



Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data

Duanchen Sun^{1,2}, Xiangnan Guan^{1,2}, Amy E. Moran^{3,4}, Ling-Yun Wu⁵, David Z. Qian⁴, Pepper Schedin^{3,4}, Mu-Shui Dai⁶, Alexey V. Danilov⁷, Joshi J. Alumkal⁸, Andrew C. Adey^{4,6}, Paul T. Spellman^{1,4,6} and Zheng Xia^{1,2,4,9} ✉

Single-cell RNA sequencing (scRNA-seq) distinguishes cell types, states and lineages within the context of heterogeneous tissues. However, current single-cell data cannot directly link cell clusters with specific phenotypes. Here we present Scissor, a method that identifies cell subpopulations from single-cell data that are associated with a given phenotype. Scissor integrates phenotype-associated bulk expression data and single-cell data by first quantifying the similarity between each single cell and each bulk sample. It then optimizes a regression model on the correlation matrix with the sample phenotype to identify relevant subpopulations. Applied to a lung cancer scRNA-seq dataset, Scissor identified subsets of cells associated with worse survival and with *TP53* mutations. In melanoma, Scissor discerned a T cell subpopulation with low *PDCD1/CTLA4* and high *TCF7* expression associated with an immunotherapy response. Beyond cancer, Scissor was effective in interpreting facioscapulohumeral muscular dystrophy and Alzheimer's disease datasets. Scissor identifies biologically and clinically relevant cell subpopulations from single-cell assays by leveraging phenotype and bulk-omics datasets.

Single-cell sequencing technologies are revolutionizing biomedical research and clinical practice by enabling the comprehensive characterization of cells from complex tissues^{1,2}. In contrast to bulk data that measure the averaged properties of whole tissue, scRNA-seq allows identifying cell types, states and lineages of different cell subpopulations in a heterogeneous tissue ecosystem^{3–5}. To recognize critical subpopulations from single-cell data, the standard approach is to perform unsupervised clustering to define cell clusters, inspect marker genes of each cluster and assess the enrichment of the marker genes in known cell types and pathways to evaluate the importance for each cell cluster^{6,7}. However, identifying cell subpopulations that drive phenotypes, such as disease stage, tumor metastasis, treatment response and survival outcome, is of indispensable importance because it will facilitate cell type targeted therapies as well as prognostic biomarker discovery^{3,8}. Unfortunately, single-cell technology is not practical in large cohorts, and most single-cell experiments involve fewer than 20 patient samples^{9–11}, which lacks the statistical power to identify the cell subpopulations driving the phenotype of interest.

In addition, valuable clinical phenotype information is widely available from big data consortia like The Cancer Genome Atlas (TCGA)¹² through a decade-long collection of clinicopathologic annotations. Clinical phenotype information is primarily collected on bulk tissue samples, especially in the form of formalin-fixed, paraffin-embedded samples, which are not feasible for single-cell profiling. Therefore, there is an unmet need to leverage such widely accessible and valuable phenotype information to guide cell subpopulation identification from single-cell data.

To our knowledge, there is no bioinformatics tool that uses external bulk phenotypes to guide the identification of key cell subpopulations in a unified framework for single-cell data analysis. Therefore, in this study, we introduce Single-Cell Identification of Subpopulations with Bulk Sample Phenotype Correlation (Scissor). By leveraging bulk data and phenotype information, this algorithm automatically selects cell subpopulations from single-cell data that are most responsible for the differences of phenotypes. The novelty of Scissor is that it uses phenotype information from bulk data to identify the most highly disease-relevant cell subsets. Our studies suggest that Scissor is a promising tool to explore and interpret single-cell data from a new perspective, which can shed fresh light on disease mechanisms and improve the diagnosis and treatment of diseases.

Results

Overview of Scissor. To use bulk data and phenotype information for assisting single-cell data analysis, we developed an algorithm, Scissor, to identify cell subpopulations from single-cell data that are most highly associated with the given phenotypes. Briefly, the three data sources for Scissor are a single-cell expression matrix, a bulk expression matrix and a phenotype of interest (Fig. 1a). The phenotype annotation of each bulk sample can be a continuous dependent variable, binary group indicator vector or clinical survival data. The key step of Scissor is to quantify the similarity between the single-cell data and bulk data through a measurement such as Pearson correlations for each pair of cells and bulk samples. After this, Scissor

¹Computational Biology Program, Oregon Health & Science University, Portland, OR, USA. ²Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA. ³Department of Cell, Developmental and Cancer Biology, Oregon Health & Science University, Portland, OR, USA.

⁴Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. ⁵Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. ⁶Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR, USA. ⁷City of Hope National Medical Center, Duarte, CA, USA. ⁸Department of Internal Medicine, Rogel Cancer Center, University of Michigan, Ann Arbor, MI, USA. ⁹Department of Molecular Microbiology and Immunology, Oregon Health & Science University, Portland, OR, USA. ✉e-mail: xiaz@ohsu.edu

optimizes a regression model on the correlation matrix with the sample phenotype (Fig. 1b). The selection of the regression model depends on the type of the input phenotype—for example, linear regression for continuous variables, logistic regression for dichotomous variables and Cox regression for clinical survival data. Because a small subset of cells could drive the phenotype of interest¹³, a sparse penalty and a graph regularization are imposed on the regression model to select similar cells that are important for the given phenotype with high confidence (Fig. 1b). Based on the signs of the estimated regression coefficients, the cells with non-zero coefficients can be indicated as Scissor positive (Scissor⁺) cells and Scissor negative (Scissor⁻) cells, which are positively and negatively associated with the phenotype of interest, respectively (Fig. 1c). The cells with coefficients of zero are indicated as background cells. Furthermore, to control the false associations between single-cell and bulk data, we designed a reliability significance test to determine whether the chosen data are suitable for our phenotype-to-cell associations (Fig. 1d). Finally, the Scissor-selected cells will be further characterized in downstream analyses, such as exploration of signature genes and functionally enriched pathways (Fig. 1e).

Capturing phenotype-associated subpopulations in simulations. We first assessed the performances of Scissor on a series of simulated datasets to test whether Scissor can recover the known phenotype-associated cell subpopulations. To achieve this, we employed Splatter¹⁴ to simulate two phenotype-specific cell subpopulations as the ground truth and one common cell type shared by the two phenotypes, as well as the corresponding bulk expression data (Fig. 1f and Methods). In one simulation, the two ground truth phenotype-specific cell subpopulations were characterized by 21 true differentially expressed genes (DEGs) between them but cannot be distinguished by standard single-cell analysis pipelines because they were in the same cluster (Supplementary Fig. 1a,b). Using bulk gene expressions to guide the cell subpopulation identification, we found that Scissor recovered 90.5% and 91.8% of the ground truth cells unique to each phenotype, with low false-positive rates of 2.1% and 0.9%, respectively (Fig. 1g,h). There were 22 DEGs between the two cell subpopulations identified by Scissor, including all 21 true DEGs between the two ground truth phenotype-associated cell subpopulations (Supplementary Fig. 1b). In contrast, the fold change signals of these 21 true DEGs at the bulk level were weak with values less than 1.1 (Supplementary Fig. 1c). Such improved phenotype-specific gene signal detection will facilitate the reliable downstream interrogation of different cell types and states within a heterogeneous cell population. When performing this simulation 100 times, Scissor had an average recall of 0.88 and precision of 0.93 in identifying phenotype-associated subpopulations.

Furthermore, to check whether the performance of Scissor depends on the true DEG signals at the bulk level, we reset the expression levels of these true DEGs as the least differentially expressed genes (Methods). We found that Scissor can still identify 87.6% and 87.3% of the ground truth cells for each phenotype,

with slightly higher false-positive rates of 4.2% and 5.9%, respectively (Supplementary Fig. 1d,e). In this setting, Scissor detected all 21 true DEGs with only two false-positive DEGs (Supplementary Fig. 1b). Finally, we demonstrated the robustness of Scissor on the batch effects corrected scRNA-seq data (Supplementary Fig. 1f–i). Overall, the results from these simulations were largely consistent with each other, indicating Scissor's capacity to identify the phenotype-associated cell subpopulations from single-cell data.

Identifying subpopulations of tumors and normal phenotypes.

We then applied Scissor to lung cancer scRNA-seq data that included tumor cells and cells from the tumor microenvironment with known cell types¹⁰ (Fig. 1i). To demonstrate the efficacy of our algorithm, we used the tumor and normal phenotypes from 577 TCGA-lung adenocarcinoma (LUAD) bulk samples¹⁵ to guide Scissor analysis. We expected that, by using these data, Scissor would infer cells that were most highly associated with cancer or normal phenotype in this heterogeneous single-cell dataset. Because of the binary phenotype settings of this application, where samples with a phenotype indicator value 1 correspond to tumor samples, Scissor⁺ cells should be associated with the cancer cells, and Scissor⁻ cells should be associated with the normal phenotype. Among 29,888 cells from different cell types (Fig. 1i), 361 Scissor⁺ cells and 534 Scissor⁻ cells were selected by Scissor, which were associated with the tumor and normal phenotypes with high confidence (Fig. 1j). As anticipated, over 98% of Scissor⁺ cells were verified to be malignant cells (Fig. 1k). Such a high proportion of cancer cells in Scissor⁺ cells cannot be selected by chance (Supplementary Fig. 1j, hypergeometric test $P < 2 \times 10^{-16}$), which confirmed that Scissor can identify the cells associated with the phenotypes of interest. As for Scissor⁻ cells, the cell types were relatively more balanced than Scissor⁺ cells because it was designed to associate with more diverse non-malignant cell types (Fig. 1k). Myeloid cells and alveolar cells were the two main selected cell types, accounting for 42.3% and 36.9% of total Scissor⁻ cells, respectively. All cell types in Scissor⁻ cells, especially the alveolar cells, are important cell types in normal lung tissues. Thus, we demonstrated that Scissor can precisely identify the most phenotype-associated cells from single-cell data with the guidance of phenotype information from bulk data.

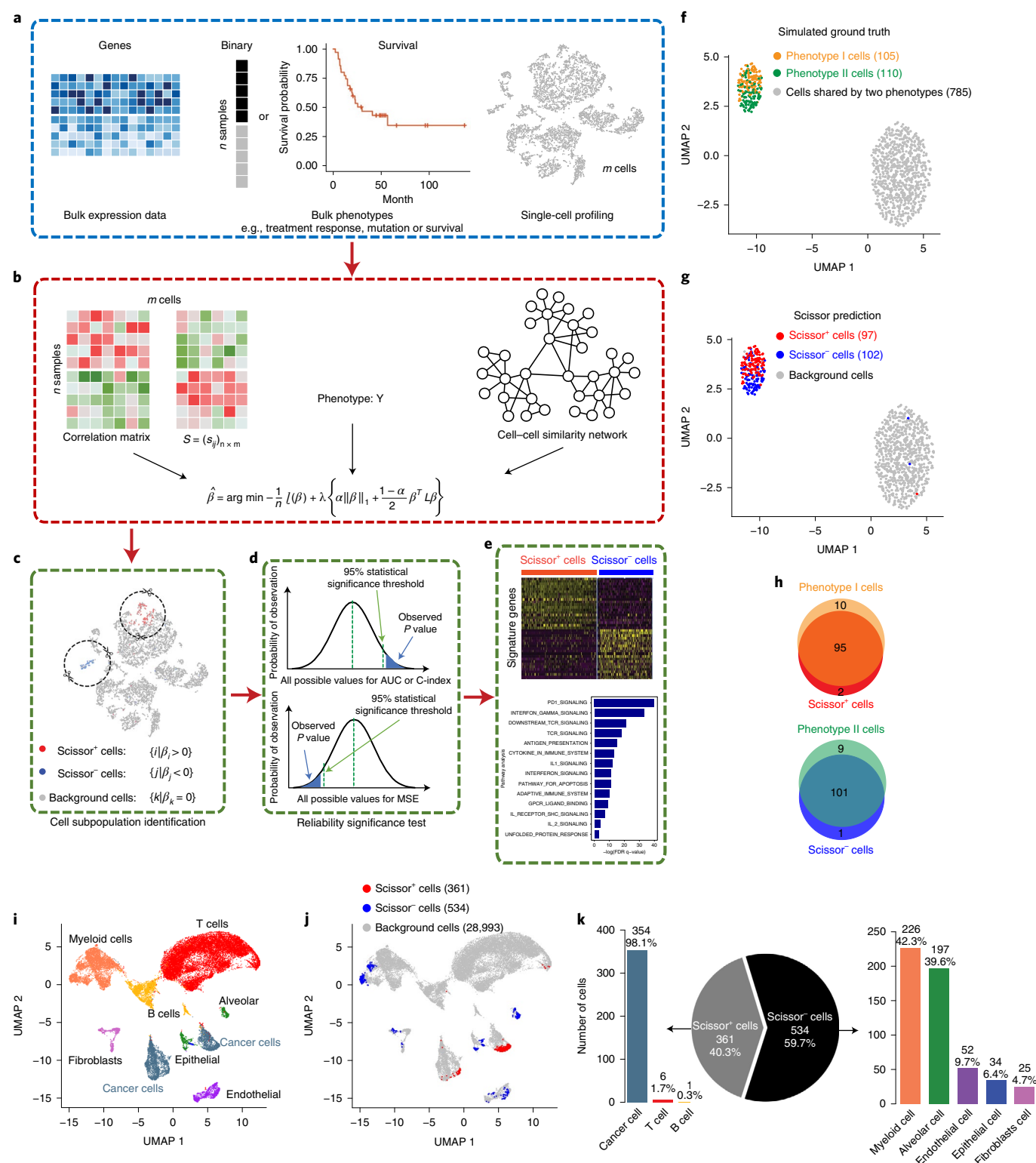
Detecting a hypoxic subpopulation related to worse survival.

Cancer cells are heterogeneous and include subpopulations such as cancer stem cells, which are known to drive tumor progression and poor prognosis^{16,17}. Therefore, we applied Scissor, guided by 471 TCGA-LUAD bulk samples with survival information¹⁵, to identify aggressive cancer cell subpopulations within 4,102 cancer cells from the same lung cancer scRNA-seq dataset¹⁰. These cells were separated into 12 clusters (Fig. 2a), which demonstrated the heterogeneous nature of the cancer cells. Of 205 Scissor-selected cells, 201 Scissor⁺ cells were associated with worse survival (defined as Scissor_WS cells thereafter), and only four Scissor⁻ cells were associated with good survival (Fig. 2b). The Scissor_WS cells were

Fig. 1 | The workflow of Scissor and its performance in applications with known phenotype-associated cell subpopulations. **a**, The inputs for Scissor are a single-cell expression matrix, a bulk expression matrix and a corresponding phenotype of interest, such as drug response or clinical information. **b**, Scissor calculates a correlation matrix and a cell-cell similarity network based on input sources, which are further integrated with the phenotype into a network regularized sparse regression model to select the most relevant cell subpopulations. **c**, The selected cells by Scissor with red and blue dots indicating the cells positively or negatively associated with the phenotype of interest. **d**, The Scissor results are evaluated by a reliability significance test to control the false associations between single-cell and bulk data. **e**, The selected cells by Scissor can be further investigated by downstream analyses. **f**, UMAP visualization of the simulated cells. The orange and green dots are the ground truth cell subpopulations specific to phenotype I and II, respectively. **g**, Visualization of the Scissor-selected cells on the same UMAP as in **a**. The red and blue dots are the Scissor-selected cells associated with phenotype I and II. **h**, Venn diagrams show overlaps between the Scissor-selected cells and the ground truth phenotype-specific cells. **i**, UMAP visualization of 29,888 LUAD cells. The color encoding the cell types is defined in the original paper. **j**, UMAP visualization of the Scissor-selected cells. The red and blue dots are cells associated with the tumor and normal phenotypes, respectively. **k**, The pie chart of the Scissor-selected cells with the corresponding bar plots showing the detailed constitutions in each type defined in **i**.

mainly from clusters 1 and 3 (Fig. 2c). To understand the underlying transcriptional patterns of Scissor_WS cells, we compared the gene expressions of those cells with all other cells. As a result, 23 upregulated genes and 205 downregulated genes were differentially expressed in Scissor_WS cells over all other cells, respectively (Fig. 2d and Supplementary Table 1). Notably, we found that multiple important hypoxia-related genes were among the above 23 overexpressed genes (Fig. 2e). Functional enrichment analysis also

confirmed that the hypoxia-related pathways, such as glycolysis and glucose metabolism processes, were activated in Scissor_WS cells (Fig. 2f and Supplementary Table 1). Consistently, motif analysis revealed that the *HIF1A* binding motif was the most enriched motif in the 23 upregulated genes (Supplementary Table 1), which is a key mediator of cellular response to lowered oxygen levels¹⁸. We also observed higher hypoxia pathway activities in an identified cell subpopulation related to higher tumor stages (Supplementary Fig. 2).



To further examine the clinical relevance of the above 23 overexpressed genes (defined as lung cancer signature; Supplementary Table 1), we chose six independent lung cancer datasets collected in PRECOG¹⁹ (Supplementary Table 1). We found that in five of six datasets, the patients with higher signature scores had significantly worse survival time than the patients with lower signature scores (Fig. 2g and Supplementary Fig. 3). This observation indicated that the lung cancer signature derived from Scissor⁺ cells was associated with worse survival and could provide potential drug targets for further investigation²⁰. Among the six chosen datasets, two have additional clinical features. We thereby investigated whether our lung cancer signature had predictive power for patient survival beyond clinical features. To achieve this, we examined pathological stage, sex, age at diagnosis and our lung cancer signature in these two datasets. We found that only the pathological stage and our signature were significantly associated with patient survival in univariate Cox survival analyses (Fig. 2h). Furthermore, our signature remained statistically significant in both datasets after adjusting for tumor stage in multivariable Cox survival analyses (Fig. 2i).

In summary, Scissor identified an aggressive cancer cell subpopulation from the LUAD scRNA-seq data that was associated with worse survival outcomes and can be characterized by the overexpression of hypoxia-related genes. The high hypoxia activity might drive LUAD progression and thereby conferred poor outcomes to patients whose tumors contained significant numbers of such cells.

Profiling a cell subpopulation associated with TP53 mutation.

To further demonstrate the high flexibility of Scissor in exploring a variety of hypotheses of cell states in single-cell data, we used other phenotypical features provided by TCGA-LUAD to guide the identification of cell subpopulations within the same 4,102 lung cancer cells. Here, we focused on TP53, a commonly mutated tumor suppressor genes found in human malignancies.

We collected TP53 mutation status (mutant or wild-type) from TCGA-LUAD as the phenotypes of 498 bulk samples. Scissor identified a total of 414 Scissor⁺ cells associated with the TP53 mutant and 318 Scissor⁻ cells associated with the wild-type (Fig. 3a). To uncover the transcriptional differences among these cells, we compared the gene expressions of Scissor⁺ cells with that of Scissor⁻ cells. As a result, 337 upregulated genes and 14 downregulated genes were differentially expressed in Scissor⁺ cells over Scissor⁻ cells (Fig. 3b and Supplementary Table 2). The 337 upregulated genes include multiple E2F target genes and cell cycle progression-related genes—for example, *AURKA*, *CDK1*, *CCNB2* and *TOP2A* (Fig. 3b). Functional enrichment analysis also confirmed that cell cycle-related pathways, such as hallmark E2F targets, were activated in Scissor⁺ cells (Fig. 3c and Supplementary Table 2). Consistently, master regulator analysis revealed that the activities of E2F transcription factor (TF) family members E2F1 and E2F4 were both highly elevated in Scissor⁺ cells (Fig. 3d). In this analysis, we also found that TFs FOXM1 and MYC, which are known to be repressed by wild-type TP53 (refs. ^{21,22}), were

activated in the cell subpopulation associated with the TP53 mutant (Fig. 3d). FOXM1 is reported as a major predictor of unfavorable outcomes in human cancers²². These observations are in line with the literature that TP53 mutation leads to tumorigenesis by involving its impaired capability to arrest cell cycle and maintain DNA repair in response to oncogenic stimuli²³. Indeed, master regulator analysis confirmed that TP53 is inactivated in Scissor⁺ cells (Fig. 3d), which further demonstrates the capability of Scissor in identifying phenotype-associated cell subpopulations.

Furthermore, by linking those 337 upregulated genes (defined as TP53 mutation signature; Supplementary Table 2) with clinical outcomes, we demonstrated that the patients with higher TP53 mutation signature scores had significantly worse survival time than the patients with lower scores (Fig. 3e). For the downregulated genes, we found that major histocompatibility complex (MHC) class-related genes *HLA-A*, *B2M* and *CD74* are downregulated in Scissor⁺ cells, and only one of them can be directly identified from the bulk level gene expression (Fig. 3f and Supplementary Fig. 4). Notably, *B2M* is a critical component of MHC class I antigen presentation, and loss-of-function mutations in *B2M* have been reported in patients with cancer who are resistant to immunotherapy²⁴. Thus, our phenotype-guided scRNA-seq analysis implied that TP53 mutation is likely a mechanism of resistance to checkpoint inhibitor treatment. In summary, this application demonstrated the versatile abilities of Scissor in integrating a variety of phenotypes for single-cell subpopulation annotations.

Identifying a T cell subpopulation related to immunotherapy.

Immune checkpoint blockade (ICB) has achieved exciting results in a wide variety of cancers^{25,26}. To understand the mechanism underlying ICB response, we performed Scissor on a melanoma scRNA-seq dataset to identify a T cell subpopulation related to ICB response.

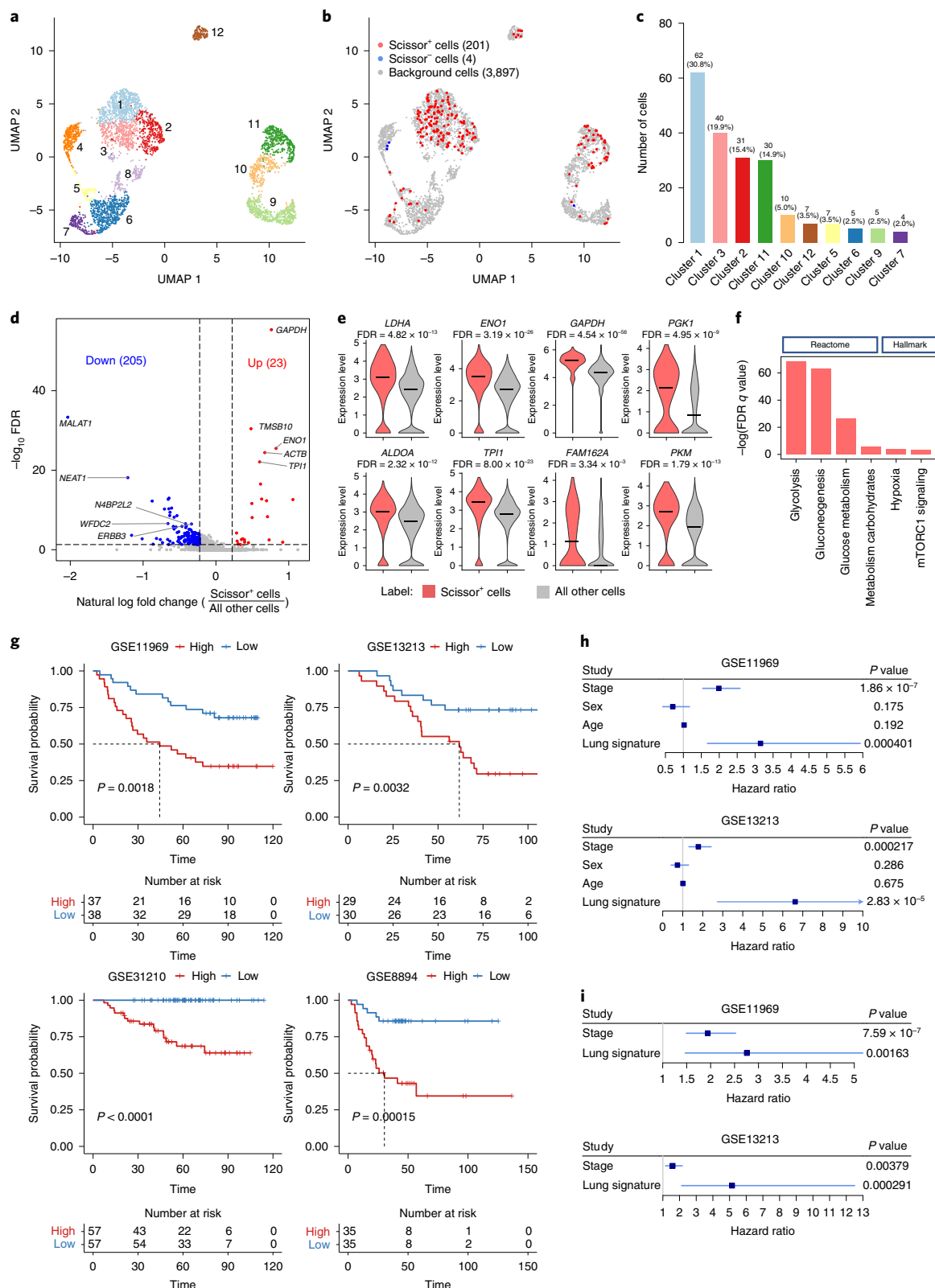
We focused our analysis on 1,894 T cells from the metastatic melanoma tumor microenvironment²⁷ and collected 70 melanoma bulk patients with known immunotherapy response information from two studies^{28,29}. In the standard scRNA-seq data analysis, these T cells were clustered into six clusters (Fig. 4a). By performing Scissor, we identified 105 T cells as Scissor⁺ cells, which were associated with a favorable immunotherapy response (defined as Scissor_{FR} cells thereafter), and did not report any cells that are associated with the unfavorable immunotherapy responses (Fig. 4b). These 105 Scissor_{FR} cells resided mainly in clusters 2 and 3 (Fig. 4c), indicating that clusters 2 and 3 were more associated with the effective responses than other clusters. To characterize the transcriptional identities of the Scissor_{FR} cells, we compared the gene expression of these cells with all other cells. In total, 17 upregulated and 120 downregulated differential expression genes were identified in Scissor_{FR} cells (Fig. 4d and Supplementary Table 3). The Scissor_{FR} cells that were associated with an effective ICB response had increased expressions of genes linked to T cell memory (*CCR7* and *SELL*) and survival (*IL7R*) as well as lower expressions of

Fig. 2 | Scissor identification results on lung cancer cells guided by TCGA-LUAD survival outcomes. **a**, The UMAP visualization of 4,102 LUAD cancer cells. **b**, UMAP visualization of the Scissor-selected cells. The red and blue dots are Scissor⁺ (worse survival) and Scissor⁻ (good survival) cells, respectively. **c**, Bar plot shows the constitution of Scissor⁺ cells in different clusters. **d**, Volcano plot of differential gene expressions in Scissor⁺ cells versus all other cells. The two vertical dashed lines represent $\pm \ln(1.25)$ fold changes in gene expression, and the horizontal dashed line denotes an FDR cutoff of 0.05. The FDR was the adjusted *P* value calculated by the two-tailed Wilcoxon rank sum test. **e**, Violin plots of expression levels of selected upregulated genes in Scissor⁺ cells. The FDR was the adjusted *P* value calculated by the two-tailed Wilcoxon rank sum test. **f**, Enrichment bar plot of selected hypoxia-related Reactome and Hallmark pathways. **g**, Kaplan-Meier survival curves show the clinical relevance of the lung cancer signature on four independent datasets. Tick marks indicate censoring events. The statistical *P* values were determined by the two-tailed log rank sum test. **h**, Forest plots show the hazard ratios and 95% confidence intervals for the lung cancer signature and three additional clinical features according to the univariate Cox model. **i**, The forest plots show the hazard ratios and 95% confidence intervals for the lung cancer signature and stage information according to the multivariable Cox model. Squares represent the hazard ratios, and the horizontal bars extend from the lower limits to the upper limits of the 95% confidence intervals of the estimates of the hazard ratios. These statistics are calculated on GSE11969 (*n* = 149 biologically independent samples) and GSE13213 (*n* = 117 biologically independent samples). The statistical *P* values were determined by the two-tailed Wald test.

inhibitory genes (*HAVCR2*, *LAG3*, *PDCD1* and *CTLA4*) and MHC II class genes (*HLA-DRB5*, *HLA-DRB1*, *HLA-DPA1*, *HLA-DQB2* and *HLA-DRB6*) (Fig. 4d,e and Supplementary Table 3). The Scissor_FR cells also exhibited enhanced expression of transcript factor *TCF7* that is associated with a favorable outcome in ICB treatment^{30,31} (Fig. 4e). In addition, pathway enrichment analysis showed that the Scissor_FR cells had higher TNF- α signaling and lower activity

of *CTLA4*, *PD1* signaling and lymphocytic choriomeningitis virus (LCMV)/tumor exhaustion pathways (Fig. 4f).

The above 137 differential expression genes related to effective immunotherapy response (defined as immunotherapy responsive signature; Supplementary Table 3) could be informative in predicting treatment success. We found that our signature scores were significantly higher in ICB responders than in non-responders



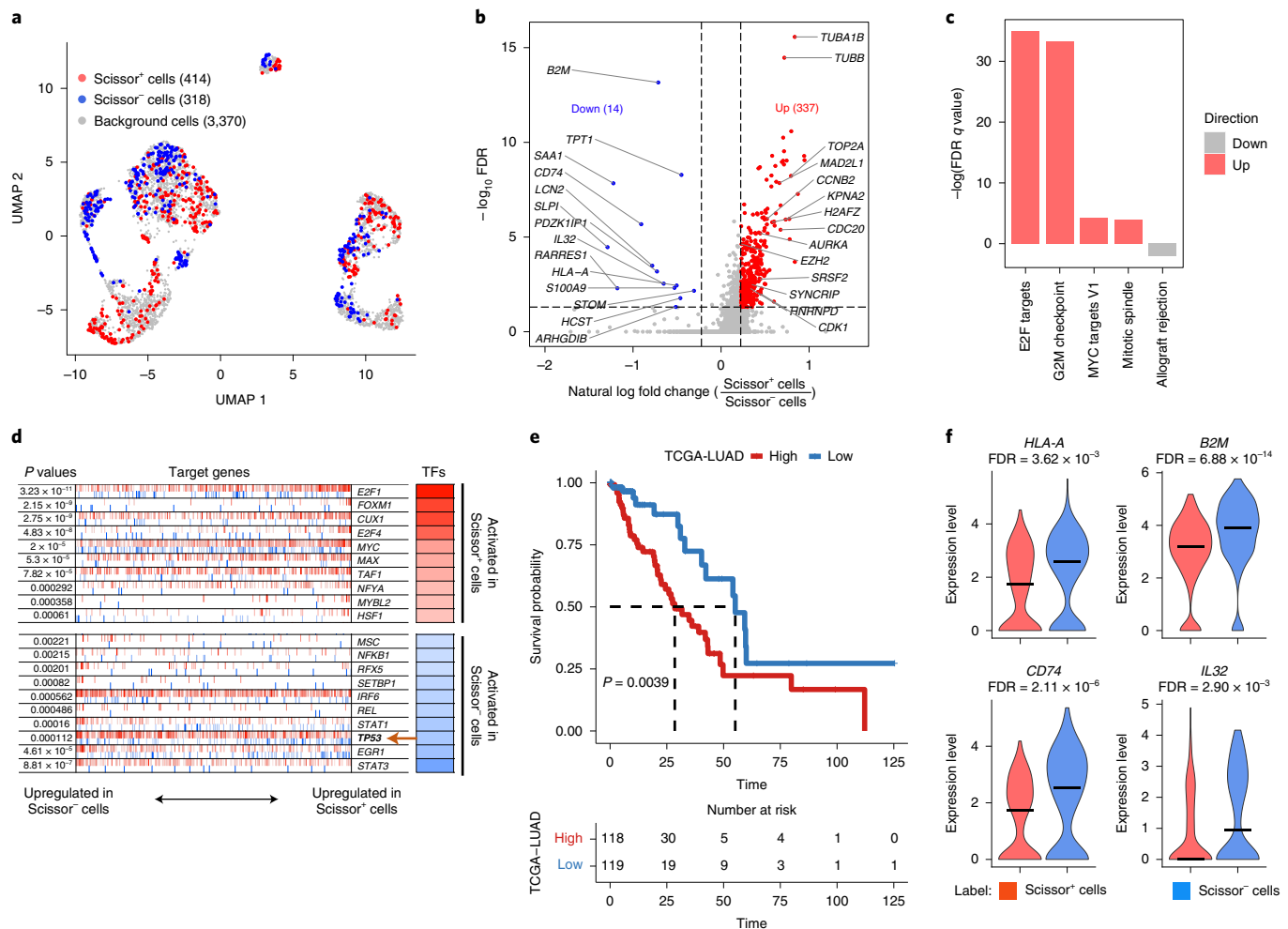


Fig. 3 | Scissor identification results on lung cancer cells guided by TP53 mutation status. **a**, UMAP visualization of the Scissor-selected cells. The red and blue dots are cells associated with the TP53 mutant and wild-type phenotypes, respectively. **b**, Volcano plot of differential gene expressions in Scissor+ cells versus Scissor- cells. Red and blue points mark the genes with significantly increased or decreased expressions in Scissor+ cells (FDR < 0.05 and absolute fold change > 1.25). The two vertical dashed lines represent $\pm \ln(1.25)$ fold changes in gene expression, and the horizontal dashed line denotes an FDR cutoff of 0.05. The FDR was the adjusted P value calculated by the two-tailed Wilcoxon rank sum test. **c**, Enrichment bar plot of selected cell cycle-related pathways in the Hallmark domain. **d**, Master regulator analysis reveals the top 10 most activated TFs in Scissor+ cells and Scissor- cells, respectively. The targets of each TF are shown as tick marks with red vertical lines representing positive targets and blue vertical lines representing negative targets. Each row also illustrates the statistical P value and the inferred differential activity for each TF. **e**, Kaplan-Meier survival curve shows the clinical relevance of the TP53 mutation signature. Tick marks indicate censoring events. A table is attached below to display the number of alive patients at given time points. The statistical P value was determined by the two-tailed log rank sum test. **f**, Violin plots of expression levels of selected downregulated genes in Scissor+ cells. The FDR was the adjusted P value calculated by the two-tailed Wilcoxon rank sum test.

in an independent ICB dataset (Fig. 4g; Student's *t*-test, $P = 5.0 \times 10^{-4}$). Additionally, the upregulated and downregulated genes in our immunotherapy responsive signature were also significantly enriched in responders and non-responders (Fig. 4h; Kolmogorov-Smirnov test, false discovery rate (FDR) = 0.002 and 0.0085, respectively). We further evaluated our immunotherapy responsive signature in five types of tumor-infiltrating lymphocytes (TILs) with distinctive differentiation states. We found that LAG3-low/PD1-low effector CD8 T cells had the highest signature scores, followed by naive CD8, bystander TIL, LAG3-high/PD1-high effector CD8, and exhausted CD8 T cells (Fig. 4i). Notably, our signature can significantly distinguish between PD1-low effector CD8, PD1-high effector CD8 and exhausted CD8 T cells (Fig. 4i and Supplementary Fig. 5a,b). Furthermore, the pseudotime analysis based on our immunotherapy responsive signature revealed the relative orders of the six clusters as cluster 2>3>4>5>6>1 (Supplementary Fig. 5c). Moreover, we found that our signature was more enriched in the

memory precursor CD8 T cells than the short-lived effector T cells (Fig. 4j; Student's *t*-test, $P = 0.02$). This observation indicated that Scissor_FR cells were more like PD1-low memory-precursor cells, which have higher *TCF7* expression and are associated with a good immunotherapy response.

Collectively, our Scissor analysis of an scRNA-seq melanoma dataset independently revealed a *PDCD1/CTLA4* low and *TCF7* high T cell subpopulation whose distinct transcriptome was essential to favorable response to immunotherapy. These results demonstrate that Scissor analysis of single-cell data is capable of identifying subpopulations associated with the specific phenotype, even though the single-cell data itself have no such phenotype information.

Identifying cell subpopulations associated with facioscapulo-humeral muscular dystrophy. We further applied Scissor on a facioscapulo-humeral muscular dystrophy (FSHD) single-cell dataset to explore the applicability of our method on non-cancer studies.

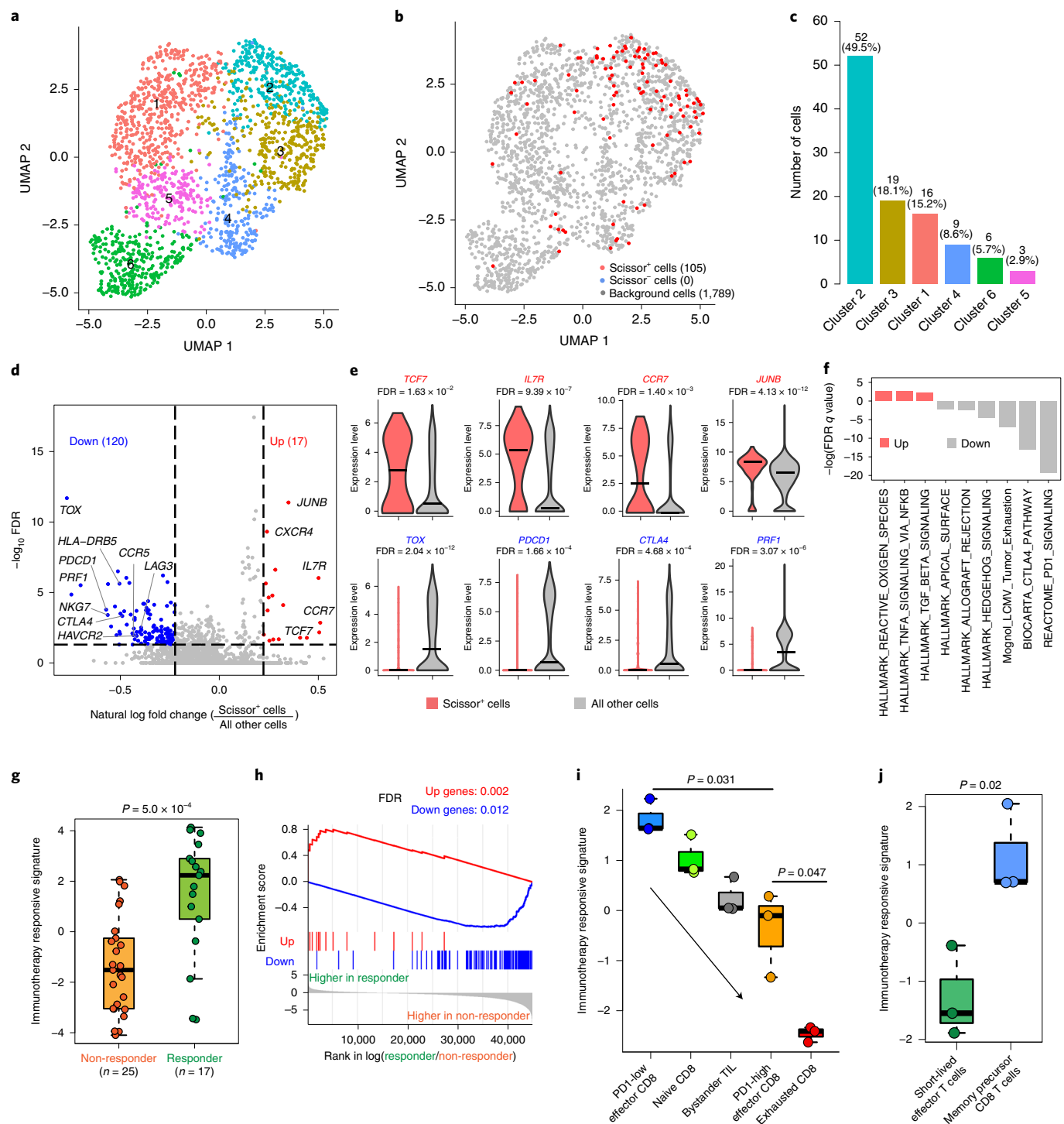


Fig. 4 | Scissor identification results on melanoma T cells. **a**, UMAP visualization of 1,894 melanoma T cells in six clusters. **b**, UMAP visualization of the Scissor-selected cells. **c**, Bar plot shows the distribution of Scissor⁺ cells across the six T cell populations. **d**, Volcano plot of differential gene expressions in Scissor⁺ cells versus all other cells. The two vertical dashed lines represent $\pm \ln(1.25)$ fold changes, and the horizontal dashed line denotes an FDR cutoff of 0.05. The FDR was the adjusted *P* value calculated by the two-tailed Wilcoxon rank sum test. **e**, Violin plots show the expression levels of important immune genes in Scissor⁺ cells. The FDR was the adjusted *P* value calculated by the two-tailed Wilcoxon rank sum test. **f**, Enrichment bar plot shows the significantly enriched pathways in Scissor⁺ cells compared to all other cells (FDR < 0.05). **g**, Box plot shows the enrichment scores of the immunotherapy responsive signature in the non-responders and responders from Sade-Feldman's cohort. **h**, Gene set enrichment analysis plot of the up and down signature genes in the responder versus non-responder comparison from Sade-Feldman's cohort. **i**, Box plot shows the enrichment scores of the immunotherapy responsive signature in five types of CD8 T cells from a mouse liver tumor model (*n* = 3 biologically independent replicates per group from left to right). **j**, Box plot shows the enrichment scores of the immunotherapy responsive signature in the memory precursor CD8 T cells and short-lived effector CD8 T cells (*n* = 3 biologically independent replicates in each condition). The box plot center line and the box limits 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 limits. The statistical *P* value was determined by the two-tailed Student's *t*-test, unless otherwise indicated.

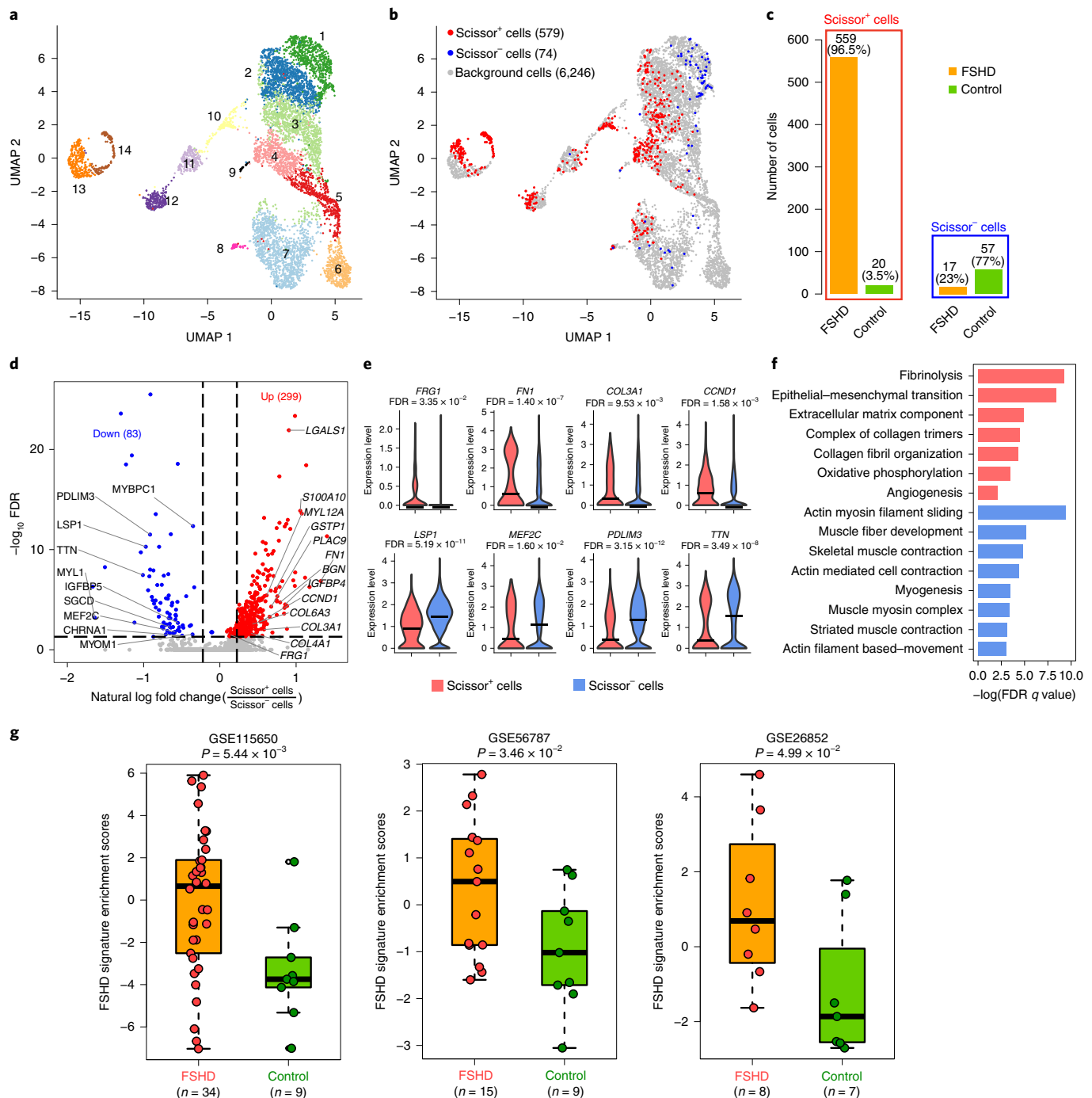


Fig. 5 | Scissor identification results on FSHD cells. **a**, UMAP visualization of 6,899 cells derived from FSHD and control samples. **b**, UMAP visualization of the Scissor-selected cells. The red and blue dots represent Scissor⁺ and Scissor⁻ cells, associated with FSHD and control phenotypes, respectively. **c**, Bar plot shows the detailed phenotypic constitutions of the Scissor-selected cells. **d**, Volcano plot of differential gene expressions in Scissor⁺ cells versus Scissor⁻ cells. Red and blue points mark the genes with significantly increased or decreased expressions in Scissor⁺ cells compared to Scissor⁻ cells (FDR < 0.05 and absolute fold change > 1.25). The two vertical dashed lines represent $\pm \ln(1.25)$ fold changes in gene expression, and the horizontal dashed line denotes an FDR cutoff of 0.05. The FDR was the adjusted P value calculated by the two-tailed Wilcoxon rank sum test. **e**, Violin plots show the expression levels of selected dysregulated genes in Scissor⁺ cells. The FDR was the adjusted P value calculated by the two-tailed Wilcoxon rank sum test. **f**, Enrichment bar plot of selected muscle-related pathways in the Hallmark, GO's biological process and cellular component domains. **g**, Box plots show the enrichment scores of the FSHD molecular signature in the FSHD and normal controls from three independent validation datasets. The statistical P values were determined by the two-tailed Student's t -test. The box plot center line and the box limits 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 limits.

In total, 6,899 cells derived from FSHD and control samples³² were analyzed by Scissor with the guidance of 35 bulk muscle biopsies (27 patients with FSHD and eight controls). These cells were initially

grouped into 14 clusters in scRNA-seq data analysis (Fig. 5a). After integrating these cells and bulk data by Scissor, 579 cells were identified as Scissor⁺ cells, which were associated with FSHD, and 74 cells

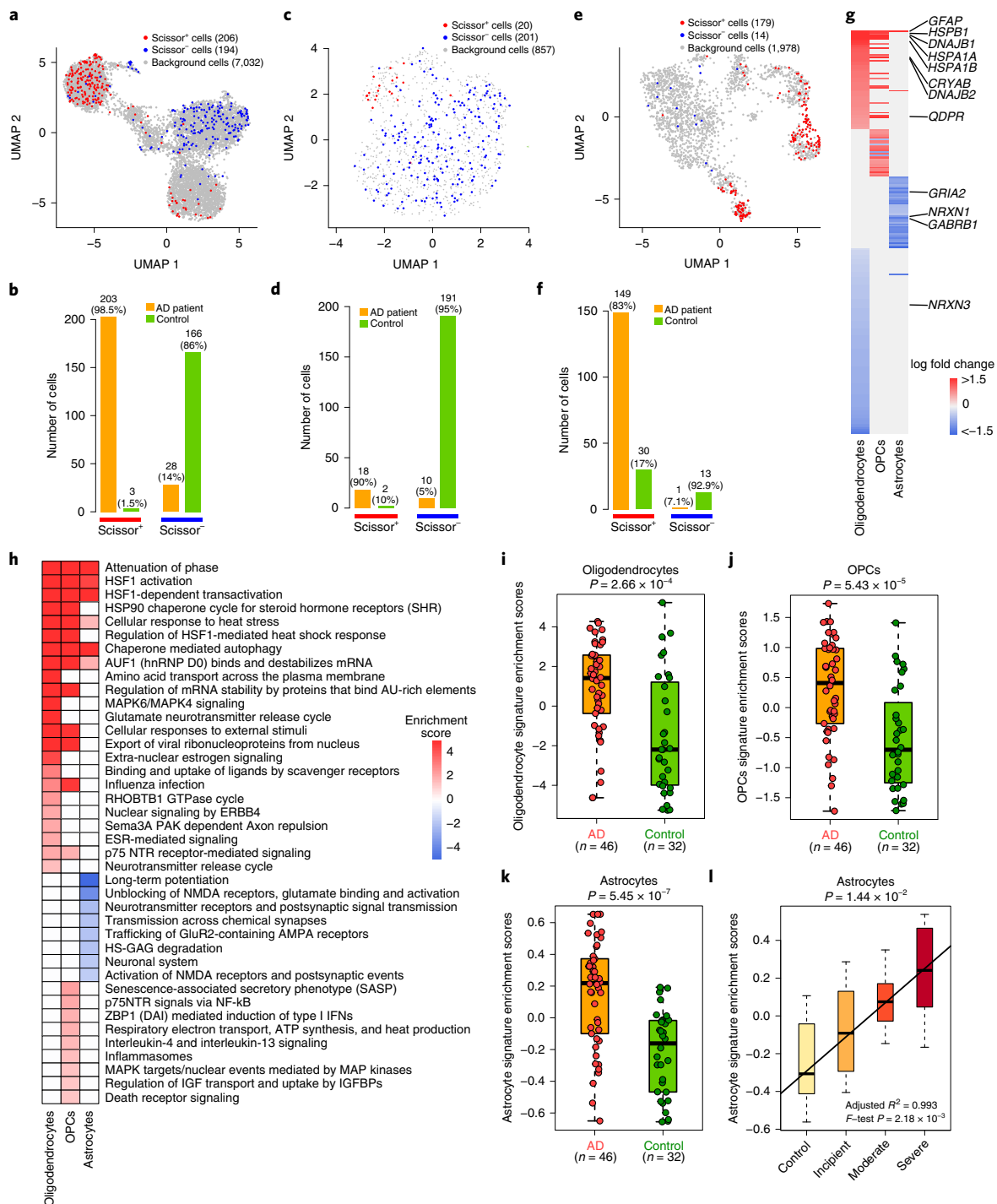


Fig. 6 | Scissor identification results on AD. a–f, UMAP visualizations of the Scissor-selected cells on oligodendrocytes (**a**), OPCs (**c**) and astrocytes (**e**) with corresponding bar plots showing the phenotypic constitutions of the Scissor-selected cells on oligodendrocytes (**b**), OPCs (**d**) and astrocytes (**f**). **g**, Heat map of differential gene fold changes in Scissor⁺ cells versus Scissor⁻ cells across all three brain cell types. Red and blue elements mark the genes with significantly increased or decreased expressions in Scissor⁺ cells (FDR < 0.05 and absolute fold change > 1.25). **h**, Heat map of enriched Reactome pathways. The red and blue elements represent the activated and repressed pathways in corresponding cell types. **i**, Box plot shows the enrichment scores of the oligodendrocytes' molecular signature in patients with AD and normal controls from GSE109887. **j**, Box plot shows enrichment scores of the OPCs' molecular signature in patients with AD and normal controls from GSE109887. **k**, Box plot shows the enrichment scores of the astrocytes' molecular signature in patients with AD and normal controls from GSE109887. **l**, Box plot shows the enrichment scores of the astrocytes' molecular signature in control, incipient-, moderate- and severe-stage patients with AD from GSE28146 ($n = 8$, $n = 7$, $n = 8$ and $n = 7$ biologically independent patients per group from left to right). The statistical P value was determined by the Kruskal-Wallis test. The linear regression line represents the relationship between median enrichment scores and AD stages. The box plot center line and the box limits represent median value and the upper and lower quartiles, respectively. Box whiskers indicate the largest and smallest values no more than 1.5 times the interquartile range from the limits. The statistical P value was determined by the two-tailed Student's t -test, unless otherwise indicated.

marked as Scissor⁻ cells were linked with the normal phenotype (Fig. 5b). Consistent with the encoded phenotype information for Scissor in this application, 559 of 579 Scissor⁺ cells (97.5%) originated from patients with FSHD, and nearly 80% of Scissor⁻ cells were from normal samples (Fig. 5c), indicating that Scissor can identify phenotype-associated subpopulations with high specificity in non-cancer cases.

Given the adequate number of the Scissor-selected cells, we could directly compare the gene expression of Scissor⁺ cells with Scissor⁻ cells to uncover the underlying transcriptional patterns of the identified cells with distinct phenotypes. As a result, 299 upregulated genes and 83 downregulated genes were differentially expressed in Scissor⁺ cells over Scissor⁻ cells, respectively (Fig. 5d and Supplementary Table 4). Among these genes, we found that many downregulated genes in Scissor⁺ cells are involved in normal muscle functions (Fig. 5e), such as myogenesis (*LSP1*, *MEF2C* and *MYL1*), muscle contraction (*IGFBP5*, *MYBPC1* and *MYOM1*) and actin filament or myosin-based functions (*PDLIM3*, *TTN* and *SGCD*). As for the upregulated genes in Scissor⁺ cells, we found that a proportion of these genes might be dysregulated due to fibrotic infiltration³³, which includes extracellular matrix proteins as collagen types III and IV (*COL3A1* and *COL4A1*) and fibronectin (*FN1*). FSHD region gene 1 (*FRG1*) was also over-expressed in Scissor⁺ cells (Fig. 5e), which is crucial for angiogenesis and epithelial-to-mesenchymal transition³⁴. The functional enrichment analysis also confirmed that Scissor⁺ cells were characterized by the reduction of muscle fibers as well as loss of myogenesis and muscle contraction functions (Fig. 5f and Supplementary Table 4), which is consistent with the molecular hallmarks of FSHD^{35,36}. To further demonstrate the characteristics of the phenotype-associated cells identified by Scissor, we built an FSHD molecular signature using the 382 differentially expressed genes between the two cell subpopulations (Supplementary Table 4) and examined the signature's ability in distinguishing patients with FSHD from normal controls. We found that, on all three independent datasets, the enrichment scores of the FSHD molecular signature were significantly higher in patients with FSHD than in normal controls (Fig. 5g; Student's *t*-test, $P = 5.44 \times 10^{-3}$, 3.46×10^{-2} and 4.99×10^{-2} , respectively).

Therefore, our Scissor analysis of FSHD scRNA-seq data identified disease-associated subpopulations characterized by the dysfunction of fibrinolysis and myogenesis as well as muscle contractions, which substantiates the utility of Scissor on non-cancer studies.

Discerning subpopulations related to Alzheimer's disease.

Single-cell technology presents immense opportunities to disentangle cellular diversity and alterations in neurological disorders. In this study, we applied Scissor on three brain cell types from an Alzheimer's disease (AD) scRNA-seq study³⁷ to explore the cell subpopulations that are most highly associated with the disease status and progression. Collectively, 7,432 oligodendrocytes, 1,078 oligodendrocyte progenitor cells (OPCs), and 2,171 astrocytes (Supplementary Fig. 6a) were analyzed separately by Scissor with the guidance of 14 bulk samples (seven patients with AD and seven healthy controls). For oligodendrocytes, Scissor identified 206 Scissor⁺ cells associated with the patients with AD and 194 Scissor⁻ cells associated with the healthy controls (Fig. 6a). When compared to the cell sources, 203 of 206 Scissor⁺ cells (99%) were from patients with AD, and 86% of Scissor⁻ cells were from normal samples (Fig. 6b). For OPCs, 20 Scissor⁺ cells and 201 Scissor⁻ cells were identified (Fig. 6c), with 90% of Scissor⁺ cells from patients with AD and 95% of Scissor⁻ cells from normal samples (Fig. 6d). As for astrocytes, the identified cell subpopulations consisted of 179 Scissor⁺ cells and 14 Scissor⁻ cells (Fig. 6e), with 83% of Scissor⁺ cells from patients with AD and 93% of Scissor⁻ cells from normal samples, respectively (Fig. 6f). Therefore, Scissor successfully

associated cell subpopulations with the desired phenotype in all three brain cell types.

We then investigated DEGs between Scissor⁺ cells and Scissor⁻ cells within each cell type (Fig. 6g, Supplementary Fig. 6b and Supplementary Table 5). In oligodendrocytes and OPCs, multiple heat shock and chaperone genes involving proper protein folding (for example, *HSPA1A*, *HSPB1* and *DNAJB1*) were upregulated in Scissor⁺ cells, consistent with the previous observation³⁸. Pathway enrichment analysis confirmed that the HSF1 pathway, a master activator of chaperone gene expression³⁹, is highly activated in Scissor⁺ cells to protect cells against protein misfolding (Fig. 6h and Supplementary Table 5). In astrocytes, most DEGs were downregulated in Scissor⁺ cells, including *GRIA2*, which increases NMDA receptor activity⁴⁰. Pathway enrichment results also suggested the impaired NMDA-dependent long-term potentiation in Scissor⁺ astrocytes (Fig. 6h and Supplementary Table 5), which has been observed in AD⁴¹. Additionally, *GFAP* (glial fibrillary acidic protein) was coordinately upregulated in Scissor⁺ cells across all three cell types (Fig. 6g). It has been reported that *GFAP* expressions correlate with amyloid- β plaque density in AD brain tissue⁴², and elevated *GFAP* levels in the blood positively correlate with cognitive impairment⁴³. All these results showed that Scissor indeed identified cell subpopulations with characteristics consistent with the corresponding bulk phenotype.

To further demonstrate the phenotypic associations of the cell subpopulations identified by Scissor, we constructed molecular signatures based on the DEGs in Scissor-identified cell subpopulations (Supplementary Table 5) and used independent AD datasets to evaluate the functions of these signatures. As a result, the enrichment scores of the corresponding molecular signatures in each cell type were significantly higher in patients with AD than in normal controls (Fig. 6i–k; Student's *t*-test, $P = 2.66 \times 10^{-4}$ for oligodendrocytes, 5.43×10^{-5} for OPCs and 5.45×10^{-7} for astrocytes). Notably, we also found that our astrocytes' molecular signature is positively associated with the progression of AD from incipient to moderate to severe conditions (Fig. 6l; Kruskal–Wallis test, $P = 1.44 \times 10^{-2}$). Thus, this Scissor⁺ cell subpopulation of astrocytes could play a vital role in the early stages of AD.

Taken together, our Scissor analysis identified cell subpopulations that are most highly associated with AD in three brain cell types, which could contribute to comprehending the underlying pathogenesis of AD and might facilitate disease diagnosis and therapy.

Discussion

Identifying the phenotype-specific cell subpopulations from single-cell data can give rise to breakthroughs in understanding disease mechanisms. Multiple efforts have been made to discern disease-relevant cells from single-cell data⁴⁴. For example, HoneyBadger⁴⁵ and inferCNV⁴⁶ can identify cancerous cells by predicting copy number variations from scRNA-seq data. However, there are a variety of external phenotypes beyond tumor-versus-normal in cancer and non-cancer diseases, such as treatment resistance, disease stages, survival outcomes and aging, which are widely available in bulk data. Therefore, there is a great need for further methodological progress to use these abundant phenotypes in single-cell data analysis. To this end, we introduce Scissor as a new computational tool to leverage the phenotype information from bulk data to identify the most highly phenotype-associated cell subpopulations.

We demonstrated the broad utilities of Scissor in a total of ten applications across different diseases, ranging from engineered simulations, cancer cells, immune cells and FSHD cells to multiple types of brain cells in AD (Supplementary Table 6). One of the advantages is that Scissor does not require any unsupervised clustering on single-cell data, which avoids subjective decisions of cell cluster numbers or clustering resolution⁴⁷. Most importantly,

Scissor provides a flexible framework to integrate various external phenotypes in bulk data to guide single-cell data analysis, enabling hypothesis-free identification of clinically and biologically relevant cell subpopulations.

The signs of the regression coefficients were used to infer the relation of the Scissor-selected cells with the phenotypes. We designed generalized criteria for proper interpretation of Scissor results (Supplementary Table 7 and Methods) because we observed little proportions of negative correlations between single-cell and bulk samples in real applications (Supplementary Table 8). Based on our criteria, those little proportions of negative correlations did not affect our current results and interpretations. Furthermore, we also performed correlation evaluations for the Scissor-selected cells in each application, which could help check the validity of the Scissor result (Supplementary Tables 9–11). In addition, we tested Scissor's performance in various perturbed bulk samples and demonstrated that the Scissor identification results were robust to noise (Supplementary Figs. 7 and 8 and Supplementary Table 12) and had no model-driven preferences to experimental noise or bias (Supplementary Fig. 9). For computational efficiency, the running time and memory usages of Scissor in real applications are acceptable, with cell numbers ranging from ~1,000 to ~30,000 (Supplementary Table 13).

The construction of the correlation matrix is a key step in Scissor to quantify the dependence or similarity between the single-cell data and bulk data. Except for the Pearson correlation applied in Scissor, we will explore other similarity measurements, such as entropy-based mutual information and hypothesis testing-based methods. Furthermore, although we only conducted the integration of bulk gene expressions with scRNA-seq data, Scissor can also be applied to other single-cell measurements, such as chromatin accessibility⁴⁸ and DNA methylation⁴⁹.

Overall, Scissor demonstrates promise for integrating single-cell data with phenotype information to dissect clinically significant subsets from heterogeneous cell populations. This strategy will boost biological discoveries and interpretation. We anticipate that Scissor will enable a broad application of widely available phenotype information on single-cell data analysis and help unravel the most disease-relevant subpopulations for cell-targeted therapies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41587-021-01091-3>.

Received: 5 June 2020; Accepted: 10 September 2021;

Published online: 11 November 2021

References

- Zhang, Q. et al. Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* **179**, 829–845 (2019).
- Yofe, I., Dahan, R. & Amit, I. Single-cell genomic approaches for developing the next generation of immunotherapies. *Nat. Med.* **26**, 171–177 (2020).
- Wagner, J. et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **177**, 1330–1345 (2019).
- Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
- Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
- Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
- Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
- Miao, Y. et al. Adaptive immune resistance emerges from tumor-initiating stem cells. *Cell* **177**, 1172–1186 (2019).
- Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308 (2018).
- Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
- Guo, X. et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).
- Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
- Karaayvaz, M. et al. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, 3588 (2018).
- Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
- Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
- Lawson, D. A. et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **526**, 131–135 (2015).
- Brady, S. W. et al. Combating subclonal evolution of resistant cancer phenotypes. *Nat. Commun.* **8**, 1231 (2017).
- Ryan, H. E. et al. Hypoxia-inducible factor-1 α is a positive factor in solid tumor growth. *Cancer Res.* **60**, 4010–4015 (2000).
- Gentles, A. J. et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
- Wilson, W. R. & Hay, M. P. Targeting hypoxia in cancer therapy. *Nat. Rev. Cancer* **11**, 393–410 (2011).
- Santoro, A. et al. p53 loss in breast cancer leads to Myc activation, increased cell plasticity, and expression of a mitotic signature with prognostic value. *Cell Rep* **26**, 624–638 (2019).
- Barsotti, A. M. & Prives, C. Pro-proliferative FoxM1 is a target of p53-mediated repression. *Oncogene* **28**, 4295–4305 (2009).
- Perri, F., Piscanti, S. & Della Vittoria Scarpati, G. P53 mutations and cancer: a tight linkage. *Ann. Transl. Med.* **4**, 522 (2016).
- Sade-Feldman, M. et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).
- Robert, C. et al. Pembrolizumab versus Ipilimumab in advanced melanoma. *N. Engl. J. Med.* **372**, 2521–2532 (2015).
- Weber, J. S. et al. Nivolumab versus chemotherapy in patients with advanced melanoma who progressed after anti-CTLA-4 treatment (CheckMate 037): a randomised, controlled, open-label, phase 3 trial. *Lancet Oncol.* **16**, 375–384 (2015).
- Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
- Hugo, W. et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* **165**, 35–44 (2016).
- Van Allen, E. M. et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).
- Chen, Z. et al. TCF-1-centered transcriptional network drives an effector versus exhausted CD8 T cell-fate decision. *Immunity* **51**, 840–855 (2019).
- Siddiqui, I. et al. Intratumoral Tcf1⁺PD-1⁺CD8⁺ T cells with stem-like properties promote tumor control in response to vaccination and checkpoint blockade immunotherapy. *Immunity* **50**, 195–211 (2019).
- van den Heuvel, A. et al. Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. *Hum. Mol. Genet.* **28**, 1064–1075 (2019).
- Cooper, D. & Upadhyaya, M. *Facioscapulohumeral Muscular Dystrophy (FSHD): Clinical Medicine and Molecular Cell Biology* (Taylor & Francis, 2004).
- Tiwari, A., Pattnaik, N., Mohanty Jaiswal, A. & Dixit, M. Increased FSHD region gene1 expression reduces in vitro cell migration, invasion, and angiogenesis, ex vivo supported by reduced expression in tumors. *Biosci. Rep.* **37**, BSR20171062 (2017).
- Lassche, S. et al. Sarcomeric dysfunction contributes to muscle weakness in facioscapulohumeral muscular dystrophy. *Neurology* **80**, 733–737 (2013).
- Banerji, C. R. S. et al. Dynamic transcriptomic analysis reveals suppression of PGC1 α /ERR α drives perturbed myogenesis in facioscapulohumeral muscular dystrophy. *Hum. Mol. Genet.* **28**, 1244–1259 (2019).
- Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
- Ashraf, G. M. et al. Protein misfolding and aggregation in Alzheimer's disease and type 2 diabetes mellitus. *CNS Neurol. Disord. Drug Targets* **13**, 1280–1293 (2014).
- Neef, D. W., Jaeger, A. M. & Thiele, D. J. Heat shock transcription factor 1 as a therapeutic target in neurodegenerative diseases. *Nat. Rev. Drug Discov.* **10**, 930–944 (2011).
- Yu, S. P., Sensi, S. L., Canzoniero, L. M., Buisson, A. & Choi, D. W. Membrane-delimited modulation of NMDA currents by metabotropic glutamate receptor subtypes 1/5 in cultured mouse cortical neurons. *J. Physiol.* **499**, 721–732 (1997).

41. Prieto, G. A. et al. Pharmacological rescue of long-term potentiation in Alzheimer diseased synapses. *J. Neurosci.* **37**, 1197–1212 (2017).
42. Muramori, F., Kobayashi, K. & Nakamura, I. A quantitative study of neurofibrillary tangles, senile plaques and astrocytes in the hippocampal subdivisions and entorhinal cortex in Alzheimer's disease, normal controls and non-Alzheimer neuropsychiatric diseases. *Psychiatry Clin. Neurosci.* **52**, 593–599 (1998).
43. Chatterjee, P. et al. Plasma glial fibrillary acidic protein is elevated in cognitively normal older adults at risk of Alzheimer's disease. *Transl. Psychiatry* **11**, 27 (2021).
44. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
45. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* **28**, 1217–1227 (2018).
46. inferCNV of the Trinity CTAT Project. <https://github.com/broadinstitute/inferCNV>
47. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).
48. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
49. Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Phenotype-guided single-cell subpopulation identification by Scissor. The workflow of Scissor is shown in Fig. 1. Denote m and n as the cell number and the bulk sample number, respectively. The three data sources for Scissor inputs are a single-cell expression matrix with m cells, a bulk profiling data with n samples and a sample phenotype Y of interest (Fig. 1a). Y annotates each bulk sample and can be a continuous dependent variable, binary group indicator vector or clinical survival data. Scissor first uses quantile normalization on the single-cell and bulk expression data to remove the underlying batch effect. After this, a Pearson correlation matrix $S = (S_{ij})_{n \times m}$ is calculated for each pair of cells and bulk samples to quantify the similarity between the single-cell data and bulk data, where s_{ij} is the correlation of sample i and cell j across common genes in normalized single-cell and bulk data.

Scissor optimizes a regression model on the correlation matrix S with the sample phenotype Y (Fig. 1b). Let β denote a vector of coefficients on cells and $l(\beta)$ denote an appropriately chosen log-likelihood function. The formula of $l(\beta)$ depends on the input phenotype Y —for example, linear regression for continuous dependent variables, logistic regression (classification) for dichotomous variables and Cox regression for clinical survival data (see more details in the next section). Because a cell subpopulation alone could drive the phenotype of interest, we impose a sparse penalty (l_1 -norm) on the model to select the high-confidence cells that are important for the given phenotype. Furthermore, the network-based penalty enforces the tightly connected nodes (cells) in the network to have more similar coefficients³⁰, making the phenotype-to-cell association results more consistent and interpretable. Inspired by this, we use the shared nearest neighbor graph calculated in Seurat³¹ to serve as a cell–cell similarity network G and impose a corresponding graph regularization on the regression model.

Overall, Scissor is formulated as the following network regularized sparse regression model (Fig. 1b):

$$\min_{\beta} -\frac{1}{n} l(\beta) + \lambda \left\{ \alpha \|\beta\|_1 + \frac{1-\alpha}{2} \beta^T L \beta \right\},$$

where L is a symmetric normalized Laplacian matrix, which is defined as

$$L = D^{-\frac{1}{2}} (D - A) D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}},$$

where $A = (a_{ij})_{m \times m}$ is a binary or weighted adjacency matrix of G . a_{ij} equals one or a value ranging from 0 to 1 if cell i and j are connected in G and $a_{ij} = 0$, otherwise. $D = (d_{ij})_{m \times m}$ is the degree matrix of G , where $d_{ii} = \sum_{j=1}^m a_{ij}$, and $d_{ij} = 0$ for $i \neq j$. The tuning parameter λ controls the overall strength of the penalty, and α balances the amount of regularization for smoothness and sparsity.

The non-zero coefficients of β solved by the above optimization model are used to select the cell subpopulations associated with the phenotype of interest (Fig. 1c). According to the sign of β , we denote the selected cells by Scissor⁺ cells and Scissor[−] cells, which are positively and negatively associated with the phenotype of interest, respectively. The cells with coefficients of zero are denoted as background cells. Furthermore, to control false associations between the single-cell and bulk data, we designed a reliability significance test to determine whether the chosen data are suitable for our phenotype-to-cell associations (Fig. 1d). Finally, the Scissor-selected cells will further be investigated by several downstream analyses, such as the differential expression gene analysis, functional enrichment analysis and motif analysis, to reveal the underlying biological mechanisms of the selected cell subpopulations (Fig. 1e).

Log-likelihood functions. The formula of log-likelihood function $l(\beta)$ depends on the type of the phenotype Y . Scissor can do the following: (1) linear regression for continuous dependent variables; (2) classification for dichotomous variables; and (3) Cox regression for clinical survival data.

In detail, denote $S_i = (s_{i1}, s_{i2}, \dots, s_{im})^T$ as the correlation coefficients for sample i across all m cells. If $Y = (y_1, y_2, \dots, y_n)^T$ is a continuous variable, the linear regression log-likelihood function is used:

$$l(\beta) = -\sum_{i=1}^n (y_i - \beta^T S_i)^2$$

If Y is a binary group indicator vector—for example, $y_i \in \{0, 1\}$, the logistic regression log-likelihood function is used:

$$l(\beta) = \sum_{i=1}^n \left[y_i \beta^T S_i - \log \left(1 + \exp(\beta^T S_i) \right) \right]$$

For time-to-event outcomes subject to independent censoring, the Cox regression is considered. Let T_i be the non-negative event time and C_i be the censoring time. Denote $\hat{T}_i = \min(T_i, C_i)$ as the observed event time or censoring time, and $\delta_i = I(T_i \leq C_i)$ as the event indicator, where $I(\cdot)$ is an indicator function. The following log-likelihood function is used:

$$l(\beta) = \sum_{i=1}^n \delta_i \left[\beta^T S_i - \log \left(\sum_{k \in R_i} \exp(\beta^T S_k) \right) \right]$$

where $R_i = \{k : \hat{T}_k \geq \hat{T}_i\}$ denotes the risk set at time \hat{T}_i .

Parameter tunings and implementations. The algorithm proposed by Li et al.³² is used to solve the above network regularized sparse regression. In Scissor, two model parameters need to be determined. First, parameter λ controls the overall strength of the whole penalty term. For a fixed α , we set 100 possible λ and applied ten-fold cross-validation to select the optimal λ with the minimum averaged error. Second, parameter $\alpha \in [0, 1]$ balances the effect of the l_1 -norm and the network-based penalties. A larger α inclines to emphasize l_1 -norm to encourage sparsity, and a smaller α gives more weight to the network term to encourage selecting similar cells. In real applications, a fixed value of α cannot work for all kinds of datasets because different datasets could have different sensitivities to the changes of α . To select as many similar cells as possible from similar cell types, we started to search α from a small value, and the default search list for α in Scissor is $\{0.005, 0.01, 0.05, 0.1, 0.2, 0.3, \dots, 0.9\}$. For each fixed α , λ is determined by grid search based on the cross-validation. The goal of Scissor is to identify a small group of cells that are most highly correlated with the specific phenotypes with high confidence. Based on this motivation as a priori, we determined α using the following criteria: the number of Scissor-selected cells should not exceed a certain percentage of total cells (default 20%) in the single-cell data. In each experiment, a search on the above searching list is performed from the smallest to the largest until a value of α meets the above criteria.

Reliability significance test. We designed a reliability significance test to exclude the false association between the identified cell subpopulations and bulk phenotypes (Fig. 1d). This statistical test can determine whether the inferred phenotype-to-cell associations are reliable (statistical P value less than 0.05) or are false positives. Our motivation for the test is as follows: if the chosen single-cell and bulk data are not suitable for the phenotype-to-cell associations, the correlations would be less informative and not well associated with the phenotype labels. Thus, the corresponding prediction performance would be poor and not be significantly distinguishable from the randomly permuted labels. Inspired by this, we used the following procedures to perform the reliability significance test.

First, we performed k -fold cross-validation (CV) on correlation matrix S and used the training sets only to estimate the coefficients of cells in Scissor. The prediction performances of the trained Scissor models were evaluated on testing sets, and an averaged evaluation measurement was obtained to serve as an actual test statistic. Second, we randomly permuted the bulk sample labels multiple times to break up the original bulk phenotype–genotype relationships. By performing the same Scissor analysis and CV evaluation using each permuted bulk data, we obtained a background distribution of the corresponding evaluation measurement quantifying the prediction performances at the random level. Finally, the actual test statistic calculated in the original data was compared to the background distribution values. The reliability significance test P value was the number of the permutation-based test statistics above (or below) the actual test statistic, divided by the permutation times. In this study, the evaluation measurements used in the reliability significance test are the mean squared error (MSE) for linear regression (smaller is better), the area under the ROC curve (AUC) for classification (higher is better) and the concordance index (C-index) for Cox regression (higher is better).

Next, we explored whether the reliability significance test can effectively detect the ‘false-positive’ associations between the single-cell and bulk data by reporting non-significant P values. We performed two kinds of randomizations on bulk data to generate false-positive examples. After Scissor analysis of those randomized bulk data, our reliability significance test reported that the P values for these counter-examples were all greater than 0.05 (Supplementary Table 14), indicating that these randomized bulk data were not suitable to use in Scissor to identify the corresponding phenotype-associated cells. We also performed the reliability significance test on the main applications used in this study and found that the test P values were all less than 0.05 (Supplementary Fig. 10), indicating that the inferred phenotype-to-cell associations are reliable.

Scissor selected cells interpretations. Both Scissor⁺ cells and Scissor[−] cells are the Scissor-selected cells that are most highly associated with the specified phenotypes, corresponding to the cells with the estimated coefficients greater than and less than zero, respectively. The associations between the Scissor-selected cells and phenotypes depend on the model in use and should be interpreted in a context-specific manner. For both linear regression and classification models, the initial values in Y will affect the interpretations. The Scissor⁺ cells will be associated with the phenotypes encoded as a higher value in Y , and Scissor[−] cells will correspond to the phenotypes encoded as a lower value. For example, in Scissor's application on the FSHD single-cell dataset, if the patients with FSHD are encoded as one, and the normal controls are encoded as zero in Y , Scissor⁺ cells will be associated with the FSHD, and Scissor[−] cells will be associated with the normal phenotype. If the encoding for the two phenotypes is reversed in Y , the interpretations of Scissor⁺ cells and Scissor[−] cells are reversed accordingly. For Cox regression, Scissor⁺ cells are always associated with worse survival, and Scissor[−] cells are associated with good survival.

Scissor can associate cells with phenotypes, and this kind of association is a relative concept between phenotypes. Namely, Scissor assigns which phenotype a cell is more likely associated with than the other phenotype. Considering the

possible negative correlations between the single-cell and bulk samples, we can further interpret a cell by assigning it to the following three categories: if the average of a cell's correlations with all bulk samples is greater than zero, and the number of positive correlations is larger than the number of the negative correlations, this cell is more similar to the associated phenotype; if the average of a cell's correlations is less than zero, and the number of negative correlations is larger than the number of positive correlations, this cell should be interpreted as more dissimilar to the other phenotype; otherwise, this cell's association with the phenotype is undeterminable (Supplementary Table 7). In most cases, the negative correlation values are very few, and the identified cells fall into the 'more similar' category.

In total, there are ten applications of Scissor in this study. The corresponding phenotype encoding and the interpretations of the Scissor-selected cells were summarized in Supplementary Table 6. We showed that Scissor was capable of detecting both Scissor⁺ and Scissor⁻ cells through a variety of applications. In some datasets, Scissor⁺ cells and Scissor⁻ cells can be largely unbalanced. Scissor can detect more cells of one phenotype than the other phenotype depending on how strong the cells in a dataset are associated with the specific phenotypes.

Simulation setup. We used Splatter¹⁴ to simulate a single-cell dataset with 1,000 cells and 5,000 expression genes. These cells were from three simulated cell types with group probabilities of 0.8, 0.1 and 0.1, respectively, leading to one large and two small cell subpopulations. The probabilities of a gene being differentially expressed in each of the three groups were set as 0.1, 0.01 and 0.01. Then, we assigned these cells to two different phenotypes (named phenotype I and phenotype II) to simulate the known relationships between cells and phenotypes as the ground truth. To achieve this, we set the large cell subpopulation as common cells shared by the two phenotypes and assigned the other two small cell subpopulations to each phenotype, respectively. These two small subset cells are unique to each phenotype and, thus, can be viewed as the ground truth phenotype-specific cell subpopulations. To simulate the expression profiles of bulk samples for each phenotype, we randomly selected 1,000 cells with replacement from the cell subpopulations contained by each phenotype and then averaged the expressions. In this way, we generated bulk gene expressions of 50 samples for each phenotype.

The differential expression genes between the two cell subpopulations were called by Seurat⁵¹ using the default two-tailed Wilcoxon rank sum test. To explore how the performance of Scissor was affected by the true DEG signals at the bulk level, we simulated another bulk data that the true DEGs are not differentially expressed. In detail, we first selected the least differentially expressed gene from bulk data and then used its expression profiles to replace the expressions of all the true DEGs. This true DEG signal-removed bulk expression data would make Scissor more challenging to capture the real phenotype-to-cell associations.

We used precision and recall measurements in supervised learning to evaluate the performance of Scissor. The precision is defined as the proportion of the Scissor-identified ground truth cells among all Scissor-selected cells, and recall is the proportion of the Scissor-identified ground truth cells among all ground truth cells.

scRNA-seq data pre-processing. We applied Seurat R package (v3.2.1)⁵¹ functions to pre-process the scRNA-seq data used in Scissor. Similarly to the standard quality control step in single-cell data analysis, we first filtered out the genes with low expressions in cells. After this, we normalized the filtered expression matrix using the NormalizeData function with the default parameters. Furthermore, we focused on the genes that exhibited high cell-to-cell variation and identified the variable genes using the FindVariableFeatures function with the default 'vst' method. We then performed the principal component analysis on the scaled variable gene expressions. Using the first ten principal components, we constructed a shared nearest neighbor graph for cells and identified cell clusters using the FindClusters function. Finally, we applied the uniform manifold approximation and projection (UMAP)⁵³ technique using the RunUMAP function to visualize cells in low dimensions.

Differential expression gene analysis. The differentially expressed genes in Scissor⁺ cells versus Scissor⁻ cells (or all other cells in some applications) were identified using the FindMarkers function in Seurat⁵¹ with the default Wilcoxon rank sum test. We required the following criteria to obtain the differential expression genes. First, the gene expression between the two groups is statistically significant with an FDR less than 0.05. Second, the absolute value of expression fold change between the two groups is greater than 1.25. We used the VlnPlot function to exhibit the relative expression levels of the selected differential expression genes.

We built several Scissor-derived signatures based on the differential expression genes in each experiment. The complete lists of signature genes can be found in Supplementary Tables.

Pathway enrichment analysis. The pre-ranked version of CAMERA⁵⁴ (function name: cameraPR) in the limma R package (v3.42.2) was used to evaluate the enriched pathways. cameraPR needs a pre-ranked gene list according to a user-defined statistic score. This study used the differential expression genes output of the FindMarkers function to obtain the pre-ranked gene list. In detail, the statistic ranking score for each gene was calculated using the following formula:

$$\log_{FC} * \log_{10} \frac{1}{FDR + 10^{-100}}$$

where logFC stands for the log-transformed fold change between the two groups, and FDR is the adjusted Wilcoxon rank sum test *P* value. The pathways used in CAMERA were downloaded from the Molecular Signatures Database (MSigDB, v7.3)⁵⁵. The Mognol_LCMV_Tumor_Exhaustion gene set contains the upregulated genes in exhausted T cells from both OT-I mouse model and LCMV-infected mice⁵⁶.

Survival analysis. We performed some survival analyses to examine the clinical relevance of the Scissor-derived signatures (lung cancer signature and TP53 mutation signature). The gene set variation analysis (GSVA) algorithm⁵⁷ with the default settings, as implemented in the GSVA R package (v1.34.0), was applied to calculate our signature score for each sample. Next, the samples were stratified into two groups based on the quantile values of the signature scores (upper quartile versus lower quartile). Survival curves of these two groups of patients were estimated by the Kaplan–Meier method, with statistical significance calculated using the log rank test.

We used the univariate Cox proportional hazard model to examine the association between survival time and clinical features (Pathological Stage: Stage I = 1, Stage II = 2, Stage III = 3; Sex: 'Male' = 1, 'Female' = 2; Age at diagnosis). The significant prognostic factors (*P* < 0.05) in the univariate Cox proportional hazard model were included in the subsequent multivariable Cox proportional hazard model. The Kaplan–Meier estimator, log rank test, Wald test and Cox proportional hazard models were calculated in the survival R package (v3.2.3).

Motif analysis. oPOSSUM software⁵⁸ (v3.0) was used to detect the motifs of the transcription factor binding sites for the lung cancer signature genes. We used the TF binding site annotations provided by the tool (JASPAR CORE Profiles). For the parameter settings, we set 2,000 base pairs in upstream and downstream sequences and kept other parameters as default values. Only motifs with a *z*-score greater than 10 and a Fisher score greater than 7 were reported in the final motif list.

Master regulator analysis. TF activity was inferred using the master regulator inference algorithm (MARINA)⁵⁹ in the viper R package (v1.20.0). A pre-ranked gene list with scores and a regulatory network are the two data sources required as the inputs. In this study, we used the same pre-ranked gene list calculated in the pathway enrichment analysis. The TF regulome was curated from several databases as previously described⁶⁰.

Signature enrichment score calculation. A pseudo-regulon was built based on the significantly upregulated and downregulated genes between the cell subpopulations identified by Scissor. These genes were served as the positive and negative target genes of the constructed pseudo-regulon, respectively. The single-sample extension of MARINA⁶¹ (function name: viper) in the viper R package (v1.20.0) was employed to infer the activity of the constructed pseudo-regulon, which was used as the signature enrichment score for each sample.

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

Data availability

All datasets analyzed in this study were published previously. The corresponding descriptions and pre-processing steps are described in the Supplementary Materials.

Software availability

The open-source Scissor R package and tutorial are available at GitHub: <https://github.com/sunduanchen/Scissor>.

References

- Li, C. & Li, H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175–1182 (2008).
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
- Li, X., Xie, S., Zeng, D. & Wang, Y. Efficient ℓ_0 -norm feature selection based on augmented and penalized minimization. *Stat. Med.* **37**, 473–486 (2018).
- Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
- Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).
- Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).

56. Mognol, G. P. et al. Exhaustion-associated regulatory regions in CD8⁺ tumor-infiltrating T cells. *Proc. Natl Acad. Sci. USA* **114**, E2776–E2785 (2017).
57. Hanzelmann, S., Castelo, R. & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
58. Kwon, A. T., Arenillas, D. J., Worsley Hunt, R. & Wasserman, W. W. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda)* **2**, 987–1002 (2012).
59. Lefebvre, C. et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* **6**, 377 (2010).
60. Robertson, A. G. et al. Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer Cell* **32**, 204–220 (2017).
61. Alvarez, M. J. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).

Acknowledgements

This work was supported by the following funding: NIH 5K01LM012877 (to Z.X.); NIH 1R21HL145426 (to Z.X.); NIH 1R01CA207377 (to D.Z.Q.); NIH NIGMS MIRA R35GM124704 (to A.C.A.); the Medical Research Foundation of Oregon (to Z.X.); NCI R01 CA251245, P50 CA097186, P50 CA186786, P50 CA186786-07S1 and Department of Defense Impact Award W81XWH-16-1-0597 (to J.J.A.); and NCI R01CA244576 (to A.V.D.). We thank W. Anderson and A. Hill for editing the manuscript. The resources of the Exacloud high-performance computing environment, developed jointly by Oregon Health & Science University (OHSU) and Intel, and the technical support of the OHSU Advanced Computing Center are gratefully acknowledged.

Author contributions

D.S. and Z.X. conceived the idea, implemented the algorithm and performed the analyses. D.S., G.X., P.T.S. and Z.X. interpreted the results. A.E.M., L.Y.W., D.Z.Q., P.S., M.D., A.V.D., J.J.A. and A.C.A. provided scientific insights on the applications. Z.X. supervised the study. D.S. and Z.X. wrote the manuscript with feedback from all other authors. All authors read and approved the final manuscript.

Competing interests

A.E.M. discloses receipt of a sponsored research agreement from AstraZeneca. A.V.D. reports consultancy from Abbvie, Beigene, Celgene, Curis, Janssen, Karyopharm, Nurix, Seattle Genetics, Teva Oncology and TG Therapeutics; research funding from Aptose Biosciences, Bristol Myers Squibb, Gilead Sciences and Takeda Oncology; and consultancy and research funding from AstraZeneca, Bayer Oncology, Genentech and Verastem Oncology. J.A.A. has received consulting income from Janssen Biotech, Merck Sharp & Dohme and Dendreon and honoraria for speaker's fees from Astellas. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41587-021-01091-3>.

Correspondence and requests for materials should be addressed to Zheng Xia.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

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

Reporting Summary

Nature Research 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 Research 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 • TCGAblinks R package (v. 2.15.3) to download TCGA data

Data analysis

We developed an R package Scissor in this study. The source code and tutorial of Scissor are freely available at <https://github.com/sunduanchen/Scissor>. In addition, we also used the following tools for analysis:

- oPOSSUM software (v.3.0) for motif analysis.
- RSEM (v1.3.0) for gene expression quantification from bulk RNA-seq.
- The cameraPR function compiled in the limma R package (v.3.42.2) was used to evaluate the enriched molecular pathways.
- R package Seurat (v.3.2.1) for single-cell preprocessing.
- R package GSVA (v.1.34.0) was applied to calculate the signature score.
- R package survival (v.3.2-3) for survival analysis.
- R package viper (v.1.20.0) for calculating the enrichment score of gene sets with both up-regulated and down-regulated genes.
- R package splatter (v.1.10.1) for the scRNA-seq simulation.
- R package scRecover (v.1.2.0) for the imputation of single-cell RNA-seq data
- R package clusterProfiler (v.3.13.0) for the enrichment plot.
- R package Slingshot (v.1.3.2) for the pseudotime analysis.
- R package loomR (v.0.2.0) to analyze .loom file.
- R package Scissor (v.2.1.0) for the analysis presented in this study.

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 Research [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 list of figures that have associated raw data
- A description of any restrictions on data availability

- Lung cancer single-cell sequencing data (E-MTAB-6149, E-MTAB-6653), <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6149/>, <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6653/>
- Lung cancer bulk data (TCGA-LUAD), <https://portal.gdc.cancer.gov/projects/TCGA-LUAD>
- Robustness testing data (GSE30219), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30219>
- Lung cancer signature validation data (GSE13213, GSE11969, GSE31210, GSE8894, GSE3141, and GSE4573), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13213>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11969>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31210>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8894>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3141>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4573>
- Melanoma single-cell sequencing data (GSE72056), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72056>
- Melanoma bulk data (GSE78220), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78220>
- Melanoma bulk data (phs000452.v3.p1), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000452.v3.p1
- Immunotherapy responsive signature validation data (GSE120575, GSE126973, and GSE8678), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120575>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126973>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8678>
- FSHD single-cell sequencing data (GSE122873), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122873>
- FSHD bulk data (GSE140261), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140261>
- FSHD molecular signature validation data (GSE115650, GSE56787, and GSE26852), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115650>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56787>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26852>
- Alzheimer's disease single-cell sequencing data, <http://adsn.ddnetbio.com>
- Alzheimer's disease bulk data (GSE39420), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39420>
- Alzheimer's disease molecular signature validation data (GSE109887, GSE28146), <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109887>, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28146>
- Pathway enrichment analysis database (Molecular Signatures Database, MsigDB), <https://www.gsea-msigdb.org/gsea/msigdb/>
- Transcription factor binding site profiles (JASPAR CORE Profiles), http://opossum.cisreg.ca/cgi-bin/oPOSSUM3/opossum_human_ssa
- Transcription factor regulome database, <https://pubmed.ncbi.nlm.nih.gov/28810145/>

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](https://www.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	We demonstrated the broad utility of Scissor in a total of ten applications, ranging from engineered simulations, cancer cells (guided by survival/tumor stage/mutations), immune cells, and a muscle disease dataset to multiple types of brain cells in Alzheimer's disease. In each application, Scissor proves its ability to identify the biologically and clinically relevant cell subpopulations and is able to reveal the underlying biological mechanisms of investigated disease.
Data exclusions	For Tirosh's melanoma T cell scRNA-seq dataset, we removed the smallest cluster with 174 cells that are characterized by the high expression of cell cycle-related genes and are contaminated with melanoma marker genes. A similar strategy was utilized previously [PMID: 30633907].
Replication	For public datasets used in this study, the replication information, if available, can be found in the original papers cited in this manuscript's reference list. For the data analysis in this study, all attempts at replication were successful.
Randomization	No new data were collected in this study. Samples were allocated into groups as defined in the original papers cited in this manuscript's reference list. When selecting bulk samples to guide single-cell data analysis, we used the dataset from the same disease and tumor type.
Blinding	No new data were collected in this study. All blinding information during the data collection if available, was described in the original papers cited in this manuscript's reference list. The investigators were not blinded to group allocation during data analysis. The Scissor algorithm needs the bulk group labels to identify the phenotype-associated subpopulations. But all the group labels were defined by the original papers.

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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging