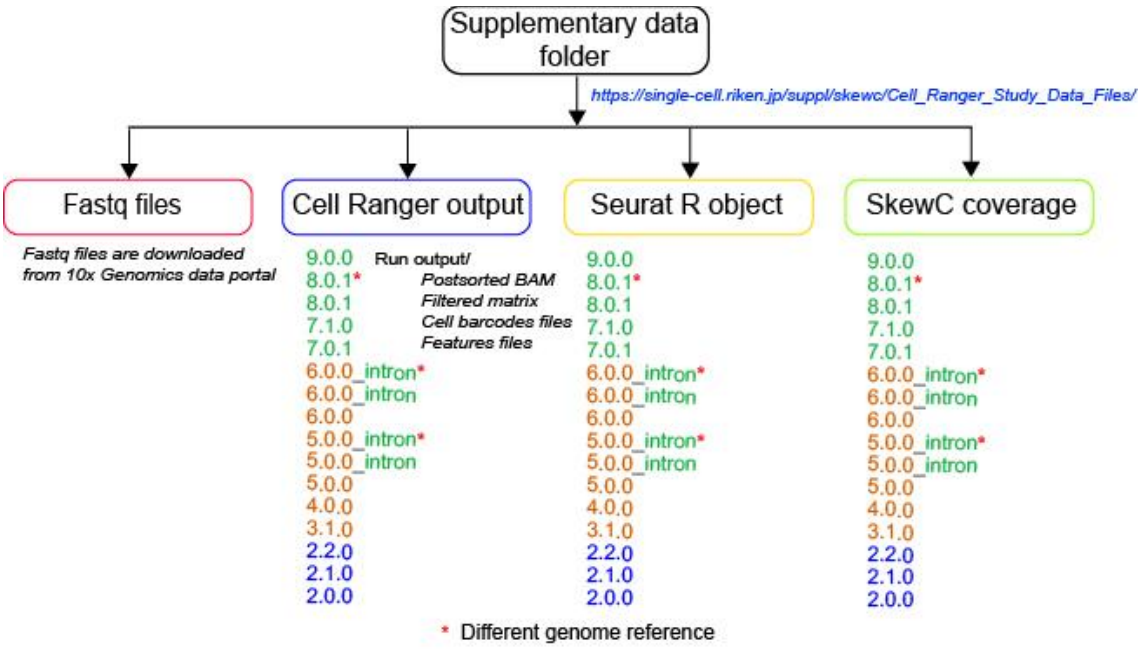
README for supplementary information for the manuscript

Barcode calling contributes to divergence in quality and downstream analysis of chromium single-cell data

Imad Abugessaisa[1,2,3], Akira Hasegawa[1], Scott James Walker[1], Shintaro Katayama[4,5,6], Juha Kere[4,5,6,✉], Takeya Kasukawa[1,✉]

[1]Laboratory for Large-Scale Biomedical Data Technology, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa, 230-0045, Japan.
[2]Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), The University of Osaka, 2-2, Yamadaoka, Suita, Osaka, 565-0871, Japan
[3]Graduate School of Medicine and Faculty of Medicine, The University of Osaka, 2-2, Yamadaoka, Suita, Osaka, 565-0871, Japan
[4]Folkhälsan Research Center, Biomedicum Helsinki, Haartmaninkatu 8, FI-00290 Helsinki, Finland
[5]Department of Biosciences and Nutrition, Karolinska Institutet, NEO, Blickagången 16, SE-141 83 Huddinge, Sweden
[6]Stem Cells and Metabolism Research Program, Biomedicum Helsinki, FI-00290 Helsinki, Finland
[5]Institute for Protein Research, Osaka University, 1-1 Yamadaoka, Suita, Osaka 565-0871, Japan
✉e-mail: juha.kere@ki.se or takeya.kasukawa@riken.jp



The folder contains the structure of online supplementary data. For each dataset used in the manuscript, a single folder was created with the name of the dataset. Under each dataset folder, we created four subfolders as shown in the figure.

1. **The raw sequence data (fastq files)** are either downloaded from 10x Genomics data portal (https://www.10xgenomics.com/datasets) or from GSO NCBI data portal (https://www.ncbi.nlm.nih.gov/geo/) and stored in **Fastq** folder.
2. **The Cell Ranger output** folder contains the output from each version of Cell Ranger pipeline (run-count) for each version a tarball file with the version number was provided (read more about the description of the run-count from (https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/outputs/cr-outputs-overview). The main files in the run-count are the postsorted BAM, filtered matrix, list of cell barcodes in text file, and the features (genes) in text file.
3. The Seurat R object folder contains one R data object per each CR version run processed by Seurat R Bioconductor package(Butler, Hoffman, Smibert, Papalexi, & Satija, 2018). The R object provided as a (.RDS) file and compressed. The RDS file can be read by R readRDS() function to reproduce the results. The R object contains; the raw results of QC and downstream analysis explained in (**Figure 1**).
4. **SkewC coverage** folder contains one (.R) file which provides gene body coverage for each cell barcode in each run of Cell Ranger version. SkewC coverage, was computed by SkewC methods(Abugessaisa, Hasegawa, Katayama, Kere, & Kasukawa, 2023; Abugessaisa et al., 2022).

References

Abugessaisa, I., Hasegawa, A., Katayama, S., Kere, J., & Kasukawa, T. (2023). Computational approach to evaluate scRNA-seq data quality and gene body coverage with SkewC. *STAR Protoc, 4*(1), 102038. doi:10.1016/j.xpro.2022.102038

Abugessaisa, I., Hasegawa, A., Noguchi, S., Cardon, M., Watanabe, K., Takahashi, M., . . . Kasukawa, T. (2022). SkewC: Identifying cells with skewed gene body coverage in single-cell RNA sequencing data. *iScience, 25*(2), 103777. doi:10.1016/j.isci.2022.103777

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol, 36*(5), 411-420. doi:10.1038/nbt.4096