scSTAR reveals hidden heterogeneity with a real-virtual cell pair structure across single cell samples

Jie Hao (jhao@fudan.edu.cn)
Institute of Clinical Science

Ke Chen
Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden

Jiaqiang Zhang
Department of Anesthesiology and Perioperative Medicine, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, Zhengzhou, Henan, 450003, China

Duojiao Wu
Zhongshan Hospital, Fudan University

Wei Cao
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 Cent

Jiawei Zou
School of Life Sciences and Biotechnology, Shanghai Centre for Systems Biomedicine, Shanghai Jiao Tong University

Guoguo Shang
Department of Pathology of Zhongshan Hospital, Fudan University

Jean Yang
University of Sydney https://orcid.org/0000-0002-5271-2603

KongFatt Wong-Lin
Intelligent Systems Research Centre, Ulster University, Magee Campus, Derry-Londonderry, Northern Ireland, UK

Hourong Sun
Department of Cardiac Surgery, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan City, Shandong, 250012, China

Wantao Chen
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 o

Xiangdong Wang
Institute of Clinical Science, Zhongshan Hospital, Fudan University

Xin Zou
scSTAR reveals hidden heterogeneity with a real-virtual cell pair structure across single cell samples

Authors: Jie Hao1#, Ke Chen2#, Jiaqiang Zhang3#, Duojiao Wu1#, Wei Cao4, Jiawei Zou5, Guoguo Shang6, Jean Y.H. Yang7, KongFatt Wong-Lin8, Hourtong Sun9, Wantao Chen10*, Xiangdong Wang1*, Xin Zou11*

Affiliations:

1 Institute of Clinical Science, Zhongshan Hospital, Fudan University, Shanghai, China
2 Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Shanghai, 201602, China
3 Department of Anesthesiology and Perioperative Medicine, Henan Provincial People’s Hospital, People’s Hospital of Zhengzhou University, Zhengzhou, Henan, 450003, 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 School of Life Sciences and Biotechnology, Shanghai Centre for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China.
6 Department of Pathology of Zhongshan Hospital, Fudan University, Shanghai, China.
7 School of Mathematics and Statistics and Charles Perkins Center, The University of Sydney, Australia
8 Intelligent Systems Research Centre, Ulster University, Magee Campus, Derry~Londonderry, Northern Ireland, UK
9 Department of Cardiac Surgery, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan City, Shandong, 250012, China
10 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.
11 Jinshan Hospital Center for Tumor Diagnosis & Therapy, Jinshan Hospital, Fudan University, Shanghai, 201508, China

* To whom correspondence should be addressed, Jie Hao, Email: jhao@fudan.edu.cn; Xin Zou, Email: x.zou@hotmail.co.uk; Xiangdong Wang. Email: xdwang@fuccb.com; Wantao Chen, Email: chenwantao196323@sjtu.edu.cn.
#These authors contributed equally to this work.

Abstract:
We present scSTAR, single-cell States Transfer Across-sample of RNA-seq data, which estimates individual cell state dynamics by creating real-virtual cell pairs across samples/conditions, therefore each cell can act as its own control when comparing between conditions. Using simulated and published scRNA-seq datasets, we show that new cell subtypes/functions which are more linked to the biological problem under investigation are discovered, especially when current methods give blur patterns.

Main Text:
Single cell RNA sequencing (scRNA-seq) data offers insights into cell-to-cell biological variation at the transcriptomic level. The mRNA expressional variations of individual cells between, e.g., pre- and post- treatments1, tumor and adjacent normal2, primary and relapse tumors3, across ages4, etc., are critical in understanding the biological progression. However, in current single-cell experimental procedure, cells are destroyed,
making it physically not possible to track the state transition of each individual cell across samples/conditions. Various analytical methods have been developed aiming to tackle this problem\(^5\)\(^6\), however, they are mainly based on the average state value over a cluster of cells, e.g., cell subtypes or cell relatives, and thus the state transition patterns extracted are limited to sub-bulk rather than single cell resolution. Furthermore, interrogating the average state difference across cell clusters can lead to the overlooking of modest biological differences between individual cells, especially when the difference is smaller than the inter-cluster variations. Inspired by bulk data analysis, where this issue is usually addressed by comparing paired samples from the same subject\(^7\)\(^8\), here we propose a paired-cell model, scSTAR (single-cell State Transition Across-samples of Rna-seq data), where for each real cell in one sample/condition, scSTAR estimates its virtual projection in the other. Such creation of real-virtual cell pairs allows each cell to act as its own control when comparing between samples/conditions.

The scSTAR *in silico* estimates of individual cell state transition is achieved by generating real-virtual cell pairs across samples/conditions, and the difference between the real-virtual cell pair should only relate to the biological problem of interest. We demonstrate that cell state dynamics can be achieved by maximising the covariance between cell states from various samples, which is the partial least squares solution (PLS)\(^9\) (see Online Methods for more details). In a typical single cell analytical workflow, scSTAR can be applied just before “clustering/trajectory” etc., step\(^10\)\(^14\) (Figure 1a) to reveal detailed sub-cluster structures of cell dynamics pattern in response to experimental condition changes, even when they are buried by other interference.
Figure 1. scSTAR schematic diagram and preliminary evaluation on scRNA-seq data. (a) The flow diagram of estimating single cell state transfer information, results of the clustering step are shown for the immunosenescence data. (b) scSTAR improves the immunotherapy outcome prediction accuracy. (c, d) scSTAR more clearly reveals the Peripheral Blood Mononuclear Cell (PBMC) discrimination between early and late stage lung cancer patients. (e, f) scSTAR better reconstructs dendritic cell transition trajectories from normal to tumor tissue.

To benchmark scSTAR, we compared the original (unprocessed), mutual nearest neighbors (MNN) and scSTAR processed data with three commonly used unsupervised clustering methods, k-means, SC3 and Seurat on a series of simulated noisy data (Methods). scSTAR-processed data outperforms unprocessed and MNN methods with
clearer clustering and better ARIs (adjusted Rand index) in almost all comparisons, especially at low signal-to-noise ratio (FC (fold change) = 1.3, p <10^{-5}) (Figure S1).

To further evaluate the performance of scSTAR for therapy response prediction, we applied it on a melanoma data\textsuperscript{1} to construct an immunotherapy-response prediction model using cell type composition patterns (Additional information Remark A). scSTAR effectively predicted the immunotherapy induced dynamics of each cell from pre-treatment samples with AUC = 0.96 which is a clear increase compared to the prediction models (AUC = 0.8) with clusters reported in the original study (Figure 1b).

We used scSTAR to uncover novel condition-specific cell types in data from a mouse immunosenescence study (Additional information Remark B). The exhausted T cell (T\textsubscript{Exh}) and effector memory T cell (T\textsubscript{EM}) were prominent in elderly and young animals with unstimulated cells from B6 mice, respectively. In contrast, the central memory T cell (T\textsubscript{CM}) was unaffected with aging (Figure S3a-d). Trajectory analysis also showed that T\textsubscript{EM} tended to be replaced by T\textsubscript{Exh} with aging; but T\textsubscript{CM} was separated from such trajectory (Figure S3e-f). The cell dynamic patterns revealed by scSTAR from B6 mice were confirmed in CAST mice (Figure S3g-j). A similar conclusion was drawn from stimulated cells (Figure S4).

In addition, beside what the original studies have found, novel cell subtypes, new discriminatory patterns, clearer progression trajectories, etc., were identified with scSTAR on four more scRNA-seq datasets\textsuperscript{2,16-18} (Additional information Remark C, D, E, F, Figure S5-10). For example, PBMC divergence between early and late stage lung cancer patients was clearly revealed by scSTAR (Figure 1c, 1d); dendritic cell state transition from normal to tumor tissues was better illustrated (Figure 1e, 1f).

Furthermore, an unknown Treg cell subtype associated with suppression of anti-tumor immune functions (Tumor Associated Treg (TAT)) was uncovered with scSTAR. The Treg cell of adjacent normal and tumor tissues from four solid tumors, liver\textsuperscript{19}, lung\textsuperscript{17}, breast\textsuperscript{20} and esophageal\textsuperscript{2} cancers (Figure 2a, Additional information Remark G) were examined. TAT can be well characterized by high expression of SQLE, HSPA5, HSP90B1 and PSMD12 from mTOC1 gene set (Figure 2b, 2c). It was also found significantly associated with poor patients’ survival (Figure S11-S23) and similar patterns were observed in 11 out of 21 cancers in TGCA database\textsuperscript{21} (Figure 2d, 2e). The patients with high TAT also tended to be nonresponding to immunotherapy on a melanoma dataset \textsuperscript{1} (Figure 2f). Its existence was verified on a human ESCC (Esophageal Squamous Cell Carcinoma) \textsuperscript{2} (Figure 2g-k), a mouse lung cancer\textsuperscript{22} scRNA-seq dataset (Methods, Figure S25) and further confirmed by immunofluorescence microscopy in human ESCC (Figure 2l). TAT suppresses immune functions similar as conventional eTreg (effector Treg) cells, but with much stronger amplitudes (Figure 2k, S24) and start to be functional at very early stage of tumorigenesis (Figure S25d). Additionally, gene enrichment analysis showed TAT has highly active glycolysis and hypoxia, which also indicated a strong immune suppression capability\textsuperscript{23} (Figure 2d). We envisage that the ubiquitous existence of TAT in various solid tumor types and species (human and mouse) endows its potential to be an effective immunotherapy target in cancer treatment.
**Figure 2.** TAT cells in human and mouse tumor samples. (a) The analysis workflow of Treg cells. (b) GSEA hallmark enrichment analysis on the markers of pro-tumor Treg clusters from different cancer types. (c) The highlighted key genes associated with the most common hallmark gene sets across the four cancer types. (d) The enriched hallmark gene sets up-regulated in putative TAT cells. (e) The number of cancers out of 21, in which significant association is found between high expression of selected genes and poor patients’ survival. (f) Significance of gene expression differences between immunotherapy responder and non-responder melanoma patients. (g) Treg clusters composition of tumor and adjacent normal in human ESCC sample. (h) Heatmap of Z-score normalized log2 (count+1) expression of selected T cell function-associated genes in each cell cluster. (i, j) Pseudotime trajectory illustrated a bifurcation transition pattern of Treg cells from normal to tumor tissues. (k) Overall interaction strength between Treg cells and other immune cell types. (l) Immunofluorescence image of ESCC sample.

scSTAR estimates the transition of cell state for each individual cell across samples/conditions. This particular property opens up a new direction of interrogating single cell heterogeneity. We have shown that the information revealed based on the state transfer pattern of each single cell are closely related to clinical metadata (patients' survival, treatment outcomes, etc.) and can bring more biological insights to the cell heterogeneity in response to condition changes even when analysis on cell static profiles fails. scSTAR can be applied directly on scRNA-seq data with various experimental setups, and it has the potential to considerably simplify comparative analysis in a variety of research and practical applications.

**Methods**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper. The scSTAR R package is available at https://github.com/XZouProjects/scSTAR.

**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

**Acknowledgement**

This work was supported in part by the National Natural Science Foundation of China [82170045 to JH, 31800253 to KC and JH, 81771672 to DJW, 81672745 to WC]; 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]; the Shanghai Sailing Program [17YF1410400 to KC]. Innovative Research Team of High-Level Local Universities in Shanghai [SSMU-ZLCX20180502 to JH, WTC].

Competing interests None

Methods

scSTAR - Cell state transit dynamic component dissection

Concept behind scSTAR

In a two-condition comparison scenario, let’s denote $X$ the cell states collected in condition A and $Y$ denote the cell states from condition B. Let us consider condition A represent the baseline and the cell states $Y$ from condition B can decomposed as:

$$Y = \hat{X} + V$$

(1)

where $\hat{X}$ represents the projected profile of the cells from condition B to condition A, therefore, $\hat{X}$ is an unknown variable and not observable from the data, and needs to be estimated. We consider this as the virtual cells paired with $Y$; $V$ represents state transfer of each cell when the condition shifted from A to B. We define $V$ is mainly dependent on the way of manipulation rather than the cell initial states, therefore, $\hat{X}$ and $V$ are not expected to the associated biologically.

For most cell-state analysis, the variable of interest is $V$. To estimate $V$, the conventional scRNA-seq data interpretation strategy is to infer the properties of $V$ by comparing $Y$ and some other observed cells $X$ from condition A. However, the cell-to-cell heterogeneity may confound such performances, especially when $V$ is relatively weak. Such obstacle is ubiquitous in real experiments, such as the changes of cell states are usually smaller than the differences between cell subtypes.

Ideally, to estimate $\hat{X}$ from $Y$, principal component regression (PCR) can be adopted. For Eq 1 and given $X$ and $Y$, principal component analysis (PCA) can decompose $X$ as $p^T t$, where $p$ is PCA loading matrix and $t$ denotes the corresponding PCA score. As $\hat{X}$ and $X$ are different cells from the same condition, therefore, they can be characterized by the same loading $P$, e.g., $X = p^T t$ and $\hat{X} = p^T t'$. Let $Y = p^T t' + V$. Give $Y$ and $p$, $t'$ can be estimated by minimizing square error

$$t' = \arg\min(||Y - p^T t'||^2)$$

(2)

The solution of Eq 2 is $t' = (p^T p)^{-1} p Y = \text{pinv}(p) Y$. And $\hat{X}$ can be estimated as $p^T (p^T p)^{-1} p Y$, further $V = Y - p^T (p^T p)^{-1} p Y$. 
However, in practice, the mismatch between X and Y may bias the results obtained using PCR. To account for such possible mismatch, partial least square (PLS) is adopted here, and the solution can well approximate the solution of Eq. 2. Suppose X and Y can be decomposed into linear combination forms $X = w^t$, $Y = cu$ using PLS. The cost function of PLS provides

$$J = \max_{w^T w = 1, c^T c = 1} \text{Cov}(Xw, Yc) = \max_{w^T w = 1, c^T c = 1} \text{Cov}(Xw, (\hat{X} + V)c) = \max_{w^T w = 1, c^T c = 1} \frac{1}{n}(w^T X^T Yc)$$

$$= \max_{w^T w = 1, c^T c = 1} \frac{1}{n} \text{Cov}(Xw, \hat{X}c) + \text{Cov}(Xw, Vc)$$

(3)

The term $\text{Cov}(Xw, \hat{X}c)$ is maximized when $w = c$ are the eigenvectors of $\text{Var}(X)$ since we assume $\hat{X} \sim X$, which is the solution of PCA of $X$. The second term $\text{Cov}(Xw, Vc) = w \text{Cov}(X, V)c^T = 0$, as $X$ and $V$ are unrelated. Therefore, the solution of Eq 3 tends to approximate PCA analysis results of $X$.

Furthermore, the cost function can be expressed as

$$J = \max_{w^T w = 1, c^T c = 1} \text{Cov}(Xw, Yc) = \max_{w^T w = 1, c^T c = 1} \sqrt{\text{Var}(Xw) \text{Var}(Yc)} \text{Corr}(Xw, Yc)$$

(4)

It can be seen that the estimation of $c$ also accounts for the maximization of correlation between the two group of cells, which may account for the mismatch between $\hat{X}$ and $X$.

Based on the above derivation, the virtual cell $\hat{X}$ can be estimated using $c^T (c^T c)^{-1} c Y$ and the difference between the real-virtual cell pair driven by the biological problem of interest, i.e., $V = Y - \hat{X}$.

Similarly, the cell dynamic components contained in $X$ when compared to $Y$ can also be extracted, i.e., $V_{signal, X} = X - w^T (w^T w)^{-1} w Y$.

**Three-step scSTAR procedure**

In real scenarios, the variations $V$ contained in each cell can be considered as a linear combination of the following components

$$V = V_{batch} + V_{noise}^{r+b} + V_{signal}$$

(5)

where $V_{batch}$ indicates the batch effect, and $V_{noise}$ indicates the interferences not related to the discrimination between the two groups, consisting of random (including technical) (r) and biological (b) noise. $V_{signal}$ represents the gene expression differences between the paired conditions studied.

To reduce the interferences of random noise, the OGFSC method proposed in our previous study is first applied with the parameter $\alpha$ fixed to 0.5, which implies that the genes with variances smaller than the expected values are removed.

**Step 1: Removal of $V_{batch}$**

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

$$V_{\text{batch}} = A_{\text{PLS1}} \text{pinv}(A_{\text{PLS1}}) Y$$

$$\hat{V}' = Y - V_{\text{batch}}$$

where $Y$ is the observed data vector of a cell. $A$ is the PLS loading matrix containing $m$ PLS components. The value of $m$ can be either artificially defined or estimated by maximizing the goodness-of-prediction $Q^2$ calculated by the 10-fold cross-validation method\(^{28}\). $\text{pinv}(A) = (A^T A)^{-1} A^T$ denotes the Moore-Penrose pseudoinverse derived by minimum square error criterion. As $A$ is constructed from the anchor cells, the term $A_{\text{PLS1}} \text{pinv}(A_{\text{PLS1}})$ only contains the variation components related to batch effect (Supplementary Texts – Section 1: Proof of concept: component dissection by PLS). Therefore, the residual variations contained in $\hat{V}'$ are dominated by $V_{\text{signal}}$ and $V_{\text{noise}}$. The batch effect can be removed from the cells of both groups using the Eqs. 2 and 3.

**Step 2: Cell state transfer estimation by decomposing $V_{\text{signal}}$ from $V_{\text{noise}}$**

Using a similar principle, a second PLSDA model (PLS2) is constructed. As noise variations $V_{\text{noise}}$ do not contribute to the discrimination of the two groups, PLS2 dedicatedly captures the variation of $V_{\text{signal}}$. Using the loading matrix $A_{\text{PLS2}}$ to estimate the virtual cell profiles, and signal can be extracted from $\hat{V}'$:

$$\hat{V}_{\text{signal}} = \hat{V}' - A_{\text{PLS2}} \ast \text{pinv}(A_{\text{PLS2}}) \ast \hat{V}'$$

In such a way, the $\hat{V}_{\text{signal}}$ is the estimation of $V_{\text{signal}}$. As a result, all the gene expression values related to irrelevant noises are shrunk towards zero in $\hat{V}_{\text{signal}}$, whilst the expression values of DE genes are retained.

**Step 3: The DE gene identification**

Based on the property of the processed data $\hat{V}_{\text{signal}}$, a new strategy is proposed to identify the DE genes. Considering the non-DE genes have been shrunk to zero, the DE genes should have higher values in $\hat{V}_{\text{signal}}$ than non-DE genes, regardless their original gene expression levels. First, the conventional two-sample rank sum test is applied by comparing the data of the two groups (after batch effect removal). The non-DE gene panel consists of the top 20% genes with the largest p-values.

Let’s define the amount of biological variation of a gene $g$ be quantified by the sum of squares (SS) of $\hat{V}_{\text{signal}}$

$$SS_g = \sum_i \hat{V}_{\text{signal}}_{g,i}^2$$

where $i$ indicates the cells from the same group. Those genes with $SS$ larger in group A than group B, and at the same time, with $SS$ values larger than the upper boundary of the confidence interval of the
non-DE genes (by default 99%), are considered up-regulated in group A. In the same way, the DE genes upregulated in group B can also be identified.

The PLS-DA model tends to only preserve the genes with much larger expression in one group than in the other group. Therefore, for those genes identified as up-regulated in one group, their processed expression values in this group approximate the change amplitudes of the same cells between the two groups.

**Simulated datasets**

To evaluate the proposed algorithm, we used the protocol presented in Haghverdi et al.\textsuperscript{10} to create simulated datasets, and mimic case-control studies, which contain batch effect, 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 sub-clusters with different proportions were simulated in the case group by manipulating the fold changes (FC) of randomly selected differential expression (DE) genes. The FC was positively correlated with the strength of signal. Each sub-cluster contained 25 DE genes, and each simulated dataset was generated for 50 times. The number-of-PLS-components in scSTAR was estimated automatically.

The simulated data contained batch effect, random noise and biological signal (with diverse patterns) at FC = 1.3, 1.5, and 2, to mimic case-control studies. In the case group, 2 to 4 sub-clusters were generated to represent the cell heterogeneity. ARI (see Evaluation metrics for definition), between 0 and 1, indicating 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 Adjusted Rand Index (ARI)

\[
ARI = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} - \sum_{i=1}^{k} a_i \sum_{j=1}^{k} b_j} {\frac{1}{2} \left( \sum_{i=1}^{k} a_i \right) \left( \sum_{j=1}^{k} b_j \right)}
\]

where \(n_{ij}\) are contingency table entry values, \(a_i\) and \(b_j\) are the sum 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, five datasets were adopted: a head and neck squamous cell carcinoma data (GSE103322); an esophageal squamous cell carcinoma tissue dataset (GSE145370); a hepatocellular cell carcinoma tissue dataset (GSE140228); a lung squamous cell carcinoma tissue dataset (GSE99254); a breast cancer tissue dataset (GSE114727).
The potential clinical application of scSTAR was demonstrated on a melanoma immunotherapy treatment dataset (GSE120575).

Validations of scSTAR by revealing new discriminatory patterns, clearer progression trajectories, etc. were adopted on: a mouse immunosenescence study dataset (E-MTAB-4888); a lung cancer peripheral blood cell dataset (GSE99254); an esophageal squamous cell carcinoma tissue dataset (GSE145370); a breast cancer chemotherapy treatment dataset (GSE135710).

**Multi-color IHC**

The clinical specimens performed 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 sample consisted a patient with esophageal squamous cell carcinoma.

Human tissue specimen was collected within 30 minutes after tumor resection and fixed in formalin for 48 h, then followed by dehydration and embedding. The paraffin tissue was cut into 4μm and fixed to glass slide. The slide was placed in 40°C oven for 30 minutes to dry the fixed tissue. Deparaffinization in xylene for 10 minutes for 3 times and then rehydrated in 100%, 95 %, 70% alcohol successively for 3 minutes each. Antigen was recovered by immersion in boiling EDTA buffer (pH9.0) for 15 minutes. Then the slide was blocked by Antibody Diluent/Block to avoid nonspecific site for 10 minutes and incubated with primary antibodies in a humidified chamber for 1h at room temperature (RT). The section was washed by TBST for 2 minutes for 3 times and incubated with HRP-conjugated second antibody for 10 min at RT. Next the section was washed by TBST for 2 minutes for 3 times and incubated with 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), HSPA5 (CST, 1:200) (Opal570) and PSMD12 (1:200) (Opal480).

**References:**

1. Sade-Feldman, M. et al. Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998-1013 e1020 (2018).
2. Zheng, Y. et al. Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment. *Nat Commun* **11**, 6268 (2020).
3. Sun, Y. et al. Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma. *Cell* **184**, 404-421 e416 (2021).
4. Martínez-Jimenez, C.P. et al. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433-1436 (2017).
5. Hormoz, S. et al. Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell systems* **3**, 419-433 e418 (2016).
6. Buder, T., Deutsch, A., Seifert, M. & Voss-Bohme, A. CellTrans: An R Package to Quantify Stochastic Cell State Transitions. *Bioinform Biol Insights* **11**, 117793217712241 (2017).
7. Loo, R.L., Zou, X., Appel, L.J., Nicholson, J.K. & Holmes, E. 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. *The American journal of clinical nutrition* **107**, 323-334 (2018).
8. van Velzen, E.J.J. et al. Multilevel data analysis of a crossover designed human nutritional intervention study. *J Proteome Res* **7**, 4483-4491 (2008).
9. Martens, H. & Naes, T. multivariate calibration (2nd edn). *Wiley: Chichester* **1** (1989).
10. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology* **36**, 421-427 (2018).
11. 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).
12. Lin, Y. et al. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 9775-9784 (2019).
13. 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).
14. Efremova, M., Vento-Tormo, M., Teichmann, S.A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* **15**, 1484-1506 (2020).
15. Kiselev, V.Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods* **14**, 483-486 (2017).
16. Puram, S.V. et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624 e1624 (2017).
17. 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).
18. Lu, Y. et al. Complement Signals Determine Opposite Effects of B Cells in Chemotherapy-Induced Immunity. *Cell* **180**, 1081-1097 e1024 (2020).
19. Zhang, Q. et al. Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell* **179**, 829-845 e820 (2019).
20. Azizi, E. et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308 e1236 (2018).
21. Nagy, A., Munkacsy, G. & Gyorffy, B. Pancancer survival analysis of cancer hallmark genes. *Sci Rep-Uk* **11** (2021).
22. Li, A. et al. IL-33 Signaling Alters Regulatory T Cell Diversity in Support of Tumor Development. *Cell Rep* **29**, 2998-3008 e2998 (2019).
23. Field, C.S. et al. Mitochondrial Integrity Regulated by Lipid Metabolism Is a Cell-Intrinsic Checkpoint for Treg Suppressive Function. *Cell Metab* **31**, 422-437 e425 (2020).

24. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* **33**, 155-160 (2015).

25. Fusi, N., Stegle, O. & Lawrence, N.D. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput Biol* **8**, e1002330 (2012).

26. Hao, J., Cao, W., Huang, J., Zou, X. & Han, Z.G. Optimal Gene Filtering for Single-Cell data (OGFSC) - a gene filtering algorithm for single-cell RNA-seq data. *Bioinformatics* (2018).

27. Bermejo, S. & Cabestany, J. Adaptive soft k-nearest-neighbour classifiers. *Pattern Recogn* **33**, 1999-2005 (2000).

28. Zou, X., Holmes, E., Nicholson, J.K. & Loo, R.L. 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* **86**, 5308-5315 (2014).
# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- NMBCSuppfinal.docx