scSTAR reveals hidden heterogeneity with a real-virtual cell pair structure across conditions in single-cell RNA sequencing data

Hao, J., Zou, J., Zhang, J., Chen, K., Wu, D., Cao, W., Shang, G., Yang, J. Y. H., Wong-Lin, K., Sun, H., Zhang, Z., Wang, X., Chen, W., & Zou, X. (2023). scSTAR reveals hidden heterogeneity with a real-virtual cell pair structure across conditions in single-cell RNA sequencing data. *Briefings in Bioinformatics*, *24*(2), 1-13. Article bbad062. Advance online publication. https://doi.org/10.1093/bib/bbad062

Link to publication record in Ulster University Research Portal

Published in:
Briefings in Bioinformatics

Publication Status:
Published online: 22/02/2023

DOI:
10.1093/bib/bbad062

Document Version
Author Accepted version

General rights
The copyright and moral rights to the output are retained by the output author(s), unless otherwise stated by the document licence.

Unless otherwise stated, users are permitted to download a copy of the output for personal study or non-commercial research and are permitted to freely distribute the URL of the output. They are not permitted to alter, reproduce, distribute or make any commercial use of the output without obtaining the permission of the author(s).

If the document is licenced under Creative Commons, the rights of users of the documents can be found at https://creativecommons.org/share-your-work/cclicenses/.

Take down policy
The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk

Download date: 09/12/2024

scSTAR reveals hidden heterogeneity with a real-virtual cell pair structure across conditions in single-cell RNA sequencing data

Authors: Jie Hao[1#*], Jiawei Zou[1#], Jiaqiang Zhang[2#], Ke Chen[3], Duojiao Wu[1], Wei Cao[4], Guoguo Shang[5], Jean Y.H. Yang[6], KongFatt Wong-Lin[7], Hourong Sun[8], Zhen Zhang[9], Xiangdong Wang[1], Wantao Chen[9*], Xin Zou[10*]

Affiliations:

[1] Institute of Clinical Science, Zhongshan Hospital, Fudan University, Shanghai, China

[2] Department of Anesthesiology and Perioperative Medicine, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, Zhengzhou, Henan, 450003, China.

[3] Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Shanghai, 201602, China

[4] Department of Oral Maxillofacial-Head and Neck Oncology, Ninth People's Hospital, Shanghai Key Laboratory of Stomatology & Shanghai Research Institute of Stomatology, National Clinical Research Center of Stomatology, Shanghai Jiao Tong University School of Medicine, Shanghai, 200011, China

[5] Department of Pathology of Zhongshan Hospital, Fudan University, Shanghai, China.

[6] School of Mathematics and Statistics and Charles Perkins Center, The University of Sydney, Australia

[7]Intelligent Systems Research Centre, Ulster University, Magee Campus, Derry~Londonderry, Northern Ireland, UK

[8]Department of Cardiac Surgery, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan City, Shandong, 250012, China

[9]Ninth People's Hospital, Shanghai Key Laboratory of Stomatology & Shanghai Research Institute of Stomatology, National Clinical Research Center of Stomatology, Shanghai Jiao Tong University School of Medicine, Shanghai, 200011, China.

[10]Jinshan Hospital Center for Tumor Diagnosis & Therapy, Jinshan Hospital, Fudan University, Shanghai, 201508, China

* To whom correspondence should be addressed, Xin Zou, Email: xzou@fudan.edu.cn; Jie Hao, Email: jhao@fudan.edu.cn; Wantao Chen, Email: chenwantao196323@sjtu.edu.cn

[#]These authors contributed equally to this work.

Abstract

Cell-state transition can reveal additional information from single-cell RNA-sequencing data in time-resolved biological phenomena. However, most of the current methods are based on the time derivative of the gene expression state, which restricts them to the short-term evolution of cell states. Here, we present scSTAR, which overcomes this limitation by constructing a paired-cell projection between biological conditions with an arbitrary time span by maximizing the covariance between two feature spaces using partial least square and minimum squared error methods. In mouse aging data, the response to stress in CD4+ memory T cell subtypes was found to be associated with aging. A novel Treg subtype characterized by mTORC activation was identified to be associated with antitumor immune suppression, which was confirmed by immunofluorescence microscopy and survival analysis in 11 cancers from TCGA. On melanoma data, scSTAR improved immunotherapy-response prediction accuracy from 0.8 to 0.96.

Keywords: cell state dynamics; partial least square regression; scRNA-seq

Abbreviations

TCGA：The Cancer Genome Atlas Program

Treg: Regulatory T cells

$T_{CM:}$ central memory T cells

T_{EM}: effector memory T cells

FC: fold change

scRNA-seq: Single-cell RNA sequencing

PLS: partial least square

AUC: areas under the receiver operating characteristic curve

ARI: adjusted Rand index

UMAP: uniform manifold approximation and projection

GO: gene oncology

LUAD: Lung adenocarcinoma

HCC: Hepatocellular carcinoma

TME: tumor microenvironment

NSCLC: non-small cell lung cancer

KNN: k-nearest neighbors

DE: differential expression

IHC: immunohistochemistry screening

Anti-PD1: immunotherapy with an anti-PD-1 monoclonal antibody

Introduction

Single-cell RNA sequencing (scRNA-seq) data offer insights into cell-to-cell biological variation at the transcriptomic level. It captures the snapshot heterogeneity of cell states in biological processes, but the cells between biological conditions are unpaired due to their destructive nature.

Despite promising progress in understanding immune cell dynamics associated with biological processes, tumorigenesis and treatment responses using scRNA-seq data, the fundamental mechanisms underlying immune cell heterogeneities remain poorly understood [1-4]. T cells have been investigated to probe aging-associated gene expression variations [5, 6]. Aging-related gene expression levels may not be conserved in cell-type-specific regulatory

programs [5], implying that more detailed cell dynamics remain to be uncovered. Regulatory T (Treg) cells can impair antitumor immune activities, but their contribution to tumor development remains unclear [7, 8]. Autoimmunity elicited by Treg-targeting immunotherapy [9] implied the possible existence of unknown substructures of Tregs. Anti-PD1 immunotherapy targeting $CD8^+$ T cells has achieved promising clinical outcomes, but only a small portion of patients can benefit from it [10, 11]. The complexity of $CD8^+$ T cells in the tumor immune microenvironment and their dynamic transitions in relation to therapy outcomes are still inadequately understood. Thus, it hindered the improvement of single-cell data interpretation in revealing cell state transition-related heterogeneities.

Analysis of the cell clusters based on their static transcriptional levels may miss the heterogeneity in subcluster cell-state dynamics. Various analytical methods have been developed to tackle this problem. The recent advance is the development of RNA velocity [12] and its extension methods, e.g., CellRank [13] and dynamo [14], which predict the future state of individual cells in developmental lineage and cellular dynamics on a timescale of hours based on the ratio between nascent and mature mRNA. Such methods have intrinsic limitations with scenarios where research interests are over larger periods, e.g., pre- and post-treatments[15], adjacent normal and tumor [4], and across ages [5], and even expression/nonexpression of certain gene(s). Furthermore, quantifying velocity may require additional information, such as metabolic labeling [14]. In more general scenarios, unwanted variations are removed before downstream analysis to reveal dynamic signals from scRNA-seq data, such as removing batch effects through factor analysis [16-18], reducing random noise via filtering methods [19-24], and eliminating specified biological noise or confounding factors by latent variable models [25], mixture model frameworks [20, 26, 27], and regression-based normalization [28]. The data imputation algorithm is one of the most widely used approaches to reduce unwanted noise [29-

31]. While various techniques have been developed to remove one or more unwanted variations, extracting the biological signals of interest may be a straightforward alternative. However, it remains a challenge to extract biological signals that are comparable or weaker in scale with unwanted variations. In bulk studies, this issue is normally addressed by comparing paired samples from the same individual [32, 33]. Nevertheless, such a strategy cannot be easily extended to single-cell experiments, as one cell can only be profiled once in either condition; hence, cells are not paired. As a consequence, cell dynamics cannot be obtained by directly comparing paired cells between conditions.

Here, we developed scSTAR (single-cell State Transition Across-samples of Rna-seq data), which uses a supervised machine learning algorithm, partial least square (PLS), to profile the differences between two biological feature spaces by maximizing the covariance of the global characteristics. Based on the PLS model, each real cell in one condition can be virtually projected to the counterpart space. The differences between each real-virtual cell pair represent the state transition of the specific cell. Systematic benchmarks and case studies demonstrate that scSTAR is accurate and robust for revealing heterogeneities in scRNA-seq data. It captures quantitative cell state transitions and identifies potential cell subtypes, biomarkers and biological processes related to aging, tumorigenesis, treatment response, etc. Activation of the response to stress and DNA repair were found in $CD4^+$ memory T cells from aged mice in two species. An enriched mTORC1 signature is a sign of a poor prognosis-related Treg cell subtype across cancers, and its existence was validated experimentally. Reclustering on scSTAR captured dynamic features of melanoma patient $CD8^+$ T cells and improved immunotherapy response prediction accuracy from 0.8 to 0.96.

Results

scSTAR *in silico* estimates individual cell state transitions by generating real-virtual cell pairs across samples/conditions, and the expression differences between the real-virtual cell pair should only relate to the biological problem of interest (Figure 1a). More details of the implementation can be found in Methods. In a typical single-cell analytical workflow, scSTAR can be applied just before "clustering/trajectory", etc. step [17, 34] (Figure 1b) to reveal detailed subcluster structures of the cell-state dynamics pattern in response to experimental condition changes, even when they are buried by other interference.

Systematic benchmarking demonstrates the superior performance of scSTAR

We first benchmarked scSTAR against the original (unprocessed) data and five existing noise reduction methods, i.e., Combat [27], MAGIC [29], SAVER [30], MNN [16] and scMerge [18], and evaluated it through three clustering methods, k-means, SC3 [35] and Seurat [17]. The noisy simulation datasets mimicked 12 case-control scenarios with different combinations of parameter settings: 1) the expression level fold changes (FCs) between case and control data: 1.3, 1.5 and 2; 2) the intragroup heterogeneity in the case group: 2, 3 and 4 subclusters (2 subclusters with cell ratio between subclusters: 1:2, and 1:1; 3 and 4 subclusters with equal proportions of cells between subclusters) (Methods). For each scenario, 10 datasets (2000 cells per dataset) were randomly generated, which led to 120 simulated datasets in total. The aim of this evaluation was to compare the capabilities in revealing intragroup heterogeneity in the presence of noise interference when using the aforementioned 21 processing procedures.

First, the areas under receiver operating characteristic curves (AUCs) were calculated for the data containing 2 subclusters to quantitatively illustrate how well different subclusters could be

separated (Figure 2a). In terms of AUC, scSTAR and SAVER have values ranging from approximately 0.9 to 1, whereas most of the rest were between 0.5 and 0.8. scSTAR showed slightly better results than SAVER. At low FCs (i.e., low signal amplitude differences), intragroup heterogeneities tended to be masked by strong noise interferences, and the scSTAR algorithm could dramatically reduce such interference without causing obvious distortions to the data. Furthermore, the clustering results by k-means, SC3 and Seurat were evaluated in terms of the adjusted Rand index (ARI) [36] (Methods). ARI is often used in cluster validation to measure the agreement between two partitions: one given by the clustering process and the other defined by external criteria. The closer its value is to 1, the better the clustering performance. With a relatively weak signal (FC=1.3), scSTAR outperformed original, Combat, MAGIC, SAVER and scMerge in all cases (Figure 2b-e). Although MNN achieved comparable results with scSTAR in some cases of 2 subclusters, the performance of MNN declined dramatically when the data complexity increased, e.g., with 3 and 4 subclusters. A similar trend was also observed for FCs of 1.5 and 2 (FC=1.5, Figure S1a-d, FC=2, Figure S1e-h). In addition, we illustrate the distribution of cells on averaged DE gene expression after being processed by each method and clustered by k-means (Figure 2f-l) in the 2 subcluster case. The heterogeneities of cells were faithfully revealed by scSTAR with very few misclassified cells (Figure 2f), whereas there were many more misclassified cells with other methods. The results demonstrate that cell heterogeneities that can be revealed by scSTAR are hardly observed by other existing methods. Finally, a visualization of cell topology structures using Uniform Manifold Approximation and Projection (UMAP) demonstrated that strong noise interference can blur the cell subcluster patterns with procedures using existing methods. In contrast, clear separation of cell heterogeneity was revealed with scSTAR processed data (Figure 2m-2s). In summary, scSTAR is the only method that can consistently achieve reliable results across various metrics. This

illustrates that scSTAR has a stronger capability to reduce noise interference than existing methods.

scSTAR reveals aging-related cell subtypes in mouse immunosenescence data

Next, we used scSTAR to uncover aging-related CD4+ T cell transcriptional dynamics in mouse immunosenescence data [5], and the observed aging patterns were verified in different mouse species. The cells were collected from two inbred mouse subspecies separated by 1 million years of divergence, Mus musculus domesticus (B6) and Mus musculus castaneus (CAST), and from both old and young mice in naïve and active conditions. The original conclusions were that no global expression profile change was found due to aging in either naïve or activated CD4+ T cells; only ~10% of genes were differentially expressed between cells from young and old mice in a small subset of cells, and these were not conserved between B6 and CAST. The reconstructed plot confirmed their conclusion (Figure S2a and b). When reanalyzing the same CD4+ T cells using scSTAR, much clearer separation patterns between young and old mice were revealed (Figure S2c and d). For the naïve B6 data (young vs old), 5 clusters were obtained (Figure 3a). The cluster with cells expressing Ccr7 and Sell suggests that they are central memory T cells (T_{CM}) [37]; those expressing Cxcr3 and Cd44 but lacking Ccr7 suggest that they are effector memory T (T_{EM}) cells [37, 38], and those expressing Tigit suggest that they are exhaustion T cells (T_{Exh}) [39]. The last two clusters were annotated as unidentified $Fem1c^{+}$ $CD4^{+}$ T cells (T_{Fem1c+}) and intermediate-stage cells (T_{int}) (Figure 3b). The T_{EM} cluster was mainly found in young mice, the T_{Exh} and T_{Fem1c+} clusters were mainly found in old mice, and the T_{CM} cluster existed in both age groups (Figure 3c). Taking Figure 3b and 3c together, scSTAR revealed a shift from effective to exhausted functions during aging.

GO analysis showed that enriched lymphocyte activation and catabolic functions were found in young mouse T_{EM} cells, and negative regulation of catabolic functions, apoptotic processes, and responses to stress were found in old mouse T_{Exh} and T_{Fem1c+} clusters (Figure 3d). Trajectory analysis illustrated that during the course of aging, T_{EM} tended to be replaced by T_{Exh} and T_{Fem1c+}. T_{int} represents an intermediate state on the transition trajectory. However, the aging-irrelevant T_{CM} was separated from this transition branch (Figure 3e and 3f). The heatmap of the marker genes identified for each of the cell clusters using the Seurat FindAllMarkers function clearly revealed the aging-induced cell dynamic pattern (Figure 3g), although such a pattern was observable but weaker in the original data (Figure 3h). Furthermore, these dynamic patterns were also confirmed in CAST. For example, within the top 2000 most variable genes of naïve B6 cells, 1034 genes passed the gene filtering processing of naïve CAST mice. It can be seen that young and old CAST cells can be well separated by these genes with scSTAR processed data (Figure 3i) but not with the original CAST expression data (Figure 3j). For the activated B6 data (young vs old), aging-related clustering patterns were also observable, with two distinct old mouse clusters, T_{Exh} and T_{reg}, and one young mouse cluster, T_{div} (in the cell division cycle, as indicated by high Cdc23) (Figure S3). Interestingly, the negative regulation of catabolic functions observed in the naïve old mice turned positive in the activated old mice.

scSTAR identifies gene expression-specific dynamics during aging

To demonstrate the capability of scSTAR in revealing gene-specific dynamics during biological processes in the presence of strong interference, we exemplify the T_{div} cells from activated B6 mice in the previous section. First, cells with positive/negative expression of Cdc23 were used as criteria for case/control in scSTAR. Unsupervised clustering identified 3 clusters in the scSTAR processed (Figure 4a) and original data (Figure 4b). For the scSTAR processed data, cluster

scS_1 was associated with young mice via $Cdc23^+$ cells, and scS_0 and scS_2 were associated with old mice via $Cdc23^-$ cells by the hypergeometric test (Figure 4c). Consistently, a similar but weaker association was also observed in the original data (Figure 4d). Aging can cause DNA damage in many aspects of dysfunction and disease [40], and Cdc23 affects the response to DNA damage [41]. The GO analysis showed molecular functions involving DNA repair, DNA metabolic processing and response to stress. enriched in scS_1 were uniquely observed in scSTAR-processed $Cdc23^+$ cells from young mice. However, the old mouse-associated scS_2 was mainly enriched in cell cycle activities (Figure 4e). This observation implied that Cdc23 may play a role in antiaging activities. However, immune activity-related functions were dominant in the enriched GO terms obtained using the original data (Figure 4f). Due to the stimulation process in the experimental settings, the gene expression variations in activated immune cells were dominated by immune function activities. The results illustrated that scSTAR can reveal subtle information of interest (aging-related gene expression variations) in the presence of strong biological interferences (immune response activities).

scSTAR uncovers a new active effector Treg (eTreg) cell subtype in pan cancers

To reveal detailed tumorigenesis-related cell heterogeneity, we analyzed scRNA-seq datasets from two cancer types, LUAD [1] and HCC [2], and found a protumorigenic eTreg subtype associated with poor patient prognosis.

The 739 Treg cells from adjacent normal and tumor tissues in the original LUAD study were annotated as nTreg (naïve Treg, high expression of SELL, low expression of FOXP3), eTreg (high expression of FOXP3) and $CD4^+$ Th (no FOXP3 expression) based on original expression data (Figure 5a, 5b, and 5m). Then, scSTAR was applied to treat normal tissue as a control and tumor tissue as a case group, and 5 clusters, C1-C5, were achieved (Figure 5c). Pseudotime

trajectory analysis [34] illustrated that clusters C1, C3 and C4 were located at the end of branches (Figure S4a, S4b). Furthermore, the DE genes upregulated in C1 and C3 (identified by 'FindAllMarkers', Table S1) were significantly associated with low probabilities of survival, which suggests that the activities of C1 and C3 might be protumorigenic; therefore, they represented active Treg subtypes in the tumor microenvironment (TME) (Figure 5i, 5j, S4c and S4d). Similar processing was also applied to 1959 Treg cells from HCC samples. Using scSTAR in combination with clustering, trajectory reconstruction and survival analyses, we identified a similar subset of eTreg cells, HCC C4, as protumorigenic (Figure 5e-h, 5k, 5n, S5, Table S2).

It was found that protumorigenic eTreg cells from both LUAD and HCC tended to have higher expression levels of HSPA5 and HSP90B1 from the mTORC1 pathway. The expression of the above markers was highly correlated with the recently reported eTreg active markers ICOS and IL1R1 [42] (Figure 5m and 5n). $HSPA5^{+}$/$HSP90B1^{+}$ eTreg cells highly expressed a set of 400 genes common to both tumor types (rank sum test false discovery rate <0.05, fold change >0, Table S3). They included the genes associated with Treg immunosuppressive functions, such as LAYN [43], REL[44], TNFRSF9 [1], ICOS and IL1R1 [42], reported from various cancers, which further supported that the identified eTreg cell subtype (referred to as $HSPA5^{+}$ eTregs) was responsible for immunosuppressive functions in the TME.

To further justify the above speculation, the top 10 GSEA hallmark gene sets enriched in the 400 genes were identified (Figure 5l, blue dots). Survival analysis of the 21 cancer TCGA datasets [45] using the identified genes in each of the top 10 gene sets was performed, and the signatures enriched in the mTORC1 gene set were highly predictive of worse patient prognosis in 11 out of 21 cancer types ($p<0.05$, Cox regression, Figure 5l read dots, Figure S6). In addition, the signatures were evaluated on a melanoma dataset [15], where patients with high expression of identified genes in mTROC1 tended to be nonresponsive to immunotherapy ($p<10^{-10}$, rank

sum test, Figure 5l bars). Immunofluorescence microscopy of human kidney and esophageal cancer samples also validated the existence of HSPA5$^+$ eTreg cells in tumor tissues (Figure 5m and S7, Methods).

The existence of Hspa5$^+$ eTregs was also found in mouse lung cancer development model scRNA-seq data [8] (Figure S8 a-c). In addition, survival analysis of lung adenocarcinoma TCGA data illustrated that the proposed HSP90B1 and HSPA5 markers can discriminate the immunosuppression Treg subtype well (Figure S8 d-e). In contrast, the Treg subtype marker TNFRSF9 discovered using conventional methods [1] failed to do so (Figure S8c). Further analysis showed that the identified marker gene set had 30 overlapping genes (Table S4) with a previously reported curated gene set of poor progression in non-small cell lung cancer (NSCLC) [46], which also confirmed the protumorigenic functions of Hspa5$^+$ eTregs.

scSTAR improves the accuracy of the immunotherapy response predictive model

To demonstrate that accurate cell subtype clustering can improve immunotherapy response prediction, we applied scSTAR to melanoma data [15] to construct an immunotherapy response prediction model using cell type composition patterns. In the original study, 5410 CD8$^+$ T cells from both pre- and posttreatment specimens were categorized into CD8_B and CD8_G subtypes, and the ratio between the two subtypes was predictive of immunotherapy response patterns. These CD8$^+$ T cells were reprocessed by scSTAR with pretreatment specimens as the control and posttreatment specimens as the case group.

With the optimized clustering parameters determined (Figure 6a, Methods), scSTAR-processed pretreatment CD8$^+$ T cells were categorized into 6 clusters (Figure 6b, Table S5), which were associated with immunotherapy-response patterns using a hypergeometric test

(Figure 6c). The prediction score was obtained as the ratio between the numbers of cells in clusters (C4 + C6) and (C1 + C2 + C5). A significant difference was found between the nonresponders and responders ($p = 0.0004$, Figure 6d), and the associated AUC was 0.96 (Figure 6f). For comparison, responder/nonresponder prediction using CD8_B and CD8_G cells from pretreatment was calculated, which showed $p = 0.03$ (Figure 6e) and AUC 0.8 (Figure 6f).

Discussion

We have shown scSTAR's ability to estimate the state transition for each individual cell during aging, gene-specific expression progression, tumor progression, immunotherapy response, etc. Apart from what the original studies have found, more biological insights into cell heterogeneity, such as novel cell subtypes, new discriminatory patterns, and clearer progression trajectories in response to biological processes of interest, were identified. scSTAR generalizes beyond the time scale of cell development and successfully reveals how aging affects the transcriptional dynamics of CD4+ T cell subtypes from two divergent mouse species. The effect of aging was comprehensively inspected from two diverse perspectives, i.e., age category and marker gene expression level. A clear and consistent observation was that aging was characterized by deactivation of DNA repair and response to stress. We validated the existence of a new activation state of the eTreg subtype HSPA5^{+} eTregs from human LUAD, HCC, and mouse lung cancer scRNA-seq datasets and experimentally in human kidney and esophageal cancer samples. Its protumorigenic character was verified in 11 tumors in the TCGA database. On a melanoma immunotherapy dataset, we showed that scSTAR revealed cell substructure heterogeneities associated with immunotherapy response patterns, which could be applied to predict patient-specific therapy outcomes. We envisaged that scSTAR-based cell heterogeneity discovery can benefit many more biological or clinical scenarios.

scSTAR is robust, as indicated by the AUC and ARI metrics, and scSTAR is the only method that can consistently achieve reliable results across various conditions. With high interference datasets (multiple clusters and low FCs), scSTAR is useful to recover the true clustering patterns that may be masked by various interferences. This illustrates that scSTAR has a stronger capability to reduce noise interference compared to existing methods. As a result, more informative cell heterogeneities can be discovered from scSTAR processed data. PLS, as a supervised method, endows the proposed scSTAR algorithm with sensitivity only to the experimental variations investigated, regardless of the amplitudes of these variations.

A limitation with scSTAR's imbedded PLS is that it is more suitable for an overdetermined system in finding subcluster cell heterogeneity locally. However, in the scenario of an underdetermined system when abundant cells are considered, the results of scSTAR might be compromised.

scSTAR brings a notable advantage of quantitative cell-state dynamics of arbitrary time spans without any extra information. It can easily fit into the current common analysis workflow; thus, many previously published scRNA-seq datasets can be reanalyzed to reveal more detailed subcluster structures. The multiangle perspective usages of scSTAR demonstrated in this work bring a number of innovations in *in silico* examining cell state- or gene expression-specific dynamics in relation to biological condition(s) using scRNA-seq data. Of note, scSTAR is not limited to multiple condition comparison experiments. The scSTAR gene expression dynamics estimation could also be applied intraconditionally. As a general framework for estimating virtual cells in a specific feature space, we anticipate that scSTAR will be useful to reveal insightful cell heterogeneities through the virtual cell pair structure, where analyzing the static state is often challenging.

Methods

Concept of scSTAR

scSTAR is designed to extract the cell state transition dynamic heterogeneities between conditions. In an ideal two-group comparison scenario, let us denote X as the cell states from the control group and Y as the cell states from the case group. X represents the baseline, and Y can be decomposed as:

$$Y = \hat{X} + V \quad (1)$$

where $\hat{X}$ is the projection of Y in the control feature space and is not directly observable from the data; V represents the state transition matrix from $\hat{X}$ to Y, which should be mainly caused by the experimental changes. Both $\hat{X}$ and V have the same dimension as Y. Characterizing $\hat{X}$ by principal component analysis (PCA), $\hat{X} = P^T S$, the loading matrix P can be obtained by

$$P = \underset{P^T P=1}{\operatorname{argmax}}\, Cov(\hat{X}P, \hat{X}P) \quad (2)$$

Since $S = (P^T P)^{-1} P\hat{X}$, we have $\hat{X} = P^T (P^T P)^{-1} P(Y - V)$. As V and $\hat{X}$ are unrelated, $P^T (P^T P)^{-1} PV \approx 0$, $\hat{X}$ can be approximated as $P^T (P^T P)^{-1} PY$, and V can be calculated as

$$V = Y - P^T (P^T P)^{-1} PY \quad (3)$$

Eq (3) indicates that given Y, V can be estimated by P.

Assuming $\hat{X}$ and X are different cells from the same feature space, we use X to replace the first $\hat{X}$ in Eq. (2), and C to replace P to represent the possible mismatch between $\hat{X}$ and X, Eq (2) can be approximated by

$$\underset{P^T P=1}{\operatorname{argmax}}\, Cov(\hat{X}P, \hat{X}P) \Leftrightarrow \underset{C^T C=1, P^T P=1}{\operatorname{argmax}}\, Cov(XC, \hat{X}P) \quad (4)$$

Considering that X and V are unrelated and $Cov(XC, VP) \approx 0$ for arbitrary P and C, the straightforward manipulation of Eq (4) provides

$$P = \underset{C^T C=1, P^T P=1}{\operatorname{argmax}}\, (Cov(XC, \hat{X}P) + Cov(XC, VP))$$

$$= \underset{C^T C=1, P^T P=1}{\operatorname{argmax}}\, Cov(XC, YP) \quad (5)$$

The solution of Eq (5) can be achieved by partial least squares (PLS), and C and P denote the PLS loading matrices of X and Y, respectively. Furthermore, the cost function can be expressed as

$\max_{P^TP=1,\ C^TC=1} \sqrt{Var(XC)Var(YP)}Corr(XC, YP)$. The estimation of P also accounts for the maximization of the correlation between the two groups of cells, which may account for the mismatch between $\hat{X}$ and X. Hence, cell state dynamics can be achieved by maximizing the covariance between cell states from various conditions (Figure 1a).

A three-step scSTAR procedure

In more realistic scenarios, variation V includes not only signals of interest but also noise. During a dynamic process, the variation V contained in each cell can be considered a linear combination of the following components [20, 25]:

$$V = V_{batch} + V_{noise}^{r+b} + V_{signal} \quad (6)$$

where V_{batch} indicates the batch effect, and V_{noise}^{r+b} consists of random (including technical) (r) and biological (b) noise and indicates the interferences unrelated to the discrimination between the two groups. V_{signal} represents the gene expression differences between the paired conditions studied.

The proposed scSTAR algorithm is designed to extract V_{signal} by dissecting different components in V in separate steps: first, removing V_{batch}, then extracting V_{signal}. As V_{batch} may have some statistical similarity with V_{signal}, e.g., both are associated with group discrimination, V_{batch} is identified and removed first using the method described in the following section: Step 1. Then, as noise is usually not correlated with group discrimination, V_{signal} can be extracted using the method presented in the following section: Step 2.

Similar to the normal scRNA-seq data analysis procedures, the scRNA-seq data were preprocessed with a gene filtering step. Here, we used the OGFSC [47] R package, where genes with variances smaller than the noise threshold curve defined by parameter α were removed. The default value of α is set to 0.5 to preserve the signal integrity to the maximum extent.

Step 1: Removal of V_{batch}

Let us define 'anchor' as the cells that can be paired between the two groups. A reasonable assumption is that the differences between a pair of anchor cells are only caused by the batch effect [16]. Here, the k-nearest neighbors (KNN) [48] method is used to identify the paired anchor cells, which should be mutually within the k nearest neighbors. By default, k is set to 3, and the similarity between cells is evaluated in terms of cosine metrics. Next, a first PLS model (PLS1) is constructed only on the anchor cells from both groups to characterize the batch effect. The component of the batch effect V_{batch} is then removed using the PLS model and the minimum square error method as follows:

$$V_{batch} = P_{PLS1}\, pinv(P_{PLS1})Y \tag{7}$$

$$V' = Y - V_{batch} \tag{8}$$

where Y is the observed data vector of a cell. P is the PLS loading matrix of Y containing m PLS components. The value of m can be either manually defined or estimated by maximizing the goodness-of-prediction Q^2 calculated by the 10-fold cross-validation method [49]. $pinv(P) = (P^T P)^{-1} P^T$ denotes the Moore-Penrose pseudoinverse derived by the minimum square error criterion. As P is constructed from the anchor cells, the term $P_{PLS1}\, pinv(P_{PLS1})$ only contains the variation components related to the batch effect. Therefore, the residual variations contained in V' are dominated by V_{signal} and V_{noise}. The batch effect can be removed from the cells of both groups using Eqs. 7 and 8.

Step 2: Extraction of cell state transfer V_{signal}

A second PLS model (PLS2) was constructed using all cells from both groups. As noise variations V_{noise} do not contribute to the discrimination of the two groups, PLS2 dedicatedly captures the variation of V_{signal}. Using the loading matrix P_{PLS2} to estimate the virtual cell profiles, the signal can be extracted from V':

$$\hat{V}_{signal} = P_{PLS2} * pinv(P_{PLS2}) * V' \tag{9}$$

where $\hat{V}_{signal}$ is the estimation of V_{signal}. As a result, irrelevant noise shrinks towards zero in $\hat{V}_{signal}$, while the expression values of DE genes are retained.

All variation components irrelevant to the signal of interest are excluded from $\hat{V}_{signal}$, and the remaining amplitudes represent the differential expression (DE) between two conditions. Now, the heterogeneities between cells indicate the diverse dynamic patterns of those cells when conditions change.

Step 3: DE gene identification

Based on the obtained $\hat{V}_{signal}$, we designed a procedure to extract such DE gene heterogeneity. First, all cells from both groups were clustered using the Seurat R package based on their $\hat{V}_{signal}$. Then, the 'FindAllMarkers' function was applied to identify the genes specifically highly expressed (FDR<0.05) in each cluster. As non-DE genes in $\hat{V}_{signal}$ tend to shrink to zero, the highly expressed genes identified by 'FindAllMarkers' should be upregulated DE genes contained in each cell cluster. A hypergeometric test is then applied to associate each cell cluster with group information. We defined the highly expressed genes of one cluster as upregulated in a group if the cells from the group dominated the cluster, as indicated by hypergeometric test p<0.05.

On some occasions, a cell cluster may fail to be associated with any group, which implies that this cluster of cells tends to be stable during condition changes.

Simulation datasets

To evaluate the proposed algorithm, we used the protocol presented in Haghverdi et al.[16] to create simulated datasets and mimic case-control studies, which contain batch effects and random and biological factors. In the simulated dataset, there were 1000 cells in each group, and each cell had 100 genes. The control group contained batch and random factors only, and various numbers of subclusters with different proportions were simulated in the case group by manipulating the fold changes (FCs) of randomly selected differentially expressed (DE) genes. FC was positively correlated with the strength of the signal. Each subcluster contained 25 DE genes, and each simulated dataset was generated 30 times. The number of PLS components in scSTAR was estimated automatically.

The simulated data contained batch effects, random noise and biological signals (with diverse patterns) at FC = 1.3, 1.5, and 2 to mimic case-control studies. In the case group, 2 to 4 subclusters were generated to represent cell heterogeneity. For the 2-subcluster datasets, two types of subcluster proportions are generated, i.e., 1:2 and 1:1. For the 3- and 4-subcluster datasets, the proportions of subclusters are equal. The batch effect and random noise are generated using the default parameters presented in [16]. The ARI (see **Evaluation**

metrics for definition), between 0 and 1, indicates the consistency between the true cell subtype annotation and the clustering results.

Evaluation metrics

The results obtained by different clustering methods on the simulated data were evaluated by the adjusted Rand index (ARI)

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{11}$$

where n_{ij} are contingency table entry values and a_i and b_j are the sums of the i^{th} row and the j^{th} column of the contingency table, respectively. The closer the ARI value is to 1, the closer it is to the true cluster.

Experimental Datasets

To illustrate the capability of scSTAR in discovering novel cell subtypes, three datasets were adopted: a human hepatocellular cell carcinoma tissue dataset (GSE140228), a human lung squamous cell carcinoma tissue dataset (GSE99254) and a mouse lung cancer model dataset (GSE129914). Validations of scSTAR by revealing new discriminatory patterns, clearer progression trajectories, etc. were adopted from a mouse immunosenescence study dataset (E-MTAB-4888). The potential clinical application of scSTAR was demonstrated on a melanoma immunotherapy treatment dataset (GSE120575).

Multicolor IHC for HSPA5^{+} eTreg cell validation

The clinical specimens obtained with IHC in this study were collected with informed consent for research use and were approved by the Medical Ethics Committee of Henan Provincial People's Hospital (2019(44)) according to the Declaration of Helsinki. The samples consisted of a patient with esophageal squamous cell carcinoma and a patient with kidney cancer.

Human tissue specimens were collected within 30 minutes after tumor resection and fixed in formalin for 48 h, followed by dehydration and embedding. The paraffin tissue was cut into 4 μm sections and fixed on glass slides. The slide was placed in a 40°C oven for 30 minutes to dry the fixed tissue. The sample was deparaffinized in xylene 3 times for 10 minutes each and then successively rehydrated in 100%, 95%, and 70% alcohol for 3 minutes each. Antigen was recovered by immersion in boiling EDTA buffer (pH 9.0) for 15 minutes. Then, the slide was blocked with Antibody Diluent/Block to avoid nonspecific sites for 10 minutes and incubated with primary antibodies in a humidified chamber for 1 h at room temperature (RT). The sections were washed with TBST 3 times for 2 minutes each and incubated with HRP-conjugated secondary antibody for 10 min at RT. Next, the sections were washed with TBST for 2 minutes 3 times and incubated with an Opal Multi-Color IHC Kit to amplify the signal. The images were captured, and analysis was conducted with Phenochart. The primary antibodies included FOX3P (Abcam, 1:200) (Opal 690), CD4 (CST, 1:200) (Opal520) and HSPA5 (CST, 1:200) (Opal570).

Construction of the immunotherapy response predictive model

The dataset contained 16291 $CD45^+$ cells collected from 48 melanoma specimens, either at baseline and/or during treatment (anti-CTLA4 and/or PD-1) [15]. The 48 patients were classified into 31 nonresponders and 17 responders. In the original study, $CD8^+$ T cells were categorized into B and G subtypes, and the ratio of the two subtypes was shown to be predictive of immunotherapy response patterns. In this study, 5410 cells, which were provided in the published dataset (GSE120575) and annotated as $CD8^+$ T cells in Table S2 of the original paper, were reanalyzed. The $CD8^+$ T cells were from 18 patients, including 7 responders and 11 nonresponders. There were 3067 cells collected from pretreatment specimens and 2343 from posttreatment specimens. The dataset was obtained using the full-length SMART-seq2 protocol.

scSTAR was first applied to profile the cell dynamics between pre- and post-treatment $CD8^+$ T cells. Then, based on the cell dynamic properties, i.e., scSTAR processed data, pre-treatment $CD8^+$ T cells were categorized into a few clusters, each of which can be associated with responders or nonresponders (response

pattern). A prediction score was calculated for each patient as the ratio of the numbers of cells associated with nonresponders over responders. The following steps were performed:

1) OGFSC was applied to perform gene filtering with the number of bins set to 30.
2) scSTAR was applied to pre- and post-treatment $CD8^+$ T cells with a predefined *m* number of PLS components. The processed pre-treatment $CD8^+$ T cells were then categorized into a predefined *K* number of cell clusters using k-means. The optimization of *m* and *K* is described in the following step 5).
3) A hypergeometric test was applied to associate each cell cluster with responders or nonresponders ($p<0.05$).
4) For each patient, a prediction score was calculated as the ratio of the cells associated with nonresponders over responders. An AUC can be calculated based on the prediction scores. By evaluating the obtained AUC on a null hypothesis model obtained by randomly shuffling the responsive labels of all patients 100 times, a p-value was obtained.
5) Different combinations of parameters m and K were tested using steps 2)-4), and their optimal values were chosen to achieve the maximum AUC.

Declarations

Acknowledgements

We thank Professor Léon Otten for critical suggestions.

Funding

This work was supported in part by the National Natural Science Foundation of China [82170045 to JH, 31800253 to KC, 81771672 to DJW, 81672745 to WC]; the Special Fund for Scientific Research of Shanghai Landscaping & City Appearance Administrative Bureau [G222410 to KC, JH and XZ]; the Translational Medicine Cross Research Fund of Shanghai Jiao

Tong University [ZH2018QNB29 to JH]; the Natural Science Foundation of Shanghai [16ZR1417900 to XZ]; the Shanghai Pujiang program [16PJ1405200 to XZ, 16PJ1405100 to JH]; and the Shanghai Sailing Program [17YF1410400 to KC]. The Innovative Research Team of High-level Local Universities in Shanghai [SHSMU-ZLCX20212301 to JH, WTC].

Author Contributions

Conceptualization: JH, XZ, KC, DJW, WTC; Methodology: WC, KFW-L, HRS, XDW, JY, WTC, JQZ, GGS; Investigation: JH, XZ, WTC, KC, WC, DJW; Visualization: JH, XZ, KC, JWZ; Project administration: JH, XZ; Supervision: JH, XZ, DJW, WTC, KC; Writing – original draft: JH, XZ, KC

Competing interests

None

Code availability

The scSTAR R package is available at https://github.com/XZouProjects/scSTAR.

References

1. Guo, X., et al., *Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing.* Nat Med, 2018. **24**(7): p. 978-985.
2. Zhang, Q., et al., *Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma.* Cell, 2019. **179**(4): p. 829-845 e20.
3. Azizi, E., et al., *Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment.* Cell, 2018. **174**(5): p. 1293-1308 e36.
4. Zheng, Y., et al., *Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment.* Nat Commun, 2020. **11**(1): p. 6268.
5. Martinez-Jimenez, C.P., et al., *Aging increases cell-to-cell transcriptional variability upon immune stimulation.* Science, 2017. **355**(6332): p. 1433-1436.

6. Luo, O.J., et al., *Multidimensional single-cell analysis of human peripheral blood reveals characteristic features of the immune system landscape in aging and frailty.* Nature Aging, 2022. **2**(4): p. 348-364.
7. Togashi, Y., K. Shitara, and H. Nishikawa, *Regulatory T cells in cancer immunosuppression - implications for anticancer therapy.* Nat Rev Clin Oncol, 2019. **16**(6): p. 356-371.
8. Li, A., et al., *IL-33 Signaling Alters Regulatory T Cell Diversity in Support of Tumor Development.* Cell Rep, 2019. **29**(10): p. 2998-3008 e8.
9. Tanaka, A. and S. Sakaguchi, *Targeting Treg cells in cancer immunotherapy.* Eur J Immunol, 2019. **49**(8): p. 1140-1146.
10. Reck, M., et al., *Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater.* J Clin Oncol, 2019. **37**(7): p. 537-546.
11. Vokes, E.E., et al., *Nivolumab versus docetaxel in previously treated advanced non-small-cell lung cancer (CheckMate 017 and CheckMate 057): 3-year update and outcomes in patients with liver metastases.* Ann Oncol, 2018. **29**(4): p. 959-965.
12. La Manno, G., et al., *RNA velocity of single cells.* Nature, 2018. **560**(7719): p. 494-498.
13. Lange, M., et al., *CellRank for directed single-cell fate mapping.* Nat Methods, 2022.
14. Qiu, X., et al., *Mapping transcriptomic vector fields of single cells.* Cell, 2022.
15. Sade-Feldman, M., et al., *Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma.* Cell, 2018. **175**(4): p. 998-1013 e20.
16. Haghverdi, L., et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.* Nat Biotechnol, 2018. **36**(5): p. 421-427.
17. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
18. Lin, Y., et al., *scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets.* Proc Natl Acad Sci U S A, 2019. **116**(20): p. 9775-9784.
19. Buettner, F., et al., *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.* Nat. Biotechnol., 2015. **33**(2): p. 155-60.
20. Fusi, N., O. Stegle, and N.D. Lawrence, *Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies.* PLoS Comput Biol, 2012. **8**(1): p. e1002330.
21. Brennecke, P., et al., *Accounting for technical noise in single-cell RNA-seq experiments.* Nature Methods, 2013. **10**(11): p. 1093-5.
22. Grun, D., L. Kester, and A. van Oudenaarden, *Validation of noise models for single-cell transcriptomics.* Nat. Methods, 2014. **11**(6): p. 637-40.
23. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, *Bayesian approach to single-cell differential expression analysis.* Nat. Methods, 2014. **11**(7): p. 740-2.
24. Hao, J., et al., *Optimal Gene Filtering for Single-Cell data (OGFSC)-a gene filtering algorithm for single-cell RNA-seq data.* Bioinformatics, 2019. **35**(15): p. 2602-2609.
25. Buettner, F., et al., *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.* Nat Biotechnol, 2015. **33**(2): p. 155-60.
26. Buttner, M., et al., *A test metric for assessing single-cell RNA-seq batch correction.* Nat Methods, 2019. **16**(1): p. 43-49.

27. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.
28. Bacher, R., et al., *SCnorm: robust normalization of single-cell RNA-seq data.* Nat Methods, 2017. **14**(6): p. 584-586.
29. van Dijk, D., et al., *Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.* Cell, 2018. **174**(3): p. 716-729 e27.
30. Huang, M., et al., *SAVER: gene expression recovery for single-cell RNA sequencing.* Nat Methods, 2018. **15**(7): p. 539-542.
31. Arisdakessian, C., et al., *DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data.* Genome Biol, 2019. **20**(1): p. 211.
32. Loo, R.L., et al., *Characterization of metabolic responses to healthy diets and association with blood pressure: application to the Optimal Macronutrient Intake Trial for Heart Health (OmniHeart), a randomized controlled study.* Am J Clin Nutr, 2018. **107**(3): p. 323-334.
33. van Velzen, E.J.J., et al., *Multilevel data analysis of a crossover designed human nutritional intervention study.* Journal of Proteome Research, 2008. **7**(10): p. 4483-4491.
34. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nat Biotechnol, 2014. **32**(4): p. 381-386.
35. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nat Methods, 2017. **14**(5): p. 483-486.
36. Santos, J.M. and M. Embrechts, *On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification.* Artificial Neural Networks - Icann 2009, Pt Ii, 2009. **5769**: p. 175-+.
37. Zhang, L., et al., *Lineage tracking reveals dynamic relationships of T cells in colorectal cancer.* Nature, 2018. **564**(7735): p. 268-272.
38. Raphael, I., R.R. Joern, and T.G. Forsthuber, *Memory CD4(+) T Cells in Immunity and Autoimmune Diseases.* Cells, 2020. **9**(3).
39. Brummelman, J., K. Pilipow, and E. Lugli, *The Single-Cell Phenotypic Identity of Human CD8(+) and CD4(+) T Cells.* Int Rev Cell Mol Biol, 2018. **341**: p. 63-124.
40. Schumacher, B., et al., *The central role of DNA damage in the ageing process.* Nature, 2021. **592**(7856): p. 695-703.
41. de Boer, H.R., S.G. Llobet, and M.A.T.M. van Vugt, *Controlling the response to DNA damage by the APC/C-Cdh1 (vol 73, pg 949, 2016).* Cellular and Molecular Life Sciences, 2016. **73**(15): p. 2985-2998.
42. Mair, F., et al., *Extricating human tumour immune alterations from tissue inflammation.* Nature, 2022. **605**(7911): p. 728-735.
43. Zheng, C.H., et al., *Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing.* Cell, 2017. **169**(7): p. 1342-+.
44. Grinberg-Bleyer, Y., et al., *NF-kappa B c-Rel Is Crucial for the Regulatory T Cell Immune Checkpoint in Cancer.* Cell, 2017. **170**(6): p. 1096-1108.
45. Nagy, A., G. Munkacsy, and B. Gyorffy, *Pancancer survival analysis of cancer hallmark genes.* Scientific Reports, 2021. **11**(1).
46. Director's Challenge Consortium for the Molecular Classification of Lung, A., et al., *Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.* Nat Med, 2008. **14**(8): p. 822-7.
47. Hao, J., et al., *Optimal Gene Filtering for Single-Cell data (OGFSC) - a gene filtering algorithm for single-cell RNA-seq data.* Bioinformatics, 2018.

48. Bermejo, S. and J. Cabestany, *Adaptive soft k-nearest-neighbour classifiers.* Pattern Recognition, 2000. **33**(12): p. 1999-2005.
49. Zou, X., et al., *Statistical HOmogeneous Cluster SpectroscopY (SHOCSY): an optimized statistical approach for clustering of (1)H NMR spectral data to reduce interference and enhance robust biomarkers selection.* Anal Chem, 2014. **86**(11): p. 5308-15.

Key Points

- Cell-state transition can reveal additional information from single-cell RNA-sequencing data in time-resolved biological phenomena.
- scSTAR constructs a paired-cell projection between biological conditions that is applicable to reveal insight into biological experiments even with a large time span.
- Detailed cell dynamic heterogeneities and novel cell subtypes were revealed in various datasets, which improved the investigation of the biological questions of interest.

Figure 2. scSTAR evaluation on simulated data. (a) The AUC obtained using the data processed by different methods for 2 clusters. (b-e) Clustering results evaluated by ARI with 7 preprocessing and 3 clustering methods at FC=1.3. For 2 subclusters in the case group, (b) with ratio 1:2, (c) 1:1; (d) 3 subclusters in the case group with equal proportions; (e) 4 subclusters in the case group with equal proportions. (f-l) The distribution of cells on normalized average DE gene expression levels for 7 preprocessing methods. The correctly and incorrectly classified cells are indicated. (m-s) The umap scatter plots illustrate the separability of cell subclusters processed by different methods. As Combat only slightly modified the data, the scatter plots of original (n) and Combat (o) look very similar.