Hyper-fast and accurate clustering of ultra-large-scale single-cell data with ensemble random projection

Shibiao Wan [1] **Junil Kim** [2 3] **Yiping Fan** [1] **Kyoung Jae Won** [2 3]

Abstract

Recent advances on single-cell RNA-sequencing (scRNA-seq) technologies have enabled parallel transcriptomic profiling of millions of cells. However, existing scRNA-seq clustering methods are lack of scalability, time-consuming and prone to information loss during dimension reduction. To address these concerns, we present SHARP, an ensemble random projection-based algorithm which is scalable to clustering 10 million cells. By adopting a divide-and-conquer strategy, a sparse random projection and two-layer meta-clustering, SHARP has the following advantages: (1) hyper-faster than existing algorithms; (2) scalable to 10-million cells; (3) accurate in terms of clustering performance; (4) preserving cell-to-cell distance during dimension reduction; and (5) robust to dropouts in scRNA-seq data. Comprehensive benchmarking tests on 20 scRNA-seq datasets demonstrate SHARP remarkably outperforms state-of-the-art methods in terms of speed and accuracy. To the best of our knowledge, SHARP is the only R-based tool that is scalable to clustering 10 million cells.

1. Introduction

To characterize novel cell types and detect intra-population heterogeneity, scRNA-seq has been widely applied in biology and medicine by enabling parallel transcriptomic profiling of millions of cells. To cluster high dimensional scRNA-seq data, dimension reduction algorithms such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), or uniform manifold approximation and projection (UMAP) are often used to process and to visualize high dimensional scRNA-seq data. However, these algorithms either require considerable computational time or are susceptilbe to losing original cell-to-cell distances in the low-dimensional space.

2. Algorithm

To effectively handle very large-scale scRNA-seq data without excessive distortion of cell-to-cell distances, we developed SHARP (Wan et al., 2020) (`https://github.com/shibiaowan/SHARP`), a hyper-fast clustering algorithm based on ensemble random projection (RP) (Fig. 1A). SHARP employed a divide-and-conquer strategy followed by RP to accommodate effective processing of large-scale scRNA-seq data (Fig. 1A). SHARP processes scRNA-seq data in 4 interconnected steps: (1) data partition, (2) RP based clustering, (3) weighted ensemble clustering and (4) similarity-based meta-clustering. During data partition, the scRNA-seq data is divided into small blocks (random size). The divide-and-conquer strategy enables SHARP to upload and process more than 1 million cells. The divided data blocks are further processed by RP followed by a hierarchical clustering algorithm. Because the performance of an individual RP-based clustering is volatile, ensemble of several runs of RPs is used. A weighted-ensemble clustering (i.e., wMetaC) algorithm merges individual RP-based clustering results. Finally, a similarity-based ensemble clustering (i.e., sMetaC) approach is to integrate clustering results of each block (Fig. 1A).

3. Advantages of SHARP

3.1. SHARP is faster than other predictors

We performed comprehensive benchmarking of SHARP against existing scRNA-seq clustering algorithms using 20 scRNA-seq datasets whose cell number ranges from 124 to 10 million cells (Fig. 1B-C). The computing cost of SHARP was substantially lower than other clustering algorithms (Fig. 1B). The required computing cost of SHARP rose roughly linearly even with the very large size of the datasets.

[1] Center for Applied Bioinformatics, St. Jude Childrens' Research Hospital, Memphis, 38105, TN, USA [2] Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Ole Maaløes Vej 5. 2200 Copenhagen N, Denmark [3] Novo Nordisk Foundation Center for Stem Cell Biology, DanStem, Faculty of Health and Medical Sciences, University of Copenhagen, Ole Maaløes Vej 5. 2200 Copenhagen N, Denmark. Correspondence to: Shibiao Wan <shibiao.wan@stjude.org>, Kyoung Jae Won <kyoung.won@bric.ku.dk>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB). Copyright 2020 by the author(s).

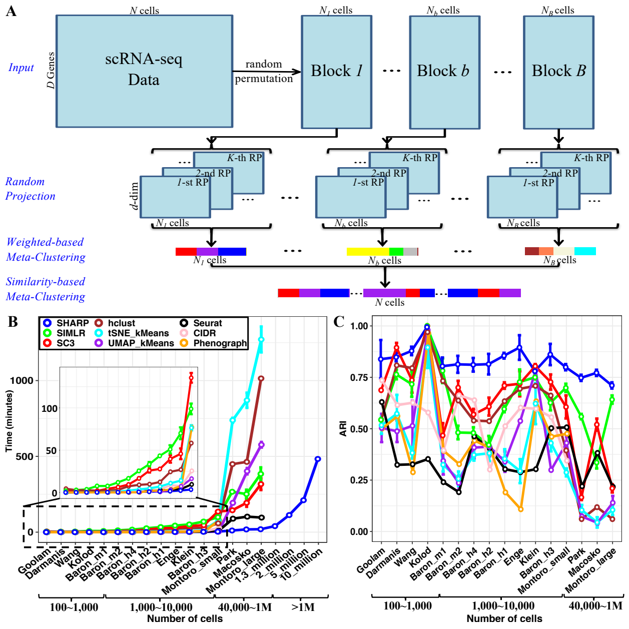

Figure 1. The framework of SHARP. (A) SHARP has 4 steps for clustering: divide-and-conquer, random projection (RP), weighted-based meta clustering, and similarity-based meta-clustering. (B) Running time and (C) clustering performance of SHARP in 20 single-cell RNA-seq datasets with numbers of single cells ranging from 124 to 10 million (The last 3 datasets were generated by randomly oversampling the dataset with 1.3 million single cells). For the datasets with >1 million cells, only SHARP can run and only the running time was provided due to lack of the ground-truth clustering labels. Refer to (Wan et al., 2020) for details of the datasets.

SHARP clustered the scRNA-seq with 1.3 million cells in 42 minutes when using a multi-core system (Fig. 1B). Due to the data loading problem (and potential exhaustive memory use), we could not show the running time of other approaches for 1.3 million cells. The running time of SHARP for 1.3 million cells is even 2 times (42 mins vs 96 mins) faster than that of Seurat for 66,255 cells. We expect far superior performance of SHARP against its competitors in case data loading is feasible.

3.2. SHARP is scalable to 10 million cells

To demonstrate the scalability of SHARP, we performed random over-sampling of the mouse brain dataset of 1.3 million cells so that we were able to construct even larger sizes of scRNA-seq datasets, e.g., 10 million cells. The running time of SHARP was simply linearly increased with the increasing of cell numbers from 1 million to 10 million. In our system using 16 cores, SHARP just needed around 8 hours (i.e., 482.8 minutes) to cluster 10 million cells into 1175 clusters (Fig 1B).

3.3. SHARP is accurate in terms of clustering performance

For almost all datasets we tested, SHARP showed better performances (Fig. 1C). The performance of other algorithms became generally worse for large datasets (>40,000 single cells). In contrast, SHARP showed an ARI (adjusted Rand index) larger than 0.7 regardless of the size of the datasets, demonstrating its robustness(Fig. 1C).

3.4. SHARP preserves cell-to-cell distance

We investigated the degree of distortion caused by dimension reduction and compared the correlation of cell-to-cell distances after reducing dimension using SHARP, PCA and t-SNE, respectively. SHARP showed almost perfect similarities in cell-to-cell distance with correlation coefficient > 0.94 even in a dimensional space which is 74 times lower (from 20862 to 279) than the original one whereas cell-to-cell distances for PCA and t-SNE were distorted when dimension reduction was performed to the same number of dimensions (Fig. 2A of (Wan et al., 2020)).

3.5. SHARP is robust to dropouts

scRNA-seq suffers a high frequency of dropouts where many of the true expressions are not captured. To evaluate the robustness of SHARP against dropouts, we tested SHARP while artificially increasing dropout rates in a scRNA-seq dataset (Fig. 2B of (Wan et al., 2020)). We found that SHARP is robust to the added dropouts, while we observed poorer performance for the added dropouts in general for other methods (Fig. 2B of (Wan et al., 2020)).

3.6. Clustering 1.3 million cell data using SHARP

Of note, SHARP provides an opportunity to study the million-cell-level dataset. Using SHARP, we identified a total of 244 clusters from this 1.3 million dataset (17 clusters with more than 1,000 cells). The top 4 clusters among them were found to have clear different expression patterns (Fig. 2E of (Wan et al., 2020)). Gene Ontology (GO) analysis show that Cluster 2 is associated with dendrites and Cluster 3 is with axon. We also identified a cluster (Cluster 8) enriched for the genes associated with "non-motile cilium assembly", which is important for brain development and function and immune cells with high IL4 expression (Cluster 14).

References

Wan, S., Kim, J., and Won, K. J. SHARP: hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection. *Genome Research*, 30(2): 205–213, 2020.