THE UNIVERSITY OF MINNESOTA

THE DEVELOPMENT AND APPLICATION OF MACHINE LEARNING FOR DRUG DISCOVERY AND DRUG RESPONSE PREDICTION FOR PERSONALIZED CANCER TREATMENT

A DISSERTATION SUBMITTED TO THE BIOINFORMATICS AND COMPUTATIONAL BIOLOGY PROGRAM IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

BY
DANIELLE MICHELLE-MAESER STOVER

ADVISOR: R. STEPHANIE HUANG

SPRING 2024

ACKNOWLEDGEMENT

I would like to acknowledge and thank my advisors at the University of Minnesota, Dr. R. Stephanie Huang, Dr. Yuk Sham, and Dr. Chad Myers for believing in me and constantly supporting me throughout my journey in the Bioinformatics and Computational Biology program. Thank you for this incredible opportunity to grow as a computational scientist and help contribute to cancer research. I'd like to thank my PhD committee for sharing their subject expertise and passion for discovery with me. Thank you Dr. R. Stephanie Huang, Dr. Chad Myers, Dr. Michael Olin, and Dr. David Largaespada. A special thanks to my incredible colleagues in Huang Lab and coworkers Yingbo Huang, Weijie Zhang, Dr. Robert Gruener, Dr. Robert Galvin, Dr. Tomo Koga, and Dr. Florina-Nicoleta Grigore. You've all been incredibly kind and wonderful to work with. I'd also like to acknowledge Dr. Clark Chen for his contributions to my doctoral preliminary examination committee. Additionally, I'd like to thank my mentors at Merck, Kenneth Lynn and Dr. Cory White, as well as my mentor at Moderna, Dr. Chiao-wen Hsiao, for providing rigorous training and internship opportunities for me to learn new methodologies and improve as a data scientist throughout my PhD. I'm deeply honored to have worked alongside each of you.

DEDICATION

This work is dedicated to my amazing and incredible family. I'm forever grateful to you. To my parents, Michelle and Robert, my brothers Jake and Luke, and my twin sister Nikki. To my husband Peter. Thank you for always supporting and inspiring my dreams. I love you.

To the one who inspired me to pursue this journey in the first place; life wouldn't be the same without you. And to Trent in heaven. I'm grateful to you for setting me on this path in life. It is my deepest wish that the work I do can help make a difference for those fighting cancer.

ABSTRACT

In the field of pharmacogenomics and precision medicine, gene expression analysis has become a crucial tool in predicting patient drug response. My contributions to this field come primarily in the development and application of two bioinformatic packages: *oncoPredict* and *scIDUC*. *oncoPredict* is a tool based in the R programming language, primarily used to predict the response of various cancer samples (cell line, patient, etc.) to different drugs. This is made possible by incorporating machine learning to analyze the complex relationships between genomic features and drug response from pan-cancer cell lines. These relationships are learned from microarray or bulk RNA sequencing (RNA-seq) data and high-throughput drug screens, then applied to patient data to generate novel drug discovery hypotheses. In turn, *oncoPredict* aids in identifying potential drug candidates, understanding mechanisms of drug resistance, and predicting the effectiveness of drugs on specific cancer types. *scIDUC* (single-cell Integration and Drug Utility Computation) is a computational framework based in python that quickly and accurately generates predictions of drug response for cells derived from single-cell RNA sequencing (scRNA-seq) data. It is a transfer learning-based approach that learns relationships between drug sensitivities and relevant gene expression patterns based on cell line bulk RNA-seq data and high-throughput drug screens, similar to *oncoPredict*. The key difference, however, is that prior to training drug response models, *scIDUC* integrates bulk RNA-seq and target scRNA-seq data to denoise and extract shared gene expression patterns between bulk and single-cell data sources. The resulting bulk data is then used to train drug response models, whose coefficients are further applied to post-integration single-cell data to infer cellular drug sensitivity scores.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 3 Tables:

Chapter 4 Tables:

Chapter 5 Tables:

LIST OF FIGURES

Chapter 2 Figures:

Chapter 3 Figures:

Chapter 4 Figures:

Chapter 5 Figures:

Chapter 6 Figures:

CHAPTER 1: CONCEPTUAL OVERVIEW OF THESIS, BACKGROUND & SIGNIFICANCE

The Relationship Between Gene Expression and Drug Response Plays a Significant Role in Improving Clinical Decision-Making

Gene expression significantly influences drug response. The expression levels of certain genes can determine how a patient's body will react to a drug, affecting both the efficacy and the potential for adverse side effects.[1-7] For example, the gene expression levels of drug transporters or proteins that help move substances across cell membranes can affect how well a drug is absorbed, distributed, metabolized, and excreted from the body.[3] Additionally, gene expression levels of drug targets or genes involved in pathways that are activated or inhibited by drugs also affect how a patient responds to such drugs.[1,3] Overall, the genome provides a dynamic view of what's happening in a cell at a given time, reflecting the influence of both genetic factors and environmental conditions.

Other genomic factors that may contribute to drug response include but are not limited to genetic mutations, copy number variations (CNVs), and epigenetic changes.[8-11] Mutations in a gene can change the structure and function of the resulting protein, which can affect how it interacts with other molecules and influences gene expression pathways. The presence of a certain mutation might make a tumor particularly susceptible or resistant to a specific drug.[9] CNVs refer to changes in the number of copies of a particular gene or region of the genome, ranging from deletions to duplications or even multiple copies. Duplications or multiple copies of a gene can lead to overexpression, while deletions can lead to a decrease in a gene's expression. If CNVs involve regulatory elements (enhancers or promoters), the extent to which disruption occurs in gene

expression is heavily broadened. As a result, CNVs can influence drug response, especially if they result in the amplification or deletion of genes that are critical for the effectiveness of the drug.[10] Epigenetic changes refer to modifications to the DNA or associated proteins that affect gene activity without altering the DNA sequence itself. They include things like DNA methylation, histone modification, and RNA associated silencing. These changes can regulate when, how and to what extent genes are expressed. These changes can affect drug metabolism and therefore cellular response to therapy.[8,10] While the informativeness of gene expression toward drug response is maximized when considered alongside these other genomic factors, gene expression may in some cases, be the most critical predictor. This is potentially the result of gene expression reflecting the impact of these genomic factors.[10,11]

Large Scale Cell Line and Drug Screening Data is Invaluable For Drug Discovery

The GDSC (Genomics of Drug Sensitivity in Cancer)[10] and CTRP (Cancer Therapeutics Response Portal)[11] are two significant databases in the field of cancer research, particularly for understanding the relationship between cancer genomics (including mutations, CNVs, and gene expression profiles for pan-cancer cell lines) and drug response. High-throughput cell line drug screening data like these have been used to train machine learning models, aiming to translate in vitro drug response to in vivo tumor response predictions.[7,12-19] Cell line drug screening data refer to the information gathered from experiments where various drugs are tested on different cell lines to assess and measure their effects. This data has proven to be an invaluable source in the field of drug development, offering insights into drug efficacy, mechanisms of action, and the genetic basis of drug response while also reducing the time and cost associated with bringing new drugs to market.

A Regression Framework Enables Accurate Drug Response Prediction From Sequencing Data

Predicting patient drug response from bulk sequencing methods such as microarray or bulk RNA

sequencing (RNA-seq) data as well as from these large scale cell line drug databases is an important application of machine learning and personalized medicine.[7,12-19] Bulk sequencing methods average gene expression in a sample of cells, making them less granular than single-cell RNA-seq (scRNA-seq) data but still valuable for predicting how patients will respond to drugs. A number of machine learning algorithms have been utilized for drug response prediction, including but not limited to regression (linear and logistic), support vector machines (SVMs), random forests, and neural networks.[7,12-19] However, we have found ridge regression to be effective for predicting drug response, and it poses a number of advantages beyond its high accuracy; its architecture is transparent, and it's appropriate for data with high-dimensionality and multicollinearity as in the case of gene expression data.[20] Specifically, it's appropriate because it adds a regularization or penalty term to the loss function, and this regularization technique helps prevent overfitting.[20]

A regression framework also enabled us to predict continuous drug response values instead of binary (dichotomous) ones, which offers several advantages. Continuous drug response values are preferred because binary categorizations (e.g. responder or sensitive vs. non-responder or resistant) can oversimplify complex biological relationships, reducing the granularity of data, potentially leading to misclassification and masking important variations, impacting research findings and clinical decisions.[21] Drug responses often exist on a spectrum rather than as binary outcomes.[22] By capturing the full spectrum of responses, continuous data captures nuanced differences in drug response, supporting the goals of personalized medicine. Additionally, continuous outcomes typically provide more statistical power than binary outcomes. This means that studies using continuous data may require fewer subjects to detect a given effect size, making research more efficient and potentially less costly. More statistical power also means a greater ability to detect real differences or associations when they exist.

oncoPredict and scIDUC Accurately Infer Drug Response from Bulk and Single-Cell

Sequencing Data, Respectively

In this body of work, we compile our own set of tools developed for the primary purpose of drug response prediction from bulk patient data into an R package called *oncoPredict*, presented in Chapter 2. *oncoPredict* utilizes a ridge regression framework applied to gene expression data to predict response on a continuous scale.[7] We then apply these tools to a rare and aggressive form of brain cancer, glioblastoma multiforme (GBM), in Chapter 3, to accomplish two primary goals: to identify drug candidates to combat this deadly disease where standard of care treatment falls short and to identify biomarkers for these drug candidates, to improve the efficacy of drug discovery by selecting the right patient populations for subsequent evaluation. GBM poses significant challenges in treatment due to its complexity and heterogeneity. While our approach identified a key relationship between MEK (mitogen-activated protein kinase) inhibitors and *PHGDH* (phosphoglycerate dehydrogenase) gene expression, monotherapy generally has limited efficacy in treating GBM due to the tumor's high level of heterogeneity and its ability to rapidly develop resistance to drugs.[23,24] As such, multimodal approaches may be more effective.[25] Therefore, there is great need for computational approaches to acknowledge the unique gene expression at the cellular level and leverage scRNA-seq data to make cellular level drug response predictions.

Predicting drug response at the single-cell level is an emerging area in cancer treatment.[26] It involves the application of machine learning techniques to analyze data at the individual cell level to understand how different cells within a patient's body might respond to a specific drug or treatment. Tumors and other disease tissues are often highly heterogeneous. By taking into account a tumor's heterogeneity in this way we can take an even greater step toward a more personalized treatment plan.[27-29] Every patient's cellular makeup is unique, and their response to drugs can vary widely. In the drug development process, understanding how various cell types respond to a drug at a granular level can help in screening and optimizing drug candidates more effectively. Thereby

devising personalized treatment plans, helping healthcare providers avoid ineffective treatments, reducing the rate of treatment failure as well as associated costs and ultimately improving treatment efficacy and patient outcomes. Traditional methods, which average responses from bulk tissue samples, can miss important nuances.[26] Unlike bulk RNA-seq data, scRNA-seq allows for the examination of this heterogeneity, providing a comprehensive view of gene expression with much finer granularity and insights into how different cell populations within the same tissue respond to treatment. Imputing drug response at the cellular level does pose several major challenges however. Firstly, there is a lack of training data to be used to learn the relationships between cellular level gene expression and drug response at the single-cell level. Secondly, single-cell data is highly sparse due to several inherent characteristics of the technology and biological processes.[30] For example, each cell contains a relatively small amount of RNA, and this low starting material can lead to incomplete representation of the RNA molecules present in the cell in the sequencing data.[30] Given the high dimensionality of single-cell RNA-seq data, it's statistically likely that only a subset of genes will be detected in each cell, leading to sparse data matrices where many entries are zero.[30] Dropout events are also common, which occurs when RNA molecules present in the cell are not detected during sequencing. This can lead to inefficiencies in the reverse or amplification steps to result in zero or very low read counts for some genes that are actually expressed in the cell. In this body of work, we develop and refine methods to address these challenges in a python package called *scIDUC*, presented in Chapter 4, helping to unlock the potential of personalized medicine at the single-cell level.[26] We have applied *scIDUC* to various cancer types, supporting its validity, versatility, and superiority over competing methods. Chapter 5 provides a systematic review of *scIDUC* along these competing approaches.

REFERENCES

---

1. Adam, G. *et al.* Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precis. Onc.* **4**, 19 (2020).

2. Lv, W. *et al.* Exploring effects of DNA methylation and gene expression on pan-cancer drug response by mathematical models. *Exp Biol Med (Maywood)* **246**, 1626–1642 (2021).

3. Sadée, W. & Dai, Z. Pharmacogenetics/genomics and personalized medicine. *Human Molecular Genetics* **14**, R207–R214 (2005).

4. Roden, D. M., Wilke, R. A., Kroemer, H. K. & Stein, C. M. Pharmacogenomics: The Genetics of Variable Drug Responses. *Circulation* **123**, 1661–1670 (2011).

5. Willyard, C. Copy number variations' effect on drug response still overlooked. *Nat Med* **21**, 206–206 (2015).

6. Tang, Y.-C., Powell, R. T. & Gottlieb, A. Molecular pathways enhance drug response prediction using transfer learning from cell lines to tumors and patient-derived xenografts. *Sci Rep* **12**, 16109 (2022).

7. Maeser, D., Gruener, R. F. & Huang, R. S. oncoPredict: an R package for predicting *in vivo* or cancer patient drug response and biomarkers from cell line screening data. *Briefings in Bioinformatics* **22**, bbab260 (2021).

8. Gibney, E. R. & Nolan, C. M. Epigenetics and gene expression. *Heredity* **105**, 4–13 (2010).

9. Plass, C. *et al.* Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat Rev Genet* **14**, 765–780 (2013).

10. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).

11. Seashore-Ludlow, B. *et al.* Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery* **5**, 1210–1223 (2015).

12. Xie, M., Lei, X., Zhong, J., Ouyang, J. & Li, G. Drug response prediction using graph representation learning and Laplacian feature selection. *BMC Bioinformatics* **23**, 532 (2022).

13. Lenhof, K., Eckhart, L., Gerstner, N., Kehl, T. & Lenhof, H.-P. Simultaneous regression and classification for drug sensitivity prediction using an advanced random forest method. *Sci Rep* **12**, 13458 (2022).

14. Kurilov, R., Haibe-Kains, B. & Brors, B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Sci Rep* **10**, 2849 (2020).

15. Park, A., Lee, Y. & Nam, S. A performance evaluation of drug response prediction models for individual drugs. *Sci Rep* **13**, 11911 (2023).

16. Liu, C. *et al.* An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Molecular Therapy - Nucleic Acids* **21**, 676–686 (2020).

17. Dong, Z. *et al.* Anticancer drug sensitivity prediction in cell lines from baseline gene

expression through recursive feature selection. *BMC Cancer* **15**, 489 (2015).

18. Aben, N., Vis, D. J., Michaut, M. & Wessels, L. F. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics* **32**, i413–i420 (2016).

19. Ali, M. & Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev* **11**, 31–39 (2019).

20. Arashi, M., Roozbeh, M., Hamzah, N. A. & Gasparini, M. Ridge regression and its applications in genetic studies. *PLoS ONE* **16**, e0245376 (2021).

21. Schmitz, S., Adams, R. & Walsh, C. The use of continuous data versus binary data in MTC models: A case study in rheumatoid arthritis. *BMC Med Res Methodol* **12**, 167 (2012).

22. Bouhaddou, M. *et al.* Drug response consistency in CCLE and CGP. *Nature*

540, E9–E10 (2016).

23. Becker, A., Sells, B., Haque, S. & Chakravarti, A. Tumor Heterogeneity in Glioblastomas: From Light Microscopy to Molecular Pathology. *Cancers* **13**, 761 (2021).

24. Dymova, M. A., Kuligina, E. V. & Richter, V. A. Molecular Mechanisms of Drug Resistance in Glioblastoma. *IJMS* **22**, 6385 (2021).

25. McBain, C. *et al.* Treatment options for progression or recurrence of glioblastoma: a network meta-analysis. *Cochrane Database of Systematic Reviews* **2021**, (2021).

26. Zhang, W. *et al. Inferring therapeutic vulnerability within tumors through integration of pan-cancer cell line and single-cell transcriptomic profiles*. http://biorxiv.org/lookup/doi/10.1101/2023.10.29.564598 (2023) doi:10.1101/2023.10.29.564598.

27. Fustero-Torre, C. *et al.* Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. *Genome Med* **13**, 187 (2021).

28. Suphavilai, C. *et al.* Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Med* **13**, 189 (2021).

29. Wu, Z. *et al.* Single-Cell Techniques and Deep Learning in Predicting Drug Response. *Trends in Pharmacological Sciences* **41**, 1050–1065 (2020).

30. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat Commun* **11**, 1169 (2020).

CHAPTER 2: ONCOPREDICT R PACKAGE FOR PREDICTING IN VIVO OR CANCER PATIENT DRUG RESPONSE AND BIOMARKERS FROM CELL LINE SCREENING DATA

Danielle Maeser[1†], Robert F. Gruener[2†], Dr. R. Stephanie Huang[3*]

*Corresponding author

[†]Equal contribution

1. Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN 55455
2. Ben May Department for Cancer Research, University of Chicago, Chicago, IL 60637, USA
3. Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, MN 55455, USA

CONTRIBUTIONS BY FIRST AUTHORS

Danielle Maeser: package developer, code and method contributor, manuscript writer/reviewer.

Robert Gruener: code and method contributor, manuscript writer/reviewer.

- Danielle developed the R package and submitted it to CRAN R repository. Developed code to overhaul and generalize previous methodologies, namely pRRophetic originally developed by Paul Geeleher, initially hard coded to specific datasets. Introduced additional functionality into the code, including computing correlation between imputed drug response and gene expression to identify expression based biomarkers of drug response, computing regression metrics such as R-squared for each drug response model, computing principal component regression as an alternative to ridge regression, etc. Overall, increasing user customization and generalizability of the code so that it can be applied to any dataset based on user defined parameters.
- Robert pre-processed and cleaned CTRP and GDSC screening data to ensure high quality data. This included removing gene duplicates, checking for unreliable drugs, investigating outliers, etc. Unreliable drugs in GDSC were defined as those whose IC50 values are consistently over its max concentration tested. This suggests that the drug's efficacy at the tested concentrations is limited.
- Danielle and Robert compared the performance across different models to reconfirm choice of ridge regression using the R package *ridge*. This included comparing performance across two R packages, *glmnet* and *ridge*, to determine whether choice of source code impacted downstream analysis; in brief, *glmnet* was observed to result in less variability in imputed drug response, and so we selected *ridge*. Specifically: in the case of *glmnet*, some drug models were so poor that a constant predictor like the null hypothesis of a horizontal line with an intercept equal to the mean of the dependent variable would fit the observed data better. In this case, the residual sum of squares exceeded the total sum of squares, and the R-squared value is negative. This was not the case for *ridge*. This may be due to a variety of factors. For example, *glmnet* uses coordinate descent for fitting its models, and *ridge* employs a direct matrix algebra solution for estimating the ridge regression coefficients. Our comparison also included comparing performance (e.g. mean squared error) across different choices of the regularization parameter (lambda), which

controls the impact of the penalty term. Therefore, comparing lasso regression, ridge regression, and elastic regression through cross-validation. While results were similar across these regression techniques, ridge regression performed best. We shared responsibility for application of GLDS, calcPhenotype, and IDWAS to the data presented in this paper.

INTRODUCTION

High-throughput cancer cell line screening datasets are important resources for drug discovery [1]. The two largest publicly available screening efforts are the Broad Institute's Cancer Therapeutics Response Portal (CTRP) [2] and Sanger's Genomics of Drug Sensitivity in Cancer (GDSC) [3]. By providing multi-omic cell line features as well as summarized drug response information, these screens enable identifying associations between a compound's activity and genomic features. These data have become important tools for both biomarker discovery and for developing predictive models of drug response (reviewed in [4–9]). We previously published work using cell line screening data for robust biomarker discovery [10], drug response imputations [11] and subsequent discovery of novel biomarkers from drug response imputations [12].

Here, we offer a new R package, oncoPredict, that provides a convenient wrapper for these three separate methodologies (Figure 1). oncoPredict enables easy implementation of each of the three methodologies, with each pipeline being summarized into a single R function: GLDS, calcPhenotype and IDWAS, respectively. calcPhenotype allows for fitting and predicting drug response based on baseline transcriptomic cell line data [4]. This functionality is similar to our original pRRophetic package [13] but has been overhauled to enable faster implementation, provides updated CTRP and GDSC screening data and allows for increased user customization. oncoPredict also integrates two methods of biomarker

identification. The first method, imputed drug-wide association study (IDWAS), identifies biomarkers directly in clinical data based on associations of predicted drug sensitivity with copy number variation (CNV) and other somatic mutation data [12]. The second method performs biomarker discovery directly in the cell line dataset and aims at identifying drug-specific biomarkers after accounting for variability in general levels of drug sensitivity (GLDS) [10].

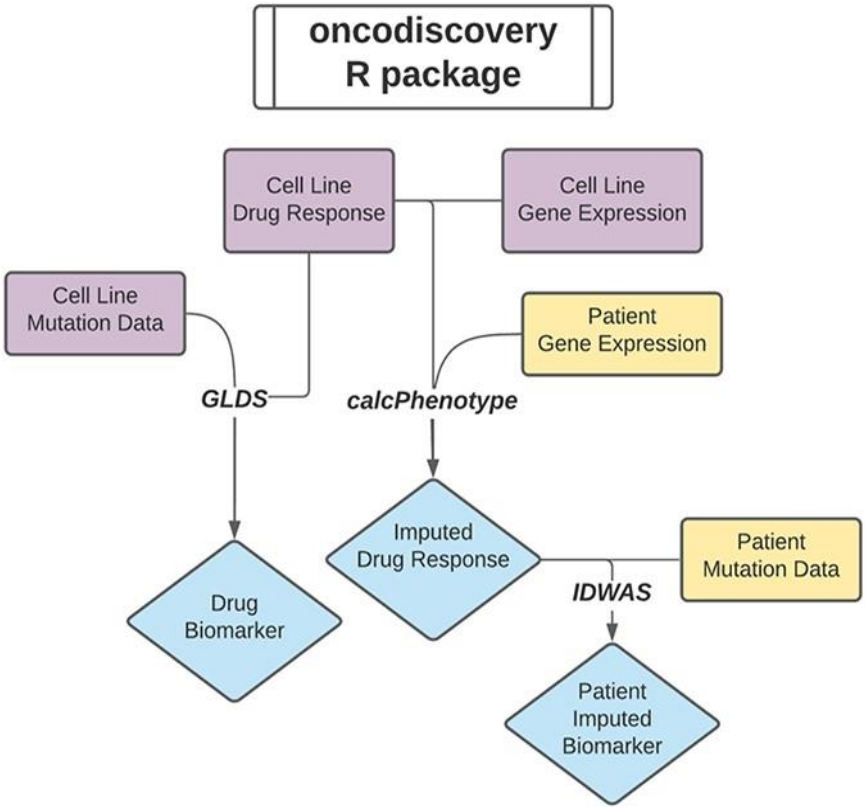


Figure 1. Overview of the oncoPredict R package. Flow diagram showing the inputs (rectangles), the functions (lines) and outputs (diamonds) from oncoPredict. This covers the three primary functionalities included in the package for drug response and biomarker prediction. Purple and yellow indicate cell line and patient input data—these inputs are anticipated to be the most common for the associated function, but the functions are flexible and capable of accepting data from other types of biological systems as discussed in the text.

By integrating these methodologies together into one R package, we enable easy implementation of these computational drug discovery tools. While previously described, implementing these methods became inaccessible to many potential users due to difficulty of finding the code, inflexibility of the code for adapting to new uses as well as R version updates. This R package, however, provides updated and convenient implementation of each pipeline. Additionally, integrating these three methods will allow users to perform simultaneous biomarker and imputations for drug discovery, similar to that done in [14], which could increase the impact of their computational discoveries.

oncoPredict bridges *in vitro* and *in vivo* data by expanding the utility of existing *in vitro* data to enable *in vivo* discovery. The primary use cases for each function are summarized below. Vignettes are provided in our Github repository to cover special use cases. All R source codes are publicly available via GitHub and on our website (see Availability section).

METHODS & RESULTS

To facilitate usability and robustness of oncoPredict, the transcriptomic and cell line response data from CTRP and GDSC have been downloaded, processed and included. For CTRP, drug response data were obtained from the Cancer Target Discovery and Development Network established by the National Cancer Institute's Office of Cancer Genomics [15]. The corresponding gene expression data are hosted by the Broad Institute's Cancer Cell Line Encyclopedia (CCLE) data portal [16]. The GDSC gene expression and drug response values were downloaded from the GDSC website [17]. GDSC consists of two datasets, GDSC1 and GDSC2 dataset, and the data from both are included in oncoPredict. These data are together referred to as GDSC in the remainder of this paper. After download, the data were processed and formatted into gene expression or drug response data matrices.

GENERAL LEVEL OF DRUG SENSITIVITY: GLDS

Method Principle

We previously reported on a GLDS phenomenon and its effect on biomarker identification. GLDS is the observation that individuals in a population (cell lines or patients) can be, in general, more sensitive, or more resistant to a therapy, no matter the treatment. As shown in the original paper, this phenomenon is related to, but not necessarily the same as, multi-drug resistance (MDR). Correcting for this variable allows for biomarker identification that is more specific to the drug of interest.

To correct for GLDS, for a given drug, a set of negative control (i.e. completely unrelated) drugs are selected. These negative control drugs are unrelated mechanistically and have a measured cellular response that is not highly correlated ($r_s < 0.7$) with the drug of interest. GLDS is then estimated as the first 10 principal components (PCs) of the set of negative control response values in a panel of cancer cell lines. These PCs were then included as covariates in a regression model for drug sensitivity against gene mutations with the goal to identify drug-specific biomarker(s). If negative control drugs are not available for a compound, a gene signature can also be utilized to correct for GLDS. This gene signature is a way of estimating GLDS without dependence on large-scale drug screening datasets.

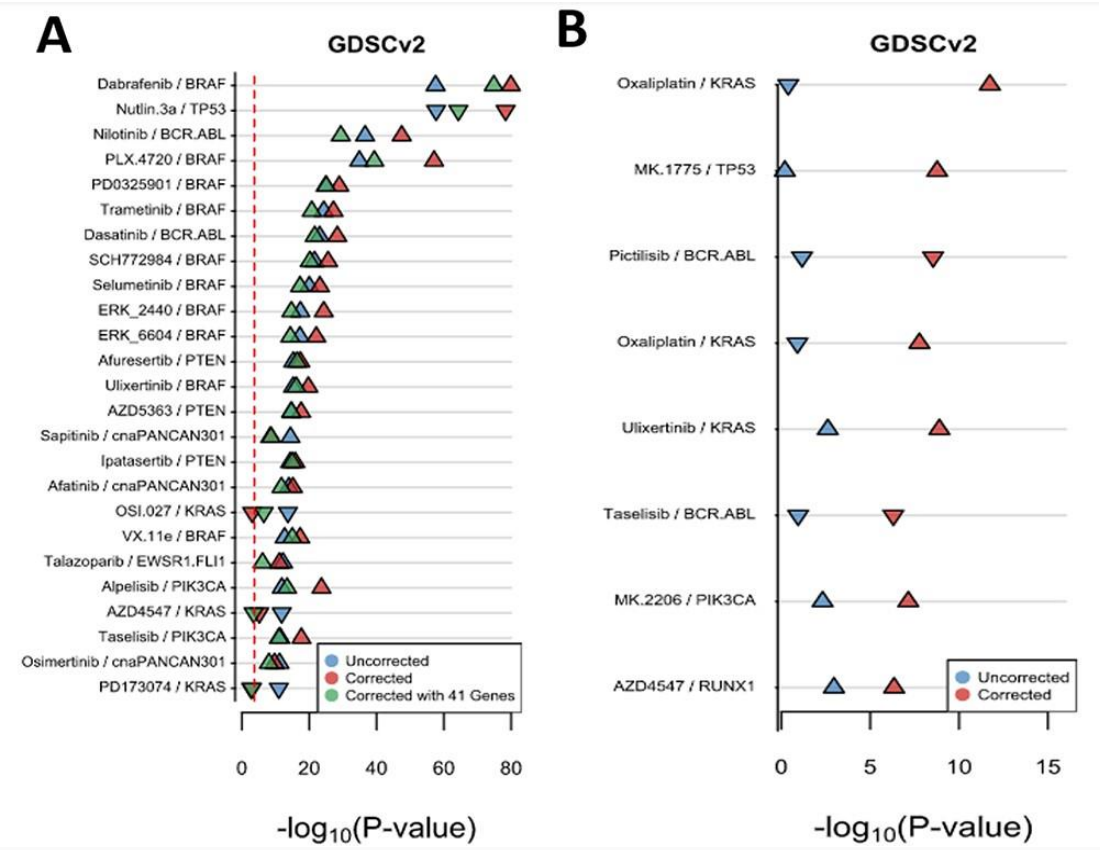


Figure 2. Correcting for GLDS Improves Biomarker Identification in GDSCv2. (**A**) Dot plot of top gene–drug associations with and without correction for GLDS. Also included is the significance of the associations after correcting for GLDS using a gene set derived from the data. The direction of the triangle indicates the direction of the association: triangle pointing up indicates the drug is more effected in the mutated setting. Red dash line represents significance. (**B**) Similar graph as in (**A**), but graphed are gene–drug associations that were not significant in the original data (blue) but were significant after correcting for GLDS.

Function Utility and Updates

The GLDS function identifies GLDS uncorrected and corrected associations between drug sensitivity and a phenotype of interest, like somatic mutation. We have included the outputs obtained from applying the function to several large-scale drug screening datasets (including GDSC and CTRP) in oncoPredict's

Open Science Framework (OSF). Also included in oncoPredict's OSF are the data required to generate these associations, including updated GDSC and CTRP drug relatedness and CNV and coding variant data for pan-cancer. This output includes all drug–gene relationships which have been uncorrected and corrected for GLDS. The GLDS function has been modified to improve usability when applied to datasets beyond GDSC and CTRP. We also modified the function to give the user the ability to adjust the relatedness parameter, selecting the threshold for high correlation when filtering negative control drugs. This provides a more reliable method of filtering negative controls than what was previously employed in GLDS.

In addition, one can now generate a gene signature as a surrogate of GLDS using the package by providing GLDS function with a gene expression profile. The expression of these genes is significantly correlated to each of the 10 PCs used to estimate GLDS across a drug screening dataset. In other words, they also represent negative controls. The performance of both the PC method and the gene signature method is shown in Figure 2 using GDSCv2 data.

Example Use Case

We originally showcased this method by correcting for GLDS in Sanger's GDSC 2014 screening data. Here, we applied our package to identify biomarkers of response with and without controlling for GLDS in the separate GDSCv2 dataset (Figure 2). In addition, we used the GLDS signature functionality to generate a 41-gene signature from this dataset, consisting of genes highly correlated with GLDS. Thus, the package enables two ways of conditioning on GLDS; one is through the top 10 PCs, which is implemented in the GLDS function, and the other is with a 41-gene signature, which users can use as a way to correct for GLDS without the use of additional large-scale drug screening data. Consistent with previous results, genes associated with MDR, such as CYR61, were included in this signature.

Figure 2A shows the top gene–drug associations with and without correcting for GLDS. The gene–drug associations shown here were corrected for GLDS both by directly estimating GLDS from the data and by estimating using the 41-gene signature set to show the concordance of the two methods and to provide an estimation for GLDS. In other case, significant gene–drug association disappears after being conditioned on GLDS. For example, PD173074 (a FGFR inhibitor) was predicted to have a significant association with *KRAS* mutations without GLDS correction; however, this association was insignificant with correction for GLDS. These observations support previous findings that controlling for the first 10 PCs or gene signature can be used to describe GLDS. In our previous work, we showed that correcting for GLDS reduced spurious findings. Similar results are shown here. On the other hand, some meaningful associations can only be found when conditioned with GLDS as shown in Figure 2B. For example, KRAS is an established biomarker for oxaliplatin [18], but this association was only identified after correcting for GLDS.

PREDICTING DRUG RESPONSE FUNCTION: CALCPHENOTYPE

Method Principle

This function implements the pipeline for the prediction of clinical chemotherapeutic response by using only baseline tumor gene expression data. A complete technical description of this pipeline is described in [11]. Briefly, large-scale gene expression and drug screening data (training dataset) are used to build ridge regression models that can then be applied to new gene expression datasets (testing dataset) to yield drug sensitivity predictions for the new dataset [19]. These drug models are built following removal or summarization of gene duplication, homogenization (batch correction) and filtering of low-varying genes. In oncoPredict, users can use preset standards that have been identified depending on the type of data presented (microarray, RNA-Sequencing, etc.; see vignettes) or specify summarization, homogenization and filtering parameters. Finally, our calcPhenotype function is applied to the processed, standardized and filtered clinical tumor expression data, yielding a drug sensitivity prediction for each patient.

Function Utility and Updates

This calcPhenotype function can be used in retrospective studies to predict drug response and compare with the observed patient response to investigate model performance. Recently, however, we showed that patient-imputed drug response can be used for drug discovery directly in the patient data [14]. That is, by looking for differences in the imputed drug response values between two related cancer subsets, we could identify compounds showing higher efficacy toward one cancer subtype. Finally, this functionality can also be used to quickly turn any patient transcriptome dataset into a dataset for biomarker discovery as discussed with the IDWAS function.

Users can either use the GDSC or CTRP data that are prepackaged into the oncoPredict as the training datasets or can supply their own training data in order to predict drug response in any gene expression matrix provided. Genes that are duplicated in either the training or testing datasets are removed or summarized by the mean expression based on user input. Vignettes describe the optimal options for preprocessing the training and testing expression data for microarray, RNA-Seq or a mix of both. Compared to the method published with the original manuscript, the calcPhenotype function was updated to give users more options for model generation and evaluation. The default predictive model provided is ridge regression, but principal component regression (PCR) is also available in calcPhenotype as this method also performs well in drug response predictions [20]. Regression evaluation metrics (R2) can easily be obtained for each model. Lastly, the function was updated to provide an option to calculate the correlations between predicted drug response and gene expression as a way of identifying expression-based biomarkers of drug response.

Example Use Case

To assess the model performance of the calcPhenotype function, we imputed drug response in an ovarian cancer clinical dataset, which had both patient response to paclitaxel and gene expression microarray data

(GSE51373) [21]. We used the CTRP dataset to train a model of paclitaxel drug response. We then stratified patients by their observed clinical outcome (defined by the trial as resistant or sensitive to paclitaxel treatment) and compared the predicted sensitivity scores (Figure 3). We see that imputed paclitaxel sensitivity is able to correctly stratify patients into their responder/non-responder categories ($P = 0.0075$ by *t*-test). Additionally, the area under the receiver operating characteristic (ROC) curve was 0.79, which is consistent with the performance of the models of the original paper [11].

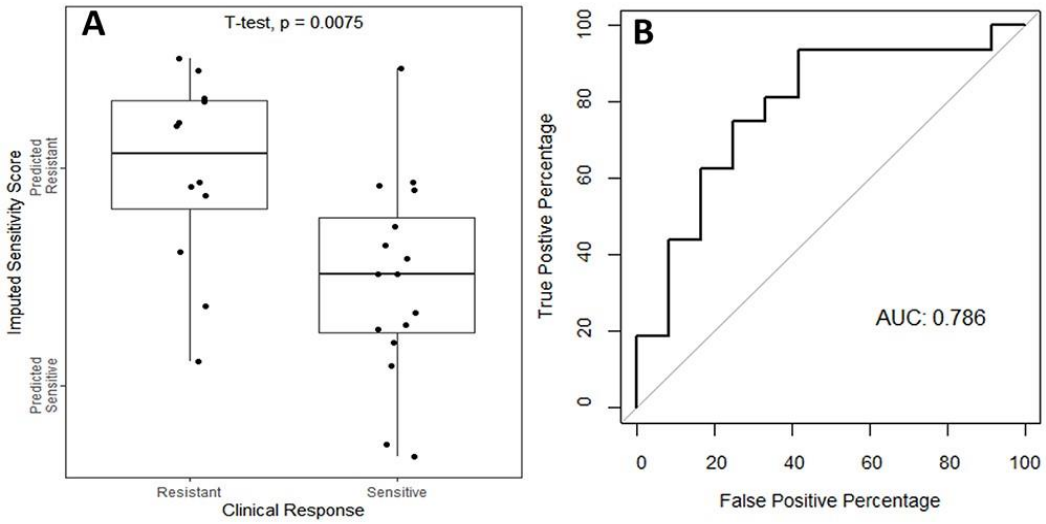


Figure 3. Paclitaxel imputed response in ovarian cancer clinical trial (GSE51373). (**A**) Patient expression data were downloaded from GEO, processed and then used as a testing expression data for the oncoPredict calcPhenotype function. Actual clinical outcomes for the patient were then plotted against the imputed drug response output (*y*-axis, a continuous response value). (**B**) ROC curve showing the percentage of true positives against the percentage of false positives as the classification threshold is varied.

IDENTIFYING BIOMARKERS FROM PATIENT DATA: IDWAS

Method Principle

The IDWAS approach is an extension of the imputing drug response values that readily enables biomarkers identification in that population. IDWAS is similar in conception and implementation to

GWAS. That is, associations between imputed drug response values and either somatic mutations or CNV are determined by using linear models in R in order to estimate drug–gene interactions and identify biomarkers of drug response. By taking advantage of the large sample size of clinical datasets, this allows for the identification of novel relationships that are not seen in cell line datasets. Additionally, since it uses patient genomic data, any biomarkers found are pertinent to that patient population. The full methodology for the imputed drug-wide association study (IDWAS) is described in [12].

Function Utility and Updates

To outline the method employed by the IDWAS function, the function takes two inputs: a matrix of patient genetic features and a matrix of imputed drug response values (e.g. the output of the calcPhenotype function). The genetic features can either be a mutation matrix containing 0/1 to describe wild-type/mutated gene status or a CNV matrix containing the CNV levels of the genes in patients. The IDWAS function allows flexible usage of either type of data. Additionally, by integrating the IDWAS and calcPhenotype functions into a single package, the use of the IDWAS methodology is greatly facilitated. Users are able to easily obtain drug response predictions in any clinical dataset by using calcphenotype and then search for biomarkers by using IDWAS. Vignettes are provided to demonstrate how to download and apply this methodology to the TCGA datasets, as described in the below use case. To aid in usability, pan-cancer TCGA mutation and CNV matrices are provided in oncoPredict. However, users could easily adapt this application to other datasets like the International Cancer Genome Consortium or cancer-specific datasets such as METABRIC.

Example Use Case

The Cancer Genome Atlas (TCGA) [22] gene expression data were downloaded from firebrowse.org, and mutation data were downloaded using the TCGAbiolinks R package [23]. TCGA mutation data were summarized on a per-gene basis by calling a gene mutated if the protein had any mutation that would

affect the amino acid sequence, which was formatted into a mutation matrix. For the genes assessed in this paper, genes mutated with a frequency above 0.5% across all patients were included.

Using oncoPredict's calcPhentope function, we built drug response models with the CTRP data. We imputed drug response for all 496 drugs across the TCGA samples ($n = 8536$). Similarly, to our previous report [12], the top significant association was that between nutlin-3a (an MDM2 inhibitor) and *TP53* status. As can be seen in Figure 4A, this is a highly significant and specific association. The direction of this association (not shown) is consistent with expectations and indicates that *TP53* mutations render nutlin-3a ineffective. Conversely, for selumetinib (a MEK inhibitor), the association was highly and specifically associated with being more effective in the *KRAS* mutated setting (Figure 4B).

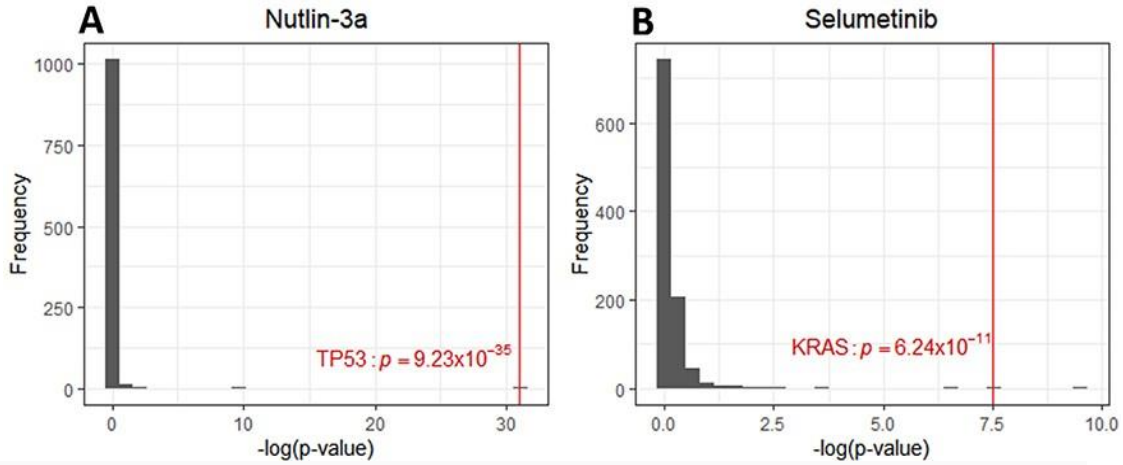


Figure 4. **IDWAS in TCGA patient data.** (A, B) Using the oncoPredict calcphenotype function, imputed sensitivities were generated for each patient against all the drugs in CTRP. The IDWAS function then allowed for testing associations between patient cancer gene mutations and patient imputed drug response values. For nutlin-3a (A) and selumetinib (B), histograms that show the significance (*P*-value) frequency across all the gene–drug associations are plotted.

ROAD MAP FOR FUTURE UPDATES

To ease use, we have downloaded and packaged the most up-to-date CTRP and GDSC data, two of the largest publicly available drug screening datasets. The package will be updated twice yearly to include any major updates to these data. Additionally, we will incorporate new and appropriate screening data resources as they become available.

Further feature requests and ongoing changes to increase usability will continue over the life of the package. For example, we have plans to further integrate calcPhenotype, IDWAS and GLDS to make the computational discovery of biomarkers and drug response predictions even more seamless based on user feedback. By integrating these distinct but separate methodologies, the package will give users a one-stop shop instead of the previously dispersed and disaggregated methods linked to individual papers.

The pipelines implemented in the package are 'state-of-the-art' and represent accurate and convenient drug response prediction methods. In general, however, the field lacks guidelines for the best use and application of drug response prediction methods. Similar to the inclusion of PCR as an option in the calcPhenotype function, as we and others in the field identify gold standards for drug response predictions, oncoPredict, will continue to update.

CONCLUSION

In conclusion, we have presented the R package oncoPredict that bridges the *in vitro* drug screening with *in vivo* drug and biomarker discovery. One can easily predict patient tumor response to a large number of

drugs screened in cancer cell lines and can perform biomarker discovery both with and without GLDS condition.

REFERENCES

---

[1] A. Ling, R. F. Gruener, J. Fessler, and R. S. Huang, "More than Fishing for a Cure: The Promises and Pitfalls of High Throughput Cancer Cell Line Screens," *Pharmacol Ther*, vol. 191, pp. 178–189, Nov. 2018, doi: 10.1016/j.pharmthera.2018.06.014.

[2] B. Seashore-Ludlow *et al.*, "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset," *Cancer Discov*, vol. 5, no. 11, pp. 1210–1223, Nov. 2015, doi: 10.1158/2159-8290.CD-15-0235.

[3] F. Iorio *et al.*, "A Landscape of Pharmacogenomic Interactions in Cancer," *Cell*, vol. 166, no. 3, pp. 740–754, Jul. 2016, doi: 10.1016/j.cell.2016.06.017.

[4] J. C. Costello *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nat Biotechnol*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014, doi: 10.1038/nbt.2877.

[5] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," *Pac Symp Biocomput*, pp. 63–74, 2014.

[6] F. Azuaje, "Computational models for predicting drug responses in cancer research," *Brief Bioinform*, vol. 18, no. 5, pp. 820–829, Sep. 2017, doi: 10.1093/bib/bbw065.

[7] M. Ali and T. Aittokallio, “Machine learning and feature selection for drug response prediction in precision oncology applications,” *Biophys Rev*, vol. 11, no. 1, pp. 31–39, Feb. 2019, doi: 10.1007/s12551-018-0446-z.

[8] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, “Machine learning approaches to drug response prediction: challenges and recent progress,” *npj Precision Oncology*, vol. 4, no. 1, Art. no. 1, Jun. 2020, doi: 10.1038/s41698-020-0122-1.

[9] B. Güvenç Paltun, H. Mamitsuka, and S. Kaski, “Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches,” *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 346–359, Jan. 2021, doi: 10.1093/bib/bbz153.

[10] P. Geeleher, N. J. Cox, and R. S. Huang, “Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models,” *Genome Biol*, vol. 17, no. 1, p. 190, Sep. 2016, doi: 10.1186/s13059-016-1050-9.

[11] P. Geeleher, N. J. Cox, and R. S. Huang, “Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines,” *Genome Biology*, vol. 15, p. R47, Mar. 2014, doi: 10.1186/gb-2014-15-3-r47.

[12] P. Geeleher *et al.*, “Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies,” *Genome Res.*, vol. 27, no. 10, pp. 1743–1751, Oct. 2017, doi: 10.1101/gr.221077.117.

[13] P. Geeleher, N. Cox, and R. S. Huang, “pRRophetic: An R Package for Prediction of Clinical Chemotherapeutic Response from Tumor Gene Expression Levels,” *PLOS ONE*, vol. 9, no. 9, p. e107468, Sep. 2014, doi: 10.1371/journal.pone.0107468.

[14] R. F. Gruener *et al.*, “Facilitating Drug Discovery in Breast Cancer by Virtually Screening Patients Using In Vitro Drug Response Modeling,” *Cancers*, vol. 13, no. 4, Art. no. 4, Jan. 2021, doi: 10.3390/cancers13040885.

[15] “https://ocg.cancer.gov/programs/ctd2/data-portal,” *Office of Cancer Genomics*. https://ocg.cancer.gov/programs/ctd2/data-portal (accessed Mar. 09, 2021).

[16] “Broad Institute Cancer Cell Line Encyclopedia (CCLE).” https://portals.broadinstitute.org/ccle/data (accessed Mar. 09, 2021).

[17] “Drug Download Page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer.” https://www.cancerrxgene.org/downloads/bulk_download (accessed Mar. 09, 2021).

[18] Y.-L. Lin *et al.*, “KRAS Mutation Is a Predictor of Oxaliplatin Sensitivity in Colon Cancer Cells,” *PLoS ONE*, vol. 7, no. 11, p. e50701, Nov. 2012, doi: 10.1371/journal.pone.0050701.

[19] E. Cule, S. Moritz, and D. Frankowski, “ridge: Ridge Regression with Automatic Selection of the Penalty Parameter. R package version 2.9.” 2021. [Online]. Available: https://CRAN.R-project.org/package=ridge

[20] L.-K. Schätzle, A. H. Esfahani, and A. Schuppert, “Methodological challenges in translational drug response modeling in cancer: A systematic analysis with FORESEE,” *PLOS Computational Biology*, vol. 16, no. 4, p. e1007803, Apr. 2020, doi: 10.1371/journal.pcbi.1007803.

[21] M. Koti *et al.*, “Identification of the IGF1/PI3K/NF κB/ERK gene signalling networks associated with chemotherapy resistance and treatment response in high-grade serous epithelial ovarian cancer,” *BMC Cancer*, vol. 13, p. 549, Nov. 2013, doi: 10.1186/1471-2407-13-549.

[22] Cancer Genome Atlas Research Network *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.

[23] A. Colaprico *et al.*, "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Res*, vol. 44, no. 8, p. e71, May 2016, doi: 10.1093/nar/gkv1507.

CHAPTER 3: INTEGRATION OF COMPUTATIONAL PIPELINE TO STREAMLINE EFFICACIOUS DRUG NOMINATION AND BIOMARKER DISCOVERY IN GLIOBLASTOMA

Danielle Maeser[1], Robert F. Gruener[2], Robert Galvin[3], Tomoyuki Koga[4], Florina-Nicoleta Grigore[4], Yuta Suzuki[4], Frank B. Furnari[5], Clark Chen[4], R. Stephanie Huang[2]

1. Department of Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN, United States
2. Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, MN, United States
3. Department of Pediatrics, University of Minnesota, Minneapolis, MN, United States
4. Department of Neurosurgery, University of Minnesota, Minneapolis, MN, United States
5. Department of Medicine, University of California San Diego, La Jolla, CA, United States

CONTRIBUTIONS BY FIRST AUTHORS

Danielle Maeser: code and method contributor, manuscript writer/reviewer.

ABSTRACT

Glioblastoma multiforme (GBM) is the deadliest, most heterogeneous and common brain cancer in adults. Despite major advancements in neurosurgery as well as chemotherapy and radiotherapy techniques, overall prognosis has improved little over the decades. Not only is there an urgent need to identify efficacious therapeutics but there is also a great need to pair these therapeutics with biomarkers that can help tailor treatment to the right patient populations given the heterogeneous nature of the disease. We implemented machine learning and causal inference pipelines, developed for drug response prediction and drug-biomarker prediction. Specifically, we built patient drug response models by integrating patient tumor transcriptome data with high-throughput cell line drug screening data as well as Bayesian networks to infer relationships between patient gene expression and drug response. Through these discovery pipelines, we identified multiple agents of interest for GBM to be effective across five independent patient cohorts and in a mouse avatar model (spanning nearly 1,000 GBM samples); among them, a number of MEK inhibitors (MEKis). We also predicted phosphoglycerate dehydrogenase enzyme (*PHGDH*) gene expression levels to be causally associated with MEKi efficacy, where knockdown of this gene increased tumor sensitivity to MEKi and overexpression led to MEKi resistance. Overall, our work demonstrated the power of integrating computational approaches into a drug development pipeline. In doing so, we quickly nominated a number of drugs with varying known mechanisms of action that can efficaciously target GBM. By simultaneously identifying biomarkers with these candidate drugs, we also improve the efficacy of drug discovery by providing tools to select the right patient populations for subsequent evaluation.

INTRODUCTION

Glioblastoma multiforme (GBM) is one of the deadliest diseases with only approximately 2% of GBM patients surviving three years or more.[1,2] Following radiation therapy and surgery, standard of care for GBM consists of chemotherapy (carmustine) with temozolomide (TMZ) or transcranial magnetic stimulation.[3] Unfortunately, a large portion of GBM patients do not respond to these treatment strategies and develop resistance to them very quickly. There is an urgent need to develop new therapeutic options for GBM patients. Yet, the heterogeneous nature of this disease and the requirement of passing through the blood-brain barrier for most of systematic treatment, have significantly limited the progress on development of efficacious drug treatment for GBM. Furthermore, traditional drug development takes on average about 12 years and 1 billion dollars to bring a new drug through regulatory approval. At this rate, hundreds and thousands of GBM patients will continue to suffer from lack of treatment. To this end, computational approaches to nominate drugs and identify biomarkers may offer a new path to significantly expedite the drug development process.

In this study, we apply a computational pipeline, *oncoPredict*[4], to enable drug sensitivity prediction in patient tumors.[5] When applied to several hard to treat cancer settings (e.g. triple negative breast cancer and castration resistant prostate cancer), *oncoPredict* has successfully nominated and pre-clinically validated drugs currently undergoing animal and clinical evaluation.[6,7] We are now applying this computational tool to the GBM setting in order to nominate efficacious therapy and help combat this highly heterogeneous disease. Specifically, the computational pipeline was applied to eight independent GBM and non-high grade glioma (non-HGG) patient datasets (spanning nearly 2,000 patients total). In addition, we generated drug sensitivity predictions in a novel GBM mouse avatar model to validate candidate drugs nominated across the GBM clinical cohorts. Furthermore, we employed causal inference framework to identify biomarkers for the candidate drugs of interest with the goal to facilitate selection of

the right patients to be treated with the right drug, which is critically important in heterogeneous disease like GBM.[8] Experimental validation of a selected biomarker was carried out in the mouse avatar model.

RESULTS

Identification of GBM Therapeutic Susceptibilities Utilizing Drug Response Prediction

In our drug discovery pipeline, we first applied *oncoPredict* to patient tumor transcriptome profiles to project patient likelihood of response to hundreds of medications. The patient tumor drug sensitivity projection was performed in five independent GBM patient datasets (a total of n=850), a mouse GBM avatar model (n=60); as well as two low-grade glioma (LGG) datasets from the Chinese Glioma Genome Atlas (CGGA-LGG, n=516)[9] and The Cancer Genome Atlas (TCGA-LGG, n=282).[10] and avatar neural progenitor cells (NPCs, n=6).[11] Details of datasets employed are listed in Supplemental Table 1. Our goal is to identify which drugs are repeatedly predicted to be more efficacious in GBM samples when compared to non-GBM samples across various datasets and modality. Hodges-Lehmann estimate (HLE) was performed with predicted drug sensitivity between GBM and non-GBM samples. Consensus across the majority (at least 3 of the 5) of the GBM clinical datasets and validation in the mouse GBM avatar data relative to NPCs was set to select candidate drugs of interest. After comparing predicted drug sensitivity scores between each of the five GBM patient datasets and CGGA-LGG dataset, with these filtering criteria, we identified 22 drugs of interest (red * marked in Figure 1A where HLE is also indicated in 1B). The difference in these predicted drug response scores, indicated by the HLE, for these 22 drug candidates between GBM and non-HGG groups are shown in Figure 1B. When a different set of non-GBM control, TCGA-LGG, was used, we identified 62 drugs (red * marked in Supplemental Figure 1A where HLE is also indicated in Supplemental 1B).

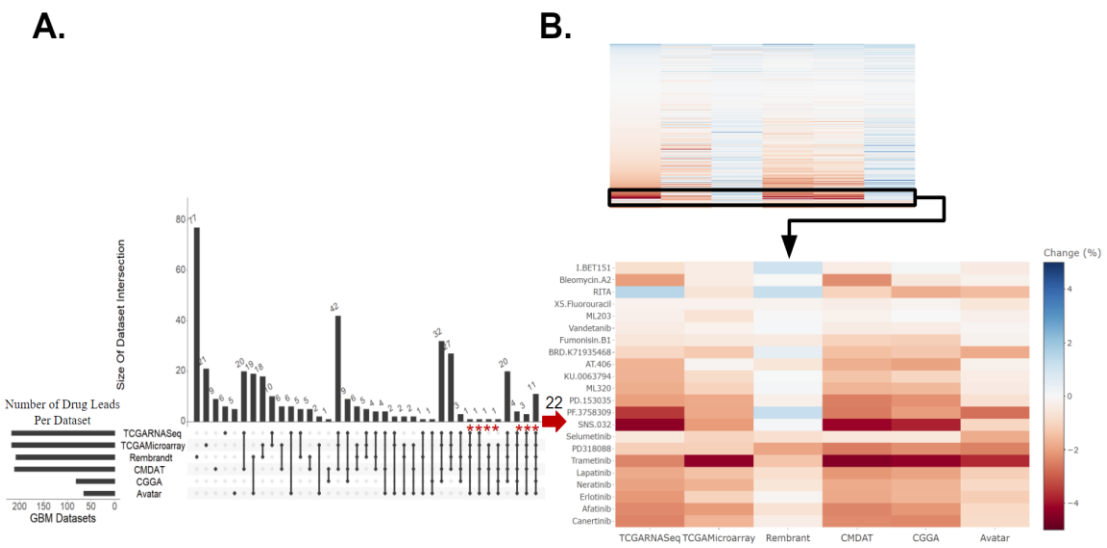


Figure 1: Drug candidates identified for GBM relative to non-HGG (CGGA-LGG and NPC). A. Upset plot displaying intersections of drugs predicted to be efficacious for GBM relative to non-high grade glioma (non-HGG) across six GBM datasets. These drugs had a Hodges-Lehmann estimate (HLE) within the top 50% of drugs with a FDR corrected p-value of <= 0.05. The red asterisk indicates 22 drugs that were identified as efficacious in more than half of the clinical datasets against CGGA-LGG and validated in the avatar dataset. **B.** The top panel heatmap displays the standardized HLE for all drugs identified to have a significantly different drug response across GBM and non-HGG samples. The bottom panel heatmap displays the HLE specifically for the 22 drug leads. Drugs predicted to be more efficacious for GBM are darker, in red. Those more efficacious for non-HGG are lighter, in blue.

Of the drug candidates predicted to be efficacious in GBM relative to non-HGG, 6 drugs overlapped across both CGGA-LGG (Figure 1) and TCGA-LGG (Supplemental Figure 1) non-HGG patient controls. These 6 drugs included mitogen-activated protein kinase inhibitors (MEKis) PD318088 and trametinib, as well as BRD.K71935468 (inducer of reactive oxygen species), Fumonisin.B1 (inhibitor of ceramide synthase), ML203 (activator of muscle pyruvate kinase), and RITA (inhibitor of p53-MDM2 interaction). Overall, the drugs identified across the LGG controls fell under a variety of classes including MEKis,

EGFR inhibitors (EGFRis), and VEGFR inhibitors (VEGFRis). Other targets include STAT3, BRAF, CDKs, etc. (Supplemental Data 2). Many of these drug leads have already been identified as potentially efficacious therapeutics in preclinical studies, supporting the validity of our computational projection. A few of these candidates have been tested in patients, yet none of them have been tested in patient populations guided by biomarker screening. As seen in Figure 1 and Supplemental Figure 1, MEKis were repeatedly identified as drug leads across independent patient datasets. For example, higher predicted sensitivity towards trametinib was observed across GBM samples relative to non-HGG (regardless the control datasets: CGGA-LGG, TCGA-LGG, or NPCs) where the heatmaps display similar effect size and directionality measured by HLE with an FDR corrected p-value <0.001. Therefore, we went forward with experimentally testing trametinib and additional MEKi selumetinib and PD318088 in the GBM avatar models. It is worth noting that while our drug discovery pipeline focused on candidate drugs predicted across the majority of clinical datasets with computational validation in avatar data, several other drugs were predicted to show higher sensitivity in all GBM clinical datasets without avatar validation. The independent nature of these patient datasets made the discovery interesting as well and these drugs information can be found in Supplemental Data 3.

The Efficacy of MEKis was Validated in GBM Avatar Model

Fundamentally, we found a number of MEKis showing higher predicted sensitivity in GBM across a number of independent datasets. We selected trametinib, selumetinib, and PD318088 for experimental testing in our GBM avatar model along standard of care agents: temozolomide (TMZ) and carmustine. After exposing to increasing concentrations of TMZ or carmustine, both standard of care agents produced a lower area under the dose response curve (AUC) for GBM cells relative to NPCs ($p<0.05$, Supplemental Figure 2), providing validity of our avatar system as an experimental model. Lower AUC values were also observed upon treatment of GBM cells with all three MEKis evaluated. In Figure 2, we plotted both

imputed and experimentally measured AUC for MEKis across avatar samples. All three MEKis performed as predicted across these samples, where GBM samples were significantly (FDR p-value<0.001) more sensitive to each MEKi tested, producing a lower AUC value relative to the NPC control samples.

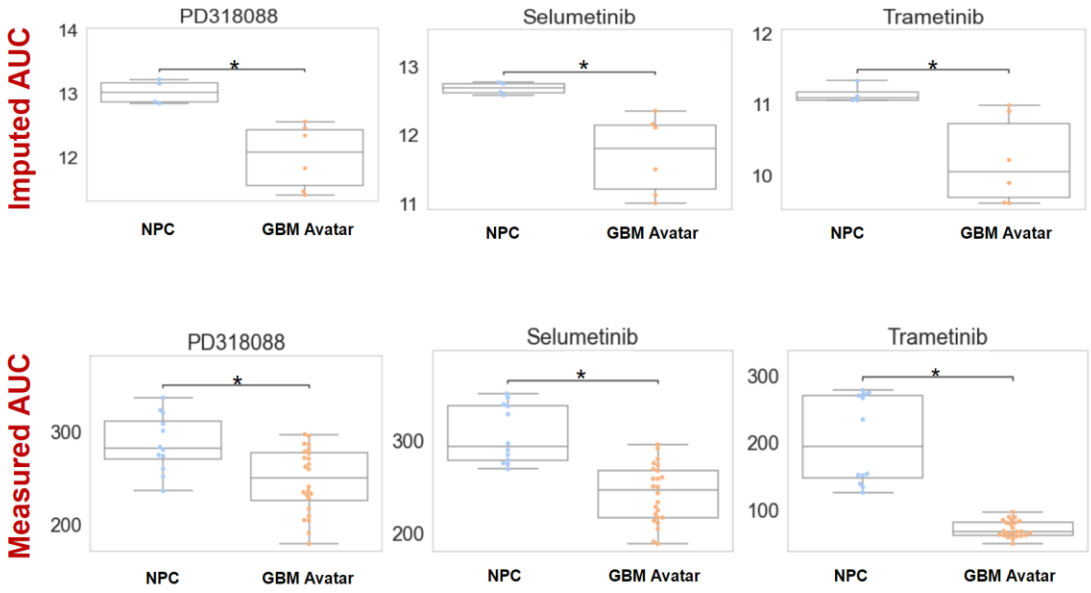


Figure 2: Imputed vs. measured drug response of MEK inhibitors (MEKis) across Glioblastoma (GBM) avatar and control samples. A. Predicted drug sensitivity scores for three MEKi (PD318088, selumetinib, trametinib) obtained from *oncoPredict*. AUC (area under the dose response curve) was predicted for neural progenitor cells (NPCs) and GBM samples for three MEKi's. **B.** Measured drug response by exposing GBM avatar or NPC cells to each MEKi separately with increasing concentrations. AUC was calculated using area under the dose-response curve after CellTiter experiments. Each sample was tested 6 times. For both imputed and measured scenarios, AUC across GBM and NPC samples were compared using a Wilcoxon sum rank test. The asterisk indicates statistical significance. For imputed drug response, the p-value was 0.006, 0.01, and 0.02 for PD318088, selumetinib, and trametinib respectively. For measured drug response, p<0.001.

Application of Causal Inference to Identify Biomarkers Indicative of MEKi Response

We employed a computational pipeline intended for causal inference in large omic data analysis for biomarker discovery for our drug of interest. This enabled us to predict drug-gene relationships for the MEKi drug candidates identified. Our pipeline contains two steps: 1. Spearman correlation coefficient (SCC) analysis, a univariate approach with a goal to filter for the most informative genes and 2. Bayesian-network learning, a multivariate approach consisting of the min-max parents-child (MMPC) and hybrid-parent and child (hybrid-PC) algorithms[12] to infer and visualize causal biomarkers. As shown in Figure 3A, through SCC analysis between TCGA-GBM RNA-Sequencing (RNASeq) data and predicted drug response for each MEKi, independently, we identified genes whose expression correlated with drug sensitivity (FDR $p<0.05$ with moderately-strongly $|r|>=0.60$). This substantially reduced the dimensionality of the transcriptomic data from tens of thousands to hundreds or less, and enabled reliable application of the Bayesian algorithms. Specifically, the data's dimensionality was reduced to 239, 72 and 54 informative genes for trametinib, PD318088 and selumetinib, respectively. Using each of these gene sets, the MMPC algorithm predicted parent and child (PC) nodes representative of genes that either directly influence or are directly impacted by drug response, where 8-12 PCs were identified for each drug. The MMPC algorithm also computed a test statistic, indicating the magnitude of the partial correlation for each PC, where a larger correlation indicates a stronger causal relationship. The hybrid-PC algorithm was also directly applied to the filtered TCGA-GBM RNASeq data to predict 1-2 parental nodes for each drug. For both trametinib and PD318088, phosphoglycerate dehydrogenase enzyme (*PHGDH*) was predicted to be a PC by the MMPC algorithm and a parental node by the hybrid-PC algorithm. In addition, SH2B adaptor protein algorithm was also predicted to be a PC for selumetinib and PD318088 as well as a parental node for selumetinib, directly influencing MEKi response (Supplemental Figure 4).

PHGDH Expression Levels Help Inform MEKi Response

Our computational pipeline identified *PHGDH* as a parental node for two MEKis with the largest test statistic from conditional independence testing (Supplemental Figure 4). The significant p-value and the large test statistic indicates a significantly strong causal relationship between *PHGDH* expression levels and MEKi response relative to the other PC nodes. In addition, in the univariate analysis, we observed significant and positive correlation between *PHGDH* gene expression and predicted or measured trametinib response across our six GBM datasets (5 patient cohorts and avatar) (Supplemental Figure 5). Taken together, we hypothesized that *PHGDH* knockdown increases cellular sensitivity to trametinib; and overexpression leads to trametinib resistance. To test this hypothesis, experimental testing was carried out by manipulating the *PHGDH* expression levels in a collection of GBM avatar samples (both proneural and mesenchymal subtype cells). We had successful knockdown and overexpression of *PHGDH* in all GBM avatar samples (Figure 3B). The GBM cells with/without *PHGDH* manipulation were then treated with increasing concentrations of trametinib and the surviving percentage was measured and compared to controls. Two control samples with unmanipulated *PHGDH* expression were used, including a parental condition. To assess knockdown, control samples also included lentivirus plasmid vector pLKO.1-puro control vector. The average of three short hairpin RNA (shRNAs) constructs with *PHGDH* knockdown demonstrated significantly increased sensitivity to trametinib (p-value < 0.0001) relative to both control samples (Figure 3C). To assess overexpression, control samples also included plasmid pLV-EF1. The *PHGDH* overexpression led to significantly increased resistance to trametinib (p-value < 0.0001) relative to both control samples as well (Figure 3D).

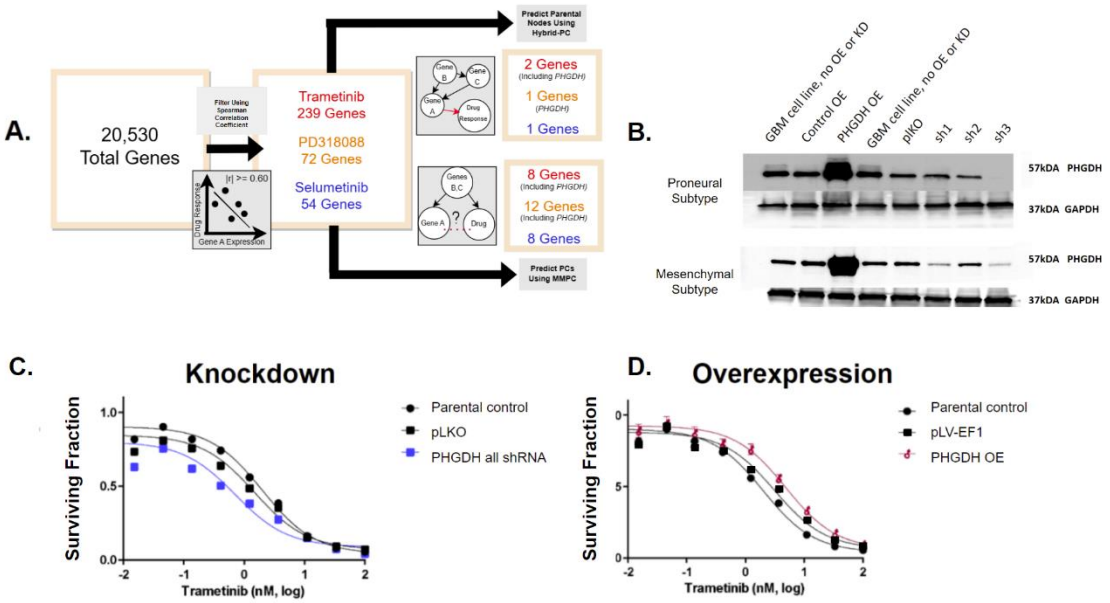


Figure 3: PHDGH was identified and validated to be a biomarker that affects MEKi treatment effect in GBM. A. Biomarker discovery pipelines and findings for trametinib (red), PD318088 (gold), and selumetinib (blue). The pipeline consisted of filtering this data using the Spearman correlation coefficient (SCC) to select genes whose significant and absolute SCC was equal to or exceeded 0.60 (FDR $p<0.05$). Then applying two Bayesian algorithms (MMPC and Hybrid-PC) to infer correlation and causality between gene expression and drug response using this filtered data. From this pipeline, *PHGDH* expression was predicted to directly influence both trametinib and PD318088 responses. **B.** Western blot of *PHGDH* and house-keeping control *GAPDH* in two GBM avatar subtype samples (the proneural and mesenchymal subtype, each carrying unique genomic modifications). The experiment confirms knockdown and overexpression of *PHGDH* were successful in our GBM avatar (across mesenchymal and proneural GBM subtypes). **C and D.** CellTiter Glo experiments after trametinib exposure for 72 hours in GBM avatar cells (of both proneural and mesenchymal subtypes) following *PHGDH* knockdown (C) and *PHGDH* overexpression (D). Each experimental condition was evaluated in triplicates, and the plots (C and D) represent average across all experiments. For knockdown experiments, 3 sets of shRNA were

employed and their average effects were plotted as shown in the blue curve in panel C. A 2-way ANOVA test was employed to test the statistical differences among conditions. There are statistical differences in the cell survival after trametinib treatment between *PHGDH* knockdown and both controls ($p<0.001$ for both parental and pLKO). There are also statistical differences in the cell survival after trametinib treatment between *PHGDH* overexpression and both controls ($p<0.001$ for both parental and pLV-EF1). A significant increase in drug resistance occurred following overexpression as well as a significant increase in drug efficacy following knockdown.

DISCUSSION

Utilizing a computational drug sensitivity prediction tool independently across five GBM patient datasets (approximately 1,000 samples), a GBM mouse avatar model and 2 control non-HGG patient datasets, we identified a collection of drugs that were predicted to have a higher sensitivity in GBM relative to non-HGG (the complete list is provided in Supplemental Data 2). While many of our nominated drugs (close to 20) have been/are being evaluated in clinical studies against GBM, nearly half (approximately 40) of our discoveries have shown efficacy in preclinical experiments reported previously. All of these justify the validity of our computational approach. Furthermore, given the heterogeneity among these independent clinical studies, the consistency in our discoveries further justify their importance and present numerous opportunities for follow up studies. These drugs achieve tumor growth inhibition through a variety of mechanisms of actions, namely VEGFRis, EGFRis, and MEKis, which have a rich history of study in the context of GBM[13-23]. Other targets identified that frequently show up in clinical studies include mammalian target of rapamycin (MTOR) and cyclin-dependent kinases (CDKs). In addition, drugs targeting STAT3, reactive oxygen species (ROS), PAR1, PAK4, apoptosis proteins, ubiquitin-specific protease 14 (USP14) are also among our top predicted efficacious candidates in treating GBM.

All of them present potential new opportunities. The one unique outcome of our approach is that we identified agents that are known to act through various mechanisms of actions yet all effective in the setting of GBM simultaneously.

To dive in deeper for the candidate drugs of interest, we found a number of VEGFRis through our pipeline, namely avastin, axitinib, and lenvatinib. VEGFRis have a long track record in GBM applications and have been identified as providing promising therapeutic opportunity if improvement of biomarkers aid in advancing the clinical efficacy of this approach.[13] Avastin is an example of such an inhibitor which has been investigated in preclinical and clinical studies for its potential to delay GBM tumor growth. It is currently approved in adult patients whose cancer has progressed after prior treatment (recurrent or rGBM).[14] Other VEGFRis include axitinib and lenvatinib, which have shown success in pre-clinical studies.[15,16] A prospective phase I and II study is currently in place to assess the effectiveness of lenvatinib in combination with pembrolizumab for GBM (NCT05973903).

EGFRis were also repeatedly identified as efficacious for treating GBM across our pipeline. In fact, EGFRis have had varying degrees of success against GBM. For example, afatinib, an EGFR inhibitor that has shown effect against GBM in preclinical studies.[17,18] However, when tested in GBM patients (NCT00727506) either alone or in combination with TMZ, only very limited efficacy was observed in unselected patients[19]. Importantly, follow-up study showed that afatinib is only effective in selected patients harboring *EGFRvIII* mutation.[20] Given the heterogeneity commonly recognized for a disease like GBM, this example highlights the need for biomarker screening as a means to direct patients to their appropriate treatment to help clinical trial design in drug development and eventually in the combat against GBM.

In our study, multiple MEKis were repeatedly projected to show higher efficacy in GBM samples regardless which control datasets were used. We experimentally validated all three MEKis (trametinib,

selumetinib, and PD318088) for their preferential activity in mouse avatar samples when compared to NPC. Surveying the literature, we found that preclinical evidence exists to support the efficacy of trametinib and selumetinib in treating glioma. For example, trametinib has been reported to exert a strong antiproliferative effect on multiple established GBM cell lines, and its inhibitory effect on cell growth was observed even after standard of care treatment.[21] Selumetinib was reported to stabilize disease progression in glioma patients during a phase II clinical trial.[22] PD318088 on the other hand, is a more novel MEKi, which has not been clinically tested against GBM.[23]

Causal inferences have been employed to study disease risks, and they are yet to be widely used in identification of drug biomarkers. In this study, we integrated multiple approaches for biomarker discovery for our drug of interest. We first apply a commonly used univariate analysis between gene expression and drug sensitivity to narrow down the list of informative genes, then two Bayesian based causal inference tools were employed to identify causal genes. Through this pipeline, we nominated *PHGDH* as our top MEKi biomarker and was able to experimentally validate its role in MEKi sensitivity. Specifically, we found *PHGDH* expression levels as having a significant and positive correlation and causal relationship with trametinib response. Interestingly, *PHGDH* has previously been proposed as a therapeutic target for melanoma in overcoming resistance to MEKis PD0325901 and trametinib,[24,25] where suppression of *PHGDH* led to sensitivity in MEKi resistant melanoma cells. These findings mirror the directionality predicted by our biomarker discovery pipeline, supporting the same biomarker for MEKi may be utilized in other cancer settings. Indeed, we envision the same biomarker discovery pipeline can be applied to any other drugs and/or phenotypes in the future. In addition, our biomarker discovery pipeline identified several other genes which may directly influence MEKi response as well. For example, *SH2B* was predicted to be a PC for both selumetinib and PD318088 as well as a parental node for selumetinib, directly influencing MEKi response. It is highly expressed in GBM, promoting progression through activating STAT3 signaling.[26] It is known to play a critical role in promoting GBM

progression and has previously been proposed as a new therapeutic target. Therefore, our discovery also warranted studying of this gene as a potential biomarker for MEKis.

Overall, the findings from this study support the approach to integrate computational and experimental models as well as multiple independent cross platform and cross species datasets, with the goal to speed up drug discovery and development for GBM. Our research supports ongoing efforts to improve selectivity of treatments applied to patients with GBM through applying discovery pipelines to GBM expression data, pairing drug candidates with biomarkers indicative of drug response.

METHODS

GBM Clinical Data Tested in Computational Modeling

In this study, we applied computational drug and biomarker discovery pipelines to nine publicly available primary LGG and adult GBM patient datasets (Supplemental Table 1) to identify compounds of interest for specific patient populations defined by the presence of biomarkers. In total, our drug imputation model was applied to almost 2,000 LGG and GBM samples, imputing almost 500 drug response scores for each sample. The LGG datasets consisted of bulk RNAseq from TCGA (n=516) and the CGGA (n=282). TCGA datasets were downloaded using the TCGAbiolinks R package,[10] and CGGA datasets were downloaded from the CGGA webpage.[9] The GBM datasets include microarray data (n=332) and bulk RNAseq (n=165) from TCGA, microarray data from CGGA (n=102), microarray data from Rembrandt (n=189, GSE108476),[27] and a personally combined dataset of published literature (referred to as CMDAT) also using affymatrix (n=62). All datasets were normalized and log transformed to yield Gaussian expression distributions.

GBM Mouse Avatar Model Utilized for Experimental Validation of Drug Candidates and Inferred Drug-Biomarker Relationships

The GBM avatar model[11] was created by introducing different genetic driver mutations using CRISPR-Cas9 into human induced pluripotent stem cells. This was followed by differentiation of GBM-associated mutations containing NPC and animal orthotopic engraftment to develop human adult GBM models. The NPC samples represent a pre-HGG state. Tumor cells were obtained and cultured to produce spheres, and this process was repeated twice to produce primary and secondary sets of tumors and spheres. Secondary tumors were obtained following engraftment of primary spheres, and secondary spheres were obtained from secondary tumors. This process is reflected in Supplemental Figure 3. This resulted in 66 total samples including NPCs, engrafted tumor avatar samples, and spheres with technical replicas. Specifically, 6/66 samples were NPCs, and the remaining 60 samples consisted of 28 mesenchymal subtype samples and 32 proneural subtype samples. The samples representative of the mesenchymal subtype were characterized by *PTEN* and *NF1* deletion, and the proneural subtype was characterized by *TP53* deletion and *PDGFRA* mutation. The bulk RNAseq expression profiles for these NPCs, tumors, and spheres were obtained, and batch effect correction was performed where appropriate using remove unwanted variation (RUV) normalization.[28]

Overview of Drug Discovery Pipeline

To identify drug leads or drugs predicted to elicit a greater response in GBM samples relative to non-HGG samples, we applied a drug discovery pipeline. This pipeline involved comparing drug response scores across different sequencing platforms (microarray and RNAseq) as well as patient and avatar data. Cancer cell line screening data has been used to train machine learning models, aiming to translate in vitro drug response to in vivo tumor response predictions and generate novel drug discovery hypotheses. To date, the Broad Institute's Cancer Therapeutics Response Portal[29] (CTRP) is one of the largest

publicly available drug screening efforts, providing drug screening for nearly 1,000 cell lines and 500 compounds. The most updated dataset from CTRP is CTRP version 2 (CTRP2), representing a variety of cancers and molecular targets. CTRP's cell line transcriptome data is provided through the Broad Institute's Cancer Cell Line Encyclopedia database[30] (CCLE). For our purposes, the names across the cancer cell lines from CTRP2 and CCLE were harmonized to Cellosaurus accession numbers, indicated by the 'CVCL' prefix.[31] Drugs screened across <40% of all the cancer cell lines were also removed, helping to ensure robust predictions. This resulted in 887 cell lines and 493 drugs. Our R package *oncoPredict*'s function *calcPhenotype()* estimates a gene's weight in determining a cell's drug sensitivity through applying linear regression with a ridge penalty. This allowed a predictive drug sensitivity score, in the form of AUC to be obtained for each sample running through *oncoPredict*. To compare cross platform drug response data, we accounted for technical variation and platform effects by transcriptome integration using Rank-In[32] prior to running *calcPhenotype()*.

The non-parametric Wilcoxon rank sum (WRS) tests were selected for the statistical comparison between GBM and non-HGG data. WRS tests were performed, as the assumption of normality did not hold for the independent samples T-test in which case the WRS holds power advantages. To reduce the probability of making one or more false discoveries or type 1 errors, which are common in multiple hypothesis tests, the statistical Benjamini and Hochberg FDR controlling procedure was implemented to adjust p-values. Drugs were filtered for statistically significant p-values, which confirmed that differences in predicted drug response scores existed between GBM and non-HGG data, and ranked by their effect size. The HLE was measured for the magnitude of significance or effect size, known as the location shift.[33] It established directionality to determine whether a given drug was recommended for GBM over non-HGG, offering a robust measure of effect size against outliers and distribution assumptions. Once drugs were ranked, the top percentage of significant drugs with the largest location shift were selected. The top 50th percentile of compounds with the largest HLE was selected as a conservative choice, as it was large enough to help

prevent efficacious drugs from slipping through the pipeline while restrictive enough to help capture compounds with the greatest magnitude in response. Due to the high variability of drugs recommended for each GBM patient dataset, drugs recommended across the majority of clinical datasets were selected for comparison with those recommended across the avatar data. Drug imputation was also performed for the avatar dataset, which served as computational verification. Drugs selected in the top 50% of HLE across both the majority of clinical data as well as avatar data were brought forward as drug candidates for experimental validation.

Overview of Biomarker Discovery Pipeline

Our biomarker discovery pipeline was employed to predict biomarkers for drug leads in an effort to uncover vulnerable patient populations. This pipeline utilized the SCC as a univariate approach to biomarker discovery with Bayesian-network learning (MMPC and hybrid-PC)[12] as a multivariate approach. Univariate approaches like the SCC are defined as those that evaluate the informativeness of each gene individually in isolation from the other genes according to a criterion. MMPC and hybrid-PC algorithms depict a more biologically accurate representation of gene interactions by testing for conditional independence. They implemented the Fisher Independence Test to measure the partial correlation coefficient between predicted drug response scores and the expression of a gene under scrutiny while conditioning on a set of the other predictor genes known as the 'conditioning set.' The TCGA-GBM RNASeq dataset was selected as the primary GBM dataset used in the biomarker discovery pipeline because relative to microarray data, RNAseq has higher specificity. It can more accurately detect differential expression as well as rare or low expressed genes, and it outperforms microarray in determining transcriptomic characteristics of cancer.[34,35] The first step in this pipeline was to filter the TCGA-GBM RNASeq patient gene expression dataset for genes that were significantly (FDR p-value <0.05) and moderately to strongly correlated (|SCC| >= 0.60) with the predicted drug response scores

under scrutiny. The SCC was computed by setting '*cc=TRUE*' in *oncoPredict*'s *calcPhenotype()* function, adjusted for Spearman correlation. .

After filtering by the SCC, we applied the MMPC algorithm to predict PCs. Filtering the gene expression data using SCC reduced the dimensionality of the MMPC algorithm's input data. This achieved reproducibility of its outputs since Bayesian network learning algorithms like MMPC can suffer from variable order dependence, which is a problem with high dimensional data like gene expression.[36] The partial correlation computed for a given gene is equal to the correlation between two sets of residuals when linear regression is applied: the first is between the drug under scrutiny and the conditioning set, and the second is between the gene under scrutiny and the conditioning set. Hence this test regresses both the target and the variable under scrutiny on the conditioning set. Drugs which are predicted to be independent have a partial correlation of zero, larger values indicate greater dependency, and smaller values indicate limited dependency. MMPC outputs a test statistic, taking into account the directionality of the partial correlation, measuring the strength and the directionality of the predicted gene-drug associations identified. As the independence tests are performed, they progressively exclude irrelevant genes (genes that are independent from the predicted drug response scores). The end result of this algorithm are genes that have survived these elimination stages known as PCs. PCs will have significantly large test statistics where the sign of the test statistic (whether it is positive or negative) indicates directionality. This sign allowed us to determine whether a gene under scrutiny was associated with drug sensitivity through up or down regulation. The MMPC algorithm does not distinguish parental and child nodes from the PCs identified, so we also applied the Hybrid PC algorithm to the filtered gene expression data to infer the structure of the Bayesian network between gene expression and predicted drug response in order to determine which PCs were parental nodes. By utilizing these univariate and multivariate approaches, we were able to predict correlation, causality, and which genes lead to drug sensitivity and how (whether through up or down regulation).

Experimental Testing of Drug Candidates GBM Mouse Avatar Model

The avatar models used to computationally validate drug leads were also used for experimental evaluation. Drug leads (PD318088, trametinib, selumetinib) and standard of care agents (TMZ, carmustine) were tested in six avatar samples. Two samples were NPCs, two were primary spheres, and two were secondary spheres. Mesenchymal and proneural avatar samples were represented equally amongst these sample types. To measure efficacy of drug leads, relative ATP was measured on day three of treatment. The concentrations used for testing were selected by referencing dose-response curves obtained from several published cell line drug screening datasets for GBM cell lines, reported in our Simplicity application. Doses selected are provided in Supplemental Data 1. The drug concentrations were measured by taking the natural log of the micromolar concentration. For each drug tested, relative ATP was captured six times per sample across nine different doses. Drug response or measured AUC was obtained by generating dose-response curves. Then measuring AUC using the trapezoidal rule to compute the area underneath the relative ATP curve through the R function *trapz()* from the package *pracma*.

Experimental Testing of Inferred Drug-Biomarker Relationships in GBM Mouse Avatar Model

Through applying our biomarker discovery pipeline to TCGA-GBM RNAseq data, we hypothesized that *PHGDH* knockdown contributes to trametinib sensitivity and overexpression contributes to resistance. To experimentally test this relationship in GBM mouse avatar samples, a secondary mesenchymal and proneural sphere with *PHGDH* knockdown and overexpression were treated with trametinib and the surviving percentage was measured and compared to negative controls (Figure 3B). Negative controls consisted of parental samples and empty vector transduced samples. To assess knockdown, control samples also included lentivirus plasmid vector pLKO.1-puro control vector.

CODE AND DATA AVAILABILITY

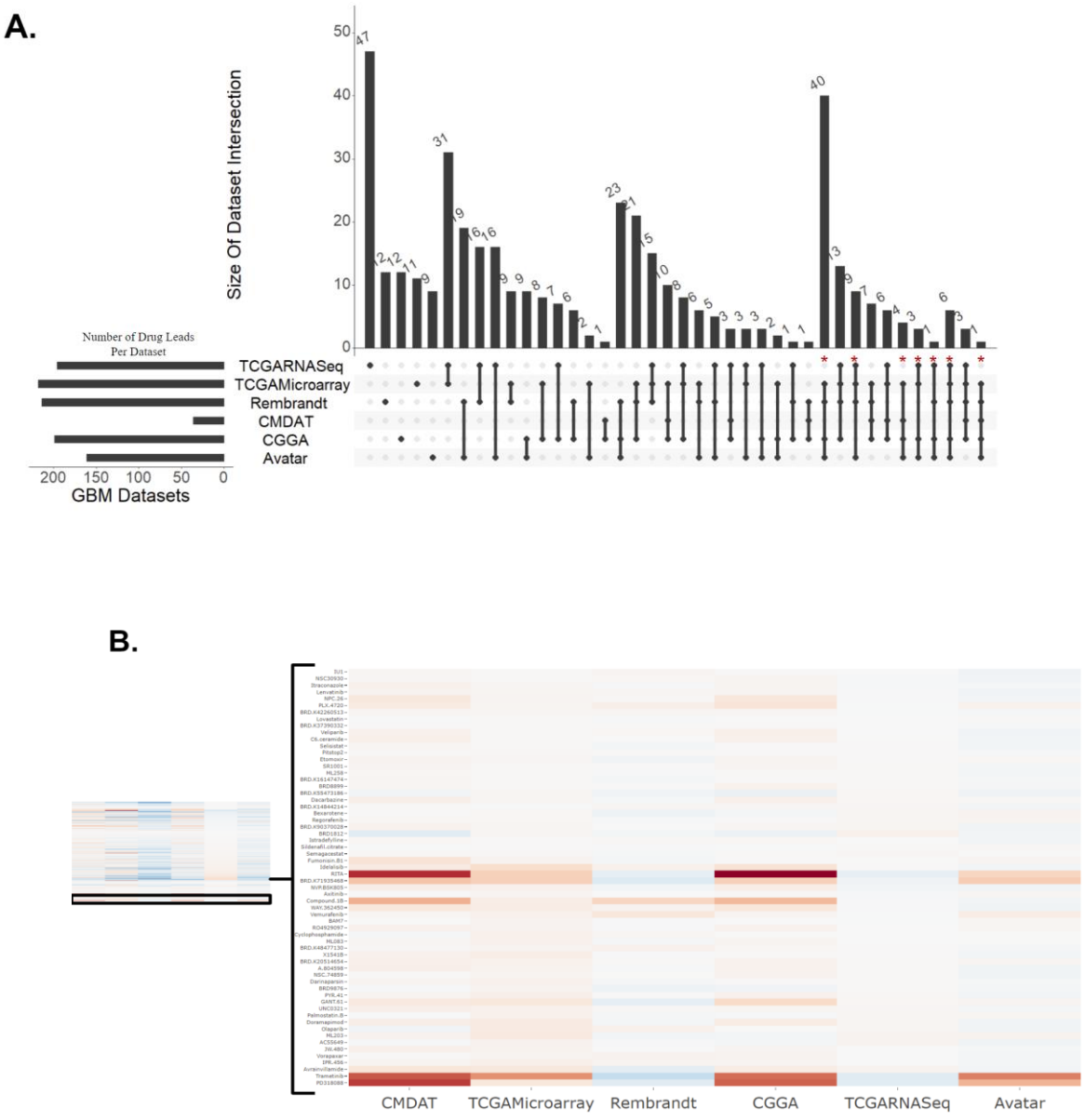
---

https://osf.io/ar9zg/

SUPPLEMENTAL

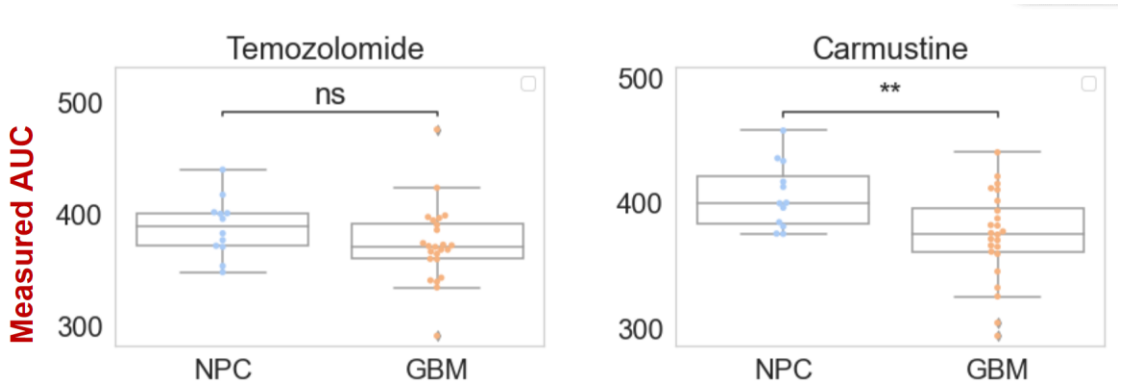
Supplemental Table 1: Overview of the clinical and avatar datasets. Bulk gene expression derived from GBM patient and non-high grade glioma (non-HGG) data are used to computationally predict efficacious compounds against GBM. Two non-HGG patient datasets, representing low-grade glioma (LGG) from The Cancer Genome Atlas (TCGA) and Chinese Glioma Genome Atlas (CGGA), and five GBM patient datasets are used to nominate potential drugs of interest against GBM. The GBM patient datasets were obtained from TCGA, the Rembrandt study, a personal archive (CMDAT), and CGGA. The avatar dataset consists of GBM totaled 60 samples, technical replicates included, as well as 6 neural progenitor cells (NPCs).

Non-HGG Datasets	GBM Patient Datasets	GBM Avatar Datasets
TCGA LGG RNAseq (n=516)	TCGA Microarray (n=332)	Tumor spheres & tumor cells (n=60)
CGGA LGG RNAseq (n=282)	TCGA RNAseq (n=165)	
Avatar NPCs (n=6)	CGGA Microarray (n=102)	
	Rembrandt Microarray (n=189)	
	CMDAT Microarray (n=62)	

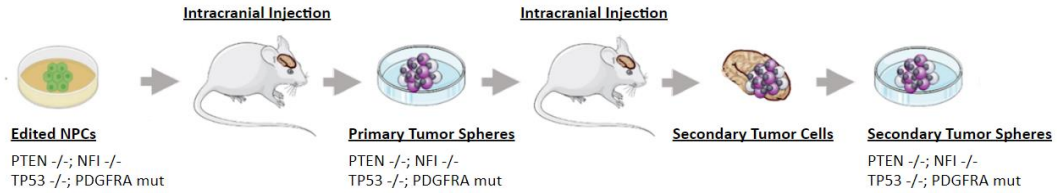


Supplemental Figure 1: Drug leads identified for GBM relative to non-HGG (TCGA-LGG and NPC). A. Upset plot displaying intersections of drugs predicted to be efficacious for GBM relative to non-high grade glioma (non-HGG) across six GBM datasets. These drugs had a Hodges-Lehmann estimate (HLE) within the top 50% of drugs with a FDR corrected p-value of <= 0.05. The red asterisk indicates 62 drugs that were identified across avatar data and more than half of the clinical datasets. **B.**

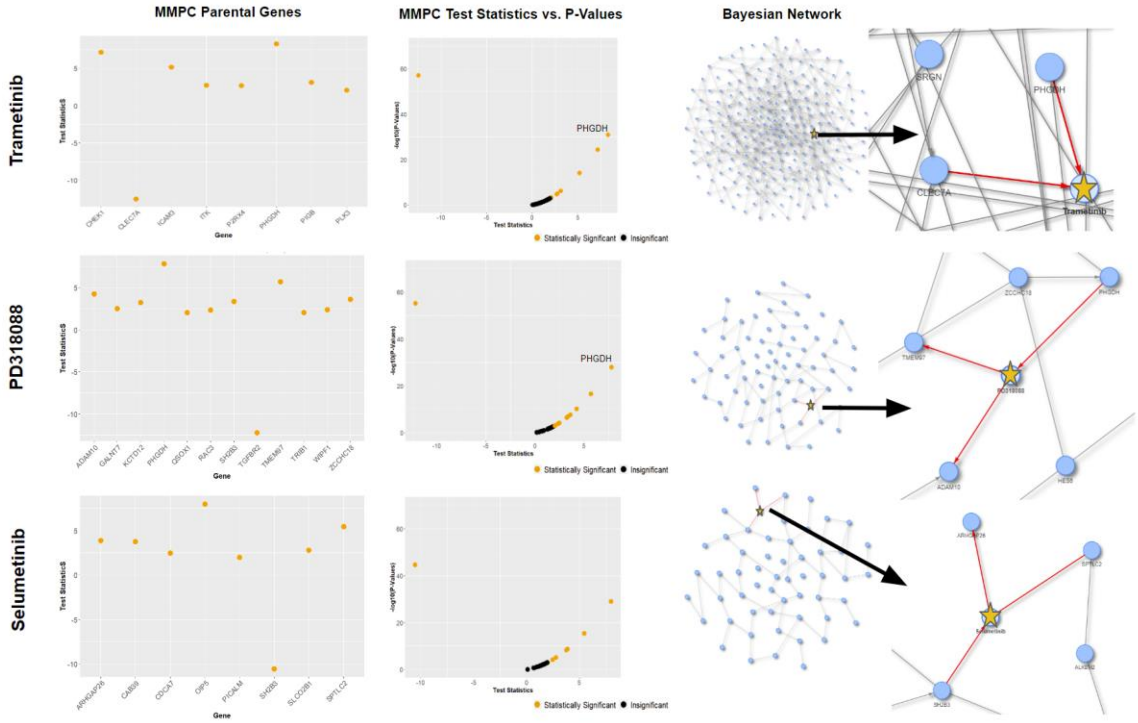
The smaller heatmap displays the standardized HLE for all significant drugs identified. The larger heatmap displays the HLE specifically for drug leads. Drugs predicted to be more efficacious for GBM are in red and include MEK inhibitor trametinib which elicited the greatest response, and those more efficacious for non-HGG are blue.



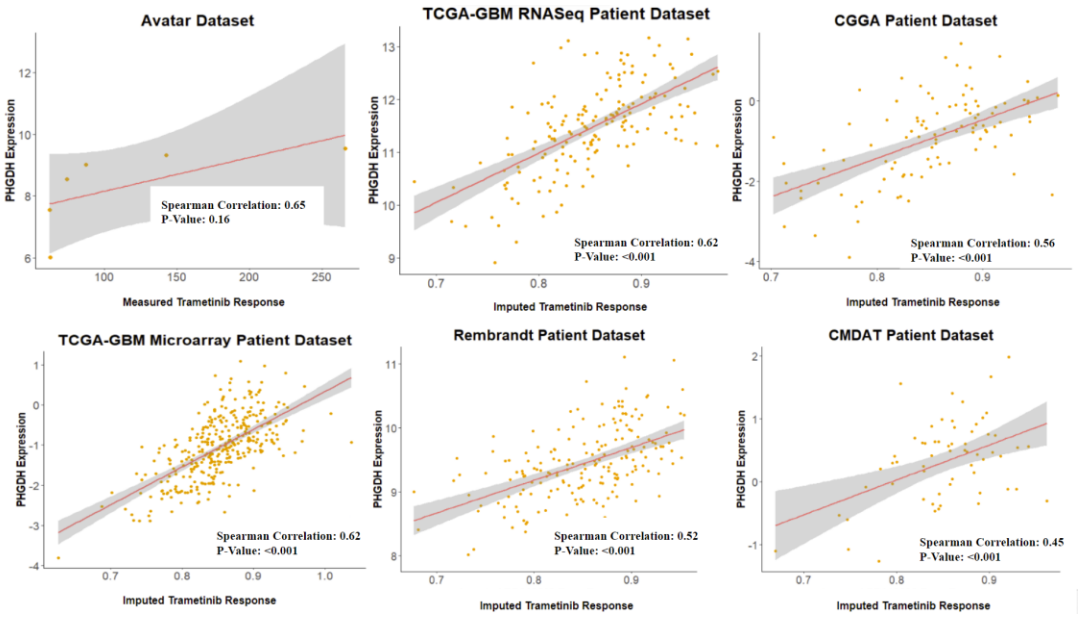
Supplemental Figure 2: Standard of care agent's measured drug response across GBM mouse avatar samples. These plots display the drug response captured from experimentally testing standard of care agents temozolomide and carmustine, at various drug concentrations, in avatar mouse models. AUC (area under the dose response curve) was obtained through measuring the relative ATP scores across nine different doses, six times per concentration. Then applying the trapezoidal method to calculate the area under the dose response curve. The drug concentrations were measured by taking the natural log of the micromolar concentration. GBM samples, including the neural progenitor cell (NPC) samples that evolved to specific subtypes, are color coated.



Supplemental Figure 3: Generation of the mouse avatar bulk RNAsequencing data. The mouse avatar models were created following a series of steps including 1) introduction of different genetic driver mutations common in two GBM (glioblastoma) subtypes into human induced pluripotent stem cells (iPSCs) using CRISPR-Cas9. 2) differentiation of iPSCs into neural progenitor cells (NPCs), which were orthotopically engrafted into mice 3) generation of tumor cells, which were cultured to produce spheres, and this process was repeated twice 4) generation of the bulk RNAsequencing gene expression profiles of these NPCs, tumors, and spheres.



Supplemental Figure 4: Multivariate method for MEK inhibitor (MEKi) biomarker discovery. MMPC parental genes were obtained through applying the Min-Max Parents Children (MMPC) algorithm to TCGA-GBM RNA sequencing data and imputed MEKi response. MMPC parental genes were indicated for each MEKi, where *PHGDH* was predicted to be a parental gene for both trametinib and PD31808 with the largest positive test statistic. The test statistic indicates increased MEKi sensitivity may result from *PHGDH* knockdown. The Bayesian network was obtained from applying the hybrid MMPC algorithm to the TCGA-GBM dataset, and *PHGDH* is confirmed to be a parental node to MEKi response.



Supplemental Figure 5. Linear plots displaying the relationship between *PHGDH* gene expression across the six GBM datasets and measured or imputed response to trametinib. The Spearman correlation coefficient and p-value between gene expression and drug response is provided. In this figure, 'Avatar' indicates the correlation between the avatar gene expression and the average area under the dose response curve measured from experimental testing.

Supplemental Data 1-3 are provided in the OSF link.

REFERENCES

1. Smoll, N. R., Schaller, K. & Gautschi, O. P. Long-term survival of patients with glioblastoma multiforme (GBM). *Journal of Clinical Neuroscience* 20, 670–675 (2013).
2. Walid, M. S. Prognostic Factors for Long-Term Survival after Glioblastoma. *TPJ* 12, 45–48 (2008).
3. Krieg, S. M. *et al.* Changing the clinical course of glioma patients by preoperative motor mapping with navigated transcranial magnetic brain stimulation. *BMC Cancer* 15, 231 (2015).
4. Maeser, D., Gruener, R. F. & Huang, R. S. oncoPredict: an R package for predicting *in vivo* or cancer patient drug response and biomarkers from cell line screening data. *Briefings in Bioinformatics* 22, bbab260 (2021).
5. Adam, G. *et al.* Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precis. Onc.* 4, 19 (2020).
6. Zhang, W. *et al.* Computational drug discovery for castration-resistant prostate cancers through in vitro drug response modeling. *Proc. Natl. Acad. Sci. U.S.A.* 120, e2218522120 (2023).
7. Gruener, R. F. *et al.* Facilitating Drug Discovery in Breast Cancer by Virtually Screening Patients Using In Vitro Drug Response Modeling. *Cancers* 13, 885 (2021).
8. Sareen, H. *et al.* Molecular Biomarkers in Glioblastoma: A Systematic Review and Meta-Analysis. *IJMS* 23, 8835 (2022).

9. Zhao, Z. *et al.* Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients. *Genomics, Proteomics & Bioinformatics* 19, 1–12 (2021).

10. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* 44, e71–e71 (2016).

11. Koga, T. *et al.* Longitudinal assessment of tumor development using cancer avatars derived from genetically engineered pluripotent stem cells. *Nat Commun* 11, 550 (2020).

12. Tsagris, M. & Tsamardinos, I. Feature selection with the R package MXM. *F1000Res* 7, 1505 (2019).

13. Robles Irizarry, L., Hambardzumyan, D., Nakano, I., Gladson, C. L. & Ahluwalia, M. S. Therapeutic targeting of VEGF in the treatment of glioblastoma. *Expert Opinion on Therapeutic Targets* 16, 973–984 (2012).

14. Fu, M. *et al.* Use of Bevacizumab in recurrent glioblastoma: a scoping review and evidence map. *BMC Cancer* 23, 544 (2023).

15. Saha, D. *et al.* Combinatorial Effects of VEGFR Kinase Inhibitor Axitinib and Oncolytic Virotherapy in Mouse and Human Glioblastoma Stem-Like Cell Models. *Clinical Cancer Research* 24, 3409–3422 (2018).

16. Li, J. *et al.* A multi-targeted tyrosine kinase inhibitor lenvatinib for the treatment of mice with advanced glioblastoma. *Molecular Medicine Reports* 16, 7105–7111 (2017).

17. Owen, S. *et al.* Genomic Analysis of Tumors from Patients with Glioblastoma with Long-Term Response to Afatinib. *OTT* Volume 15, 367–380 (2022).

18. Vengoji, R. *et al.* Afatinib and Temozolomide combination inhibits tumorigenesis by targeting EGFRvIII-cMet signaling in glioblastoma cells. *J Exp Clin Cancer Res* 38, 266 (2019).

19. Cruz Da Silva, E., Mercier, M.-C., Etienne-Selloum, N., Dontenwill, M. & Choulier, L. A Systematic Review of Glioblastoma-Targeted Therapies in Phases II, III, IV Clinical Trials. *Cancers* 13, 1795 (2021).

20. Owen, S. *et al.* Genomic Analysis of Tumors from Patients with Glioblastoma with Long-Term Response to Afatinib. *OTT* Volume 15, 367–380 (2022).

21. Selvasaravanan, K. D. *et al.* The limitations of targeting MEK signalling in Glioblastoma therapy. *Sci Rep* 10, 7401 (2020).

22. Fangusaro, J. *et al.* A phase II trial of selumetinib in children with recurrent optic pathway and hypothalamic low-grade glioma without NF1: a Pediatric Brain Tumor Consortium study. *Neuro-Oncology* 23, 1777–1788 (2021).

23. Cheng, Y. & Tian, H. Current Development Status of MEK Inhibitors. *Molecules* 22, 1551 (2017).

24. Doepner, M., Lee, I. Y. & Ridky, T. W. Drug Resistant Melanoma May Be Vulnerable to Inhibitors of Serine Synthesis. *Journal of Investigative Dermatology* 140, 2114–2116 (2020).

25. Nguyen, M. Q. *et al.* Targeting PHGDH Upregulation Reduces Glutathione Levels and Resensitizes Resistant NRAS-Mutant Melanoma to MAPK Kinase Inhibition. *Journal of Investigative Dermatology* 140, 2242-2252.e7 (2020).

26. Cai, S. *et al.* SH2B3, Transcribed by STAT1, Promotes Glioblastoma Progression Through Transducing IL-6/gp130 Signaling to Activate STAT3 Signaling. *Front. Cell Dev. Biol.* 9, 606527 (2021).

27. Gusev, Y. *et al.* The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci Data* 5, 180158 (2018).

28. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 32, 896–902 (2014).

29. Basu, A. *et al.* An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell* 154, 1151–1161 (2013).

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

31. Bairoch, A. The Cellosaurus, a Cell-Line Knowledge Resource. *J Biomol Tech* 29, 25–38 (2018).

32. Tang, K. *et al.* Rank-in: enabling integrative analysis across microarray and RNA-seq for cancer. *Nucleic Acids Research* 49, e99–e99 (2021).

33. Hodges, J. L. & Lehmann, E. L. Estimates of Location Based on Rank Tests. *Ann. Math. Statist.* 34, 598–611 (1963).

34. Zhang, W., Yu, Y., Hertwig, F. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol* 16, 133 (2015). https://doi.org/10.1186/s13059-015-0694-1

35. Corchete, L.A., Rojas, E.A., Alonso-López, D. *et al.* Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep* 10, 19737 (2020). https://doi.org/10.1038/s41598-020-76881-x

36. Hsu, W. H. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning. *Information Sciences* 163, 103–122 (2004).

CHAPTER 4: INFERRING THERAPEUTIC VULNERABILITY WITHIN TUMORS THROUGH INTEGRATION OF PAN-CANCER CELL LINE AND SINGLE-CELL TRANSCRIPTOMIC PROFILES

Weijie Zhang[1,2,†], Danielle Maeser[1,2,†], Dr. Adam Lee[2], Yingbo Huang[2], Dr. Robert F. Gruener[2], Israa Gamal Aly Abdelbar[2,3], Dr. Sampreeti Jena[2], Dr. Anand G. Patel[4,5], Dr. R. Stephanie Huang[1,2,*]

*Corresponding author

[†]Equal contribution

1. Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN 55455
2. Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, MN 55455
3. Clinical Pharmacy Practice Department, The British University in Egypt, El Sherouk, 11837, Egypt
4. Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105
5. Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN 38105

CONTRIBUTIONS BY FIRST AUTHORS

Weijie Zhang: code and method contributor, manuscript writer/reviewer.

Danielle Maeser: code and method contributor, manuscript writer/reviewer.

- Weijie initiated the idea of testing canonical correlation analysis (CCA) as the primary framework in *scIDUC* and proposed the idea of utilizing drug relevant genes (DRGs) to maximize the likelihood of retaining shared transcriptomic patterns associated with drug response; Weijie also implemented selecting drug response relevant features (canonical correlation vectors) to map original data sources to a shared pharmacogenomic subspace for downstream drug response modeling. Weijie led TME study, RMS drug discovery, and CRPC drug nomination/experimental validation. Weijie developed the R Shiny App to provide easy access to scIDUC.
- Danielle initiated the idea of testing non-negative matrix factorization (NMF) as the primary framework in *scIDUC* and tested various methods for optimizing selection of the k/metagene parameter; this involved testing methods commonly used in the truncated SVD performed during integration as well utilizing coassignment probability, consensus matrices, Wasserstein distance, etc. Danielle generated the Windows executable to provide easy access to scIDUC.
- Weijie and Danielle contributed to method development and application to the single-cell datasets provided. This includes discussing batch effects in scRNA-seq data and their impact on use of gold standards for method evaluation, testing integration methods and methods for model training (e.g. ridge, kernel, linear regression), determining metrics for comparing similarity between bulk and single cell data (binary classification metrics, p-value, Cohen's D, rho) and ease of

integration, parameter fine tuning (e.g. k, DRGs, number of genes in general), application of competing methods, and locating data sources.

ABSTRACT

Single-cell sequencing techniques have greatly advanced our current understanding of intratumoral heterogeneity through identifying tumor subpopulations with distinct biologies and therapeutic responses. However, translating biological differences into treatment strategies is challenging, as we still lack tools to facilitate efficient drug discovery that tackles heterogeneous tumors. One key step in development of such approaches centers around accurate prediction of drug response at the single-cell level to offer therapeutic options to specific cell subpopulations. Here, we present a transparent computational framework (nicknamed *scIDUC*) to predict therapeutic efficacies on an individual-cell basis by integrating single-cell transcriptomic profiles with large, data-rich pan-cancer cell line screening datasets. Our method detects shared expression patterns between the two data sources and utilizes such information to project cellular drug response. This method achieves high accuracy, with predicted sensitivities easily able to separate cells into their true cellular drug resistance status as measured by effect size (Cohen's d > 1.0); this holds when using single-cell RNA-seq from both cell line and *in vivo* models. More importantly, we examine our method's utility with three distinct prospective tests, and in each our predicted results are accurate and mirrored biological expectations. In the first two tests, we investigated predicting drugs for cell subpopulations that are resistant to standard-of-care (SOC) therapies due to intrinsic resistance or effects of tumor microenvironments. In both, our results showed high consistency with experimental findings from the original studies. In the third test, we generated SOC therapy resistant cell lines, used scIDUC to identify drugs predicted effective on the resistant line, and validated the predictions with in

vitro experiments. Together, scIDUC quickly and directly translates scRNA-seq data into meaningful cellular drug response for individual cells, displaying the potential to be used by researchers as a first-line tool for nuanced and heterogeneity-aware drug discovery.

INTRODUCTION

Heterogeneity within tumors, where distinct cell subpopulations display varying phenotypic features, has been causally linked to therapy resistance and disease recurrence in many cancers[1–3]. Tumor cells unresponsive to standard-of-care (SOC) pharmacological interventions continue to proliferate and cascade disease progression under the selective pressure. Such phenotypic aberrations often correlate with molecular variations in cellular mutational and transcriptional profiles[4–6]. Thanks to the quickly evolving single-cell (SC) sequencing technologies, genomic and transcriptomic landscapes within tumors in many patient populations have been continuously characterized[7,8]. On the other hand, the increasingly available single-cell sequencing data confers opportunities for development of new treatment strategies that tackle problematic cell groups, address clonal heterogeneity, and eventually help achieve curability in cancers[9,10].

In addition to traditional pharmaceutical research and development pipelines, computational frameworks have emerged as an indispensable tool for drug discovery for their cost-efficient nature and more importantly, the ability to screen many drugs for various indications[11–13]. Current *in silico* drug discovery models are largely constructed based on openly available high-throughput drug screens on pan-cancer cell lines (CCLs)[14,15], whose transcriptomic profiles are systematically evaluated through bulk RNA sequencing (RNA-seq)[16]. While computational tools utilizing relationships between CCL gene expression

from RNA-seq and drug response have demonstrated practicality in predicting efficacious treatments[13,17–19], such relationships cannot be directly applied to generate predictions of drug response at the cellular level, as RNA-seq is limited to measuring average expression across a diverse set of cells, which obscures cell type and composition, as well as temporal and spatial distributions. Thus, inferring cellular drug response requires specialized tools to transfer current bulk-learned drug-gene information to single-cell RNA sequencing (scRNA-seq) data that encapsulate cell level gene expression patterns[20–24].

In recent years, such computational tools have been conceptualized, and a few implementations have also been proposed[10,20,24]. The common crucial functionality among the proposed methods relates to overcoming fundamental differences in properties of bulk and SC RNA-seq data to enable predictions of drug response at the SC resolution using learned drug-gene information from bulk data. To achieve this, Beyondcell, DREEP, and scDr choose to learn a fixed number of drug-specific signature genes from bulk CCL data and apply learned signatures independently in scRNA-seq data to calculate signature scores which indicate drug response[22,25,26]. However, considering that genes may harbor varying predictability for response to different drugs and scRNA-seq data are notorious for their low detection rates as well as stochastic drop-outs[14,27], it is not guaranteed that gene signatures always deliver reliable predictions of sensitivities to various drugs[28]. In comparison, SCAD and scDEAL directly tackle differences between bulk and SC data and emphasize integration of the two domains via neural network based approaches[23,29]. While data-hungry deep learning (DL) routes could benefit from large scRNA-seq data and model complex drug-gene relationships, the availability of CCL bulk data could pose continued limits against accurate parameter estimation. Also, it has been shown that DL methods offer comparable performances as classical machine learning frameworks in drug response modeling[19], while in general consist of more parameters and request more computing resources. CaDRReS-Sc also conducts bulk-SC data integration but through projecting original data into a fixed-dimensional subspace[21]. These integration-embedded methods by default use dichotomous labels for drug response (sensitive or resistant), and such arbitrary

cutoffs often may not reflect pharmacological properties and could mask variation of drug response among heterogeneous cells. Furthermore, insufficient evidence has been presented thus far to demonstrate the translational value of these predictive models in aiding cell-type aware early development for diverse biology models.

Therefore, to fill the current gap and to establish an adaptable virtual SC drug screen platform tailored toward clinically meaningful predictions, we present scIDUC (single-cell Integration and Drug Utility Computation), a novel and transparent transfer learning based framework that quickly and accurately generates predictions of drug responses for scRNA-seq data. scIDUC learns relationships between drug sensitivities and relevant gene expression patterns based on CCL RNA-seq data and CCL high-throughput drug screens. Integration of CCL RNA-seq dataset and target scRNA-seq dataset is performed to denoise and extract shared gene expression patterns between bulk and SC data sources; the resulted bulk data is then used to train drug response models, whose coefficients are further applied to post-integration SC data to infer cellular drug sensitivity scores. We evaluated our method using a variety of scRNA-seq datasets with known cellular drug sensitivity status. Through prospective analysis in three distinct scenarios, we further demonstrated the versatility of our framework in various biological models addressing research questions and generating meaningful therapeutic predictions with potential clinical impact. Validation of predictions yielded from scIDUC substantiates its potential as a first-line research tool in the field of computational drug discovery for addressing intratumoral heterogeneity and to facilitate hypothesis formulation for various oncology research topics.

RESULTS

Overall framework of scIDUC

An overview of our scIDUC computational pipeline is outlined in Figure 1. Drug screen results from the CTRPv2 in normalized area-under-the-dose-response-curve (nAUC) were used as CCL drug response (see Methods). Prior to main computation steps, we aimed to identify drug response relevant genes (DRGs, see Methods) for each drug (Step 0). Instead of filtering for a certain number of DRGs using an arbitrary threshold, a typical DRG list for a drug still consists of all genes, which are ranked from most drug response informative to the least. This is to partially address that different drugs may display different predictability in different scRNA-seq data. The input bulk and SC RNA-seq datasets were then subsetted to retain the same DRGs to facilitate data integration.

Given the distinct properties between bulk RNA-seq and scRNA-seq data, the next step is to integrate the CCL bulk RNA-seq dataset and the target scRNA-seq dataset, while preserving shared gene expression patterns and parsing out less relevant noise (Step 1). The rationale of imposing DRGs is to maximize the likelihood of retaining shared transcriptomic patterns that are associated with drug response. Many data integration methods have been proposed to merge multiple scRNA-seq datasets. The scRNA-seq analysis R package Seurat has incorporated canonical correlation analysis (CCA) as one of the core algorithms to combine multiple SC datasets, based on the rationale that CCA will preserve similarities between data sources[30]. Also, non-negative matrix factorization (NMF) has been used for joining scRNA-seq datasets, partially given the interpretation of its inner decomposition factors as "metagenes"[31,32]. Thus, in Step 1, we examined both CCA and NMF algorithms for integrating bulk and SC datasets, which in theory contain less commonalities compared with merging SC datasets only. We also designed experiments to further evaluate performances of CCA and NMF for accurate drug response predictions.

The integration step generates embeddings of the two input RNA-seq datasets and projects them into a low dimensional space. Next in Step 2, we utilized regression-based approaches to model drug response. We trained models using CCL (bulk) embeddings as predictors and measured drug response as the response. Coefficients of the subspace features were then applied to SC embeddings to generate inferred cellular drug response. This yields predicted nAUC values for all cells in the scRNA-seq data as inferred response to drugs in the CTRPv2.

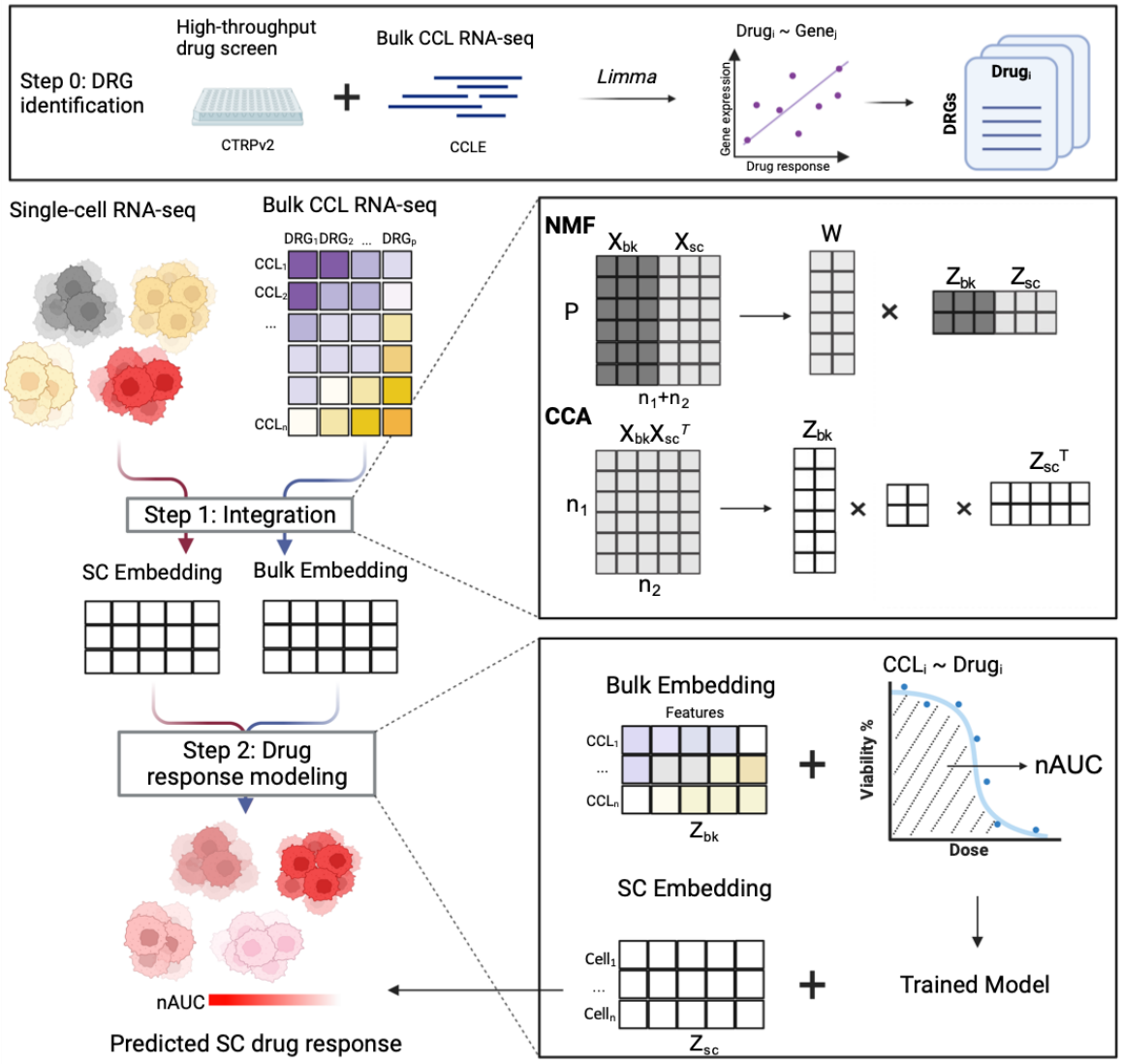


Figure 1. Schematic overview of scIDUC. Drug response relevant genes (DRGs) are generated for use in the scIDUC pipeline (Step 0). Step 1 integrates input CCL bulk RNA-seq data (X_{bk}) and scRNA-seq data (X_{sc}) to preserve shared expression patterns while reducing noise. The resulting embeddings of bulk RNA-seq (Z_{bk}) are used with CCL drug response to construct regression models in Step 2. Learned coefficients are applied to scRNA-seq embeddings (Z_{sc}) to infer cellular drug sensitivity scores. DRG: drug response relevant gene. CCL: cancer cell line. SC: single-cell. NMF: nonnegative matrix factorization. CCA: canonical correlation analysis. nAUC: normalized area under the dose response curve.

Selection of parameters and evaluation of pipeline performances

Based on the overall design of scIDUC, one crucial parameter is the number of DRGs, as it directly affects data integration and drug response performances. Since the input bulk RNA-seq data mostly remains constant (CCL expression profile) while the target scRNA-seq varies, we sought to determine this parameter in a data-dependent way, codifying this parameter as the ratio between number of SCs and number of DRGs (or SC-DRG ratio). In addition, the flexible pipeline also allows us to evaluate different means of integration (Step 1) and drug response modeling (Step 2).

Thus, to establish an optimal structure of our pipeline, we applied scIDUC with different parameters or settings to three independent scRNA-seq datasets with known sensitivity status to specific drugs. To be broadly applicable, we chose datasets that represent various diseases and biological origins. In these data, drug resistance was established through chronic exposing model system(s) to drugs of interest, followed by scRNA-seq of both parent drug-sensitive and derived drug-resistant model(s). Specifically, in the Lung-PC9 dataset, Kong et al. chronically exposed PC9 lung cancer cells to Gefitinib[33] to establish resistance. In the second dataset (Breast-MCF7)[34], Ben-David et al. developed Bortezomib resistant cells derived from the MCF7 breast cancer cell line; Bell et al. generated cells resistant to BET inhibitors from murine acute myeloid leukemia patient derived xenografts (AML-PDX) models (Supplementary Table S1)[35,36] .

We compared predicted cellular drug response from scIDUC with the true sensitive/resistant labels. Specifically, we designed two experiments investigating impacts from (1) means of data integration and drug response modeling as well as from (2) the SC-DRG ratio parameter.

Predicted cellular drug sensitivities from scIDUC were evaluated via calculating two effect sizes between the true resistant and sensitive groups, namely the common-language effect size (Rho statistic) and the

Cohen's D effect size. Rho statistics indicate the probability of a randomly selected cell from the true resistant group having higher predicted nAUC than a randomly selected cell from the true sensitive group (see Methods). A value less than 0.5 indicates contradictory predicted drug response status, a value around 0.5 implies random chance, and a value above 0.5 is ideal. Similarly, Cohen's D describes differences between predicted nAUCs between the two groups while considering variability (see Methods). A Cohen's D can generally be interpreted as having a small effect size at 0.2, a medium effect size at 0.5, and a large effect size at 0.8[37]. Higher values of either criterion therefore signify better performances.

(1). Selection of integration methods (CCA or NMF) and drug response modeling strategies

We first probed different formulae in both data integration (Step 1) as well as in modeling drug response (Step 2). Given that one crucial parameter for both methods is the inner dimension k (number of latent factors for NMF and number of canonical correlation vectors, or CCVs, for CCA), we conducted extensive investigations into the robustness of each integration approach with varying k values ($k = \{1,2,\ldots,50\}$). For drug response modeling, we incorporated linear regression (Lm) and non-parametric regression models based on a Gaussian kernel (Kernel) to ascertain the optimal pipeline. We used the first k CCVs (for CCA) and the first k latent factors (or metagenes, for NMF) respectively and examined predicted single-cell drug sensitivities against the truth (Supplementary Figure S2). Two-sample t-tests using predicted nAUCs were performed between the true resistant cells and sensitive cells, based on which a positive t-statistic indicates correct predicted directions. For Lung-PC9, both methods were able to generate correct sensitivity trends towards gefitinib, while CCA based integration shows superiority in terms of p-values and t-statistics. For Breast-MCF7 and AML-PDX, CCA consistently predicted correct cellular drug response status, whereas volatile test statistics were observed with NMF, especially NMF with downstream linear models for drug response prediction. Although NMF coupled with kernelized

regression resulted in more stable results than NMF and linear models, CCA continued to show superior results regardless of regression models (Supplementary Figure S2). Taken together, CCA integration showed superiority over NMF regarding both accuracy and robustness. We also found that nonparametric modeling of drug response seemed to work better for NMF compared with linear models. When coupled with CCA, both regression models gave comparable results. Additionally, while the selection of the optimal k in either CCA or NMF plays a central role in algorithm performance[30,38], given that the computational goal is to model drug response, we implemented feature selection on post-integration embeddings to include subspace features that correlate with drug response (see Methods). Through this, scIDUC was able to quickly select only a few meaningful features for model training and prediction without screening for optimal inner dimensions in an unsupervised fashion.

(2). Identification of optimal SC-DRG ratios:

Given results from (1), we further tested varying SC-DRG ratios when using (a) no integration, (b) CCA+Lm, and (c) NMF+Kernel. An SC-DRG ratio range spanning from 10 to 0.2 was examined. For each SC-DRG, in Step 2, a subset of 95% of available CCLs were randomly selected as a bootstrap sample to train prediction models. This process is repeated 50 times, which allowed us to test scIDUC's stability and robustness. Evaluation results were reported at each SC-DRG ratio (Figure 2). When integration was performed, median Cohen's D surpassed 0.8 and median Rho statistics surpassed 0.5 for the majority of ratios tested, implying that scIDUC can in general recapitulate known cellular response to drugs across three scRNA-seq datasets regardless of sc-DRG ratios. No clear trend was observed from results without integration. Predicted nAUCs with integration also showed higher robustness, indicated by less variability in the results. While with both integration algorithms the prediction accuracy tends to peak when sc-DRG ratios fell between 1 and 0.2, NMF showed higher variability compared to CCA and had poorer performances outside the ideal sc-DRG ratio range (Figure 2 and Supplementary Table S2). We

observed that a sc-DRG ratio between 1 to 0.2 in general gave good results, supported by stable Rho statistics close to 1 and Cohen's D larger than 1 across all three scRNA-seq data.

Taking results from (1) and (2) together, we prioritize CCA as scIDUC's core integration method. We employ linear regression to model drug response for its simplicity and interpretability. A SC-DRG ratio between 0.2 and 1 was recommended for obtaining optimal predictions. However, the final build of scIDUC package does allow users to explore NMF and kernel regression as alternative settings.

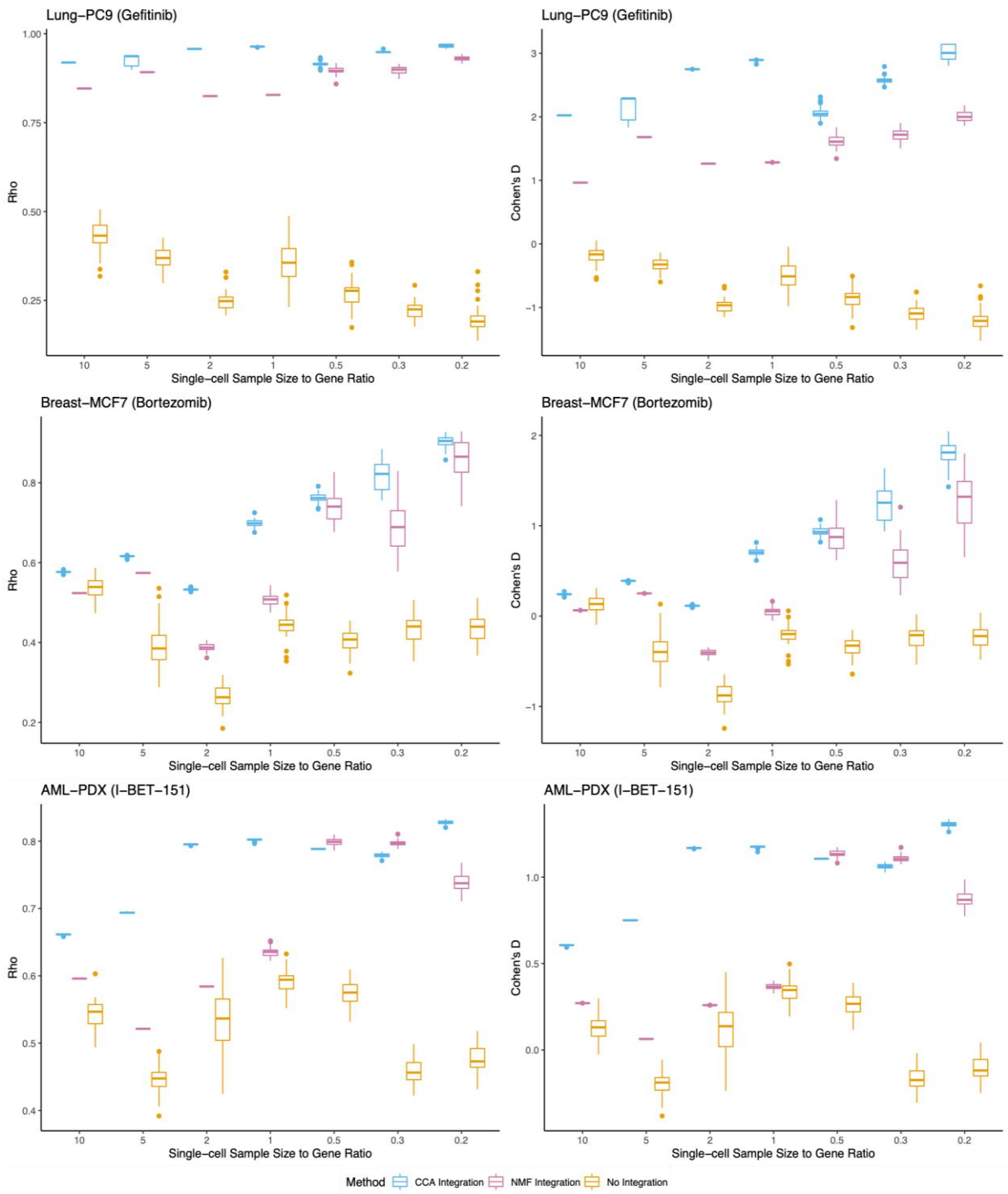


Figure 2 scIDUC recapitulates cellular drug sensitivity status in scRNA-seq data. We applied scIDUC with different data integration settings using varying SC-DRG ratios. 50 bootstrap samples were generated within each ratio. Common-language effect Rho (left column) and Cohen' D (right column)

comparing predicted cell response between the true resistant vs. sensitive cell groups across three scRNA-seq datasets. A larger value indicates a higher separation between the groups. Top row: Lung-PC9; mid row: Breast-MCF7; bottom row: AML-PDX. Basic information of these dataset can be found in Supplementary Table S1. Summarized numeric results are shown in Supplementary Table S2.

scIDUC outperforms other methods in scRNA-seq data from various sources

Next, scIDUC was benchmarked against other methods, namely Beyondcell[22] and CaDRReS-Sc[21], which also aim to predict cellular drug response (Supplementary Information 1). Apart from the three datasets used to evaluate the scIDUC pipeline, we included three additional scRNA-seq data, representing various biology backgrounds as benchmarking datasets (Supplementary Table S1). In CRPC-CCLs, Schnepp et al. exposed castration-resistant prostate cancer (CRPC) cell lines PC3 and DU145 to incremental doses of docetaxel to acquire resistance[39]. The PDAC-CFPAC1 dataset describes ductal pancreatic adenocarcinoma CFPAC1 cells whose drug response was profoundly altered by tumor microenvironment (TME)[40]. When growing in complete classical organoid media, CFPAC1 cells lost basal properties and became responsive to several treatments including SN-38 and paclitaxel. In RMS-oPDX, Patel et al. discovered a mesoderm-like cell colony using scRNA-seq on orthotopic patient-derived xenografts (oPDX) from pediatric patients with rhabdomyosarcoma (RMS). These cells were shown to be highly resistant to the chemotherapy irinotecan (whose active metabolite is SN-38) compared to myoblast cells but sensitive to EGFR inhibitors[41]. These data allowed us to investigate performances of scIDUC and competing methods in diverse biology models and indications.

We benchmarked prediction performances of scIDUC, Beyondcell[22], and CaDRReS-Sc[21] using the same evaluation criteria, namely Cohen's D and Rho statistics. Bootstrapping was implemented for all three methods (see Methods). For scIDUC, since a [0.2,1] SC-DRG range generally produced good results based on our experiment results from Figure 2, we used two SC-DRG ratios (0.2 and 0.9) within this

range to demonstrate the minimal parameter tuning needs for scIDUC, while avoiding biased results. Across all datasets, our method achieves the highest effect sizes across all three benchmarking datasets under both SC-DRG specifications, with median Rho-statistics above 0.8 and median Cohen's D above 1 (Figure 3). Beyondcell was able to separate the true resistant and sensitive cells and showed high consistency, however its results were less accurate (median Rho-statistics around 0.7 and median Cohen's D less than 1) compared with scIDUC. CaDRReS-Sc did not generate meaningful predictions to recollect cellular drug response (median Rho-statistics around 0.5 and median Cohen's D close to 0). Additionally, across datasets Lung-PC9, Breast-MCF7, and AML-PDX, we observed similar results: scIDUC in general had the highest Rho statistics and Cohen's D; Beyondcell provided meaningful predictions though less accurate. Interestingly, CaDRReS-Sc showed good performance with the Breast-MCF7 dataset, though failed to recapitulate true cellular drug response status in the other two. Notebly, for AML-PDX, neither Beyondcell or CaDRReS-Sc was able to recollect true cell drug response status (Supplementary Figure S3). The suboptimal performances of the two other methods is not unanticipated given their methodological designs. For Beyondcell, though applying bulk-learned signatures to SC data could bypass integrating bulk and SC datasets and to some extent reflect cellular response to drugs, these signatures are swayed by quality of scRNA-seq data. Signature score calculating can be subordinate to random factors such as drop-outs and low expression values, resulting in less ideal predictions. CaDRReS-Sc centers its modeling strategies around IC50 instead of AUC, which has been demonstrated to be more reliable for response prediction[42]. Here, our benchmarking results could suggest that modeling strategies used by CaDRReS-Sc lack adaptivity to account for AUCs as drug response. Furthermore, a fixed set of drug response essential genes were used by CaDRReS-Sc for all drugs[43]. Since different drugs might correlate with different genes, an invariant gene set might be insufficient to capture drug-gene relationships. In comparison, scIDUC and Beyondcell are both sensitive to drug-specific genes and showed superior results.

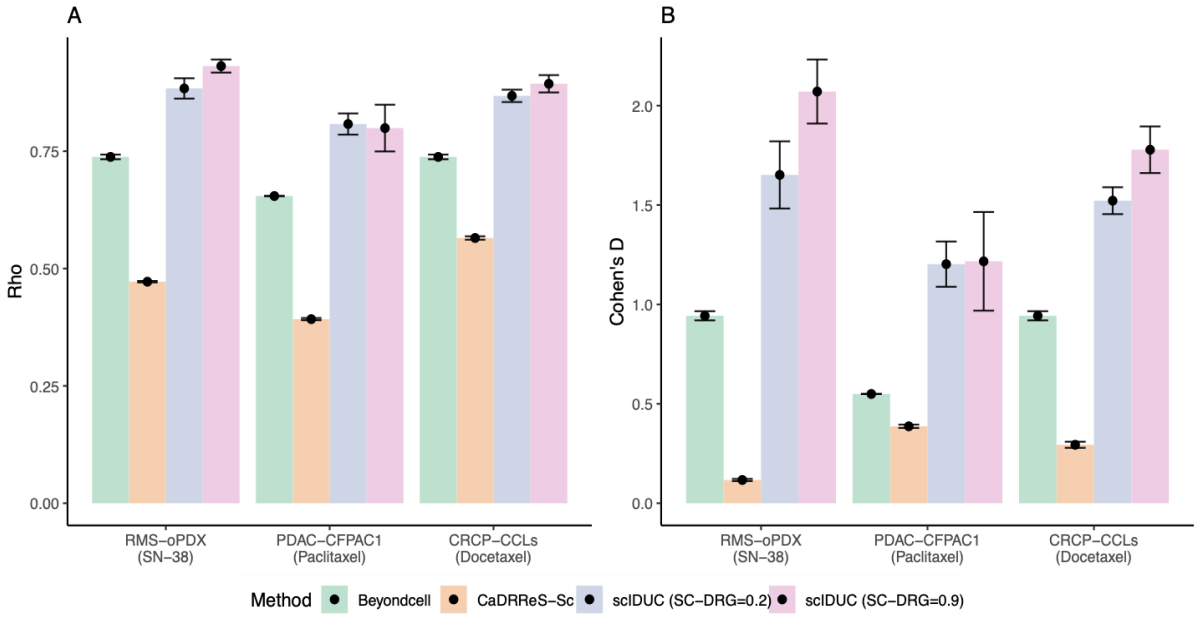


Figure 3 scIDUC outperforms other methods across three scRNA-seq datasets. For each method, 50 bootstrap samples were generated (see Methods). In all three datasets, scIDUC shows higher Common-language effect Rho (A) and Cohen' D (B) comparing predicted cell response between the true resistant vs. sensitive cell groups than that of other methods (CaDRReS-Sc or Beyondcell). Summarized numeric results are shown in Supplementary Table S3.

scIDUC enables clinically meaningful drug discovery

We next showcase the utility of our scIDUC framework as a trailblazer in aiding hypothesis development for clinically meaningful drug discovery. We conducted prospective analyses for three different scenarios utilizing the RMS-oPDX, PDAC-CFPAC1, and CRPC-CCLs datasets, representing diverse biology models including patient-derived xenografts (PDX), tumor microenvironments (TME), and acquired drug resistance *in vitro*. In RMS-oPDX, we applied scIDUC to screen efficacious drugs targeting the SOC (SN-38) resistant mesoderm-like cells. Our nominated drugs showed high consistency with findings from

the original study. In PDAC-CFPAC1, scIDUC was used to predict sensitivities of CFPAC1 cells grown in different TMEs to various drugs. The resulting drug differential efficacy profile between the two TMEs were highly comparable to drug panel results reported by Raghavan et al. Finally, in CRPC-CCLs, we utilized scIDUC to screen drugs showing efficacies in docetaxel-resistant CRPC cells and successfully validated our nomination through in vitro experiments.

Nominating drugs targeting mesoderm cells in RMS patients

In RMS-oPDX, Patel et al. delineated that mesoderm-like cells in pediatric RMS tumors bore profound resistance to SOC chemotherapy irinotecan (SN-38)[41]. Therefore, targeting therapy-resistant mesoderm cells constitutes a cornerstone for curbing the current high rates of disease recurrence. To this end, we applied scIDUC to RMS-oPDX, aiming to discover drugs showing high efficacy in mesoderm cells. To increase likelihood of finding actionable therapeutics, we expanded the pool of candidate drugs by predicting cellular sensitivities to various compounds from not only the CTRPv2 but in addition the Genomics of Drug Sensitivity in Cancer 2 (GDSC2)[44] databases. Two-sample testing was performed between mesoderm (SOC resistant) cells and myoblast (SOC sensitive) cells, through which drugs displaying lower predicted nAUC (suggesting higher sensitivity) in mesoderm cells were included as candidates. To ensure prediction robustness, this pipeline was applied independently to oPDX from each patient (11 in total) in the dataset, and frequency of each nominated drug was summarized over all patients. We considered drugs with a frequency higher than 50%, or nominated from at least 6 out of 11 patients independently as robust candidates. We identified drugs with diverse mechanisms and ranked their target pathways by the number of drugs belonging to the same class (Figure 4A). In both CTRPv2 and GDSC2, epidermal growth factor receptor (EGFR) is among the most frequent targets (Figure 4A). Furthermore, 16 drugs were simultaneously proposed to be efficacious against mesoderm cells by comparing prediction results from both databases. The top five target pathways of the 16 identified compounds were presented in Figure 4B, in which EGFR is ranked the first, targeted by three drugs

(Afatinib, Gefitinib, and Erlotinib). Our drug nomination is strongly supported by the original study, where EGFR was validated as an actionable drug target for chemo-resistant mesoderm cells. The application of scIDUC recapitulates such a finding through independent, data-driven analysis. In addition, scIDUC has proposed other drugs and target pathways as potential strategies for inhibiting mesoderm cells in RMS patients, such as MEK inhibitors (Trametinib and Selumetinib). MEK inhibitors were previously shown to induce tumor differentiation in RMS and potently inhibit RMS in both cell line models and xenograft models[45,46]. These findings warranted the further evaluation of these newly nominated drugs in SOC resistant RMS patients.

Depicting therapeutic susceptibility in PDAC cells affected by different TMEs

Raghavan et al. showed that TME dramatically altered sensitivities to various therapeutics in the PDAC cell line model CFPAC1 as well as in PDAC organoids[40]. For example, as previously recapitulated by scIDUC, CFPAC1 cells grown in classical complete organoid media (scClassical) were re-sensitized to SN-38 and Paclitaxel compared to their counterparts grown in the basal cell line media (scBasal). An additional measured drug screen panel was performed by the authors to obtain a broader differential drug response profile between scBasal and scClassical states induced by different TMEs (Supplementary Figure S3A or originally Figure 6G from Raghavan et al.). To study predictability of scIDUC in this situation, we predicted sensitivities of scBasal and scClassical cells to the same treatments tested in the original drug panel. Based on the drug panel, treatments with demonstrated differential efficacy (i.e., treatments whose mean differential efficacy deviated from zero and confidence interval did not contain zero) were kept, among which seven drugs were found in the CTRPv2 database and used by scIDUC to generate cellular response predictions. Two-sample t-tests were conducted to examine differences in predicted nAUCs between scBasal and scClassical cells (Supplementary Figure S3B). Cohen's D was calculated to demonstrate differences in predicted drug response between scBasal and scClassical cells; a positive Cohen's D indicates more resistance in scBasal cells, whereas a negative value implies the

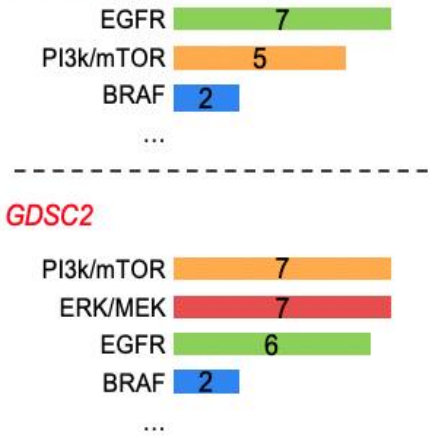
contrary. We plotted Cohen's D of each treatment by its actual value to illustrate the extent of predicted differential efficacy between the two cell states (Figure 4C, Left). Corresponding drug panel results from Raghavan et al. were shown in their original ranking for comparison (Figure 4C, Right). Our results achieved high consistency with drug panel data by Raghavan et al. Gemcitabine, SN-38, and Paclitaxel had the highest Cohen's D, indicating profound discrepancies between resistant scBasal cells and sensitive scClassical cells. MK 1775 and 5-Fluorouracil showed modest and slight differences, whereas Trametinib and Afatinib were predicted to be more sensitive in scBasal cells. Our results were strongly supported by the original experimentally measured drug response in different TMEs, which highlights the capability of scIDUC at capturing TME-shaped drug response at the single-cell level. Furthermore, scIDUC can be applied to hundreds of drugs screened in CCL drug screenings to decipher potential differential drug response in different TMEs, which is extremely expensive if not impossible to carry out at this scale experimentally.

Discovering and validating drugs for docetaxel-resistance in CRPC

Though docetaxel was approved for CRPC, resistance among patients is prevalent[39], highlighting a need to identify new therapeutics targeting the non-responsive cells. We utilized the CRPC-CCLs dataset where docetaxel sensitive and resistant cells have been experimentally defined and supported by our scIDUC prediction. Here we employed scIDUC again to predict cellular sensitivities to hundred other treatments in the CTRPv2 and the GDSC2 databases and prospectively identified drugs showing predicted efficacy in docetaxel resistant cells. We conducted two-sample t-tests comparing docetaxel resistant and sensitive cell groups, through which we selected drugs showing higher effects (lower nAUCs) in the resistant group using an adjusted p-value threshold of less than 0.05. Resulting drugs from each database were ranked based on their adjusted p-values from the lowest to the highest. We selected the top five drugs with the smallest adjusted p-values from each data source and listed their molecular targets (Figure 4D). A number of BRAF inhibitors were identified from the CTRPv2 (vemurafenib and

PLX-4720) and the GDSC2 (PLX-4720 and dabrafenib). In CTRPv2, vemurafenib showed the highest differential efficacy between the two CRPC cell groups, with docetaxel resistant cells having significantly predicted sensitivity than docetaxel resistant cells (Figure 4E, adjusted p-value=5.34×10^{-14}). To experimentally validate this prediction, we chose to test vemurafenib using a previously developed *in vitro* cell line model system containing docetaxel sensitive and resistant DU145 cells. We first exposed these established cells to docetaxel to ensure presence of differential response to the drug (Supplementary Figure S5). To evaluate our candidate drug, we treated cells with increasing concentrations of vemurafenib and generated dose-response curves for both cell groups (Figure 4F). Vemurafenib showed significantly higher inhibitory activity among docetaxel resistant DU145 cells with a mean half maximal inhibitory concentration (IC50) of 12.9 μM/L compared to its IC50 of 27.9 μM/L in sensitive cells (Figure 4G, two-way ANOVA $p<0.0001$). Taken together, our *in vitro* experiment results matched our computational predictions, further supporting the reliability of prospective results generated by scIDUC. Overall, through applying scIDUC in three different scenarios fulfilling varying research needs, we demonstrate its ability to enable drug development targeting heterogeneous tumors. Further examination and experimental validation of our prediction results underscore the potential clinical impact of identified drugs. The diverse biological backgrounds in these cases, including PDX, TMEs, and in vitro cell models, support broad adaptations of scIDUC to enable clinically meaningful drug discovery.

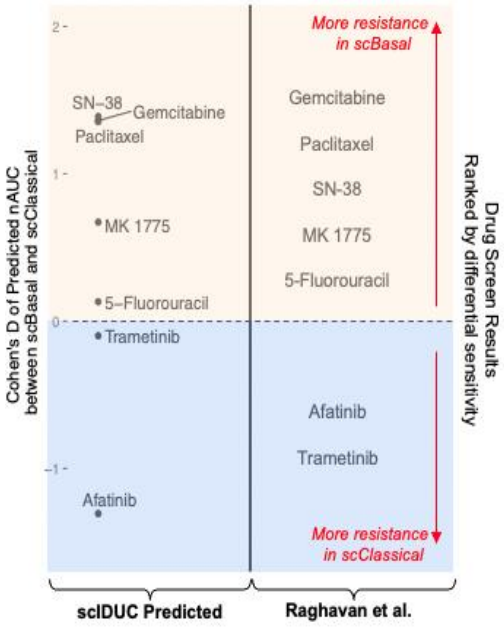
A



B

Target	Drug
EGFR	3: Afatinib, Gefitinib, Erlotinib
MEK	2: Trametinib, Selumetinib
BRAF	1: Dabrafenib, PLX-4720
VEGFR	1: Cediranib
PI3K/mTOR	1: BYL-719
...	...

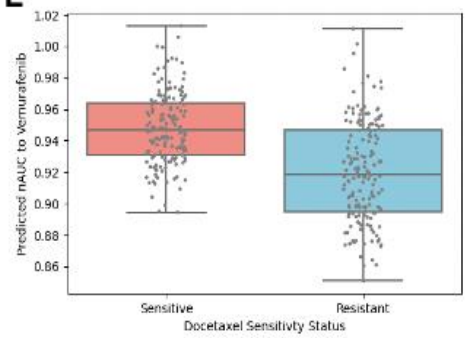
C



D

Drug	Adjusted P-value	Target
CTRPv2		
Vemurafenib	5.34E-14	BRAF
Erlotinib	7.30E-10	EGFR
PLX-4720	2.66E-09	BRAF
Dasatinib	1.66E-08	ABL
Trametinib	2.27E-06	MEK
GDSC2		
PLX-4720	1.85E-12	BRAF
Dabrafenib	8.37E-12	BRAF
AZD3759	1.48E-11	EGFR
ERK_2440	3.37E-11	ERK
Trametinib	1.13E-10	MEK

E



F

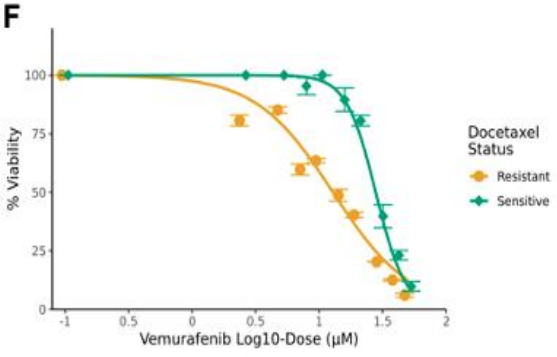


Figure 4 scIDUC facilitates drug discovery in various models by identifying cell-type-specific drug candidates. A. Top drug targets predicted by scIDUC for SOC-resistant mesoderm cells in RMS-oPDX. B. Efficacious drugs for mesoderm cells in RMS-oPDX concurrently identified from the CTRPv2 and the GDSC2 and their targets. C. ScIDUC recapitulates differential efficacies shaped by tumor microenvironments (TME) in the PDAC model CFPAC1 cells. D. Top candidate drugs predicted to be efficacious against docetaxel resistant cells from each high-throughput drug screen E. BRAF inhibitor vemurafenib is predicted to be effective against docetaxel-resistant DU145 cells (adjusted $p<0.0001$). For each cell group, the box shows the median, the first, and the third quartiles of predicted nAUCs. F. Docetaxel-resistant DU145 cells show higher sensitivity to vemurafenib compared to their docetaxel-sensitive counterparts in vitro (two-way ANOVA $p<0.0001$). At each concentration, mean percent viability ± standard deviation is plotted. SOC: Standard-of-Care. CTRPv2: Cancer Therapeutic Response Portal Version 2. GDSC: Genomics of Drug Sensitivity in Cancer. PDAC: Pancreatic Ductal Adenocarcinoma.

DISCUSSION

Assessment of gene expression at the SC level offers detailed mappings of cell compositions and substantially advances understanding of complex diseases such as cancer that involves heterogeneous cell types. Origins of therapy resistance and disease recurrence have been linked to such heterogeneity in many malignancies, often attributed to existence of insusceptible cells thriving under selective pressures[1–3]. Accordingly, cell-type-aware drug discovery using scRNA-seq data has demonstrated potentials to curb resistance and improve treatment outcomes[47–49]. Though computational drug discovery tools have been proposed for this goal amid increasing public access to scRNA-seq datasets, most pipelines still focus on target identification and validation, a procedure often involves generation of scRNA-seq data, or

inference of cellular changes under therapeutic perturbations (instead of cellular response to a potential treatment given its expression profile)[10,50]. These methods so far have not been demonstrated to be able to utilize existing scRNA-seq data to conduct virtual drug screens, facilitate hypothesis formulation, and propose drug candidates addressing tumor heterogeneity. To fill this research need, we have developed scIDUC, which integrates the pan-cancer CCL bulk RNA-seq and scRNA-seq data and infers cellular response to various drugs. By coercing both datasets to have the same DRGs, the CCA-based integration preserves the similarities between bulk and scRNA-seq data in a drug response relevant context, allowing accurate predictions to be made at the single cell level without the need for scRNA-seq drug screen training data, which is difficult to acquire. More importantly, through prospective analysis and validation in three distinct scenarios, we demonstrated the versatility of scIDUC for quickly generating cell-type-specific predictions. The resulting predictions showed high concordance with previous findings and experiment results, further bolstering the utility of scIDUC for providing therapeutic opportunities with clinical impact.

To configure the scIDUC pipeline for optimal results, we evaluated its performances with varying factors including SC-DRG ratios, integration methods, and drug response models against known cell drug response status in independent scRNA-seq datasets. Our results spotlighted an SC-DRG ratio range between 0.2 to 1, data integration via CCA, and linear regression models for accurate predictions. Since this data consisted of parental and resistant cell populations receiving different treatment, we gathered additional evidence showing that scIDUC predictions were not confounded by potential batch effects. First, we compared predicted sensitivities to a variety of drugs between true resistant and true sensitive cell groups in AML-PDX (Supplementary Figure S5). Considering the established resistance against I-BET-151, predicted cellular response toward another BET inhibitor, I-BET-762, also showed significant differential sensitivity. However, predicted response to other drugs such as histone deacetylase inhibitors (entinostat and vorinostat), NAMPT inhibitor (daporinad), and EGFR/HER2 inhibitor (Lapatinib) showed

minimal to no separation between the two groups. Moreover, in the RMS-oPDX dataset, cells from an oPDX sample underwent sequencing altogether and therefore precluded batch effects. In this scenario, scIDUC not only recapitulated resistance to the SOC therapy SN-38 but provided drug nomination highly in line with the original findings. Collectively, our results show that scIDUC predictions are not hindered by potential existence of batch effects.

To date, a few other methods have been proposed to computationally predict single cell drug response using drug screen data on CCLs, such as Beyondcell, CaDRRes-sc, scDR, and scDEAL[21–23,26]. Each of these methods employ a unique transfer learning based approach, utilizing relationships between CCL expression data and drug response to predict single cell level drug sensitivity. However, there are several differences in the actual methodology among these methods, which is reviewed in Supplementary Information 1. We compared scIDUC with CaDRRes-sc and Beyondcell and demonstrated its superiority in prediction accuracy. Across the six scRNA-seq datasets with known cellular drug response labels (gold standards), cellular drug sensitivities predicted by scIDUC not only reiterated true drug response but also had highest precision (evidenced by highest rho statistics and Cohen's D). In comparison, results by Beyondcell echoed true cellular drug sensitivities but lacked accuracy, whereas CaDRRes-sc in general failed to provide meaningful predictions. Notably, CaDRRes-sc utilizes an invariant set of essential genes generated from CRISPR screens in its data integration process[43], while both scIDUC and Beyondcell derive drug-specific marker genes. Our benchmarking results support the rationale of using genes whose expression correlates with measured drug response. Since CRISPR screens detect genes altered by therapeutic perturbations, they may not reflect drug-gene relationships at the baseline level. On the other hand, using signatures alone is susceptible to varying quality of scRNA-seq data which are known to have low detection rates. Meanwhile, calculated scores based on these signatures reflect only relative sensitivities within a scRNA-seq data and lack pharmacological meanings. Taken together, scIDUC achieves desirable results through incorporating both drug-specific features and bulk-SC integration. In

addition, scDR and scDEAL both employ DL methodologies to perform bulk-SC integration and drug response prediction. Both methods embody binarized labels as drug response and train corresponding models as classification problems. Given that drug response across CCLs is better described by spectrum values such as AUCs[51], it is unclear if binary cellular drug labels can capture varying degrees of sensitivity in a heterogeneous tumor. Furthermore, insufficient evidence was given by either method to demonstrate how resulting SC drug response will benefit hypothesis generation and drug discovery. It is also noteworthy that compared with other methods, scIDUC requires minimal parameter tuning, enabling adaptations to a broad user base for various oncology therapy research topics.

Successful characterization of drug response profiles at the SC level plays a fundamental role in advancing precision medicine in cancers[52,53]. Learned cellular sensitivity to various drugs can greatly benefit studies tackling topics such as heterogeneity and cancer drug resistance by providing cell type specific therapy vulnerability information. Aided by the robust prediction results from scIDUC, we spearheaded hypothesis generation and drug candidate identification in three distinct scenarios. For RMS-oPDX, we applied scIDUC independently in 11 oPDX samples to identify drugs showing efficacy against the SOC therapy resistant mesoderm cells. In the majority of the samples EGFR inhibitors were nominated as one of the potential drug classes, reiterating original findings from Patel et al[41]. Moreover, we also discovered a number of other drug classes as potential targets. For example, MEK inhibitors showed high occurrences combating mesoderm-like cells. Previous studies have established evidence that MEK inhibitors effectively inhibit RMS both in vitro and in vivo[45,46]. Moreover, inhibition of the MAPK/ERK pathway by MEK inhibitors have been shown to downregulate mesodermal genes in embryonic stem cells[54]. Given these findings, further evaluation of MEK inhibitors in RMS patients with disease recurrences is warranted. In the second scenario, we showcased that scIDUC was able to capture TME-shaped differential drug response in CFPAC1 cells, a PDAC cell line model. Tumor microenvironmental niche factors have been shown to drive aggressive PDAC progression from a

therapy-responsive “classical” state to a less differentiated “basal” state[40,55]. Thus, characterizing PDAC cellular drug response in different TME-driven states is a crucial first step to develop treatments to curb the current high mortality rate (five-year survival ~9%). Differential cellular drug efficacies between scBasal and scClassical cells resulted from scIDUC accurately recapitulate drug panel testing results reported by Raghavan et al. (Figure 4C). In addition, Shinkawa et al. also reported similar findings where basal-like, poorly progressed PDAC organoids showed higher resistance to Gemcitabine compared to classical-like organoids [55]. These discoveries substantially support usage of scIDUC to streamline drug discovery under different TMEs without the need to conduct large scale drug screens. Finally, we utilized scIDUC again to screen for alternative therapeutics against docetaxel resistance in CRPC CCLs including DU145 and PC3. Our top candidate, namely the BRAF inhibitor vemurafenib, showed consistent higher efficacy in docetaxel resistant DU145 cells than sensitive ones when evaluated in vitro (Figure 4F). A previous trial of vemurafenib has reported an average maximum serum concentration (Cmax) of 61.4 μg/mL, or equivalent to 125.3 μM/L, was well tolerated by patients[56]. Our in vitro experiments exposing vemurafenib in docetaxel resistant cells estimated an IC50 of 12.9 μM/L, which sits well below the safety dose, highlighting its clinical potential to be used in combination with docetaxel to control CRPC progression. Furthermore, two EGFR inhibitors have been proposed to combat docetaxel resistance by our computational pipeline (Figure 4D), consistent with previous studies suggesting that EGFR inhibitors mediate docetaxel resistance in CRPC[57,58]. To sum up, our scIDUC powered computational pipelines were able to quickly propose drug candidates with clinical impact. Our method was able to pinpoint a selective collection of actionable drugs that are ready to be evaluated for different purposes. Serving as an alternative to traditional drug screens and target identification, scIDUC was able to provide rationalized and streamlined drug nomination for therapeutic development. When combined with experimental testing, it can speed up development of efficacious treatments.

In addition, knowledge of intratumoral therapy vulnerability has been explored to inform formulation of drug combinations that target multiple cell groups to help eliminate heterogeneous tumors[20,59]. Given the vast number of potential combination therapies, computational frameworks have been proposed to conduct virtual systematic screens for specific indications[60,61]. To this end, cellular drug response scores are key components for modeling combination efficacies. For example, nAUC might be perceived as the probability of an organism surviving a certain drug treatment; under such an assumption, cellular nAUCs can be used to infer potential drug synergy under the various statisical models[62]. Cellular drug response scores from scIDUC provide key variables for drug combination modeling strategies; our future work will incorporate scIDUC and computational drug combination discovery pipelines to establish a virtual screen platform for various complex cancers.

In sum, we showcase a computational method to depict SC vulnerability to various drugs, establishing a foundation for cell-type-aware drug discovery combating the prevailing issue of treatment failure due to tumor heterogeneity. Our case studies not only provide therapeutic options for various diseases but also substantiate the necessity of our proposed method in aiding efficient sdevelopment of clinical meaningful treatments.

METHODS

Data acquisition

The pan-cancer cell line (CCL) transcriptomic data (bulk RNA-seq) was downloaded from the Cancer Dependency Map (DepMap, https://depmap.org/portal/)[63] and originally from the Cancer Cell Line Encyclopedia (CCLE)[16]. CCLE expression data were downloaded in raw count and log2(TMP+1) formats. CCL drug response data was downloaded from DepMap, originally from the Cancer

Therapeutics Response Portal (CTRPv2) generated at the Broad Institute[15,63,64]. We utilized raw drug screen data to refit dose-response curves and retain robust drug sensitivity profiles[65]. For each drug-CCL pair, area-under-the-dose-response-curve (AUC) was divided by its tested dose range to generate normalized AUC (nAUC), which was then used as the drug response in scIDUC; nAUC is continuous and ranges between 0 to 1 with 0 implying complete cell kill and 1 implying no cell kill.

The single-cell RNA-seq datasets used in this study were downloaded from various sources, depending on the availability of original data provided by the authors. A detailed description of each data source and its properties can be found in Supplementary Table S1.

Preprocessing

CCL names in downloaded CCLE transcriptome and CTRPv2 drug response data were harmonized to Cellosaurus accession numbers, which make use of the prefix "CVCL"[14,66]. Given that a drug was only screened in a subset of CCLs, we calculated percentages of missing values for each drug and excluded those screened in less than 40 percent of all CCLs in the database. This results in a total of 493 treatments (and 887 CCLs) in the CTRPv2 dataset.

For scRNA-seq data with raw counts, each dataset was pre-processed using the Scanpy Python module[67]. A threshold was imposed on all datasets to filter for cells with at least 200 genes detected and genes detected in at least 3 cells. Each cell was then normalized to have the same total counts of 1 million (counts per million, CPM) and log-transformed with a pseudo-count of 1, i.e., log2(CPM+1).

Drug response relevant genes generation

We used the R package limma to detect drug response relevant genes (DRGs) given the continuous nature of nAUC. As recommended by the package, for each drug, CCLE raw expression counts were used to

construct linear models. Resulting genes were ranked by B-statistics which indicates probabilities of differentially expressed from most significantly associated with drug response to least.

Integration of bulk and sc data

We implemented two different approaches to integrate bulk and SC RNA-seq data, namely canonical correlation analysis (CCA) and non-negative matrix factorization (NMF). Let $X_{bk} \in \mathbb{R}^{n_1 \times p}$ be the CCL bulk RNA-seq matrix with n_1 samples and p genes; let $X_{sc} \in \mathbb{R}^{n_2 \times p}$ be a scRNA-seq matrix with n_2 cells and p genes. The bulk- and SC-matrices have the same DRGs.

Integration via CCA

We expanded the CCA integration pipeline proposed in the Seurat package[30]. Briefly, we conducted singular value decomposition (SVD) on the matrix derived based on the multiplication of $X_{bk}{X_{sc}}^T$. $X_{bk}{X_{sc}}^T$ captures similarities between CCLs from bulk RNA-seq and cells from scRNA-seq based on the shared DRGs. Therefore, resulting singular vectors through SVD, i.e.,

$$\text{SVD}(X_{bk}{X_{sc}}^T) = \text{USV}^T = \sum_{i=1}^{k} u_i s_i v_i^T,$$

can be viewed as canonical correlation vectors (CCVs). The left singular vectors $U = u_1, u_2, \dots, u_{n_1}$ correspond to CCVs for bulk data; the right singular vectors $V = v_1, v_2, \dots, v_{n_1}$ correspond to CCVs for SC data.

Accordingly, $Z_{bk} = U\sqrt{S} \in \mathbb{R}^{n_1 \times k}$ and $Z_{sc} = V\sqrt{S} \in \mathbb{R}^{n_2 \times k}$ provides embeddings of X_{bk} and X_{sc} to a subscape where similarities between bulk and single-cell data are preserved. A visualization of this process is provided in Supplementary Figure S1.

Integration via NMF

Since several studies have reported NMF-based methods to capture common gene expression patterns and adjust for discrepancies between batches[31,32], we included NMF as an alternative means to integrate bulk

and single-cell data. Let $Y = [{X_{bk}}^T, {X_{sc}}^T] \in \mathbb{R}^{p\times(n_1+n_2)}$ be a concatenated matrix containing both data sources, then

$$NMF(Y) = WH,$$

Where $W \in \mathbb{R}^{p\times k}$is a common factor matrix whose columns can be viewed as metagenes; $H \in \mathbb{R}^{k\times(n_1+n_2)}$ describes metagene expression profiles for bulk samples and single cells. Within $H, H_{bk} \in \mathbb{R}^{k\times n_1}$ and $H_{sc} \in \mathbb{R}^{k\times n_2}$ are metagene expression matrices for bulk data and SC data, respectively.

Drug-response relevant feature extraction

We performed ad hoc feature extraction to select a pharmacogenomic subspace (if CCA) or pharmacogenomic metagenes (if NMF) for accurately inferring single-cell drug response without the need to determine the inner dimensionality within the matrix decomposition tasks. For embeddings resulted from CCA integration, we correlate each dimension (feature) in Z_{bk} (bulk embeddings) with measured drug response. Resulting p-values are adjusted via the Benjamini–Hochberg procedure to control false discovery rates (FDRs)[68]. Dimensions that have FDRs less than a threshold δ are retained. In other words, such pharmacogenomic subspace comprises dimensions $r = \{r_1, r_2, \ldots, r_j\} \subset \{1, 2, \ldots, k\}$, where $FDR(r_j) < \delta$ and $\delta \in \{0.05, 0.1\}$ by default. Training data is then defined as $X_{train} = {Z_{bk}}^{n_1\times r}$, and predictions of cellular drug response will be made on $X_{test} = {Z_{sc}}^{n_2\times r}$.

For NMF, we select metagenes by correlating each metagene in H_{bk} with measured drug response. Metagenes that have FDRs less than a threshold π are kept. We retain a set of metagenes $m = \{m_1, m_2, \ldots, m_l\} \subset \{1, 2, \ldots, k\}$, where $FDR(m_l) < \pi$. Training data is then defined as $X_{train} = ({H_{bk}}^{m\times n_1})^T$, and predictions of cellular drug response will be made on $X_{test} = ({H_{sc}}^{m\times n_2})^T$.

Prediction

To predict single-cell drug response, we formulate a linear regression model using X_{train} as predictors and measured drug response as the dependent variable. Learned coefficients are then applied to X_{test} to generate nAUCs for cells. We also included an alternative non-parametric regression model based upon the radial basis function (RBF) kernel.

Evaluation metrics

To evaluate the performances of algorithms, we compared predicted cellular nAUCs against true drug sensitivity status (resistant or sensitive) via two-sample t-tests. To better illustrate, we incorporated two additional metrics showing the effect sizes of predicted drug response differences between resistant and sensitive cell groups.

Let the predicted nAUCs of resistant cells be $L_r = (l_{r1}, l_{r2}, \ldots, l_{rp})$ and that of sensitive cells be $L_s = (l_{s1}, l_{s2}, \ldots, l_{sq})$. The common language effect size Rho (ρ) is a non-parametric statistic describing the probability that a randomly selected cell from L_r will have a greater nAUC than a randomly sampled cell from the L_s[69]. Thus, ρ can be directly calculated via the Mann-Whitney U-statistic:

$$\rho = \frac{U}{p \times q}$$

This is also equivalent to the area-under-the-receiver-operating-characteristic-curve (AUC-ROC). Therefore a larger ρ indicates a more accurate prediction result.

We also calculate Cohen's D as a parametric effect size which provides a measure of robustness and variation in addition to differences between two cell groups:

$$Cohen's\ D = \frac{\bar{L_r} - \bar{L_s}}{s}$$

where s is the pooled standard deviation from the two groups, i.e., $s = \sqrt{\frac{(p-1)s^2_{Lr}+(q-1)s^2_{Ls}}{p+q-2}}$. A value of 0.8 or higher is typically viewed as a large effect size and above[37].

Cellular drug response prediction via CaDDReS-Sc and Beyondcell

CaDDReS-Sc by default supports only the GDSC drug screen. While the CaDRReS-Sc github repository suggests flexibility to train a new model based on other drug response datasets like CTRP, this pipeline was not applied to CTRP in neither the original manuscript nor github. To allow for comparison between other similar methods, we used CTRPv2. To calculate the weight for each sample-drug pair that is determined by the logistic weight function, we used max concentrations and AUC as the original pipeline learned weights from max concentrations and IC50s. To generate predictions on the single cell samples, we defined the kernel features as the correlations between CCLs and single cell samples. For Beyondcell, we utilized gene signatures generated from only the CTRPv2 database to infer cellular drug sensitivity scores (Beyondcell Scores or BCS, Supplementary Information 1). Since a higher BCS indicates higher sensitivity, we compared sensitive cells against resistant cells to ensure consistent directions with the rest of the methods. Bootstrap aggregation was performed, and performance was summarized across 50 applications of scIDUC and CaDRReS-Sc where 95% of bulk samples were randomly selected for each application. Given that Beyondcell provides pre-trained signatures, we sampled 95% of up- and down-regulated genes without replacement as a bootstrap experiment.

Cell Culture and Reagents

The DU145 prostate cancer cell line was obtained from American Type Culture Center (ATCC) and cultured in RPMI 1640 medium (Thermo Fisher Scientific), supplemented with 10% fetal bovine serum (FBS) (Gibco, Thermo Fisher Scientific) and maintained at 37 °C with 5% CO_2. A docetaxel-resistant cell line model for DU145 was established by chronically exposing the parent cell line to stepwise increasing

concentrations of docetaxel as previously described[70,71]. Both cell lines were periodically monitored for mycoplasma using the Universal Mycoplasma Detection Kit following the manufacturer's protocol (ATCC). In vitro drug screening in both the docetaxel-resistant and control DU145 models was performed using either vemurafenib (PLX4032; CAS No. 918504-65-1) or docetaxel (RP-56976; CAS No. 114977-28-5) obtained from MedChem Express (Monmouth Junction, NJ, USA) dissolved in dimethylsulfoxide (DMSO) to obtain stock concentrations of 100mM for vemurafenib or 5mM for docetaxel.

Drug Screening in Docetaxel-Resistant and Control DU145 Cell Lines

Docetaxel-resistant and control DU145 cells were trypsinized, harvested, counterstained with Hoechst 33342 Fluorescent Stain (Thermo Scientific, Pierce Biotechnology, Rockford, IL) and resuspended in full growth media to $5x10^4$ cells per mL prior to plating in 96-well microplates (Thermo Scientific) using a seeding density of $5x10^3$ cells per well and allowed to attach for 24 hours. Following incubation, cells were treated with different concentrations of either docetaxel ranging from 0.92nM to 6uM, or vemurafenib ranging from 2.5uM to 50uM. Cell viability for each well was measured following a 72-hour drug exposure using the WST-1 assay [(Roche Applied. Science, Penzberg, Upper Bavaria, Germany) following the manufacturer's protocol. Absorbance at the 450 nm wavelength was assessed using the Synergy HTX Multi-Mode Plate Reader (BioTek, Winooski, VT)]. Absorbance values for each well were used to calculate percent viability relative to the no drug condition. Results are reported as a mean and standard deviation of three independent biological experiments, each containing three technical replicates for each experimental condition.

SUPPLEMENTARY INFORMATION

---

1. Brief overview of other single-cell drug response prediction algorithms

We briefly introduce the methodologies of current competing methods. We compared performances of scIDUC against that of Beyondcell and CaDRReS-Sc in our study using a variety of single-cell (SC) datasets with known drug sensitivity information[1,2]. Additionally, Chen et al. have developed a deep learning (DL) approach–scDEAL based on a variational autoencoder–to infer sensitivities to drugs in scRNA-seq data[3]. Similarly, Zheng et al. used an adversarial learning approach and developed SCAD[4]. These DL approaches utilize strictly binarized drug response (sensitive vs. resistant) instead of a continuous value, which better reflects drug response propertities[5]. ASGARD imputes SC drug sensitivities by scanning for drugs that associate with reversed gene expression from disease samples to normal samples[6]. As a result, it requires both disease and normal scRNA-seq data from the same subjects, which is a limiting factor for its utility.

In scIDUC, the CCA integration process involves singular value decomposition (SVD) on the similarity matrix between SC:

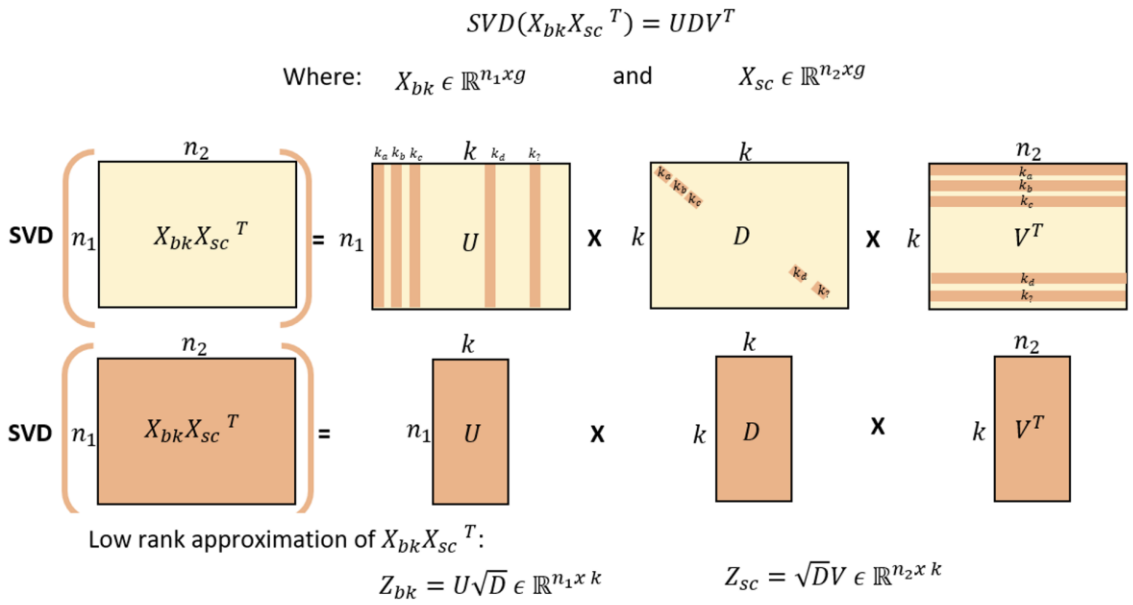


Figure S1 Integration of CCL RNA-seq and single-cell RNA-seq datasets via CCA.

CaDRReS-Sc: CaDRReS-Sc is a machine learning framework used for cancer drug response prediction based on single-cell RNA-sequencing data. It extends a previously established method, CaDRReS, calibrated for higher accuracy of drug response prediction based on single cell data. In brief, it is a matrix factorization model, learning a latent pharmacogenomic space that captures the relationship between drug response profiles and transcriptomic data derived from cells, cell clusters, cell lines, or patients. Cell line data screened across a panel of drugs are used for model training to obtain a more robust drug response prediction through sharing information across drugs. The objective function learns the pharmacogenomic space, incorporating a logistic weight function (C_{iu}) to assign a weight for each sample-drug pair, reducing noise from extrapolation errors in IC50 values from the dose-response curve fitting step. The objective function that is minimized is defined below.

$$Minimize\ \frac{1}{2}\frac{\sum_i\sum_u\left(d_u c_{iu}\left(s_{iu}-\hat{s}_{iu}\right)^2\right)}{k} + \text{regularization}$$

Where:

$$\hat{s}_{ui} = \mu + b_i^Q + b_u^P + q_i * p_u \text{ which is equivalent to } \hat{s}_{ui} = \mu + b_i^Q + b_u^P + q_i * (X_u * W_p)^T C_{iu} = min(f(s_{iu}, o_i, l), f(\hat{s}_{iu}, o_i, l))$$

$$s_{ui} = -\log log(IC50)_2,\ The\ observed\ sensitivity\ score\ of\ sample\ u\ for\ drug\ i$$

$$k = The\ total\ number\ of\ drug\ sample\ pairs$$

$$\mu = The\ overall\ mean\ drug\ response$$

$$b_i^Q = The\ bias\ for\ drug\ i$$

$$b_u^p = The\ bias\ of\ the\ unseen\ sample\ u$$

$$q_i \epsilon R^f = Drug\ i\ in\ the\ f\ dimensional\ latent\ space$$

$$p_u \epsilon R^f = Sample\ u\ in\ the\ f\ dimensional\ latent\ space$$

$$W_p \epsilon R^{dxf} = A\ transformation\ matrix,\ projecting\ transcriptomic\ kernel\ features\ X_u R^d\ for\ each\ sample\ onto\ a\ pharmacogenetic\ space$$

Beyondcell: Beyondcell requires a scRNAseq expression matrix and a collection of drug signatures to compute a scaled Beyondcell enrichment score (BCS), ranging from 0-1. This score indicates the activity of a signature in the expression matrix or how susceptible each cell is based on the analyzed gene signature where a high and low BCS indicates concordance and discordance between the signature and the analyzed cell, respectively. The BCS is defined in Additional file 1 of the original publication. Alternatively, it is included below.

A signature is obtained from a differential expression analysis and consists of drug perturbation, containing transcriptional changes induced by a drug, or drug sensitivity, reflecting the transcriptional status of sensitivity or resistance prior to drug treatment. Alternatively, the user can provide a GMT file/ranked matrix. If functional signatures are applied, the BCS can be used to evaluate the cell's functional status. Therefore, depending on what collection is used, the BCS can measure the cell

perturbation susceptibility or the predicted sensitivity to a given drug. If a gene signature has separate sets of upregulated and downregulated genes and is therefore bidirectional, the BCS is calculated for each signature mode. The individual sum of the expression is calculated and divided by the number of genes in the given signature that are present in the scRNAseq expression matrix.

The BCS is calculated for each drug-cell pair, resulting in a BCS matrix used to determine the presence of therapeutic clusters within the scRNAseq data, visualized using a UMAP. These clusters represent tumor cell subpopulations with distinct shared drug behavior, and the therapeutic differences among the cell populations guide drug selection to nominate cancer-specific treatments. Drug selection is performed using a sensitivity-based ranking, which prioritizes the best drug hits.

For each signature analyzed, a switch point (SP) is calculated, which represents the value in the 0-1 scale where cells switch from down-regulated to up-regulated status. Tumors which are the most therapeutically homogenous will either have a SP of 0, indicating all cells are sensitive to a drug, or 1, indicating all cells are resistant to a drug. A heterogeneous response toward a drug is indicated by a SP between 0-1.

Let $X = (x_{ij})$ be a single-cell expression matrix with n genes contained in $I = \{i_1, \dots, i_n\}$ and m cells contained in $J = \{j_i, \dots, j_m\}$. Also, consider we have a geneset $GS_M = \{S_{M,1}, \dots, S_{M,p}\}$ which consists in a set of p sets of genes (signatures, denoted as S) per mode in $M = \{UP \vee DN\}$.

To compute the BCS, the following steps are taken:

First, we calculate the intersection between I and each S_M in GS_M:

$$G_{M,S_M} = I \cap S_M \quad for\ S_M\ in\ GS_M$$

Second, we obtain the submatrix of X that contains only the genes present in G_{M,S_M}. We denote this submatrix as $Y = (y_{gj})$, with q genes contained in $G_{M,S_M} = \{g_1, \dots, g_q\}$. The raw score for each cell j and signature S_M within a mode M is equal to the mean expression of the genes belonging to G_{M,S_M}.

$$raw_{M,S_M,j} = \bar{y_j} = \frac{1}{|G_{M,S_M}|} \cdot \sum_{k=1}^{q} y_{kj}$$

Then, the raw scores are normalized, using the mean and the standard deviation of the gene expression.

$$norm_{M,S_M,j} = raw_{M,S_M,j} \cdot f$$

The normalization factor f can be decomposed as follows:

$$f = \frac{\sum_{k=1}^{q} y_{kj} - \sqrt{\frac{\sum_{k=1}^{q}(y_{kj} - \bar{y_j})^2}{q-1}}}{\bar{y_j} + \sqrt{\frac{\sum_{k=1}^{q}(y_{kj} - \bar{y_j})^2}{q-1}}}$$

$$f = \frac{sumexpr - sd}{mean + sd}$$

Thus, the higher the standard deviation the lower f. This results in a higher penalization of the raw BCS for cells with outlier genes within G_{M,S_M}. Moreover, the lower the $mean$ and $sumexpr$, the lower f, which further penalizes cells with a great number of zeros.

Finally, we operate on the normalized scores and scale the results between [0, 1] for each signature S_M.

$$BCS_{S_M,j} = \begin{cases} (norm_{\ UP,S_M,j} - norm_{\ DN,S_M,j})\ [0,1] & \text{if } M = \{UP, DN\} \\ norm_{\ UP,S_M,j}\ [0,1] & \text{if } M = \{UP\} \\ -norm_{\ DN,S_M,j}\ [0,1] & \text{if } M = \{DN\} \end{cases}$$

scDEAL: scDEAL integrates a bulk RNA-seq dataset (typically pan-cancer cell line profiles) and a scRNA-seq dataset by minimizing a loss function of maximum mean discrepancy (MMD):

$Loss_{MMD}(E_b(X_b), E_s(X_s)) = |\frac{1}{n}\sum_{i=1}^{n} \quad \phi({x^i}_b) - \frac{1}{m}\sum_{j=1}^{m} \quad \phi({x_s}^j)|_H$, where $X_b = \{{x_b}^i\}_{i\in\{1,2,\ldots,n\}}$ is the bulk RNA-seq with n samples and $X_s = \{{x_s}^j\}_{j\in\{1,2,\ldots,m\}}$ is the scRNA-seq with m cells. ϕ maps original data into a universal reproducing kernel Hilbert space (RKHS) and $|.|_H$indicates the RKHS norm measuring distances between two vectors. Such a loss function pursues similar distributions between the two data sources. By incorporating the additional MMD loss into the optimization of an encoder model for predicting drug response, continuous probability scores Y_s are produced for scRNA-seq data. scDEAL then binarizes cellular drug response using a 0.5 probability threshold. The available scDEAL Python package currently only supports the five scRNA-seq experiment datasets appeared in Chen et al.

2. Information of scRNA-seq data used in the paper

Basic properties of the scRNA-seq used in this manuscript is shown in Table S1. We have preprocessed each dataset using the scanpy package. For each one we included cells with at least 200 RNAs detected and genes detected in at least 3 cells. Cells with high percentages of mitochondria genes were filtered out. We did not select any high variability genes as scIDUC utilizes drug response relevant genes (DRGs) in the input datasets.

Table S1 Information of the scRNA-seq datasets

Data Name	Authors	No. of Cells	Sample Source	Disease	Drug Name	Availability
Lung-PC9	Kong et al.	507	PC9 cell line	Lung cancer	Gefitinib	GSE112274
AML-PDX	Bell et al.	1472	MLL-AF9	Acute Myeloid Leukemia	I-BET-151	GSE110894
Breast-MCF7	Ben-David et al.[a]	2899	MCF7 cell line	Breast cancer	Bortezomib	GSE114462
RMS-oPDX	Patel et al.	5643	Patient-derived xenografts	Rhabdomyosarcoma (RMS)	SN-38 and EGFRis	GSE174376[b]
CRPC-CCLs	Schnepp et al.	324	PC3 and DU145 cell lines	Castration-resistant prostate cancer (CRPC)	Docetaxel	GSE140440
PDAC-CFPAC1	Raghavan et al.	2042	CFPAC1 cell line	Pancreatic ductal adenocarcinoma (PDAC)	SN-38 and Paclitaxel	Single Cell Portal #1644[c]

[a] For the Ben-David dataset, we treated t0 cells as Bortezomib sensitive cells and t96 cells as resistant cells.

[b] The RMS data in our study was obtained directly from Dr. Anand G. Patel.

[c] CFPAC1 cell line data presented in Figure 4 in Raghavan et al. (2021) was used in our paper.

3. Performances of scIDUC and other competing methods

Table S2 scIDUC performance with different integration algorithms and different cell-to-DRG ratios.

Data	Method	Metric	SC-DRG Ratio						
			10	**5**	**2**	**1**	**0.5**	**0.3**	**0.2**
AML-PDX	CCA Integration	-LOG10(P-Value)	15.11 (13.23)	23.23 (19.39)	49.56 (46.08)	53.23 (43.31)	46.6 (40.16)	41.86 (38.96)	58.24 (56.72)
AML-PDX	CCA Integration	Cohen's D	0.61 (0)	0.75 (0)	1.17 (0)	1.17 (0.01)	1.11 (0)	1.06 (0.02)	1.31 (0.02)
AML-PDX	CCA Integration	Rho	0.66 (0)	0.69 (0)	0.8 (0)	0.8 (0)	0.79 (0)	0.78 (0)	0.83 (0)
AML-PDX	NMF Integration	-LOG10(P-Value)	0.55 (0)	1.6 (0)	0.33 (0)	0.96 (1.34)	4.58 (4.72)	6.93 (7.77)	6.72 (5.96)
AML-PDX	NMF Integration	Cohen's D	0.27 (0)	0.06 (0)	0.26 (0)	0.37 (0.02)	1.14 (0.02)	1.11 (0.02)	0.87 (0.05)
AML-PDX	NMF	Rho	0.6 (0)	0.52 (0)	0.58 (0)	0.64 (0.01)	0.8 (0)	0.8 (0)	0.74 (0.01)

	Integration								
AML-PDX	No Integration	-LOG10(P-Value)	15.11 (13.23)	23.23 (19.39)	49.56 (46.08)	53.23 (43.31)	46.6 (40.16)	41.86 (38.96)	58.24 (56.72)
AML-PDX	No Integration	Cohen's D	0.12 (0.06)	-0.2 (0.06)	0.12 (0.16)	0.34 (0.06)	0.26 (0.06)	-0.16 (0.06)	-0.1 (0.07)
AML-PDX	No Integration	Rho	0.54 (0.02)	0.45 (0.02)	0.53 (0.05)	0.59 (0.02)	0.57 (0.02)	0.46 (0.02)	0.48 (0.02)
Breast-MCF7	CCA Integration	-LOG10(P-Value)	6.98 (3.75)	26.78 (15.94)	55 (56.14)	41.5 (33.58)	72.17 (54.36)	106.57 (103.98)	50.39 (100.46)
Breast-MCF7	CCA Integration	Cohen's D	0.24 (0.01)	0.39 (0.01)	0.11 (0.01)	0.71 (0.04)	0.93 (0.05)	1.24 (0.2)	1.79 (0.14)
Breast-MCF7	CCA Integration	Rho	0.58 (0)	0.62 (0)	0.53 (0)	0.7 (0.01)	0.76 (0.01)	0.82 (0.04)	0.9 (0.02)
Breast-MCF7	NMF Integration	-LOG10(P-Value)	3.33 (0)	0.28 (0)	9.61 (12.26)	5.99 (6.41)	12.65 (13.8)	11.72 (13.09)	16.16 (19.22)
Breast-	NMF	Cohen's D	0.06 (0)	0.25 (0)	-0.41 (0.04)	0.05 (0.05)	0.88 (0.16)	0.58 (0.21)	1.27 (0.31)

MCF7	Integration								
Breast-MCF7	NMF Integration	Rho	0.52 (0)	0.57 (0)	0.39 (0.01)	0.51 (0.02)	0.74 (0.04)	0.68 (0.06)	0.86 (0.05)
Breast-MCF7	No Integration	-LOG10(P-Value)	6.98 (3.75)	26.78 (15.94)	55 (56.14)	41.5 (33.58)	72.17 (54.36)	106.57 (103.98)	50.39 (100.46)
Breast-MCF7	No Integration	Cohen's D	0.13 (0.09)	-0.4 (0.2)	-0.87 (0.12)	-0.21 (0.1)	-0.34 (0.1)	-0.23 (0.12)	-0.23 (0.12)
Breast-MCF7	No Integration	Rho	0.54 (0.02)	0.39 (0.05)	0.27 (0.03)	0.44 (0.03)	0.41 (0.03)	0.43 (0.03)	0.44 (0.03)
Lung-PC9	CCA Integration	-LOG10(P-Value)	13.58 (12.84)	14.62 (13.19)	22.71 (14.62)	21.9 (18.97)	14.26 (8.64)	22.27 (11.38)	28.71 (14.94)
Lung-PC9	CCA Integration	Cohen's D	2.02 (0)	2.15 (0.18)	2.75 (0)	2.89 (0.01)	2.06 (0.09)	2.58 (0.05)	3.01 (0.11)
Lung-PC9	CCA Integration	Rho	0.92 (0)	0.93 (0.01)	0.96 (0)	0.96 (0)	0.91 (0.01)	0.95 (0)	0.97 (0)
Lung-PC9	NMF	-LOG10(P-	1.68 (0)	2.92 (0)	24.71 (0)	3.88 (0)	0.81 (0.89)	0.96 (1.03)	0.74 (0.59)

	Integration	Value)							
Lung-PC9	NMF Integration	Cohen's D	0.96 (0)	1.68 (0)	1.26 (0)	1.28 (0)	1.61 (0.09)	1.71 (0.1)	2.01 (0.08)
Lung-PC9	NMF Integration	Rho	0.85 (0)	0.89 (0)	0.82 (0)	0.83 (0)	0.9 (0.01)	0.9 (0.01)	0.93 (0.01)
Lung-PC9	No Integration	-LOG10(P-Value)	13.58 (12.84)	14.62 (13.19)	22.71 (14.62)	21.9 (18.97)	14.26 (8.64)	22.27 (11.38)	28.71 (14.94)
Lung-PC9	No Integration	Cohen's D	-0.19 (0.13)	-0.34 (0.1)	-0.97 (0.1)	-0.52 (0.22)	-0.86 (0.16)	-1.09 (0.12)	-1.2 (0.16)
Lung-PC9	No Integration	Rho	0.43 (0.04)	0.37 (0.03)	0.25 (0.02)	0.36 (0.06)	0.27 (0.04)	0.22 (0.02)	0.2 (0.03)

[a] values are presented as mean (SD).

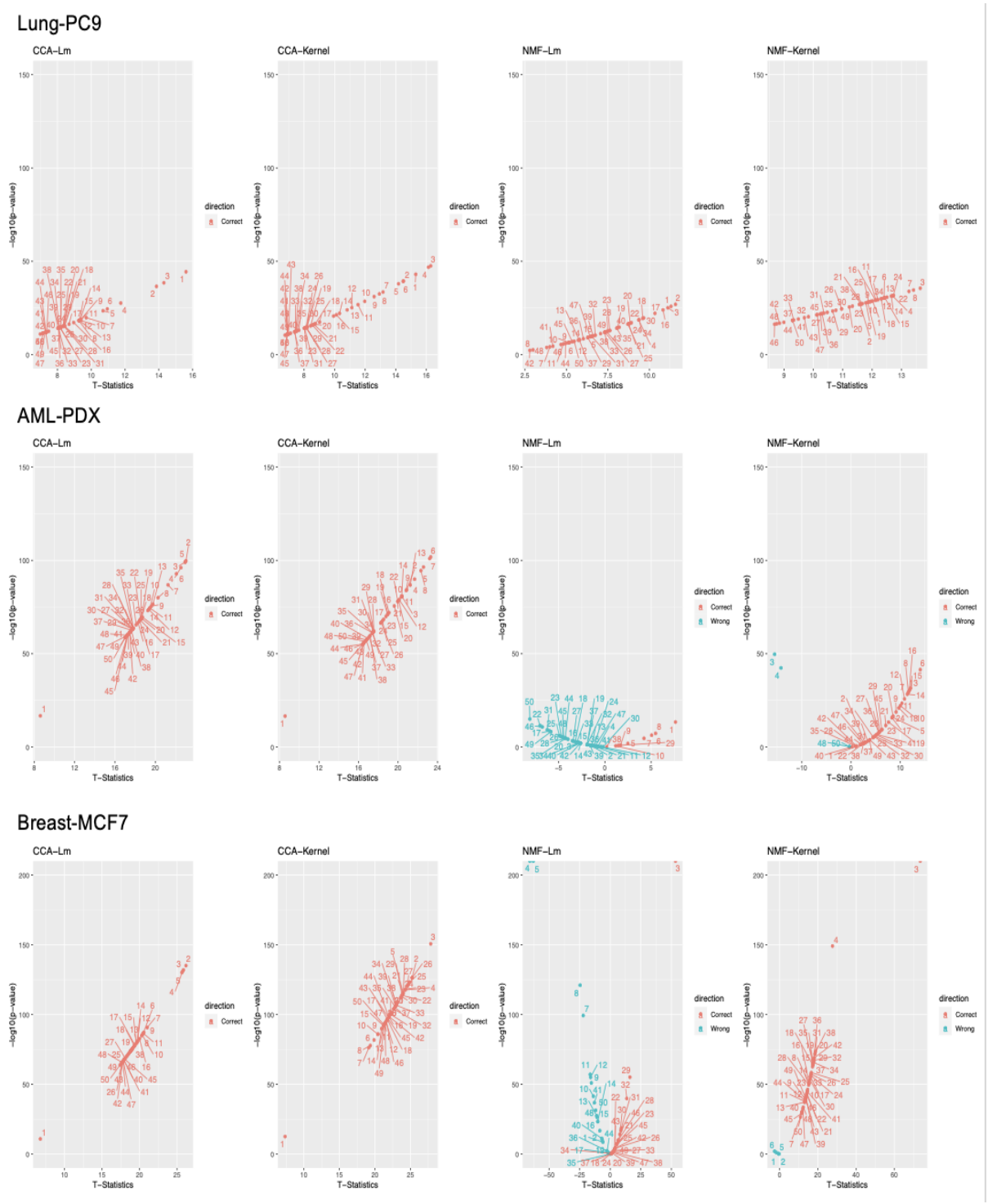


Figure S2 Robustness of CCA and NMF based integration for single-cell drug response prediction.

Combinations of integration methods (CCA or NMF) and drug response models (Lm: linear model;

Kernel: kernelized nonparametric regression) were tested. Inner dimensions of CCA or NMF ranging from 1 to 50 were tested for robustness. T-tests results comparing predicted nAUC between true resistant and sensitive groups are shown; a t-statistic larger than 0 indicates correct direction.

Table S3 Benchmarking scIDUC against other competing methods[a].

Data	Method	Cohen's D	-LOG10(P-Value)	Rho
CRPC-CCLs	Beyondcell	0.94 (0.02)	14.32 (0.6)	0.74 (0)
CRPC-CCLs	CaDRReS-Sc	0.29 (0.02)	2.01 (0.17)	0.57 (0)
CRPC-CCLs	scIDUC (SC-DRG=0.2)	1.52 (0.07)	33.25 (2.28)	0.87 (0.01)
CRPC-CCLs	scIDUC (SC-DRG=0.9)	1.78 (0.12)	42.11 (4.07)	0.89 (0.02)
PDAC-CFPAC1	Beyondcell	0.55 (0)	33.37 (0)	0.65 (0)
PDAC-CFPAC1	CaDRReS-Sc	0.39 (0.01)	17.36 (0.71)	0.39 (0)
PDAC-CFPAC1	scIDUC (SC-DRG=0.2)	1.2 (0.11)	138.46 (22.31)	0.81 (0.02)
PDAC-CFPAC1	scIDUC (SC-DRG=0.9)	1.22 (0.25)	143.98 (44.6)	0.8 (0.05)
RMS-oPDX	Beyondcell	0.94 (0.02)	14.32 (0.6)	0.74 (0)
RMS-oPDX	CaDRReS-Sc	0.12 (0.01)	2.66 (0.21)	0.47 (0)
RMS-oPDX	scIDUC (SC-DRG=0.2)	1.65 (0.17)	300 (0)	0.88 (0.02)

RMS-oPDX	scIDUC (SC-DRG=0.9)	2.07 (0.16)	300 (0)	0.93 (0.01)
AML-PDX	Beyondcell	0.55 (0)	33.37 (0)	0.65 (0)
AML-PDX	CaDRReS-Sc	0.39 (0.01)	17.36 (0.71)	0.39 (0)
AML-PDX	scIDUC (SC-DRG=0.2)	1.2 (0.11)	138.46 (22.31)	0.81 (0.02)
AML-PDX	scIDUC (SC-DRG=0.9)	1.22 (0.25)	143.98 (44.6)	0.8 (0.05)
Breast-MCF7	Beyondcell	1.37 (0.07)	238.46 (19.84)	0.83 (0.01)
Breast-MCF7	CaDRReS-Sc	1.66 (0.01)	Inf (NaN)	0.88 (0)
Breast-MCF7	scIDUC (SC-DRG=0.2)	1.77 (0.15)	300 (0)	0.9 (0.02)
Breast-MCF7	scIDUC (SC-DRG=0.9)	0.77 (0.04)	300 (0)	0.72 (0.01)
Lung-PC9	Beyondcell	2.05 (0.05)	15.33 (0.48)	0.92 (0)
Lung-PC9	CaDRReS-Sc	0.46 (0.06)	23.04 (10.79)	0.36 (0.03)
Lung-PC9	scIDUC (SC-DRG=0.2)	2.94 (0)	41.47 (0)	0.97 (0)
Lung-PC9	scIDUC (SC-DRG=0.9)	2.94 (0)	41.47 (0)	0.97 (0)

[a] The values are presented as mean (SD).

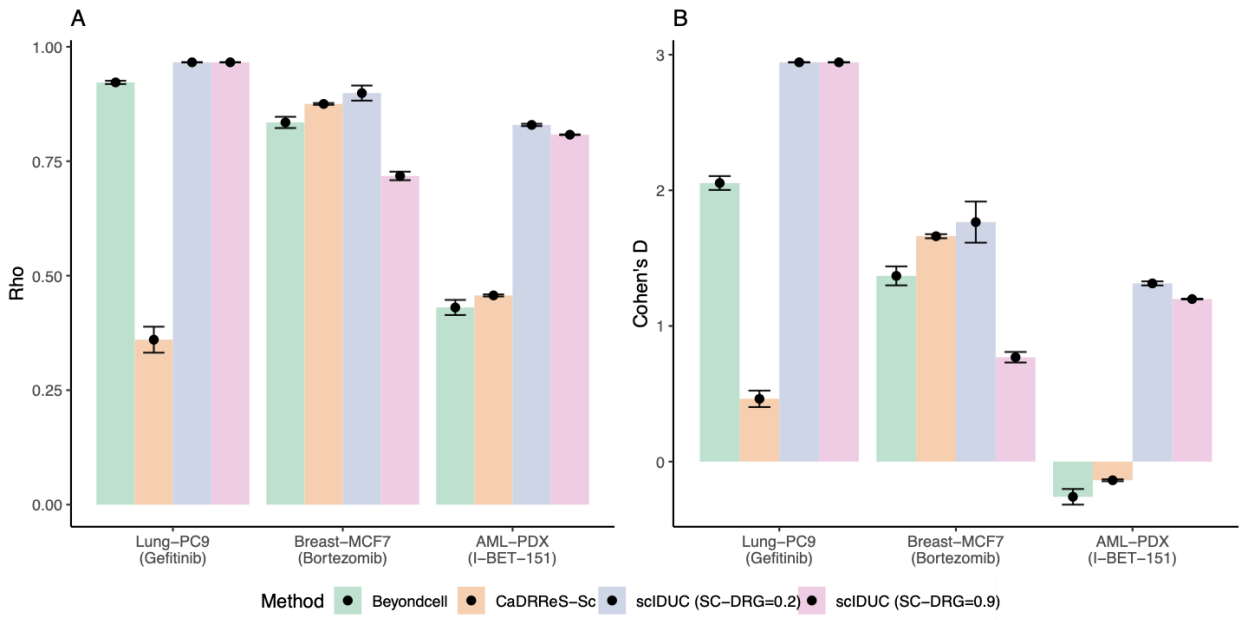


Figure S3 scIDUC outperforms other methods across three additional scRNA-seq datasets. For each method, 50 bootstrap samples were generated (see Methods). In all three datasets, scIDUC shows higher Common-language effect Rho (A) and Cohen' D (B) comparing predicted cell response between the true resistant vs. sensitive cell groups than that of other methods (CaDRReS-Sc or Beyondcell).

4. Prospective analysis to capture therapeutic vulnerabilities in PDAC-CFPAC1 cells shaped by TMEs

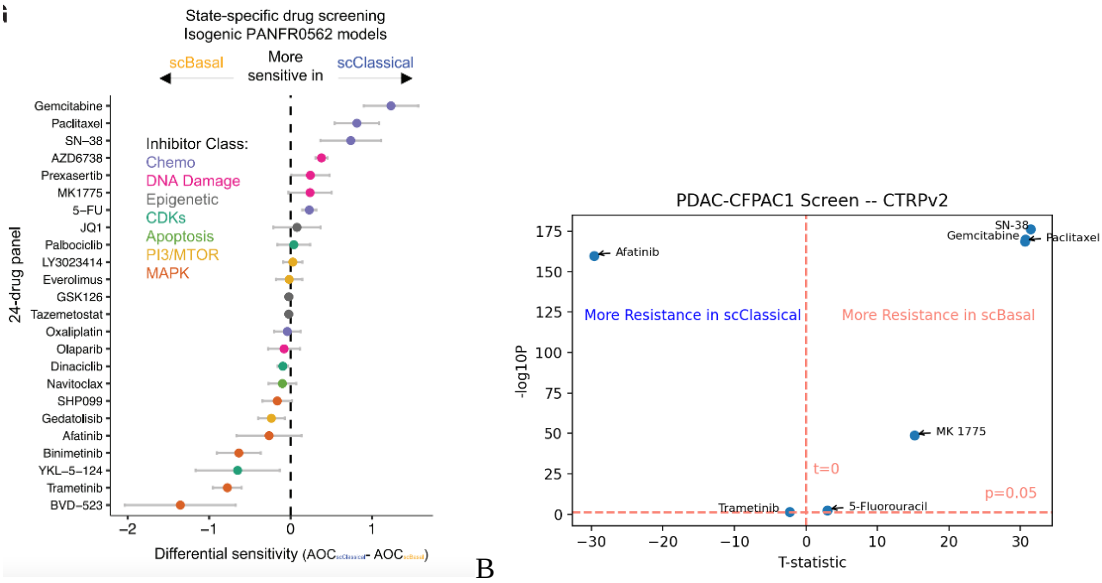


Figure S4 A. Drug panel screens showing differential efficacy in a pancreatic ductal carcinoma patient-derived xenograft model with two different subtypes due to different TMEs. This figure was originally generated by Raghavan et al. (Figure 6G in DOI:https://doi.org/10.1016/j.cell.2021.11.017). **B.** Predicted cellular response to various drugs using scIDUC. T-tests results comparing predicted nAUCs between scBasal and scClassical cells are shown.

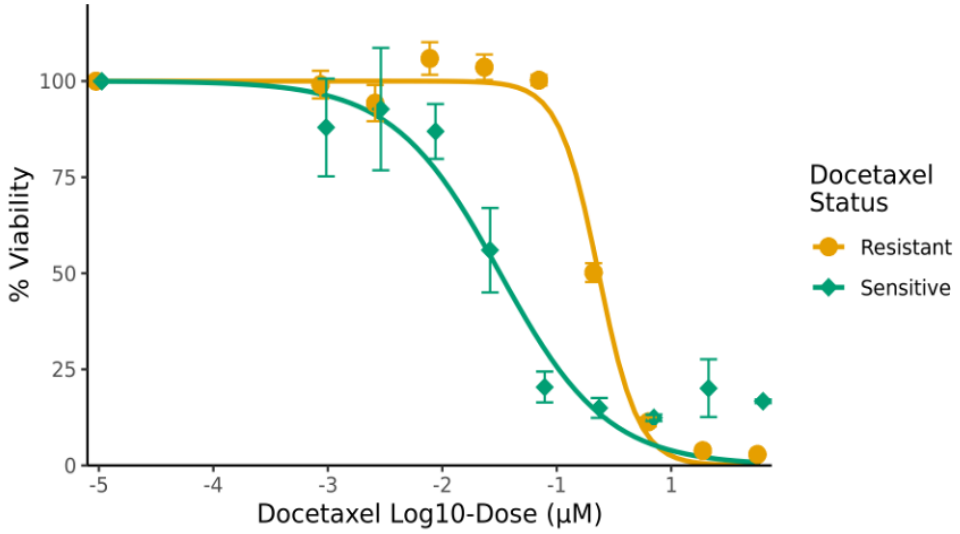


Figure S5 WST assay showing differential sensitivity to docetaxel among DU145 cells. Docetaxel-resistant DU145 cells show higher resistance to docetaxel compared to their docetaxel-sensitive counterparts in vitro (two-way ANOVA $p<0.0001$). At each concentration, mean percent viability ± standard deviation is plotted.

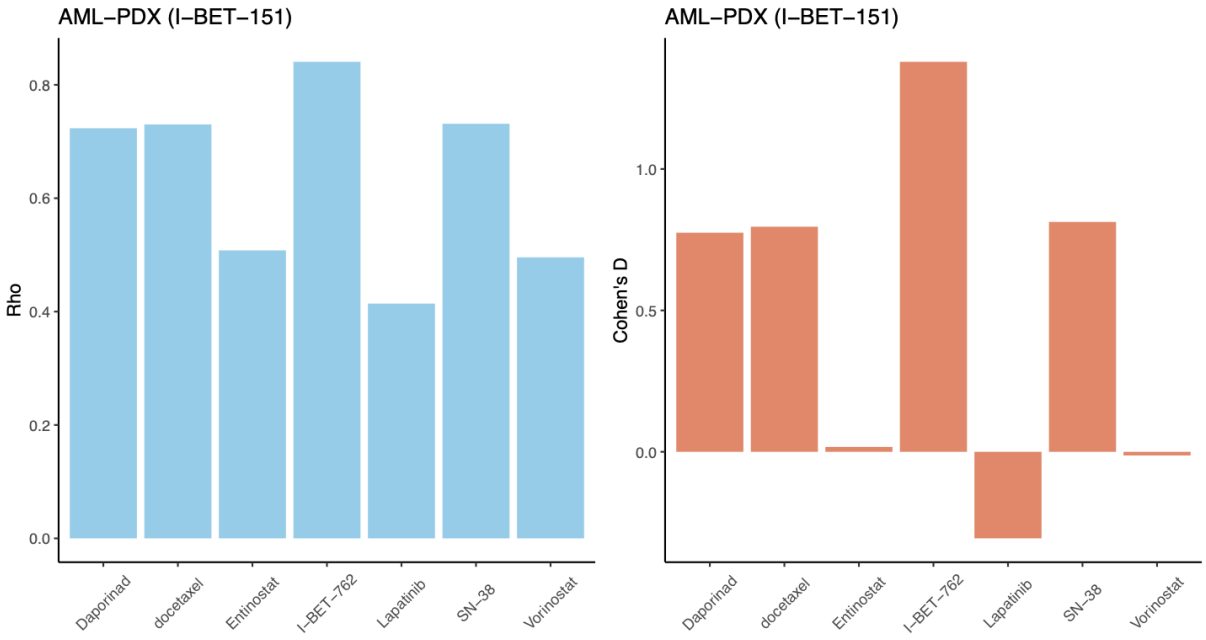


Figure S6 scIDUC predicts accurate results in the presence of potential bath effects. Left: Rho statistics comparing predicted nAUCs between I-BET sensitive cells and resistant cells over seven drugs. Right: Cohen's D comparing predicted nAUCs between I-BET sensitive cells and resistant cells over seven drugs.

REFERENCE

---

1. Jamal-Hanjani M, Quezada SA, Larkin J, Swanton C. Translational Implications of Tumor Heterogeneity. *Clin Cancer Res*. 2015;21(6):1258-1266. doi:10.1158/1078-0432.CCR-14-1429

2. Schmidt F, Efferth T. Tumor Heterogeneity, Single-Cell Sequencing, and Drug Resistance. *Pharmaceuticals*. 2016;9(2):33. doi:10.3390/ph9020033

3. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;168(4):613-628. doi:10.1016/j.cell.2017.01.018

4. Kim KT, Lee HW, Lee HO, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol*. 2015;16(1):127. doi:10.1186/s13059-015-0692-3

5. Liu J, Dang H, Wang XW. The significance of intertumor and intratumor heterogeneity in liver cancer. *Exp Mol Med*. 2018;50(1):e416-e416. doi:10.1038/emm.2017.165

6. Saeed K, Ojamies P, Pellinen T, et al. Clonal heterogeneity influences drug responsiveness in renal cancer assessed by ex vivo drug testing of multiple patient-derived cancer cells. *International Journal of Cancer*. 2019;144(6):1356-1366. doi:10.1002/ijc.31815

7. Ortega MA, Poirion O, Zhu X, et al. Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clinical and Translational Medicine*. 2017;6(1):46. doi:10.1186/s40169-017-0177-y

8. Wu F, Fan J, He Y, et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat Commun*. 2021;12(1):2540. doi:10.1038/s41467-021-22801-0

9. Heath JR, Ribas A, Mischel PS. Single-cell analysis tools for drug discovery and development.

Nat Rev Drug Discov. 2016;15(3):204-216. doi:10.1038/nrd.2015.16

10. Van de Sande B, Lee JS, Mutasa-Gottgens E, et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov*. Published online April 28, 2023:1-25. doi:10.1038/s41573-023-00688-4

11. Karaman B, Sippl W. Computational Drug Repurposing: Current Trends. *Current Medicinal Chemistry*. 2019;26(28):5389-5409. doi:10.2174/0929867325666180530100332

12. Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov*. 2019;18(1):41-58. doi:10.1038/nrd.2018.168

13. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precis Onc*. 2020;4(1):1-10. doi:10.1038/s41698-020-0122-1

14. Ling A, Gruener RF, Fessler J, Huang RS. More than Fishing for a Cure: The Promises and Pitfalls of High Throughput Cancer Cell Line Screens. *Pharmacol Ther*. 2018;191:178-189. doi:10.1016/j.pharmthera.2018.06.014

15. Rees MG, Seashore-Ludlow B, Cheah JH, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol*. 2016;12(2):109-116. doi:10.1038/nchembio.1986

16. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603-607. doi:10.1038/nature11003

17. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol*. 2014;15(3):R47. doi:10.1186/gb-2014-15-3-r47

18. Suphavilai C, Bertrand D, Nagarajan N. Predicting Cancer Drug Response using a Recommender System. Wren J, ed. *Bioinformatics*. 2018;34(22):3907-3914. doi:10.1093/bioinformatics/bty452

19. Smith AM, Walsh JR, Long J, et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*. 2020;21(1):119. doi:10.1186/s12859-020-3427-8

20. Wu Z, Lawrence PJ, Ma A, Zhu J, Xu D, Ma Q. Single-Cell Techniques and Deep Learning in Predicting Drug Response. *Trends in Pharmacological Sciences*. 2020;41(12):1050-1065. doi:10.1016/j.tips.2020.10.004

21. Suphavilai C, Chia S, Sharma A, et al. Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Medicine*. 2021;13(1):189. doi:10.1186/s13073-021-01000-y

22. Fustero-Torre C, Jiménez-Santos MJ, García-Martín S, et al. Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. *Genome Medicine*. 2021;13(1):187. doi:10.1186/s13073-021-01001-x

23. Chen J, Wu Z, Qi R, et al. Deep Transfer Learning of Drug Responses by Integrating Bulk and Single-cell RNA-seq data. Published online August 6, 2021:2021.08.01.454654. doi:10.1101/2021.08.01.454654

24. Ji Y, Lotfollahi M, Wolf FA, Theis FJ. Machine learning for perturbational single-cell omics. *Cell Systems*. 2021;12(6):522-537. doi:10.1016/j.cels.2021.05.016

25. Gambardella G, Viscido G, Tumaini B, Isacchi A, Bosotti R, di Bernardo D. A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. *Nat Commun*. 2022;13(1):1714. doi:10.1038/s41467-022-29358-6

26. Lei W, Yuan M, Long M, et al. scDR: Predicting Drug Response at Single-Cell Resolution. *Genes*. 2023;14(2):268. doi:10.3390/genes14020268

27. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat Commun*. 2020;11(1):1169. doi:10.1038/s41467-020-14976-9

28. Qi X, Shen M, Fan P, et al. The Performance of Gene Expression Signature-Guided Drug–

Disease Association in Different Categories of Drugs and Diseases. *Molecules*. 2020;25(12):2776. doi:10.3390/molecules25122776

29. Zheng Z, Chen J, Chen X, et al. Enabling Single-Cell Drug Response Annotations from Bulk RNA-Seq Using SCAD. *Advanced Science*. 2023;10(11):2204113. doi:10.1002/advs.202204113

30. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031

31. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat Protoc*. 2020;15(11):3632-3662. doi:10.1038/s41596-020-0391-8

32. Peng M, Li Y, Wamsley B, Wei Y, Roeder K. Integration and transfer learning of single-cell transcriptomes via cFIT. *Proceedings of the National Academy of Sciences*. 2021;118(10):e2024383118. doi:10.1073/pnas.2024383118

33. Kong SL, Li H, Tai JA, et al. Concurrent Single-Cell RNA and Targeted DNA Sequencing on an Automated Platform for Comeasurement of Genomic and Transcriptomic Signatures. *Clinical Chemistry*. 2019;65(2):272-281. doi:10.1373/clinchem.2018.295717

34. Ben-David U, Siranosian B, Ha G, et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*. 2018;560(7718):325-330. doi:10.1038/s41586-018-0409-3

35. Fong CY, Gilan O, Lam EYN, et al. BET inhibitor resistance emerges from leukaemia stem cells. *Nature*. 2015;525(7570):538-542. doi:10.1038/nature14888

36. Bell CC, Fennell KA, Chan YC, et al. Targeting enhancer switching overcomes non-genetic drug resistance in acute myeloid leukaemia. *Nat Commun*. 2019;10(1):2723. doi:10.1038/s41467-019-10652-9

37. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol*. 2013;4:863. doi:10.3389/fpsyg.2013.00863

38. Maisog JM, DeMarco AT, Devarajan K, Young SS, Fogel P, Luta G. Assessing Methods for

Evaluating the Number of Components in Non-Negative Matrix Factorization. *Mathematics (Basel)*. 2021;9(22):2840. doi:10.3390/math9222840

39. Schnepp PM, Shelley G, Dai J, et al. Single-Cell Transcriptomics Analysis Identifies Nuclear Protein 1 as a Regulator of Docetaxel Resistance in Prostate Cancer Cells. *Mol Cancer Res*. 2020;18(9):1290-1301. doi:10.1158/1541-7786.MCR-20-0051

40. Raghavan S, Winter PS, Navia AW, et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell*. 2021;184(25):6119-6137.e26. doi:10.1016/j.cell.2021.11.017

41. Patel AG, Chen X, Huang X, et al. The myogenesis program drives clonal selection and drug resistance in rhabdomyosarcoma. *Developmental Cell*. 2022;57(10):1226-1240.e8. doi:10.1016/j.devcel.2022.04.003

42. Kurilov R, Haibe-Kains B, Brors B. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Sci Rep*. 2020;10(1):2849. doi:10.1038/s41598-020-59656-2

43. Wang T, Birsoy K, Hughes NW, et al. Identification and characterization of essential genes in the human genome. *Science*. 2015;350(6264):1096-1101. doi:10.1126/science.aac7041

44. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*. 2013;41(D1):D955-D961. doi:10.1093/nar/gks1111

45. Yohe ME, Gryder BE, Shern JF, et al. MEK inhibition induces MYOG and remodels super-enhancers in RAS-driven rhabdomyosarcoma. *Sci Transl Med*. 2018;10(448):eaan4470. doi:10.1126/scitranslmed.aan4470

46. Danielli SG, Porpiglia E, De Micheli AJ, et al. Single-cell profiling of alveolar rhabdomyosarcoma reveals RAS pathway inhibitors as cell-fate hijackers with therapeutic relevance. *Science Advances*. 2023;9(6):eade9238. doi:10.1126/sciadv.ade9238

47. Sade-Feldman M, Yizhak K, Bjorgaard SL, et al. Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell*. 2018;175(4):998-1013.e20.

doi:10.1016/j.cell.2018.10.038

48. Cohen YC, Zada M, Wang SY, et al. Identification of resistance pathways and therapeutic targets in relapsed multiple myeloma patients through single-cell sequencing. *Nat Med*. 2021;27(3):491-503. doi:10.1038/s41591-021-01232-w

49. Abdelfattah N, Kumar P, Wang C, et al. Single-cell analysis of human glioma and immune cells identifies S100A4 as an immunotherapy target. *Nat Commun*. 2022;13(1):767. doi:10.1038/s41467-022-28372-y

50. Lotfollahi M, Susmelj AK, Donno CD, et al. Learning interpretable cellular responses to complex perturbations in high-throughput screens. Published online May 18, 2021:2021.04.14.439903. doi:10.1101/2021.04.14.439903

51. Bouhaddou M, DiStefano MS, Riesel EA, et al. Drug response consistency in CCLE and CGP. *Nature*. 2016;540(7631):E9-E10. doi:10.1038/nature20580

52. Filipp FV. Opportunities for Artificial Intelligence in Advancing Precision Medicine. *Curr Genet Med Rep*. 2019;7(4):208-213. doi:10.1007/s40142-019-00177-4

53. Qi R, Zou Q. Trends and Potential of Machine Learning and Deep Learning in Drug Study at Single-Cell Level. *Research*. 2023;6:0050. doi:10.34133/research.0050

54. Kimura T, Kaga Y, Ohta H, et al. Induction of Primordial Germ Cell-Like Cells From Mouse Embryonic Stem Cells by ERK Signal Inhibition. *Stem Cells*. 2014;32(10):2668-2678. doi:10.1002/stem.1781

55. Shinkawa T, Ohuchida K, Mochida Y, et al. Subtypes in pancreatic ductal adenocarcinoma based on niche factor dependency show distinct drug treatment responses. *J Exp Clin Cancer Res*. 2022;41:89. doi:10.1186/s13046-022-02301-9

56. Zhang W, Heinzmann D, Grippo JF. Clinical Pharmacokinetics of Vemurafenib. *Clin Pharmacokinet*. 2017;56(9):1033-1043. doi:10.1007/s40262-017-0523-7

57. Hour TC, Chung SD, Kang WY, et al. EGFR mediates docetaxel resistance in human castration-

resistant prostate cancer through the Akt-dependent expression of ABCB1 (MDR1). *Arch Toxicol*. 2015;89(4):591-605. doi:10.1007/s00204-014-1275-x

58. Lin JZ, Hameed I, Xu Z, Yu Y, Ren ZY, Zhu JG. Efficacy of gefitinib-celecoxib combination therapy in docetaxel-resistant prostate cancer. *Oncology Reports*. 2018;40(4):2242-2250. doi:10.3892/or.2018.6595

59. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*. 2018;15(2):81-94. doi:10.1038/nrclinonc.2017.166

60. Ling A, Huang RS. Computationally predicting clinical drug combination efficacy with cancer cell line screens and independent drug action. *Nat Commun*. 2020;11(1):5848. doi:10.1038/s41467-020-19563-6

61. Kong W, Midena G, Chen Y, et al. Systematic review of computational methods for drug combination prediction. *Computational and Structural Biotechnology Journal*. 2022;20:2807-2814. doi:10.1016/j.csbj.2022.05.055

62. Plana D, Palmer AC, Sorger PK. Independent Drug Action in Combination Therapy: Implications for Precision Oncology. *Cancer Discovery*. 2022;12(3):606-624. doi:10.1158/2159-8290.CD-21-0212

63. Basu A, Bodycombe NE, Cheah JH, et al. An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell*. 2013;154(5):1151-1161. doi:10.1016/j.cell.2013.08.003

64. Seashore-Ludlow B, Rees MG, Cheah JH, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov*. 2015;5(11):1210-1223. doi:10.1158/2159-8290.CD-15-0235

65. Simplicity v1.0. Accessed June 20, 2023. https://oncotherapyinformatics.org/simplicity/

66. Bairoch A. The Cellosaurus, a Cell-Line Knowledge Resource. *J Biomol Tech*. 2018;29(2):25-38. doi:10.7171/jbt.18-2902-002

67. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*. 2018;19(1):15. doi:10.1186/s13059-017-1382-0

68. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.

69. McGraw KO, Wong SP. A common language effect size statistic. *Psychological Bulletin*. 1992;111:361-365. doi:10.1037/0033-2909.111.2.361

70. Zhang W, Lee AM, Jena S, et al. Computational drug discovery for castration-resistant prostate cancers through in vitro drug response modeling. *Proceedings of the National Academy of Sciences*. 2023;120(17):e2218522120. doi:10.1073/pnas.2218522120

71. Shan Y, Huang Y, Lee AM, Mentzer J, Ling A, Huang RS. A Long Noncoding RNA, GAS5 Can Be a Biomarker for Docetaxel Response in Castration Resistant Prostate Cancer. *Frontiers in Oncology*. 2021;11. Accessed May 9, 2022. https://www.frontiersin.org/article/10.3389/fonc.2021.675215

CHAPTER 5: A REVIEW OF COMPUTATIONAL METHODS FOR PREDICTING CANCER DRUG RESPONSE AT THE SINGLE-CELL LEVEL THROUGH INTEGRATION WITH BULK RNASEQ DATA

Danielle Maeser[1,2], Weijie Zhang[1,2], Yingbo Huang[2], R. Stephanie Huang[2]

[1]Department of Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN, United States

[2]Department of Experimental and Clinical Pharmacology, University of Minnesota, Minneapolis, MN, United States

CONTRIBUTIONS BY FIRST AUTHORS

Danielle Maeser: code contributor, manuscript writer/reviewer.

Weijie Zhang: code contributor, manuscript writer/reviewer.

Yingbo Huang: code contributor, manuscript writer/reviewer.

- Danielle, Weijie, and Yingbo contributed methods toward performing the literature review as well as keyword suggestions. Initially, Yingbo provided a R based method for web scraping PubMed, and Weijie suggested referencing Google Scholar and identified several methods to

include in the review. Danielle ran the executable used to generate the final results and summarized them. All three authors contributed ideas toward method organization as well as drafts of the final figure, and the final figures were generated by Yingbo.

ABSTRACT

Cancer treatment failure is often attributed to tumor heterogeneity, where diverse malignant cell clones exist within a patient. Despite a growing understanding of heterogeneous tumor cells depicted by single-cell RNA sequencing (scRNA-seq), there is still a gap in the translation of such knowledge into treatment strategies tackling the pervasive issue of therapy resistance. In this review, we survey methods leveraging large-scale drug screens to generate cellular sensitivities to various therapeutics. These methods enable efficient drug screens in scRNA-seq data and serve as the bedrock of drug discovery for specific cancer cell groups. We envision that they will become an indispensable tool for tailoring patient care in the era of heterogeneity-aware precision medicine.

INTRODUCTION

In an ongoing effort to curb cancer morbidity and mortality, computational methods have shown great potential at aiding drug discovery.[1,2] Based on artificial intelligence (AI) principles including both classical machine learning (ML) and deep learning (DL) frameworks, many computational approaches leverage high-throughput drug screens (HTS) on cancer cell lines (CCLs) to infer sensitivities to various drugs in a target dataset.[1] These HTS data, in combination with CCL molecular profilings, offer insights

into connections between CCL expression profiles and drug response phenotypes. Models trained on these data can be applied to target expression datasets (e.g., patient tumors) to predict vulnerabilities to various anti-cancer drugs and further lead off drug selection in specific settings.

In recent years, an increasing volume of studies through single cell RNA-seq (scRNA-seq) analysis suggest heterogeneous tumors, in which cancer cells display temporal and spatial diversity, are causally related to a critical challenge in cancer treatment: disease progression through drug resistance.[3] Within a heterogeneous tumor, various subpopulations of cells, each with distinct genetic and phenotypic traits, can coexist. When exposed to therapeutic interventions, these diverse cell subclones may respond differently, with some intrinsically resistant to treatment. The selective pressure exerted by the therapy can lead to the proliferation of these therapy-resistant cell subclones, ultimately contributing to treatment failure.[4] Recognizing the pivotal role of tumor heterogeneity, this necessitates cell-type aware drug discovery to target problematic cell groups within patient tumors and provide treatment opportunities with clinical impact. In this context, computational frameworks are expected to enable efficient early development by generating drug response profiles at the single-cell (SC) level.[5] Meanwhile, given that 1) large-scale CCL expression profiles are systematically surveyed by bulk RNA-seq which captures an averaged estimation across within-tumor cell subtypes and 2) the fundamental differences in RNA-seq and scRNA-seq techniques, specialized methods are needed to utilize large-scale drug screens of CCLs to infer drug activity at the SC level.

One key embodiment of such methods relates to the successful application of bulk-learned drug-gene relationships to SC datasets.[6] The construct of such methods involves an essential process that converts learned relationships between CCL molecular profiles and drug response into references for scRNA-seq to generate cellular drug sensitivity status. Several transfer learning approaches have been proposed to

facilitate such a process, including data integration through matrix factorization, variational autoencoder networks, and biomarker/signature based frameworks.

Successful prediction of cellular response is the keystone of cell type-aware drug discovery. It will facilitate development of therapeutics for targeting therapy-resistant cells. Furthermore, speciality treatments can be used with standard-of-care (SOC) therapies as a combination to reach all malignant cells and help eliminate heterogeneous tumors. In this review, we present computational methods that are designed to infer drug activity at the SC level. We focus on how inferring cellular drug response facilitates pre-clinical research through enabling hypothesis generation and providing actionable drug candidates. Moreover, we also reason that predicted cellular drug response can be used to design combination therapies to target heterogeneous tumors and help achieve curability.

LITERATURE DATA COLLECTION

Harzing's Publish or Perish software was used to query work with Google Scholar between the years 2020-2023, pulling the first 200 entries whose title contained the keyword 'single cell drug' or 'single cell therapeutic.[7] In Figure 1, a total of 374 entries were identified and subsequently filtered to remove review articles, abstracts, and preprint archives. Duplicate papers were removed as well as papers whose title or abstract did not contain the keyword 'drug response'. From the 20 papers obtained at the end of filtering, we zoomed in and focused on eight publications that described computational drug response prediction at the single-cell level through integration with bulk RNAseq data. A summary of these works is depicted in Figure 2 and include neural network/deep-learning methods (scDEAL[8], SCAD[9]), biomarker or signature based methods (DREEP[10,11], Beyondcell[12], ASGARD[13], and scDr[14]), and traditional machine learning based approaches (CaDRReS-Sc[15], scDRUG[16]).

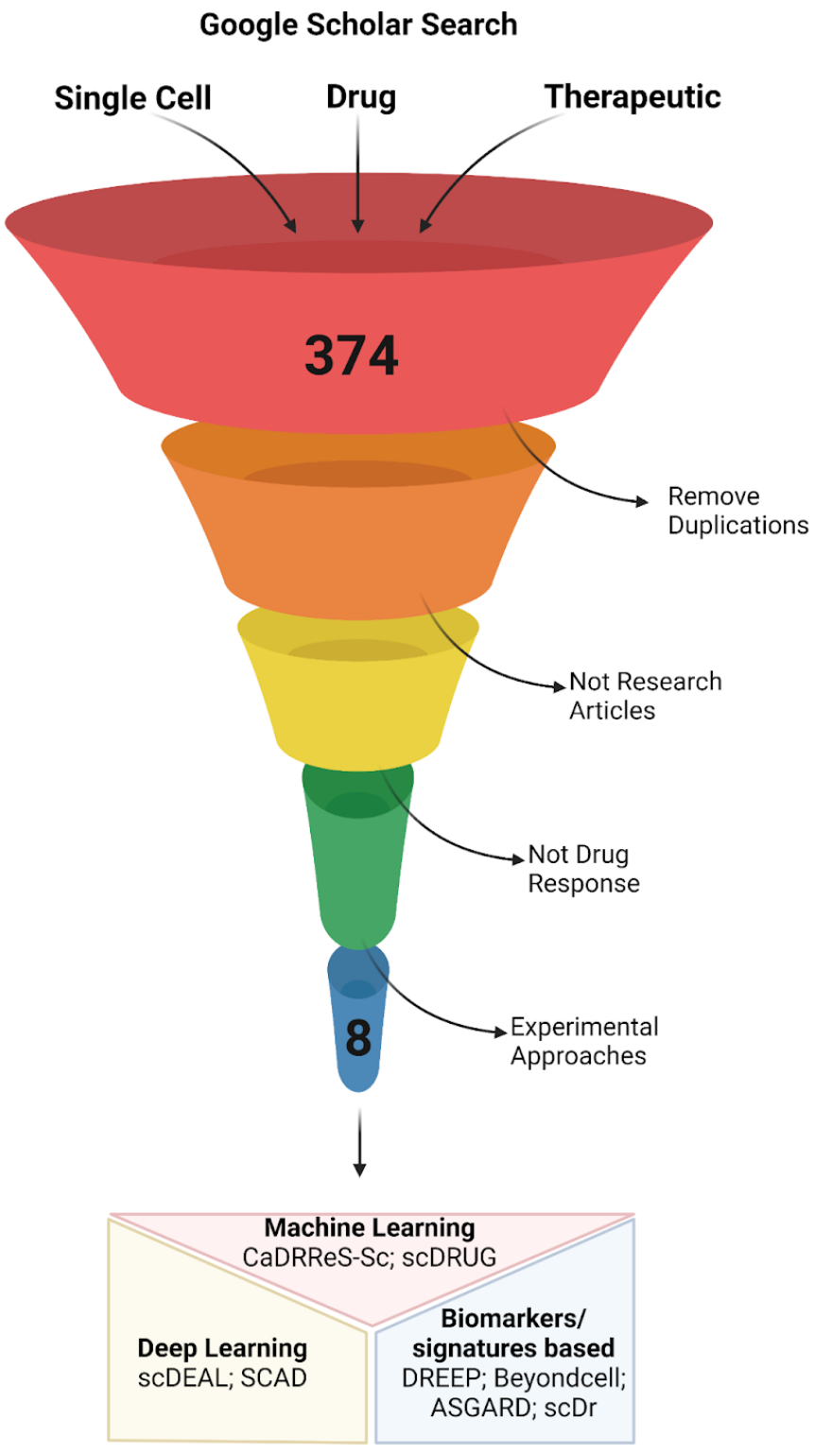


Figure 1. Literature selection workflow

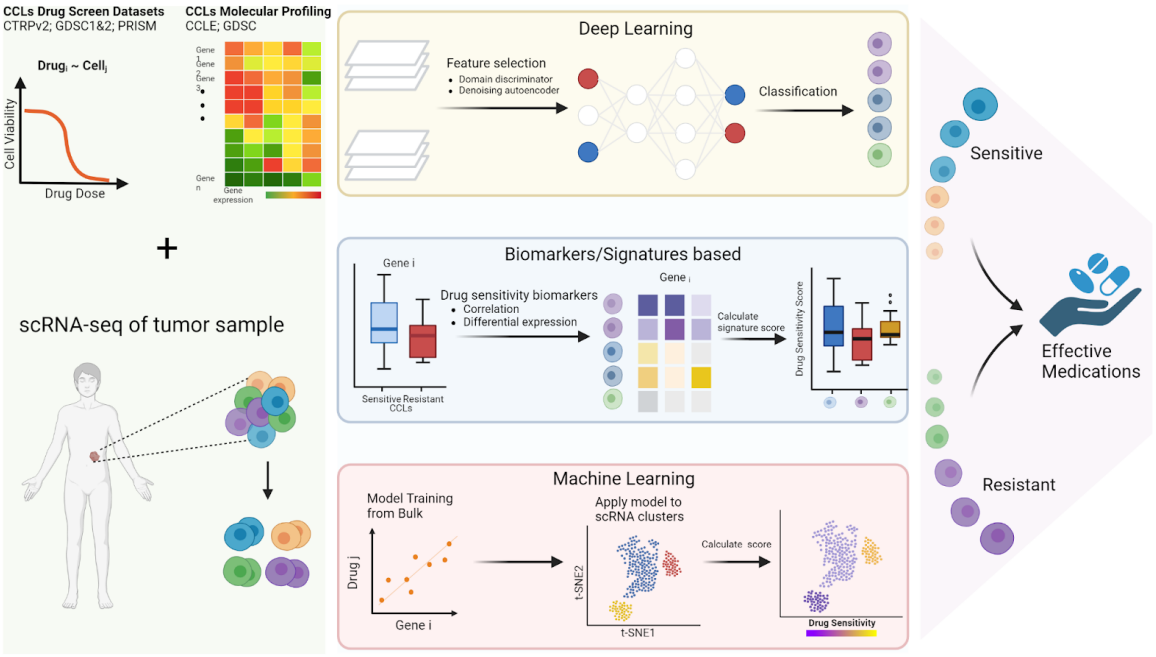


Figure 2. Overview of single-cell drug sensitivity prediction methods

HTS drug screens and CCL molecular data sources

All eight published works reviewed have been developed to infer drug response at the single-cell level using a collection of CCL HTS databases. To date, the largest publicly available HTS screening efforts are the Sanger's Genomics of Drug Sensitivity in Cancer (GDSC)[17], the Broad Institute's Cancer Therapeutics Response Portal (CTRP),[18] and the PRISM Repurposing dataset (PRISM).[19] In addition, the Library of Integrated Network-Based Cellular Signatures (LINCS) consortium has generated encyclopedic profiles of cellular response (termed cellular signatures) under drug perturbations.[20] The cell lines and compounds included in these datasets represent a variety of cancers and molecular targets.[21] The CTRP evaluates 481 drugs on 860 CCLs; the GDSC tests roughly 300 to 400 drugs across two initiatives; the PRISM expands screened drugs to include various non-oncology molecules and covers thousands of treatments.The LINCS data contains CCL perturbation profiles under tens of thousands small molecules.

For CCL transcriptomic data, the Broad Institute's Cancer Cell Line Encyclopedia database (CCLE) provides a comprehensive RNA-seq on approximately 1000 CCLs.[22] In addition, the GDSC platform measures CCL gene expression through microarray probes. CCL molecular features and drug response phenotypes together provide a gateway for unearthing drug-gene relationships and constructing predictive models. To describe drug sensitivity, HTS data include summary statistics in the form of half-maximal inhibitory concentration (IC_{50}) values in GDSC and area-under-the-dose-response-curve (AUC) values for CTRP and PRISM.

SC DRUG RESPONSE PREDICTION METHODS

The aforementioned HTS and CCL data constitutes the majority of data sources used by computational methods to derive drug-gene relationships. A breakdown of specific data used by each SC drug response prediction method is provided in Table 1. Since the CCL transcriptomic profiles were analyzed at the bulk level, where expression of a gene is aggregated over the entire sample, learned drug-gene information cannot be directly applied to scRNA-seq data for meaningful drug response projection. To overcome this, several methods attempt to unify bulk and SC data to maximize similarities between the two sources. On the other hand, some methods seek to identify informative markers from bulk CCL and subsequently apply such markers in SC data to infer cellular drug response. An overview of the key methodologies used in these works are listed in Table 1 and further discussed in details.

Table 1. Key aspects of SC drug response prediction methods.

Original Method	HTS Implemented	Feature and Biomarker Selection	Data Integration Approach	Drug Response Statistics	Drugs Used for Benchmarking
scDEAL https://github.com/OSU-BMBL/scDEAL	GDSC	Selects variable genes. Detects critical genes for drug response.	Neural Networks	Converts AUC to binary labels of sensitive and resistant	Cisplatin Docetaxel Erlotinib Gefitinib I-BET-762
SCAD	GDSC	Extracts invariant features between bulk and SC domains	Neural Networks	Converts IC50 to binary labels	Afatinib AR-42 Cetuximab Gefitinib NVP-TAE684 PLX4720 Sorafenib Vorinostat
CaDDReSS-Sc	GDSC	Adopts a predefined essential gene set	Matrix Factorization	Uses IC50 derived binary labels and refits	Docetaxel Doxorubicin Epothilone B Gefitinib

https://github.com/CSB5/CaDRReS-Sc				drug response curves	Obatoclax Mesylate PHA-793887 PI-103 Vorinostat
Beyondcell https://github.com/cnio-bu/beyondcell	GDSC, CTRP, LINCS	Identifies drug response biomarkers from bulk data	Relies on bulk-based biomarkers.	Calculates a unit-free signature score	321 anticancer drugs from Ben-David et al.[23]
scDR	CTRP	Identifies drug response biomarkers from bulk data	Relies on bulk-based biomarkers	Calculates a unit-free signature score	77 FDA-approved drugs.
DREEP https://github.com/gambalab/DREEP	CTRP	Identifies drug response biomarkers from bulk data	Relies on bulk-based biomarkers	Calculates enrichment scores via GSEA	450 drugs from the CTRP (bulk level).
ASGARD	LINCS	Identifies genes altered	Relies on bulk-based	Calculates a customized	150 drugs from different

https://github.com/lanagarmire/Asgard		by drug perturbations in bulk data	perturbation biomarkers	score based on signature reversion	diseases. Regards a drug as positive if it is: 1) FDA-approved, 2) used in advanced clinical trials, or 3) proven effective in animal models

[a]scDrug adopted the entirety of CaDDReSS-Sc and thus not listed here as an original approach.

Neural Network/ DL methods

A number of deep learning (DL) methods based on neural networks have been proposed to calibrate drug response modeling for more precise predictions.[24-26] Adaptations of these methods for predicting SC drug sensitivities have been discussed and implemented in recent years.[6] Such computational pipelines leverage the abundance of scRNA-seq data as DL algorithms often include many parameters and consume a lot of data for adequate training of these parameters. Aided by flexibility in finding latent space and features, specialized neural networks are designed to minimize distributional discrepancies between the input bulk RNA-seq and scRNA-seq sources, such that drug-gene relationships learned solely from bulk RNA-seq can be meaningfully applied to the target scRNA-seq dataset.

To this end, SCAD adopted an adversarial learning approach[9] training a domain discriminator to counter cross-domain bias between the two data sources. This forces invariant feature extraction across bulk and SC RNA-seq domains in order to integrate the two sources. For drug response learning, bulk sample labels from GDSC were binarized and used to supervise a prediction network which minimizes a binary cross-entropy (BCE) loss and generates predicted labels in scRNA-seq data.[9] While possessing similar functional compartments, scDEAL, on the other hand, employs denoising autoencoders for feature selection from bulk and SC RNA-seq data.[8] Given the nature of generative models, the input dataset is compressed into a "bottleneck" on which data integration is performed. To facilitate transfer learning, a loss function incorporating the probability measurement, maximum mean discrepancy (MMD), is minimized to retain similarity between bulk and SC data. ScDEAL also utilizes binary bulk sample drug response to train a prediction network by minimizing the BCE. Ultimately, both methods assemble multiple specialized DL networks to fulfill the bulk-to-SC drug label prediction. DL models have the potential to concurrently account for heterogeneous scRNA-seq and provide fine-grained cellular drug sensitivity labels thanks to their elasticity and capability at learning complex relationships. However, it is unclear if dichotomization of continuous drug response measurements can always render pharmacological meanings, as response is better characterized as a spectrum.[27] Also, optimal model structures can be drug dependent, and learning such structures often involves intensive computation, which can sometimes mask their practicality among the pharmacological research community for hypothesis generation and early development. Both methods require neural network parameter tuning for each drug to calibrate an optimal structure for transfer learning using measured drug sensitivity. This may consequently limit their feasibility to evaluate against many drugs or conduct large-scale drug screens.

Biomarkers or signatures based methods

While most methods involve recognition of molecules whose activities associate closely with sensitivities to drugs, a few methods center around identification and application of these biomarkers to maximize

their predictability for drug response.[28-30] Genes or gene signature sets are learned from the paired data of CCL bulk RNA-seq and drug response from HTS, frequently through correlation analysis.[31,32] Next, instead of relying on detection of shared expression patterns between bulk and SC data, the same markers are directly selected from the scRNA-seq data and their normalized expression values are coalesced into a scalar describing relative likelihood of a cell having the desired phenotype (e.g., high sensitivity to a drug). As discovery of biomarkers depends only on bulk RNA-seq and bulk sample labels, generating cellular drug response predictions can be done *ad hoc* in a separate manner.

To predict drug response in breast cancer cells, Gambardella et al. developed DREEP, which learns drug specific biomarkers correlated with either sensitive or resistant phenotype from the CTRP and GDSC. At the SC level, the top 250 most expressed genes are compared against a drug's biomarker profile through the Gene Set Enrichment Analysis (GSEA), through which an enrichment score (ES) was calculated for each cell-drug pair. A cell is classified to be sensitive to the drug with the extreme ES.[10,11] Beyondcell derives sensitivity signature sets (SS) using the differential expression (DE) analysis R package limma from the CTRP and GDSC. Additionally, it compares expression levels before and after a drug perturbation from the LINCS dataset to offer perturbation signature sets (PS). With PS, drug response can be inferred based on the "signature reversion" principle, which prioritizes drugs that induce reverse-to-normal expression changes in disease models.[12] A drug's signature is divided to up- or down-regulated sets, each having 250 most significant genes, and applied to the target SC data to calculate a cell specific "Beyondcell Score" (BCS) indicating relative sensitivity to the drug.[12] In scDr, CCLs in CTRP are dichotomized into sensitive and resistant ones. Differential expression analysis is then carried out between the two CCL groups, through which top 200 biomarkers based on their log2 fold change (log2FC) in either up- or down-direction are identified. The log2FC of the marker genes are used in conjunction with gene expression Z-scores in the SC data to generate drug response scores.[14] Also demonstrated in breast cancer scRNA-seq, ASGARD requires the input of both disease samples and normal tissue samples and pairs cell identity clusters between the two types. To identify drugs for a specific cell cluster, DE analysis

is first carried out between normal and disease clusters. DE genes are then used to screen for drugs that significantly reverse expression patterns in the disease cluster to that of normal in the LINCS dataset.[13]

Unlike the deep learning methods discussed, no training data or intense computation is required, as learned gene signatures can be applied to any SC data independently. However, due to low and sparse expression levels in a single cell and the stochastic nature of drop-outs, predictive power of a pre-defined gene set is not always guaranteed. Beyondcell addresses this potential pitfall by penalizing cells with high sparsity in their corresponding signature genes. Depending on the availability of paired disease-normal scRNA-seq from the same cohort, applications of ASGARD can be limited. Furthermore, compared to normal tissues, expression changes in certain advanced cancers are not unidirectional, which will greatly convolute the signature reversion principle.[33]

Machine learning methods

Traditional machine learning approaches have a rich history in the area of drug discovery.[34,35] They have commonly been used to integrate various genomic spaces, including drug-gene interactions, disease-gene interactions, and gene-gene interactions.[34,36,37] Great effort has been taken toward applying these approaches to drug response prediction.[33,38,39] CaDDReSS-Sc is a machine learning framework used for cellular level cancer drug response prediction.[14] It incorporates a set of 1856 essential genes[41] identified through CRISPR screens as encoding components of fundamental pathways. CaDDReSS-Sc is an extension of CaDDReS,[40] calibrated for single-cell transcriptomic profiles. The purpose of the factorization is to learn a latent pharmacogenomic space of 10 dimensions, projecting relationships between cell line gene expression with known drug response information. The dot product between the latent space's cell-line vector and drug vector indicate specific cell-line drug responses and is then used to impute drug response of unseen samples (e.g. patients or cell-lines). Thus, the factorization allows for model training. Unlike CaDDReS, CaDDReSS-Sc computes kernel features using both bulk and single-

cell RNA seq data prior to the model training, as essentially the Pearson correlation coefficient between their per-sample gene expression. scDrug is another cellular level drug response prediction technique based on unsupervised machine learning, leveraging CaDDReSS-Sc coupled with an automated pipeline to cluster scRNAseq data.[16] The resolution selected is associated with the optimal silhouette coefficient or distance between clusters. For each cluster, differentially expressed genes are ranked and cell types are annotated using scMatch.[42] This data is then used together with bulk RNA profiles to predict how each cluster will affect patient survival using CaDRReS-Sc. Unlike the biomarker based methods discussed, the essential genes comprising the signature used for CaDDReSS-Sc and scDrug were originally identified for their essential roles in cellular livelihood and are not compound specific.[41] However, drug-gene associations can encompass a wide variety of genes; therefore, using only essential genes does not guarantee adequate gene-drug relationship information that can be used to infer drug response in an unseen data.

Combination of biomarker/signature and machine learning based methods

scIDUC leverages drug-gene signatures in a machine learning based method to infer cellular level drug response.[43] Prior to integrating bulk and single-cell data, the R package limma is used to identify drug response relevant genes (DRGs) from the known bulk sample phenotype. Then the datasets are integrated utilizing canonical correlation analysis to capture common gene expression patterns across the two datasets, adjusting for discrepancies. Non-negative matrix factorization can also be used to find correlations between these datasets as an alternative approach to integration. Lastly, linear regression is formulated using the integrated training and single-cell datasets. Due to the ease of parameter tuning, scIDUC lacks the intensive computation potentially associated with deep learning methods discussed previously; these methods are highly dependent on the existence of scRNAseq data with measured drug sensitivity to allow for parameter fine-tuning whereas scIDUC can more easily be applied to a large collection of drugs. Unlike the signature based methods, scIDUC does not also require corresponding

normal tissue samples, which are not always readily available, or dichotomizing CCL training data into resistant or sensitive labels, which do not capture the nuances and spectrum of drug response. Additionally, traditional machine learning approaches do not consider drug specific genes and instead utilize an essential gene set that may not be specific enough to accurately project each drug sensitivity. scIDUC's DRGs, on the other hand, are learned directly from the bulk data used to train the model, and the gene sets identified for each drug are most significantly associated with drug response.

PROMISING TRENDS AND POTENTIAL PITFALLS

Facilitating heterogeneity-aware drug discovery

Most of the 8 reviewed works enable projection of cell specific vulnerability to therapeutics, which can be associated with cell identities to facilitate drug nomination for certain cell types. For example, Fustero-Torre et al. used Beyondcell to generate sensitivities of 451HLu human melanoma cells to various drugs. They screened for drugs showing high predicted efficacy among BRAF inhibitor (BRAFi) resistant cell clusters as candidate therapeutics to combat BRAFi resistance in melanoma.[12] With increasing availability of scRNA-seq data from varying diseases, such a strategy can be adapted to facilitate drug discovery targeting specific cell clones in many indications. However, most examples provided by the current works only referenced existing studies to justify reliability of prediction results. To truly evaluate the utility of these methods, experimental analysis of the proposed therapeutics should be carried out in vitro or in vivo.

These methods also model drug response through using either resistant or sensitive labels or a continuous spectrum; they do not attempt to predict drug dosage, which plays a role in improving our understanding of cancer tumor heterogeneity. Specifically, tumor heterogeneity is linked to the emergence of therapy

resistant cell subclones. By predicting and adjusting drug dosages, it may be possible to more effectively target and inhibit these therapy resistant subclones, reducing the likelihood of resistance development. Methods are actively developing now.[44]

In addition, scRNAseq allows for a detailed distinction between normal and malignant cell sub-populations.[45-48] This distinction, in turn, greatly facilitates heterogeneity aware drug discovery and avoids inhibition of normal populations, minimizing toxicity. It is imperative for future research to harness these full capabilities of scRNAseq to optimize treatment outcomes.

Spearheading drug combination efficacy prediction

At an individual level, therapeutic vulnerability profiles within a patient may enable modeling of drug combinations as tailored treatments to combat heterogeneous tumors. First, identified drugs for therapy-resistant cell groups can be used with SOC treatments to form drug combinations to help eliminate a tumor. This is a direct extension of cell specific drug discovery. Moreover, predictions of drug combination efficacy may be achieved using cellular drug response. Combination efficacy can be inferred probabilistically by assuming predicted cellular drug response indicates likelihood of cell kill.[49] In this case, methods including continuous variables for drug response show advantages over those using only binary labels. For example, AUCs can be scaled to indicate percentages of cell death under a treatment or probabilities of cell kill; for a combination with multiple agents, their cell kill probabilities can be aggregated to generate a probability of cell kill for the whole population. This can be done at the cellular level or at the cell cluster level. An averaged combination probability therefore estimates efficacy over the whole heterogeneous population. However, it is unclear if current drug combination prediction rationales grant desirable predictive powers, especially given the discrepancy between complex intratumoral structures and information reflected by current scRNA-seq techniques.

Data availability as a limiting factor

The key limiting factor against predicting drug response at the single-cell level is insufficient training power due to the lack of public benchmarked data. Understanding how to appropriately integrate bulk and scRNA-seq data alleviates this limitation. Ultimately, it will enable the design of more precise therapeutic regimens, taking into account a patient's specific microenvironment and tumor heterogeneity. A pitfall of this approach, however, is the ability to impute cellular level drug response is contingent on the type and quality of bulk CCL data used in the integration. Specifically, the techniques covered in this review largely leverage response across chemotherapeutics, and both chemotherapy and immunotherapy may be used alone, together ('chemoimmunotherapy'), or in combination with other treatments (e.g. radiation therapy or surgery), which these approaches can't generate predictions for yet.

Conclusion

We carefully review eight latest existing methods that leverage HTS data to project cell-level drug sensitivity in given scRNA-seq data. They employ a variety of computational principles including deep learning frameworks, more traditional machine learning based approaches, as well as biomarkers. These methods center around techniques for transferring bulk-learned information into SC prediction anchors, directly or indirectly. Applications of these methods demonstrate their utility at generating and testing hypotheses for heterogeneity-aware drug discovery. Depending on specific research questions and biological models, different methods might be preferred for hypothesis generation. Eventually, application of these methods will help reduce occurrence of drug resistance, cancer relapse, and potentially lead to complete tumor regression.

REFERENCES

---

1. Adam, G. *et al.* Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precis. Onc.* 4, 19 (2020).
2. Jarada, T. N., Rokne, J. G. & Alhajj, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Cheminform* 12, 46 (2020).
3. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol* 15, 81–94 (2018).
4. Brady, S. W. *et al.* Combating subclonal evolution of resistant cancer phenotypes. *Nat Commun* 8, 1231 (2017).
5. Qi, R. & Zou, Q. Trends and Potential of Machine Learning and Deep Learning in Drug Study at Single-Cell Level. *Research* 6, 0050 (2023).
6. Wu, Z. *et al.* Single-Cell Techniques and Deep Learning in Predicting Drug Response. *Trends in Pharmacological Sciences* 41, 1050–1065 (2020).
7. *Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish*
8. *Chen, J. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat Commun* 13, 6494 (2022).

scDEAL is a deep transfer learning approach based on a neural network architecture to predict cancer drug response through integrating bulk and single-cell RNA sequencing (RNA-seq) data. It predicts single-cell drug response based on a trained model and the single-cell RNA-seq data, utilizing the maximum mean discrepancy as a loss function to produce binary drug response labels. Results from the application of scDEAL to six drug-treated single-cell data with experimental validation indicated

that scDEAL is robust. The available scDEAL Python package currently only supports the datasets appeared in Chen et al.

9. Zheng, Z. *et al.* Enabling Single-Cell Drug Response Annotations from Bulk RNA-Seq Using SCAD. *Advanced Science* 10, 2204113 (2023).
10. Gambardella, G. *et al.* A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. *Nat Commun* 13, 1714 (2022).
11. Pellecchia, S., Viscido, G., Franchini, M. & Gambardella, G. *Predicting drug response from single-cell expression profiles of tumours*. http://biorxiv.org/lookup/doi/10.1101/2023.06.01.543212 (2023) doi:10.1101/2023.06.01.543212.
12. *Fustero-Torre, C. *et al.* Beyondcell: targeting cancer therapeutic heterogeneity in single-cell RNA-seq data. *Genome Med* 13, 187 (2021).

Beyondcell is a method aimed to predict cellular drug response from single-cell RNA sequencing data. It generates a Beyondcell enrichment score (BCS) for each drug-cell pair, ranging from 0-1, and it is indicative of the activity of a gene-drug signature in a given gene expression matrix. This score also measures how susceptible each single-cell is to a drug under scrutiny where a high BCS represents high concordance between the gene signature and the single-cell analyzed. This method was applied to a breast cancer single-cell dataset where a BCS was computed for a panel of drugs. It was able to identify distinct drug-response cell subpopulations before and after bortezomib treatment, drug-resistant cellular populations, and single-cell variability in drug response across cancer patients. Through deconvolving tumor heterogeneity, it was also able to propose drug treatments and therefore can help design more precise treatment regimens.

13. *He, B. *et al.* ASGARD is A Single-cell Guided Pipeline to Aid Repurposing of Drugs. *Nat Commun* 14, 993 (2023).

ASGARD is a method for imputing single-cell drug sensitivities through identifying drugs that can reverse single-cell gene expression data from a diseased state to a normal state. Thus, it requires both diseased and normal single-cell data from the same subjects as input data. Drug efficacy is measured by a drug score at the individual patient level and takes into account the significance of the reversed differential gene expression pattern between diseased and normal samples. ASGARD was compared to other drug repurposing methods using bulk RNA-seq samples by summarizing single-cell RNA sequencing (RNA-seq) data into pseudo-bulk RNA-seq data and was found to predict drugs more accurately. It was also found to have robust performance across different single-cell populations with clinical validation. In short, ASGARD demonstrated itself to be a promising drug recommendation pipeline.

14. Lei, W. *et al.* scDR: Predicting Drug Response at Single-Cell Resolution. *Genes* 14, 268 (2023).
15. *Suphavilai, C. *et al.* Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Med* 13, 189 (2021).

Through leveraging a recommender system and information across drugs, CaDRReS-Sc can predict drug accuracy across single-cell RNA sequencing (RNA-seq) data with high accuracy (80%). Thus, it can capture transcriptomic heterogeneity, predicting drug response in unseen cell types. Its latent pharmacogenomic model also eases visualization and interpretation in order to examine key drug pathways. This method was extended to combinations of drugs where drug pairs identified were more effective than individual drugs in vitro.

16. Hsieh, C.-Y. *et al.* scDrug: From single-cell RNA-seq to drug response prediction. *Computational and Structural Biotechnology Journal* 21, 150–157 (2023).
17. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research* 41, D955–D961 (2012).

18. Seashore-Ludlow, B. *et al.* Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery* 5, 1210–1223 (2015).

19. Yu, C. *et al.* High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol* 34, 419–423 (2016).

20. Keenan, A. B. *et al.* The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Systems* 6, 13–24 (2018).

21. Ling, A., Gruener, R. F., Fessler, J. & Huang, R. S. More than fishing for a cure: The promises and pitfalls of high throughput cancer cell line screens. *Pharmacology & Therapeutics* 191, 178–189 (2018).

22. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature* 483, 603–607 (2012).

23. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* 560, 325–330 (2018).

24. Ji, Y., Lotfollahi, M., Wolf, F. A. & Theis, F. J. Machine learning for perturbational single-cell omics. *Cell Systems* 12, 522–537 (2021).

25. Zhang, H., Chen, Y. & Li, F. Predicting Anticancer Drug Response With Deep Learning Constrained by Signaling Pathways. *Front. Bioinform.* 1, 639349 (2021).

26. Jia, P. *et al.* Deep generative neural network for accurate drug response imputation. *Nat Commun* 12, 1740 (2021).

27. Bouhaddou, M. *et al.* Drug response consistency in CCLE and CGP. *Nature* 540, E9–E10 (2016).

28. Ben-Hamo, R. *et al.* Predicting and affecting response to cancer therapy based on pathway-level biomarkers. *Nat Commun* 11, 3296 (2020).

29. Miranda, S. P., Baião, F. A., Fleck, J. L. & Piccolo, S. R. Predicting drug sensitivity of cancer cells based on DNA methylation levels. *PLoS ONE* 16, e0238757 (2021).

30. Chen, Y. *et al.* Response prediction biomarkers and drug combinations of PARP inhibitors in prostate cancer. *Acta Pharmacol Sin* 42, 1970–1980 (2021).

31. Lee, J. H., Park, Y. R., Jung, M. & Lim, S. G. Gene regulatory network analysis with drug sensitivity reveals synergistic effects of combinatory chemotherapy in gastric cancer. *Sci Rep* 10, 3932 (2020).

32. Rees, M. G. *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 12, 109–116 (2016).

33. Koudijs, K. K. M., Böhringer, S. & Guchelaar, H.-J. Validation of transcriptome signature reversion for drug repurposing in oncology. *Briefings in Bioinformatics* 24, bbac490 (2023).

34. Dai, W. *et al.* Matrix Factorization-Based Prediction of Novel Drug Indications by Integrating Genomic Space. *Computational and Mathematical Methods in Medicine* 2015, 1–9 (2015).

35. Poleksic, A. Hyperbolic matrix factorization improves prediction of drug-target associations. *Sci Rep* 13, 959 (2023).

36. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16, 1289–1296 (2019).

37. Hsu, L. L. & Culhane, A. C. Impact of Data Preprocessing on Integrative Matrix Factorization of Single Cell Data. *Front. Oncol.* 10, 973 (2020).

38. Ammad-ud-din, M. *et al.* Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* 32, i455–i463 (2016).

39. Wang, L., Li, X., Zhang, L. & Gao, Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* 17, 513 (2017).

40. Suphavilai, C., Bertrand, D. & Nagarajan, N. Predicting Cancer Drug Response using a Recommender System. *Bioinformatics* 34, 3907–3914 (2018).

41. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* 350, 1096–1101 (2015).

42. Hou, R., Denisenko, E. & Forrest, A. R. R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* 35, 4688–4695 (2019).

43. Zhang, W. *et al. Inferring therapeutic vulnerability within tumors through integration of pan-cancer cell line and single-cell transcriptomic profiles*. http://biorxiv.org/lookup/doi/10.1101/2023.10.29.564598 (2023) doi:10.1101/2023.10.29.564598.

44. Ianevski, A. *et al. Single-cell transcriptomes identify patient-tailored therapies for selective co-inhibition of cancer clones*. http://biorxiv.org/lookup/doi/10.1101/2023.06.26.546571 (2023) doi:10.1101/2023.06.26.546571.

45. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol* **39**, 599–608 (2021).

46. Van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281.e24 (2019).

47. Sh, Y. *et al.* CaSee: A lightning transfer-learning model directly used to discriminate cancer/normal cells from scRNA-seq. *Oncogene* **41**, 4866–4876 (2022).

48. Nofech-Mozes, I., Soave, D., Awadalla, P. & Abelson, S. Pan-cancer classification of single cells in the tumour microenvironment. *Nat Commun* **14**, 1615 (2023).

49. Pomeroy, A. E., Schmidt, E. V., Sorger, P. K. & Palmer, A. C. Drug independence and the curability of cancer by combination chemotherapy. *Trends in Cancer* 8, 915–929 (2022).

CHAPTER 6: CONCLUSION

SUMMARY

Overall, the pharmacogenomic methods described in this body of work aim to help improve our ability to achieve personalized cancer treatment; to tailor treatments to the individual, to increase the efficacy of therapies, and to reduce the likelihood of adverse drug reactions. In turn, this may help empower patients with more information to make informed decisions about their health, reduce the burden of healthcare costs, and give patients the quality of life they deserve. While *oncoPredict* has demonstrated its ability to identify a range of promising drugs with diverse mechanisms in the context of GBM, its capabilities can be enhanced through utilizing Bayesian networks. Through inferring connections between patient gene expression and drug response we can pinpoint biomarkers. In turn, these biomarkers can help select appropriate patient populations for further evaluations and tailor treatment. When applied to five independent patient cohorts and a mouse avatar model, encompassing nearly 1,000 GBM samples, a causal relationship between the expression levels of *PHGDH* gene and the efficacy of MEKis were observed and experimentally confirmed. Specifically, reduced expression of this gene increased tumor sensitivity to MEKis while increased expression led to resistance. While *oncoPredict* and Bayesian networks are one means of combating the heterogeneity of GBM to identify a more personalized treatment approach, *scIDUC* has also demonstrated its superior ability to enable drug development targeting heterogeneous tumors. Not only did *scIDUC* recapitulate resistance to the standard of care therapy (SN-38) typically ascribed to RMS (rhabdomyosarcoma), it provided drug nomination highly in line with drugs that have shown potential in previous studies. Additionally, because cells from each RMS sample underwent sequencing altogether and therefore precluded batch effects, *scIDUC* predictions are

not hindered by potential existence of these technical effects. When *scIDUC* was applied to drugs included in a drug screen panel that compared drug responses between pancreatic cancer cells grown across differing media conditions, *scIDUC* was highly consistent with the original drug screen panel, which means scIDUC can capture tumor microenvironment influences on drug response. *scIDUC* was also able to identify experimentally supported drug nominations against CRPC (castration resistant prostate cancer) docetaxel resistant cells. Therefore, *scIDUC* is potentially eligible for helping to identify new therapeutic targets and improve drug discovery for GBM patients as well as other challenging cancer types.

FUTURE DIRECTIONS

Our future work can be focused on several key areas. First, application of *scIDUC* to GBM data. Ultimately, the goal is to predict drug combinations to be tested against GBM. While *scIDUC* does not directly predict combination therapy yet, its ability to infer cellular level drug response can enable drug combination nomination, and this is currently a work in progress. Monotherapy or the use of a single drug or therapeutic method, has generally not been sufficient for the effective treatment of GBM due to several reasons such as its aggressive and heterogenous nature, and complex microenvironment (its blood supply, immune evasion mechanisms, and interaction with normal brain tissue).[1-3] This makes them less likely to respond to monotherapy and more likely to respond to combination therapy. While the methodology described in Chapter 3 aims to counter these challenges through predicting therapy for specific patient populations who are particularly susceptible to that therapy, additional investigation through application of *scIDUC* is warranted. Both potential new monotherapies guided by biomarker screening and an understanding of cellular level drug sensitivity for eventual combination therapy prediction is important for future investigation. In an effort to extend the GBM project, we have explored databases including the

Human Tumor Atlas (HTAN)[4] for prospective GBM scRNA-seq and cellular level drug response data without success. We have, however, come across a unique opportunity to apply *scIDUC* to small cell lung cancer (SCLC) data from HTAN, with known vitality status outcomes and pre-treatment scRNA-seq data.

Like GBM, SCLC is also known for its aggressiveness, heterogeneity, and poor prognosis.[5] Therefore, this presents an exciting opportunity to dive deeper into tumor heterogeneity, expanding *scIDUC* to other cancer types and data to validate and enhance the generalizability of our method, and continuously improve the algorithm such that it can directly infer drug combinations. To explore potential application of *scIDUC* to this data, we imputed cisplatin and etoposide response and compared response across the known vitality status of five patients. Both therapies are commonly used in combination and serve as a common treatment regimen for SCLC.[6] For comparison purposes, we also imputed response to EGFR inhibitors, gefitinib and afatinib, which are not the primary choice for SCLC due to the rarity of EGFR mutation, as SCLC rarely harbors the types of mutations that would respond well to these inhibitors. We can hypothesize that *scIDUC* can detect these relationships, and Figure 1 shows this to be true. Not only do both cisplatin and etoposide imputed drug response scores differ significantly between vitality status (p-value<0.001), but the effect size as measured by Cohen's D ranges from 0.53-0.61 and the Rho ranges from 62-63%. This suggests a difference in drug response between the two groups is noticeable where the samples in the ‘Alive’ group were predicted to have a greater sensitivity to treatment. This was not the case for the EGFR inhibitors, whose Cohen’s D was reversed and Rho<50%, indicating the samples in the ‘Alive’ group would not have responded well to standard treatment. Since the same trends are not seen across drug imputations, this helps confirm that the probability of batch effect influencing the differences in drug response is low.

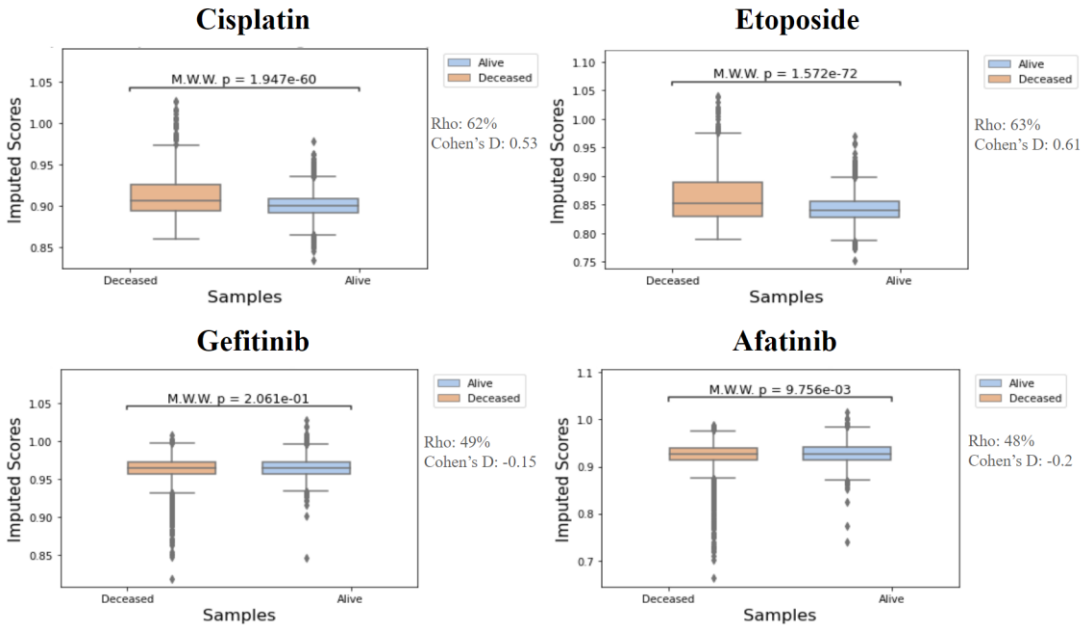


Figure 1. Application of *scIDUC* to Lung Cancer Data to Impute Cisplatin and Etoposide Drug Response. scRNA-seq data was collected from 5 SCLC patients previously untreated where vitality status is known. Cellular level drug response was imputed using *scIDUC* for treatment (cisplatin and etoposide) hypothesized to show efficacy in the 'Alive' vitality group and for treatment (EGFR inhibitors) hypothesized to show reverse trends. These hypotheses were confirmed, where p-value<0.001, Rho>60%, and Cohen's D>0.5 for both cisplatin and etoposide, indicating the samples in the 'Alive' group would respond well to standard of care treatment but not to EGFR inhibitors. These metrics derive from the Mann-Whitney U test.

Another future direction we are currently pursuing is enabling *scIDUC* to fulfill its intended purpose: to predict drug combinations using its cellular level drug response prediction data. Tumors can consist of a diverse mix of cancer cell types, each with its unique genetic and molecular profile, and differences in vulnerabilities. By analyzing tumor samples, we can identify distinct subpopulations of cancer cells within a tumor and infer drugs that will best target those populations. Based on a comprehensive analysis,

a combination of drugs can be chosen that targets the major subpopulations. Unfortunately, there are several challenges with this approach. Firstly, a given tumor sample may, depending on the resolution, consist of 2 or 10 cell types, for example. Obviously, identifying more than 2 or 3 drugs in a combination is not practical, in which case, we may likely have to cluster the 10 or so cell types into 2 or 3 broad yet similar clusters. Secondly, comparing the area under the dose response curve (AUC score) for multiple drugs is not possible, as these values may likely derive from tests using different dosing ranges for each drug. As such, a number of challenges must be addressed before direct combination imputation from scRNA-seq data alone is possible. Additionally, the concept behind the approach described here is more based on independent drug action, where the effects of multiple drugs in a combination are independent of each other.[7] In this model, each drug in the combination exerts its effects without influencing or being influenced by the action of the other drugs in the mixture. This approach simplifies the analysis by assuming no interaction between the drugs. That being said, it is not always applicable. The interactions between different drugs and cancer cell types can be complex, and unexpected resistance, toxicity, or synergistic effects may occur, which independent drug action doesn't take into account. Fortunately, we were recently presented with a pediatric acute myeloid leukemia (AML) scRNA-seq dataset, where cellular level drug response is known as well as cell phenotype (normal or tumor) and cell type. This provides us with an exciting opportunity to leverage this dataset for testing *scIDUC*'s ability to directly infer drug combinations and help pave the way to a transparent, efficient, and accurate computational method for directly predicting combination drugs and personalized treatment plans.

REFERENCES

---

1. Becker, A., Sells, B., Haque, S. & Chakravarti, A. Tumor Heterogeneity in Glioblastomas: From Light Microscopy to Molecular Pathology. Cancers 13, 761 (2021).

2. Dymova, M. A., Kuligina, E. V. & Richter, V. A. Molecular Mechanisms of Drug Resistance in Glioblastoma. IJMS 22, 6385 (2021).
3. McBain, C. et al. Treatment options for progression or recurrence of glioblastoma: a network meta-analysis. Cochrane Database of Systematic Reviews 2021, (2021).
4. Rozenblatt-Rosen, O. et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell 181, 236–249 (2020).
5. Shue, Y. T., Lim, J. S. & Sage, J. Tumor heterogeneity in small cell lung cancer defined and investigated in pre-clinical mouse models. Transl. Lung Cancer Res. 7, 21–31 (2018).
6. Jiang, S., Huang, L., Zhen, H. *et al.* Carboplatin versus cisplatin in combination with etoposide in the first-line treatment of small cell lung cancer: a pooled analysis. *BMC Cancer* 21, 1308 (2021). https://doi.org/10.1186/s12885-021-09034-6
7. Plana, D., Palmer, A. C. & Sorger, P. K. Independent Drug Action in Combination Therapy: Implications for Precision Oncology. Cancer Discovery 12, 606–624 (2022).