Methods

Single cell sample prep and sequencing Cell preparation

10x Genomics® Cell Preparation Guide describes best practices and general protocols for washing, counting and concentrating cells from both abundant and limited cell suspensions (greater than or less than 100000 total cells, respectively) in preparation for use in 10x Genomics Single Cell Protocols.

We recommend that users refer to the official documentation : 10x genomics single-cellprotocols-cell-preparation-guide

Single-Cell RNA Sequencing（scRNA-seq）

ScRNA-seq was performed using Chromium Single Cell 5' Gel Bead and Library Construction Kit (10× Genomics, #PN-1000165) and Single Cell V(D)J Amplification Kit (10× Genomics, #PN1000252). Reverse transcription, cDNA recovery, cDNA amplification, and library construction were performed according to the manufacturer's protocol. The constructed libraries were sequenced on Illumina NovaSeq at Novogene.

Data analysis

scRNA processing Quality control

We use fastp to perform basic statistics on the quality of the raw reads. Generally, cellranger- count support FASTQ files from raw base call (BCL) files generated by Illumina sequencers as input file. 10x Genomics® not recommend additional processing of the sequence.

If clean reads is indispensable:

Then, raw read sequences produced by the Illumina pipeline in FASTQ format were pre-processed through Trimmomatic software which can be summarized as below:

(1)Remove low-quality reads: scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 10 (SLIDINGWINDOW: 4:10)

(2)Remove trailing low quality or N bases (below quality 3) (TRAILING:3) (3)Remove adapters : there are two modes to remove the adapter sequence:

a. alignment with the adapter sequence, the number of matching bases were greater than 7 and mismatch=2;

b. when read1 and read2 overlapping base scoring greater than 30, removed non-overlapping portions (ILLUMINACLIP: adapter.fa: 2: 30: 7)

(4)Drop reads below the 26 bases long

(5)Discard those reads that can not form paired.The remaining reads that passed all the filtering steps was counted as clean reads and all subsequent analyses were based on this. At last, we use fastp to perform basic statistics on the quality of the clean reads.

Generation and Analysis of Single-Cell Transcriptomes

Raw reads were demultiplexed and mapped to the reference genome by 10X Genomics Cell Ranger pipeline(https://support.10xgenomics.com/single-cell-gene expression/software/pipelines/latest/what-is-cell-ranger) using default parameters. All downstream single-cell analyses were performed using Cell Ranger-ARC and Seurat(Macosko et al., 2015; Satija et al., 2015) unless mentioned specifically. In brief, for each gene and each cell barcode (filtered by CellRanger), unique molecule identifiers were counted to construct digital expression matrices. Secondary filtration by Seurat : A gene with expression in more than 3 cells was considered as expressed, and each cell was required to have at least 200 expressed genes. And filter out some of the foreign cells.

In detail:

cellranger count takes FASTQ files performs alignment, filtering, barcode counting, and UMI counting. It uses the Chromium cellular barcodes to generate feature barcode matrices, determine clusters, and perform gene expression analysis. The count pipeline can take input from multiple sequencing runs on the same GEM well. Add -–nosecondary option to skip secondary analysis of the feature-barcode matrix(dimensionality reduction, clustering and visualization).

When doing large studies involving multiple GEM wells, run cellranger count on FASTQ data from each of the GEM wells individually, and then pool the results using cellranger aggr:

cellranger aggr aggregates outputs from multiple runs of cellranger count, normalizing those runs to the same sequencing depth and then recomputing the feature-barcode matrices and analysis on the combined data. The aggr pipeline can be used to combine data from multiple samples into an experiment-wide feature-barcodematrix and analysis.

Before Secondary Analysis of Gene Expression we use Seurat Second QC it(as above). Subsequent analysis of cellranger reanalyze and Seurat were all performed basing on this output gene expression matrix .

Secondary Analysis of Gene Expression CellRanger

Cellranger reanalyze takes feature-barcode matrices produced by cellranger count or cell ranger aggr and reruns the dimensionality reduction, clustering, and gene expression algorithms using cell ranger default parameter settings.

Seurat

The Seurat package was used to normalise data, dimensionality reduction, clustering, differential expression. we used Seurat alignment method canonical correlation analysis (CCA) [Nat. Biotechnol. 36, 411–420 (2018).] for integrated analysis of datasets. For clustering, highly variable

genes were selected and the principal components based on those genes used to build a graph, which was segmented with a resolution of 0.6.

Global Analysis Between Samples

Based on filtered gene expression matrix by Seurat ,between samples differential expression analysis was carried out using the edgeR package (McCarthy et al., 2012) to obtain zone-specific marker genes.

Enrichment analysis of marker genes

Gene Ontology (GO) enrichment analysis of marker genes was implemented by the clusterProfiler R package, in which gene length bias was corrected. GO terms with corrected Pvalue less than 0.05 were considered significantly enriched by marker gene.

KEGG (Kanehisa M, 2008) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecularlevel information, especially large-scale molecular datasets generated by genome sequencing and other high-through put experimental technologies (http://www.genome.jp/kegg/). We usedclusterProfiler R package to test the statistical enrichment of markergenes in KEGG pathways.

Reactome pathway-based analysis of marker genes was implemented by the ReactomePA R package. REACTOME is an open-source, open access, manually curated and peer-reviewed pathway database(https://reactome.org/)

Protein-Protein Interaction networks(PPI) analysis of marker genes was implemented by the STRINGdb R package

VDJ processing

Assembled and annotated contigs of BCRs and TCRs were generated by the Cellranger ‘‘vdj’’ function by comparing to the reference genome. Only efficient contigs characterized by highconfidence and productive UMIs R 2 were reserved for clonotype constitution. Only cells with at least one heavy chain (IGH for BCR and TRB for TCR) and one light chain (IGL/IGK for BCR and TRA for TCR) were retained for further analysis. In terms of B cells or T cells with two or more assembled contigs, the heavy or light chain with the highest UMI level was regarded as the dominant chain in each cell. Each unique IGH-IGL/IGK or TRB-TRA pair was defined as a clonotype whose expansion degree was indicated by the frequency of B cells or T cells harboring this clonotype. Clonotypes with identical VDJ sequences and rearranged VDJ genes were integrated for different samples belonging to one category. Immunarch scRepertoire software was used for subsequent data analysis （https://immunarch.com/articles/web_only/v11_db.html）.

References

1.Baglama, J. & Reichel, L. Augmented Implicitly Restarted Lanczos Bidiagonalization Methods. SIAM Journal on Scientific Computing 27, 19–42 (2005).

2.Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, (2008).

3.Lun, A., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes. T., Participants in the 1st Human Cell Atlas Jamboree, Marioni, J. Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Biorxiv 2018

4.Robinson, M. D. & Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9, 321–332 (2007). Link to edgeR source.

5.Van der Maaten, L.J.P. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15, 3221-3245 (2014).

6.Yu, D., Huber, W. & Vitek, O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. Bioinformatics29, 1275–1282 (2013).

7.Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. Batch effects in single cell RNAsequencing data are corrected by matching mutual nearest neighbors.Nature biotechnology, 36(5), 421 (2018).

8.Park, J. E., Polanski, K., Meyer, K., & Teichmann, S. A. Fast Batch Alignment of Single Cell Transcriptomes Unifies Multiple Mouse Cell Atlases into an Integrated Landscape. bioRxiv, 397042 (2018).

9.Hie, B. L., Bryson, B., & Berger, B. Panoramic stitching of heterogeneous single-cell transcriptomic data. bioRxiv, 371179 (2018).

10.Young M D, Wakefield M J, Smyth G K, et al. Method Gene ontology analysis for RNA-seq: accounting for selection bias[J]. Genome Biol, 2010, 11: R14.

11.Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1):27-30.(KEGG)

12.McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40: 4288–4297

13.Butler A , Hoffman P , Smibert P , et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species[J]. Nature Biotechnology, 2018.

14.YuG , He Q Y . ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization[J]. Molecular BioSystems, 2016, 12.

15.Franceschini A , Szklarczyk D , Frankild S , et al. STRING V9.1: Protein-Protein Interaction Networks, with Increased Coverage and Integration[J]. Nucleic Acids Research, 2012, 41(D1) 16. Lloyd-Price J, Arze C, Ananthakrishnan AN, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655-662.

17. Danni Cheng , Ke Qiu, Yufang Rao, et al. Proliferative exhausted CD8+ T cells exacerbate longlasting anti-tumor effects in human papillomavirus-positive head and neck squamous cell carcinoma.eLife. 2023; 12: e82705.