Scedar: a scalable Python package for single-cell RNA-seq exploratory data analysis

Yuanchao Zhang[1,2], Man S. Kim[1], Erin R. Reichenberger[1], Ben Stear[1], and Deanne M. Taylor[1,3]
1. Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA 19041, USA
2. Department of Genetics, Rutgers University, Piscataway, NJ 08854, USA
3. Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA

Abstract: In single-cell RNA-seq (scRNA-seq) experiments, the number of individual cells has increased exponentially, and the sequencing depth of each cell has decreased significantly. As a result, analyzing scRNA-seq data requires extensive considerations of program efficiency and method selection. In order to reduce the complexity of scRNA-seq data analysis, we present scedar, a scalable Python package for scRNA-seq exploratory data analysis. The package provides a convenient and reliable interface for performing visualization, imputation of gene dropouts, detection of rare transcriptomic profiles, and clustering on large-scale scRNA-seq datasets. The analytical methods are efficient, and they also do not assume that the data follow certain statistical distributions. The package is extensible and modular, which would facilitate the further development of functionalities for future requirements with the open-source development community. The scedar package is distributed under the terms of the MIT license at https://pypi.org/project/scedar.

1 Introduction

Cost-effective large-scale transcriptomic profiling of individual cells is enabled by the development of microfluidic, nanodroplet, and massively parallel sequencing technologies. Using these technologies, single-cell RNA-seq (scRNA-seq) experiments usually generate transcriptomic profiles of thousands to millions of individual cells (Svensson, Vento-Tormo, and Teichmann 2018). Therefore, scRNA-seq has become more commonly used to either study specific biological questions or comprehensively profile certain tissues or organisms (Filbin et al. 2018; Cao et al. 2017; Regev et al. 2017).

Analyses of scRNA-seq datasets therefore require efficient computational programs and sophisticated statistical methods. The programs should be able to manage memory efficiently, exploit multiple cores of the processing units, and handle errors and exceptions gracefully. The

statistical methods must be able to function against high dimensionality, low signal-to-noise ratio, and different characteristics of data generated from different technologies and protocols (Kiselev, Andrews, and Hemberg 2019; Dueck et al. 2016; Ziegenhain et al. 2017). Such requirements can become a barrier between experimental design and biological interpretations of the results.

In order to be scalable, methods have been designed to minimize the usage of hardware resources, so that a large-scale scRNA-seq dataset can be analyzed using a desktop computer, such as Seurat v3.0 (Butler et al. 2018) and Scanpy (Wolf, Angerer, and Theis 2018). Seurat is an R package providing visualization and robust statistical methods to explore and interpret the heterogeneity of the dataset. Scanpy is a Python package providing efficient reimplementations of pre-existing statistical methods and analytical workflows, which can be used to perform general exploratory data analysis and special inferences.

Here, we attempt to achieve the scalability of scRNA-seq data analysis through an alternative approach, which is to enable the user to exploit powerful analytical methods using modern high-performance computing architectures (Zaharia et al. 2016; Dean and Ghemawat 2008; Abadi et al. 2016), such as servers and clusters with large amount of memory, multiple central and graphical processing unit cores, and solid-state drives (SSD) with higher read and write speed than traditional hard disk drive (HDD). Such computational resources have been made easily accessible by cloud computing services like Amazon Web Services and Microsoft Azure. Applying this approach, we designed the analytical methods to be able to run in parallel processes with minimized memory overhead.

Additionally, we also aim to develop robust exploratory data analysis (EDA) methods, so that they could be applied to various datasets generated by different experimental designs, technologies, and protocols. Single-cell RNA-seq datasets have distinct statistical characteristics if generated from different scRNA-seq technologies and platforms (Zappia, Phipson, and Oshlack 2017; Dueck et al. 2016; Ziegenhain et al. 2017), such as SMARTer (Pollen et al. 2014), Drop-seq (Macosko et al. 2015) and 10x Genomics GemCode (Zheng et al. 2017). In order to avoid assumptions on the statistical distributions of the datasets, we incorporated efficient implementations of machine learning methods into the data exploration process. Through extensive exploration, the statistical properties of the data could be observed and used to guide the selection of appropriate methods for biological interpretation (Zappia, Phipson, and Oshlack 2018).

Therefore, we developed a scalable and reliable Python package, **s**ingle-**c**ell **e**xploratory **d**ata analysis for **R**NA-seq (scedar), to facilitate the exploration of large-scale scRNA-seq datasets. Scedar provides analytical routines for visualization, gene dropout imputation, rare transcriptomic profile detection, clustering, and identification of cluster separating genes. The visualization methods are integrated with the efficient scRNA-seq data structures to provide intuitive, convenient, and flexible plotting interfaces. We implemented methods to impute gene dropouts (Supplementary Algorithm S1) and detect rare transcriptomic profiles (Supplementary

Algorithm S2) based on the k-nearest neighbor (KNN) algorithm. The detected rare transcriptomic profiles could be compared with their nearest neighbors in detail to identify rare cell states or types. For clustering analysis, we provide a novel cell clustering algorithm named MIRAC (Algorithm 1), minimum description length (**M**DL) **i**teratively **r**egularized **a**gglomerative **c**lustering. In order to identify genes that are able to distinguish the clusters, we provide a method using a sparsity-aware gradient boosted tree system, XGBoost (Chen and Guestrin 2016).

2 Methods

2.1 scedar package design

We designed scedar in an object-oriented manner for quickly exploring large-scale scRNA-seq transcription level matrices on a remote server utilizing parallel computing techniques, in order to provide a robust and extensible platform for EDA, rather than surpassing the analytical performance of any of ≥ 275 existing scRNA-seq data analysis methods (Zappia, Phipson, and Oshlack 2018). The application programming interface (API) is designed to be intuitive to users familiar with the R programming language. The standard analysis workflow is to explore the dataset, cluster cells, and identify cluster separating genes (**Figure 1**), which is implemented as four main modules: data structure, KNN, clustering and visualization.

The core data structure stores each transcription level matrix as a standalone sparse matrix or full array instance, and it can easily be extended to support customized analytical procedures. The built-in extensions include common procedures like pairwise distance computation, Principal Component Analysis (PCA), t-SNE (Maaten and Hinton 2008), UMAP (McInnes and Healy 2018), and k-nearest neighbor graph (Jacomy et al. 2014). We optimized time and memory efficiency with the following design patterns: parallel computing, lazy loading, caching and copy-on-write.

The KNN and clustering modules utilize the data structure and parallel computing to efficiently perform analytical procedures, and the results are stored in the data structure for further reference.

The visualization module contains plotting methods optimized for large datasets, especially for plotting heatmaps. For example, it takes less than two minutes to generate a heatmap image from a 50,000 x 20,000 matrix containing random standard normal entries.

Preprocessing is implemented as selection and transformation routines of the core data structure, which is not a focus of scedar. The package is designed to identify the necessity of certain preprocessing procedures by extensively exploring the original state of the data. Although preprocessing, such as batch effect correction and normalization, could facilitate the detection of biological variances between cells, applying preprocessing methods correctly

requires careful validation of their assumptions, otherwise they may introduce unwanted bias or variability (Hicks and Irizarry 2015).

2.2 Minimum description length iteratively regulated agglomerative clustering

Minimum description length (MDL) iteratively regulated agglomerative clustering (MIRAC) extends hierarchical agglomerative clustering (HAC) (Müllner 2011) in a divide and conquer manner for scRNA-seq data. Input with raw or dimensionality reduced scRNA-seq data, MIRAC starts with one round of bottom-up HAC to build a tree with optimal linear leaf ordering (Bar-Joseph, Gifford, and Jaakkola 2001), and the tree is then divided into small sub-clusters, which are further merged iteratively into clusters. Because each individual cluster becomes more homogenous with higher number of clusters, the iterative merging process is regularized with the MDL principle (Hansen and Yu 2001). The asymptotic time complexity of the MIRAC algorithm is $O(n^4 + mn^2)$, where n is the number of samples, and m is the number of features. The space complexity is $O(n^2 + mn)$. Relevant mathematical theories and notations of MDL are briefly described in the following section. The pseudo-code of MIRAC is shown in **Algorithm 1**.

input : X, a numeric data matrix of size $n \times m$
$n_{\text{min}}^{\text{cluster}}$, minimum number of rows for a cluster
$r_{\text{min}}^{\text{mdl}}$, minimum ratio of MDL reduction for increasing the number of cluster
output: Clusters of n rows a list of n $\langle$row index, cluster index$\rangle$ pairs

```
// Phase 1: divide n rows into sub-clusters
// build HAC tree with linear optimal leaf ordering
hacTree ← HAClinearOptLeafOrd(X)
// divide the tree leaves into sub-clusters with less than n_min^cluster rows
subClusters ← DivideHACTree (hacTree, n_min^cluster)
// Phase 2: iteratively merge sub-clusters into clusters
// Merge beginning and end sub-clusters to form two minimum sub-clusters
// with more than n_min^cluster rows
MergeSubCls (subClusters, start = 1, minSize =n_min^cluster)
n_sub-clusters ← Length (subClusters)
MergeSubCls (subClusters, start =n_sub-clusters, step = −1, minSize =n_min^cluster)
i ← 2
while i ≤ n_sub-clusters do
    left ← subClusters [i − 1]
    this ← subClusters [i]
    // merge this with left or right until this has ≥ n_min^cluster rows
    while Length (this) < n_min^cluster and i + 1 ≤ n_sub-clusters do
        right ← subClusters [i + 1]
        // minimum merged sub-cluster with ≥ n_min^cluster rows starting from i + 1
        rightMM ← MergeSubCls (subClusters, start = i + 1, minSize =n_min^cluster, inPlace =false)
        if EncodeMDL (left, this) < EncodeMDL (rightMM, this) then // merge left and this
            left ← MergeSubCls (subClusters, selected = [i − 1, i])
            this ← right
        else // merge this and right
            this ← MergeSubCls (subClusters, selected = [i, i + 1])
        end
    end
    // decide whether merge left and this or not
    merged ← Concatenate (left, this)
    mergedMDL ← MDL (merged)
    splitMDL ← MDL (left) + MDL (this) + SplitOverheadMDL (left, this)
    if splitMDL < MaxSplitMDL (mergedMDL, r_min^mdl) then // split
        i ← i + 1
    else // merge
        MergeSubCls (subClusters, selected = [i − 1, i])
    end
end
return subClusters
```

Algorithm 1: Minimum description length iteratively regulated agglomerative clustering

Comparing to HAC, MIRAC is not designed to be faster but rather to improve the cluster robustness. The asymptotic time complexity of MIRAC is the same as HAC with optimal leaf ordering. However, the use of MDL rather than deterministic similarity metrics, improves the noise tolerance by estimating similarity with probabilistic models to give more weight on signal and less weight on noise, assuming that the signal is stronger than noise.

2.2.1 Rationale

The rationale behind MIRAC is to reduce the number of cell partitions that need to be evaluated and find a good one among them.

The number of all possible partitions of a set with n elements is the Bell number, which could be computed with the following recurrence

$$B_n = \sum_{k=0}^{n-1} B_k \binom{n-1}{k},$$

where $n \geq 1$ and $B_0 = 1$ (Wilf 2005).It is computationally intractable to compute the code lengths of B_n partitions, so we reduced the number of cell partitions to evaluate with the following steps:

1. We only evaluate the partitions of n cells that consecutively divide the optimal HAC tree T leaf ordering. The HAC tree with optimal leaf ordering maximizes the sum of similarity measurements between adjacent items in the leaf ordering (Bar-Joseph, Gifford, and Jaakkola 2001), which could be computed with an algorithm developed by Bar-Joseph *et al*. that runs in time O(n^4). The number of partitions that consecutively divide an ordering is the same as the number of compositions of n, which is $C_n = 2^{n-1}$. Because this number still grows exponentially with regard to n, it is necessary to further reduce the set of partitions for evaluation.
2. Within step 1 partitions, we only evaluate those with all clusters having $\geq t$ cells. The number of such partitions is the same as the compositions of n with all summands $\geq t$, which is

   $$C_n^{\{t,t+1,\ldots,n\}} = \sum_{k=1}^{\lfloor n/t \rfloor} \binom{n-k(t-1)-1}{k-1}$$

   given by (Abramson 1976). The growth rate of $C_n^{\{t,t+1,\ldots,n\}}$ with regard to n is smaller, but we still need to reduce it further. For example, $C_n^{\{20,21,\ldots,n\}}$ for n in an ⟨20, 40, 60, 80, 100, 120, 140, 160, 180, 200⟩ are ⟨1, 2, 23, 274, 2695, 24941, 232016, 2184520, 20628613, 194570810⟩, and the values of 2^n are ⟨1048576, 1099511627776, ..., 1.607 x 10^{60}⟩.
3. Within step 2 partitions, we only evaluate the ones that could be generated by merging adjacent clusters of a specific partition $P^s = \{P_1^s, P_2^s, \ldots, P_k^s\}$, where P^s is generated by recursively divide the root HAC tree T until the partitioned subtree has $\leq t-1$

leaves. Thus, $\lceil n/(t-1) \rceil \leq k \leq n$. Let $M(P^s)$ denote the number of step 2 partitions that could be generated by merging adjacent clusters of P^s. The upper bound of $M(P^s)$ is the same as the number of step 2 partitions, which could be reached when $k = n$. The lower bound of $M(P^s)$ is not straightforward, since the merged partitions should have all clusters with $\geq t$ cells.

In order to find a good cell partition P^g in the subset of all possible ones, we iteratively inspect each cluster of P^s and merge it with either its left or right adjacent cluster according to the similarity determined by the two-stage MDL scheme described in the following sections. The procedure has the following steps:

1. Merge P^s clusters in the beginning of the optimal ordering until the merged set contains $\geq t$ cells (**line 9 Algorithm 1**).
2. Similarly, merge P^s clusters at the end of the optimal ordering until the merged set contains $\geq t$ cells (**line 11 Algorithm 1**).
3. For every cluster P_i^s in the middle, determine the similarity between P_i^s and the cluster on the left and right, and merge P_i^s with the more similar cluster (**line 20 - 26 Algorithm 1**). Although the cluster on the left P_{i-1}^s always has $\geq n_{\min}^{\text{cluster}}$ cells, the cluster on the right P_{i+1}^s may have a minimum of one cell, so that the determined similarity between P_i^s and P_{i+1}^s is sensitive to noise. In order to improve the robustness of similarity comparison, instead of determining the similarity between P_i^s and P_{i+1}^s, we determine the similarity between P_i^s and $P_{\text{rmm}}^s = P_{i+1}^s \cup P_{i+2}^s \cup ... \cup P_{i+m}^s$ (**line 20 Algorithm 1**), where $|P_{\text{rmm}}^s| \geq n_{\min}^{\text{cluster}}$, $|P_{\text{rmm}}^s| - P_{i+m}^s < n_{\min}^{\text{cluster}}$, and rmm is the shorthand for right minimax.
4. Once $|P_i^s| \geq n_{\min}^{\text{cluster}}$, determine the similarity between P_{i-1}^s and P_i^s, and merge them if their similarity is above a certain threshold, otherwise split them (**line 32 - 36 Algorithm 1**). The code length of merged $X_{\text{list}(P_{i-1}^s \cup P_i^s)}$ is
$$L_m = L(X_{\text{list}(P_{i-1}^s \cup P_i^s)}) \ .$$
The code length of divided $X_{\text{list}(P_{i-1}^s \cup P_i^s)}$ is
$$L_d = L(X_{\text{list}(P_{i-1}^s \cup P_i^s)}, \{P_{i-1}^s, P_i^s\}) \ .$$
If $L_d - L_m \geq -r_{\min}^{\text{mdl}} \text{abs}(L_m)$, merge P_{i-1}^s and P_i^s, otherwise split them. This conditional statement generalizes the situations where L_m is either non-negative or negative.
5. Continue inspecting the next cluster in P^s.
6. After inspecting all middle clusters, P^g is the final updated P^s.

We use a standard HAC tree to perform MIRAC when the number of samples is larger than ten thousand, which usually results in decreased but still acceptable performances (**Figures 2**). Although the optimal leaf ordering of the HAC tree is an important assumption, its computation

takes too long when the number of samples is large. For example, it takes more than a week to compute the optimal leaf ordering of a dataset with about 68,000 samples. However, the time complexity of computing a good HAC linear ordering could be significantly reduced by implementing other ordering techniques (Aydin, Bateni, and Mirrokni 2016).

2.2.2 Analysis of complexity

The time complexity of MIRAC is $O(n^4 + mn^2)$, where $n, m \geq 1$. The time complexity of computing the HAC tree with optimal leaf ordering is $O(n^4)$ (Bar-Joseph, Gifford, and Jaakkola 2001). The time complexity of finding P^g is $O(mn^2)$, which is briefly analyzed as following.

The time complexity is $O(nm)$ for computing the code length $L(X_{n\times m}, P)$. In computing $L(X_{n\times m}, P)$, we compute the code lengths of each cluster and their labels. Since it takes $O(n)$ time to compute the code length of n observations with a single feature, computing the code length of all clusters and features individually takes $O(nm)$ time, and computing the code length of n labels take $O(n)$ time.

Similarly, the time complexity is $O(n_1 m + n_2 m)$ for computing the code length of $X_{n_1\times m}$ encoded by a model fitted with $X_{n_2\times m}$. Fitting a model on $X_{n_2\times m}$ takes $O(n_2 m)$ amount of time. Using the model to encode $X_{n_1\times m}$ takes $O(n_1 m)$ amount of time.

The time complexity of searching for P^g is $O(mn^2)$. In the worst case scenario, $|P^s| = n$, and the inspecting cluster is always merged with the left hand side cluster in step 4. However, the impact of $n_{\min}^{\text{cluster}}$ and step 3 on the time complexity is not straightforward, so we divide step 3 code length computation into three parts and analyze the upper bound of them individually. The divided three parts are the left cluster, inspecting cluster, and right minimax cluster. Because we keep increasing the size of the left cluster, the upper bound of computing the left cluster code length throughout the execution is $O(mn^2)$. The maximum size of the inspecting cluster is $n_{\min}^{\text{cluster}} - 1$, so the upper bound of computing the inspecting cluster code length is also $O(mn^2)$, when every middle singleton is evaluated $n_{\min}^{\text{cluster}} - 1$ times before merging with the left. The maximum size of the right minimax cluster is $2n_{\min}^{\text{cluster}} - 1$,so the upper bound of evaluating right minimax cluster is also $O(mn^2)$, when every increment of the left cluster takes $n_{\min}^{\text{cluster}}$ times of evaluating the right minimax cluster. Summarizing these three parts, the overall upper bound is $O(mn^2)$.

The space complexity of MIRAC is $O(nm + n^2)$, which could be decomposed into the following three parts. The space complexity of storing the $n \times m$ data matrix X is $O(nm)$. The space

complexity of storing the HAC tree with optimal leaf ordering is $O(n)$. The space complexity of storing the pairwise distance matrix is $O(n^2)$.

2.2.3 Extension with community detection

We extended MIRAC with community detection to improve scalability. We apply MIRAC on a relatively large number of detected single-cell communities to identify final clusters. We used the Leiden algorithm for community detection on KNN graphs (Traag, Waltman, and van Eck 2019). We also provide a KNN graph construction method that supports approximate nearest neighbor (ANN) search using a Hierarchical Navigable Small World (HNSW) graph (Malkov and Yashunin 2016).

2.3 Cluster separating genes identification

We use XGBoost (Chen and Guestrin 2016), a scalable and sparsity-aware boosted tree system, to identify genes that are able to separate a specific set of clusters. This method is designed for data exploration after applying any one of a number of various statistical approaches that have been developed to identify differentially expressed genes (Vallejos, Marioni, and Richardson 2015; Kharchenko, Silberstein, and Scadden 2014; Qiu et al. 2017; Soneson and Robinson 2018; Korthauer et al. 2016), in order to quickly identify genes or sets of genes that are able to separate an arbitrary set of clusters for further inspection.

Rather than providing a meticulous p-value for each gene among the compared clusters, we rank the genes by their importances on separating the clusters under comparison. The importance of a gene in the trained classifier is the number of times it has been used as an inner node of the decision trees. We use cross validation to train an XGBoost classifier on the compared clusters, and the classifier is essentially a bag of decision trees (Chen and Guestrin 2016). In order to alleviate the influences of stochasticity on interpretation, we use bootstrap with feature shuffling to better estimate the importance of genes in separating the compared clusters. The obtained list of important genes could be further explored by inspecting transcription level fold changes and decision tree structures.

Comparing to NSForest (Aevermann et al. 2018), a method based on random forest (Breiman 2001; Pedregosa et al. 2011) to identify a parsimonious set of cluster separating genes from scRNA-seq data, our method identifies all possible cluster separating genes using gradient boosting. Practically, NSForest version 1.3 is distributed on GitHub as a Python script without encapsulation or testing (checked on Oct 11, 2018), but our method is distributed through the Python Package Index, comprehensively tested, and easy to use through the user friendly API. With regard to scalability, our method uses a scalable implementation of gradient boosting algorithm (Chen and Guestrin 2016), whereas NSForest uses the implementation of random forest in scikit-learn (Pedregosa et al. 2011).

2.4 K-nearest neighbor methods

A k-nearest neighbor analytical strategy exploits the similarity between data points for classification, regression, and imputation (Cover and Hart 2006; Bezdek, Chuah, and Leep 1986). In k-nearest neighbor methods, samples are considered as points in a space with each dimension representing a measured property, which is often referred to as a feature, of the samples. The similarity between samples can be evaluated with various distance metrics, which is extensively reviewed by Bellet *et al*. (Bellet, Habrard, and Sebban 2013). Generally, a distance metric is a function that takes two samples and output a numeric value to represent the distance between the two samples in their feature space. For example, the Euclidean distance metric is the following function,

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2},$$

where the p and q are two samples, and q_i and p_i are the data values on their ith dimension. The k-nearest neighbors of a sample are the k number of other samples that have the smallest distances from the sample. The k-nearest neighbors of a sample are informative, due to their similarity, to determine the category of the sample in classification, the relevant continuous property in regression, and the missing values in imputation.

We developed two methods based on the KNN algorithm to facilitate the exploration of scRNA-seq datasets. With relative large number of cells profiled in each scRNA-seq experiment, we assume that each one of the non-rare cells is similar to at least k other cells in their transcriptomic profiles. With this assumption, we impute gene dropouts and detect rare transcriptomic profiles. In the scedar implementation, we also support approximate nearest neighbor (ANN) search using Hierarchical Navigable Small World (HNSW) graph (Malkov and Yashunin 2016), which greatly improves the scalability of KNN graph construction.

2.4.1 Impute gene dropouts

In an scRNA-seq experiment, if a truly expressed gene is not detected in a cell, the gene is considered a “dropout”, and such events are called gene dropouts (Kharchenko, Silberstein, and Scadden 2014). Gene dropouts may be caused by biological and technical reasons (Kharchenko, Silberstein, and Scadden 2014; Risso et al. 2018; Pierson and Yau 2015). The rationale behind the possible causes of biological dropouts mainly involves transcriptional bursting (Suter et al. 2011; Tantale et al. 2016) and RNA degradation. With regard to technical dropouts, the main concerns are the relatively small number of RNA transcripts of a gene, amplification efficiency, and batch effect (Zappia, Phipson, and Oshlack 2017).

We exploit the transcriptomic profiles of the k-nearest neighbors of a cell to impute the gene dropouts in the cell (**Algorithm S1**). The algorithm could take multiple iterations, so that the dropped-out genes that are expressed in all k-nearest neighbors could be imputed at first, and

the ones that are expressed in most but not all of k-nearest neighbors could be imputed in the following iterations.

2.4.2 Detecting rare transcriptomic profiles

We mark transcriptomic profiles as rare if they are distinct from their k-nearest neighbors, according to the pairwise similarity between cells (**Algorithm S2**). The algorithm could take multiple iterations, so that the most distinct transcriptomic profiles could be marked at first and less distinct ones in the following iterations.

This method is provided mainly to facilitate detailed inspection of rare transcriptomic profiles rather than removing outliers from the data. Because rare transcriptomic profiles may have various biological and technical causes, samples and features in a dataset should only be removed after extensive exploratory data analysis and rigorous reasoning with domain specific knowledge. Closely comparing rare transcriptomic profiles with their nearest neighbors may also yield insights into their biological differences, which may further facilitate the identification of rare cell types and states.

2.5 Benchmark

We benchmarked the clustering and KNN performances of scedar on simulated and experimental scRNA-seq datasets. We obtained previously published experimental scRNA-seq datasets (**Table 1**). We also generated 50 simulated scRNA-seq read count datasets with Splatter (Zappia, Phipson, and Oshlack 2017). The simulation parameters are estimated by Splatter according to a Drop-seq dataset (Macosko et al. 2015) and a 10x Genomics GemCode dataset (Zheng et al. 2017). Within each simulated dataset, the cells have 8 clusters taking approximately the following proportions: ⟨0.3, 0.2, 0.15, 0.15, 0.05, 0.05, 0.05, 0.05⟩, with a gene dropout rate around 5%.

Table 1. Real scRNA-seq datasets for benchmark

Publication	# cells	# genes	Organism	Tissue	Protocol	Raw Data Type
Deng *et al.* (2014)	268	22,431	Mus Musculus	Embryo	Smart-Seq/Smart-Seq2	Read count
Pollen *et al.* (2014)	301	23,730	Homo sapiens	Dermal, blood, pluripotent and neural	SMARTer	TPM
Kolodziejczyk *et al*. (2015)	704	38,653	Mus Musculus	Embryonic stem cell	SMARTer	Read count
Zeisel *et al.* (2015)	3005	19,972	Mus Musculus	Brain	STRT-Seq UMI	Read count
Macosko *et al*. (2015)	44,808	23,288	Mus Musculus	Retina	Drop-seq	Read count
Zheng *et al*. (2015)	68,579	33,694	Homo sapiens	Peripheral blood mononuclear cell	GemCode	Read count
Han *et al*. (2018)	405,191	39,855	Mus Musculus	Various tissues	Microwell-Seq	Read count
Cao *et al*. (2019)	2,058,652	26,183	Mus Musculus	Embryo	sci-RNA-seq3	Read count

Source: https://hemberg-lab.github.io/scRNA.seq.datasets/. TPM represents Transcripts Per Million in the raw data type column.

We performed all benchmark analyses on a high-performance computing cluster, of which the computing resources are strictly managed by Univa Grid Engine. The cluster nodes have CPUs of Intel Xeon E5-2680 v3 or Intel Xeon E7-8880 v3. The memory sizes are either 128GB, 256GB, or 1TB. When scheduling analytical jobs for benchmarking, we make sure that the number of cores and allocated memory are enough for the program.

2.5.1 Clustering

The clustering accuracy and stability of MIRAC were benchmarked together with several other clustering methods on experimental scRNA-seq datasets listed in **Table 1**.

The following clustering methods are directly applied on the original data without preprocessing.

- MIRAC on 2D t-SNE projection.
- Single-cell consensus clustering (SC3) version 1.7.7 (Kiselev et al. 2017). SC3 is selected for comparison because it was extensively compared with other methods on different experimental datasets.

- K-means clustering on 2D t-SNE projection.
- Hierarchical agglomerative clustering on 2D t-SNE projection.
- Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al. 1996) on 2D t-SNE projection.
- Community clustering on KNN graph constructed using the Euclidean distances of 100 principal components. We used the Leiden algorithm for community detection (Traag, Waltman, and van Eck 2019).

Although MIRAC could be directly applied on the expression matrix, dimensionality reduction is able to improve the performance of similarity and density based clustering methods when the number of features is high (Aggarwal, Hinneburg, and Keim 2001). The mathematical influences of the high number of features are briefly described in the previous sections.

Although t-SNE projections are stochastic and influenced by the perplexity parameter, t-SNE is proved to be able to recover well-separated clusters (Linderman and Steinerberger 2019)and t-SNE has been extensively used as dimensionality reduction method for scRNA-seq data (Macosko et al. 2015; Cao et al. 2017). Thus we chose to use t-SNE for r demonstration in this report, but we note that scedar also supports the use of PCA and UMAP for MIRAC clustering, which can be applied just as easily as the t-SNE method.

When benchmarking for accuracy, we cluster the cells in each experimental dataset using the compared clustering methods with a grid of parameters. We use the maximum similarities between the clustering results and the cell types from the publications in order to compare the accuracy of different clustering methods. Although taking the maximum increases the chance of overfitting, it resembles the procedure of clustering analysis in practice.

We also recorded the runtime of clustering methods on different datasets. In **Figure 2A**, SC3 was performed with 20 cores, and community clustering is performed with 60 cores. Other clustering methods were performed with a single core. When running MIRAC, we did not require the hierarchical tree to be optimal. In **Figure S8C**, the community-detection-extended MIRAC clustering method was performed with 25 CPU cores, and other methods were performed with 60 CPU cores.

We also benchmarked the influences of t-SNE random states on the clustering results (Figure S1). T-SNE embeddings generated with different random states may be distinct from each other, although current mathematical results have shown that clustering using t-SNE embeddings is able to recover well separated clusters (Linderman and Steinerberger 2019). We characterized the stability to random states of clustering methods by running them on each experimental dataset with the same parameters but ten different random states. The similarity of clustering results between different random states are used to compare the stability of different clustering methods. The stability of community clustering is not characterized, because it is based on PCA.

2.5.2 Cluster similarity metrics

We use two cluster similarity metrics, cluster consistent ratio (CCR) and adjusted Rand index (ARI) (Hubert and Arabie 1985) for measuring the accuracy and stability of clustering methods respectively. When we have a coarse reference partition P_r and a finer clustering partition P_c, the CCR is computed as the ratio of pairs within each cluster of P_c that are also in the same cluster of P_r, with the number of clusters kept the same across compared methods. The ARI is computed with the Python package scikit-learn (Pedregosa et al. 2011) using the mathematical formula given by Hubert and Arabie (Hubert and Arabie 1985).

The reference partitions P_r of real datasets are obtained from their original publications (**Table S1**). The clusters in Deng *et al*. dataset (Deng et al. 2014) are experimentally determined by manual single cell isolation. The clusters in Pollen *et al.* (Pollen et al. 2014) are experimentally determined by the cell line or tissue type of the isolated single cells. The clusters in Kolodziejczyk *et al*. (Kolodziejczyk et al. 2015) are determined by the embryonic stem cell culturing conditions and batches. The clusters in Zeisel *et al.* (Zeisel et al. 2015) are determined by the BackSPIN clustering method (Zeisel et al. 2015) and further inspected with domain specific knowledge.

We choose CCR to measure clustering accuracy rather than ARI, because ARI greatly penalizes the split of a large cluster in P_r into multiple smaller ones in P_c. This behavior of ARI could prevent the evaluation of transcriptomic variabilities within a large group of cells in clustering analysis, which is an important goal of scRNA-seq experiments that cannot be easily achieved by bulk RNA-seq experiments. In addition, we also provide a method in scedar to easily merge multiple clusters together, in case the users found the sub-types of a cell type are very similar to each other.

However, we used ARI to measure clustering stability rather than CCR, because the differences between P_r and P_c are completely caused by different random states, hence splitting a cluster in P_r should be penalized.

MDL is not used as a cluster similarity metric, even though MDL is used in MIRAC to guide the process of finding good partitions. MDL is only used to guide and regularize the merging process of local adjacent sub-clusters by evaluating their similarity, but it is not used as an objective to be optimized globally. Using MDL as a benchmarking metric would introduce a bias as among the various methods, MIRAC is the only method using MDL. Moreover, interpreting relative global MDL differences is not straightforward as it is not only affected by cluster labels, but also transcription levels.

2.5.3 Detection of rare transcriptomic profiles

We visualize real datasets before and after removing rare transcriptomic profiles in t-SNE projection and pairwise distance heatmaps, without quantitative evaluations like receiver operating characteristic (ROC) curve, since rare transcriptomic profiles are not well defined with a large number of genes (Aggarwal, Hinneburg, and Keim 2001). Approaches to identify rare data points in low-dimensional spaces (≤ 3) do not scale well to higher dimensions due to the exponentially decreased density of data points in space and increased instability of distances, which is elaborated in the previous section about mathematical theories on high-dimensional data analysis. In **Figure S8B**, KNN rare transcriptomic profile detection is performed with 25 CPU cores.

It is important to note that rare transcriptomic profiles are detected to facilitate detailed inspection rather than the removal of them from the data. The visualizations are only used to illustrate the capability of the KNN method for detecting rare transcriptomic profiles.

2.5.4 Gene dropout imputation

We simulate gene dropouts using Splatter (Zappia, Phipson, and Oshlack 2017) to obtain a dropout rate around 5%. Then, we benchmark the performance of imputing gene dropout as two parts: detection and smoothing. On simulated data, because the true gene dropouts are known, we use a ROC curve and mean squared errors (MSEs) to characterize the performance of gene dropout detection and smoothing respectively. In **Figure 3A, S7A, and S8A**, all imputation methods were executed with a single core. On real data, we visualize the cells in 2D t-SNE space before and after imputation. The compared methods are KNN gene dropout imputation (KNNGDI) version 0.1.5, SAVER version 1.0.0 (Huang et al. 2018), MAGIC version 1.1.0 (van Dijk et al. 2018), and scImpute version 0.0.6 (Li and Li 2017).

2.5.5 Method speed-up by parallel processing

We characterized the analysis speed-up by parallel processing on different datasets (**Figure S9**). KNN dropout imputation is not evaluated on larger datasets, because the speed limiting step has not yet been parallized. In addition to parallelizing the computation in the implemented methods, we also provide an easy-to-use utility function (scedar.utils.parmap) to parallely run analytical methods with different parameters, which would significantly facilitate exploratory parameter search.

3 Results

3.1 Basic workflow of scedar

We illustrate the basic workflow of using scedar for scRNA-seq exploratory data analysis with the dataset published by Zeisel *et al.* (Zeisel et al. 2015) (**Figure 1**). The dataset contains the RNA unique molecule identifier (UMI) counts of 19,972 genes in 3005 cells from the mouse brain. We selected this dataset for demonstration because it could be clearly visualized in small graphs, although our package is capable of analyzing much larger datasets with over 2 million cells (**Figures S2**, **S3**, **S4** and **S5**).

In **Figure 1**, each box represents a data analysis step, and they are consecutively executed according to the arrow. The purpose and runtime are listed in the upper ribbon. The code that is essential to the step is listed in the box, and their results are also shown.

The preparation step imports required packages and loads the data. The class SampleDistanceMatrix is one of the core data structures in the package that is used to store the read counts and pairwise distances of an scRNA-seq dataset. Because the pairwise distance computation is delayed until necessary, i.e. lazily loaded, this step only takes 12 seconds. We use cosine distance rather than correlation distance to greatly speed up the computation, since we implemented the computation procedure of pairwise cosine distances completely with numpy linear algebra operations with OpenBLAS backend (Xianyi, Qian, and Yunquan 2012; Wang et al. 2013).

The t-Distributed Stochastic Neighbor Embedding (t-SNE) scatter plot and KNN graph are used to explore the dataset. The cell type labels published by Zeisel *et al.* (Zeisel et al. 2015) are truncated to fit in the space. We also provide methods to visualize arbitrary statistics, e.g. number of expressed genes, of individual cells as color gradient. The layouts of cells in t-SNE and KNN graph are similar to each other. Although KNN graph is faster than t-SNE, the runtime for t-SNE could be reduced by optimizing its parameters, e.g. lowering the number of iterations.

The MIRAC step clusters the cells and visualizes them with t-SNE scatter plot and pairwise distance matrix heatmap. The heatmap generation procedure in scedar is optimized for large-scale datasets, which is able to generate a heatmap with tens of thousands of columns and rows in a few minutes. Users could also generate heatmaps for the read count matrix to directly inspect the sparsity of datasets (**Figures S2B**, **S3B**, **S4B** and **S5B**).

The last step identifies cluster separating genes with XGBoost (Chen and Guestrin 2016). Users could choose an arbitrary set of clusters to compare, and the genes are ranked by their importance in separating the clusters. Then, the read counts of a gene across clustered labels could easily be visualized by t-SNE scatter plot.

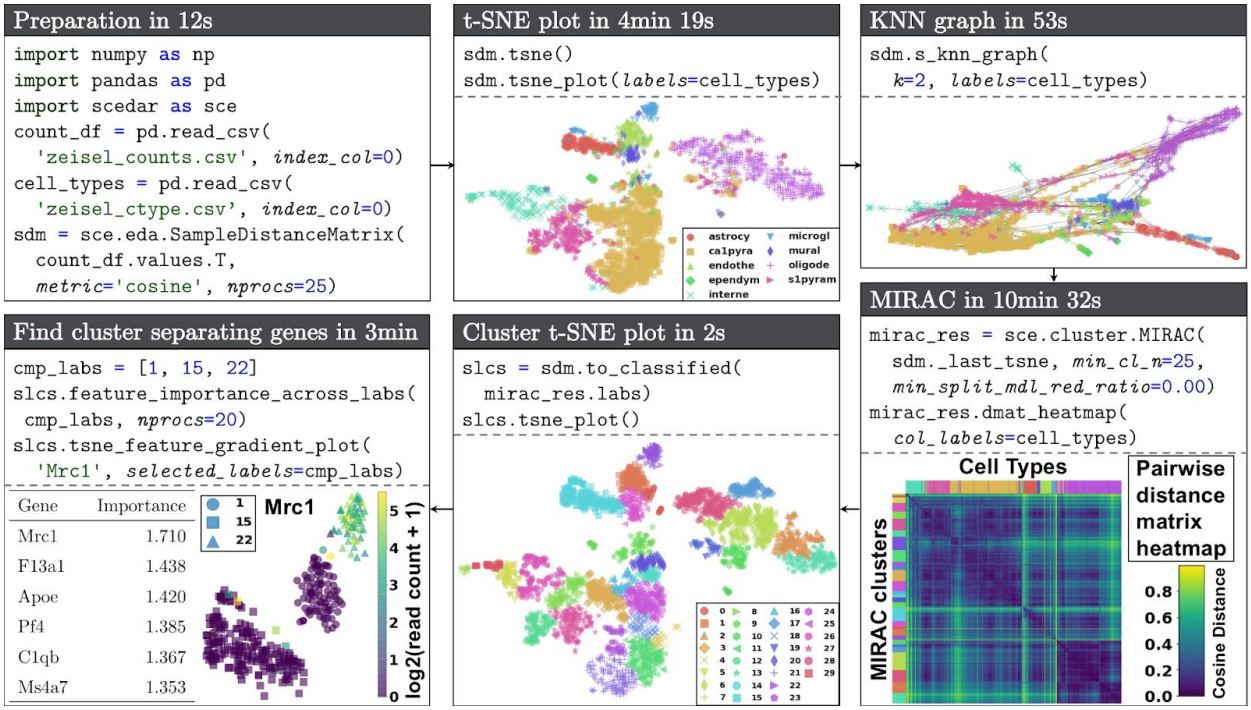

Figure 1. Demo of scedar. Workflow of using scedar to analyze an scRNA-seq dataset with 3005 mouse brain cells and 19,972 genes generated using the STRT-Seq UMI protocol by Zeisel *et al.* (Zeisel et al. 2015). Procedures and parameters that are not directly related to data analysis are omitted. The full version of the demo is available at https://github.com/logstar/scedar/tree/master/docs/notebooks.

3.2 Performance of MIRAC clustering

We benchmarked several clustering methods on the datasets listed in **Table 1** (**Figure 2**). Each dataset is clustered multiple times with each clustering method to obtain different numbers of clusters (**Table S1**).

The t-SNE based clustering methods are faster than SC3 (**Figure 2A**). Also, the t-SNE based clustering methods have similar runtimes (**Figure 2A)**, since the time limiting step is the computation of t-SNE projection, when optimal hierarchical clustering ordering is not required in MIRAC. When the optimal ordering is required, the time limiting step is the computation of optimal ordering (**Figure 1**). With regard to practical usage, we recommend the users to explore different parameters of MIRAC without optimal ordering. Once appropriate parameters have been identified, users can perform MIRAC with optimal ordering. However, for datasets with more than 10,000 cells, we recommend that users perform MIRAC without optimal ordering (**Figure S2C and S3C**). For datasets with more than 100,000 cells, we recommend that users perform community-detection-extended MIRAC without optimal ordering (**Figure S8C**), which

takes for instance 2885.5 ± 176.0 (mean ± standard deviation) seconds on the mouse organogenesis cell atlas (MOCA) dataset with 2,058,652 single cells using 25 CPU cores (**Figure S8C**) (Cao et al. 2019). We also performed clustering with the community clustering methods implemented in scedar, Seurat (Butler et al. 2018) and Scanpy (Wolf, Angerer, and Theis 2018) on relatively large scRNA-seq datasets with more than 10,000 cells (**Figure S8C**). The runtime of community clustering methods Seurat and Scanpy is comparable to community clustering and community-detection-extended MIRAC (**Figure S8C**), except that Scanpy and Seurat were not able to cluster the MOCA dataset that contain 2,058,652 single cells on a server with 1TB memory due to a mandatory conversion of the sparse read count matrix into dense matrix.

The cluster consistent ratios (CCRs) of t-SNE based clustering methods are comparable to SC3 (**Figure 2B**). The PCA based community clustering results showed relatively lower CCRs than other clustering methods (**Figure 2B**), which might due to the non-optimal representation of similarities between high-dimensional single cell read counts in the linear PCA space, comparing to t-SNE embeddings that are optimized to capture the similarities between high-dimensional single cell read counts. The representative MIRAC clustering results of Zheng *et al.* (Zheng et al. 2017) and Macosko *et al.* (Macosko et al. 2015) datasets are visualized with t-SNE scatter plots and pairwise distance matrix heatmaps (**Figures S2C** and **S3C**). For smaller datasets, the representative MIRAC results are shown in **Figure S6**. Although t-SNE projections obtained with different random states are distinct from each other, the consistency of clustering results is comparable to SC3 (**Figure S1**).

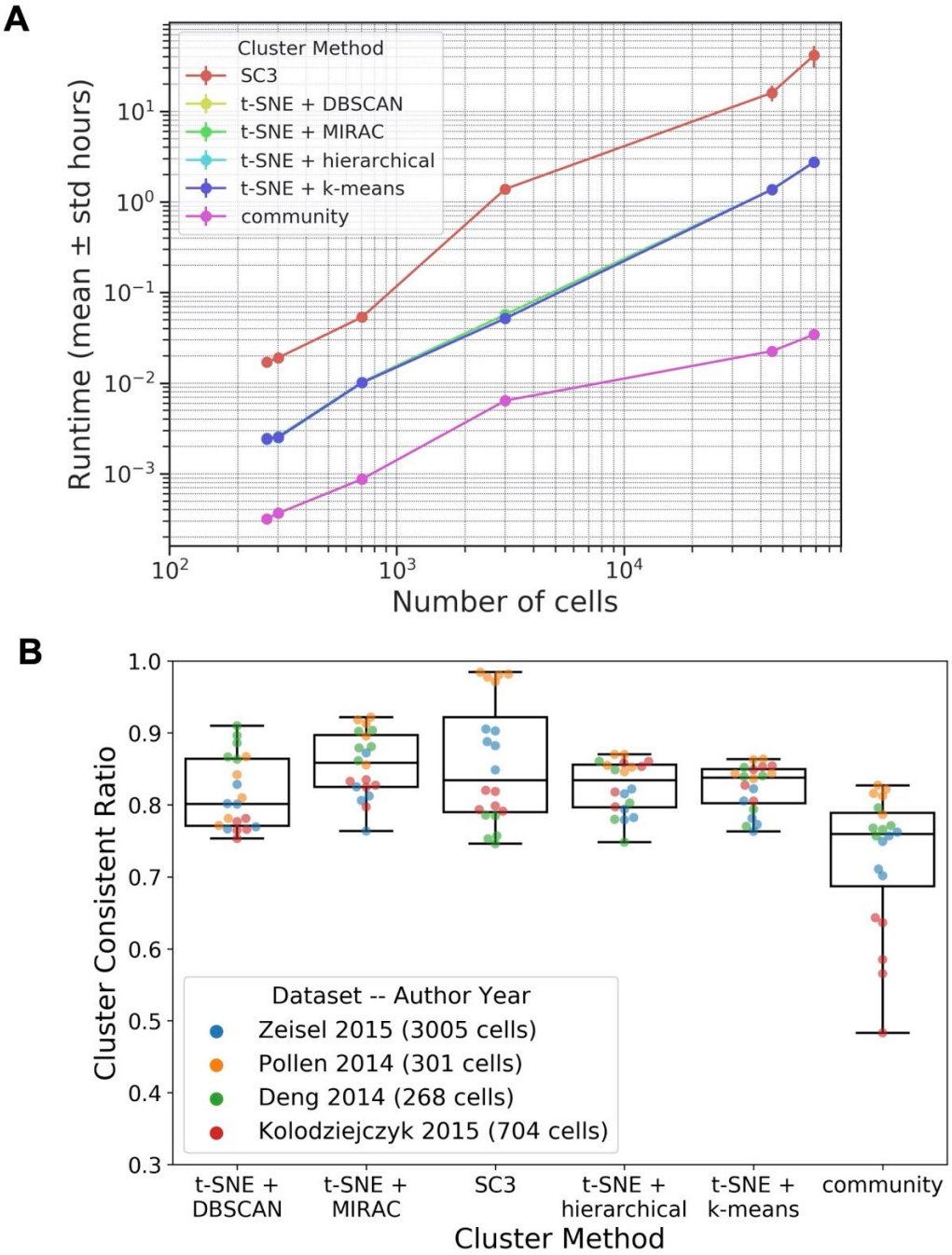

Figure 2. Clustering method benchmarks on experimental datasets. (**A**) Runtimes. (**B**) CCRs on different datasets, with different points of each dataset representing different numbers of clusters. For each dataset, the numbers of clusters are the same across all compared clustering methods.

3.3 Performance of imputing gene dropouts

We benchmarked several gene dropout imputation methods on the simulated 10x Genomics (**Figure 3**) and Drop-seq (**Figure S7**) datasets. K-nearest neighbor gene dropout imputation (KNNGDI) is faster than other compared methods on relatively small datasets (**Figure 3A**) and is capable of handling scRNA-seq datasets with over 10,000 cells (**Figure S8A**). The runtime of KNNDGI on 2,058,652 cells is 71.56 ± 1.71 (mean ± standard deviation) hours.

The performance of KNNGDI on detecting gene dropouts is comparable to SAVER and better than scImpute and MAGIC. The ROC curve of scImpute sharply turns around TPR = 0.25 (**Figure 3B** and **S7B**) because its threshold parameters, determining whether a zero entry is a dropout or not, are not sensitive enough to achieve any higher TPRs. Although the AUCs of KNNGDI and SAVER are higher than scImpute and MAGIC, they all have comparable performances when FPRs are lower than 0.05 (**Figure 3B**), except that MAGIC has worse performance on the simulated Drop-seq datasets that have higher sparsity than the Zeisel *et al.* (Zeisel et al. 2015) dataset (**Figure S7B**).

The MSEs of KNNGDI on correcting gene dropouts are comparable to SAVER and smaller than scImpute and MAGIC (**Figure S7C** and **S7D**). However, these methods all have MSEs many times higher than the MSEs of the observed counts to the true counts, which implies that these methods all introduced many times more reads than the true dropout reads. Despite different levels of MSEs, the t-SNE embeddings of the imputed read counts are consistent with the t-SNE embeddings of original read counts (**Figure 3C**, **S2E**, **S3E, S4E and S5E**).

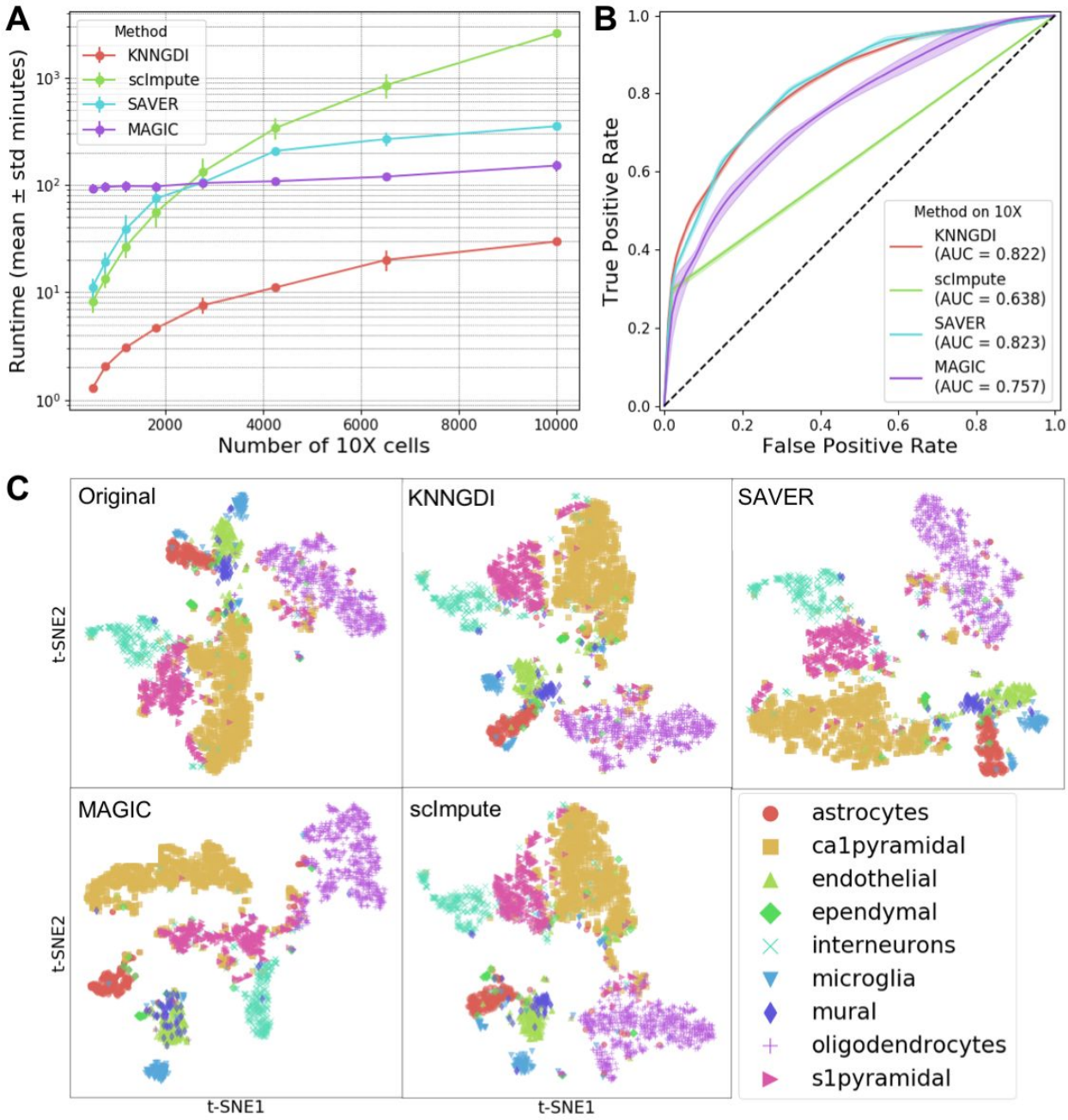

Figure 3. Gene dropout imputation method benchmarks. (**A**) Runtimes on 40 simulated 10x Genomics datasets. (**B**) ROC curves (± standard deviation) of dropout detection on the simulated 10x Genomics datasets. (**C**) t-SNE scatter plots of the Zeisel *et al*. (Zeisel et al. 2015) dataset after gene dropout imputations.

3.4 Performance of detecting rare transcriptomic profiles

We detected rare transcriptomic profiles in datasets listed in **Table 1** with the KNN method (**Figure 4, S2F**, **S3F**, **S4F**, **S5F** and **S6**). Although the detection method has many limitations,

rare transcriptomic profiles tend to be points on the t-SNE scatter plots either far away from the majority of their same types or along the edges of a group of agglomerated points. On the pairwise distance matrix heatmap, rare transcriptomic profiles tend to be small chunks that are distinct from their neighbors, and the heatmap becomes smoother after removing the rare transcriptomic profiles. The detected rare transcriptomic profiles could be further inspected as potential rare cell types and states by comparing with their nearest neighbors. The KNN rare transcriptomic profile detection method is also capable of handling scRNA-seq datasets with over 10,000 cells (**Figure S8B**), with a runtime of 22.14 ± 0.87 (mean ± standard deviation) minutes on 2,058,652 cells.

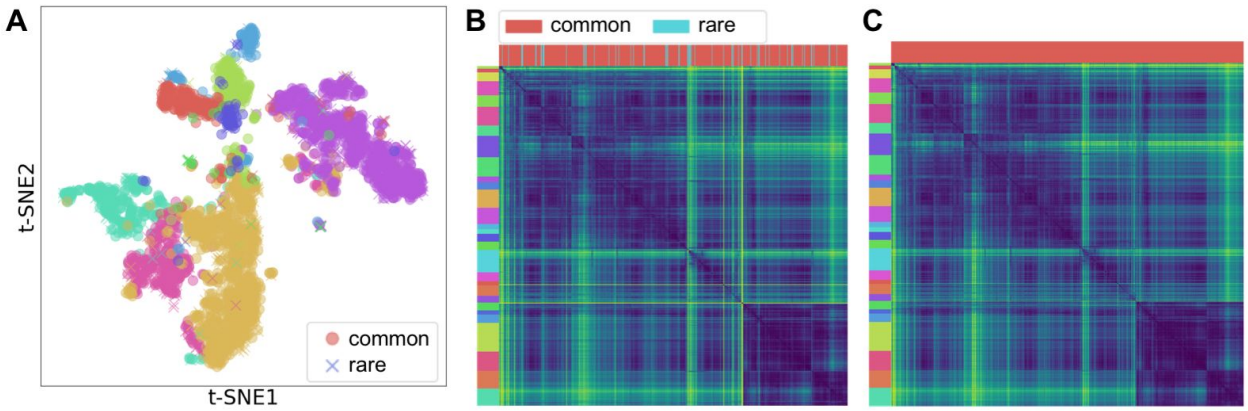

Figure 4. KNN rare transcriptomic profile detection on the Zeisel *et al*. (Zeisel et al. 2015) dataset. (**A**) t-SNE scatter plot with colors labeling cell types and markers labeling common or rare transcriptomic profiles. 9.3% cells are marked as rare. (**B**) Pairwise cosine distance heatmap with left strip as MIRAC labels and upper strip as common or rare transcriptomic profiles labels. (**C**) Pairwise cosine distance heatmap with rare transcriptomic profiles removed.

3.5 Identification of cluster separating genes

We used scedar to identify the genes distinguishing the MIRAC cluster 1, 15, and 22 of the Zeisel *et al*. (Zeisel et al. 2015) dataset (**Figure 5** and **Table S2**). In the original publication, the upper to lower MIRAC clusters 22, 1, and 15 in the t-SNE scatter plot are assigned to microglia, endothelial cells, and astrocytes respectively (**Figure 3C**). We choose these three clusters to inspect one of the discrepancies between cell types and MIRAC clustering results, where the small isolated upper part of the MIRAC cluster 15 is assigned to microglia instead of endothelial cells in the original publication.

The smaller upper isolated part of MIRAC cluster 15 might be a distinct cell sub-type of microglia. Although it expresses a microglia marker gene C1qb (**Figure 5B**) (Beutner et al. 2013), it does not express Mrc1 or Apoe in the same pattern as the MIRAC cluster 22 (**Figure 5B** and **5C**). According to the transcription levels of the cluster separating genes (**Figure 5C**), the smaller upper isolated part of MIRAC cluster 15, which is located at the top of the cluster 15

rows, has some genes expressed in the same pattern as the microglia, but it also has some other genes expressed distinctly from the microglia.

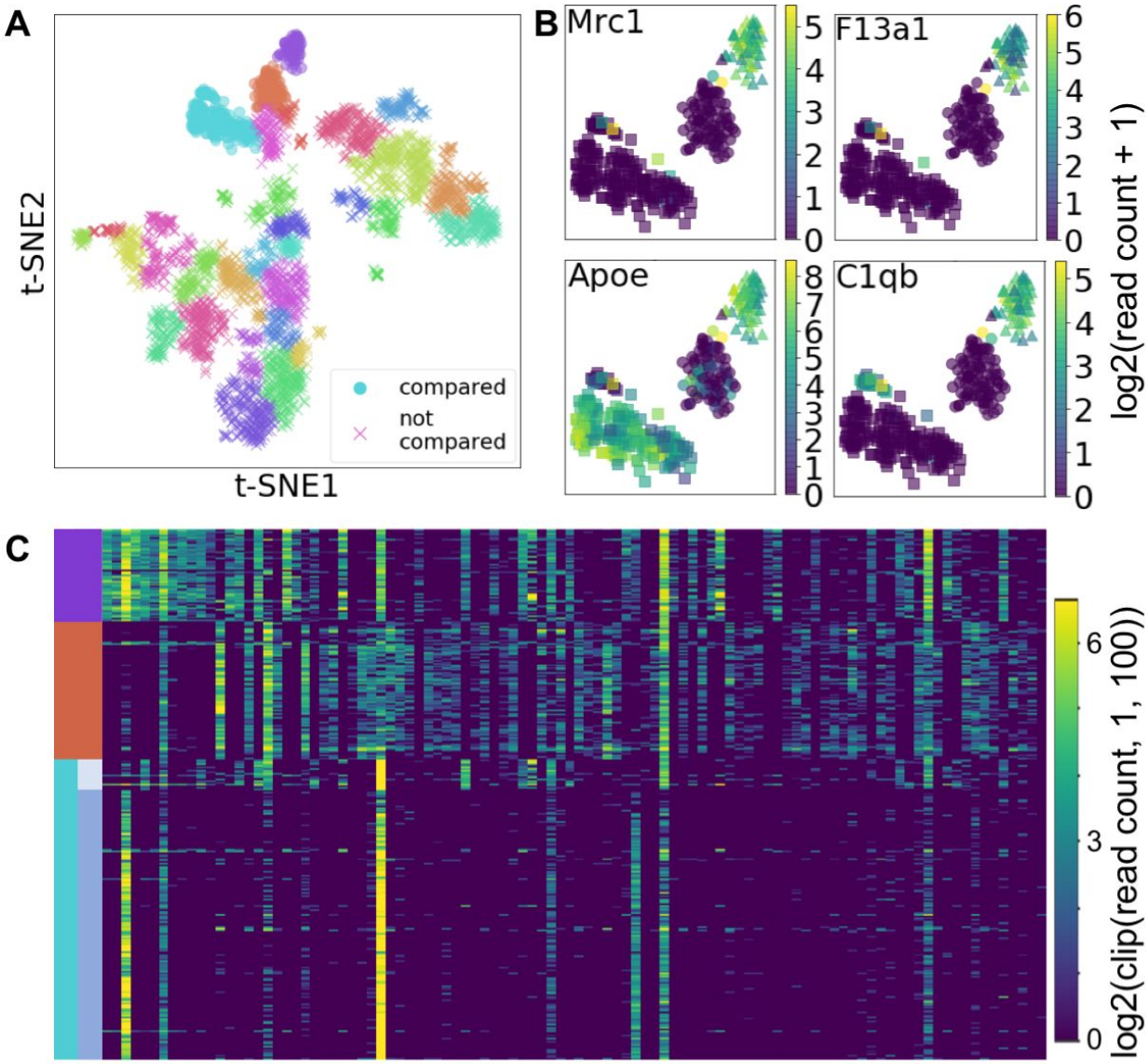

Figure 5. Identified genes separating the MIRAC clusters 1, 15, and 22 of the Zeisel *et al*. (Zeisel et al. 2015) dataset. (**A**) t-SNE scatter plot with color as MIRAC cluster labels and marker shape as compared or not compared. (**B**) t-SNE scatter plots of the compared clusters with color as $\log_2$(read count + 1) of the corresponding gene and marker shape as MIRAC clusters. (**C**) Transcription level heatmap of the top 100 important cluster separating genes in the compared cells, with rows as cells ordered by cluster labels and columns as genes ordered by importance. The color gradient is $\log_2$(clip(read count, 1, 100)), where the clip(read count, 1, 100) function changes any read count below 1 to 1 and above 100 to 100, in order to better compare genes at different transcription levels.

4 Discussion

Comprehensive profiling of the transcriptomes of individual cells within an organism or tissue by scRNA-seq is able to facilitate the systematic study of physiological or pathological states (Villani et al. 2017; Baslan and Hicks 2017; Cao et al. 2017). Previous scRNA-seq experiments identified novel cell types or states (Villani et al. 2017), obtained insights into the regulation of differentiation process (Cao et al. 2017; Rizvi et al. 2017; Cao et al. 2019; Schiebinger et al. 2019), and inferred molecular mechanisms of tissue functions (Zeisel et al. 2015; Lake et al. 2018; Haber et al. 2017).

The biological results of scRNA-seq experiments are obtained from extensive data analyses, which could take more time than doing the experiments. As the size of scRNA-seq datasets increases rapidly (Svensson, Vento-Tormo, and Teichmann 2018; Rozenblatt-Rosen et al. 2017), computational methods also need to be scalable by exploiting hardware and software techniques for big data analytics, in order to complete an analysis within a reasonable amount of time.

Scedar is able to facilitate gene dropout imputation, rare transcriptomic profile detection, clustering, and identification of cluster separating genes for scRNA-seq data by exploiting scalable system design patterns and high-performance computing architectures. We parallelized time-consuming computational procedures by multi-processing without excessive copies of shared data in memory. We also decompose the whole analytical procedure into multiple steps, so that certain steps could be specifically optimized without repeating others. In addition, intermediate results, such as pairwise distances and t-SNE projections, are lazy loaded and cached in a unified data structure to speed up analytical routines, prevent repeated computations, and alleviate the burden on users to keep track of all intermediate results.

Comparing to other computational tools that were developed or updated for large scale scRNA-seq data analysis, like PAGODA2 (Fan et al. 2016), Seurat v2.0 (Butler et al. 2018), and Scanpy (Wolf, Angerer, and Theis 2018), scedar distinguishes itself with an additional research focus on developing new analytical methods, including those based on machine learning. In scedar, we adapted KNN, a typical machine learning algorithm, to impute gene dropouts and detect rare transcriptomic profiles. MDL principle, an important concept in computational learning, is applied to cluster single cells (Hansen and Yu 2001). A scalable and sparsity-aware gradient boosted tree system XGBoost (Chen and Guestrin 2016), which implements a typical machine learning algorithm, is used to identify genes that are able to distinguish different clusters. In addition, these machine learning methods are able to exploit modern high-performance computing architecture, which improves the scalability of the package.

In scedar, we developed a clustering algorithm, MIRAC, for scRNA-seq data. MIRAC clusters observations in three steps: 1) build a tree by hierarchical clustering, 2) divide the tree into sub-trees, and 3) merge similar sub-trees into individual clusters. This clustering strategy adapts the basic ideas of BIRCH (Zhang, Ramakrishnan, and Livny 1996) and BackSPIN (Zeisel et al. 2015). Instead of building a balanced tree structure, like the clustering feature tree in BIRCH for further partitioning, MIRAC divides the tree structure built by hierarchical clustering optionally in a balanced manner (**Figure S10** and Supplementary Method section 1.4), which simplifies the clustering procedure. In BackSPIN, a sorted pairwise correlation matrix is recursively bi-partitioned into clusters according to criteria based on the normalized sums of correlation coefficients. In contrast, MIRAC iteratively merges the relatively small sub-clusters according to criteria based on the MDL principle, which increases the robustness for determining whether two groups of observations should be put in the same cluster or not. Especially, when the number of observations is large within a group, finding an optimal bi-partition is not straightforward, since there may be multiple distinct sub-groups. Although the performance of MIRAC under certain metrics is comparable to other clustering algorithms, it is able to provide distinct clusters that are sensitive to local structures, which could be used as alternative perspectives to interpret the source of heterogeneity within the dataset.

There are still many possible improvements on scedar. To improve the scalability of MIRAC, we could provide more efficient methods to obtain the optimal leaf ordering using linear embedding techniques (Aydin, Bateni, and Mirrokni 2016). To improve the scalability of the backend data structure, we could extend it with distributed analytic systems such as Apache Spark (Zaharia et al. 2016). To visualize the differences between different clusters, we could plot the fold changes of gene transcription levels on the pathway maps in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa et al. 2016).

5 Data availability

In scedar benchmark, the real scRNA-seq datasets can be obtained from the following websites:

- https://hemberg-lab.github.io/scRNA.seq.datasets
- https://support.10xgenomics.com/single-cell-gene-expression/datasets
- https://figshare.com/articles/MCA_DGE_Data/5435866
- https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/downloads

The simulated datasets can be generated using scripts at https://github.com/logstar/scedar/tree/master/docs/r_scripts.

6 Code Availability

The scedar package is distributed through Pypi https://pypi.org/project/scedar
The code and other resources are available in Github at https://github.com/logstar/scedar

7 Supplementary Text

We have provided a Supplementary file with text, algorithms and figures referenced in this work.

8 References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. "Tensorflow: A System for Large-Scale Machine Learning." In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–83. usenix.org.

Abramson, Morton. 1976. "Restricted Combinations and Compositions." *The Fibonacci Quarterly. The Official Journal of the Fibonacci Association* 14 (5): 439–52.

Aevermann, Brian D., Mark Novotny, Trygve Bakken, Jeremy A. Miller, Alexander D. Diehl, David Osumi-Sutherland, Roger S. Lasken, Ed S. Lein, and Richard H. Scheuermann. 2018. "Cell Type Discovery Using Single-Cell Transcriptomics: Implications for Ontological Representation." *Human Molecular Genetics* 27 (R1): R40–47.

Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. 2001. "On the Surprising Behavior of Distance Metrics in High Dimensional Space." In *Database Theory — ICDT 2001*, 420–34. Springer Berlin Heidelberg.

Aydin, Kevin, Mohammadhossein Bateni, and Vahab Mirrokni. 2016. "Distributed Balanced Partitioning via Linear Embedding." In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 387–96. WSDM '16. New York, NY, USA: ACM.

Bar-Joseph, Z., D. K. Gifford, and T. S. Jaakkola. 2001. "Fast Optimal Leaf Ordering for Hierarchical Clustering." *Bioinformatics* 17 Suppl 1: S22–29.

Baslan, Timour, and James Hicks. 2017. "Unravelling Biology and Shifting Paradigms in Cancer with Single-Cell Sequencing." *Nature Reviews. Cancer* 17 (9): 557–69.

Bellet, Aurélien, Amaury Habrard, and Marc Sebban. 2013. "A Survey on Metric Learning for Feature Vectors and Structured Data." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1306.6709.

Beutner, Clara, Bettina Linnartz-Gerlach, Susanne V. Schmidt, Marc Beyer, Michael R. Mallmann, Andrea Staratschek-Jox, Joachim L. Schultze, and Harald Neumann. 2013. "Unique Transcriptome Signature of Mouse Microglia." *Glia* 61 (9): 1429–42.

Bezdek, James C., Siew K. Chuah, and David Leep. 1986. "Generalized K-Nearest Neighbor Rules." *Fuzzy Sets and Systems. An International Journal in Information Science and Engineering* 18 (3): 237–56.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. "Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species." *Nature Biotechnology*, April. https://doi.org/10.1038/nbt.4096.

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, et al. 2017. "Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism." *Science* 357 (6352): 661–67.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. "The Single-Cell Transcriptional Landscape of Mammalian

Organogenesis." *Nature*, February, 1.
Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–94. New York, New York, USA: ACM Press.
Cover, T., and P. Hart. 2006. "Nearest Neighbor Pattern Classification." *IEEE Transactions on Information Theory / Professional Technical Group on Information Theory* 13 (1): 21–27.
Dean, Jeffrey, and Sanjay Ghemawat. 2008. "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM* 51 (1): 107–13.
Deng, Qiaolin, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. 2014. "Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells." *Science* 343 (6167): 193–96.
Dijk, David van, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, et al. 2018. "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion." *Cell*, June. https://doi.org/10.1016/j.cell.2018.05.061.
Dueck, Hannah R., Rizi Ai, Adrian Camarena, Bo Ding, Reymundo Dominguez, Oleg V. Evgrafov, Jian-Bing Fan, et al. 2016. "Assessing Characteristics of RNA Amplification Methods for Single Cell RNA Sequencing." *BMC Genomics* 17 (1): 377.
Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Kdd*, 96:226–31.
Fan, Jean, Neeraj Salathia, Rui Liu, Gwendolyn E. Kaeser, Yun C. Yung, Joseph L. Herman, Fiona Kaper, et al. 2016. "Characterizing Transcriptional Heterogeneity through Pathway and Gene Set Overdispersion Analysis." *Nature Methods* 13 (3): 241–44.
Filbin, Mariella G., Itay Tirosh, Volker Hovestadt, Mckenzie L. Shaw, Leah E. Escalante, Nathan D. Mathewson, Cyril Neftel, et al. 2018. "Developmental and Oncogenic Programs in H3K27M Gliomas Dissected by Single-Cell RNA-Seq." *Science* 360 (6386): 331–35.
Haber, Adam L., Moshe Biton, Noga Rogel, Rebecca H. Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, et al. 2017. "A Single-Cell Survey of the Small Intestinal Epithelium." *Nature* 551 (7680): 333–39.
Hansen, Mark H., and Bin Yu. 2001. "Model Selection and the Principle of Minimum Description Length." *Journal of the American Statistical Association* 96 (454): 746–74.
Hicks, Stephanie C., and Rafael A. Irizarry. 2015. "Quantro: A Data-Driven Approach to Guide the Choice of an Appropriate Normalization Method." *Genome Biology* 16 (June): 117.
Huang, Mo, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I. Murray, Arjun Raj, Mingyao Li, and Nancy R. Zhang. 2018. "SAVER: Gene Expression Recovery for Single-Cell RNA Sequencing." *Nature Methods*, June, 1.
Hubert, Lawrence, and Phipps Arabie. 1985. "Comparing Partitions." *Journal of Classification* 2 (1): 193–218.
Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." *PloS One* 9 (6): e98679.
Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. "KEGG as a Reference Resource for Gene and Protein Annotation." *Nucleic Acids Research* 44 (D1): D457–62.
Kharchenko, Peter V., Lev Silberstein, and David T. Scadden. 2014. "Bayesian Approach to Single-Cell Differential Expression Analysis." *Nature Methods* 11 (7): 740–42.
Kiselev, Vladimir Yu, Tallulah S. Andrews, and Martin Hemberg. 2019. "Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data." *Nature Reviews. Genetics*,

January. https://doi.org/10.1038/s41576-018-0088-9.
Kiselev, Vladimir Yu, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N. Natarajan, et al. 2017. “SC3: Consensus Clustering of Single-Cell RNA-Seq Data.” *Nature Methods* 14 (5): 483–86.
Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Jason C. H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, et al. 2015. “Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation.” *Cell Stem Cell* 17 (4): 471–85.
Korthauer, Keegan D., Li-Fang Chu, Michael A. Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. 2016. “A Statistical Approach for Identifying Differential Distributions in Single-Cell RNA-Seq Experiments.” *Genome Biology* 17 (1): 222.
Lake, Blue B., Song Chen, Brandon C. Sos, Jean Fan, Gwendolyn E. Kaeser, Yun C. Yung, Thu E. Duong, et al. 2018. “Integrative Single-Cell Analysis of Transcriptional and Epigenetic States in the Human Adult Brain.” *Nature Biotechnology* 36 (1): 70–80.
Linderman, George C., and Stefan Steinerberger. 2019. “Clustering with T-SNE, Provably.” *SIAM Journal on Mathematics of Data Science* 1 (2): 313–32.
Li, Wei Vivian, and Jingyi Jessica Li. 2017. “scImpute: Accurate And Robust Imputation For Single Cell RNA-Seq Data.” *bioRxiv*. https://doi.org/10.1101/141598.
Maaten, Laurens van der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9 (Nov): 2579–2605.
Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14.
Malkov, Yu A., and D. A. Yashunin. 2016. “Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs.” *arXiv [cs.DS]*. arXiv. http://arxiv.org/abs/1603.09320.
McInnes, Leland, and John Healy. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.
Müllner, Daniel. 2011. “Modern Hierarchical, Agglomerative Clustering Algorithms.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1109.2378.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–30.
Pierson, Emma, and Christopher Yau. 2015. “ZIFA: Dimensionality Reduction for Zero-Inflated Single-Cell Gene Expression Analysis.” *Genome Biology* 16 (1): 618.
Pollen, Alex A., Tomasz J. Nowakowski, Joe Shuga, Xiaohui Wang, Anne A. Leyrat, Jan H. Lui, Nianzhen Li, et al. 2014. “Low-Coverage Single-Cell mRNA Sequencing Reveals Cellular Heterogeneity and Activated Signaling Pathways in Developing Cerebral Cortex.” *Nature Biotechnology* 32 (10): 1053–58.
Qiu, Xiaojie, Andrew Hill, Jonathan Packer, Dejun Lin, Yi-An Ma, and Cole Trapnell. 2017. “Single-Cell mRNA Quantification and Differential Analysis with Census.” *Nature Methods* 14 (3): 309–15.
Regev, Aviv, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, et al. 2017. “The Human Cell Atlas.” *eLife* 6 (December). https://doi.org/10.7554/eLife.27041.
Risso, Davide, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. 2018. “A General and Flexible Method for Signal Extraction from Single-Cell RNA-Seq Data.” *Nature Communications* 9 (1): 284.

Rizvi, Abbas H., Pablo G. Camara, Elena K. Kandror, Thomas J. Roberts, Ira Schieren, Tom Maniatis, and Raul Rabadan. 2017. “Single-Cell Topological RNA-Seq Analysis Reveals Insights into Cellular Differentiation and Development.” *Nature Biotechnology* 35 (6): 551–60.

Rozenblatt-Rosen, Orit, Michael J. T. Stubbington, Aviv Regev, and Sarah A. Teichmann. 2017. “The Human Cell Atlas: From Vision to Reality.” *Nature* 550 (7677): 451–53.

Schiebinger, Geoffrey, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, et al. 2019. “Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming.” *Cell* 176 (6): 1517.

Soneson, Charlotte, and Mark D. Robinson. 2018. “Bias, Robustness and Scalability in Single-Cell Differential Expression Analysis.” *Nature Methods* 15 (4): 255–61.

Suter, David M., Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. 2011. “Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics.” *Science* 332 (6028): 472–74.

Svensson, Valentine, Roser Vento-Tormo, and Sarah A. Teichmann. 2018. “Exponential Scaling of Single-Cell RNA-Seq in the Past Decade.” *Nature Protocols* 13 (4): 599–604.

Tantale, Katjana, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean-Marc Victor, Marie-Cécile Robert, Serena Capozi, et al. 2016. “A Single-Molecule View of Transcription Reveals Convoys of RNA Polymerases and Multi-Scale Bursting.” *Nature Communications* 7 (July): 12248.

Traag, V. A., L. Waltman, and N. J. van Eck. 2019. “From Louvain to Leiden: Guaranteeing Well-Connected Communities.” *Scientific Reports* 9 (1): 5233.

Vallejos, Catalina A., John C. Marioni, and Sylvia Richardson. 2015. “BASiCS: Bayesian Analysis of Single-Cell Sequencing Data.” *PLoS Computational Biology* 11 (6): e1004333.

Villani, Alexandra-Chloé, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, et al. 2017. “Single-Cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors.” *Science* 356 (6335): eaah4573.

Wang, Q., X. Zhang, Y. Zhang, and Q. Yi. 2013. “AUGEM: Automatically Generate High Performance Dense Linear Algebra Kernels on x86 CPUs.” In *SC ’13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 1–12.

Wilf, Herbert S. 2005. *Generatingfunctionology*. AK Peters/CRC Press.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis.” *Genome Biology* 19 (1): 15.

Xianyi, Z., W. Qian, and Z. Yunquan. 2012. “Model-Driven Level 3 BLAS Performance Optimization on Loongson 3A Processor.” In *2012 IEEE 18th International Conference on Parallel and Distributed Systems*, 684–91.

Zaharia, Matei, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, et al. 2016. “Apache Spark: A Unified Engine for Big Data Processing.” *Communications of the ACM* 59 (11): 56–65.

Zappia, Luke, Belinda Phipson, and Alicia Oshlack. 2017. “Splatter: Simulation of Single-Cell RNA Sequencing Data.” *Genome Biology* 18 (1): 174.

———. 2018. “Exploring the Single-Cell RNA-Seq Analysis Landscape with the scRNA-Tools Database.” *PLoS Computational Biology* 14 (6): e1006245.

Zeisel, Amit, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, et al. 2015. “Brain Structure. Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-Cell RNA-Seq.” *Science* 347 (6226):

1138–42.
Zhang, T., R. Ramakrishnan, and M. Livny. 1996. “BIRCH: An Efficient Data Clustering Method for Very Large Databases.” *ACM Sigmod Record*. https://dl.acm.org/citation.cfm?id=233324.
Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January): 14049.
Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. 2017. “Comparative Analysis of Single-Cell RNA Sequencing Methods.” *Molecular Cell* 65 (4): 631–43.e4.