

Method

Marker-free characterization of full-length transcriptomes of single live circulating tumor cells

Sarita Poonia,¹ Anurag Goel,^{2,3} Smriti Chawla,¹ Namrata Bhattacharya,² Priyadarshini Rai,¹ Yi Fang Lee,^{4,11} Yoon Sim Yap,⁵ Jay West,^{6,12} Ali Asgar Bhagat,^{4,13,14} Juhi Tayal,⁷ Anurag Mehta,⁸ Gaurav Ahuja,¹ Angshul Majumdar,^{2,9,10} Naveen Ramalingam,⁶ and Debarka Sengupta^{1,2,9}

¹Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India; ²Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India; ³Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India; ⁴Biolidics Limited, Singapore 118257, Singapore; ⁵National Cancer Centre Singapore, Singapore 169610, Singapore; ⁶Fluidigm Corporation, South San Francisco, California 94080, USA; ⁷Department of Research, ⁸Department of Laboratory Services and Molecular Diagnostics, Rajiv Gandhi Cancer Institute and Research Centre-Delhi (RGCIRC-Delhi), New Delhi 110085, India; ⁹Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India; ¹⁰Department of Electronics & Communications Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), New Delhi 110020, India

The identification and characterization of circulating tumor cells (CTCs) are important for gaining insights into the biology of metastatic cancers, monitoring disease progression, and medical management of the disease. The limiting factor in the enrichment of purified CTC populations is their sparse availability, heterogeneity, and altered phenotypes relative to the primary tumor. Intensive research both at the technical and molecular fronts led to the development of assays that ease CTC detection and identification from peripheral blood. Most CTC detection methods based on single-cell RNA sequencing (scRNA-seq) use a mix of size selection, marker-based white blood cell (WBC) depletion, and antibodies targeting tumor-associated antigens. However, the majority of these methods either miss out on atypical CTCs or suffer from WBC contamination. We present unCTC, an R package for unbiased identification and characterization of CTCs from single-cell transcriptomic data. unCTC features many standard and novel computational and statistical modules for various analyses. These include a novel method of scRNA-seq clustering, named deep dictionary learning using *k*-means clustering cost (DDLK), expression-based copy number variation (CNV) inference, and combinatorial, marker-based verification of the malignant phenotypes. DDLK enables robust segregation of CTCs and WBCs in the pathway space, as opposed to the gene expression space. We validated the utility of unCTC on scRNA-seq profiles of breast CTCs from six patients, captured and profiled using an integrated ClearCell FX and Polaris workflow that works by the principles of size-based separation of CTCs and marker-based WBC depletion.

[Supplemental material is available for this article.]

Cancer ranks as a prime reason for death and a vital barrier to longer life expectancy in every country of the world (Sung et al. 2021). According to World Health Organization (WHO) estimates, in 2019 (Mathers 2020) among 183 countries, cancer ranked as the first or second cause of death of people below the age of 70 yr and ranked third or fourth in 23 countries (Sung et al. 2021). The primary reason for 90% of cancer-related deaths is metastasis (Bittner et al. 2020), the process in which the cancer cells detach from the primary tumor, enter into the circulation, and eventually

colonize distant organs, causing the spread of disease (Krebs et al. 2014; Siegel et al. 2015). To metastasize, cancer cells secrete chemokines to attract immune cells (McAllister and Weinberg 2014; Liu and Cao 2016), facilitating tumor proliferation and intravasation (Gajewski et al. 2013; Kitamura 2018). After cancer cells enter the bloodstream, they are subjected to various stressors, including the lack of cell-cell and cell-matrix adhesion, shear pressures, and immune response. Despite this, a few cancer cells make it through the tortuous journey and leave the vasculature to a secondary site (Shenoy and Lu 2016; Follain et al. 2018).

Circulating tumor cells (CTCs) have recently attracted a lot of attention owing to their critical role in tumor metastasis. Around 40% to 80% of patients with metastatic breast cancers have been found to have CTCs in their blood (Kwa and Esteva 2018). The

Present addresses: ¹¹Thermo Fisher Scientific, Singapore 739256, Singapore; ¹²BioSkrby Corporation, Durham, NC 27701, USA; ¹³Department of Biomedical Engineering, Faculty of Engineering, National University of Singapore, Singapore 117575, Singapore; ¹⁴Institute for Health Innovation and Technology (iHealthtech), National University of Singapore, Singapore 117599, Singapore
Corresponding authors: debarka@iiitd.ac.in, naveen.ramalingam@fluidigm.com, angshul@iiitd.ac.in
Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.276600.122>.

© 2023 Poonia et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <https://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

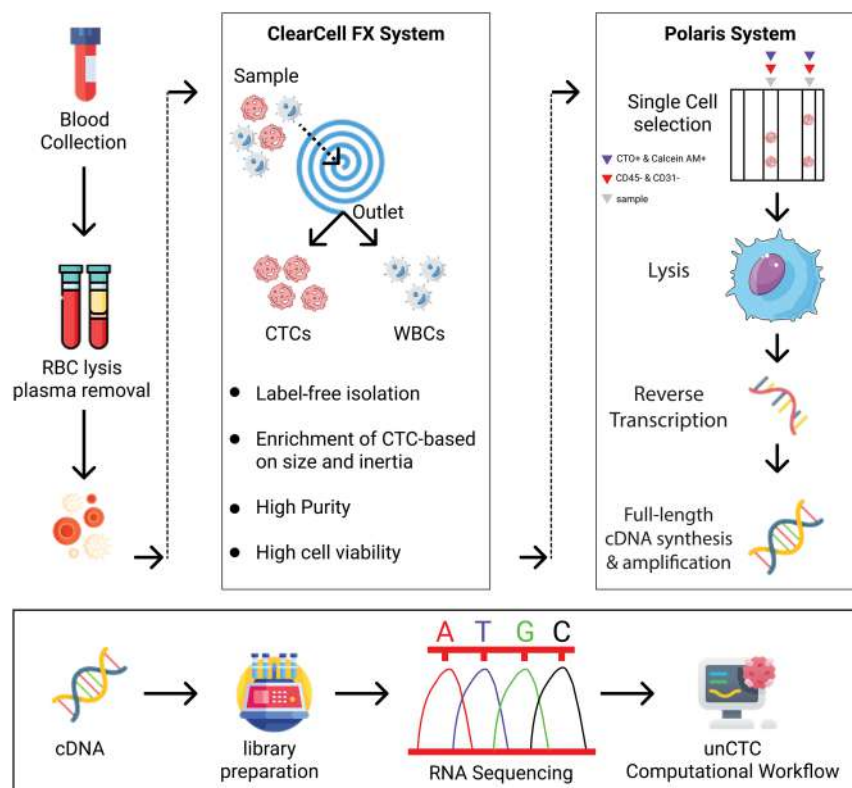


Figure 2. ClearCell FX and Polaris workflow for marker-free enrichment of CTCs. The schematic diagram depicts the key steps involved in the capture and isolation of CTCs using a two-pronged system. ClearCell FX uses a spiral chip to size-sort CTCs. Polaris performs single-cell capture and cDNA synthesis of potential CTCs after depletion of cells that are PTPRC/CD31-positive. Finally, cDNA thus received is subjected to library preparation and RNA sequencing.

unCTC recognizes CTCs selected by the ClearCell FX and Polaris workflow

We have recently demonstrated CTC characterization using supervised machine learning methods (Iyer et al. 2020). Given the dynamic nature of CTC phenotypes, it is however useful to characterize single CTC transcriptomes by unsupervised means. Further, classification-based characterization approaches are fallible in scenarios in which the obtained CTCs are of atypical phenotypes. unCTC alleviates this shortcoming by bringing to bear a spectrum of unbiased single-cell characterization tools. As an extended validation, we subjected the 72 filtered single-cell transcriptomes associated with potential CTCs captured by the ClearCell FX and Polaris workflow. These come from a total of six women entailing three major subtypes of breast cancer: ER⁻/PR⁻/HER2⁻, ER⁺/PR⁺/HER2⁻, and ER⁻/PR⁻/HER2⁺. As a control, we also considered the CTC data set published by Ebright et al. (2020) that comprises 824 cells from 45 patients with breast cancer of the ER⁺/PR⁺/HER2⁻ subtype. For WBCs, we considered 752 scRNA-seq profiles processed in two distinct runs using the Smart-seq2 protocol (Supplemental Table S3; Ding et al. 2020). This integrative analysis task is referred to as Study 2. Figure 5 summarizes the performance of all four methods, Vanilla Seurat, fastMNN, Harmony, and unCTC, in terms of their ability to segregate WBCs and CTCs. We found a mixture of CTCs and WBCs across most clusters with fastMNN and Harmony (Fig. 5D–I). Vanilla Seurat identified a number of clusters with CTCs alone;

however, the embeddings appear to be confounded by batch effects (Fig. 5A–C). Clusters returned by unCTC grouped the CTCs into three clusters, whereas the WBCs were clumped into one large cluster (Fig. 5J–L). Notably, we found ClearCell FX and Polaris selected CTCs clustered with one of the ER⁺ subgroups sourced from the study by Ebright et al. (2020). Supplemental Figure S2 depicts the PCA-, MNN-, and Harmony-based visualizations of the clusters, detected by all four methods. Out of the 72 CTCs that finally qualified the filtering criteria (obtained from ClearCell FX and Polaris workflow), ER⁺ cells were most prevalent (54 out of 72). Among the rest, there were seven and 11 cells of the HER2⁺ and triple-negative categories, respectively. One possible reason for not detecting HER2⁺ and triple-negative CTCs as separate categories is their inadequate numbers, which makes it difficult for unCTC to retain relevant genes/pathways through several upstream filtering steps such as gene filtering and pathway selection.

Marker-dependent characterization of CTC clusters

Cell lineages are best understood through the enrichment of well-characterized lineage markers. Two approaches can be adopted for this: investigating differential expression for single markers and for marker panels. For the second study (comprising ClearCell FX and Polaris), we analyzed lineage identities for clusters identified by DDLK (Fig. 6A). Multiple well-known immune cell markers were spotted among the top 200 differentially up-regulated genes (Supplemental Table S5) among cells in cluster 0 that predominantly contains WBCs from the Ding et al. data set. These are *NKG7*, *PTPRC*, *PTPRCAP*, *IL32*, *CD74*, and *CD48*. The remaining clusters (clusters 1, 2, 3) comprise mostly CTCs (from the Poonia et al. and Ebright et al. data sets).

Cluster 1 among these is found to have elevated expression levels of integrins (*ITGA2B* and *ITGB5*). Integrins are principal adhesion molecules and play a central role in platelet function and hemostasis. Recent studies have postulated CTC–platelet interaction based on RNA extracts of single and clustered CTCs (Ting et al. 2014; Szczerba et al. 2019; Aceto 2020). CTCs constantly interact with factors in the blood such as platelets, circulating nucleic acids, and extracellular vesicles, which influence their molecular profiles (Ward et al. 2021). We observed elevated expression levels of the platelet degranulation markers *CLU* and *SPARC*, which are known for regulating *PF4* (Beck et al. 2019), a critical endocrine factor previously described to be associated with worse outcomes in patients with lung cancer (Pucci et al. 2016). *PF4* was also found to have elevated expression in cluster 1-specific cells. Cluster 1-specific CTCs showed elevated expression of numerous oncogenes with well-known roles in breast cancer progression. *CDKN1A* (Koch et al. 2020), *TIMP1* (Abreu et al. 2020), and *PGRMC1* (Clark et al. 2016) are notable among these. Cluster 2 that harbored

The popular way to express k -means clustering is via the following formulation:

$$\sum_{i=1}^k \sum_{j=1}^n h_{ij} \|z_j - \mu_i\|_2^2 \quad (1)$$

$$h_{ij} = 1, \quad \text{if } x_j \in \text{Cluster } i$$

$$h_{ij} = 0, \quad \text{otherwise}$$

where z_j denotes the j th sample and μ_i the i th cluster.

An alternate formulation for k -means is via matrix factorization (Bauckhage 2015):

$$\|Z - ZH^T(HH^T)^{-1}H\|_F^2, \quad (2)$$

where Z is the data matrix formed by stacking z_j 's as columns, and H is the matrix of binary indicator variables h_{ij} . We prefer expressing k -means as Equation 2 in this work.

Because DDL is not a popular framework, we review it briefly. Dictionary learning (Tošić and Frossard 2011) learns a basis (D) such that the data (X) can be generated/synthesized from the coefficients (Z):

$$X = DZ. \quad (3)$$

The term dictionary learning is relatively new. The same problem has been known as matrix factorization in the past. One can see that Equation 3 is factoring the data matrix X into D and Z . In its most basic form, dictionary learning/matrix factorization is solved via the following:

$$\min_{D,Z} \|X - DZ\|_F^2, \quad (4)$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm defined as the sum of the squares of all the terms in the matrix.

In DDL, instead of learning one layer of the dictionary, multiple layers are learned instead. This is expressed as

$$X = D_3\phi(D_2\phi(D_1Z)). \quad (5)$$

Here D_1, D_2, D_3 are three layers of dictionaries, and ϕ is the activation function between two layers. It is shown for three layers as an example; it can be more than three.

The solution to the unsupervised formulation is expressed as follows:

$$\min_{D_1,D_2,D_3,Z} \|X - D_3\phi(D_2\phi(D_1Z))\|_F^2. \quad (6)$$

In Equation 9, a greedy solution to Equation 6 was proposed. This was not optimal in the sense that there was feedback from shallow to deeper layers but not vice versa. To overcome this, the joint solution was proposed by Singhal and Majumdar (2018) based on the majorization minimization (MM) approach.

In this work, we will use the ReLU activation function for two reasons: (1) it is easier to incorporate as an optimization constraint, and (2) ReLU has been proven to have better function approximation capabilities. Therefore, our basic framework for DDL (with ReLU) will be expressed as follows:

$$\min_{D_1,D_2,D_3,Z,H} \|X - D_1D_2D_3Z\|_F^2 \quad \text{s.t. } D_2D_3 \geq 0, D_3Z \geq 0, Z \geq 0. \quad (7)$$

ReLU activation

We propose to incorporate the k -means cost (Eq. 2) into the DDL formulation (Eq. 7). The basic idea is to use the features generated by DDL as inputs for clustering. However, instead of solving it

piecemeal, we jointly optimize the following cost function:

$$\min_{D_1,D_2,D_3,Z,H} \|X - D_1D_2D_3Z\|_F^2 + \mu \|Z - ZH^T(HH^T)^{-1}H\|_F^2 \quad (8)$$

DDL k -means

$$\text{s.t. } D_2D_3 \geq 0, D_3Z \geq 0, Z \geq 0.$$

We solve Equation 8 using alternating minimization. Initially, we ignore the nonnegativity constraints in Equation 8; later on, we will discuss how they can be handled. The updates for different variables are as follows:

$$D_1 \leftarrow \min_{D_1} \|X - D_1D_2D_3Z\|_F^2 \quad (9)$$

$$D_1^k = XZ_1^\dagger, \quad \text{where } Z_1 = D_2^{k-1}D_3Z^{k-1}.$$

Here in the Z_1^\dagger , the superscript cross stands for Pseudoinverse.

$$D_2 \leftarrow \min_{D_2} \|X - D_1D_2D_3Z\|_F^2 \quad (10)$$

$$D_2^k = (D_1^k)^\dagger XZ_2, \quad \text{where } Z_2 = D_3^{k-1}Z^{k-1}$$

$$D_3 \leftarrow \min_{D_3} \|X - D_1D_2D_3Z\|_F^2 \quad (11)$$

$$D_3^k = (D_1^kD_2^k)^\dagger X(Z^{k-1})^\dagger$$

$$Z \leftarrow \min_Z \|X - D_1D_2D_3Z\|_F^2 + \mu \|Z - ZH^T(HH^T)^{-1}H\|_F^2. \quad (12)$$

To solve Z , we need to take the gradient of the expression in Equation 12 and equate it to zero. The derivation is given below.

$$\begin{aligned} \nabla \left(\|X - D_1D_2D_3Z\|_F^2 + \mu \|Z - ZH^T(HH^T)^{-1}H\|_F^2 \right) &= 0 \\ \Rightarrow (D_1D_2D_3)^T X - (D_1D_2D_3)^T (D_1D_2D_3)^T Z - Z(\mu I - \mu H^T(HH^T)^{-1}H) &= 0 \\ \Rightarrow (D_1D_2D_3)^T X = (D_1D_2D_3)^T (D_1D_2D_3)^T Z + Z(\mu I - \mu H^T(HH^T)^{-1}H). \end{aligned}$$

The last step of the derivation implies that Z is a solution to Sylvester's equation of the form $AX + XB = C$. There are many efficient solvers for the same.

The final step is to update H . This is obtained by solving

$$H^k \leftarrow \min_H \|Z - ZH^T(HH^T)^{-1}H\|_F^2. \quad (13)$$

This is the k -means algorithm applied on Z .

In the derivation so far, we have not accounted for the ReLU nonnegativity constraints. Ideally, imposing the constraints would require solving them via forward-backward type splitting algorithms; such algorithms are iterative and hence would increase the run-time of the algorithm. We account for these constraints by simply putting the negative values in Z, Z_1 , and Z_2 to zeroes in every iteration.

The algorithm is shown in a succinct fashion below. Once the convergence is reached, the clusters can be found from H . Because Equation 8 is a nonconvex function, we do not have any guarantees for convergence. We stop the iterations when the H does not change significantly in subsequent iterations.

Algorithm: DDL $\pm k$ -means

Initialize: $D_1^0, D_2^0, D_3^0, Z^0, H^0$

Repeat till convergence

Update D_1^k, D_2^k, D_3^k using (9), (10), (11)

Update Z^k by solving Sylvester's eqn

Update H^k by k -means clustering

End



Gene expression based inference of cancer drug sensitivity

Received: 18 November 2021

Accepted: 12 September 2022

Published online: 27 September 2022

 Check for updates

Smriti Chawla¹, Anja Rockstroh², Melanie Lehman^{2,3}, Ellca Ratther², Atishay Jain⁴, Anuneet Anand⁴, Apoorva Gupta⁵, Namrata Bhattacharya^{2,4}, Sarita Poonia¹, Priyadarshini Rai¹, Nirjhar Das⁶, Angshul Majumdar^{4,7,8}, Jayadeva⁶, Gaurav Ahuja¹, Brett G. Hollier², Colleen C. Nelson²  & Debarka Sengupta^{1,4,7} 

Inter and intra-tumoral heterogeneity are major stumbling blocks in the treatment of cancer and are responsible for imparting differential drug responses in cancer patients. Recently, the availability of high-throughput screening datasets has paved the way for machine learning based personalized therapy recommendations using the molecular profiles of cancer specimens. In this study, we introduce Precily, a predictive modeling approach to infer treatment response in cancers using gene expression data. In this context, we demonstrate the benefits of considering pathway activity estimates in tandem with drug descriptors as features. We apply Precily on single-cell and bulk RNA sequencing data associated with hundreds of cancer cell lines. We then assess the predictability of treatment outcomes using our in-house prostate cancer cell line and xenografts datasets exposed to differential treatment conditions. Further, we demonstrate the applicability of our approach on patient drug response data from The Cancer Genome Atlas and an independent clinical study describing the treatment journey of three melanoma patients. Our findings highlight the importance of chemo-transcriptomics approaches in cancer treatment selection.

Cancer is a highly complex disease exhibiting varying degrees of genetic and phenotypic heterogeneity within individuals. Despite the apparent overall improvement in prognosis, responses to cancer treatment are often unpredictable. This is primarily attributable to the clonal diversity of cancer cells and associated phenotypically altered non-malignant cells in the tumor microenvironment. These pose a substantial hindrance to the optimal therapeutic management of the disease^{1,2}. Contemporary therapeutic strategies use cancer drugs, with lower toxicity that specifically target aberrantly expressed or mutated proteins and, in general, *EGFR* expression and mutations, *KRAS* mutations, *BCR-ABL* fusions, and *HER2* overexpression are such examples of common therapeutic targets in cancer³. Unfortunately, not all cancers and anti-cancer drugs are known to be associated with strong targetable genetic biomarkers. As such, it is concluded that the simple

relationship of drug targets or mutational status alone is comprehensive for predicting the efficacy of specific targeted therapies^{4,5}. Furthermore, administering a targeted therapy without considering drug resistance as a consequence may lessen patient survival. The drug resistance might be manifested through clonal expansion under treatment-induced selective pressure or from alternative signaling pathways that sustain tumor growth⁶. As such, early inference of drug response based on pretreatment molecular portraits of cancer has become a necessity^{2,7}.

In recent years, the availability of large-scale pharmacogenomic databases has propelled predictive personalized oncology research². Cancer Cell Line Encyclopedia (CCLE)⁸, Genomics of Drug Sensitivity in Cancer (GDSC)⁹, and Cancer Therapeutics Response Portal v2 (CTRPv2)¹⁰ are noteworthy among these. These high-throughput

A full list of affiliations appears at the end of the paper.  e-mail: colleen.nelson@qut.edu.au; debarka@iitd.ac.in

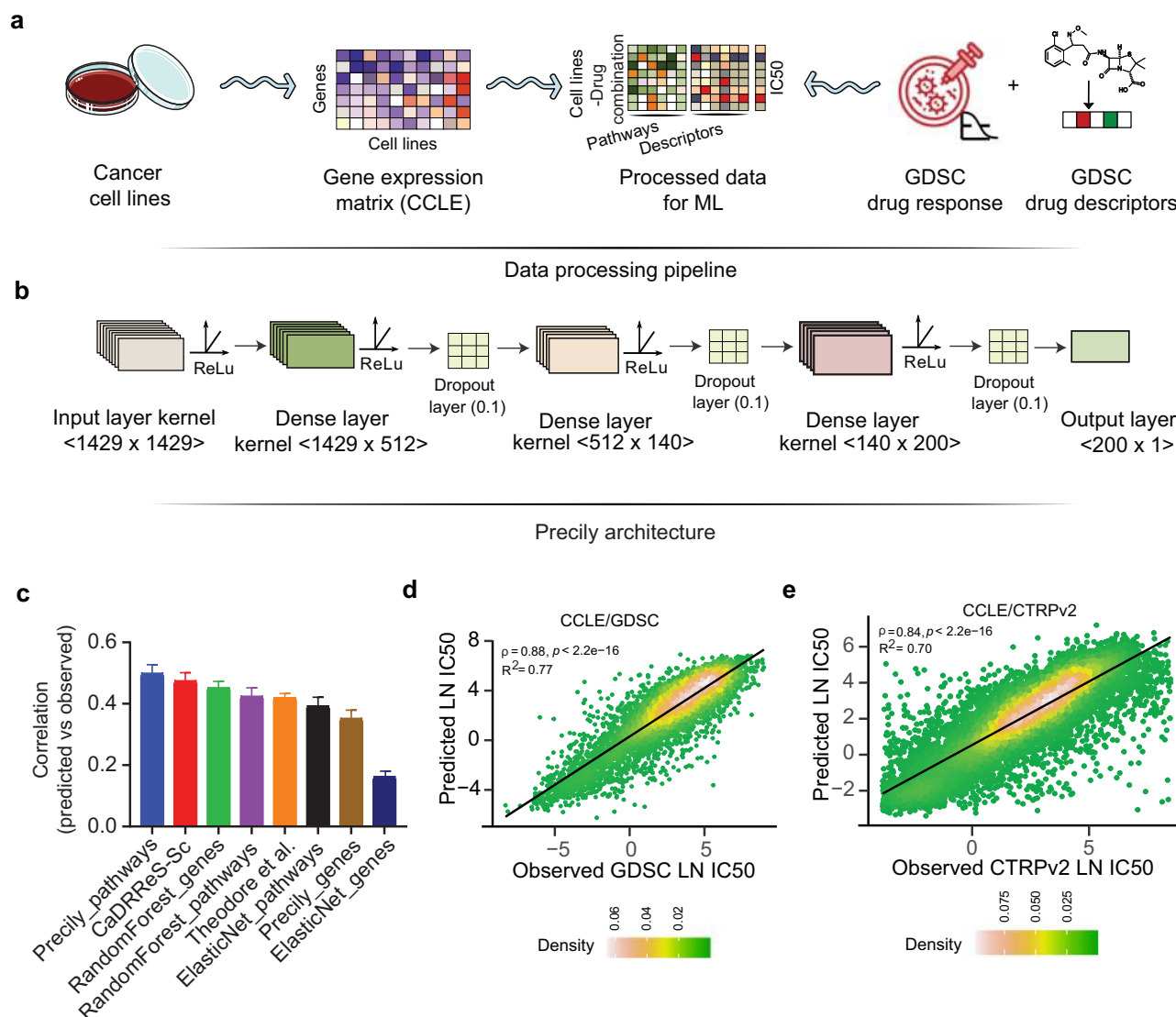


Fig. 1 | Illustration of the predictive analysis workflow of Precily. a Schematic workflow depicting the data processing pipeline of Precily. The first step involved the processing of training data. The RNA-seq gene expression (RSEM TPM) profiles from Cancer Cell Line Encyclopedia (CCLE) were subjected to pathway score transformation using GSVA. This GSVA score matrix was integrated with the drug descriptors obtained in the form of SMILES embedding for each compound. **b** Model architecture. The second step was the training of the ML model on this data, comprising GSVA scores and drug descriptors as an explanatory variable set and natural log-transformed IC50 values sourced from the GDSC database as the response variable. A deep neural network (DNN) from

the Keras platform was used to perform the regression task of predicting drug response. **c** Comparison of drug response prediction across different approaches. Barplot shows the distribution of Pearson's correlation coefficients for predicted vs. observed LN IC50 values for individual drugs ($n = 173$). Data are presented as mean values \pm SEM (Standard Error of the Mean). **d** Scatter plot demonstrating the performance of Precily across all cell line-drug pairs in the CCLE/GDSC test data. P -value was calculated using a two-sided t -test. **e** Scatter plot demonstrating the performance of Precily across all cell line-drug pairs in the CCLE/CTRPv2 test data. P -value was calculated using a two-sided t -test. Source data are provided in the Source Data file.

based on the squared coefficient of variation (CV^2). On held out data, Precily based predictions attained the highest correlation with ground truth, closely followed by CaDRReS-Sc. Figure 1c shows distributions of Pearson's correlation coefficients (ρ) across drugs, indicating the coherence between predictions by different methods and the ground truth LN IC50 values. The reason for presenting correlations at the level of drugs is that a major work¹⁶, we considered for comparisons, trained drug specific models using off-the-shelf H2O machine learning modules. In this way, one needs to manage one model for each compound. This approach is suboptimal since it does not leverage structural information of the compounds for prediction. For a global picture, we pooled our predictions across drugs and cell lines and obtained a Pearson's correlation coefficient value of 0.88 ($R^2 = 0.77$; P -value $< 2.2e-16$) (Fig. 1d).

While GDSC primarily catalogs anti-cancer drugs, the Cancer Therapeutics Response Portal v2 (CTRPv2) database features an assorted set of small molecules comprising tool compounds, probes and drugs, including US Food and Drug Administration (FDA)-approved cancer therapeutics¹⁰. We reciprocated a similar analysis of CCLE/GDSC on the CCLE/CTRPv2 combination. Notably, only 68 drugs were found common between GDSC2 and CTRPv2 datasets (Supplementary Fig. 1a). Further, we compared the distribution of LN IC50 values from GDSC2 and CTRPv2 datasets (Supplementary Fig. 1b). Gross differences were observed, which prevented us from integrating the two. Precily yielded a Pearson's correlation coefficient value of 0.84 ($R^2 = 0.70$; P -value $< 2.2e-16$) (Fig. 1e), whereas CaDRReS-Sc obtained $\rho = 0.83$ ($R^2 = 0.68$; P -value $< 2.2e-16$) (Supplementary Fig. 1c). Taken together, our analyses suggest that sensitivity to