Supplementary Figure 1: Comparison of CellMinerCDB to other pharmacogenomics tools in terms of datasets and features.

Each of the compared projects has unique attributes, datasets, and objectives. The compared attributes below tend to be of a general nature. We endeavored to find at least a single example of any particular attribute to assign that a project possessed the characteristic (a green checkmark) in the absence of a characteristic a red "X" is given; dashes are used for entries that were not applicable (e.g., GDSCTools is a software package and not a web-based application). This should be emphasized especially given that for many of these projects aggregate collections of datasets and only some of the datasets might contain a particular characteristic.

	CellMinerCDB	DepMap	PharmacoDB	cBioPortal	CMap	CancerRxGene	GDA	CancerResource	SYNERGxDB	DrugComb	DrugCombDB	GDSCTools	PharmacoGx
COMMON CELL LINE DATASETS													
CCLE	✓	✓	✓	✓	✓	✖	✓	✓	✖	✓	✖	✖	✓
NCI-60	✓	✖	✖	✓	✖	✖	✓	✓	✓	✓	✓	✖	✖
GDSC	✓	✓	✓	✖	✖	✓	✖	✓	✖	✖	✖	✓	✓
Other cell line datasets	✓	✓*	✓	✖	✓	✖	✖	✖	✓	✓	✖	✖	✓
OMIC DATA TYPES													
Expression	✓	✓	✓	✓	✓	✓	✓	✓	✓	✖	✖	✓	✓
Mutation	✓	✓	✓	✓	✖	✓	✓	✓	✖	✖	✖	✓	✓
Copy Number	✓	✓	✓	✓	✖	✓	✖	✖	✓	✖	✖	✓	✓
Methylation	✓	✓	✓	✓	✖	✓	✖	✖	✖	✖	✖	✓	✓
Proteomics	✓	✓	✓	✓	✓	✖	✖	✖	✖	✖	✖	✖	✓
MicroRNA	✓	✓	✖	✓	✖	✖	✖	✖	✖	✖	✖	✖	✖
Metabolomics	✓	✓	✖	✖	✖	✖	✖	✖	✓	✖	✖	✖	✖
Gene Dependency	✓	✓	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖
Phenotypic Data	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DRUG DATA													
Single Drug Responses	✓	✓	✓	✖	✓	✓	✓	✓	✖	✓	✓	✓	✓
Drug Combination Responses	✓	✖	✖	✖	✓	✖	✖	✖	✓	✓	✓	✖	✖
Drug Mechanism Information	✓	✓	✓	✓	✓	✓	✓	✓	✖	✖	✓	✖	✓
Availability of Structures Information	✓†	✓	✓	✖	✖	✖	✖	✓	✓	✓	✓	✖	✓
FDA Approval Information	✓	✓	✓	✓	✖	✖	✖	✖	✖	✖	✓	✖	✓
WEB-BASED FEATURES													
Web-Based Analysis Tool	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	—
Cross-Database Analysis	✓	✓	✓	✓	✓	✖	✖	✓	✓	✓	✖	—	—
Univariate Analysis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	—
Multivariate Analysis	✓	✖	✖	✖	✓	✖	✓	✓	✖	✖	✖	—	—
Not Restricted to Pre-Computed Results	✓	✓	✖	✓	✓	✖	✓	✓	✖	✓	✖	—	—
Filter Analysis by Tissue	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	—
Search by Drug Synonyms	✓	✖	✓	✖	✓	✖	✖	✓	✖	✖	✓	—	—
Search by Gene Synonyms	✓	✓	✖	✓	✖	✖	✖	✖	✖	✖	—	—	—
Data Plots	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	—
Download Data for Query Results	✓	✓	✓	✓	✓	✓	✓	✖	✓	✓	✖	—	—
Download Button for Plots	✓	✓	✓	✓	✓	✖	✓	✖	✖	✖	✓	—	—
SOFTWARE													
Official Programmatic Interface	✓‡	✓§	✓	✓	✓	✓	✖	✖	✓	✓	✓	✓	✓
Analysis Website Open-Source	✓	✖	✓	✓	✖	✓¶	✖	✖	✓	✖	✖	—	—
On-Site Version (or Last Update) Information	✓	✓	✓	✓	✓	✓	✖	✓	✓	✓	✓	✓	✓

Footnotes

* Other cell lines beyond CCLE/GDSC are included, but not other "sets"

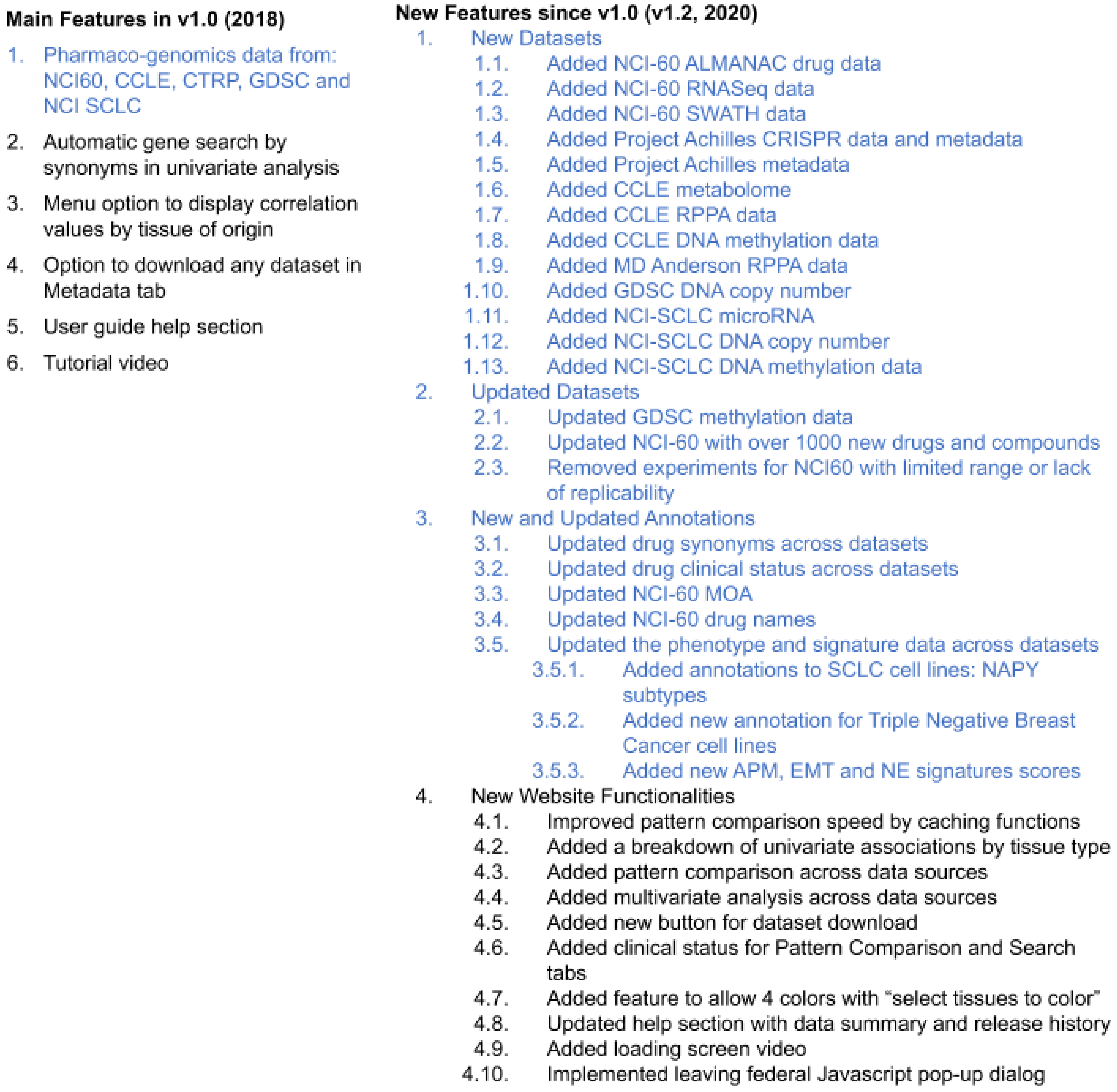
† Only for the NCI-60 via the programmatic interface

‡ Only allows access of NCI-60 data

§ An unofficial R interface exists

¶ R Shiny and Python Dash explorer code exists

Supplementary Figure 2



Main Features in v1.0 (2018)

1. Pharmaco-genomics data from: NCI60, CCLE, CTRP, GDSC and NCI SCLC
2. Automatic gene search by synonyms in univariate analysis
3. Menu option to display correlation values by tissue of origin
4. Option to download any dataset in Metadata tab
5. User guide help section
6. Tutorial video

New Features since v1.0 (v1.2, 2020)

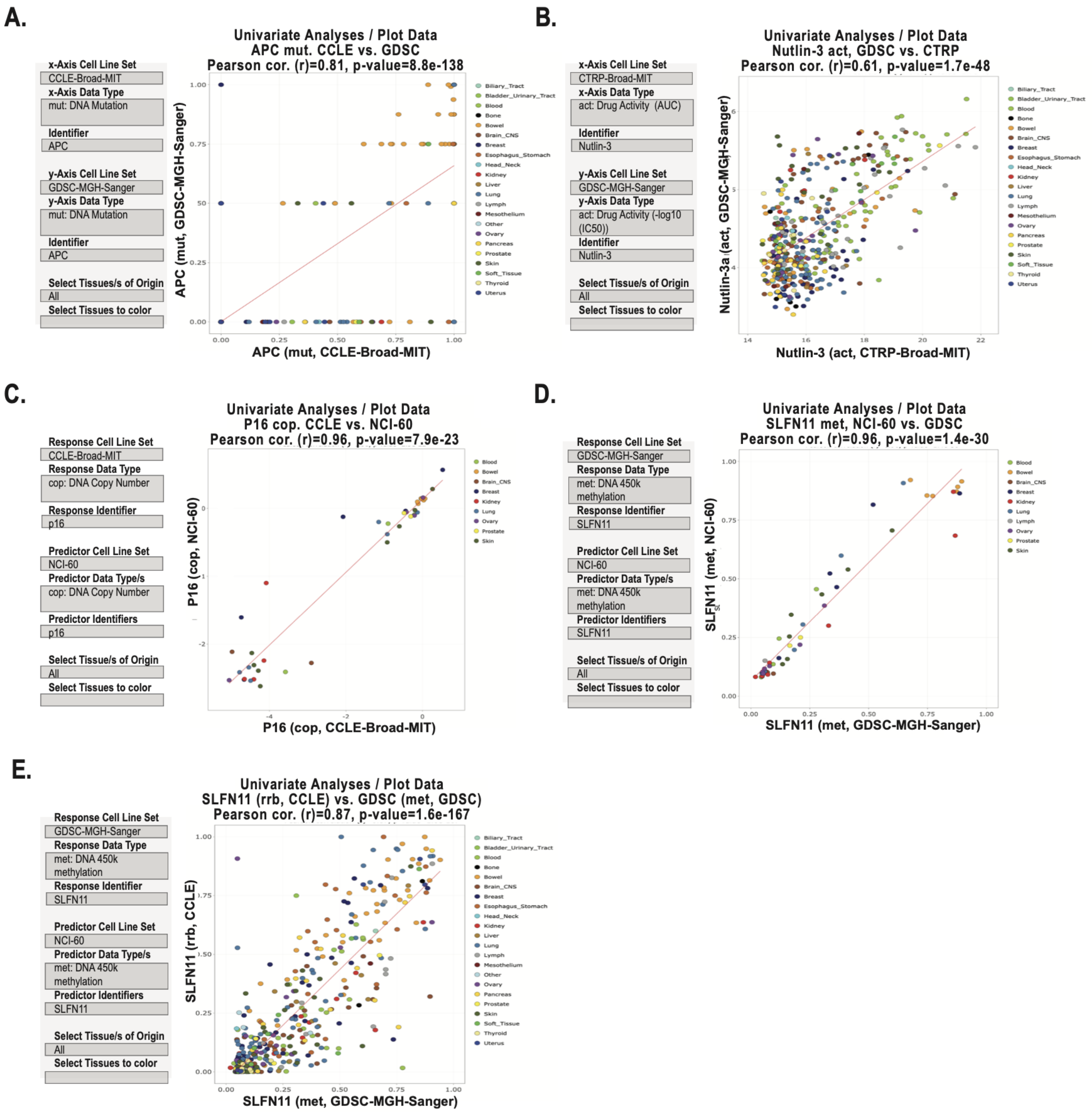
1. New Datasets
    - 1.1. Added NCI-60 ALMANAC drug data
    - 1.2. Added NCI-60 RNASeq data
    - 1.3. Added NCI-60 SWATH data
    - 1.4. Added Project Achilles CRISPR data and metadata
    - 1.5. Added Project Achilles metadata
    - 1.6. Added CCLE metabolome
    - 1.7. Added CCLE RPPA data
    - 1.8. Added CCLE DNA methylation data
    - 1.9. Added MD Anderson RPPA data
    - 1.10. Added GDSC DNA copy number
    - 1.11. Added NCI-SCLC microRNA
    - 1.12. Added NCI-SCLC DNA copy number
    - 1.13. Added NCI-SCLC DNA methylation data
2. Updated Datasets
    - 2.1. Updated GDSC methylation data
    - 2.2. Updated NCI-60 with over 1000 new drugs and compounds
    - 2.3. Removed experiments for NCI60 with limited range or lack of replicability
3. New and Updated Annotations
    - 3.1. Updated drug synonyms across datasets
    - 3.2. Updated drug clinical status across datasets
    - 3.3. Updated NCI-60 MOA
    - 3.4. Updated NCI-60 drug names
    - 3.5. Updated the phenotype and signature data across datasets
        - 3.5.1. Added annotations to SCLC cell lines: NAPY subtypes
        - 3.5.2. Added new annotation for Triple Negative Breast Cancer cell lines
        - 3.5.3. Added new APM, EMT and NE signatures scores
4. New Website Functionalities
    - 4.1. Improved pattern comparison speed by caching functions
    - 4.2. Added a breakdown of univariate associations by tissue type
    - 4.3. Added pattern comparison across data sources
    - 4.4. Added multivariate analysis across data sources
    - 4.5. Added new button for dataset download
    - 4.6. Added clinical status for Pattern Comparison and Search tabs
    - 4.7. Added feature to allow 4 colors with “select tissues to color”
    - 4.8. Updated help section with data summary and release history
    - 4.9. Added loading screen video
    - 4.10. Implemented leaving federal Javascript pop-up dialog

Supplementary Figure 3

CelMinerCDB Explores & Validates	Main Steps	Examples of Findings
Cell line reproducibility, & consistency	Univariate Analyses: Plot Data: Expression of the same gene across different datasets (X & Y)	Cell lines are highly reproducible across datasets
Omic data robustness & reproducibility	Univariate Analyses: Plot Data: Expression, copy number variation, promoter methylation, mutations for the same gene across datasets (X & Y)	Transcripts, promoter methylation, gene copy number are highly reproducible across datasets
Drug data robustness & reproducibility	Univariate Analyses: Plot Data: Activity of the same drug across datasets (X & Y)	Warning: Not all drugs are consistent across dataset
Select and compare subsets of cell lines based or tissue of origin or metadata: Breast, Kidney, Lung	Univariate Analyses: select Y axis: Select Tissue/s of Origin or Select Tissues to color (Breast, Kidney, Lung)	Genes are also selectively expressed in particular cancer cell lines
Test Phenotypic data (mda): NE, APM, EMT	Univariate Analyses: select Data Type mda: NE, APM, EMT. Additional selection can be done for subset	Cell lines have low Antigen Presenting Machinery score (APM)
Epigenetics: promoter methylation for any given gene	Univariate analyses: Plot Data: Expression of a given gene vs its methylation (X & Y Data Type) within a given Cell Line Set or across datasets (independent datasets can be tested for missing Data Type and confirmation)	Promoter methylation is a driver for gene expression (SLFN11; MGMT)
Gene amplification and deletions for any given gene	Univariate analyses: Plot Data: Expression of a given gene vs copy number (X & Y Data Type) within a given Cell Line Set or across datasets (independent datasets can be tested for validation and missing Data Type)	MYC genes and other oncogenes are often driven by copy number variation (CNV)
Integrate and complement different datasets for common cell lines	Univariate Analyses: Plot Data: Plot different parameters (Data Type for genomic or drug response) across Cell Line Sets (X & Y) to counter missing data in one dataset	Drug response data in one dataset can be correlated with genomics of another dataset
Genomic pathway discovery (coregulated genes and microRNAs)	Univariate analyses: Plot Data: expression of a given gene (X or Y Data Type) within a given dataset or across datasets; also use the Compare Patterns tab.	Genes that that are coexpressed with the input genes
Discover determinants of drug response and targeted drug delivery	Univariate Analyses: Plot Data: Compare Patterns: Coregulated genes for a given gene (X or Y) within a given dataset (independent datasets can be tested for confirmation)	Resistance of cell lines to chemotherapy and potential response to kinase inhibitors
Validate genomic determinant of drug response	Univariate Analyses: Plot Data: Compare Patterns: plot genomic parameter vs drug (X or Y Data Type)	Validation of SLFN11 for DNA damaging chemotherapy
Examine drug correlations: COMPARE analyses	Univariate Analyses: Plot Data: Data Type: drug vs drug (X or Y); also select Compare Patterns to identify drug-drug correlations	Cell lines sensitive to etoposide are cross-sensitive to topotecan
Multivariate models of drug response & genomic features	Multivariate Analyses: Cell Line Set; Response Data Type; Predictor Data Type/s; Predictor Identifier: enter drug and genomic parameters to be tested as indentifier or use LASSO to discover additional non-redundant determinants of response	Discover independent omic or drug parameters to build a molecular signature for drug response or gene expression
Data download	Univariate Analyses: View Data: Download tabs or Multivariate Analyses: Download tab	Allow further in depth analyses and data download in Excel
Drug identifier conversion	Not applicable; automatically occurs	Allow drug identification across different

Highlighted in red characters are the option tabs of CellMinerCDB: (https://discover.nci.nih.gov/cellminercdb/)

Supplementary Figure 4



A.

x-Axis Cell Line Set
CCLE-Broad-MIT
x-Axis Data Type
mut: DNA Mutation
Identifier
APC
y-Axis Cell Line Set
GDSC-MGH-Sanger
y-Axis Data Type
mut: DNA Mutation
Identifier
APC
Select Tissues/s of Origin
All
Select Tissues to color

B.

x-Axis Cell Line Set
CTRP-Broad-MIT
x-Axis Data Type
act: Drug Activity (AUC)
Identifier
Nutlin-3
y-Axis Cell Line Set
GDSC-MGH-Sanger
y-Axis Data Type
act: Drug Activity (-log10 (IC50))
Identifier
Nutlin-3
Select Tissue/s of Origin
All
Select Tissues to color

C.

Response Cell Line Set
CCLE-Broad-MIT
Response Data Type
cop: DNA Copy Number
Response Identifier
p16
Predictor Cell Line Set
NCI-60
Predictor Data Type/s
cop: DNA Copy Number
Predictor Identifiers
p16
Select Tissue/s of Origin
All
Select Tissues to color

D.

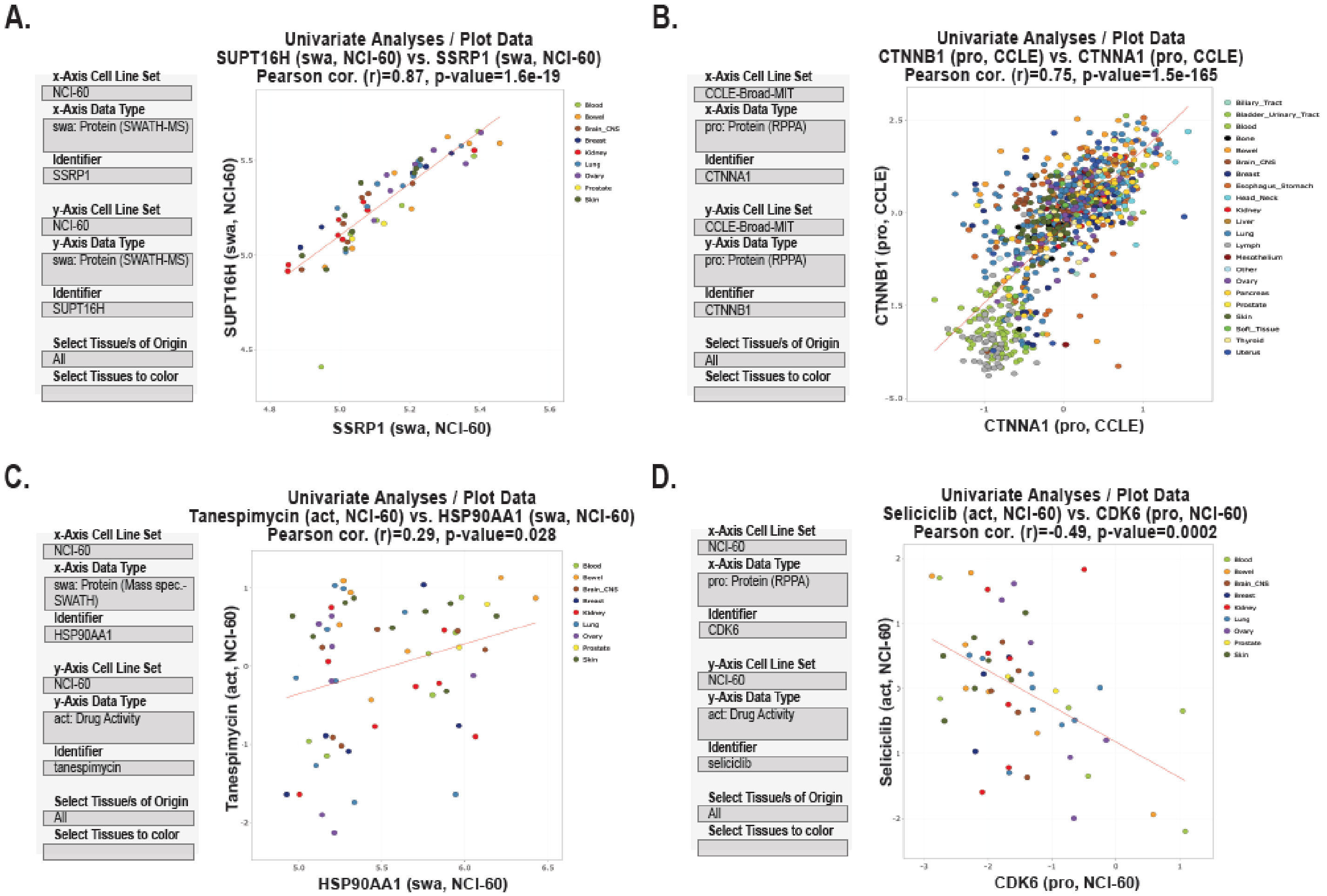
Response Cell Line Set
GDSC-MGH-Sanger
Response Data Type
met: DNA 450k methylation
Response Identifier
SLFN11
Predictor Cell Line Set
NCI-60
Predictor Data Type/s
met: DNA 450k methylation
Predictor Identifiers
SLFN11
Select Tissue/s of Origin
All
Select Tissues to color

E.

Response Cell Line Set
GDSC-MGH-Sanger
Response Data Type
met: DNA 450k methylation
Response Identifier
SLFN11
Predictor Cell Line Set
NCI-60
Predictor Data Type/s
met: DNA 450k methylation
Predictor Identifiers
SLFN11
Select Tissue/s of Origin
All
Select Tissues to color

Figure legend. CellMinerCDB Univariate Analyses scatter plot quality control examples. A. DNA mutation levels of APC as measured by CCLC versus GDSC. B. Drug activity levels of Nutlin-3 as measured by CTRP versus GDSC. C. DNA copy number of P16 as measured by CCLE versus NCI-60. D. DNA methylation levels of SLFN11 as measured by GDSC versus NCI-60. E. DNA methylation levels of SLFN11 as measured by GDSC versus CCLE. All examples are created in CellMinerCDB at https://discover.nci.nih.gov/cellminercdb/ using the selections detailed in the input box (on the left). Each dot is a cell line, with tissues of origin indicated in the legend (on right). The regression line is in red. X- and y-axes, correlations (r) and p-values are as defined within each panel.

Supplementary Figure 5



A.

x-Axis Cell Line Set: NCI-60
x-Axis Data Type: swa: Protein (SWATH-MS)
Identifier: SSRP1
y-Axis Cell Line Set: NCI-60
y-Axis Data Type: swa: Protein (SWATH-MS)
Identifier: SUPT16H
Select Tissue/s of Origin: All
Select Tissues to color:

B.

x-Axis Cell Line Set: CCLE-Broad-MIT
x-Axis Data Type: pro: Protein (RPPA)
Identifier: CTNNA1
y-Axis Cell Line Set: CCLE-Broad-MIT
y-Axis Data Type: pro: Protein (RPPA)
Identifier: CTNNB1
Select Tissue/s of Origin: All
Select Tissues to color:

C.

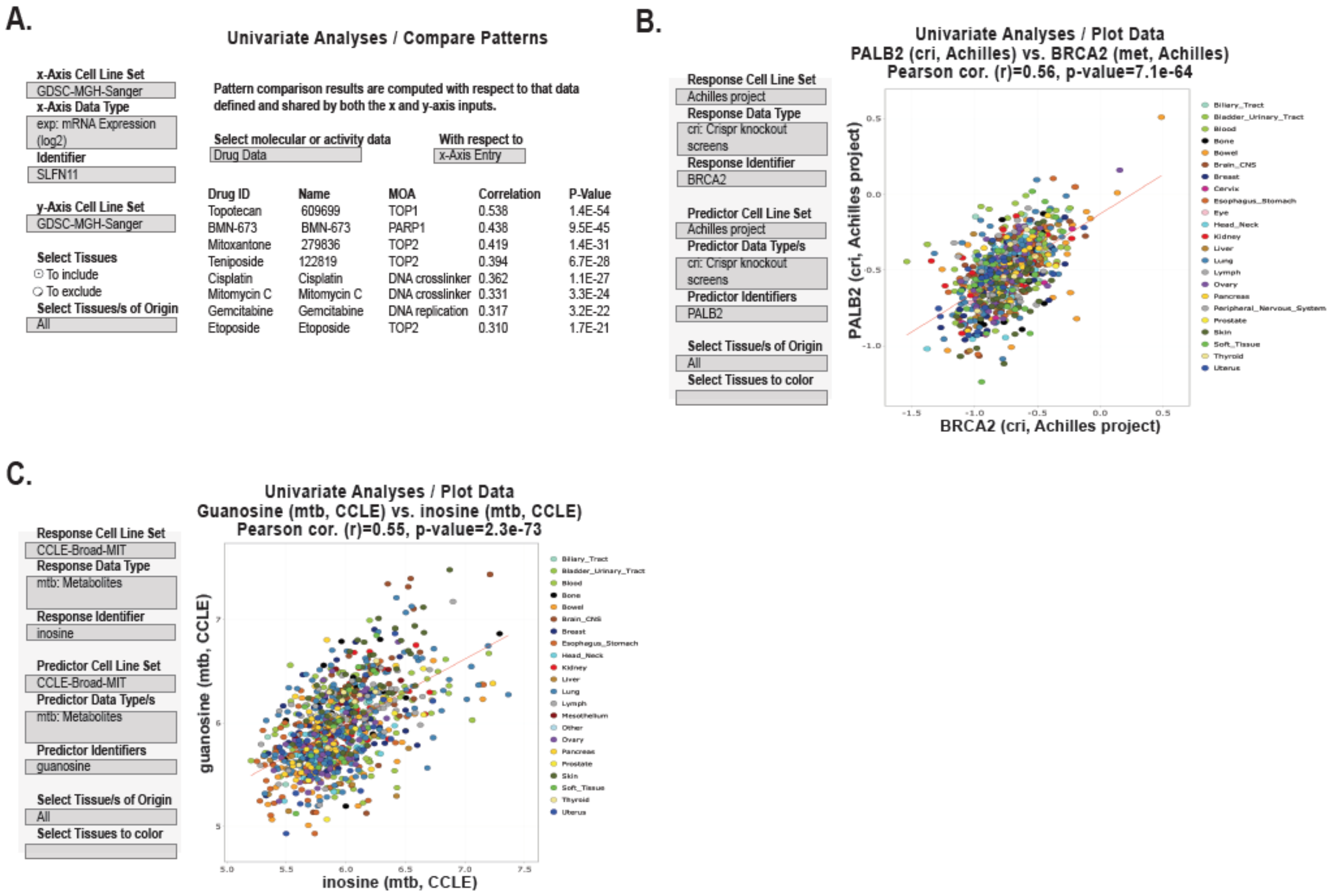
x-Axis Cell Line Set: NCI-60
x-Axis Data Type: swa: Protein (Mass spec.-SWATH)
Identifier: HSP90AA1
y-Axis Cell Line Set: NCI-60
y-Axis Data Type: act: Drug Activity
Identifier: tanespimycin
Select Tissue/s of Origin: All
Select Tissues to color:

D.

x-Axis Cell Line Set: NCI-60
x-Axis Data Type: pro: Protein (RPPA)
Identifier: CDK6
y-Axis Cell Line Set: NCI-60
y-Axis Data Type: act: Drug Activity
Identifier: seliciclib
Select Tissue/s of Origin: All
Select Tissues to color:

Figure legend. CellMinerCDB Univariate Analyses scatter plot protein use examples. A. FACT (facilitates chromatin transcription) complex members SSRP1 (structure specific recognition protein 1) versus SUPT16H (SPT16 homolog) total protein levels (PMID: 26378236). B. CDH1-CTNN (E-cadherin-catenin) complex members CTNNA1 total protein versus CTNNB1 protein levels (PMID: 10692769). C. HSP90AA1 (heat shock protein 90 alpha family class A member 1) total protein (novel candidate biomarker) versus the HSP90 inhibitor tanespimycin activity levels. D. CDK6 (cyclin dependent kinase 6) total protein (novel candidate biomarker) versus the CDK inhibitor seliciclib activity levels. All examples are created in CellMinerCDB at https://discover.nci.nih.gov/cellminercdb/ using the selections detailed in the input box (on the left). Each dot is a cell line, with tissues of origin indicated in the legend (on right). For the scatter plots, the regression line is in red. X- and y-axes, correlations (r) and p values are as defined within each panel. Drugs are at clinical trial level.

Supplementary Figure 6



A. Univariate Analyses / Compare Patterns

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

Select molecular or activity data: Drug Data
With respect to: x-Axis Entry

Drug ID	Name	MOA	Correlation	P-Value
Topotecan	609699	TOP1	0.538	1.4E-54
BMN-673	BMN-673	PARP1	0.438	9.5E-45
Mitoxantone	279836	TOP2	0.419	1.4E-31
Teniposide	122819	TOP2	0.394	6.7E-28
Cisplatin	Cisplatin	DNA crosslinker	0.362	1.1E-27
Mitomycin C	Mitomycin C	DNA crosslinker	0.331	3.3E-24
Gemcitabine	Gemcitabine	DNA replication	0.317	3.2E-22
Etoposide	Etoposide	TOP2	0.310	1.7E-21

B. Univariate Analyses / Plot Data
PALB2 (cri, Achilles) vs. BRCA2 (met, Achilles)
Pearson cor. (r)=0.56, p-value=7.1e-64

Response Cell Line Set: Achilles project
Response Data Type: cri: Crispr knockout screens
Response Identifier: BRCA2
Predictor Cell Line Set: Achilles project
Predictor Data Type/s: cri: Crispr knockout screens
Predictor Identifiers: PALB2
Select Tissue/s of Origin: All
Select Tissues to color:

C. Univariate Analyses / Plot Data
Guanosine (mtb, CCLE) vs. inosine (mtb, CCLE)
Pearson cor. (r)=0.55, p-value=2.3e-73

Response Cell Line Set: CCLE-Broad-MIT
Response Data Type: mtb: Metabolites
Response Identifier: inosine
Predictor Cell Line Set: CCLE-Broad-MIT
Predictor Data Type/s: mtb: Metabolites
Predictor Identifiers: guanosine
Select Tissue/s of Origin: All
Select Tissues to color:

Figure legend. CellMinerCDB Univariate Analyses use examples. A. Compare Pattern: FDA-approved drugs whose activities are significantly correlated to SLFN11 expression, using "FDA" for Clinical Status. B. Scatter plot of Crispr knockouts of BRCA2 versus PALB2 as measured by the Achilles project. C. Scatter plot of metabolite levels of inosine versus guanosine as measured by CCLE. All examples created in CellMinerCDB at https://discover.nci.nih.gov/cellminercdb/ using the selections detailed in the input box (on the left). For the scatter plots, each dot is a cell line, with tissues of origin indicated in the legend (on right). The regression line is in red. x- and y-axes, correlations (r) and p-values are as defined within each panel.