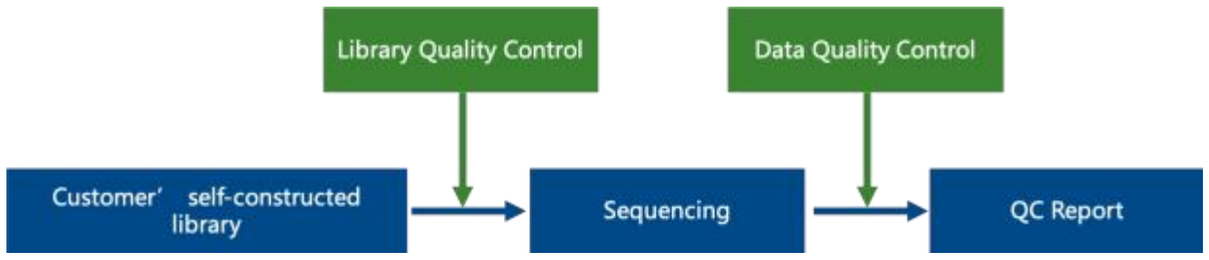
10X Single Cell Standard Analysis Method

A. Library Preparation and Sequencing

From the RNA samples to the final data, each step (including sample testing , library preparation and sequencing) will influence the data quality. The quality of data would have direct impacts on the analysis results . To guarantee the reliability of the data, Quality Control will be performed on each step of the procedure . For customer' self-constructed libraries , the quality detection will be performed before sequencing to ensure producing high quality of data.



1 Library Quality Control

There are mainly three methods in QC for library quality control:

(1) Qubit 2 .0 : tests the library concentration preliminarily.

(2) Agilent 2100 : tests the insert size .

(3) Q- PCR: quantifies the library effective concentration precisely.

2 Sequencing

The qualified libraries are fed into Illumina sequencers after pooling according to its effective concentration and expected data volume .

B. Data Analysis

1. 10x Sequencing data processing and quality control

Once the FASTQ files for each samples is generated , the data analysis begins . Reads from scRNA-seq were processed with Cell Ranger software (10x Genomics) with the default parameters for each sample separately. Briefly, N nucleotides (nt) of Read1s or Read2s were aligned against the reference genome with STAR. Barcodes and UMIs were filtered and corrected . PCR duplicates were marked using the barcode , UMI and gene ID. Only confidently mapped , non- PCR duplicates with valid barcodes and UMIs were used to generate a gene-barcode matrix for further analysis .

The filtered gene expression matrices containing only cellular barcodes were generated in data processing step . Then the Seurat R package was futher utilized to achieve quality control.

Alignment to Reference Genome

Cell Ranger uses an aligner called STAR, which peforms splicing-aware alignment of reads to the genome . It then uses the transcript annotation GTF to bucket the reads into exonic, intronic, and intergenic, and by whether the reads align (confidently) to the genome . A read is exonic if at least 50% of it intersects an exon , intronic if it is non-exonic and intersects an intron , and intergenic otherwise .

The exonic reads were further aligned to the existing transcriptome with annotation as well. A read that is compatible with the exons of an annotated transcript, and aligned to the same strand , is considered mapped to the transcriptome . Among mapped reads , these uniquely mapped reads are the only ones considered for UMI counting .

Cell Calling and UMI Counting

Before counting UMIs , Cell Ranger attempts to correct for sequencing errors in the UMI sequences . The UMI of the less supported read group is corrected to the UMI with higher support. Cell Ranger again groups the reads by barcode , UMI (possibly corrected) , and gene annotation . If two or more groups of reads have the same barcode and UMI, but different gene annotations , the gene annotation with the most supporting reads is kept for UMI counting , and the other read groups are discarded .

The cell-calling algorithm based on the EmptyDrops method in Cell Ranger determines cell-associated barcodes based on their UMI count or by their RNA profiles . The algorithm has two key steps:

1 . It uses a cutoff based on total UMI counts of each barcode to identify cells . This step identifies the primary mode of high RNA content cells .

2 . Then the algorithm uses the RNA profile of each remaining barcode to determine if it is an “ empty" or a cell containing partition . This second step captures low RNA content cells whose total UMI counts may be similar to empty GEMs .

Quality Control

Low-quality libraries in 10x scRNA-seq data can arise from a variety of sources such as cell damage during dissociation or failure in library preparation . These usually manifest as ‘cells’ with low total counts , few expressed genes and high mitochondrial read proportions .

The "filtered_feature_bc_matrix" generated by Cellranger were read into the Seurat R package . For each cell, quality control metrics such as the total number of counts and the proportion of counts in mitochondrial genes were calculated .

Cells that met any one of the following criteria were filtered out for downstream processing in each sample: < 200 genes with present in each cell, genes with non-zero counts in at most 3 cells , > 5,000 feature count (potential multiplets) , the proportion of the feature count attributable to mitochondrial genes was greater than 50%, or the proportion of the feature count attributable to Heparin-Binding genes was greater than 5%.

Doubletfinder is used to filter doublets. Then, singlets are using for further analysis.

2. Identification of Highly Variable Genes (HVGs)

Feature selection removes the uninformative genes and identifies the most relevant features to reduce the number of dimensions used in downstream analysis . The highly variable genes (HVGs) methods rely on the assumption that the genes with highly variable expression across cells are resulted from biological effects rather than technical noise .

After cell filtering , 10x gene expression matrices from each sample were loaded into Seurat. The expressions of each gene were normalized by total counts for that cell, multiplied by a scale factor (the median UMI counts for all cells within the sample) , and natural-log transformed using log1p . Then the expressions of each gene were scaled .

Next, highly variable genes were identified using FindVariableGenes function with the highest standardized variance selected by selection .method = 'vst' . The top 3000 most variable genes of each sample selected by Seurat were used to compute the PCs .

Normalization

The Cell Ranger reported UMI count value for each gene i in each cell j was divided by the sum of the total UMI counts in cell j to normalize for differences in library size , and then multiplied by M, the median UMI counts for all cells within sample , resulting in Counts -per-median ($CPM_{i,j}$) values . $E_{i,j}$ was then calculated as log(CPMi,j + 1) .

Expression values for $E_{i,j}$ for gene i in cell j were calculated following:

$$\text{The counts -per-median}(CPM_{i,j}) = Count_{i,j} / Totalcount_j * M$$

$$\text{Expression values } E_{i,j} = \log_{10}(CPM_{i,j} + 1)$$

3. Cell Subpopulation Identification

A key goal of 10x scRNA-seq data analysis is to identify cell subpopulations (different populations are often distinct cell types) within a certain condition or tissue to unravel the heterogeneity of cells . To identify a gene expression signature associated with this sample or group of cells , HVGs previously determined are used as input for dimensionality reduction via principal component analysis (PCA) . The resulting PCs were then used as input for clustering analysis .

P rincipal components analysis (PCA) discovers axes in high-dimensional space that capture the largest amount of variation . The PCA on the log-normalized expression values is performed using runPCA function with setting total 50 PCs .

To partition the data into clusters of transcriptionally related cells , a shared nearest neighbor (SNN) modularity optimization based **clustering algorithm** is used to identify clusters of cells . In this process , the top PCs retained as input for clustering , generally ranging from 5 to 20 , are determined by elbow point.

For visualization purposes , dimensionality was further reduced to 2D using **t - distributed stochastic neighbor embedding (t - SNE) and uniform manifold approximation and projection (UMAP)** . Both of them are try to find a low-dimensional representation that preserves relationships between neighbors in high-dimensional space . Compared to t-SNE, the UMAP visualization tends to have more compact visual clusters with more empty space between them.

Novogene Co ., Ltd .

4 . Marker Gene Detection

Identification of marker genes is usually based around the retrospective detection of differential expression between clusters . These marker genes allow us to assign biological meaning to each cluster based on their functional annotation . In the most obvious case , the marker genes for each cluster are a priori associated with particular cell types , allowing for cluster- ing to serve as a proxy for cell-type identity. Significantly differential expressed genes for each cluster were identified using the Wilcox test with the thredhold qvalue <0 .05 and log2foldchange >0 .25 .

5. Enrichment Analysis and Annotation of Marker Genes

G ene O ntology (GO) is a standardized classification system widely used for gene function , which supplies a set of controlled vocabulary to describe the property of genes and gene products comprehensively. There are 3 ontologies in GO system: molecular function , cellular component and biological process . The basic unit of GO is GO-term, each of which belongs to one type of ontology. This method firstly maps all source genes to GO terms in the database (http://www.geneontology.org/) , calculating gene numbers for each term, then using Wallenius non-central hyper-geometric distribution to find significantly enriched GO terms in source genes comparing to the reference genes background .

The interactions of multiple genes may be involved in certain biological functions . **KEGG (Kyoto E ncyclopedia of G enes and G enomes)** is a collection of manually curated databases dealing with genomes , biological pathways , diseases , drugs , and chemical substances . KEGG is utilized for bioinformatics research and education , including data analysis in genomics , metagenomics , metabolomics and other omics studies . Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways associated with differentially expressed marker genes compared with the whole genome background .

The **Reactome** (http://www.reactome .org) is a database of reactions , pathways and biological processes , which can be used to browse pathways and submit data to a suite of data analysis tools , containing curated annotations that cover a diverse set of topics in molecular and cellular biology. Reactome terms with padj < 0 .05 are significant enrichment.

TFCat is a curated catalog of mouse and human **transcription factors (TF)** based on a reliable core collection of annotations obtained by experts' review of the scientific literature . Annotated genes are assigned to a functional category and confidence level. We use the differentially expressed marker genes in each cluster to search the TFCat, then provides the annotation of the TF and the corresponding reference (PubMed ID) .

The **protein- protein interaction network** is constructed for differentially expressed genes in each cluster by using STRING protein interaction database (http://string-db .org (http://string-db.org)).Protein-protein interaction is provided as network file which can be imported into Cytoscape software and visualized and edited . The central organizing metaphor of Cytoscape is a network graph , with molecular species represented as nodes and intermolecular interactions represented as links , that is , edges , between nodes .

Reference

1 . Z heng , G. X. Y., Terry, J. M., Belgrader , P., Ryvkin , P., Bent, Z .W. , & Wilson , R., et al. (2017) . Massively parallel digital transcriptional profiling of single cells . Nature Communications , 8 , 14049 .

2 . Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, III WMM, Hao Y, Stoeckius M, Smibert P, Satija R. (2019) . Comprehensive Integration of Single-Cell Data. Cell, 177 , 1888-1902 .

3 . Ayyaz , A., Kumar , S., Sangiorgi, B. et al. (2019) . Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. Nature , 569 , 121– 125 .

4 . Peng , Y. R., Shekhar , K., Yan , W., Herrmann , D., Sappington , A., & Bryman , G. S., et al. (2019) . Molecular classification and comparative taxonomics of foveal and peripheral cells in primate retina. Cell, 176 , 1222-1237 .e22 .

5 . Kim, D. W., Yao , Z ., Graybuck, L. T., Kim, T. K., & Anderson , D. J,. (2019) . Multimodal analysis of cell types in a hypothalamic node controlling social behavior . Cell, 179(3) , 713-728 .e17 .

6 . Amezquita, R. A., Lun , A.T.L., Becht, E. et al. (2020) . Orchestrating single-cell analysis with Bioconductor . Nature Methods , 17 , 137– 145 .

7 . Geng , Chen , Baitang , Ning , Tieliu , & Shi. (2019) . Single-cell rna-seq technologies and related computational data analysis . Frontiers in Genetics , 10 , 317 .

8 . Dobin , A., Davis , C. A., Schlesinger , F., et al. (2013) . STAR: ultrafast universal RNA-seq aligner . Bioinformatics , 29(1): 15-21 .

9 . Becht, E., McInnes , L., Healy, J. et al. (2019) . Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnol 37 , 38–44 .

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

11 . Butler , A., Hoffman , P., Smibert, P., Papalexi, E., & Satija, R. (2018) . Integrating single-cell transcriptomic data across different conditions , technologies , and species . Nature Biotechnology, 36(5) , 411–420 .

12 . Yu , G., Wang , L. G., Han , Y., & He , Q.Y. . (2012) . Clusterprofiler: an r package for comparing biological themes among gene clusters . Omics A Journal of Integrative Biology, 16(5) , 284 -287 .

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

Yu , G., & He , Q. Y. . (2016) . Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization . Molecular BioSystems , 12(2) , 477-479 .

14 . Schlesner , Matthias , Eils , Roland , Gu , & Z uguang . (2016) . Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics , 32 , 2847– 2849 .