UniverSC: a flexible cross-platform single-cell data processing pipeline

S. Kelly
RIKEN
Kai Battenberg
RIKEN
Nicola Hetherington
RIKEN
Makoto Hayashi
RIKEN
Aki Minoda (akiko.minoda@riken.jp)
RIKEN https://orcid.org/0000-0002-2927-5791

Methods Brief Communication

Keywords: RNA-sequencing, universal single-cell processing tool, UniverSC

Posted Date: March 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-244461/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on November 11th, 2022. See the published version at https://doi.org/10.1038/s41467-022-34681-z.

UniverSC: a flexible cross-platform single-cell data processing pipeline

S. Thomas Kelly[1†], Kai Battenberg[1,2†], Nicola A. Hetherington[1], Makoto Hayashi[2], Aki Minoda[1]*

* Corresponding author. Email: akiko.minoda@riken.jp

† These authors contributed equally to this work.

1. Laboratory for Cellular Epigenomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

2. Plant Symbiosis Research Team, RIKEN Center for Sustainable Resource Science, Yokohama, Japan

Abstract

Single-cell RNA-sequencing analysis to quantify RNA molecules in individual cells has become popular owing to the large amount of information one can obtain from each experiment. We have developed UniverSC (https://github.com/minoda-lab/universc), a universal single-cell processing tool that supports any UMI-based platform. Our

19 command-line tool enables consistent and comprehensive integration, comparison, and
20 evaluation across data generated from a wide range of platforms.
21

22 # Main

23 Single-cell genomics technologies has led to a recent surge in studies of cellular
24 heterogeneity. A common workflow for many of the single-cell RNA-sequencing
25 (scRNA-seq) technologies involves capturing individual cells, either in gel emulsion
26 with beads or in wells, followed by the addition of a unique molecular identifier (UMI)
27 to each RNA molecule to reduce the effect of amplification bias. Throughput has
28 increased over the years and current technologies generate scRNA-seq data for
29 thousands to hundreds of thousands of cells from one run, some of which are
30 commercially available (e.g., Chromium by 10x Genomics, Nadia by Dolomite Bio, and
31 ddSEQ by Bio-Rad). This increase in throughput has made it possible for researchers to
32 apply scRNA-seq to a whole range of tissues as well as whole organisms[1,2,3]. It is
33 expected that scRNA-seq will become more accurate, more reliable, and with lower cost
34 per cell, being feasible for a wide range of studies as the technology is developed
35 further[4].
36

Leveraging on the fact that most scRNA-seq technologies utilize the same concept of cell barcodes and UMIs, we developed UniverSC, a shell utility that works as a wrapper for Cell Ranger (10x Genomics), which can handle datasets generated by a wide range of single-cell technologies. Cell Ranger was chosen as a unifying pipeline for several reasons: 1) it is optimized to run in parallel on a cluster and the installation is fairly straightforward, 2) many labs working on single-cell analysis are likely to already be familiar with the outputs, 3) many tools have already been released for downstream analysis of the output format due to its popularity, 4) the rich summary information and post-processing is useful for further optimization and troubleshooting if necessary, and 5) the latest open-source release (version 3.0.2) has been optimized further by adapting open-source techniques such as third-party "EmptyDrops"[5] for cell calling or filtering, which does not assume thresholds specific for the Chromium platform (10x Genomics).

Conceptually, UniverSC carries out its entire process in seven steps (Fig. 1). Given a set of paired-end sequence files in FASTQ format (R1 and R2), a genome reference (as required by Cell Ranger), and the name of the selected technology, UniverSC reformats the whitelist barcodes and sequence files to fit what is expected by Cell Ranger. Additionally, UniverSC provides a file with summary statistics including mapping rate,

read and UMI counts for each barcode, and averages for the filtered cells. Sequence trimming based on adapter contamination or sequencing quality is not included in the pipeline and no trimming is required to pass files to UniverSC. However, trimming is highly recommended particularly on R2 files from Illumina platforms, as this generally improves the mapping quality. This requires careful data handling to ensure that all Read 1 and Read 2 are strictly in pairs while only trimming Read 2. We provide a script for convenience that filters Read 1 and Read 2 by quality of Read 2 and avoids mismatching cell barcodes. In principle, UniverSC can be run on any droplet-based or well-based technologies (see the software documentation and Table S1 for more details), provided that they contain UMIs. Settings can also be restored to run 10x Genomics Chromium samples as changes made on Cell Ranger by UniverSC are reversible.

At initial release, UniverSC has pre-set parameters for 19 technologies (Table S1). Further technologies can be used with "custom" input parameters for any barcode and UMI lengths or by requesting a feature to be added to the GitHub repository. Testing datasets for the following settings are provided: 10x Genomics version 2 and 3 (default), DropSeq, and ICELL8. UniverSC is freely available at GitHub

(https://github.com/minoda-lab/universc), and at DockerHub (https://hub.docker.com/repository/docker/tomkellygenetics/universc). See methods for details on how to install and run UniverSC.

We demonstrate our method using published test data from human cell lines that we provide with the package for 10x Genomics (version 3)[6], DropSeq[7], and ICELL8[8]. DropSeq is an example of a droplet-based single-cell technology that does not have known barcodes so a whitelist needs to be generated for compatibility. ICELL8 is a well-based technology that has a known barcode whitelist and allows selecting subsets of wells by known barcodes. These represent two different classes of technologies with different configurations for processing cell barcodes. To assess the degree of similarity between UniverSC and other pipelines, we processed the three test datasets through UniverSC and platform-specific pipeline: Cell Ranger (version 3.0.2) (10x Genomics: https://www.10xgenomics.com) for 10x Genomics data, dropSeqPipe (version 0.6)[9] for DropSeq data, and CogentAP (version 1.0) (Takara Bio Inc.: https://www.takarabio.com) for ICELL8 data (Fig. 2). For all three test datasets, the gene-barcode matrix (GBM) generated through UniverSC with the GRCh38 (hg38) was highly correlated with GBM generated by the coupled pipeline (Fig. 2). As expected, it

was identical (r=1) in the case of Cell Ranger 3.0.2, and was 0.96 or higher in the two other sets of GBMs. Likewise, clustering results were also highly similar: Adjusted Rand Index (ARI) was 1 in the case of 10x Genomics data, 0.74 and 0.85 for DropSeq data and ICELL8 data, respectively. The UniverSC outputs for each of these technologies are provided in the supplementary materials. These results should not be interpreted as biologically meaningful and we recommend running UniverSC on full datasets from experiments or public databases to gain biological insights.

By analyzing datasets from different platforms with the same data processing criteria, we envision UniverSC will contribute in several ways. Firstly, it will enable a fair comparison when evaluating the best platform for a specific sample type, which may be especially important with challenging samples, such as those containing large cells. Secondly, novel single-cell technologies are developed, they will not require a dedicated data processing pipeline for their own technology. Lastly, as single-cell technologies become integral to a wide range of studies, mitigation of technical errors and integration of scRNA-seq data generated across different groups and platforms will be necessary. Processing data that contains various barcode and UMI configurations under consistent framework will most likely become not only convenient but essential. While there are

pipelines that can be configured for a variety of technologies (dropSeqPipe[9]; dropEst[10], Kallisto/BUStools[11]), Cell Ranger performs well in a server or cluster environment and generates a rich and informative output summary. UniverSC enables cross-platform single-cell data integration with a command-line interface, eliminating the need to install or configure separate pipelines for each platform. We believe our tool will also facilitate the generation of reproducible results. We provide this tool for free and open-source to democratize single-cell analysis for a wide range of scientific applications.

Acknowledgements

We acknowledge contributions from Tommy Terooatea of RIKEN IMS for testing UniverSC, Jonathon Moody and Chung-Chau Hon of RIKEN IMS for their insightful discussion. We also acknowledge Shuwen Chen, Tsuyoshi Okumo, Max Sanchez, and Karthik Swaminathan (Takara Bio) for support analysing data from ICELL8 platform with their CogentAP pipeline. We thank developers at 10x Genomics of Cell Ranger and dependencies for making their code publicly available. We also thank Marcus Kinsella (CZI) for releasing a docker image of an open-source version of Cell Ranger 2.0.2.

Funding

This work was supported by a JSPS KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas "Principles of pluripotent stem cells underlying plant vitality" (JP17H06470 to A.M. and 17H06472 to M. H.). *Conflict of Interest*: None to declare. We have no affiliation with 10x Genomics, Dolomite Bio, Takara Bio, or any other vendor.

Author contributions

STK and KB generated the UniverSC script, carried out the comparative analysis, and wrote the manuscript. NAH generated datasets and tested the script. MH and AM supervised the project. AM edited the manuscript.

References

1. Cao, J., et al. *Science* **357**, 661-667 (2017).

2. Regev, A., et al. *eLife* **6**, e27041 (2017).

3. The Tabula Muris Consortium. *Nature* **562**, 367-372 (2018).

4. Kulkarni, A., Anderson, A. G., Merullo, D. P. & Konopka, G. *Curr. Opin. Biotech*. **58**, 129-136 (2019).

5. Lun, A.T.L., et al. *Genome Biol.* **20**, 63 (2019).

6. Zheng, G. X. Y., et al. *Nat Commun.* **8**, 14049 (2017).

7. Macosko, E. Z., et al. *Cell* **161**(5), 1202-1214 (2015).

8. Goldstein, L.D., et al. *BMC Genomics* **18**, 519 (2017).

9. Roeilli, P., Mueller, S., Girardot, C. & Kelly, S. T. *GitHub repository* https://github.com/Hoohm/dropSeqPipe/tree/develop (Accessed 13 January, 2021)

10. Petukhov, V., et al. *Genome Biol.* **19**, 78 (2018).

11. Melsted, P., Ntranos, V. & Pachter, L. *Bioinformatics* **35(21)**, 4472-4473 (2019).

Methods

The set of input parameters for UniverSC is similar to that required by Cell Ranger, with a few additions. The UniverSC workflow requires paired-end FASTQ input files and reference data as prepared by Cell Ranger. By default, UniverSC assumes Read 1 of the FASTQ to contain the cell barcode and UMI and Read 2 to contain the transcript sequences which will be mapped to the reference, as is common in 3' scRNA-seq protocols. Given a known barcode and UMI length, UniverSC will check the file name and barcodes, altering the configurations to match that of Chromium as needed. The "chemistry" appropriate for each single-cell technology for 3' scRNA-seq is determined automatically (technologies for 5' scRNA-seq other than that of Chromium are not supported at time of writing). Data from multiple lanes is supported and so is a custom set of barcodes specific to a given technology other than 10x Genomics.

Test data for 10x Genomics[6] has been downloaded from the 10x Genomics website (10x Genomics: https://www.10xgenomics.com). Test data for DropSeq[7] has been downloaded from GEO (Accession GSE63473) and prepared to match the same reference data. The Nadia technology uses the same barcode design as DropSeq (beads

supplied by ChemGenes). Test data for ICELL8[8] were obtained from EGA (Accession EGAD00001003443) and filtered for reads matching the same loci.

Each test data was processed in parallel to generate two GBMs consisting of UMI counts. The pair of GBMs were adjusted to have matching sets of barcodes and genes: only barcodes found in both GBMs were kept, and genes only found in one GBM was added to the other with 0 UMIs assigned. The adjusted pair of GBMs were then used to carry out clustering analysis with an R package Seurat[12]. The adjusted pair of GBMs were then used to carry out clustering analysis with an R package Seurat. Finally, the Pearson correlation between the GBMs and the adjusted ARI between the two clustering outcomes were calculated.

We provide documentation for UniverSC accessible as a manual and help system in the terminal and a user-interface which checks file inputs and gives error messages to identify potential problems. UniverSC can be run on any Unix-based system in the shell and the source code is publicly available along with installation instructions at GitHub (https://github.com/minoda-lab/universc), and a docker image is also available at DockerHub with all dependencies installed from source

https://hub.docker.com/repository/docker/tomkellygenetics/universc). We recommend installing UniverSC in a local directory (to a home directory) or somewhere with write access, it can be run on any system with Cell Ranger installed (i.e. added to the PATH environment variable). We also recommend running UniverSC on a server with sufficient memory to run the STAR alignment algorithm. Submission to a cluster in parallel with a job scheduler is supported, but note that UniverSC can only run on one technology at a time due to the different barcode whitelist requirements. A continuous integration service will be used to test updates and maintain the software. See the manual for further details. Note that UniverSC was developed by a third-party unrelated to 10x genomics, and an open-source version of Cell Ranger 3.0.2 is used with Cloupe (a portion of Cell Ranger) inactivated to comply with 10x Genomics End User Software License Agreement.

Supplementary Data

Zip files for samples of output are provided for each technology as an example of the processed data returned by UniverSC. This data can also be used to reproduce the results in Figure 2.

Supplementary file 1: Results for UniverSC 1.0.0 on10x Genomics version 3 data (Zheng et al., 2017).

Supplementary file 2: Results for Cell Ranger 3.0.2 on10x Genomics version 3 data (Zheng et al., 2017).

Supplementary file 3: Results for UniverSC 1.0.0 on DropSeq data (Macosko et al., 2015).

Supplementary file 4: Results for dropSeqPipe 0.6 on DropSeq data (Macosko et al., 2015).

Supplementary file 5: Results for UniverSC 1.0.0 on ICELL8 data (Goldstein et al., 2017).

Supplementary file 6: Results for CogentAP 1.0 on ICELL8 data (Goldsein et al., 2017).

Glossary

Here follows a summary of definitions used in here and in the package documentation to avoid ambiguity.

Bioinformatics Procedure The steps taken to process the data, typically formatting data using existing algorithms and scripting languages.

Cell Barcode A short nucleotide sequence incorporated as a part of each fragment in a library that is used to determine the corresponding cell.

Chemistry The term used by 10x Genomics to refer to different "versions" of their experimental kits. Here we use it to describe different parameters for Cell Ranger to account for these differences in the chemistry used to prepare the samples.

Index The index adapter sequence used in multiplexed sequencing, typically for identifying each sample, using the i7 or i5 indexes for Illumina platforms. Some technologies use these for additional cell barcodes.

Library In the context of genomics this refers to the genomic or complementary DNA prepared for sequencing on an NGS sequencing platform.

Platform The instrument used to perform single-cell encapsulation (e.g., Chromium, Nadia, ICELL8) or NGS (e.g., HiSeq2500, NovaSeq6000, MGISEQ-2000).

Single-Cell Encapsulation An experiment to capture individual cells in wells or droplets.

Single-Cell RNA-seq Combining single-cell encapsulation with next-generation sequencing of complementary RNA to gain an expression profile of individual cells.

Whitelist The appropriate list of expected barcodes used as reference.

Unique Molecular Identifier (UMI) A short nucleotide sequence incorporated as a part of each fragment in a library that is used to determine the corresponding molecule.

References (supplementary material)

12. Stuart, T., Butler, A., et al. *Cell* **177**(7), 1888-1902.e21 doi: 10.1016/j.cell.2019.05.031 (2019).

13. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Reports* **2**, 666-673 (2012).

14. Hashimshony, T., et al. *Genome Biol*. **17**, 77 (2016).

15. Yan, Y. *GitHub repository* https://github.com/yanailab/celseq2 (Accessed July 10, 2020).

16. Veres, A. & Lee, C. H. *GitHub repository* https://github.com/indrops/indrops (Accessed July 10, 2020)

17. Klein, A. M., et al. *Cell* **161(5)**, 1187-1201 (2015).

18. Zilionis, R., et al. *Nat Protoc.* **12**, 44-73 (2017).

19. Jaitin, D. A., et al. *Science* **343(6172)**, 776-779 (2014).

20. Keren-Shaul, H., et al. *Nat Protoc.* **14**, 1841-1862 (2019).

21. Sasagawa, Y., et al. *Genome Biology* **114**, R31 (2013).

22. Vitak, S. A., et al. *Nat Methods* **14(3)**, 302-308 (2017).

23. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T., S. *bioRxiv* doi:10.1101/003236 (2014).

24. Bagnoli, J. W., et al. *Nat Commun.* **9**, 2937 (2018).

25. Gierahn, T. M., et al. *Nat. Methods* **14(4)**, 395-398 (2017).

26. Hagemann-Jensen, M., et al. *bioRxiv* doi:10.1101/817924 (2019).

27. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellman, I. *GigaScience* **7(6)** doi:10.1093/gigascience/giy059 (2018).

28. Rosenberg, A.B., et al. *Science* **360(6385)**, 176-182 (2018).

29. Romagnoli, D., et al. *BMC Genomics* **19**, 960 (2018).

30. Teichmann Group. *GitHub repository* https://teichlab.github.io/scg_lib_structs/methods_html/SureCell.html (Accessed July 10, 2020)

Figures and Tables

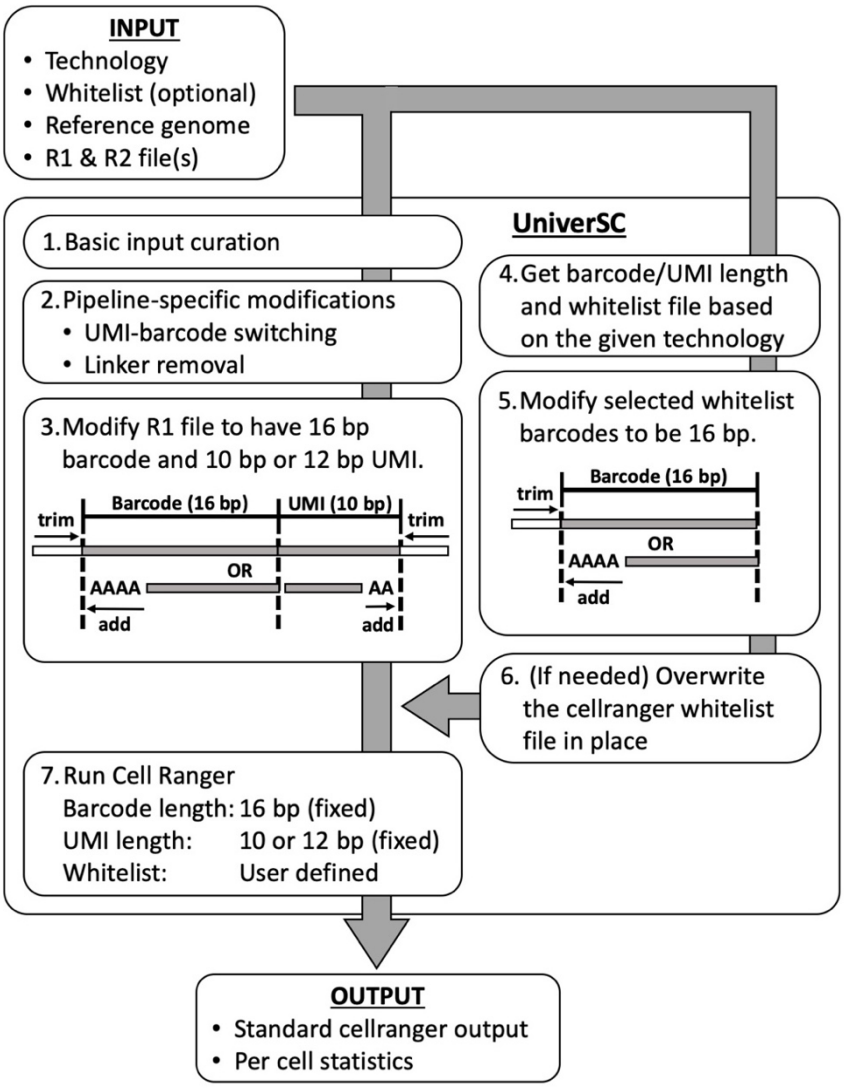

Figure 1. **Overview of UniverSC**. Given a pair of FASTQ files (R1 and R2), a genome reference (as required by Cell Ranger), and the name of the technology UniverSC first runs a basic input curation (step-1). The curated input files are then adjusted for pipeline-specific modification (step-2) and subsequently reformatted to match the expected barcode and UMI lengths (step-3). In parallel, the barcode whitelist suited for the technology (if unspecified by the user) is determined (step-4), and the whitelist

barcodes are modified to 16 bp (step-5). If the selected whitelist is different from the whitelist in place for Cell Ranger at the moment, the whitelist is replaced (step-6). Finally, the modified sample data is processed by Cell Ranger against the modified whitelist (step-7) to generate a standard output along with a summary file with per cell statistics.

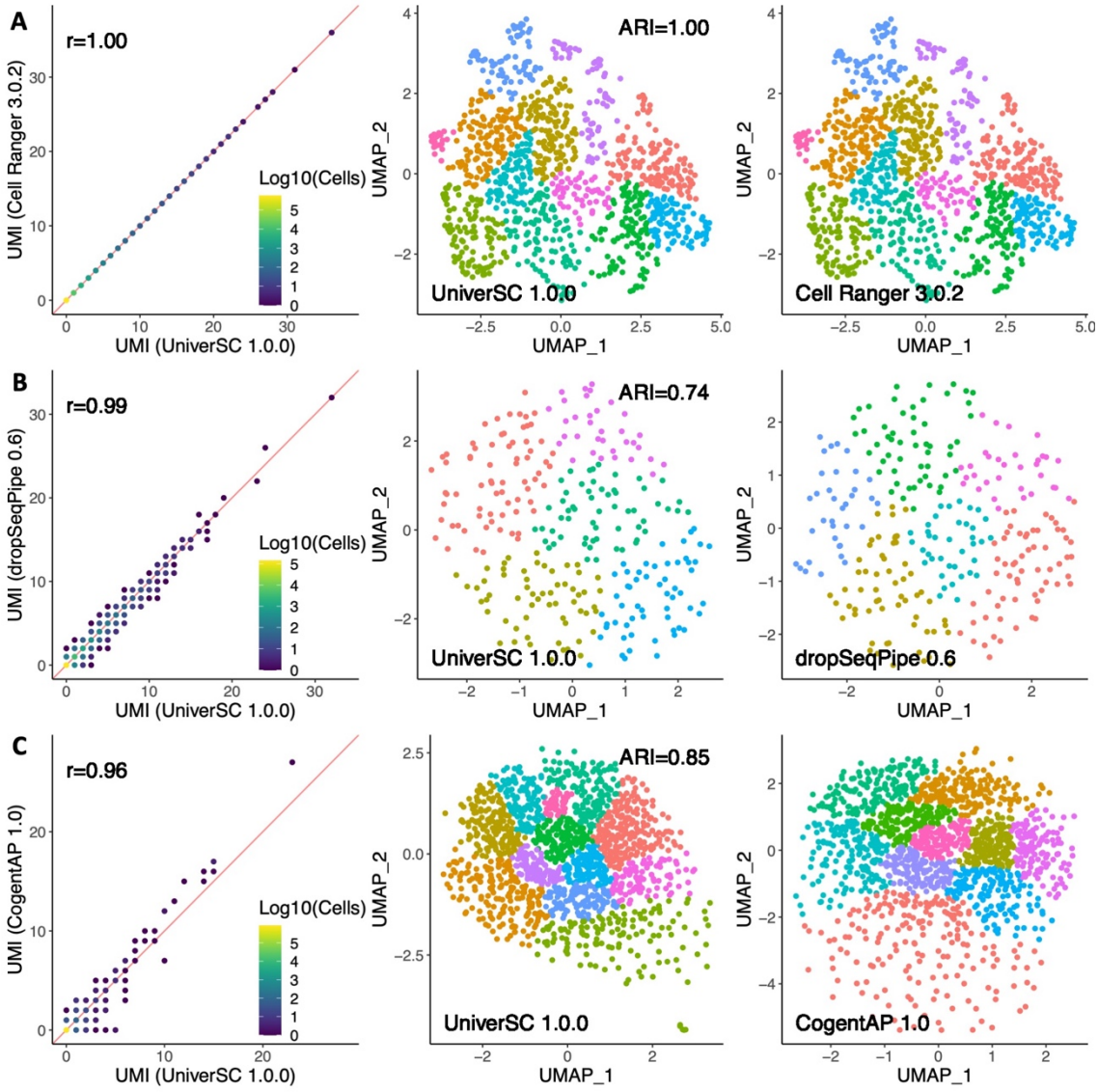

Figure 2. **Similarity assessment of UniverSC against other pipelines.** Comparisons between the GBM generated by UniverSC 1.0.0 against Cell Ranger 3.0.2 (A), dropSeqPipe 0.6 (B), and CogentAP 1.0 (C). Direct comparison of GBMs is on the left column followed by the clusters resulting from UniverSC 1.0.0 (center column) and its counterpart (right column). The processed data compared here is provided in the supplementary data.

Table S1. **Technologies currently available and settings used by UniverSC.**

Parameter value	Technology	Barcode length[†]	UMI length	Reference
10x-v2 (or 10x)	10x (version 2) Instrument: Chromium Vendor: 10x Genomics	16	10	10x Genomics; [6]*[6]
10x-v3 (or 10x)	10x (version 3) Instrument: Chromium Vendor: 10x Genomics	16	12	10x Genomics
celseq	CEL-Seq	8	4	[13, 14]
celseq2*[1]	CEL-Seq2	6	6	[14, 15]
icell8	ICELL8 Instrument: ICELL8 Vendor: Takara Bio	11	14	Takara Bio Inc.; [8]*[6]
indrops-v1*[2,3]	inDrop (version 1)	19	6	[16, 17]

indrops-v2*[2,3]	inDrop (version 2) Vendor: 1CellBio	19	6	[16, 18]
indrops-v3*[3,4]	inDrop (version 3)	11	6	[18]
Nadia (or dropseq)	Nadia or DropSeq Instrument: Nadia Vendor: Dolomite Bio	12	8	Dolomite Bio; [7]*[6]
marsseq-v1	MARS-Seq	6	10	[19]
marsseq-v2	MARS-Seq 2.0	7	8	[20]
quartz-seq2-1536	Quartz-Seq2 (1536 wells)	15	8	[21]
quartz-seq2-384	Quartz-Seq2 (384 wells)	14	8	[21]
sciseq*[1,4]	SCI-seq	10	8	[1, 22]
scrbseq	SCRB-Seq, mcSCRB-Seq	6	10	[23, 24]
seqwell	plexWell vendor: seqWell	12	8	[25]
smartseq	SMART-Seq (version 3)	11	8	[26, 27]

	Vendor: Takara Bio (v2)			
splitseq*[1,2,5]	SPLiT-Seq	18	10	[27, 28]
surecell*[5]	SureCell Instrument: ddSEQ Vendor: BioRad	18	8	Bio-Rad; [29, 30]

† Barcode length is max or total (linkers are removed automatically where needed) excluding barcodes in the index files which requires demultiplexing.

*[1] These technologies have their UMIs before their barcodes. The positions of UMIs and barcodes are automatically inverted when these technologies are selected as options.

*[2] These technologies have their barcodes and UMIs in R2 rather than R1. The functional role of R1 and R2 are automatically inverted when these technologies are selected as options.

*[3] These technologies have their barcode in two segments with Barcode-1 (8-11 bp) and Barcode-2 (8 bp). The first 19 bp of the adjusted R1 file is originally recognized as barcode, but only the first 16 bp are used upon assigning reads to cells.

*[4] These technologies have dual indexes (I1 and I2 from the i7 and i5 indexes from Illumina), which contain additional information on cell barcode rather than sample. These require demultiplexing with bcl2fastq before running UniverSC.

*[5] These technologies have their barcode is in three segments with Barcode-1 (6 bp), Barcode-2 (6 bp), and Barcode-3 (6 bp). The first 18 bp of the adjusted R1 file is originally recognized as barcode, but only the first 16 bp are used upon assigning reads to cells.

*[6] Test data used in our study was generated from data this paper originally published.

Figures

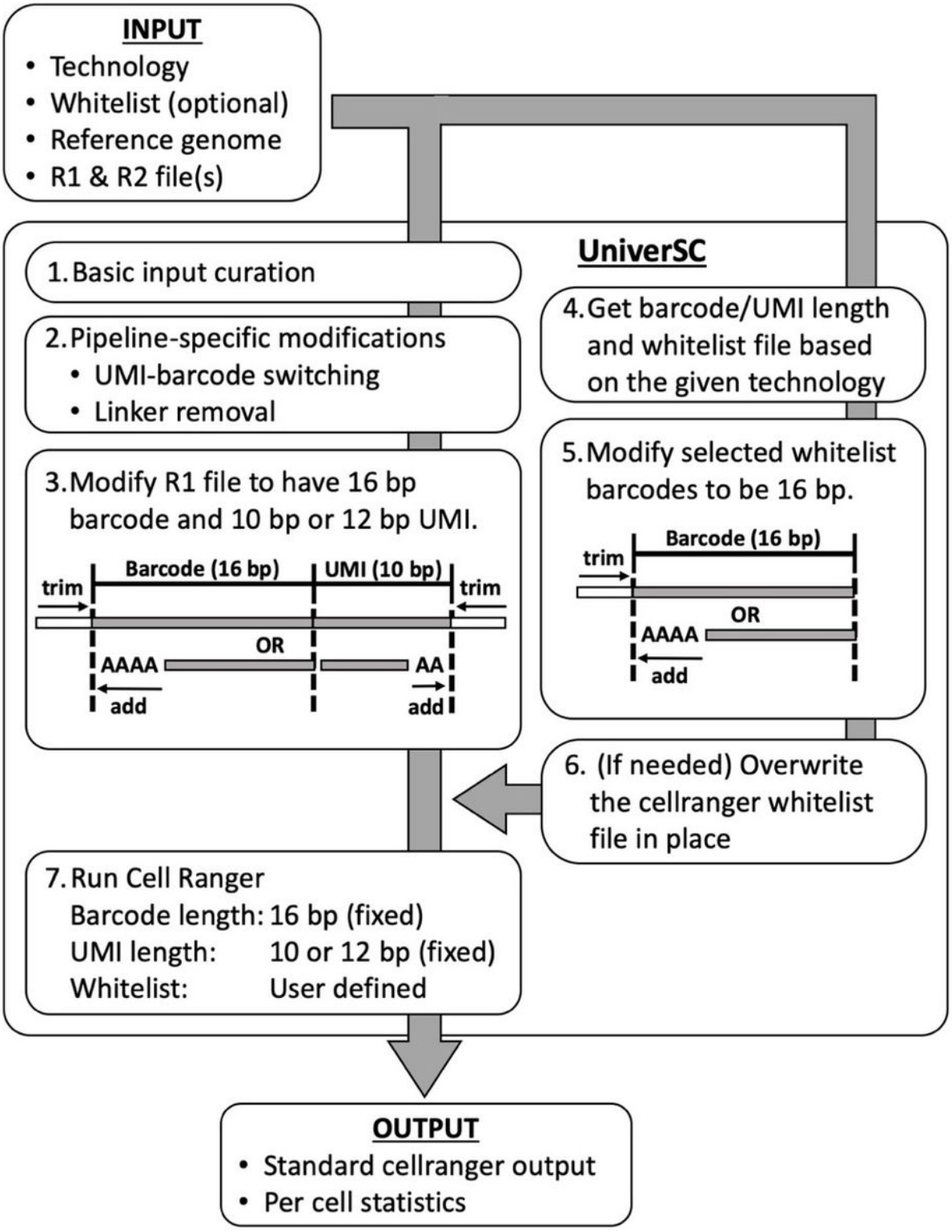

Figure 1

Overview of UniverSC. Given a pair of FASTQ files (R1 and R2), a genome reference (as required by Cell Ranger), and the name of the technology UniverSC first runs a basic input curation (step-1). The curated input files are then adjusted for pipeline-specific modification (step-2) and subsequently reformatted to

match the expected barcode and UMI lengths (step-3). In parallel, the barcode whitelist suited for the technology (if unspecified by the user) is determined (step-4), and the whitelist barcodes are modified to 16 bp (step-5). If the selected whitelist is different from the whitelist in place for Cell Ranger at the moment, the whitelist is replaced (step-6). Finally, the modified sample data is processed by Cell Ranger against the modified whitelist (step-7) to generate a standard output along with a summary file with per cell statistics.

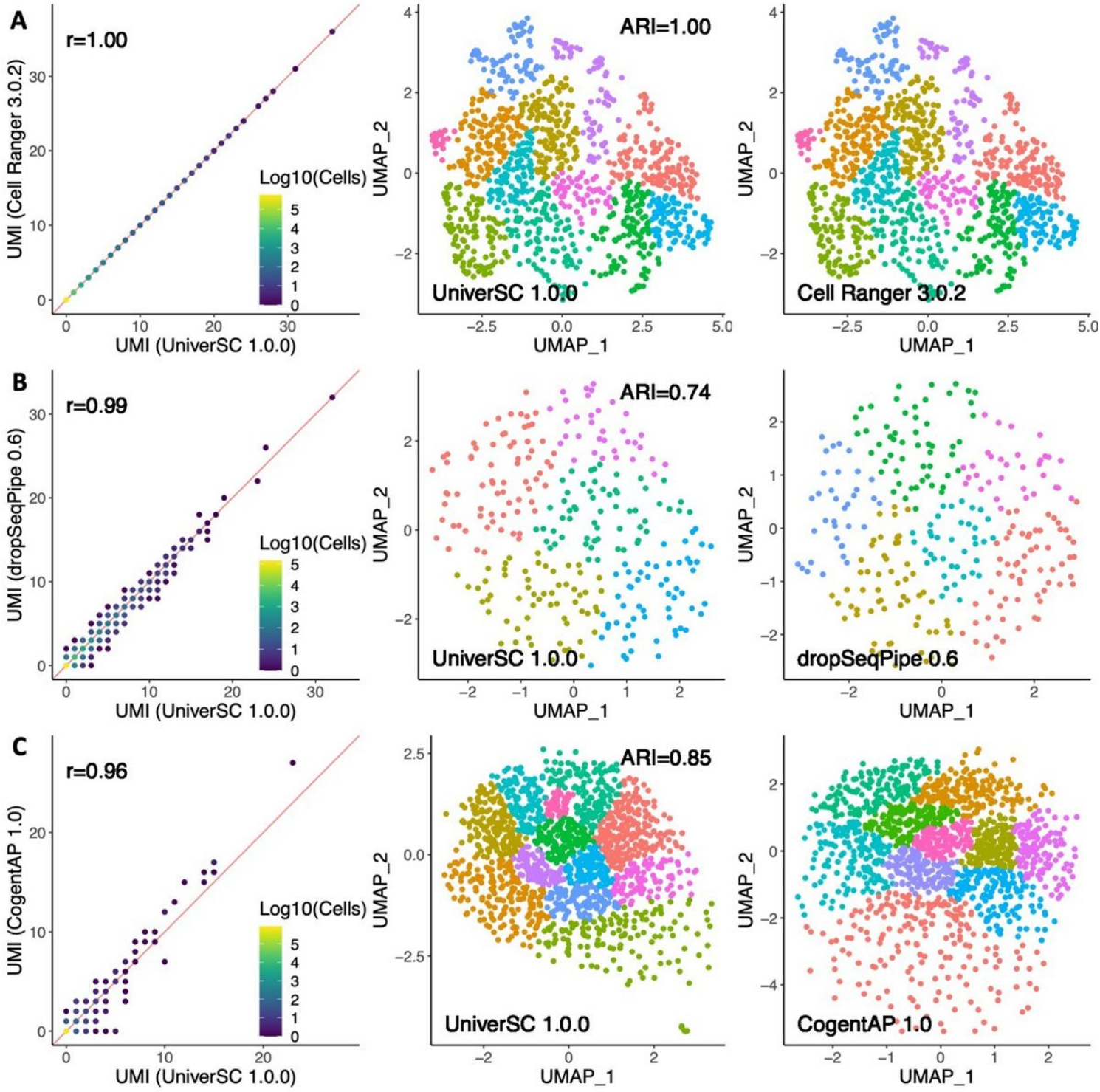

Figure 2

Similarity assessment of UniverSC against oth 289 er pipelines. Comparisons between the GBM generated by UniverSC 1.0.0 against Cell Ranger 3.0.2 (A), dropSeqPipe 0.6 (B), and CogentAP 1.0 (C). Direct comparison of GBMs is on the left column followed by the clusters resulting from UniverSC 1.0.0 (center column) and its counterpart (right column). The processed data compared here is provided in the supplementary data.