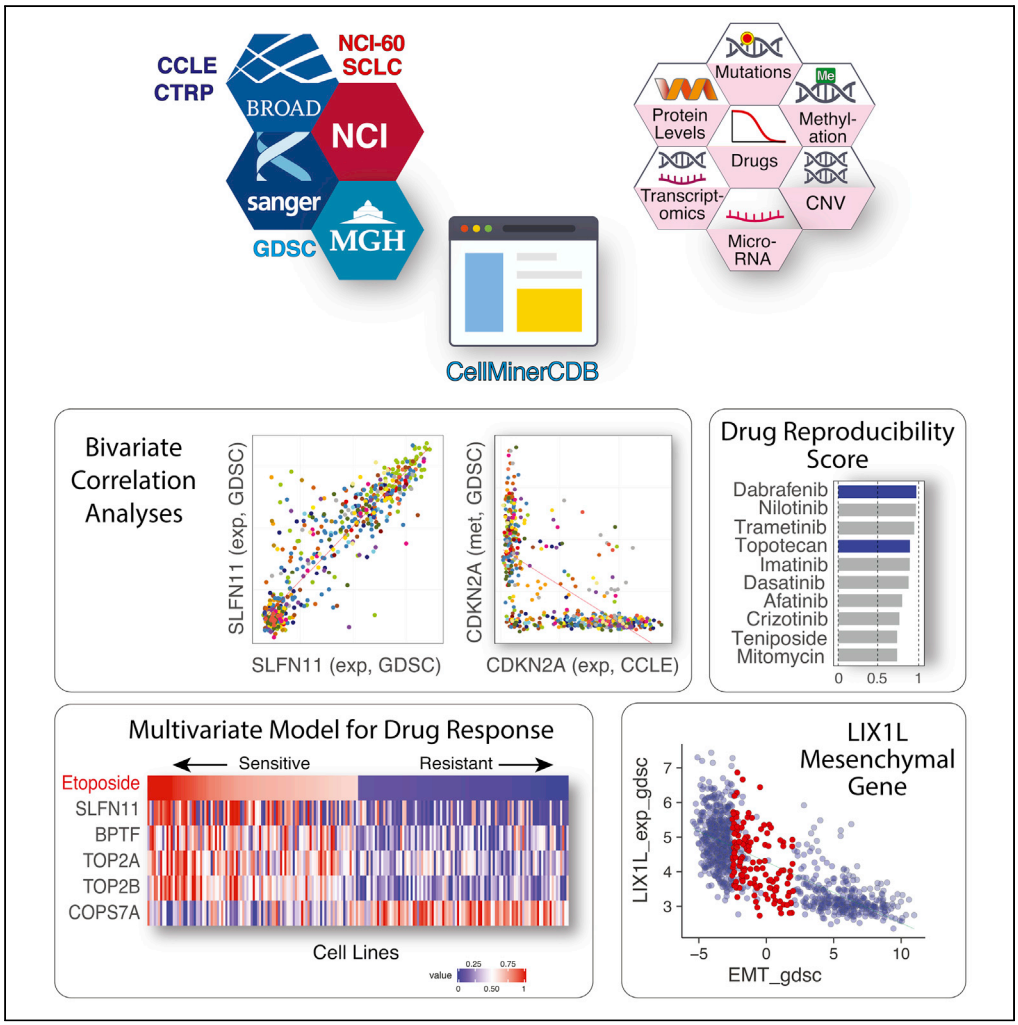
iScience

CellPress

Article

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




Vinodh N. Rajapakse, Augustin Luna, Mihoko Yamade, ..., Mathew Garnett, William C. Reinhold, Yves Pommier

vinodh.rajapakse@nih.gov (V.N.R.)
augustin_luna@hms.harvard.edu (A.L.)
pommier@nih.gov (Y.P.)

## HIGHLIGHTS

CellMinerCDB integrates pharmacogenomic data of the major cancer cell line databases

It seamlessly enables genomic and drug data exploration within and across databases

It tests genomic data reproducibility and proposes drug response determinants

We expand the GDSC drug panel and advance LIX1L as a novel mesenchymal gene

Rajapakse et al., iScience *10*, 247–264
December 21, 2018
https://doi.org/10.1016/j.isci.2018.11.029

iScience

CellPress

Article

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

Vinodh N. Rajapakse,[1,8,*] Augustin Luna,[2,8,*] Mihoko Yamade,[5,8] Lisa Loman,[1] Sudhir Varma,[1] Margot Sunshine,[1,7] Francesco Iorio,[3] Fabricio G. Sousa,[6] Fathi Elloumi,[1,7] Mirit I. Aladjem,[1] Anish Thomas,[1] Chris Sander,[2] Kurt W. Kohn,[1] Cyril H. Benes,[4] Mathew Garnett,[3] William C. Reinhold,[1] and Yves Pommier[1,9,*]

## SUMMARY

**CellMinerCDB provides a web-based resource (https://discover.nci.nih.gov/cellminercdb/) for integrating multiple forms of pharmacological and genomic analyses, and unifying the richest cancer cell line datasets (the NCI-60, NCI-SCLC, Sanger/MGH GDSC, and Broad CCLE/CTRP). CellMinerCDB enables data queries for genomics and gene regulatory network analyses, and exploration of pharmacogenomic determinants and drug signatures. It leverages overlaps of cell lines and drugs across databases to examine reproducibility and expand pathway analyses. We illustrate the value of CellMinerCDB for elucidating gene expression determinants, such as DNA methylation and copy number variations, and highlight complexities in assessing mutational burden. We demonstrate the value of CellMinerCDB in selecting drugs with reproducible activity, expand on the dominant role of SLFN11 for drug response, and present novel response determinants and genomic signatures for topoisomerase inhibitors and schweinfurthins. We also introduce *LIX1L* as a gene associated with mesenchymal signature and regulation of cellular migration and invasiveness.**

## 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 the typically encountered diversity of patient experiences, incomplete knowledge of the multiple molecular determinants of response and resistance factors downstream of the primary drug targets, and tumor heterogeneity. In this setting, the relative homogeneity of cell lines is advantageous, making them model systems for resolving and establishing cellular intrinsic drug response mechanisms. These features motivated the development of cancer cell line pharmacogenomic databases.

Building on the NCI-60 paradigm (Abaan et al., 2013; Reinhold et al., 2012, 2015, 2017; Zoppoli et al., 2012), pharmacogenomic data portals such as the Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett et al., 2012; Iorio et al., 2016), the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012; Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium, 2015), and the Cancer Therapeutics Response Portal (CTRP) (Rees et al., 2016) have expanded to span ~1,400 cancer cell lines. Each database provides a readily available resource for translational research, and proposals have been advanced to further enrich them to over 10,000 cancer cell lines for better coverage of tumor type diversity (Boehm and Golub, 2015). The NCI-60 dataset includes drug activity data for over 21,000 compounds, together with a wide range of molecular profiling data (gene expression, mutations, copy number, methylation, and protein expression). The GDSC and CCLE collections focus on drug activity data for clinically relevant drugs over larger cell line sets, together with an array of molecular profiling data that match the NCI-60 and clinical genomic analyses. The CTRP provides independent drug activity data for nearly 500 compounds over cell lines spanning most of the CCLE and GDSC collections. Each source-specific portal allows deep exploration of its associated datasets, but does not allow immediate cross-database analyses. Yet, substantial overlaps in both cell lines and drugs have the potential to empower integrative analyses, building on the complementarity of the cancer cell line datasets. However, data complexity and mundane (but significant) sources of friction, such as differences in entity naming (cell lines, drugs) and data preparation, have until now made working across databases challenging, even for those with informatics training.

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

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

[3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

[4]Massachusetts General Hospital Cancer Center and Department of Medicine, Harvard Medical School, Charlestown, MA 02129, USA

[5]First Department of Medicine, Hamamatsu University School of Medicine, Hamamatsu 431-3192, Japan

[6]Centro De Estudos Em Células Tronco, Terapia Celular E Genética Toxicológica, Programa De Pós-Graduação Em Farmácia, Universidade Federal De Mato Grosso Do Sul, Campo Grande, MS 79070-900, Brazil

[7]General Dynamics Information Technology Inc., 3211 Jermantown Road, Fairfax, VA 22030, USA

[8]These authors contributed equally

[9]Lead Contact

*Correspondence: vinodh.rajapakse@nih.gov (V.N.R.), augustin_luna@hms.harvard.edu (A.L.), pommier@nih.gov (Y.P.)

https://doi.org/10.1016/j.isci.2018.11.029

To enable integrative analyses within and across data sources, we are introducing CellMinerCDB (https://discover.nci.nih.gov/cellminercdb/), a web application allowing immediate, interactive exploration of the richest cancer cell line genomic and pharmacogenomic databases (Figure 1). In CellMinerCDB, named entities are transparently matched across sources, allowing cell line molecular features and drug responses to be readily compared using bivariate scatterplots and correlation analyses. Multivariate models of drug response or any genomic cell line attribute 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. CellMinerCDB is built using the publicly available rcellminer R/Bioconductor package, which provides analyses and a standard data representation format (Luna et al., 2016). The latter also allows CellMinerCDB to be readily updated to include additional data. Although the rcellminer package (Luna et al., 2016) is available for bioinformaticists, it requires knowledge of the R programming language to install, configure, and conduct analyses. CellMinerCDB, by contrast, is accessible via a web-based interface meant for direct, general use. Furthermore, CellMinerCDB is enhanced with new data sources and analyses, including a wide range of fully interoperable pharmacogenomics datasets, as well as multivariate analyses that can be used to explore the biological complexity of these data. The accessibility of these analyses and breadth of available data make CellMinerCDB a unique resource for cancer cell line pharmacogenomic data exploration and hypothesis generation.

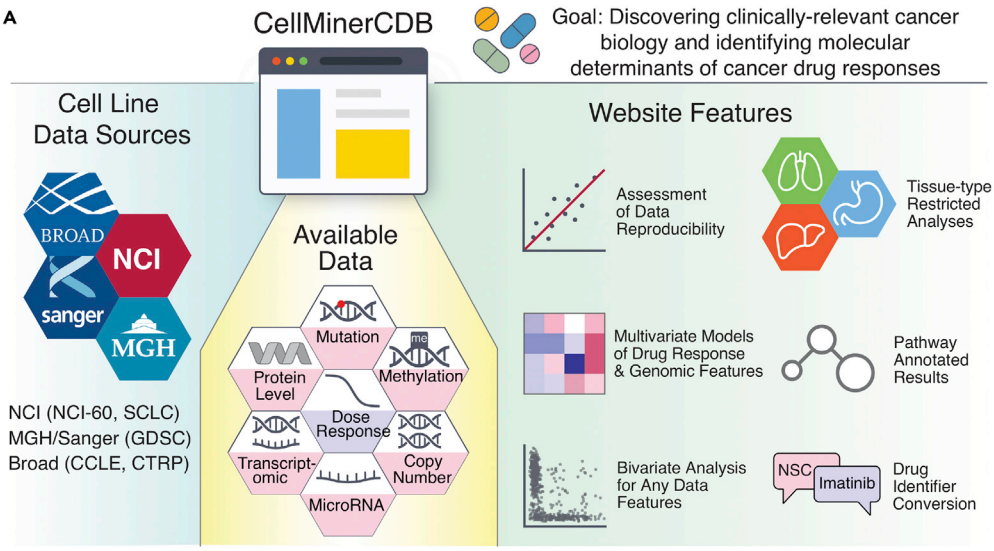
Here we present CellMinerCDB (https://discover.nci.nih.gov/cellminercdb/), highlighting key features of molecular and drug data reproducibility, and complementarity across sources. We provide examples illustrating cancer biology explorations and drug response determinants. We propose the potential repurposing of oxyphenisatin acetate (acetalax; NSC59687) as an anticancer agent for triple-negative breast cancer. We demonstrate multivariate analyses for the exploration of genomic response determinants for topoisomerase inhibitors and schweinfurthins, a class of National Cancer Institute (NCI)-developed compounds derived from natural products. CellMinerCDB also provides phenotypic genomic signatures for cancer cell lines, including a gene-expression-based measure of epithelial-to-mesenchymal (EMT) transition status. We demonstrate the use of the latter to assess EMT stratification within specific tissues of origin, leading to the identification of a novel EMT gene, *LIX1L*. Detailed use of CellMinerCDB is described in a video tutorial (https://youtu.be/XljXazRGkQ8).

## RESULTS

### Data Source Comparisons

CellMinerCDB integrates four prominent cancer cell line data sources: the CellMiner NCI-60 (Abaan et al., 2013; Luna et al., 2016; Reinhold et al., 2015, 2017), Sanger/Massachusetts General Hospital GDSC (Garnett et al., 2012), the Broad/Novartis CCLE, the Broad CTRP (Barretina et al., 2012; Rees et al., 2016), and a tissue-specific dataset encompassing 66 small cell lung cancer lines (NCI-SCLC) (Polley et al., 2016) (Figure 1). Collectively, these databases provide drug activity and molecular profiling data for approximately 1,400 distinct cancer cell lines (Figure 1B, Supplemental Information). Each source has particular strengths. The NCI-60 is unmatched with respect to the breadth of molecular profiling data, as well as the number of tested drugs, compounds, and natural products (>20,000). It also includes replicate data readily accessible via the established CellMiner data portal (Reinhold et al., 2015). The GDSC, CCLE, and CTRP sources feature much larger numbers of cell lines, spanning tissues of origin not included in the NCI-60. The range of tested compounds in these expanded cell line panels is narrow relative to the NCI-60, although the GDSC and CTRP focus on a wide range of clinically relevant anticancer drugs. The CTRP provides data for 170 US Food and Drug Administration (FDA)-approved or investigational anticancer drugs and 196 other compounds with mechanism of action information. The CTRP molecular data in CellMinerCDB are from the CCLE (Figure 1B).

Despite ongoing data acquisition and processing efforts, gaps exist with respect to genomic profiling data (Figure 1B, dark gray table entries). For the GDSC gene mutation and methylation data, we took advantage of processing pipelines developed for the NCI-60 (Reinhold et al., 2014, 2017) to compute gene-level summary data from publicly available raw data. Remaining source-specific molecular profiling data gaps can be filled within CellMinerCDB by effectively extending data provided by one source to another. This is possible because of extensive overlaps between tested cell lines and drugs (Figure 1). For example, gene-level methylation data are not publicly available for the CCLE, but GDSC methylation data are available for the matching 671 CCLE lines and 597 CTRP lines (Figure 1C).




B

Summary of Molecular Drug Activity Data

| Source | # Lines | # Lines (Median) | # Drugs | DNA Variants | mRNA Exp | DNA Copy | DNA Meth. | Mir Exp | Protein Exp | # Molecular |
|---|---|---|---|---|---|---|---|---|---|---|
| NCI-60 | 60 | 57 | 21768 | 9307 | 25040 | 23232 | 17553 | 417 | 162 | 75711 |
| GDSC | 1080 | 900 | 297 | 16532 | 19562 | * | 17580 | NA | NA | 53674 |
| CCLE | 1036 | 491 | 24 | 1667 | 19851 | 23316 | * | * | * | 44834 |
| CTRP | 823 | 751 | 481 | 1667 | 19851 | 23316 | * | * | * | 44834 |
| NCI-SCLC | 67 | 66 | 526 | NA | 17804 | NA | NA | NA | NA | 17804 |

Overlapping Cell Lines and Drugs

C

| CELL LINES | NCI-60 | GDSC | CCLE | CTRP |
|---|---|---|---|---|
| NCI-60 | 60 | 55 | 44 | 41 |
| GDSC | | 1080 | 671 | 597 |
| CCLE | | | 1036 | 823 |
| CTRP | | | | 823 |

D

| DRUGS | NCI-60 | GDSC | CCLE | CTRP |
|---|---|---|---|---|
| NCI-60 | 21768 | 57 | 12 | 63 |
| GDSC | | 256 | 13 | 74 |
| CCLE | | | 24 | 16 |
| CTRP | | | | 481 |

E

| CELL LINES | NCI-SCLC | GDSC | CCLE | CTRP |
|---|---|---|---|---|
| NCI-SCLC | 67 | 40 | 36 | 27 |
| GDSC | | 59 | 39 | 29 |
| CCLE | | | 53 | 39 |
| CTRP | | | | 39 |

F

| DRUGS | NCI-SCLC | GDSC | CCLE | CTRP |
|---|---|---|---|---|
| NCI-SCLC | 526 | 109 | 21 | 119 |
| GDSC | | 297 | 15 | 71 |
| CCLE | | | 24 | 14 |
| CTRP | | | | 481 |

**Figure 1. CellMinerCDB Overview**

(A) CellMinerCDB integrates cancer cell line information from principal resources and provides powerful, user-friendly analysis tools.

(B) Summary of molecular and drug activity data for the five data sources currently included in CellMinerCDB. For molecular data types, the numbers indicate the number of genes with a particular data type. GDSC gene-level mutation and methylation data (numbers in red) were prepared from raw data as part of the development of CellMinerCDB. Asterisks indicate molecular data under development, but not publicly available. Protein expression was determined by reverse-phase protein array.

(C) Cell line and drug overlaps between data sources.

(D) Drug overlaps between data sources.

(E) Small cell lung cancer (SCLC) cell line overlaps between data sources.

(F) SCLC cell line-tested drug overlaps between data sources.

CellMinerCDB automatically matches synonymous cell line and drug names (https://discover.nci.nih.gov/cellminercdb/), freeing users from a mundane but time-consuming impediment to work across data sources.

## Molecular Data Reproducibility

Integrative analyses presuppose data concordance across sources. Such analyses can be readily performed with CellMinerCDB because of the extensive overlaps across the cancer cell line databases: 55 of the NCI-60 lines are in GDSC and 44 are in CCLE, 671 lines (~60%) are shared between CCLE and GDSC (Figure 1C), 40 of the 67 NCI-SCLC lines are in GDSC and 36 are in CCLE (Figure 1E), 74 drugs are in both GDSC and CTRP, and 63 drugs are in both NCI-60 and CTRP (Figure 1D).

For the genomic data, we assessed concordance by computing Pearson's correlations between gene-specific molecular profiles over matched cell lines for all pairs of sources and comparable data types. The distributions of expression, copy number, and methylation data correlations indicate highly significant concordance across sources (Figure 2A). Concordance was also evident based on non-parametric Spearman's rank correlations (Figure S1, related to Figure 2A). For these analyses, gene-level transcript expression and methylation patterns with uniformly low values across matched cell lines were excluded due to their lack of meaningful pattern (Transparent Methods). The median correlations exceed 0.7 in all cases (Figure 2A). The striking concordance between NCI-60 and GDSC methylation data (median R = 0.97, median n = 52) may derive in part from the use of same technology platform (Reinhold et al., 2017) and gene-level data summarization approach (Transparent Methods). Examples for specific genes are displayed in Figure S2 (related to Figure 2A), demonstrating the high data reproducibility for *SLFN11* (Schlafen 11) expression in the NCI-60 versus GDSC, *CDH1* (E-cadherin) expression in GDSC versus CCLE, *SLFN11* methylation in the GDSC versus NCI-60, and *CDKN2A* ($p16^{INK4}$/$p19^{ARF}$) copy number in NCI-60 versus CCLE. Readers are invited to explore their own queries at https://discover.nci.nih.gov/cellminercdb/ by selecting a genomic feature for any given gene in two different datasets of their choice.

Gene-level mutation values in CellMinerCDB indicate the probability that an observed mutation is homozygous and is function impacting. For genes with multiple deleterious mutations in a given cell line, values are converted to cumulative probability values (Reinhold et al., 2014), and are available in graphical and tabular forms at https://discover.nci.nih.gov/cellminercdb/. To compare mutation profiles across sources, we binarized the matched cell line data by assigning a value of 1 to lines with an aforementioned probability value greater than 0.3. This value was selected to be below the formally expected value of 0.5 for a heterozygous mutation to allow for technical variability.

Entirely matched mutation profiles across sources should have a Jaccard index value of 1. As such, the similarity index distributions indicate greater discordance for the mutation data (Figure 2B) than for the other types of genomic data (Figure 2A). The similarity distribution values are higher for the NCI-60 (NCI-60/GDSC median J = 0.5, n = 55; NCI-60/CCLE median J = 0.71, n = 39) than for the GDSC/CCLE comparison (median J = 0.38, n = 593). One caveat, however, is that the large cell line database comparisons entail far larger numbers of matched cell lines. Indeed, the Jaccard similarity values approaching 1 with the NCI-60 comparisons often derive from just one or two matched mutant cell lines. We used similar processing steps to derive gene-level mutation data from variant call data for the NCI-60, GDSC, and CCLE (Transparent Methods). Still, inconsistencies were notable.

Differences between the underlying sequencing technologies and initial data preparation methods are likely to account for the observed discrepancies between the gene mutation data across the datasets. For example, the CCLE mutation data were obtained for a selected set of 1,667 cancer-associated genes subject to high-depth exome capture sequencing (Barretina et al., 2012). They consistently yielded the largest numbers of cell lines with function-impacting mutations. The greater number of mutations found for *KRAS*, *PTEN*, *BRAF*, *NRAS*, or *MSH6* in CCLE relative to the GDSC or NCI-60 databases (evaluated by global exome sequencing; Figure S3, related to Figures 2C and 2D) reflects the importance of sequencing depth for accurate assessment of mutations.

For a more focused and translational assessment of mutation data concordance, we examined the overlap between sources for established oncogenes and tumor suppressor genes (Figures 2C and 2D, Table S1). For the tumor suppressors, we binarized the data using a probability threshold of 0.7 (to account for the

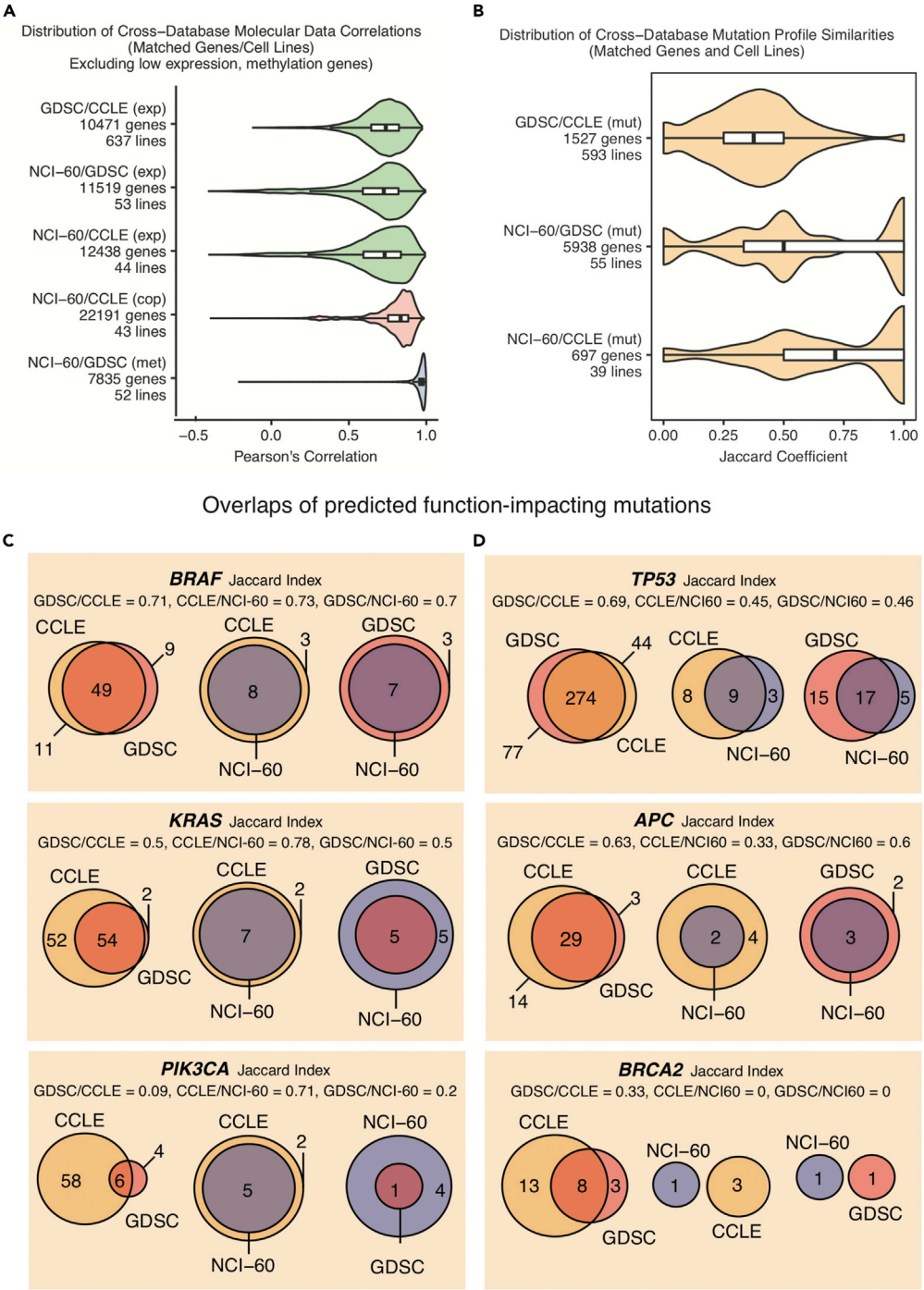



**Figure 2. Molecular Data Reproducibility across Sources**

Comparison of the available genomic features of the cell lines shared between the CellMinerCDB data sources. Bar plots indicate the median and inter-quartile range.

(A) Pearson's correlation distributions for comparable expression (exp), DNA copy number (cop), and DNA methylation (met) data.

(B) Jaccard coefficient distributions for comparable binary mutation (mut) data. The Jaccard coefficient for a pair of gene-specific mutation profiles is the ratio of the number of mutated cell lines reported by both sources to the number of mutated lines reported by either source.

**Figure 2. *Continued***

(C and D) Overlaps of function-impacting mutations as predicted using SIFT/PolyPhen2 for selected tumor suppressor genes and oncogenes. Matched cell line mutation data were binarized by assigning a value of 1 to lines with a homozygous mutation probability greater than a threshold, which was set to 0.3 for (B) and for oncogenes in (C) and to 0.7 for tumor suppressor genes in (D).

recessive nature of such mutations), whereas for the oncogenes, a 0.3 threshold was used (to account for the dominance of oncogene-activating mutations). These values were set below the formally expected values of 1 and 0.5 for homozygous and heterozygous mutations, respectively, to allow for technical variability. As expected, the most frequently mutated genes were TP53, KRAS, BRAF, APC, RB1, NF1, PTEN, SMARCA4, and MLH1 (Table S1, related to Figure 2). BRAF mutation profiles showed the expected overlap ($J > 0.7$) across datasets, as was the case for the TP53 gene across the GDSC and CCLE ($J = 0.69$). On the other hand, PIK3CA, BRCA2, BRCA1, MLH1, MSH6, and MSH2 mutation comparisons were largely divergent. These discrepancies reflect the ongoing challenges and trade-offs with mutation profiling technologies and mutation calling procedures. The ability of CellMinerCDB to compare and integrate data across sources highlights the fundamental research efforts and technological standards still required for the accurate identification of mutations. As a practical matter, CellMinerCDB users can readily compare cell line mutation calls across sources for any given gene of interest. For follow-up studies, they can then select either cell lines that are consistently identified as mutant across sources or the larger set of mutant lines (according to one or more sources).

## Drug Activity Data Reproducibility and Enrichment

Independent studies have examined drug data reproducibility, noting potential sources of data divergence such as assay type and duration of drug treatments (Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium, 2015; Haibe-Kains et al., 2013; Haverty et al., 2016). To explore the reproducibility and the ability of CellMinerCDB to identify genomic signatures over a larger number of cell lines from different tissues of origin, we tested a selected set of NCI-60-screened compounds in the larger GDSC panel (Table S2, related to Figure 3). Noting that the GDSC and the NCI/Developmental Therapeutics Program (DTP) used different assays to determine their IC50 values (Cell Titer Glo measurements of ATP at 72 hr post-treatment versus sulforhodamine B measurement of total protein at 48 hr post-treatment, with additional differences in cell seeding densities and drug dose ranges), we tested in parallel 19 drugs referenced by their NSCs (National Service Center identifiers) and associated with a range of mechanisms of action.

Two drugs with the strongest correlations were oxyphenisatin acetate (acetalax) and bisacodyl (Figure 3A, $R = 0.84$, $p = 8.6 \times 10^{-13}$, $N = 44$ and $R = 0.80$, $N = 43$, $p = 1.0 \times 10^{-10}$, respectively). These FDA-approved laxatives were included in our comparative analysis based on their range of antiproliferative activity in the NCI-60 (further corroborated by NCI-60 activity data for several derivatives), unique pattern of activity compared with the FDA-approved anticancer drugs, outstanding activity in two of the three NCI-60 triple-negative breast cancer cell lines, and lack of pre-existing data in the CTRP, CCLE, or GDSC. The GDSC results confirmed that oxyphenisatin acetate (acetalax) elicits a broad range of cytotoxic responses in the expanded GDSC cell line collection. Extending our NCI-60 observations, it is more active than any of the 15 tested oncologic drugs by a significant margin ($p < 7 \times 10^{-10}$) in the 22 GDSC triple-negative breast cancer lines (Table S3, related to Figure 3).

Overall, 16 of the 19 newly tested compounds across the NCI-60 and GDSC gave significant correlations (Table S2, related to Figure 3). Technical discrepancies were evident for three drugs. Dacarbazine, an alkylating agent related to temozolomide, and vincristine, an anti-tubulin, both had poor reproducibility even within DTP assay replicates. Fulvestrant appeared to be out of the proper concentration range in the DTP assay (Figure S4, related to Figure 3). The non-camptothecin indenoisoquinoline-based topoisomerase I inhibitor in clinical trial, LMP744 (NSC 706744; MJ-III-65) (Burton et al., 2018), was also included in our 19-compound test set to assess the similarity of its activity profile with that of topotecan over a larger cell line collection and to enrich the genomic signature associated with its activity (see section Exploring Drug Response Determinants). Consistent with its activity as a topoisomerase I inhibitor (Antony et al., 2003; Burton et al., 2018), LMP744 is highly correlated with topotecan in the GDSC testing ($R = 0.83$, $p = 4.2 \times 10^{-187}$, $N = 715$) (Figure S5, related to Figure 3), and exhibits significant activity data concordance between NCI-60 and GDSC ($R = 0.66$, $p = 9.8 \times 10^{-7}$, $N = 44$) (Figure 3B).

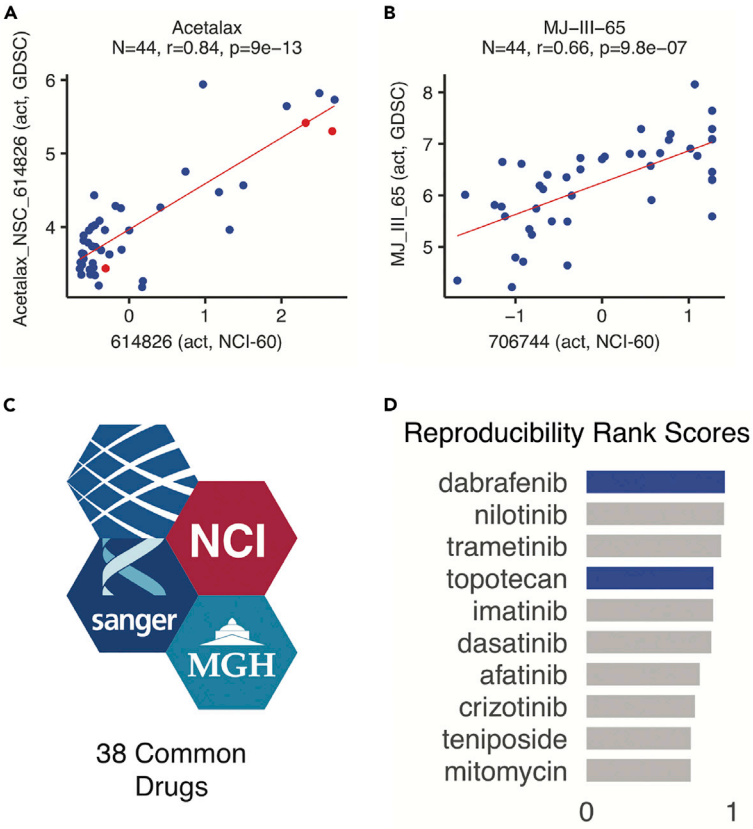

A
Acetalax
N=44, r=0.84, p=9e−13
Acetalax_NSC_614826 (act, GDSC)
614826 (act, NCI-60)
B
MJ−III−65
N=44, r=0.66, p=9.8e−07
MJ_III_65 (act, GDSC)
706744 (act, NCI-60)
C
NCI
sanger
MGH
38 Common Drugs
D
Reproducibility Rank Scores
dabrafenib
nilotinib
trametinib
topotecan
imatinib
dasatinib
afatinib
crizotinib
teniposide
mitomycin
0
1
Specifically Active
E
Dabrafenib
N=248, r=0.71, p=6e−39
dabrafenib (act, CTRP)
Dabrafenib (act, GDSC)
F
Dabrafenib
N=46, r=0.92, p=2.1e−19
764134 (act, NCI-60)
Dabrafenib (act, GDSC)
Broadly Active
G
Topotecan
N=428, r=0.62, p=9.5e−48
topotecan (act, CTRP)
Topotecan (act, GDSC)
H
Topotecan
N=44, r=0.62, p=5.8e−06
609699 (act, NCI-60)
Topotecan (act, GDSC)

**Figure 3. Drug Activity Data Reproducibility**

(A and B) GDSC versus NCI-60 drug activity data in matched cell lines for (A) oxyphenisatin acetate (acetalax; NSC59687) and (B) MJ-III-65 (LMP744; NSC706744). Each point represents a matched cell line. Red points in (A) indicate triple-negative breast cancer cell lines.

(C–H) (C and D) A total of 38 drugs were tested in the NCI-60, GDSC, and CTRP. CCLE was excluded because of its small drug dataset (24 drugs), which is largely included in CTRP. For each of the three inter-source comparisons, drugs were ranked by activity correlation strength (q-value), with ranks scaled between 0 (lowest) and 1 (highest). Specifically active compounds, such as the BRAF inhibitor dabrafenib, show strong correlations based on the response of melanoma lines shown in red (E and F), whereas broadly active compounds, such as the topoisomerase I inhibitor topotecan, show strong correlations based on broad response patterns (G and H). The NCI-60-matched data in (F) and (H) capture the pattern observed with matched data between the larger GDSC and CTRP collections. The full data table excerpted in (D) is shown in Figure S6.

Further focusing on drug activity data reproducibility, we analyzed the 38 drugs previously tested in each of the three databases with larger numbers of tested drugs (NCI-60, GDSC, and CTRP) (Figure 3C). For each of the three inter-source comparisons, drugs were ranked by activity correlation strength (q-value, scaled between 0 [lowest] and 1 [highest]). The drugs were then ordered by the average of the three inter-source comparison rank scores (Figures 3D and S6, related to Figure 3). As noted in earlier studies of drug activity data reproducibility (Haverty et al., 2016), strong activity correlations were observed for *specifically active* compounds (Figures 3E and 3F), such as the BRAF inhibitor dabrafenib, wherein outstanding response occurs in cell lines with the activated kinase target. Notably, we also observed high correlations for *broadly active* drugs, such as the topoisomerase I inhibitor topotecan (Figures 3G and 3H), indicating that the cancer cell line responses are reproducible across databases and assays and are not limited to protein kinase inhibitors. Still, for many of the 38 assessed drugs (see lower half of Figure S6, related to Figure 3), there were discordant activity data between one or more pairs of sources. The inter-source activity data comparisons enabled by CellMinerCDB allow individual researchers to identify drugs with concordant data, so they can pursue reliable molecular pharmacology and translational genomic analyses (see below and Figures 5 and 6).

## Exploring Gene Regulatory Determinants

Cancer-specific gene expression is known to be affected by DNA copy number variations (CNVs) and epigenetic alterations such as promoter methylation. CellMinerCDB makes it easy to explore these and other potential gene regulatory determinants. For example, in the NCI-60, reduced expression of the tumor suppressor gene *CDKN2A* ($p16^{INK4a}$) is associated with both DNA copy loss (Figure 4A) and promoter hypermethylation (Figure 4B) across tissue types. Notably, Figure 4C shows that approximately 25% of NCI-60 cell lines show both alterations, consistent with biallelic, "two-hit" suppression of *CDKN2A* expression. Integration of matched cell line GDSC methylation and CCLE copy number data illustrates the same *CDKN2A* regulatory relationships in a larger cell line collection (Figures 4D–4F). Table S4 (related to Figure 4) shows that *CDKN2A* stands out with respect to the high proportion of cell lines showing co-occurrence of promoter methylation and DNA copy loss. Conversely, the impact of copy gain on increased oncogene expression can be similarly assessed with CellMinerCDB. Figure 4G shows that a subset of *MYC*-driven CCLE small cell lung cancer lines (red dots) exhibits both *MYC* copy gain and increased *MYC* gene expression. KRAS activation, typically regarded as mutation driven, can also occur by copy gain, as evident in a subset of CCLE lines (Figure 4H), consistent with clinical studies (Wagner et al., 2011).

## Exploring Drug Response Determinants

CellMinerCDB allows correlation analyses and scatterplots for testing and visualizing potential response-determinant relationships (univariate analyses) as well as multivariate linear regression methods for integrating multiple determinants (multivariate analyses; see Figures 5D and 6B). CellMinerCDB also enables the discovery of candidate genomic determinants of drug response as well as drug-drug correlations ("Compare Pattern" tab in the "Univariate analyses" tool; https://discover.nci.nih.gov/cellminercdb/). This method led to the discovery of Schlafen 11 (*SLFN11*) expression as a causal determinant of response to widely used DNA-targeted anticancer agents, including topoisomerase inhibitors, platinum derivatives, PARP inhibitors, and antimetabolites (Barretina et al., 2012; Murai et al., 2016, 2018; Zoppoli et al., 2012). Starting with target expression profiles, CellMinerCDB correlation analyses can identify compounds with matching activity profiles. For example, CellMinerCDB can be used to demonstrate that epidermal growth factor receptor (EGFR) expression is significantly correlated with the activity of erlotinib and afatinib in all major cell line databases, as well as with the activity of other established EGFR inhibitors available in one or

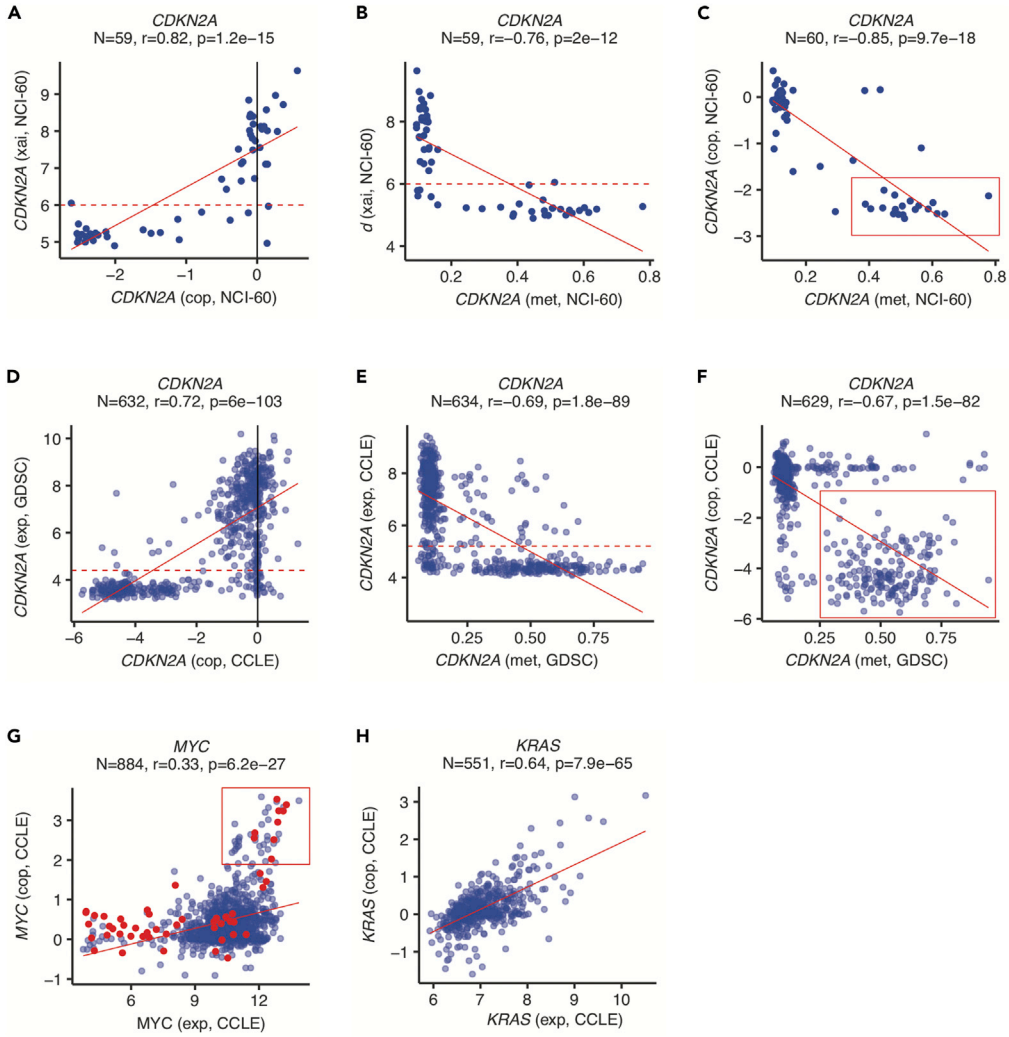
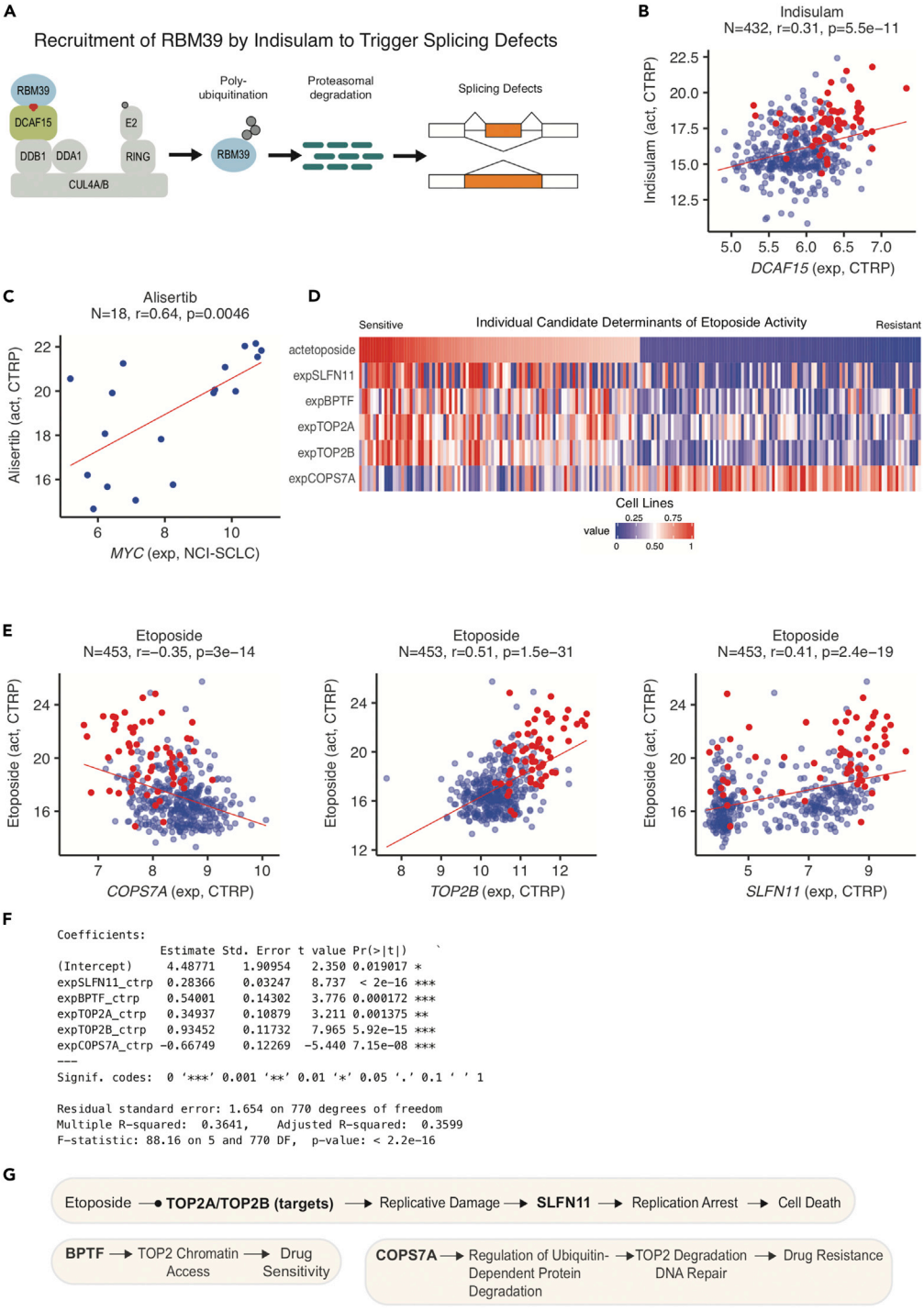



**Figure 4. Exploring Gene Expression Determinants**

Reduced mRNA expression (xai, average log2 intensity) of the cell cycle inhibitor and tumor suppressor CDKN2A (p16) is associated with DNA copy loss (cop) (A) and promoter methylation (met) (B) in the NCI-60 lines. In a subset of NCI-60 lines, enclosed in the red box, (C), DNA copy loss accompanies higher levels of promoter methylation. DNA copy number and promoter methylation data from the CCLE and GDSC, respectively, can be also be visualized over matched cell lines to verify a similar pattern in larger cell line collections (D–F). Note that the corroboration of the NCI-60 regulatory relationships in a far larger and more diverse cell line set is uniquely enabled by CellMinerCDB, which allows gene-level methylation data only available in the GDSC to complement gene-level DNA copy number data only available in the CCLE (for automatically matched cell lines). DNA copy number gain is associated with increased expression (exp, *Z* score microarray log2 intensity data) of the oncogenes MYC (G) and KRAS (H) in selected CCLE cell lines. In (G), small cell lung cancer lines are indicated in red to highlight a subset potentially derived from MYC-driven tumors (within red box).

more data sources. CellMinerCDB correlation analyses also allow direct evaluation of drug resistance determinants. For example, potential substrates of drug efflux ABC transporters can be recovered because strong negative activity correlates with the expression of ABC drug transporters, as in the case of paclitaxel and ABCB1 in the GDSC ($r = -0.33$; p value = $5.7 \times 10^{-12}$).

CellMinerCDB also allows users to assess on the spot the generality of results presented in the literature, and iteratively explore evidence for multifactorial mechanistic models. Figure 5A shows an example for indisulam, which targets the splicing factor RBM39 for proteasomal degradation by forming a ternary complex with RBM39 and the E3 ubiquitin ligase receptor DCAF15. A report of increased indisulam sensitivity in hematopoietic cell lines with high *DCAF15* expression is readily verified with CellMinerCDB (Figure 5B, red




F

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     4.48771    1.90954   2.350 0.019017 *  
expSLFN11_ctrp  0.28366    0.03247   8.737  < 2e-16 ***
expBPTF_ctrp    0.54001    0.14302   3.776 0.000172 ***
expTOP2A_ctrp   0.34937    0.10879   3.211 0.001375 ** 
expTOP2B_ctrp   0.93452    0.11732   7.965 5.92e-15 ***
expCOPS7A_ctrp -0.66749    0.12269  -5.440 7.15e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.654 on 770 degrees of freedom
Multiple R-squared:  0.3641,    Adjusted R-squared:  0.3599 
F-statistic: 88.16 on 5 and 770 DF,  p-value: < 2.2e-16
```

G




**Figure 5. Exploring Drug Response Determinants**

(A and B) Response to the pre-mRNA splicing inhibitor indisulam versus expression of its target complex component DCAF15 in the CTRP. Drug response in (B) is measured by the activity area above the dose-response curve, with higher values indicating relative drug sensitivity. A report of increased indisulam sensitivity in hematopoietic cell lines (shown in red) with high DCAF15 expression is readily verified (Han et al., 2017).

(C) Response to the aurora kinase inhibitor alisertib is associated with increased MYC expression in small cell lung cancer lines (Mollaoglu et al., 2017).

(D) Heatmap indicating etoposide drug activity and candidate determinant gene expression in the 100 most sensitive and resistant CTRP cell lines.

**Figure 5. *Continued***

(E) Scatterplots of etoposide activity versus candidate determinant gene expression in CTRP cell lines, with hematopoietic cell lines shown in red.

(F) A statistical summary of a multivariate linear model of etoposide response in the CTRP.

(G) A mechanistic scheme indicating how the selected determinants may influence etoposide drug response.

dots) (Han et al., 2017). CellMinerCDB also corroborates a report of *MYC*-driven small cell lung cancer exhibiting vulnerability to aurora kinase inhibition (Mollaoglu et al., 2017) (Figure 5C).

As determinants of drug responses are multifactorial, CellMinerCDB includes a multivariate analysis tool under the "Regression Models" tab. Figures 5D–5G illustrate its use for the topoisomerase II inhibitor etoposide. Starting with expression of the drug target (*TOP2A* and *TOP2B*) (Pommier et al., 2016) and *SLFN11* (Zoppoli et al., 2012), users can select additional determinants based on biological knowledge. Determinant selection can be further guided by pathway annotations, as well as partial correlation analyses, which measure the capacity of additional features to improve the current model (Figure 5G). Additional determinants can be found using the "LASSO" tool in the "Algorithm" dropdown menu of the "Regression Models" tab of the CellMinerCDB website. The use of the multivariate modeling tools included in the "Regression Models" tab is outlined in the video tutorial (https://youtu.be/XljXazRGkQ8) and will be exemplified below in the A Multivariate Model of Schweinfurthin A Drug Activity section.

## Benefits of Analyses over Multiple Data Sources

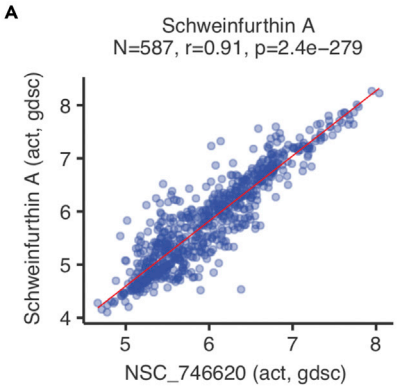
The uniform data representation, accessibility, and interoperability provided by CellMinerCDB allows direct exploratory analyses across the datasets (NCI-60, CCLE, GDSC, CTRP, NCI/DTP-SCLC). This is critically important to identify molecular and drug response determinants with consistent data across sources for specific analyses. The automatic management of cell line overlaps also enables comprehensive analyses encompassing all databases, using complementary data to supplement source-associated gaps in molecular and drug activity data (see Figure 1, and prior section). These application features are highlighted in the examples presented below.

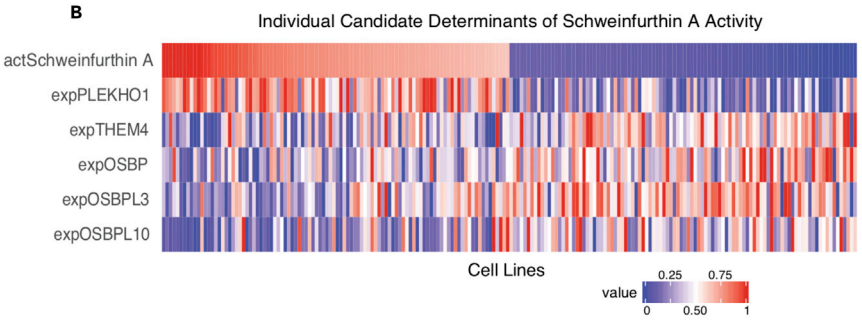
The main ABC transporter, ABCB1 (PgP), is a dominant factor conferring resistance to multiple classes of clinically relevant drugs. Because CellMinerCDB integrates different databases with different drugs tested in each database, it can reliably test the relationship between drug resistance and ABCB1 expression. The first step (https://discover.nci.nih.gov/cellminercdb/) is to ensure that ABCB1 expression exhibits a high dynamic range (i.e., cells with and without expression and high expression in the positive cells) and that ABCB1 expression is highly correlated across databases. Figure S7A (expression of ABCB1 across GDSC and CCLE) shows very high correlation between the two databases ($r = 0.91$; p value $= 1.4 \times 10^{-238}$). The next step is to use the "Compare Patterns" tool of CellMinerCDB by entering "ABCB1" as the "x-Axis Data Type" for each of the databases (selected via the "x-Axis Cell Line Set"). Table S5 (related to Figure 5) shows the top 50 drugs with activity negatively correlated with ABCB1 for the three datasets: GDSC, CTRP (CCLE), and NCI-60. Overlapping and established drugs effluxed by ABCB1 in each dataset are highlighted in yellow. In addition, each dataset includes many additional drugs. Therefore, if a drug is not in one dataset, it may be found in others. Figure S8 (related to Figure 5) shows that adding ABCB1 to SLFN11 enhances the prediction of doxorubicin activity. This analysis is readily done using the "Regression Models" tool of CellMinerCDB. Finally, Figure S7 (related to Figure 4) shows that ABCB1 is epigenetically regulated by promoter hypermethylation Figure S7B rather than by copy number alteration Figure S7C.

Figure S9 (related to Figure 5) presents an example of cross-database exploration to identify cyclin D1 (CCND1) as a potential determinant of response to the HDAC inhibitor belinostat, together with evidence of CCND1 expression regulation by DNA copy number and promoter methylation.

## A Multivariate Model of Schweinfurthin A Drug Activity

Schweinfurthin A was discovered by the NCI natural products initiative (Thornburg et al., 2018) to identify compounds with distinctive NCI-60 activity profiles indicative of novel targets (via COMPARE analysis; Paull et al., 1989). Its wide activity range with notable potency in leukemia and CNS lines (<10 nmol/L) motivated the synthesis of a series of derivatives (Kodet et al., 2014). Because the development of schweinfuthins has been hampered by limited understanding of their molecular pharmacology, we tested Schweinfurthin A and 5'-methylschweinfurthin G (NSC 746620) in the GDSC panel and applied the various features of




C

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.16271    0.47333  15.132  < 2e-16 ***
expPLEKHO1_gdsc     0.15593    0.01805   8.639  < 2e-16 ***
expTHEM4_gdsc      -0.11133    0.02914  -3.820 0.000148 ***
expOSBP_gdsc       -0.22318    0.06004  -3.717 0.000221 ***
expOSBPL3_gdsc     -0.06131    0.02505  -2.447 0.014684 *
expOSBPL10_gdsc    -0.08570    0.04139  -2.071 0.038847 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6581 on 576 degrees of freedom
Multiple R-squared:  0.2752,         Adjusted R-squared:  0.2689
F-statistic: 43.74 on 5 and 576 DF,  p-value: < 2.2e-16
```

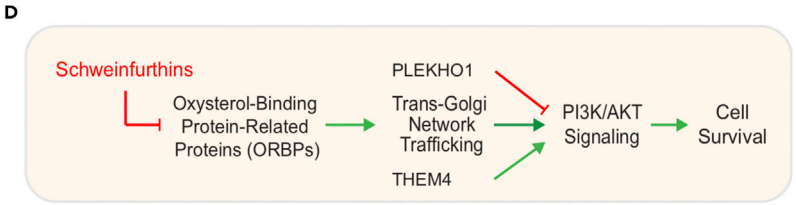



**Figure 6. A Multivariate Model of Schweinfurthin A Drug Activity**

(A) Reproducibility of the data for the two schweinfurthin derivatives tested in the GDSC.

(B) Heatmap indicating Schweinfurthin A drug activity and candidate determinant gene expression in the 100 most sensitive and resistant non-hematopoietic GDSC cell lines.

(C) A statistical summary of a multivariate linear model of Schweinfurthin A response in the GDSC.

(D) Scheme of the proposed molecular pharmacology of the schweinfurthins. Schweinfurthins have been shown to inhibit PI3K/AKT signaling and cell survival by binding oxysterol-binding-protein-related proteins (ORPs) to disrupt *trans*-Golgi

**Figure 6. *Continued***

network trafficking required for robust pathway activity (Bao et al., 2015). Together with the ORPs OSBP, OSBPL3, and OSBPL10, the other candidate determinants, PLEKHO1 and THEM4, have also been implicated in PI3K/AKT signaling (Liu et al., 2013; Tokuda et al., 2007).

Plots and analyses in panels B–D are based on non-hematopoietic GDSC cell lines.

CellMinerCDB to reveal the molecular pathways for response. After confirming that the activities of both compounds were highly correlated ($R = 0.87$, $p = 8.8 \times 10^{-182}$, $N = 585$, Figure 6A), we explored the genomic correlates with activity for the ≈700 GDSC cell lines tested. The CellMinerCDB Univariate Analysis tool ("Compare Patterns" tab) indicates that the leading molecular correlate (by lowest p value) of schweinfurthin activity is the expression of *PLEKHO1*, a negative regulator of phosphatidylinositol 3-kinase (PI3K)/AKT signaling ($R = 0.47$, $p = 1.95 \times 10^{-33}$, $N = 582$) (Tokuda et al., 2007). This result is consistent with a recent study showing that schweinfurthins inhibit mammalian target of rapamycin (mTOR)/AKT signaling by interfering with *trans*-Golgi network (TGN) trafficking (Bao et al., 2015). In particular, schweinfurthins bind to oxysterol-binding proteins, which regulate TGN trafficking (Burgett et al., 2011; Mesmin et al., 2017), thereby arresting lipid-raft-mediated PI3K activation and functional mTOR/RheB complex formation.

Next, using the multivariate analysis feature of CellMinerCDB ("Regression Models" tab; https://discover.nci.nih.gov/cellminercdb/), we developed a linear predictive model for schweinfurthin response integrating the expressions of *PLEKHO1*, *THEM4*, a positive regulator of AKT signaling (Liu et al., 2013), and the genes encoding the oxysterol-binding protein family members *OSBP*, *OSBPL3*, and *OSBP10* (Figures 6C and 6D). Increased expression of the oxysterol-binding protein family members conceivably sustains TGN trafficking and PI3K/AKT signaling, in keeping with their negative regression coefficient weights as resistance determinants in the model. The negative weighting of *THEM4* expression and positive weighting of *PLEKHO1* expression are similarly consistent with their respective roles in activating and suppressing PI3K/AKT signaling. These analyses give molecular insight into the cholesterol trafficking and intracellular membrane pathways as the targets of schweinfurthins and open new opportunities for testing the potential activity of schweinfurthins with genomic and molecular signatures.

## Relating EMT Status with Gene Expression to Identify *LIX1L* Expression and Schweinfurthin Activity in Mesenchymal Cells

EMT is a fundamental process in development, wound healing, and cancer progression, characterized by the loss of cell-cell adhesion and the acquisition of motile and invasive properties (Figure 7A) (Kalluri and Weinberg, 2009; Lamouille et al., 2014). EMT is driven by dominant transcription factors, including ZEB1/2, SNAI1/2, and TWIST1/2, and is reversible through a continuum of states from epithelial to mesenchymal. These attributes have motivated the development of gene-expression-based EMT signatures to identify cell line state and understand drug resistance.

We applied a 37-gene EMT signature initially developed in the NCI-60 (Kohn et al., 2014) to derive a numerical index of EMT status as a weighted sum of cell-line-specific EMT expression values (Transparent Methods). Figure 7B shows the bimodal distribution for the EMT index values across the 821 non-hematopoietic GDSC cell lines, allowing cell line stratification into epithelial, mesenchymal, and epithelial-mesenchymal categories. EMT stratification within particular tissues of origin also shows a substantial proportion of intermediate epithelial-mesenchymal lines in liver, ovary, and lung cancer cell lines (Figure 7C). The numerical EMT index is available for all CellMinerCDB-integrated data sources as the variable KOHN_EMT_PC1 ("Metadata"), allowing its correlation with any chosen molecular or drug response feature.

The EMT index identified *LIX1L* as a novel mesenchymal gene whose expression is highly correlated with the EMT index signature (Kohn et al., 2014) across multiple cancer types (Figure 7D, $R = -0.75$, $p = 8.9 \times 10^{-179}$, $N = 823$). *LIX1L* is also broadly expressed in TCGA tumor samples (Figure S10A, related to Figure 7). Knockdown analyses in the breast cancer MDA-MB-231 cell line suggest that *LIX1L* expression reduces cell migration and invasiveness (Figures 7F–7H, S10C, and S10D).

We also correlated the EMT index with the activity profiles of the 297 compounds in the GDSC database, including the 19 additional compounds tested for the current study (Figure S11, related to Figure 7). Schweinfurthin A is the strongest negative correlate, indicating its selective antiproliferative activity in mesenchymal cancer cell lines, such as those derived from bone or soft tissue (Figure S11C). The second strongest negative correlate with the EMT index is the RHO-associated kinase 1 inhibitor GSK269962A

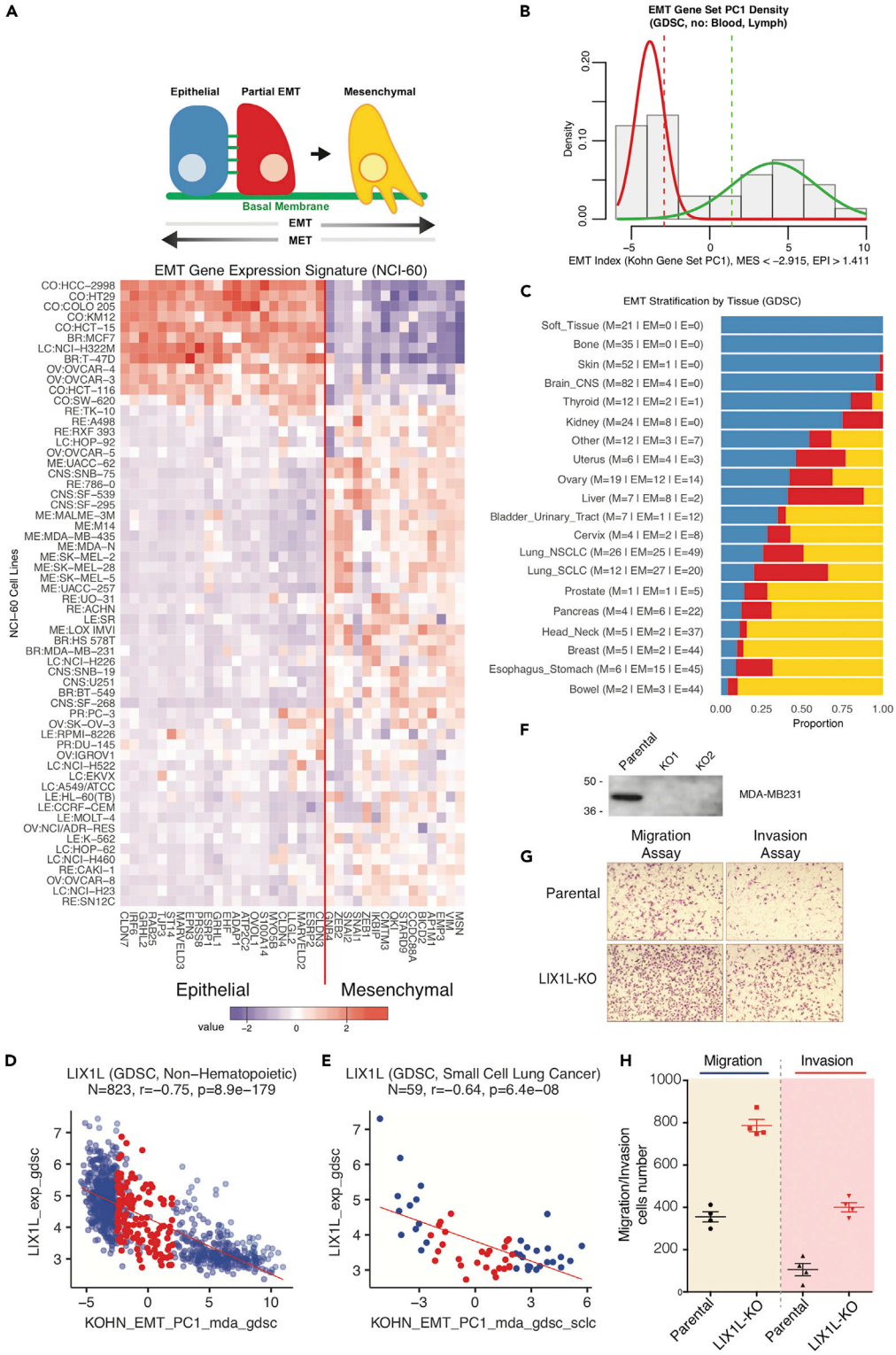



**Figure 7. Relating Epithelial Mesenchymal Transition (EMT) Status with Gene Expression to Identify *LIX1L* as a Novel EMT Gene**

(A) A 37-gene EMT signature developed in (Kohn et al., 2014) was used to derive a numerical index of EMT status as a weighted sum of cell-line-specific EMT gene expression values (see Transparent Methods for details). Epithelial and mesenchymal statuses are associated with positive and negative index values, respectively.

**Figure 7. *Continued***

(B) For 821 non-hematopoietic cell lines in the GDSC collection, the EMT index values show a bimodal distribution, which can be modeled as a normal mixture. Cell lines with EMT index values less than (greater than) 1 standard deviation above (below) the putative mesenchymal (epithelial) group mean are annotated as mesenchymal (epithelial).

(C) EMT stratification by tissue of origin.

(D and E) Expression of *LIX1L*, a novel mesenchymal gene, is strongly correlated with the EMT index signature. "Epithelial-mesenchymal" lines with intermediate EMT index values are indicated in red. Mesenchymal lines are in blue at the left, and epithelial are in blue at the right.

(F) Western blot showing the efficient knockdown of LIX1L in MDA-MB231 cells.

(G) Representative image showing increased migration and invasion after LIX1L knockdown.

(H) Quantitation of the increased migration and invasion of cells after LIX1L knockdown. Individual experiments are shown as dots. Error bars indicate the standard error of the mean.

(Figure S11C), whose target regulates actin dynamics and cell motility associated with EMT (Kalluri and Weinberg, 2009; Lamouille et al., 2014). On the opposite side, the drug most highly correlated with epithelial cell line status was acetalax (oxyphenisatin), with its two independent samples (NSC59687 and 614826) at the top of the list above afatinib and lapatinib (Figure S11C), consistent with its potential activity in epithelial breast cancer cells (Morrison et al., 2013).

## DISCUSSION

CellMinerCDB (https://discover.nci.nih.gov/cellminercdb/) allows researchers to interact directly with an unparalleled breadth of cancer cell line genomic and pharmacologic data. The examples described here, spanning data assessment, integration, and discovery, demonstrate the value of working within and across data sources.

The CellMinerCDB analyses support the maturity and essential reproducibility of most molecular profiling technology platforms, such as gene transcript expression and DNA CNV. Mutation data are prominently featured in translational studies, and CellMinerCDB exposes the issue of discrepancies between matched cell line mutation profiles across sources. This provides a foundation for understanding and mitigating sources of variability, which reflect the ongoing technical challenges and trade-offs with the acquisition and interpretation of genome sequencing data. Somatic variant calling in cancer cell lines is inherently challenging because of the absence of matched normal tissue for comparison, as well as the potentially higher mutation burden in cell lines relative to primary tumor tissue. One approach for excluding potentially cell-line-specific passenger mutations is to filter variants based on frequency in patient populations (Iorio et al., 2016). Variability in cell line mutation data across sources may also arise from differences in variant calling algorithms, as well as data sources used for filtering likely germline variants. Indeed, the reproducibility of matched cell line gene expression, DNA copy number, and methylation signatures across databases (see Figures 2A, S1, and S2) indicates that the mutation data inconsistencies are likely technical. An existing strategy for acquiring more robust mutation data is to pursue higher-depth targeted sequencing of a restricted gene set. Indeed, we noted that the CCLE data, derived from the latter approach, consistently identified more mutant cell lines for prominent oncogenes and tumor suppressor genes than genome-wide exome sequencing (as in the NCI-60 and GDSC databases). CellMinerCDB can directly integrate mutation data across sources to identify cell lines with consistent mutation calls for a gene of interest, together with other potentially impacted lines. Given the primarily technical basis for the mutation data discrepancies, the best course for users remains this sort of comparison of data across sources. CellMinerCDB enables researchers to focus on consistently mutation-impacted cell lines for further bioinformatic analyses and experimental use (including targeted high-depth sequencing for specific genes of interest).

Regarding drug reproducibility, assays across the major cancer cell line data sources measure different biochemical features at different time points. Still, CellMinerCDB demonstrates significant concordance between drug activity data generated at the NCI (NCI-60 and NCI/DTP-SCLC), at the Broad Institute (CCLE, CTRP), and at the Sanger Institute and Massachusetts General Hospital (GDSC), including data for widely used anticancer drugs. The cross-database comparison features of CellMinerCDB allow researchers to explore potential drug reproducibility issues and focus on drugs with reliable data. Scatterplots of matched cell line activity data can highlight problem areas with particular assays, such as inappropriate concentration ranges, as illustrated for fulvestrant in the NCI/DTP assay.

In addition, among the 19 NCI-60 drugs tested for reproducibility and expansion of genomic correlates, we found significant consistency for 16 drugs, including the novel non-camptothecin topoisomerase I inhibitor

LMP744 (Burton et al., 2018), and identified the FDA-approved laxatives oxyphenisatin acetate (acetalax) and bisacodyl as potential novel anticancer drug candidates. CellMinerCDB shows (https://discover.nci.nih.gov/cellminercdb/) that oxyphenisatin exhibits a wide range of concordant antiproliferative activity across the NCI-60 cell lines and across the 710 GDSC cell lines tested, being substantially more active in triple-negative breast cancer lines relative to other cancer drugs tested on the GDSC panel. These findings suggest the potential of oxyphenisatin derivatives for repurposing as anticancer drugs.

With both drug activity reproducibility and broader associations between molecular features, such as *CDKN2A* expression and gene copy/methylation, we noted that the NCI-60 could effectively capture relationships evident in larger cell line sets. The latter better reflect tissue type diversity and context-specific molecular features. Still, for dominant associations, such as *SLFN11* expression and DNA-targeted drug responses, representative cell line sets such as the NCI-60 are often sufficient (Zoppoli et al., 2012). In addition, the NCI-60 provides drug responses for over 21,000 individual agents, making it an unmatched resource for the discovery of new chemotypes based on correlations with genomic data and response patterns for drugs with known targets. CellMinerCDB makes correlation-based COMPARE analyses readily accessible for drug discovery (https://dtp.cancer.gov/databases_tools/compare.htm [Paull et al., 1989]). It also enables a direct visualization of activity data for compounds retrieved in the analysis together with data for the queried entity, using the "Univariate Analysis" tool (https://discover.nci.nih.gov/cellminercdb/). The NCI databases are also a tractable starting point for molecular data expansion with leading-edge technologies. RNA sequencing data with isoform-specific transcript expression and SWATH mass spectrometry-based protein expression data have been generated for the NCI-60, and will be made available within CellMinerCDB (Shao et al., 2015). We are committed to sustain development of CellMinerCDB as an ongoing resource in the mold of the existing NCI CellMiner data portal, which has steadily integrated new data and analyses since its inception (Reinhold et al., 2012, 2014, 2015, 2017). These developments will expand the current features of CellMinerCDB. For example, as existing and emerging data sources provide novel proteomic and isoform-specific transcript expression data, we are planning to integrate these with regular updates of the same website.

CellMinerCDB (https://discover.nci.nih.gov/cellminercdb/) ultimately aims to provide a seamless platform for data exploration and hypothesis generation, integrating previously isolated data sources and enhancing their interpretation through the intuition and expertise of experimental scientists and clinicians. The present publication provides only a sample of the potential uses of CellMinerCDB. CellMinerCDB uniquely complements existing data portals that provide detailed information on their associated data, together with specialized analyses. By empowering researchers to easily build on the strengths of individual databases and pursue their own questions, CellMinerCDB aims to advance the potential of cancer cell line pharmacogenomic data to lay the foundation, validate, and focus experimental and, ultimately, clinical drug development and precision medicine.

## Limitations of the Study

We see CellMinerCDB as primarily a data exploration and hypothesis generation tool. Its selection of analyses reflects what is practically manageable in this context, both computationally and conceptually. For example, we do not provide analyses with extended runtimes that are less suitable for interactive data exploration. We do, however, make all data integrated within the application easily downloadable, to expedite more specialized or computationally intensive analyses. CellMinerCDB still provides interactive access to the most fundamental methods, including regression-based predictive models, which have been prominently featured in highly cited studies of cancer cell line pharmacogenomic data.

We have also attempted to minimize the conceptual barrier for basic exploratory analyses by making reasonable default choices for this setting. In keeping with existing CellMiner tools (http://discover.nci.nih.gov/cellminer) (Reinhold et al., 2012, 2014, 2015, 2017) and leading studies, we use Pearson's correlations to measure association between molecular or drug response variables. We do note for users that statistical significance results for these correlations presuppose approximately multivariate normal data; substantial deviations from this assumption can be readily noted through the provided scatterplots. Still, a comparative study of CCLE and GDSC data favored Pearson correlations over non-parametric Spearman correlations, showing that the latter often failed to detect patterns in which responses are restricted to a relatively small fraction of cell lines (as will often be the case for pathway-targeted drugs) (Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium, 2015).

In keeping with the exploratory focus of CellMinerCDB, we do not enforce formal adjustments for multiple hypothesis testing (although our pattern comparison and related results do include tabulation of q-values to allow false-discovery-rate-based filtering). The strictest adjustment for multiple testing would require inaccessible knowledge of all analyses conducted by a user toward a particular aim. This sort of application of statistical filters would likely exclude many experimentally established relationships that involve more than one determinant.

Our goal is to strike a balance between providing statistical measures (with reasonable caveats), and allowing scientific experts to apply their judgment when exploring data. We finally note that it is possible to consider additional levels of inter-source data integration. For example, pooling molecular or drug data for distinct cell lines across sources (as compared to strictly overlapping cell lines) could increase the power of statistical analyses. This approach, although potentially valuable, would require careful assessment and adjustment for source-specific effects and is outside the scope of the current study.

## METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, 11 figures, and 5 tables and can be found with this article online at https://doi.org/10.1016/j.isci.2018.11.029.

## ACKNOWLEDGMENTS

We would like to thank Dr. 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. The work was supported by the Center for Cancer Research, Intramural Program of the National Cancer Institute (Z01 BC006150 to Y.P.), Ruth L. Kirschstein National Research Service Award (F32 CA192901 to A.L.), and the National Resource for Network Biology (NRNB) from the National Institute of General Medical Sciences (NIGMS) (P41 GM103504 to C.S.). M.J.G was funded by the Wellcome Trust (086375 and 102696). This study was also supported by fellowships from the Japanese Society of Clinical Pharmacology and Therapeutics and the Japan Society for the Promotion of Science (to M.Y).

## AUTHOR CONTRIBUTIONS

Conception, design, and development, V.N.R., A.L., F.E., L.L., W.C.R., and Y.P.; acquisition and preparation of data, V.N.R., S.V., M.S., F.I., C.H.B, M.J.G., and W.C.R.; analysis and interpretation of data, development of results, V.N.R., A.L., M.Y., S.V., F.G.S., M.I.A., A.T., K.W.K., C.H.B., M.G., W.C.R., and Y.P.; experimental validation studies, M.Y., K.W.K, and Y.P.; writing, review, and revision of the manuscript, V.N.R., A.L., M.Y., F.I., F.G.S., M.I.A., A.T., C.S., C.H.B., M.G., W.C.R., and Y.P.; study supervision, W.C.R.and Y.P.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 17, 2018
Revised: October 11, 2018
Accepted: November 15, 2018
Published: December 12, 2018

## REFERENCES

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.

Antony, S., Jayaraman, M., Laco, G., Kohlhagen, G., Kohn, K.W., Cushman, M., and Pommier, Y. (2003). Differential induction of topoisomerase I-DNA cleavage complexes by the indenoisoquinoline MJ-III-65 (NSC 706744) and camptothecin: base sequence analysis and activity against camptothecin-resistant topoisomerases I. Cancer Res. *63*, 7428–7435.

Bao, X., Zheng, W., Hata Sugi, N., Agarwala, K.L., Xu, Q., Wang, Z., Tendyke, K., Lee, W., Parent, L., Li, W., et al. (2015). Small molecule schweinfurthins selectively inhibit cancer cell proliferation and mTOR/AKT signaling by interfering with trans-Golgi-network trafficking. Cancer Biol. Ther. *16*, 589–601.

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.

Boehm, J.S., and Golub, T.R. (2015). An ecosystem of cancer cell line factories to support a cancer dependency map. Nat. Rev. Genet. *16*, 373–374.

Burgett, A.W., Poulsen, T.B., Wangkanont, K., Anderson, D.R., Kikuchi, C., Shimada, K., Okubo, S., Fortner, K.C., Mimaki, Y., Kuroda, M., et al. (2011). Natural products reveal cancer cell dependence on oxysterol-binding proteins. Nat. Chem. Biol. *7*, 639–647.

Burton, J.H., Mazcko, C.N., LeBlanc, A.K., Covey, J.M., Ji, J.J., Kinders, R.J., Parchment, R.E., Khanna, C., Paoloni, M., Lana, S.E., et al. (2018). NCI comparative oncology program testing of non-camptothecin indenoisoquinoline topoisomerase i inhibitors in naturally occurring canine lymphoma. Clin. Cancer Res. https://doi.org/10.1158/1078-0432.CCR-18-1498.

Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84–87.

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.

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.

Han, T., Goralski, M., Gaskill, N., Capota, E., Kim, J., Ting, T.C., Xie, Y., Williams, N.S., and Nijhawan, D. (2017). Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15. Science *356*, https://doi.org/10.1126/science.aal3755.

Haverty, P.M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., Neve, R.M., Martin, S., Settleman, J., Yauch, R.L., et al. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. Nature *533*, 333–337.

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.

Kalluri, R., and Weinberg, R.A. (2009). The basics of epithelial-mesenchymal transition. J. Clin. Invest. *119*, 1420–1428.

Kodet, J.G., Beutler, J.A., and Wiemer, D.F. (2014). Synthesis and structure activity relationships of schweinfurthin indoles. Bioorg. Med. Chem. *22*, 2542–2552.

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.

Lamouille, S., Xu, J., and Derynck, R. (2014). Molecular mechanisms of epithelial-mesenchymal transition. Nat. Rev. Mol. Cell Biol. *15*, 178–196.

Liu, Y.-P., Liao, W.-C., Ger, L.-P., Chen, J.-C., Hsu, T.-I., Lee, Y.-C., Chang, H.-T., Chen, Y.-C., Jan, Y.-H., Lee, K.-H., et al. (2013). Carboxyl-terminal modulator protein positively regulates Akt phosphorylation and acts as an oncogenic driver in breast cancer. Cancer Res. *73*, 6194–6205.

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.

Mesmin, B., Bigay, J., Polidori, J., Jamecna, D., Lacas-Gervais, S., and Antonny, B. (2017). Sterol transfer, PI4P consumption, and control of membrane lipid order by endogenous OSBP. EMBO J. *36*, 3156–3174.

Mollaoglu, G., Guthrie, M.R., Böhm, S., Brägelmann, J., Can, I., Ballieu, P.M., Marx, A., George, J., Heinen, C., Chalishazar, M.D., et al. (2017). MYC drives progression of small cell lung cancer to a variant neuroendocrine subtype with vulnerability to aurora kinase inhibition. Cancer Cell *31*, 270–285.

Morrison, B.L., Mullendore, M.E., Stockwin, L.H., Borgel, S., Hollingshead, M.G., and Newton, D.L. (2013). Oxyphenisatin acetate (NSC 59687) triggers a cell starvation response leading to autophagy, mitochondrial dysfunction, and autocrine TNFalpha-mediated apoptosis. Cancer Med. *2*, 687–700.

Murai, J., Feng, Y., Yu, G.K., Ru, Y., Tang, S.-W., Shen, Y., and Pommier, Y. (2016). Resistance to PARP inhibitors by SLFN11 inactivation can be overcome by ATR inhibition. Oncotarget *7*, 76534–76550.

Murai, J., Tang, S.-W., Leo, E., Baechler, S.A., Redon, C.E., Zhang, H., Al Abo, M., Rajapakse, V.N., Nakamura, E., Jenkins, L.M.M., et al. (2018). SLFN11 blocks stressed replication forks independently of ATR. Mol. Cell *69*, 371–384.e6.

Paull, K.D., Shoemaker, R.H., Hodes, L., Monks, A., Scudiero, D.A., Rubinstein, L., Plowman, J., and Boyd, M. (1989). Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of a mean graph and COMPARE algorithm. J. Natl. Cancer Inst. *81*, 1088–1092.

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*, https://doi.org/10.1093/jnci/djw122.

Pommier, Y., Sun, Y., Huang, S.N., and Nitiss, J.L. (2016). Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nat. Rev. Mol. Cell Biol. *17*, 703–721.

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.

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.

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.

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.

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.

Shao, S., Koh, C.C., Gillessen, S., Joerger, M., Jochum, W., and Aebersold, R. (2015). Minimal sample requirement for highly multiplexed protein quantification in cell lines and tissues by PCT-SWATH mass spectrometry. Proteomics *5*, 3711–3721.

Thornburg, C.C., Britt, J.R., Evans, J.R., Akee, R.K., Whitt, J.A., Trinh, S.K., Harris, M.J., Thompson, J.R., Ewing, T.L., Shipley, S.M., et al. (2018). NCI program for natural product discovery: a publicly-accessible library of natural product fractions for high-throughput screening. ACS Chem. Biol. *13*, 2484–2497.

Tokuda, E., Fujita, N., Oh-hara, T., Sato, S., Kurata, A., Katayama, R., Itoh, T., Takenawa, T., Miyazono, K., and Tsuruo, T. (2007). Casein kinase 2–interacting protein-1, a novel Akt pleckstrin homology domain-interacting protein, down-regulates PI3K/Akt signaling and suppresses tumor growth in vivo. Cancer Res. *67*, 9666–9676.

Wagner, P.L., Stiedl, A.-C., Wilbertz, T., Petersen, K., Scheble, V., Menon, R., Reischl, M., Mikut, R., Rubin, M.A., Fend, F., et al. (2011). Frequency and clinicopathologic correlates of KRAS amplification in non-small cell lung carcinoma. Lung Cancer *74*, 118–123.

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.

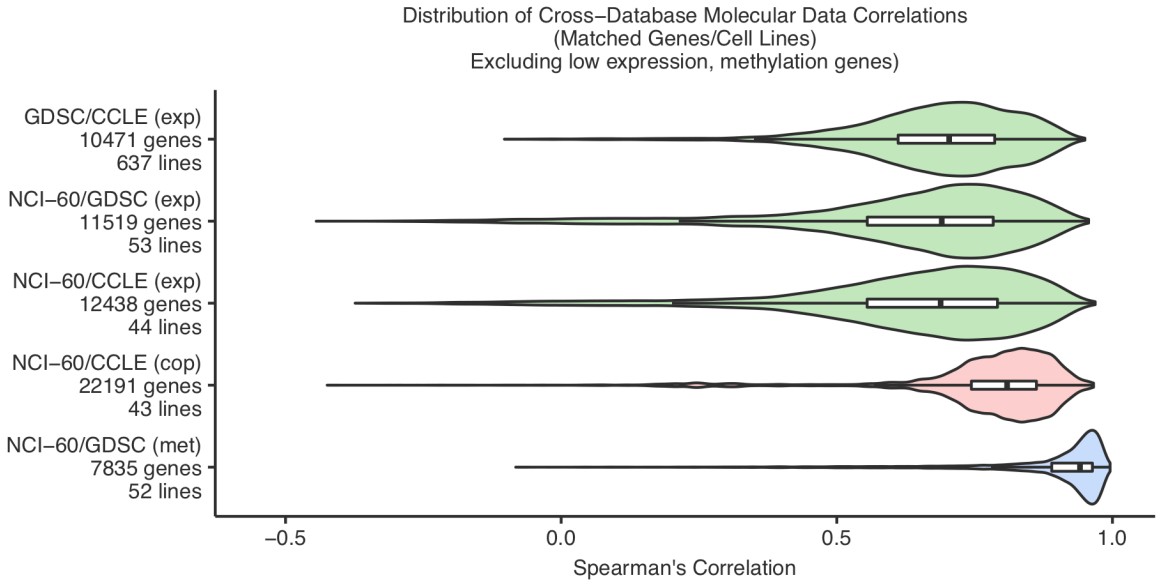
ISCI, Volume *10*

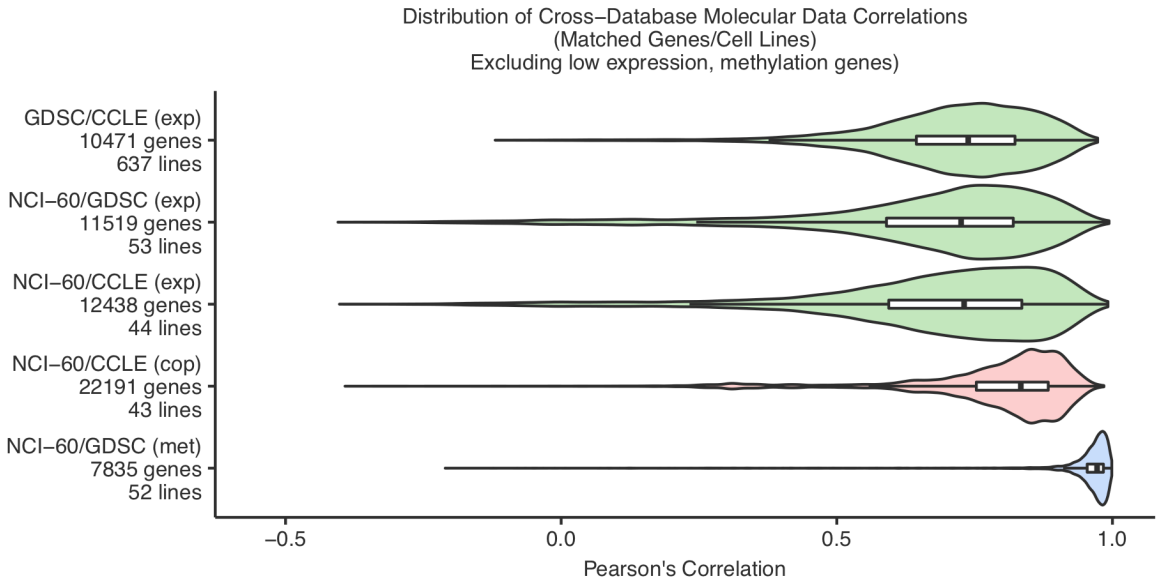
## Supplemental Information

# 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**

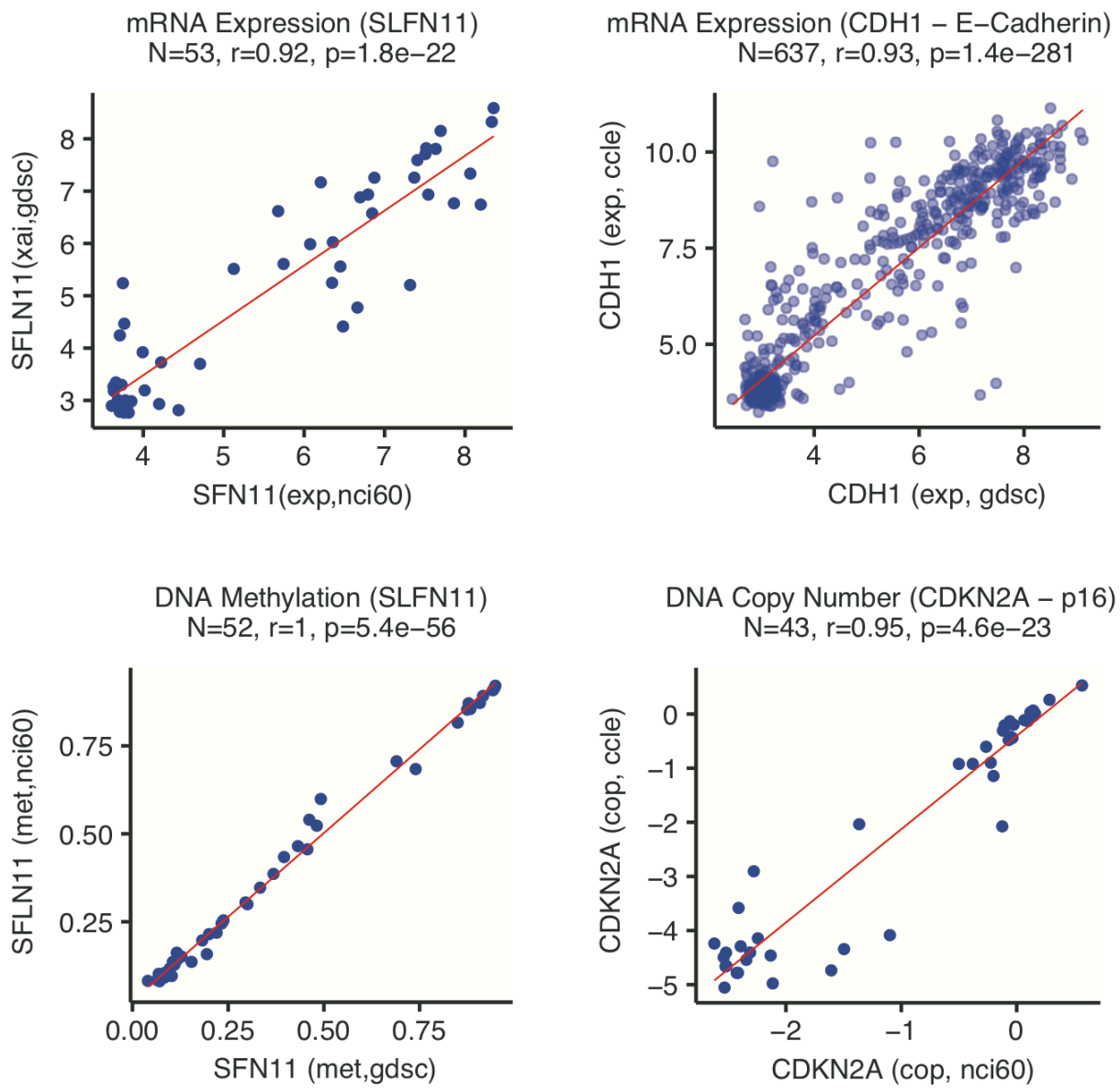
## Supplemental Information - Figures:



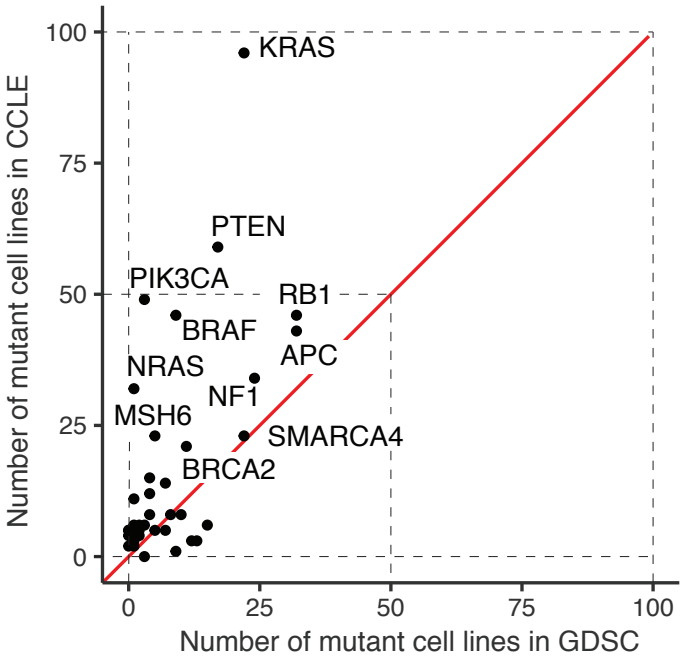




| Comparison | Median Correlation (Spearman) | Median Correlation (Pearson) |
|---|---|---|
| GDSC/CCLE (exp) | 0.704 | 0.738 |
| NCI-60/CCLE (cop) | 0.809 | 0.834 |
| NCI-60/CCLE (exp) | 0.688 | 0.731 |
| NCI-60/GDSC (exp) | 0.690 | 0.725 |
| NCI-60/GDSC (met) | 0.941 | 0.972 |

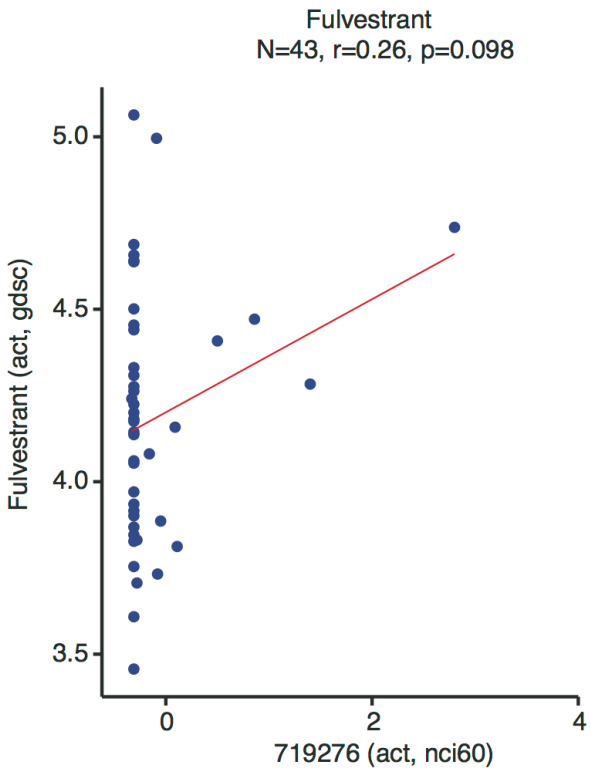
**Supplementary Figure 1, Related to Figure 2**: Spearman's and Pearson's correlation distributions for comparable, matched cell line transcript expression (exp), DNA copy number (cop), and DNA methylation (met) data across CellMinerCDB-integrated data sources.



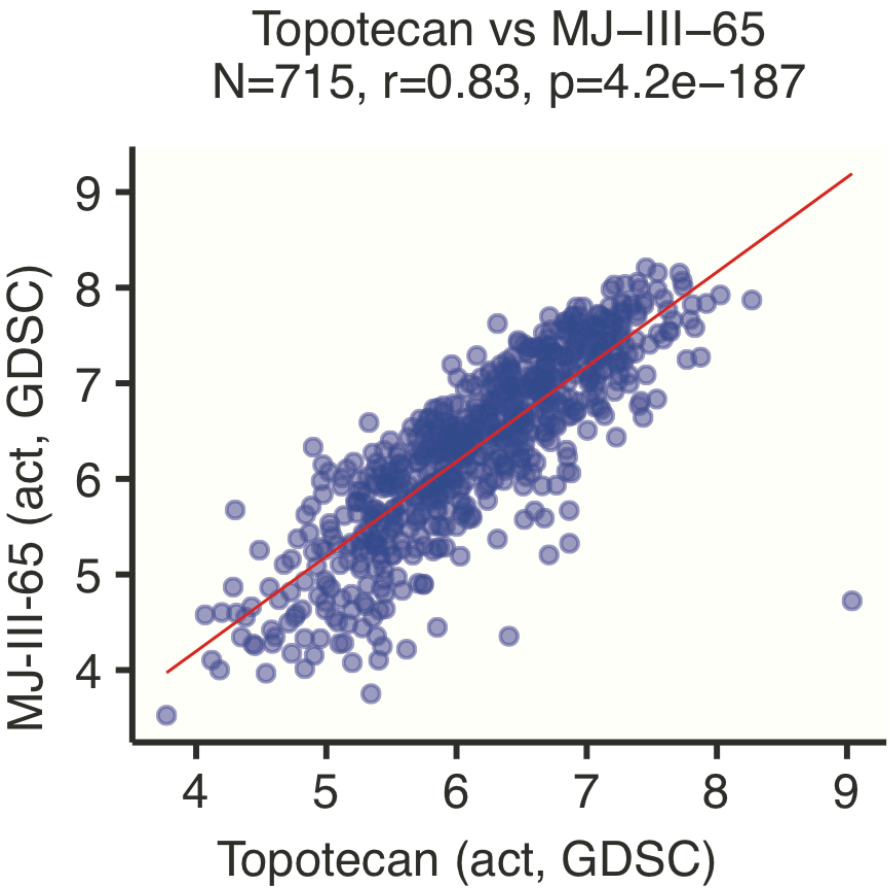

**Supplementary Figure 2, Related to Figure 2**: Inter-source data reproducibility examples for selected genes and molecular data types.



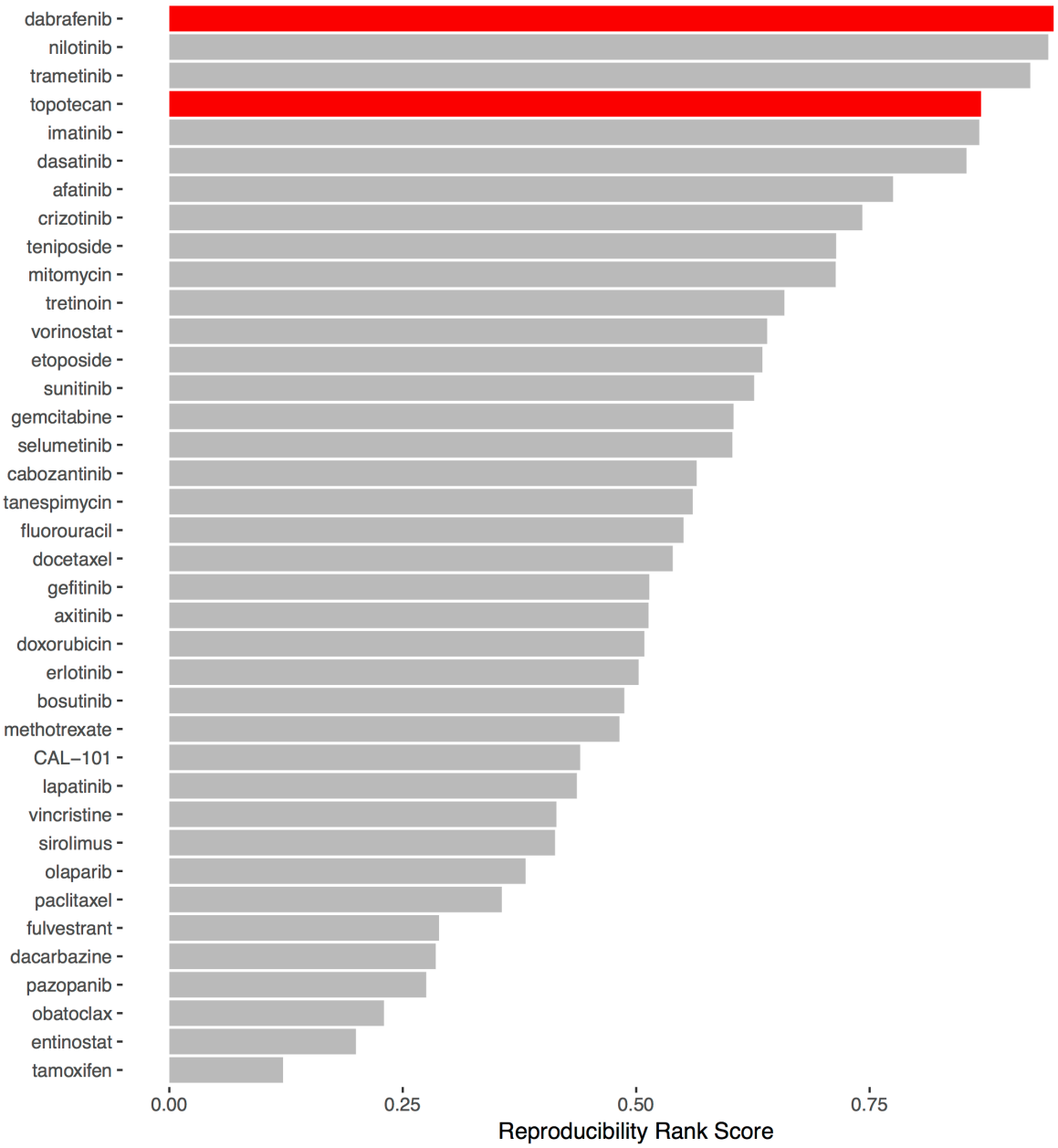

**Supplementary Figure 3, Related to Figure 2:** Importance of sequencing depth for retrieving mutant cell lines. CCLE vs. GDSC mutant cell line counts for selected oncogenes and tumor suppressor genes.



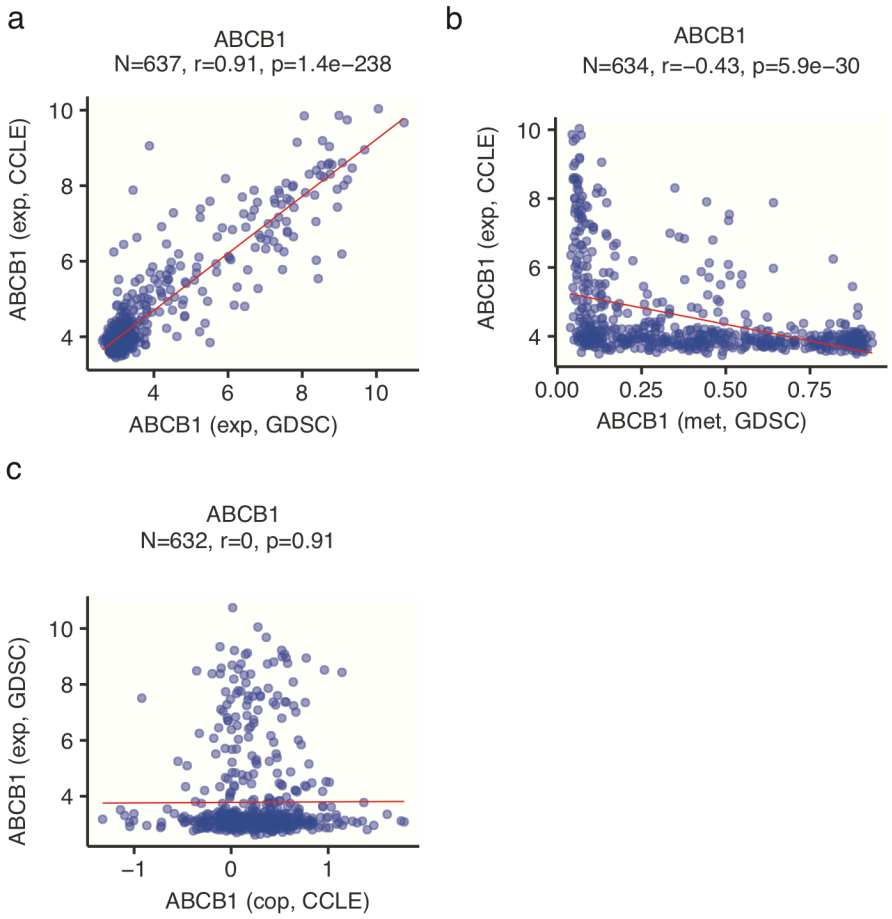

**Supplementary Figure 4, Related to Figure 3:** GDSC versus NCI-60 drug activity for fulvestrant, indicating inappropriate drug concentration range in NCI-60 activity assay.




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

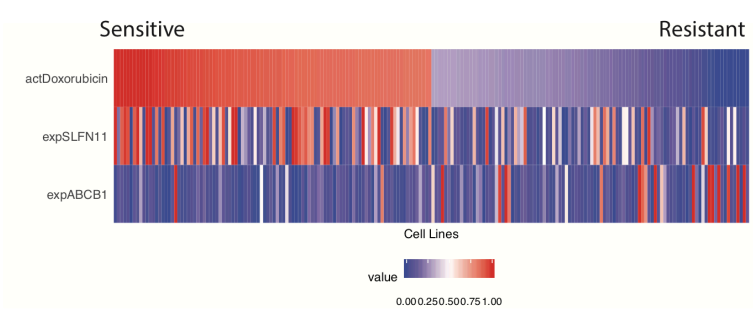



**Supplementary Figure 6, Related to Figure 3:** Overall drug activity data reproducibility rankings for 38 compounds tested in the NCI-60, GDSC, and CTRP, integrating pairwise activity correlations between the sources.




**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



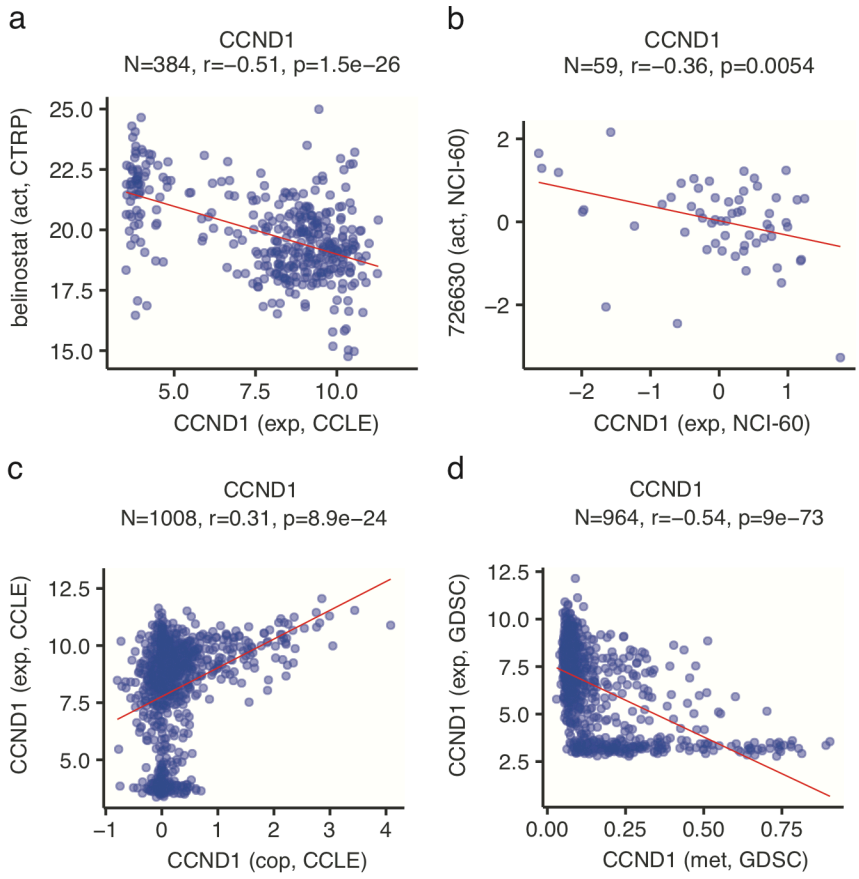

b

```
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
```

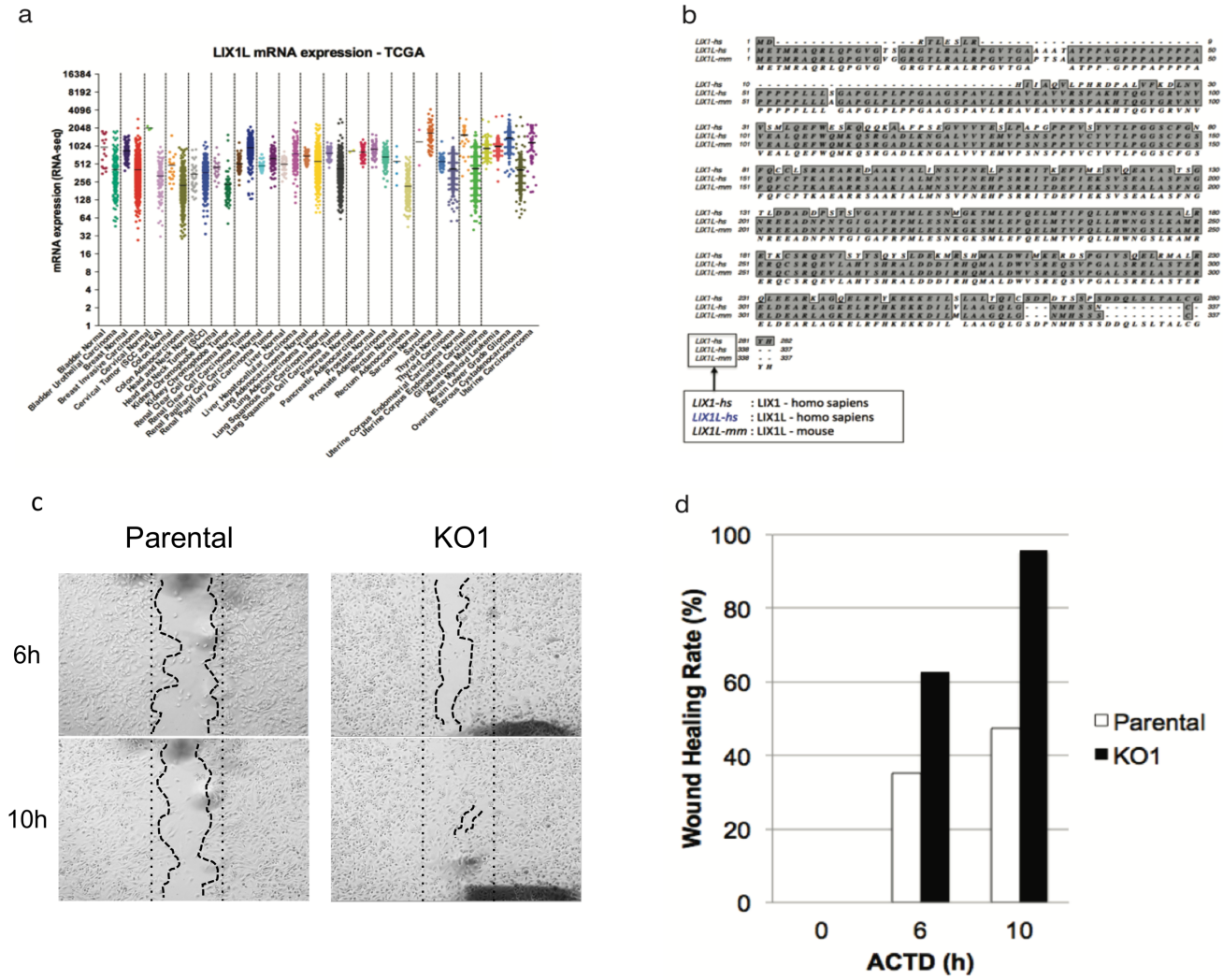
c

```
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
```

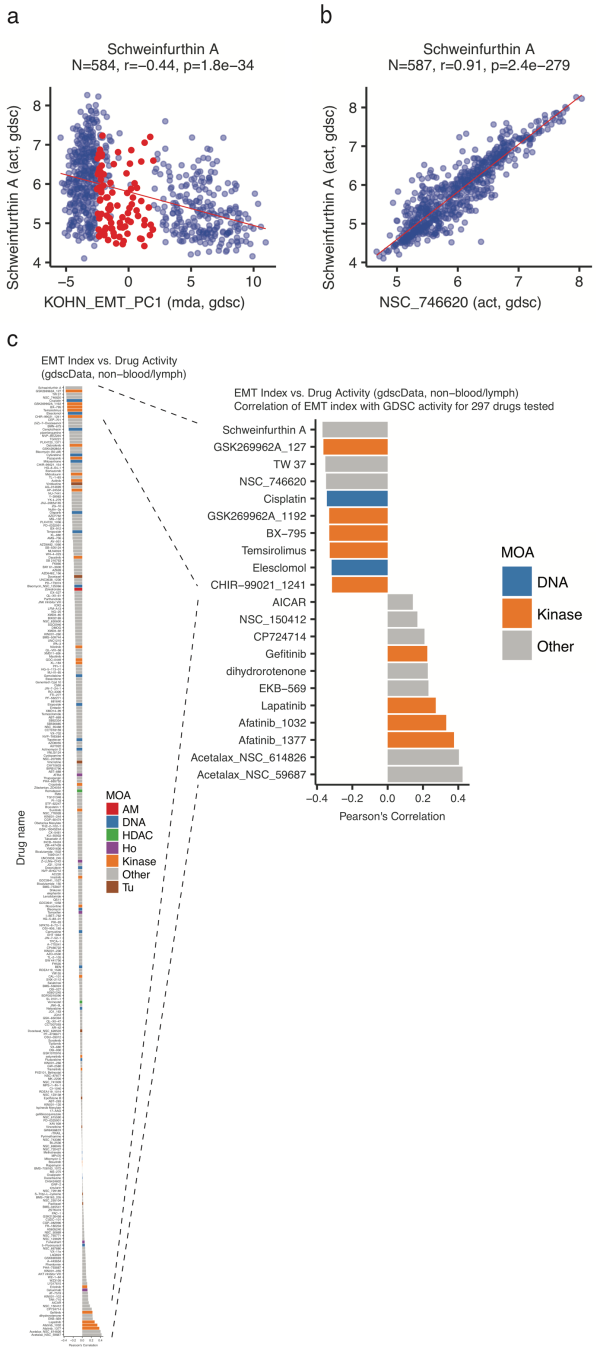
**Supplementary Figure 8, Related to Figure 5:** ABCB1 expression complements SLFN11 expression in predicting doxorubicin drug activity in GDSC cell lines (b, c), with high ABCB1 expression evident in several highly resistant cell lines indicated at the right of the heatmap in (a).




**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.