# Shaoxia Documentation

**Table of Content**

[Shaoxia Documentation 1](#_Toc160631383)

[Features 2](#_Toc160631384)

[Demo and Contact 2](#_Toc160631385)

[Installation 2](#_Toc160631386)

[Use Shanxia Platform 6](#_Toc160631387)

[Parameter Setting Guidelines 14](#_Toc160631388)

[The design of MySQL database 16](#_Toc160631389)

Shaoxia is an open-source cluster-based analysis plateform for single cell RNA-seq data. It is crafted from scratch, with love. In recent years, with advance of technology and reduction in costs, Single-cell RNA sequencing is increasingly accessible to researchers of many fields. scRNA-Seq enables researchers to profile the transcriptomes of individual cells. Combined with clear scRNA-seq analysis framework, Shaoxia aims to provide user an easy-to-use, web-based, interactive GUI while harnessing modern compute system to accelerate big data ananlysis of scRNA-seq.

### Features

with some considerations in design, Shaoxia platform have following features:

1. Clear, intuitive framework, covering all aspects of scRNA-seq data analysis.
2. Easy-to-use, interactive graphical user interface (GUI).
3. Disk-free design, Shaoxia does not store the intermediate files produced by the analysis, but stores the parameters.
4. Releases the power of modern compute system (such as HPC) to accelerate the analysis of scRNA-seq data (using SLURM).
5. Interactive analysis for cell quality control and cell type annotation.
6. Support three types of input data: FASTQ files, cellranger results, cell-gene count matrix.

### Demo and Contact

The Shaoxia platform’s demo: http://www.shaoxia.cloud. This website is publicly accessible for all people, but it has limit computing resource. if you are interested in trying to use Shaoxia, please contact me (weiweideng0106@outlook.com) to apply for user registration.

### Installation

**Note:** system terminal prompt is $, R terminal prompt is >.

Prerequisite (on Unix-like systems):

1. **Python**. Install miniconda with python 3.9.12, and then:
2. Using Conda to create Shaoxia environment:

$ *conda create --name shaoxia python=3.9.12*

$ *conda activate shaoxia*

$ *conda install aiofiles aiohttp pymysql pyrserve aiomysql pyyaml*

1. Using Conda to create InferCNV environment:

$ *conda create --name inferCNV*

$ *conda activate inferCNV*

$ *conda install -c conda-forge jags*

1. Using Conda to create cellphoneDB environment:

$ *conda create -n cpdb python=3.8*

$ *conda activate cpdb*

$ *pip install cellphonedb*

1. Using Conda to create SCENIC environment:

$ *conda create -y -n scenic python=3.10*

$ *conda activate scenic*

$ *pip install pyscenic*

$ *pip install numpy==1.23.5*

$ *pip install pandas==1.5.3*

$ *pip install numba==0.56.4*

1. Using Conda to create scvelo environment:

$ *conda create --name scvelo*

$ *conda activate scvelo*

$ *pip install -U scvelo*

1. **R**. Install R version 4.2.0., and then:
2. Install Rserve package:

$ *sudo R CMD INSTALL Rserve\_1.8-11.tar.gz*

1. Install Seurat package:

> *install.packages('Seurat')*

1. Install InferCNV package:

> *install.packages("rjags")*

> *if (!requireNamespace("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

> *BiocManager::install("infercnv")*

1. Install SeuratWrappers package:

> *install.packages('remotes')*

> *remotes::install\_github('satijalab/seurat-wrappers')*

1. Install SeuratDisk package:

> *if (!requireNamespace("remotes", quietly = TRUE)) {*

*install.packages("remotes")*

*}remotes::install\_github("mojaveazure/seurat-disk")*

1. Install ggplot2 package:

> *install.packages("ggplot2")*

1. Install ComplexHeatmap package:

> *if (!require("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

> *BiocManager::install("ComplexHeatmap")*

1. Install dplyr package

> *install.packages("dplyr")*

1. Install clusterProfiler package:

> *if (!require("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

> *BiocManager::install("clusterProfiler")*

1. Install reactome.db package:

> *if (!require("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

> *BiocManager::install("reactome.db")*

1. Install org.Hs.eg.db package:

> *if (!require("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

> *BiocManager::install("org.Hs.eg.db")*

1. Install org.Mm.eg.db package

> *if (!require("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

> *BiocManager::install("org.Mm.eg.db")*

1. Install tidyverse package:

> *install.packages("tidyverse")*

1. Install GSVA package:

> *if (!require("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

> *BiocManager::install("GSVA")*

1. Install pheatmap package:

> *install.packages("pheatmap")*

1. Install monocle3 package:

> *BiocManager::install(c('BiocGenerics', 'DelayedArray', 'DelayedMatrixStats',*

*'limma', 'lme4', 'S4Vectors', 'SingleCellExperiment',*

*'SummarizedExperiment', 'batchelor', 'HDF5Array',*

*'terra', 'ggrastr'))*

> *install.packages("devtools")*

> *devtools::install\_github('cole-trapnell-lab/monocle3')*

1. Install nichenetr package:

> *devtools::install\_github("saeyslab/nichenetr")*

1. Install stringr package

> *install.packages("stringr")*

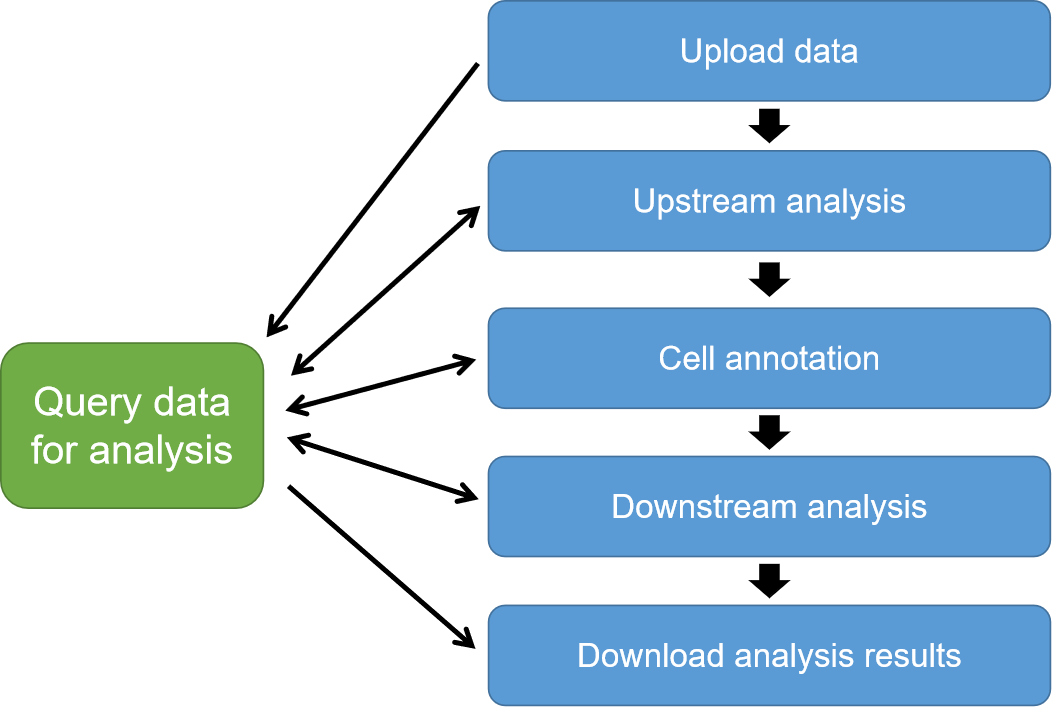
1. **MySQL** installation, please follow the instructions from <https://dev.mysql.com/doc/mysql-installation-excerpt/8.0/en/binary-installation.html> to install MySQL.
2. **SLURM** installation, please follow the instructions from https://slurm.schedmd.com/quickstart\_admin.html to install SLURM.
3. **NGINX** installation, please follow the instructions from <https://www.nginx.com/resources/wiki/start/topics/tutorials/install/> to install nginx, its configuration file can be download from github.
4. **Supervisor** installation, please follow the instructions from <http://supervisord.org/installing.html> to install on your system. its configuration file can be download from github.
5. **NPM** installation, please follow the instructions from <https://nodejs.org/en/download/package-manager> to install it on your system.

Setup Shaoxia platform (On Unix-like systems’ shell terminal):

1. Create some dir: $ *mkdir shaoxia && cd shaoxia && mkdir data python\_backend slurm\_job\_logs slurm\_job\_scripts tmpImg*
2. Download Shaoxia backend code files from github, and move them to *python\_backend* dir
3. Change configuration in *config.yaml* file.
4. Download Shaoxia frontend code files, and run cmds: $ *npm run install && npm run build*, to generates *dist* diretory, and move the diretory to shaoxia diretory.
5. Activate shaoxia environment and run $ ./db.py to initialize mysql database.
6. Run$ *sudo systemctl start nginx.service* to setup nginx.
7. Run$ *sudo systemctl start supervisor.service* to setup shaoxia platform.
8. Visit http:192.168.1.110 (the URL depends on you nginx configuration and local network environment) to use Shaoxia platform.

### Use Shanxia Platform

Shaoxia’s workflow is very simple:

Figure 1. workflow of Shaoxia platform.

1. **Upload Data**

**NOTE:** Only one sequencing sample’s data can be uploaded at a time.

Shaoxia platform support three types of input data for analysis:

1. Fastq files, that generated by 10X Genomics scRNA-seq protocol. Just move mouse to Upload data -> 10X fastq button and click, select all fastq files of one sample to upload.
2. Cellranger results, that generated by 10X Genomics’ cellranger software. Just move mouse to Upload data -> 10X cellranger results button and click, select result files of one sample to upload.
3. Count matrix, that can be generated by any scRNA-seq protocol. Just move mouse to Upload data -> cell-gene matrix button and click, select the matrix file of one sample to upload. The matrix file is a tab-seperate txt file, which has cell barcode in first row and gene name on first column. A sample file can be downlaod on github.
4. **Upstream analysis**

As mentioned on figure one, uptream analysis contains serval parts:

手机屏幕的截图

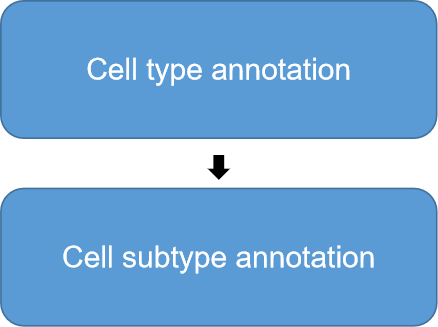
描述已自动生成

Figure 2. Upstream analysis.

**NOTE:** After upload data to the server, you can click “Query data for analysis” button to select uploaded data for analysis.

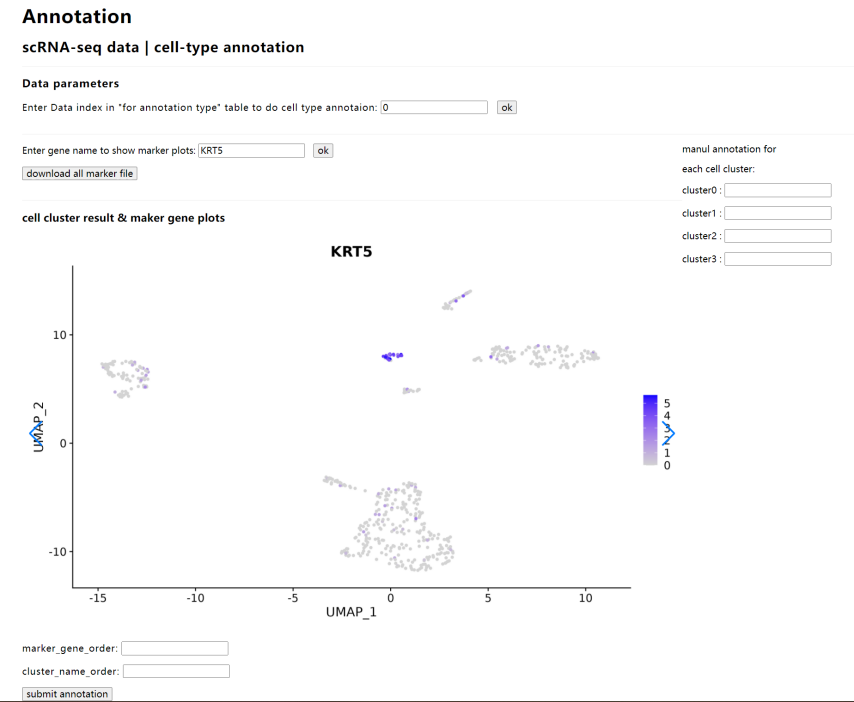
1. Reads mapping, only works for fastq upload data, click “Query data for analysis -> for reads mapping” to get the index of fastq upload data in the upload table, and then enter some cellranger parameters in “Upstream analysis -> reads mapping” page to carry out reads mapping.
2. Generate loom, only works for reads mapping results, and if you want to get RNA vilocity downstream result, you should run this step. Click “Query data for analysis -> for generate loom” to get the index of reads mapping results in the table, and then enter the index on “Upstream analysis -> generate loom” to carry out generate loom.
3. Quality control, works for all types of input data (including: cellranger result upload data, count matrix upload data, reads mapping results, generate loom results), click “Query data for analysis -> for quality control” to get the index of input data in the table, and then go to “Upstream analysis -> quality control” page, select appropriate parameters to view quality control result images, and at last upload the parameters.
4. Merge/integrate, only supports for quality control results of reads mapping and cellranger results upload data currently. If you want to combine multiple sample’s data, Shaoxia provied two type of methods -- merge and integrate, both of them come form Seurat R package. Click “Query data for analysis -> for merge/integrate” to get the indexes of input data in the table, which you want to combine them together, and then go to “Upstream analysis -> multi-sample integration” or “Upstream analysis -> multi-sample merge” to merge/integrate serveral sample.
5. Dimension reduction & cell clustering, works for quality control and merge/integrate results, Click “Query data for analysis -> for dc” to get the index of input data in the table, and then go to “Upstream analysis -> dimension reduction & cell clustering” page, enter appropriate parameters to view cell clustering result images, and at last upload the parameters.
6. **Cell annotation**

Cell annotation has two steps:

Figure 3. cell annotation.

1. Cell annotation, works for results of dimension reduction & cell clustering, Click “Query data for analysis -> for annotation type” to get the index of input data in the table, and then enter the index to start cell type annotation (Fig. 4).

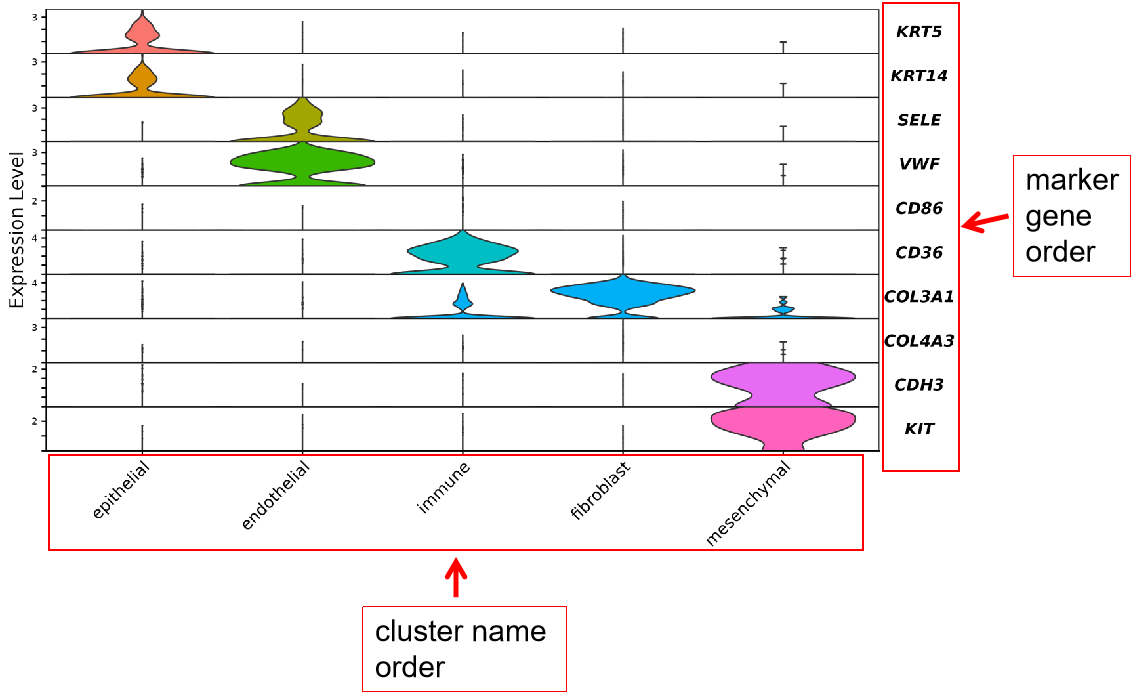
Click “download all marker file” button to download marker files; enter marker gene and click “ok” button to show cell cluster result and marker gene plots; enter cell cluter annotation on right side; enter marker gene order and cluster name order (sperate by ‘:’) on the bottom; click “submit annotation” to complete cell type annotation.

Figure 4. cell type annotation page.

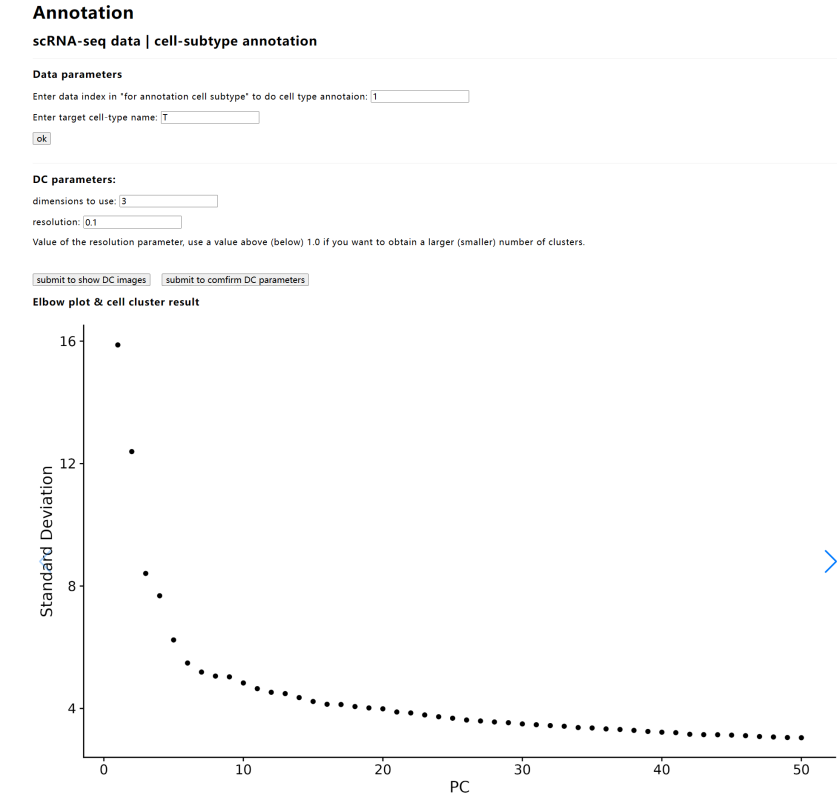
**NOTE:** marker gene order and cluster name order are used for generating annotation results (Fig. 5)

Example of cluster name order -- epithelial:endothelial:immune:fibroblast:mesenchymal

Example of marker gene order -- KRT5:KRT14:SELE:VWF:CD86:CD36:COL3A1:COL4A3:CDH3:KIT

Figure 5. marker gene violin plot.

1. Cell subtype annotation, only work for cell type annotation results, and this step is optional. For example, if you want to identify cell subtypes among immune cells, such as B cell or T cell, you can do cell subtype annotation. Similar to cell type annotation, click “Query data for analysis -> for annotation subtype” to get the index of input data in the table, and then enter the index to start cell type annotation. Cell subtype annotation has two steps -- DC (dimension reduction & cell clustering) (Fig. 6) and annotation (Fig. 7).

Figure. 6. DC part of cell subtype annotation.

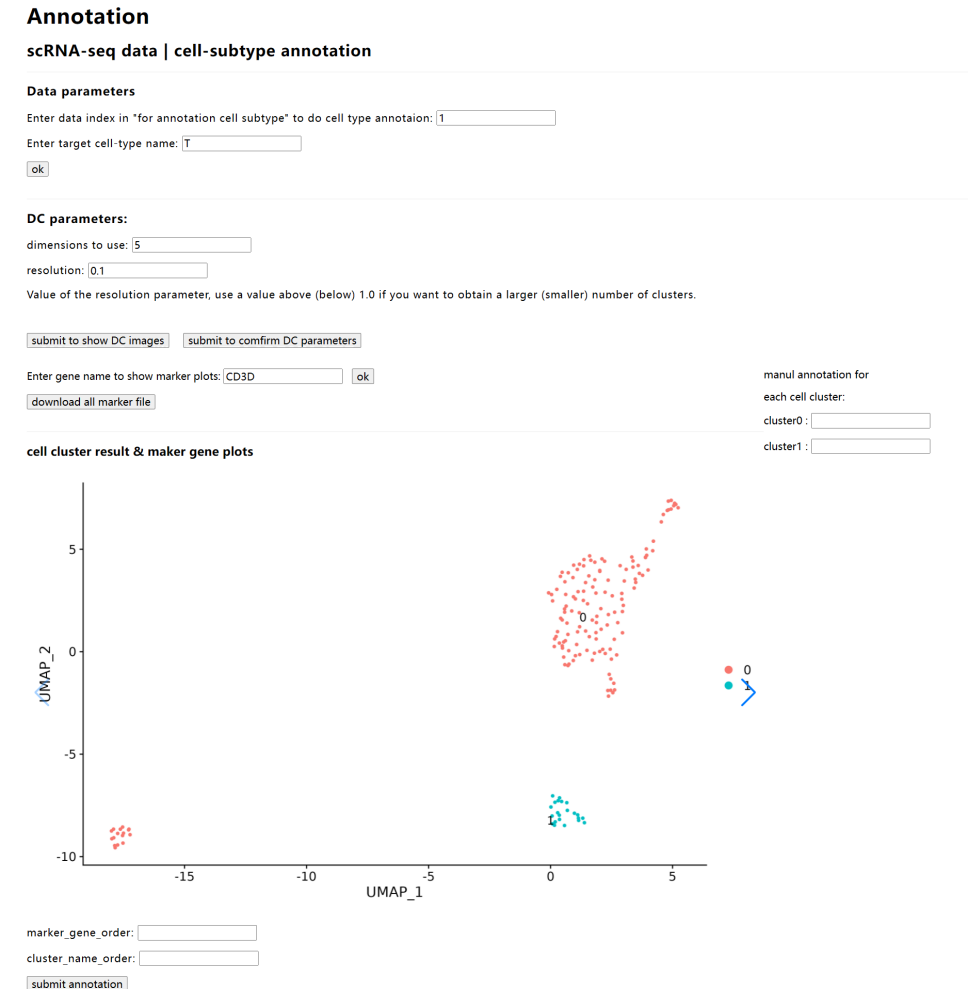


Figure 7. cell subtype annotation.

1. **Downstream analysis**

Currently, Shaoxia provide 11 type of downstream analysis . Downstream analysis, works for cell type annotation and subtype annotation results. Click “Query data for analysis -> for downstream analysis” to get the index of input data in the table, and then click “Downstream analysis” to select one type of analysis, enter data index (and some parameters) to do it (Fig. 8).

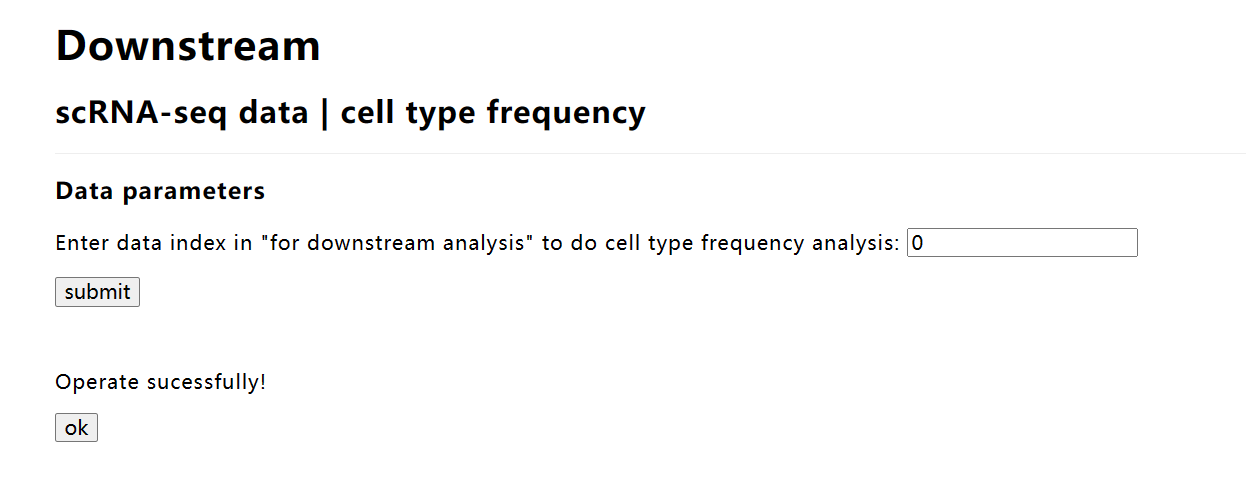


Figure 8. one example of downstream analysis: cell type frequency analysis.

In downstream analysis, Shaoxia uses SLURM to schedule analysis jobs, thus if you want to cancel/suspend/resume analysis jobs, click “Management -> manage analysis jobs” and click “Query data for analysis -> for slurm job management” to get job id to manage jobs, and you also can view status of analysis jobs on “Query data for analysis -> for slurm job management” page.

1. **Download analysis results**

If any downstream analysis jobs have completed, you can download results of upstream, annotation and downstream analysis of that data. click “Query data for analysis -> for download results” to get annotation id in the “Downstream analysis results” table, and then click “Download result” button, select annotation type and enter annotation id to download analysis results (Fig. 9).

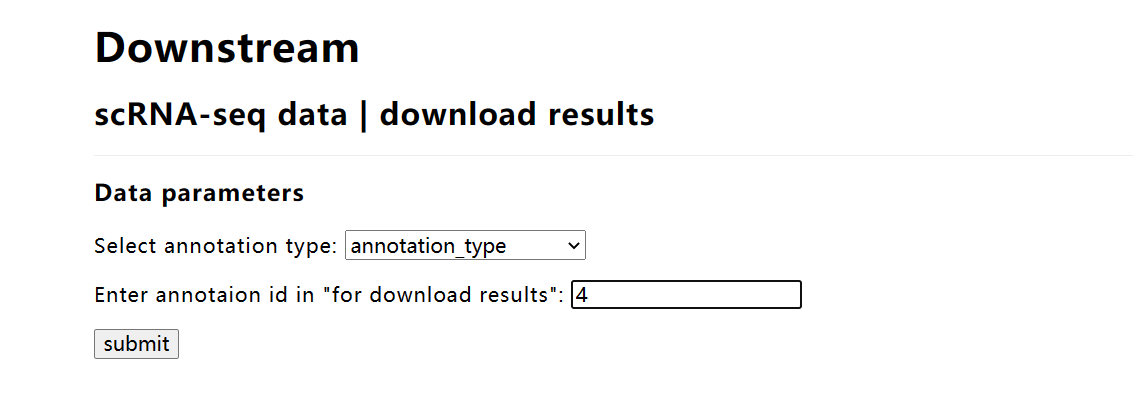


Figure. 9. Download results page.

1. **Management**

Shaoxia provide two types of management, 1) job management, user can stop, suspend and resume his (or her) own analysis jobs, 2) user management, administrator user can add regular users to the platform.

1. job management, “Click Query data for analysis -> for slurm job management” to get slurm job id and click “Management -> manage analysis jobs”, and then select a command and enter the job id, click “submit” to complete the operation (Fig. 10).

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 10. job management.

2. add user, click “Management -> add user”, and then enter the user information, click “submit” to complete the operation (Fig. 11 ).

图形用户界面, 应用程序

描述已自动生成

Figure 11. add user.

1. **Login and Logout**

1 Login, just click any button of analysis step to login, for example, click “Query data for analysis”

2 Logout, just click “Logout” button to logout.

### Parameter Setting Guidelines

In addition to established best practices for scRNA-seq data analysis, we would like to share some insights into Shaoxia platform parameter settings based on our experience. We encourage you to explore these suggestions and experiment to find the optimal configuration for your specific datasets. We also welcome any feedback or advice you may have. Please feel free to reach out to us via email at weiweideng0106@gmail.com.

**1) Upstream analysis**

1. Quality control parameters. The parameters and its descriptions can see in the following table, user can check quality control images (before and after quality control) to adjust parameters.

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| lower limiting number of genes that detected in a cell | nFeature\_RNA in Seurat, extremely low number of detected genes could indicate loss-of-RNA. For example: set to 200 indicate cells that expressed less than 200 genes are removed. |
| upper limiting number of genes that detected in a cell | nFeature\_RNA in Seurat, extremely high number of detected genes could indicate doublets. For example: set to 6000 indicate cells that expressed more than 6000 genes are removed. |
| minimum number of reads that need detected in a cell | nCount\_RNA in Seurat, low number of reads could indicate bad data. For example: set to 2000 indicate cells that contained less than 2000 reads are removed. |
| maximum percentage of mitochondrial gene reads in a cell | (percent.mt) High percentage of mitochondrial gene reads could indicate dying cells. For example: set to 5 indicate cells which percentage of ofribosomal reads is more than 5% are removed. |
| minimum percentage of low proportion ofribosomal gene reads in a cell | (percent.rb) Extremely low proportion of ofribosomal reads could indicate loss-of-RNA. For example: set to 20 indicates cells which percentage of ofribosomal reads is less than 20% are removed. |

2. Dimension reduction & clutering. The parameters and its descriptions can see in the following table, user can check elbow plot and cell clustering results to adjust parameters. For example, set dimensions to 8 and resolution to 0.1 when you want to do cell type annotation (get small number of cell clusters), set dimensions to 15 and resolution to 0.5 when you want to do cell subtype annotation (get larger number of cell clusters).

|  |  |
| --- | --- |
| **Parameters** | **Descriptions** |
| dimensions to use | dims in Seurat FindNeighbors and RunUMAP functions, number of principle components (dimensions) to use. |
| resolution | resolution in Seurat FindClusters function, larger value will generate larger number of cells clusters. |

**2) Cell type and subtype annotation.**

To assist with cell type or subtype identification, users can leverage resources like the CellMarker database (http://xteam.xbio.top/CellMarker/) and relevant scientific literatures to research specific gene markers. These resources can inform manual annotation within the platform.

**3) Downstream analysis**

For 11 downstream analyzes, default parameters are used.

### The design of MySQL database

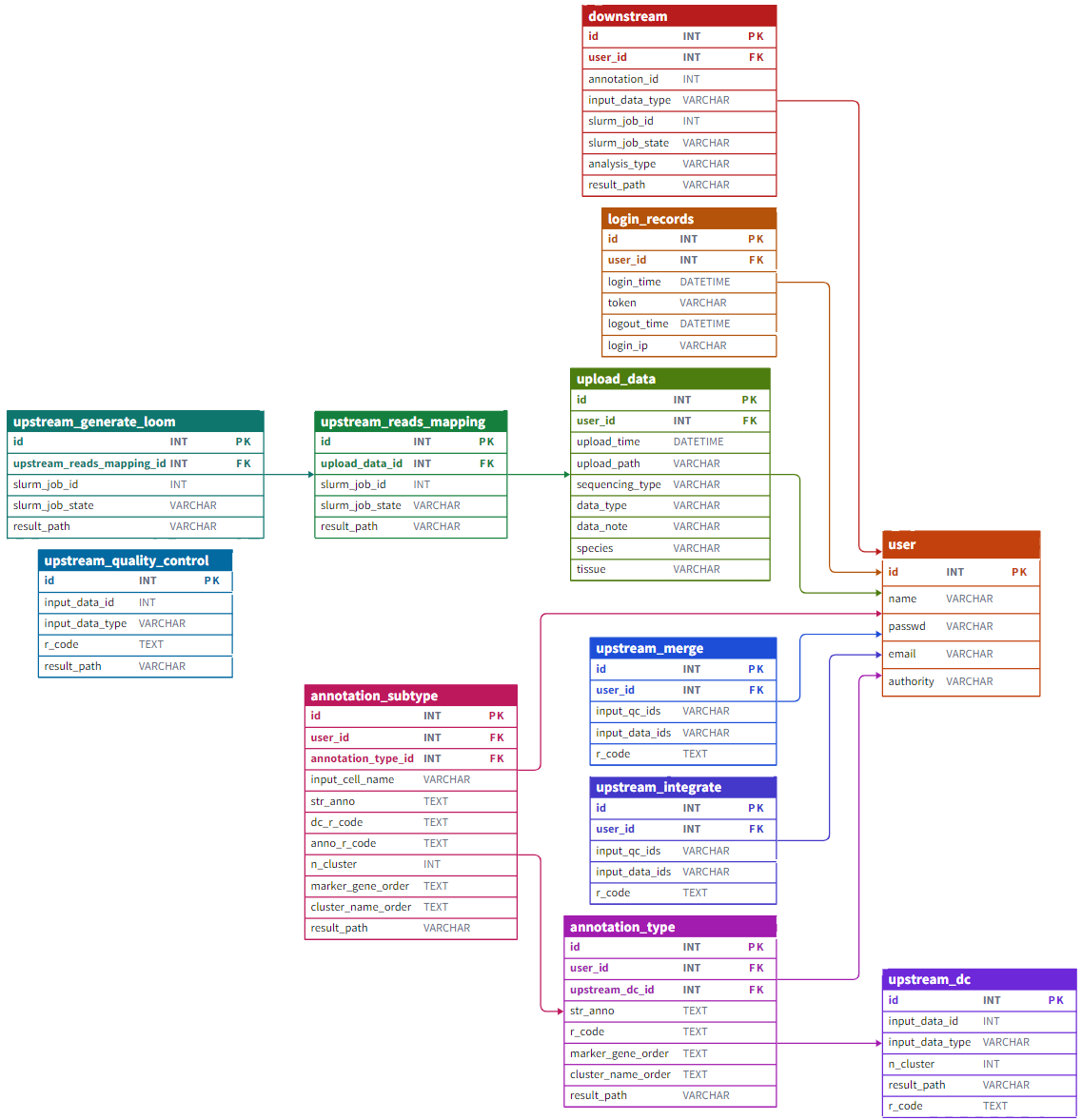


Figure 11. the schema of MySQL database.