Subject Section

A cloud-based platform for the analysis of single cell RNA sequencing data.

Nithin Joshy¹, and Kyuson Yun^{2,*}

¹Farmington High School, 10 Montith Dr, Farmington CT-06032, ²Department of Neurology, Houston Methodist Research Institute, Houston TX 77030.

*To whom correspondence should be addressed.

Abstract

Motivation: Single-cell RNA sequencing (scRNA-seq) is a recent technology that has provided many valuable biological insights. Notable uses include identifying novel cell-types, measuring the cellular response to treatment, and tracking trajectories of distinct cell lineages in time. The raw data generated in this process typically amounts to hundreds of millions of sequencing reads and requires substantial computational infrastructure for downstream analysis, a major hurdle for a biological research lab. Fortunately, the preprocessing step that converts this huge sequence data into manageable cell-specific expression profiles is standardized and can be performed in the cloud. We demonstrate how a cloud-based computational framework can be used to transform the raw data into biologically interpretable cell-type-specific information, using either 3' or 5' transcriptome libraries from 10x Genomics. The processed data which is an order of magnitude smaller in size can be easily downloaded to a laptop for customized analysis to gain deeper biological insights.

Results: We produced an automated and easily extensible pipeline in the cloud for the analysis of single-cell RNA-seq data which provides a convenient method to handle post-processing of scRNA sequencing using next generation sequencing platforms. The basic step provides the transformation of the scRNA-seq data to cell-type-specific expression profiles and computes the quality control metrics for the dataset. The extensibility of the platform is demonstrated by adding a doublet-removal algorithm and recomputing the clustering of the cells. Any additional computational steps that take a cell-type expression counts matrix as input can be easily added to this framework with minimal effort.

Availability: The framework and its documentation for installation is available at the Github repository <http://github.com/nj3252/CB-Source/>

Contact: kyun@houstonmethodist.org

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

Introduction

Single-cell RNA sequencing is used to precisely determine expressed genes, and by extension, cell types and states among heterogeneous cell types present within a sample (1). Application of this technology is rapidly growing and some of these are listed below to illustrate its power. Comparing the differences in gene expression between individual cells has identified novel cell-types in a tissue (2). In cancer research, the ability to find and characterize outlier cells has potential implications for furthering our understanding of drug resistance mechanisms and relapse in cancer treatment (3). scRNA-seq is also being utilized to delineate cell lineage relationships in early development, cell-type differentiation, and cell fate determination (4).

As with many genomics approaches, a major hurdle in using this technology is the amount of raw data that it generates and computational power needed to analyze the data. Higher-end computational infrastructure is required to process the raw sequencing data into biologically

meaningful cell-type-specific expression profiles (5). However, the need for high power computational infrastructure is limited only for the first step for sequence alignment and counts matrix generation. Purchasing high-end computational power for such a small portion of a study is a massive waste of resources. One can easily overcome this need and far more effectively and efficiently process the data in the cloud. Using cloud computing provides cheap and convenient access to the necessary computing power and data sharing. The resources used can be easily customized to optimize the system for cost and time. It is also easy to process scRNA-seq data in the cloud because although the steps take a long time, the initial sequence processing steps are fixed, especially when using the single cell sequencing platform from 10X Genomics, and thus easily automated. The transformed dataset is an order of magnitude smaller which means that further analysis can be done without high-end computational resources. Importantly, these processed data are ready for analysis any downstream analysis tools; for example, an interactive free software Loupe Cell Browser from 10X Genomics is easy to use and can provide first pass analysis.

Methods

Detailed methods for library preparation and sequencing are as described in another manuscript (Abdelfattah et al., in review).

The components for processing a scRNA-seq data was containerized and stored in the Google cloud repository. The widely used platform for generating scRNA-seq dataset is droplet based single cell capture from 10x Genomics (6). The reference implementation provides a containerized version of Cell Ranger from 10X genomics that can be used to process scRNA-seq data available in the cloud storage.

To demonstrate the versatility of this method, an online pipeline was developed with Flask (Figures 2 and 3), a comprehensive web framework in python. After data is uploaded to Google Cloud Storage (GCS), it can be submitted for post-processing by entering the URI to the data along with basic details such as a name and organism type into an online form. The framework utilizes dsub, a Google Cloud Life Sciences API, to create new VMs to process the submitted data using the Cell Ranger software (available from 10X Genomics). The web interface of a portal that captures these input variables is shown in Figure 1. To demonstrate the extensibility of this framework, we have integrated doublet removal into the pipeline, an essential next step in the scRNA-seq analysis process. Doublet detection and removal is done through Scrublet (7), a computational framework to identify doublets in a single cell dataset. Furthermore, the initial steps of downstream analysis have been automated by incorporating Seurat, an R package to process and visualize scRNA-seq data. For example, Seurat was used in the pipeline to identify cells that likely belong to the same cell types using a clustering algorithm (Figure 4).

Once sample specific information is entered, data processing will commence without any extra steps needed from the user. Deploying the web interface through the google application engine (GAE) is important to access the stored data on GCS as well as necessary cloud computing power. Computing power is obtained through the google cloud engine (GCE) which provides customizable virtual machines. Virtual machine specs are easily modified which allows the

process to be optimized for cost and time. A 5000 cell PBMC dataset from 10x Genomics was used to benchmark pipeline speeds on various preset VMs from the Google Cloud Platform (Table 1). Results from this analysis revealed that processing speed is limited first by memory and then by cores, meaning that higher memory machines are more efficient. Contrary to our initial assumption, a more powerful VM is actually cheaper because it greatly reduced time for processing and lowered cost.

We have developed a cloud-based framework for processing scRNA-seq data that will make scRNA-seq analysis amenable for biological research groups without rich computational resources or designated computational/bioinformatics expert. Through this pipeline, we have demonstrated how automation in the cloud is the most effective method to deal with large scRNA-seq datasets. It allows research teams to easily deal with the problem of required post-processing, allowing for greater focus on drawing biological conclusions from the data.

Machine Type	Time (Hours)	Cost (\$)
N1-standard-8	57	15.22
N1-highcpu-32	19	15.11
N1-standard-16	26	13.86
N1-highmem-16	10	10.48
N1-highmem-8	17	8.91
N1-standard-32	6.75	7.12
N1-highmem-32	2.5	5.24

Table 1. Using a high-end machine with more memory saves on cost and time.

The screenshot shows a web interface for a 'CB Pipeline'. At the top, there are navigation links: 'Home', 'Submit Job', 'Aggregate', and a 'Console' link on the right. The main heading is 'Enter details to submit pipeline job'. Below this, there are several input fields and options:

- Project Name:** A text input field.
- Input Data Location:** A text input field.
- Choose a zone to submit the job:** A dropdown menu with 'us-central1' selected.
- Please note that there is a limit of 24 CPU per zone, so please choose a different zone when submitting multiple jobs at the same time.**
- Organism:** Radio buttons for 'Human' (selected) and 'Mouse'.
- Submit:** A green button at the bottom left.

Figure 1. Submitting single cell data for processing is easy through the automated pipeline.

Figure 3. Pipeline Sample Flow Diagram

Figure 4. Seurat UMAP clustering of scRNA-seq data

References

1. Artegiani, Benedetta, Anna Lyubimova, Mauro Muraro, Johan H. van Es, Alexander van Oudenaarden, and Hans Clevers. "A Single-Cell RNA Sequencing Study Reveals Cellular and Molecular Dynamics of the Hippocampal Neurogenic Niche." *Cell Reports* 21, no. 11 (December 12, 2017): 3271–84. <https://doi.org/10.1016/j.celrep.2017.11.050>.
2. Villani, Alexandra-Chlo, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, et al. "Single-Cell RNA-Seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors." *Science (New York, N.Y.)* 356, no. 6335 (21 2017). <https://doi.org/10.1126/science.aah4573>.
3. Eyler, Christine E., Hironori Matsunaga, Volker Hovestadt, Samantha J. Vantine, Peter van Galen, and Bradley E. Bernstein. "Single-Cell Lineage Analysis Reveals Genetic and Epigenetic Interplay in Glioblastoma Drug Resistance." *Genome Biology* 21, no. 1 (July 15, 2020): 174. <https://doi.org/10.1186/s13059-020-02085-1>.
4. Qiu, Xiaojie, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. "Reversed Graph Embedding Resolves Complex Single-Cell Trajectories." *Nature Methods* 14, no. 10 (October 2017): 979–82. <https://doi.org/10.1038/nmeth.4402>.
5. Angerer, Philipp, Lukas Simon, Sophie Tritschler, F. Alexander Wolf, David Fischer, and Fabian J. Theis. "Single Cells Make Big Data: New Challenges and Opportunities in Transcriptomics." *Current Opinion in Systems Biology* 4 (August 2017): 85–91. <https://doi.org/10.1016/j.coisb.2017.07.004>.
6. <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger>
7. Wolock, Samuel L., Romain Lopez, and Allon M. Klein. "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data." Preprint. Bioinformatics, July 9, 2018. <https://doi.org/10.1101/357368>.