Published online 16 November 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D1083–D1093***
doi: 10.1093/nar/gkaa968

CellMiner Cross-Database (CellMinerCDB) version 1.2: Exploration of patient-derived cancer cell line pharmacogenomics

Augustin Luna[1,*,†], Fathi Elloumi[2,3,†], Sudhir Varma[2,4], Yanghsin Wang[2,3], Vinodh N. Rajapakse[2], Mirit I. Aladjem[2], Jacques Robert[5], Chris Sander[1], Yves Pommier[2,*] and William C. Reinhold[2,*]

[1]cBio Center, Dana-Farber Cancer Institute and Department of Cell Biology, Harvard Medical School, Boston, MA 02215, USA, [2]Developmental Therapeutics Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20892, USA, [3]General Dynamics Information Technology Inc., Fairfax, VA 22042, USA, [4]HiThru Analytics LLC, Princeton, NJ 08540, USA and [5]Inserm unité 1218, Université de Bordeaux, Bordeaux 33076, France

Received August 14, 2020; Revised September 25, 2020; Editorial Decision October 08, 2020; Accepted October 19, 2020

ABSTRACT

CellMiner Cross-Database (CellMinerCDB, discover.nci.nih.gov/cellminercdb) allows integration and analysis of molecular and pharmacological data within and across cancer cell line datasets from the National Cancer Institute (NCI), Broad Institute, Sanger/MGH and MD Anderson Cancer Center (MDACC). We present CellMinerCDB 1.2 with updates to datasets from NCI-60, Broad Cancer Cell Line Encyclopedia and Sanger/MGH, and the addition of new datasets, including NCI-ALMANAC drug combination, MDACC Cell Line Project proteomic, NCI-SCLC DNA copy number and methylation data, and Broad methylation, genetic dependency and metabolomic datasets. CellMinerCDB (v1.2) includes several improvements over the previously published version: (i) new and updated datasets; (ii) support for pattern comparisons and multivariate analyses across data sources; (iii) updated annotations with drug mechanism of action information and biologically relevant multigene signatures; (iv) analysis speedups via caching; (v) a new dataset download feature; (vi) improved visualization of subsets of multiple tissue types; (vii) breakdown of univariate associations by tissue type; and (viii) enhanced help information. The curation and common annotations (e.g. tissues of origin and identifiers) provided here across pharmacogenomic datasets increase the utility of the individual datasets to address multiple researcher question types, including data reproducibility, biomarker discovery and multivariate analysis of drug activity.

INTRODUCTION

A critical aim of precision medicine is to match drugs with genomic determinants of response. Identifying tumor molecular features that affect response to specific drug treatments is especially challenging because of patient diversity, incomplete knowledge of the multiple molecular determinants of response and tumor heterogeneity. The relative homogeneity and ease of experimentation of patient-derived cell lines are advantageous, making them widely used model systems for establishing and resolving intrinsic drug response mechanisms as well as performing synthetic lethality screens (1). These features motivated the development of cancer cell line pharmacogenomic databases. Following the omics characterization of a 60-cell line panel developed by the National Cancer Institute (NCI), known as the NCI-60 (2–7), several large cancer cell line sets have been developed with pharmacogenomic data, including the Sanger/Massachusetts General Hospital Genomics of Drug Sensitivity in Cancer (GDSC) (8,9), the Broad/Novartis Cancer Cell Line Encyclopedia (CCLE) (10) and the Broad Cancer Therapeutics Response Portal (CTRP) datasets (11); together they provide information on ∼1400 cancer cell lines. Complementarity and overlap in the strengths of the cancer cell line datasets (e.g. number of drugs or total cell lines) provide the potential for valuable comparisons and integrative analyses in which the total value of the data is greater than the sum of the parts. However, data complexity and sources of inconsistency, such as differences in entity naming (e.g. cell lines, drugs) and

[*]To whom correspondence should be addressed. Tel: +1 240 760 7339; Fax: +1 240 541 4475; Email: aluna@jimmy.harvard.edu
Correspondence may also be addressed to William C. Reinhold. Tel: +1 240 760 7339; Fax: +1 240 541 4475; Email: wcr@mail.nih.gov
Correspondence may also be addressed to Yves Pommier. Tel: +1 240 760 7339; Fax: +1 240 541 4475; Email: pommier@nih.gov
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

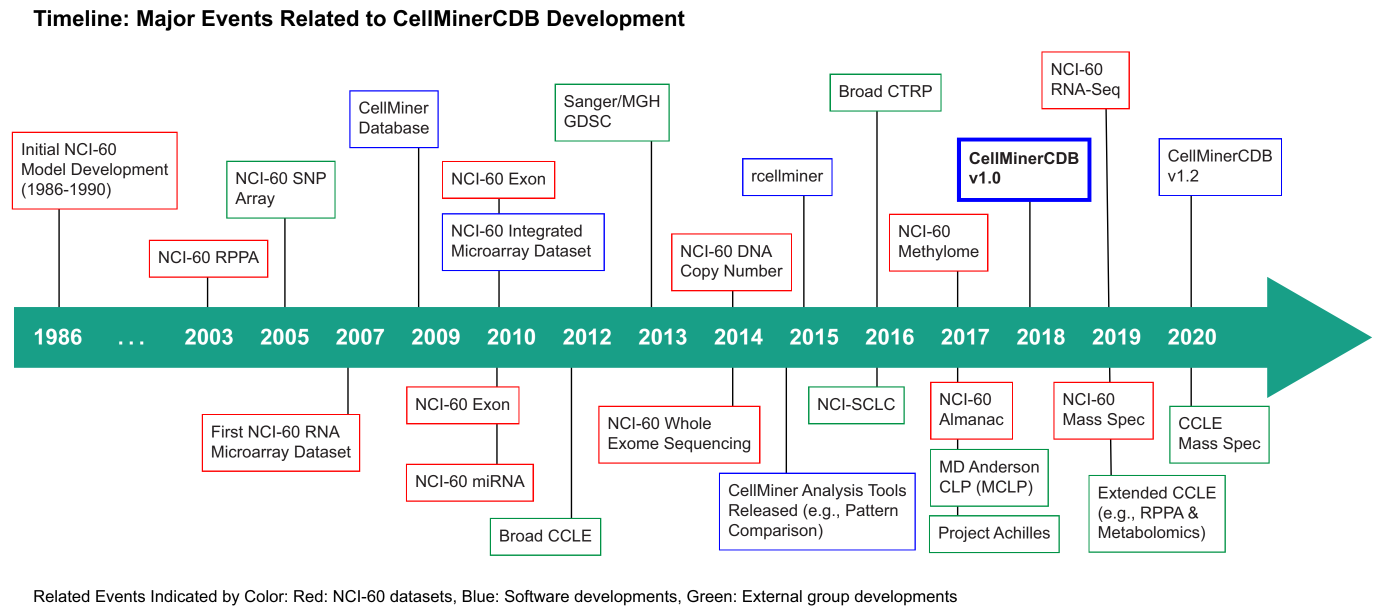

Timeline: Major Events Related to CellMinerCDB Development
1986
. . .
2003
2005
2007
2009
2010
2012
2013
2014
2015
2016
2017
2018
2019
2020
Initial NCI-60 Model Development (1986-1990)
NCI-60 RPPA
NCI-60 SNP Array
CellMiner Database
NCI-60 Exon
NCI-60 Integrated Microarray Dataset
Sanger/MGH GDSC
NCI-60 DNA Copy Number
rcellminer
Broad CTRP
NCI-60 Methylome
CellMinerCDB v1.0
NCI-60 RNA-Seq
CellMinerCDB v1.2
First NCI-60 RNA Microarray Dataset
NCI-60 Exon
NCI-60 miRNA
Broad CCLE
NCI-60 Whole Exome Sequencing
CellMiner Analysis Tools Released (e.g., Pattern Comparison)
NCI-SCLC
NCI-60 Almanac
MD Anderson CLP (MCLP)
Project Achilles
NCI-60 Mass Spec
Extended CCLE (e.g., RPPA & Metabolomics)
CCLE Mass Spec
Related Events Indicated by Color: Red: NCI-60 datasets, Blue: Software developments, Green: External group developments

Figure 1. Timeline of CellMinerCDB development. Related events are indicated by color: NCI-60 datasets (red), software developments (blue) and external group developments (green).

data processing, make working across databases challenging; current, source-specific portals do not provide cross-database analyses.

CellMiner Cross-Database (CellMinerCDB) combines the expertise and approaches developed within the Genomics and Pharmacology Facility (discover.nci.nih.gov), Developmental Therapeutics Branch, Center for Cancer Research, National Institutes of Health (NIH). The group has an established history (Figure 1) in the area of bioinformatics and pharmacogenomics, beginning with studies involving the NCI-60 cell line screen developed by the Developmental Therapeutics Program (12). These studies included pioneering efforts to (i) predict drug mechanism of action (MOA) from activity patterns (13), (ii) introduce the now ubiquitously used cluster image maps (i.e. clustered heatmaps) for comparison of molecular alterations and drug activity patterns (14) and (iii) integrate gene expression and compound activity into a single database (3,4,15). CellMiner (discover.nci.nih.gov/cellminer) was set up to allow direct data access and exploration (4,15), and became the platform for additional profiling technologies: protein expression (16,17), transcript microarrays (18,19), microRNA expression (19), DNA methylation (6,20), DNA whole-exome sequencing (5), DNA copy number (21), H2AX protein and phosphorylation levels (22) and RNA sequencing (23). Signatures of each of these data types were developed for integrative analysis (24). The first use of many assaying techniques has occurred on the NCI-60 and then later replicated on other cell line collections such as the GDSC and CCLE (Figure 1) (25–30).

First released in 2018, CellMinerCDB (discover.nci.nih.gov/cellminercdb) is a database accessible through a web interface that enables integrative analyses within and across cancer cell line pharmacogenomic databases (31) (Figure 1). CellMinerCDB integrates pharmacogenomic datasets both generated by the NCI (2–7) and downloaded from other project sites. Those include the CCLE/CTRP (11,32) and GDSC (both GDSC1 and GDSC2) datasets (8,9). CellMinerCDB differs from related cancer genomics aggregation efforts such as cBioPortal (33), a data portal focused on clinical sample cancer genomics. It focuses on cancer patient-derived human cell line molecular and pharmacological data, and it differs from the NCI Genomic Data Commons (34) and the data portals of the CCLE and GDSC in that the data are standardized to allow cross-database analysis; no raw data files are available. CellMinerCDB most closely resembles PharmacoDB (35), which also aggregates pharmacogenomic datasets. However, CellMinerCDB places an emphasis on web-based cross-database analyses of the available datasets. Supplementary Figure S1 showcases the distinctive value of CellMinerCDB to other pharmacogenomic tools (28,33,35–44) in terms of datasets available and capabilities.

In CellMinerCDB, named entities (e.g. drugs and genes) are transparently matched across sources, allowing cell line molecular features and drug responses to be readily compared using bivariate scatter plots and correlation analyses. Multivariate models of factors affecting drug responses or genomic cell line attributes can also be assessed. Analyses can be restricted to tissues of origin, with cell lines across all sources mapped to a uniform tissue type hierarchy. Gene pathway annotations allow assessment and filtering of analysis results. The analyses that CellMinerCDB makes accessible along with the breadth of available data make CellMinerCDB a unique resource for cancer cell line pharmacogenomic data exploration and hypothesis generation. In several cancer types (e.g. pancreatic, prostate, small cell lung cancer, etc.), recent studies have utilized CellMinerCDB to reveal associations between cancer drug sensitivity and (i) gene expression (45–48), (ii) genomic alterations (46) and (iii) cell line subgroups (49).

Here, we present the updates to the data and software infrastructure available in CellMinerCDB (Supplementary Figure S2; discover.nci.nih.gov/cellminercdb). This includes the addition of 12 new datasets and updates to several others. Within these new datasets, new data types are included (e.g. two-drug combinations, CRISPR genetic dependencies, mass spectrometry proteomics and metabolome data) as well as the addition of new gene signatures broadening the scope of CellMinerCDB. Functionally, CellMinerCDB has been improved with features to (i) speed up analysis, (ii) perform a broader range of cross-database analyses, (iii) make it easier to download corresponding data, (iv) present additional drug annotation information, (v) simplify the analysis of subsets of multiple tissue types and (vi) break down univariate associations by tissue type. We provide the reader with use cases highlighting the research potential and complementarity of the composite data.

CELLMINERCDB DATASETS

Over the last 2 years, CellMinerCDB (discover.nci.nih.gov/cellminercdb) has been expanded and now includes additional and updated datasets with a focus on genomic, proteomic and metabolomic data (collected prior to treatment), integrated with drug responses and CRISPR-based gene dependency data for these cell lines. Figure 2 summarizes CellMinerCDB content as well as the number of overlaps in cell lines and drugs across the different data sources.

Updates to existing datasets

In our newest release of the NCI-60 dataset within CellMinerCDB, over 1000 compound activities, including 27 FDA-approved and 412 clinical trial drugs (more than doubled), have been added, with processing as previously reported (3). We removed (curated) drug experiments that had a limited response range or were inconsistent across replicates. We updated the NCI-60 drug MOA and drug names as well as the drug clinical status and synonyms for all datasets. We also reprocessed the existing GDSC methylation data (9).

Newly included datasets

The new release integrates three new cell line sets with new data: the NCI-ALMANAC data for a recent paired drug activity for 105 FDA-approved drugs (50), the RPPA-based protein data from MDACC CLP, which is also known as MCLP (tcpaportal.org/mclp), and the CRISPR–Cas9 gene dependency map with phenotypic data from Project Achilles (28).

Additional data from current cell line sets have been included: (i) RNA-seq and SWATH proteomic data for the NCI-60 cell line set (17,23); (ii) microRNA, methylation and copy number data for the NCI-SCLC cell lines (51,52); (iii) copy number data for GDSC cell lines (9); and (iv) RPPA-based protein data, metabolome data and bisulfite sequencing promoter methylation data for CCLE cell lines (26,29).

Phenotypic gene signatures

Additionally, we have added phenotypic gene signature scores for all data sources for several key cancer phenotypes, including epithelial–mesenchymal transition (EMT) (53), antigen presentation machinery (APM) (54) and neuroendocrine (NE) signatures (55).

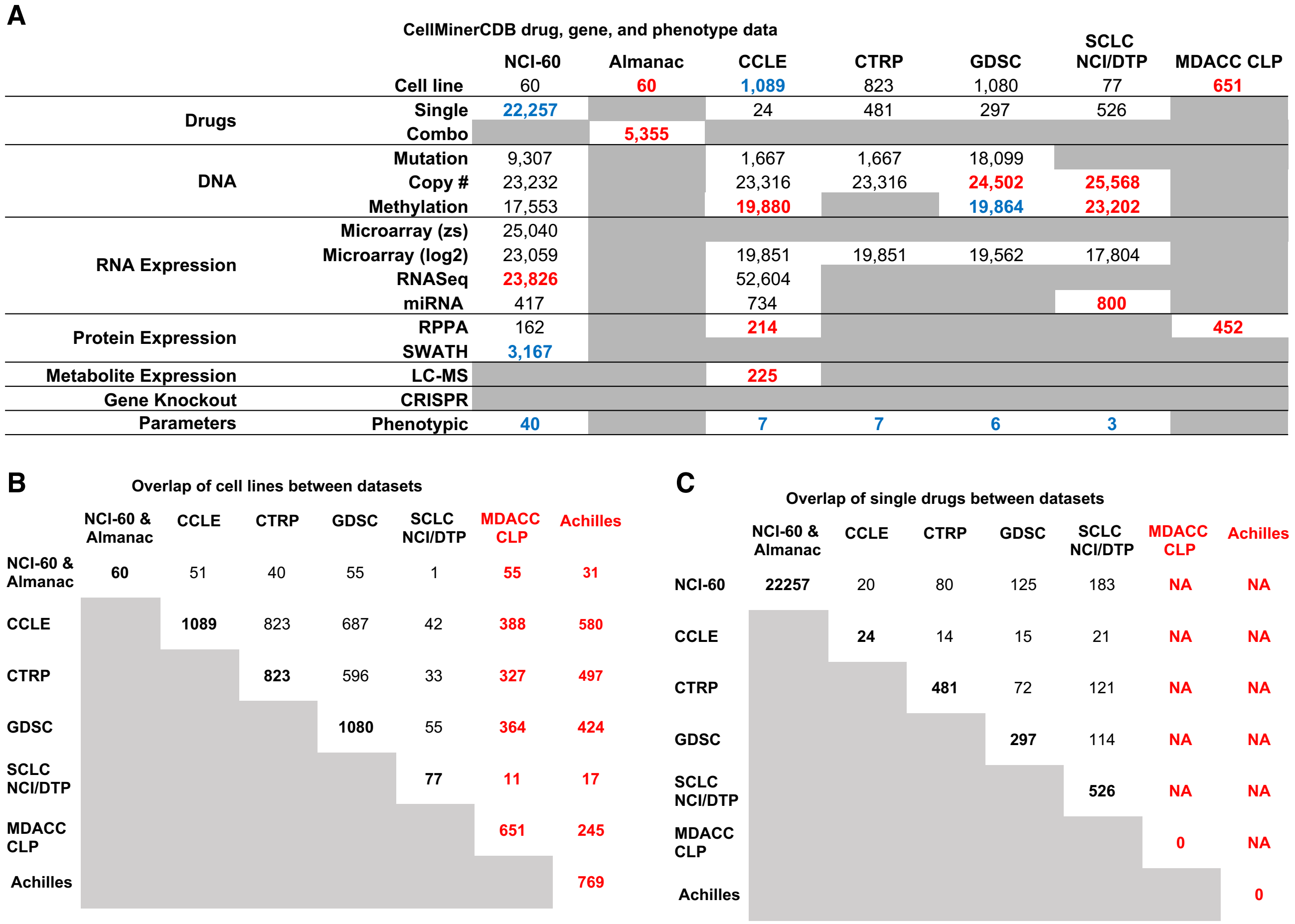
SOFTWARE INFRASTRUCTURE

To create CellMinerCDB, cell line datasets are processed and undergo an additional curation step. This allows us to provide cross-database web-accessible analyses and download features (Figure 3). The different components of CellMinerCDB are described in the following sections.

Data curation

For each data source and cell line, we manually curate a reference table to match cell lines across different sources. Additionally, we assign common terms for the tissue of origin (from the data source) based on the OncoTree (oncotree.mskcc.org; github.com/cBioPortal/oncotree) ontology structure and terms. OncoTree is an attempt to standardize cancer type diagnosis from a clinical perspective. OncoTree was developed and expertly curated by basic researchers and clinicians as a multi-institutional committee including the Memorial Sloan Kettering Cancer Center and Dana-Farber Cancer Institute. We assign OncoTree levels 1–4 (from least to most specific) for each cell line; in cases of ambiguity, resources such as Cellosaurus are consulted during the curation step (56). When available, we record additional cell line annotations such as patient sample information (e.g. gender or age) or curated information such as EMT status (53), SCLC subtypes (57) and triple negative breast cancer status.

For the NCI-60 drug curation, drug information (obtained from the NCI/DTP, dtp.cancer.gov) including National Service Center identifiers is the starting point for drug annotations, including preferred names, aliases, MOAs and FDA approval status. Additional and updated database identifiers are added from multiple sources, including PubChem, the NCI Thesaurus and the scientific literature. We compiled a list of terms that are present in 'chemical' names rather than common or generic names (e.g. 'methyl', 'cyclo', 'pyrrol'). We scored each drug name on the number of such terms present in it (the 'chemical name score') and on the number of characters in the name. In cases where the NCI/DTP provided drug has a chemical name score >3 or the length of the name is >20 characters, we use the PubChem (58) preferred name (if that name has a chemical name score ≤ 3 and length ≤ 20). MOAs and clinical status (i.e. FDA approval status) are manually curated with information from NCI/DTP, FDA alerts (fda.gov) and journal articles as new data become available from external data providers. For other drug datasets (e.g. CCLE), we retrieve drug annotation information from our data providers and external sources (e.g. PubChem) including the identifiers, names, synonyms and MOA when available. Using these identifiers, CellMinerCDB matches drugs across data sources.



A

CellMinerCDB drug, gene, and phenotype data

		NCI-60	Almanac	CCLE	CTRP	GDSC	SCLC NCI/DTP	MDACC CLP
	Cell line	60	60	1,089	823	1,080	77	651
Drugs	Single	22,257		24	481	297	526	
	Combo		5,355					
DNA	Mutation	9,307		1,667	1,667	18,099		
	Copy #	23,232		23,316	23,316	24,502	25,568	
	Methylation	17,553		19,880		19,864	23,202	
RNA Expression	Microarray (zs)	25,040						
	Microarray (log2)	23,059		19,851	19,851	19,562	17,804	
	RNASeq	23,826		52,604				
	miRNA	417		734			800	
Protein Expression	RPPA	162		214				452
	SWATH	3,167						
Metabolite Expression	LC-MS			225				
Gene Knockout	CRISPR							
Parameters	Phenotypic	40		7	7	6	3	

B

Overlap of cell lines between datasets

	NCI-60 & Almanac	CCLE	CTRP	GDSC	SCLC NCI/DTP	MDACC CLP	Achilles
NCI-60 & Almanac	60	51	40	55	1	55	31
CCLE		1089	823	687	42	388	580
CTRP			823	596	33	327	497
GDSC				1080	55	364	424
SCLC NCI/DTP					77	11	17
MDACC CLP						651	245
Achilles							769

C

Overlap of single drugs between datasets

	NCI-60 & Almanac	CCLE	CTRP	GDSC	SCLC NCI/DTP	MDACC CLP	Achilles
NCI-60	22257	20	80	125	183	NA	NA
CCLE		24	14	15	21	NA	NA
CTRP			481	72	121	NA	NA
GDSC				297	114	NA	NA
SCLC NCI/DTP					526	NA	NA
MDACC CLP						0	NA
Achilles							0

Figure 2. CellMinerCDB dataset overview. (**A**) Summary of molecular and drug activity data for the cell line sets included in CellMinerCDB. For the molecular and drug data types, the numbers indicate the number of genes or drugs. Blue numbers indicate a change in the number of features compared to the previous release, red numbers indicate new features and gray boxes denote entries for which there are no data in CellMinerCDB. (**B**) Number of cell line overlaps between data sources. (**C**) Number of single-drug activity overlaps between data sources. Abbreviations: datasets: National Cancer Institute (NCI-60), Cancer Cell Line Encyclopedia (CCLE), Cancer Therapeutics Response Portal (CTRP), Genomics of Drug Sensitivity in Cancer (GDSC), the NCI Small Cell Lung Cancer (SCLC) and the MD Anderson Cancer Center (MDACC) Cell Line Project (CLP); other: *z*-scores (zs), multiplatform microarray average $\log_2$ intensities (log2) and liquid chromatography–mass spectrometry (LC–MS).

Data representation and processing

CellMinerCDB datasets are created using R data packages that use two S4 class objects for data representation: molData and drugData as defined by the rcellminer R analysis package (59). The molData object contains results for molecular assays (e.g. genomics, proteomics, etc.) and drugData contains results for drug responses. The core of these objects is a list of ExpressionSet objects. This R data structure is made available by the Biobase Bioconductor package. This data representation allows molecular profiling and drug response data to be conveniently stored with sample metadata using a well-documented and widely available format [details are available in the documentation for the rcellminer and rcellminerData Bioconductor packages (59)].

Each CellMinerCDB data package contains the data from the data provider (i.e. Excel spreadsheets, tab-delimited files, etc.) along with R scripts used to process the data into the molData and drugData objects. In recent work, an instructional data package has been made available at github.com/CBIIT/rcellminerData using a subset of the NCI-60 rcellminerData Bioconductor package data to guide data package developers (59).

Software architecture

The CellMinerCDB web interface is built using the Shiny framework (shiny.rstudio.com) in the R programming language. This simplifies development, contribution and extension by developers with bioinformatic experience (where R is widely used) (60), in contrast to projects such as cBioPortal that depend on Java and JavaScript knowledge (61). CellMinerCDB architecture (Figure 4A) is based on pharmacogenomic data packages and project-specific R packages, including rcellminer (core data structure functionality), rcellminerUtilsCDB (gene signature and cross-database mapping), rcellminerElasticNet (multivariate analysis) and geneSetPathwayAnalysis (annotation information from GeneCards and Pathway Commons) (62,63). It should be noted that in this update, the rcellminer Bioconductor package removes previous Java

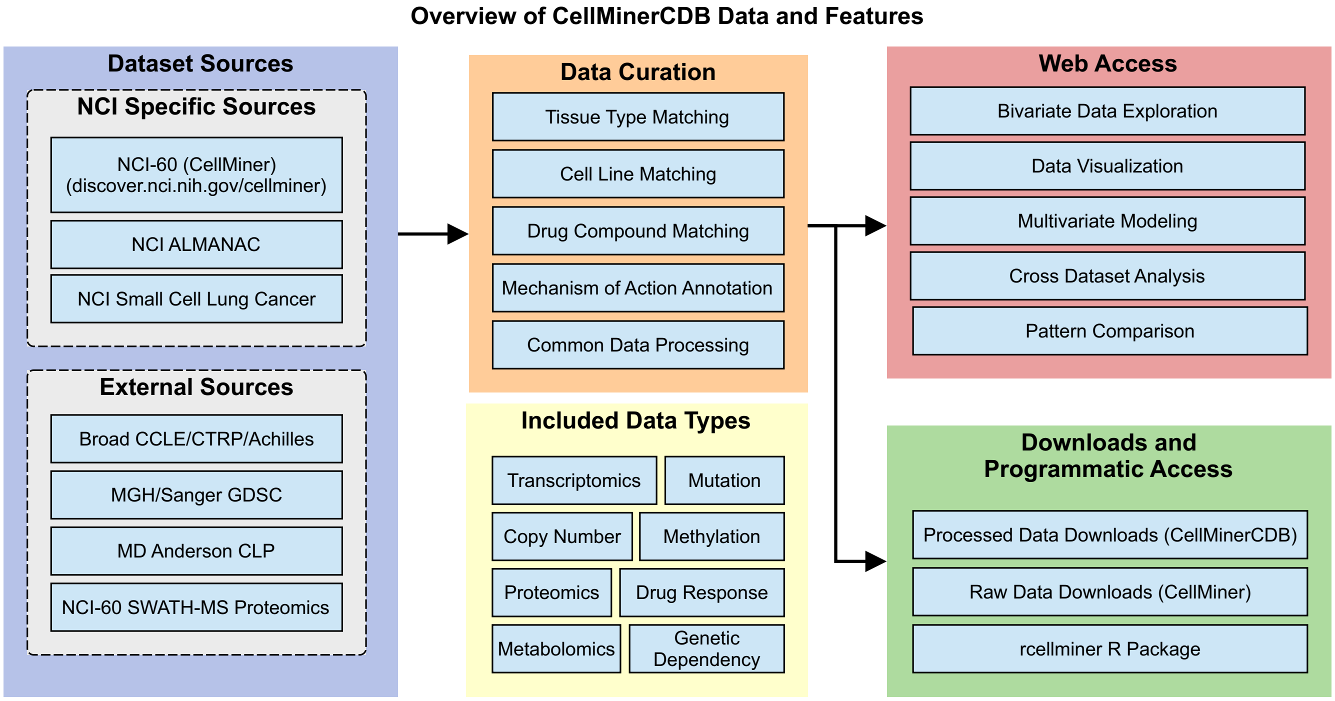


Figure 3. CellMinerCDB overview: CellMinerCDB integrates cancer cell line information from multiple sources and provides ready-made, user-friendly analysis tools as well as data download features for further analyses.

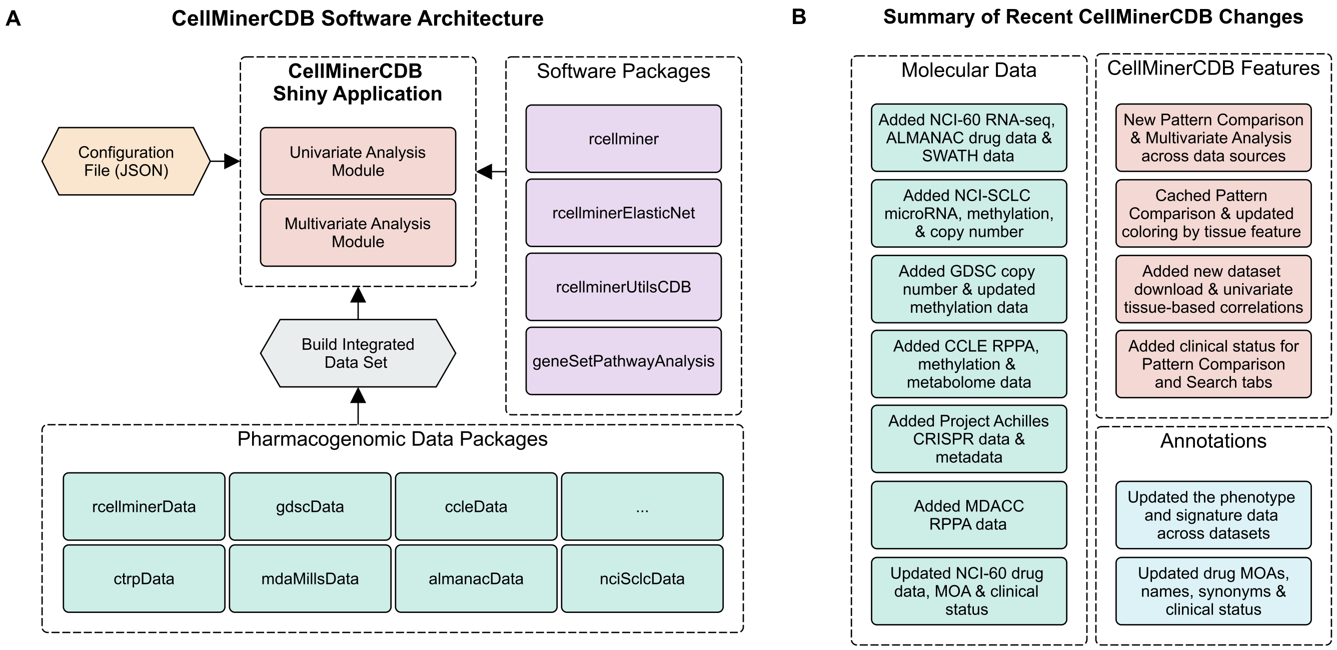

A
CellMinerCDB Software Architecture
Configuration File (JSON)
CellMinerCDB Shiny Application
Univariate Analysis Module
Multivariate Analysis Module
Software Packages
rcellminer
rcellminerElasticNet
rcellminerUtilsCDB
geneSetPathwayAnalysis
Build Integrated Data Set
Pharmacogenomic Data Packages
rcellminerData
gdscData
ccleData
...
ctrpData
mdaMillsData
almanacData
nciSclcData
B
Summary of Recent CellMinerCDB Changes
Molecular Data
Added NCI-60 RNA-seq, ALMANAC drug data & SWATH data
Added NCI-SCLC microRNA, methylation, & copy number
Added GDSC copy number & updated methylation data
Added CCLE RPPA, methylation & metabolome data
Added Project Achilles CRISPR data & metadata
Added MDACC RPPA data
Updated NCI-60 drug data, MOA & clinical status
CellMinerCDB Features
New Pattern Comparison & Multivariate Analysis across data sources
Cached Pattern Comparison & updated coloring by tissue feature
Added new dataset download & univariate tissue-based correlations
Added clinical status for Pattern Comparison and Search tabs
Annotations
Updated the phenotype and signature data across datasets
Updated drug MOAs, names, synonyms & clinical status

Figure 4. (**A**) CellMinerCDB software architecture. (**B**) Summary of recent CellMinerCDB changes. Colors: CellMinerCDB application elements (red), data packages (green), software packages (purple), configuration files (yellow), annotations/metadata (blue) and pre-processing steps (gray).

code dependencies to simplify installation. Various properties (e.g. datasets to make available and interface properties) of CellMinerCDB are configurable using a JavaScript Object Notation file. CellMinerCDB uses Shiny modules to support independent analysis modules (e.g. the multivariate analysis module) to allow developers to customize the web interface.

DATA AVAILABILITY

New functionality allows users to download data from any of the available -omic data via the CellMinerCDB (discover.nci.nih.gov/cellminercdb) Metadata tab as compressed (zip) files. Additionally, users can download drug synonym information or cell line annotation information from this tab.

The NCI-60 dataset has an additional method for download via the rcellminerData Bioconductor package.

UPDATES TO CELLMINERCDB FUNCTIONALITY AND USE CASES

In addition to recent dataset updates, CellMinerCDB functionality includes (i) updated drug annotations, (ii) expanded help information, (iii) improved pattern comparison speed, (iv) added functionality that provides correlation results by tissue type, (v) facilitated pattern comparison and (vi) improved multivariate and gene signature-based analyses across data sources (Figure 4B; complete release history: Supplementary Figure S2).

Training materials

CellMinerCDB provides several training materials, including detailed documentation on the CellMinerCDB website (discover.nci.nih.gov/cellminercdb) from the Help tab, as well as a video tutorial (youtube.com/watch?v=XljXazRGkQ8).

Use cases

This section presents several CellMinerCDB use cases, workflow and analysis examples. Figure 5 and Supplementary Figures S4–S6 provide visualizations across multiple use case examples. Each figure presents the input values (in the left-side gray boxes) to allow the exact reproduction of the result shown. Additional use case workflows are in the Help section of the CellMinerCDB website, and we provide a summary of available workflows in Supplementary Figure S3.

Quality control across database example. Figure 5A provides an example of data reproducibility across datasets. There has been recent discussion regarding the reliability and reproducibility of cell line data, especially when produced at different institutions (32,64,65). Reproducibility issues can arise from multiple factors including variability in what are erroneously perceived to be identical cell lines (66), and from the individuals doing the work and platforms used. CellMinerCDB allows direct comparison of omics profiles or drug activities. The example given shows that ATM expression values as measured by both the CCLE and GDSC are significantly correlated ($r = 0.77$, $P = 5.3e{-128}$). It illustrates an example where users are able to assess the reproducibility of the genomic and pharmacological data across cell lines processed independently with different platforms at different institutions (31). Additional data reproducibility examples are shown for DNA mutation, drug activity, DNA copy number and DNA methylation in Supplementary Figure S4A–E.

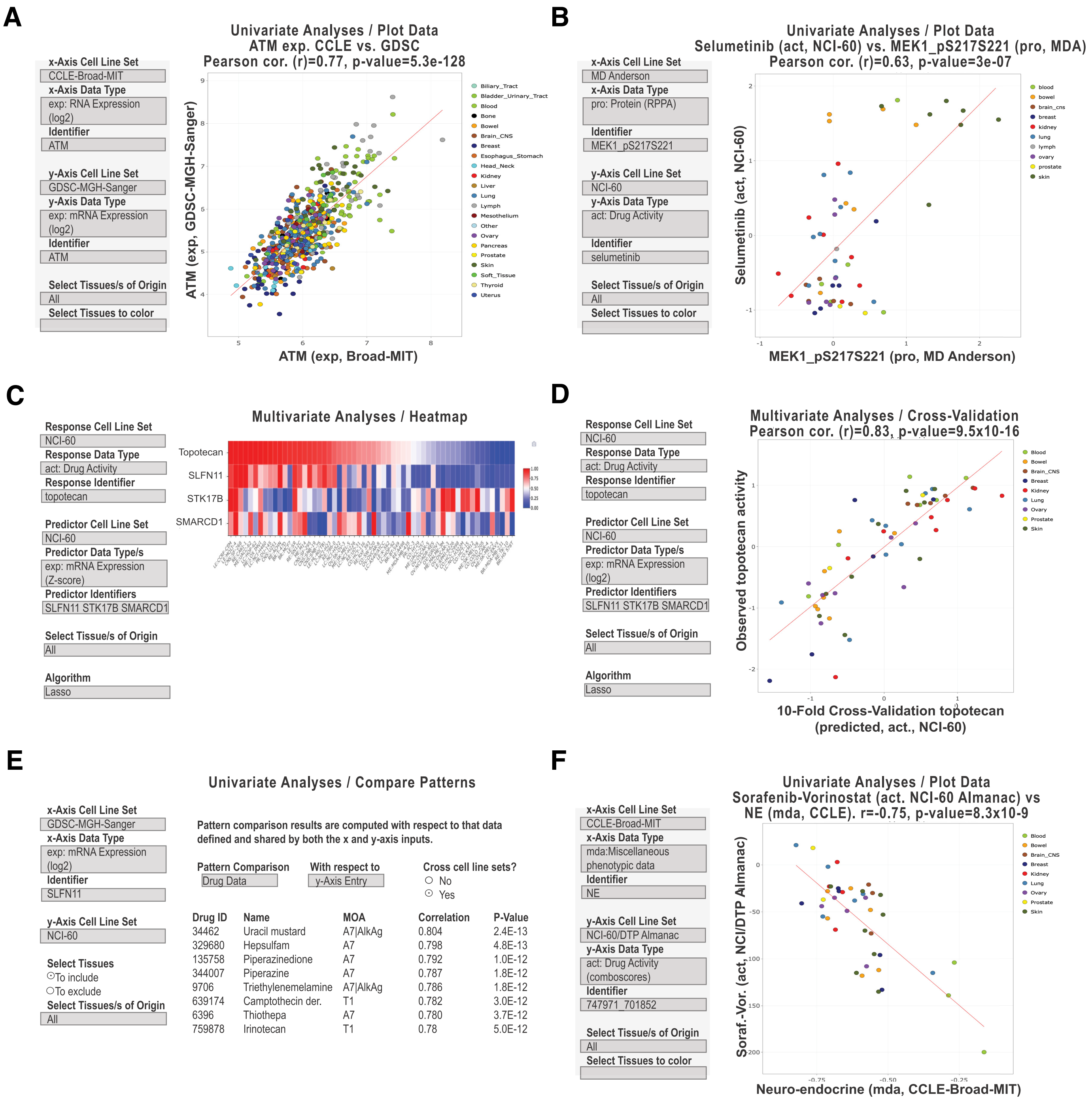
Identification of candidate drug biomarkers. Cancer cell lines are the logical starting place for identifying molecular features of relevance as part of precision medicine research. Figure 5B provides an example visualizing the relationship between MAP2K1 (using the name MEK1) phosphorylation status and the MAP2K1 inhibitor selumetinib. MAP2K1 is not presently a qualified biomarker for this FDA-approved drug. Here, we show a significant correlation between the selumetinib drug response in the NCI-60 and phosphorylation status in the MDACC RPPA dataset ($r = 0.63$, $P = 3e{-07}$), suggesting further investigation into patient response predictability. It should be noted that when searching for selumetinib, the drug synonyms (aliases) 'AZD6244' and '741078' can alternatively be used; available drug synonyms are downloadable as a table from the Metadata tab of the website. Supplementary Figure S5C and D provides additional examples of the types of analyses to connect drug activity to -omic features with examples using HSP90 and CDK inhibitors, respectively. Correlations for such analyses may not always be extremely high (>0.7) but may nonetheless be informative for further investigations.

Exploration of proteomic complexes. The addition of proteomic data allows users to explore the stoichiometric relationship of protein complex subunits. Supplementary Figure S5A shows the high correlation between protein levels of both subunits of the chromatin remodeling complex FACT (facilitates chromatin transcription): SSRP1 versus SUP16H ($r = 0.87$, $P = 1.6e{-19}$) in the NCI-60 SWATH mass spectrometry data (17,67). Supplementary Figure S5B also shows the correlation between the two components of the catenin complex involved in cell adhesion CTNNA1 and CTNNB1 (α- and β-catenin, respectively) (68).

Multivariate analysis example. In the majority of cases, multiple factors determine drug response in cell lines as well as in patients. Multivariate analysis allows investigation of drug response with respect to multiple -omic features simultaneously. CellMinerCDB allows users to either (i) manually input a multivariate model of their own design or (ii) have the system automatically generate a multivariate model for a molecular entity (i.e. an -omic or a drug response profile); here we have chosen to show the latter.

Figure 5C presents the CellMinerCDB 'Multivariate Analysis' functionality using the NCI-60 data. In the example, topotecan activity (a response variable) is analyzed as a function of gene transcript levels (potential response predictors, i.e. features). Using the LASSO (least absolute shrinkage and selection operator) feature selection algorithm (69), which is integrated into CellMinerCDB, three genes are identified automatically: SLFN11, STK17B and SMARCD1. The example puts forward a well-known predictor of topotecan response (SLFN11) (7,70) along with two predictors unknown in the topotecan literature (STK17B and SMARCD1) that may be the genesis of new hypotheses.

The relationship of the drug and selected features (i.e. predictors) of the LASSO-derived regression model is displayed as a heatmap (Figure 5C). The observed experimental topotecan response values and predicted response values (after 10-fold cross-validation) are highly correlated ($r = 0.83$, $P = 9.5e{-16}$; Figure 5D). Predicted response values are obtained (over 10 iterations) by successively holding out 10% of the cell lines and predicting their response using a linear regression model fit to the remaining 90% of the data. After all 10 iterations have been done, each sample has one cross-validated prediction. Details for


A
Univariate Analyses / Plot Data
ATM exp. CCLE vs. GDSC
Pearson cor. (r)=0.77, p-value=5.3e-128
ATM (exp, GDSC-MGH-Sanger)
ATM (exp, Broad-MIT)
Biliary_Tract
Bladder_Urinary_Tract
Blood
Bone
Bowel
Brain_CNS
Breast
Esophagus_Stomach
Head_Neck
Kidney
Liver
Lung
Lymph
Mesothelium
Other
Ovary
Pancreas
Prostate
Skin
Soft_Tissue
Thyroid
Uterus
B
Univariate Analyses / Plot Data
Selumetinib (act, NCI-60) vs. MEK1_pS217S221 (pro, MDA)
Pearson cor. (r)=0.63, p-value=3e-07
Selumetinib (act, NCI-60)
MEK1_pS217S221 (pro, MD Anderson)
blood
bowel
brain_cns
breast
kidney
lung
lymph
ovary
prostate
skin
C
Multivariate Analyses / Heatmap
Topotecan
SLFN11
STK17B
SMARCD1
D
Multivariate Analyses / Cross-Validation
Pearson cor. (r)=0.83, p-value=9.5x10-16
Observed topotecan activity
10-Fold Cross-Validation topotecan
(predicted, act., NCI-60)
Blood
Bowel
Brain_CNS
Breast
Kidney
Lung
Ovary
Prostate
Skin
E
Univariate Analyses / Compare Patterns
F
Univariate Analyses / Plot Data
Sorafenib-Vorinostat (act. NCI-60 Almanac) vs
NE (mda, CCLE). r=-0.75, p-value=8.3x10-9
Soraf.-Vor. (act, NCI/DTP Almanac)
Neuro-endocrine (mda, CCLE-Broad-MIT)
Blood
Bowel
Brain_CNS
Breast
Kidney
Lung
Ovary
Prostate
Skin

A

x-Axis Cell Line Set: CCLE-Broad-MIT
x-Axis Data Type: exp: RNA Expression (log2)
Identifier: ATM
y-Axis Cell Line Set: GDSC-MGH-Sanger
y-Axis Data Type: exp: mRNA Expression (log2)
Identifier: ATM
Select Tissues/s of Origin: All
Select Tissues to color:

B

x-Axis Cell Line Set: MD Anderson
x-Axis Data Type: pro: Protein (RPPA)
Identifier: MEK1_pS217S221
y-Axis Cell Line Set: NCI-60
y-Axis Data Type: act: Drug Activity
Identifier: selumetinib
Select Tissue/s of Origin: All
Select Tissues to color:

C

Response Cell Line Set: NCI-60
Response Data Type: act: Drug Activity
Response Identifier: topotecan
Predictor Cell Line Set: NCI-60
Predictor Data Type/s: exp: mRNA Expression (Z-score)
Predictor Identifiers: SLFN11 STK17B SMARCD1
Select Tissue/s of Origin: All
Algorithm: Lasso

D

Response Cell Line Set: NCI-60
Response Data Type: act: Drug Activity
Response Identifier: topotecan
Predictor Cell Line Set: NCI-60
Predictor Data Type/s: exp: mRNA Expression (log2)
Predictor Identifiers: SLFN11 STK17B SMARCD1
Select Tissue/s of Origin: All
Algorithm: Lasso

E

x-Axis Cell Line Set: GDSC-MGH-Sanger
x-Axis Data Type: exp: mRNA Expression (log2)
Identifier: SLFN11
y-Axis Cell Line Set: NCI-60
Select Tissues: ⦿ To include ○ To exclude
Select Tissues/s of Origin: All

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

Pattern Comparison: Drug Data
With respect to: y-Axis Entry
Cross cell line sets? ○ No ⦿ Yes

Drug ID	Name	MOA	Correlation	P-Value
34462	Uracil mustard	A7|AlkAg	0.804	2.4E-13
329680	Hepsulfam	A7	0.798	4.8E-13
135758	Piperazinedione	A7	0.792	1.0E-12
344007	Piperazine	A7	0.787	1.8E-12
9706	Triethylenemelamine	A7|AlkAg	0.786	1.8E-12
639174	Camptothecin der.	T1	0.782	3.0E-12
6396	Thiothepa	A7	0.780	3.7E-12
759878	Irinotecan	T1	0.78	5.0E-12

F

x-Axis Cell Line Set: CCLE-Broad-MIT
x-Axis Data Type: mda:Miscellaneous phenotypic data
Identifier: NE
y-Axis Cell Line Set: NCI-60/DTP Almanac
y-Axis Data Type: act: Drug Activity (comboscores)
Identifier: 747971_701852
Select Tissue/s of Origin: All
Select Tissues to color:

Figure 5. CellMinerCDB analysis examples. (**A**) Univariate analysis scatter plot of ATM transcript expression levels as measured by CCLE (Broad) versus GDSC (Sanger/MGH). (**B**) Univariate analysis scatter plot of MAP2K1 (MEK1) phosphoprotein levels versus the MAP2K1 inhibitor selumetinib activity levels. (**C**) Multivariate analysis and heatmap to identify and visualize molecular predictors for topotecan activity. (**D**) Multivariate analysis cross-validation scatter plot to visualize the observed versus predicted activities. (**E**) Univariate analysis 'Compare Patterns' output to identify drugs whose activities are most significantly correlated to SLFN11 expression. (**F**) Univariate analysis scatter plot of the NE transcript expression signature versus the two-drug activity NCI-ALMANAC ComboScore of sorafenib–vorinostat. All examples are captured images from CellMinerCDB (discover.nci.nih.gov/cellminercdb) using the selections detailed in the input box (on the left). Each dot in the scatter plots is a cell line, with tissues of origin indicated in the legend (on right). For the scatter plots, the regression trend line is in red. The *x*- and *y*-axes, correlations (Pearson's *r*) and *P*-values are as defined within each panel.

the LASSO algorithm and the multivariate analysis are in the CellMinerCDB Help section. Biologically, SLFN11 is known to affect topotecan response (7,70). STK17B is a serine/threonine kinase associated with apoptosis, and SMARCD1 is a SWI/SNF subunit of chromatin remodeling factors that could enhance the topoisomerase DNA accessibility necessary for topotecan activity (71,72). While the biological functions of STK17B and SMARCD1 plausibly affect topotecan response, neither has been previously reported in the scientific literature.

It should be noted for users that functionality to fix a model trained on one dataset and apply this to another dataset is not currently available and is better suited for more formal predictive analyses outside of CellMinerCDB to test model generalizability. Also, the interpretation of multivariate models (especially those that are algorithmically generated) requires scrutiny by individual researchers and should be guided by a well-formed understanding of the biological activities of the identified predictors and the MOA of the input drug to guide any follow-up effort. Besides manual and automated model creation, CellMinerCDB can be used to supplement a base model with additional features using partial correlation analysis.

Pattern comparison for the identification of biomarkers and drug targets. Figure 5E shows an example of broad-spectrum biomarker identification; here, a broad-spectrum biomarker is defined as a gene with a molecular profile having predictive strength across cancer types. Using the CellMinerCDB 'Compare Patterns' functionality, the transcript expression of SLFN11 from GDSC is compared to the activity profiles of all drugs and compounds in the NCI-60. High correlations are found for multiple DNA-damaging drugs, with TOP1 inhibitors and alkylating agents at the top, as well as TOP2, PARP1 and DNA synthesis inhibitors (not shown in the figure). These results provide unbiased evidence for SLFN11 as a potential biomarker for whole classes of commonly used standard of care chemotherapies (73). This example demonstrates the ability of CellMinerCDB to compare a pattern of interest via a correlation analysis to patterns of other types (transcript expression versus activity in this case). Furthermore, it demonstrates the cross-database capabilities of CellMinerCDB. The values presented here are not corrected for multiple testing though false discovery rate adjusted values are available via the website. Additionally, the pattern comparison feature provides a way to identify co-regulation between genes as previously shown to be involved in EMT (31) and lineage transcription factor pathways in small cell lung cancers (49). Furthermore, the 'Compare Patterns' panel includes gene location annotations to help reveal possible gene regulation due to proximity to neighbors and gene copy number alterations. Supplementary Figure S6A provides a visualization of a pattern comparison result (to that in Figure 5E) not employing the cross-cell line sets functionality update, that is GDSC SLFN11 transcription as compared to GDSC FDA-approved drug activities. In both cases, DNA-damaging drugs are recognized.

Exploring drug combinations and gene signatures. Figure 5F provides an example showing the availability of (i) phenotypic gene signatures and (ii) drug combination response data in CellMinerCDB. Correlation of a 50-gene NE status signature (55) with the response of cell lines to a combination of sorafenib and vorinostat shows that NE-like cells (higher NE score) potentially respond less well to the drug combination.

Exploring other data types: genetic dependencies and metabolomic data. The inclusion and integration of the CRISPR–Cas9 knockout data from Project Achilles allow users to search for co-dependent genes across cell lines. For instance, BRCA2 and PALB2 (Partner And Localizer of BRCA2) show high correlation ($r = 0.56$, $P = 7.1e{-64}$; Supplementary Figure S6B). This observation is consistent with the abrogation of homologous recombination (HR) activity that occurs when BRCA1–PALB2 binding is disrupted (74). A second CRISPR–Cas9 example is the correlation between silencing *BRAF* (*on cell survival*) and selective activity of vemurafenib ($P < 2.9e{-15}$), primarily in the melanoma cell lines, as expected due to their large presence of the V600E mutations (75,76). An example from the CCLE metabolite data, providing a comparison of inosine and guanosine levels ($r = 0.55$, $P = 2.3e{-73}$), is given in Supplementary Figure S6C. Inosine monophosphate (IMP) is converted to xanthosine monophosphate (XMP) by IMP dehydrogenase. XMP is converted to guanosine monophosphate (GMP) by GMP synthase (77).

Exploring drug structures. Currently, CellMinerCDB does not provide access to drug chemical structures. However, structures can be accessed as SMILES representations of NCI-60 compounds via the rcellminer R package on Bioconductor.

CONCLUSION

The past two decades have seen a surge of molecular and pharmacological data for overlapping sets of patient-derived cancer cell lines by multiple institutions (e.g. NCI, Broad Institute and Sanger/MGH). CellMinerCDB (discover.nci.nih.gov/cellminercdb) allows researchers to explore data across institutions using their expertise without bioinformatic support while benefiting from curated genomic and pharmacological annotations. The use case examples presented here are meant to provide a representative selection of explorations and analyses. They include a diverse group of analysis type databases: (i) data reproducibility: the uniquely compiled datasets here allow users to continue the discussion on the reproducibility of pharmacogenomic data (65,78) including the possibility of conducting further systematic analyses; (ii) candidate biomarker discovery; (iii) multivariate analysis of molecular features for drug activity; (iv) exploration of the relationship of composite features (multigene signatures versus response to drug combinations); (v) exploration of protein complexes, genetic dependencies and metabolomic pathways; and (vi) complementary analyses where data from one dataset supplement another (e.g. use of Broad CCLE gene expression data for matching Broad CTRP drug tested cell lines to calculate drug–gene expression correlations). Hence, a wide range of univariate and multivariate explorations can be undertaken. CellMinerCDB seamlessly allows this form of data assessment and integration. It opens the door to hypothesis generation and validation in non-isogenic cancer cell lines across multiple tissues of origin as well as the discovery of potential novel genomic regulatory networks and drug response determinants.

CellMinerCDB supports the analysis of many profiling platforms, including molecular parameters (e.g. gene/protein expression, alteration, metabolomic and genetic dependency), drug compound activity and phenotypic signatures (such as NE status, EMT and subsets of molecularly defined cancer subtypes as in the case of the small cell lung cancer cell lines). Yet, CellMinerCDB does not include drug perturbation analyses, which are provided by other platforms (e.g. the Connectivity Map). CellMinerCDB focuses on the molecular and genomic analyses of cancer cells at their steady-state levels prior to drug response. The goal of CellMinerCDB is to provide accessibility and analytical methods across previously disconnected datasets in a manner that allows its utilization by bench scientists, clinicians and others with domain knowledge, as well as traditional informaticists by combining the strengths and unique features of each dataset.

Going forward, CellMinerCDB will continue to (i) expand datasets [e.g. the PRISM drug repurposing collection (79)] and annotations, (ii) improve analysis speed and (iii) simplify installation for local reuse by other groups. Future software developments are being planned to (i) further advance the accuracy and coverage of annotations, for example through the use of algorithmic prediction [e.g. use of similarity analyses (4) to assess the possible MOA based on the similarity of drug response to known MOA compounds], (ii) use network biology tools and approaches to create models of drug response, (iii) extend analysis features to include comparisons between cell lines and patient cohorts starting with datasets from The Cancer Genome Atlas (TCGA) (80) and (iv) provide future customization by users of available signatures. We anticipate that the modular architecture design of CellMinerCDB will enable these additional analytical features.

CellMinerCDB complements pharmacogenomic data portals through the additional layer of curation and specialized analyses we provide across datasets. This enhances the value of individual pharmacogenomic datasets by its availability for comparison across projects to further our understanding of the combinatorial influences that affect pharmacological outcomes as we move toward personalized medicine.

SOFTWARE AVAILABILITY

All software developed as part of CellMinerCDB is freely available, open source and hosted GitHub in the following repositories: github.com/CBIIT/cellminercdb, github.com/CBIIT/rcellminerUtilsCDB, github.com/CBIIT/geneSetPathwayAnalysis, github.com/CBIIT/rcellminerElasticNet and github.com/CBIIT/rcellminer. Users of CellMinerCDB can provide feedback and ask questions of the development team using webadmin@discover.nci.nih.gov or users can submit developer feedback, file bug reports and request new features using project-specific issue trackers (e.g. github.com/CBIIT/cellminercdb/issues).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the many data providers and collaborators who have made this project possible, specifically Mahmoud Ghandi (CCLE), Paul Clemons (CTRP), Howard Lightfoot, Mathew J. Garnett (GDSC), Beverly Teicher (NCI, CCR), Paul Meltzer (NCI, CCR) and John Minna (UT Southwestern Medical Center). We would like to thank David Goldstein of the NCI Office of Science and Technology Resources for supporting the purchase of software required to enable efficient, multi-user access to the CellMinerCDB site.

FUNDING

National Cancer Institute [Z01 BC006150, ZIC BC011475, ZIA BC010411 and ZIC BC011509]; National Institute of General Medical Sciences [P41 GM103504 to A.L. and C.S.]. Funding for open access charge: National Cancer Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Cowley,G.S., Weir,B.A., Vazquez,F., Tamayo,P., Scott,J.A., Rusin,S., East-Seletsky,A., Ali,L.D., Gerath,W.F., Pantel,S.E. *et al.* (2014) Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data*, **1**, 140035.
2. Abaan,O.D., Polley,E.C., Davis,S.R., Zhu,Y.J., Bilke,S., Walker,R.L., Pineda,M., Gindin,Y., Jiang,Y., Reinhold,W.C. *et al.* (2013) The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.*, **73**, 4372–4382.
3. Reinhold,W.C., Sunshine,M., Liu,H., Varma,S., Kohn,K.W., Morris,J., Doroshow,J. and Pommier,Y. (2012) 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.*, **72**, 3499–3511.
4. Reinhold,W.C., Sunshine,M., Varma,S., Doroshow,J.H. and Pommier,Y. (2015) Using CellMiner 1.6 for systems pharmacology and genomic analysis of the NCI-60. *Clin. Cancer Res.*, **21**, 3841–3852.
5. Reinhold,W.C., Varma,S., Sousa,F., Sunshine,M., Abaan,O.D., Davis,S.R., Reinhold,S.W., Kohn,K.W., Morris,J., Meltzer,P.S. *et al.* (2014) NCI-60 whole exome sequencing and pharmacological CellMiner analyses. *PLoS One*, **9**, e101670.
6. Reinhold,W.C., Varma,S., Sunshine,M., Rajapakse,V., Luna,A., Kohn,K.W., Stevenson,H., Wang,Y., Heyn,H., Nogales,V. *et al.* (2017) The NCI-60 methylome and its integration into CellMiner. *Cancer Res.*, **77**, 601–612.
7. Zoppoli,G., Regairaz,M., Leo,E., Reinhold,W.C., Varma,S., Ballestrero,A., Doroshow,J.H. and Pommier,Y. (2012) Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents. *Proc. Natl Acad. Sci. U.S.A.*, **109**, 15030–15035.
8. Garnett,M.J., Edelman,E.J., Heidorn,S.J., Greenman,C.D., Dastur,A., Lau,K.W., Greninger,P., Thompson,I.R., Luo,X., Soares,J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

9. Iorio,F., Knijnenburg,T.A., Vis,D.J., Bignell,G.R., Menden,M.P., Schubert,M., Aben,N., Gonçalves,E., Barthorpe,S., Lightfoot,H. *et al.* (2016) A landscape of pharmacogenomic interactions in cancer. *Cell*, **166**, 740–754.
10. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
11. Rees,M.G., Seashore-Ludlow,B., Cheah,J.H., Adams,D.J., Price,E.V., Gill,S., Javaid,S., Coletti,M.E., Jones,V.L., Bodycombe,N.E. *et al.* (2016) Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.*, **12**, 109–116.
12. Stinson,S.F., Alley,M.C., Kopp,W.C., Fiebig,H.H., Mullendore,L.A., Pittman,A.F., Kenney,S., Keller,J. and Boyd,M.R. (1992) Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.*, **12**, 1035–1053.
13. Weinstein,J.N., Kohn,K.W., Grever,M.R., Viswanadhan,V.N., Rubinstein,L.V., Monks,A.P., Scudiero,D.A., Welch,L., Koutsoukos,A.D. and Chiausa,A.J. (1992) Neural computing in cancer drug development: predicting mechanism of action. *Science*, **258**, 447–451.
14. Weinstein,J.N., Myers,T.G., O'Connor,P.M., Friend,S.H., Fornace,A.J. Jr, Kohn,K.W., Fojo,T., Bates,S.E., Rubinstein,L.V., Anderson,N.L. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science*, **275**, 343–349.
15. Scherf,U., Ross,D.T., Waltham,M., Smith,L.H., Lee,J.K., Tanabe,L., Kohn,K.W., Reinhold,W.C., Myers,T.G., Andrews,D.T. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.
16. Nishizuka,S., Charboneau,L., Young,L., Major,S., Reinhold,W.C., Waltham,M., Kouros-Mehr,H., Bussey,K.J., Lee,J.K., Espina,V. *et al.* (2003) Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc. Natl Acad. Sci. U.S.A.*, **100**, 14229–14234.
17. Guo,T., Luna,A., Rajapakse,V.N., Koh,C.C., Wu,Z., Liu,W., Sun,Y., Gao,H., Menden,M.P., Xu,C. *et al.* (2019) Quantitative proteome landscape of the NCI-60 cancer cell lines. *iScience*, **21**, 664–680.
18. Reinhold,W.C., Mergny,J.-L., Liu,H., Ryan,M., Pfister,T.D., Kinders,R., Parchment,R., Doroshow,J., Weinstein,J.N. and Pommier,Y. (2010) Exon array analyses across the NCI-60 reveal potential regulation of TOP1 by transcription pausing at guanosine quartets in the first intron. *Cancer Res.*, **70**, 2191–2203.
19. Liu,H., D'Andrade,P., Fulmer-Smentek,S., Lorenzi,P., Kohn,K.W., Weinstein,J.N., Pommier,Y. and Reinhold,W.C. (2010) mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol. Cancer Ther.*, **9**, 1080–1091.
20. Reinhold,W.C., Reimers,M.A., Lorenzi,P., Ho,J., Shankavaram,U.T., Ziegler,M.S., Bussey,K.J., Nishizuka,S., Ikediobi,O., Pommier,Y.G. *et al.* (2010) Multifactorial regulation of E-cadherin expression: an integrative study. *Mol. Cancer Ther.*, **9**, 1–16.
21. Varma,S., Pommier,Y., Sunshine,M., Weinstein,J.N. and Reinhold,W.C. (2014) High resolution copy number variation data in the NCI-60 cancer cell lines from whole genome microarrays accessible through CellMiner. *PLoS One*, **9**, e92047.
22. Ji,J., Zhang,Y., Redon,C.E., Reinhold,W.C., Chen,A.P., Fogli,L.K., Holbeck,S.L., Parchment,R.E., Hollingshead,M., Tomaszewski,J.E. *et al.* (2017) Phosphorylated fraction of H2AX as a measurement for DNA damage in cancer cells and potential applications of a novel assay. *PLoS One*, **12**, e0171582.
23. Reinhold,W.C., Varma,S., Sunshine,M., Elloumi,F., Ofori-Atta,K., Lee,S., Trepel,J.B., Meltzer,P.S., Doroshow,J.H. and Pommier,Y. (2019) RNA sequencing of the NCI-60: integration into CellMiner and CellMinerCDB. *Cancer Res.*, **79**, 3514–3524.
24. Gmeiner,W.H., Reinhold,W.C. and Pommier,Y. (2010) 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.*, **9**, 3105–3114.
25. Garraway,L.A., Widlund,H.R., Rubin,M.A., Getz,G., Berger,A.J., Ramaswamy,S., Beroukhim,R., Milner,D.A., Granter,S.R., Du,J. *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, **436**, 117–122.
26. Li,H., Ning,S., Ghandi,M., Kryukov,G.V., Gopal,S., Deik,A., Souza,A., Pierce,K., Keskula,P., Hernandez,D. *et al.* (2019) The landscape of cancer cell line metabolism. *Nat. Med.*, **25**, 850–860.
27. Meyers,R.M., Bryan,J.G., McFarland,J.M., Weir,B.A., Sizemore,A.E., Xu,H., Dharia,N.V., Montgomery,P.G., Cowley,G.S., Pantel,S. *et al.* (2017) Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.*, **49**, 1779–1784.
28. Tsherniak,A., Vazquez,F., Montgomery,P.G., Weir,B.A., Kryukov,G., Cowley,G.S., Gill,S., Harrington,W.F., Pantel,S., Krill-Burger,J.M. *et al.* (2017) Defining a cancer dependency map. *Cell*, **170**, 564.e16–576.e16.
29. Ghandi,M., Huang,F.W., Jané-Valbuena,J., Kryukov,G.V., Lo,C.C., McDonald,E.R. 3rd, Barretina,J., Gelfand,E.T., Bielski,C.M., Li,H. *et al.* (2019) Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, **569**, 503–508.
30. Nusinow,D.P., Szpyt,J., Ghandi,M., Rose,C.M., McDonald,E.R. 3rd, Kalocsay,M., Jané-Valbuena,J., Gelfand,E., Schweppe,D.K., Jedrychowski,M. *et al.* (2020) Quantitative proteomics of the Cancer Cell Line Encyclopedia. *Cell*, **180**, 387.e16–402.e16.
31. Rajapakse,V.N., Luna,A., Yamade,M., Loman,L., Varma,S., Sunshine,M., Iorio,F., Sousa,F.G., Elloumi,F., Aladjem,M.I. *et al.* (2018) CellMinerCDB for integrative cross-database genomics and pharmacogenomics analyses of cancer cell lines. *iScience*, **10**, 247–264.
32. Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium (2015) Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, **528**, 84–87.
33. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
34. Jensen,M.A., Ferretti,V., Grossman,R.L. and Staudt,L.M. (2017) The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, **130**, 453–459.
35. Smirnov,P., Kofia,V., Maru,A., Freeman,M., Ho,C., El-Hachem,N., Adam,G.-A., Ba-Alawi,W., Safikhani,Z. and Haibe-Kains,B. (2018) PharmacoDB: an integrative database for mining *in vitro* anticancer drug screening studies. *Nucleic Acids Res.*, **46**, D994–D1002.
36. Yang,W., Soares,J., Greninger,P., Edelman,E.J., Lightfoot,H., Forbes,S., Bindal,N., Beare,D., Smith,J.A., Thompson,I.R. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–D961.
37. Caroli,J., Sorrentino,G., Forcato,M., Del Sal,G. and Bicciato,S. (2018) GDA, a web-based tool for genomics and drugs integrated analysis. *Nucleic Acids Res.*, **46**, W148–W156.
38. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437.e17–1452.e17.
39. Ahmed,J., Meinel,T., Dunkel,M., Murgueitio,M.S., Adams,R., Blasse,C., Eckert,A., Preissner,S. and Preissner,R. (2011) CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res.*, **39**, D960–D967.
40. Cokelaer,T., Chen,E., Iorio,F., Menden,M.P., Lightfoot,H., Saez-Rodriguez,J. and Garnett,M.J. (2018) GDSCTools for mining pharmacogenomic interactions in cancer. *Bioinformatics*, **34**, 1226–1228.
41. Smirnov,P., Safikhani,Z., El-Hachem,N., Wang,D., She,A., Olsen,C., Freeman,M., Selby,H., Gendoo,D.M.A., Grossmann,P. *et al.* (2016) PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, **32**, 1244–1246.
42. Zagidullin,B., Aldahdooh,J., Zheng,S., Wang,W., Wang,Y., Saad,J., Malyutina,A., Jafari,M., Tanoli,Z., Pessia,A. *et al.* (2019) DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.*, **47**, W43–W51.
43. Liu,H., Zhang,W., Zou,B., Wang,J., Deng,Y. and Deng,L. (2020) DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res.*, **48**, D871–D881.

44. Seo,H., Tkachuk,D., Ho,C., Mammoliti,A., Rezaie,A., Madani Tonekaboni,S.A. and Haibe-Kains,B. (2020) SYNERGxDB: an integrative pharmacogenomic portal to identify synergistic drug combinations for precision oncology. *Nucleic Acids Res.*, **48**, W494–W501.
45. Perkail,S., Andricovich,J., Kai,Y. and Tzatsos,A. (2020) BAP1 is a haploinsufficient tumor suppressor linking chronic pancreatitis to pancreatic cancer in mice. *Nat. Commun.*, **11**, 3018.
46. Hsieh,Y.-Y., Liu,T.-P., Chou,C.-J., Chen,H.-Y., Lee,K.-H. and Yang,P.-M. (2019) Integration of bioinformatics resources reveals the therapeutic benefits of gemcitabine and cell cycle intervention in SMAD4-deleted pancreatic ductal adenocarcinoma. *Genes*, **10**, 766.
47. Cheteh,E.H., Sarne,V., Ceder,S., Bianchi,J., Augsten,M., Rundqvist,H., Egevad,L., Östman,A. and Wiman,K.G. (2020) Interleukin-6 derived from cancer-associated fibroblasts attenuates the p53 response to doxorubicin in prostate cancer cells. *Cell Death Discov.*, **6**, 42.
48. Kriegsman,B.A., Vangala,P., Chen,B.J., Meraner,P., Brass,A.L., Garber,M. and Rock,K.L. (2019) Frequent loss of IRF2 in cancers leads to immune evasion through decreased MHC class I antigen presentation and increased PD-L1 expression. *J. Immunol.*, **203**, 1999–2010.
49. Tlemsani,C., Pongor,L., Girard,L., Roper,N., Elloumi,F., Varma,S., Luna,A., Rajapakse,V.N., Sebastian,R., Kohn,K.W. *et al.* (2020) SCLC_CellMiner: integrated genomics and therapeutics predictors of small cell lung cancer cell lines based on their genomic signatures. *Cell Rep.*, **33**, 108296.
50. Sidorov,P., Naulaerts,S., Ariey-Bonnet,J., Pasquier,E. and Ballester,P.J. (2019) Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front. Chem.*, **7**, 509.
51. Polley,E., Kunkel,M., Evans,D., Silvers,T., Delosh,R., Laudeman,J., Ogle,C., Reinhart,R., Selby,M., Connelly,J. *et al.* (2016) Small cell lung cancer screen of oncology drugs, investigational agents, and gene and microRNA expression. *J. Natl Cancer Inst.*, **108**, djw122.
52. Krushkal,J., Silvers,T., Reinhold,W.C., Sonkin,D., Vural,S., Connelly,J., Varma,S., Meltzer,P.S., Kunkel,M., Rapisarda,A. *et al.* (2020) Epigenome-wide DNA methylation analysis of small cell lung cancer cell lines suggests potential chemotherapy targets. *Clin. Epigenetics*, **12**, 93.
53. Kohn,K.W., Zeeberg,B.M., Reinhold,W.C. and Pommier,Y. (2014) Gene expression correlations in human cancer cell lines define molecular interaction networks for epithelial phenotype. *PLoS One*, **9**, e99269.
54. Wang,S., He,Z., Wang,X., Li,H. and Liu,X.-S. (2019) Antigen presentation and tumor immunogenicity in cancer immunotherapy response prediction. *eLife*, **8**, e49020.
55. Zhang,W., Girard,L., Zhang,Y.-A., Haruki,T., Papari-Zareei,M., Stastny,V., Ghayee,H.K., Pacak,K., Oliver,T.G., Minna,J.D. *et al.* (2018) Small cell lung cancer tumors and preclinical models display heterogeneity of neuroendocrine phenotypes. *Transl. Lung Cancer Res.*, **7**, 32–49.
56. Bairoch,A. (2018) The Cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech.*, **29**, 25–38.
57. Rudin,C.M., Poirier,J.T., Byers,L.A., Dive,C., Dowlati,A., George,J., Heymach,J.V., Johnson,J.E., Lehman,J.M., MacPherson,D. *et al.* (2019) Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nat. Rev. Cancer*, **19**, 289–297.
58. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213.
59. Luna,A., Rajapakse,V.N., Sousa,F.G., Gao,J., Schultz,N., Varma,S., Reinhold,W., Sander,C. and Pommier,Y. (2016) rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. *Bioinformatics*, **32**, 1272–1274.
60. Pittard,W.S. and Li,S. (2020) The essential toolbox of data science: Python, R, Git, and Docker. *Methods Mol. Biol.*, **2104**, 265–311.
61. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, l1.
62. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Stein,T.I., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*, **54**, 1.30.1–1.30.33.
63. Rodchenkov,I., Babur,O., Luna,A., Aksoy,B.A., Wong,J.V., Fong,D., Franz,M., Siper,M.C., Cheung,M., Wrana,M. *et al.* (2020) Pathway Commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic Acids Res.*, **48**, D489–D497.
64. Safikhani,Z., El-Hachem,N., Quevedo,R., Smirnov,P., Goldenberg,A., Juul Birkbak,N., Mason,C., Hatzis,C., Shi,L., Aerts,H.J. *et al.* (2016) Assessment of pharmacogenomic agreement. *F1000Res.*, **5**, 825.
65. Haibe-Kains,B., El-Hachem,N., Birkbak,N.J., Jin,A.C., Beck,A.H., Aerts,H.J.W.L. and Quackenbush,J. (2013) Inconsistency in large pharmacogenomic studies. *Nature*, **504**, 389–393.
66. Ben-David,U., Siranosian,B., Ha,G., Tang,H., Oren,Y., Hinohara,K., Strathdee,C.A., Dempster,J., Lyons,N.J., Burns,R. *et al.* (2018) Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, **560**, 325–330.
67. Huang,H., Santoso,N., Power,D., Simpson,S., Dieringer,M., Miao,H., Gurova,K., Giam,C.-Z., Elledge,S.J. and Zhu,J. (2015) FACT proteins, SUPT16H and SSRP1, are transcriptional suppressors of HIV-1 and HTLV-1 that facilitate viral latency. *J. Biol. Chem.*, **290**, 27297–27310.
68. Debruyne,P., Vermeulen,S. and Mareel,M. (1999) The role of the E-cadherin/catenin complex in gastrointestinal cancer. *Acta Gastroenterol. Belg.*, **62**, 393–402.
69. Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **58**, 267–288.
70. Reinhold,W.C., Thomas,A. and Pommier,Y. (2017) DNA-targeted precision medicine; have we been caught sleeping? *Trends Cancer Res.*, **3**, 2–6.
71. Doherty,G.A., Byrne,S.M., Austin,S.C., Scully,G.M., Sadlier,D.M., Neilan,T.G., Kay,E.W., Murray,F.E. and Fitzgerald,D.J. (2009) Regulation of the apoptosis-inducing kinase DRAK2 by cyclooxygenase-2 in colorectal cancer. *Br. J. Cancer*, **101**, 483–491.
72. Priam,P., Krasteva,V., Rousseau,P., D'Angelo,G., Gaboury,L., Sauvageau,G. and Lessard,J.A. (2017) SMARCD2 subunit of SWI/SNF chromatin-remodeling complexes mediates granulopoiesis through a CEBPε dependent mechanism. *Nat. Genet.*, **49**, 753–764.
73. Murai,J., Thomas,A., Miettinen,M. and Pommier,Y. (2019) Schlafen 11 (SLFN11), a restriction factor for replicative stress induced by DNA-targeting anti-cancer therapies. *Pharmacol. Ther.*, **201**, 94–102.
74. Foo,T.K., Tischkowitz,M., Simhadri,S., Boshari,T., Zayed,N., Burke,K.A., Berman,S.H., Blecua,P., Riaz,N., Huo,Y. *et al.* (2017) Compromised BRCA1–PALB2 interaction is associated with breast cancer risk. *Oncogene*, **36**, 4161–4170.
75. Patrawala,S. and Puzanov,I. (2012) Vemurafenib (RG67204, PLX4032): a potent, selective BRAF kinase inhibitor. *Future Oncol.*, **8**, 509–523.
76. Chapman,P.B., Hauschild,A., Robert,C., Haanen,J.B., Ascierto,P., Larkin,J., Dummer,R., Garbe,C., Testori,A., Maio,M. *et al.* (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.*, **364**, 2507–2516.
77. Gu,J.J., Tolin,A.K., Jain,J., Huang,H., Santiago,L. and Mitchell,B.S. (2003) Targeted disruption of the inosine 5′-monophosphate dehydrogenase type I gene in mice. *Mol. Cell. Biol.*, **23**, 6702–6712.
78. Safikhani,Z., Smirnov,P., Freeman,M., El-Hachem,N., She,A., Rene,Q., Goldenberg,A., Birkbak,N.J., Hatzis,C., Shi,L. *et al.* (2016) Revisiting inconsistency in large pharmacogenomic studies. *F1000Res.*, **5**, 2333.
79. Corsello,S.M., Nagari,R.T., Spangler,R.D., Rossen,J., Kocak,M., Bryan,J.G., Humeidi,R., Peck,D., Wu,X., Tang,A.A. *et al.* (2020) Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer*, **1**, 235–248.
80. Sanchez-Vega,F., Mina,M., Armenia,J., Chatila,W.K., Luna,A., La,K.C., Dimitriadoy,S., Liu,D.L., Kantheti,H.S., Saghafinia,S. *et al.* (2018) Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell*, **173**, 321.e10–337.e10.