Bioinformatics, 37(21), 2021, 3964–3965
doi: 10.1093/bioinformatics/btab420
Advance Access Publication Date: 7 June 2021
Applications Note

Gene expression

HGC: fast hierarchical clustering for large-scale single-cell data

Ziheng Zou [1,†], Kui Hua [1,†,*] and Xuegong Zhang [1,2,*]

[1]MOE Key Laboratory of Bioinformatics, Division of Bioinformatics, BNRIST and Department of Automation, Tsinghua University, Beijing 100084, China and [2]School of Life Sciences, Tsinghua University, Beijing 100084, China

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Valentina Boeva

Received on February 9, 2021; revised on May 26, 2021; editorial decision on May 27, 2021

Abstract

Summary: Clustering is a key step in revealing heterogeneities in single-cell data. Most existing single-cell clustering methods output a fixed number of clusters without the hierarchical information. Classical hierarchical clustering (HC) provides dendrograms of cells, but cannot scale to large datasets due to high computational complexity. We present HGC, a fast Hierarchical Graph-based Clustering tool to address both problems. It combines the advantages of graph-based clustering and HC. On the shared nearest-neighbor graph of cells, HGC constructs the hierarchical tree with linear time complexity. Experiments showed that HGC enables multiresolution exploration of the biological hierarchy underlying the data, achieves state-of-the-art accuracy on benchmark data and can scale to large datasets.

Availability and implementation: The R package of HGC is available at https://bioconductor.org/packages/HGC/.

Contact: zhangxg@tsinghua.edu.cn or stevenhuakui@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The development of single-cell RNA sequencing (scRNA-seq) and bioinformatics technologies have accelerated the understanding of cell heterogeneity (Aldridge and Teichmann, 2020). The current practice for studying the multi-level cell heterogeneity is to first produce a fixed number of clusters and then adjust the clustering resolutions in an *ad hoc* manner (Hua and Zhang, 2019; Luecken and Theis, 2019). This workflow loses the underlying hierarchical information and requires multi-rounds of re-clustering to find a suitable resolution. Hierarchical clustering (HC) enables direct multi-resolution exploration of the heterogeneity, but classical HC methods are only suitable for small datasets due to the high computational complexity.

We propose a fast Hierarchical Graph Clustering tool HGC for large-scale single-cell data. The key idea of HGC is to construct a dendrogram of cells on their shared nearest-neighbor (SNN) graph. This combines the advantages of graph-based clustering and HC. We applied HGC on both synthetic and real scRNA-seq datasets. Results showed that HGC can recover the biological hierarchy underlying the data, can achieve high clustering accuracy at fixed resolution and can scale well to large datasets.

2 Materials and methods

HGC contains two major steps: graph construction and dendrogram construction. For the graph construction step, HGC adopts the standard procedure of building the SNN graph, which is to first apply principal component analysis on the expression data and then build the k nearest neighbor graph and the SNN graph in the PC space (Fig. 1a). For the step of dendrogram construction on the graph, HGC uses a recursive procedure of finding the nearest-neighbor node pairs and updating the graph by merging the node pairs (Fig. 1a).

The key in finding the nearest-neighbor pair on a graph is the distance measure. HGC uses the node-pair sampling distance introduced in (Bonald *et al.*, 2018). For a weighted, undirected graph $G = (V, E)$, let A be the weighted adjacent matrix. If we sample node pairs or edges at random in proportion to their weights, the probability that node pair or edge (i, j) being sampled is:

$$p(i,j) = \frac{A_{ij}}{\sum_{i,j} A_{ij}} \tag{1}$$

Similarly, sampling nodes in proportion to their weighted degrees results in the probability of node i being sampled:

$$p(i) = \frac{\sum_j A_{ij}}{\sum_{i,j} A_{ij}} = \sum_j p(i,j) \tag{2}$$

The node-pair sampling distance is defined as the ratio between individual sampling and the pair sampling:

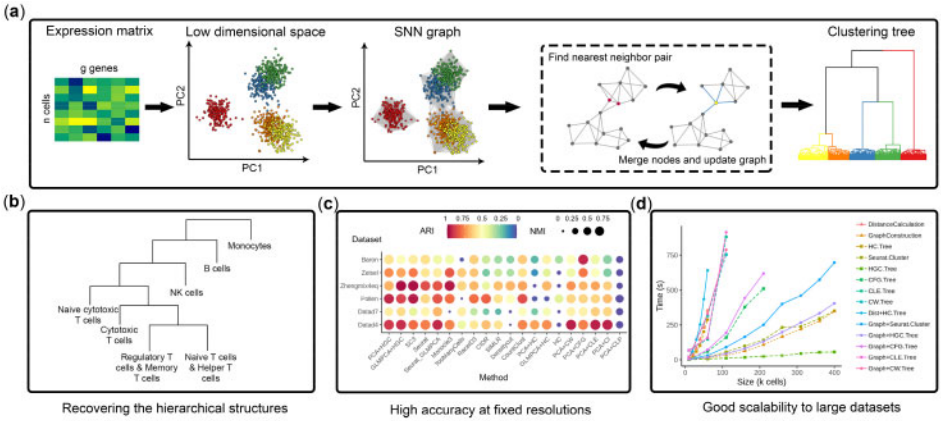

Fig. 1. Overview of HGC. (**a**) Workflow of HGC, including the construction of SNN graph, and the recursive procedure for constructing dendrogram on the graph. (**b**) Experiment of HGC on the PBMC dataset showing the recovery of the hierarchical structure underlying the data. (**c**) Benchmarking of 20 clustering methods on 6 datasets, including HGC applied on different feature spaces, representative clustering methods designed for scRNA-seq data and some general clustering methods such as classical HC and 5 graph-based clustering methods (CW, CFG, CLE, CI and CLP, Supplementary Material). HGC, Seurat and SC3 achieved comparable clustering accuracy and significantly outperformed other methods. (**d**) Benchmarking of the scalability of HGC, HC, three graph-based methods and Seurat on datasets with different sizes. HGC is significantly faster than other methods

$$d(i,j) = \frac{p(i)p(j)}{p(i,j)} \tag{3}$$

After finding two nearest-neighbor nodes, the graph is updated by merging them into a new node (Fig. 1a). The weights of the new nodes are the sum of weights of the original nodes. Weights of edges between new nodes are the sum of weights of edges between original nodes forming the two new nodes. The whole HC procedure can be accelerated using the nearest-neighbor chain algorithm because the node-pair sampling distance is reducible (Bonald *et al.*, 2018).

We implemented HGC as a software in R, with the key function written in Rcpp. It provides Seurat-style function to enable seamlessly usage in the popular pipeline. It also includes tools to assist downstream analysis such as dynamicTreeCut for cutting the dendrogram into specific clusters and plotting functions for visualizing HC results.

3 Results

3.1 HGC reveals hierarchical structure of cell heterogeneity

We experimented HGC on two datasets with known hierarchical structures (Supplementary Figs S1 and S7). We visualized the clustering results with dendrograms and t-SNE plots. Results showed that HGC well recovered the hierarchical structures (Fig. 1b and Supplementary Figs S2, S3, S8 and S9). We cut the dendrogram at certain level to get fixed-number clustering results, and used adjusted rand index (ARI) to compare with known results on the data. Clusters given by HGC agreed well with the known labels at different levels (Supplementary Figs S2c and S8c).

3.2 Benchmarking at fixed resolutions

To further benchmark HGC's performance on revealing cell heterogeneity at fixed levels, we collected 6 scRNA-seq datasets with known labels and compared results of 20 clustering methods with known labels using ARI and normalized mutual information. Results showed that HGC, Seurat and SC3 achieved the best clustering accuracy and significantly outperformed other methods (Fig. 1c, Supplementary Tables S3 and S4).

3.3 Scalability

We tested the time efficiency of HGC on datasets of different sizes sampled from Mouse Cell Atlas dataset (Han *et al.*, 2018). On the data of 5000 cells, HGC was 33 times faster than HC (13s versus 434s on a laptop computer). Running HC on a laptop became infeasible on datasets with more than 10 000 cells. HGC completed the HC on the data of 400 000 cells in 404s, ~70% faster even than Seurat which only gives a fixed number of clusters and much faster than some existing graph-based hierarchical methods (Fig. 1d and Supplementary Fig. S15).

4 Conclusion

We developed a new method HGC and its R package for fast HC of single-cell data. It can reveal the hierarchical structure underlying the data, achieves state-of-the-art clustering accuracy and can scale to very large single-cell datasets.

Funding

This work was supported by the NSFC Projects (61721003 and 62050178) and National Key R&D Program of China (2018YFC0910401).

Conflict of Interest: none declared.

References

Aldridge,S. and Teichmann,S.A. (2020) Single cell transcriptomics comes of age. *Nat. Commun.*, **11**, 4307.

Bonald,T. *et al.* (2018) Hierarchical graph clustering using node pair sampling. arXiv preprint arXiv:1806.01664.

Han,X. *et al.* (2018) Mapping the mouse cell atlas by microwell-seq. *Cell*, **173**, 1307.

Hua,K. and Zhang,X. (2019) A case study on the detailed reproducibility of a Human Cell Atlas project. *Quan. Biol.*, **7**, 162–169.

Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.