RESEARCH ARTICLE

Dr.seq2: A quality control and analysis pipeline for parallel single cell transcriptome and epigenome data

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

Translational Medical Center for Stem Cell Therapy & Institute for Regenerative Medicine, Shanghai East Hospital, School of Life Science and Technology, Shanghai Key Laboratory of Signaling and Disease Research, Tongji University, Shanghai, China

* yzhang@tongji.edu.cn

Citation: Zhao C, Hu S, Huo X, Zhang Y (2017) Dr.seq2: A quality control and analysis pipeline for parallel single cell transcriptome and epigenome data. PLoS ONE 12(7): e0180583. https://doi.org/10.1371/journal.pone.0180583

Editor: Zhaohui Qin, Emory University Rollins School of Public Health, UNITED STATES

Received: February 22, 2017

Accepted: June 16, 2017

Published: July 3, 2017

Data Availability Statement: The MARS-seq files were available from NCBI Gene Expression Omnibus (GEO) database under accession GSE54006. The 10x genomics datasets were available from 10x genomic data support (https://support.10xgenomics.com/single-cell/datasets). The scATAC-seq datasets were available from NCBI Gene Expression Omnibus (GEO) database under accession GSE65360. The Drop-seq samples were available from NCBI Gene Expression Omnibus (GEO) database under accession GSM1626793.

Abstract

An increasing number of single cell transcriptome and epigenome technologies, including single cell ATAC-seq (scATAC-seq), have been recently developed as powerful tools to analyze the features of many individual cells simultaneously. However, the methods and software were designed for one certain data type and only for single cell transcriptome data. A systematic approach for epigenome data and multiple types of transcriptome data is needed to control data quality and to perform cell-to-cell heterogeneity analysis on these ultra-high-dimensional transcriptome and epigenome datasets. Here we developed Dr.seq2, a Quality Control (QC) and analysis pipeline for multiple types of single cell transcriptome and epigenome data, including scATAC-seq and Drop-ChIP data. Application of this pipeline provides four groups of QC measurements and different analyses, including cell heterogeneity analysis. Dr.seq2 produced reliable results on published single cell transcriptome and epigenome datasets. Overall, Dr.seq2 is a systematic and comprehensive QC and analysis pipeline designed for parallel single cell transcriptome and epigenome data. Dr.seq2 is freely available at: http://www.tongji.edu.cn/~zhanglab/drseq2/ and https://github.com/ChengchenZhao/DrSeq2.

Introduction

To better understand cell-to-cell variability, an increasing number of transcriptome technologies, such as Drop-seq [1, 2], Cyto-seq [3], 10x genomics [4], MARS-seq [5], and epigenome technologies, such as Drop-ChIP [6], single cell ATAC-seq (scATAC-seq) [7], have been developed in recent years. These technologies can easily provide a large amount of single cell transcriptome information or epigenome information at minimal cost, which makes it possible to perform analysis of cell heterogeneity on the transcriptome and epigenome levels, deconstruction of a cell population, and detection of rare cell populations. However, different single cell transcriptome technologies have their own features given their specific experimental design, such as cell sorting methods, RNA capture rates, and sequencing depths. But the methods and software such as Dr.seq [8] were developed for one single cell data type with certain functions

Funding: This work was supported by National Natural Science Foundation of China (31571365, 31322031 and 31371288), National Key Research and Development Program of China (2016YFA0100400), Specialized Research Fund for the Doctoral Program of Higher Education (20130072110032), and Program of Shanghai Academic Research Leader (17XD1403600). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

(S1 File). Furthermore, the quality control step of single cell epigenome data is more challenging than for transcriptome data given the amplification noise caused by the limit number of DNA copy in single cell epigenome experiments. But few quality control and analysis method was developed specific for single cell epigenome data. Thus a comprehensive QC pipeline suitable for multiple types of single cell transcriptome data and epigenome data is urgently needed. Here, we provide Dr.seq2, a QC and analysis pipeline for multiple types of parallel single cell transcriptome and epigenome data, including recently published scATAC-seq data. Dr.seq2 can systematically generate specific QC, analyze, and visualize unsupervised cell clustering for multiple types of single cell data. For single cell transcriptome data, the QC steps of Dr.seq2 are primarily derived from Dr.seq [8] and the output of Dr.seq2 on these data will not be described in details in this paper.

Materials and methods

Drop-seq data

The Drop-seq samples were obtained from NCBI Gene Expression Omnibus (GEO) database under accession GSM1626793.

MARS-seq data

The MARS-seq samples were obtained from NCBI Gene Expression Omnibus (GEO) database under accession GSE54006. These samples were combined as a MARS-seq dataset and analyzed by Dr.seq2 using three different dimension reduction methods.

10x genomics data

The 10x genomics datasets were obtained from 10x genomic data support (https://support.10xgenomics.com/single-cell/datasets). The sample named "50%: 50% Jurkat: 293T Cell Mixture" was analyzed by Dr.seq2 using three different dimension reduction methods.

scATAC-seq data

The scATAC-seq datasets were obtained from NCBI Gene Expression Omnibus (GEO) database under accession GSE65360. We combined 288 scATAC datasets (GSM1596255 ~ GSM1596350, GSM1596735 ~ GSM1596830, GSM1597119 ~ GSM1597214) from three cell types and analyzed by Dr.seq2. Cell clustering was conducted for the combined scATAC-seq dataset. We also plotted the cell type labels using different colors on the clustering plot and found consistent classifications with the clustering results.

Drop-ChIP data

The Drop-ChIP datasets were obtained from NCBI Gene Expression Omnibus (GEO) database under accession GSE70253.

Implementation of Dr.seq2

Dr.seq2 was implemented using Python and R. Linux or MacOS environment with Python (version = 2.7) and R (version> = 2.14.1) was suitable for Dr.seq2. It was distributed under the GNU General Public License version 3 (GPLv3). A detailed tutorial was provided on the Dr.seq2 webpage (http://www.tongji.edu.cn/~zhanglab/drseq2) and source code of Dr.seq2 was available on github (https://github.com/ChengchenZhao/DrSeq2).

Quality control components

Dr.seq2 conducted four groups of QC measurements on single cell epigenome data: (i) reads level QC; (ii) bulk-cell level QC; (iii) individual-cell level QC; and (iv) cell-clustering level QC.

Reads level QC and bulk-cell level QC. We used a published package called RseQC [9] for reads level QC of Drop-ChIP data and scATAC-seq data to measure the general sequence quality. In bulk-cell level QC, a Drop-ChIP dataset (or scATAC-seq datasets combined from several scATAC-seq samples) was regarded as a bulk-cell ChIP-seq (or bulk-cell ATAC-seq) data. Next, "combined peaks" were detected with total reads from the "bulk-cell" data using MACS[10] for output and the following steps. Different MACS parameters were applied to Drop-ChIP and scATAC-seq data. We used the published package CEAS to measure the performance of ChIP for ChIP-seq data (or Tn5 digestion for scATAC-seq data) [11].

Individual-cell level QC. The reads number distribution was calculated by counting the number of reads assigned to each single cell. A single cell referred to a unique cell barcode in Drop-ChIP data. For scATAC-seq data, the peak number in each cell was defined as the number of "combined peaks" occupied by the reads in the cell. The distribution of different peak numbers in each cell indicated the different amount of information the cell contains.

Cell-clustering level QC. Cells were first clustered based on their occupancy of "combined peaks" using hierarchical clustering. Next, cells in each cluster were regarded as the same cell type (or same cell sub-type), and reads from the same cell type were merged. For each cell type, unique peaks from other cell types were defined as specific peaks in this cell type. Specific peaks in different cell types were displayed with different colors according to genomic locations. Silhouette method is used to interpret and validate the consistency within clusters defined in previous steps.

Note that reads with no overlap with "combined peaks" were discarded in this step and the following steps. Clusters containing less than 3 single cells were also discarded.

Simulation of scATAC-seq datasets

To measure the tolerance of Dr.seq2 for low sequencing depth and small numbers of cells of a certain cell type, we simulated datasets from 3 cell types with different cell proportions and sequencing depths using scATAC-seq data (Table 1). To test the effect of low sequencing depth, we sampled the reads count from 10,000 reads to 100,000 reads for each cell and compared these results with the Goodman-Kruskal's lambda index [12] of clustering results using cells with a certain number of reads.

To test the effect of low cell numbers of a certain cell type (defined as a target cell type) on cell clustering, we defined 1 of the 3 cell types as the "target cell type", whereas the other cell types were defined as the "regular cell type", and sampled cells with following compositions: 10:70:70 (10 for target cell type, 70 for the two regular cell types), 15:67:67, 20:65:65, 25:62:62, 30:60:60, 35:57:57, 40:55:55, 45:52:52 and 50:50:50. Then, we called "combined peaks" and clustered cells on the simulated dataset. The Goodman-Kruskal's lambda index [12] was

Table 1. Meta data and accession ID for the scATAC-seq data used in simulation for pipeline tolerance evaluation.

Accession ID	Cell line	Cell type	Target/regular cell
GSM1596255~GSM1596350	H1	human embryonic stem cell line	Target
GSM1596735~GSM1596830	GM12878	lymphoblastoid cells	Regular
GSM1597119~GSM1597214	K562	chronic myeloid leukemia cells	Regular

We defined the 1 out of 3 cell types as "target cell type", while the other cell types were defined as "regular cell type".

https://doi.org/10.1371/journal.pone.0180583.t001 calculated to evaluate the cell clustering performance. The average Goodman-Kruskal's lambda index and 95% confidence intervals were calculated from 20 simulations.

Results and discussion

Dr.seq2 overview

The Dr.seq2 QC and analysis pipeline is suitable for both single cell transcriptome data and epigenome data. Multiple types of single cell transcriptome data (including scRNA-seq, Drop-seq, inDrop, MARS-seq and 10x genomics data) and epigenome data (including scATAC-seq and Drop-ChIP) are acceptable for Dr.seq2 with relevant functions (S1 Fig).

Recently many methods and software were developed for single cell RNA-seq data. However most of them were suitable for certain data types with limited functions. We compared the major function of Dr.seq2 to existing state-of-the-art methods (Table 2). Dr.seq2 provides two advantages: 1) Dr.seq2 supports different types of single cell transcriptome data and single cell epigenome data. 2) Dr.seq2 provides both multifaceted QC reports and cell clustering

Table 2. Comparison of functions between Dr.seq2 and other software developed for single cell transcriptome data.

Name	Supporting single cell epigenome data (e.g. scATAC and Drop-ChIP)	Reads level QC	Individual cell level QC	Highly variable gene detection	Noise reduction	Informative cell selection	Cell clustering (sub cell type identification)	Differential expressed gene detection	Pseudo-temporal ordering	Reference
Dr.seq2	✓	✓	✓	✓	✓	✓	✓	✓		-
Dr.seq		✓	✓	✓	✓	✓	✓			[8]
BASiCS				✓	✓		✓			[13]
scLVM				✓	✓		✓	✓		[14]
SINCERA		✓	✓	✓	✓		✓	✓		[15]
OEFinder				✓	✓					[16]
ZIFA					✓		✓			[17]
Destiny					✓		✓			[18]
SNN-Cliq					✓		✓			[19]
RaceID					✓		✓	✓		[20]
SCUBA					✓		✓	✓	✓	[21]
BackSPIN							✓	✓		[22]
PAGODA				✓	✓		✓			[23]
MAST								✓		[24]
SCDE							✓	✓		[25]
scDD								✓		[26]
Monocle							✓		✓	[27]
Waterfall							✓		✓	[28]
Sincell							✓		✓	[29]
Oscope							✓		✓	[30]
Wanderlust							✓		✓	[31]
CellTree							✓		✓	[32]
SinQC		✓	✓		✓					[33]
ASAP					✓		✓	✓		[34]

We compare the major function of Dr.seq2 to existing state-of-the-art methods. Each column shows different functions of these methods and software.

https://doi.org/10.1371/journal.pone.0180583.t002

results. Then We used the simulated single cell RNA-seq data from seven RNA-seq datasets from ENCODE (S2 File) to estimate the performance of our Dr.seq2 pipeline (using different dimensional reduction methods: SIMLR and t-SNE) in cell clustering comparing to three existing methods (SINCERA, SNN-Cliq, BackSPIN). We applied these five methods on ten datasets with different numbers of reads per cell range from 100 to 10,000 to measure the accuracy and time cost of each method on different sequencing depth. SIMLR shows more accurate clustering results than t-SNE on the datasets with small number of reads per cell and comparable clustering results on the datasets with large number of reads per cell. And Dr.seq2 (using either SIMLR or t-SNE) shows better clustering accuracy than SNN-Cliq, and comparable clustering accuracy with BackSPIN and SINCERA on the datasets with large number of reads per cell. On the datasets with small number of reads per cell, SINCERA clustering result shows better accuracy than Dr.seq2 (using either SIMLR or t-SNE) and SNN-Cliq. However SINCERA takes a great mount of time on all these datasets comparing with Dr.seq2. As for BackSPIN, it does not support for these datasets with small number of reads per cell. Overall, Dr.seq2 (using either SIMLR or t-SNE) provides reliable cell clustering results with acceptable time cost (S2 Fig).

QC and analysis workflow

Dr.seq2 uses raw sequencing files in FASTQ format or alignment results in SAM/BAM format as input with relevant commands and generates four steps of QC measurements and analysis results (Fig 1).

For transcriptome data, the QC steps of Dr.seq2 are primarily derived from Dr.seq [8]. However, almost all data types are now supported, and more dimension reduction methods, including PCA, t-SNE and SIMLR[35], are supported. For single cell epigenome data, technologies like scATAC-seq and Drop-ChIP are increasingly common. However few quality control and analysis approaches have been developed for these data. Dr.seq2 conducts QC measurements on single cell epigenome data from four aspects: (i) reads level QC, including sequence quality, nucleotide composition and GC content of reads inherited from previous work; (ii) bulk-cell level QC, including genomic distribution of "combined peaks" and average profile on regulatory regions; (iii) individual-cell level QC, including the distribution of the number of reads and the peak number distribution; and (iv) cell-clustering level QC, including Silhouette score[36] and cell type-specific peak detection.

Cell clustering for different single cell transcriptome data types using different dimension reduction methods

We applied our pipeline to three different types of single cell transcriptome data (Drop-seq, MARS-seq and 10x genomics data) using three different dimension reduction methods (PCA, t-SNE and SIMLR[35]) to evaluate the performance of Dr.seq2 on different types of single cell transcriptome data (Fig 2). Due to the different distance calculation method and kernel function the method used, Dr.seq2 represented cluster results from different dimensions.

Bulk-cell level QC of scATAC-seq data to measure the performance of Tn5 digestion

To evaluate the performance of Dr.seq2 on single cell epigenome data, we combined 288 scATAC datasets (GSM1596255 ~ GSM1596350, GSM1596735 ~ GSM1596830, GSM1597119 ~ GSM1597214) from three cell types and applied Dr.seq2 to it. "Combined peaks" were detected

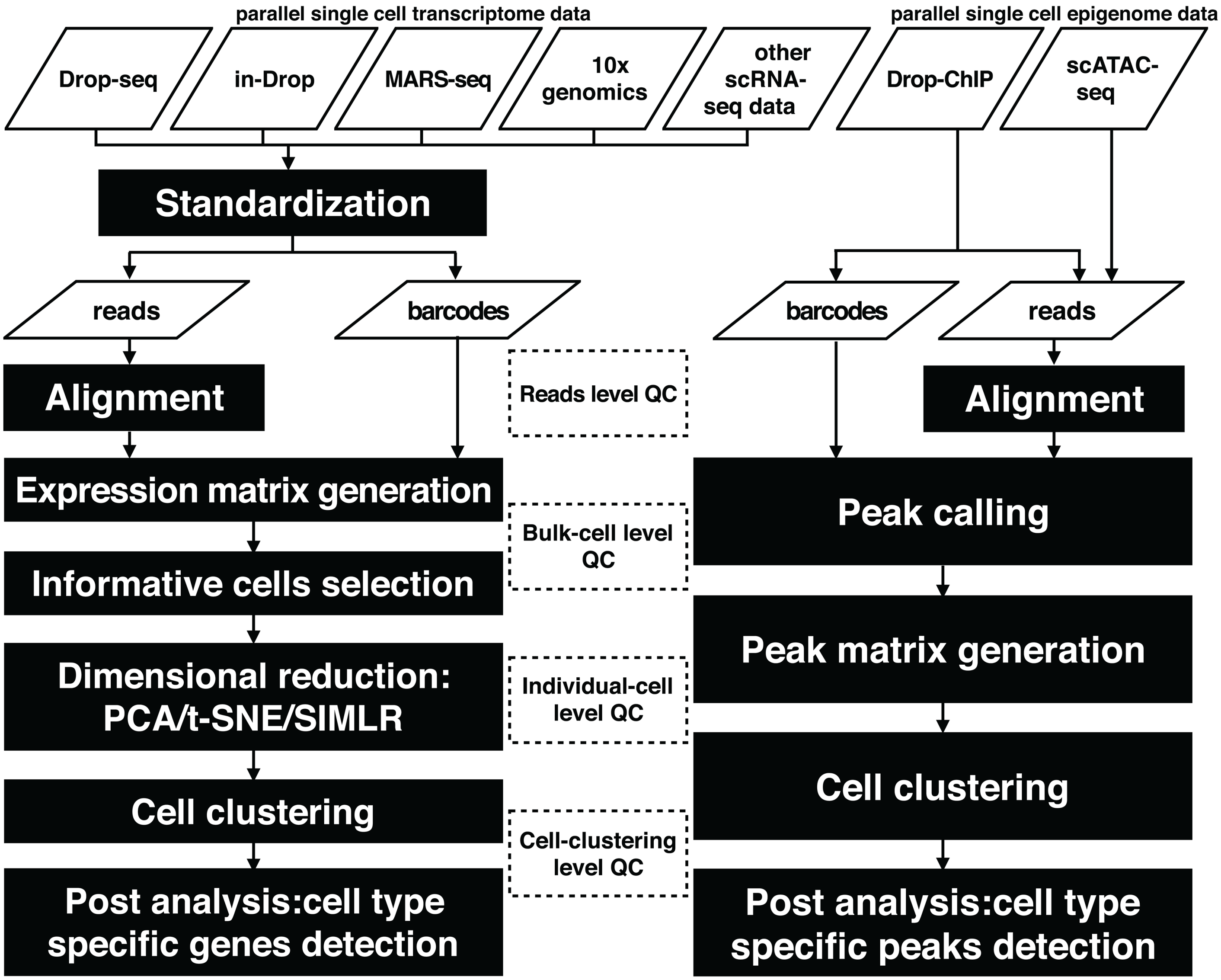

Fig 1. Flowchart illustrating the Dr.seq2 pipeline with default parameters. The workflow of the Dr.seq2 pipeline includes QC and analysis components for parallel single cell transcriptome and epigenome data. The QC component contains reads level, bulk-cell level, individual-cell level and cell-clustering level QC.

https://doi.org/10.1371/journal.pone.0180583.g001

with total reads from the combined dataset using MACS for output and the following steps. We measured the scATAC data quality in bulk-cell level from 4 aspects (Fig 3): 1) Peak distribution on each chromosome; 2) Open regions distributed over the genome along with their scores; 3) Average profiling on different genomic features; 4) Fragment length distribution. The peak distribution on each chromosome and the open region distributed over the genome showed the general quality of Tn5 digestion. The average profiling on different genomic features represented the quality of Tn5 digestion around specific regions. And the periodicity fragment length distribution indicated factor occupancy and nucleosome positions due to different Tn5 digestion degrees.

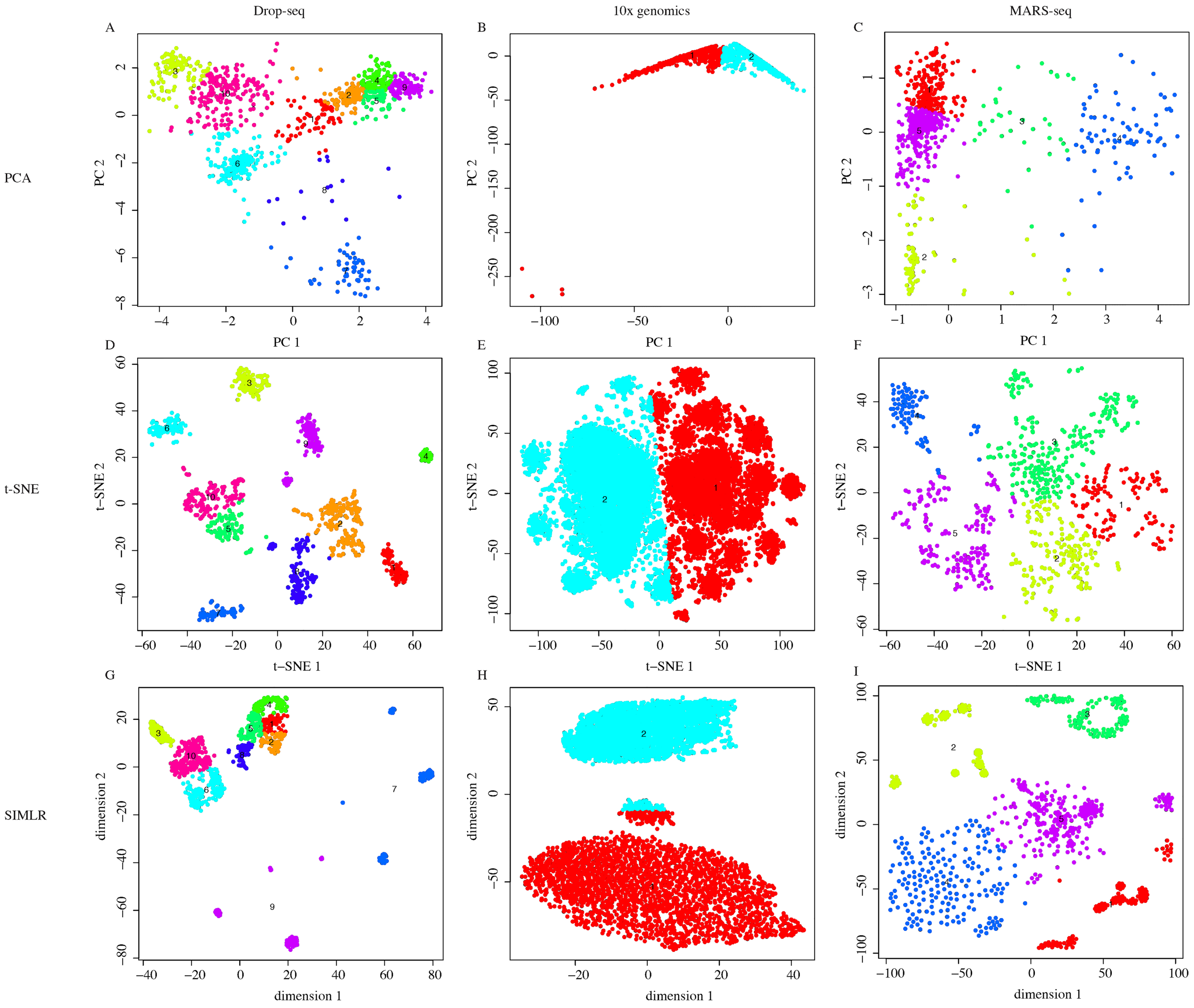

Fig 2. Dimensional reduction results for different single cell transcriptome data types. (A-I) Cell clustering results using dimensional reduction methods (PCA, t-SNE and SIMLR) on different types of single cell transcriptome data (Drop-seq, 10x genomics and MARS-seq).

https://doi.org/10.1371/journal.pone.0180583.g002

Cell clustering for scATAC-seq datasets with three clusters that were consistent with the cell type labels

To measure the cell clustering performance of Dr.seq2 on epigenome data, cells from the combined scATAC-seq dataset were firstly clustered based on their occupancy of "combined peaks" using hierarchical clustering. Then cell type labels were marked by different colors according to the original cell type information (red stand for H1 cells, yellow stand for GM12878 cells and blue stand for K562 cells). Cells were clearly separated into different groups that were consistent with the cell type labels by Dr.seq2 (Fig 4A).

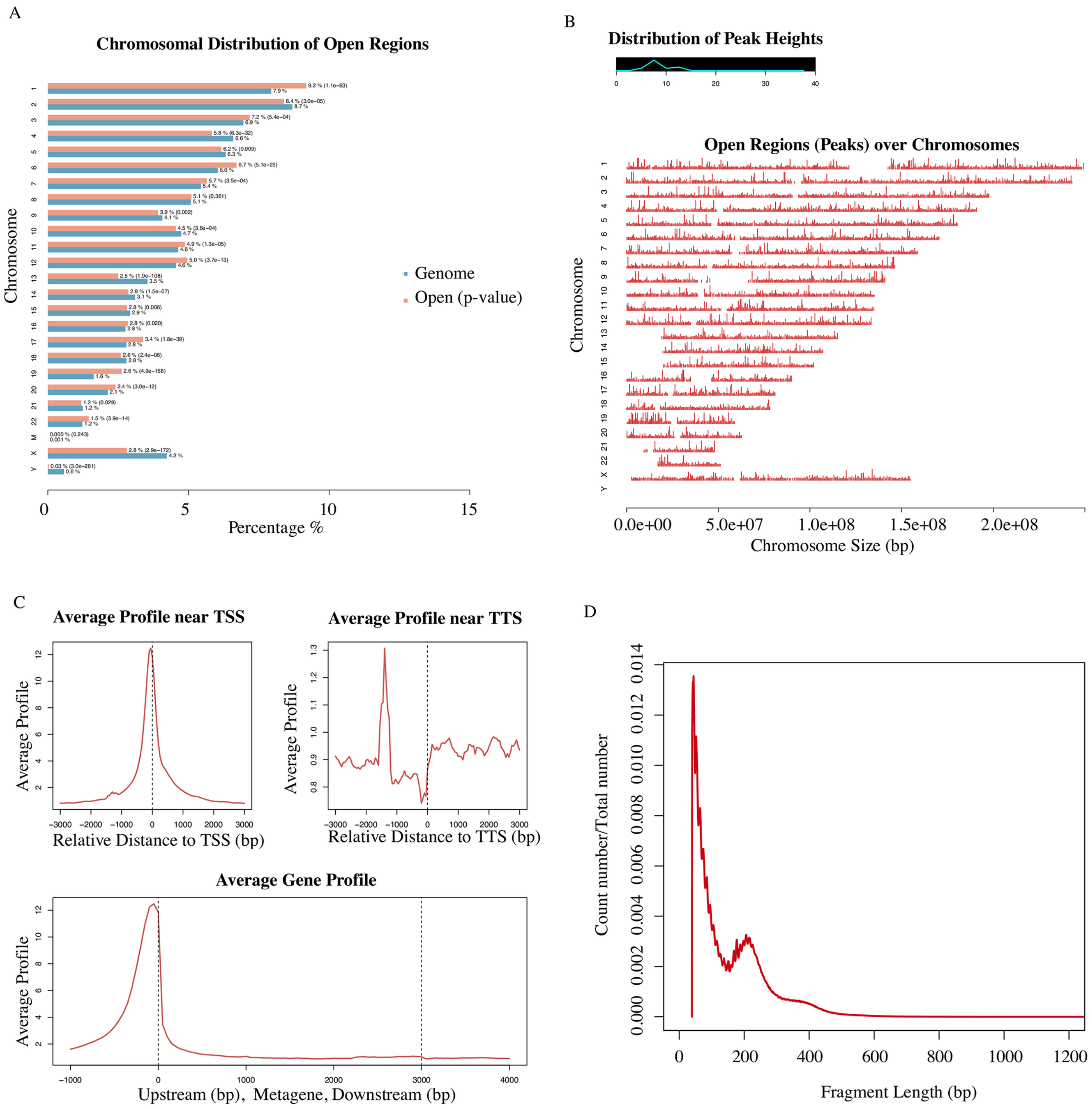

Fig 3. Bulk-cell level QC for scATAC-seq datasets. A) Peak region number distribution on each chromosome. The blue bars represent the percentages of the whole tiled or mappable regions in the chromosomes (genome background) and the red bars showed the percentages of the whole open region. These percentages are also marked right next to the bars. P-values for the significance of the relative enrichment of open regions with respect to the gnome background are shown in parentheses next to the percentages of the red bars. **B)** Open region distribution over the genome along with their scores or peak heights. The line graph on the top left corner illustrates the distribution of peak score. The x-axis of the main plot represents the actual chromosome sizes. **C)** Average profiling on different genomic features. The panels on the first row display the average enrichment signals around TSS and TTS of genes, respectively. The bottom panel represents the average signals on the meta-gene of 5 kb. **D)** Red line shows number distribution of different fragment length.

https://doi.org/10.1371/journal.pone.0180583.g003

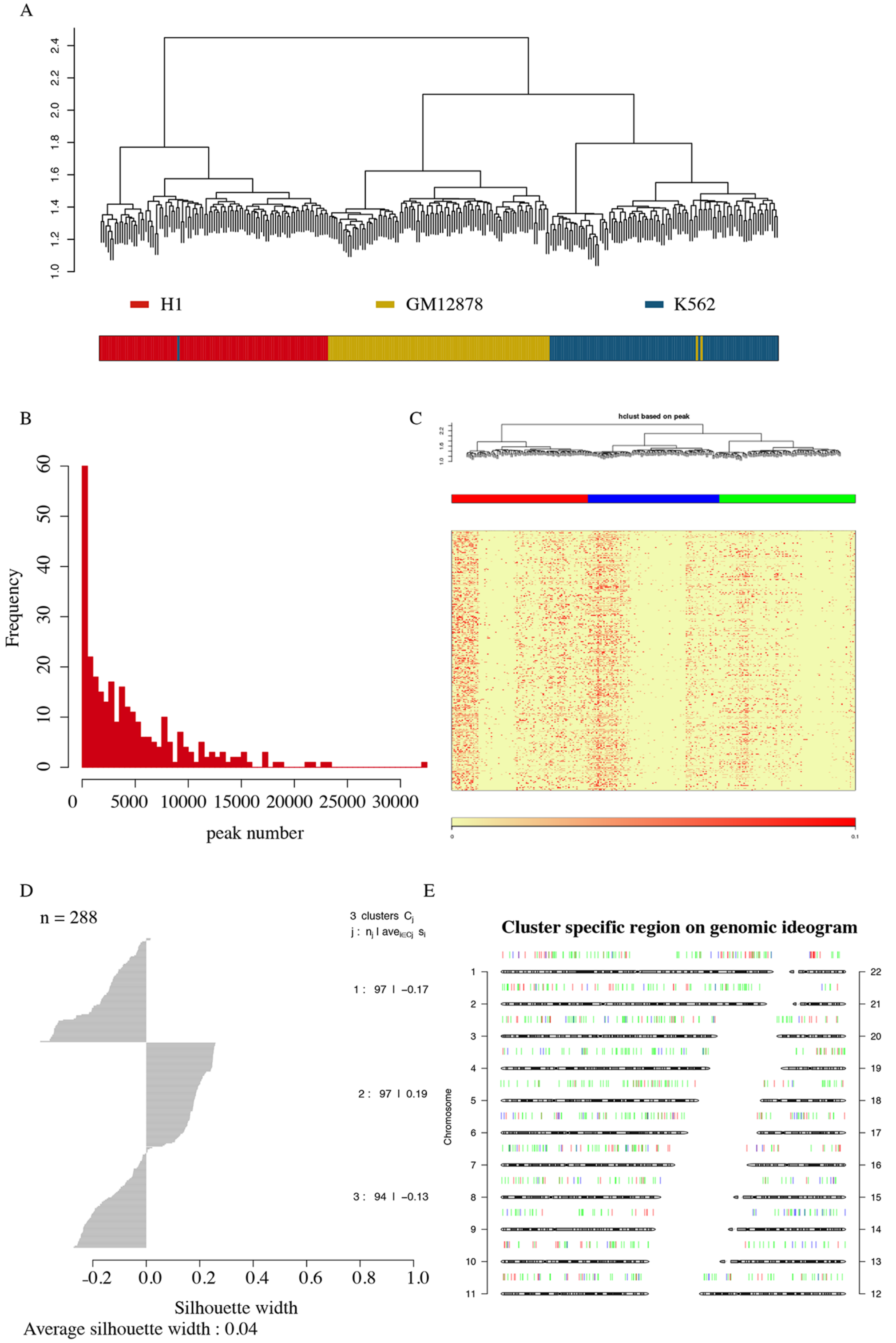

Fig 4. Cell-clustering level QC and single-cell level QC for scATAC-seq data. A) Upper panel shows cell-clustering results for combined scATAC samples generated from 3 different cell types. Bottom panel shows corresponding cell type labels of each cell marked by different colors (red stand for H1 cells, yellow stand for GM12878 cells and blue stand for K562 cells). The clustering step of Dr.seq2 clearly separated the scATAC-seq samples from three different cell types into different groups that were consistent with the cell type labels. B) Distribution of peak number for each single cell. C) Cell Clustering tree and peak region in each cell. The

upper panel represents the hieratical clustering results based on each single cell. The second panel with different colors represents decision of cell clustering. The bottom two panels (heatmap and color bar) represent the "combined peaks" occupancy of each single cell. D) Barplot shows Silhouette score of each cluster. Silhouette method is used to interpret and validate the consistency within clusters defined in previous steps. E) Cluster specific regions in each chromosome. Specific regions for different cell clusters are marked by different colors and ordered according to genomic loci.

https://doi.org/10.1371/journal.pone.0180583.g004

Single-cell level QC and post analysis of scATAC-seq data

In the single-cell level QC of Dr.seq2 on scATAC-seq data, the peak number of in each cell was defined as the number of "combined peaks" occupied by the reads in the cell. The distribution of different peak numbers in each cell indicated the different amount of information the cell contains (Fig 4B). Cell clustering was conducted based on the peak information in each cell using hierarchical clustering and open region was shown in the order of genomic location (Fig 4C). And Silhouette score [36] validated the consistency of each cluster (Fig 4D). Then cells in the same clusters were considered as cells in the same cell type and combined for the detection of cell type specific regions, which were defined as the peak regions that only covered in this cell type. Specific regions for different cell clusters were marked by different colors and ordered according to genomic loci (Fig 4E).

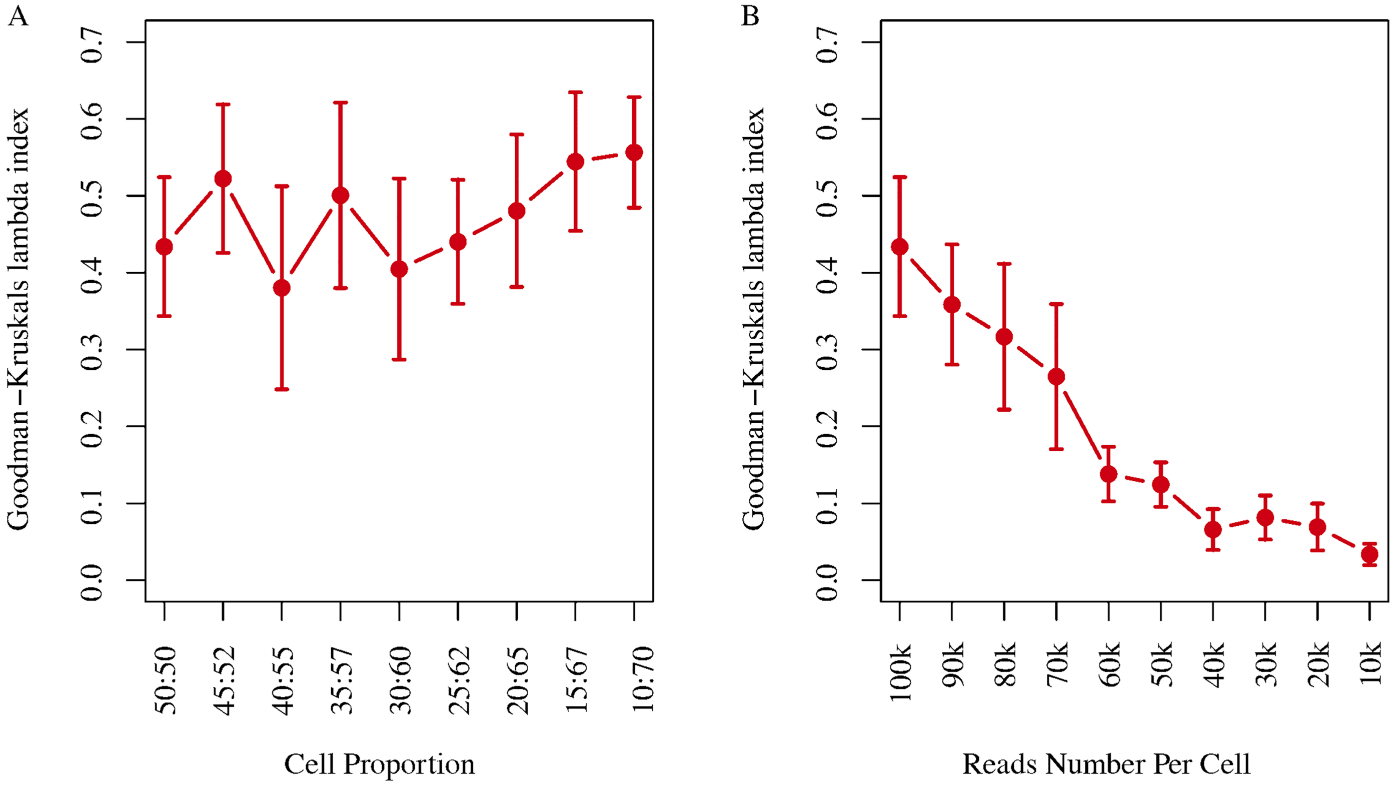

Fig 5. Cell clustering stability on simulated scATAC-seq data. A) Clustering stability of Dr.seq2 on simulated data with different numbers of reads per cell. The lambda index (y-axis) is plotted as a function of the number of reads per cell (x-axis). Error bars represent 95% confidence intervals calculated from 20 simulations. **B)** Clustering stability of Dr.seq2 on simulated data with different cell proportion depths. The lambda index (y-axis) is plotted as a function of the target cell number (x-axis). Error bars represent 95% confidence intervals calculated from 20 simulations.

https://doi.org/10.1371/journal.pone.0180583.g005

Table 3. Running time of each QC and analysis step for scATAC datasets.

Steps	Time (s/CPU)	Percentage (%)
Merge Cells	1507	39.72
Bulk-cell level QC	1654	43.60
Individual-cell level QC and cell-clustering QC	626	16.50
Post-analysis	5	0.13
Summary Report	2	0.05

288 scATAC datasets from three cell types were used to evaluate the runtime of Dr.seq2. The running time for each step was calculated using a single CPU (Intel® Xeon® CPU E5-2640 v2 @ 2.00 GHz).

https://doi.org/10.1371/journal.pone.0180583.t003

Cell clustering stability on simulated scATAC-seq data

To measure the tolerance of Dr.seq2 for low sequencing depth and small numbers of cells of a certain cell type, we simulated datasets with different cell proportions and sequencing depths by using scATAC-seq datasets from three cell types (Table 1).

We selected cells in different proportion with 100,000 reads per cell and then performed cell clustering using Dr.seq2. The performance of cell clustering methods was evaluated by Goodman-Kruskal's lambda index. And the average Goodman-Kruskal's lambda index calculated from 20 simulations indicated that Dr.seq2 was suitable for cell clustering with different cell proportions (Fig 5A). We also selected fifty cells from each cell type with the reads count range from 10,000 reads to 100,000 reads for each cell to measure the tolerance of Dr.seq2 on low sequence depth. Dr.seq2 produced stable clustering results with greater than 40,000 reads per cell (Fig 5B).

Computational cost of Dr.seq2

We also measured the computational time cost of Dr.seq2 by applied Dr.seq2 on combined scATAC-seq datasets (Table 3). The running time of each step was calculated using a single CPU (Intel® Xeon® CPU E5-2640 v2 @ 2.00 GHz).

Conclusions

In summary, Dr.seq2 is designed for QC and analysis components of parallel single cell transcriptome and epigenome data. Parallel single cell transcriptome data generated by different technologies can be transformed to the standard input for Dr.seq2 with contained functions. Using relevant commands, Dr.seq2 can also be used to report quality measurements based on four aspects and generate detailed analysis results for scATAC-seq and Drop-ChIP datasets.

Supporting information

S1 Fig. Workflow displays the software structure and detailed QC steps of Dr.seq2. A) Dr. seq2 provides QC and analysis for three major data types: single cell transcriptome data (DrSeq part), Drop-ChIP data (DrChIP part) and scATAC-seq data (ATAC part). For single cell RNA-seq data, two additional step-by-step functions are included: 1. Expression matrix generation for amounts of single cell RNA-seq datasets (GeMa step) and 2. Cell clustering and analysis for the single cell expression matrix (comCluster step). For different parallel single cell RNA-seq technologies, input data are standardized for DrSeq part. **B)** Four groups of QC measurements are conducted on single cell transcriptome data and epigenome data: 1.Reads level QC including reads quality, reads nucleotide composition and reads GC content 2.Bulk-cell level QC including reads alignment summary and gene body coverage for transcriptome data;

peak distribution; average profile on regulatory region and the distribution of different numbers of fragment length for epigenome data. 3. Individual-cell level QC including duplicate rate distribution, covered gene number and intron rate distribution and intron rate distribution for transcriptome data; peak number distribution and fragment length distribution for epigenome data. 4. Cell-clustering level QC including Gap statistics score and Silhouette score for transcriptome data, h-clustering and cluster specific peaks for epigenome data.
(TIF)

S2 Fig. Comparing the performance of Dr.seq2 and three existing state-of-the art methods on cell clustering. A) Clustering accuracy measured by the Goodman-Kruskal's lambda index of Dr.seq2 t-SNE, Dr.seq2 SIMLR methods and three published methods on simulated data with different numbers of reads per cell. The lambda index (y-axis) is plotted as a function of the number of reads per cell (x-axis). **B)** Running time of Dr.seq2 t-SNE, Dr.seq2 SIMLR methods and three published methods on simulated data with different numbers of reads per cell. The running time (y-axis) is plotted as a function of the number of reads per cell (x-axis). The running time for each method was calculated using a single CPU (Intel® Xeon® CPU E5-2640 v2 @ 2.00 GHz).
(TIF)

S1 File. Comparison of functions between Dr.seq2 and other software developed for single cell transcriptome data.
(XLSX)

S2 File. Meta data and accession ID for the bulk-cell RNA-seq data used in simulation.
(XLSX)

S3 File. Dr.seq2 QC and analysis output report for the scATAC-seq dataset.
(PDF)

S4 File. Dr.seq2 QC and analysis output report for the Drop-ChIP dataset.
(PDF)

S5 File. Dr.seq2 QC and analysis output report for the 10x genomics dataset.
(PDF)

Acknowledgments

We thank Shiyang Zeng and Yiying Lang for their suggestions.

Author Contributions

Conceptualization: CZ SH XH YZ.

Data curation: CZ.

Formal analysis: CZ.

Funding acquisition: YZ.

Methodology: CZ SH YZ.

Resources: CZ.

Software: CZ.

Supervision: YZ.

Writing – original draft: CZ.

Writing – review & editing: CZ SH XH YZ.

References

1. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015; 161(5):1202–14. https://doi.org/10.1016/j.cell.2015.05.002 PMID: 26000488.
2. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015; 161(5):1187–201. https://doi.org/10.1016/j.cell.2015.04.044 PMID: 26000487.
3. Fan HC, Fu GK, Fodor SP. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015; 347(6222):1258367. https://doi.org/10.1126/science.1258367 PMID: 25657253.
4. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017; 8:14049. https://doi.org/10.1038/ncomms14049 PMID: 28091601.
5. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014; 343(6172):776–9. https://doi.org/10.1126/science.1247651 PMID: 24531970.
6. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015; 33(11):1165–72. https://doi.org/10.1038/nbt.3383 PMID: 26458175.
7. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015; 523(7561):486–90. https://doi.org/10.1038/nature14590 PMID: 26083756.
8. Huo X, Hu S, Zhao C, Zhang Y. Dr.seq: a quality control and analysis pipeline for droplet sequencing. Bioinformatics. 2016; 32(14):2221–3. https://doi.org/10.1093/bioinformatics/btw174 PMID: 27153611.
9. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012; 28 (16):2184–5. https://doi.org/10.1093/bioinformatics/bts356 PMID: 22743226.
10. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9(9):R137. https://doi.org/10.1186/gb-2008-9-9-r137 PMID: 18798982.
11. Shin H, Liu T, Manrai AK, Liu XS. CEAS: cis-regulatory element annotation system. Bioinformatics. 2009; 25(19):2605–6. https://doi.org/10.1093/bioinformatics/btp479 PMID: 19689956.
12. Goodman LA, Kruskal WH. Measures of association for cross-classification. J Am Stat Assoc. 1954; 49:732–64.
13. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS Comput Biol. 2015; 11(6):e1004333. https://doi.org/10.1371/journal.pcbi.1004333 PMID: 26107944.
14. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015; 33(2):155–60. https://doi.org/10.1038/nbt.3102 PMID: 25599176.
15. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. PLoS Comput Biol. 2015; 11(11):e1004575. https://doi.org/10.1371/journal.pcbi.1004575 PMID: 26600239.
16. Leng N, Choi J, Chu LF, Thomson JA, Kendziorski C, Stewart R. OEFinder: a user interface to identify and visualize ordering effects in single-cell RNA-seq data. Bioinformatics. 2016; 32(9):1408–10. https://doi.org/10.1093/bioinformatics/btw004 PMID: 26743507.
17. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015; 16:241. https://doi.org/10.1186/s13059-015-0805-z PMID: 26527291.
18. Angerer P, Haghverdi L, Buttner M, Theis FJ, Marr C, Buettner F. destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics. 2016; 32(8):1241–3. https://doi.org/10.1093/bioinformatics/btv715 PMID: 26668002.
19. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics. 2015; 31(12):1974–80. https://doi.org/10.1093/bioinformatics/btv088 PMID: 25805722.

20. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015; 525(7568):251–5. https://doi.org/10.1038/nature14966 PMID: 26287467.

21. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, et al. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. Proc Natl Acad Sci U S A. 2014; 111(52):E5643–50. https://doi.org/10.1073/pnas.1408993111 PMID: 25512504.

22. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015; 347(6226):1138–42. https://doi.org/10.1126/science.aaa1934 PMID: 25700174.

23. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016; 13(3):241–4. https://doi.org/10.1038/nmeth.3734 PMID: 26780092.

24. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015; 16:278. https://doi.org/10.1186/s13059-015-0844-5 PMID: 26653891.

25. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014; 11(7):740–2. https://doi.org/10.1038/nmeth.2967 PMID: 24836921.

26. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016; 17(1):222. https://doi.org/10.1186/s13059-016-1077-y PMID: 27782827.

27. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014; 32 (4):381–6. https://doi.org/10.1038/nbt.2859 PMID: 24658644.

28. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, et al. Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. Cell Stem Cell. 2015; 17(3):360–72. https://doi.org/10.1016/j.stem.2015.07.013 PMID: 26299571.

29. Julia M, Telenti A, Rausell A. Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. Bioinformatics. 2015; 31(20):3380–2. https://doi.org/10.1093/bioinformatics/btv368 PMID: 26099264.

30. Leng N, Chu LF, Barry C, Li Y, Choi J, Li X, et al. Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. Nat Methods. 2015; 12(10):947–50. https://doi.org/10.1038/nmeth.3549 PMID: 26301841.

31. Bendall SC, Davis KL, Amir el AD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014; 157 (3):714–25. https://doi.org/10.1016/j.cell.2014.04.005 PMID: 24766814.

32. duVerle DA, Yotsukura S, Nomura S, Aburatani H, Tsuda K. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. BMC Bioinformatics. 2016; 17(1):363. https://doi.org/10.1186/s12859-016-1175-6 PMID: 27620863.

33. Jiang P, Thomson JA, Stewart R. Quality control of single-cell RNA-seq by SinQC. Bioinformatics. 2016; 32(16):2514–6. https://doi.org/10.1093/bioinformatics/btw176 PMID: 27153613.

34. Gardeux V, David F, Shajkofci A, Schwalie PC, Deplancke B. ASAP: a Web-based platform for the analysis and inter-active visualization of single-cell RNA-seq data. bioRxiv. 2016. https://doi.org/10.1101/096222

35. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017; 14(4):414–6. https://doi.org/10.1038/nmeth.4207 PMID: 28263960.

36. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math. 1987; 20:53–65.