# scDrugAtlas: an integrative single-cell drug response atlas for unraveling tumor heterogeneity in therapeutic efficacy

Wei Huang[1], Xinda Ren[1], Yinpu Bai[1], Hui Liu[1*]

[1*]College of Computer and Information Engineering, Nanjing Tech University, Nanjing, 211800 , Jiangsu, China.

*Corresponding author(s). E-mail(s): hliu@njtech.edu.cn;

**Abstract**

Tumor heterogeneity often leads to substantial differences in responses to same drug treatment. The presence of pre-existing or acquired drug-resistant cell subpopulations within a tumor survive and proliferate, ultimately resulting in tumor relapse and metastasis. The drug resistance is the leading cause of failure in clinical tumor therapy. Therefore, accurate identification of drug-resistant tumor cell subpopulations could greatly facilitate the precision medicine and novel drug development. However, the scarcity of single-cell drug response data significantly hinders the exploration of tumor cell resistance mechanisms and the development of computational predictive methods. In this paper, we propose scDrugAtlas, a comprehensive database devoted to integrating the drug response data at single-cell level. We manually compiled more than 100 datasets containing single-cell drug responses from various public resources. The current version comprises large-scale single-cell transcriptional profiles and drug response labels from more than 1,000 samples (cell line, mouse, PDX models, patients and bacterium), across 66 unique drugs and 13 major cancer types. Particularly, we assigned a confidence level to each response label based on the tissue source (primary or relapse/metastasis), drug exposure time and drug-induced cell phenotype. We believe scDrugAtlas could greatly facilitate the Bioinformatics community for developing computational models and biologists for identifying drug-resistant tumor cells and underlying molecular mechanism. The scDrugAtlas database is available at: http://drug.hliulab.tech/scDrugAtlas/.

**Keywords:** Single-cell transcriptome, Drug response, Drug resistance, Tumor heterogeneity, Pseudotime trajectory

# 1 Introduction

The existence of intrinsic and acquired drug-resistant cells is the main cause of failure in tumor therapy. Despite the fact that routine drug treatment can effectively eliminate the majority of malignant cells, a small subset of tumor cells would survive and continue to proliferate, ultimately leading to tumor recurrence and progression. Some scholars have developed bioinformatic methods to predict patients' clinical response to drugs based on the bulk RNA-seq of cell lines already with large-scale drug sensitivity data, such as GDSC [1, 2], CCLE [3, 4] and CTRP [5]. However, these bulk-level drug sensitivity data can not identify the response or resistance of individual cells to specific drugs, thereby failing to adequately dissect the tumor heterogeneity in drug responses.

In recent years, the advent of single-cell RNA-sequencing (scRNA-seq) technology had generated a wealth of single-cell transcriptional data, providing valuable opportunity to explore the intratumoral heterogeneity in drug response [6]. However, no large-scale experimental data for single-cell drug sensitivity is available, resulting in a significant gap in our knowledge regarding drug response at the single-cell level and bulk level. This gap press a challenging but urgent task to predict the drug sensitivity of individual cells. Several studies have employed transfer learning techniques to align bulk RNA-seq and single-cell RNA-seq (scRNA-seq) data for predicting drug responses at the single-cell level [7, 8]. scAdaDrug employs adaptive-weighted multi-source domain adaptation to align scRNA-seq data with multiple bulk RNA-seq data domains, resulting in improved performance in predicting drug responses of individual cells [9]. SSDA4Drug, on the other hand, utilizes semi-supervised domain adaptation to enhance domain alignment by leveraging few-shot samples from the target scRNA-seq domain, thereby achieving superior accuracy in identifying individual drug-resistant cells [10]. However, the scarcity of single-cell drug response data still significantly impede the exploration of tumor cell resistance mechanisms, and also restricts the development of high-performance computational methods for predicting drug-resistant individual cells.

So, we have worked hard to develop an integrative single-cell drug response database, referred to as scDrugAtlas. We manually compiled single-cell drug response datasets from various public resources, including GEO, PubMed publications and external databases. To ensure the reliability of the drug response labels, we carefully examined that the cell tissue originated from primary, recurrent, or metastatic lesions, whether the cell line was exposed to the drug for a sufficient duration, and whether the cells acquired resistance and continued to proliferate. Based on these assessments, we assigned response labels to each cell regarding specific drug. After data preprocessing and filtering of low-quality data, our database now comprises the transcriptional profiles and drug response labels of 1,014,771 cells came from more than 1,000 samples (cell line, mouse, PDX models, patients and bacterium), across 64 unique drugs and 13 major cancer types. Furthermore, we have built quite a few integrated datasets for some cancer type of interest by removal of batch effects. To our best knowledge, scDrugAtlas is the largest comprehensive database devoted to single-cell drug response to

date. So, we strongly believe scDrugAtlas will draw widespread reader interest, especially for those who are interested in exploring drug-resistant cell subpopulations in tumors.

## 2 Data resources and preprocessing

First, we retrieved the PubMed publications using the keywords "((single cell) OR (scRNA-seq) " AND "drug" AND "(response) OR (sensitivity) OR (resistance) OR (tolerance)", and subsequently extracted associated data and determined the response labels of individual cells by manually reviewing the content of these publications. Also, we used the same keywords to retrieve datasets from GEO [11] and determined the single-cell response labels based on the corresponding publications. Finally, we also selected some datasets from external databases, including CeDR [12] and DRMref [13]. Figure 2 illustrated the data resources, preprocessing and integration procedures, as well as data analysis functionalities.

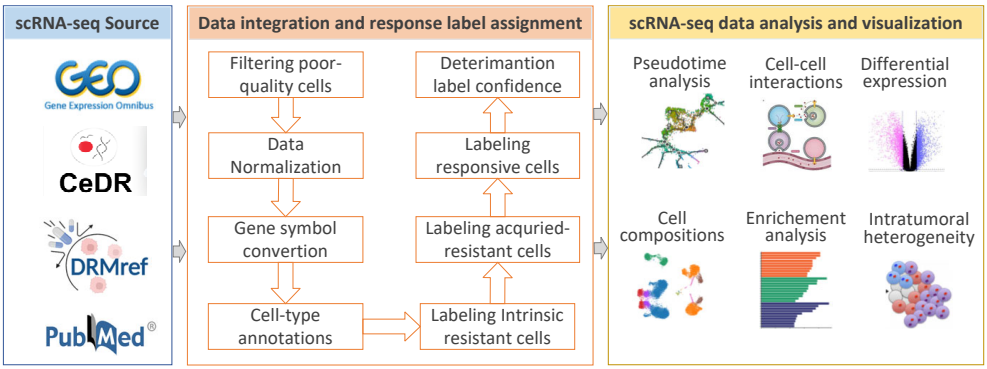


**Fig. 1** Overview of scDrugAtlas data resource, data integration and drug response label assignment, and functional modules for single-cell transcriptomic data analysis and visualization.

After data collection, we excluded cells with the number of expressed genes <500 and with the mitochondrial gene ratio >10% from each dataset. Subsequently, we utilized pyensembl to convert gene identifiers to Ensemble IDs. As a result, scDrugAtlas now comprises the single-cell transcriptional profiles and drug response labels of 1,014,771 cells came from more than 1,000 samples (cell line, mouse, PDX models, patients and bacterium), across 64 distinct drugs and 13 major cancer types. The drugs covered chemotherapy, targeted therapy and immunotherapy. Detailed drug information was extracted from DrugBank (version 5.1.10) [14].

To facilitate the usage of our data resource, we built a few combined datasets for some cancer type of interest, by integrating the single-cell transcriptional profiles of same type cancer cells that were subjected to specific drug treatment. Specifically, we combined the datasets and labeled the batches, and then corrected batch effects

using the batch balanced kNN algorithm from scanpy [15]. For example, one combined dataset comprised 23,314 melanoma cells subjected to Vemurafenib treatment, in which 11,955 cells are responsive and 11,359 cells are resistant. Another combined dataset consists of 9,274 breast cancer cells, in which 6,235 cells are responsive and 3,039 cells are resistant when subjected to Paclitaxel treatment. These combined datasets are ready for the usage in training machine learning models for predicting single-cell drug response.

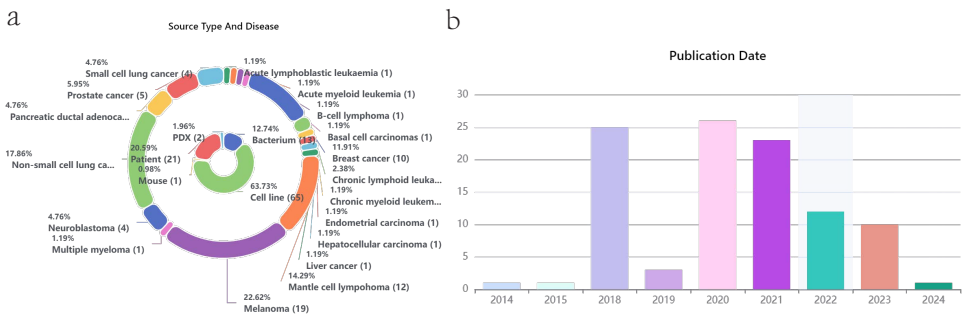


**Fig. 2** Statistics of scDrugAtlas datasets. (a) Proportion of datasets according to cell source types and diseases. (b) Frequency of datasets according to publication date.

## 3 Assignment of drug response labels

Drug resistance can be categorized as intrinsic or acquired based on its development timeline. Intrinsic resistance is present before drug treatment, whereas acquired resistance develops following therapy. Each type is observed in approximately 50% of cancer patients with drug resistance [16]. Pre-existing resistance refers to the innate capability of a cell to resist the effects of a drug without prior exposure or selective pressure. For instance, the HepG2 hepatocellular carcinoma cells demonstrates intrinsic resistance to Doxorubicin, while the A549 non-small cell lung cancer (NSCLC) cells shows intrinsic resistance to certain tyrosine kinase inhibitors (TKIs) such as Erlotinib. For the pre-existing drug-resistant cells, we assign them the label resistance.

In response to sustained drug pressure, tumor cells may develop resistance through genetic mutations, transcriptional reprogramming, or epigenetic silencing. Additionally, tumor cells can adapt by activating alternative survival pathways or exploiting microenvironmental factors that protect them from the drug. As a result, we assigned the resistance label to those cells derived from patients, mice, or patient-derived xenograft (PDX) models, only if tumor recurrence or metastasis occurs. Conversely, cells obtained prior to the development of resistance are categorized as sensitive. For individual cells derived from cell lines, the development of drug resistance primarily depends on the duration of drug exposure. Therefore, the drug response labels are assigned based on their phenotypic changes or the time elapsed post-exposure.

Cells that continue to proliferate during treatment are deemed fully resistant and are assigned high-confidence resistance labels. Cells that survive treatment but show no significant proliferation are labeled as resistant with relatively low confidence. The pre-treatment cells are classified as sensitive.

# 4 Functionalities

A website with user-friendly data visualization and analysis tools is provided to help users explore the wealth of data (Fig 4a). Users can input a gene of interest to explore the scRNA-seq data with each cell colored by its drug response label. Other popular analysis tools includes differential expression analysis between sensitive vs resistant cell populations, pseudotime analysis for visualizing the cancer cell progression to drug resistance along with evolutionary trajectory. We summarized the functional modules as below:

- Dataset listing and search: Users can search for single-cell drug response datasets of interest through keywords on the homepage. Each dataset has been associated with tissue or cell source, cell count, drug, and quality level of response labels. In particular, users can filter datasets by setting source type, species, release time, and disease on the Dataset Listing page (Fig 4b).
- scRNA-seq data visualization: UMAP is used to display single-cell sequencing data, with cell types annotated using the SingleR tool [17]. Here users can browse the proportion and number of sensitive and resistant cells within a dataset (Fig 4c). Additionally, the expression profiles of genes of interest in different cells can be displayed through a heatmap on the Dataset Detail Page.
- Gene search across datasets: user can search for genes of interest across multiple scRNA-seq datasets and narrow down the scope by species and tissue type. We use UMAP to display the expression levels of the gene across different datasets.
- Differential expression analysis: differential expression analysis between sensitive and resistant cells in a specific dataset or across two datasets can be easily conducted. The volcano plot is used to display differentially expressed genes. Moreover, users can launch functional enrichment analysis using the up-regulated or down-regulated gene sets here.
- Evolutionary trajectory analysis: Pseudotime analysis is performed using monocle3 tool [18]. The visualization of drug sensitivity changes along the evolutionary trajectory of cells is helpful to discover the critical state of tumor cells transitioning from sensitivity to resistance and thus identify key drug-resistant genes (Fig 4d).
- Gene set variation analysis: Calculating the signaling pathway activity levels of each cell using the C2 gene set of the MSigDB database [19], which allows users to compare the signaling pathway activity levels between sensitive and resistant cells and discover key signaling pathways leading to drug resistance (Fig 4e).
- Functional enrichment analysis: user can easily conduct GO and KEGG functional enrichment analysis for the up-regulated or down-regulated gene sets between sensitive and resistant cells.

- Intra-tumoral heterogeneity scoring: The DEPTH tool [20] is used to calculate the heterogeneity score within tumors, which is helpful for explore the differences of transcriptional state between sensitive and resistant cells (Fig 4f).
- Cell-cell communication: Calculate and display cell-cell communications between different types of cells using CellphoneDB tool [21]. Specifically, users are guided to focus on the strength of ligand-receptor interactions between resistant and sensitive cells (Fig 4g).
- Dataset download: Users can download datasets of interest from the Dataset detail page. More importantly, we have constructed a batch of combined datasets regarding specific drugs, by using the Harmony tool to remove batch effects. These combined datasets are ready for the usage in training deep learning models for predicting single-cell drug response.

# 5 A case study

To demonstrate the utility of our dataset, we built a machine learning model using a combined dataset and validated its effectiveness on single-cell sequencing data from patient samples. Specifically, We employed the integrated dataset comprised 9,274 breast cancer cells, with 6,235 cells categorized as responsive and 3,039 cells as resistant to Paclitaxel treatment. The UMAP projection of this dataset is illustrated in Figure 4(a). To build a classification model, we labeled the sensitive cells as 1 and resistant cells as 0. The dataset was partitioned into training and testing subsets at an approximate ratio of 3:1, and then a random forest model was trained. This model achieved an area under the curve (AUC) value of 0.95 on the test set, as shown in Figure 4(a).

To verify the performance on independent test set, we obtained single-cell sequencing data of patient-derived xenograft (PDX) models generated using dissociated tumor cells obtained from a patient with triple-negative breast cancer from the GEO database (accession number: GSE226329). Two samples were treated with paclitaxel and DMSO (control) for 15 days, and the tumors were harvested after administration of the last dose. We applied the trained random forest model to predict drug sensitivity for individual cells from these two samples. The results showed that the predicted score of the resistant cells was significantly lower than that of the control cells (p-value$<0.05$). Although the case is simple, its capability to discriminate the drug sensitivity in patient-derived single-cell sequencing data effectively validates the practical application of our dataset.

# 6 Discussion and Conclusion

The phenotype-based high-throughput screening (HTS) primarily captures the overall drug response of tumor cells but falls short in revealing the difference of drug response at single-cell level. This limitation arises because bulk RNA sequencing (bulk RNA-seq) often produces signals dominated by growth-advantageous cell clones, thereby obscuring signals of the minority of drug-resistant cell subpopulations. As

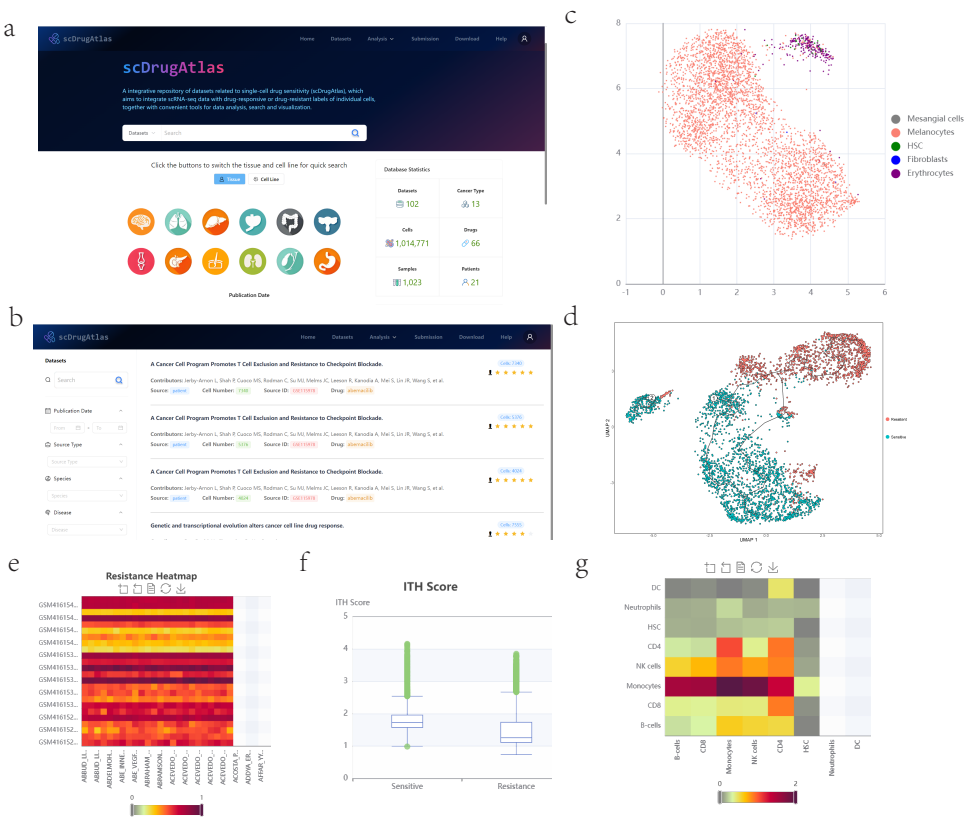


**Fig. 3** Screenshots of website homepage, dataset listing page, and functional modules. (a) Website Homepage, (b) scRNA-seq dataset listing and filtering page, (c) UMAP plot of scRNA-seq data with cell-type annotations. (d) Pseudotime trajectory of tumor cells colored by drug response labels (responsive vs resistant). (e) Pathway activity levels of single cells. (f) Intra-tumoral heterogeneity scores of cell subgroups split by responsive and resistant labels. (g) Cell-cell interactions between different types of cells.

a result, the molecular mechanisms underlying tumor drug resistance remain largely uncharacterized.

Despite the DRMref [13] database also paid attention to the single-cell drug resistance, the scale of data and the diversity of covered drugs and species remain significantly limited. Compared to DRMref, our database possessed nearly three times the amount of data, as well as more reliable drug response labels. To ensure the reliability of the drug response labels, we meticulously verified several factors: the origin of the cell tissue (whether from primary, recurrent, or metastatic lesions), the duration of drug exposure, and the presence of acquired resistance and continued cell proliferation.

In addition to the advantages in dataset volume and label quality, we have also prioritized the development of functionality for comparative analysis between

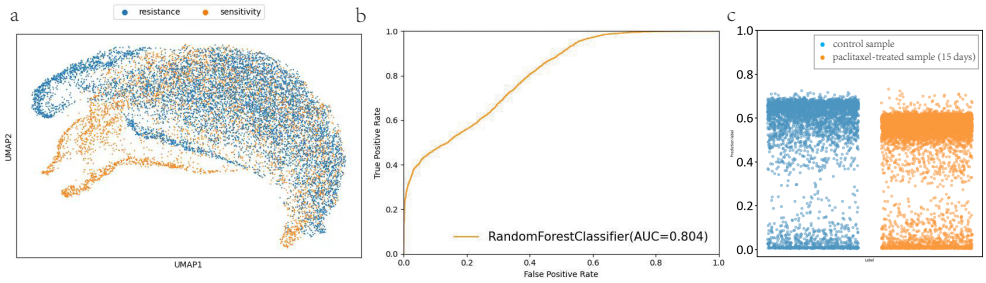


**Fig. 4** Visualization of integrated scRNA-seq data regarding breast cancer response to paclitaxel, as well as performance evaluation of random forest classifier trained on the integrated dataset. (a) ROC curve achieved by random forest classifier on training set. (b) UMAP plot of scRNA-seq data in training set. (d) Predicted drug sensitivity for individual cells of two PDX samples

drug-sensitive and drug-resistant cells. By correlating the tumor cell evolutionary trajectories with the progression of cells towards drug-resistant state, users can easily identify the critical state of cellular transition from sensitivity to drug resistance during pseudo-time, which greatly facilitate the discovery of new drug resistance genes. Furthermore, we calculate the activity levels of signaling pathways in drug-sensitive and drug-resistant cell subpopulations, and conduct differential analysis based on the activity levels to reveal the drug resistance mechanisms of tumors at the single-cell level.

# References

[1] 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).

[2] Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).

[3] Li, H. *et al.* The landscape of cancer cell line metabolism. *Nature medicine* **25**, 850–860 (2019).

[4] Ghandi, M. *et al.* Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).

[5] Rees, M. G. *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology* **12**, 109–116 (2016).

[6] Vitale, I., Shema, E., Loi, S. & Galluzzi, L. Intratumoral heterogeneity in cancer progression and response to immunotherapy. *Nature medicine* **27**, 212–224 (2021).

[7] Chen, J. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell rna-seq data. *Nature Communications* **13**, 6494 (2022).

[8] Zheng, Z. *et al.* Enabling single-cell drug response annotations from bulk rna-seq using scad. *Advanced Science* **10**, 2204113 (2023).

[9] Duan, W. & Liu, H. Predicting single-cell drug sensitivity by adaptive weighted feature for adversarial multi-source domain adaptation. *arXiv preprint arXiv:2403.05260* (2024).

[10] Liu, H. & Huang, K. Identifying drug-resistant individual cells within tumors by semi-supervised domain adaptation. *bioRxiv* 2024–08 (2024).

[11] Barrett, T. *et al.* Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research* **41**, D991–D995 (2012).

[12] Wang, Y.-Y. *et al.* Cedr atlas: a knowledgebase of cellular drug response. *Nucleic acids research* **50**, D1164–D1171 (2022).

[13] Liu, X. *et al.* Drmref: comprehensive reference map of drug resistance mechanisms in human cancer. *Nucleic Acids Research* **52**, D1253–D1264 (2024).

[14] Knox, C. *et al.* Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research* **52**, D1265–D1275 (2024).

[15] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1–5 (2018).

[16] Kelderman, S., Schumacher, T. N. & Haanen, J. B. Acquired and intrinsic resistance in cancer immunotherapy. *Molecular oncology* **8**, 1132–1139 (2014).

[17] Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology* **20**, 163–172 (2019).

[18] Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381–386 (2014).

[19] Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).

[20] Li, M., Zhang, Z., Li, L. & Wang, X. An algorithm to quantify intratumor heterogeneity based on alterations of gene expression profiles. *Communications biology* **3**, 505 (2020).

[21] Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols* **15**, 1484–1506 (2020).