Nucleic Acids Research, 2025, gkaf1230
https://doi.org/10.1093/nar/gkaf1230
Database issue

CellMiner cross-database (CellMinerCDB) version 2.2 for explorations of patient-derived cancer cell line pharmacogenomics

Fathi Elloumi[1,*], William C. Reinhold[1,*], Sudhir Varma[1,2], Yanghsin Wang[1,3], Meric Kinali[4], Yasuhiro Arakawa[5], Yoshitaka Inoue[1,6,7], Mirit I. Aladjem[1], Yves Pommier[1,*], Augustin Luna[1,7,*]

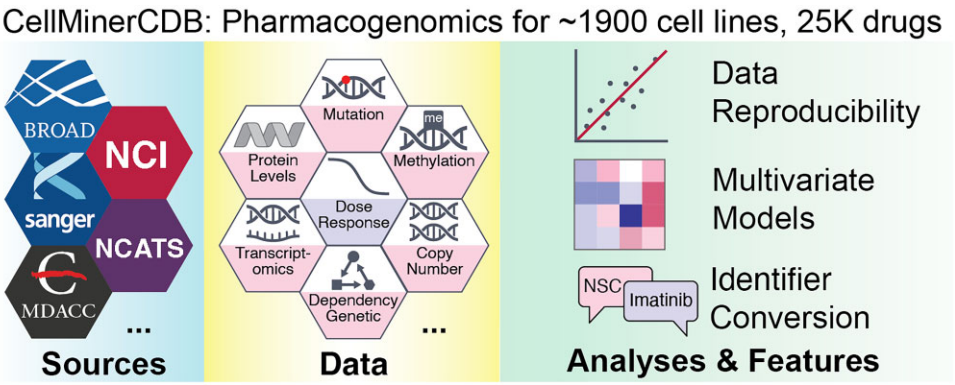
[1]Developmental Therapeutics Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20814, United States
[2]HiThru Analytics LLC, Princeton, NJ 08540, United States
[3]Essex Management, LLC, Rockville, MD 20850, United States
[4]Computer Science Department, University of Massachusetts Boston, Boston MA 02125, United States
[5]Department of Clinical Pharmacology and Therapeutics, Jikei University School of Medicine, Tokyo 105-8461, Japan
[6]Department of Computer Science & Engineering, College of Science & Engineering, University of Minnesota, Minneapolis, MN 55455, United States
[7]Computational Biology Branch, National Library of Medicine, NIH, Bethesda, MD 20894, United States

*To whom correspondence should be addressed. Tel: +1 240 760 6601; Fax: +1 240 541 4475; Email: fathi.elloumi@nih.gov
Correspondence may also be addressed to William C. Reinhold: Email: wcr@mail.nih.gov
Correspondence may also be addressed to Yves Pommier: Email: pommiery@nih.gov
Correspondence may also be addressed to Augustin Luna: Email: augustin @nih.gov

Abstract

CellMiner Cross-Database (CellMinerCDB) (https://discover.nci.nih.gov/cellminercdb/) is an established interactive application providing direct access and enabling exploration of cancer cell line pharmacogenomics without extensive programming experience. Data are compiled from many sources, including the National Cancer Institute(NCI), Broad Institute Dependency Map (DepMap), Sanger/MGH Genomics of DrugSensitivity in Cancer (GDSC), MD Anderson Cell Lines Project (MCLP), andNational Center for Advancing Translational Sciences (NCATS). In the version 2.2 update, our collection has expanded to pharmacogenomics data for 1916 cancer cell lines and over 25 000 drugs. Drug screening data include many additional compounds for potential drug repurposing from the Broad PRISM, NCATS, and NCI. The user interface facilitates uncovering specific samples of interest and identifying drug and cell lines across databases. We also expanded the annotations for cross-referencing other databases and downloading our data for further cancer biology and drug discovery studies. Herein, we provide use cases for CellMinerCDB, including (i) data reproducibility given overlaps of cell lines, genes, and drugs across databases; (ii) candidate biomarker discovery; and (iii) cross-dataset analyses.

Graphical abstract



Introduction

Cancer cell lines are key models for investigating the biological characteristics that drive cancer development, proliferative signaling, evasion of growth suppression, drug discovery, and resistance [1, 2]. Studies of these processes necessitate examination of multiple layers of omics (e.g. transcriptomics, proteomics, metabolomics, etc.), identification of interactions that arise between cellular gene products, and analyses of pharmacological response with the end goals of understanding cancer biology, discovering new drugs,

Received: September 17, 2025. **Revised:** October 10, 2025. **Accepted:** October 13, 2025
Published by Oxford University Press 2025.

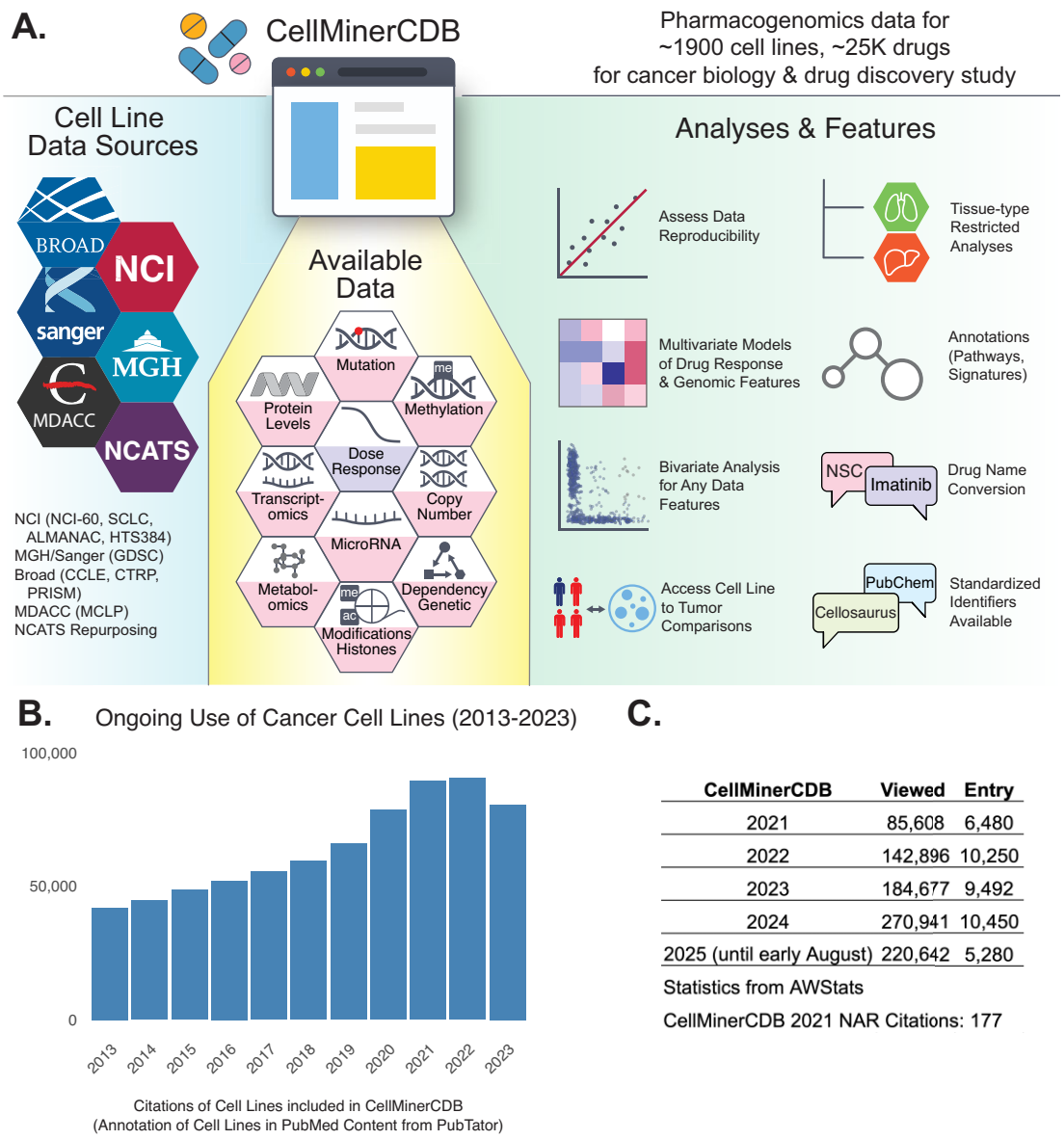


Figure 1. (**A**) Overview of CellMinerCDB. (**B**) Citation counts for cell lines included in CellMinerCDB from 2013 to 2023; data for 2024 are not included due to incompleteness; (**C**) CellMinerCDB web usage using AWStats tool (awstats.org). The table presents the number of viewed pages and unique entries to CellMinerCDB.

elucidating their molecular pharmacology, and identifying predictive biomarkers. CellMiner Cross-Database (CellMinerCDB) has been developed at the National Cancer Institute (NCI) to provide access to molecular profiling and drug screening data beyond the 60 cancer cell lines from 9 cancer types known as the NCI-60 (https://discover.nci.nih.gov/cellminer/) [3–5]. CellMinerCDB () integrates large-scale cell line cancer pharmacogenomics data from institutes outside the NCI, including the Broad, Sanger, MD Anderson Cancer Center (MDACC), and NIH NCATS (National Center for Advancing Translational Sciences) [6–13].

As an interactive data analysis platform, CellMinerCDB enables researchers to explore and analyze data across institutions without bioinformatic support while benefiting from curated -omic and pharmacological annotations (Fig. 1A). Examples of potential use cases are extensive, and include: (i) analyzing data reproducibility; (ii) multivariate analyses of molecular features for prediction of drug activity; (iii) explorations of multigene signatures; (iv) explorations of genetic dependency effects; and (v) complementary analyses where data from one dataset supplements data from another (for instance, use of Broad CCLE gene expression data to match Broad CTRP chemosensitivity data from drug tested cell lines to calculate drug–gene expression correlations).

Since its initial release, CellMinerCDB has expanded, now including information for 1916 cancer cell lines from the initial approximately 1400. Collective use of these cell lines for research is extensive, as shown in Fig. 1B, based on PubTator cell line annotations of PubMed articles [14]. With the continued use and expansion of CellMinerCDB (Fig. 1C), we have updated both its data and interface functionalities, including the addition of new drug and cancer cell line datasets (including new drug repurposing datasets) and updates to several others. Concomitantly, we provide improved annotations of the

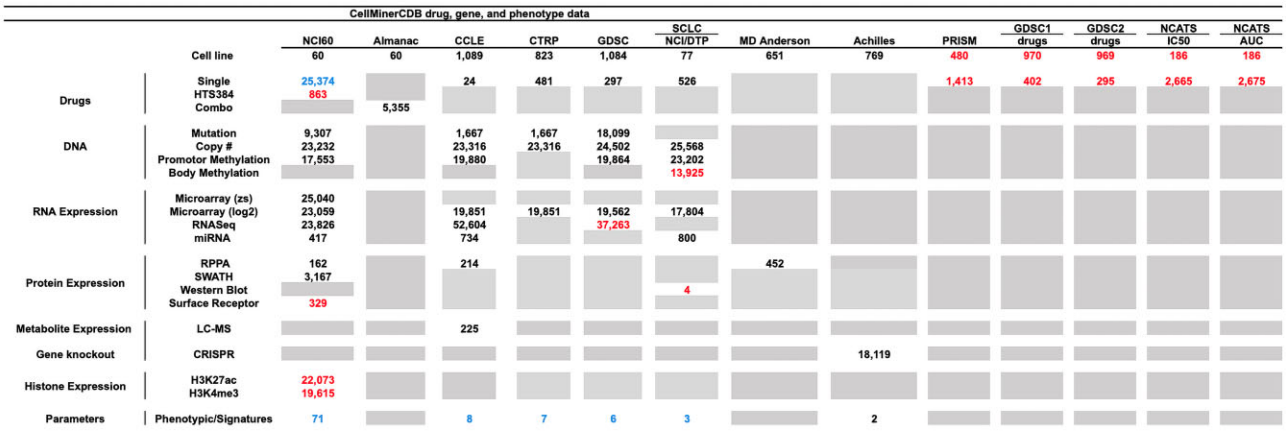


Figure 2. Summary of molecular and drug activity data for the cell line sets included in CellMinerCDB version 2.2. For the molecular and drug data types, the numbers indicate the number of genes and drugs, respectively. For the drug data, the "Combo" entry is for two-drug combinations from the NCI60 Almanac. A grey box indicates that no data are available. Numbers highlighted in blue indicate a change in the number of features compared to the previous version, whereas numbers highlighted in red indicate the number of new features.

data for the harmonization of identifiers across datasets. This information is available to users through new cell line or drug information pages detailing data availability across datasets. Functionally, CellMinerCDB has been enhanced with features that enable a broader range of cross-database analyses by: (i) allowing users to select or search for cell lines during their exploration (e.g. within scatter plots); (ii) hierarchical selection of tissue types of interest; and (iii) simplified selection of metadata or signatures in univariate analysis. As part of this update, we provide use case examples that illustrate the research potential and complementarity of the integrated data, focused on newly added datasets.

Cellminercdb datasets

Since our previous publication [7], CellMinerCDB (https://discover.nci.nih.gov/cellminercdb/) has been updated with new pharmacogenomics data. Fig. 2 summarizes the content of the new release. Fig. 3A shows the number of drugs and assay type per dataset, as well as the overlap across the datasets (new datasets are in red), and Fig. 3B shows the number of cell lines per dataset and the overlap of cell lines.

CellMinerCDB new drug screening data

The new drug datasets include the NIH NCATS (National Center for Advancing Translational Sciences) drug data, the new "NCI-60 HTS384" drug activity data, and the Broad Institute PRISM (Profiling Relative Inhibition Simultaneously in Mixtures) secondary screening drug data (i.e. "MTS010") (https://depmap.org/repurposing/) [5, 11, 15, 16]. The NCATS and PRISM (Profiling Relative Inhibition Simultaneously in Mixtures) include 19% and 27% non-oncology drugs, respectively (based on source annotations), which can be explored both for molecular pharmacology insights as well as for prospective drug repurposing studies in anti-cancer therapies. Additionally, we make available separate collections of the GDSC (Genomics of Drug Sensitivity in Cancer) phase 1 (GDSC1) and phase 2 (GDSC2) drug data (https://www.cancerrxgene.org/) [17–19].

Cell line viability measurements included in the datasets that comprise CellMinerCDB utilize various methodologies (Fig. 3A). Drug activity for most of the CellMinerCDB datasets was measured by adenosine triphosphate (ATP) levels (CellTiter-Glo assay). Data for GDSC1 (Phase 1) were collected using both the rezaurin (oxidation-reduction) and Syto60 (nucleic acid binding). New NCI datasets (NCI-60 and SCLC) use total protein (Sulforhodamine B assay), and PRISM uses a DNA barcoding-based approach.

Updates to existing datasets

The new CellMinerCDB release (version 2.2) integrates the latest version (version 2.14) of the NCI-60 Developmental Therapeutics Program (DTP) drug response screening data using the previous sulforhodamine B protocol. We also integrated recent profiling data for surface receptors on the NCI-60 using flow cytometry for 332 receptor proteins [15]. Separately, we have included ChIP-seq (chromatin immunoprecipitation followed by sequencing) data measuring activating histone H3K4me3 and H3K27ac modifications [20]. Data have been included for other datasets, such as RNAseq data for GDSC (https://cellmodelpassports.sanger.ac.uk/downloads); previously, only microarray gene expression measurements were available in CellMinerCDB for GDSC. For the NCI SCLC dataset, we also provide gene body methylation and protein levels (measured by Western blotting) and data for NCI SCLC [21].

We have now incorporated gene signatures for fundamental cancer processes. For example, we generated and have included the 18-gene signature for DNA replication stress (RepStress) computed as a weighted sum of standardized (Z-score) transcript expression scores for all datasets that have gene expression data [22]. Also, the CCLE dataset now includes an additional Centromere and kinetochore gene Expression Score (CES) signature score [23].

Software infrastructure

Core software

The core backend software infrastructure remains the same from our previous publication [7]. Briefly, CellMinerCDB is built using the R Shiny framework (shiny.rstudio.com). This simplifies development and extension by developers with

A.

Overlap of single drugs between datasets

	NCI-60	CCLE	CTRP	GDSC	SCLC NCI/DTP	PRISM	GDSC1 drugs	GDSC2 drugs	NCI60 HTS384	NCATS IC50	NCATS AUC	Measurement/Assay*
NCI-60	25374	21	131	146	384	395	162	140	607	689	689	Total protein (sulforhodamine B) assay, 48 hrs
CCLE		24	14	16	21	22	16	14	21	22	22	ATP (CellTiter-Glo), 72 hrs
CTRP			481	77	128	134	83	66	112	165	165	ATP (CellTiter-Glo), 72 hrs
GDSC				297	115	134	235	114	112	201	201	NUC-AC (Syto 60). OXID-RD (Resazurin); 72 hrs
SCLC NCI/DTP					526	286	131	106	373	400	400	Total protein (sulforhodamine B) assay, 48 hrs
PRISM						1413	149	116	287	795	797	DNA barcode (PRISM)
GDSC1 drugs							402	120	135	245	245	Resazurin or Syto60 assay, 72 hrs
GDSC2 drugs								295	117	181	181	ATP (CellTiter-Glo), 72 hrs
NCI60 HTS384									863	522	522	ATP (CellTiter-Glo), 72 hrs
NCATS IC50										2665	2657	ATP (CellTiter-Glo), 48 hrs
NCATS AUC											2675	ATP (CellTiter-Glo), 48 hrs

* the number of hours is post-treatment

Abbreviations: ATP: adenosine triphosphate, OXID-RD: oxidation reduction, NUC-AC: nucleic acid

B.

Overlap of cell lines between datasets

	NCI-60, HTS384 & Almanac	CCLE	CTRP	GDSC	SCLC NCI/DTP	MD Anderson	Achilles	PRISM	GDSC1 drugs	GDSC2 drugs	NCATS IC50	NCATS AUC
NCI-60, HTS384 & Almanac	60	52	40	55	1	55	31	34	55	55	8	8
CCLE		1089	823	696	42	401	581	480	651	652	102	102
CTRP			823	601	33	334	498	441	566	566	90	90
GDSC				1084	55	380	432	362	970	966	82	82
SCLC NCI/DTP					77	14	17	9	46	45	12	12
MD Anderson						651	250	202	359	356	63	63
Achilles							769	343	403	403	56	56
PRISM								480	346	346	30	30
GDSC1 drugs									970	961	80	80
GDSC2 drugs										969	79	79
NCATS IC50											186	186
NCATS AUC												186

Figure 3. CellMinerCDB dataset overview. (**A**) The number of drugs overlaps between data sources. For GDSC, we kept the previous version of drug data and added the GDSC phase 1 (GDSC1) and phase 2 (GDSC2) drug data. For NCATS, we provide both IC50 and AUC drug data. (**B**) The number of cell line overlaps between data sources. For both panels, the updated datasets are highlighted in red.

bioinformatic experience, where R programming is widely used. The CellMinerCDB architecture is based on pharmacogenomic data packages (for each dataset) and a series of core R packages that we develop and maintain: rcellminer (core data structure functionality), rcellminerUtilsCDB (cross-database mapping), rcellminerElasticNet (multivariate analysis), and geneSetPathwayAnalysis (annotation information from GeneCards and Pathway Commons) [8, 24, 25]. In a recent work, we provide a detailed protocol describing software installation and data formats used [26].

Annotation updates

As part of this release, we update the rcellminerUtilsCDB and geneSetPathwayAnalysis packages for gene sets primarily used as part of the multivariate analysis, as well as update cell lines and drug annotations (e.g. external identifiers). We now use Cellosaurus (https://www.cellosaurus.org) accession identifiers as the common identifier for cell lines across the different datasets. Also, for download, we now provide PubChem Compound Identifiers (CID) when it is present in the original data source (e.g. NCI-60, GDSC1, and GDSC2); this includes over 1000 of our common drugs (i.e. drugs in multiple datasets) and the majority of the NCI-60 DTP compounds. This will allow users to cross-reference compounds in our data collection to the rich metadata available in PubChem that includes information such as specific clinical trials, patents, disease indications, toxicity, as well as additional identifiers linking to resources such as DrugBank, Chemical Abstracts Service (CAS) registry, KEGG, MeSH (Medical Subject Headings), and PubMed publications.

Given that cell lines are routinely used as models of patient tumors, it is important to understand the similarity of particular cell lines to tumor samples. Thus, we now include precalculated similarity scores from the TumorComparer study of ~8000 tumor samples of 24 cancer types [27]. For download, we provide the averaged similarity scores across gene expression, copy number, and mutation profiles for 575 cell lines that overlap between the TumorComparer and CellMinerCDB cell lines.

Examples of updates to CellMinerCDB functionality and use cases

Interface changes

The CellMinerCDB interface has been updated to follow the new mandatory U.S. Web Design System (USWDS, https://designsystem.digital.gov/), which provides greater consistency across federal web services. Visual presentation of the site has mostly been tested with mid-20-inch desktop monitors (e.g. base font sizes); typical browser zoom features can be used to scale text based on user need. Separate from the global design edits, we have made additional interface changes to simplify the use of the site. We have made it easier to select a signature (e.g. multi-gene signatures) or miscellaneous data (e.g. cell line doubling time) by providing a list of all options, as users may not be familiar with the available options. Likewise, the tissue selection has been upgraded using a tree structure for display and selection purposes; this allows users to see all available options as well as understand the hierarchical nature of the options.

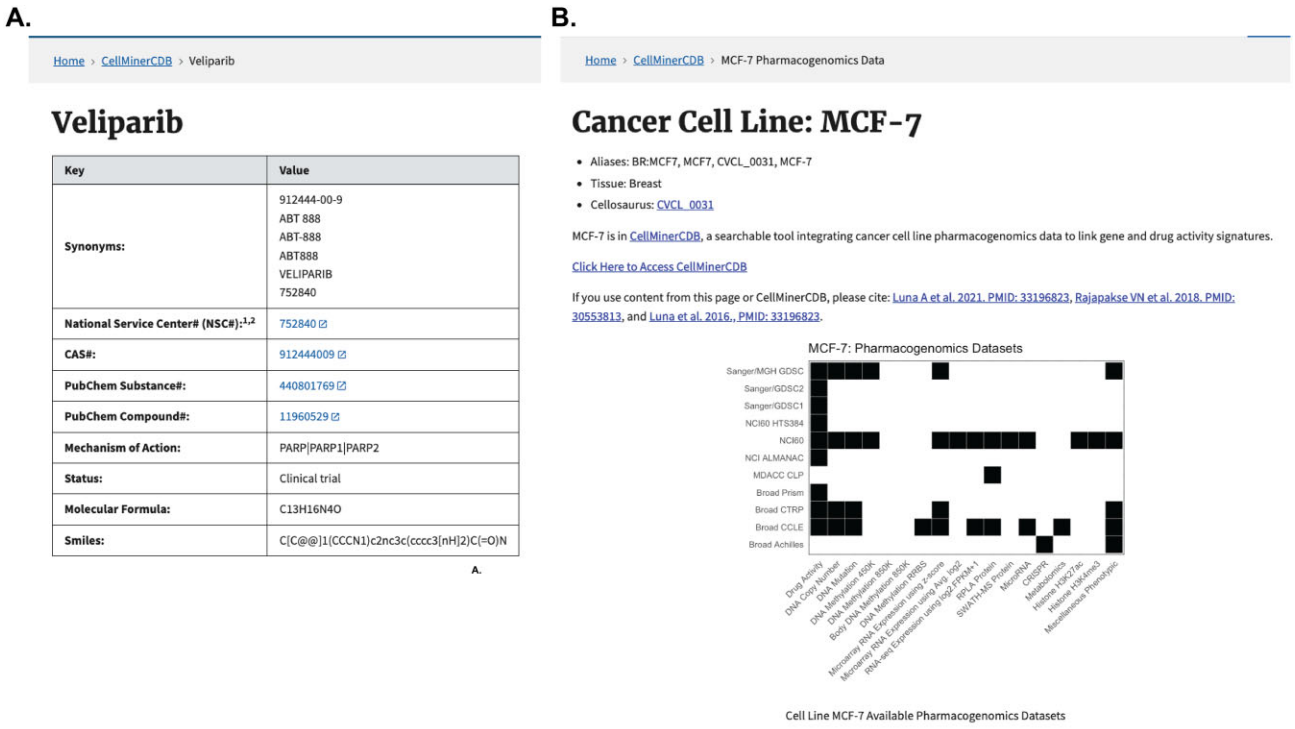


Figure 4. New cell line and drug information page overview. (**A**) Veliparib drug data information and its availability, and matching names across all datasets. (**B**) Example for the MCF-7 cell line showing the available pharmacogenomic data across all datasets.

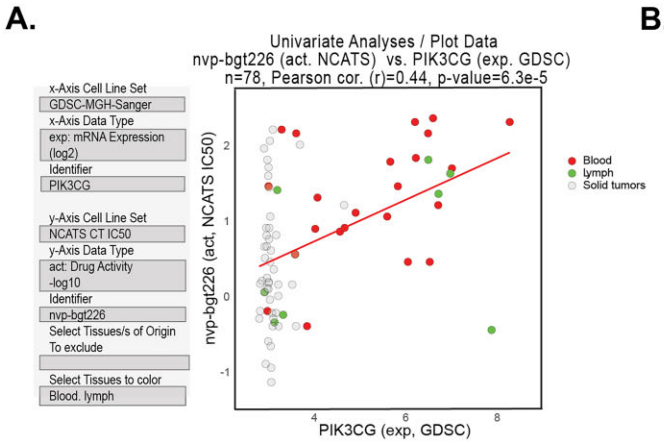
We provide two new features in the univariate analysis. The scatter plot now displays the number of samples and a text box to search and highlight specific cell line samples within the plot. We have updated the gene set list for the multivariate Lasso model. Aditionally, CellMinerCDB data can be used to facilitate sophisticated machine learning studies, such as the recent drGT graph neural network-based algorithm for drug response prediction that uses CellMinerCDB data [28]. Furthermore, we added a new cell surface gene set from which the automated Lasso feature (predictor) selection can choose genes. The cell surface genes were taken from Waas *et al.* [29] and manually curated internally. To help users identify content of interest, we have added two tabs to query the list of cell lines and common drugs (i.e. drugs that tend to be observed in multiple datasets). Detailed links from this query provide pages for a given cell line or drug with datasets and data type availability across all the datasets in CellMinerCDB. Fig. 4 presents the "Veliparib" drug information (Fig. 4A) and the "MCF-7" cell line information (Fig. 4B).

Use cases

Here we present several use cases and analysis examples. Scatter plots and associated data tables were generated using CellMinerCDB (https://discover.nci.nih.gov/cellminercdb/).

All drugs included in Fig. 5A and B are either FDA-approved or in clinical trials, and the genes listed are established targets for those drugs. Of the 11 drugs presented, seven are used for cancer and four for non-cancer treatments. In Fig. 5A, the scatter plot presents a correlation between the expression of PIK3CG (Phosphoinositide-3-Kinase Gamma Catalytic Subunit) and the sensitivity of cancer cells to the anticancer drug currently in clinical trials, NVP-BGT226 (BTG226), targeting both the PIK3CG and MTOR proteins. This scatterplot shows the overall correlation between PIK3CG gene expression and drug activity. It also shows that solid tumors exhibit low expression levels for PIK3CG (the target), essentially at background levels. From this analysis, we can conclude that (i) leukemias and lymphomas exhibit a wide range of PIK3CG transcript expression, with many cell lines expressing the gene at much higher levels than the solid tumors; and (ii) the average NVP-BGT226 activity levels are higher for the leukemias and lymphomas than for the solid tumors, which do not appear plausible targets to PIK3CG inhibitors. Fig. 5B provides 10 more analogous examples of drug targets (based on their expression in GDSC) with significant correlations to the drugs' activities. Drugs shown for this example come from the updated NCATS drug screening data collection.

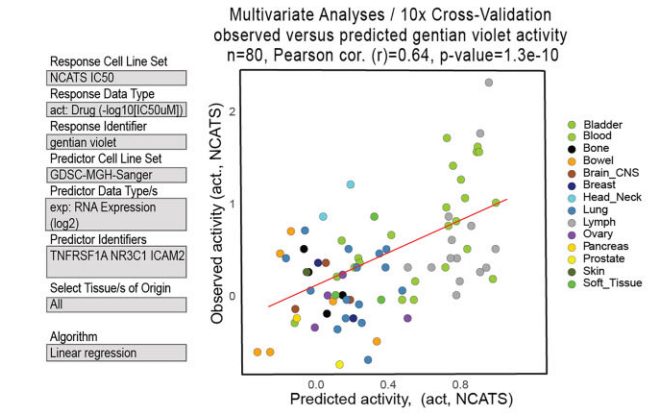
The drug NVP-BGT226 from Fig. 5A is currently in clinical trial for solid tumors, including breast, head and neck, endothelial cells, and lung cancer [30]. As our scatter plot indicates, the drug may potentially be more active in leukemias and lymphomas. In addition, we find that PIK3CG expression may serve as a biomarker for NVP-BGT226 activity in leukemias and lymphomas, potentially making patients with these diseases better candidates for this drug. Fig. 5A and B also illustrate how the use of the NCATS data facilitates doing the same type of pharmacogenomics analysis in the non-cancer arena as for known cancer treatments. An example is the iloperidone and DRD2 drug-gene pairing, in which the target DRD2 is shown to be significantly correlated to the drug; Iloperidone is an FDA-approved schizophrenia drug, and DRD2 is a D2 subtype of the dopamine receptor (Fig. 5B). The drug is shown here to be more active when the gene is expressed at higher levels. Gene expression data for these examples comes from the GDSC microarray gene transcript measurements.

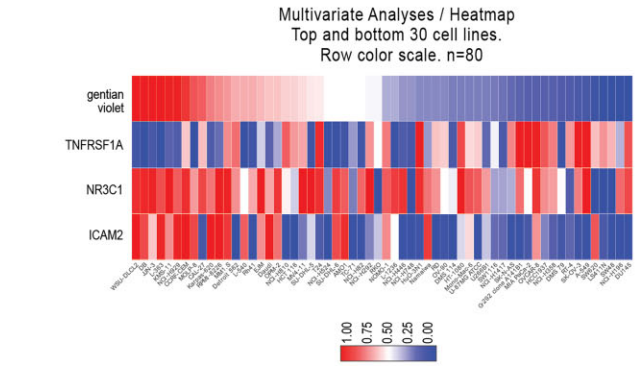


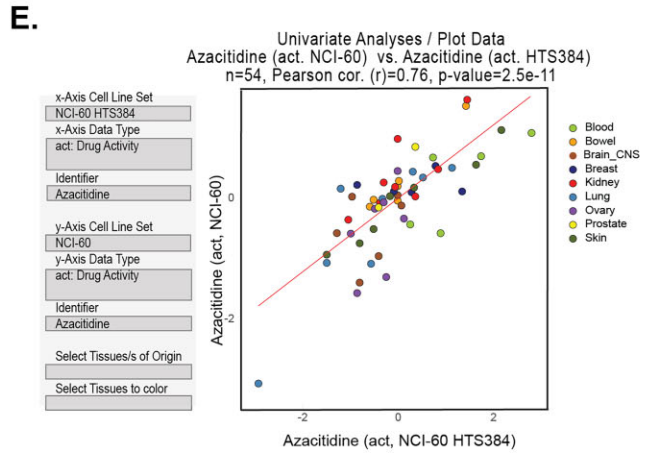
B.

NCATS drug activities versus drug target expression correlations

Drug			Activity vs transcript expression		
Name	Mech. of action	Use	Gene	Correlation	p value
jnj-28312141	CSF1R inhibitor	cancer	CSF1R	0.54	2.1e-5
dabigatran etex.	F2 inhibitor	anticoagulant	F2	0.77	6.1e-10
gentian violet	TNFRSF1A mod.	antiseptic	TNFRSF1A	-0.40	2.3e-4
amg-232	MDM2 inhibitor	cancer	MDM2	0.70	0.001
ast-487	FLT3 inhibitor	cancer	FLT3	0.60	0.002
budesonide	NR3C1 ant.	anti-inflamatory	NR3C1	0.55	0.002
iloperidone	DRD2 ant.	schizophrenia	DRD2	0.59	0.004
neratinib	EGFR inhibitor	cancer	EGFR	0.34	0.004
sar-405838	MDM2 inhibitor	cancer	MDM2	0.59	0.004
pictilisib	PIK3CG inhibitor	cancer	PIK3CG	0.36	0.004







F.

HTS384 (ATP) versus NCI-60 (protein) drug activity correlations

Drug		Activity vs activity		
Name	Mechanism of action	Correlation	p value	n
Melphalan	Alkalating agent	0.821	6.1e-14	54
Mithramycin	DNA binding / RNA syn. inh.	0.93	4.8e-25	54
Chelerythrine	BCL2 inhibitor	0.69	6.0e-8	47
Cytarabine	DNA syn. inhibitor	0.72	3.3e-9	51
Vincristine	Tubulin fragmenter	0.29	0.039	51
Daunorubicin	TOP2 inhibitor	0.79	1.4e-12	54
Wortmannin	PIK3 inhibitor	0.68	6.6e-9	56
Irinotecan	MTOR inhibitor	0.65	1.1e-7	53
Vorinostat	Histone deacetylase	0.74	8.8e-11	56
Erlotinib	EGFR inhibitor	0.86	1.4e-17	56

Figure 5. Examples of correlations between drug activity and drug targets. (**A**) Univariate scatter plot of PIK3CG (Phosphoinositide-3-Kinase Gamma Catalytic Subunit) transcript expression (log2) in GDSC-MGH-Sanger (*x*-axis) versus IC50 drug activity of nvp-bgt226 in NCATS (*y*-axis) for 78 overlapping cell lines. The red line is the regression line, and each circle is a cell line. (**B**) Additional examples of significantly correlated and biologically linked NCATS IC50 drug activities versus GDSC transcript expression levels. (**C**) Multivariate analysis predicting the activity of gentian violet versus expression of three inflammation/immune response genes (TNFRSSF1A, NR3C1, and ICAM2). The predicted and observed activity of gentian violet is on the *x*-axis and *y*-axis, respectively. The plots were created using the CellMinerCDB\Multivariate Analyses\Plot Data tab selections with the specific inputs detailed in the input boxes to the left. (**D**) Heatmap of gentian violet activity as the response variable with transcript expression of the predictors (*y*-axis). The top and bottom 30 cell lines are displayed. (**E**) Scatter plot of NCI-60 drug activity as measured by ATP levels versus total protein. The red line is the regression line. Each circle is a cell line. (**F**) Additional examples of 10 significantly correlated drug activities as measured by ATP levels versus total protein. All drug mechanisms of action are unique. Included are eight FDA-approved and two clinical trial drugs.

Next, we highlight the multivariate analysis features of CellMinerCDB. Fig. 5C and D (10x cross-validation and heatmap, respectively) characterize the multivariate analysis prediction of the antiseptic gentian violet as the response variable, as predicted by three genes: TNFRSF1A, NR3C1, and ICAM2. Note that for either standard linear regression models on CellMinerCDB, 10-fold cross-validation is applied to fit model coefficients and predict response, while withholding portions of the data to better estimate robustness; more methodological details are available in our previous publication [9]. Using transcript expression, each of the three genes significantly adds predictive capability to the model. The first is the known drug target, TNFRSF1A, and the second and third variables, NR3C1 and ICAM2, are inflammation and immune response genes, respectively; each adds to the drug activity prediction significantly (i.e. as indicated by p-value). The combination of the three provides a more accurate prediction of the drug activity signature than any of the genes individually. The p-value for the combined model utilizing the three genes TNFRSF1A, NR3C1, ICAM2 is 1.3e-10 shown in Fig. 5C. The individual correlation of TNFRSF1A to gentian violet is *P*-value 2.3e-4 (shown in Fig. 5B); the correlation p-values of gentian violet with NR3C1 is 4.8e-3 and 8.0e-7 for ICAM2, respectively (not shown but accessible via the Univariate Analysis tab); this comparison uses the GDSC microarray gene transcript measurements. Similar to the previous example, this example makes use of our newly added NCATS drug activity data for this CellMinerCDB version.

With Fig. 5E and F, we highlight another of our dataset additions: the HTS384 NCI-60 drug screening dataset, which now uses the ATP (by CellTiter Glo) viability assay. This is an updated protocol from the NCI over the previous protein level (by sulforhodamine B) based measurement. With Fig. 5E, we show the reproducibility of the drug responses for azacitidine (also spelled as azacytidine) as measured by either ATP or protein levels on NCI-60 cell lines; the correlation for azacitidine is 0.76. Fig. 5F provides a similar comparison for 11 other cancer FDA-approved or clinical trial drug activities as measured by either CellTiter-Glo or sulforhodamine B on the same NCI-60 cell lines. Patterns of protein and ATP-measured viability levels are correlated with one another (470 compounds correlate at 0.6 or better) for these two screening protocols performed on the NCI-60 [5]. These figures illustrate how users can easily compare different drug assay types to determine comparability or lack thereof. This process can also be extended to drug measurements from different institutes.

To showcase recent profiling data additions to CellMinerCDB, Fig. 6A focuses on histone modification (methylation and acetylation status) profiling conducted on the NCI-60 [20]. Histone modifications are critical regulators of epithelial to mesenchymal transition (EMT) [31, 32]. Specifically, histone H3 lysine 27 acetylation (H3K27ac)-marked enhancers are essential for the EMT process [33] and H3K4 trimethylation at EMT-related transcription factor (TFs) promoters enhances these TFs expression via positive feedback loops [34]. As shown in Fig. 6A, a relationship is observed between CDH1 transcript expression and two epigenetic modifications, H3K4me3 (trimethylation) and H3K27ac (acetylation), which are indicative of transcriptional activation. Using linear regression, each of these has a contribution *P*-value $\leq$ 0.0001. These p-values are from individual tests of each regression coefficient, assessing whether a specific predictor has a significant effect on the outcome after accounting for the other variables in the model. The multivariate prediction of CDH1 expression is better than either of the individual predictions using the H3K4me3 or H3K27ac modifications alone. Fig. 6B provides ten analogous examples of epithelial-mesenchymal transition genes with significant correlations to their histone methylation and acetylation status [35]. In all cases, the combination of the two provides a more accurate prediction than either of the two individually, as indicated by the significant p-values for the combined predictive variable. The epithelial-mesenchymal transition genes included in Fig. 6A and B have been defined previously as part of the 38 EMT gene signature [9, 35].

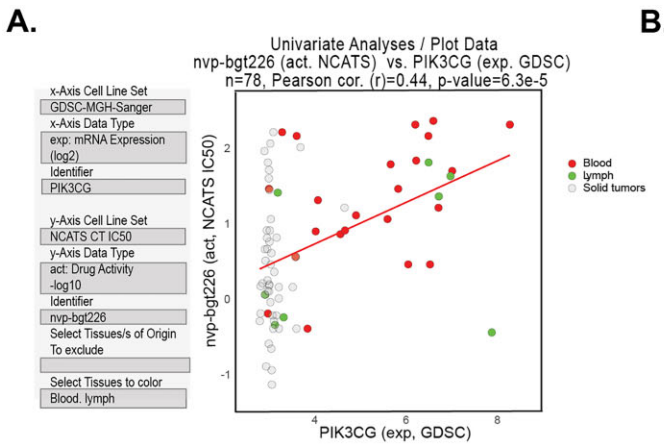
Our last use case highlights the use of multi-gene signatures in CellMinerCDB and how multivariate analysis can be used to expand genes included in pre-existing signatures to understand other biologically related genes. Fig. 6C and D utilize our 18-gene signature for replication stress (RepStress) [22] as the response variable. Each of the four predictive variable genes adds to the predictive strength of the model with a *P*-value of $\leq$0.0001 and was chosen due to the involvement of the gene in either the process of DNA damage response or replication. The combination of the 4 genes provides a more accurate prediction of the RepStress signature than any of the 4 genes individually. Both the RNAseq expression data from GDSC, used in Fig. 6C and D, and the RepStress signature are new updates in the most recent CellMinerCDB version.

Fig. 6C and D again illustrate how users may explore signatures that are precomputed within the updated CellMiner web application. The replication stress (RepStress) gene expression signature is significantly associated with the outcome of ATR inhibition, which enhances the activity of clinical DNA-damaging agents commonly used in cancer treatment [36–38].

Conclusion

Cell lines remain major experimental models used in laboratories and drug companies around the world. The multiple cancer cell lines included in CellMinerCDB continue to be mentioned in tens of thousands of publications per year (Fig. 1B). The wide use of cell lines stems from their extensive characterization and relative ease of use; this is a trend we expect to persist even as newer experimental models such as organoids and patient-derived xenografts (PDXs) gain use. With CellMinerCDB, we have sought to make large systematic data collections for these cell lines accessible to researchers, continuing efforts pioneered by the NIH initially through the NCI-60 cell line collection since the 1990s [4, 39].

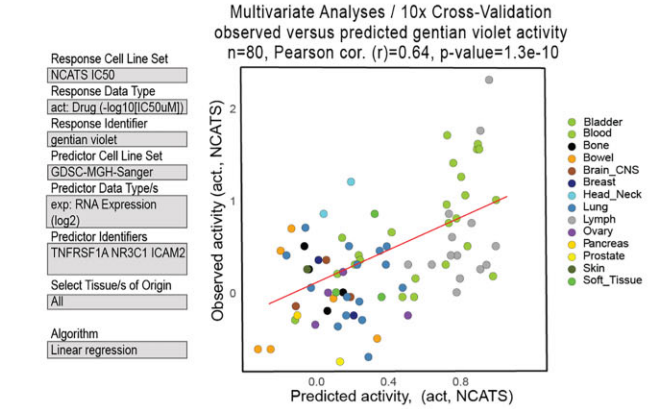
Use of cell lines by the research community has seen ongoing developments. The past decade has seen major advancements in the efforts to efficiently, systematically, and reproducibly characterize cell lines, advancements captured in this release of CellminerCDB. For example, we see community-wide trends emerging, such as the use of the ATP-based CellTiter Glo assay with major data-generating institutes, such as the NCI, NCATS, Broad, and Sanger institutes. As such, we see some convergence with these groups rather than continuing with their previously varied methods that included sulforhodamine B, Syto60, and rezaurin assays. But we also see developments to increase the efficiency of drug screening through the PRISM platform, which makes use of DNA

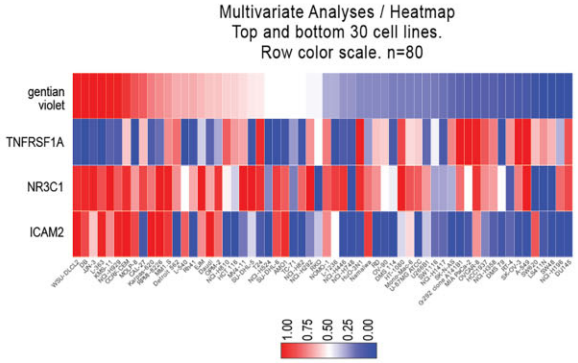


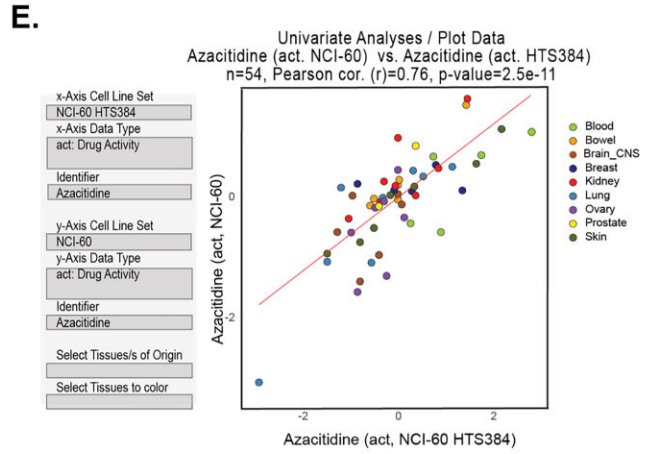
B.

NCATS drug activities versus drug target expression correlations

Drug			Activity vs transcript expression		
Name	Mech. of action	Use	Gene	Correlation	p value
jnj-28312141	CSF1R inhibitor	cancer	CSF1R	0.54	2.1e-5
dabigatran etex.	F2 inhibitor	anticoagulant	F2	0.77	6.1e-10
gentian violet	TNFRSF1A mod.	antiseptic	TNFRSF1A	-0.40	2.3e-4
amg-232	MDM2 inhibitor	cancer	MDM2	0.70	0.001
ast-487	FLT3 inhibitor	cancer	FLT3	0.60	0.002
budesonide	NR3C1 ant.	anti-inflamatory	NR3C1	0.55	0.002
iloperidone	DRD2 ant.	schizophrenia	DRD2	0.59	0.004
neratinib	EGFR inhibitor	cancer	EGFR	0.34	0.004
sar-405838	MDM2 inhibitor	cancer	MDM2	0.59	0.004
pictilisib	PIK3CG inhibitor	cancer	PIK3CG	0.36	0.004







F.

HTS384 (ATP) versus NCI-60 (protein) drug activity correlations

Drug		Activity vs activity		
Name	Mechanism of action	Correlation	p value	n
Melphalan	Alkalating agent	0.821	6.1e-14	54
Mithramycin	DNA binding / RNA syn. inh.	0.93	4.8e-25	54
Chelerythrine	BCL2 inhibitor	0.69	6.0e-8	47
Cytarabine	DNA syn. inhibitor	0.72	3.3e-9	51
Vincristine	Tubulin fragmenter	0.29	0.039	51
Daunorubicin	TOP2 inhibitor	0.79	1.4e-12	54
Wortmannin	PIK3 inhibitor	0.68	6.6e-9	56
Irinotecan	MTOR inhibitor	0.65	1.1e-7	53
Vorinostat	Histone deacetylase	0.74	8.8e-11	56
Erlotinib	EGFR inhibitor	0.86	1.4e-17	56

Figure 6. Multivariate analysis of predicted versus observed epithelial-mesenchymal transition (EMT) gene expression. (**A**) Multivariate analysis for CDH1 transcript expression as the response variable and histone H3K4me3 (trimethylation; "hs4" data prefix) and H3K27ac (acetylation; "his" data prefix) as predictor variables. The circles are cell lines colored by tissue of origin (right). The red line is the regression line. The plots were created using the CellMinerCDB\Multivariate Analyses\Plot Data tab selections with the specific inputs used, as detailed in the input boxes to the left. (**B**) Additional examples of multivariate predictions for significant correlations between EMT genes vs. histone H3K4me3 (methylation) and H3K27ac (acetylation). H3K4me3 is the tri-methylation of histone 3 at lysine 4. H3K27ac is the acetylation of histone 3 on lysine 27. (**C**) Heatmap of replication stress (RepStress) as the response variable with transcript expression of four DNA damage response or replication genes as the predictors. The top and bottom 30 cell lines are included. Red indicates high expression and blue low. (**D**) Multivariate analysis 10x cross-validation for the RepStress signature as the response variable and transcript expression of the four genes from panel C as predictor variables. In the scatter plot**,** the predicted RepStress signature is on the *x*-axis and the observed on the y-axis. The circles are cell lines colored by tissue of origin (right). The red line is the regression line. The plots were created using the CellMinerCDB\Multivariate Analyses\Plot Data tab selections with the specific inputs detailed in the input boxes to the left.

barcodes to allow screening of single drug agents on multiple cell lines at once.

Here, we illustrate use cases using the additional datasets that we have incorporated since our last publication to focus on drug exploration of non-cancer drugs through datasets from the NCATS and Broad PRISM. Such work can yield interesting results, for example, our recent publications show the repurposing potential of two non-cancer drugs, oxyphenisatin acetate (acetalax) and bisacodyl, for use in triple-negative breast cancer that has since been replicated in PDX models [40, 41]. Other major use cases, in general, for CellMinerCDB include: (i) data reproducibility; (ii) candidate biomarker discovery; (iii) prediction of molecular features indicative of drug activity; (iv) exploration of protein complexes, genetic dependencies, and biological pathways; and (v) cross-dataset analyses.

Ongoing developments in the areas of machine learning and artificial intelligence depend on well-annotated datasets with common representations, and we attempt to simplify the use of our datasets for these types of studies. Additionally, the core software infrastructure of CellMinerCDB has been well-tested with almost a decade of use, which has allowed the infrastructure to be used in several dedicated type-specific cancer portals that are useful for specific cancer disease communities, including sarcoma, adrenocortical carcinoma, and small cell lung cancer [6, 12, 13].

For future work, CellMinerCDB will continue to integrate additional datasets while making updates to existing datasets, refining annotations as well, and making visual presentation modifications to improve user experience. We plan to add new functionalities, including a module for differentially expressed genes and pathway enrichment. Also, owing to the recent advancements in the area of artificial intelligence (AI) (specifically large language models, LLMs), we hope to bring some of this technology to CellMinerCDB users to help expand (i) analyses available to CellMinerCDB, (ii) customization for data visualization, and (iii) potentially, curation of datasets [42, 43].

Software availability

Archived versions of this code are available on Zenodo at the following link (https://zenodo.org/records/15150484), and CellMinerCDB data packages are available at . All software developed as part of CellMinerCDB is freely available, open-source, and developed on the GitHub collaboration platform in the following repositories: https://github.com/CBIIT/cellminercdb, https://github.com/CBIIT/rcellminerUtilsCDB, and https://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. https://github.com/CBIIT/cellminercdb/issues).

Acknowledgements

We thank external data provider teams that have made this work possible.

Author contributions: F.E. (Conceptualization [lead], Data curation [supporting], Methodology [lead], Software [lead], Writing – original draft [equal], Writing – review & editing [equal]), W.R. (Data curation [supporting], Supervision [equal], Writing – original draft [equal], Writing – review & editing [equal]), S.V. (Data curation [lead], Methodology [supporting]), Y.W. (Software [supporting]), M.K. (Data curation [supporting], Writing – review & editing [supporting]), Y.A. (Data curation [supporting]), Y.I. (Data curation [supporting]), M.A. (Writing – review & editing [equal]), Y.P. (Funding acquisition [lead], Supervision [equal], Writing – review & editing [equal]), A.L. (Conceptualization [equal], Data curation [equal], Funding acquisition [supporting], Software [supporting], Supervision [equal], Writing – original draft [equal], Writing – review & editing [equal]).

Conflict of interest

The authors declare no competing interests.

Funding

Our work is funded by the Center for Cancer Research, the Intramural Program of the National Cancer Institute, NIH (Z01-BC 006150). In addition, this work is supported by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM) (ZIALM240126). Funding to pay the Open Access publication charges for this article was provided by NLM.

Data availability

Users can download all data of CellMinerCDB (https://discover.nci.nih.gov/cellminercdb) using the Metadata tab as compressed (zip) files. Additionally, users can download drug synonym information or cell line annotations from this tab. The NCI-60 dataset is available for download via the rcellminerData Bioconductor package. Separately, we have provided CellMinerCDB datasets as R data packages on Zenodo (https://zenodo.org/records/17088217). These datasets are structured in an equivalent manner to the rcellminerData data package [8].

References

1. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov* 2022;12:31–46. https://doi.org/10.1158/2159-8290.CD-21-1059
2. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000;100:57–70. https://doi.org/10.1016/S0092-8674(00)81683-9
3. Weinstein JN, Myers TG, O'Connor PM *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997;275:343–9. https://doi.org/10.1126/science.275.5298.343
4. Paull KD, Hamel E, Malpeis L. In: Foye WO (ed.), *Cancer Chemotherapeutic Agents*. ACS, Washington, DC, 1995, 8–45.
5. Kunkel MW, Coussens NP, Morris J *et al.* HTS384 NCI60: the next phase of the NCI60 screen. *Cancer Res* 2024;84:2403–16. https://doi.org/10.1158/0008-5472.CAN-23-3031
6. Arakawa Y, Elloumi F, Varma S *et al.* A database tool integrating genomic and pharmacologic data from adrenocortical carcinoma cell lines, PDX, and patient samples. *Cancer Res Commun* 2024;4:2384–98. https://doi.org/10.1158/2767-9764.CRC-24-0100
7. Luna A, Elloumi F, Varma S *et al.* CellMiner Cross-Database (CellMinerCDB) version 1.2: exploration of patient-derived cancer

cell line pharmacogenomics. *Nucleic Acids Res* 2021;49:D1083–93. https://doi.org/10.1093/nar/gkaa968
8. Luna A, Rajapakse VN, Sousa FG *et al.* rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. *Bioinformatics* 2016;32:1272–4. https://doi.org/10.1093/bioinformatics/btv701
9. Rajapakse VN, Luna A, Yamade M *et al.* CellMinerCDB for integrative cross-database genomics and pharmacogenomics analyses of cancer cell lines. *iScience* 2018;10:247–64. https://doi.org/10.1016/j.isci.2018.11.029
10. Reinhold WC, Sunshine M, Liu H *et al.* CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res* 2012;72:3499–511. https://doi.org/10.1158/0008-5472.CAN-12-1370
11. Reinhold WC, Wilson K, Elloumi F *et al.* CellMinerCDB: NCATS is a web-based portal integrating public cancer cell line databases for pharmacogenomic explorations. *Cancer Res* 2023;83:1941–52. https://doi.org/10.1158/0008-5472.CAN-22-2996
12. Tlemsani C, Heske CM, Elloumi F *et al.* Sarcoma_CellminerCDB: a tool to interrogate the genomic and functional characteristics of a comprehensive collection of sarcoma cell lines. *iScience* 2024;27:109781. https://doi.org/10.1016/j.isci.2024.109781
13. Tlemsani C, Pongor L, Elloumi F *et al.* SCLC-CellMiner: a resource for small cell lung cancer cell line genomics and pharmacology based on genomic signatures. *Cell Rep* 2020;33:108296. https://doi.org/10.1016/j.celrep.2020.108296
14. Wei CH, Allot A, Lai PT *et al.* PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Res* 2024;52:W540–6. https://doi.org/10.1093/nar/gkae235
15. Heumos S, Dehn S, Brautigam K *et al.* Multiomics surface receptor profiling of the NCI-60 tumor cell panel uncovers novel theranostics for cancer immunotherapy. *Cancer Cell Int* 2022;22:311. https://doi.org/10.1186/s12935-022-02710-y
16. Yu C, Mannan AM, Yvone GM *et al.* High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol* 2016;34:419–23. https://doi.org/10.1038/nbt.3460
17. Iorio F, Knijnenburg TA, Vis DJ *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;166:740–54. https://doi.org/10.1016/j.cell.2016.06.017
18. Yang W, Soares J, Greninger P *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013;41:D955–61. https://doi.org/10.1093/nar/gks1111
19. Garnett MJ, Edelman EJ, Heidorn SJ *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;483:570–5. https://doi.org/10.1038/nature11005
20. Gopi LK, Kidder BL. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. *Nat Commun* 2021;12:1419. https://doi.org/10.1038/s41467-021-21707-1
21. Pongor LS, Tlemsani C, Elloumi F *et al.* Integrative epigenomic analyses of small cell lung cancer cells demonstrates the clinical translational relevance of gene body methylation. *iScience* 2022;25:105338. https://doi.org/10.1016/j.isci.2022.105338
22. Takahashi N, Kim S, Schultz CW *et al.* Replication stress defines distinct molecular subtypes across cancers. *Cancer Res Commun* 2022;2:503–17. https://doi.org/10.1158/2767-9764.CRC-22-0168
23. Zhang W, Mao JH, Zhu W *et al.* Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nat Commun* 2016;7:12619. https://doi.org/10.1038/ncomms12619
24. Safran M, Dalah I, Alexander J *et al.* GeneCards Version 3: the human gene integrator. *Database* 2010;2010:baq020. https://doi.org/10.1093/database/baq020
25. Rodchenkov I, Babur O, Luna A *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res* 2020;48:D489–97.
26. Elloumi F, Tlemsani C, Heske CM *et al.* Protocol to set up the CellMinerCDB pharmacogenomics analysis web application with custom cell line data. *STAR Protocols* 2025;6:104076. https://doi.org/10.1016/j.xpro.2025.104076
27. Sinha R, Luna A, Schultz N *et al.* A pan-cancer survey of cell line tumor similarity by feature-weighted molecular profiles. *Cell Rep Methods* 2021;1:100039. https://doi.org/10.1016/j.crmeth.2021.100039
28. Inoue Y, Lee H, Fu T *et al.* drGAT: attention-guided gene assessment of drug response utilizing a drug-cell-gene heterogeneous network. arXiv, https://arxiv.org/abs/2405.08979v2, 14 May 2024, preprint: not peer reviewed.
29. Waas M, Snarrenberg ST, Littrell J *et al.* SurfaceGenie: a web-based application for prioritizing cell-type-specific marker candidates. *Bioinformatics* 2020;36:3447–56. https://doi.org/10.1093/bioinformatics/btaa092
30. Simioni C, Cani A, Martelli AM *et al.* The novel dual PI3K/mTOR inhibitor NVP-BGT226 displays cytotoxic activity in both normoxic and hypoxic hepatocarcinoma cells. *Oncotarget* 2015;6:17147–60. https://doi.org/10.18632/oncotarget.3940
31. Noguchi S, Eitoku M, Moriya S *et al.* Regulation of gene expression by sodium valproate in epithelial-to-mesenchymal transition. *Lung* 2015;193:691–700. https://doi.org/10.1007/s00408-015-9776-9
32. Li D, Sun H, Sun WJ *et al.* Role of RbBP5 and H3K4me3 in the vicinity of Snail transcription start site during epithelial-mesenchymal transition in prostate cancer cell. *Oncotarget* 2016;7:65553–67. https://doi.org/10.18632/oncotarget.11549
33. Qiao Y, Wang Z, Tan F *et al.* Enhancer reprogramming within pre-existing topologically associated domains promotes TGF-beta-induced EMT and cancer metastasis. *Mol Ther* 2020;28:2083–95. https://doi.org/10.1016/j.ymthe.2020.05.026
34. Lindner P, Paul S, Eckstein M *et al.* EMT transcription factor ZEB1 alters the epigenetic landscape of colorectal cancer cells. *Cell Death Dis* 2020;11:147. https://doi.org/10.1038/s41419-020-2340-4
35. Kohn KW, Zeeberg BM, Reinhold WC *et al.* Gene expression correlations in human cancer cell lines define molecular interaction networks for epithelial phenotype. *PLoS One* 2014;9:e99269. https://doi.org/10.1371/journal.pone.0099269
36. Jo U, Arakawa Y, Zimmermann A *et al.* The novel ATR inhibitor M1774 induces replication protein overexpression and broad synergy with DNA-targeted anticancer drugs. *Mol Cancer Ther* 2024;23:911–23,https://doi.org/10.1158/1535-7163.MCT-23-0402
37. Taniguchi H, Chakraborty S, Takahashi N *et al.* ATR inhibition activates cancer cell cGAS/STING-interferon signaling and promotes antitumor immunity in small-cell lung cancer. *Sci. Adv.* 2024;10:eado4618. https://doi.org/10.1126/sciadv.ado4618
38. Abel ML, Takahashi N, Peer C *et al.* Targeting replication stress and chemotherapy resistance with a combination of sacituzumab govitecan and berzosertib: a phase I clinical trial. *Clin Cancer Res* 2023;29:3603–11. https://doi.org/10.1158/1078-0432.CCR-23-0536
39. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer* 2006;6:813–23. https://doi.org/10.1038/nrc1951
40. Mizunuma M, Redon CE, Saha LK *et al.* Acetalax (Oxyphenisatin Acetate, NSC 59687) and bisacodyl cause oncosis in triple-negative breast cancer cell lines by poisoning the ion exchange membrane protein TRPM4. *Cancer Res Commun* 2024;4:2101–11. https://doi.org/10.1158/2767-9764.CRC-24-0093

41. Reinhold WC, Marangoni E, Elloumi F *et al.* Acetalax and bisacodyl for the treatment of triple-negative breast cancer: a combined molecular and preclinical study. *Cancer Res Commun* 2025;5:375–88. https://doi.org/10.1158/2767-9764.CRC-24-0435
42. Liu W, Li J, Tang Y *et al.* DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis. *Nat Commun* 2025;16:2256. https://doi.org/10.1038/s41467-025-57430-4
43. Tiwari K, Matthews L, May B *et al.* ChatGPT usage in the Reactome curation process. *bioRxiv*, https://doi.org/10.1101/2023.11.08.566195, 8 November 2023, preprint: not peer reviewed.

Received: September 17, 2025. **Revised:** October 10, 2025. **Accepted:** October 13, 2025
Published by Oxford University Press 2025.