Bioinformatics, 32(16), 2016, 2514–2516
doi: 10.1093/bioinformatics/btw176
Advance Access Publication Date: 10 April 2016
Applications Note

OXFORD

Genome analysis

Quality control of single-cell RNA-seq by SinQC

Peng Jiang[1,*], James A. Thomson[1,2,3] and Ron Stewart[1,*]

[1]Regenerative Biology Laboratory, Morgridge Institute for Research, Madison, WI 53707, USA, [2]Department of Cell and Regenerative Biology, University of Wisconsin, Madison, WI 53706, USA and [3]Department of Molecular, Cellular and Developmental Biology, University of California, Santa Barbara, CA 93106, USA

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on October 23, 2015; revised on March 7, 2016; accepted on March 28, 2016

Abstract

Summary: Single-cell RNA-seq (scRNA-seq) is emerging as a promising technology for profiling cell-to-cell variability in cell populations. However, the combination of technical noise and intrinsic biological variability makes detecting technical artifacts in scRNA-seq samples particularly challenging. Proper detection of technical artifacts is critical to prevent spurious results during downstream analysis. In this study, we present 'Single-cell RNA-seq Quality Control' (SinQC), a method and software tool to detect technical artifacts in scRNA-seq samples by integrating both gene expression patterns and data quality information. We apply SinQC to nine different scRNA-seq datasets, and show that SinQC is a useful tool for controlling scRNA-seq data quality.

Availability and Implementation: SinQC software and documents are available at http://www.morgridge.net/SinQC.html

Contacts: PJiang@morgridge.org or RStewart@morgridge.org

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

1 Introduction

Single-cell RNA-seq (scRNA-seq) provides a relatively unbiased approach to characterize and dissect the heterogeneity of cells in complex mixtures (Eberwine *et al.*, 2014). However, one of the major challenges of this technology is distinguishing biological heterogeneity from technical artifacts (cells with substantial technical noise that makes their gene expression patterns distinguished from other cells) (Sandberg, 2014; Stegle *et al.*, 2015).

To detect potential technical artifacts in scRNA-seq, previous studies have used various strategies that can be generally grouped into three categories. The first category involves using housekeeping genes to perform QC. For example, cells not expressing housekeeping genes (e.g. Actb, Gapdh) or abnormally expressing them are filtered out (Ting *et al.*, 2014; Treutlein *et al.*, 2014). The assumption of methods in this category is that housekeeping genes are highly and consistently expressed, which is not necessarily true for single cells. For example, a study using single-cell qPCR not only showed that the gene expression of housekeeping genes had high variation between individual cells but also that gene expression expression of housekeeping genes can even distinguish cell types (Oyolu *et al.*, 2012). Thus, a reliance on housekeeping genes to perform QC can result in removing cells with real biological variation. The second category involves using overall gene expression patterns to define technical artifacts. For example, cells are excluded from further analysis if they cluster separately from the rest of the cells (Zeisel *et al.*, 2015) or if their median expression values fall below a certain threshold (Pollen *et al.*, 2014). The major problem of the methods in this category is that they can potentially discard cells with real biological variation. The third category involves using the number of genes detected (per some defined expression threshold) and/or the reads mapping rate to define technical artifacts (Kumar *et al.*, 2014). However, the number of genes detected and mapping rate vary among experiments depending on the quality of a particular library, cell type, or RNA-protocol. Hence, the cutoff settings are typically arbitrary. Thus, although single-cell approaches hold great promise in exploring heterogeneity within a cell population or complex mixture, QC still remains a major challenge (Stegle *et al.*, 2015).

In this study, we present 'Single-cell RNA-seq Quality Control' (SinQC), a method and software tool for detecting technical artifacts in scRNA-seq samples. SinQC assumes that if gene expression outliers are also associated with poor sequencing library quality (poor

data quality, e.g. low mapped reads, low mapping rate or low library complexity), then they are more likely to be technical artifacts than to be cells with real biological variation.

2 Method

A detailed description of the SinQC algorithm can be found in Supplementary Data. Briefly, given a batch of scRNA-seq data, SinQC first uses gene expression patterns to detect outliers. SinQC assumes that gene expression outliers contain both cells with real biological variation and technical artifacts, but the rest of the cells (main population cells) in general, are more likely to contain good quality cells. Thus, SinQC uses cells of the main population as controls to estimate data quality cutoffs and a corresponding false positive rate (FPR). For each sample, SinQC calculates two data quality meta-scores: Minimal Quantile Score (MQS) and Weighted Combined Quality Score (WCQS) by combining a set of data quality metrics (total number of mapped reads, mapping rate and library complexity). These two data quality meta-scores represent whether a sample has significant deficiency in any of the three quality metrics or the overall quality metrics are low, respectively. SinQC determines these two data quality meta-score cutoffs by allowing a minimal fraction (user-defined) of cells of the main population to fail to pass these cutoffs. The technical artifacts are defined as gene expression outliers with poor data quality (Supplementary Fig. S1A). A more detailed and comprehensive study of SinQC (e.g. comparison with other QC methods) can be found in Supplementary Data (Supplementary results and discussion).

3 Results and discussion

We applied SinQC to a highly heterogeneous scRNA-seq dataset containing 301 cells (mixture of 11 different cell types) (Pollen *et al.*, 2014). SinQC detected 12 technical artifacts (FPR < 5%). These 12 artifacts showed a significantly lower mapping rate but not fewer mapped reads nor lower library complexity if compared with the QC pass cells (Supplementary Fig. S2). We calculated the number of genes detected (TPM > 1) for each cell. The artifacts detected have significantly fewer genes detected if compared with QC pass cells ($P = 3.83e\text{-}07$, 1-sided Wilcoxon rank sum test; Supplementary Fig. S3). As shown in Figure 1, technical artifacts detected by SinQC overall have fewer genes detected and/or lower mapping rates, which is similar to using a 'genes detected and/or mapping rate' method to do quality control (Kumar *et al.*, 2014). However, SinQC has the advantage of not having to arbitrarily choose thresholds for the number of genes detected or mapping rate. Moreover, SinQC uses data quality meta-scores (Section 2) instead of only using mapping rate to represent data quality for each sample. The following examples will further demonstrate that integrating more universal data quality metrics is helpful to detect technical artifacts.

We next applied SinQC to another eight scRNA-seq datasets, including six batches of human H1 ES cells (H1-Exp1 = 72 cells, H1-Exp2 = 81 cells, H1-Exp3 = 75 cells, G1 = 91 cells, S = 80 cells and G2 = 76 cells) (Leng *et al.*, 2015) and two mouse datasets (ES = 48 cells and MEF = 44 cells) (Islam *et al.*, 2011). We ran SinQC separately for each dataset. As shown in Supplementary Figure S4, the technical artifacts (FPR < 5%) identified by SinQC either have significantly fewer mapped reads, and/or lower mapping rate, and/or lower library complexity compared with cells that pass QC. Among eight low-heterogeneity scRNA-seq datasets tested, all except Human G1 show significantly fewer number of genes detected in

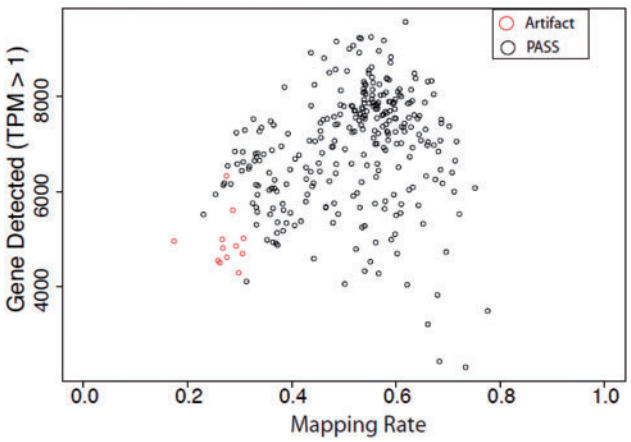

Fig. 1. Technical artifacts detected by SinQC (FPR < 5%) in a highly heterogeneous dataset containing a mixture of 11 cell types

artifacts if compared with the QC pass cells (Supplementary Table S1). Human G1 shows marginal overlap of 95% of CI between artifacts and the QC pass cells (Supplementary Table S1). This indicates that artifacts identified by SinQC overall have lower number of genes detected and are also associated with poor data quality. The technical artifacts detected by SinQC overall have fewer genes detected if compared with QC pass cells. But this does not mean that the cells with fewer genes detected are technical artifacts. The number of genes detected is determined by both data quality and cell type (a detailed discussion can be found in Supplementary Data (Supplementary Results and Discussion)). By integrating both gene expression and data quality information, SinQC maximizes the probability that the technical artifacts are correctly detected while also minimizing the false positives by using cells of the main population as data quality controls.

If a single-cell RNA-seq experiment contains hundreds or thousands cells, it is likely that they are processed in several experimental batches. For our lab-generated human embryonic stem cell (ES) datasets (Leng *et al.*, 2015), we processed them in three different experimental batches (H1-Exp1 = 72 cells, H1-Exp2 = 81 cells and H1-Exp3 = 75 cells). We further compared the technical artifacts detected (FPR < 5%) if we run SinQC on each individual experiment alone or we run SinQC on pooled experimental batches. We applied SinQC to three pooled batches of human H1 single cell RNA-seq data and detect 15 technical artifacts (H1-Exp1 = 12, H1-Exp2 = 3, H1-Exp3 = 0) (Supplementary Fig. S9). However, if we run SinQC batch by batch, we detect not only these 15 artifacts but also 11 additional ones (Supplementary Table S1) (H1-Exp1 = 12, H1-Exp2 = 11, H1-Exp3 = 3). This suggests that SinQC is more sensitive if run batch by batch. This is because pooling batches will increase the diversity of the population being studied owing to batch effects in scRNA-seq datasets. Since SinQC uses relative measurements to determine data quality cutoffs, the increased diversity in pooled batches will relax the absolute data quality cutoffs thus allowing more gene expression outliers to pass these cutoffs.

We then further investigated the sensitivity and specificity of SinQC when different types of cell are mixed. We mixed two mouse datasets (48 ES cells and 44 MEF cells) (Islam *et al.*, 2011) in different ways to simulate datasets containing different portions of subpopulations. As shown in Supplementary Figure S5, the artifacts detected by SinQC using different combinations of datasets are overall consistent with each other. However, we observe that SinQC increases specificity at the cost of dropping sensitivity when the extent

of heterogeneity in a dataset is high. For example, if we ran SinQC on each individual ES or MEF dataset, we can detect more artifacts, if compared to running SinQC on pooled mixture datasets (e.g. 'All'). However, the two artifacts (ESC_46 and ESC_32) which were detected by pooled mixture datasets ('All') can be robustly detected by running SinQC either on ES and MEF datasets separately or on 'ES + 1/5 (MEF)' or '1/5 (ES) + MEF'. In highly heterogeneous cell populations, detecting technical artifacts carries a higher risk of dropping real biological variation cells. The increased specificity and decreased sensitivity of SinQC for highly heterogeneous cell populations can minimize the false positives.

Acknowledgements

We thank Joe Phillips, Ning Leng, Scott Swanson and Mackenzie Holland for editorial assistance.

Funding

This work is supported by Morgridge Institute for Research and NIH grant 5U01HL099773-02 (to J.A.T).

Conflict of Interest: none declared.

References

Eberwine,J. *et al*. (2014) The promise of single-cell sequencing. *Nat. Methods*, **11**, 25–27.

Islam,S. *et al*. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res*., **21**, 1160–1167.

Kumar,R.M. *et al*. (2014) Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, **516**, 56–61.

Leng,N. *et al*. (2015) Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods*, **12**, 947–950.

Oyolu,C. *et al*. (2012) Distinguishing human cell types based on housekeeping gene signatures. *Stem Cells*, **30**, 580–584.

Pollen,A.A. *et al*. (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol*., **32**, 1053–1058.

Sandberg,R. (2014) Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods*, **11**, 22–24.

Stegle,O. *et al*. (2015) Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, **16**, 133–145.

Ting,D.T. *et al*. (2014) Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep*., **8**, 1905–1918.

Treutlein,B. *et al*. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.

Zeisel,A. *et al*. (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.