intramural Program of the US National Cancer Institute, National Institutes of Health, Bethesda, Maryland (Z01 BC 006150).

References

1. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ Jr., Kohn KW, et al. An information-intensive approach to the molecular pharmacology of cancer. Science 1997;275:343–9
2. Luna A, Elloumi F, Varma S, Wang Y, Rajapakse VN, Aladjem MI, et al. CellMiner Cross-Database (CellMinerCDB) version 1.2: Exploration of patient-derived cancer cell line pharmacogenomics. Nucleic Acids Res 2021;49:D1083–D93
3. Rajapakse VN, Luna A, Yamade M, Loman L, Varma S, Sunshine M, et al. CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines. iScience 2018;10:247–64
4. 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 2012:13
5. Reinhold WC, Sunshine M, Varma S, Doroshow JH, Pommier Y. Using CellMiner 1.6 for Systems Pharmacology and Genomic Analysis of the NCI-60. Clinical cancer research : an official journal of the American Association for Cancer Research 2015;21:3841–52
6. Reinhold WC, Thomas A, Pommier Y. DNA-Targeted Precision Medicine; Have we Been Caught Sleeping? Trends in Cancer 2017;3:2–6
7. 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
8. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet 2000;24:236–44

9. Tlemsani C, Takahashi N, Pongor L, Rajapakse VN, Tyagi M, Wen X, et al. Whole-exome sequencing reveals germline-mutated small cell lung cancer subtype with favorable response to DNA repair-targeted therapies. Sci Transl Med 2021;13

10. Pongor LS, Tlemsani C, Elloumi F, Arakawa Y, Jo U, Gross JM, et al. Integrative epigenomic analyses of small cell lung cancer cells demonstrates the clinical translational relevance of gene body methylation. iScience 2022;25

11. Allison M NCATS launches drug repurposing program. Nat Biotechnol 2012;30:571–2

12. Huang R, Zhu H, Shinn P, Ngan D, Ye L, Thakur A, et al. The NCATS Pharmaceutical Collection: a 10-year update. Drug Discov Today 2019;24:2341–9

13. Mathews Griner LA, Guha R, Shinn P, Young RM, Keller JM, Liu D, et al. High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell-like diffuse large B-cell lymphoma cells. Proc Natl Acad Sci U S A 2014;111:2349–54

14. Heske CM, Davis MI, Baumgart JT, Wilson K, Gormally MV, Chen L, et al. Matrix Screen Identifies Synergistic Combination of PARP Inhibitors and Nicotinamide Phosphoribosyltransferase (NAMPT) Inhibitors in Ewing Sarcoma. Clinical cancer research : an official journal of the American Association for Cancer Research 2017;23:7301–11

15. Ju W, Zhang M, Wilson KM, Petrus MN, Bamford RN, Zhang X, et al. Augmented efficacy of brentuximab vedotin combined with ruxolitinib and/or Navitoclax in a murine model of human Hodgkin’s lymphoma. Proc Natl Acad Sci U S A 2016;113:1624–9

16. Lin GL, Wilson KM, Ceribelli M, Stanton BZ, Woo PJ, Kreimer S, et al. Therapeutic strategies for diffuse midline glioma from high-throughput combination drug screening. Sci Transl Med 2019;11

17. Wilson KM, Mathews-Griner LA, Williamson T, Guha R, Chen L, Shinn P, et al. Mutation Profiles in Glioblastoma 3D Oncospheres Modulate Drug Efficacy. SLAS Technol 2019;24:28–40

18. Holbeck SL, Camalier R, Crowell JA, Govindharajulu JP, Hollingshead M, Anderson LW, et al. The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. Cancer Res 2017;77:3564–76

19. 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

20. Reinhold WC, Varma S, Sunshine M, Rajapakse V, Luna A, Kohn KW, et al. The NCI-60 Methylome and its Integration into CellMiner. Cancer Res 2017;77

21. Reinhold WC, Varma S, Sunshine M, Elloumi F, Ofori-Atta K, Lee S, et al. RNA Sequencing of the NCI-60: Integration into CellMiner and CellMiner CDB. Cancer Res 2019;79:3514–24

22. Liu H, D’Andrade Petula, Fulmer-Smentek Stephanie, Lorenzi Philip, Kohn Kurt W., Weinstein John N., et al. mRNA and microRNA expression profiles integrated with drug sensitivities of the NCI-60 human cancer cell lines MCT 2010;9(5):1080–1091.

23. Nishizuka S, Chen ST, Gwadry FG, Alexander J, Major SM, Scherf U, et al. Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. Cancer Res 2003;63:5243–50

24. Guo T, Luna A, Rajapakse VN, Koh CC, Wu Z, Liu W, et al. Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines. iScience 2019;21:664–80

25. Gopi LK, Kidder BL. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. Nat Commun 2021;12:1419

26. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7

27. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 2019;569:503–8

28. Heimerdinger P, Rosin A, Danzer MA, Gerdes T. A Novel Method for Humidity-Dependent Through-Plane Impedance Measurement for Proton Conducting Polymer Membranes. Membranes (Basel) 2019;9

29. Thomas A, Takahashi N, Rajapakse VN, Zhang X, Sun Y, Ceribelli M, et al. Therapeutic targeting of ATR yields durable regressions in small cell lung cancers with high replication stress. Cancer Cell 2021;39:566–79 e7

30. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47:D930–D40

31. Siramshetty VB, Grishagin I, Nguyen Eth T, Peryea T, Skovpen Y, Stroganov O, et al. NCATS Inxight Drugs: a comprehensive and curated portal for translational research. Nucleic Acids Res 2022;50:D1307–D16

32. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–D82

33. Bairoch A The Cellosaurus, a Cell-Line Knowledge Resource. J Biomol Tech 2018;29:25–38

34. Zhao C, Jiang T, Ju JH, Zhang S, Tao J, Fu YL, J., et al. TruSight Oncology 500: Enabling Comprehensive Genomic Profiling and Biomarker Reporting with Targeted Sequencing. bioRxiv 2020

35. Luna A, Rajapakse VN, Sousa FG, Gao J, Schultz N, Varma S, et al. rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. Bioinformatics 2016;32:1272–4

36. Thomas A, Pommier Y. Targeting Topoisomerase I in the Era of Precision Medicine. Clinical cancer research : an official journal of the American Association for Cancer Research 2019;25:6581–9

37. Niepel M, Hafner M, Mills CE, Subramanian K, Williams EH, Chung M, et al. A Multi-center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. Cell Syst 2019;9:35–48 e5

38. 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

39. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol 2016;12:109–16

40. Jo U, Murai Y, Takebe N, Thomas A, Pommier Y. Precision Oncology with Drugs Targeting the Replication Stress, ATR, and Schlafen 11. Cancers (Basel) 2021;13

Significance

CellMinerCDB: NCATS provides activity information for 2,675 drugs in 183 cancer cell lines and analysis tools to facilitate pharmacogenomic research and identify determinants of response.

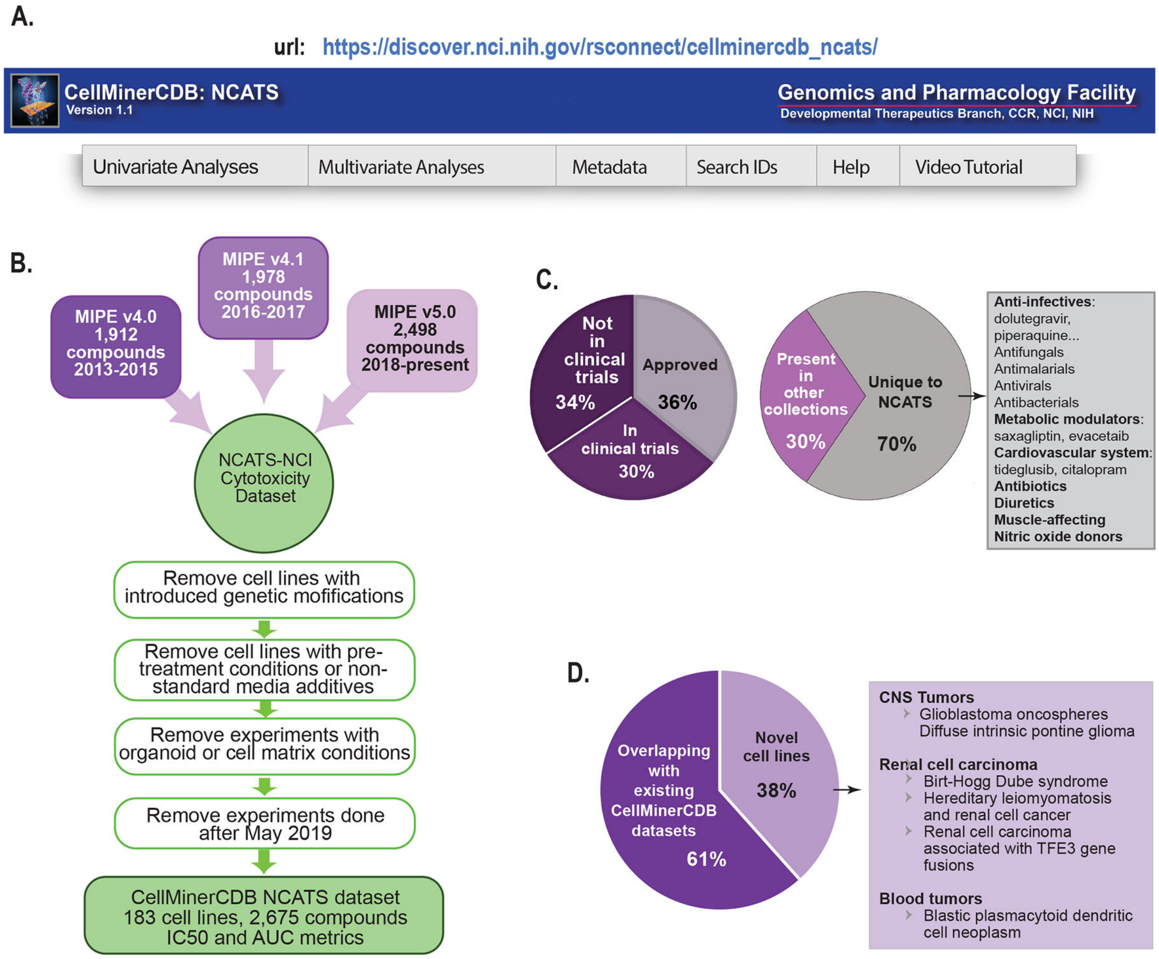
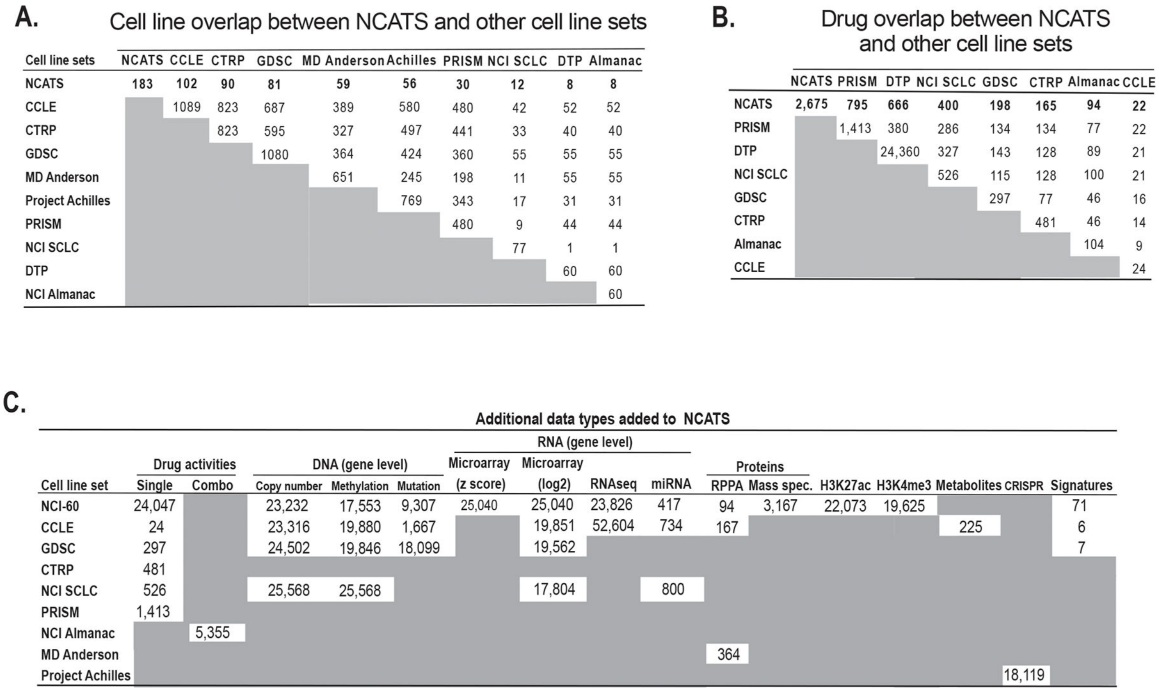

A.
url: https://discover.nci.nih.gov/rsconnect/cellminercdb_ncats/
CellMinerCDB: NCATS
Version 1.1
Genomics and Pharmacology Facility
Developmental Therapeutics Branch, CCR, NCI, NIH
Univariate Analyses
Multivariate Analyses
Metadata
Search IDs
Help
Video Tutorial
B.
MIPE v4.0 1,912 compounds 2013-2015
MIPE v4.1 1,978 compounds 2016-2017
MIPE v5.0 2,498 compounds 2018-present
NCATS-NCI Cytotoxicity Dataset
Remove cell lines with introduced genetic mofifications
Remove cell lines with pre-treatment conditions or non-standard media additives
Remove experiments with organoid or cell matrix conditions
Remove experiments done after May 2019
CellMinerCDB NCATS dataset 183 cell lines, 2,675 compounds IC50 and AUC metrics
C.
Not in clinical trials 34%
Approved 36%
In clinical trials 30%
Present in other collections 30%
Unique to NCATS 70%
Anti-infectives: dolutegravir, piperaquine...
Antifungals
Antimalarials
Antivirals
Antibacterials
Metabolic modulators: saxagliptin, evacetaib
Cardiovascular system: tideglusib, citalopram
Antibiotics
Diuretics
Muscle-affecting
Nitric oxide donors
D.
Overlapping with existing CellMinerCDB datasets 61%
Novel cell lines 38%
CNS Tumors
Glioblastoma oncospheres
Diffuse intrinsic pontine glioma
Renal cell carcinoma
Birt-Hogg Dube syndrome
Hereditary leiomyomatosis and renal cell cancer
Renal cell carcinoma associated with TFE3 gene fusions
Blood tumors
Blastic plasmacytoid dendritic cell neoplasm

Figure 1: The CellMinerCDB: NCATS web application, NCATS dataset, drugs and cell lines.
A. Url, banner and tabs for the CellMinerCDB: NCATS web-application. **B.** Schematic of the creation of the NCATS-NCI cytotoxicity dataset. Multiple versions of the MIPE library were combined into a single dataset to make the "NCATS-NCI" "Cytotoxicity Dataset". This dataset was trimmed down to remove cell lines with introduced genetic modifications, pre-treatment conditions, non-standard media additives and data not meeting the sharing embargo date of 18 months. **C.** Pie chart on left showing the clinical status of the 2,675 CellMinerCDB: NCATS compounds: 36% are FDA-approved, 30% have entered clinical trials and 34% are experimental. Pie chart on right showing the compounds overlapping between CellMinerCDB: NCATS and all other datasets included in CellMinerCDB 1.4. Thirty percent (837) of NCATS compounds overlap with at least one of the other CellMinerCDB datasets and 70% (1,860) do not. Of those compounds found only in the NCATS datasets, there are multiple non-cancer drug types included (see box). **D.** Pie chart showing the cell line overlaps between CellMinerCDB: NCATS, and all other datasets included in CellMinerCDB 1.4.



A. Cell line overlap between NCATS and other cell line sets

Cell line sets	NCATS	CCLE	CTRP	GDSC	MD Anderson	Achilles	PRISM	NCI SCLC	DTP	Almanac
NCATS	**183**	**102**	**90**	**81**	**59**	**56**	**30**	**12**	**8**	**8**
CCLE		1089	823	687	389	580	480	42	52	52
CTRP			823	595	327	497	441	33	40	40
GDSC				1080	364	424	360	55	55	55
MD Anderson					651	245	198	11	55	55
Project Achilles						769	343	17	31	31
PRISM							480	9	44	44
NCI SCLC								77	1	1
DTP									60	60
NCI Almanac										60

B. Drug overlap between NCATS and other cell line sets

	NCATS	PRISM	DTP	NCI SCLC	GDSC	CTRP	Almanac	CCLE
NCATS	**2,675**	**795**	**666**	**400**	**198**	**165**	**94**	**22**
PRISM		1,413	380	286	134	134	77	22
DTP			24,360	327	143	128	89	21
NCI SCLC				526	115	128	100	21
GDSC					297	77	46	16
CTRP						481	46	14
Almanac							104	9
CCLE								24

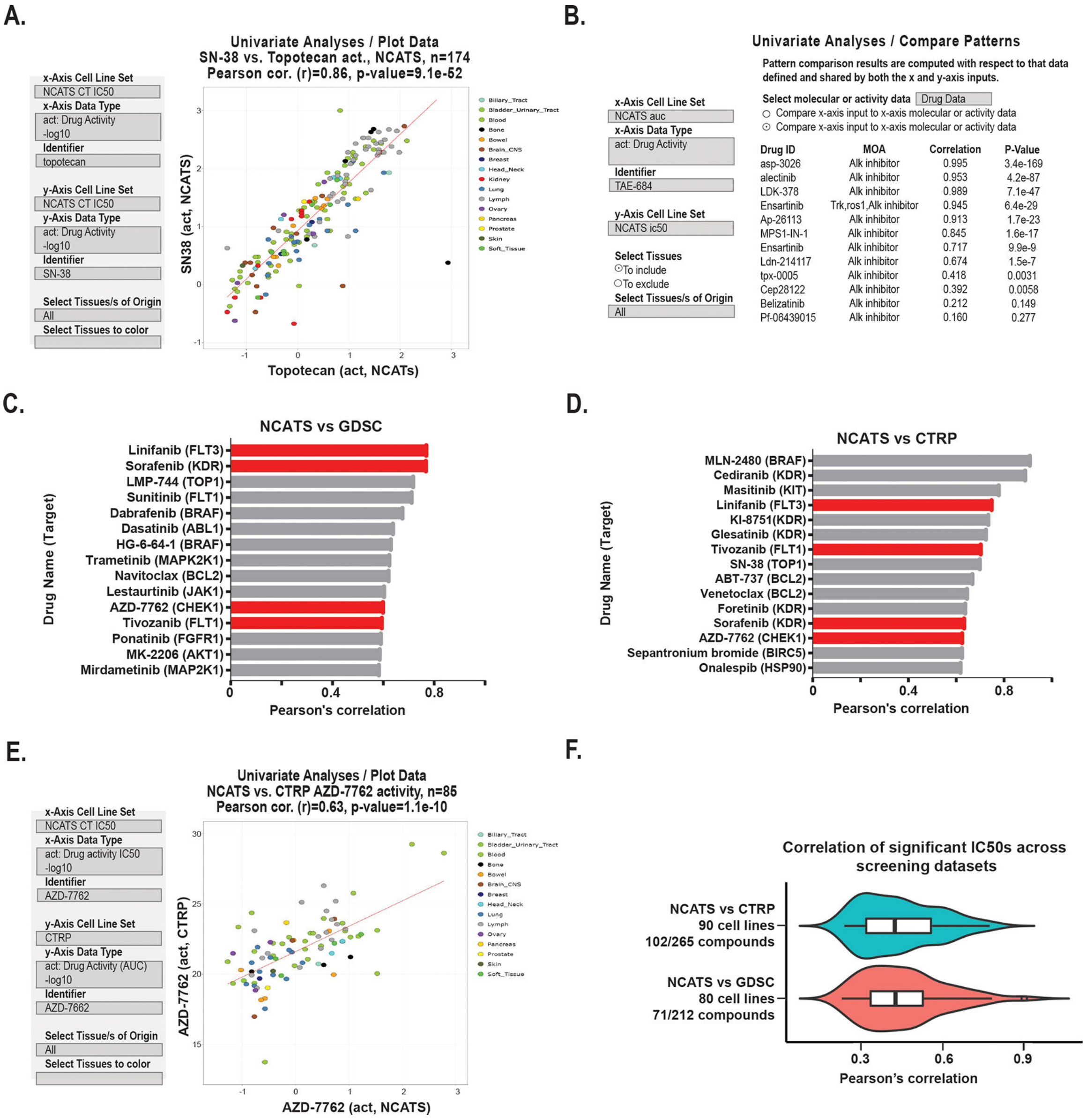
C.

Additional data types added to NCATS

						RNA (gene level)										
	Drug activities		DNA (gene level)			Microarray	Microarray			Proteins						
Cell line set	Single	Combo	Copy number	Methylation	Mutation	(z score)	(log2)	RNAseq	miRNA	RPPA	Mass spec.	H3K27ac	H3K4me3	Metabolites	CRISPR	Signatures
NCI-60	24,047		23,232	17,553	9,307	25,040	25,040	23,826	417	94	3,167	22,073	19,625			71
CCLE	24		23,316	19,880	1,667		19,851	52,604	734	167				225		6
GDSC	297		24,502	19,846	18,099		19,562									7
CTRP	481															
NCI SCLC	526		25,568	25,568			17,804		800							
PRISM	1,413															
NCI Almanac		5,355														
MD Anderson										364						
Project Achilles															18,119	

Figure 2: Cell line and drug overlap, and data types in CellMinerCDB-NCATS.

A. Cell lines overlap between NCATS and the 9 other cell line datasets. CCLE is the Cancer Cell Line Encyclopedia, CTRP the Cancer Therapeutics Response Portal, GDSC the Genomics of Drug Sensitivity in Cancer, Project Achilles is from the Cancer Dependency Map Portal (DepMap), PRISM is the Profiling Relative Inhibition Simultaneously in Mixtures from Broad-MIT, NCI SCLC is the National Cancer Institute small cell lung cancer, DTP is the Developmental Therapeutics Program and NCI Almanac is the NCI60-DTP Almanac. **B.** Drug overlap between NCATS and the 7 other cell line datasets. Number of drugs is as based on the comparison of NCATS area under the curve (AUC) overlap and the 7 other cell line sets. The MD Anderson and DepMap Achilles cell line datasets are not included as they have no drug activities. The NCI Almanac has two-drug activities measurements. The drugs with data for inhibitory concentration 50% (IC50's) are slightly less in number. For acronym definitions see panel A. **C.** Available data in CellMinerCDB: NCATS. For the "Drug activities" columns, the "Single" numbers are compounds or drugs. The "Combo" drugs are 2-drug combinations for 105 FDA-approved drugs. For the DNA, RNA, and CRISPR columns, the numbers are genes with information for that cell line set. For the "Protein" columns, the numbers are epitopes for the RPPA (reverse phase protein arrays) and protein fragments for the mass spectrometry. For the "Metabolite" column, the numbers are metabolites. For the "Signatures" column, the number is signatures of various types. CTRP DNA copy number and mutation, microarray log2, and signatures data is identical to that in CCLE, and so is not included here.



A.

x-Axis Cell Line Set: NCATS CT IC50
x-Axis Data Type: act: Drug Activity -log10
Identifier: topotecan

y-Axis Cell Line Set: NCATS CT IC50
y-Axis Data Type: act: Drug Activity -log10
Identifier: SN-38

Select Tissues/s of Origin: All
Select Tissues to color:

Univariate Analyses / Plot Data
SN-38 vs. Topotecan act., NCATS, n=174
Pearson cor. (r)=0.86, p-value=9.1e-52

B.

x-Axis Cell Line Set: NCATS auc
x-Axis Data Type: act: Drug Activity
Identifier: TAE-684

y-Axis Cell Line Set: NCATS ic50

Select Tissues
⊙To include
○To exclude
Select Tissues/s of Origin: All

Univariate Analyses / Compare Patterns

Pattern comparison results are computed with respect to that data defined and shared by both the x and y-axis inputs.

Select molecular or activity data: Drug Data
○ Compare x-axis input to x-axis molecular or activity data
⊙ Compare x-axis input to x-axis molecular or activity data

Drug ID	MOA	Correlation	P-Value
asp-3026	Alk inhibitor	0.995	3.4e-169
alectinib	Alk inhibitor	0.953	4.2e-87
LDK-378	Alk inhibitor	0.989	7.1e-47
Ensartinib	Trk,ros1,Alk inhibitor	0.945	6.4e-29
Ap-26113	Alk inhibitor	0.913	1.7e-23
MPS1-IN-1	Alk inhibitor	0.845	1.6e-17
Ensartinib	Alk inhibitor	0.717	9.9e-9
Ldn-214117	Alk inhibitor	0.674	1.5e-7
tpx-0005	Alk inhibitor	0.418	0.0031
Cep28122	Alk inhibitor	0.392	0.0058
Belizatinib	Alk inhibitor	0.212	0.149
Pf-06439015	Alk inhibitor	0.160	0.277

E.

x-Axis Cell Line Set: NCATS CT IC50
x-Axis Data Type: act: Drug activity IC50 -log10
Identifier: AZD-7762

y-Axis Cell Line Set: CTRP
y-Axis Data Type: act: Drug Activity (AUC) -log10
Identifier: AZD-7662

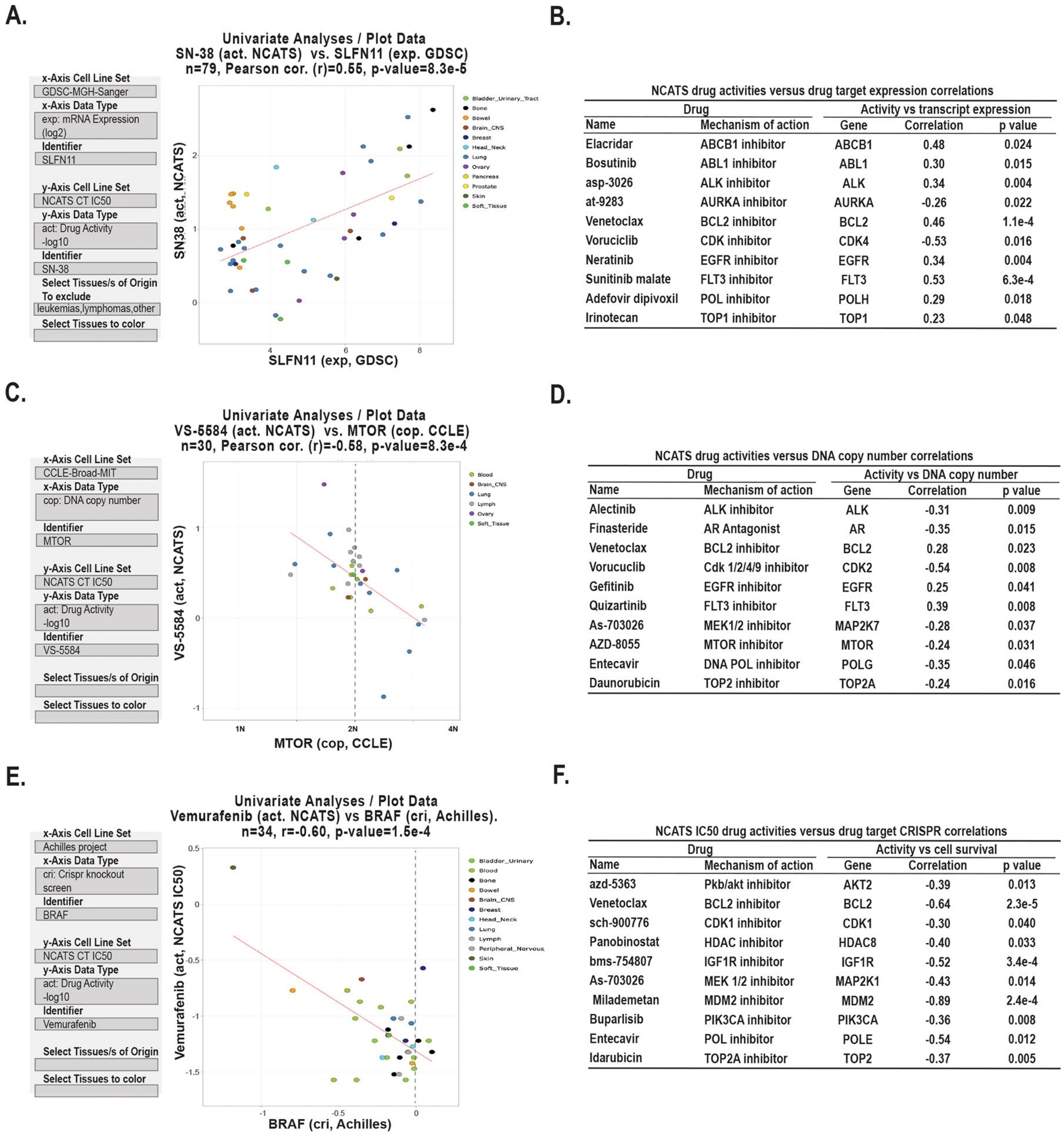
Select Tissue/s of Origin: All
Select Tissues to color:

Univariate Analyses / Plot Data
NCATS vs. CTRP AZD-7762 activity, n=85
Pearson cor. (r)=0.63, p-value=1.1e-10

Figure 3: Comparisons of drugs in CellMinerCDB: NCATS.

A. Scatter plot of the activities of topotecan (x-axis) versus SN38 (y-axis), both measured by NCATS. The plot is a screenshot from CellMinerCDB-NCATS (see Figure 1A: Univariate Analyses). **B.** Comparison of the ALK inhibitor TAE-684 with the other ALK inhibitors tested by NCATS. The results were generated using CellMinerCDB-NCATS (Univariate Analyses \ Compare Patterns tab selections) including a filter to output only “ALK inhibitor” in the MOA (mechanism of action) column and ordered by p-value. **C.** Bar graph showing the top 15 compounds with the highest positive correlation for IC50 value comparisons between NCATS and GDSC. Red bars highlight the compounds highly correlated between NCATS and CTRP (panel D): linifanib, sorafenib, AZD-7762 and tivozanib. The primary target of each compound is shown in parenthesis. **D.** Bar graph showing the top 15 compounds with the highest positive correlation for IC50 values between

NCATS and CTRP. Red bars highlight the compounds highly correlated between NCATS and GDSC (panel C). The primary target of each compound is shown in parenthesis. **E.** A scatter plot of AZD-7762 activity as measured by NCATS (x-axis) vs. CTRP (y-axis). The plot is a screenshot generated using the Univariate Analyses \ Plot Data tab selections. For the scatter plots A, B and E, individual dots are cell lines with color coding by tissue of origin. **F.** Violin plot showing all compounds with IC50 with positive correlations and with p-values < 0.05 between either between NCATS and CTRP or NCATS and GDSC. All compounds shown had a minimum of 16 cell lines overlap between datasets. The box plot overlay shows a median correlation of 0.4. All correlations presented are Pearson's.



x-Axis Cell Line Set
GDSC-MGH-Sanger
x-Axis Data Type
exp: mRNA Expression (log2)
Identifier
SLFN11

y-Axis Cell Line Set
NCATS CT IC50
y-Axis Data Type
act: Drug Activity -log10
Identifier
SN-38
Select Tissues/s of Origin To exclude
leukemias,lymphomas,other
Select Tissues to color

NCATS drug activities versus drug target expression correlations

Drug		Activity vs transcript expression		
Name	Mechanism of action	Gene	Correlation	p value
Elacridar	ABCB1 inhibitor	ABCB1	0.48	0.024
Bosutinib	ABL1 inhibitor	ABL1	0.30	0.015
asp-3026	ALK inhibitor	ALK	0.34	0.004
at-9283	AURKA inhibitor	AURKA	-0.26	0.022
Venetoclax	BCL2 inhibitor	BCL2	0.46	1.1e-4
Voruciclib	CDK inhibitor	CDK4	-0.53	0.016
Neratinib	EGFR inhibitor	EGFR	0.34	0.004
Sunitinib malate	FLT3 inhibitor	FLT3	0.53	6.3e-4
Adefovir dipivoxil	POL inhibitor	POLH	0.29	0.018
Irinotecan	TOP1 inhibitor	TOP1	0.23	0.048

C.
Univariate Analyses / Plot Data
VS-5584 (act. NCATS) vs. MTOR (cop. CCLE)
n=30, Pearson cor. (r)=-0.58, p-value=8.3e-4
VS-5584 (act, NCATS)
MTOR (cop, CCLE)
-1
0
1
1N
2N
4N
Blood
Brain_CNS
Lung
Lymph
Ovary
Soft_Tissue

x-Axis Cell Line Set
CCLE-Broad-MIT
x-Axis Data Type
cop: DNA copy number
Identifier
MTOR

y-Axis Cell Line Set
NCATS CT IC50
y-Axis Data Type
act: Drug Activity -log10
Identifier
VS-5584

Select Tissues/s of Origin
Select Tissues to color

NCATS drug activities versus DNA copy number correlations

Drug		Activity vs DNA copy number		
Name	Mechanism of action	Gene	Correlation	p value
Alectinib	ALK inhibitor	ALK	-0.31	0.009
Finasteride	AR Antagonist	AR	-0.35	0.015
Venetoclax	BCL2 inhibitor	BCL2	0.28	0.023
Vorucuclib	Cdk 1/2/4/9 inhibitor	CDK2	-0.54	0.008
Gefitinib	EGFR inhibitor	EGFR	0.25	0.041
Quizartinib	FLT3 inhibitor	FLT3	0.39	0.008
As-703026	MEK1/2 inhibitor	MAP2K7	-0.28	0.037
AZD-8055	MTOR inhibitor	MTOR	-0.24	0.031
Entecavir	DNA POL inhibitor	POLG	-0.35	0.046
Daunorubicin	TOP2 inhibitor	TOP2A	-0.24	0.016

E.
Univariate Analyses / Plot Data
Vemurafenib (act. NCATS) vs BRAF (cri, Achilles).
n=34, r=-0.60, p-value=1.5e-4
Vemurafenib (act, NCATS IC50)
BRAF (cri, Achilles)
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
Bladder_Urinary
Blood
Bone
Bowel
Brain_CNS
Breast
Head_Neck
Lung
Lymph
Peripheral_Nervous
Skin
Soft_Tissue

x-Axis Cell Line Set
Achilles project
x-Axis Data Type
cri: Crispr knockout screen
Identifier
BRAF

y-Axis Cell Line Set
NCATS CT IC50
y-Axis Data Type
act: Drug Activity -log10
Identifier
Vemurafenib

Select Tissues/s of Origin
Select Tissues to color

NCATS IC50 drug activities versus drug target CRISPR correlations

Drug		Activity vs cell survival		
Name	Mechanism of action	Gene	Correlation	p value
azd-5363	Pkb/akt inhibitor	AKT2	-0.39	0.013
Venetoclax	BCL2 inhibitor	BCL2	-0.64	2.3e-5
sch-900776	CDK1 inhibitor	CDK1	-0.30	0.040
Panobinostat	HDAC inhibitor	HDAC8	-0.40	0.033
bms-754807	IGF1R inhibitor	IGF1R	-0.52	3.4e-4
As-703026	MEK 1/2 inhibitor	MAP2K1	-0.43	0.014
Milademetan	MDM2 inhibitor	MDM2	-0.89	2.4e-4
Buparlisib	PIK3CA inhibitor	PIK3CA	-0.36	0.008
Entecavir	POL inhibitor	POLE	-0.54	0.012
Idarubicin	TOP2A inhibitor	TOP2	-0.37	0.005

Figure 4: NCATS:CDB Univariate comparisons of drug activities to transcript, DNA copy number and CRISPR signatures.

A. Scatter plot of SLFN11 transcript expression from GDSC (x-axis) versus SN-38 activity measured by NCATS (y-axis). The plot is a snapshot from CellMinerCDB-NCATS (Univariate Analyses). **B.** Additional examples of significantly correlated and biologically linked NCATS IC50 drug activities versus GDSC transcript expression levels. All gene examples are targets for the corresponding drugs. **C.** Scatter plot of MTOR DNA copy number as measured by CCLE (x-axis) versus VS-5584 activity as measured by NCATS (y-axis). The plot is a screenshot from CellMinerCDB-NCATS (Univariate Analyses \ Plot Data tab selections) with the specific inputs used detailed in the boxes to the left. The vertical line has been added at 0 intensity or 2N DNA copy number. The units for the x axis have been converted from intensity to ploidy (Copy Number = 2 × $2^{\text{intensity}}$) for biological clarity. **D.** Additional examples of significantly correlated and biologically linked NCATS IC50 drug activities versus CCLE DNA copy number from plots generated as in D. Genes are targets of the corresponding drugs. **E.** Scatter plot of BRAF CRISPR knockdown cell survival from the Achilles Project (x-axis) versus vemurafenib activity as measured by NCATS (y-axis). The plot is a screenshot from CellMinerCDB (Univariate Analyses \ Plot Data tab) with the specific inputs used detailed in the input boxes to the left. The vertical line has been added at 0 to indicate that the cell lines to the left of line have decreased survival following knocking down BRAF. **F.** Additional examples of significant correlations of drug activities versus CRISPR knockdown of the target genes. The CRISPR knockdown cell survival data are from the Achilles Project. All correlations presented in the figure are Pearson's. For all scatter plots, dots are cell lines with color coding by tissue of origin indicated to the right.

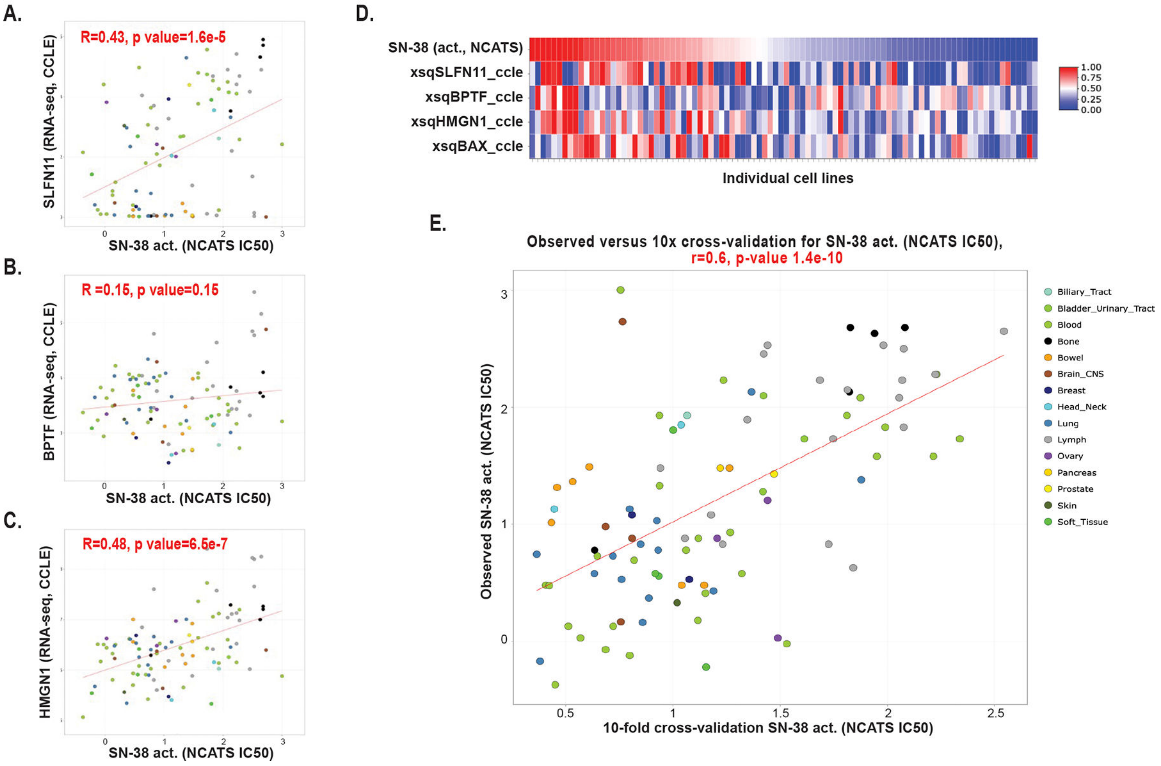
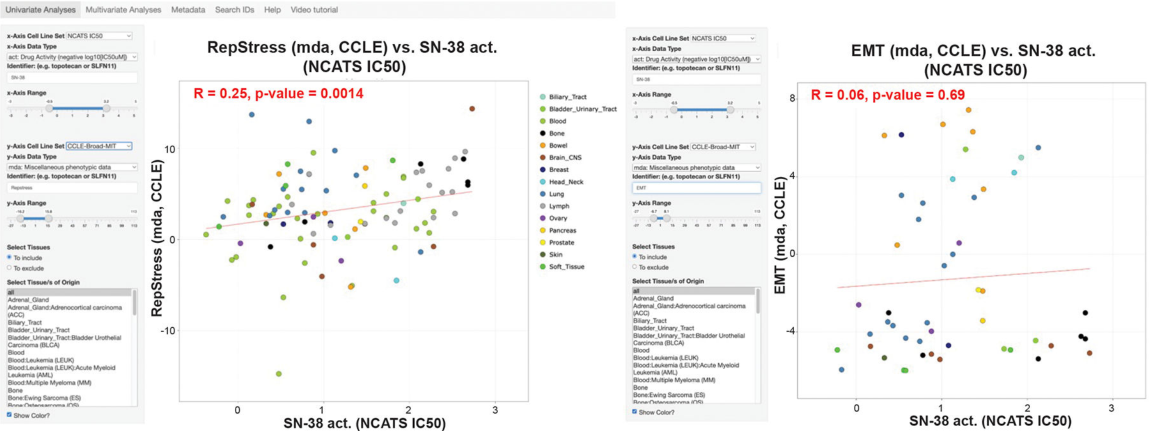


Figure 5. Multivariate analysis of SN-38 activity in NCATS using the expression of SLFN11, BPTF, HMGN1 and BAX in the overlapping cell lines in CCLE are better predictors of SN-38 activity than any of the 4 genes taken individually.

A. Predictive value of SLFN11 expression. **B**. Predictive value of BPTF (encoding a protein regulating chromatin remodeling as a regulator of ATP hydrolysis of the NURF complex). **C**. Predictive value of HMGN1 (encoding High Mobility Group Nucleosome-binding domain-containing protein 1) associated with active transcription. **D**. Cluster image map of the multivariate analysis of SN-38 activity predicted by the expression of 4 genes together. See Supplementary Fig. 1 for BAX univariate data. E. Scatter plot of the observed versus 10-fold cross-validation for SN-38 using the same predictor genes as in D.



x-Axis Cell Line Set: NCATS IC50
x-Axis Data Type: act: Drug Activity (negative log10[IC50uM])
Identifier: (e.g. topotecan or SLFN11): SN-38
x-Axis Range: -0.5 – 3.2

y-Axis Cell Line Set: CCLE-Broad-MIT
y-Axis Data Type: mda: Miscellaneous phenotypic data
Identifier: (e.g. topotecan or SLFN11): Repstress
y-Axis Range: -16.2 – 15.8

Select Tissues
● To include
○ To exclude

Select Tissue/s of Origin
all
Adrenal_Gland
Adrenal_Gland:Adrenocortical carcinoma (ACC)
Biliary_Tract
Bladder_Urinary_Tract
Bladder_Urinary_Tract:Bladder Urothelial Carcinoma (BLCA)
Blood
Blood:Leukemia (LEUK)
Blood:Leukemia (LEUK):Acute Myeloid Leukemia (AML)
Blood:Multiple Myeloma (MM)
Bone
Bone:Ewing Sarcoma (ES)
Bone:Osteosarcoma (OS)

☑ Show Color?

x-Axis Cell Line Set: NCATS IC50
x-Axis Data Type: act: Drug Activity (negative log10[IC50uM])
Identifier: (e.g. topotecan or SLFN11): SN-38
x-Axis Range: -0.5 – 3.2

y-Axis Cell Line Set: CCLE-Broad-MIT
y-Axis Data Type: mda: Miscellaneous phenotypic data
Identifier: (e.g. topotecan or SLFN11): EMT
y-Axis Range: -6.7 – 8.1

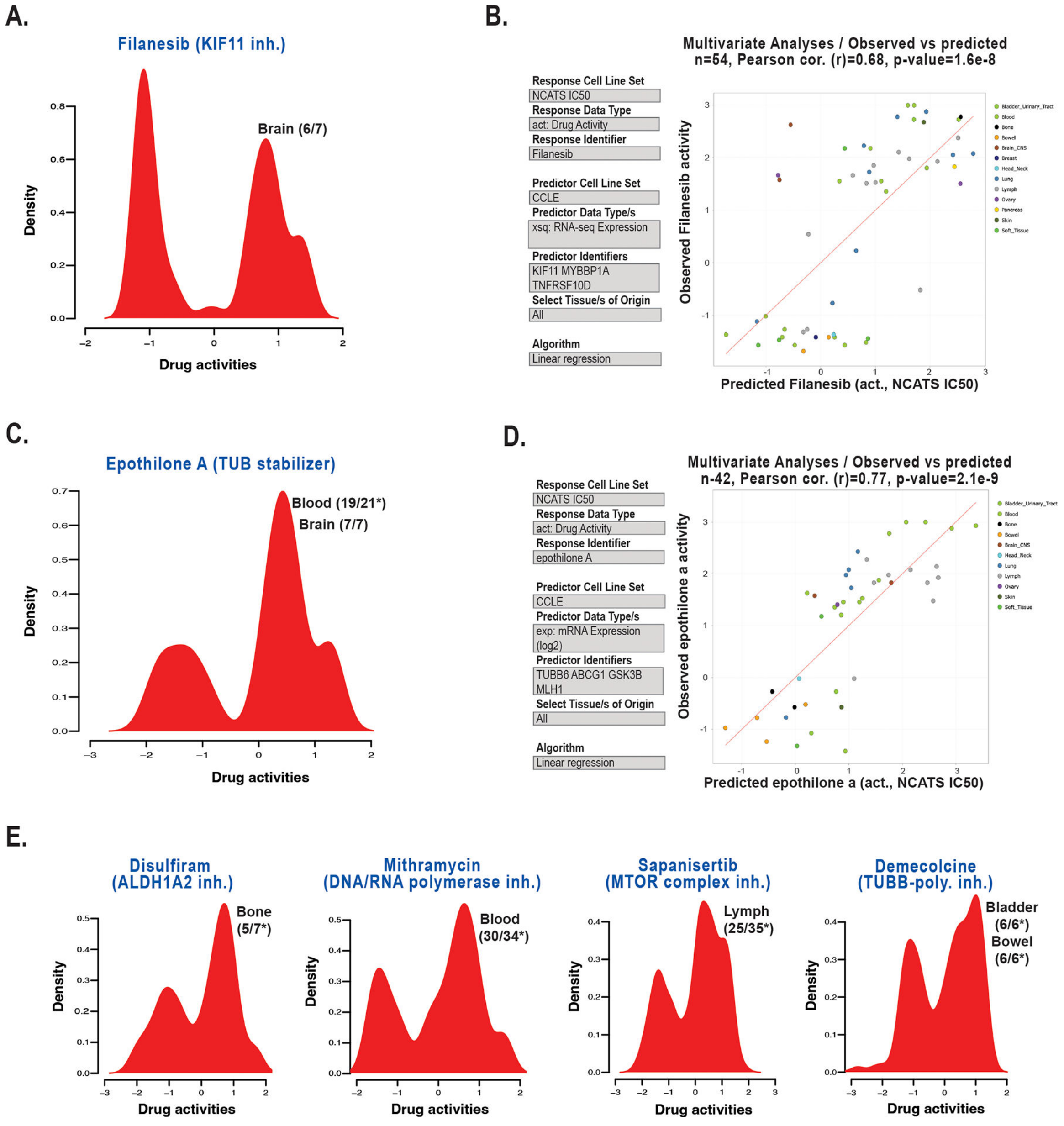
Select Tissues
● To include
○ To exclude

Select Tissue/s of Origin
all
Adrenal_Gland
Adrenal_Gland:Adrenocortical carcinoma (ACC)
Biliary_Tract
Bladder_Urinary_Tract
Bladder_Urinary_Tract:Bladder Urothelial Carcinoma (BLCA)
Blood
Blood:Leukemia (LEUK)
Blood:Leukemia (LEUK):Acute Myeloid Leukemia (AML)
Blood:Multiple Myeloma (MM)
Bone
Bone:Ewing Sarcoma (ES)
Bone:Osteosarcoma (OS)

☑ Show Color?

Figure 6. Genomic signature analysis identifies replication stress (RepStress) but not EMT (Epithelial-Mesenchymal Transition) as predictor of SN-38 activity in the overlapping cell lines of NCATS and CCLE.

Left and right panels: snapshots of CellMinerCDB: NCATS for RepStress and EMT, respectively.



B.

Response Cell Line Set: NCATS IC50
Response Data Type: act: Drug Activity
Response Identifier: Filanesib
Predictor Cell Line Set: CCLE
Predictor Data Type/s: xsq: RNA-seq Expression
Predictor Identifiers: KIF11 MYBBP1A TNFRSF10D
Select Tissue/s of Origin: All
Algorithm: Linear regression

D.

Response Cell Line Set: NCATS IC50
Response Data Type: act: Drug Activity
Response Identifier: epothilone A
Predictor Cell Line Set: CCLE
Predictor Data Type/s: exp: mRNA Expression (log2)
Predictor Identifiers: TUBB6 ABCG1 GSK3B MLH1
Select Tissue/s of Origin: All
Algorithm: Linear regression

Figure 7: Drug distributions, tissue of origin enrichments and molecular predictors of drug activity.

A. A density plot of filanesib activity (IC50 z-scores from NCATS) (x-axis) versus distribution of the cell lines plotted as density (y-axis). **B.** Multivariate analysis for filanesib activity as the response variable and CCLE transcript expression of three genes as predictor variables. **C.** Density plot of epothilone A activity (x-axis) versus density (y-axis). The brain enrichment p-value is 0.082. **D.** Multivariate analysis for epothilone A activity as the response variable and CCLE transcript expression of four genes as predictor variables. **E.** Density plots for 4 NCATS drugs showing drug activity IC50 z-scores vs. distribution of the cell lines plotted as density (y-axis). For the density plots in **A, C** and **E**, drug activities are z-scores calculated across cell lines for IC50s (x-axis). Enriched tissue of origins are included (if present) with both the number of cell lines present within the peak (first number) and total number of cell lines of that type (second number). The asterisks indicate significant p-values <0.05. All other p-values are less than 0.07. In the scatter plots **B** and **D,** the predicted drug activity is on the x-axis and the observed drug activity is on the y-axis. All correlations presented are Pearson's. Dots are cell lines with color coding by tissue of origin. The plots were created using the CellMinerCDB: NCATS \ Multivariate Analyses \ Plot Data tab selections with the specific inputs used detailed in the input boxes to the left