Bioinformatics, 35(24), 2019, 5306–5308
doi: 10.1093/bioinformatics/btz601
Advance Access Publication Date: 2 August 2019
Applications Note

Genome analysis

scRNABatchQC: multi-samples quality control for single cell RNA-seq data

Qi Liu[1,2,†], Quanhu Sheng[1,2,†], Jie Ping[1,2], Marisol Adelina Ramirez[1,2], Ken S. Lau[2,3], Robert J. Coffey[3] and Yu Shyr[1,2,*]

[1]Department of Biostatistics, [2]Center for Quantitative Sciences and [3]Department of Cell and Developmental Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

*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: Mark Robinson

Received on March 6, 2019; revised on July 26, 2019; editorial decision on July 29, 2019; accepted on July 30, 2019

Abstract

Summary: Single cell RNA sequencing is a revolutionary technique to characterize inter-cellular transcriptomics heterogeneity. However, the data are noise-prone because gene expression is often driven by both technical artifacts and genuine biological variations. Proper disentanglement of these two effects is critical to prevent spurious results. While several tools exist to detect and remove low-quality cells in one single cell RNA-seq dataset, there is lack of approach to examining consistency between sample sets and detecting systematic biases, batch effects and outliers. We present scRNABatchQC, an R package to compare multiple sample sets simultaneously over numerous technical and biological features, which gives valuable hints to distinguish technical artifact from biological variations. scRNABatchQC helps identify and systematically characterize sources of variability in single cell transcriptome data. The examination of consistency across datasets allows visual detection of biases and outliers.

Availability and implementation: scRNABatchQC is freely available at https://github.com/liuqivandy/scRNABatchQC as an R package.

Contact: yu.shyr@vanderbilt.edu

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

1 Introduction

Single cell RNA-sequencing (scRNA-seq) is a powerful technique of whole-transcriptome profiling at the resolution of individual cells. It has been successfully used to discover rare and heterogeneous cell populations, and reconstruct developmental trajectories (Carter *et al.*, 2018; Karaayvaz *et al.*, 2018).

One major challenge in scRNA-seq analysis is to detect technical artifacts and remove poor quality cells. Previous studies have employed different strategies to detect technical artifacts (Ilicic *et al.*, 2016; Jiang *et al.*, 2016; Lun *et al.*, 2016; McCarthy *et al.*, 2017; Tian *et al.*, 2018). They generally use features such as overall gene expression patterns, number of genes detected and housekeeping genes or spike-in RNA. For example, SinQC integrated both gene expression patterns and sample sequencing library qualities, such as total number of mapped reads, mapping rate and library complexity to detect technical artifacts (Jiang *et al.*, 2016). Ilicic *et al.* used biological and technical features, including number of genes detected, number of mapped reads and percentage of mitochondrial genes to train a SVM model to distinguish low from high quality cells (Ilicic *et al.*, 2016). The Scater and scPipe packages provided methods to compute a variety of QC metrics and visualization to diagnose potential issues (McCarthy *et al.*, 2017; Tian *et al.*, 2018). These strategies successfully identify compromised cells within a single dataset. For integrated or comparative analysis of large collections of scRNA-seq experiments, quality assessment across datasets is crucial to detect outliers, potential batch effects, or systematic biases since they will mask underlying biology and result in misleading conclusions.

Here, we present scRNABatchQC, an R package designed to assess the similarity/difference across scRNA-seq datasets over numerous technical factors, biological features, expression profiles and related pathways. By comparing technical and biological metrics across datasets, scRNABatchQC enables the detection of systematic errors, batch effects or outlier samples. It will greatly improve quality control and reproducibility analysis for single-cell RNA sequencing.

2 Implementation

scRNABatchQC is written in R. It is easy to implement even for users with limited programming experience. There is only one required input, gene-by-cell count matrices, which can be supplied by any delimited files or compressed files (ending gz or .bz2), or read from 10X, SingleCellExperiment or Seurat v3 object. Besides, there are optional arguments that users can specify or adjust, such as the organism, the number of highly variable genes (HVGs), the number of principle components (PCs), scale factor, etc.scRNABatchQC summarizes QC report in one html file, which includes six sections: Overview, QC summary, Technical View, Biological View, Expression Similarity and Pairwise Difference (Fig. 1). The Overview gives a brief introduction of the software. QC summary provides a table listing a variety of QC metrics, such as the number of total counts/cells/genes, the cutoff for filtering cells, the number of cells removed due to low quality in each sample. Technical View presents diagnostic graphics on 11 technical features, such as the distribution of total counts and the variance explained by total counts in each sample, mean-variance trend, etc. Biological View compares the HVGs, the genes related to one specific principal component (PC-related genes), and their enriched pathways if the organism is supported by WebGestalt (Zhang *et al.*, 2005). Expression Similarity provides the global expression correlation across datasets, and two low-dimensional embedding of all cells, principal component analysis and t-distributed stochastic nearest-neighbor embedding. Pairwise Difference identifies differentially expressed genes between all pairs of samples and pathways associated with these genes (Methods Description in the Supplementary File S1). In addition to the html file, scRNABatchQC stores the gene-by-cell count matrices and QC metadata in SingleCellExperiment objects, which ensures the output of scRNABatchQC compatible with other Bio-conductor workflows.

3 Application

For demonstration, we used scRNA-seq data of mouse retinal bipolar cells, which includes a total of 44 994 cells in 6 replicates prepared from 2 experimental batches (Shekhar *et al.*, 2016). Batch 1 has 4 replicates (S1–S4), while batch 2 consists of 2 replicates (S5 and S6). Supplementary File S2 is the output generated by scRNABatchQC.

The report summarizes that batch 1 has ∼4500 cells/replicate, and batch 2 has ∼13 000 cells/replicate. Compared to batch 1, batch 2 has more cells with high percentage of mitochondrial RNA genes (the maximum percentage is 70%∼80%), which were removed from downstream analysis due to low quality (Supplementary File S2: QC Summary). Batches 1 and 2 present different distributions on all 11 technical features (Supplementary File S2: Figs S1–11), suggesting the existence of batch effects. For example, batch 2 shows a more rapid increase in the expression cumulative plot, suggesting a library with lower complexity (Supplementary File S2: Fig. S5). Biological View illustrates that six samples are very similar in their HVGs, all enriched for processes related to photo-transduction. These results suggest that HVGs mainly capture the biological variations across cell types, and six samples share similar biological heterogeneity and cellular compositions. Although being very similar, samples are still clustered by their batch, indicating that batch effects have some minor effect on variations within samples (Supplementary File S2: Figs S12 and 13). Pairwise Difference identifies differentially expressed genes between two batches, such as Xist, Hopx, mt-Rnr1 and mt-Rnr2 (Supplementary File S2: Fig. S19). No enriched pathways are detected for these differential genes, therefore Supplementary Figure S20 is not generated. Xist, Hopx are sex-related genes, while mt-Rnr1 and mt-Rnr2 are mitochondrial RNA genes. These results suggest that differences in the sex populations and library preparation between the two batches are likely to contribute to batch effects, which is consistent with the original paper (Shekhar *et al.*, 2016).

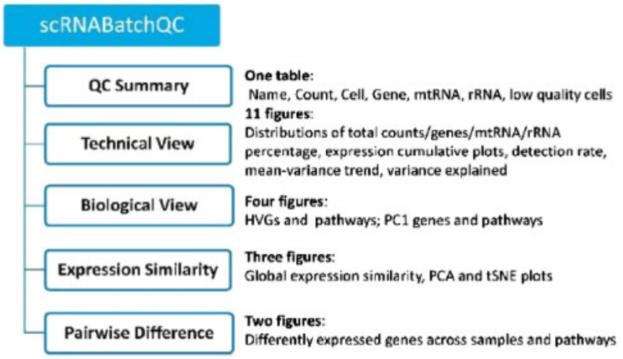
scRNABatchQC
QC Summary
One table:
Name, Count, Cell, Gene, mtRNA, rRNA, low quality cells
11 figures:
Distributions of total counts/genes/mtRNA/rRNA percentage, expression cumulative plots, detection rate, mean-variance trend, variance explained
Technical View
Biological View
Four figures:
HVGs and pathways; PC1 genes and pathways
Expression Similarity
Three figures:
Global expression similarity, PCA and tSNE plots
Pairwise Difference
Two figures:
Differently expressed genes across samples and pathways

Fig. 1. The outline of scRNABatchQC

4 Conclusion

Large-scale projects, such as Human Cell Atlas, are now generating comprehensive collections of scRNA-seq datasets. Understanding the existence and the sources of experimental noise is very important to integration and interpretation of the data. scRNABatchQC, a quality assessment tool over numerous technical and biological features, not only provides a global overview of all experiments, but also enables the examination of technical or biological origin of discrepancies between experiments and detect possible outliers and batch effects. scRNABatchQC can also be applied to other single cell experiments, such as single nuclei RNA-seq.

Funding

This work was supported by the National Cancer Institute grants (U2C CA233291 and U54 CA217450).

Conflict of Interest: none declared.

References

Carter,R.A. *et al.* (2018) A single-cell transcriptional atlas of the developing murine cerebellum. *Curr. Biol.*, **28**, 2910–2920.e2.

Ilicic,T. *et al.* (2016) Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.*, **17**, 29.

Jiang,P. *et al.* (2016) Quality control of single-cell RNA-seq by SinQC. *Bioinformatics*, **32**, 2514–2516.

Karaayvaz,M. *et al.* (2018) Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.*, **9**, 3588.

Lun,A.T. *et al.* (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, **5**, 2122.

McCarthy,D.J. *et al.* (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186.

Shekhar,K., *et al.* (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323.e30.

Tian,L. *et al.* (2018) scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput. Biol.*, **14**, e1006361.

Zhang,B. *et al.* (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–W748.