*Bioinformatics*, 32(14), 2016, 2221–2223
doi: 10.1093/bioinformatics/btw174
Advance Access Publication Date: 15 March 2016
Applications Note

OXFORD

Gene expression

# Dr.seq: a quality control and analysis pipeline for droplet sequencing

**Xiao Huo†, Sheng'en Hu†, Chengchen Zhao and Yong Zhang***

School of Life Science and Technology, Shanghai Key Laboratory of Signaling and Disease Research, Tongji University, Shanghai 20092, China

*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: Ziv Bar-Joseph

Received on December 22, 2015; revised on February 26, 2016; accepted on March 25, 2016

## Abstract

**Motivation:** Drop-seq has recently emerged as a powerful technology to analyze gene expression from thousands of individual cells simultaneously. Currently, Drop-seq technology requires refinement and quality control (QC) steps are critical for such data analysis. There is a strong need for a convenient and comprehensive approach to obtain dedicated QC and to determine the relationships between cells for ultra-high-dimensional datasets.
**Results:** We developed Dr.seq, a QC and analysis pipeline for Drop-seq data. By applying this pipeline, Dr.seq provides four groups of QC measurements for given Drop-seq data, including reads level, bulk-cell level, individual-cell level and cell-clustering level QC. We assessed Dr.seq on simulated and published Drop-seq data. Both assessments exhibit reliable results. Overall, Dr.seq is a comprehensive QC and analysis pipeline designed for Drop-seq data that is easily extended to other droplet-based data types.
**Availability and Implementation**: Dr.seq is freely available at: http://www.tongji.edu.cn/~zhanglab/drseq and https://bitbucket.org/tarela/drseq
**Contact:** yzhang@tongji.edu.cn
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

In recent years, the advances of RNA-seq technology have allowed single-cell RNA-seq (scRNA-seq) to emerge as a powerful and popular technique to study the cell-to-cell expression variability of thousands genes (Tang *et al.*, 2009). However, the time and cost constraints of preparing libraries from many individual cells have prevented scRNA-seq for more broad applications. Recently, Drop-seq and inDrop, two new, easy, fast and low-cost technologies to quickly profile thousands of individual cells for RNA-seq, have been developed (Klein *et al.*, 2015; Macosko *et al.*, 2015). As both technologies are based on the combination of droplet-based microfluid and individual-cell barcode approaches, and their data structures are largely similar, we termed both technologies as Drop-seq in this study. Drop-seq only requires 12 h to prepare 10 000 single-cell libraries for sequencing (Macosko *et al.*, 2015). The breakthrough of this technology provides many applications, such as the deconstruction of a cell population, the detection of rare cell types and the inference of interactions between genes. However, Drop-seq technology still suffers from some restrictions. Currently, the mRNA capture rate, conversion efficiency and sequencing depth are not sufficient to cover thousands of genes from a large fraction of cells, leading to increased levels of noise, which can influence the interpretation of Drop-seq data. Therefore, quality control (QC) should be the first step in the analysis pipeline to assure the data quality and to provide a solid basis for subsequent analysis. Although there are already some QC algorithms for bulk RNA-seq (Wang *et al.*, 2012) and scRNA-seq (Grun *et al.*, 2015) data, a QC pipeline specifically for Drop-seq data is still necessary given the low capture efficiency, high dimensionality and sequenced cell number. Here, we present

Dr.seq, an integrated Drop-seq QC and analysis pipeline. Dr.seq can systematically evaluate Drop-seq data quality together with the visualization of unsupervised cell clustering.

## 2 Methods

In reads level QC and bulk-cell level QC, we regarded a Drop-seq data as a bulk-cell RNA-seq data and measured the general quality of the data. In the individual-cell level QC step, we grouped reads with identical cell barcode, and for each cell barcode, we calculated the unique reads count, the reads duplicate rate and the number of covered genes. We selected cell barcodes with covered gene numbers larger than a user-defined cutoff as those arising from STAMPs (single-cell transcriptomes attached to microparticles, i.e. informative single-cell transcriptomes), and used the cell-clustering level QC step to provide measurements for sample heterogeneity of the selected STAMP barcodes. See Supplementary Methods for details.

## 3 Results

The Dr.seq QC and analysis pipeline (Fig. 1) uses two paired sequencing files (FASTQ or SAM format) as input. One file contains transcript information, and the other contains cell barcode and UMI information. The transcript-sequencing file was aligned to the reference genome, and the mapped reads with high sequencing quality were kept. Dr.seq provides four groups of QC measurements (Supplementary Methods): (i) Reads level QC including quality, nucleotide composition and GC content of the reads. (ii) Bulk-cell level QC including gene body coverage and reads alignment summary (Supplementary Fig. S1). (iii) Individual-cell level QC including the distributions of reads duplicate rate, intron reads rate and covered gene number in selected STAMPs, for the evaluation of mRNA capture efficiency at individual-cell level (Supplementary Fig. S2). (iv) Cell-clustering level QC including gap statistics and silhouette score (Rousseeuw, 1987) for the evaluation of the sample heterogeneity (Supplementary Figs S3A and S3B). In addition to a QC report document describing the above QC measurements, Dr.seq also generated a series of analysis results, including: (i) expression index, (ii) paired-wise correlation table, (iii) PCA and t-SNE dimensional reduction output, (iv) cluster assignment for selected STAMPs and (v) visualization of t-SNE and clustering output (Supplementary Fig. S3C, Supplementary Methods).

To evaluate the performance of our pipeline, we simulated Drop-seq data from 10 cell types by sampling reads from bulk-cell RNA-seq data in 10 human cell lines (Supplementary Table S1). To assess the performance of cell clustering on the tolerance of low sequencing depth and small cell number in certain cell types, we simulated Drop-seq data with different sequencing depths per cell and different numbers of individual cells in certain cell types, respectively (Supplementary Methods in Supplementary file 1). To estimate the stability of cell clustering, Goodman-Kruskal's lambda index (Goodman and Kruskal, 1954; Tichy *et al.*, 2011) was conducted to compare the clustering results with predefined cell type labels. Dr.seq showed reliable cell clustering results with 1000 or more reads per cell and 20 or more individual cells per cell type (Supplementary Fig. S4).

We applied Dr.seq on published Drop-seq data in mouse retinal cells (Macosko *et al.*, 2015), and obtained a multifaceted and detailed QC reports (Supplementary file 2) together with a series of analysis results. The running time of each QC and analysis step of Dr.seq was also displayed (Supplementary Table S2).

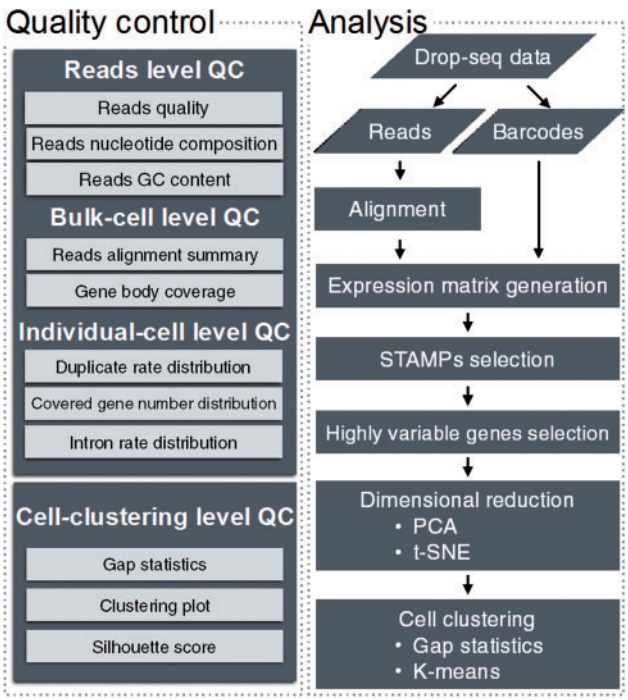


**Fig. 1.** Flowchart illustrating the Dr.seq pipeline with default parameters. The workflow of the Dr.seq pipeline includes QC and analysis components. The QC component contains reads level, bulk-cell level, individual-cell level and cell-clustering level QC

## 4 Conclusion

In summary, Dr.seq is specifically designed for QC and analysis of Drop-seq data. It takes standard-format input files via simple commands, reports informative QC measurements from four levels, and provides detailed analysis results. It also displays tolerance of low sequencing depth and small cell number in certain cell types. Besides, Dr.seq has the potential to be easily extended to other droplet-based data types, such as Drop-ChIP (Rotem *et al.*, 2015).

## Acknowledgements

We thank Jianxing Feng, Shaojuan Li, Qian Qin, Shenglin Mei and Yiqing Chen for their suggestions.

## Funding

This work was supported by National Natural Science Foundation of China (31571365, 31322031, 31371288) and Specialized Research Fund for the Doctoral Program of Higher Education (20130072110032).

*Conflict of Interest*: none declared.

## References

Goodman,L.A. and Kruskal,W.H. (1954) Measures of association for cross-classification. *J. Am. Stat. Assoc.*, **49**, 732–764.

Grun,D. *et al.* (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**, 251–255.

Klein,A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Rotem,A. *et al.* (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, **33**, 1165–1172.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Comput. Appl. Math.*, **20**, 53–65.

Tang,F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.

Tichy,L. *et al.* (2011) Evaluating the stability of the classification of community data. *Ecography*, **34**, 807–813.

Wang,L. *et al.* (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.