Computational and Structural Biotechnology Journal

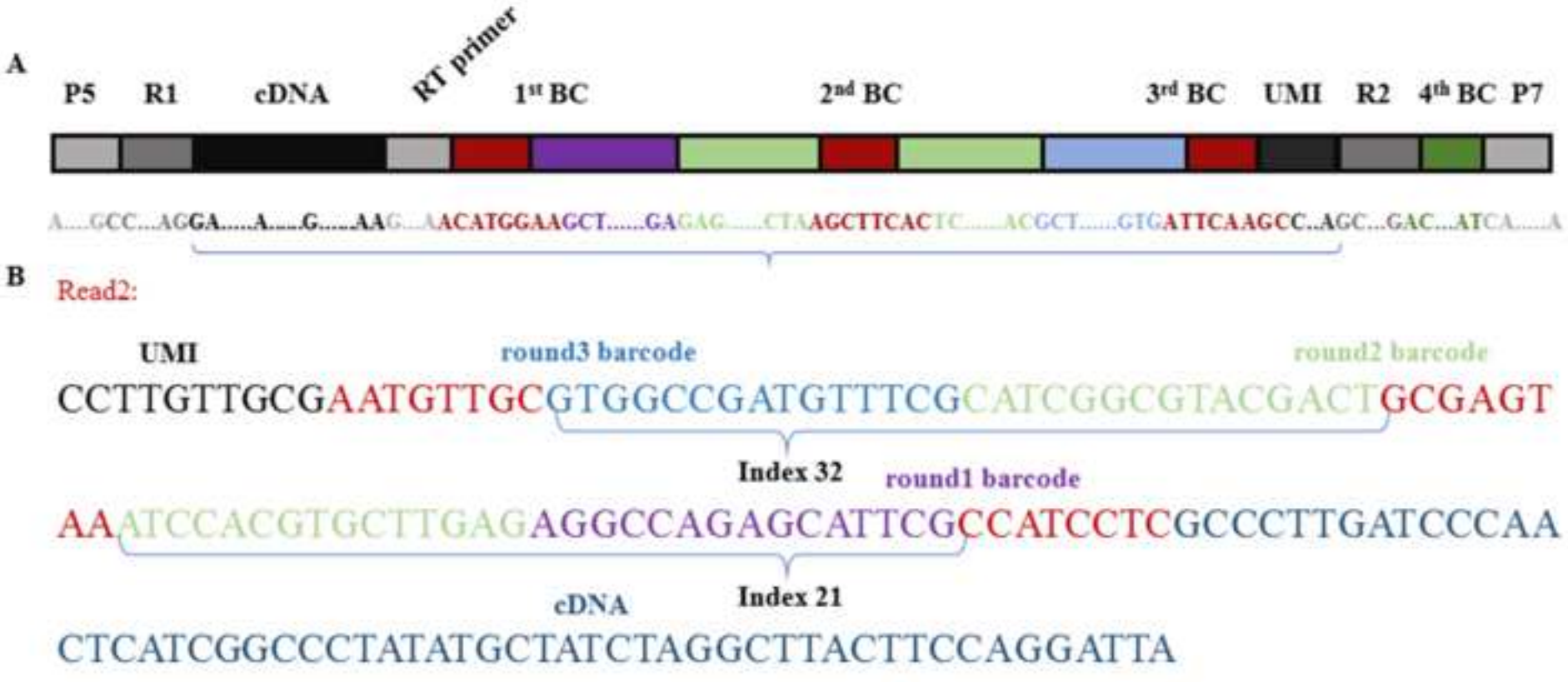
SCSit: A high-efficiency preprocessing tool for single-cell sequencing data from SPLiT-seq

--Manuscript Draft--

Manuscript Number:	CSBJ-D-21-00352
Article Type:	Research Paper
Section/Category:	Genomics
Keywords:	SCSit; single cell sequencing; SPLiT-seq; preprocessing tool; cell type identification
Abstract:	SPLiT-seq provides a low-cost platform to generate single-cell data by labeling the cellular origin of RNA through four rounds of combinatorial barcoding. However, an automatic and rapid method for preprocessing and classifying single-cell sequencing (SCS) data from SPLiT-seq is currently lacking. Here, we develop a high-efficiency preprocessing tool for single-cell sequencing data from SPLiT-seq (SCSit), which can directly identify combinatorial barcodes and UMI of cell types and obtain more labeled reads, and remarkably enhance the retained data from SCS due to the exact alignment of insertion and deletion. Compared with the original method used in SPLiT-seq, the consistency of identified reads from SCSit increases to 97%, and mapped reads are twice than the original. Furthermore, the runtime of SCSit is less than 10% of the original. It can accurately and rapidly analyze SPLiT-seq raw data and obtain labeled reads, as well as effectively improve the single-cell data from SPLiT-seq platform. The data and source of SCSit are available on the GitHub website https://github.com/shang-qian/SCSit .

Graphical Abstract

Click here to download Graphical Abstract Figure1.png



SCSit: A high-efficiency preprocessing tool for single-cell sequencing data from SPLiT-seq

Mei-Wei Luan[1,#], Jia-Lun Lin[2,#], Yu-Xiao Liu[2], Chuan-Le Xiao[3], Rongling Wu[4], Shang-Qian Xie[1,*]

[1] Key Laboratory of Genetics and Germplasm Innovation of Tropical Special Forest Trees and Ornamental Plants (Ministry of Education), School of Life Science, Hainan University, Haikou 570228, China

[2] College of Biomedical Information and Engineering, Hainan Medical University, Haikou 571199, China

[3] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China

[4] Public Health Sciences and Statistics and Center for Statistical Genetics, Pennsylvania State University, Hershey, PA, USA

# These authors contributed equally to this work

* To whom correspondence should be addressed:

Shang-Qian Xie, Email: sqianxie@foxmail.com

Abstract

SPLiT-seq provides a low-cost platform to generate single-cell data by labeling the cellular origin of RNA through four rounds of combinatorial barcoding. However, an automatic and rapid method for preprocessing and classifying single-cell sequencing (SCS) data from SPLiT-seq is currently lacking. Here, we develop a high-efficiency preprocessing tool for single-cell sequencing data from SPLiT-seq (SCSit), which can directly identify combinatorial barcodes and UMI of cell types and obtain more labeled reads, and remarkably enhance the retained data from SCS due to the exact alignment of insertion and deletion. Compared with the original method used in SPLiT-seq, the consistency of identified reads from SCSit increases to 97%, and mapped reads are twice than the original. Furthermore, the runtime of SCSit is less than 10% of the original. It can accurately and rapidly analyze SPLiT-seq raw data and obtain labeled reads, as well as effectively improve the single-cell data from SPLiT-seq platform. The data and source of SCSit are available on the GitHub website https://github.com/shang-qian/SCSit.

Keywords: SCSit, single cell sequencing, SPLiT-seq, preprocessing tool, cell type identification

Introduction

Cellular heterogeneity is a widespread phenomenon in biology by which cells vary in genetic and genomic factors [1]. Cell heterogeneity may have dramatic impact on biological processes and diseases, such as bacterial sepsis, cell immunity, aneurysm and cardiomyocyte [2-5]. High-throughput single-cell sequencing (SCS) technologies, such as next generation sequencing (NGS) and third generation sequencing (TGS), have been developed and widely used to identify cell types from morphologically similar cell populations and multi-cellular tissues [6-10]. Compared with conventional sequencing technologies, SCS have an obvious advantage in cell type identification at the single-cell level, especially for low-abundance gene information that may be easily neglected in previous tissue-level studies [11]. SCS provide a cutting-edge technology to measure the real-time expression of genes in a single cell [12-14] and reveal inter-cellular heterogeneity [15-17], which play an important role in understanding cellular features and functions in tumors [18], developmental biology [19, 20], microbiology [21], and neuroscience [22, 23].

At present, several single-cell sequencing technologies have been developed, including DroNC-seq [24], CROP-seq [25], LIANTI [26], and scCOOL-seq [27], scSLAM-seq [28], DART-seq [29] and TAP-seq [30]. DroNC-seq combines single nucleus RNA-seq (sNuc-seq) and Drop-seq using microfluidic beads marking up single-cell DNA, showing efficient and sensitive capabilities to identify single-cell types [24]. CROP-seq, called CRISPR droplet sequencing, enables pooled CRISPR-Cas9 screening with

single-cell droplet, which facilitates high-throughput single-cell sequencing in a cost-effective way [25]. LIANTI linearly amplifies the whole genome DNA sequence by inserting the transposons in single cells, which significantly increases the depth and resolution of single-cell DNA sequencing [26]. scCOOL-seq is a single-cell complex sequencing technology that simultaneously characterizes the chromatin state, nucleosome location, DNA methylation, copy number variation and chromosome ploidy [27]. scSLAM-seq is a single-cell, thiol-(SH)-linked alkylation of RNA for metabolic labelling sequencing which records transcriptional activity directly integrating metabolic RNA labelling and biochemical nucleoside conversion [28]. DART-seq enables multiplexed amplicon sequencing and transcriptome profiling in single cells [29]. TAP-seq, called targeted Perturb-seq, focuses single-cell RNA-seq coverage on genes of interest and permits a routine analysis of thousands of CRISPR-mediated perturbations within a single experiment [30]. Although the SCS technologies mentioned above have their own advantages and characteristics, they all require custom microfluidics or microwells for cell sorting to obtain single cells, resulting in the high cost of single-cell sequencing.

Recently, Rosenberg *et al*. developed a single-cell RNA-seq method, split-pool ligation-based transcriptome sequencing (SPLiT-seq), which labeled the cellular origin of RNA through four rounds of combinatorial barcoding and unique molecular identifier (UMI) (Fig. 1A) [12]. SPLiT-seq eliminated the need of single cells isolation because of the index information of DNA barcodes. The alignment of cell barcodes could be used to

identify cell types from SPLiT-seq data, and this principle greatly reduced SCS cost and experimental requirements, making it to be widely used in single cell research. However, there is currently no automatic and rapid preprocessing method that enables the classification of single-cell sequencing data from SPLiT-seq. The existing methods simply based on ordinary alignment tools, such as BLAST or BWA, are time-consuming and fallibility for simultaneous determination of all three barcodes in different regions of each sequence. Therefore, we develop a high-efficiency preprocessing tool for single-cell sequencing data from SPLiT-seq (SCSit), which automatically identifies three rounds of barcode and UMI and significant increase the clean SCS reads due to the accurate detection of insertion and deletion of barcodes in the alignment. SCSit effectively solves the classification and extraction of cell type labels from SPLiT-seq and achieves more accurate single-cell data.

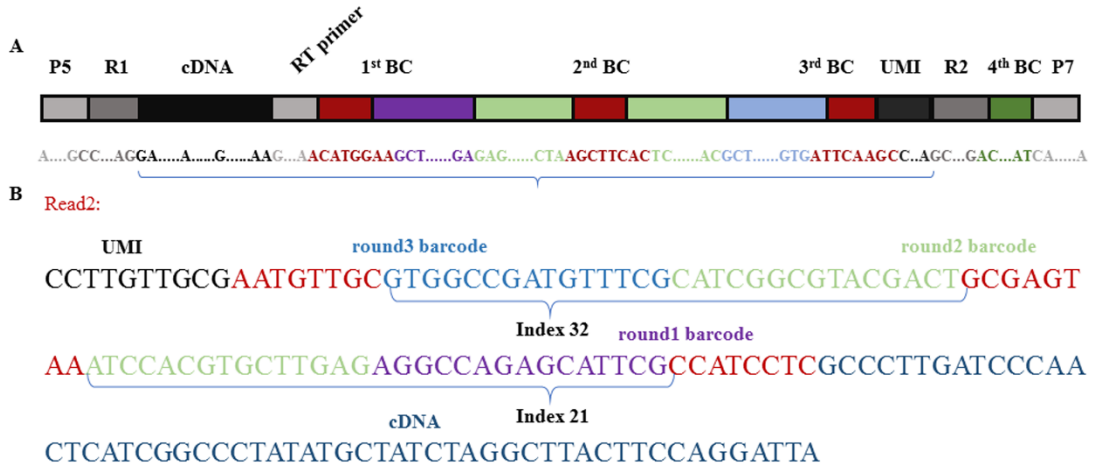


Figure 1. Overview of barcoded cDNA molecules from SPLiT-seq data (A: labeling transcriptome with SPLiT-seq, B: Read2 containing three barcodes, UMI and cDNA).

Materials and methods

Feature of SCS data

Raw data of SPLiT-seq was sequenced on Illumina platform using 150 nucleotide (nt) kits and paired-end sequencing. Read1 included the transcript (cDNA) and R1 primer sequences, and Read2 covered UMI, three BC barcode combinations and cDNA (Fig. 1B). Thus, the identification of Read2 determines the accuracy and efficiency of cell type classification in SCS data, and it is a key step in SCSit.

Identification of index position of barcodes in Read2

Five contents contain UMI, three BC barcodes and cDNA, and the UMI and three barcodes in Read2 were used as specific tag to obtain labeled reads that identified different cell types (Fig. 1B). Each barcode is composed of indicator sequences of cell type (8 nt) and index sequences of barcodes (Table S1). The index sequences of barcode 1 and 2 (*index21*, 30 nt), barcode 2 and 3 (*index32*, 30 nt) were joined each end to end (Fig. 1B and Table S1), and the joint sequences (*index21* and *index32*) were used to identify each round barcode in Read2. The sequences of *index21* and *index32* were divided into 23 segments by 8 nt k-mer. Then the Read2 were mapped to 23 segments by sliding window (8 nt) and detected the position of index sequences of three barcodes. There were three situations of detection of index sequence in Read2:

(1) Complete match of index sequence: sliding window sequence (8 nt) of Read2 was completely and continuously matched to index sequence of barcode, the position of index sequence of Read2 was determined by the start of matched window1 (W_1,

122 *Ms*) and the end of matched window23 (W_{23}, *Me*), and *Me-Ms* was equal to the
123 length of index sequence (30 nt) (Fig. 2A).
124 (2) Considerate one mismatch: i) If one mismatch exists in the position *m* of sequences
125 ($9 \leq m \leq 21$), and W_1 and W_{23} are all completely matched. The index sequence of
126 Read2 could be identified while *Me-Ms* = 30, and the start and end position of
127 index sequences are determined from *Ms* to *Me* (Fig. 2B). ii) If the mismatch occurs
128 in the position *m* ($1 \leq m \leq 8$) of matched W_1, the index sequence of Read2 could be
129 identified while *Me-Ms* = 30-*m*, and the start and end position of index sequences
130 are determined from *Ms-m* to *Me* (Fig. 2B). iii) If the mismatch occurs in the
131 position *m* ($22 \leq m \leq 30$) of matched W_{23}, the index sequence of Read2 could be
132 identified while *Me-Ms* = 30-*m*, and the start and end position of index sequences
133 were determined from *Ms* to *Me*+*m* (Fig. 2B).
134 (3) Considerate one INDEL (length of index sequence in Read2 ≠ 30 nt): i) If INDEL
135 is present in the position *m* ($9 \leq m \leq 21$) of between matched $W_1$and W_{23}. The index
136 sequence of Read2 could be identified while *Me-Ms* is equal to the length of
137 matched index sequence containing insertion (31 nt) or deletion (29 nt), and the
138 start and end position of index sequences are determined from *Ms* to *Me* (Fig. 2C).
139 ii) If the INDEL exists in the position *m* ($1 \leq m \leq 8$) of matched W_1, the index
140 sequence of Read2 could be identified while *Me-Ms* = 30-*m*, and the start and end
141 position of index sequences with insertion are determined from *Ms-m*-1 to *Me* and
142 with deletion from *Ms-m*+1 to *Me* (Fig. 2C). iii) If the INDEL exists in the position
143 *m* ($22 \leq m \leq 30$) of matched W_{23}, the index sequence of Read2 could be identified

while *Me-Ms* = 30-*m*, and the start and end position of index sequences with insertion are determined from *Ms* to *Me*+*m*+1 and with deletion from *Ms* to *Me*+*m* (Fig. 2C).

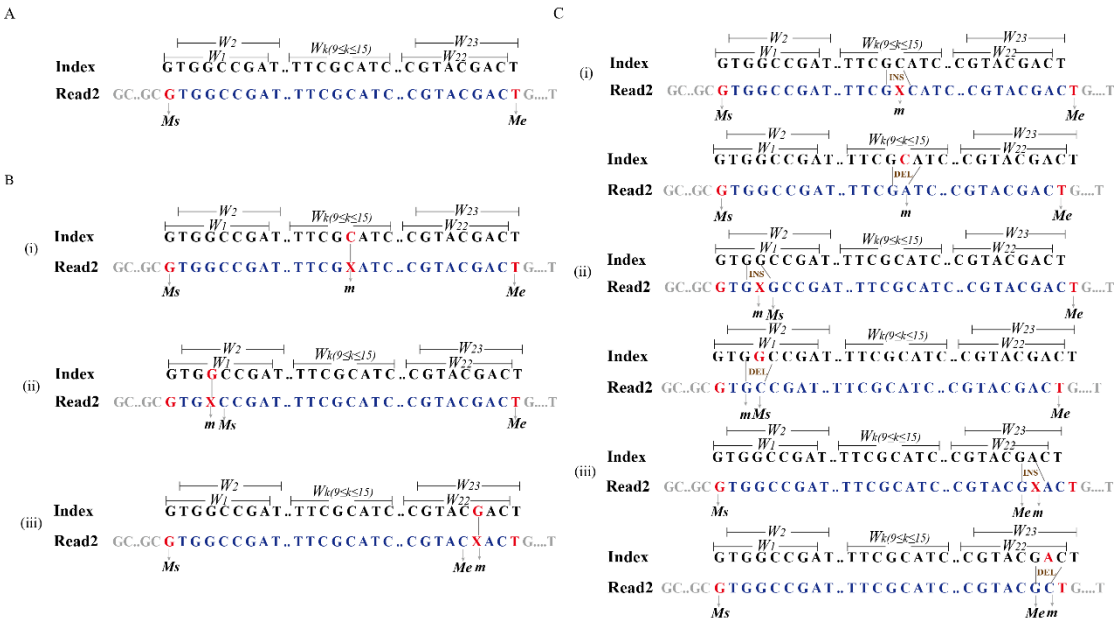


Figure 2. Index sequence identification of barcode in Read2 (INS: insertion, DEL: deletion, A: complete match of index sequence. B: considerate one mismatch. C: considerate one INDEL).

Identification of indictor sequences of barcodes in Read2

Based on the identification of round index of barcodes, the labeled reads could be obtained, which cell type could be classified by the alignment of indictor sequences of barcodes. For the identification of indictor sequences, we executed a quad to decimal conversion, and used 0, 1, 2, and 3 to present A, T, G, and C. Then the fragment sequences were converted to decimal number (*intSeq*), which was calculated by

$$intSeq = \sum_{i=0}^{n} trans(seq[i]) * 4^{(n-i-1)}$$

where *seq[i]* denotes the *i*th base of a fragment, *n* is the length of one fragment. The three round barcodes were converted and stored in decimal number lists, and the *intSeq* values were used as the index of three lists (Fig 3A). The indictor sequences of barcodes

in Read2 were mapped to the barcode lists by using the *intSeq* index, thus labeled reads could rapidly be marked that cell types could be rapidly identified in SCSit.

To further make the SCSit more applicable, we used the distance function to allow mismatch between indictor sequence of Read2 and three round barcodes. The distance function is defined by

$$Dist(x,y) = \sum_{i=0}^{n-1} D(x[i], y[i]) * \left(1 + \frac{i}{n}\right),$$

$$D(x,y) = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

where $x[i]$ and $y[i]$ are referred to as the *i*th base of sequence *x* and *y*, *n* referred to as the length of them. For the indicator sequence of barcode, *n*=8. If the value of *Dist* is less than threshold (2), the barcode identification of Read2 is valid, otherwise Read2 is abandoned.

A

Barcode round1 list		
No.	Sequence	*intSeq*
1	AACGTGAT	3681
2	AAACATCG	798
3	ATGCCTAA	7120
4	AGTGGTCA	9884
5	ACCACTGT	15577
6	ACATTGGC	12651
...	...	...
96	CCATCCTC	61943

Barcode round2 list		
No.	Sequence	*intSeq*
1	ACAGATTC	12823
2	ATTGGCTC	5815
3	CAAGGAGC	49803
4	CACCTTAC	53075
5	CCATCCTC	61943
6	CCGACAAC	63683
...	...	...
96	AAGACGGA	2280

Barcode round1 list		
No.	Sequence	*intSeq*
1	AACCGAGA	3976
2	AAGACGGA	2280
3	ACACAGAA	13088
4	ACGTATCA	14620
5	AGCAGGAA	11424
6	ATCCTGTA	8036
...	...	...
96	AGTGGTCA	9884

B

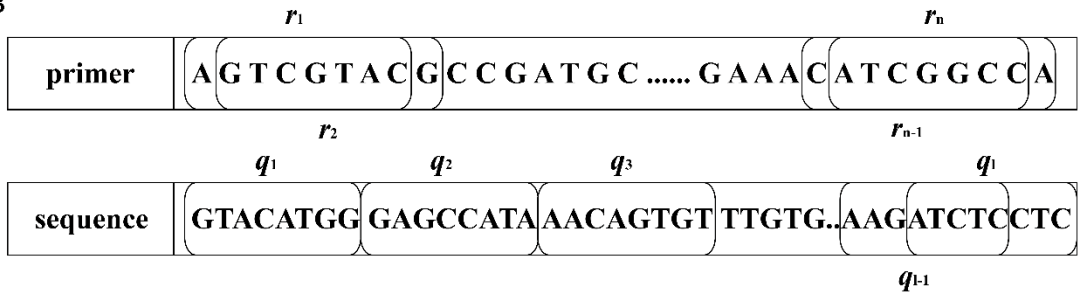


Figure 3. Decimal conversion of barcodes and segmentation of sequences (A: decimal conversion, B: the sequence was divided into *n* segments with 8 nt).

Identification and preprocessing of Read1

Raw Read1 sequences were retained with the valid Read2, and trim the reads by PrimerList (forward and reverse, Table S2) used in the SPLiT-seq literature. The sequences of PrimerList were divided into *n* segments (r_1, …, r_n) by k-mer (8 nt), and similarly converted to the decimal number (*intSeq*) (Fig. 3A). Read1 sequence was divided into *l* (q_1, …, q_l) segments. If the last segment is less than 8 bases, then it achieves the 8 bases backwards (q_l) (Fig. 3B). Then the fragment was direct inquiry by *intSeq* index of PrimerList. The first and last matched segment of Read1 were recorded and used to trim the primer sequences of Read1.

Implementation and validation

SCSit was developed by C program and parallel compute by multithread to quickly identify the cell type of SPLiT-seq reads. The raw reads FASTQ format of SPLiT-seq data as input executed SCSit program and output labeled reads with combinatorial barcodes and UMI. The output Read1 FASTQ format file was composed of combinatorial barcodes and UMI, and Read2 file was corresponding sequencing data. To validate the accuracy and reliability of SCSit, we collected and compared the treated five samples from mouse (SAMN08567263) and mixed human and mouse cells containing HEK293, HelaS3, and NIH/3T3 (SAMN08567259, SAMN08567260, SAMN08567261, and SAMN08567264) used in the SPLiT-seq literature [12]. The original method used in SPLiT-seq discarded reads that last 6 bases of them did not match barcode sequence, UMI region were then filtered based on quality score that read with >1 low-quality base (phred $\leq$ 10) and three 8 nt cell barcodes with more than one mismatch. We compared the SCSit and original method by using the filtered clean data. The source of SCSit and validation data were available on the GitHub website https://github.com/shang-qian/SCSit.

Results

Five datasets of different species from SPLiT-seq were used to perform the assessment of SCSit. The identified labeled reads and runtime were compared between SCSit and original methods (Table 1). The results indicate the identified labeled reads of SCSit in SAMN08567263 (56,833,343), SAMN08567261 (163,780,622), SAMN08567260

(163,502,984), SAMN08567259 (148,423,169) and SAMN08567264 (160,257,239) were more than those in the original method (Table 1). And the rate of identified reads in SCSit were all more than 65% that were distinctly higher than those by the original method. The consistency of SCSit identified reads was 96~97% in the original (Table 1). The reads uniquely identified by SCSit are all more than 13 percent in five samples (Fig. 4A-E). Especially, almost double increase rates of identified reads are found in SAMN08567260 and SAMN08567264 (Table 1). The main reason for the obvious improvement of SCSit in labeled reads identification is the consideration of indel and mismatch of barcodes alignment and UMIs (Table S3), which further illustrates the necessity of developing proper method to obtain labeled reads from SPLiT-seq data. Besides, we assessed the runtime of SCSit for five datasets under 4 cores of CPU. Results demonstrate that the runtime of five samples in SCSit is less than the original method with blastn-short (Table 1). The runtime of SCSit was mainly used to identify the indictor of barcodes and trimming primer, while the original took two part time that contains barcodes alignment with blastn-short and UMI identification.

To further validate the accuracy of obtain labeled reads from SCSit, we mapped the identified reads to either the combined mm10-hg19 genome or mm10 genome using STAR [31]. The mapped reads number of SCSit are more than these in the original method in five samples, SAMN08567264 has the most incremental mapped reads (114,732,417) and twice than the original method (Table 1 and Fig. 4F). The 93~98% of uniquely mapped reads by SCSit are consistent with the reads in the original, which directly enhances the number of mapped reads (Table S3). The above results illustrate

that SCSit is an accurate and efficiency tool to obtain labeled reads from SPLiT-seq.

Table 1. The statistical assessment of SCSit in five samples.

Sample ID	Method	Raw reads No.	Identified reads No. (ratio*)	Consistency rate in the original (%)	Runtime (h)	Mapped reads No.
SAMN08567263	SCSit	77,621,181	56,833,328 (73.22%)	96.10	0.62	44,541,508
	Original		51,706,161 (66.61%)		17.68	41,067,548
SAMN08567261	SCSit	218,683,580	163,780,622 (74.89%)	97.18	1.90	123,956,003
	Original		145,809,694 (66.68%)		50.34	112,226,501
SAMN08567259	SCSit	221,577,898	148,423,169 (66.98%)	97.17	1.98	109,834,181
	Original		131,707,053 (59.44%)		51.01	98,810,811
SAMN08567260	SCSit	215,597,675	163,502,984 (75.84%)	96.81	1.87	122,062,984
	Original		82,387,120 (38.21%)		49.58	62,516,718
SAMN08567264	SCSit	241,868,411	160,257,239 (66.26%)	96.88	2.15	114,732,417
	Original		75,844,129 (31.36%)		55.67	54,887,436

ratio*: the percentage of identified reads in raw reads. Original referred to original method

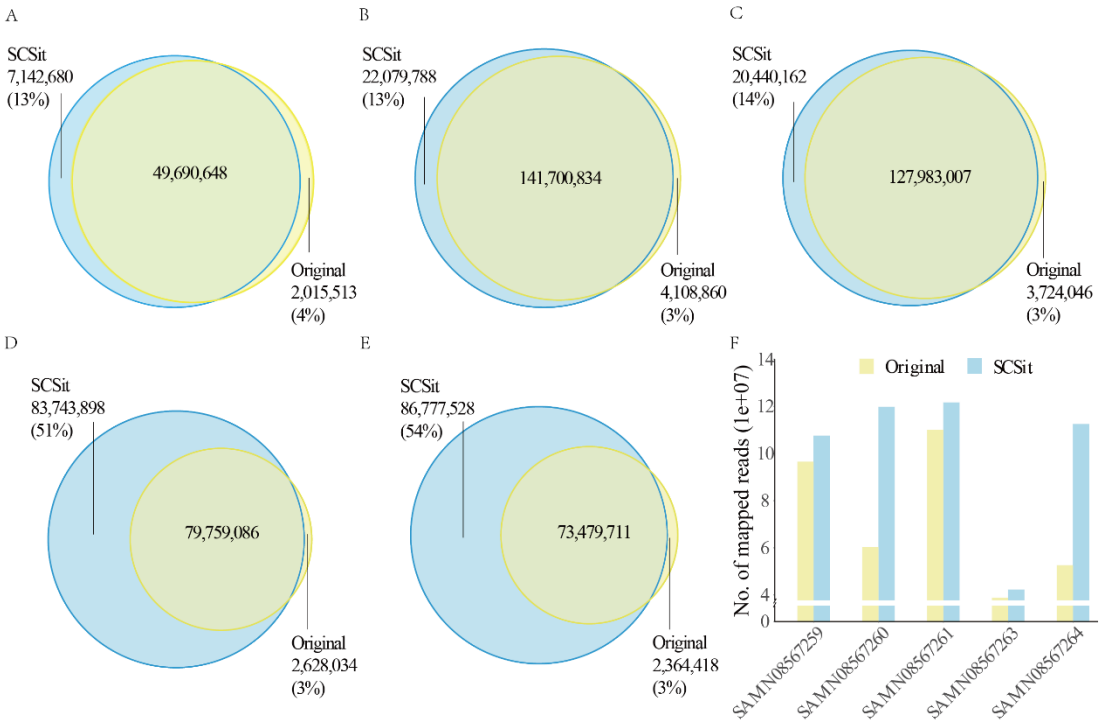


Figure 4. Comparison SCS reads from SCSit and original method (A-E: SAMN08567263, SAMN08567261, SAMN08567259, SAMN08567260, and SAMN08567264, F: comparison of mapped reads between SCSit and original).

Discussion

SCSit, an automatic, rapid and accurate preprocessing tool for single-cell sequencing data, and obtain labeled reads for SPLiT-seq data which can directly identifies cell types. SPLiT-seq labels cell types by four rounds of combinatorial barcoding and UMI [12]. K-mer alignment algorithm that completely considers the mismatch and indel in barcode sequences and UMI are used to obtain labeled reads and classify cell types in SCSit, and conversion index of decimal conversion greatly improves the efficiency of alignment. The comparison of identified reads and consistency ratio with the original illustrates that SCSit has a high-efficiency preprocessing performance for cell type's identification of SCS data from SPLiT-seq.

SCSit identifies more labeled reads in five samples, the uniquely identified reads were 7,142,680, 22,079,788, 20,440,162, 83,743,898 and 86,777,528 in sample SAMN08567263, SAMN08567261, SAMN08567259, SAMN08567260 and SAMN 08567264, respectively (Table S3). The unique additional reads of SCSit contain the barcodes with indel, unidentified UMI and barcodes in the original, and the barcodes absence that one of the three barcodes was missing (Table S3). The UMI and barcodes unidentified reads in original are 90~97% of uniquely identified reads in SCSit (Table S3). In order to facilitate comparison of SCSit and original method that discard the barcode sequence with more than one mismatch [12], we used one mismatch and one indel for barcodes alignment in this study. Actually, SCSit is also suitable for the case of more than one mismatch and indel in the identification of Read2. However, evaluating results from SCSit with more than one mismatch and indel show that the identified reads only increase 4~5% but takes 20~40% more runtime (Table S4). Considering the optimal balance of efficiency, we use one mismatch and one indel as default setting in SCSit.

The alignment of barcodes with fault-tolerant and indel is the main reason for the high-efficiency and rapid preprocessing in SCSit. Although the fault-tolerant and indel alignment were proposed for SPLiT-seq data in this study, the core principle could be widely used in other single-cell sequencing data similar to SPLiT-seq that using barcode sequence information. As new single-cell sequencing technology are proposed in the

future, SCSit will be updated and improved in time to accommodate more single-cell sequencing platform data.

Conclusions

SCSit, a rapid and high-efficiency preprocessing tool for single-cell sequencing data was developed in this study. It could accurately analyze SPLiT-seq raw data and labeled reads, and effectively improved the single-cell data from SPLiT-seq platform.

Competing interests

There were no competing interests.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (31760316, 32060149 and 31871326), Hainan Provincial Natural Science Foundation of China (320RC500 and ZDKJ201815), Priming Scientific Research Foundation of Hainan University (KYQD(ZR)1721).

Contributions

SQX conceived the project and designed the experiments, MWL and YXL collected datasets and performed the bioinformatics analysis, JLL validated the method, MWL, CLX, and SQX wrote the manuscript. RW critically read and revised the manuscript. All authors read and approved the final manuscript.

References

[1] E.Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, et al., (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets, Cell 161(5) 1202-1214.

[2] M. Reyes, M.R. Filbin, R.P. Bhattacharyya, K. Billman, T. Eisenhaure, et al., (2020) An immune-cell signature of bacterial sepsis, Nature medicine 26(3) 333-340.

[3] E. Papalexi, R. Satija, (2018) Single-cell RNA sequencing to explore immune cell heterogeneity, Nature reviews. Immunology 18(1) 35-45.

[4] Z. Liu, L. Wang, J.D. Welch, H. Ma, Y. Zhou, et al., (2017) Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte, Nature 551(7678) 100-104.

[5] G. Zhao, H. Lu, Z. Chang, Y. Zhao, T. Zhu, et al., (2020) Single cell RNA sequencing reveals the cellular heterogeneity of aneurysmal infrarenal abdominal aorta, Cardiovascular research.

[6] N. Tung, C. Battelli, B. Allen, R. Kaldate, S. Bhatnagar, et al., (2015) Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel, Cancer 121(1) 25-33.

[7] C. Genomes Project, G.R. Abecasis, D. Altshuler, A. Auton, L.D. Brooks, et al., (2010) A map of human genome variation from population-scale sequencing, Nature 467(7319) 1061-73.

[8] S.B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R.P. Lach, C. Ingle, et al., (2010) Transcriptome genetics using second generation sequencing in a Caucasian population, Nature 464(7289) 773-7.

[9] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, et al., (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data, Genome research 24(11) 1881-93.

[10] C. Gawad, W. Koh, S.R. Quake, (2016) Single-cell genome sequencing: current state of the science, Nature reviews. Genetics 17(3) 175-88.

[11] D. Ramskold, S. Luo, Y.C. Wang, R. Li, Q. Deng, et al., (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells, Nature biotechnology 30(8) 777-82.

[12] A.B. Rosenberg, C.M. Roco, R.A. Muscat, A. Kuchina, P. Sample, et al., (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding, Science 360(6385) 176-182.

[13] Z. Xue, K. Huang, C. Cai, L. Cai, C.Y. Jiang, et al., (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing, Nature 500(7464) 593-7.

[14] T. Voet, P. Kumar, P. Van Loo, S.L. Cooke, J. Marshall, et al., (2013) Single-cell paired-end genome sequencing reveals structural variation per cell cycle, Nucleic acids research 41(12) 6119-38.

[15] E.F. Davis-Marcisak, P. Orugunta, G. Stein-O'Brien, S.V. Puram, E.R. Torres, et al., (2018) Expression variation analysis for tumor heterogeneity in single-cell RNA-sequencing data, bioRxiv 479287.

[16] P. Angerer, L. Simon, S. Tritschler, F.A. Wolf, D. Fischer, et al., (2017) Single cells make big data: New challenges and opportunities in transcriptomics, Current Opinion in Systems Biology 4 85-91.

[17] F.A. Vieira Braga, S.A. Teichmann, X. Chen, (2016) Genetics and immunity in the era of single-cell genomics, Human molecular genetics 25(R2) R141-R148.

[18] D.T. Ting, B.S. Wittner, M. Ligorio, N. Vincent Jordan, A.M. Shah, et al., (2014) Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells, Cell reports 8(6) 1905-1918.

[19] M.G. Filbin, I. Tirosh, V. Hovestadt, M.L. Shaw, L.E. Escalante, et al., (2018) Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq, Science 360(6386) 331-335.

[20] J.C. Marioni, D. Arendt, (2017) How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology, Annual review of cell and developmental biology 33 537-553.

[21] M. Hess, A. Sczyrba, R. Egan, T.W. Kim, H. Chokhawala, et al., (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen, Science 331(6016) 463-7.
[22] M.J.C. Jordao, R. Sankowski, S.M. Brendecke, Sagar, G. Locatelli, et al., (2019) Single-cell profiling identifies myeloid cell subsets with distinct fates during neuroinflammation, Science 363(6425).
[23] I. Tirosh, A.S. Venteicher, C. Hebert, L.E. Escalante, A.P. Patel, et al., (2016) Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma, Nature 539(7628) 309-313.
[24] N. Habib, I. Avraham-Davidi, A. Basu, T. Burks, K. Shekhar, et al., (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq, Nature methods 14(10) 955-958.
[25] P. Datlinger, A.F. Rendeiro, C. Schmidl, T. Krausgruber, P. Traxler, et al., (2017) Pooled CRISPR screening with single-cell transcriptome readout, Nature methods 14(3) 297-301.
[26] C. Chen, D. Xing, L. Tan, H. Li, G. Zhou, et al., (2017) Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI), Science 356(6334) 189-194.
[27] F. Guo, L. Li, J. Li, X. Wu, B. Hu, et al., (2017) Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells, Cell research 27(8) 967-988.
[28] F. Erhard, M.A.P. Baptista, T. Krammer, T. Hennig, M. Lange, et al., (2019) scSLAM-seq reveals core features of transcription dynamics in single cells, Nature 571(7765) 419-423.
[29] M. Saikia, P. Burnham, S.H. Keshavjee, M.F.Z. Wang, M. Heyang, et al., (2019) Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells, Nature methods 16(1) 59-62.
[30] D. Schraivogel, A.R. Gschwind, J.H. Milbank, D.R. Leonce, P. Jakob, et al., (2020) Targeted Perturb-seq enables genome-scale genetic screens in single cells, Nature methods 17(6) 629-635.
[31] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, et al., (2013) STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29(1) 15-21.