Quantitative Biology 2022, 10(4): 333–340
https://doi.org/10.15302/J-QB-022-0295

RESEARCH ARTICLE

Comparative analysis of NovaSeq 6000 and MGISEQ 2000 single-cell RNA sequencing data

Weiran Chen[1,†], Md Wahiduzzaman[1,†], Quan Li[1], Yixue Li[1,2,*], Guangyong Zheng[1,*], Tao Huang[1,*]

[1] Bio-Med Big Data Center, Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China
[2] School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
* Correspondence: yxli@sibs.ac.cn; gyzheng@picb.ac.cn; huangtao@sibs.ac.cn

Received December 3, 2021; Revised March 13, 2022; Accepted March 29, 2022

***Background*: Single-cell RNA sequencing (scRNA-seq) technology is now becoming a widely applied method of transcriptome exploration that helps to reveal cell-type composition as well as cell-state heterogeneity for specific biological processes. Distinct sequencing platforms and processing pipelines may contribute to various results even for the same sequencing samples. Therefore, benchmarking sequencing platforms and processing pipelines was considered as a necessary step to interpret scRNA-seq data. However, recent comparing efforts were constrained in sequencing platforms or analyzing pipelines. There is still a lack of knowledge of analyzing pipelines matched with specific sequencing platforms in aspects of sensitivity, precision, and so on.**

***Methods*: We downloaded public scRNA-seq data that was generated by two distinct sequencers, NovaSeq 6000 and MGISEQ 2000. Then data was processed through the Drop-seq-tools, UMI-tools and Cell Ranger pipeline respectively. We calculated multiple measurements based on the expression profiles of the six platform-pipeline combinations.**

***Results*: We found that all three pipelines had comparable performance, the Cell Ranger pipeline achieved the best performance in precision while UMI-tools prevailed in terms of sensitivity and marker calling.**

***Conclusions*: Our work provided an insight into the selection of scRNA-seq data processing tools for two sequencing platforms as well as a framework to evaluate platform-pipeline combinations.**

Keywords: Single-cell RNA sequencing; cell-type; data processing; pipeline; platform

Author summary: We proposed that evaluating scRNA-seq data processing pipelines should aim at comparing the sequencer-pipeline combinations rather than benchmarking between either sequencers or pipelines. We compared sequencer-pipeline combinations in aspect of gene detection, dropout rates, number of markers and cell types. Based on results above we made recommendations for different purposes of research such as finding more marker genes or gaining maximum precision.

INTRODUCTION

Unlike conventional bulk RNA sequencing technology, offering average expression levels of a heterogeneous mixture of cells, single-cell RNA sequencing (scRNA-seq) technology is a cutting-edge approach that enables transcriptomic profiling at single-cell resolution [1]. Through transcripts filtering, dimensionality reduction

† These authors contributed equally this work.

and cell clustering of scRNA-seq data, cells can be separated into distinct cell types [2]. With sufficient sequencing depth one could even identify cells with different states which facilitate investigation of homeostasis, cell communications and tumor heterogeneity [3].

Currently, there exist diverse sequencing protocols as well as sequencers [4]. Among all the sequencers, Illumina series platforms are famous due to their high quality and high throughput on data generation capability. MGI Tech provides alternative sequencing choices by launching BGISEQ and MGISEQ series of sequencers. BGISEQ and MGISEQ platforms are based on the cPAS approach that evolved from the cPAL method published by Complete Genomics in 2009 [5]. Illumina platforms use bridge structures which consist of cDNA fragments and primers for amplification. BGISEQ and MGISEQ platforms generate circular templates, after rolling circle amplification (RCA) the replicates of templates form DNA nanoballs for the cPAS-based sequencing [6].

Aside from single-cell RNA sequencing protocols and platforms, there is also a diversity of processing pipelines. At least 6 processing pipeline software have been published since 2015. Cell Ranger was developed by the 10X Genomics company and published in 2017 which now becomes one of the most popular processing pipelines [7], while the famous software Drop-seq-tools was firstly introduced by broad institute in 2015 [8]. Other processing pipelines such as UMI-tools [9], zUMIs [10], dropEst [11], scPipe [12] have been published in more recent years. A standard processing pipeline generally consists of several steps, including barcode identification, quality check, reads mapping and sequence annotation. Despite the common procedures involved, each processing pipeline contained its unique features. For instance, the Cell Ranger pipeline was only designed for 10X Genomics sequencing data, by contrast, the Drop-seq-tools and UMI-tools pipeline can handle data produced by various sequencing protocols. In practical terms, the Cell Ranger pipeline can estimate cell number automatically while the Drop-seq-tools and UMI-tools pipeline require an assignment of cell numbers for barcode extraction. By default, the Drop-seq-tools pipeline provides a trimming step to remove adaptor, which is not included in the Cell Ranger or UMI-tools pipeline.

In order to discuss the potential biases that may be introduced by different processing pipelines, Gao *et al*. used public scRNA-seq data to compare seven upstream pipelines with respect to time consumption, computational usage and downstream analysis results [13]. In 2017, Zieganhain *et al*. systematically evaluated six scRNA-seq protocols by measuring both sensitivity and precision of the methods [14]. Based on the measurements of sensitivity and precision, several evaluation analyses between sequencing platforms have been published. In 2019, Natarajan *et al*. analyzed scRNA-seq data produced by BGISEQ500 and Hiseq2500 to find whether there is a comparable performance between the two platforms [15]. Similarly, it is reported that single-cell transcriptomic profiles produced by MGISEQ 2000 platform and Illumina HiSeq 4000 platform were highly correlated [16].

Despite that comparing analysis of different platforms or processing pipelines for scRNA data have emerged in recent years, no research has systematically evaluated the match performance of sequencing platform coupled with processing pipeline for scRNA-seq data. Since wisely choosing platform-pipeline combinations could be helpful to meet the diverse needs of scRNA-seq based researches, for example, looking for more cell markers or cell types. To this aim, we used the public data from a benchmarking project in which each library was equally extracted from the same single-cell RNA pool of peripheral blood mononuclear cells (PBMCs) and sequenced by either NovaSeq 6000 platform or MGISEQ 2000 platform (see Materials and Methods), sequencing data was down sampled and subjected to three popular processing pipelines (the Cell Ranger, Drop-seq-tools, and UMI-tools pipeline) and thus six combinations of sequencing platform and analyzing pipeline were generated. Then we compared their performance in terms of quality, sensitivity, precision and downstream analysis outcomes to determine the suitable processing pipeline for each sequencing platform.

RESULTS

General comparison between platform-pipeline combinations

Firstly, we processed the FASTQ format data by down sampling the raw reads, because different sequencers conduced to divergent total numbers of reads (see Methods). Reads quality of data generated by the NovaSeq 6000 platform and MGISEQ 2000 platform analyzed by FastQC was shown in Fig.1 [17]. Quality scores of reads produced by the two platforms were similar and both in-between 34 and 37. Three processing pipelines Cell Ranger, Drop-seq-tools, UMI-tools were applied for sequencing data to acquire expression profiles. We then converted the outputs of Drop-seq-tools and UMI-tools into Cell Ranger output format for unified downstream analysis. Platform-pipeline combination was introduced to refer to the expression profile of each condition, for example, we claimed the expression profile that was produced by MGISEQ 2000 platform and processed by UMI-tools as "MGISEQ

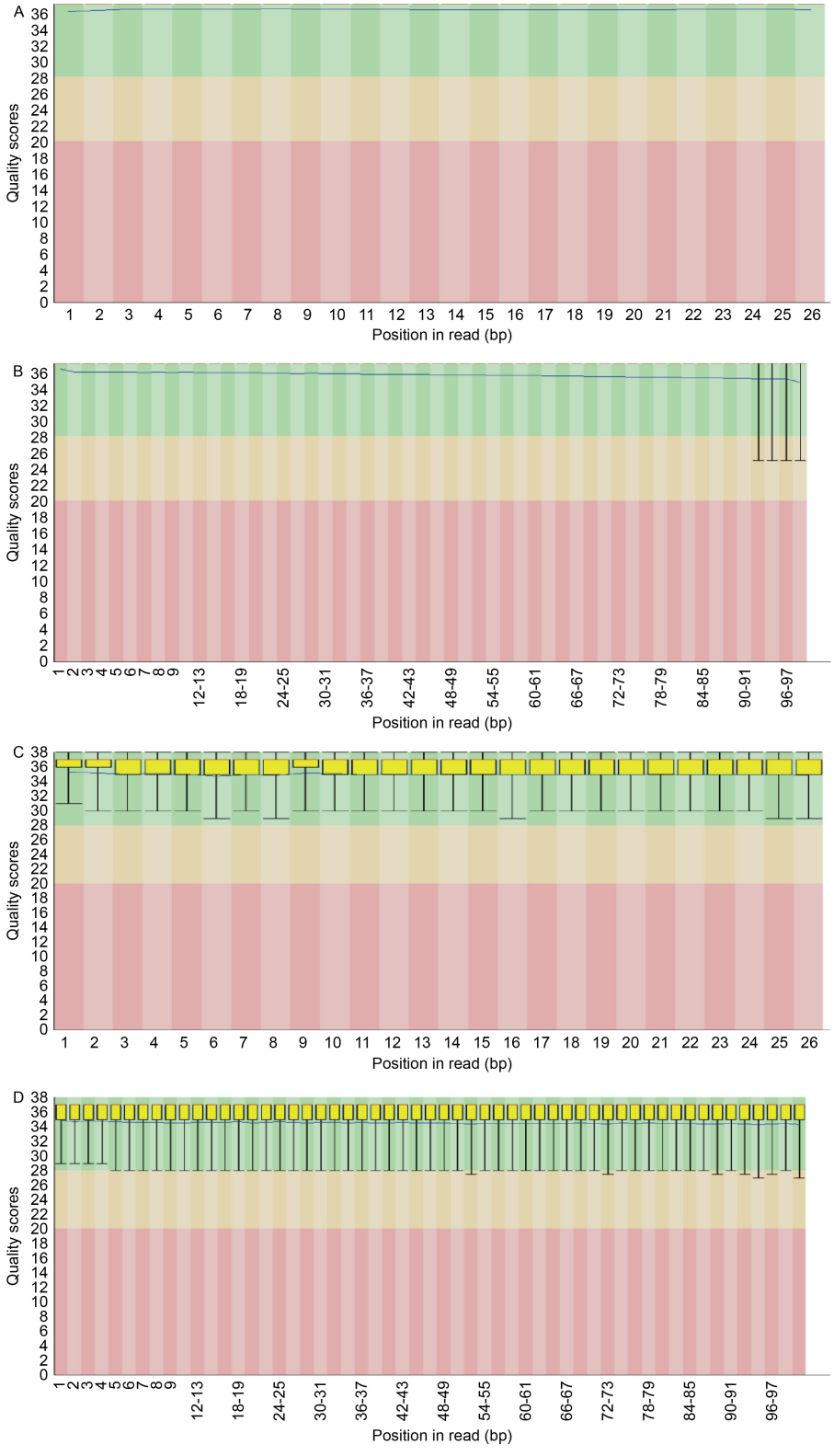

Figure 1. Comparing reads quality between the two sequencing platforms. (A) Quality score of R1 reads generated by NovaSeq 6000 platform. (B) Quality score of R1 reads generated by MGISEQ 2000 platform. (C) Quality score of R2 reads generated by NovaSeq 6000 platform. (D) Quality score of R2 reads generated by MGISEQ 2000 platform.

2000&UMI-tools". Considering that Cell Ranger is the mainstream tool released by 10X Genomics we randomly down sampled cells to the lowest number that has been detected by Cell Ranger to reduce batch effects and to gain 16,256 cells in each profile. Within this scope we then conducted correlation analysis (Fig. 2). Generally, correlation coefficients between platforms were higher than that between pipelines. Among the three pipelines, UMI-tools and Drop-seq-tools demonstrated the higher correlation coefficient in each platform.

UMI-tools has the highest sensitivity

To investigate the sensitivity, we down sampled the expression profiles to count the number of genes detected across depth ranging from 0 to 3 million reads. For both platforms, UMI-tools detected the highest number of genes (Fig. 3). When increasing the sequencing depth, the number of detected genes of UMI-tools were higher than that of Drop-seq-tools and Cell Ranger, but the difference was not so prominent between Drop-seq-tools and Cell Ranger. One possible reason may be the lack of the quality control step in UMI-tools by default. In consistent with the results of correlation analysis, sensitivity analysis show similar performance between the platforms rather than between pipelines.

Dropout possibility as a measurement of precision

Since amplification noise was assumed to be the same given that the libraries were from the same RNA pool, we calculated dropout rates to compare the precision. For each condition, we focused on genes which had non-zero expression within no less than 25% cell population. 1000 cells were randomly sampled and we calculated the faction of cells with zero transcript for each gene as dropout rate and estimated overall dropout possibility as the mean of dropout rate across all genes. The distribution of dropout rate was presented in Fig. 4. A consistent performance with respect to dropout possibility was observed across platforms. Drop-seq-tools has the highest dropout possibility for both platforms while Cell Ranger showed the lowest dropout possibility. Thus, it suggested that in term of precision, Cell Ranger outperformed the other two pipelines for both platforms.

Downstream analysis

As marker identification is the core of scRNA-seq analysis, we conducted downstream analysis for expression profiles of platform-pipeline combinations with Seurat package [18]. After filtering, PCA and cell clustering (Fig. 5A, Supplementary Figs. S1−S6), we found marker genes that distinguished between different

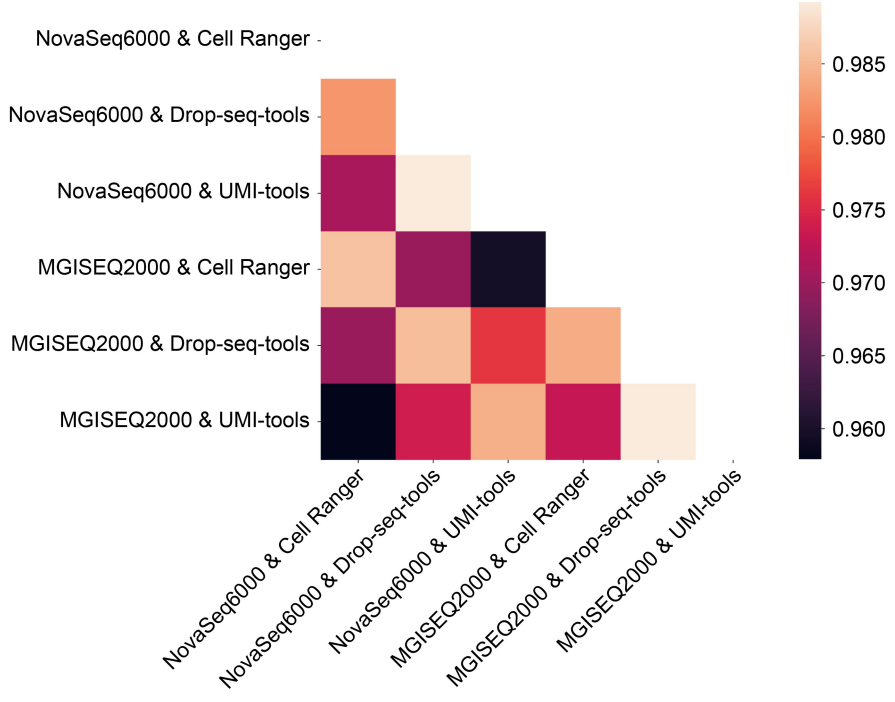

Figure 2. Correlation of expression between platform-pipeline combinations. Cells and features for correlation analysis were based on the intersection of all six combination. (Color bar indicating correlation coefficient.)

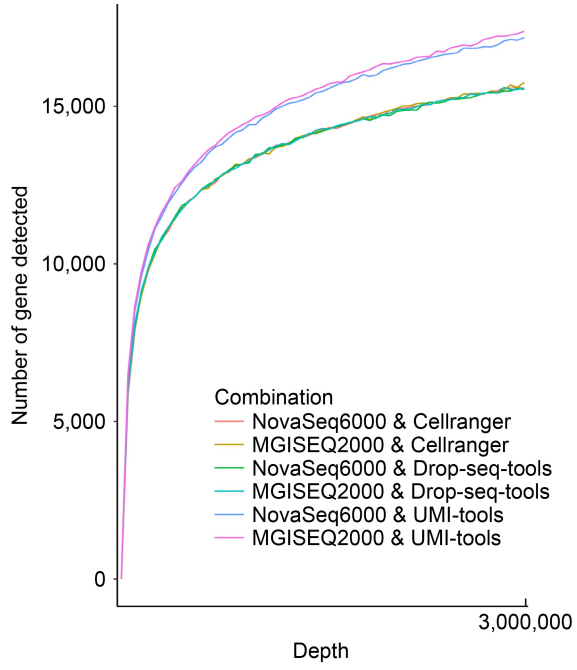

Figure 3. Sensitivity analysis of platform-pipeline combinations.

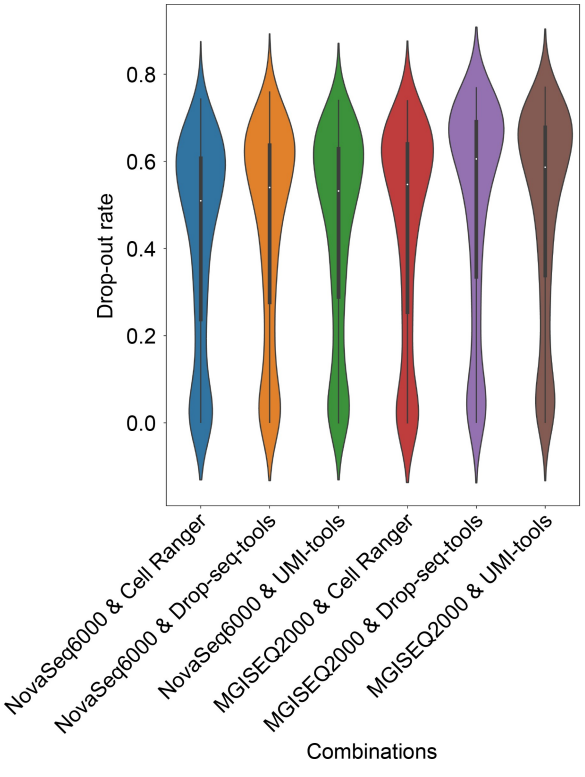

Figure 4. Precision comparison among platform-pipeline combinations.

cell types. In aspect of processing tools, UMI-tools showed the highest number of markers in both platforms which was coincident with its performance in sensitivity (Fig. 5B). Though sensitivity curves of Cell Ranger and Drop-seq-tools appeared to be undistinguishable, the number of detected markers of Cell Ranger was higher than that of Drop-seq-tools. Given the data-transforming steps included in Seurat analysis, it indicated that sensitivity may not be able to predict the results of downstream analysis such as marker detection.

Aside from the detection of marker genes, another issue which is intensively investigated in scRNA-seq analysis lies in cell-type classification. By cooperating marker genes identified by Seurat, we applied scCATCH package [19] to automatically annotate cell clusters. Almost every cell clusters had been assigned a cell identity (Fig. 5A). Figure 6A shows a major overlap of cell types between platform-pipeline combinations. Comparing with UMI-tools fewer markers had been identified by Drop-seq-tools, however, Drop-seq-tools contained more cell types for NovaSeq 6000 platform than the other tools, despite the fact that UMI-tools had more exclusive markers (Fig. 6B). All together indicated that the three pipelines had exerted similar effects on cell-type calling and Drop-seq-tools may be more sensitive in downstream analysis in term of cell-type detection.

DISCUSSION

Previous researches have been constrained within comparisons either between sequencing platforms or processing tools, none of them provided the suitable tool for a particular platform. To evaluate the performance of combinations of distinct sequencer and processing pipelines, we compared platform-pipeline combinations with respect to sensitivity, precision and downstream analysis outputs. We found MGISEQ 2000 platform demonstrating a comparable performance comparing with the NovaSeq6000 platform as the numbers of detected markers were close in two platforms and most cell types identified by Seurat were identical.

For data of low coverage, it is important to maintain a sufficient sensitivity. UMI-tools consistently performed well in aspects of sensitivity and marker identification while Drop-seq-tools showing inferior to UMI-tools and Cell Range. As to precision, Drop-seq-tools combined with both platforms had shown the worst performance in dropout rates which suggested that Drop-seq-tools may be not suitable for NovaSeq 6000 platform and MGISEQ 2000 platform comparing with the other two pipelines. UMI-tools displayed a comparable dropout possibility with Cell Ranger and the best performance in marker detection for both platforms, downstream analysis of data generated by UMI-tools also identified a satisfying number of cell types (Fig. 6). Combined with the performance of Cell Ranger and Drop-seq-tools in

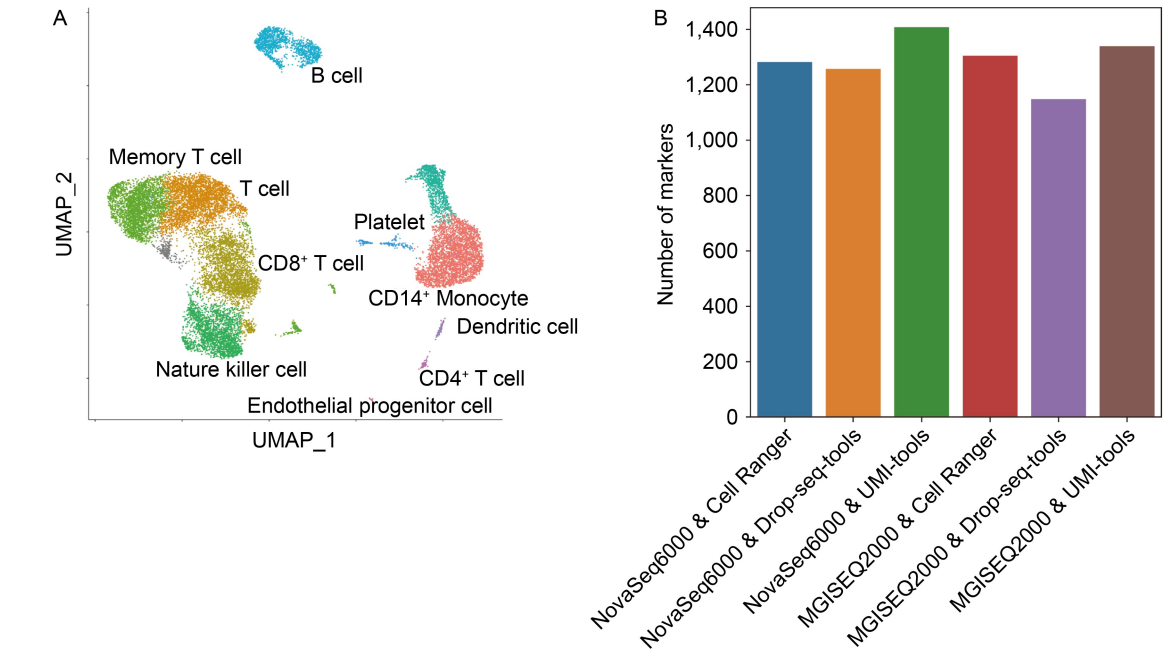

Figure 5. Downstream analysis. (A) UMAP graph showing the distribution of cells from different cell types (NovaSeq 6000 & Cell Ranger). UMAP graph was based on first 20 PC. (B)The number of markers of all platform-pipeline combinations detected by Seurat.

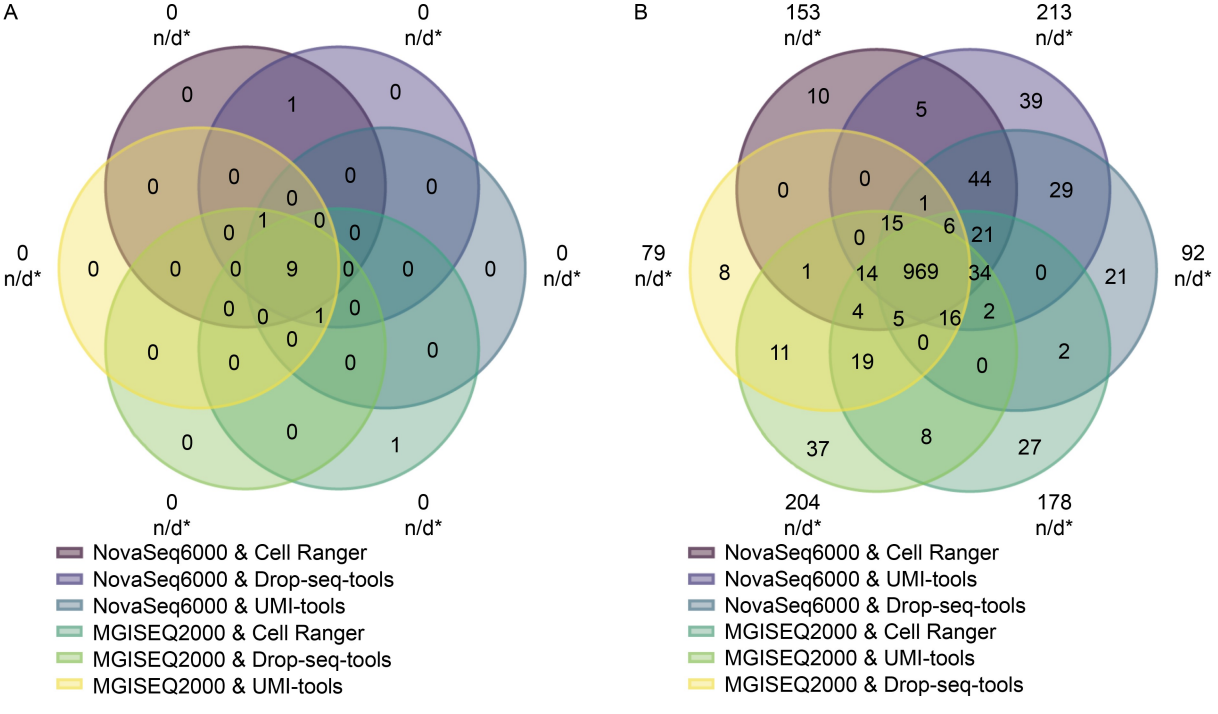

Figure 6. Inclusion relation in terms of cell types and cell markers. (A) Venn diagram of cell types across all platform-pipeline combinations. (B) Inclusion relation in terms of detected markers. (*elements of set in intersections that are not displayed, such as shared only between NovaSeq 6000 & Cell Ranger and MGISEQ 2000 & Cell Ranger)

cell-type detection, it implied that the number of markers does not necessary determine the recognition of cell types.

The high performance of Cell Ranger in precision and its comparable number of detected markers to that of UMI-tools made it a suitable choice for processing both

platforms. Moreover, Cell Ranger was characterized as highspeed, when encountering large dataset Drop-seq-tools has been reported to be one of the slowest processing tools [9] and in previous studies UMI-tools was described to be a relatively low speed software. To summarize, we suggest to use Cell Ranger to process data for large dataset produced by either NovaSeq 6000 platform or MGISEQ 2500 platform. For small dataset, we recommend UMI-tools to process sequencing results if more cell marker genes are needed otherwise we recommend Cell Ranger (see Table 1).

The comparisons above have its limits with regards to sample size, but it provided a frame work for evaluations of platform-pipeline combinations. Considering that sequencing results may differ in tissues and species, we expected further comparison of processing tools for different species as well as different tissues in the future.

MATERIALS AND METHODS

Data preprocessing

As the description of the dataset provided by the article of data source [20], peripheral blood samples were collected in vacutainer cell preparation tubes (BD Biosciences: 362753), PBMCs library was generated using the 10X Genomics Chromium system containing a pool of PBMCs from 14 donors. Then library was delivered to NovaSeq 6000 platform and MGISEQ 2000 platform for sequencing. Since MGISEQ 2000 platform produced more reads than NovaSeq 6000 sequencer, after downloading FASTQ data from ArrayExpress (E-MTAB-9024, PBMC2) we down sampled the raw reads to the depth of 2 billion reads.

Analysis of scRNA-seq data

First, we applied FastQC to check the reads quality for each platform, then delivered fastq format data to the three processing pipelines Drop-seq-tools, UMI-tool and Cell Ranger with cell number automatically detected. To reduce batch effects on cell detection brought by different pipelines, we randomly down sampled the cells to 16,256 which was also the cell number detected by the Cell Ranger. We transformed the expression tables output by Drop-seq-tools and UMI-tools into Cell Ranger output format for analyzing steps. Downstream analyses were implemented using R version 3.5.2 and package Seurat (version 3.2.2). For each expression matrix we removed cells containing less than 500 features or more than 10% mitochondrial expression, the remaining profiles were normalized and scaled. Principle component analysis (PCA) was performed with the first 20 principle components selected for cell clustering by shared nearest neighbor (SNN) graph. After dimensional reduction by UMAP, we applied scCATCH package (2.1) for cell-type classification and marker identification. For each cluster with multiple assignments we preserved only one cell identity.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.15302/ J-QB-022-0295.

ACKNOWLEDGEMENTS

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (Nos. XDB38050200 and XDA26040304).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Weiran Chen, Md Wahiduzzaman, Quan Li, Yixue Li, Guangyong Zheng and Tao Huang declare that they have no conflict of interest or financial conflicts to disclose.

This article does not contain any studies with human or animal subjects performed by any of the authors.

OPEN ACCESS

Table 1 Summarizing evaluation of all platform-pipeline combinations and our recommendations

	Cell Ranger		Drop-seq-tools		UMI-tools	
Platform	NovaSeq 6000	MGISEQ 2000	NovaSeq 6000	MGISEQ 2000	NovaSeq 6000	MGISEQ 2000
Sensitivity	Medium	Medium	Medium	Medium	High	High
Dropout possibility	0.420984	0.443619	0.446616	0.497096	0.442754	0.488321
Number of marker	1282	1305	1257	1148	1408	1339
Number of cell types	11	11	12	11	11	11
Speed	High	High	Low	Low	Relatively low	Relatively low
Large dataset	Recommended	Recommended	–	–	–	–
Small dataset	Recommended (good precision)	Recommended (good precision)		–	Recommended (more markers)	Recommended (more markers)

REFERENCES

1. Hwang, B., Lee, J. H. and Bang, D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med., 50, 1−14
2. Trapnell, C. (2015) Defining cell types and states with single-cell genomics. Genome Res., 25, 1491−1498
3. Michalopoulos, G. K. (2021) Novel insights into liver homeostasis and regeneration. Nat. Rev. Gastroenterol. Hepatol., 18, 369−370
4. Chen, G., Ning, B. and Shi, T. (2019) Single-cell RNA-seq technologies and related computational data analysis. Front. Genet., 10, 317
5. Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., *et al*. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science, 327, 78−81
6. Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet., 17, 333−351
7. Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., *et al*. (2017) Massively parallel digital transcriptional profiling of single cells. Nat. Commun., 8, 14049
8. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., *et al*. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161, 1202−1214
9. Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res., 27, 491−499
10. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. and Hellmann, I. (2018) zUMIs—A fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience, 7, 6
11. Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D. T., Samsonova, M. G. and Kharchenko, P. V. (2018) dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. Genome Biol., 19, 78
12. Tian, L., Su, S., Dong, X., Amann-Zalcenstein, D., Biben, C., Seidi, A., Hilton, D. J., Naik, S. H. and Ritchie, M. E. (2018) scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLOS Comput. Biol., 14, e1006361
13. Gao, M., Ling, M., Tang, X., Wang, S., Xiao, X., Qiao, Y., Yang, W. and Yu, R. (2021) Comparison of high-throughput single-cell RNA sequencing data processing pipelines. Brief. Bioinform., 22, bbaa116
14. Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I. and Enard, W. (2017) Comparative analysis of single-cell rna sequencing methods. Mol. Cell, 65, 631−643.e4
15. Natarajan, K. N., Miao, Z., Jiang, M., Huang, X., Zhou, H., Xie, J., Wang, C., Qin, S., Zhao, Z., Wu, L., *et al*. (2019) Comparative analysis of sequencing technologies for single-cell transcriptomics. Genome Biol., 20, 70
16. Jeon, S. A., Park, J. L., Kim, J. H., Kim, J. H., Kim, Y. S., Kim, J. C. and Kim, S. Y. (2019) Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing. Genomics Inform., 17, e32
17. Andrews, S. (2014) FastQC: A quality control tool for high throughput sequence data
18. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. 3rd, Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019) Comprehensive integration of single-cell data. Cell, 177, 1888−1902.e21
19. Shao, X., Liao, J., Lu, X., Xue, R., Ai, N. and Fan, X. (2020) Sccatch: automatic annotation on cell types of clusters from single-cell rna sequencing data. iScience, 23, 100882
20. Senabouth, A., Andersen, S., Shi, Q., Shi, L., Jiang, F., Zhang, W., Wing, K., Daniszewski, M., Lukowski, S. W., Hung, S. S. C., *et al*. (2020) Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. NAR Genom. Bioinform., 2, lqaa034

