Gene expression

scSampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data

Dongyuan Song1,†, Nan Miles Xi 2,†, Jingyi Jessica Li 3,* and Lin Wang 4,*

1Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles, CA 90095-7246, USA, 2Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL 60660-1601, USA, 3Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA and 4Department of Statistics, The George Washington University, Washington, DC 20052-0086, USA

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

*To whom correspondence should be addressed.

Abstract

Summary: The number of cells measured in single-cell transcriptomic data has grown fast in recent years. For such large-scale data, subsampling is a powerful and often necessary tool for exploratory data analysis. However, the easiest random subsampling is not ideal from the perspective of preserving rare cell types. Therefore, diversity-preserving subsampling is required for fast exploration of cell types in a large-scale dataset. Here, we propose scSampler, an algorithm for fast diversity-preserving subsampling of single-cell transcriptomic data.

Availability and implementation: scSampler is implemented in Python and is published under the MIT source license. It can be installed by “pip install scsampler” and used with the Scanpy pipeline. The code is available on GitHub: https://github.com/SONGDONGYUAN1994/scsampler. An R interface is available at: https://github.com/SONGDONGYUAN1994/scsampler.

Contact: jli@stat.ucla.edu or linwang@gwu.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have undergone rapid development in recent years. A remarkable achievement is the generation of large-scale datasets, and there are even datasets containing a million cells (see Supplementary Table S1). Such massive scRNA-seq datasets have impeded exploratory data analysis (e.g., visualization) on standard computers.

An intuitive solution to this ‘big data’ challenge is to subsample (downsample) a large-scale dataset, i.e. to select a subset of representative cells. Random subsampling is fast and has been implemented in popular pipelines such as Seurat (Satija et al., 2015) and Scanpy (Wolff et al., 2018). However, random subsampling may miss rare cell types and is thus not ideal for preserving the transcriptome diversity. To overcome this drawback, Hie et al. (2019) proposed the algorithm Geosketch for ‘intelligently selecting’ a subset (called ‘sketching’) of single cells. Geosketch aims to evenly sample cells across the transcriptome space by approximately minimizing the Hausdorff distance between the subsample and the original sample. In the follow-up algorithm Hopper (DeMeo and Berger, 2020), the authors improved the performance of Geosketch in terms of reducing the Hausdorff distance. In fact, prior to Geosketch and Hopper and outside of the single-cell field, this ‘intelligent subsampling’ problem has been well studied in the field of computer experiment design, in which the ‘space-filling design’ implements the idea of even subsampling (Joseph, 2016). Popular space-filling designs include the minimax and maximin distance designs (Johnson et al., 1990). Geosketch and Hopper conceptually belong to the minimax distance design, which can be approximated by, but is much more computationally intensive than, the maximin distance design (Johnson et al., 1990). Here, we propose scSampler, a Python package for fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. By ‘diversity-preserving sampling’, scSampler implements the maximin distance design to make cells in the subsample as separative as possible. We show that scSampler outperforms existing subsampling methods in minimizing the Hausdorff distance between the subsample and the original sample. Moreover, scSampler is fast and scalable for million-level data.

2 Implementation

The input is a matrix \( X \in \mathbb{R}^{n \times p} \). The columns of \( X \) correspond to \( p \) features (by default, top \( p \) PCs from a cell-by-gene log(count + 1) matrix and scaled to \([0, 1]\)), and rows correspond to \( n \) cells. Then we have a set of cells \( X = \{ x_1, \ldots, x_n \} \), where \( x_i \in \mathbb{R}^p \) is a row vector of \( X \). Our goal is to find a size \( n^\prime \) subset \( X^\prime \subseteq X \) that minimizes the Hausdorff distance.
datasets and subsample sizes each method. The table shows the mean ranks and standard deviations across all Hausdorff distance against computation time. (c) Summary of the performance of subsampling methods in the two measures. Notably, scSampler-B1 consistently yields the smallest Hausdorff distances across all datasets and all subsample sizes. Moreover, scSampler is fast: on the largest cortex dataset (more than 1 million cells), scSampler-B1 finishes within 15 min, and scSampler-B16 takes only 1 min and still outperforms Geosketch and Hopper by achieving a lower Hausdorff distance. Figure 1c shows that scSampler is consistently ranked the top (smaller ranks are better) across the 18 datasets.

To verify if rare cell types are better captured by scSampler than by other methods, we calculate the Gini coefficient of cell type proportions in each subsample; a smaller Gini coefficient indicates more balanced cell types (Supplementary Section S3.3). In more than 60% of the combinations of 18 datasets and four subsample sizes, the fastest scSampler-B16 leads to the smallest Gini coefficient (Supplementary Section S5). Considering that the real datasets may not have accurately annotated cell types, we examine the simulated datasets and find that scSampler-B16 leads to the smallest Gini coefficient in 90% of the combinations of eight simulated datasets and four sample sizes, confirming that scSampler preserves rare cell types well.

**Funding**

This work was supported by National Science Foundation DBI-1846216 and DMS-2113754, NIH/NIGMS R01GM120507 and R35GM140888, Johnson and Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

**Conflict of Interest**: The authors declare no conflict of interest.

**Data availability**

The real datasets are described in Supplementary Materials S6 with access and preprocessing information. The code to reproduce the analyses is available at https://doi.org/10.5281/zenodo.5811787.

**References**


