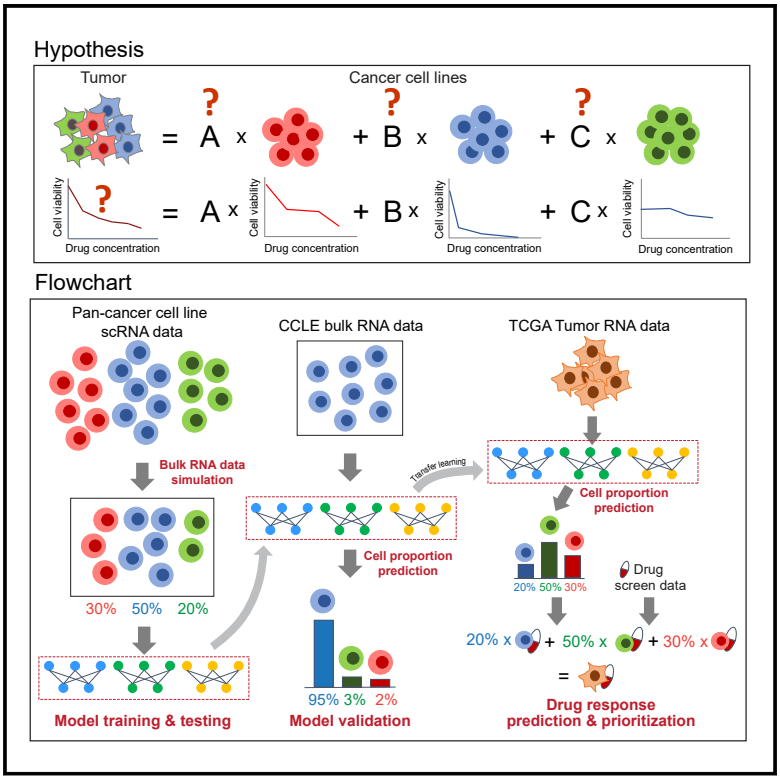
Patterns

Predicting drug response through tumor deconvolution by cancer cell lines

Graphical abstract



Highlights

- Scaden-CA is a deep learning model for tumor deconvolution
- Tumor deconvolution facilitates the use of a drug response prediction algorithm
- The model explores drug response mechanisms via DNA mutations and/or expression changes

Authors

Yu-Ching Hsu, Yu-Chiao Chiu, Tzu-Pin Lu, Tzu-Hung Hsiao, Yidong Chen

Correspondence

d93921032@gmail.com (T.-H.H.), cheny8@uthscsa.edu (Y.C.)

In brief

In this study, the authors sought to identify potential new applications of existing anti-cancer drugs beyond their original medical indications. A deep learning-based deconvolution model and drug response prediction method using cell line data were developed and then transferred to deconvolute tumor samples. By deconvoluting tumors into cancer-type-specific cell lines, the model predicted drug responses of tumors using the proportions and drug sensitivity data from those cell lines. The findings resulted in suggestions for future investigation of drug repurposing.

Hsu et al., 2024, Patterns 5, 100949
April 12, 2024
https://doi.org/10.1016/j.patter.2024.100949

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Predicting drug response through tumor deconvolution by cancer cell lines

Yu-Ching Hsu,[1,2,3,4] Yu-Chiao Chiu,[5,6] Tzu-Pin Lu,[3] Tzu-Hung Hsiao,[7,*] and Yidong Chen[4,8,9,*]
[1]Bioinformatics Program, Taiwan International Graduate Program, National Taiwan University, Taipei 115, Taiwan
[2]Bioinformatics Program, Institute of Statistical Science, Taiwan International Graduate Program, Academia Sinica, Taipei 115, Taiwan
[3]Institute of Health Data Analytics and Statistics, Department of Public Health, College of Public Health, National Taiwan University, Taipei 100, Taiwan
[4]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[5]Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15261, USA
[6]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15232, USA
[7]Department of Medical Research, Taichung Veterans General Hospital, Taichung 40705, Taiwan
[8]Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA
[9]Lead contact
*Correspondence: d93921032@gmail.com (T.-H.H.), cheny8@uthscsa.edu (Y.C.)
https://doi.org/10.1016/j.patter.2024.100949

THE BIGGER PICTURE Drug repurposing involves utilizing approved or investigational drugs beyond the scope of the original medical application. This approach offers several advantages compared with developing entirely new drugs, including a lower risk of failure, a shorter development timeline, and lower investment costs. Using previous knowledge of drug sensitivity of cancer cell lines, we created a tumor convolution model to predict how tumors will respond to anti-cancer drugs. In this approach, a deep learning model used single-cell gene expression profiles to deconvolute tumors into their constituent cancer cell lines; in other words, the model represented tumors as a mixture of different cancer-type-specific cell lines in varying proportions. Subsequently, the deconvoluted proportions facilitated the prediction of tumors' drug responses. Ultimately, these observations highlight the potential for deep learning applications in the area of drug repurposing.

SUMMARY

Large-scale cancer drug sensitivity data have become available for a collection of cancer cell lines, but only limited drug response data from patients are available. Bridging the gap in pharmacogenomics knowledge between *in vitro* and *in vivo* datasets remains challenging. In this study, we trained a deep learning model, Scaden-CA, for deconvoluting tumor data into proportions of cancer-type-specific cell lines. Then, we developed a drug response prediction method using the deconvoluted proportions and the drug sensitivity data from cell lines. The Scaden-CA model showed excellent performance in terms of concordance correlation coefficients (>0.9 for model testing) and the correctly deconvoluted rate (>70% across most cancers) for model validation using Cancer Cell Line Encyclopedia (CCLE) bulk RNA data. We applied the model to tumors in The Cancer Genome Atlas (TCGA) dataset and examined associations between predicted cell viability and mutation status or gene expression levels to understand underlying mechanisms of potential value for drug repurposing.

INTRODUCTION

Tumor heterogeneity is associated with cancer progression, recurrence, and responses to drug treatments.[1] Determining the cell composition of tumors is the key to stratifying treatments for cancer patients and developing personalized therapies.[2] Pharmacogenomics, the science of uncovering the genetic determinants of drug responses,[3] is a potential solution for stratification and customization of cancer treatments. However, large-scale drug response datasets are mainly derived from cancer cell lines, with limited screening data from patient-derived xenograft (PDX) models. Drug treatment data derived directly from patients are either scarce, inconsistent, or hard to quantify due to complex regimens for individual patients. Transferring

pharmacogenomics knowledge from *in vitro* to *in vivo* settings has been attempted to accelerate and improve drug treatment in clinical settings.[4–13]

Gene expression profiling by RNA sequencing (RNA-seq) is commonly used to characterize molecular traits. However, it only measures the average gene expression across different types of cells within samples and ignores the biological implications underlying differences among cells. Although changes in cellular composition are among the main factors affecting changes in gene expression, bulk RNA-seq methods cannot offer accurate characterizations of this phenomenon.

Unlike conventional RNA-seq, single-cell technology is a powerful tool for dissecting out distinct cell populations within one sample. Nevertheless, the cost of single-cell experiments is still much higher than conventional RNA-seq technology. It would be more cost effective to dissect the composition of cells within bulk RNA-seq data by using the knowledge from existing single-cell RNA-seq data. Therefore, deconvolution of bulk RNA-seq to infer cell populations could elucidate the biological meaning of changes in gene expression. Previous studies have developed a few computational deconvolution methods to address this issue.[14–17] One study, using a deep learning-based model called Scaden, reported excellent performance in cell type deconvolution.[14] Another recent paper explored tumor heterogeneity by deconvoluting breast cancers into breast cancer cell lines and further applied the results to drug response prediction.[18] As more large-scale single cell datasets become available, using cell deconvolution models to analyze tumor samples may bridge the gap of pharmacogenomics knowledge between cell line data and tumor data.

In this study, we aimed to predict drug response through the deconvolution of tumors by cancer cell lines. We developed Scaden-CA, which we trained with a collection of single-cell RNA (scRNA) data from cancer cell lines, to break down tumor data from The Cancer Genome Atlas (TCGA) into cancer cell lines with corresponding cancer types. We then implemented an algorithm to combine the deconvoluted proportions with drug sensitivity data from the Profiling Relative Inhibition Simultaneously in Mixtures (PRISM) dataset for predicting drug response. Overall, our model addresses the need to transfer pharmacogenomics knowledge from cancer cell lines to tumor samples for predicting drug responses, thereby advancing the development of cancer therapeutics.

RESULTS

Performance evaluation and improvement of the deconvolution models

We adapted the deep learning-based model called Scaden[14] to deconvolute tumors into cancer cell lines and then used drug response data from those cell lines to predict drug responses of the tumors. We hypothesized that diverse cancer cell lines may contain necessary genomic/cell type information to explain the heterogeneity of patient tumor samples. Our model, termed Scaden-CA, was developed in Python (under the TensorFlow/Keras environment). We trained 18 models for 18 cancer types, using simulated bulk RNA-seq data and a training process provided by the original Scaden model (Figure 1A). The number of cell lines used in the models and the performance evaluated by the concordance correlation coefficient (CCC), mean absolute error (MAE), and root-mean-square error (RMSE) are shown in Table 1. Overall, Scaden-CA showed good deconvolution performance in all cancers before cell line selection (CCC > 0.95 when number of cell lines $\leq$ 16, MAE < 0.02 and RMSE < 0.03 across all cancer types, except for lung cancer; Table 1). In addition, we checked loss value during the steps of model training and found that all cancer types converged, except for the model for lung cancer (Figure S1).

The performance of Scaden-CA deteriorated when more cell lines were included in data simulation, particularly for lung cancer. To improve performance with respect to lung cancer, we first determined the optimal number of cell lines to be included in the model and then applied mutation-guided selection criteria to choose representative cell lines. The optimal number of cell lines was determined by the interval of true proportion of simulation data that showed low and stabilized percent error (Figure 1B). Percent error decreased markedly between intervals of true proportion of 0–0.1 and 0.1–0.2. Since the optimal number of cell lines would likely be within this interval, we further analyzed this interval and observed a decreasing and stabilizing trend as the true proportions increased (Figure 1B).

To determine the optimal number of cell lines, we set a fluctuation of percent error of less than 5% as the criterion for stabilized percent error, which starts around the interval of a true proportion of 0.04–0.05 (Figure 1B). If the average proportion of the cell lines is between 0.04 and 0.05, then the number of cell lines is about 20–25. We did not consider the last interval (0.9–1.0) because, although it had the lowest percent error, this interval implies the use of only one cell line, which does not meet our goal of representing tumor heterogeneity by cancer cell lines. Therefore, we decided to use 20–25 cell lines.

To further examine how numbers of cell lines affected the accuracy of proportion estimation, we applied three mutation-guided filtering criteria to reduce the number of cell lines (see experimental procedures). The resulting model performance is shown in Table 1. For most cancer types, the CCC, MAE, and RMSE were maintained at similar levels, but the CCC for the lung cancer model showed substantial improvement in all three selection criteria compared with the model performance with 40 lung cancer cell lines. Among all three criteria, the one that required cell lines to include actionable mutations defined by the Oncology Knowledge Base (oncoKB) database, which was reduced to 11 lung cancer cell lines, showed the largest improvement (CCC = 0.986). However, since the Scaden-CA model can resolve up to 20 cell lines, we decided to use the set of 19 cell lines that covered the overlapped mutations between TCGA and Cancer Cell Line Encyclopedia (CCLE) mutation data also included in oncoKB (CCC = 0.941) for our lung cancer model.

Validation of the deconvolution model and assessment of simulation method by CCLE

To validate whether the trained Scaden-CA model can accurately predict cell line compositions, we applied the models to CCLE bulk RNA data with corresponding cancer types to test whether the cell lines could be correctly deconvoluted to themselves. The correctly deconvoluted rates (CDR; see experimental procedures) are summarized in Figure 1C for each cancer type. The best models were trained for sarcoma (>90%) and

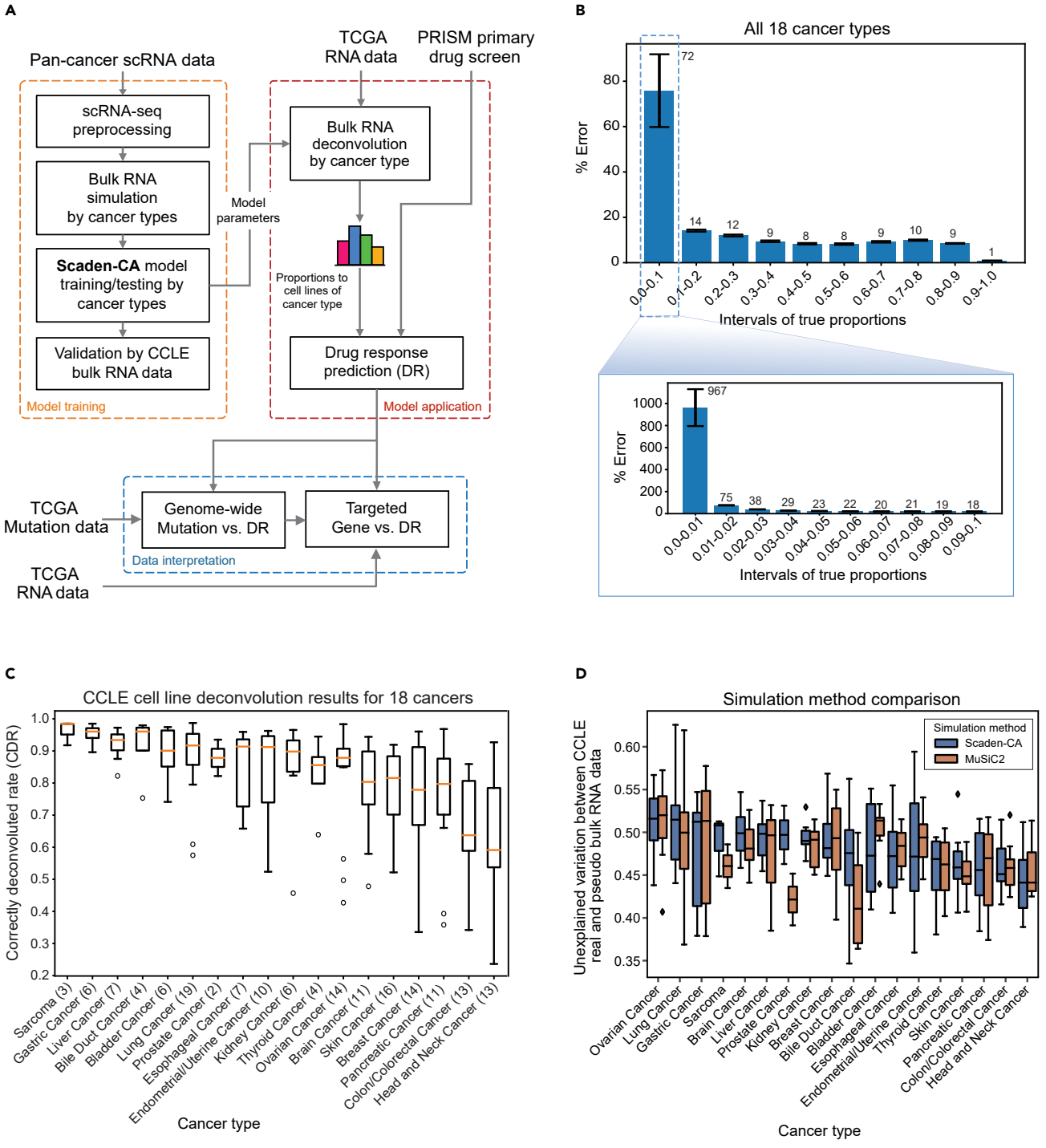


Figure 1. Model design and assessment of model performance

(A) Flowchart of the Scaden-CA deconvolution model and the drug prediction algorithm.

(B) Predictions of percent error across the 18 cancer types at different true cell proportions. The error bars indicate the standard error of the percent error.

(C) Boxplot for the validation results by CCLE bulk RNA data. Numbers in parentheses are the numbers of cell lines used in data simulation.

(D) Boxplot for the comparison results between the Scaden-CA and MuSiC2 simulation methods.

For the boxplots in (C) and (D), the box covers 50% of data from the first quartile to the third quartile. The line splitting the box in two represents the median value, the whiskers extend from the box to the farthest data point lying within 1.5 times the interquartile range from the box, and the points outside of these ranges are outliers.

Table 1. Performance of Scaden-CA models by concordance correlation coefficient (CCC), mean absolute error (MAE), and root-mean-square error (RMSE) under different criteria of cell line selection

	All cell lines				Include mutations in oncoKB or mutations on COSMIC cancer driver genes				Include mutations that are included in oncoKB				Include actionable mutations that are included in oncoKB			
Cancer type	Number of cell lines	CCC	MAE	RMSE	Number of cell lines	CCC	MAE	RMSE	Number of cell lines	CCC	MAE	RMSE	Number of cell lines	CCC	MAE	RMSE
Prostate cancer	2	0.999	0.0099	0.0164	0	N/A	N/A	N/A	0	N/A	N/A	N/A	0	N/A	N/A	N/A
Thyroid cancer	4	0.998	0.0114	0.0158	1	N/A	N/A	N/A	1	N/A	N/A	N/A	1	N/A	N/A	N/A
Bile duct cancer	4	0.998	0.0107	0.0149	2	0.999	0.0091	0.0152	2	0.999	0.0100	0.0165	2	0.999	0.0107	0.0169
Sarcoma	3	0.998	0.0097	0.0144	2	0.999	0.0089	0.0157	2	0.999	0.0103	0.0168	1	N/A	N/A	NA
Bladder cancer	6	0.996	0.0147	0.0193	4	0.998	0.0104	0.0145	3	0.999	0.0113	0.0161	2	0.999	0.0075	0.0133
Gastric cancer	6	0.996	0.0114	0.0155	6	0.996	0.0121	0.0162	3	0.998	0.0145	0.0203	2	0.999	0.0106	0.0177
Liver cancer	7	0.995	0.0117	0.0163	1	N/A	N/A	N/A	0	N/A	N/A	N/A	0	N/A	N/A	N/A
Kidney cancer	6	0.995	0.0172	0.0233	4	0.997	0.0153	0.0206	2	0.999	0.0106	0.0171	1	N/A	N/A	N/A
Esophageal cancer	7	0.991	0.0137	0.0187	4	0.998	0.0121	0.0159	3	0.998	0.0135	0.0187	0	N/A	N/A	N/A
Endometrial/uterine cancer	10	0.988	0.0146	0.0214	9	0.991	0.0131	0.0178	7	0.995	0.0125	0.0171	7	0.991	0.0155	0.0216
Pancreatic cancer	11	0.987	0.0129	0.0181	4	0.998	0.0114	0.0155	3	0.999	0.0089	0.0130	3	0.999	0.0093	0.0133
Brain cancer	11	0.985	0.0147	0.0212	6	0.995	0.0145	0.0199	4	0.995	0.0198	0.0261	1	N/A	N/A	N/A
Ovarian cancer	14	0.980	0.0134	0.0195	0	N/A	N/A	N/A	0	N/A	N/A	N/A	0	N/A	N/A	N/A
Colon/colorectal cancer	13	0.976	0.0135	0.0203	11	0.986	0.0132	0.0195	9	0.991	0.0126	0.0182	8	0.993	0.0124	0.0179
Breast cancer	14	0.975	0.0125	0.0190	9	0.992	0.0121	0.0172	8	0.994	0.0114	0.0163	6	0.995	0.0131	0.0181
Head and neck cancer	13	0.969	0.0185	0.0276	7	0.983	0.0219	0.0304	4	0.996	0.0157	0.0223	2	0.999	0.0074	0.0124
Skin cancer	16	0.956	0.0178	0.0266	14	0.964	0.0169	0.0248	7	0.993	0.0138	0.0193	6	0.995	0.0132	0.0188
Lung cancer	40	0.649	0.0164	0.0258	30	0.845	0.0169	0.0245	19	0.941	0.0163	0.0240	11	0.986	0.0137	0.0197

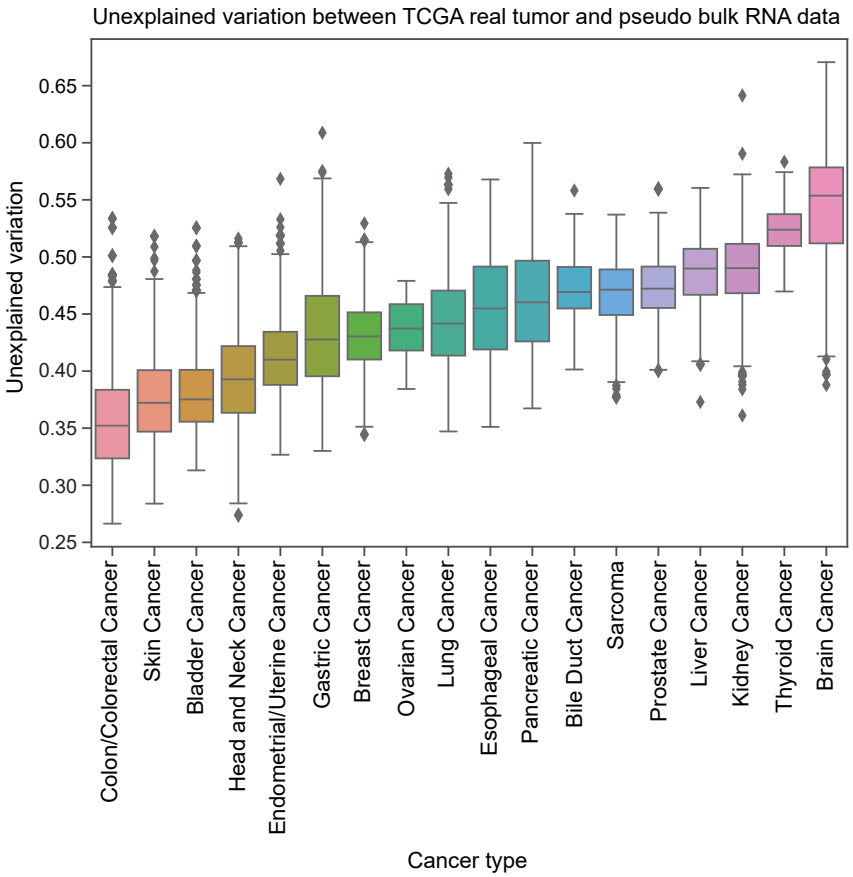


Figure 2. Boxplot for the unexplained variations between TCGA pseudo-bulk and real-bulk RNA data for evaluation of the Scaden-CA model

The definition of the boxplot (box range, whiskers, and outliers) is the same as described in Figure 1.

gastric, liver, and bile duct cancers and the least accurate models for colon/colorectal and head and neck cancers (<60%). Almost all cell lines accounted for the largest proportion of their own deconvolution results except for SNU1076 (a head and neck cancer cell line; Figure S2).

Another essential part of the Scaden-CA model is the generation of training and testing pseudo-bulk RNA-seq data. We compared the simulation methods used by Scaden-CA and MuSiC2[19] (see experimental procedures), as shown in Figure 1D. The unexplained variation (around 0.40–0.55) is mainly due to the capability of the deconvolution model. The Scaden-CA simulation method showed results comparable with the simulation method used by MuSiC2, suggesting that our simulation method can produce representative data from real-bulk RNA data for model training. Taken together, these results suggested that the Scaden-CA model provides good approximation of the potential compositions of cancer cell lines based on RNA data.

TCGA tumor deconvolution to estimate unexplained variation and to predict drug response

We applied the Scaden-CA model to TCGA samples to deconvolute to proportions of cell lines and then inferred potential therapeutic treatments from cell line drug screening data. The resulting proportions were hypothesized to represent tumor heterogeneity by a diverse collection of cancer cell lines, which can be used to predict drug responses (see experimental procedures).

Before applying our drug response prediction algorithm to TCGA tumors, we first evaluated the capability of the deconvolution model. We calculated the unexplained variation between the pseudo-bulk and real-bulk RNA data in TCGA, as shown in Figure 2. The unexplained variation ranges from about 0.35 to 0.55 across 18 cancer types, with colon/colorectal cancer showing the lowest mean unexplained variation (~0.35) and brain cancer exhibiting the highest mean unexplained variation (~0.55). This may result from the existence of other cells (e.g., immune cells and normal cells) within TCGA tumors. We will look for novel solutions in future studies to improve the Scaden-CA model.

Predicting drug response of TCGA tumor samples via deconvolution and examination of mutation-drug response associations

We then applied the drug response prediction algorithm to deconvoluted TCGA tumors and calculated the predicted drug response by multiplying the deconvoluted proportions of cell lines for each tumor with the drug response of the corresponding cell lines (see experimental procedures). We analyzed 7,781 tumor samples from 18 cancer types (Table S1), and responses to 4,518 drugs were predicted by using the PRISM screen dataset.

After obtaining predicted drug responses, we first compared all drug responses between wild-type or mutant samples in the same cancer type. We split samples into wild-type or mutant groups based on the mutation status of cancer driver genes and then inspected the associations between the predicted cell viability of the two groups with the mutation status, which led to multiple cancer-gene-drug combinations. The mutant group was defined as those with at least one mutation on the gene being analyzed, and the wild-type group was the remaining samples. We repeated these analyses for the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC) genes, and the numbers of samples analyzed for these combinations are listed in Table S1.

The results of all cancer-gene-drug combinations are shown in Figure 3A. We defined the threshold of statistical significance for cancer-gene-drug combinations as an adjusted p value less than 0.05. In Figure 3A, 182,565 (0.5%) of 37,442,652 combinations passed the threshold across the 18 cancers, or ~10,143 combinations per cancer (Table S2). Brain cancer had the largest number of significant combinations (n = 30,032) while ovarian cancer showed no significant combinations (Table S2), possibly due to the limited sample size. We listed the top 5 statistically significant cancer-gene-drug combinations for the cancers

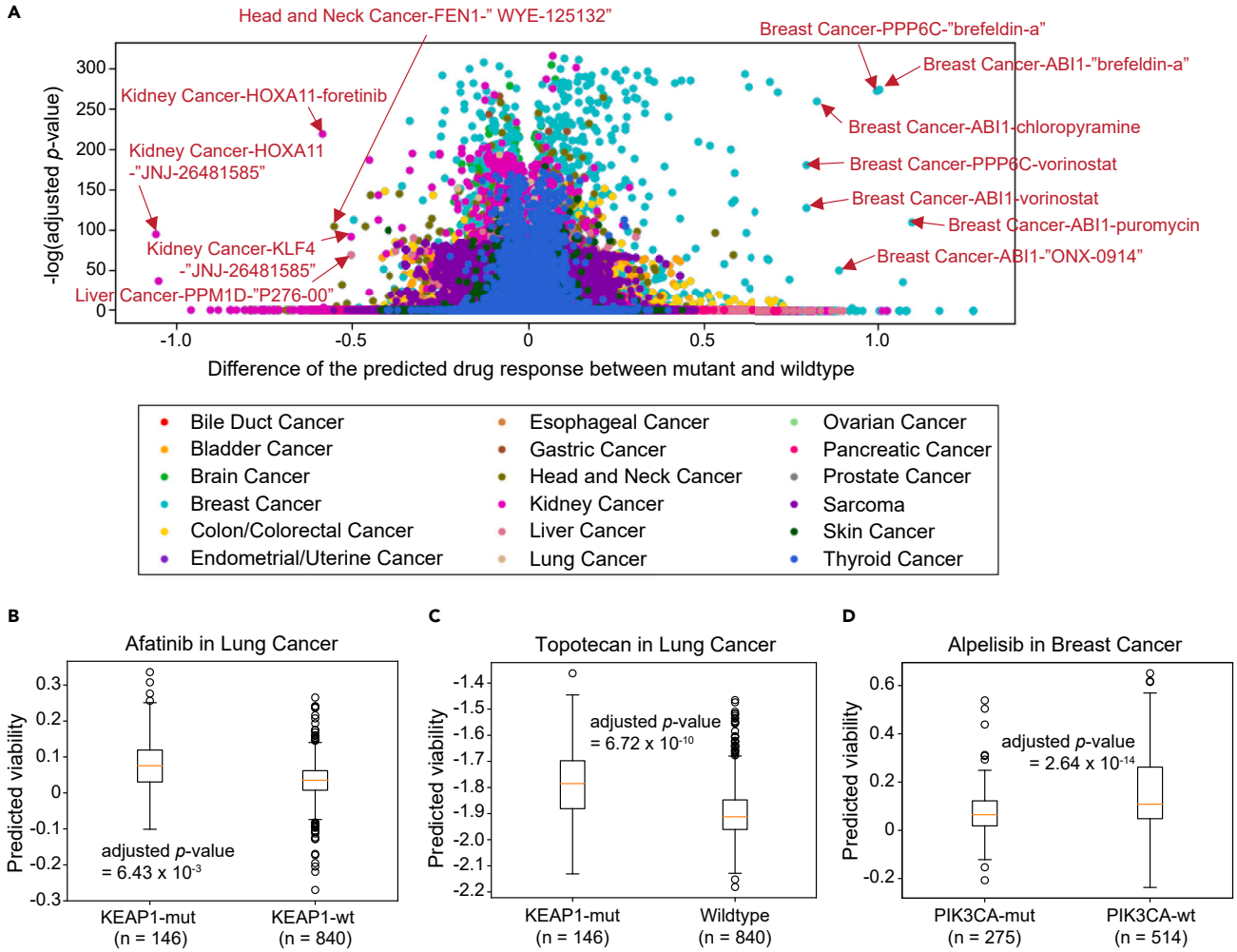


Figure 3. Exploration of the impact of mutational alteration at the gene level to drug response through cancer-gene-drug combinations

(A) Cancer-gene-drug combinations across the 18 cancer types. The x axis indicates the drug response differences of log2 fold change between mutant and wild type (WT) (Δlog 2FC), and the y axis indicates the negative logarithm of adjusted p values. The dots near the top left or top right are the combinations showing statistical significance and higher differences of log2 fold change. Marked in red are the combinations showing extreme statistical significance. Numbers of samples analyzed for these combinations are listed in Table S1 (column 4).

(B) Boxplot for the drug response data for the lung cancer-KEAP1-afatinib combination.

(C) Boxplot for the drug response data for the lung cancer-KEAP1-topotecan combination.

(D) Boxplot for the drug response data for the breast cancer-PIK3CA-alpelisib combination.

The definition of boxplots (box range, whiskers, and outliers) is the same as described in Figure 1. The p values are from t test; p values adjusted by Bonferroni correction.

with combinations that passed the filtering criteria (adjusted p <0.05 and ≥ 50 samples per group) in Table 2. These combinations represent possible drug-target associations, and the ones showing statistical significance are worth further experimental validations.

Among the results shown in Figure 3A and with an adjusted p value less than 10^{-50}, 7 combinations showed statistical significance with drug response differences of log2 fold change between mutant and wild-type groups (Δlog 2FC) greater than 0.75 and 5 combinations with Δlog 2FC less than −0.5 (Table 2 and Figure S3). Among the 3 kidney cancer cases in the 5 combinations mentioned above, the *HOXA11* gene is a tumor suppressor gene that is frequently hypermethylated in renal cell carcinoma (RCC) cell lines and primary RCC tumors,[21] and the *KLF4* gene may also suppress the growth of clear cell RCC.[22] As for the two drugs in the 3 combinations, both foretinib[23] and JNJ-26481585[24] have been reported as having anti-tumor effects in kidney cancers. Foretinib is an oral multikinase inhibitor with activity in patients with papillary RCC,[23] and JNJ-26481585,[24] a pan-histone deacetylase inhibitor, has antiproliferative activity against a variety of solid tumors, including breast cancers.

For the head and neck cancer-*FEN1*-WYE-125132 combination, *FEN1* overexpression was associated with poor survival in head and neck cancers,[25] but the associations among head and neck cancer, *FEN1*, and the drug WYE-125132 remain unknown. *PPM1D* expression is high in liver cancer and is significantly

Table 2. Top 5 cancer-gene-drug combinations for the 13 cancer types that passed the selection criteria (≥50 samples per group and $p < 0.05$) and combinations with statistical significance (adjusted $p < 10^{-50}$ and Δlog 2FC < −0.5 or > 0.75) across all 18 cancers

Cancer type	Gene	Drug	Number of mutated samples	Number of WT samples	Δlog2FC	Adjusted p of t test
Top 5 cancer-gene-drug combinations for the 13 cancers						
Bladder cancer	FGFR3	L-citrulline	61	345	0.0362	1.15×10^{-5}
	FGFR3	nolatrexed	61	345	−0.0729	1.27×10^{-5}
	FGFR3	colforsin	61	345	0.0455	1.44×10^{-5}
	FGFR3	GSK1070916	61	345	−0.1420	1.62×10^{-5}
	FGFR3	lafutidine	61	345	−0.0504	1.66×10^{-5}
Brain cancer	IDH1	BF2.649	404	257	−0.0188	4.94×10^{-52}
	IDH1	acebutolol	404	257	0.0235	8.61×10^{-50}
	IDH1	alogliptin	404	257	−0.0366	9.07×10^{-50}
	IDH1	silodosin	404	257	0.0194	9.71×10^{-50}
	IDH1	JDTic	404	257	−0.0196	1.30×10^{-49}
Breast cancer	TP53	bazedoxifene	266	523	0.0593	1.84×10^{-37}
	TP53	MM77	266	523	−0.0355	5.83×10^{-37}
	TP53	milacemide	266	523	0.0486	9.17×10^{-37}
	TP53	tedizolid	266	523	0.1269	1.79×10^{-36}
	TP53	chloramphenicol	266	523	0.0506	3.98×10^{-36}
Colon/colorectal cancer	BRAF	alverine	53	326	−0.0343	7.10×10^{-13}
	BRAF	homosalate	53	326	0.0245	2.85×10^{-12}
	BRAF	sertindole	53	326	0.0489	3.09×10^{-12}
	BRAF	disulfiram	53	326	−0.1245	4.61×10^{-12}
	BRAF	idazoxan	53	326	−0.0963	5.24×10^{-12}
Endometrial/uterine cancer	PTEN	deferiprone	292	201	−0.0257	1.04×10^{-32}
	PTEN	propranolol-(R)	292	201	−0.0260	1.25×10^{-30}
	PTEN	metocurine	292	201	−0.0246	5.13×10^{-30}
	PTEN	KY02111	292	201	−0.0132	4.01×10^{-28}
	PTEN	CGS-21680	292	201	−0.0260	6.63×10^{-27}
Gastric cancer	KMT2D	bopindolol	73	339	−0.0156	1.97×10^{-7}
	KMT2D	talarozole	73	339	0.0286	2.97×10^{-7}
	KMT2D	phthalylsulfathiazole	73	339	0.0149	5.76×10^{-7}
	KMT2D	ADX-47273	73	339	−0.0078	5.95×10^{-7}
	KMT2D	CNX-774	73	339	0.0589	6.38×10^{-7}
Head and neck cancer	NSD1	acamprosate	58	441	−0.0718	3.78×10^{-12}
	NSD1	meisoindigo	58	441	−0.0875	4.11×10^{-11}
	NSD1	terconazole	58	441	−0.0459	5.56×10^{-11}
	NSD1	captopril	58	441	−0.0550	7.88×10^{-11}
	NSD1	cidofovir	58	441	0.0377	9.21×10^{-11}
Kidney cancer	VHL	exatecan-mesylate	175	537	0.0583	3.45×10^{-47}
	VHL	IB-MECA	175	537	0.0203	5.89×10^{-46}
	VHL	eslicarbazepine-acetate	175	537	0.0171	1.41×10^{-45}
	VHL	epacadostat	175	537	0.0228	2.06×10^{-45}
	VHL	empagliflozin	175	537	0.0207	3.70×10^{-45}
Liver cancer	CTNNB1	SKF-77434	94	264	−0.0353	2.81×10^{-13}
	CTNNB1	EBPC	94	264	−0.0157	3.39×10^{-13}
	CTNNB1	prednisolone-hemisuccinate	94	264	−0.0269	4.35×10^{-13}
	CTNNB1	LY364947	94	264	−0.0428	9.02×10^{-13}
	CTNNB1	ADL5859	94	264	−0.0323	1.55×10^{-12}

(*Continued on next page*)

Table 2. *Continued*

Cancer type	Gene	Drug	Number of mutated samples	Number of WT samples	Δlog2FC	Adjusted p of t test
Lung cancer	KRAS	setiptiline	157	829	0.0337	6.93×10^{-31}
	KRAS	L-152804	157	829	0.0454	3.07×10^{-30}
	KRAS	mirtazapine	157	829	0.0152	1.73×10^{-29}
	KRAS	docosanol	157	829	0.0310	9.99×10^{-29}
	KRAS	*trans*-4-hydroxycrotonic-acid	157	829	0.0310	2.03 × 10-[20]
Pancreatic cancer	KRAS	eliglustat	111	59	−0.0212	3.35×10^{-10}
	KRAS	AT13387	111	59	−0.1168	8.19×10^{-9}
	KRAS	BVD-523	111	59	−0.0966	2.83×10^{-8}
	KRAS	suramin	111	59	−0.0115	8.99×10^{-8}
	KRAS	gedunin	111	59	0.0505	1.98×10^{-7}
Skin cancer	BRAF	PHA-793887	242	223	−0.0780	2.26×10^{-7}
	BRAF	WAY-170523	242	223	−0.0256	8.66×10^{-6}
	BRAF	barasertib	242	223	−0.0413	3.33×10^{-5}
	BRAF	2,6-dimethylpiperidine	242	223	0.0136	1.84×10^{-4}
	BRAF	apafant	242	223	−0.0139	2.66×10^{-4}
Thyroid cancer	BRAF	AZD5438	291	199	0.0614	9.89×10^{-8}
	BRAF	paliperidone	291	199	0.0316	1.21×10^{-7}
	BRAF	betaxolol	291	199	0.0259	1.26×10^{-7}
	BRAF	PHA-767491	291	199	0.0247	1.28×10^{-7}
	BRAF	cefozopran	291	199	0.0135	1.41×10^{-7}
Cancer-gene-drug combinations with statistical significance across 18 cancer types						
Breast cancer	ABI1	brefeldin A	2	787	1.004	1.88×10^{-274}
	PPP6C	brefeldin A	2	787	0.999	9.98×10^{-273}
	ABI1	chloropyramine	2	787	0.829	4.36×10^{-259}
	PPP6C	vorinostat	2	787	0.799	2.56×10^{-180}
	ABI1	vorinostat	2	787	0.799	2.18×10^{-128}
	ABI1	puromycin	2	787	1.095	1.95×10^{-109}
	ABI1	ONX-0914	2	787	0.891	1.18×10^{-50}
Head and neck cancer	FEN1	WYE-125132	2	497	−0.555	2.51×10^{-105}
Kidney cancer	HOXA11	foretinib	2	710	−0.588	5.24×10^{-219}
	HOXA11	JNJ-26481585	2	710	−1.059	3.91×10^{-96}
	KLF4	JNJ-26481585	2	710	−0.504	8.42×10^{-92}
Liver cancer	PPM1D	P276-00	2	356	−0.505	2.15×10^{-69}

associated with poor prognosis,[26] but the effects of P276-00 in liver cancer and its association with *PPM1D* have not been explored.

Among the 7 combinations with positive Δlog 2FC involving breast cancers, *PPP6C*[27] and *ABI1*[28] have low and high expression in breast cancers, respectively, and each is linked to poor survival.[27,28] Brefeldin A, chloropyramine, and vorinostat have been tested in either breast cancer cell lines or clinical trials and may have anti-tumor effects in breast cancers.[20,29–34] The associations between these genes and drugs in the combinations represent potential novel drug repurposing and merit further investigation and validation.

Aside from the novel combinations mentioned above, we also confirmed known combinations (Figures 3B–3D). The mutation status of the *KEAP1* gene modulates the response of afatinib (adjusted $p = 6.43 \times 10^{-3}$) in lung cancers[35] and confers drug resistance to drugs such as topotecan in patients with lung cancer (adjusted $p = 6.72 \times 10^{-10}$).[36,37] *PIK3CA* mutations may determine the responses of breast cancers to alpelisib (adjusted $p = 1.64 \times 10^{-14}$).[38] These results were consistent with previous studies, showing that our drug response prediction algorithm can identify some known mechanisms affecting drug responses.

We further examined statistically significant cancer-gene-drug combinations to prioritize drug repurposing worthy of further investigation (Table 3). Among these pairs, the breast cancer-TP53-CYC116 combination may be worthy of further investigation. A previous study suggested that an aurora kinase A/B/C inhibitor,

Table 3. Top 4 gene-drug pairs showing significance across all cancer-gene-drug combinations

Gene	Drug	Cancers showing significance with the gene-drug pair
SMO	SR-33805	Liver cancer, Head and neck cancer, kidney cancer, esophageal cancer
TP53	CYC116	endometrial/uterine cancer, colon/colorectal cancer, lung cancer, breast cancer
TP53	CD-437	endometrial/uterine cancer, colon/colorectal cancer, lung cancer, breast cancer
TP53	dolastatin-10	endometrial/uterine cancer, colon/colorectal cancer, lung cancer, breast cancer

AMG900,[39] preferentially inhibits the growth of breast cancer cell lines with TP53 loss-of-function mutations. Since CYC116 is an aurora kinase A/B inhibitor,[40] it may also be highly active in breast cancers or cell lines with TP53 loss-of-function mutations, but further experiments are needed to validate the results.

Examination of somatic mutations and their influence in drug response in cancers

To examine the effects of a specific amino acid (aa) change in a cancer type, for each of the point mutations in all TCGA samples, we compared patients with and without a specific somatic mutation, resulting in multiple cancer-[gene, aa change]-drug combinations. We identified 6,338,962 (1.7%) significant combinations (of 382,034,170) with a p value less than 0.05 (Figure 4A), and endometrial/uterine cancer had the largest number of significant combinations (n = 2,265,656) (Table S2). Again, ovarian cancer showed no significant combinations (Table S2). Examples of top 5 statistically significant cancer-[gene, aa change]-drug combinations passed the filtering criteria (adjusted $p < 0.05$ and ≥ 50 samples per group) are shown in Table 4, including somatic mutations of R132H in IDH1 (brain cancer), E545K in PIK3CA (breast cancer), G12C in KRAS (lung cancer), and V600E in BRAF (skin and thyroid cancers). Among the results shown in Figure 4A with an extreme p value threshold less than 10^{-50}, there are 7 cancer-[gene, aa change]-drug combinations with Δlog 2FC greater than 0.8 and 6 combinations with Δlog 2FC less than -0.7 (Table 4 and Figure S4). These combinations are all novel findings and may need further investigations to validate mutation-drug associations.

We further confirmed that some known cancer-[gene, aa change]-drug combinations recorded in the oncoKB database showed statistical significance in our results, such as breast cancer-[*PIK3CA*, H1047R]-alpelisib (Figure 4B), breast cancer-[*PIK3CA*, E545K]-alpelisib (Figure 4C), and lung cancer-[KEAP1, R320Q]-afatinib (Figure 4D) combinations. The former two combinations were included in the oncoKB database and provided evidence that breast cancer patients who carry the H1047R or E545K mutation on the *PIK3CA* gene can be treated with the FDA-approved drug alpelisib plus the selective estrogen receptor degrader fulvestrant. KEAP1/R320Q is known to be oncogenic, and our findings concerning the lung cancer-[KEAP1, R320Q]-afatinib combination were consistent with previous finding that certain mutations on the KEAP1 gene modulate the effectiveness of afatinib through disrupting the Keap1-Nrf2 pathway.[41,42] Taken together, our algorithms reproduce some evidence on known effective drug treatments and shed light on unknown mechanisms potentially useful for drug repurposing.

Associations between gene expression levels and drug responses

To understand whether differences in drug response are due to gene expression changes, we split the tumor samples into groups with high or low expression of CGC cancer driver genes (n = 736) and then assessed the differences in drug responses between two groups (Figure 5A, volcano plot of drug-response differences versus log10(p value)). Table 5 shows the top 5 statistically significant combinations for the 16 cancer types with combinations that passed the filtering criteria. Our results for the previously reported combination, lung cancer-KEAP1-afatinib (Figure 3B), are consistent with the results in Figure 5B if we consider that both down-regulation of KEAP1 and/or KEAP1 mutants constitute the loss of KEAP1 function, which leads to higher cell viability and, hence, increased drug resistance to afatinib. However, the results from the breast cancer-PIK3CA-alpelisib combination in Figure 3D did not hold in Figure 5C, indicating that the mechanism of alpelisib resistance may be solely associated with mutations and not altered PIK3CA expression.

Concordance of predicted drug response to TCGA data

To explore potential clinical applications of our predicted cell viability, we partitioned TCGA samples into responder and non-responder groups to examine to our results. The lung cancer-gemcitabine (Figure 6A) and esophageal cancer-cisplatin (Figure 6B) combinations showed statistical significance and concordant trends, with responders having less cell viability and non-responders having more cell viability. These results showed our model's potential to provide evidence of drug efficacy or repositioning recommendations in clinical settings.

DISCUSSION

We devised a method for cellular deconvolution coupled with drug response prediction to explore potential drug repurposing and the underlying mechanisms using deep learning models. While our model shows potential in predicting drug response based on individual cell types, we recognize limitations in our implementation. First, the numbers of cell lines used for the 18 cancer types were variable, ranging from 2 cancer cell lines for prostate cancer to 19 cancer cell lines for lung cancer. The numbers of cell lines were largely determined by the cell lines available in the scRNA, CCLE, and PRISM datasets. We recognize that these numbers may not reflect the extent of tumor heterogeneity. Therefore, our method could be improved if more cell lines were available for each cancer.

Second, the performance of the Scaden-CA model was degraded when large numbers of cell lines were used for model training. The Scaden-CA model showed less percent error and

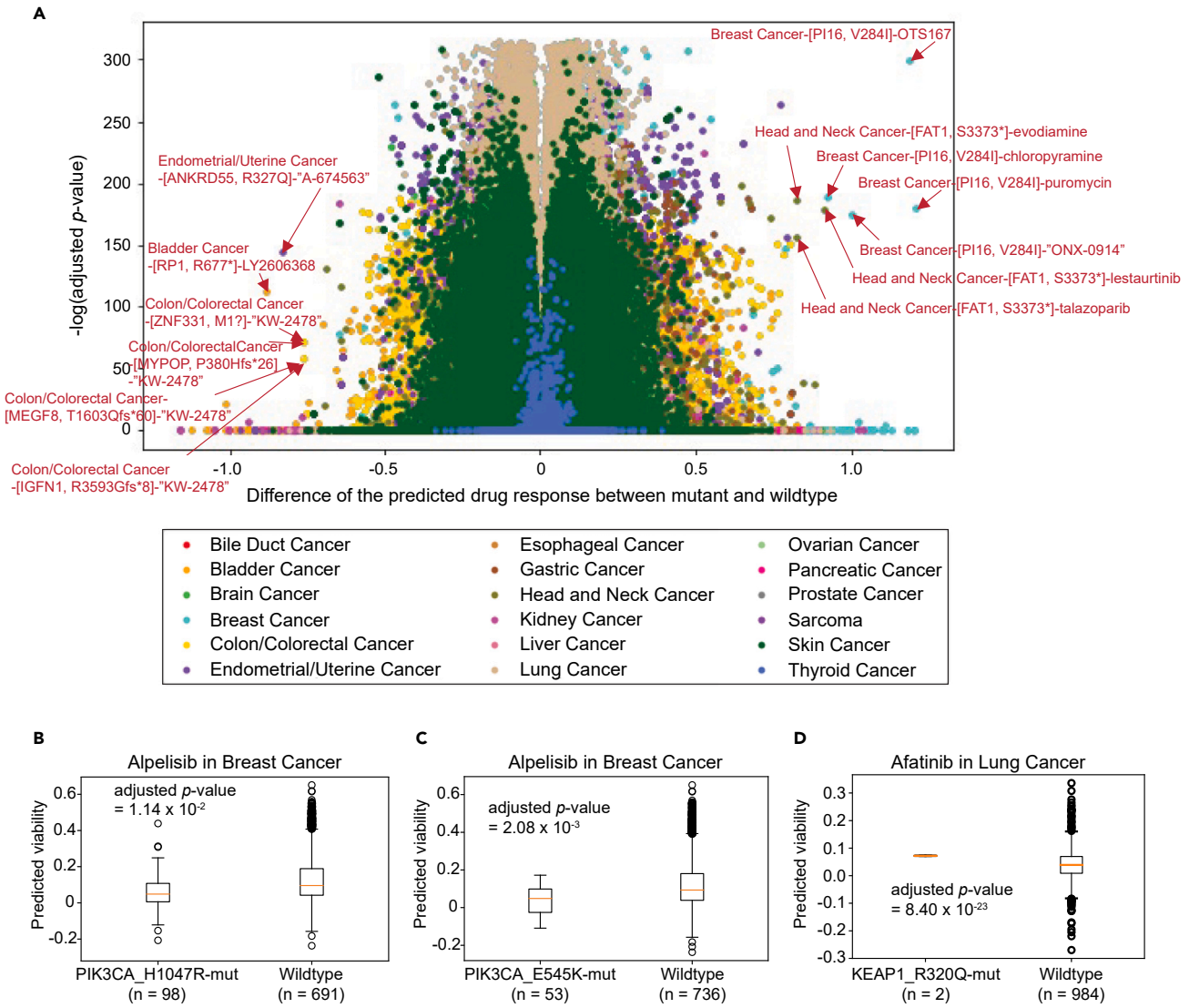


Figure 4. Exploration of the impact of specific aa changes to drug response through cancer-[gene, aa change]-drug combinations

(A) Cancer-[gene, aa change]-drug combinations across the 18 cancer types. The x axis indicates the drug response differences of log2 fold change between mutant and WT groups (Δlog 2FC), and the y axis indicates the negative logarithm of adjusted p values. The dots near the top left or top right are the combinations showing statistical significance and higher differences of log2 fold change. Marked in red are the combinations showing extreme statistical significance. Numbers of samples analyzed for these combinations are listed in Table S1.

(B) Boxplot for the drug response data for the breast cancer-[PIK3CA, H1047R]-alpelisib combination.

(C) Boxplot for the drug response data for the breast cancer-[PIK3CA, E545K]-alpelisib combination.

(D) Boxplot for the drug response data for the lung cancer-[KEAP1, R320Q]-afatinib combination.

The aa changes specified in (A)–(D) follow the Human Genome Variation Society (HGVS) nomenclature. The definition of boxplots (box range, whiskers, and outliers) is the same as described in Figure 1, and the definition of p values are the same as Figure 3.

good performance in tumor deconvolution when including 20–25 cell lines. This can hinder the incorporation of more cell lines to represent complexity at the cellular level due to tumor heterogeneity. Therefore, a more sensitive deconvolution model that can predict ultra-low cell proportions in certain tumors is desirable.

Third, in the deconvolution models of all 18 cancers, some dominant cell lines accounted for the largest proportions for each cancer, which could lead to biases in subsequent drug response predictions. In addition, in the CCLE bulk RNA deconvolution results, colon/colorectal cancer and head and neck cancer samples showed lower correctly deconvoluted rates, which might mean that they are not representative of the tumors under study. Furthermore, the unexplained variation we observed from the deconvolution results of TCGA tumors was around 0.35–0.55, which indicates that our model cannot fully capture the characteristics of certain cell populations, such as normal cells and immune cells, within tumors and, thus, may not fully capture the heterogeneity of tumors. In addition, perhaps other factors should be considered in tumor deconvolution, such as sequencing technology, additional scRNA datasets, or other omics datasets.

Table 4. Top 5 cancer-[gene, aa change]-drug combinations for the 5 cancer types that passed selection criteria (≥50 patients per group and $p < 0.05$) and combinations with statistical significance ($p < 10^{-150}$ and $\Delta log2FC > 0.8$, $p < 10^{-50}$ and $\Delta log2FC < -0.7$) across the 18 cancer types

Cancer type	Gene	Aa change[a]	Drug	Number of mutated samples	Number of WT samples	Δlog2FC	Adjusted p value of t test
Top 5 cancer-[gene, aa change]-drug combinations for the 5 cancer types							
Brain cancer	IDH1	R132H	remoxipride	367	294	0.0129	3.28×10^{-47}
	IDH1	R132H	Alogliptin	367	294	−0.0337	1.48×10^{-46}
	IDH1	R132H	PD-318088	367	294	0.0191	1.57×10^{-45}
	IDH1	R132H	JDTic	367	294	−0.0178	2.01×10^{-45}
	IDH1	R132H	BF2.649	367	294	−0.0169	2.97×10^{-45}
Breast cancer	PIK3CA	E545K	benproperine	53	736	−0.0286	2.54×10^{-11}
	PIK3CA	E545K	semapimod	53	736	−0.0369	4.50×10^{-10}
	PIK3CA	E545K	4-HQN	53	736	−0.0243	1.32×10^{-9}
	PIK3CA	E545K	nitrendipine	53	736	−0.0286	4.05×10^{-9}
	PIK3CA	E545K	fosfomycin	53	736	0.0357	5.45×10^{-9}
Lung cancer	KRAS	G12C	aurora-a-inhibitor-i	59	927	0.0678	8.77×10^{-10}
	KRAS	G12C	JNJ-26481585	59	927	0.0975	4.58×10^{-9}
	KRAS	G12C	CGH2466	59	927	0.0298	1.60×10^{-8}
	KRAS	G12C	pevonedistat	59	927	0.1397	2.74×10^{-8}
	KRAS	G12C	SB-939	59	927	0.1145	3.12×10^{-8}
Skin cancer	BRAF	V600E	PHA-793887	203	262	−0.0905	2.64×10^{-10}
	BRAF	V600E	barasertib	203	262	−0.0483	4.63×10^{-8}
	BRAF	V600E	zileuton	203	262	−0.0207	1.36×10^{-7}
	BRAF	V600E	methoxamine	203	262	−0.0219	1.38×10^{-7}
	BRAF	V600E	WAY-170523	203	262	−0.0286	2.16×10^{-7}
Thyroid cancer	BRAF	V600E	AZD5438	287	203	0.0618	3.68×10^{-7}
	BRAF	V600E	paliperidone	287	203	0.0318	4.30×10^{-7}
	BRAF	V600E	betaxolol	287	203	0.0261	4.52×10^{-7}
	BRAF	V600E	PHA-767491	287	203	0.0249	4.61×10^{-7}
	BRAF	V600E	cefozopran	287	203	0.0136	5.18×10^{-7}
Cancer-[gene, aa change]-drug combinations with statistical significance across 18 cancer types							
Bladder cancer	RP1	R677*	LY2606368	2	404	−0.8817	5.45×10^{-113}
Breast cancer	PI16	V284I	OTS167	2	787	1.1894	6.58×10^{-300}
	PI16	V284I	chloropyramine	2	787	0.9219	9.68×10^{-190}
	PI16	V284I	puromycin	2	787	1.2108	2.91×10^{-181}
	PI16	V284I	ONX-0914	2	787	1.0003	6.74×10^{-176}
Colon/colorectal cancer	ZNF331	M1?[b]	KW-2478	2	377	−0.7647	4.02×10^{-59}
	MYPOP	P380Hfs*26	KW-2478	2	377	−0.7647	1.65×10^{-58}
	MEGF8	T1603Qfs*60	KW-2478	2	377	−0.7644	1.84×10^{-72}
	IGFN1	R3593Gfs*8	KW-2478	2	377	−0.7644	1.84×10^{-72}
Endometrial/uterine cancer	ANKRD55	R327Q	A-674563	2	491	−0.8304	1.01×10^{-145}
Head and neck cancer	FAT1	S3373*	evodiamine	2	497	0.8212	3.90×10^{-187}
	FAT1	S3373*	lestaurtinib	2	497	0.9142	1.68×10^{-179}
	FAT1	S3373*	talazoparib	2	497	0.8241	5.23×10^{-157}

[a]The aa changes included in the table are defined by the Human Genome Variation Society (HGVS) nomenclature.
[b]The consequence of the mutation affecting the translation initiation codon is unknown (HGVS nomenclature).

Fourth, domain differences between cancer cell lines and real tumors were not addressed in our study, which means that the direct application of the deconvolution model trained on simulated bulk RNA to real tumors may not be suitable and may introduce biases to the deconvolution results.

Fifth, only very limited data were available in TCGA. Furthermore, the process of grouping patients into responders or

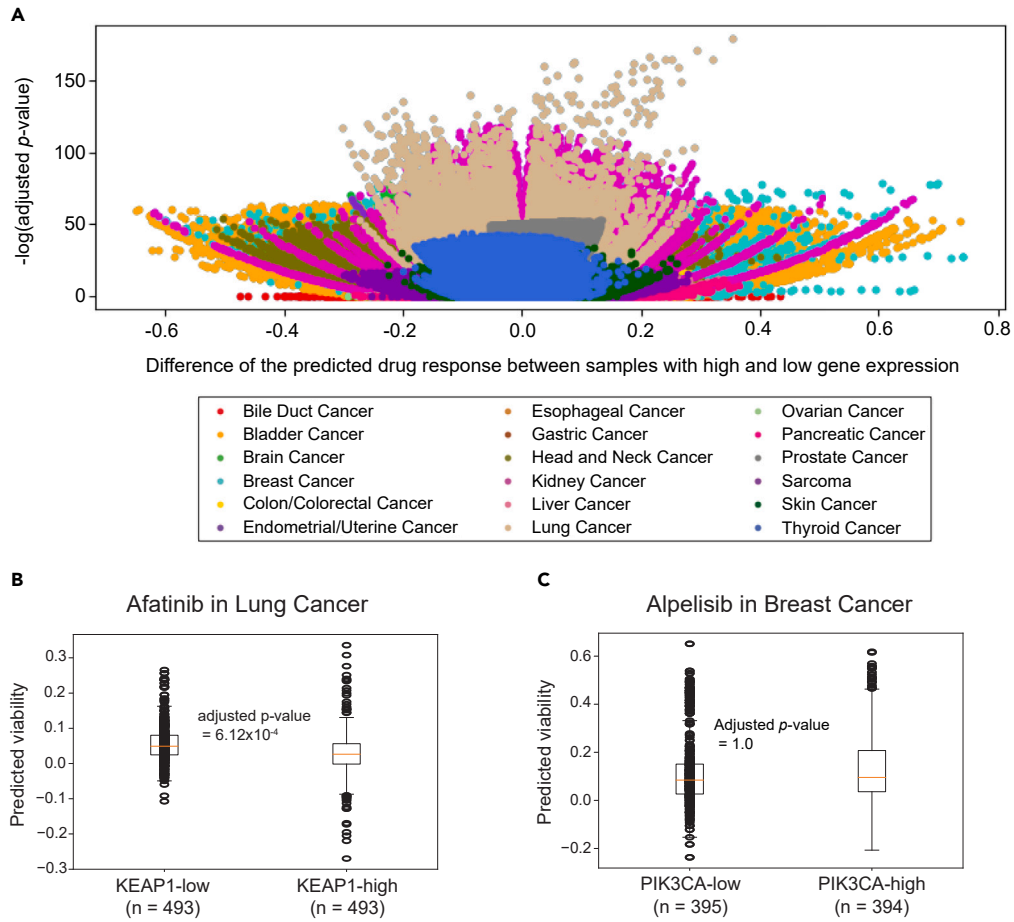


Figure 5. Exploration of the impact of gene expression alteration to drug response

(A) Cancer-gene-drug combinations analyzed by gene expression profiles.

(B) Boxplot for the drug response data for the lung cancer-KEAP1-afatinib combination.

(C) Boxplot for the drug response data for the breast cancer-PIK3CA-alpelisib combination.

The definition of boxplots (box range, whiskers, and outliers) is the same as described in Figure 1, and the definition of p values are the same as Figure 3.

non-responders can be complicated. Other factors, such as the chronological order of treatments, previously prescribed drugs, and the medical history may need to be considered for a clearer definition of responders and non-responders. In this study, we did not consider the chronological order of the treatments. Therefore, the validation results may be biased.

The algorithm we used to predict drug responses is based on the hypothesis that there is a linear relationship between the deconvoluted cell line proportions and the overall responses of tumors to drug treatments. If this linear relationship does not hold for all combinations of drugs and tumors, then a more capable deep learning model may be trained to handle nonlinearity, which we will explore in a future study.

Regardless of these limitations, the potential use of predicted cell viability from tumor samples is feasible. For example, based on results for afatinib (Figures 3B and 4D), mutations on the KEAP1 gene will result in worse treatment outcomes. Similar conclusions can be inferred from Figure 5B by the worse outcomes observed in patients with lower expression of KEAP1. Based our *in silico* predictions, perhaps patients with KEAP1 mutations or lower gene expression levels should consider treatments other than afatinib. Given all the combinations of 736 CGC genes and 4,518 drugs, a number of alternatives may be considered based on our prediction results.

In conclusion, we trained the Scaden-CA model with excellent deconvolution performance, and the drug response prediction method we implemented offers insights into both drug repurposing and the underlying mechanisms. Our model and algorithm effectively bridge the gap for translating cell line drug sensitivity data into tumor drug response through tumor deconvolution into cancer cell lines. Overall, our study has demonstrated the feasibility of characterizing tumor heterogeneity using cancer cell lines, which could enable more precise prediction of responses to drug treatment, consequently paving the way to personalized medicine.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for data should be directed to and will be fulfilled by the lead contact, Dr. Yidong Chen (cheny8@uthscsa.edu).

Materials availability

The study did not generate new unique reagents.

Table 5. Top 5 cancer-gene-drug combinations of high/low gene expression for the 16 cancer types that passed the selection criteria (≥50 patients per group and p < 0.05)

Cancer type	Gene	Drug	Number of samples of high gene expression	Number of samples of low gene expression	Δlog2FC	Adjusted p value of t test
Bladder cancer	TBX3	fipexide	203	203	−0.2058	3.61×10^{-65}
Bladder cancer	TBX3	nifurtimox	203	203	−0.2472	4.11×10^{-65}
Bladder cancer	TBX3	CCT018159	203	203	−0.3359	4.86×10^{-65}
Bladder cancer	TBX3	theophylline	203	203	−0.1756	4.97×10^{-65}
Bladder cancer	TBX3	XAV-939	203	203	−0.1530	5.81×10^{-65}
Brain cancer	AKT3	N-acetylglycyl-D-glutamic-acid	330	331	−0.0483	4.78×10^{-83}
Brain cancer	AKT3	pelitinib	330	331	−0.1349	2.04×10^{-82}
Brain cancer	TSC1	TAK-901	330	331	−0.1422	3.95×10^{-82}
Brain cancer	AKT3	narasin	330	331	−0.1774	5.25×10^{-82}
Brain cancer	CNTNAP2	narasin	330	331	−0.1766	4.66×10^{-81}
Breast cancer	ESR1	EPZ004777	394	395	−0.0930	3.37×10^{-87}
Breast cancer	ESR1	dabigatran	394	395	−0.0721	5.42×10^{-86}
Breast cancer	ESR1	TAK-285	394	395	−0.0940	7.31×10^{-86}
Breast cancer	GATA3	UNC1999	394	395	−0.0842	1.11×10^{-83}
Breast cancer	ESR1	prednisolone-hemisuccinate	394	395	0.0690	1.20×10^{-83}
Colon/colorectal cancer	AKT3	prednisolone-acetate	189	190	0.0698	1.01×10^{-49}
Colon/colorectal cancer	AKT3	PF-5274857	189	190	−0.0575	9.87×10^{-46}
Colon/colorectal cancer	AKT3	RG2833	189	190	−0.0352	6.61×10^{-45}
Colon/colorectal cancer	ZEB1	PF-5274857	189	190	−0.0568	5.24×10^{-44}
Colon/colorectal cancer	ZEB1	celiprolol	189	190	−0.0327	9.85×10^{-44}
Endometrial/uterine cancer	FOXR1	1-acetyl-4-methylpiperazine	246	247	−0.0537	1.99×10^{-79}
Endometrial/uterine Cancer	FOXR1	cefuroxime	246	247	−0.0502	3.72×10^{-78}
Endometrial/uterine cancer	PWWP2A	cefuroxime	246	247	−0.0501	3.75×10^{-78}
Endometrial/uterine cancer	ZNF479	1-acetyl-4-methylpiperazine	246	247	−0.0527	3.69×10^{-76}
Endometrial/uterine cancer	FOXR1	Y-27152	246	247	−0.0503	4.47×10^{-76}
Esophageal cancer	TP63	adatanserin	91	92	0.0304	7.12×10^{-53}
Esophageal cancer	TP63	amylene-hydrate	91	92	0.0534	3.29×10^{-52}
Esophageal cancer	TP63	sirolimus	91	92	−0.0580	1.34×10^{-51}
Esophageal cancer	TP63	ABT-702	91	92	−0.0368	7.18×10^{-51}
Esophageal cancer	TP63	monastrol	91	92	−0.0580	8.22×10^{-51}
Gastric cancer	AKT3	AT13387	206	206	−0.2148	7.19×10^{-50}
Gastric cancer	DDR2	CTS-1027	206	206	−0.0282	3.25×10^{-48}
Gastric cancer	RUNX1T1	monensin	206	206	−0.1472	4.13×10^{-48}
Gastric cancer	AKT3	tubastatin-a	206	206	−0.0237	1.34×10^{-47}
Gastric cancer	DDR2	acetanilide	206	206	0.0168	1.52×10^{-47}
Head and neck cancer	DNMT3A	nalfurafine	249	250	0.0344	6.42×10^{-65}
Head and neck cancer	PRKD1	TPCA-1	249	250	0.1473	7.13×10^{-64}
Head and neck cancer	DNMT3A	vatalanib	249	250	−0.0566	9.18×10^{-64}
Head and neck cancer	DNMT3A	S18986	249	250	0.0389	1.59×10^{-63}
Head and neck cancer	ZEB1	SGI-1776	249	250	0.1152	4.42×10^{-62}
Kidney cancer	KDR	tacrolimus	356	356	−0.0555	6.13×10^{-120}
Kidney cancer	KDR	pizotifen	356	356	−0.0350	2.42×10^{-119}

(*Continued on next page*)

Table 5. *Continued*

Cancer type	Gene	Drug	Number of samples of high gene expression	Number of samples of low gene expression	Δlog2FC	Adjusted p value of t test
Kidney cancer	KDR	pheniramine	356	356	−0.0601	1.74×10^{-118}
Kidney cancer	KDR	sotrastaurin	356	356	−0.0768	2.43×10^{-118}
Kidney cancer	KDR	NPC-01	356	356	0.0189	4.64×10^{-118}
Liver cancer	LYL1	dasatinib	179	179	0.0921	9.73×10^{-33}
Liver cancer	LYL1	enprofylline	179	179	−0.0320	7.20×10^{-32}
Liver cancer	LYL1	IB-MECA	179	179	0.0211	1.09×10^{-31}
Liver cancer	LYL1	KU-60019	179	179	−0.0316	2.59×10^{-31}
Liver cancer	LYL1	metaraminol	179	179	0.0336	1.25×10^{-29}
Lung cancer	NKX2-1	ouabain	493	493	0.3524	1.82×10^{-180}
Lung cancer	NKX2-1	dovitinib	493	493	0.2937	2.92×10^{-172}
Lung cancer	NKX2-1	daunorubicin	493	493	0.2145	2.13×10^{-167}
Lung cancer	NKX2-1	YM-155	493	493	0.2525	3.56×10^{-166}
Lung cancer	NKX2-1	doxorubicin	493	493	0.3202	7.36×10^{-166}
Pancreatic cancer	PREX2	zaldaride	85	85	0.0835	1.81×10^{-19}
Pancreatic cancer	PREX2	meglitinide	85	85	0.0811	3.33×10^{-19}
Pancreatic cancer	PREX2	Ro-25-6981	85	85	−0.0308	1.98×10^{-18}
Pancreatic cancer	PREX2	butylated-hydroxyanisole	85	85	0.0570	2.04×10^{-18}
Pancreatic cancer	PREX2	SB-408124	85	85	0.0431	2.29×10^{-18}
Prostate cancer	ZEB1	SC-144	246	248	0.1358	2.13×10^{-53}
Prostate cancer	ZEB1	AR-42	246	248	0.1324	2.29×10^{-53}
Prostate cancer	ZEB1	oridonin	246	248	0.1314	2.33×10^{-53}
Prostate cancer	ZEB1	M-344	246	248	0.1251	2.70×10^{-53}
Prostate cancer	ZEB1	alvespimycin	246	248	0.1157	3.40×10^{-53}
Sarcoma	AFF3	pumosetrag	117	117	−0.0109	3.30×10^{-18}
Sarcoma	AFF3	SB-203186	117	117	−0.0214	3.30×10^{-18}
Sarcoma	AFF3	methylatropine-nitrate	117	117	−0.0170	3.31×10^{-18}
Sarcoma	AFF3	emorfazone	117	117	−0.0382	3.31×10^{-18}
Sarcoma	AFF3	imidafenacin	117	117	−0.0437	3.33×10^{-18}
Skin cancer	MITF	pemirolast	232	233	−0.0325	4.26×10^{-40}
Skin cancer	MITF	clorotepine	232	233	−0.0527	1.68×10^{-39}
Skin cancer	MITF	semagacestat	232	233	−0.0675	6.40×10^{-38}
Skin cancer	MITF	lonidamine	232	233	−0.0502	1.19×10^{-37}
Skin cancer	MITF	enflurane	232	233	−0.0405	1.53×10^{-37}
Thyroid cancer	ASXL2	serdemetan	245	245	−0.0835	2.38×10^{-44}
Thyroid cancer	ASXL2	iproniazid	245	245	−0.0962	2.53×10^{-44}
Thyroid cancer	ASXL2	FPL-55712	245	245	−0.0403	1.66×10^{-43}
Thyroid cancer	ASXL2	eperezolid	245	245	−0.0324	3.31×10^{-43}
Thyroid cancer	ASXL2	bergapten	245	245	−0.0234	3.50×10^{-43}

Data and code availability

This study analyzes existing, publicly available data. The scRNA-seq data are available at https://singlecell.broadinstitute.org/single_cell/study/SCP542/pan-cancer-cell-line-heterogeneity.[43] The TCGA data are available at https://xenabrowser.net/datapages/.[44] The CCLE data are available at https://depmap.org/portal/ (2022Q2 version).[45] The PRISM dataset is available at https://depmap.org/repurposing/.[46] All original code has been deposited at https://github.com/ychsu2014/Predicting_drug_response_through_tumor_deconvolution_by_cancer_cell_lines and archived at Zenodo.[47]

Datasets and preprocessing

The scRNA unique molecular identifier (UMI) count data, which represent the absolute number of observed transcripts per cell from cancer cell lines, were downloaded from the Broad Institute Single Cell Portal (http://singlecell.broadinstitute.org/single_cell/study/SCP542/pan-cancer-cell-line-heterogeneity),[43] which contains 56,982 cells from 207 cell lines of 22 cancer types. The Python package Scanpy was used to preprocess the scRNA dataset. We first filtered out cells with fewer than 200 expressed genes and removed genes expressed in fewer than 3 cells. To remove cell doublets or

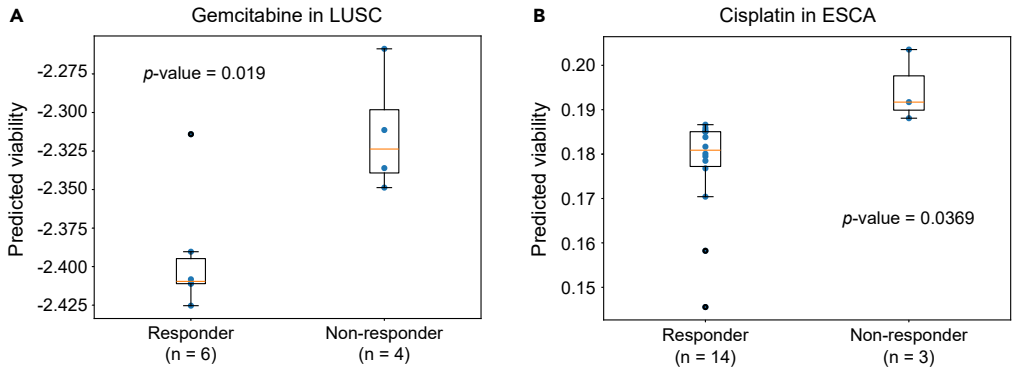


Figure 6. Validation of the drug response prediction algorithm by clinical drug treatment data

(A) Boxplot for the predicted cell viability in patients with lung squamous cell carcinoma (LUSC) who did or did not respond to gemcitabine.

(B) Boxplot for the predicted cell viability in patients with esophageal carcinoma (ESCA) among those who did or did not respond to cisplatin.

Samples used were from TCGA. The definition of boxplots (box range, whiskers, and outliers) is the same as described in Figure 1, and the definition of p values are the same as Figure 3.

multiplets, we filtered out cells with more than or equal to 6,000 expressed genes. Also, to avoid low-quality cells with extensive mitochondrial contamination, we kept cells with fewer than 13% mitochondrial counts. This data filtering resulted in a total of 53,887 cells and 25,385 genes being included in the subsequent analysis. The data were further normalized to 10,000 counts per cell to make read counts comparable among cells.

TCGA RNA, somatic mutation, and phenotypic data were downloaded from the University of California, Santa Cruz (UCSC) Xena browser (https://xenabrowser.net/datapages/).[44] The downloaded TCGA RNA data were normalized and log2 transformed, and the somatic mutation data contained only point mutations. The phenotypic data contained the primary cancer in each sample, which we used for deciding which samples to include in subsequent analyses. For further validation, we downloaded the TCGA clinical drug treatment data by the R/Bioconductor package, TCGAbiolinks (https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html).

CCLE RNA and mutation data were downloaded from the DepMap portal (v.2022Q2; https://depmap.org/portal/).[45] CCLE RNA data were represented as $\log_2$(TPM+1) values, where TPM stands for transcripts per million. The mutation data included both point mutations and indels, and only point mutations were used in our study.

The PRISM drug screening dataset was downloaded from the DepMap website (https://depmap.org/repurposing/),[46] which includes a large dataset of viability assays from a novel DNA barcode-based approach capable of performing high-throughput viability analyses for thousands of drugs against over 900 human cancer cell lines. We downloaded and used the PRISM primary dataset, which contains drug sensitivity data of 568 cancer cell lines treated with 4,518 drugs. The replicate-collapsed log-transformed fold change drug response data from the PRISM dataset were used in our analyses.

For further analyses, we used only TCGA samples with corresponding cancer types in the scRNA dataset and cell lines found in all scRNA, CCLE, and PRISM datasets, which resulted in the inclusion of 7,781 TCGA tumor samples and 187 cell lines of 18 cancers (Table S1) and drug response data of the 187 cell lines treated with 4,518 drugs. The 18 cancer types are lung cancer, breast cancer, kidney cancer, brain cancer, head and neck cancer, prostate cancer, endometrial/uterine cancer, thyroid cancer, skin cancer, gastric cancer, bladder cancer, colon/colorectal cancer, liver cancer, sarcoma, esophageal cancer, pancreatic cancer, bile duct cancer, and ovarian cancer.

To infer drug-related or cancer-related information from mutations of tumors, we used curated data from the oncoKB database (https://www.oncokb.org/)[48] and the COSMIC website (https://cancer.sanger.ac.uk/census).[49] The oncoKB database contains information on drugs that are associated with clinical applications and mutation events, and they are further stratified by different levels of evidence that support the use of certain drugs.[48] The COSMIC CGC (v.97) lists the expert-curated driver genes (n = 736) of human cancers and is widely used in cancer research.

Drug response prediction via a cancer cell line-guided deconvolution scheme

We aimed to explore drug repurposing via a novel deconvolution methodology to predict drug responses in tumors by deconvoluting tumors into cancer cell lines from which the drug response of the tumor sample was predicted. The analysis included deconvolution model training, model application, and data interpretation (Figure 1A). First, we trained, validated, and tested a deep learning-based deconvolution model (Scaden-CA) using simulated bulk RNA data from an scRNA dataset for each of the 18 cancers. Then, the Scaden-CA models (for different cell types) were applied to the CCLE bulk RNA dataset (new to the model) to validate whether the model can accurately deconvolute the cell lines to themselves. We transferred the trained model to tumor samples (i.e., TCGA samples) to obtain the proportions of the cancer cell lines by cancers, and these proportions were then input into the drug response predictor along with PRISM screening data. To further explore potential drug repurposing and the relevant mechanisms, we visualized and interpreted the drug response results for all cancer-gene-drug and cancer-[gene, aa change]-drug combinations. The details of the three main parts of the analyses are described below.

Bulk RNA data simulation and cell deconvolution model training/testing

We adopted a previously published deep learning-based cell composition analysis model, Scaden, for bulk RNA data simulation and cancer cell line deconvolution,[14] which we termed Scaden-CA. The normalized scRNA data were used to generate 8,000 artificial bulk RNA samples using 500 single cells per sample for each cancer type by the Scaden "simulate" function. This step simulates the artificial bulk RNA samples based on randomly generated proportions for cell lines of the specific cancer type. The simulated samples were split into training and testing data (80:20 ratio). Then the training and testing data were log2 transformed and scaled to the range between 0 and 1, and genes that overlapped between simulation and TCGA RNA data were preserved by using the Scaden "process" function. The resulting genes included in the Scaden-CA model for each cancer type are summarized in Table S3. After training the deconvolution models by cancer types, the models were tested separately on the individual testing data of different cancer types. The performance of the deconvolution models was first assessed by calculating Lin's CCC[50] as in the formula

$$CCC = \frac{2r\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} \quad \text{(Equation 1)}$$

where x, y stands for the true and predicted proportions, r is the correlation coefficient, σ_x and σ_y are the standard deviations of x and y, and μ_x and μ_y are the means of x and y. The CCC is 1 when true and predicted proportions are identical. However, CCC will be less than 1 if true and predicted proportions

possessed the same profile but with an offset ($\mu_x - \mu_y \neq 0$). In addition, for an overall evaluation of the model performance, we also calculated MAE and RMSE for the deconvoluted results.

Combining oncoKB and COSMIC data in cell line selection

The performance of the Scaden-CA model depends on the number of cancer cell lines used for deconvolution. Thus, we estimated the percent error of the Scaden-CA model by different ground-truth proportions from the simulated data. We first calculated the percent error for each deconvolution proportions by the following mathematical formula:

$$\%\ Error = \left|\frac{y - x}{x}\right| \times 100 \qquad \text{(Equation 2)}$$

Then, we grouped the percent errors by intervals of ground-truth proportions and visualized the results by bar plots, with error bars indicating the standard error of the percent errors within the same interval. The approximate number of cell lines suitable for model training was determined by the interval that showed small and stabilized percent errors.

To reduce the number of cell lines, we first used clusters of scRNA data to select representative cell lines. However, we observed no obvious clusters in the lung cancer cell lines (Figure S5). We then explored the possibility of selecting cell lines by other omics data. Since mutations may affect the choice of clinical therapeutics and the outcomes, we selected representative cell lines using mutation data from CCLE and TCGA. We used the following three conditions to select representative cell lines that harbor overlapping mutations between TCGA tumors and CCLE cell lines: (1) either in oncoKB or on COSMIC cancer driver genes, (2) included in oncoKB, or (3) actionable targets in oncoKB. We sought to keep the cell lines with as many clinically associated or cancer driver gene-associated mutations as possible, since drug treatments are determined by specific mutations carried by cancer patients.

The above three filtering criteria are actually the set covering problem, a classical nondeterministic polynomial-time (NP)-hard problem in combinatorial optimization. We applied the greedy algorithm to solve the set covering problem. After we applied the filtering criteria to all of the cancer types, we reran all the model training and testing steps and assessed the performance of the modified Scaden-CA models. We decided to reduce the numbers of cell lines used in the deconvolution model of lung cancer, while other cancer models remained relatively unchanged in classification errors. Also, we chose the second filtering criterion for lung cancer to include more cell lines (n = 19) and achieved good performance (CCC = 0.941).

Deconvolution model validation by the CCLE dataset

The trained deconvolution models were further validated using the CCLE bulk RNA dataset. For each of the 18 cancer types, the corresponding cell lines were deconvoluted to assess whether the models could produce accurate deconvolution results. We defined the CDR as the proportion of the cell line being deconvoluted by the cell line itself. In other words, if one cell line was deconvoluted mostly into itself among the different cell lines included in the Scaden-CA model, then the CDR from the deconvolution results will be near 1 to constitute an excellent deconvolution.

Simulation method assessment

Since our deconvolution model was trained on the simulated bulk RNA data, it is crucial to evaluate data generated by the simulation method representing real-bulk RNA data. To achieve this objective, we first used the simulation method to create pseudo-bulk RNA data from the deconvolution results from the CCLE real-bulk RNA. Then, we normalized the pseudo-bulk RNA data to $\log_2$(TPM+1) values and compared them with the real CCLE bulk RNA data by calculating the unexplained variation, $V_{unexplained}$, with the mathematical formula

$$V_{unexplained} = 1 - R^2 \qquad \text{(Equation 3)}$$

where R represents the correlation coefficient calculated by Pearson correlation. Next, we compared our results with one previously developed simulation method used for evaluation of their deconvolution model, MuSiC2.[19] Their simulation method for obtaining artificial bulk RNA data was through sampling from Poisson distribution of the inferred condition-specific mean expression and summation of read counts from all of the simulated single cells.[19] We created another set of pseudo-bulk RNA data from our deconvolution results from CCLE real-bulk RNA data, normalized the values to $\log_2$(TPM+1), and then compared it with the results of our simulation method to calculate the unexplained variation using Equation 3.

TCGA data deconvolution by the Scaden-CA model

To infer potential drug repurposing based on patients' transcriptomic traits, the TCGA RNA data were deconvoluted to proportions of cancer cell lines by the models trained for the 18 cancer types included in the study. To estimate how much the deconvoluted proportions explained the original data from tumors, we created pseudo-bulk RNA data for TCGA tumors and compared them with TCGA real-bulk RNA data to calculate the unexplained variation using Equation 3.

Drug response prediction algorithm for deconvoluted TCGA tumors

By linking these deconvoluted cancer cell line fractions with the treatment results of cancer cell lines from the PRISM dataset, we devised a simple average for predicting drug responses, $R_i = \log_2 F_i$, where F_i is the fold change of i-th tumor in response to treatment. Let N and $\tilde{N}$ be numbers of cells before and after the treatment, and we have

$$F_i = \frac{\tilde{N}}{N} = \frac{1}{N}\sum\nolimits_k n_k f_k = \sum\nolimits_k p_k f_k \qquad \text{(Equation 4)}$$

where p_k is the fraction of the k-th cell line estimated by the Scaden-CA model of the i-th tumor (and $\sum p_k = 1$), and f_k is the fold change of the k-th cell line derived from the PRISM dataset (where $\log_2 f_k$ was provided).

Further exploration and visualization of the drug response prediction

To identify possible drug repurposing and explore the underlying mechanisms, we first used mutation profiles to group TCGA tumors at gene or mutation levels. We then performed t tests on all cancer-gene-drug and cancer-[gene, aa change]-drug combinations. We also used gene expression data to group TCGA tumors to test whether genetic alternations at the RNA level affect the drug response. We split TCGA tumors into high and low gene expression groups on one of the CGC cancer driver genes, performed t tests on the cancer-gene-drug combinations related to this gene to assess significance in drug response differences, and repeated this process for all of the CGC genes. We selected t test to evaluate the significance based on our observation that the log-transformed cell viability measurements appear to possess normal distribution. Bonferroni correction was applied to adjust for multiple testing. An adjusted p value less than 0.05 was used as the threshold for statistical significance.

Validation of the predicted drug response by TCGA clinical data

To validate our predicted drug response, we analyzed the TCGA clinical drug treatment data to compare the predicted cell viability between responders and non-responders of TCGA tumors. We included entries with recorded drug responses, such as "complete response," "partial response," "stable disease," and "clinical progressive disease" and defined the former two as responders and the latter two as non-responders. The cell viability differences between responders and non-responders were assessed using t tests with statistical significance set as a p value less than 0.05.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100949.

ACKNOWLEDGMENTS

This study was supported by the Graduate Students Study Abroad Program sponsored by the National Science and Technology Council, Taiwan (to Y.-C.H.); the Cancer Prevention and Research Institute of Texas (RP220662 to Y.C. and RP190346 to Y.C.); and NIH NCI P30 CA054174 (to Y.C.).

AUTHOR CONTRIBUTIONS

Y.-C.H., Y.-C.C., T.-H.H., and Y.C. conceived the study. Y.-C.H. performed the data analysis and contributed to writing of the manuscript draft. Y.C. contributed to the review and editing of the manuscript. Y.C., T.-P.L., and T.-H.H. provided supervision and administrative support.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 4, 2023
Revised: February 7, 2024
Accepted: February 12, 2024
Published: March 5, 2024

REFERENCES

1. Stanta, G., and Bonin, S. (2018). Overview on Clinical Relevance of Intra-Tumor Heterogeneity. Front. Med. *5*, 85. https://doi.org/10.3389/fmed.2018.00085.
2. Fedele, C., Tothill, R.W., and McArthur, G.A. (2014). Navigating the challenge of tumor heterogeneity in cancer therapy. Cancer Discov. *4*, 146–148. https://doi.org/10.1158/2159-8290.CD-13-1042.
3. Evans, W.E., and Relling, M.V. (2004). Moving towards individualized medicine with pharmacogenomics. Nature *429*, 464–468. https://doi.org/10.1038/nature02626.
4. Chiu, Y.-C., Chen, H.-I.H., Zhang, T., Zhang, S., Gorthi, A., Wang, L.-J., Huang, Y., and Chen, Y. (2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med. Genom. *12*, 18. https://doi.org/10.1186/s12920-018-0460-9.
5. Chiu, Y.C., Chen, H.I.H., Gorthi, A., Mostavi, M., Zheng, S., Huang, Y., and Chen, Y. (2020). Deep learning of pharmacogenomics resources: moving towards precision oncology. Briefings Bioinf. *21*, 2066–2083. https://doi.org/10.1093/bib/bbz144.
6. Fan, J., Feng, Y., Cheng, Y., Wang, Z., Zhao, H., Galan, E.A., Liao, Q., Cui, S., Zhang, W., and Ma, S. (2021). Multiplex gene quantification as digital markers for extremely rapid evaluation of chemo-drug sensitivity. Patterns *2*, 100360. https://doi.org/10.1016/j.patter.2021.100360.
7. Huang, S., Hu, P., and Lakowski, T.M. (2021). Predicting breast cancer drug response using a multiple-layer cell line drug response network model. BMC Cancer *21*, 648. https://doi.org/10.1186/s12885-021-08359-6.
8. Wang, C., Zhang, M., Zhao, J., Li, B., Xiao, X., and Zhang, Y. (2023). The prediction of drug sensitivity by multi-omics fusion reveals the heterogeneity of drug response in pan-cancer. Comput. Biol. Med. *163*, 107220. https://doi.org/10.1016/j.compbiomed.2023.107220.
9. Rydzewski, N.R., Peterson, E., Lang, J.M., Yu, M., Laura Chang, S., Sjöström, M., Bakhtiar, H., Song, G., Helzer, K.T., Bootsma, M.L., et al. (2021). Predicting cancer drug TARGETS - TreAtment Response Generalized Elastic-neT Signatures. NPJ Genom. Med. *6*, 76. https://doi.org/10.1038/s41525-021-00239-z.
10. Li, Y., Umbach, D.M., Krahn, J.M., Shats, I., Li, X., and Li, L. (2021). Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. BMC Genom. *22*, 272. https://doi.org/10.1186/s12864-021-07581-7.
11. Park, A., Lee, Y., and Nam, S. (2023). A performance evaluation of drug response prediction models for individual drugs. Sci. Rep. *13*, 11911. https://doi.org/10.1038/s41598-023-39179-2.
12. Jia, P., Hu, R., Pei, G., Dai, Y., Wang, Y.Y., and Zhao, Z. (2021). Deep generative neural network for accurate drug response imputation. Nat. Commun. *12*, 1740. https://doi.org/10.1038/s41467-021-21997-5.
13. Liu, Q., Hu, Z., Jiang, R., and Zhou, M. (2020). DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics *36*, i911–i918. https://doi.org/10.1093/bioinformatics/btaa822.
14. Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., Heutink, P., and Bonn, S. (2020). Deep learning-based cell composition analysis from tissue expression profiles. Sci. Adv. *6*, eaba2619. https://doi.org/10.1126/sciadv.aba2619.
15. Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics *34*, 1969–1979. https://doi.org/10.1093/bioinformatics/bty019.
16. Shen-Orr, S.S., and Gaujoux, R. (2013). Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr. Opin. Immunol. *25*, 571–578. https://doi.org/10.1016/j.coi.2013.09.015.
17. Yadav, V.K., and De, S. (2015). An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. Briefings Bioinf. *16*, 232–241. https://doi.org/10.1093/bib/bbu002.
18. Gambardella, G., Viscido, G., Tumaini, B., Isacchi, A., Bosotti, R., and di Bernardo, D. (2022). A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. Nat. Commun. *13*, 1714. https://doi.org/10.1038/s41467-022-29358-6.
19. Fan, J., Lyu, Y., Zhang, Q., Wang, X., Li, M., and Xiao, R. (2022). MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. Briefings Bioinf. *23*, bbac430. https://doi.org/10.1093/bib/bbac430.
20. Tseng, C.N., Hong, Y.R., Chang, H.W., Yu, T.J., Hung, T.W., Hou, M.F., Yuan, S.S.F., Cho, C.L., Liu, C.T., Chiu, C.C., and Huang, C.J. (2014). Brefeldin A reduces anchorage-independent survival, cancer stem cell potential and migration of MDA-MB-231 human breast cancer cells. Molecules *19*, 17464–17477. https://doi.org/10.3390/molecules191117464.
21. Wang, L., Cui, Y., Sheng, J., Yang, Y., Kuang, G., Fan, Y., Jin, J., and Zhang, Q. (2017). Epigenetic inactivation of HOXA11, a novel functional tumor suppressor for renal cell carcinoma, is associated with RCC TNM classification. Oncotarget *8*, 21861–21870. https://doi.org/10.18632/oncotarget.15668.
22. Song, E., Ma, X., Li, H., Zhang, P., Ni, D., Chen, W., Gao, Y., Fan, Y., Pang, H., Shi, T., et al. (2013). Attenuation of kruppel-like factor 4 facilitates carcinogenesis by inducing g1/s phase arrest in clear cell renal cell carcinoma. PLoS One *8*, e67758. https://doi.org/10.1371/journal.pone.0067758.
23. Choueiri, T.K., Vaishampayan, U., Rosenberg, J.E., Logan, T.F., Harzstark, A.L., Bukowski, R.M., Rini, B.I., Srinivas, S., Stein, M.N., Adams, L.M., et al. (2013). Phase II and biomarker study of the dual MET/VEGFR2 inhibitor foretinib in patients with papillary renal cell carcinoma. J. Clin. Oncol. *31*, 181–186. https://doi.org/10.1200/JCO.2012.43.3383.
24. Venugopal, B., Baird, R., Kristeleit, R.S., Plummer, R., Cowan, R., Stewart, A., Fourneau, N., Hellemans, P., Elsayed, Y., McClue, S., et al. (2013). A phase I study of quisinostat (JNJ-26481585), an oral hydroxamate histone deacetylase inhibitor with evidence of target modulation and antitumor activity, in patients with advanced solid tumors. Clin. Cancer Res. *19*, 4262–4272. https://doi.org/10.1158/1078-0432.CCR-13-0312.
25. Shanchun, H., You, P., Sujuan, N., Xuebing, Z., Yijie, B., Xiaohui, X., Jianming, H., La, N., Zhehui, B., Qi, L., and Wulong, J. (2023). Integrative analyses of biomarkers and pathways for metformin reversing cisplatin resistance in head and neck squamous cell carcinoma cells. Arch. Oral Biol. *147*, 105637. https://doi.org/10.1016/j.archoralbio.2023.105637.
26. Yu, Z., Song, Y., Cai, M., Jiang, B., Zhang, Z., Wang, L., Jiang, Y., Zou, L., Liu, X., Yu, N., et al. (2021). PPM1D is a potential prognostic biomarker and correlates with immune cell infiltration in hepatocellular carcinoma. Aging *13*, 21294–21308. https://doi.org/10.18632/aging.203459.
27. Xie, W., Sun, Y., Zeng, Y., Hu, L., Zhi, J., Ling, H., Zheng, X., Ruan, X., and Gao, M. (2022). Comprehensive analysis of PPPCs family reveals the clinical significance of PPP1CA and PPP4C in breast cancer. Bioengineered *13*, 190–205. https://doi.org/10.1080/21655979.2021.2012316.
28. Wang, C., Tran-Thanh, D., Moreno, J.C., Cawthorn, T.R., Jacks, L.M., Wang, D.Y., McCready, D.R., and Done, S.J. (2011). Expression of Abl interactor 1 and its prognostic significance in breast cancer: a tissue-array-based investigation. Breast Cancer Res. Treat. *129*, 373–386. https://doi.org/10.1007/s10549-010-1241-0.
29. Kurenova, E.V., Hunt, D.L., He, D., Magis, A.T., Ostrov, D.A., and Cance, W.G. (2009). Small molecule chloropyramine hydrochloride (C4) targets

the binding site of focal adhesion kinase and vascular endothelial growth factor receptor 3 and suppresses breast cancer growth in vivo. J. Med. Chem. *52*, 4716–4724. https://doi.org/10.1021/jm900159g.

30. Burgess, J.T., Bolderson, E., Saunus, J.M., Zhang, S.D., Reid, L.E., McNicol, A.M., Lakhani, S.R., Cuff, K., Richard, K., Richard, D.J., and O'Byrne, K.J. (2016). SASH1 mediates sensitivity of breast cancer cells to chloropyramine and is associated with prognosis in breast cancer. Oncotarget *7*, 72807–72818. https://doi.org/10.18632/oncotarget.12020.
31. Abdel-Ghany, S., Raslan, S., Tombuloglu, H., Shamseddin, A., Cevik, E., Said, O.A., Madyan, E.F., Senel, M., Bozkurt, A., Rehman, S., and Sabit, H. (2020). Vorinostat-loaded titanium oxide nanoparticles (anatase) induce G2/M cell cycle arrest in breast cancer cells via PALB2 upregulation. 3 Biotech *10*, 407. https://doi.org/10.1007/s13205-020-02391-2.
32. Palczewski, M.B., Kuschman, H.P., Bovee, R., Hickok, J.R., and Thomas, D.D. (2021). Vorinostat exhibits anticancer effects in triple-negative breast cancer cells by preventing nitric oxide-driven histone deacetylation. Biol. Chem. *402*, 501–512. https://doi.org/10.1515/hsz-2020-0323.
33. Wawruszak, A., Borkiewicz, L., Okon, E., Kukula-Koch, W., Afshan, S., and Halasa, M. (2021). Vorinostat (SAHA) and Breast Cancer: An Overview. Cancers *13*, 4700. https://doi.org/10.3390/cancers13184700.
34. Foggetti, G., Ottaggio, L., Russo, D., Mazzitelli, C., Monti, P., Degan, P., Miele, M., Fronza, G., and Menichini, P. (2019). Autophagy induced by SAHA affects mutant P53 degradation and cancer cell survival. Biosci. Rep. *39*. https://doi.org/10.1042/BSR20181345.
35. Krall, E.B., Wang, B., Munoz, D.M., Ilic, N., Raghavan, S., Niederst, M.J., Yu, K., Ruddy, D.A., Aguirre, A.J., Kim, J.W., et al. (2017). KEAP1 loss modulates sensitivity to kinase targeted therapy in lung cancer. Elife *6*, e18970. https://doi.org/10.7554/eLife.18970.
36. Jeong, Y., Hellyer, J.A., Stehr, H., Hoang, N.T., Niu, X., Das, M., Padda, S.K., Ramchandran, K., Neal, J.W., Wakelee, H., and Diehn, M. (2020). Role of KEAP1/NFE2L2 Mutations in the Chemotherapeutic Response of Patients with Non-Small Cell Lung Cancer. Clin. Cancer Res. *26*, 274–281. https://doi.org/10.1158/1078-0432.CCR-19-1237.
37. Solis, L.M., Behrens, C., Dong, W., Suraokar, M., Ozburn, N.C., Moran, C.A., Corvalan, A.H., Biswal, S., Swisher, S.G., Bekele, B.N., et al. (2010). Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. Clin. Cancer Res. *16*, 3743–3753. https://doi.org/10.1158/1078-0432.CCR-09-3352.
38. André, F., Ciruelos, E., Rubovszky, G., Campone, M., Loibl, S., Rugo, H.S., Iwata, H., Conte, P., Mayer, I.A., Kaufman, B., et al. (2019). Alpelisib for PIK3CA-Mutated, Hormone Receptor-Positive Advanced Breast Cancer. N. Engl. J. Med. *380*, 1929–1940. https://doi.org/10.1056/NEJMoa1813904.
39. Kalous, O., Conklin, D., Desai, A.J., Dering, J., Goldstein, J., Ginther, C., Anderson, L., Lu, M., Kolarova, T., Eckardt, M.A., et al. (2013). AMG 900, pan-Aurora kinase inhibitor, preferentially inhibits the proliferation of breast cancer cell lines with dysfunctional p53. Breast Cancer Res. Treat. *141*, 397–408. https://doi.org/10.1007/s10549-013-2702-z.
40. Yiliyaer, and Maimaiti, Y. (2019). Aurora kinases: novel anti-breast cancer targets. Oncology and Translational Medicine *5*, 43–48. https://doi.org/10.1007/s10330-018-0315-5.
41. Park, S.H., Kim, J.H., Ko, E., Kim, J.Y., Park, M.J., Kim, M.J., Seo, H., Li, S., and Lee, J.Y. (2018). Resistance to gefitinib and cross-resistance to irreversible EGFR-TKIs mediated by disruption of the Keap1-Nrf2 pathway in human lung cancer cells. Faseb. J. *32*, 5862–5873. https://doi.org/10.1096/fj.201800011R.
42. Hast, B.E., Cloer, E.W., Goldfarb, D., Li, H., Siesser, P.F., Yan, F., Walter, V., Zheng, N., Hayes, D.N., and Major, M.B. (2014). Cancer-derived mutations in KEAP1 impair NRF2 degradation but not ubiquitination. Cancer Res. *74*, 808–817. https://doi.org/10.1158/0008-5472.CAN-13-1655.
43. Kinker, G.S., Greenwald, A.C., Tal, R., Orlova, Z., Cuoco, M.S., McFarland, J.M., Warren, A., Rodman, C., Roth, J.A., Bender, S.A., et al. (2020). Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. Nat. Genet. *52*, 1208–1218. https://doi.org/10.1038/s41588-020-00726-6.
44. Goldman, M.J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. Nat. Biotechnol. *38*, 675–678. https://doi.org/10.1038/s41587-020-0546-8.
45. DepMap (2022). DepMap 22Q2 Public. Figshare. https://doi.org/10.6084/m9.figshare.19700056.v2.
46. Corsello, S.M., Nagari, R.T., Spangler, R.D., Rossen, J., Kocak, M., Bryan, J.G., Humeidi, R., Peck, D., Wu, X., Tang, A.A., et al. (2020). Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. Nat. Can. (Ott.) *1*, 235–248. https://doi.org/10.1038/s43018-019-0018-6.
47. Hsu, Y., Chiu, Y., Lu, T., Hsiao, T., and Chen, Y. (2024). Codes for the paper "Predicting drug response through tumor deconvolution by cancer cell lines". Zenodo. https://doi.org/10.5281/zenodo.10528026.
48. Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis. Oncol. *2017*, 1–16. https://doi.org/10.1200/PO.17.00011.
49. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/s41568-018-0060-1.
50. Lin, L.I. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics *45*, 255–268. https://doi.org/10.2307/2532051.