nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00656-y

Supervised learning of high-confidence phenotypic subpopulations from single-cell data

Received: 22 September 2022

Accepted: 6 April 2023

Published online: 8 May 2023

Check for updates

Tao Ren[1,2], Canping Chen[3,4], Alexey V. Danilov[5], Susan Liu[3,4], Xiangnan Guan[6], Shunyi Du[3,4], Xiwei Wu[5], Mara H. Sherman[7,8,9], Paul T. Spellman[8,10], Lisa M. Coussens[7,8], Andrew C. Adey[8,10], Gordon B. Mills[11], Ling-Yun Wu[1,2] ✉ & Zheng Xia[3,4,8] ✉

Accurately identifying phenotype-relevant cell subsets from heterogeneous cell populations is crucial for delineating the underlying mechanisms driving biological or clinical phenotypes. Here by deploying a Learning with Rejection strategy, we developed a novel supervised learning framework called PENCIL to identify subpopulations associated with categorical or continuous phenotypes from single-cell data. By embedding a feature selection function into this flexible framework, for the first time, we were able to simultaneously select informative features and identify cell subpopulations, enabling accurate identification of phenotypic subpopulations otherwise missed by methods incapable of concurrent gene selection. Furthermore, the regression mode of PENCIL presents a novel ability for supervised phenotypic trajectory learning of subpopulations from single-cell data. We conducted comprehensive simulations to evaluate PENCIL's versatility in simultaneous gene selection, subpopulation identification and phenotypic trajectory prediction. PENCIL is fast and scalable to analyse one million cells within 1 h. Using the classification mode, PENCIL detected T-cell subpopulations associated with melanoma immunotherapy outcomes. Moreover, when applied to single-cell RNA sequencing of a patient with mantle cell lymphoma with drug treatment across multiple timepoints, the regression mode of PENCIL revealed a transcriptional treatment response trajectory. Collectively, our work introduces a scalable and flexible infrastructure to accurately identify phenotype-associated subpopulations from single-cell data.

Heterogeneous cellular systems alter cell states and compositions in response to development, perturbations, pathological change and clinical intervention, resulting in phenotypically distinct cell subpopulations[1–4]. Rapidly accumulating single-cell studies are profiling samples from different experimental or pathological conditions, such as wild-type versus knockout conditions[5], treatment resistance versus responder groups[6], and disease progression graded with scores[7]. Distinguishing subpopulations associated with phenotypes of interest from heterogeneous cell populations will improve phenotype-specific signal detection and facilitate reliable downstream analysis.

A full list of affiliations appears at the end of the paper. ✉e-mail: lywu@amss.ac.cn; xiaz@ohsu.edu

For categorical phenotypes, phenotype-associated subpopulations can be identified through differential abundance analysis. A straightforward method is to cluster cells first and then compare the ratios of conditions in each cluster[8]. Furthermore, recent developments have proposed clustering-free strategies such as DAseq[9], Milo[10] and MELD[11] by examining phenotype labels of cells connected through the k-nearest neighbour (KNN) graph. Nevertheless, KNN graphs require gene selection beforehand, which is determined separately in an unsupervised manner, for example, the top most variable genes (MVGs). Such unsupervised gene selection approaches[12,13] may not capture phenotype-associated cell subpopulations hidden in a latent gene space. Therefore, to accurately detect cells of interest, gene selection must be embedded in the subpopulation identification process. However, given the cell–cell similarity matrix as input, KNN-based tools cannot incorporate gene selection into subpopulation identification.

Moreover, beyond detecting static categorical cell subsets, we need to order the selected cells along the continuous phenotypic trajectory to reveal transitions and relationships during dynamic biological processes, such as tissue development and disease progression[14–19], a critical task for single-cell analysis[20]. However, although Milo[10] can input continuous phenotypes, it only interprets subpopulations changing with the phenotype qualitatively without ordering cells in a trajectory manner. Thus, further methodological development of new frameworks beyond cell–cell similarity is necessary.

In this Article, we propose a new tool that uses the Learning with Rejection (LWR) strategy to detect high-confidence phenotype-associated subpopulations from single-cell data (PENCIL). LWR includes a prediction function (Fig. 1a) along with a rejection function (Fig. 1b) to reject low-confidence cells. Then, by embedding a feature selection term into this LWR framework, PENCIL can perform gene selection during the training process, which allows learning proper gene spaces that facilitate accurate subpopulation identifications from single-cell data. Furthermore, the regression mode of PENCIL can order cells to reveal subpopulations undergoing continuous transitions between conditions.

Results

Overview of PENCIL

We developed a new supervised framework named PENCIL for phenotypic subpopulation identification from single-cell data. Inspired by the LWR strategy in machine learning, we convert differential abundance analysis into a supervised-learning application (Fig. 1a,b). Intuitively, from the LWR perspective, cell subpopulations enriched by a particular phenotype will be easier to classify/fit, while those with similar abundances of phenotype labels will result in more classification/fitting errors and should therefore be rejected (Extended Data Figs. 1 and 2). Then, by incorporating the supervised feature selection technique into LWR, PENCIL can simultaneously select informative genes and identify phenotype-associated cell subpopulations.

Specifically, the data sources for PENCIL input include a single-cell quantification matrix and condition labels for all cells (Fig. 1c,d). Condition labels can take various forms, such as multiple experimental perturbations, disease stages, timepoints and so on. In brief, PENCIL consists of three modules: gene weights w, predictor h and rejector r (Fig. 1e). The input gene expression vector x of a particular cell is first multiplied by the gene weights and then fed to the predictor and rejector, respectively. Then the predictor will predict the cell label and the rejector will produce a confidence score that quantifies the credibility of the predicted label from the predictor (Fig. 1f). The gene weights are learnable, and the predictor and rejector are ordinary trainable models. The parameters of all three modules are trained by minimizing an objective function on the input expression matrix with condition labels (Fig. 1g), where the gene weights are penalized with a sparse penalty (l_1-norm) to select informative genes. Minimizing this total loss is essentially choosing a smaller value between the fitting cost l and the predetermined rejection cost c for each cell. For easy-to-fit cells, the former will be chosen with confidence scores $r(x) > 0$. In turn, for hard-to-fit cells, the rejection will be chosen ($r(x) < 0$). Finally, the combination of the predicted labels and the confidence scores ($r(x) > 0$) from the rejection function will output the selected subpopulations with predicted labels (Methods).

PENCIL can take either categorical phenotypes or continuous variables as inputs. For example, Fig. 1h shows a simulated single-cell RNA sequencing (scRNA-seq) dataset with binary phenotype labels in a Uniform Manifold Approximation and Projection (UMAP)[21] using the top 5,000 MVGs. The standard top 5,000 MVGs-based clustering analysis cannot distinguish the two phenotypic clusters contained in cluster 0 (Fig. 1i). In contrast, our classification mode of PENCIL with gene selection can identify the two subtle phenotypic subpopulations, as shown by the UMAP based on the PENCIL-selected genes (Fig. 1j). Furthermore, by setting the predictor module as a regressor, PENCIL can handle continuous phenotype labels like disease stages, which performs a fundamentally different task than the differential abundance analysis. For instance, in a simulated single-cell dataset from two conditions[22] (Fig. 1k), category-based subpopulation identification methods, such as Milo[10], only identify the differentially abundant subpopulations (Fig. 1l). Intriguingly, the regression-based PENCIL reconstructs the phenotypic trajectory to reveal subpopulations that are undergoing a continuous transition between conditions (Fig. 1m).

PENCIL simultaneously selects genes and cells

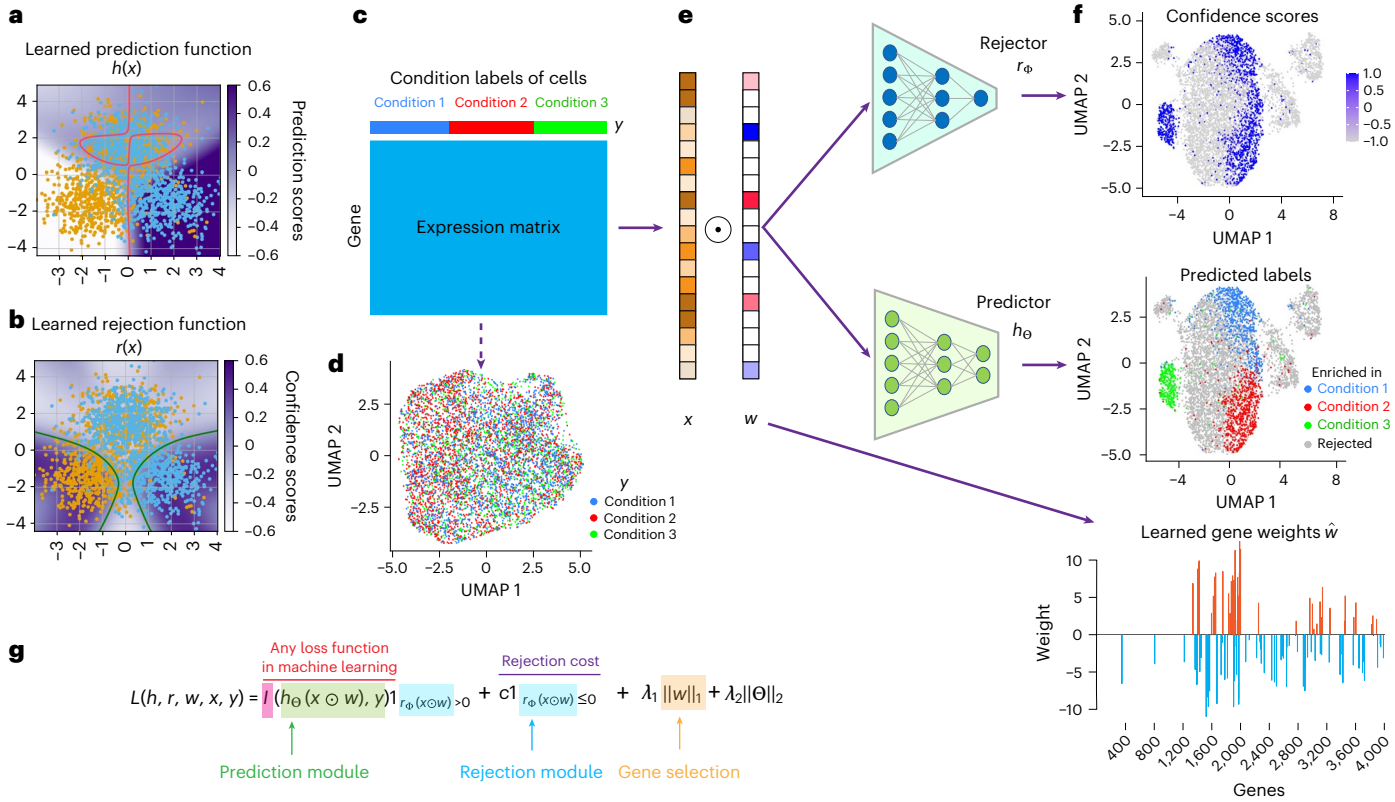
To test the effectiveness of PENCIL, we set up a series of simulated datasets, and performed comprehensive comparisons with existing methods, including DAseq[9], Milo[10] and MELD[11]. We exploited a real T-cell scRNA-seq dataset[6] with 6,350 cells to generate various simulation settings by picking gene sets and simulating condition labels accordingly (Extended Data Fig. 3 and Methods). For two conditions, we first pre-selected a subset of genes from the top 2,000 MVGs for clustering and picked two clusters as the ground truth phenotypic subpopulations (Fig. 2a). After setting up the simulation (Methods), we used the gene expression matrix of the top 2,000 MVGs and the simulated conditions labels as the same inputs for all four methods (Fig. 2b,c). Since the genes used to generate the clustering are only a subset of the total genes, the standard scRNA-seq analysis pipeline using the top 2,000 MVGs will not capture the proper cell similarities, resulting in ambiguous aggregation patterns for cell label information

Fig. 1 | The workflow of PENCIL and its main functions. a,**b**, A simulated example to show the learned prediction model with the red line as the boundary with prediction scores $h(x) = 0$ to separate the two predicted classes; and the learned rejection model with the green lines as the boundary with confidence scores $r(x) = 0$ to reject cells. **c**, The inputs for PENCIL are a single-cell data matrix and condition labels of all cells y. **d**, The single-cell expression matrix is visualized by the UMAP using the top 2,000 MVGs with cells coloured by the condition labels. **e**, The three trainable components of PENCIL: gene weights w, rejector module and predictor module. **f**, The outputs of PENCIL are confidence scores, predicted labels and learned gene weights. The UMAPs are generated by the PENCIL-selected genes with $\hat{w} \neq 0$. **g**, The rejection-based total loss function of PENCIL for optimization. **h**, UMAP using the top 5,000 MVGs showing a dataset with two conditions coloured by their condition labels. **i**, Standard clustering analysis based on the top 5,000 MVGs. **j**, UMAP based on the PENCIL-selected genes showing the identified phenotype-enriched cell subpopulations. **k**, UMAP visualization of a simulated scRNA-seq data with cells coloured by the conditions. The designated regions enriched in each condition are denoted by the dashed ovals. **l**, Differential abundance analysis such as Milo and classification mode of PENCIL can only identify static phenotype-associated cell subpopulations from the data shown in **k**. **m**, Continuous phenotype regression PENCIL analysis rejected the irrelevant cells and predicted the time orders of phenotypic cells to reveal continuous transition states as indicated by the red dashed circle.

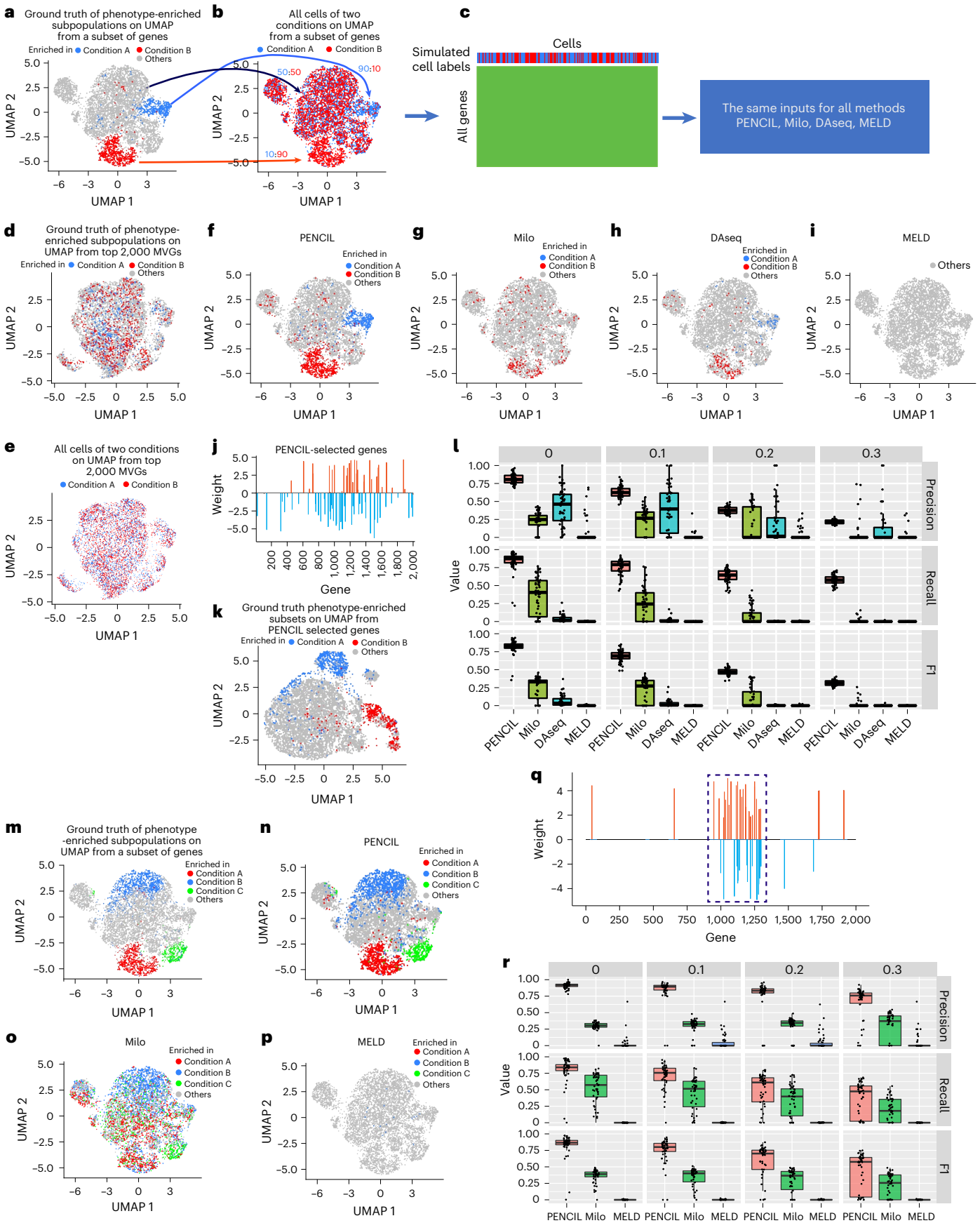
https://doi.org/10.1038/s42256-023-00656-y

(Fig. 2d,e), thus making it difficult for the methods using the KNN based on the top 2,000 MVGs to identify subpopulations of interest.

Due to its unique ability to simultaneously select genes and identify subpopulations, PENCIL recovered 84.5% of the ground truth phenotype-enriched cells while maintaining a high precision (0.833) (Fig. 2f and Extended Data Fig. 4a–c). In contrast, because the top 2,000 MVGs were not able to capture the proper similarities of the ground truth phenotypic subpopulations (Fig. 2d,e), the other three KNN-based methods did poorly, especially MELD, which did not select any cells (Fig. 2g–i and Extended Data Fig. 4d). Indeed, the feature selection in PENCIL contributes to improving the performance of this process, as illustrated by the UMAP generated from the PENCIL-selected genes, which captured an appropriate cell–cell similarity structure of the designed ground truth subpopulations (Fig. 2j,k). We repeated



https://doi.org/10.1038/s42256-023-00656-y



this experiment 50 times, each time with 300 randomly selected key genes from the top 2,000 MVGs to cluster cells. As the mixing rate increased, the performances of all the methods decreased, but PENCIL consistently provided better performances than other methods (Fig. 2l). In addition, integrating cells from different samples and conditions must address the batch-effect issue[23]. PENCIL can take the

Fig. 2 | Evaluation of PENCIL's classification mode for simultaneously selecting genes and cells in simulations. a, Ground truth of phenotype-enriched subpopulations on UMAP generated from a manually pre-selected gene set (1,000–1,300th MVGs) for the simulation with two conditions. **b**, The two phenotypic subpopulations were assigned to the two conditions accordingly with a mixing rate of 0.1 and all remaining cells are evenly assigned with condition labels, as shown by the arrows and ratios. **c**, A cartoon to show the expression matrix and simulated condition labels for all cells. **d**, Ground truth of phenotype-enriched subpopulations in **a** visualized on the UMAP using the top 2,000 MVGs. **e**, Cells with condition labels in **b** visualized on the UMAP using top 2,000 MVGs. **f**–**i**, Results of PENCIL, Milo, DAseq and MELD visualized on the same UMAP as in **a**. **j**, The learned gene weights by PENCIL. **k**, Ground truth of phenotype-enriched subpopulations in **a** visualized on the UMAP using the PENCIL-selected genes. **l**, The box plots showing the results of the four methods ($n = 50$ simulations) with four mixing rates 0, 0.1, 0.2 and 0.3. **m**, Ground truth of phenotype-enriched subpopulations and background cells on UMAP generated from a manually pre-selected gene set (1,000–1,300th MVGs) for the simulation with three conditions. **n**–**p**, Results of PENCIL, Milo and MELD visualized on the same UMAP as in **m**. **q**, The learned gene weights by PENCIL. Dashed region indicates pre-selected genes for UMAP in **m. r**, Box plots of performance for PENCIL, Milo and MELD in the simulations with three conditions and four mixing rates ($n = 50$). In box plots, the median and upper and lower quartiles are represented by the centre line and box bounds, respectively. Box whiskers display the largest and smallest values within 1.5 times the interquartile range from the quartiles.

batch-corrected matrix as input. We employed Splatter[24] to simulate expression data with batch effects. The results suggested that PENCIL can be integrated successfully with classic batch correction methods implemented in the Seurat[25] Package (Extended Data Fig. 5 and Methods). We repeated the simulations 50 times with four mixing rates for the batch-effects and showed that PENCIL consistently performed better than existing KNN-based methods (Extended Data Fig. 5g).

Moreover, PENCIL is flexible to address more than two conditions. Therefore, we did similar evaluations on simulation datasets with three conditions using the same T-cell scRNA-seq dataset[6], showing that PENCIL is more effective than other methods (Fig. 2m–r, Extended Data Fig. 6 and Methods).

Indeed, the feature selection function embedded in the PENCIL framework selects informative genes and improves the performance in identifying phenotype-enriched subpopulations hidden in a latent gene space (Extended Data Fig. 7), which cannot be accurately detected by methods lacking gene selection during training.

PENCIL enables supervised phenotypic trajectory learning

In addition to categorical phenotypes, increasingly single-cell datasets are designed to profile tissues from multiple timepoints and continuous disease stages[14–16,26]. Our LWR-based PENCIL framework can also easily incorporate those continuous phenotypes into the regression mode by updating the prediction loss function (Methods). Herein, we conducted a series of simulations to demonstrate the performance and utilities of PENCIL in regression tasks. In the first simulation, we used data from a real scRNA-seq T-cell dataset[9] (16,291 cells with 10 principal components (PCs)) that had been processed by PC analysis to generate timepoint labels. Three overlapping timepoints on the selected cell trajectory were set as the ground truth for this simulation experiment (Fig. 3a and Extended Data Fig. 8a), and cell labels were simulated accordingly, with the other cells being randomly assigned a time label as background noise (Fig. 3b). Regressing the simulated timepoints as continuous variables, PENCIL captured practically the entire track of cells defined in the simulated ground truth (Fig. 3c and Extended Data Fig. 8b). Though Milo claims to be able to handle continuous variables, it only picked out the cells at the beginning and end of the trajectory, omitting the middle cells (Fig. 3d). The Venn diagram comparisons showed that PENCIL did allocate more ground truth cells (92% versus 54%) with higher precision (90% versus 80%) than Milo (Fig. 3e). More importantly, the most unique characteristics of regression-based PENCIL is its ability to predict continuous time orders for the selected cells (Fig. 3f), whereas Milo merely tests for a decrease or increase (negative or positive) in abundance over time (Fig. 3g). Intriguingly, in this example, the histogram plot of the distribution of the time orders predicted by PENCIL showed two additional peaks at timepoints 1.5 and 2.5, suggesting hidden cell transition stages between the three designed timepoints (t1.5 and t2.5) (Fig. 3h). Thus, the predicted continuous time scores can reveal new critical timepoints or phenotypic stages between designated timepoints that would otherwise be unnoticed by experimental plans or clinical definitions.

Next, we examined the gene selection function of PENCIL in the regression task. Like in the previous regression-based experiments, we designated ground truth phenotypic subpopulation based on the clusters, but this time the clusters were generated from a pre-selected gene set to necessitate feature selection. Consistently, the regression-based PENCIL-learned informative genes facilitate subpopulation identification and predict continuous timepoints for the selected cells (Fig. 3i–o, Extended Data Fig. 8 and Methods).

By incorporating supervised regression and feature selection techniques, PENCIL identifies high-confidence phenotype-associated subpopulations and orders them along a phenotypic trajectory. Exploiting this continuous resolution from the inferred trajectory, tools such as tradeSeq[27] or functions implemented in Monocle[28] can discover dynamic gene expression patterns depending on this phenotypic trajectory.

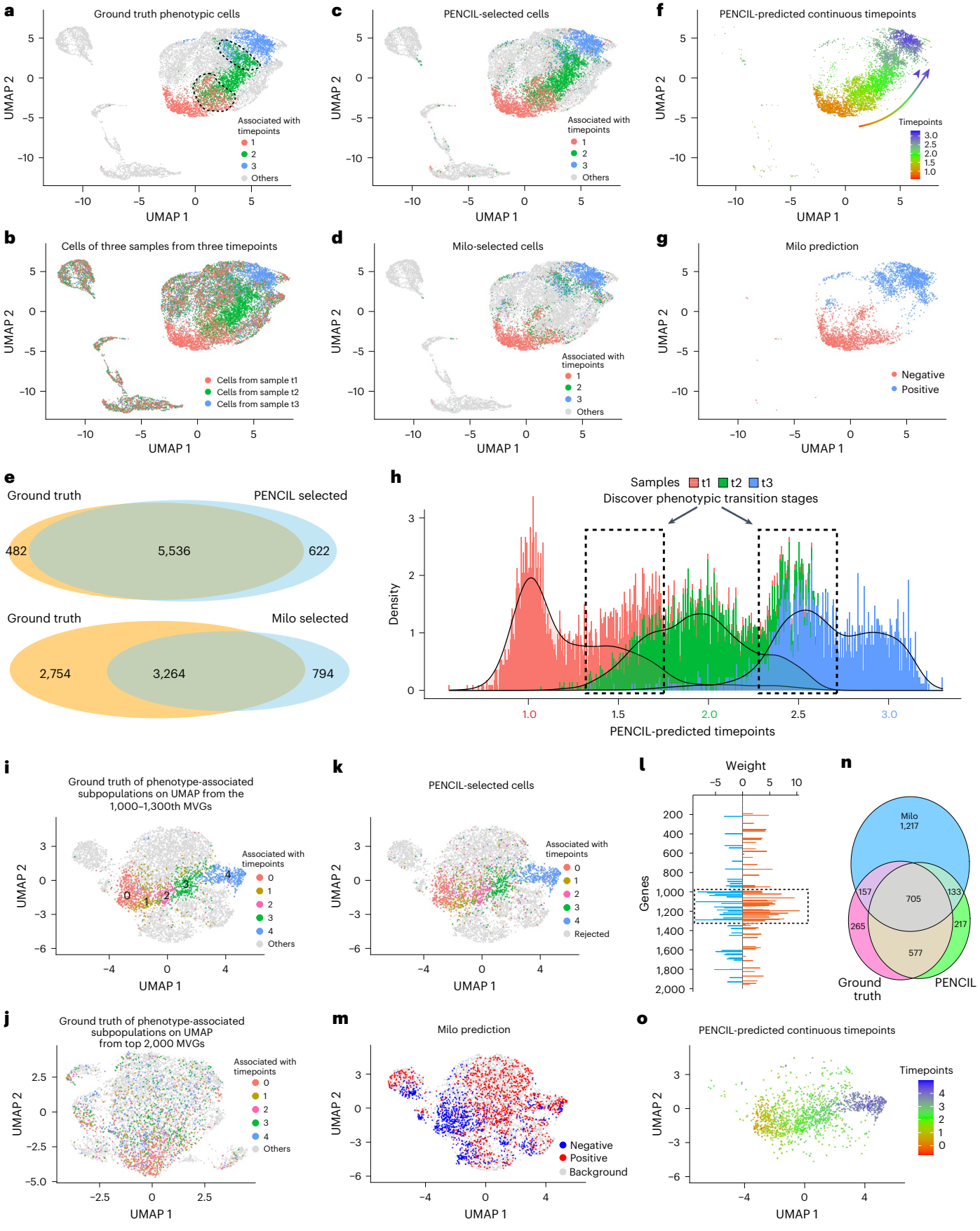
PENCIL speed and scalability

With the increasing number of sequenced cells[4,29], tools for efficient analysis of large-scale single-cell experiments are critical. Therefore, we implemented PENCIL using the powerful PyTorch framework. First, we simulated a large dataset with 1,000,000 cells and 2,000 genes from three conditions. We then randomly selected cells from this large dataset to simulate different datasets with varying cell numbers ranging

Fig. 3 | Evaluation of regression mode of PENCIL on the simulated datasets. **a**, For the first simulation, UMAP showing cells from a real scRNA-seq dataset assigned with three simulated ground truth phenotypic subpopulations and background cells. The regions within dashed lines indicate cells with labels evenly mixed by two adjacent timepoints. **b**, The three phenotypic subpopulations are assigned to the three samples accordingly, and all other cells are evenly assigned to the three samples to form the sample labels for all cells. **c**, PENCIL-selected cells. **d**, Milo-selected cells. **e**, Venn diagrams comparing the cells selected by PENCIL and Milo with the ground truth phenotypic cells, respectively. **f**, PENCIL-predicted continuous timepoints for the selected cells. **g**, Milo only assigned the selected cells as negatively and positively associated with the time course, corresponding to subpopulations decreasing and increasing with time, respectively. **h**, Histogram of PENCIL-predicted time scores of selected cells coloured by the sample labels. Dashed rectangles indicating potential transition stages. **i**, For the second simulation, UMAP from a manually pre-selected gene set (1,000–1,300th MVGs) to show cells with simulated ground truth of phenotype- associated subpopulations of five timepoints. **j**, Ground truth of phenotype-associated subpopulations in **i** visualized on UMAP using top 2,000 MVGs. **k**, PENCIL-selected cells. **l**, PENCIL-selected genes. The dashed rectangle region indicates the pre-selected gene set (1,000–1,300th MVGs) to set up the simulation in **i**. **m**, Milo-predicted cells increase and decrease with the time course. **n**, Venn diagram comparing the cells selected by PENCIL and Milo with the ground truth phenotypic cells. **o**, The PENCIL-predicted continuous timepoints for the selected cells in the second simulation.

https://doi.org/10.1038/s42256-023-00656-y

from 1,000 to 1,000,000 to evaluate PENCIL in both regression and classification modes. The elapsed time, central processing unit (CPU) and graphics processing unit (GPU) memory usages increase linearly with the number of input cells to PENCIL (Fig. 4). When the full set of 1,000,000 cells was analysed, the regression mode of PENCIL took less than 1 h, while the classification mode took 30 min. Both runtimes



https://doi.org/10.1038/s42256-023-00656-y

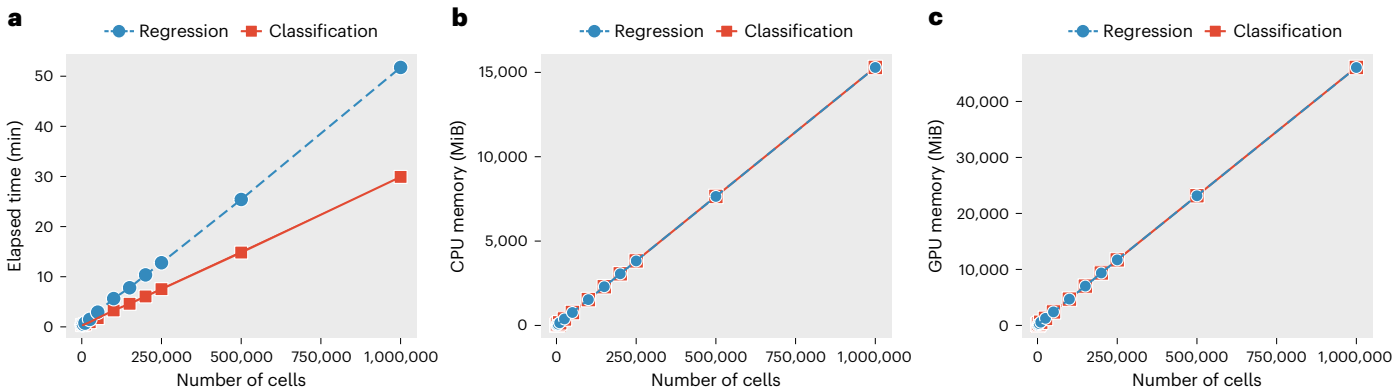


Fig. 4 | The running time and memory usages of PENCIL against the number of cells. a, Runtime of the PENCIL pipeline from inputting the normalized data to the final selected cells. **b**,**c**, Overall memory usage of CPU (**b**) and GPU (**c**) across the PENCIL workflow. MiB, mebibyte.

are acceptable for analysing such a large dataset (Fig. 4a). As CPU and GPU memory were used to load data, regression and classification modes used the same amount for the same number of input cells (Fig. 4b,c). Additionally, we found that the computational cost increases linearly with the number of input genes but remains constant against the number of conditions (Extended Data Fig. 9). The runtime evaluations were performed using an AMD EPYC 7502 32-core processor and an NVIDIA A100 GPU.

PENCIL identified subpopulations related to immunotherapy

For real data analysis, we first applied PENCIL to a $CD8^+$ T-cell scRNA-seq dataset (6,350 cells) from patients with melanoma consisting of 17 responders and 31 non-responders to immune checkpoint blockade (ICB) therapy[6] (Fig. 5a).

Targeting the ICB outcome phenotypes, the classification-based PENCIL identified 2,663 cells and 1,243 cells associated with non-responders and responders, respectively (Fig. 5b). Simultaneously, PENCIL selected 88 informative genes (Supplementary Fig. 1). On the basis of those selected genes, the UMAP exhibited a clear aggregation pattern for the PENCIL-selected cells (Fig. 5c). To catalogue transcription patterns underlying ICB outcomes, we executed a differentially expressed gene (DEG) analysis between the two subpopulations specific to ICB response and resistance. This analysis revealed 1,216 DEGs between the PENCIL-selected phenotypic subpopulations (Fig. 5d), which included 950 new DEGs in addition to the ones derived from the original all responder versus non-responder cells (Fig. 5d and Supplementary Table 1). Those DEGs unique to PENCIL are enriched in more than 200 pathways related to $CD8^+$ T cells (Supplementary Tables 2 and 3). Notably, the subpopulation associated with ICB responders has higher expressions of genes related to T-cell memory and survival, such as *IL7R*, *CCR7*, *LEF1*, *SELL* and *TCF7* (Fig. 5e,f). In contrast, the subpopulation associated with non-responders is marked by the expression of T-cell exhaustion and dysfunction genes such as *TOX*, *LAG3*, *PDCD1* and *CTLA4* (refs.30,31) (Fig. 5e,f).

Moreover, distinct from other strategies, our LWR-based supervised learning framework has an additional unique utility in that the trained PENCIL model from a given dataset can directly predict cell phenotypes for new single-cell samples. To demonstrate this utility, in the same dataset with 48 samples, we conducted a leave one patient out (LOPO) evaluation. In this approach, cells from the 47 patients were used to train the PENCIL model, which was applied to predict cell phenotypes from the left-out patient. This single-cell level phenotype prediction was further used to infer the patient-level phenotype by comparing the numbers of cells associated with the two phenotypes (Supplementary Fig. 2). We then considered each 'left-out' patient as a responder if more than 50% of cells were predicted as responder cells and evaluated this status against the actual clinical annotation. As a result, the patient-level inference based on single-cell level prediction by PENCIL correctly determined the ICB outcomes in 40 out of 48 samples (Fig. 5g) and achieved an area under precision–recall curve of 0.935 in the LOPO evaluation, which is comparable to the original study for the 48 samples[6]. In addition, given the PENCIL model trained on this T-cell melanoma ICB dataset, we applied it to three independent ICB-treated patients with melanoma with known ICB response status[32,33]. PENCIL rejected irrelevant cells and predicted the phenotypes for the selected cells (Fig. 5h–j). Then, by comparing the numbers of phenotypic cells of the two conditions predicted by PENCIL (Supplementary Fig. 2), we correctly inferred the patient-level ICB outcomes for the three new samples (Fig. 5h–j). Thus, we demonstrated a unique function of PENCIL to transfer labels to new samples, which further independently confirmed the performance of PENCIL for phenotype-enriched subpopulation analysis.

PENCIL learned trajectory of subpopulations upon treatment

As previously discussed, PENCIL's regression mode can resolve the phenotypic trajectory of subpopulations in a supervised manner that differs fundamentally from differential abundance analysis. To illustrate this utility in real data, we next applied regression-based PENCIL to a scRNA-seq dataset with samples collected from various timepoints.

In a clinical trial to evaluate a NEDD8-activating enzyme (NAE) inhibitor in treating a patient with mantle cell lymphoma (MCL)[34], a subtype of B-cell non-Hodgkin lymphoma, peripheral blood mononuclear cells were collected from the patient at baseline and after 3 and 24 h after drug infusion. Standard clustering of 3,236 peripheral blood mononuclear cells detected four clusters with three B-cell clusters and one CD4 cell cluster (Supplementary Fig. 3a). The largest B-cell-1 cluster with 2,329 cells can be characterized by the deletions of chromosomes 6 and 9 through inferCNV[35] analysis (Supplementary Fig. 3b), two recurrently affected genomic regions in MCLs[36]. Thus, we focused our analysis on the largest malignant B-cell cluster. In this cluster, standard clustering analysis based on the top 2,000 MVGs did not find any cluster dominated by a specific timepoint (Fig. 6a and Supplementary Fig. 3c,d). We then performed regression-mode PENCIL supervised by the cell labels 1, 2 and 3, corresponding to 0 h, 3 h and 24 h conditions, respectively. PENCIL identified high-confidence treatment-associated subpopulations, selecting 516 out of 1,064 cells, 445 out of 583 cells and 340 out of 682 cells from the 0 h, 3 h and 24 h conditions, respectively (Fig. 6b). Additionally, PENCIL selected 44 informative genes (Supplementary Fig. 3e). The UMAP plot based on these PENCIL selected genes clearly displayed the treatment response trajectory upon NAE inhibition (Fig. 6c,d). Then, correlating gene expressions with the predicted time orders of selected cells, we found 145 genes changing as cells

https://doi.org/10.1038/s42256-023-00656-y

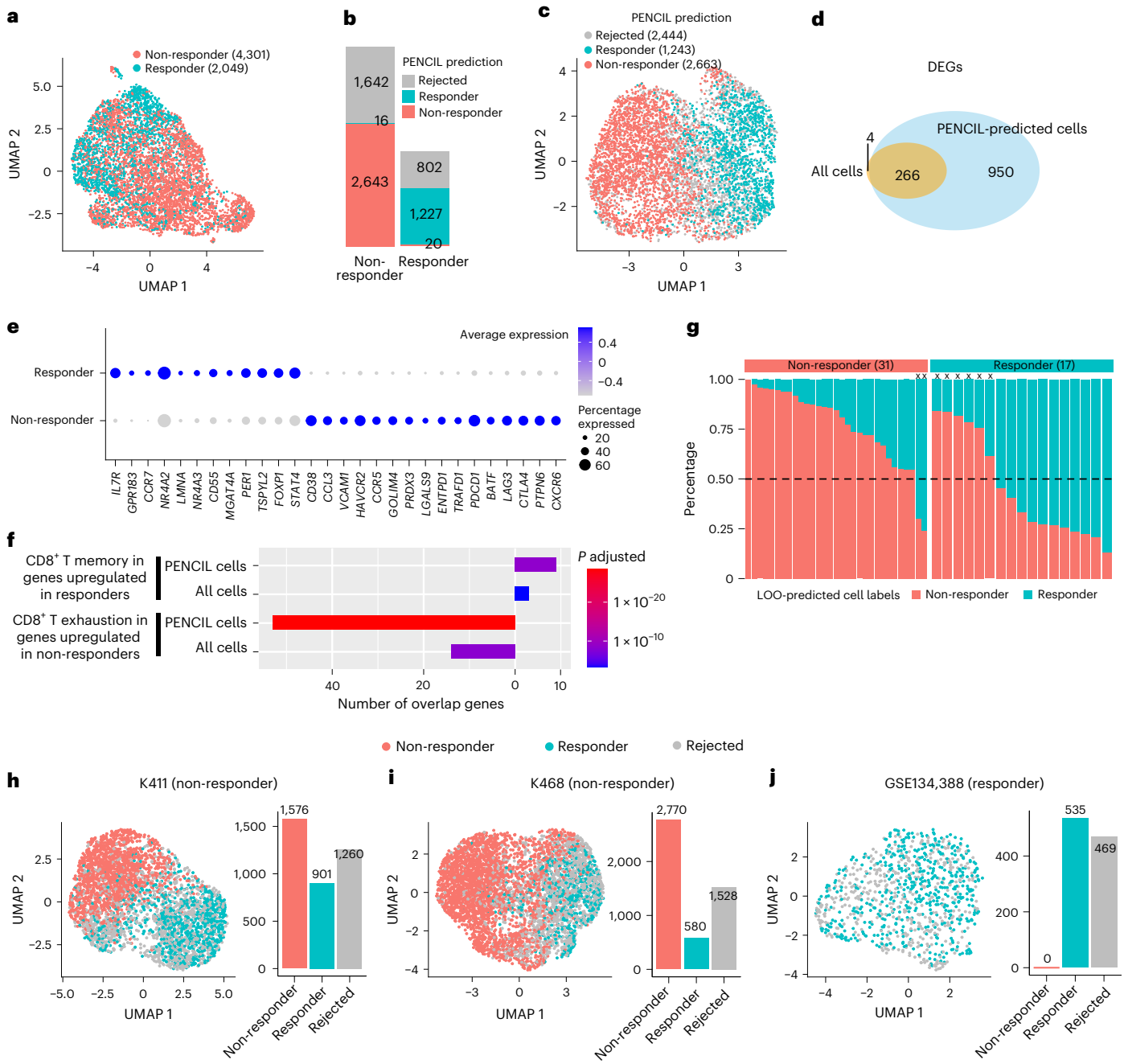


Fig. 5 | PENCIL analysis of T-cell subpopulations associated with melanoma immunotherapy outcomes. a, UMAP showing the cells using the top 2,000 MVGs. Cell number in parentheses. **b**, The PENCIL-predicted cell labels over the two conditions. **c**, PENCIL results on the UMAP based on PENCIL-selected genes. Cell number in parentheses. **d**, Venn diagram comparing the DEGs of two conditions using all cells and the DEGs of PENCIL-predicted labels of selected cells. **e**, Dot plots showing the expression levels of selected signature genes of PENCIL-predicted phenotypes. The size of the dot encodes the percentage of cells expressing each gene, and the colour encodes the average expression level. **f**, The enrichments of the two pathways in the signature genes identified from all cells and PENCIL-predicted cells. P.adjust: multiple testing-corrected *P* values from the one-sided hypergeometric test were calculated using the Benjamini–Hochberg method. **g**, LOPO prediction of responder and non-responder cells in the testing patient. The horizontal dashed line representing the cut-off to predict patients as responders or non-responders, and 'x' indicating the LOPO predictions inconsistent with the true condition. Sample number in parentheses. **h**–**j**, The PENCIL prediction results of three new melanoma samples are visualized on the UMAPs based on the genes selected by PENCIL when trained on the Sade-Feldman dataset. Bar plot showing the numbers of cell phenotypes predicted by PENCIL for each new sample. The sample name is placed at the top, along with the true ICB outcome indicated in parentheses.

progress along the treatment trajectory[17] (Fig. 6e and Supplementary Table 4). Specifically, *JUNB* and *JUN*, whose overexpression is a hallmark of lymphoma cells[37], had reduced expression following NAE inhibition (Fig. 6e). Overall, our PENCIL-predicted time course analysis resulted in more signature genes than the DEGs of each timepoint from all cells (Fig. 6f). For example, gene *JUND* is positively correlated with malignant cell proliferation in non-Hodgkin lymphoma[38], and PENCIL analysis found that NAE inhibitor repressed its expression along the predicted time course during treatment (Fig. 6g), which was not detected by the DEG analysis (Supplementary Fig. 3f).

https://doi.org/10.1038/s42256-023-00656-y

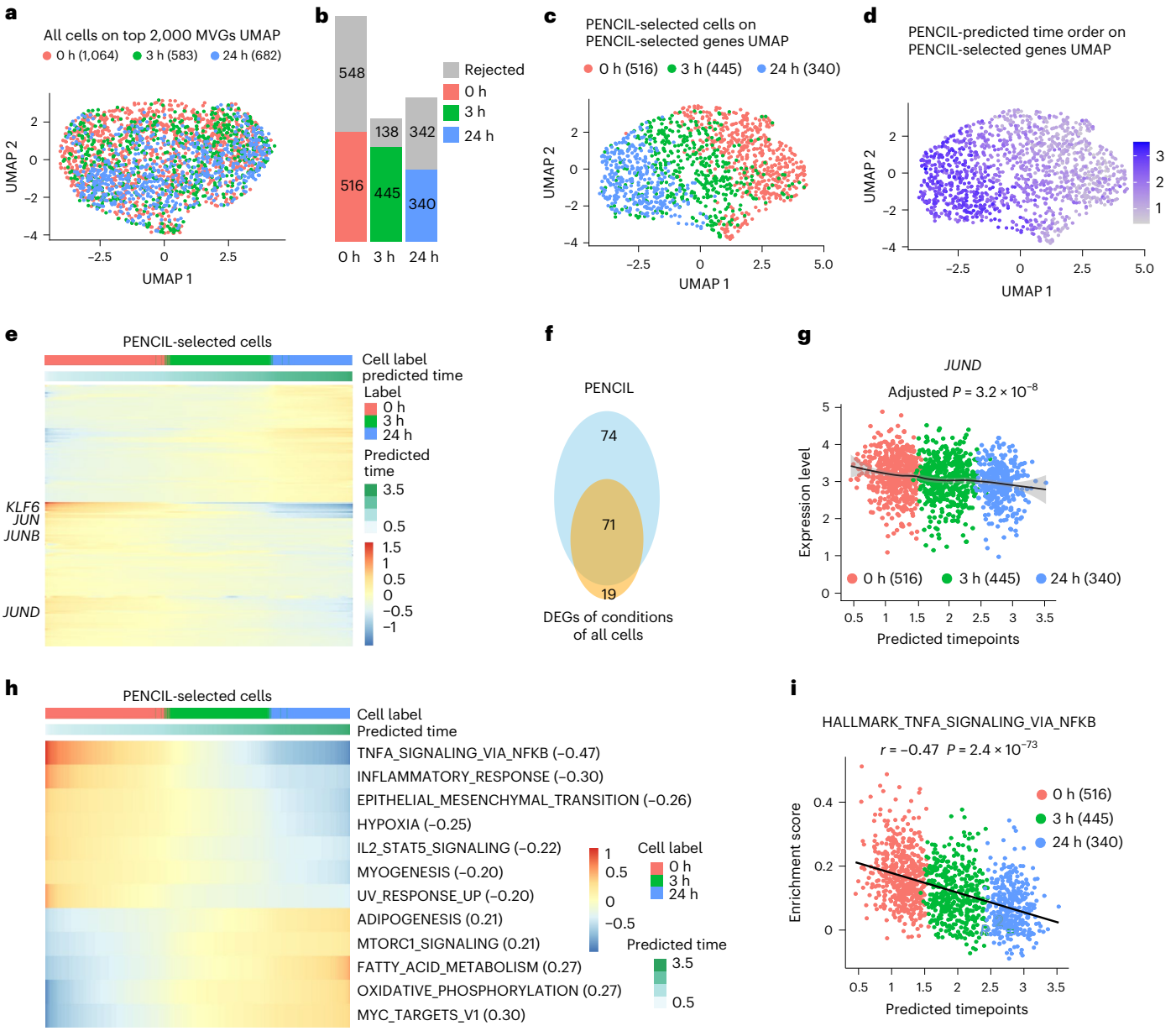


Fig. 6 | Regression mode of PENCIL analysis of scRNA-seq malignant B cells across three timepoints from a patient with MCL. a, UMAP based on the top 2,000 MVGs showing all cells of three conditions. Cell number in parentheses. **b**, PENCIL-selected cells across conditions. **c**, UMAP based on the PENCIL selected genes showing PENCIL-selected cells coloured by conditions. Cell number in parentheses. **d**, PENCIL-predicted time orders of PENCIL-selected cells on the same UMAP in **c**. **e**, Genes significantly associated with the PENCIL-predicted timepoints. **f**, Venn diagram comparing the DEGs of conditions using all cells and the genes associated with PENCIL-predicted time orders. **g**, The scatter plot shows JUND as an example of genes significantly associated with predicted timepoints that were not detected by the DEG analysis. The trend line was produced by the locally estimated scatter plot smoothing (LOESS) regression, and the error band represents the 95% confidence interval from the Student's *t*-distribution. The adjusted *P* value was calculated by the two-sided Wald test. **h**, Hallmark pathways significantly associated with the predicted time orders with absolute correlation values greater than 0.2. Pearson correlation values in parentheses. **i**, The scatter plot between the NFKB pathway activities and the predicted treatment timepoints predicted by PENCIL on cell subpopulations selected by PENCIL. The Pearson correlation coefficient and the corresponding *P* value from the two-sided *t*-test are indicated. The cell number is in parentheses.

Next, we explored the impacts of NAE inhibition at the pathway level. The proliferation and growth of MCL cells depend on NFKB signalling[39]. Interestingly, in our pathway analysis, the NFKB signalling pathway was the most negatively correlated with predicted time orders, suggesting NAE inhibition downregulated NFKB signalling along the trajectory to induce apoptosis in the MCL cells (Fig. 6h,i). This observation is consistent with our pre-clinical data that NAE inhibitor abrogates NFKB pathway activity in chronic lymphocytic leukaemia B cells[40].

Discussion

By leveraging supervised LWR, we have developed PENCIL to simultaneously select genes, select cells and predict categorical labels or continuous orders, thereby providing a new paradigm for identifying high-confidence phenotype-associated subpopulations from single-cell data (Extended Data Fig. 10).

The classification mode of PENCIL identifies subpopulations enriched by specific phenotypes, which has the same application as differential abundance testing algorithms. However, our supervised

learning-based PENCIL framework provides a more flexible way to select genes and identify subpopulations simultaneously. To demonstrate this unique feature, simulations for the comparison with other methods were designed to necessitate gene selection. However, we have to point out that our effort was not intended to develop a new method to improve the performance over existing methods incrementally, but to demonstrate that PENCIL can perform gene selection to assist subpopulation identification. Actually, when disabling the feature selection function, PENCIL and other methods performed similarly (Extended Data Fig. 7). Furthermore, the genes selected by PENCIL can be inputs for other methods to construct proper KNN graphs, which will be complementary to existing KNN-based approaches to improve their performances (Fig. 2g–i,o,p and Extended Data Fig. 7a,d) and utilize their advantages.

The extension of PENCIL to regression leads to novel applications in single-cell analysis. In the LWR framework, this switch in loss function will affect not only the predictor but also the rejector term, causing it to accept the cells transitioning between conditions (Fig. 1l,m), which fundamentally differs from differential abundance testing. Thus, the regression mode of PENCIL extends beyond detecting static categorical cell states to reveal transitions during dynamic biological processes. Even though Milo can evaluate continuous inputs, it tends to select subpopulations where phenotypic abundance monotonically changes, which usually misses phenotypic subpopulations in the middle of the time course (Fig. 3d,g). Most importantly, existing methods cannot assign time scores for the selected cells to reflect the dynamic course of phenotypes. Therefore, we believe the regression mode of PENCIL addresses a new application to supervised learning of the phenotypic trajectory of subpopulations.

PENCIL assigns cells from the same replicate with the same group label, so technical variability between samples is not considered, which is an inherited limitation in machine learning frameworks. In contrast, the statistics-based Milo can elegantly handle replication using the generalized linear model. Since PENCIL is complementary to other methods, we can provide the PENCIL-learned genes to Milo to exploit generalized linear model's statistical advantages. Furthermore, to address condition/sample unbalanced cell numbers, we introduced condition/sample weights to the loss function.

Although we only demonstrated the applications of PENCIL in scRNA-seq datasets, it can also handle other types of single-cell omics assays such as single-cell assay for transposase-accessible chromatin using sequencing[7,41–43].

Methods

Learning phenotype-associated high-confidence cell subpopulations by PENCIL

We build our method on the basis of a concept known as LWR, a machine learning strategy that introduces rejection labels in the prediction results (Fig. 1a,b). An insightful analysis of binary classification models with rejection was given in several previous studies[44–46], and a general learning model with rejection has also been implemented experimentally[47]. For this application, we further develop a more robust and theoretically supported generic rejection-based learning method and apply it to single-cell data analysis to identify phenotype-associated subpopulations with high confidence. Moreover, we incorporate feature selection into this LWR framework to achieve the unique function of simultaneously selecting genes and detecting phenotype-associated subpopulations from single-cell data.

The workflow of PENCIL is represented in Fig. 1c–g. The inputs for PENCIL are a quantified single-cell matrix and a label set of interest for each cell. Adhering to the general machine learning narrative conventions, let us denote the dataset combination to $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in R^d$ is the d-dimensional gene expression vector of the ith cell and y_i is the corresponding target label of the ith cell, such as condition, phenotype, stage and so on (Fig. 1c).

Let w be a trainable weight vector on genes, r_Φ be a learnable model called rejector parametrized by Φ to determine the confidence score for the cells ($r_\Phi(x) \le 0$ means the cell has low confidence and it will be rejected, and conversely, it will be accepted), h_Θ denote the predictor to be learned with parameters set Θ (Fig. 1e,f) and l be the learning loss function for a specific supervised learning task. For any sample (x,y) in dataset D, PENCIL's goal is to minimize the following rejection loss with gene weights (Fig. 1g):

$$L(h_\Theta, r_\Phi, w, x, y) = l(h_\Theta(x \odot w), y) 1_{r_\Phi(x\odot w)>0} + c 1_{r_\Phi(x\odot w)\le 0} + \lambda_1 ||w||_1 + \lambda_2 ||\Theta||_2,$$

where $\odot$ is the element-wise multiplication, $1_{r_\Phi>0}$ and $1_{r_\Phi\le 0}$ are indicator functions and c is the cost of rejection. We impose a sparse penalty (l_1-norm) on gene weights w to choose informative genes and l_2-norm on Θ to control the model complexity of the predictor h_Θ and enable PENCIL to pick out high-confidence cells that can be readily explained by a simple predictor.

The supervised loss l could come from a wide range of learning tasks, making PENCIL a flexible framework to be applicable in various scenarios. For example, if the target labels are multiple discrete categories, l can be a loss function for multi-classification (MC); thus, PENCIL can identify high-confidence cell subpopulations related to multi-conditions or phenotypes (Fig. 1j). When the labels are continuous variables, such as timepoints or disease stages, l can be a regression loss, so that PENCIL can determine a trajectory of selected cells highly correlated with the labels (Fig. 1m).

Differentiable surrogate and model setup

The total loss function L cannot be optimized directly using the gradient-like algorithm, due to the inclusion of indicators $1_{r_\Phi>0}$ and $1_{r_\Phi\le 0}$. We use $l(h_\Theta)$ to denote $l(h_\Theta(w \odot x), y)$ without causing ambiguity and temporarily ignoring the regularization terms. Drawing on the relaxation method in Cortes et al.[45]:

$$\begin{aligned} L(h_\Theta, r_\Phi, w, x, y) &= l(h_\Theta) 1_{r_\Phi>0} + c 1_{r_\Phi\le 0} \\ &= \max\left(l(h_\Theta) 1_{r_\Phi>0}, c 1_{r_\Phi\le 0}\right) \\ &\le \max\left(l(h_\Theta) 1_{-r_\Phi\le 0}, c 1_{r_\Phi\le 0}\right) \\ &\le \max\left(l(h_\Theta)\Psi(r_\Phi), c\Psi(-r_\Phi)\right) \\ &\le l(h_\Theta)\Psi(r_\Phi) + c\Psi(-r_\Phi), \end{aligned}$$

we can obtain the maximum surrogate (MS) and plus surrogate (PS) of L as:

$$L_{\mathrm{Rej}}^{\mathrm{MS}}(h_\Theta, r_\Phi, w, x, y) = \max\left(l(h_\Theta)\Psi(r_\Phi), c\Psi(-r_\Phi)\right)$$

$$L_{\mathrm{Rej}}^{\mathrm{PS}}(h_\Theta, r_\Phi, w, x, y) = l(h_\Theta)\Psi(r_\Phi) + c\Psi(-r_\Phi)$$

respectively, where $\Psi(\cdot)$ can be any one of the forms mentioned in Charoenphakdee et al.[48]. Furthermore, the total loss on the whole dataset D can be formulated as:

$$\hat{L}_{\mathrm{Rej}}(h_\Theta, r_\Phi, w, X, Y) = \hat{E}_{x,y\sim D}\left[L_{\mathrm{Rej}}(h_\Theta, r_\Phi, w, x, y)\right] = \frac{1}{N}\sum_{i=1}^{N} L_{\mathrm{Rej}}(h_\Theta, r_\Phi, w, x_i, y_i),$$

where $X = (x_1, \ldots, x_N)$, $Y = (y_1, \ldots, y_N)$, and $\hat{E}[\cdot]$ is the sample mean.

We substitute $w \odot x$ with x in the latter part for narrative simplicity. In the context of an MC task with M classes, the classifier $h_\Theta(x)$ is set to

a linear classifier:

$$o(x) = \theta_1 x + \theta_2$$

$$h_\Theta(x) = \mathrm{softmax}(o(x))$$

where $o(x) \in R^M$. And $r(x)$ is a two-layer neural network using the activation function $\sigma(x) = x \cdot \tanh(\mathrm{softplus}(x))$[49], that is:

$$r_\Phi(x) = \tanh(\varphi_3 \sigma(\varphi_1 x + \varphi_2) + \varphi_4) \in (-1, 1)$$

We use misclassification rate (MR) as the loss function for the MC task, and set $\Psi(r) = \mathrm{Sigmoid}(r) = \frac{1}{1+\exp(-r)}$ (ref. 48), and use PS-type rejection. So, for MC, our final implementation is:

$$L_{\mathrm{MC}}(h_\Theta, r_\Phi, w, x, y) \triangleq \frac{l_{\mathrm{MR}}(h_\Theta(x), y)}{1 + \exp(-r_\Phi(x))} + \frac{c}{1 + \exp(r_\Phi(x))}$$

where $l_{\mathrm{MR}}(h_\Theta(x), y) = 1 - h_\Theta(x)_y$; hence, the selection range of c can be restricted to $\left(0, \frac{1}{2}\right)$.

Though binary classification is a special case of MC and is included in MR, we have also implemented some other losses dedicated to binary classification, such as hinge loss[44,47].

In the regression (Reg) task, the regressor $h_\Theta(x)$ is set to a nonlinear neural network with a dropout layer:

$$h_\Theta(x) = \theta_3 \cdot \mathrm{dropout}(\sigma(\theta_1 x + \theta_2)) + \theta_4$$

while $h_\Theta(x) \in R$. The rejector $r_\Phi(x)$ is the same as one in the classification task. The loss function for regression is Huber loss, $\Psi(r) = \mathrm{Hinge}(r) = \max(1 + r, 0)$[48], and MS-type rejection is used, then:

$$L_{\mathrm{Reg}}(h_\Theta, r_\Phi, w, x, y) \triangleq \max(l_{\mathrm{Huber}}(h_\Theta(x), y)(1 + r_\Phi(x)), c(1 - r_\Phi(x)), 0),$$

where:

$$l_{\mathrm{Huber}}(h_\Theta(x), y) = \begin{cases} 0.5(h_\Theta(x) - y)^2, & |h_\Theta(x) - y| < 1, \\ |h_\Theta(x) - y| - 0.5, & \text{otherwise}, \end{cases}$$

which is insensitive to outliers and gives more robust regression results than mean square error loss.

Adjust cell numbers

In addition, we introduce class weights in the sample loss to overcome the class-imbalanced cell numbers, which is as follows:

$$\hat{L}_{u-\mathrm{MC}}(h_\Theta, r_\Phi, w, X, Y) = \frac{1}{\sum_{j=1}^{M} N_j u_j} \sum_{i=1}^{N} u_{I(i)} L_{\mathrm{MC}}(h_\Theta, r_\Phi, w, x_i, y_i),$$

where N_j is the number of cells in the jth class, u_j is the weight for the jth category and $I(i)$ indicates the index of the category to which cell i belongs. Similarly, we can also define the weight of each sample to adjust sample-imbalanced cell numbers to have higher weights to keep the cells from samples with smaller cell numbers.

Hyperparameter search

The rejection cost c is an important hyperparameter in the model. It directly affects the proportion of rejected cells and, hence, the final result. To eliminate the hassle of manual selection, we devised an algorithm to automatically select the hyperparameter c. The core principle is that, when the labels are disrupted, the result of the rejection model should reject the vast majority of cells. Otherwise, it implies that the current cost of rejection is excessive, that is, c is too large, and hence a smaller c should be picked. On the other hand, to reject as few samples as possible on the original dataset, the rejection cost should be as high as possible. Thereby, we can take as the final choice the maximum cost that can reject the majority of samples on the dataset when the labels are disrupted. This search process can be accomplished by a bisection flow as shown in Algorithm 1.

Algorithm 1.

Input: $c_{\max}$, $c_{\min}$, termination error bound ε, disruption rate r_d, and a small acceptance ratio threshold t.

Output: a proper cost of rejection c.

1. Randomly select Nr_d samples from the dataset D.
2. Randomly permute the labels of selected samples from step 1 → $(X, \tilde{Y})$.
3. While $c_{\max} - c_{\min} > \varepsilon$:
4. $\quad c = \frac{c_{\max} + c_{\min}}{2}$
5. $\quad$ Train the rejection model on the disrupted dataset $(X, \tilde{Y})$ with cost c.
6. $\quad$ Count the samples non-rejected → n.
7. $\quad$ If $\frac{n}{N} > t$:
8. $\quad\quad c_{\min} = c$.
9. $\quad$ Else:
10. $\quad\quad c_{\max} = c$.
11. Return $c_{\min}$.

Pre-train for faster convergence. The prediction model pre-trained on a purely learning task without the rejection module can converge faster in subsequent training. So, we first optimize $l(h_\Theta)$ to pre-train the predictor h_Θ, and then optimize the rejection loss $\mathcal{L}$ to train $h_\Theta(x)$ and $r_\Phi(x)$.

PENCIL implementation. PENCIL is implemented in Python and employs the powerful PyTorch framework to accelerate the optimization with GPUs, allowing direct integration with other Python-based single-cell analysis platforms such as SCNAPY[50]. Alternatively, data pre-processed by R packages such as Seurat can be saved as intermediate files for loading into Python. To streamline the analysis, we incorporated both native R and Python codes into a single document using 'R Markdown', which allows us to seamlessly transfer objects between them. Thus, PENCIL can easily interact with Seurat[25] and SCANPY[50], two popular single-cell analysis frameworks. We provided tutorials on GitHub to run PENCIL with SCANPY and Seurat.

Simulation setup

In simulations for the classification mode of PENCIL, we exploited a real T-cell scRNA-seq dataset[6] with 6,350 cells and 55,737 genes. Since scRNA-seq data are noisy and sparse, we selected the top 2,000 MVGs using the default function in the Seurat[25] package as the source data for PENCIL and other methods. For specific simulations with two or three conditions, we first selected a subset of genes from the top 2,000 MVGs for downstream clustering and visualization in UMAP to generate ground truth phenotypic subpopulations. For example, as shown in Fig. 2a,m, the 1,000–1,300th MVGs were manually pre-selected as informative genes, then all cells were visualized and clustered on the basis of the expressions of these pre-selected genes to generate ground truth phenotypic subpopulations. After that, we picked out two or three clusters and designated them to be enriched in specific conditions, respectively. All other cells were set as background cells. Next, we assigned simulated sample labels to the cells based on the ground truth phenotypic subpopulations and background cells. We used a number α called mixing rate to control the ratio between the majority and minority sample labels. Within each ground truth phenotypic condition, we assigned $(1 - \alpha)$ of the total cells of this condition with the designated majority condition labels, and the remaining cells with other labels. For the background cells, each cell was randomly assigned a condition label. In this way, we generated condition labels for all cells

for our analysis, as shown in Fig. 2b with a mixing rate $\alpha = 0.1$. We also depict this simulation process in Extended Data Fig. 3.

Second, to repeat simulations multiple times, we randomly selected 300 key genes from the top 1,000 MVGs and subsequently clustered cells according to these pre-selected key genes. After that, we randomly picked out two or three clusters and designated them as the ground truth of phenotype-enriched subpopulations and placed other cells as background cells. Next, using the same procedure as before, we generated condition labels for cells according to their designated ground truth phenotypes for four mixing rates (Fig. 2l). We performed label assignments for four mixing rates to mimic the varying components within subpopulations. We utilized precision, recall and F1 scores between the identified cells and ground truth cells to evaluate the four methods.

For the simulation with batch effects, we employed Splatter[24] to simulate an expression matrix with 9,000 cells and 8,000 genes in two batches. Six-thousand of these cells are from one batch, and 3,000 are from the other batch. And these cells are from three simulated groups with group probability of 0.6, 0.6 and 0.2. The probabilities of differential gene expression among the three groups were set as 0.1, 0.1 and 0.1. To produce expression data that necessitate gene selection, we selected 500 genes and disrupted them six times along the cell orientation, resulting in 3,000 highly variable random noisy genes. Then, we merged these noisy genes with the original remaining 7,500 genes into a new gene expression matrix of size 10,500 × 9,000. Following the default Seurat pipeline for finding MVGs[25], we got the new top 3,000 MVGs. As expected, most of these 3,000 genes are shuffled noisy genes, and only a very small fraction of them are key genes differentiating ground truth phenotype-associated subpopulations. Simulated groups can be completely separated under these differential genes (Extended Data Fig. 5a) and batch correction using Seurat revealed the three simulated groups (Extended Data Fig. 5b). However, this did not work when using the top 3,000 MVGs (Extended Data Fig. 5c). Thus, we obtained a simulated expression matrix comprising potential key genes, groups and batches. Next, we generated the condition labels for all cells by setting the cells of group 1 as background cells, cells of group 2 and group 3 as two ground truth phenotypic conditions and labelled them accordingly with a mixing rate of 0.1. After batch removal by Seurat[51], using the batch-corrected and scaled expression matrix as an input, PENCIL selected the genes (Extended Data Fig. 5d) and identified 91.0% of the ground truth cells with a precision 0.914, as shown in the UMAP generated from the PENCIL-selected genes and Venn diagram (Extended Data Fig. 5e,f). To repeat this simulation, we conducted the simulations 50 times with 4 mixing rates and showed that PENCIL has better performance than other methods (Extended Data Fig. 5g).

In the simulations for the regression mode of PENCIL analyses, we employed two types of single-cell expression data. In the first simulation, we used a scRNA-seq dataset pre-processed by PC analysis dimensional reduction[9], which comprises 16,291 cells and 10 PCs. On the basis of these PCs, we performed clustering and UMAP visualization following the standard pipeline in the Seurat[25] package and selected five clusters (denoted as clusters 1–5) as the ground truth trajectory (Extended Data Fig. 8a). We then set timepoint labels for each of these selected clusters, where clusters 1, 3 and 5 were assigned timepoint labels of t1, t2 and t3 respectively, while clusters 2 and 4 are set to be an equal mix of the two adjacent timepoint labels to mimic the transition stages (Fig. 3a). All of other cells were set as the background cells, which were randomly assigned timepoint labels as noise (Fig. 3b). Then, we used the expression matrix with ten PCs along with the simulated timepoint labels to perform PENCIL analysis without the feature selection function. In the second simulation, because we wanted to demonstrate the feature selection of PENCIL in the regression mode, we employed the raw gene-level expression scRNA-seq matrix that was used in the classification tasks. We still pre-selected a subset of genes to necessitate gene selection, which was further used for clustering and UMAP visualization to generate the ground truth subpopulations. For example, the top 1,000th–1,300th MVGs were used for clustering and UMAP visualization, which was used to select the clusters as ground truth subpopulations for the simulation case shown in Fig. 3i. All cells' timepoint labels were set up similarly as before by assigning timepoint labels according to their designated condition labels. To further demonstrate the regression mode of PENCIL's capability in simultaneous feature selection, cell selection and continuous timepoint prediction, we performed two more simulation cases by manually pre-selecting different key genes (Extended Data Fig. 8e–n).

PENCIL for analysing simulations with more than two conditions. As noted before, PENCIL can naturally be extended to address more than two conditions. Therefore, we did similar evaluations on simulation datasets with three conditions (Fig. 2m and Extended Data Fig. 6a–c) using the same T-cell scRNA-seq dataset[6] as the two conditions. For the comparisons, we included Milo and MELD because they can easily address more than two conditions, whereas DAseq can handle only two conditions. Consistently, PENCIL outperformed other methods with 0.815 recall and 0.884 precision (Fig. 2n and Extended Data Fig. 6d,e), compared with 0.816, 0.001 (recall) and 0.418, 0.176 (precision) for Milo and MELD (Fig. 2o,p and Extended Data Fig. 6f,g), respectively. A total of 80.4% of the PENCIL-selected genes came from the manually pre-selected genes (1,000–1,300th MVGs), which were used to generate this simulation (Fig. 2q), confirming its capability in feature selection to facilitate subpopulation identification. We repeated this simulation with three experimental conditions 50 times, demonstrating better performance for PENCIL than other methods (Fig. 2r).

Regression mode of PENCIL with simultaneous gene selection in simulations. We examined the gene selection function of PENCIL in the regression task. We employed the same scRNA-seq data of T cells[6] in the classification tasks to simulate a time-series dataset. First, like in the previous experiment, we picked a subset of genes (the top 1,000–1,300th MVGs) from the top 2,000 MVGs for clustering and UMAP visualization to set up the simulated ground truth. Then we selected five subpopulations as ground truth cells for five timepoints and background cells based on the clusters generated from the selected genes (Fig. 3i). The standard top 2,000 MVGs-based analysis cannot correctly capture the structures of the five ground truth subpopulations (Fig. 3j). Then, we assigned the condition labels accordingly for phenotypic subpopulations and randomly assigned condition labels for background cells (Extended Data Fig. 8c). With the top 2,000 MVGs expression matrix and the simulated labels as input source data, the regression mode of PENCIL found most of the ground truth cells (Fig. 3k and Extended Data Fig. 8d) and the genes learned by PENCIL mainly located in the pre-defined 1,000–1,300th MVG regions, as indicated by the dashed rectangle (Fig. 3l). In contrast, Milo selected many false positive cells (Fig. 3m). Specifically, PENCIL achieved 0.75 sensitivity and 0.79 precision, while Milo achieved 0.51 sensitivity and 0.39 precision (Fig. 3n). As before, the regression model of PENCIL can predict continuous timepoints for selected cells to construct the trajectory (Fig. 3o). Additional simulations can be found in Extended Data Fig. 8e–n.

Evaluation metric based on precision, recall and F1 score. In all simulations, the ground truth benchmark is defined as the groups of cells that generate phenotype-associated subpopulations. The true positive (TP) is the number of cells that are identified by both the evaluated methods and the ground truth cell set. The false positive (FP) is the number of cells selected by the methods but not included in the ground truth. The false negative (FN) is the number of cells rejected by the methods but belonging to the ground truth. Then, we use precision, recall and F1 score to assess the performance of all methods, where precision is defined as TP/(TP + FP), recall is defined as TP/(TP + FN) and the F1 score is the harmonic mean of precision and recall, calculated by (2 × precision × recall)/(precision + recall).

Genes significantly associated with predicted timepoints. We employed the functions implemented in Monocle3 (v1.2.9)[28] to identify genes significantly dependent on the timepoints predicted by PENCIL's regression mode. The gene expression levels were first fitted to the timepoints. Then, the Wald test calculated the *P* value by checking whether each coefficient differs significantly from zero, which was further adjusted by the Benjamini and Hochberg method[28]. The genes were considered significant if their adjusted *P* values were less than 0.05.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Publicly available scRNA-seq studies can be accessed via the following accession numbers or the link provided: GSE120575 (ref. 6), GSE159251 (ref. 32), GSE134388 (ref. 33) and https://zenodo.org/record/7761954 (ref. 34). More detailed description of these datasets can be found in Supplementary Material.

Code availability

The open-source PENCIL program and its tutorials are freely available at GitHub (https://github.com/cliffren/PENCIL) and Zenodo (https://doi.org/10.5281/zenodo.7762054).

References

1. Miao, Y. et al. Adaptive immune resistance emerges from tumor-initiating stem cells. *Cell* **177**, 1172–1186 e1114 (2019).
2. Wagner, J. et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **177**, 1330–1345 e1318 (2019).
3. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
4. Stephenson, E. et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
5. Ekiz, H. A. et al. MicroRNA-155 coordinates the immunological landscape within murine melanoma and correlates with immunity in human cancers. *JCI Insight* **4**, e126543 (2019).
6. Sade-Feldman, M. et al. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* **175**, 998–1013 e1020 (2018).
7. Eksi, S. E. et al. Epigenetic loss of heterogeneity from low to high grade localized prostate tumours. *Nat. Commun.* **12**, 7292 (2021).
8. Lun, A. T. L., Richard, A. C. & Marioni, J. C. Testing for differential abundance in mass cytometry data. *Nat. Methods* **14**, 707–709 (2017).
9. Zhao, J. et al. Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proc. Natl Acad. Sci. USA* **118**, e2100293118. (2021).
10. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using *k*-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
11. Burkhardt, D. B. et al. Quantifying the effect of experimental perturbations at single-cell resolution. *Nat. Biotechnol.* **39**, 619–629 (2021).
12. Sheng, J. & Li, W. V. Selecting gene features for unsupervised analysis of single-cell gene expression data. *Brief. Bioinform.* **22**, bbab295 (2021).
13. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).
14. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
15. Zhong, S. et al. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
16. Baran-Gale, J. et al. Ageing compromises mouse thymus function and remodels epithelial cell differentiation. *eLife* **9**, e56221 (2020).
17. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
18. Chen, H. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).
19. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
20. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
21. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4314 (2018).
22. Cannoodt, R., Saelens, W., Deconinck, L. & Saeys, Y. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nat. Commun.* **12**, 3942 (2021).
23. Chen, W. et al. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. *Nat. Biotechnol.* **39**, 1103–1114 (2021).
24. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
25. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 e3529 (2021).
26. Ruan, X. et al. Progenitor cell diversity in the developing mouse neocortex. *Proc. Natl Acad. Sci. USA* **118**, e2018866118 (2021).
27. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
28. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
29. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
30. Li, H. et al. Dysfunctional CD8 T cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell* **176**, 775–789 e718 (2019).
31. Scott, A. C. et al. TOX is a critical regulator of tumour-specific T cell differentiation. *Nature* **571**, 270–274 (2019).
32. Pauken, K. E. et al. Single-cell analyses identify circulating anti-tumor CD8 T cells and markers for their enrichment. *J. Exp. Med.* **218**, e20200920 (2021).
33. Li, N. et al. ALKBH5 regulates anti-PD-1 therapy response by modulating lactate and suppressive immune cell accumulation in tumor microenvironment. *Proc. Natl Acad. Sci. USA* **117**, 20159–20170 (2020).
34. Torka, P. et al. Pevonedistat, a Nedd8-activating enzyme inhibitor, in combination with ibrutinib in patients with relapsed/refractory B-cell non-Hodgkin lymphoma. *Blood Cancer J.* **13**, 9 (2023).
35. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. inferCNV of the Trinity CTAT Project. *Klarman Cell Observatory, Broad Institute of MIT and Harvard.* https://github.com/broadinstitute/inferCNV (2019).
36. Hartmann, E. M. et al. Pathway discovery in mantle cell lymphoma by integrated analysis of high-resolution gene expression and copy number profiling. *Blood* **116**, 953–961 (2010).
37. Mathas, S. et al. Aberrantly expressed c-Jun and JunB are a hallmark of Hodgkin lymphoma cells, stimulate proliferation and synergize with NF-kappa B. *EMBO J.* **21**, 4104–4113 (2002).
38. Papoudou-Bai, A. et al. The expression levels of JunB, JunD and p-c-Jun are positively correlated with tumor cell proliferation in diffuse large B-cell lymphomas. *Leuk. Lymphoma* **57**, 143–150 (2016).

39. Balaji, S. et al. NF-kappaB signaling and its relevance to the treatment of mantle cell lymphoma. *J. Hematol. Oncol.* **11**, 83 (2018).
40. Godbersen, J. C. et al. The Nedd8-activating enzyme inhibitor MLN4924 thwarts microenvironment-driven NF-kappaB activation and induces apoptosis in chronic lymphocytic leukemia B cells. *Clin. Cancer Res.* **20**, 1576–1589 (2014).
41. Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
42. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
43. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
44. Bartlett, P. L. & Wegkamp, M. H. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.* **9**, 1823–1840 (2008).
45. Cortes, C., DeSalvo, G. & Mohri, M. Learning with Rejection. *Lect. Notes Artif. Intell.* **9925**, 67–82 (2016).
46. Herbei, R. & Wegkamp, M. H. Classification with reject option. *Can. J. Stat.* **34**, 709–721 (2006).
47. Asif, A. & Minhas, F. U. A. Generalized neural framework for learning with rejection. *International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/IJCNN48605.2020.9206612 (IEEE, 2020).
48. Charoenphakdee, N., Cui, Z. H., Zhang, Y. A. & Sugiyama, M. Classification with rejection based on cost-sensitive classification. *Proc. Mach. Learn. Res.* **139**, 1507–1517 (2021).
49. Misra, D. Mish: a self regularized non-monotonic activation function. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1908.08681 (2019).
50. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
51. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 e1821 (2019).

Acknowledgements

This work was supported by the following funding: the National Key Research and Development Program of China 2020YFA0712400 (to T.R. and L.-Y.W.); NIH 1R21HL145426 (to Z.X.); Department of Defense Idea Development Award W81XWH2110539 (to Z.X.); Breast Cancer Research Foundation and NIH U01CA253472 and U01CA217842 (to G.B.M.); NIH 1R01CA244576 (to A.V.D.); NIH R35GM124704 (to A.C.A.); NIH R01CA250917 (to M.H.S.). We thank J. Zeng (University of Macau), and all the members of his bioinformatics team for generously sharing their experience and codes. We thank W. Anderson for helping edit the manuscript.

Author contributions

Z.X. conceived the idea. T.R., L.-Y.W. and Z.X. implemented the method and performed the analyses. T.R., C.C., A.V.D, S.L., X.G., S.D., L.-Y.W. and Z.X. interpreted the results. X.W., M.H.S., A.C.A., P.T.S., L.M.C. and G.B.M. provided scientific insights on the applications. A.C.A. and G.B.M. contributed to the analytic strategies. L.-Y.W. and Z.X. supervised the study. T.R., L.-Y.W. and Z.X. wrote the manuscript with feedback from all other authors. All the authors read and approved the final manuscript.

Competing interests

A.V.D. has received consulting fees from Abbvie, AstraZeneca, Bayer Oncology, BeiGene, Bristol Meyers Squibb, Genentech, Incyte, Lilly Oncology, Morphposys, Nurix, Oncovalent, Pharmacyclics and TG Therapeutics and has ongoing research funding from Abbvie, AstraZeneca, Bayer Oncology, Bristol Meyers Squibb, Cyclacel, MEI Pharma, Nurix and Takeda Oncology. X.G. is a Genentech employee and Roche shareholder. G.B.M. is SAB/Consultant for AstraZeneca, BlueDot, Chrysallis Biotechnology, Ellipses Pharma, ImmunoMET, Infinity, Ionis, Lilly, Medacorp, Nanostring, PDX Pharmaceuticals, Signalchem Lifesciences, Tarveda, Turbine and Zentalis Pharmaceuticals; stock/options/financial: Catena Pharmaceuticals, ImmunoMet, SignalChem, Tarveda and Turbine; licenced technology: HRD assay to Myriad Genetics, and DSP patents with Nanostring. L.M.C. provides consulting services for Cell Signaling Technologies, AbbVie, the Susan G Komen Foundation and Shasqi, received reagent and/or research support from Cell Signaling Technologies, Syndax Pharmaceuticals, ZelBio Inc., Hibercell Inc. and Acerta Pharma, and participates in advisory boards for Pharmacyclics, Syndax, Carisma, Verseau, CytomX, Kineta, Hibercell, Cell Signaling Technologies, Alkermes, Zymeworks, Genenta Sciences, Pio Therapeutics Pty Ltd, PDX Pharmaceuticals, the AstraZeneca Partner of Choice Network, the Lustgarten Foundation and the NIH/NCI-Frederick National Laboratory Advisory Committee. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00656-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00656-y.

Correspondence and requests for materials should be addressed to Ling-Yun Wu or Zheng Xia.

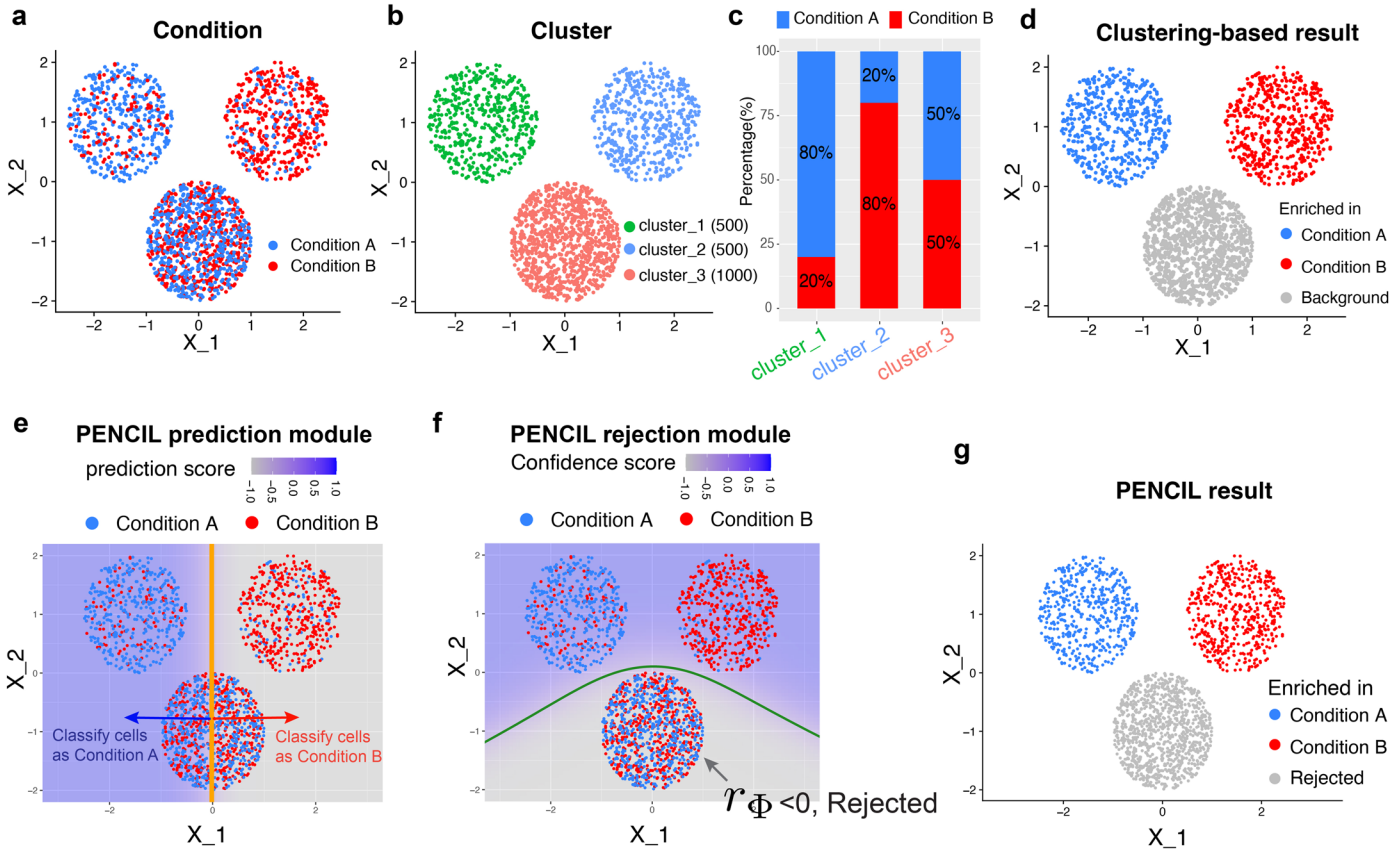
Peer review information *Nature Machine Intelligence* thanks Yun Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

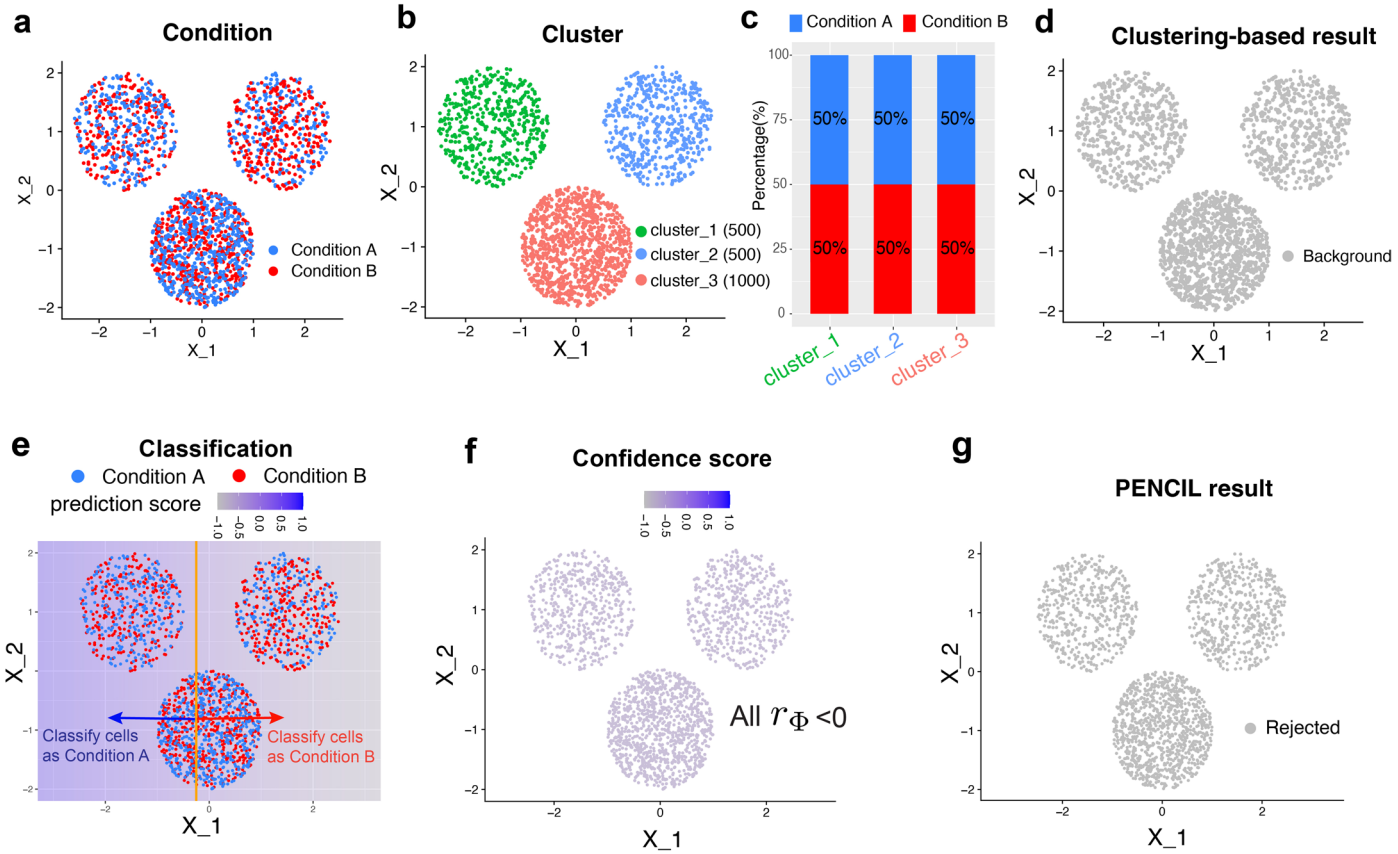
[1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. [2]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China. [3]Computational Biology Program, Oregon Health & Science University, Portland, OR, USA. [4]Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA. [5]City of Hope National Medical Center, Duarte, CA, USA. [6]Department of Oncology Biomarker Development, Genentech Inc, South San Francisco, CA, USA. [7]Department of Cell, Developmental & Cancer Biology, Oregon Health & Science University, Portland, OR, USA. [8]Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. [9]Cancer Biology & Genetics Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [10]Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. [11]Division of Oncological Sciences Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. ✉e-mail: lywu@amss.ac.cn; xiaz@ohsu.edu

https://doi.org/10.1038/s42256-023-00656-y



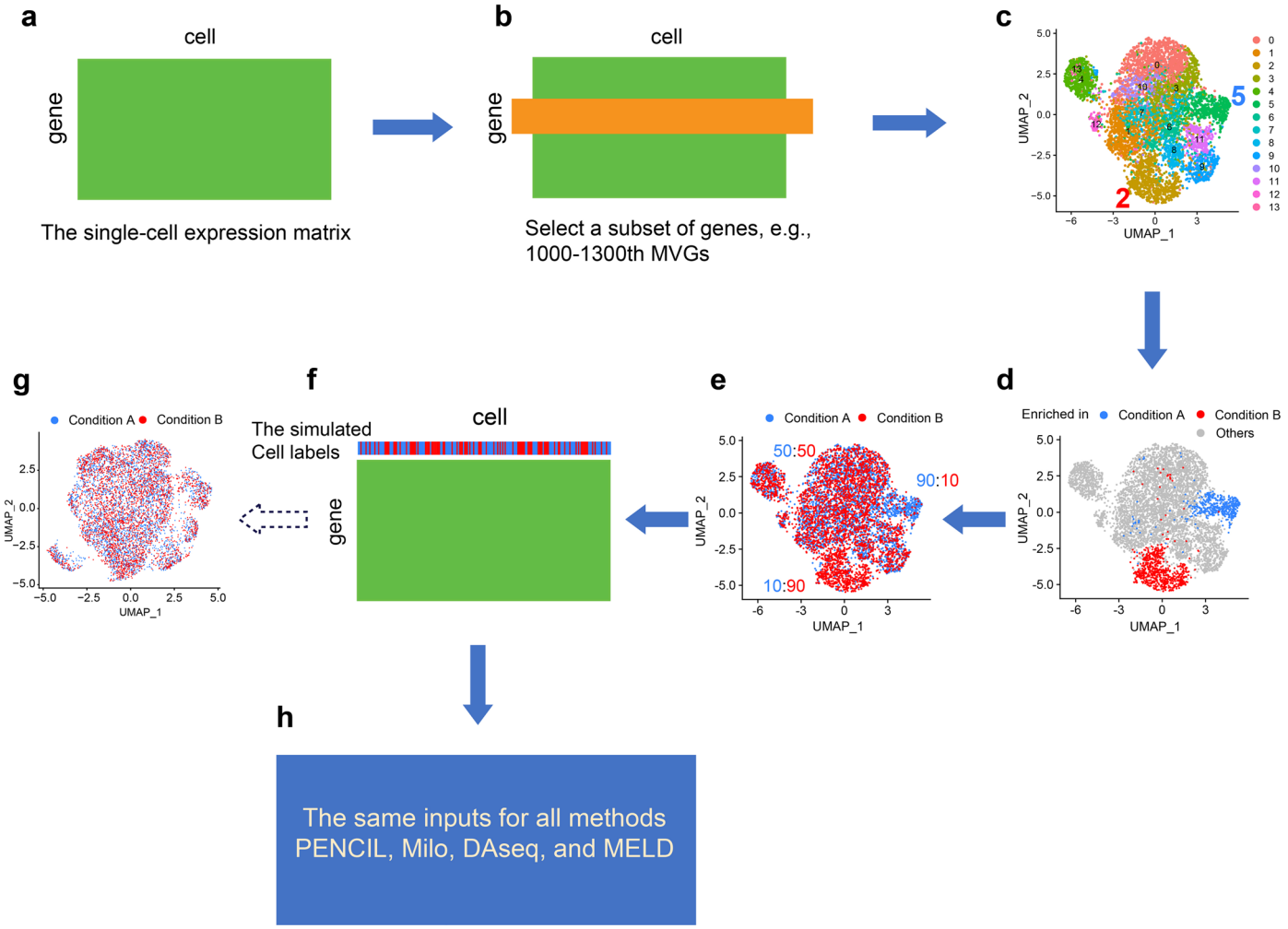
Extended Data Fig. 1 | A simple simulation consists of cells from two conditions and three cell types, each containing only two genes (X_1 and X_2). a, Visualizing cells from two conditions colored by condition labels using the two genes. **b**, Standard clustering of the cells. Cell number in parentheses. **c**, Percentage of cell condition labels within each cluster. **d**, The identified phenotypic subpopulations from the clustering-based method. **e**, The learned prediction model from PENCIL with the orange line as the boundary with prediction scores $h(x) = 0$ to classify the two conditions. Cells colored by the condition labels as in **a**. **f**, The learned rejection model from PENCIL with the green curve as the boundary with confidence scores $r(x) = 0$ to reject cells. Cells colored by the condition labels as in **a**. **g**, PENCIL identified phenotypic subpopulations.

https://doi.org/10.1038/s42256-023-00656-y



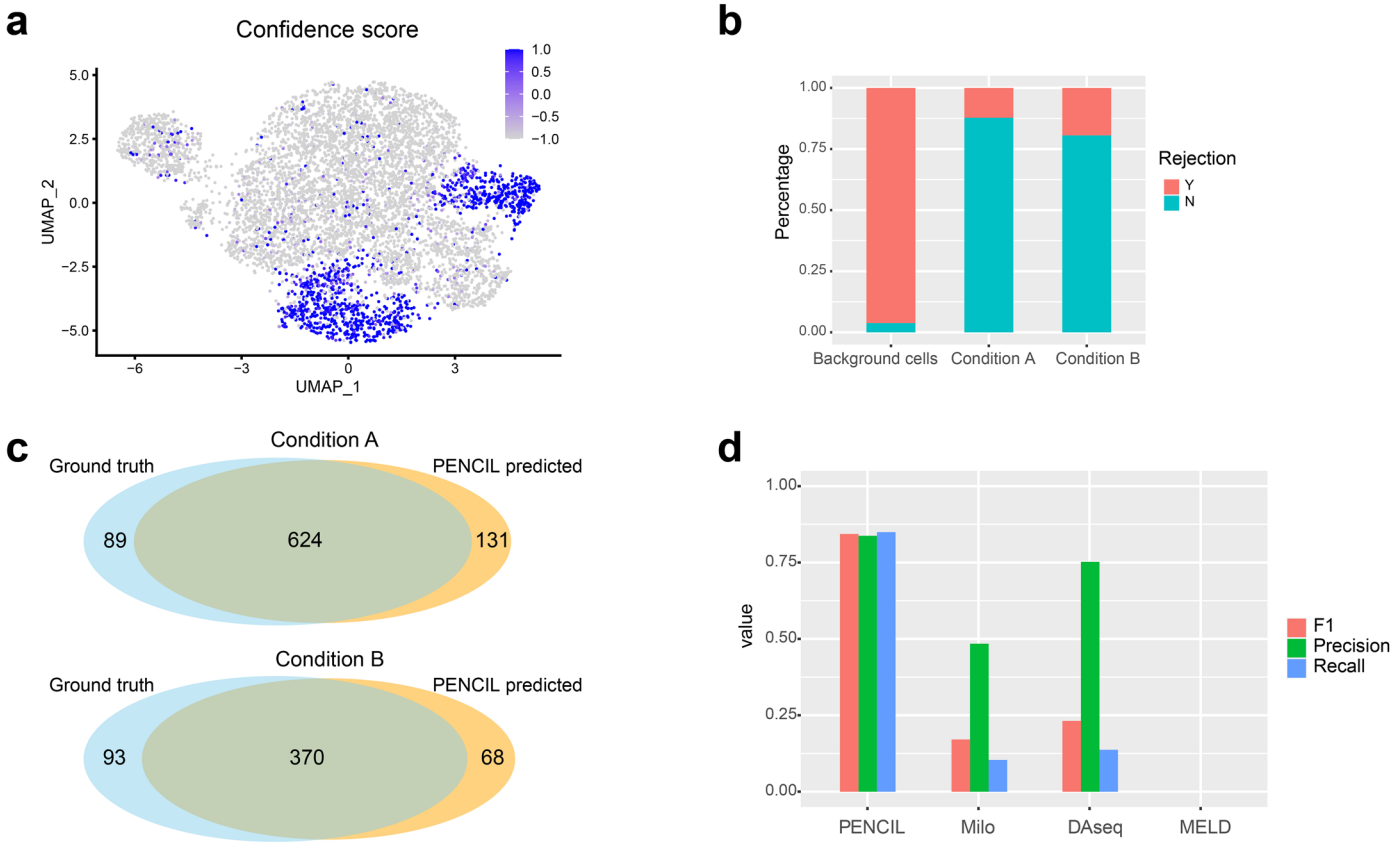
Extended Data Fig. 2 | A simple simulation includes cells from two conditions and three cell types, each containing only two genes (X_1 and X_2), but lacks enriched phenotypic subpopulations. a, Visualizing cells from two conditions colored by condition labels using the two genes. **b**, Standard clustering of the cells. Cell number in parentheses. **c**, The equal percentages of cell condition labels within each cluster. **d**, The result of the clustering-based method showing no subpopulations associated with the phenotypes. **e**, The learned prediction model from PENCIL with the orange line as the boundary with prediction scores $h(x) = 0$ to classify the two conditions. Cells colored by the condition labels as in **a**. **f**, The rejection module in PENCIL with all confidence scores $r(x) < 0$ to reject all cells. **g**, PENCIL rejected all cells.

https://doi.org/10.1038/s42256-023-00656-y



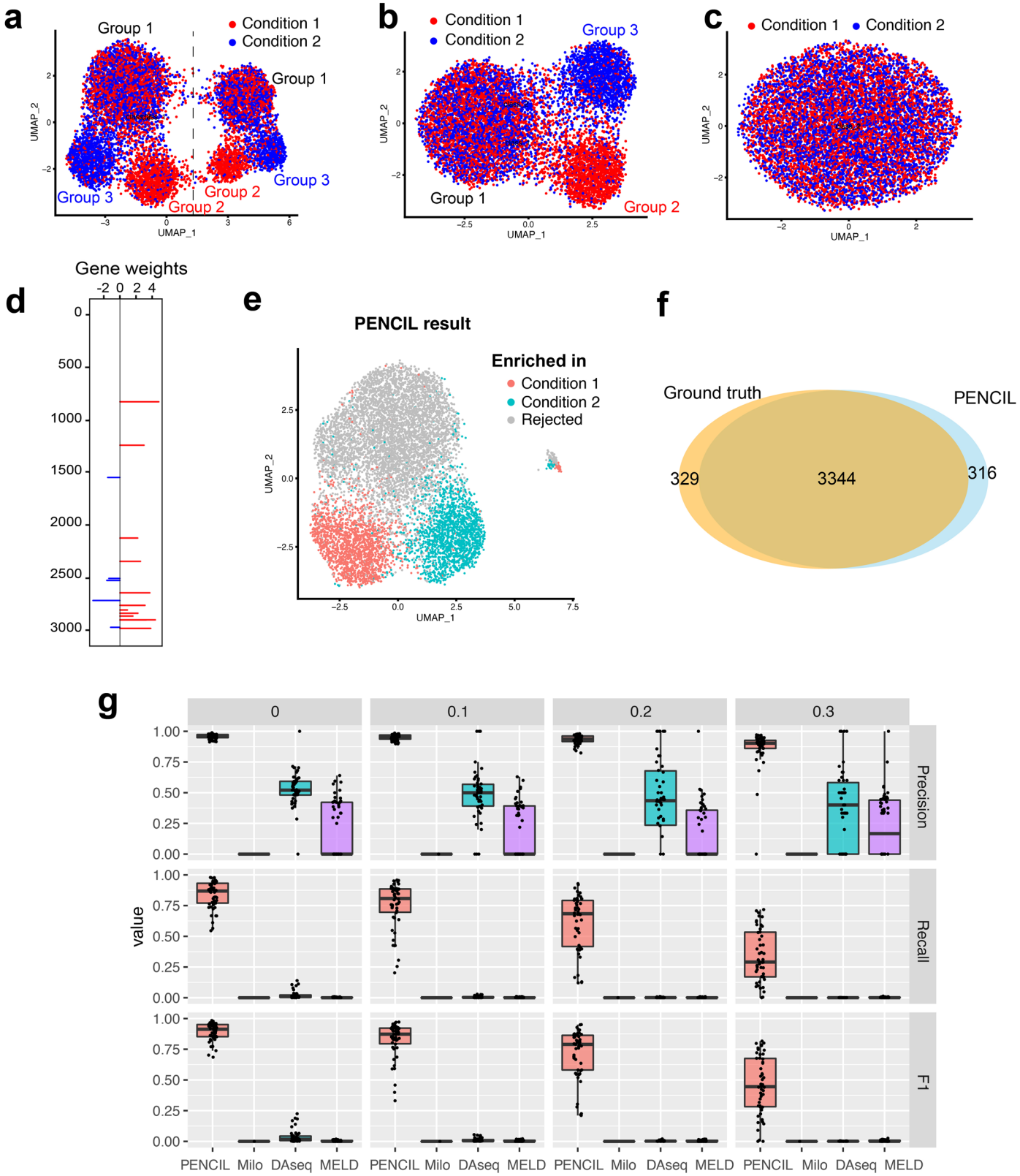
Extended Data Fig. 3 | The simulation flowchart. a, The matrix from a real scRNA-seq dataset. **b**, Selecting a submatrix with a subset of genes as indicated by the orange rectangle for the following clustering. **c**, UMAP visualization and standard clustering based on the submatrix from the previous step. **d**, Selecting two clusters from panel **c** as the ground truth subpopulations enriched in the phenotypes, respectively. **e**, Assigning cells with condition labels based on the designed conditions in panel **d** and the given mixing rate. **f**, The raw matrix with each cell assigned with a condition label as indicated on the top bar. **g**, The UMAP using the top 2000 MVGs colored by the condition labels of cells. **h**, The raw expression matrix and cell condition labels as the same inputs for all the methods.

https://doi.org/10.1038/s42256-023-00656-y



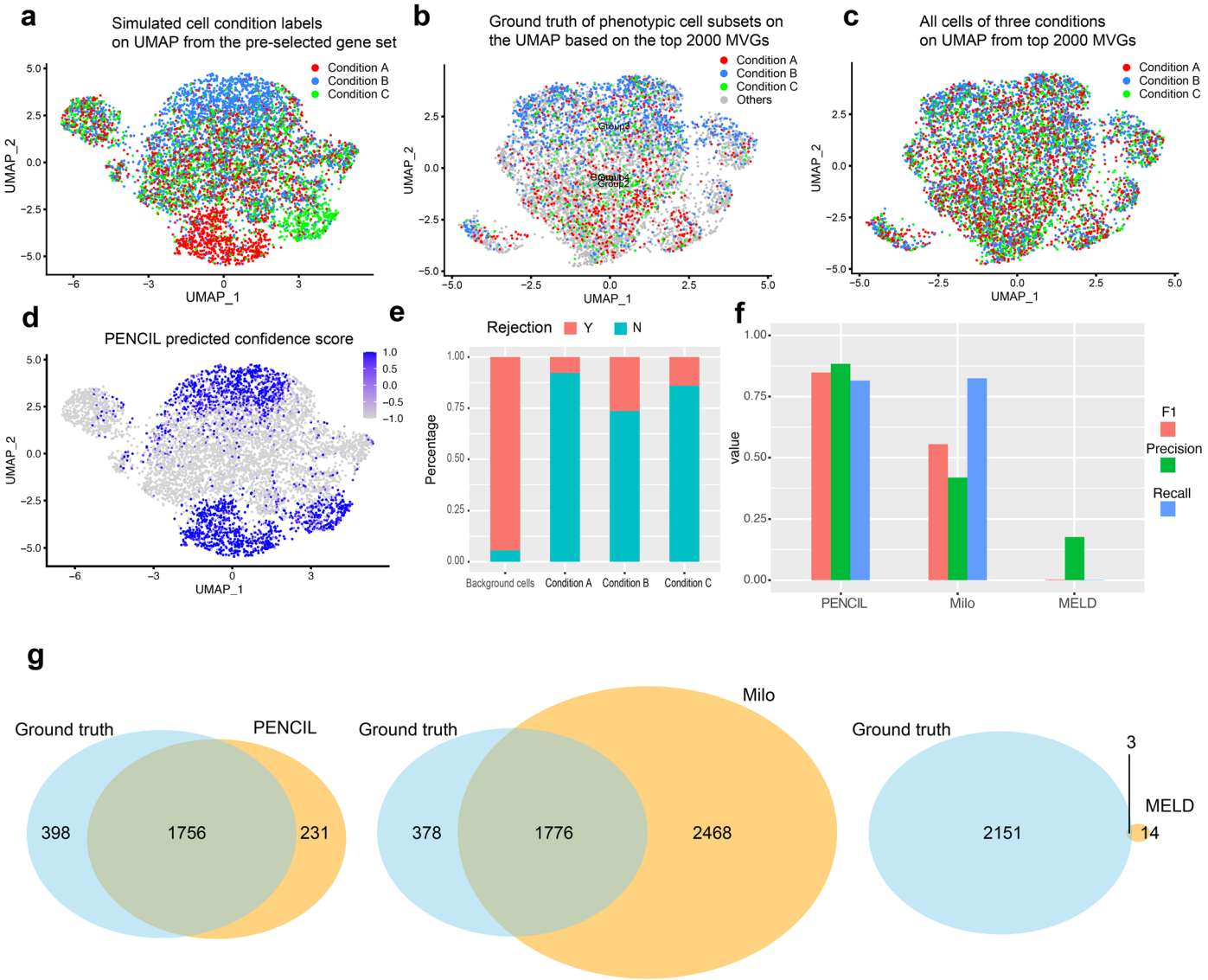
Extended Data Fig. 4 | PENCIL classification analysis of simulated datasets with two conditions. a, The confidence scores output by PENCIL. **b**, The distribution of the selected and rejected cells over the simulated ground truth of the conditions. **c**, The Venn diagram showing the overlap between the PENCIL selected cells and the ground truth phenotypic cells for the two conditions, respectively. **d**, The F1, precision and recall scores comparing the performances of the four methods.

https://doi.org/10.1038/s42256-023-00656-y



Extended Data Fig. 5 | Evaluating PENCIL on the simulated datasets with batch-effect. a, UMAP based on the manually curated genes showing the cells of two conditions from two batches separated by the dashed line. **b**, UMAP based on the manually curated genes showing the cells of two conditions after batch corrections. **c**, UMAP based on the top 3000 MVGs showing all cells. **d**, PENCIL selected genes. **e**, UMAP based on the PENCIL selected genes showing the PENCIL selected cells. **f**, The Venn diagram showing the overlap between the ground truth phenotype-enriched subpopulations and the PENCIL selected cells. **g**, The box plots comparing the performances of PENCIL, Milo, DAseq and MELD in simulated batch effects datasets with mixing rates 0, 0.1, 0.2 and 0.3 ($n = 50$ simulations). In the box plots, the center line and the box bounds represent median value and upper and lower quartiles, respectively. Box whiskers indicate the largest and smallest values no more than 1.5 times the interquartile range from the quartiles.

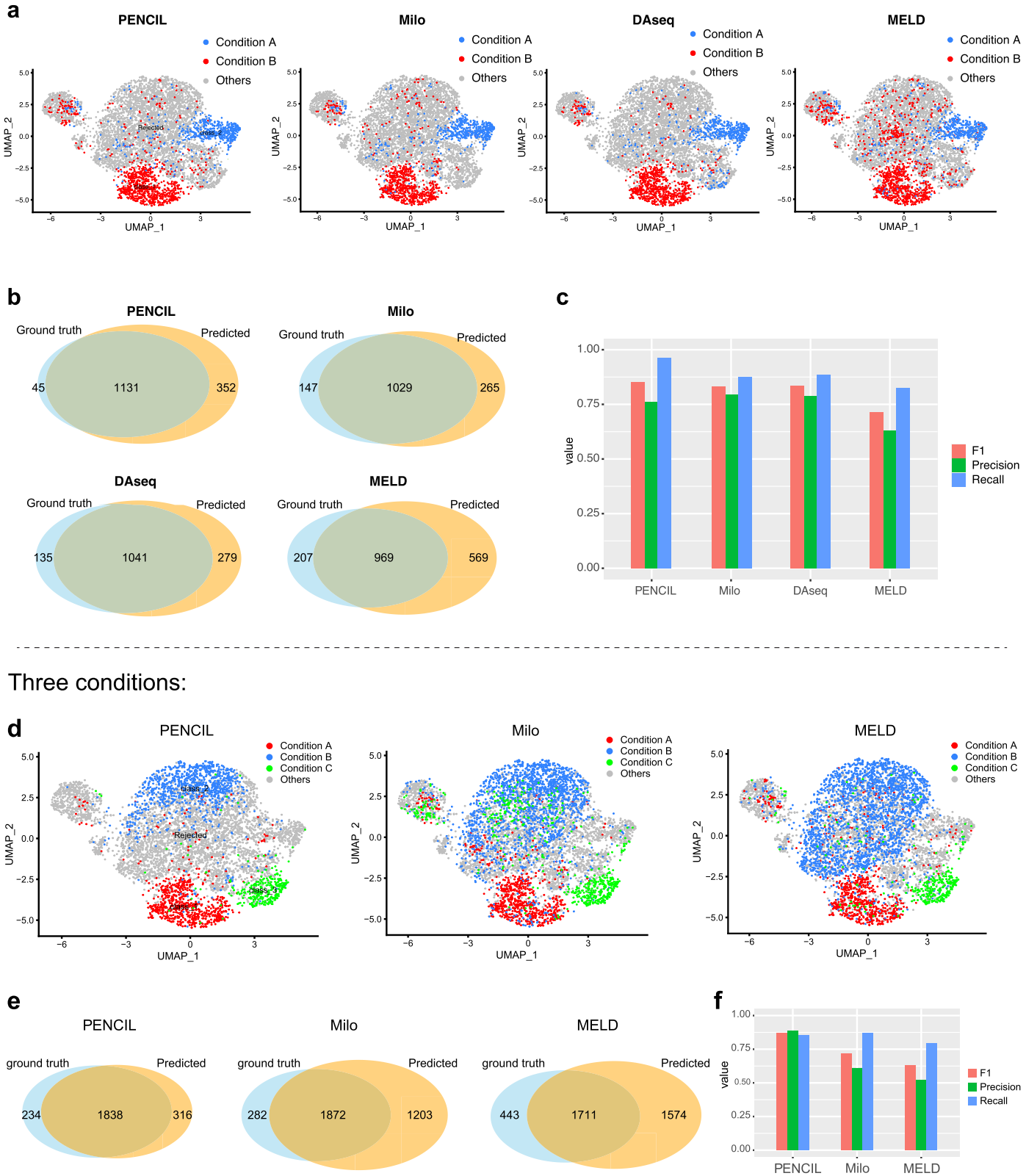
https://doi.org/10.1038/s42256-023-00656-y



Extended Data Fig. 6 | Evaluating PENCIL on the simulated datasets with three conditions. a, The UMAP based on the pre-selected gene set colored by the cell condition labels generated from ground truth cell subsets with a mixing rate of 0.1. **b**, The ground truth phenotype-associated subpopulations visualized on the UMAP using the top 2000 MVGs. **c**, Cells with the same condition labels as the ones in the panel **a** visualized on the UMAP using the top 2000 MVGs. **d**, The UMAP based on the pre-selected genes colored by the PENCIL predicted confidence scores. **e**, The distribution of PENCIL selected cells over the ground truth cell conditions. **f**, The F1, precision and recall scores comparing the performances of the three methods on this simulated dataset with three conditions. **g**, The Venn diagrams depicting the overlap between ground truth cell subpopulations and cell subsets selected by the three methods, respectively.

https://doi.org/10.1038/s42256-023-00656-y

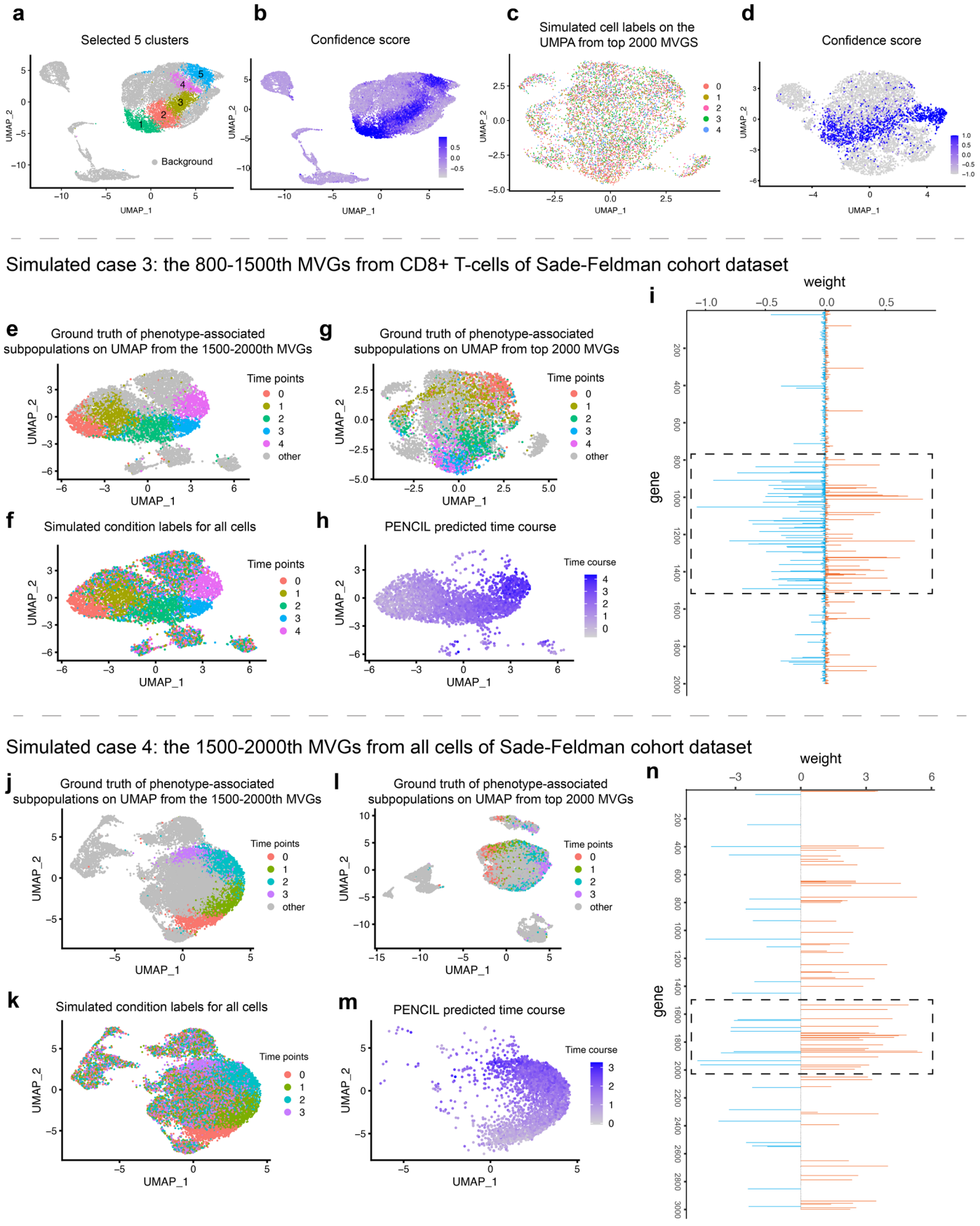
Two conditions:



Three conditions:

Extended Data Fig. 7 | Evaluating the four methods using the PENCIL selected genes as inputs. a, The results of PENCIL, Milo, DAseq and MELD when inputting the genes selected by PENCIL. **b**, The Venn diagrams comparing the result of each method with the ground truth phenotypic cell subpopulations. **c**, The F1, precision, and recall scores comparing the performances of the four methods when inputting the genes selected by PENCIL. **d-f**, A simulation for the three conditions. **d**, The UMAP plots showing the results of PENCIL, Milo, and MELD when inputting the genes selected by PENCIL. **e**, The Venn diagrams comparing the result of each method with the ground truth phenotypic cell subpopulations. **f**, The F1, precision and recall scores comparing the performances of the three methods when inputting the genes selected by PENCIL in this simulated example with three conditions.

https://doi.org/10.1038/s42256-023-00656-y

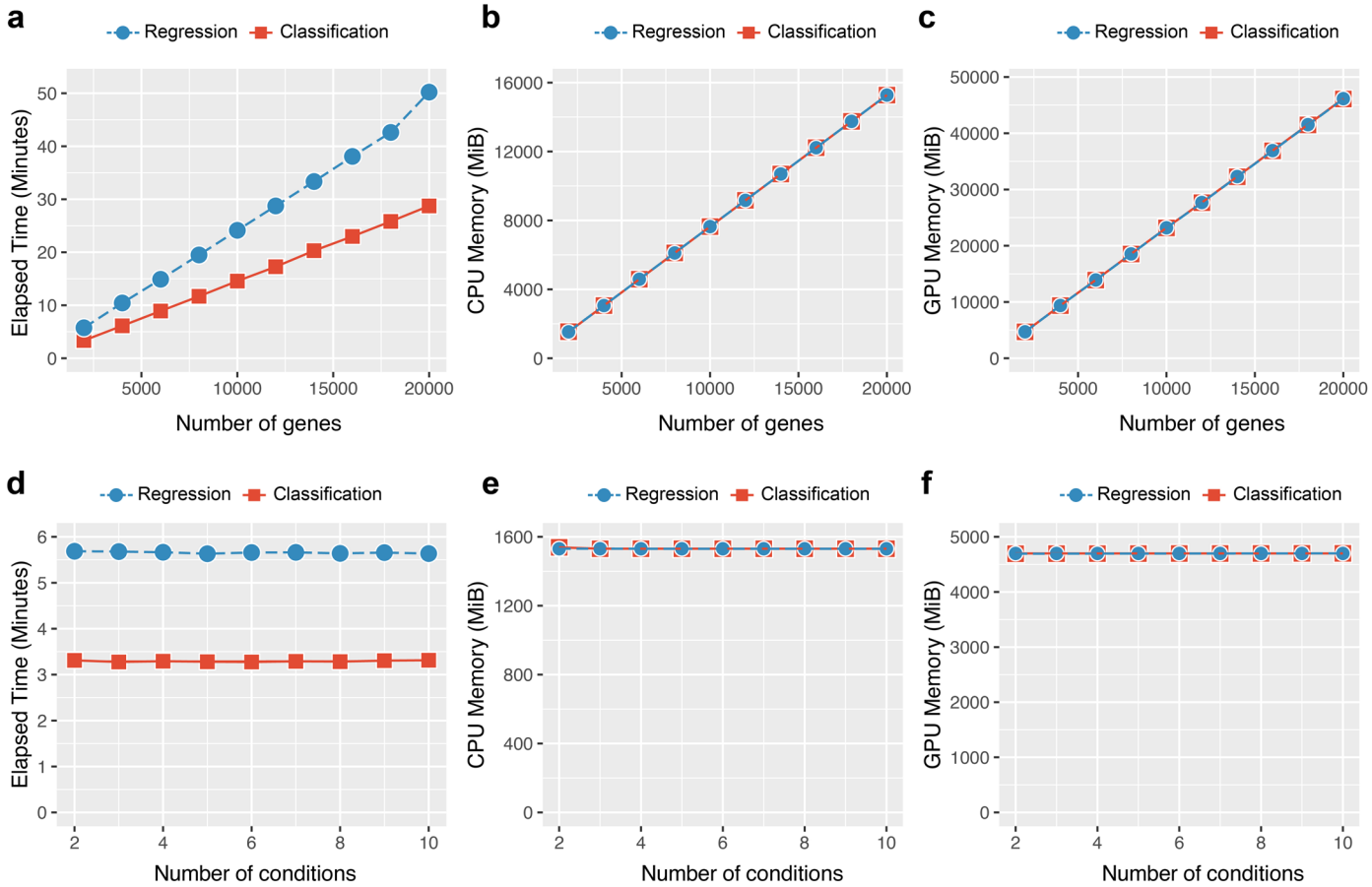


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s42256-023-00656-y

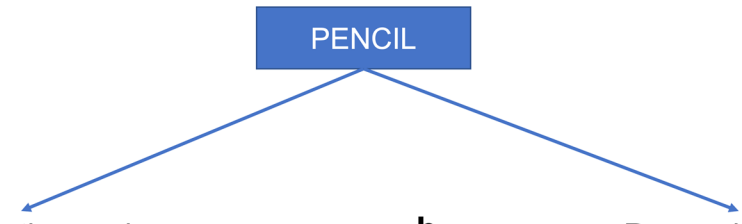
Extended Data Fig. 8 | Evaluating the regression model of PENCIL in simulated datasets. a, UMAP showing the cells of 5 clusters selected as ground truth subpopulations corresponding to main Fig. 3a. **b**, PENCIL predicted confidence scores corresponding to main Fig. 3c. **c**, The cells with simulated condition labels visualized on the UMAP using the top 2000 MVGs. **d**, PENCIL predicted confidence scores corresponding to main Fig. 3k. **e**-**i**, A simulated dataset for PENCIL regression analysis from the Feldman T-cell dataset. **e**, UMAP from a pre-selected gene set (800-1500th MVGs) to show cells with simulated ground truth phenotypic subpopulations of five time points. **f**, The five subpopulations are assigned to the five samples accordingly, and all remaining cells are evenly assigned to the five samples to simulate the sample labels. The UMAP is the same as the panel **e** colored by cell condition labels. **g**, Ground truth of phenotype-associated subpopulations in panel **e** visualized on the UMAP using the top 2000 MVGs. **h**, PENCIL predicted continuous time points for the selected cells. **i**, PENCIL selected genes. Genes within the dashed rectangle region were the gene-set to generate UMAPs in panels **e**, **f** and **h**. **j**-**n**, A simulated dataset for PENCIL regression analysis from the Sade-Feldman cohort dataset. **j**, UMAP from a pre-selected gene set (1500-2000th MVGs) to show cells with simulated ground truth subpopulations of four time points. **k**, The four subpopulations are assigned to the four samples accordingly and all remaining cells are evenly assigned to the four samples. The UMAP is the same as the panel **j** colored by simulated condition labels. **l**, Ground truth of phenotype-associated subpopulations in panel **j** visualized on the UMAP using top 2000 MVGs. **m**, PENCIL predicted continuous time points for the selected cells. **n**, PENCIL selected genes. Genes within the dashed rectangle region were the gene set to generate UMAPs in panels **j**, **k** and **m**.

https://doi.org/10.1038/s42256-023-00656-y



Extended Data Fig. 9 | PENCIL's runtime and memory usages with varying numbers of genes and conditions. a-c, For datasets with 10,000 cells and three conditions, the runtime, overall memory usage of CPU and GPU against the number of genes, respectively. **d-f**, For datasets with 10,000 cells and 2000 genes, the runtime, overall memory usage of CPU and GPU against the number of conditions, respectively. MiB, mebibyte.

https://doi.org/10.1038/s42256-023-00656-y



a Classification mode

For an existing application:

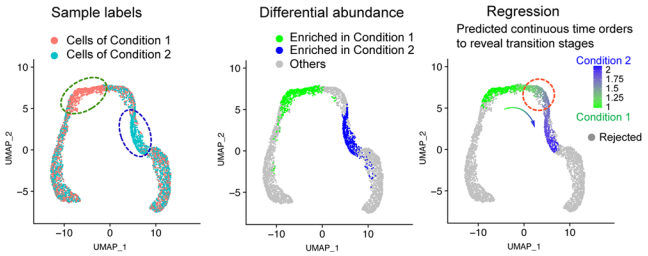
Differential abundance (DA) of subpopulations

Method	DA quantification	Clustering-free	KNN-free	Prediction on new samples	Feature selection	GPU speedup	Analyze one million cells in 1h
PENCIL	Learning with rejection	✓	✓	✓	✓	Fully	✓
Milo	Negative binomial GLM	✓	✗	✗	✗	✗	✗
DAseq	Logistic classifier prediction	✓	✗	✗	✗	Partially	✗
MELD	Kernel density estimation	✓	✗	✗	✗	✗	✗

b Regression mode

For a **new** application:

Supervised phenotypic trajectory learning of subpopulations



Extended Data Fig. 10 | A summary of PENCIL's two modes. a, The advantages of classification-based PENCIL. **b**, The regression mode of PENCIL formulates a new application to reveal a continuous dynamic process.

nature portfolio

Corresponding author(s): Zheng Xia; Ling-Yun Wu

Last updated by author(s): Mar 22, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Python (v3.9.7), R (v4.1.0), R package Seurat (v4.0.5) and Python package Scanpy (v1.9.1)
Data analysis	We developed an python package PENCIL in this study. The source code and tutorial of PENCIL are freely available at https://github.com/cliffren/PENCIL and https://doi.org/10.5281/zenodo.7762054. In addition, we also used the following tools for analysis: • R Markdown (v2.11) for performing data analysis and creating the tutorials in R. • R package Seurat (v4.0.5) for single-cell preprocessing, the differential gene expression analysis, calculating the enrichment scores of the pathways and visualization in R. • The fora function in fgsea (v1.20.0) was used to calculate the enrichments of pathways in the DEGs. • R package monocle3 (v1.2.9) for the identification of the genes significantly depending on the time points. • R package splatter (v.1.10.1) for the scRNA-seq simulation. • R package infercnv(v1.6.0) was applied to predict the segmented copy-number alterations (CNAs) in scRNA-seq data. • R package DoubletFinder (v2.0.3) for detecting the doublets. • R package SingleR (v2.0.3) was used to annotate cell types. • Python packages torch (v1.10.0), numpy (v1.20.3), pandas (v1.3.4), seaborn (v0.11.2), umap-learn (v0.5.2) and mlflow (v1.23.1) for developing the PENCIL model. • IPython jupyter notebook (v6.1.12) for performing data analysis and creating the tutorials in python. • Python package scanpy (v1.9.1) for single-cell preprocessing and visualization in python.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Only publicly available datasets were analyzed in this study. The detailed descriptions of those datasets along with their accession numbers are provided on the online methods section.
• The Sade-Feldman cohort data of melanoma immunotherapy was downloaded from GEO database (GSE120575): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120575
• Two scRNA-seq melanoma samples not responding to immunotherapy were directly downloaded from GEO (accession number: GSE159251): https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159251
• One scRNA-seq melanoma sample responding to immunotherapy (GSE134388) were directly downloaded from TISCH2: https://biostorage.s3.ap-northeast-2.amazonaws.com/TISCH_2022/SKCM_GSE134388_aPD1/SKCM_GSE134388_aPD1_expression.h5
• The scRNA-seq dataset across three treatment time points from an mantle cell lymphoma patient can be downloaded from https://zenodo.org/record/7761954
• In pathway enrichment analysis database (Molecular Signatures Database, MsigDB v7.2), hallmark gene sets and immunologic signature gene sets were downloaded from https://www.gsea-msigdb.org/gsea/msigdb/

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A. No new data were collected in this study, and no sex and gender based analysis was performed.
Population characteristics	N/A. No new data were collected in this study, and no population based analysis was performed.
Recruitment	N/A. No new data were collected in this study
Ethics oversight	N/A. No new data were collected in this study

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	1. The Sade-Feldman cohort data of melanoma immunotherapy has 48 patients (17 responders and 31 non-responders), with a total of 16291 cells including 6350 CD8+ T-cells. 2. The mantle cell lymphoma (MCL) dataset includes a total of 3236 cells from one MCL patient at three different treatment time points. 3. The CD8+ cells gene expression matrices of the two immunotherapy non-responder samples (GSE120575) were retained with 3737 cells (K411) and 4878 cells (K468) for PENCIL prediction, respectively. 4. One scRNA-seq melanoma sample responded to immunotherapy treatment has 1004 CD8+ T-cells (GSE134388). 5. During simulation, we generated datasets with cell numbers ranging from 1000 cells to 1,000,000 cells to evaluate the scalability of PENCIL.
Data exclusions	No data was excluded.
Replication	To validate the performance of PENCIL, we repeated each simulation setting 50 times. Additionally, for the immunotherapy data, we performed leave-one-patient-out to evaluate patient-level prediction of immunotherapy outcomes. We then also evaluated this patient-level immunotherapy response prediction on 3 new scRNA-seq samples.
Randomization	In the multiple replicate experiments, we synthesized the simulation data by randomly selecting a subset of genes with a fixed number of genes from the given genes. And we simulated multiple single-cell data with batch effects using different random number seeds.
Blinding	In the simulation experiments, the subset of genes used to generate the simulated data and ground truth phenotypic subpopulations are invisible to both our method and other methods used for comparison, thus ensuring fairness of comparison. When evaluating the predictive

power of the model, all data used for testing and validation are unseen during model training. There is no information leakage in all performance evaluation sessions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging