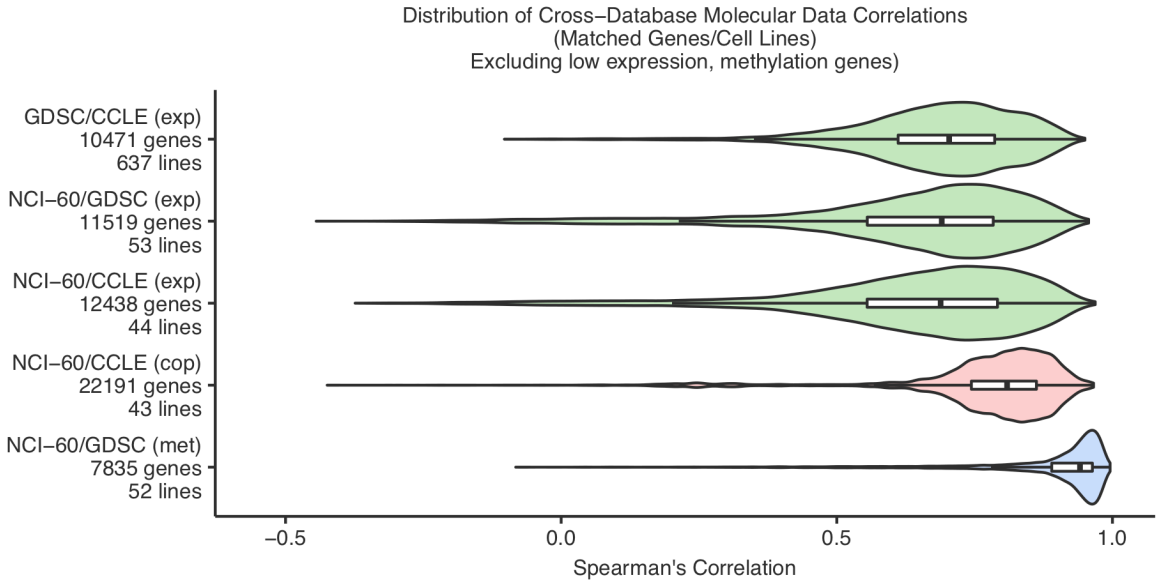
ISCI, Volume *10*

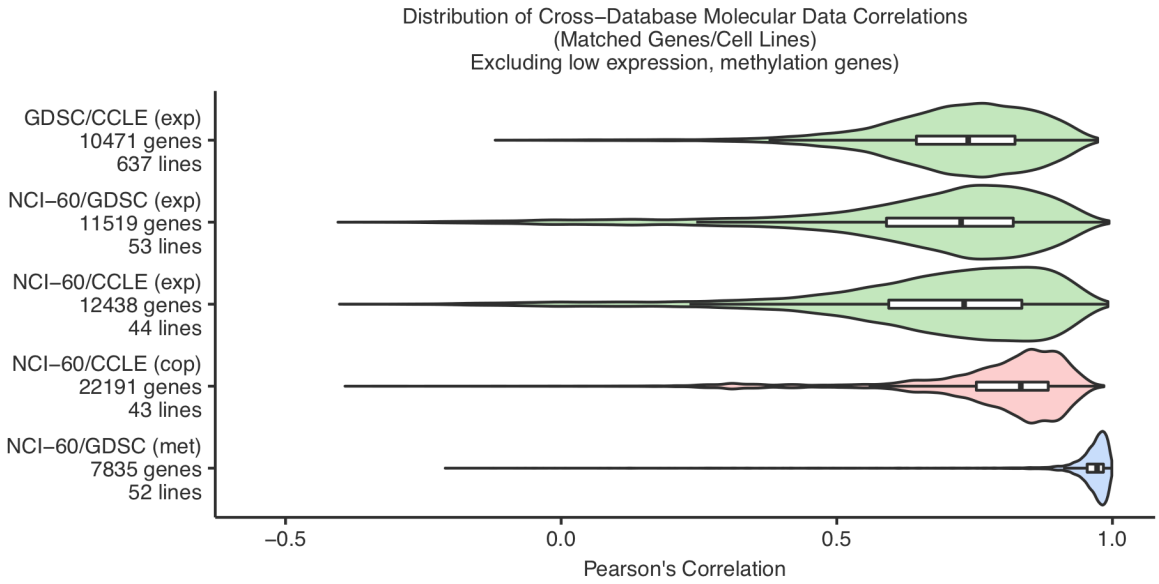
Supplemental Information

CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines

Vinodh N. Rajapakse, Augustin Luna, Mihoko Yamade, Lisa Loman, Sudhir Varma, Margot Sunshine, Francesco Iorio, Fabricio G. Sousa, Fathi Elloumi, Mirit I. Aladjem, Anish Thomas, Chris Sander, Kurt W. Kohn, Cyril H. Benes, Mathew Garnett, William C. Reinhold, and Yves Pommier

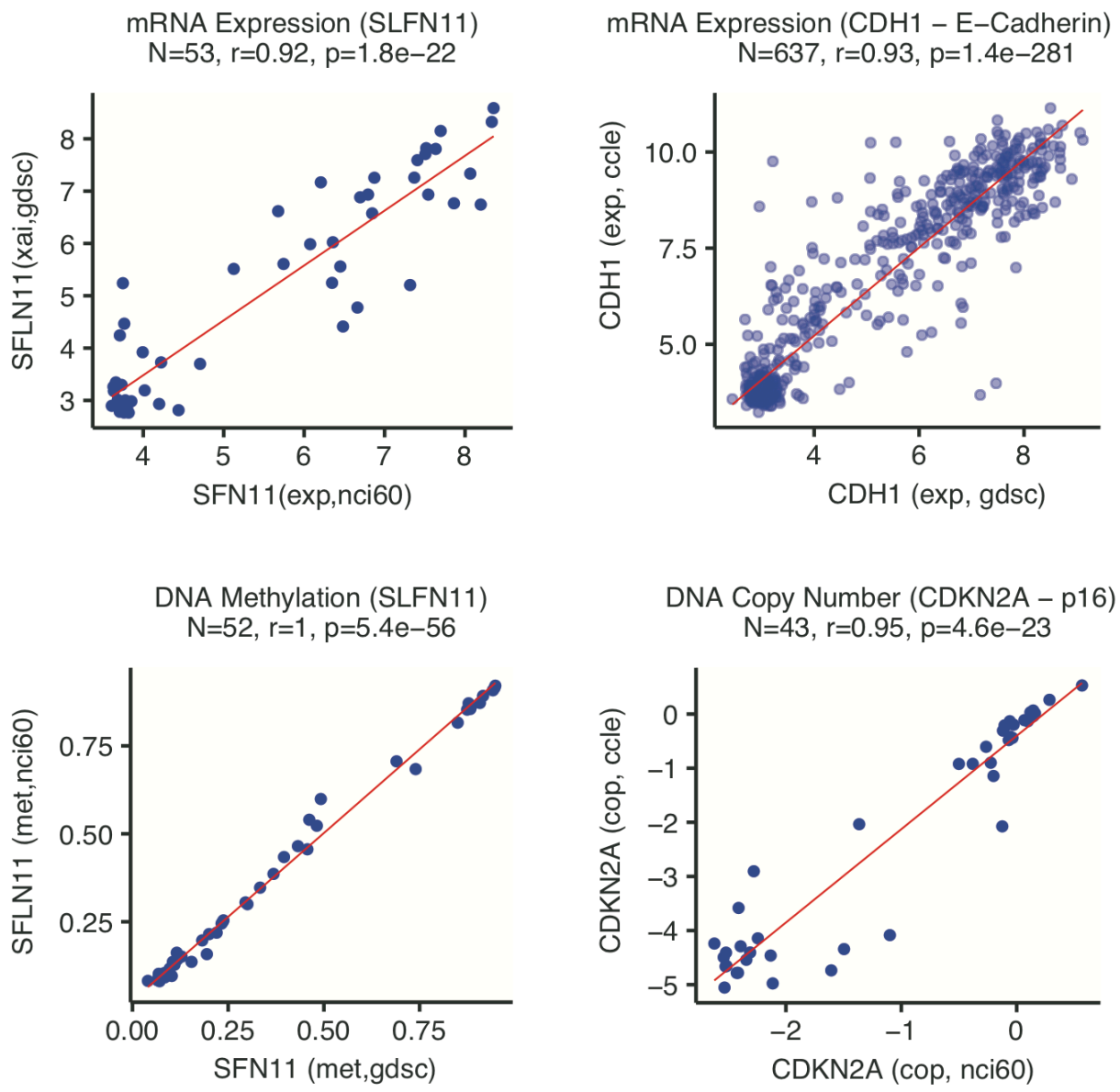
Supplemental Information - Figures:



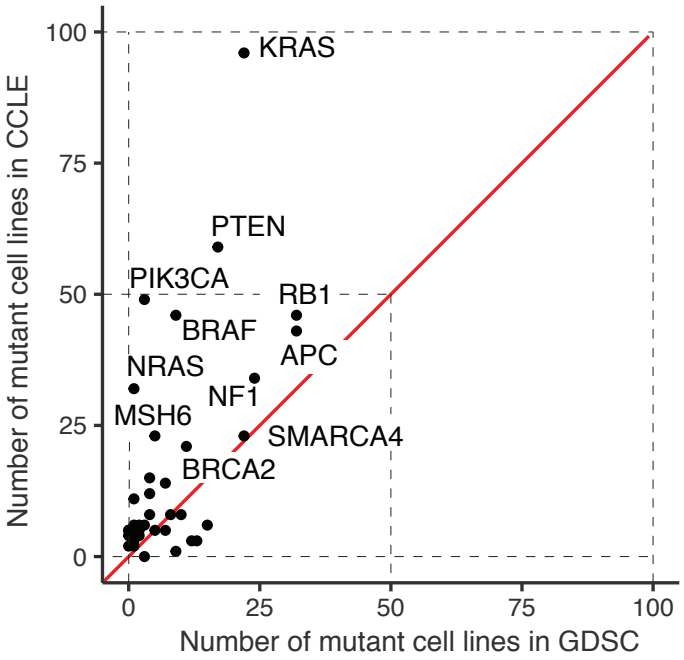


Comparison	Median Correlation (Spearman)	Median Correlation (Pearson)
GDSC/CCLE (exp)	0.704	0.738
NCI-60/CCLE (cop)	0.809	0.834
NCI-60/CCLE (exp)	0.688	0.731
NCI-60/GDSC (exp)	0.690	0.725
NCI-60/GDSC (met)	0.941	0.972

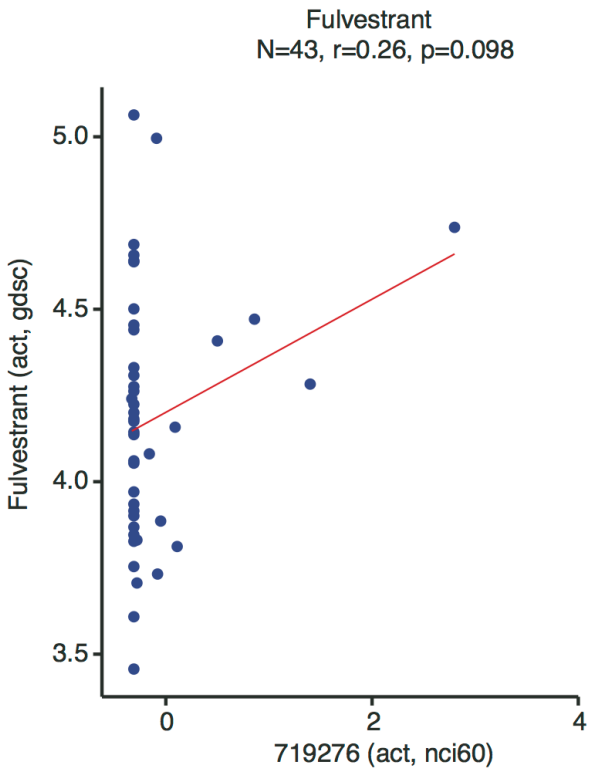
Supplementary Figure 1, Related to Figure 2: Spearman's and Pearson's correlation distributions for comparable, matched cell line transcript expression (exp), DNA copy number (cop), and DNA methylation (met) data across CellMinerCDB-integrated data sources.



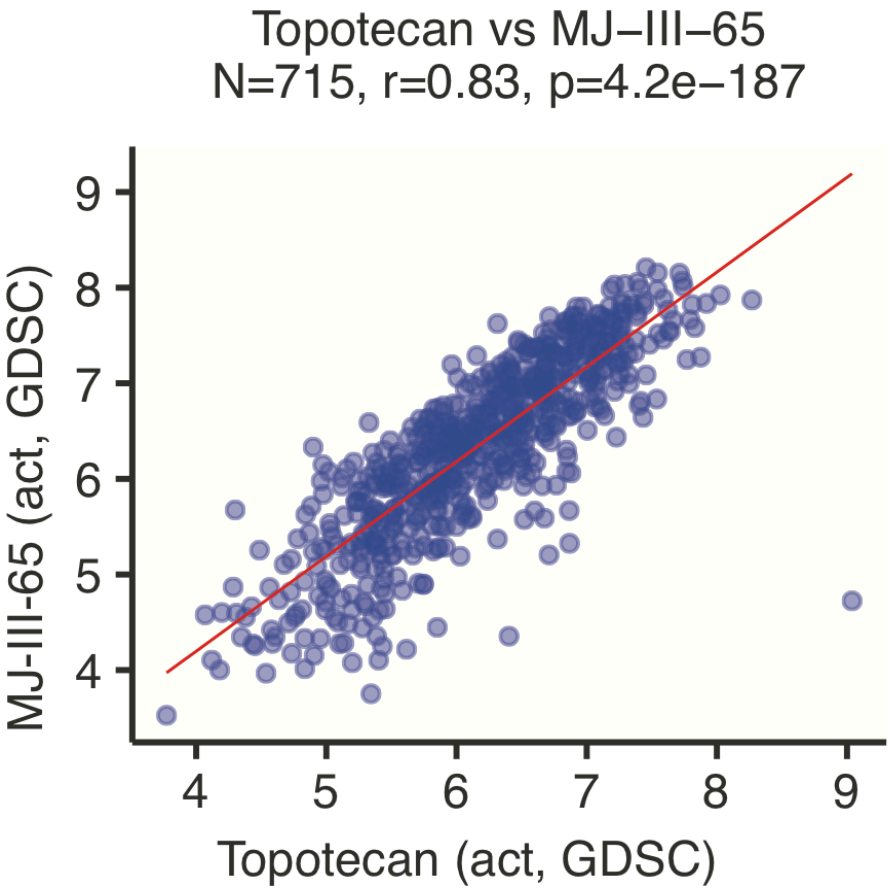
Supplementary Figure 2, Related to Figure 2: Inter-source data reproducibility examples for selected genes and molecular data types.



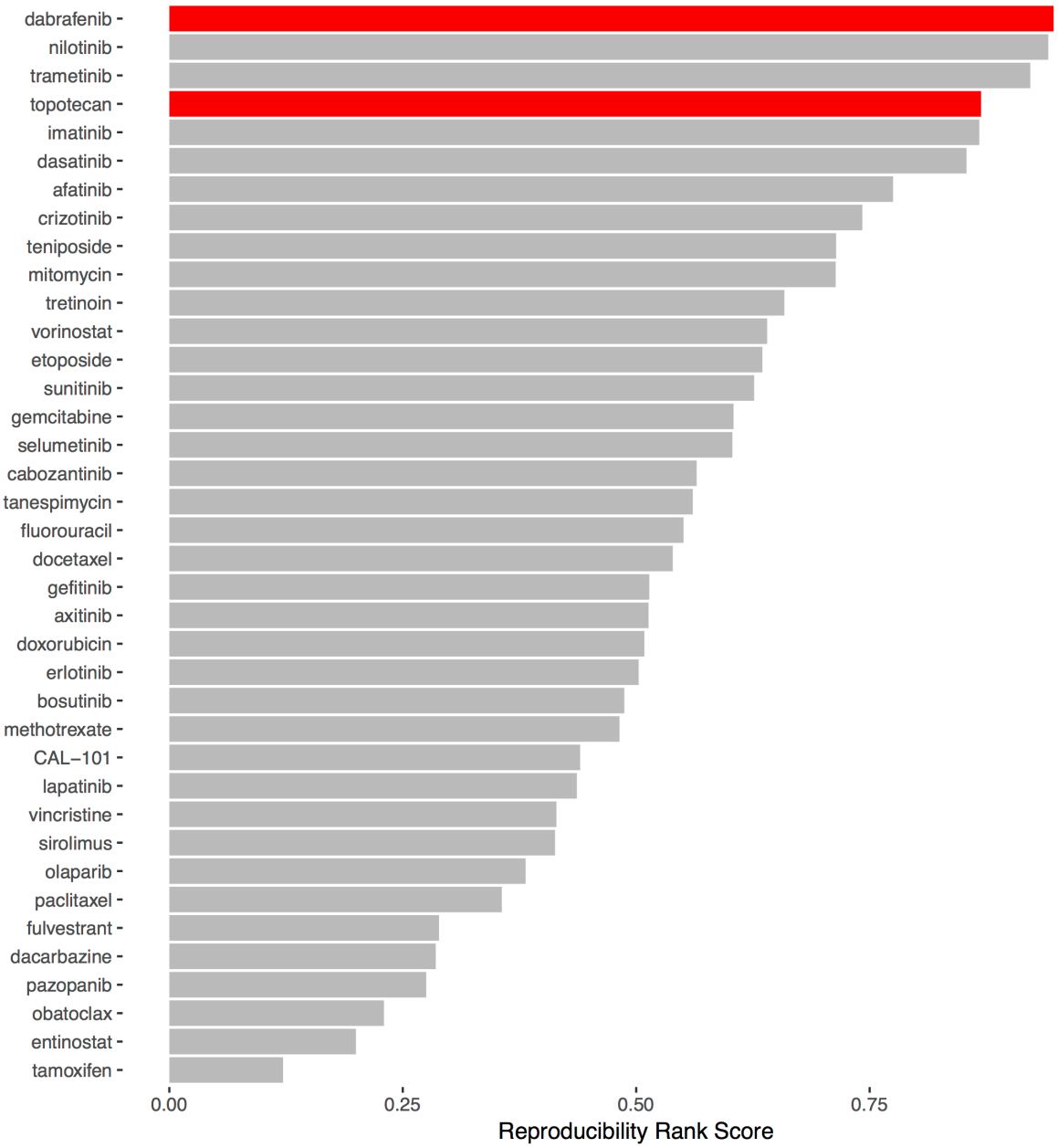
Supplementary Figure 3, Related to Figure 2: Importance of sequencing depth for retrieving mutant cell lines. CCLE vs. GDSC mutant cell line counts for selected oncogenes and tumor suppressor genes.



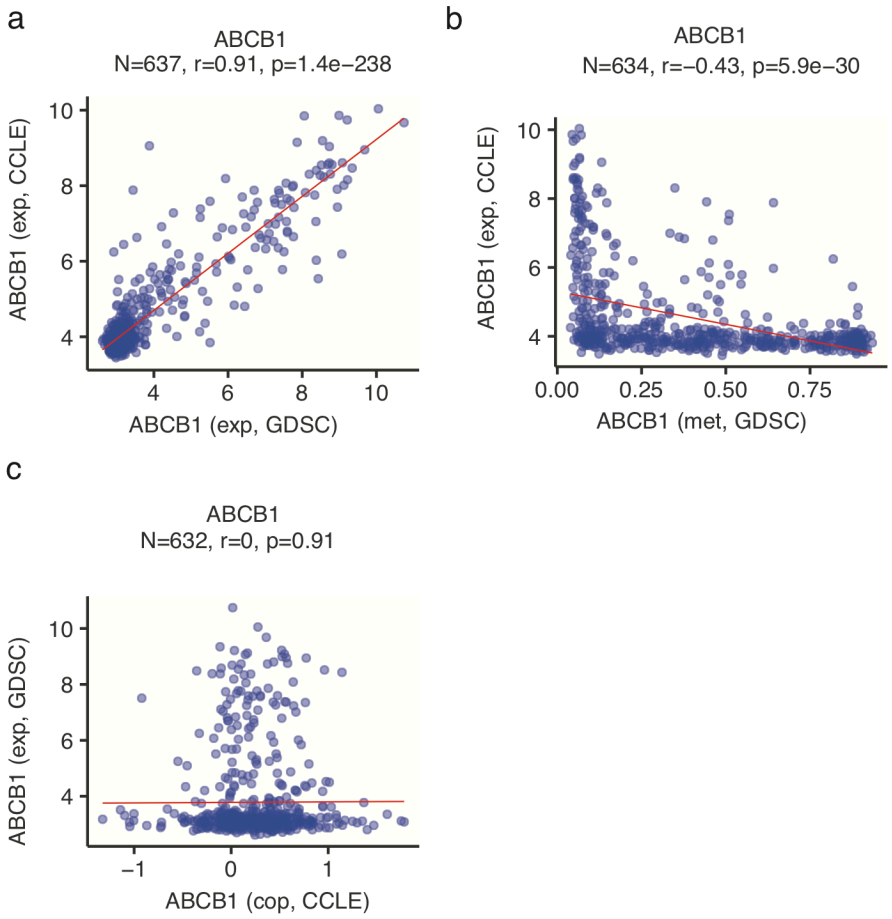
Supplementary Figure 4, Related to Figure 3: GDSC versus NCI-60 drug activity for fulvestrant, indicating inappropriate drug concentration range in NCI-60 activity assay.



Supplementary Figure 5, Related to Figure 3: LMP744 (MJ-III-65) versus topotecan drug activity in the GDSC.

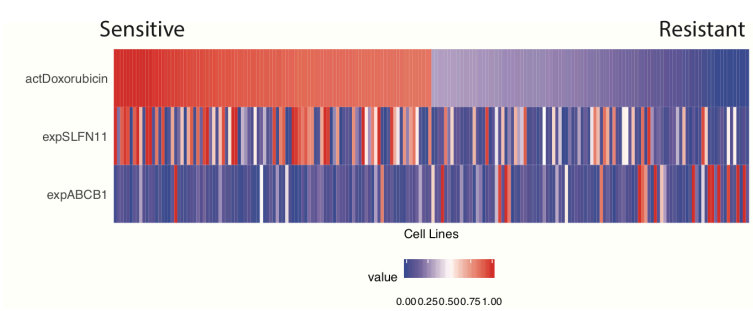


Supplementary Figure 6, Related to Figure 3: Overall drug activity data reproducibility rankings for 38 compounds tested in the NCI-60, GDSC, and CTRP, integrating pairwise activity correlations between the sources.



Supplementary Figure 7, Related to Figure 4: (a) ABCB1 transcript expression is consistently measured in matched cell lines from the CCLE and GDSC sources. Integrating gene-level methylation data provided by the GDSC and gene-level copy number data provided by the CCLE, ABCB1 expression can be seen to be regulated in part by promoter methylation (b) rather than DNA copy number (c).

a



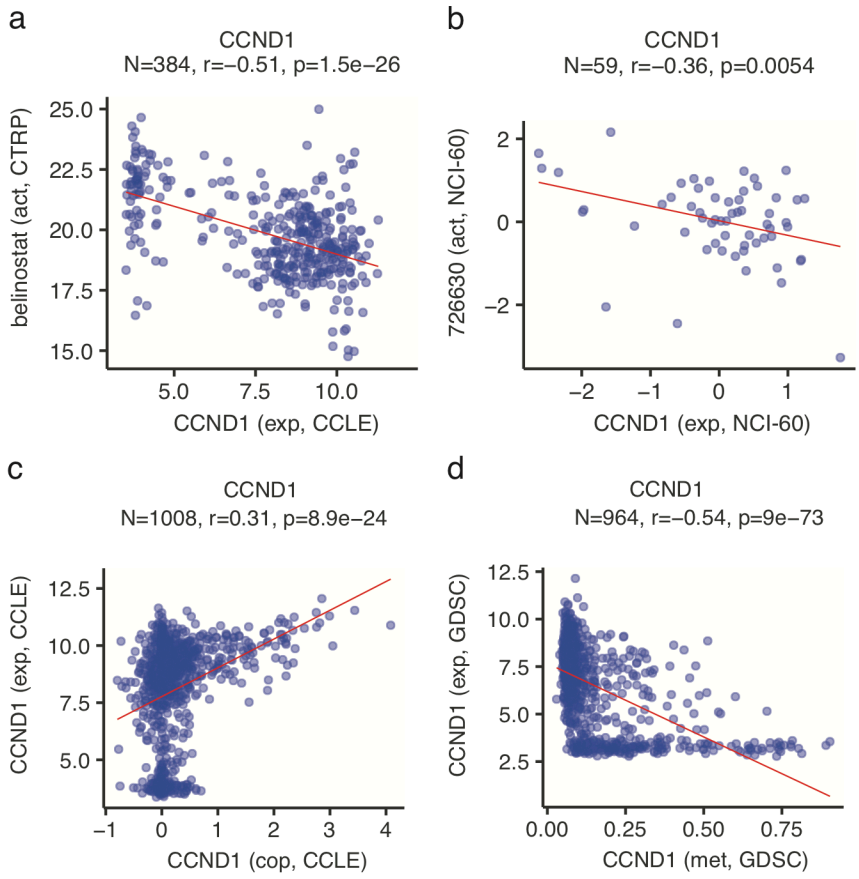
b

```
PREDICTED RESPONSE AS A FUNCTION OF INPUT VARIABLES:

Y = 6.38 + (0.0757*expSLFN11_gdscDec15)


Call:
lm(formula = lmFormula, data = lmData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15063 -0.42040  0.06639  0.48932  1.66794

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.37628    0.06382   99.92  < 2e-16 ***
expSLFN11_gdscDec15  0.07567    0.01168    6.48 1.51e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.707 on 889 degrees of freedom
Multiple R-squared:  0.04511,  Adjusted R-squared:  0.04403
F-statistic:    42 on 1 and 889 DF,  p-value: 1.514e-10
```

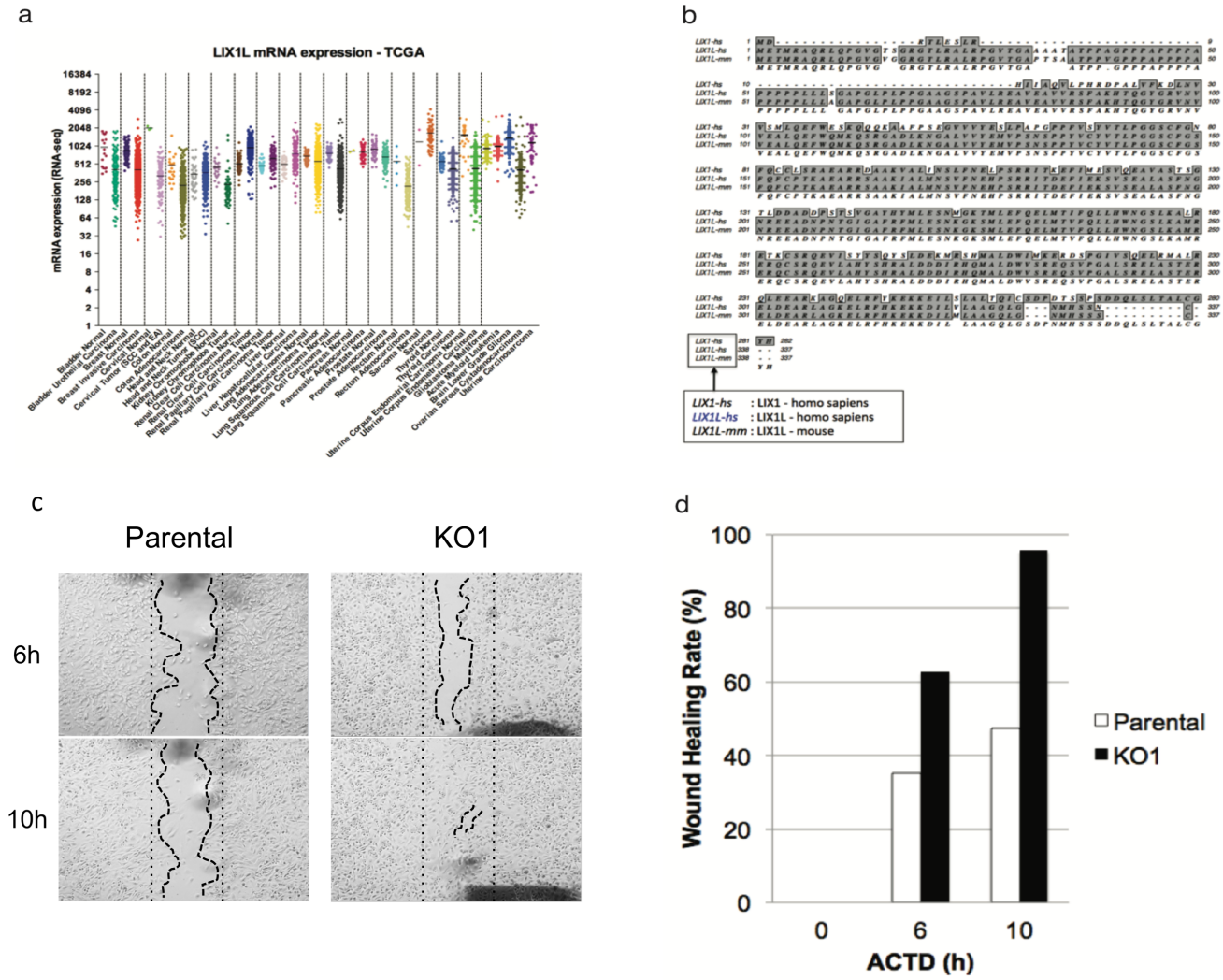
c

```
PREDICTED RESPONSE AS A FUNCTION OF INPUT VARIABLES:

Y = 6.77 + (-0.103*expABCB1_gdscDec15) + (0.0704*expSLFN11_gdscDec15)


Call:
lm(formula = lmFormula, data = lmData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19444 -0.43551  0.06364  0.47969  1.90165

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.77124    0.09907  68.348  < 2e-16 ***
expSLFN11_gdscDec15  0.07040    0.01156   6.091 1.67e-09 ***
expABCB1_gdscDec15  -0.10335    0.02002  -5.161 3.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6971 on 888 degrees of freedom
Multiple R-squared:  0.07292,  Adjusted R-squared:  0.07083
F-statistic: 34.92 on 2 and 888 DF,  p-value: 2.517e-15
```

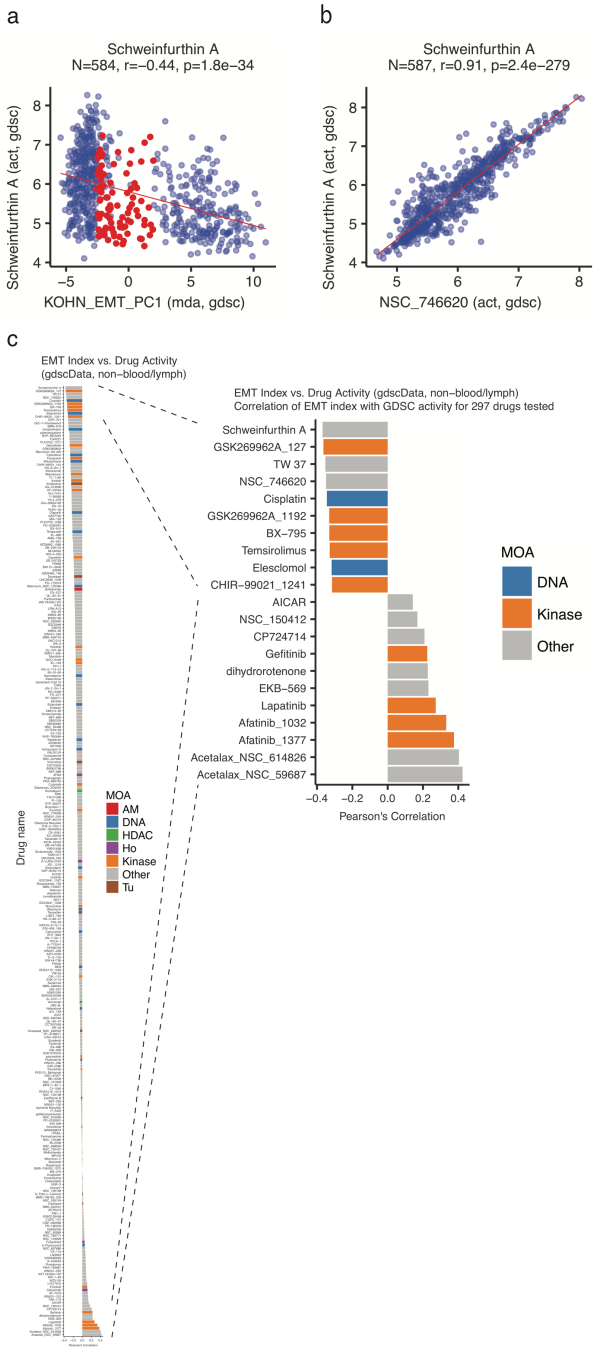
Supplementary Figure 8, Related to Figure 5: ABCB1 expression complements SLFN11 expression in predicting doxorubicin drug activity in GDSC cell lines (b, c), with high ABCB1 expression evident in several highly resistant cell lines indicated at the right of the heatmap in (a).



Supplementary Figure 9, Related to Figure 5: Activity of the HDAC inhibitor belinostat (NSC 726630 in the NCI-60) is negatively correlated with CCND1 transcript expression in both the CTRP/CCLE (a) and the NCI-60 (b). CCLE and GDSC data additionally indicate that both DNA copy number and promoter methylation regulate CCND1 transcript expression (c, d).



Supplementary Figure 10, Related to Figure 7: (a) LIX1L transcript expression across TCGA tumor samples. (b) Sequence alignment showing homology between LIX1, LIX1L (human), LIX1L (mouse). (c, d) Scratch-wound assay results showing increased cell migration with LIX1L knockout.



Supplementary Figure 11, Related to Figure 7: (a) GDSC schweinfurthin A activity versus gene expression-based EMT index value. Red points indicated cell lines with intermediate 'epithelial-mesenchymal' status, while remaining points on the left and right are classified as mesenchymal and epithelial, respectively. (b) Activity of schweinfurthin A vs. activity of 5-methylschweinfurthin G in a subset of GDSC cell lines. (c) Bar plot of Pearson's correlations between GDSC drug activities and EMT index.

Transparent Methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
rcellminer	Luna et al., 2016	http://bioconductor.org/packages/release/bioc/html/rcellminer.html
Other		
CellMiner NCI-60	Reinhold et al., 2015; Reinhold et al., 2017	https://discover.nci.nih.gov/cellminer/
Sanger/Massachusetts General Hospital GDSC	Garnett et al., 2012	https://www.cancerrxgene.org/
Broad/Novartis CCLE	Barretina et al., 2012	https://portals.broadinstitute.org/ccle
Broad CTRP	Rees et al., 2016	https://portals.broadinstitute.org/ctrp.v2.1/
NCI SCLC	Polley et al., 2016	https://sclccelllines.cancer.gov/sclc/

CONTACT FOR RESOURCE AND REAGENT SHARING

For current information about CellMinerCDB or to report site-related issues or feedback, please contact webadmin@discover.nci.nih.gov. Questions about this study should be directed to Vinodh Rajapakse (vinodh.rajapakse@nih.gov), Augustin Luna (augustin_luna@hms.harvard.edu), and Yves Pommier (pommier@nih.gov).

METHOD DETAILS

Overview

CellMinerCDB was implemented using the R programming language, with interactive features developed using RStudio's Shiny web application framework (https://shiny.rstudio.com/). Application deployment and management is enabled by RStudio's Shiny Server Pro production environment. Underlying analyses and data representations were built with functionality provided by our publicly available rcellminer R/Bioconductor package (Luna et al., 2016). For each data source, R data packages were constructed, using software components defined within rcellminer to integrate drug activity data, molecular profiling data, and associated cell line, drug, and gene annotations. This standard data representation allowed diverse data sources to be readily integrated within CellMinerCDB. Source-specific data used in CellMinerCDB data package construction are described in the sections below.

NCI-60 Data Preparation

NCI-60 drug activity, molecular profiling, and annotation data was obtained from CellMiner (Database Version 2.1). The latest versions of these data can also be downloaded from https://discover.nci.nih.gov/cellminer/loadDownload.do. Detailed information is provided in (Abaan et al., 2013; Reinhold et al., 2012; Reinhold et al., 2017; Varma et al., 2014). Essential attributes made available within CellMinerCDB are summarized below.

Compound activity. Standardized, 'z-score' values were derived from measurement of 50% growth-inhibitory (GI50) concentrations using the sulforhodamine B total protein cytotoxicity assay. For each compound, the mean and standard deviation of -log10[molar GI50] values over the NCI-60 lines are used to center and scale the data.

Gene expression. Integration of relevant probe-level data from 5 microarray platforms (Reinhold et al., 2012) is provided in both standardized 'z-score' form, derived as described above for the drug activity data, and as average log2 intensities.

Gene-level mutation. The mutation data value for a given gene and cell line is derived from computed probability of a homozygous function-impacting mutation, which is then expressed as a percentage. NCI-60 exome sequencing data was obtained and processed as described (Abaan et al., 2013). Missense mutations were functionally categorized using ANNOVAR (Wang et al., 2010). Missense mutations with a frequency > 0.005 in either the ESP6500 or 1000 Genomes normal population datasets (i.e., potential germline variants) were excluded, together with mutations predicted not to impact protein function by the SIFT and PolyPhen2 algorithms (SIFT > 0.05 or PolyPhen2 HDIV < 0.85 or PolyPhen2 HVAR < 0.85). To obtain a summarized, gene-specific mutation value for each cell line, the probability of both alleles having at least one of the variants was computed. Specifically, let $x = (x_1, \ldots, x_n)$ be a vector of gene-associated mutation conversion fraction values for a given cell line. The summary gene mutation probability value for this cell line is computed as $1 - (1 - x_1) \ldots (1 - x_n)$, and then converted to a percentage value.

DNA copy number. DNA copy data were integrated from four array-CGH platforms (Varma et al., 2014). Numerical values indicate the average log2 probe intensity ratio for the cell line (gene-specific chromosomal segment) DNA relative to normal DNA.

DNA methylation. Data were obtained using the Illumina Infinium Human Methylation 450 platform as described (Reinhold et al., 2017).Values lie between 0 (lack of methylation) and 1 (complete methylation).

microRNA expression. Data were obtained using the Agilent Technologies Human miRNA Microarray V2 (Liu et al., 2010). Numerical values indicate average log2 probe intensity.

Protein expression. Reverse phase protein array (RPPA) data were obtained as described (Nishizuka et al., 2003). Numerical values indicate probe intensities.

GDSC Data Preparation

Compound activity. Preprocessed activity data for 256 compounds were downloaded from http://www.cancerrxgene.org/downloads. GDSC-provided activity values were converted to indicate the -log10[molar IC50].

Gene expression. Raw Affymetrix Human Genome U219 microarray data deposited in ArrayExpress (E-MTAB-3610) were processed using RMA normalization. Probe-to-gene mapping was performed using the BrainArray CDF file for the Affymetrix HG-U219 platform, available at http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/17.1.0/entrezg.downloa

d/HGU219_Hs_ENTREZG_17.1.0.zip. Numerical values summarize gene-specific log2 probe intensities. Additional platform and processing details are provided in (Iorio et al., 2016).

Gene-level mutation. A tab-separated table listing variants detected in GDSC cell lines was downloaded from COSMIC (release v79). Variants indicated as heterozygous and homozygous were assigned values of 0.5 and 1, respectively. After this, gene-level mutation values were computed as described for the NCI-60 mutation data, except that the final, gene and cell line-specific mutation probabilities were retained (rather than converted to percentage values).

DNA methylation. The table of pre-processed beta values for all CpG islands across the GDSC cell lines was downloaded from the supplementary resources site http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/ (Iorio et al., 2016). DNA methylation data were obtained using the Illumina Infinium Human Methylation 450 platform, and gene-level methylation values were computed using the approach utilized with the NCI-60 data (Reinhold et al., 2017).

Determination of prospective triple negative breast cancers. Expression levels of *ERBB2*, *ESR1*, *ESR2* and *PGR* were assessed by GDSC using the Affymetrix Human Genome U219 Array and accessed in CellMinerCDB. Cell lines with a low value for all 3 genes were classified as triple negative. The log2 intensity thresholds used were *ERBB2*<5, *ESR1*< 3.5, and *PGR*<3.

CCLE Data Preparation

CCLE data were downloaded from https://portals.broadinstitute.org/ccle/home (Barretina et al., 2012).

Compound activity. Activity profiles are available for 24 compounds. CCLE-provided activity values were converted to indicate the -log10[molar IC50].

Gene expression. Raw CEL file data derived from the Affymetrix U133+2 platform were downloaded from the CCLE portal. Normalization was performed using the frma method, implemented by the corresponding Bioconductor package (McCall et al., 2010). Numerical values are the average of gene-specific log2 probe intensities, with the gene-to-probe-set mapping obtained from the hgu133plus2.db Bioconductor package.

Gene-level mutation. The table of targeted sequencing-based mutation data for 1651 genes was downloaded from the CCLE portal. Using the provided allelic fraction information for individual variants, gene-level mutation values were computed as described for the NCI-60.

DNA copy number. Data derived from the Affymetrix SNP 6.0 array were downloaded from the CCLE portal. Numerical values are normalized log2 ratios, i.e., log2(CN/2), where CN is the estimated copy number.

CTRP Data Preparation

Activity data for 481 compounds across 823 cell lines were obtained from Supplementary Tables S2, S3, and S4 of reference (Rees et al., 2016). Activity data originally indicated as the area under a 16-point dose response curve (AUC) were subtracted from the maximum observed AUC value (over all cell lines and drugs) to represent activity by the estimated area above the dose-response curve. This transformation allows increased drug sensitivity to be associated with larger values of the activity measure, consistent with other source activity data integrated within CellMinerCDB. The above CTRP cell line set is included in the CCLE, and CCLE molecular data are thus used for CTRP analyses in CellMinerCDB.

NCI-SCLC Data Preparation

Compound activity and transcript expression data for the NCI-SCLC data set were downloaded from https://sclccelllines-dev.cancer.gov/sclc/downloads.xhtml. Activity values were converted to indicate the -log10[molar IC50]. Transcript expression values are derived from log2 microarray probe intensities.

Cell Line and Gene Set Annotations

Cell lines of particular tissue or tumor types can be highlighted in two-variable plots. In addition, correlation and regression analyses can be restricted to cell line subsets by either inclusion or exclusion of selected tissue or tumor types. To enable this, all cell lines across data sources were mapped to the four-level OncoTree cancer tissue type hierarchy developed at Memorial Sloan-Kettering Cancer Center (http://www.cbioportal.org/oncotree/). Every cell line has an OncoTree top level specification, such as 'Lung', indicating its tissue of origin. Additional OncoTree levels provide more detailed annotation, distinguishing, for example, small cell lung cancer and various types of non-small cell lung cancer. Within the 'Regression Models' tab set, LASSO and partial correlation analyses can be restricted to gene sets curated by the NCI/DTB Genomics and Bioinformatics Group.

Filtering of gene-level molecular profiling data for inter-source reproducibility analyses

In pairwise (source A vs. source B) comparisons of gene expression and methylation data, genes which were essentially not expressed or methylated in the inter-source matched cell line set were excluded from correlation analyses (since these cases, the latter would be over noisy data near technical detection thresholds). In particular, in inter-source transcript expression data comparisons, we excluded genes for which the 90th percentile expression value, across matched cell lines from both compared sources, was below 6 (microarray, log2 intensity). Similarly, in the methylation data comparisons, genes for which the corresponding 90th percentile methylation value was below 0.3 (average probe beta value) were excluded.

Derivation of Epithelial-Mesenchymal Transition (EMT) index and cell line stratification

For each data source, the following steps were taken to obtain a numerical measure of EMT status.

(1) Microarray expression data (log2 intensity) over non-hematopoietic cell lines were selected for a subset of EMT genes identified in (30); these included 22 epithelial genes (ADAP1, ATP2C2, CLDN3, CLDN4, CLDN7, EHF, EPN3, ESRP1, ESRP2, GRHL1, GRHL2, IRF6, LLGL2, MARVELD2, MARVELD3, MYO5B, OVOL1, PRSS8, RAB25, S100A14, ST14, TJP3) and 15 mesenchymal genes (AP1M1, BICD2, CCDC88A, CMTM3, EMP3, GNB4, IKBIP, MSN, QKI, SNAI1, SNAI2, STARD9, VIM, ZEB1, ZEB2).

(2) Data for each gene was centered and scaled by subtracting the mean expression value over the cell line set and then dividing by the corresponding standard deviation.

(3) A principal component analysis was performed, with the EMT index obtained as the first principal component.

For a given cell line, the described EMT index is a weighted sum of EMT gene expression values. For all data sources, mesenchymal gene expression values are associated with negative weights, while epithelial gene expression values are associated with positive weights. EMT index values for non-hematopoietic cell lines in each data source show a bimodal distribution (as in Figure 8b), with putative mesenchymal and epithelial lines having negative and positive index values, respectively. The mixtools R package function nomalmixEM was used to fit a 2-component Gaussian mixture model using the source-specific EMT index data. Cell lines with EMT index

values less than (greater than) one standard deviation above (below) the putative mesenchymal (epithelial) group mean are annotated as mesenchymal (epithelial); the remaining non-hematopoietic lines are classified as epithelial-mesenchymal. Hematopoietic cell lines were excluded from EMT index value computations and associated classifications.

QUANTIFICATION AND STATISTICAL ANALYSES

Data types across sets of cell lines can be plotted with respect to one another within the 'Univariate Analyses - Plot Data' tab. From the 'Univariate Analyses - Compare Patterns' tab, additional molecular and drug response correlates can be tabulated, with respect to either the plotted x-axis or y-axis variable. Pearson's correlations are provided, with reported p-values not adjusted for multiple comparisons. The 'Regression Models' tab set allows construction and assessment of multivariate linear models. The response variable can be set to any data source-provided feature (e.g., a drug response or gene expression profile across cell lines). Basic linear regression models are implemented using the R stats package lm() function, while lasso (penalized linear regression models) are implemented using the glmnet R package (Friedman et al., 2009). The lasso performs both variable selection and linear model coefficient fitting (Tibshirani, 1996). The lasso lambda parameter controls the tradeoff between model fit and variable set size. Lambda is set to the value giving the minimum error with 10-fold cross-validation. For either standard linear regression or LASSO models, 10-fold cross validation is applied to fit model coefficients and predict response, while withholding portions of the data to better estimate robustness. The plot of cross-validation-predicted vs. actual response values can also be viewed within CellMinerCDB, to assess model generalization beyond the training data.

Additional predictive variables for a multivariate linear model can be selected using the results provided within the 'Regression Models - Partial Correlation' tab. Conceptually, the aim is to identify variables that are independently correlated with the response variable, after accounting for the influence of the existing predictor set. Computationally, a linear model is fit, with respect to the existing predictor set, for both the response variable and each candidate predictor variable. The partial correlation is then computed as the Pearson's correlation between the resulting pairs of model residual vectors (which capture the variation not explained by the existing predictor set). The p-values reported for the correlation and linear modeling analyses assume multivariate normal data. The two-variable plot feature of CellMinerCDB allows informal assessment of this assumption, with clear indication of outlying observations. The reported p-values are less reliable as the data deviate from multivariate normality.

DATA AND SOFTWARE AVAILABILITY

CellMinerCDB is accessible at https://discover.nci.nih.gov/cellminercdb/. To support users pursuing specialized or computationally intensive analyses, several data download options are available. From the 'Metadata' tab, complete data tables can be downloaded as tab-delimited text files for any source and data type of interest. Download buttons are also provided for analysis-specific data on their associated panels. These allow 2D plot, heatmap, correlation analysis ('Compare Patterns', 'Partial Correlation'), and regression model-associated data to be

downloaded to tab-delimited text files that can be imported into Excel or other analysis environments.