Briefings in Bioinformatics, 23(1), 2022, 1–12

https://doi.org/10.1093/bib/bbab366
Problem Solving Protocol

Integrating single-cell datasets with ambiguous batch information by incorporating molecular network features

Ji Dong†, Peijie Zhou†, Yichong Wu†, Yidong Chen†, Haoling Xie, Yuan Gao, Jiansen Lu, Jingwei Yang, Xiannian Zhang, Lu Wen, Tiejun Li and Fuchou Tang

Corresponding author. Tiejun Li, LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, China. E-mail: tieli@pku.edu.cn; Fuchou Tang, Beijing Advanced Innovation Center for Genomics, College of Life Sciences, Peking University, Beijing 100871, China. E-mail: tangfuchou@pku.edu.cn
†Ji Dong, Peijie Zhou, Yichong Wu and Yidong Chen contributed equally to this work.

Abstract

With the rapid development of single-cell sequencing techniques, several large-scale cell atlas projects have been launched across the world. However, it is still challenging to integrate single-cell RNA-seq (scRNA-seq) datasets with diverse tissue sources, developmental stages and/or few overlaps, due to the ambiguity in determining the batch information, which is particularly important for current batch-effect correction methods. Here, we present SCORE, a simple network-based integration methodology, which incorporates curated molecular network features to infer cellular states and generate a unified workflow for integrating scRNA-seq datasets. Validating on real single-cell datasets, we showed that regardless of batch information, SCORE outperforms existing methods in accuracy, robustness, scalability and data integration.

Key words: single-cell RNA-seq; molecular network; protein–protein interaction; data integration

Introduction

The improvements in single-cell RNA sequencing (scRNA-seq) techniques have led to the launch of several large-scale cell atlas projects [1, 2]. Since these projects are cooperated by different research groups and involve diverse tissue sources, developmental stages and/or experimental platforms, serious batch effects and technical variations are therefore introduced, bringing challenges for integrating these datasets.

To correct batch effects or technical variations and thus integrate the datasets, many state-of-the-art algorithms (e.g. MNN [3], CCA [4], Harmony [5], LIGER [6], Scanorama [7], etc.) require exact batch information as the prior and guide the data

Ji Dong is a Professor at the Guangzhou Laboratory, Guangzhou, China.
Peijie Zhou is a Postdoctoral researcher at the Department of Mathematics, University of California at Irvine, Irvine, CA, USA.
Yichong Wu is a PhD student at the School of Mathematical Sciences, Peking University, Beijing, China.
Yidong Chen is an Assistant researcher at the Center for Reproductive Medicine, Department of Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China.
Haoling Xie is a PhD student at the Beijing Advanced Innovation Center for Genomics, College of Life Sciences, Peking University, Beijing, China.
Yuan Gao is a PhD student at the Beijing Advanced Innovation Center for Genomics, College of Life Sciences, Peking University, Beijing, China.
Jiansen Lu is a PhD student at the Beijing Advanced Innovation Center for Genomics, College of Life Sciences, Peking University, Beijing, China.
Jingwei Yang is a PhD student at the Beijing Advanced Innovation Center for Genomics, College of Life Sciences, Peking University, Beijing, China.
Xiannian Zhang is an Associate Professor at the School of Basic Medical Sciences, Capital Medical University, Beijing, China.
Lu Wen is an Associate Professor at the Beijing Advanced Innovation Center for Genomics, College of Life Sciences, Peking University, Beijing, China.
Tiejun Li is a Professor at the School of Mathematical Sciences, Peking University, Beijing, China.
Fuchou Tang is a Professor at the Beijing Advanced Innovation Center for Genomics, College of Life Sciences, Peking University, Beijing, China.
Submitted: 31 May 2021; **Received (in revised form):** 17 August 2021

integration based on the shared identical cell types. Nevertheless, due to the co-existence of multiple sources of varations (patients, tissues, time-points and sequencing platforms), the batch information in complex atlas datasets tends to be ambiguous or even ill-defined. The adhoc selection of batch labels may result in the over-correction of one aspect while under-correction of the others, yielding unsatisfactory results [8].

Recently, algorithms such as reference similarity spectrum (RSS) and cluster similarity spectrum (CSS) are also proposed by using external references or intrinsic reference for integration [9–11], while either a complete reference cell-type library (in RSS) or exact batch information (in CSS) is still required. It is therefore urgent to develop a unified framework for unbiased integration of diverse scRNA-seq datasets without exact batch information.

Biological features may serve as the unbiased guidance for scRNA-seq analysis. Several algorithms utilize the transcription factor (TF)-based regulatory networks (e.g. in SCENIC [12]), functional modules (e.g. in PAGODA [13]), cell-type-specific networks (e.g. in CSN [14]) or protein–protein interaction (PPI) networks (e.g. in SCENT and netSmooth [15, 16]) to facilitate the biological interpretation. The robust and scalable frameworks to accurately infer the gene–gene or cell–cell relationship from the sparse scRNA-seq datasets are still under development [17].

Here, we introduce a unified framework, SCORE (Single-Cell mOleculaR nEtwork), which enables integration of scRNA-seq datasets regardless of batch information. We hypothesize that instead of executing function in isolation, genes tend to form complex molecular networks and function in conjunction to determine the cellular or organismal phenotypes [18, 19], and during the organismal development and differentiation, the transition from a cellular state to another is accompanied by the repression of critical modules characterizing the former cellular state and the activation of new modules (Figure 1A). In view of this, SCORE simulates the dynamic changes of molecular networks from scRNA-seq datasets by incorporating the experimentally validated and high-confidence molecular interaction information from public databases. We validated the ability of SCORE on integrating five diverse human fetal datasets and another dataset comprised of 15 adult human major organs. We also validated the accuracy, robustness and scalability of SCORE to uncover cellular states using gold-standard single-cell datasets. SCORE is freely available in https://github.com/wycwycpku/RSCORE.

Results

The workflow of SCORE

To integrate diverse scRNA-seq datasets, we need to transform the noisy gene-expression features to the unified features that are robust against technical variations, while sensitive to biological hetergeneities during development. As we reason that genes usually execute functions through interacting with other genes in molecular networks, we proposed to extract the effective molecular modules that undergo dynamic changes among different cellular states (Figure 1B and Methods) and inspect their activities during the developmental process.

The input to SCORE is the single-cell data matrix analyzed and a prior molecular interaction network. The network can be downloaded from public databases such as BioGRID PPI. Following the SCORE pipelines (Figure 1B): (i) using single-cell data matrix, we first generate a weighted gene–gene correlation network. (ii) To reduce the false positive rates of gene correlations, we use the PPI network to trim the gene–gene relationships; likewise, data-irrelevant interactions of PPI network are pruned by the weighted correlation network. This is accomplished by retaining the highly variable genes (HVGs) as the nodes and assigning correlations as the weights on PPI edges. (iii) The obtained weighted network is then decomposed into numbers of small molecular networks by random walk algorithm, termed as 'modules'. (iv) Next, each module is scored within each individual cell using the AUCell algorithm [12] and (v) a cell-module matrix is obtained, which represents the activity of individual module within each cell. (vi) Finally, this cell-module matrix can be utilized to perform downstream analyses such as visualization, clustering and cell lineage analysis (see Methods).

In addition, inspired by the concept of Steiner Tree in graph theory [20], SCORE also constructs the characteristic molecular interaction network (CMIN) to annotate a certain cellular state (Figure 1B and Methods). For the convenience of application, the SCORE package is seamlessly compatible with the popular R package Seurat for comprehensive scRNA-seq analysis [21].

SCORE is able to integrate scRNA-seq datasets with ambiguous batch information

It is often difficult to provide batch information for datasets generated from diverse tissue sources, developmental stages, etc., because the boundaries between technical variations and biologically meaningful differences are often blurred in such cases. To test the integration ability of SCORE, we utilized two scRNA-seq datasets with relatively ambiguous batch information.

The first dataset is comprised of five high-quality human fetal datasets previously generated by our group, including fetal gonads (overy and testis) [22], heart [23], kidney [24], prefrontal cortex (PFC) [25] and cerebral cortex [26], spanning from 4 to 26 weeks of fetal development (Figure 2A). In such case, widely applied batch correction methods might not be suitable to be applied directly, since the batch information is difficult to be determined: it could be inappropriate to regard each library, each tissue or each developmental stage as the distinct batch labels. Hence, we made comparison between results obtained from SCORE and Seurat pipeline on individual gene expression matrix (denoted as 'Raw' below).

To evaluate the integration result, we calculated the silhouette width and cluster purity for each analysis (see Methods). As shown in Figure 2B, cells of SCORE clustering achieved larger silhouette widths than cells of Raw pipeline clustering, indicating that they were closer in the same cluster than in different clusters. Meanwhile, the cluster purity of SCORE clustering was also higher than that of Raw pipeline clustering. Moreover, we inspected the mixability of common cell types shared by these datasets, such as immune cell, endothelial cells (ECs) and erythroid cells across different organs, expressing *PTPRC*, *CDH5* and *ALAS2*, respectively. Notably, SCORE obtained a reasonable result with cell types grouped by their own identities, while the cells tend to scatter separately in the Raw pipeline (Figure 2C, Supplementary Figure 1A and C, see Supplementary Data available online at http://bib.oxfordjournals.org/). Taking the human fetal kidney dataset as an example, Raw pipeline showed a slight batch effect in the dataset of renal interstitium (RI) that there was a distinctly isolated cell population of 19-week fetal renal cells (Supplementary Figure 2A and B, see Supplementary Data available online at http://bib.oxfordjournals.org/). In contrast, SCORE presented a more biologically reasonable result (Supplementary Figure 2C and D, see Supplementary Data available online at http://bib.oxfordjournals.org/). All fetal cells were first separated into renal cell populations and non-renal

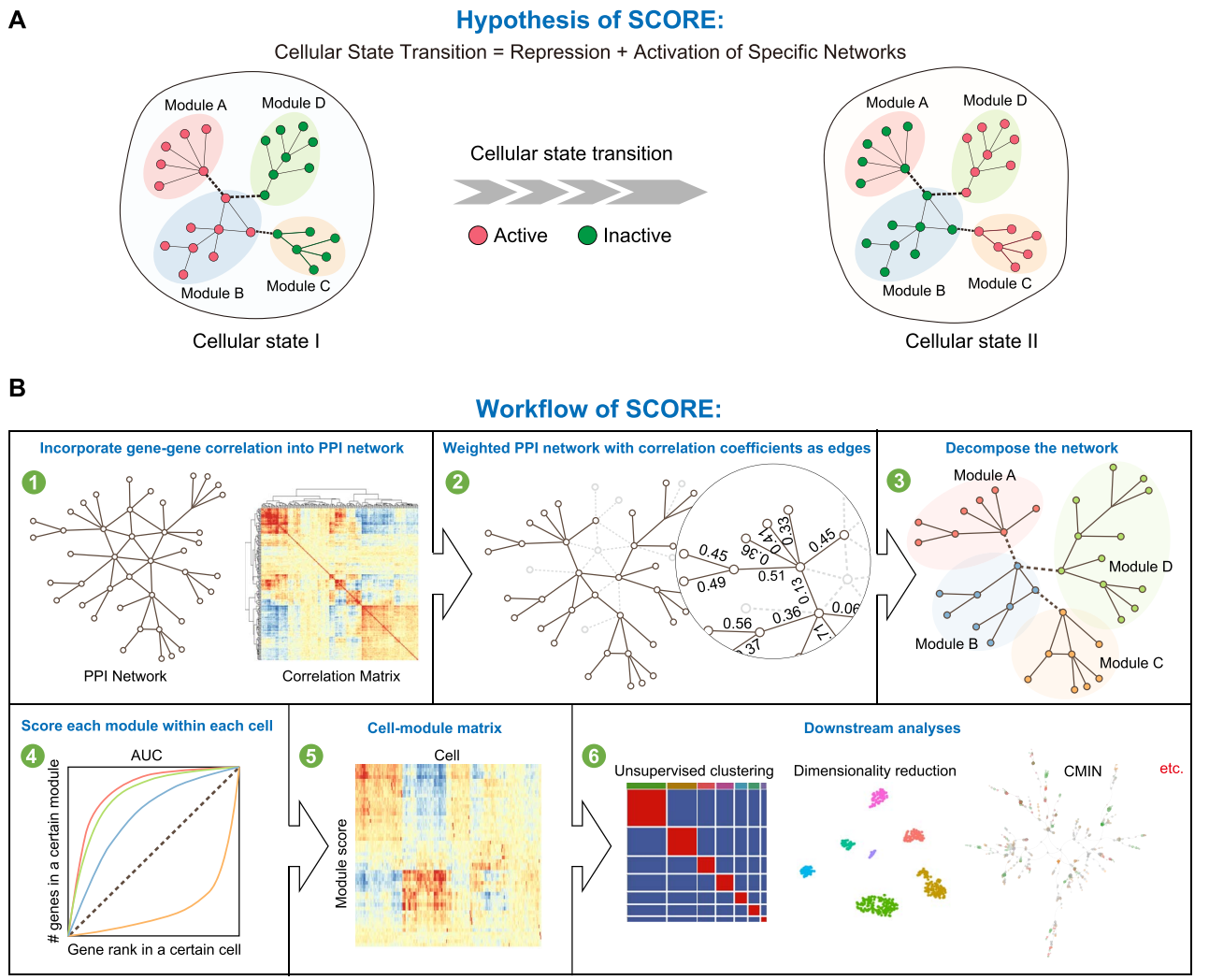

Figure 1. The workflow of SCORE. (**A**) SCORE assumes that the cellular state transition is associated with the activation/repression of functional modules of molecular interaction network, which can be inferred from single-cell transcriptome data. (**B**) In the SCORE workflow, the input PPI network from public database and gene correlation inferred from single-cell dataset are trimmed to construct a weighted molecular interaction network (WMIN). The random walk approach is then applied to WMIN to decompose molecular interaction modules via a consensus strategy, and the activation score of modules for each cell is calculated from AUCell. Downstream analysis is performed based on the obtained cell-module activity matrix to cluster and visualize the cells against technical variations and constructs the CMIN for each cellular state.

cell populations, including immune cells, erythrocytes and ECs. Renal cells comprising glomerular cells, renal capsule cells and renal tubular cells were then classified into nine clusters according to the anatomic structure of nephron, namely, cap mesenchyme (CM), podocytes, proximal tubule, loop of Henle, distal convoluted tubule, collecting duct, intraglomerular mesangium, extraglomerular mesangium and RI. Importantly, the UMAP plot based on SCORE features displayed an accurate developmental trajectory of nephrons from the CM to the formation of epithelial tubules (Supplementary Figure 2E, see Supplementary Data available online at http://bib.oxfordjournals.org/).

The second dataset includes 15 adult human major organs where batch information is also ambiguous [27], since each library is comprised of cells sampled from one organ, and we cannot reagard each library as a batch (Figure 2D). Like the fetal dataset, SCORE successfully grouped cell types across different organs by their own identities, including fibroblasts, NK/T cells, B cells, ECs, macrophages, keratinocytes, smooth muscle, epithelial cells, satellite cells, plasma cells and erythroid cells (Figure 2D and Supplementary Figure 1D, see Supplementary Data available online at http://bib.oxfordjournals.org/). SCORE clustering also achieved larger silhouette widths and higher cluster purity than Raw pipeline clustering (Figure 2E). In contrast, when we mapped RSCORE clustering result to the result of Raw pipeline, we could find clear boundaries in cells sampled from different organs. This pattern was especially obvious in fibroblasts, NK/T cells, keratinocytes and epithelial cells as shown in Figure 2F and Supplementary Figure 1B, see Supplementary Data available online at http://bib.oxfordjournals.org/.

Considering the above analyses were performed with the non-specific BioGRID PPI, we wondered whether a tissue-specific PPI would improve the performance of SCORE when dealing with the corresponding tissue related datasets. SCORE was then tested on both a human stomach dataset [27] and a human lung dataset [28] using three tissue-specific PPI from TissueNet [29], namely Geneotype-Tissue Expression (GTE-x) dataset-RNA-Sequencing PPI, Human Protein Atlas (HPA)-RNA-sequencing PPI and HPA-Protein-Expression PPI. Compared with the UMAP dimensionality reduction result using the non-specifc BioGRID PPI, analyses with these three PPIs also

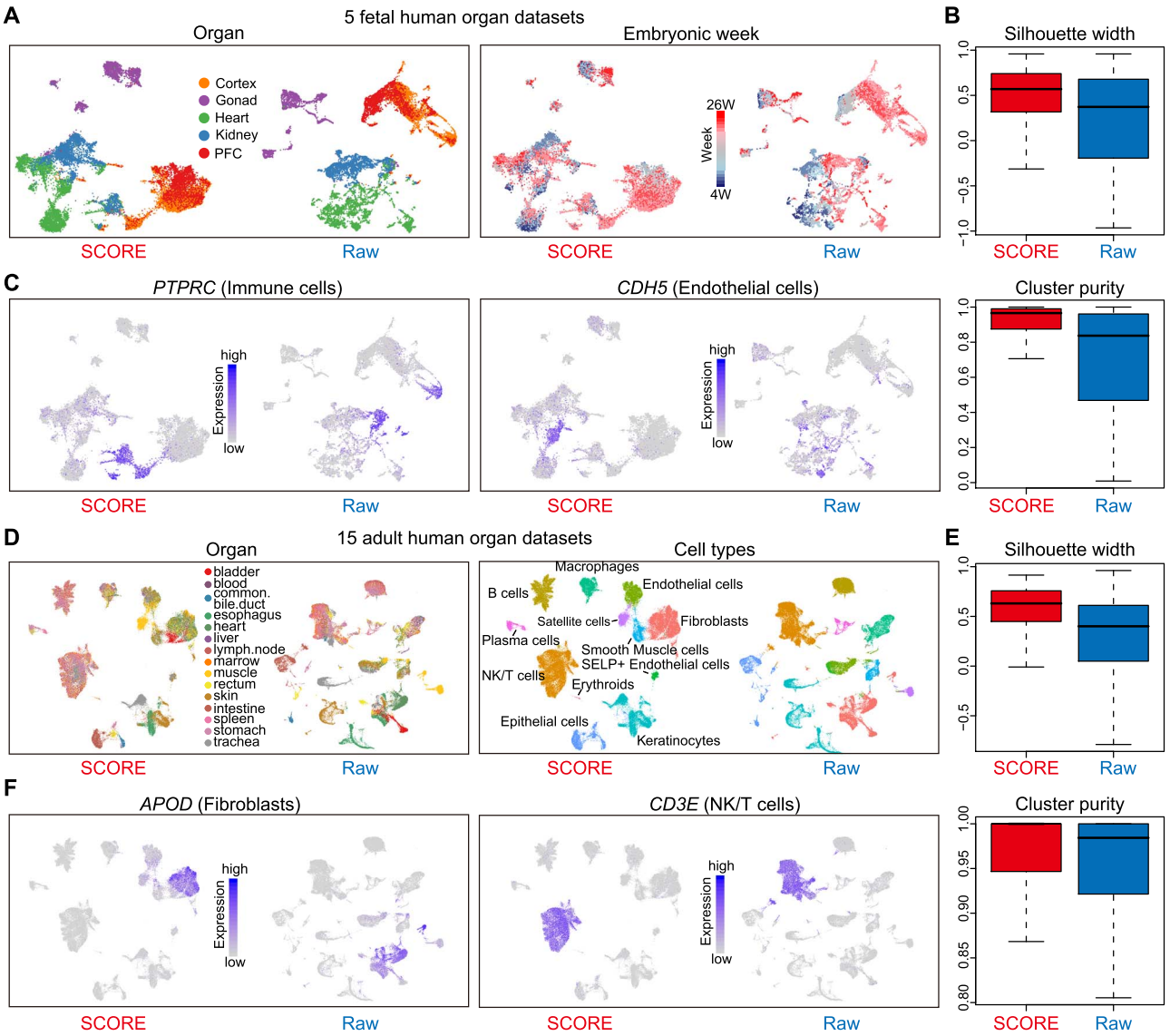

Figure 2. Integration of scRNA-seq datasets with ambiguous batch information. (**A**) UMAP visualization of the 5 fetal human ogran datasets integration results implemented by SCORE and direct merge of datasets within Seurat pipeline (Raw). The cells are colored by organ information (left) and developmental week information (right), respectively. (**B**) Box plots exhibiting the silhouette width (top) and cluster purity (bottom) of SCORE clustering and Raw pipeline clutering in the 5 fetal human ogran datasets. (**C**) UMAP visualization of the 5 fetal human ogran datasets integration results by SCORE and Raw, with cells colored by the expression level of representative marker genes for immune (left) and endothelial (right) cells. (**D**) UMAP visualization of the 15 adult human organ datasets integration results by SCORE and Raw. The cells are colored by organ information (left) and cell type information (right). (**E**) Box plots exhibiting the silhouette width (top) and cluster purity (bottom) of SCORE clustering and Raw pipeline clutering in the 15 adult human organ datasets. (**F**) The UMAP visualization of the 15 adult human organ datasets integration results by SCORE and Raw, with cells colored by the expression level of marker genes for fibroblasts (left) and NK/T (right) cells.

achieved desirable results, especially for HPA-RNA PPI and GTEx-RNA PPI, where similar cells were grouped more tightly (Supplementary Figure 3A–C, see Supplementary Data available online at http://bib.oxfordjournals.org/).

In conclusion, both the fetal and adult datasets indicate that SCORE provides more biological insights toward the cell states inference, when integrating datasets collected from both different stages and various systems during human development. Moreover, the tissue-specific PPI would be a preferable choice for a certain tissue-related dataset.

SCORE is able to integrate multi-platform datasets

To verify the performance of the algorithm to integrate datasets from different single-cell platforms, we applied SCORE to two gold-standard scRNA-seq datasets: one was generated by Smart-seq2 technique [11], while the other was generated by multiple platforms including 10× Genomics 3′, 10× Genomics 5′, cel-seq2 and drop-seq techniques [30]. Here, SCORE does not require batch labels as the input for such integration task.

The Smart-seq2 dataset has 561 cells derived from seven human cell lines (GSE81861) [11]. Two experimental batches exist in the cell lines for GM12878 lymphoblastoid cells and H1 embryonic stem cells. In the previous literature, several clustering methods were benchmarked, and the reference component analysis achieved the highest adjusted rand index (ARI) of 0.91, while other methods showed inferior performance (All = HC: 0.66; HiLoadG-HC: 0.53; BackSPIN: 0.64; RaceID2: 0.15; Seurat: 0.70) (Figure 3A). Here, we intend to continue the benchmarking in the same set-up, by comparing the performance of SCORE

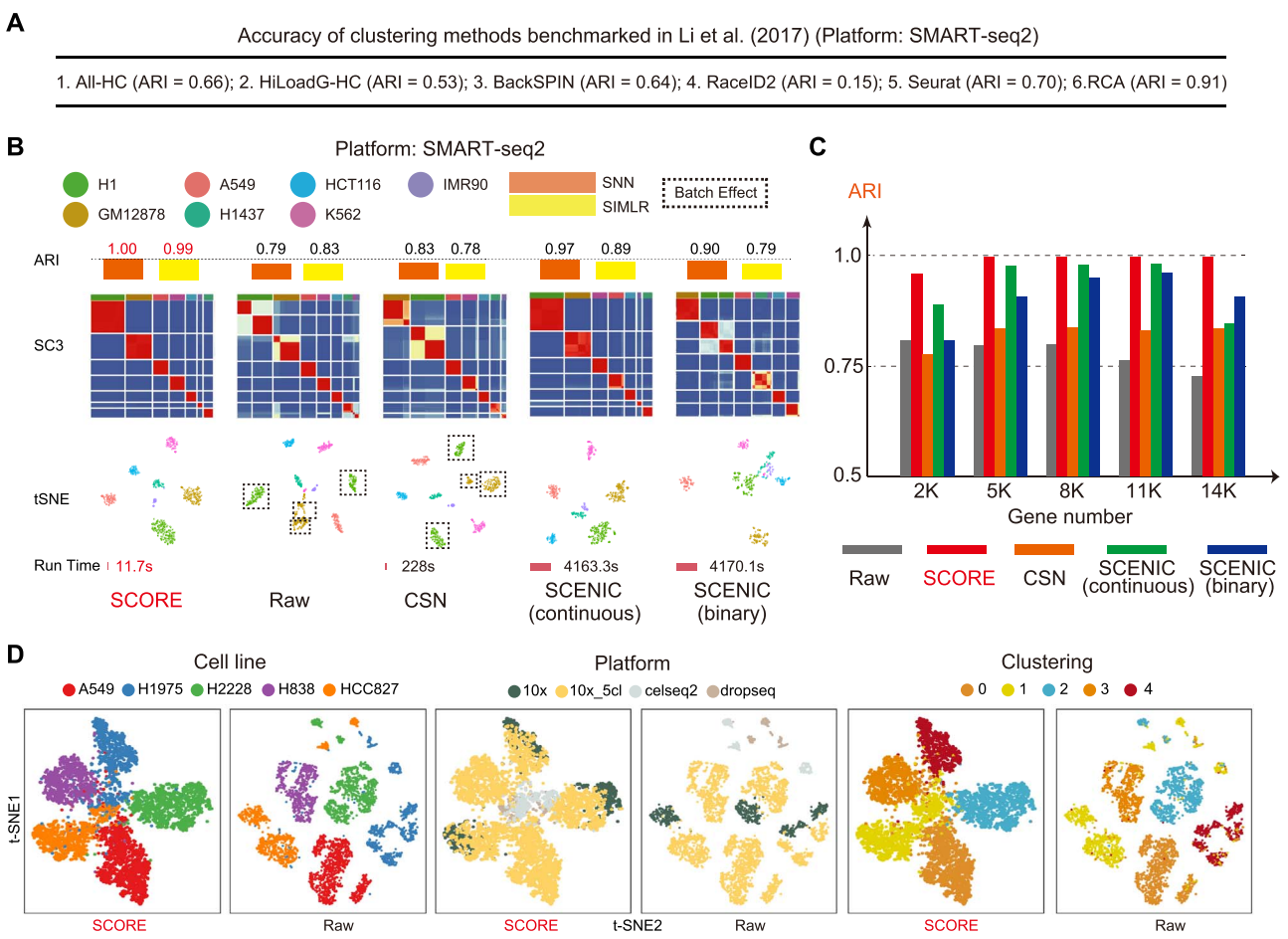

Figure 3. Integration of multi-platform datasets. (**A**) Accuracy of clustering methods evaluated on the gold-standard cell-line benchmarking dataset. (**B**) The performance of SCORE on the Smart-seq2 dataset composed of seven cell lines compared with the direct analysis on raw expression matrix within Seurat pipeline (denoted as Raw) and other GRN-based methods (i.e. CSN and SCENIC with continuous/binary features). SCORE eliminates the batch effects (denoted by dotted box) and outperforms other methods in terms of both clustering accuracy indicated by ARI and running time. (**C**) SCORE outperforms other methods in terms of clustering accuracy indicated by ARI regardless of the number of selected genes. (**D**) The performance of SCORE on the dataset with multiple platforms compared with the direct analysis within Seurat pipeline. The cells are colored by cell line information (the left column), platform information (middle) and clustering results based on integrated data (right), respectively.

with other state-of-the-art feature extraction methods based on gene regulatory network (GRN). The performance of SCORE was robust under the selection of input variable genes for analysis, and SCORE consistently outperformed other methods for the selected gene numbers ranging from 2000 to 14 000 (Figure 3B and C, Supplementary Figure 4, see Supplementary Data available online at http://bib.oxfordjournals.org/). Both the t-Distributed Stochastic Neighbor Embedding (t-SNE) plot and SC3 [31] similarity matrix demonstrated that the batch discrepancy was removed for the same cell lines (Figure 3B). In comparison, while SCENIC achieved satisfactory clustering results with 5000–11 000 genes, the clustering ARI was significantly reduced for the case of 14 000 genes due to the severe batch effects.

For the dataset with multiple platforms, there are 5319 cells derived from five human cell lines (Figure 3D) [30]. The direct clustering based on expression matrix within Raw pipeline confronted with limitations, by blurring the distinct cell identities and discriminating the unwanted experimental batches. In comparison, SCORE yielded a more reasonable result, in which the t-SNE plot showed that the batch effect was eliminated for the same cell lines, while the sharp distinction among different cell lines was still retained (Figure 3D). Overcall, application of SCORE to the gold-standard dataset validates our rationale: the incorporation of molecular interaction network in scRNA-seq analysis can facilitate the accurate cell identity dissection and reduce the technical artifacts in experiments.

Importantly, the implementation of SCORE is also efficient: with 14 000 top variable genes, SCORE finalized the analysis within 2 min, while the implementation of SCENIC lasted for 6 h in R (Figure 4A). Meanwhile, SCORE is also able to deal with the zero counts which are very common in scRNA-seq datasets (Figure 4B). We randomly replaced the expression value by zero with proportions from 10 to 90% in two scRNA-seq datasets [11, 32] and found that the ARI scores were still relatively high even if the zero count ratios were higher than 50%, suggesting the robustness of SCORE. Besides, with SCORE, we could obtain not only the differentially expressed genes (DEGs) but also the differentially activated modules (DAMs), and their dinstictive activities in different cell states supported the clustering accuracy of SCORE (Supplementary Figure 5A, see Supplementary Data available online at http://bib.oxfordjournals.org/). We could then construct the CMIN of a certain cluster to further explore the relationships and interactions of crucial genes (Figure 4C

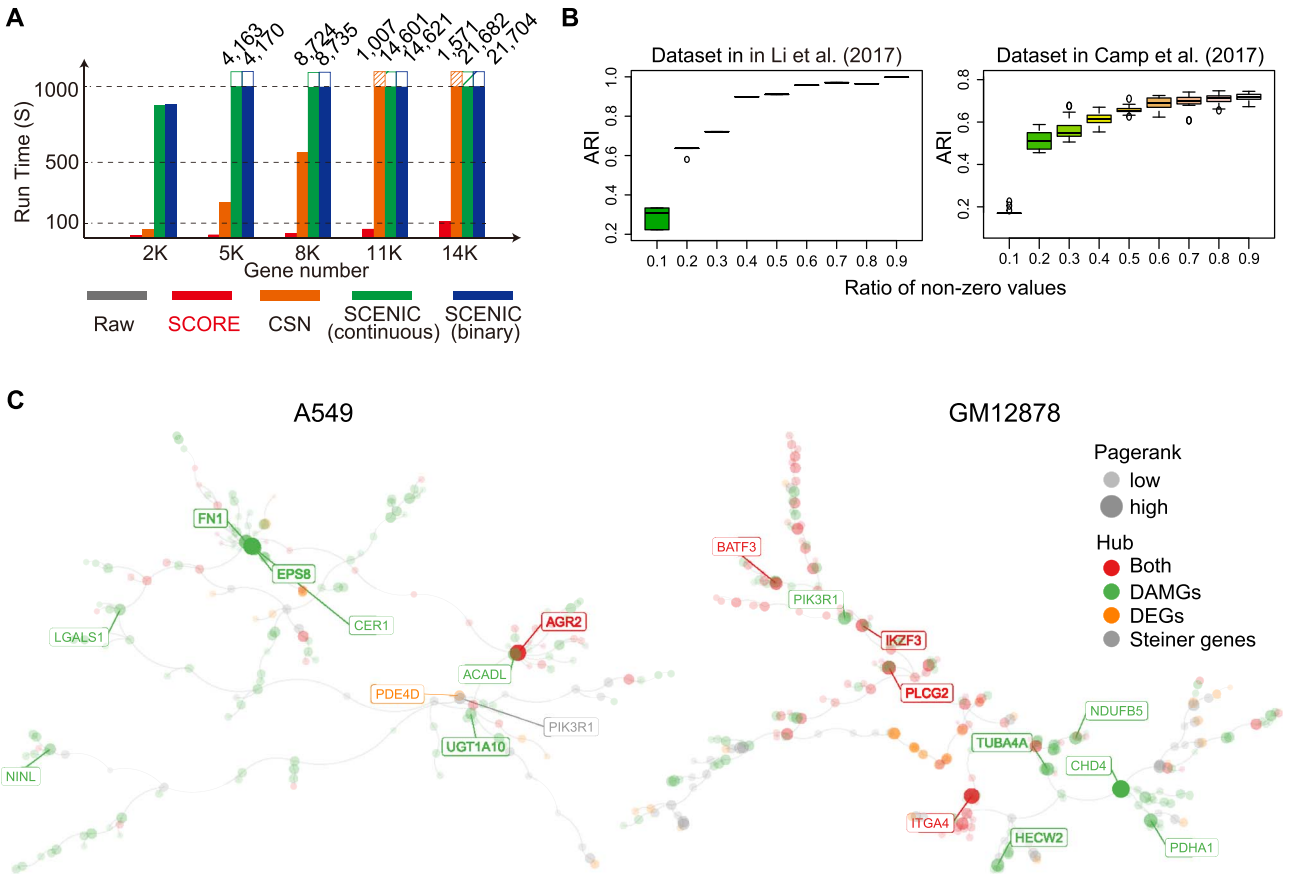

Figure 4. The assessment of SCORE on efficiency, zero gene expression values and gene modules. (**A**) SCORE outperforms other methods in terms of running time regardless of the number of selected genes. (**B**) Boxplots displaying the performance when randomly replacing the expression value by zero in two scRNA-seq datasets. (**C**) CMINs of the A549 (left) and GM12878 (right) celllines as an example. The node sizes in CMIN represent the PageRank score of each gene and the genes of top 10 PageRank scores are displayed with their names in CMIN. The node colors denote the classification of genes as DAMGs (green), DEGs (orange), both (red) and the connecting Steiner genes (gray) in CMIN.

and Supplementary Figure 5B, see Supplementary Data available online at http://bib.oxfordjournals.org/). Unlike traditional methods that focus on the differences in expression levels, CMIN ranks genes based on their topological importance in the optimal steiner tree. Therefore, the CMIN provides a simplified representation of original PPI network and highlights the significance of non-marker interacting genes surpassing traditional marker gene analysis.

A curated PPI network is important to the performance of SCORE

We performed several tests to evaluate the importance of the PPI network chosen in SCORE, a key assumption of our algorithm. The results validated the indispensable role of a curated PPI network in SCORE.

Firstly, we randomly removed the nodes or edges of the PPI network with proportions from 90 to 10% in two scRNA-seq datasets [11, 32]. As shown in Figure 5A and B, we found that 'down-sampling' of the PPI network would significantly affect the ARI score of clustering, especially when similar proportion of the nodes within the PPI is removed. Moreover, we also observed that the effects of node or edge removal within the PPI varied in different datasets. In the first dataset, ARI score could keep higher than 0.8 even if the coverage ratio of the nodes or edges is just slightly higher than 20%, while for the second dataset, ARI score drastically decreased when the coverage ratios of the nodes are below 50% or when those of the edges were below 30%.

Next, we constructed the 'fake' PPI by randomly creating links between gene nodes. We used the faked PPI to replace the PPI from Biogrid or STRING as the input to SCORE. As shown in Figure 5C, the replaced PPI network generated less satisfactory results. Besides, as above, 'down-sampling' of the PPI also significantly decreased the ARI scores.

Lastly, we adopted the gene sets in KEGG, uniprot and GO to replace the modules by decomposing the PPI in SCORE. As shown in Figure 5D, these gene sets yielded inferior results compared with the PPI models, therefore suggesting the significance of PPI to enhance the cell state discovery.

SCORE is applicable to multi-modal and multi-species datasets

Having shown that our method can be effective for the scRNA-seq analyses, we wondered whether SCORE has the potential to extend its application ranges beyond mere scRNA-seq datasets. Firstly, we tested SCORE on a dataset generated by both scRNA-seq and single-nucleus RNA-seq (snRNA-seq) techniques [33]. There existed significant batch effects between cells generated by these two techniques, as shown in the clustering result of the Raw pipeline, where cells were clearly separated by each

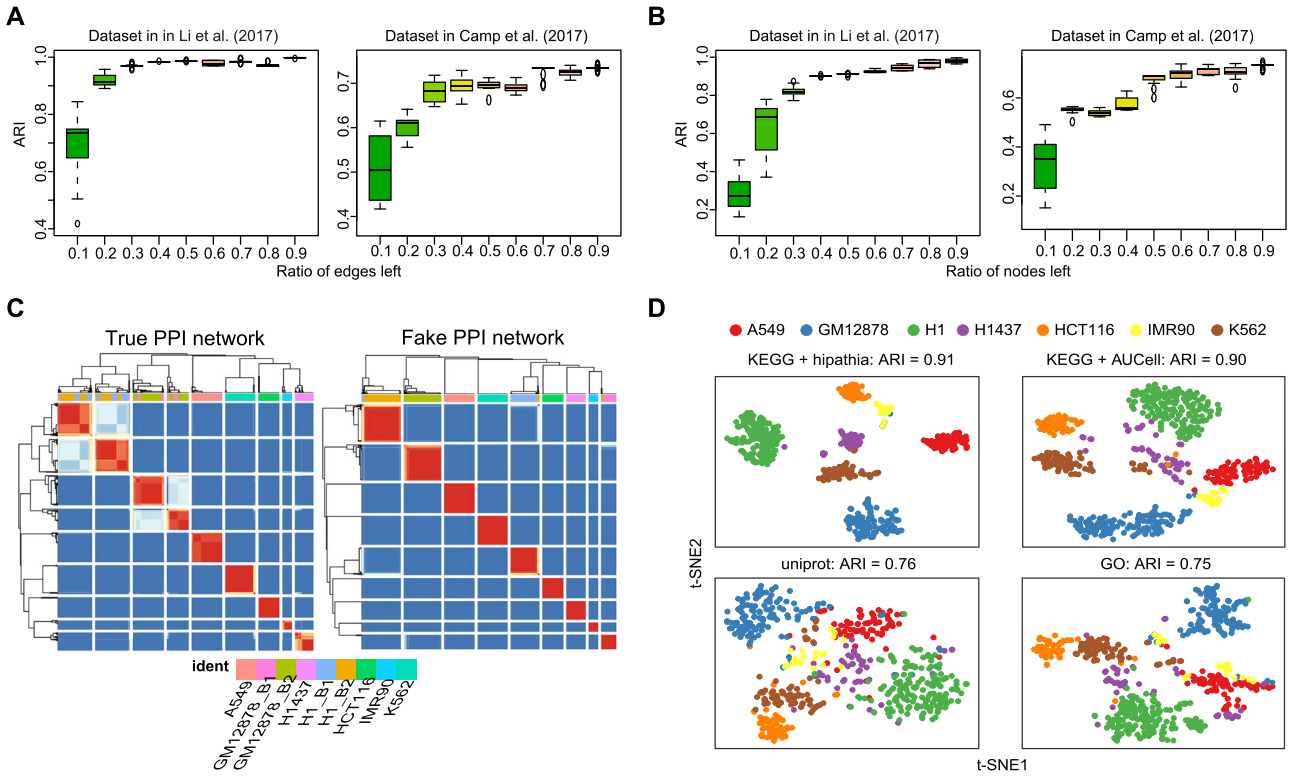

Figure 5. Importance of a curated PPI network to SCORE. (**A**) Boxplots displaying the performance when randomly deleting the edges of PPI in two scRNA-seq datasets. (**B**) Boxplots displaying the performance when randomly deleting the nodes of PPI in two scRNA-seq datasets. (**C**) The SC3 similarity plots of the results under ture and fake PPI network, respectively, when the cluster number is 9. The result of fake PPI network will distinguish 9 clusters clearly, while the true network will blurs the two batches, which means that fake PPI network cannot remove batch effect. (**D**) *t*-SNE visualization based on different network information such as KEGG, uniprot and GO.

technique (Figure 6A). In contrast, SCORE successfully grouped cells by their own identities rather than by the techniques.

Secondly, we applied SCORE on a mouse spermatogenesis dataset [34] and a human spermatogenesis dataset [35] and tried to integrate these two datasets. As shown in Supplementary Figure 6A and B, see Supplementary Data available online at http://bib.oxfordjournals.org/, the developmental trajectories obtained by SCORE were more continuous than thoses by Raw pipeline in both human and mouse datasets. In terms of the integration result of these two datasets, human and mouse datasets were clearly separated in the Raw Seurat pipeline analysis (Figure 6B). In contrast, SCORE, followed by Harmony correction, eliminated the species differences, grouped the same cell types in different species and separated the somatic and germ cells correctly.

Finally, we tested SCORE on the scATAC-seq and multi-modal datasets. From the scATAC-seq measurements, we took the transformed gene activation score matrix as the input to SCORE (see Methods). Compared with the analysis on raw gene activation matrix (Figure 6C and Supplementary Figure 6C, see Supplementary Data available online at http://bib.oxfordjournals.org/), the SCORE analysis of 10 k human PBMC dataset from 10× scATAC-seq yielded improved clustering and dimension reduction results for major blood cell types, especially for T cells and monocytes, which were scattered in the raw analysis. For the multi-modal datasets, we used SCORE to compute module activity scores for each assay and used them as the input to the weighted nearest neighbors (WNN) method for the joint analysis of different assays (see Methods) [36]. For the human PBMC multi-modal dataset by 10× paried profiling of scRNA-seq and scATAC-seq, we also observed the enhanced performance for B cells and T cells after processing with SCORE (Figure 6D and Supplementary Figure 6D, see Supplementary Data available online at http://bib.oxfordjournals.org/).

In summary, SCORE has the potential to integrate scRNA-seq and snRNA-seq datasets, merge cross-species datasets and can be extended to enhance the analysis of fast-growing multi-modal single-cell datasets.

Discussion

The close cooperation of different genes forms modules to perform specific cellular functions, and a certain cellular state or cell type can be well depicted by the activities of various gene modules [18]. As cellular states transit rapidly during organismal development and differentiation, how to simulate these dynamic processes and uncover the corresponding cell fates becomes a fundamental biology issue. In this study, we present a new computational framework, SCORE, to infer this dynamic change and reveal cell development trajectory from the molecular network point of view. There are several advantages of SCORE compared with currently widely used methods.

Firstly, SCORE can integrate scRNA-seq datasets regardless of batch information. Since scRNA-seq analyses are often hampered by ubiquitous batch effects introduced by various dissociation protocols, experiment conditions, etc., a number of batch correction methods have been developed to eliminate this issue. However, such methods usually rely on the explicit batch

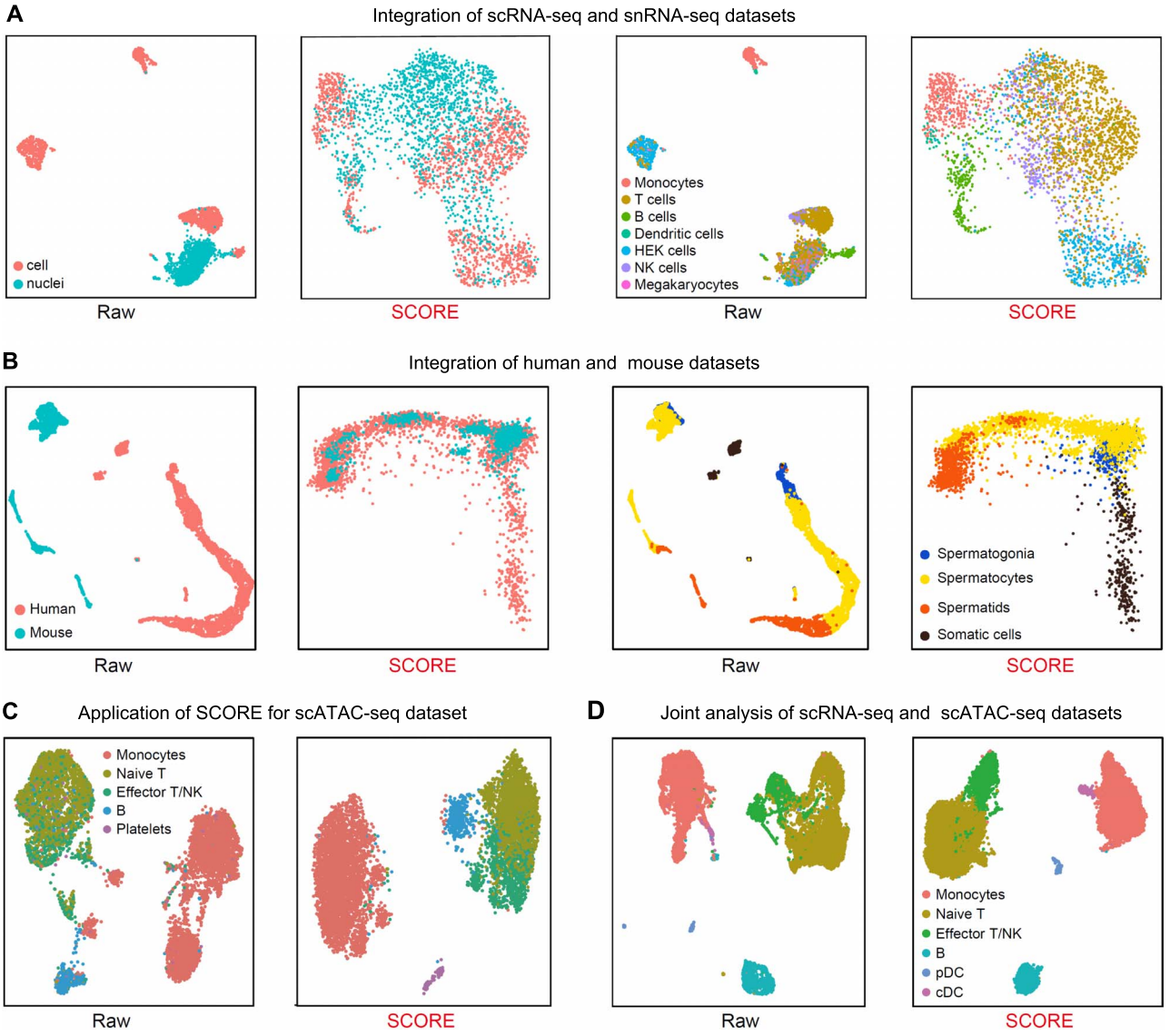

Figure 6. The extended applications of SCORE. (**A**) UMAP visualization of the performance of SCORE (right) on the integration of scRNA-seq and snRNA-seq datasets compared with Raw pipeline (left). Cells are colored by technique information or cell types. (**B**) UMAP visualization of the performance of SCORE (right) on the integration of human and mouse datasets compared with Raw pipeline (left). Cells are colored by species information or cell types. (**C**) UMAP visualization of the performance of SCORE (right) on the scATAC-seq datasets compared with Raw pipeline (left). Cells are colored by cell types. (**D**) UMAP visualization of the performance of SCORE (right) on the multi-modal datasets compared with Raw pipeline (left). Cells are colored by cell types.

information and confront with over- or under-correction challenges for datasets generated from diverse tissue sources and developmental stages, etc. In contrast, SCORE presents a unified data integration framework based on the molecular network information, which is able to integrate scRNA-seq datasets even with the ambiguous batch information.

Secondly, SCORE exhibits more biological rather than statistical significance. SCORE is based on the biology-informed assumptions about the underlying generation mechanism of scRNA-seq data: the genes usually function in conjunction through molecular networks rather than in isolation. It explicitly utilizes the available molecular network information from public databases, which improves the efficiency and accuracy in gene module inference. Some existing methods might lack the information of molecular networks, which calculate the transcriptional similarity inferred from the expression level of individual genes [4, 31, 37]. In addition, the SCENIC algorithm [12] utilizes TF-based networks to retrieve regulatory activity patterns, which can also be used to annotate cellular states. SCENIC measures the activity patterns of ∼1500 TFs while leaving tens of thousands of non-TF genes, thus the resolution of SCENIC might be not fine enough in analyzing highly heterogeneous single-cell data. In comparison, SCORE infers the molecular network from most of the expressed genes and adopts the network signatures rather than individual genes to define a certain cellular state, which improves the resolution and also provides new insights on the considered biological process in molecular level.

Thirdly, the usage of database in SCORE can reduce the false positive rates of gene interaction inference and yield more

accurate results. SCORE utilizes the curated PPI network to correct the inferred gene–gene relationship, in which each interaction pair has experimental supports. As the results shown in Figure 3B and C, SCORE possesses higher accuracy and robustness compared with other existing methods.

Fourthly, SCORE has the promising scalability to extend its application scope, based on the analysis framework of molecular network module activity. It was not only able to merge multi-species datasets from both fetus and adults, but also applicable to integrate scRNA-seq and snRNA-seq datasets, and improve the anaylsis for multi-modal single-cell datasets. We anticipate that SCORE can also be useful for the single-cell proteomics data in the future.

However, as SCORE relies on the curated PPI network and the available gene number in the dataset, we highly recommend researchers to apply SCORE to analyze high-quality scRNA-seq datasets with high-confidence molecular interaction information and greater sequencing depth.

In summary, with high accuracy, robustness and scalability, SCORE can help to integrate and analyze the scRNA-seq datasets from various sources and gain more insights to the considered complex biological processes.

Methods

Overview of SCORE

For the convenience of users, SCORE is seamlessly compatible with the Seurat pipeline. The input for SCORE workflow includes a single-cell gene expression matrix and a PPI network. Both the cells and genes in the expression matrix could be pre-filtered by users, while it is highly recommended that enough number of genes should be retained to achieve more robust analysis of SCORE, as shown via the gold-standard cell line dataset (see main text and SI). The nodes of PPI network should overlap considerably with the gene names in the expression matrix, and the edges of the network shall represent the corresponding molecular interactions with relatively high confidence. In the R implementation of SCORE, the procedure supports the automatic download of PPI from public database such as Biogrid and STRING, which is recommended in the standard workflow. Besides, SCORE was also tested using three tissue-specific PPI datasets obtained from TissueNet [29], namely GTE-x-RNA-Sequencing PPI, HPA-RNA-sequencing PPI and HPA-Protein-Expression PPI. The output of SCORE is a cell-module matrix representing the activity of individual dynamic module within each cell, which can be utilized for downstream visualization, clustering and cell lineage analysis.

Dissection of dynamic molecular networks

To simultaneously reduce the false-positive interactions from correlation inference, and prune data-irrelevant interactions of PPI network, SCORE constructs a weighted molecular interaction network (WMIN) $\mathcal{G}(V, E, W)$ by combining the data-driven and knowledge-based approaches. The vertex set V of the network only consists of the genes in the input expression matrix, and the edge set E obtained from the input PPI network. The weight W_{ij} on the edge E_{ij} is the Pearson correlation coefficients between the corresponding nodes i and j calculated from the single-cell gene expression matrix. To improve the interpretability of molecular networks and highlight the co-expression features, by default, we only keep the edges with positive weight in the WMIN.

To achieve fast and robust identification of dynamic molecular networks, SCORE utilizes the consensus detection of the weighted network community through random walk approach. Given the weighted network $\mathcal{G}(V, E, W)$, a random walk on the network is naturally induced, whose transition probability matrix (TPM) P is defined by

$$P_{ij} = \frac{W_{ij}}{d_i}, d_i = \sum_{j \in \mathcal{N}(i)} W_{ij},$$

where $\mathcal{N}(i)$ denotes the neighbors of the node i. SCORE constructs an ensemble of random walks with different step lengths on the WMIN, with the TPMs $P^{l_1}, P^{l_2}, \ldots, P^{l_R}$, where the power of matrix l_m denotes the length of time step.

For each random walk, the walktrap algorithm is applied to detect network community, respectively [38]. The algorithm partitions the network in terms of the deterministic distance induced by the TPM of random walk, based on the intuition that the random walker will be 'trapped' in the closely-connected sub-networks or communities (termed as modules). All modules with molecule number larger than 3 in each run will be kept as the final dynamic modules, resulting in the module set $\{M_k\}_{k=1}^{L}$. The set is not subject to noise for the given network and chosen step size.

It would be possible that certain modules occur repeatedly in different runs of walktrap algorithm, indicating their stability to form a closely-connected community. SCORE strengthens the weight of such modules automatically in the subsequent analyses, by restoring all the modules without deletion of repeated items, forming an ensemble of modules generated by the network random walks with various step sizes, as the input to downstream analysis of module activity.

Quantification of the module activity

For each detected module from the ensemble, SCORE utilizes AUCell to quantify its activation level within each individual cell. Given cell x, genes are ranked in descending order according to their expression level in x. By default, the z-score is adopted in SCORE to rank the genes in order to remove the effect of scaling. The recovery curve (ROC) for module M_k is then derived by counting the top ranked genes enriched in M_k. The activity measure $A_k(x)$ of module M_k in cell x (consists of the final output matrix of SCORE) is defined as the area under the curve (AUC) for the top ranked genes. Intuitively, modules with higher activity in the biochemical process tend to possess the high-ranking gene expression level, therefore associate with higher activity measure. The AUCell procedure is independent of gene expression unit or normalization method, therefore achieving the effective removal of batch effect in the single-cell experiments.

Downstream analysis

The obtained module activity $A_k(x)$ matrix from SCORE (whose rows represent modules and columns represent cells) can replace the raw gene expression matrix as the input for downstream analysis, such as dimension reduction, clustering and lineage inference. As shown in various datasets of the main text, the downstream analysis based on SCORE module activity features outperforms the raw expression matrix, in terms of clustering accuracy, development lineage trend and removal of experimental batch effects. Hence, the workflow of SCORE can be understood as the extraction of biologically meaningful

and robust features, guided by molecular interactions in the single-cell transcriptome data.

For the convenience of downstream analysis, the R implementation of SCORE is deeply fused with the workflow of Seurat v3.0 package. The input expression matrix to SCORE can be a Seurat object with RNA assay, and the output module activity features are returned as the Net assay in the same Seurat object. Users may conveniently conduct dimension reduction and cell-clustering based on the Net assay and perform marker gene analysis based on the RNA assay, by switching the default assay of Seurat object.

Construction of cell state-specific CMIN

To annotate a certain cell state, beyond the marker genes, SCORE constructs the CMIN with the concept of Steiner Tree in graph theory [20]. Given a graph $G = (V, E)$ and a subset of vertices $T \subset V$ (called terminal vertices), a Steiner tree $S \subset G$ is a connected tree that spans through the given terminal vertices T. The Steiner tree S may contain vertices not presented in T, known as the Steiner vertices, serving as the interchange node to connect the vertices in T. In molecular interaction network, the marker genes of a certain cell state are typically selected as the terminal vertices, and the mediating Steiner vertices, although not necessarily differentially expressed, are supposed to play important roles in formulating the specific cell state through molecular interactions. Therefore, the Steiner tree provides a simplified representation of original PPI network and highlights the significance of non-marker interacting genes surpassing traditional marker gene analysis.

In the downstream analysis of SCORE, given the specific cell cluster identified from module activation features, we first construct the set of terminal genes T^* by detecting the marker genes from two different levels:

(i) Union of all genes in the SCORE-extracted modules that are differentially activated in the cluster, denoted as differentially activated module genes (DAMGs).

(ii) Individual genes that are significantly up-regulated in the cluster denoted as DEGs.

While the DEGs are commonly referred as the 'markers' of cellular states, the DAMGs also represent the key molecular interaction modules to mark the cell cluster in the network resolution.

Next, to infer the CMINs that possibly formulate the cellular states rather than the genes solely marking the cellular states, we propose to calculate a Steiner tree S^* that spans the terminal gene set T with some optimal property, defined on the constructed WMIN $\mathcal{G}(V, E, W)$ by SCORE in the first step of workflow. We require that the separate DAMGs or DEGs in S^* are linked by the most relevant genes, as well as through the most likely interaction path derived from the dataset. To this end, we define the distance D_{ij} on the edge E_{ij} of WMIN by $D_{ij} = 1/(W_{ij} + \varepsilon)$, where the small number ε is added to avoid zero in dominator. Highly correlated gene pairs in PPI tend to possess much closer distance from the definition. Then, S^* can be optimized as the Steiner tree with the least sum of edge distances, which can be tackled efficiently by the greedy algorithm.

For a better visualization of CMIN, in R implementation of SCORE, we mark DAMGs, DEGs and Steiner connecting genes with different colors and also use the PageRank algorithm to measure the topological importance of the genes in optimal Steiner tree S^* as shown by the size of the nodes. We can also provide any two genesets to construct the CMIN.

Human fetal datasets

To evaluate the integration performance of SCORE, five human fetal datasets were collected from our previously published studies, including fetal gonads (overies and testis) (GEO number: GSE86146), heart (GEO number: GSE106118), kidney (GEO number: GSE109488), PFC (GEO number: GSE104276) and cerebral cortex (GEO number: GSE103723), spanning from 4 to 26 weeks of fetal development. Importantly, we re-organized the five datasets using uniform pipeline and format, which provided a rich and convenient resource for studying human fetal development (https://github.com/zorrodong/HECA). In brief, barcode and UMI information were extracted by UMI-tools from raw reads [39]. After discarding the poly A bases, TSO sequences and low-quality sequences, the clean reads were mapped to GRCh38 reference using STAR aligner [40]. We used featureCounts [41] to annotate the mapped reads and quantified the UMI counts through UMI-tools. We provided the pipeline for users (https://github.com/zorrodong/HECA).

To analyze the human fetal datasets, we first discard cells with gene number below 1000 and UMI counts below 10 000. HVGs were chosen using Seurat (mean ≥ 0.1, dispersion ≥ 0.1), and about 8000 HVGs were selected for SCORE to perform the evaluation.

Human adult datasets

The second dataset was collected with the GEO number (GSE159929), which includes 15 adult human major organs [27]. To analyze the human adult dataset, we first discard cells with gene number below 500; UMI counts >1000 while $<40\,000$; and the percent of mitochondrial genes below 25%; 4000 HVGs were selected for SCORE to perform the evaluation.

Calculation of silhouette width and cluster purity

To quantitatively assess the performance of SCORE clustering, we conducted silhouette width and cluster purity analyses with cluster and bluster packages in R. Silhouette width analysis can indicate the separation degrees of the clusters. For each cell, the average distance to all cells in the same cluster was calculated. Meanwhile, the average distance to all cells in another cluster was also calculated. The silhouette width indicates the differences between these two values. The larger silhouette widths of cells indicate that they are closer in the same cluster than in different clusters. For each cell, clustering purity measured the fraction of neighboring cells within the same cluster, and higher purity value indicates higher cluster separation.

Settings in the benchmarking with multiple scRNA-seq platforms

For the SMART-seq2 platform. The gene expression matrix was processed as fragments per kilobase per million reads (FPKM) as in the original literature. To test the robustness of different methods, we first adopted the vst method in Seurat v3.0 package to select five groups of HVGs (with the number of genes 2000, 5000, 8000, 11 000 and 14 000, respectively). We performed SCORE, as well as two other network or biological information based methods, CSN and SCENIC to further compress and extract the features, respectively, from the groups of HVGs. For the second dataset generated by multiple platforms including 10× Genomics 3′, 10× Genomics 5′, celseq2 and dropseq techniques [30], we used the '10×' mode.

For SCORE, we downloaded the *Homo sapiens* PPI network (version 3.5.173) from the BioGRID database. The top ranked genes included in the calculation of AUC values varied with the sizes of input HVGs, with 250 and 200 for 2000 and 5000 variable genes, respectively, and 400 in other cases. The parameters in implementing CSN and SCENIC were chosen with default values. The SCENIC yields both continuous and binary features as the outputs.

The running time comparison was conducted on the 2.50 GHz Xeon E5–2680 machine with 128G RAM, 12 cores and Linux OS. For SCORE and SCENIC, the CPU core number was set as 10. CSN was automatically paralleled with MATLAB 2019b. The wall time of implementing each procedure was recorded as the running time in the main text.

In the downstream clustering analysis, three methods, SNN, SIMLR and SC3, were performed on the extracted features by different methods, and the ARI as well as the similarity matrix of SC3 were used to evaluate the accuracy and the effect of batch removal. The direct analysis on raw expression matrix with selected HVGs was also performed for the comparison. The true labels were the seven collected cell lines identity (H1, GM12878, A549, HCT116, H1437, K562 and IMR90) without batch information. For SNN, we tuned the resolution parameter to obtain the optimal ARI value. As to SC3 and SIMLR, we set the number of clusters to 7. The *t*-SNE plot was produced based on the top 10 principal components of the extracted features.

Integration of scRNA-seq and snRNA-seq datasets, multi-species datasets and muti-modal datasets

We evaluated SCORE performance on a dataset generated by both scRNA-seq and snRNA-seq techniques [33] of 10× Genomics Chromium V2 platform. These two datasets were combined together and scaled separtately. Considering the low-quality of these two datasets, 3000 HVGs were selected for SCORE integration analysis. To integrate human and mouse spermatogenesis scRNA-seq datasets, only human and mouse homologous genes within the PPI network were taken into account. About 8000 genes were selected and the top 12 significant PCs were used to perform UMAP analysis after correction by Harmony.

To assess SCORE performance on scATAC-seq dataset, we downloaded the 10 k PBMC scATAC-seq dataset from 10× official website (https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_pbmc_10k). To accommodate for the molecular network nodes, the peak measurements were first transformed to gene activity matrix by counting the fragments in mapping regions, using the GeneActivity function in package Signac [42]. Based on the gene activity matrix, 5000 HVGs were selected as the input to SCORE, runned with default parameters. With the SCORE output, UMAP was computed on the first 40 PCs of the module activation scores. As the comparison, UMAP was also performed on first 20 PCs of gene activity matrix with 2000 HVGs, denoted as 'Raw'.

Moreover, we also downloaded the PBMC multi-modal dataset of paired scATAC-seq and scRNA-seq profiling from 10× official website (https://cf.10xgenomics.com/samples/cell-arc/1.0.0/pbmc_granulocyte_sorted_10k/pbmc_granulocyte_sorted_10k_filtered_feature_bc_matrix.h5,https://cf.10xgenomics.com/samples/cell-arc/1.0.0/pbmc_granulocyte_sorted_10k/pbmc_granulocyte_sorted_10k_atac_fragments.tsv.gz). The peak measurements of scATA-seq were first transformed to gene activity matrix using the GeneActivity function in package Signac; 8000 HVGs were selected for gene expression and 5000 HVGs for gene activity matrix as the input to SCORE, resulting in two matrices of module activiation scores on gene expression and chromatin accessibility, respectively. We then used first 10 PCs from the SCORE output of scRNA-seq and 20PCs from SCORE output of scATAC-seq to construct the WNN graph [36], using FindMultiModalNeighbors function in Seurat V4.0. UMAP was then conducted based on the WNN. As the comparison, UMAP was also performed on WNN from the first 20 PCs of both raw RNA-seq gene expression and raw ATAC-seq gene activity matrix with 2000 HVGs, denoted as 'Raw'.

Key Points

- The hypothesis of SCORE is fundamentally more biologically reasonable.
- SCORE can integrate scRNA-seq datasets regardless of batch information.
- SCORE is able to integrate multi-platform datasets.
- SCORE outperforms existing methods in accuracy, robustness, scalability and batch effect removal.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

SCORE is freely available in https://github.com/wycwycpku/RSCORE. The five human fetal datasets are available in https://github.com/zorrodong/HECA.

Authors' contributions

F.T., T.L., J.D. and P.Z. conceived the project; Y.G. and L.W. performed the experiments; J.D., P.Z., Y.W., Y.C., H.X., J.L., J.Y. and X.N.Z. conducted the bioinformatics analyses; J.D., P.Z., T.Li and F.T. wrote the manuscript with the help of all the authors. J.D., P.Z., Y.W. and Y.C. contributed equally to this work.

Funding

National Natural Science Foundation of China (31625018 and 81521002 to F.T., 11825102 and 11421101 to T.L.); Beijing Academy of Artificial Intelligence (BAAI) to T.L.

References

1. Regev A, Teichmann SA, Lander ES, *et al*. The human cell atlas. *Elife* 2017;**6**:e27041.
2. Rajewsky N, Almouzni G, Gorski SA, *et al*. LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* 2020;**587**:377–86.
3. Haghverdi L, Lun ATL, Morgan MD, *et al*. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**:421–7.
4. Butler A, Hoffman P, Smibert P, *et al*. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.
5. Korsunsky I, Millard N, Fan J, *et al*. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96.

6. Welch JD, Kozareva V, Ferreira A, *et al*. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**:1873–87 e1817.
7. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* 2019;**37**:685–91.
8. Tran HTN, Ang KS, Chevrier M, *et al*. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**:12.
9. He Z, Brazovskaja A, Ebert S, *et al*. CSS: cluster similarity spectrum integration of single-cell genomics data. *Genome Biol* 2020;**21**:224.
10. Kanton S, Boyle MJ, He Z, *et al*. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* 2019;**574**:418–22.
11. Li H, Courtois ET, Sengupta D, *et al*. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* 2017;**49**:708–18.
12. Aibar S, Gonzalez-Blas CB, Moerman T, *et al*. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6.
13. Fan J, Salathia N, Liu R, *et al*. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods* 2016;**13**:241–4.
14. Dai H, Li L, Zeng T, *et al*. Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res* 2019;**47**:e62.
15. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* 2017;**8**:15599.
16. Ronen J, Akalin A. netSmooth: network-smoothing based imputation for single cell RNA-seq. *F1000Res* 2018;**7**: 8.
17. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;**20**:273–82.
18. Arendt D, Musser JM, Baker CVH, *et al*. The origin and evolution of cell types. *Nat Rev Genet* 2016;**17**: 744–757.
19. Achim K, Arendt D. Structural evolution of cell types by step-wise assembly of cellular modules. *Curr Opin Genet Dev* 2014;**27**:102–8.
20. Hwang FK, Richards DS. Steiner tree problems. *Networks* 1992;**22**:55–89.
21. Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–1902 e1821.
22. Li L, Dong J, Yan L, *et al*. Single-cell RNA-Seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* 2017;**20**:891–2.
23. Cui Y, Zheng Y, Liu X, *et al*. Single-cell transcriptome analysis maps the developmental track of the human heart. *Cell Rep* 2019;**26**:1934–1950 e1935.
24. Wang P, Chen Y, Yong J, *et al*. Dissecting the global dynamic molecular profiles of human fetal kidney development by single-cell RNA sequencing. *Cell Rep* 2018;**24**:3554–3567 e3553.
25. Zhong S, Zhang S, Fan X, *et al*. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 2018;**555**:524–8.
26. Fan X, Dong J, Zhong S, *et al*. Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Res* 2018;**28**:730–45.
27. He S, Wang LH, Liu Y, *et al*. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol* 2020;**21**:294.
28. Travaglini KJ, Nabhan AN, Penland L, *et al*. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* 2020;**587**:619–25.
29. Basha O, Barshir R, Sharon M, *et al*. The TissueNet v.2 database: a quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Res* 2017;**45**:D427–D431.
30. Tian L, Dong X, Freytag S, *et al*. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* 2019;**16**:479–87.
31. Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**:483–6.
32. Camp JG, Sekine K, Gerber T, *et al*. Multilineage communication regulates human liver bud development from pluripotency. *Nature* 2017;**546**:533–8.
33. Mereu E, Lafzi A, Moutinho C, *et al*. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* 2020;**38**:747–55.
34. Chen Y, Zheng Y, Gao Y, *et al*. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res* 2018;**28**:879–96.
35. Wang M, Liu X, Chang G, *et al*. Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis. *Cell Stem Cell* 2018;**23**:599–614 e594.
36. Hao Y, Hao S, Andersen-Nissen E, *et al*. Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87 e3529.
37. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 2016;**5**:2122.
38. Pons P, Latapy M. Computing communities in large networks using random walks. In *Computer and Information Sciences - ISCIS 2005* (eds Yolum, P., Güngör, T., Gürgen, F. & Özturan, C.), Springer, Berlin, Heidelberg, 2005;284–293.
39. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 2017;**27**:491–9.
40. Dobin A, Davis CA, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
41. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
42. Stuart T, Srivastava A, Lareau C, *et al*. Multimodal single-cell chromatin analysis with Signac. bioRxiv 2020;2020.11.09.373613