Effects of sample size on plant single-cell RNA profiling

Longbiao Guo (✉ guolongbiao@caas.cn)

China National Rice Research Institute <https://orcid.org/0000-0003-1016-049X>

Hongyu Chen

Zhejiang University

Xinxin Yin

Zhejiang University

Xi Chen

Zhejiang University

Qinjie Chu

Zhejiang University

Qianhao Zhu

CSIRO Manufacturing and Infrastructure Technology: CSIRO Australian Manufacturing and Materials Precinct

Longbiao Fan

Zhejiang University

Research article

Keywords: single-cell RNA (scRNA), cell number, sampling coverage

Posted Date: May 24th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-549533/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Single-cell RNA (scRNA) profiling or scRNA-sequencing (scRNA-seq) is a rapidly developing technology and an important frontier of molecular biology science. scRNA profiling makes it possible to parallelly investigate diverse molecular features of multiple types of cells in a given plant tissue, and promotes elucidation of cellular heterogeneity and discovery of developmental processes underpinning cell differentiation. While it is assumed that the power of scRNA profiling in uncovering cellular heterogeneity largely depends on the depth of scRNA-seq, no study about the effect of the sequenced cell numbers on the power of plant scRNA-seq has ever been reported.

Results: In this study, on the basis of analyzing the sample coverage of 1,244 available scRNA-seq studies (including 30 in plants) and the effect of sample coverage on cell clustering and identification of cell types, we evaluated the effects of sample size (i.e., cell number) on the outcome of single cell transcriptome analysis by sampling different number of cells from a pool of ~57,000 *Arabidopsis thaliana* root cells integrated from five published studies. Our results indicated that the most significant principle components could be achieved when 20,000-30,000 cells were sampled, a relatively high reliability of cell clustering could be achieved by using ~20,000 cells with little further improvement by using more cells, 96% of the differentially expressed genes could be successfully identified with no more than 20,000 cells, and a relatively stable pseudotime could be estimated in the sub-sample with 5,000 cells.

Conclusions: Our results imply that ~20,000 (or 10,000 - 30,000[1]) cells are enough for profiling *Arabidopsis* root cells using scRNA-seq, although the applicability of this number to other *Arabidopsis* tissues and other plants is yet to be further determined by analyzing scRNA-seq data generated from diverse tissues of different plant species. Nevertheless, our results provide a general guide for optimizing sample size to be used in plant scRNA-seq studies. Change to “or up to 300000”?

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

Figure 1

Distribution of sample sizes (i.e., cell numbers) in the recent 686 studies on single-cell RNA profiling. Technique means the methodology applied in profiling single-cell gene expression. Distribution of the 686 studies in different species are shown in Table SI.

Figure 2

Significant principal component (PC) numbers in the nine sub-samples with different cell numbers.

Figure 3

Effect of cell numbers on cell clustering. (A) ISNE maps of the nine sub-samples with different cell numbers. (B) Changes of the number of cell clusters with the increase of cell numbers. (C) Changes of the annotated cell types with the increase of cell numbers. (D) Composition of cell types annotated in different sub-samples. (E) The scores of the five statistical measures for evaluating cluster similarity of different sample size. The five indices are Rand index (RI), Morey and Agrest's adjusted Rand index (MA), Hubert and Arabie's adjusted Rand index (HA), Fowlkes and Mallows index (FM), and Jaccard index (JI).

Figure 4

Effect of cell numbers on cell trajectory inference. (A) The UMAP map of nine sub-samples with different cell numbers. (B) Density distribution map of meristem and hair cells on the pseudotime axis in the nine sub-samples with different cell numbers. (C) The number of differentially expressed genes that change as a function of pseudotime.

Supplementary Files

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

- [FigureS1.pdf](#)

- [FigureS2.pdf](#)
- [FigureS3.pdf](#)
- [FigureS4.pdf](#)
- [TableS1.docx](#)
- [TableS2.docx](#)