# 数据分析平台设计概述

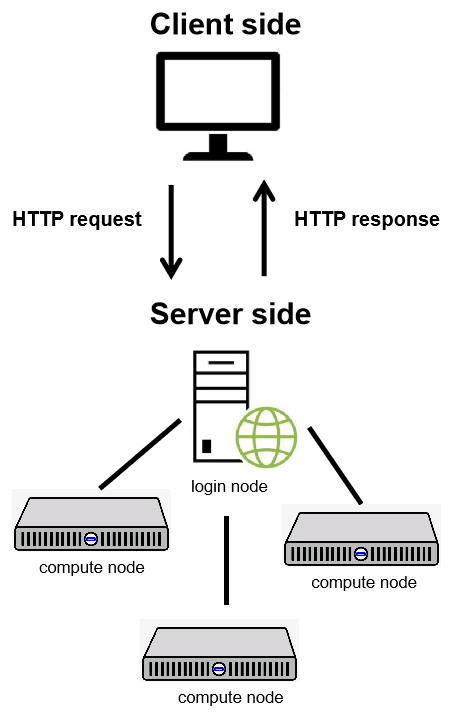
本平台是一个单细胞RNA测序数据分析平台，根据我提出的单细胞RNA测序数据分析框架， 它以网站的形式为用户提供可交互式的数据分析服务，它包含了以下四个部分：

1. 前端界面设计，使用vue 3前端框架，加速开发，界面简约实用。
2. 后端服务设计，使用python的 aiohttp等框架开发，支持异步操作，服务能力较强。
3. 单细胞RNA测序数据分析框架，我根据自身多年的单细胞分析经验，提出简单明了的分析框架，具有较高的普适性。
4. 数据库设计，根据数据分析框架，使用MySQL特别设计了数据库，高效地存储数据分析产生的结果。

本平台具有以下特点：

1. 明确、简单的单细胞分析框架，覆盖了单细胞RNA数据分析的所有方面。
2. 方便使用的、可交互性的图形化界面。
3. 硬盘友好型数据库设计，数据库只保存分析所用的参数以及最后分析结果图片，不保存分析产生的中间文件。
4. 通过使用SLURM，利用高性能计算（HPC）系统加速单细胞分析，也使得平台具有较强的伸缩性。
5. 支持三种格式的输入数据：FASTQ文件、cellranger 软件结果文件、tab分隔的reads计数矩阵的文本文件。

本平台的整体设计为：

图一、少侠单细胞RNA测序数据分析平台的整体架构

本平台的开发语言以及及软件平台：

表一、开发信息

|  |  |  |
| --- | --- | --- |
| 项目 | 名称 | 说明 |
| 操作系统 | Linux系统 | 开源 |
| 网络反向代理软件 | Nginx | 开源 |
| 开发语言 | Python, R, Typescript, Shell | 开源 |
| 开发工具 | 微软Visual Studio Code | 开源 |

本文将侧重于介绍单细胞RNA测序数据分析流程、界面（结合功能）、数据库的设计，而忽略后端服务，因为它包含了许多算法实现，超出了本文的范围。

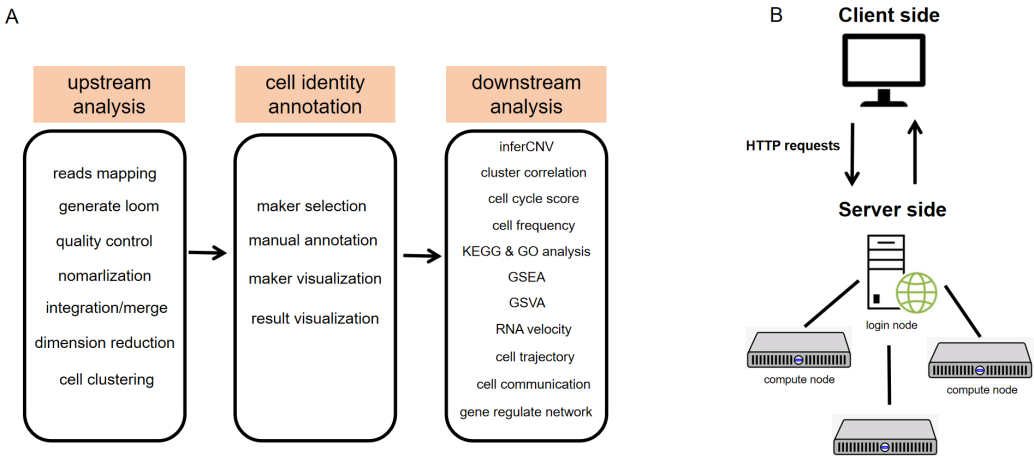
# 单细胞RNA测序数据分析流程

细胞身份注释是单细胞分析中最为关键以及最为基础的工作，如果没有细胞身份标注， 基因差异等分析就没有生物学意义，甚至可能基于错误的细胞注释得出错误的结论。考虑到这一点，我将scRNA-seq分析分为三个阶段: 1)上游分析 (upstream analysis)，2)细胞注释 (cell annotation)，3)下游分析 (downstream analysis)。通过细胞身份注释将上下游分析分开，直观清晰。

上游分析包含的步骤为：1) 测序reads 回贴到基因组 (reads mapping)；2) 生成loom文件 (generate-loom)；3) 细胞质控 (quality control)；4)不同样本数据的融合或者整合 (merge/integrate)；5) 数据降维以及细胞聚类 (dimension reduction & cell clustering)。

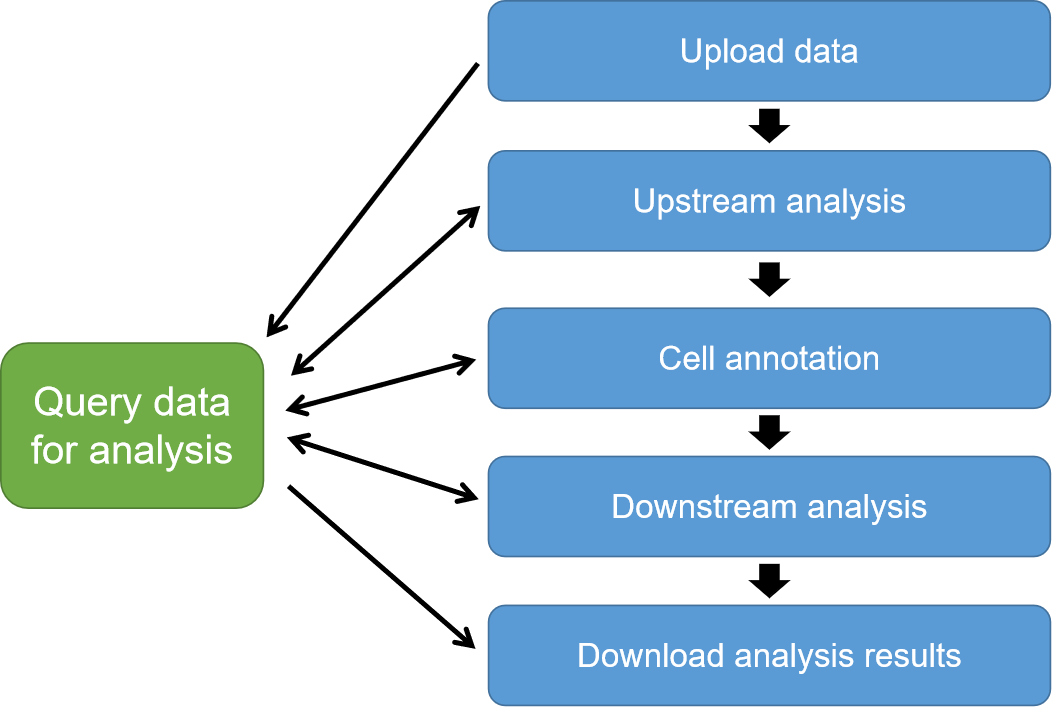
细胞身份注释包含两个步骤：1) 细胞主要类群注释，注释主要细胞类型，如B细胞；2) 细胞亚型注释，如注释B细胞的亚型——Naive B细胞等。

下游分析覆盖了11个分析：拷贝数变异分析 (inferCNV)、 细胞类群相关性分析(cell cluster correlation)、细胞类型组份分析(cell type frequency)、细胞周期评分 (cell cycle analysis)、KEGG & GO 富集分析、GSVA分析、GSEA分析、拟时分析、RNA速率分析、细胞通讯分析、基因调控网络分析。

图二、单细胞RNA测序数据分析流程

# 分析平台功能介绍

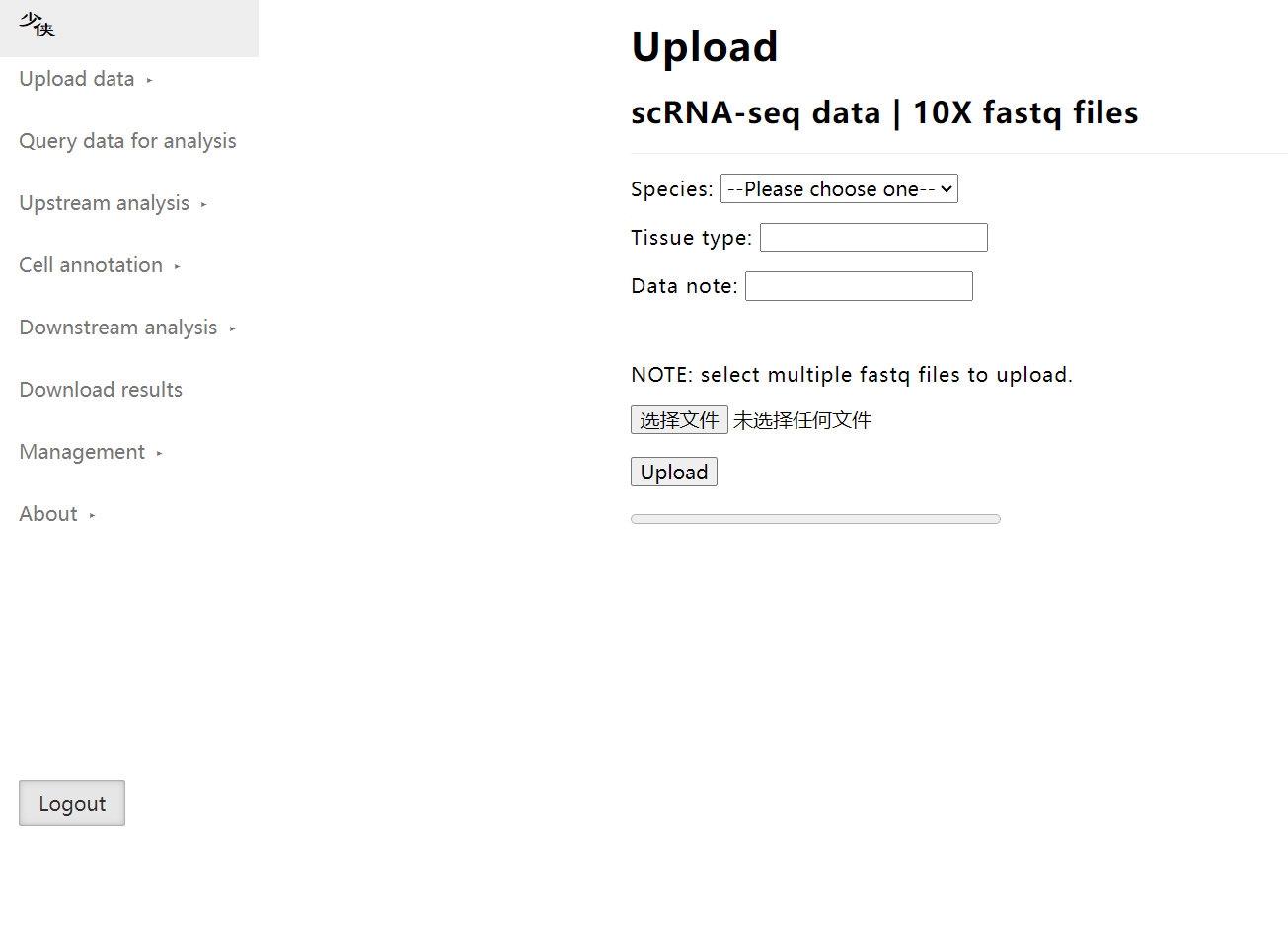
本部分将结合软件的用户界面来介绍其功能，用户界面以及功能与数据分析流程相互对应（如图三所示），查询数据页面 (Query data for analysis) 将数据上传(upload data)、上游分析、细胞身份注释、下游分析，分析结果下载各个功能联系起来 (先查询数据，而后对数据进行操作)。用户必须登录才能使用各个功能模块。

图三、少侠分析平台工作流程

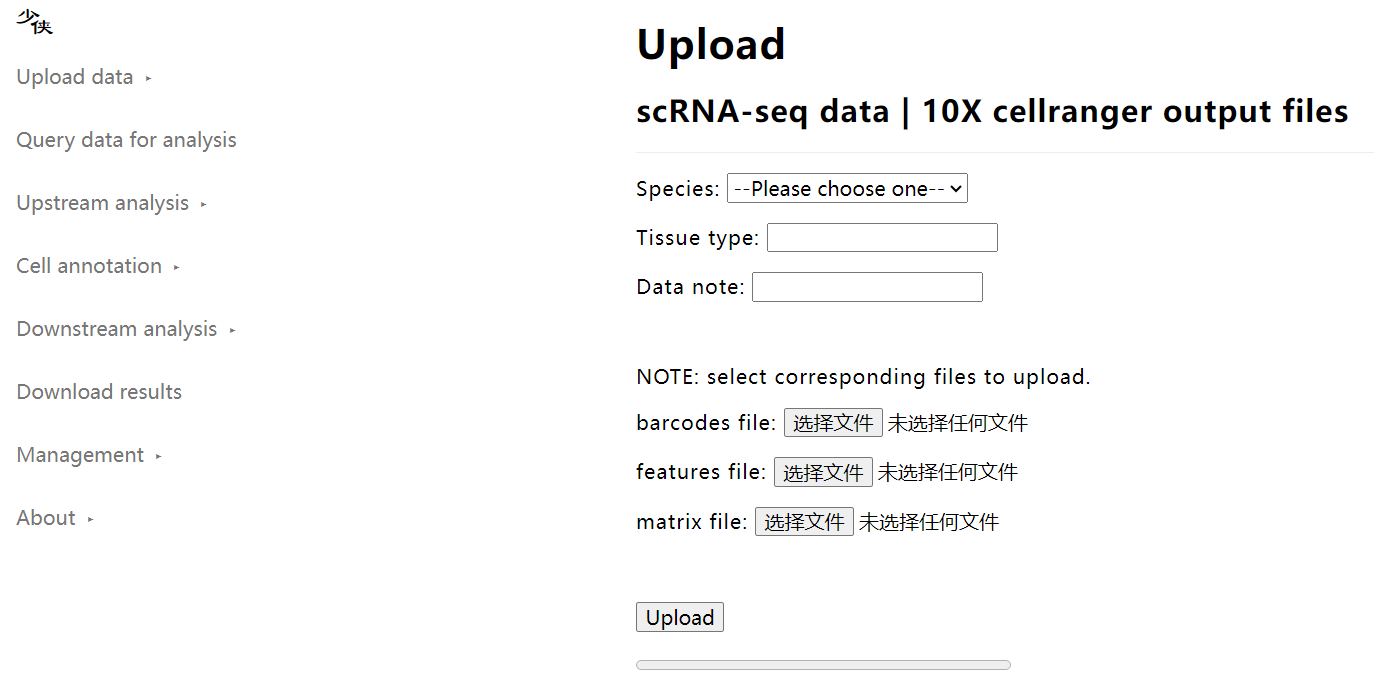
## 数据上传

本平台支持三种类型的数据上传作为分析的输入，但是用户每次上传只能上传一个样本的测序数据：

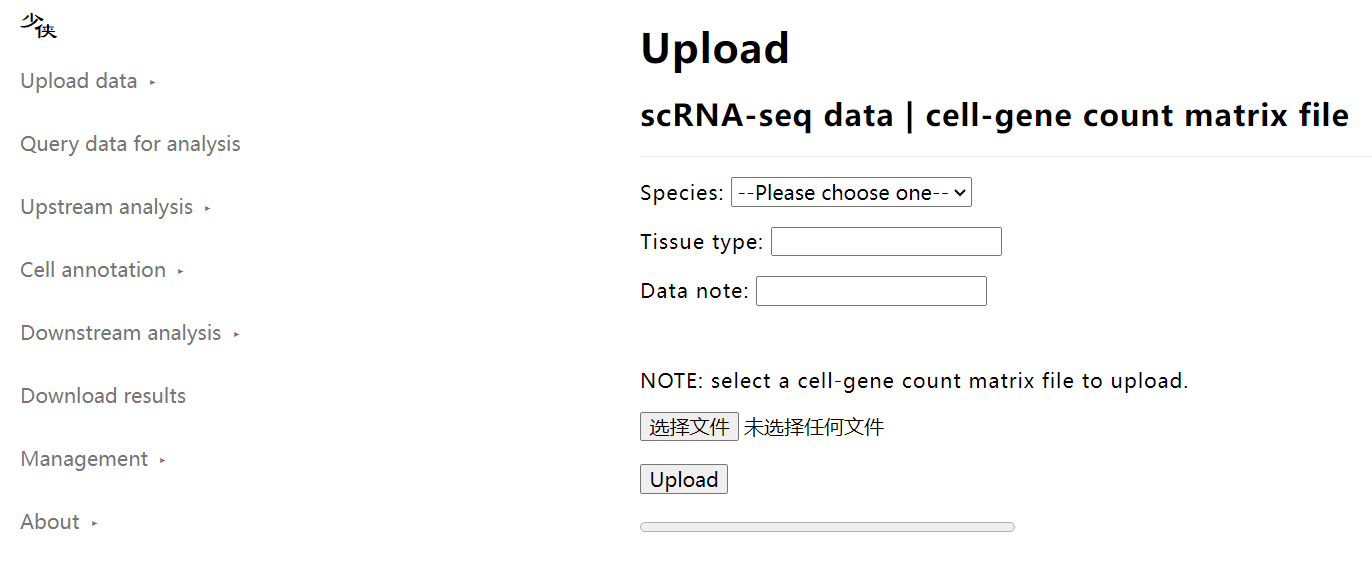
* 1. FASTQ 文件上传：鼠标点击 “Upload data -> 10X fastq” 按钮进入上传页面，在对应输入框内输入样本信息，然后点击“选择文件”选择多个FASTQ测序结果文件，最后点击上传。

图四、上传FASTQ文件页面

* 1. 10X genomics 公司cellranger 软件的结果文件上传：依次点击 “Upload data -> 10X cellranger results” 按钮进入上传页面，输入相应样本信息，选择相应结果文件，最后点击上传。

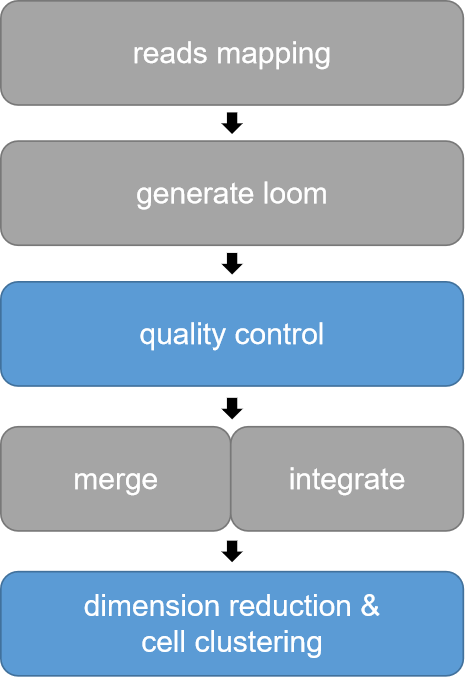
图五、上传cellranger结果文件的页面

* 1. 制表符分隔的reads计数的矩阵文本文件上传：该矩阵文件的第一行为细胞表形码 (cell barcode)，标识一个细胞，第一列为基因名，表示每一个基因；而每一个单元里的数字表示某个基因在该细胞中的表达量。点击 “Upload data -> 10X cellranger results”，进入页面，输入相应信息，选择单个矩阵文件，最后点击上传。

图六、上传矩阵文件页面

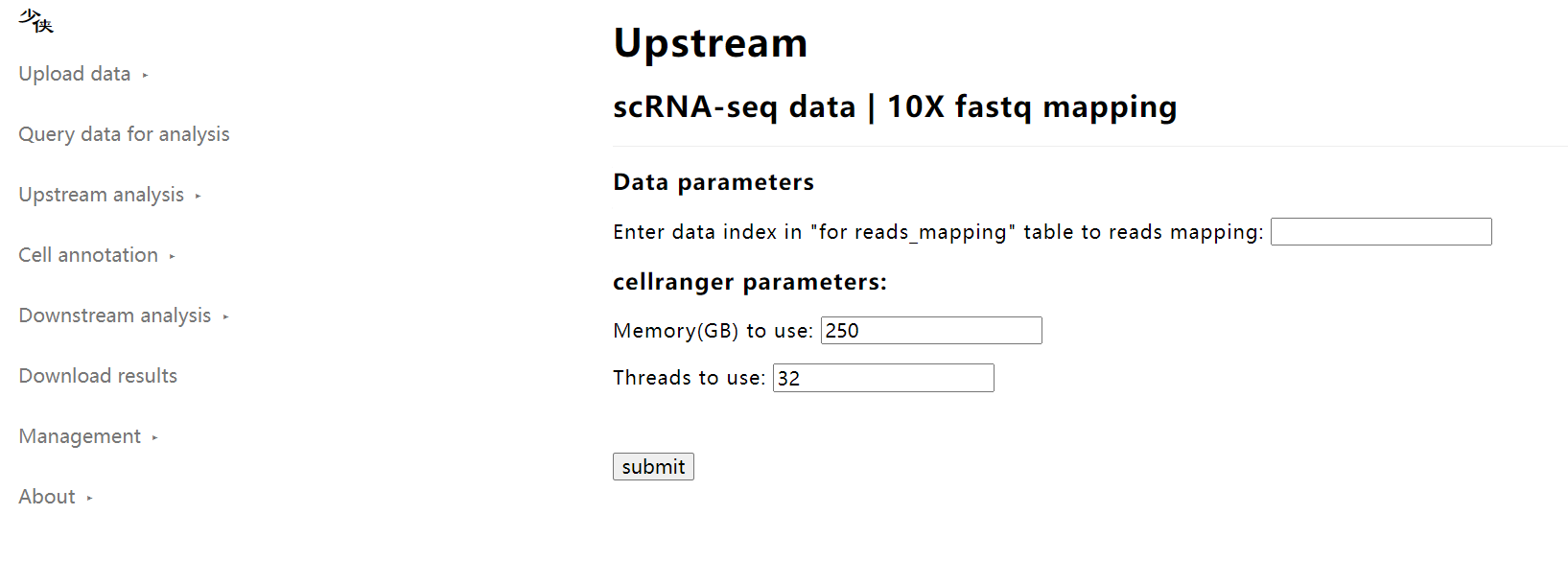
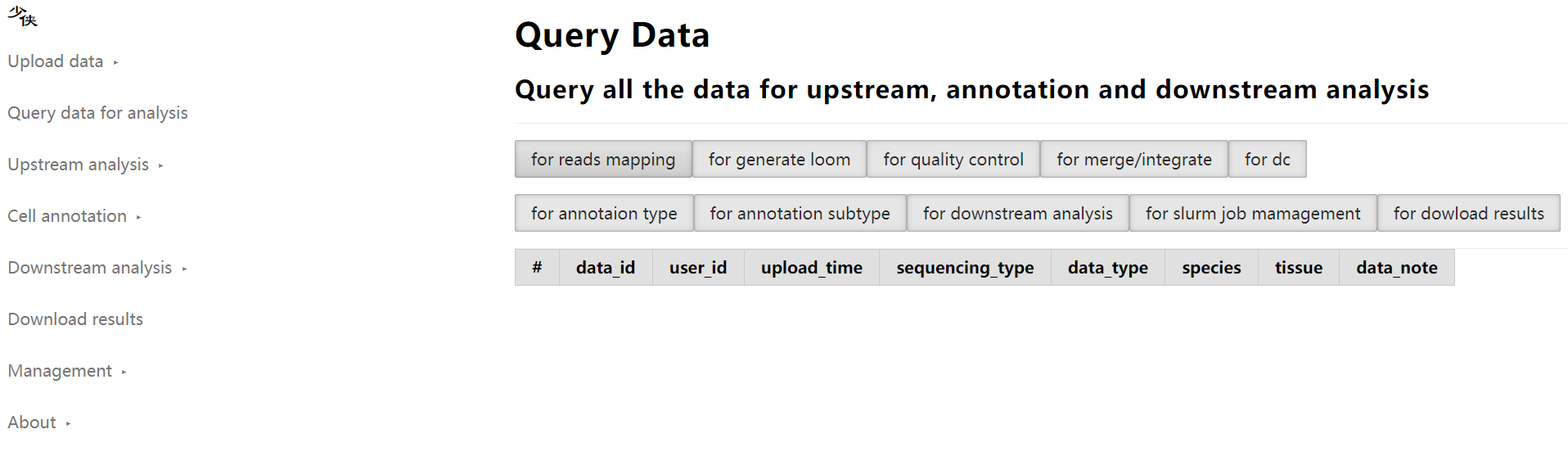
## 上游分析

上游分析包含的步骤如图七所示。图中灰色的部分是可选的步骤，蓝色的部分是必须的步骤。

图七、上游分析

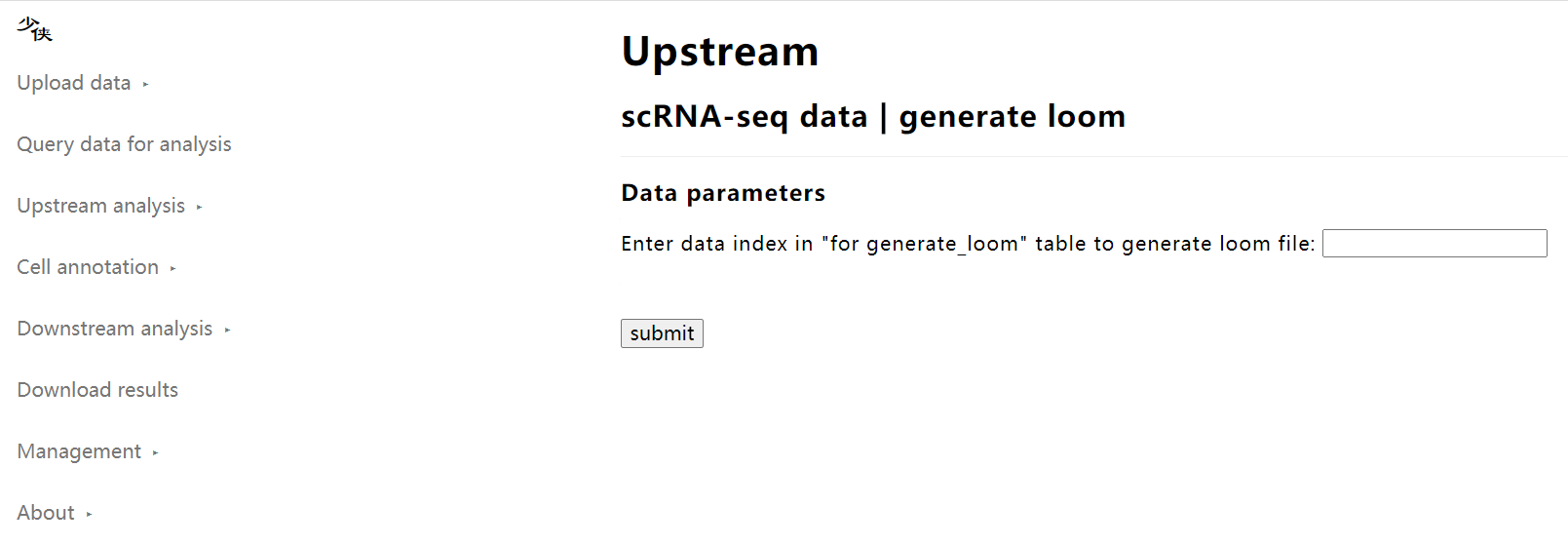
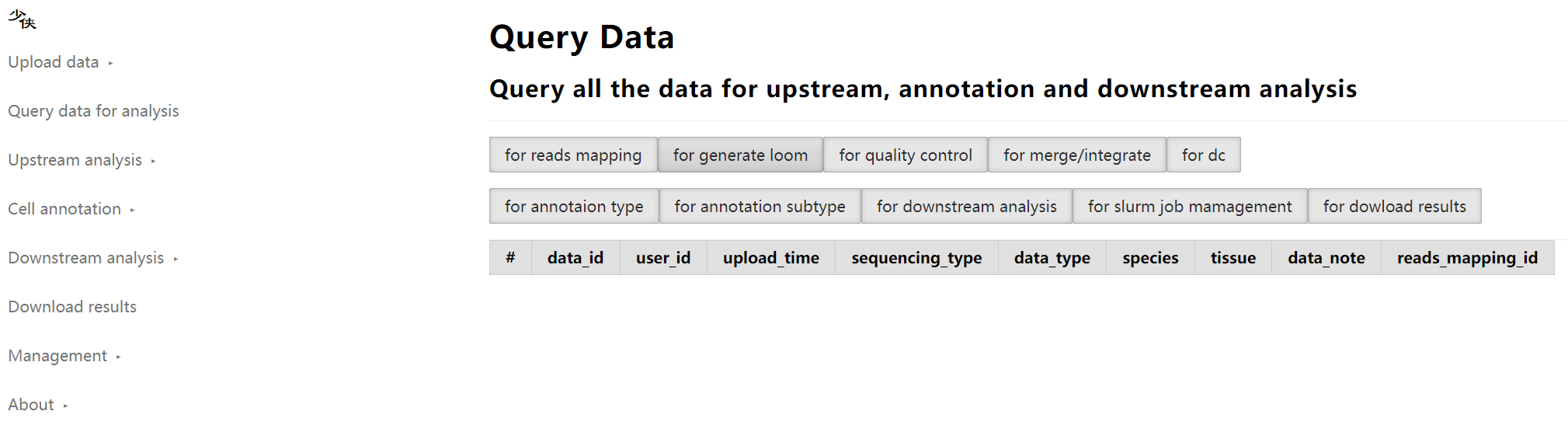
2.1) 测序reads 回贴到基因组：上传数据类型为FASTQ文件的可以进行该步骤，依次点击“Query data for analysis -> for reads mapping”，从表中得到上传的FASTQ数据的序号，然后依次点击 “Upstream analysis -> reads mapping” 进入分析页面，输入数据序号以及参数，最后点击上传，完成分析。

图八、查询上传的FASTQ数据并进行reads mapping 分析

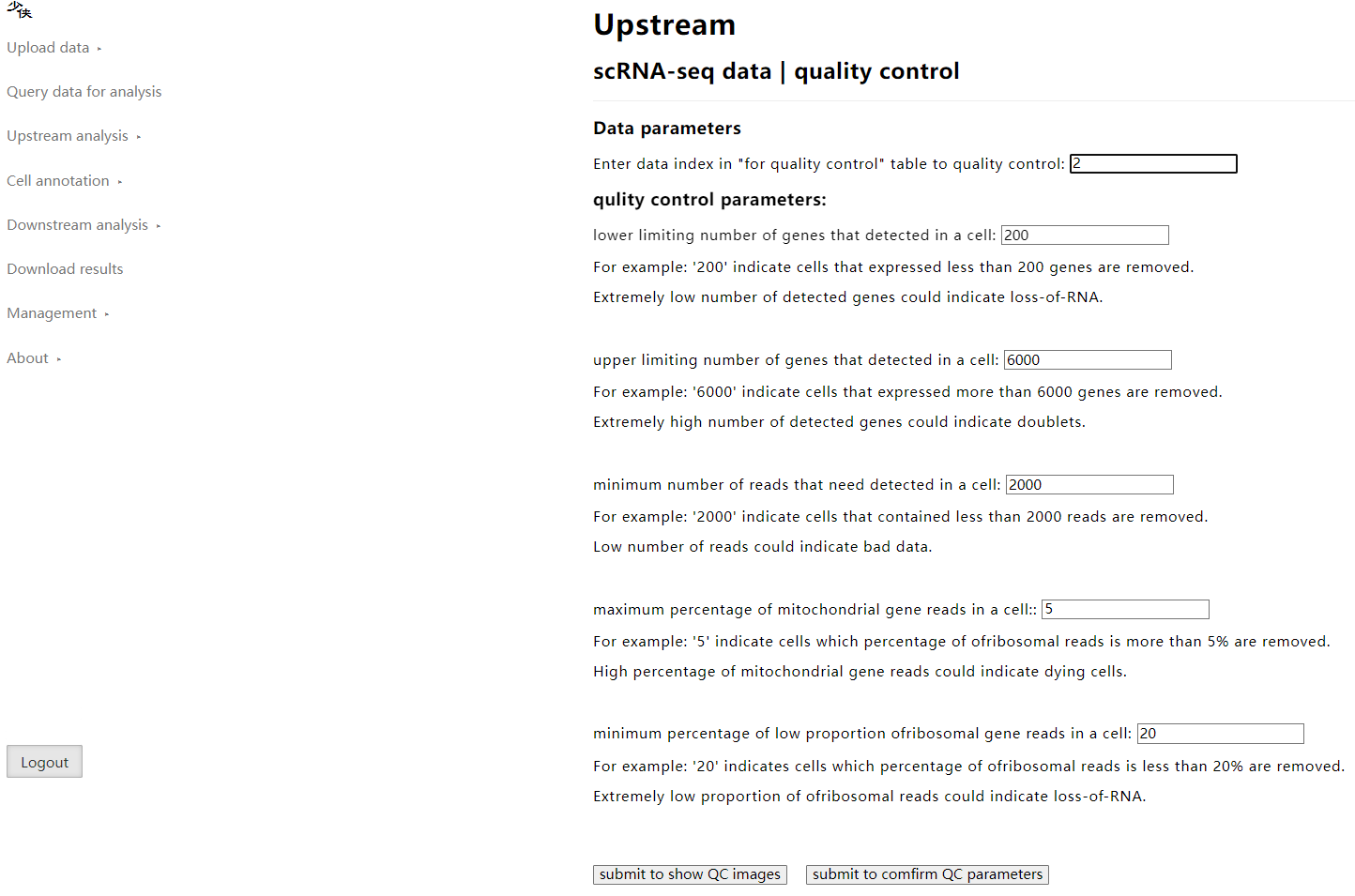
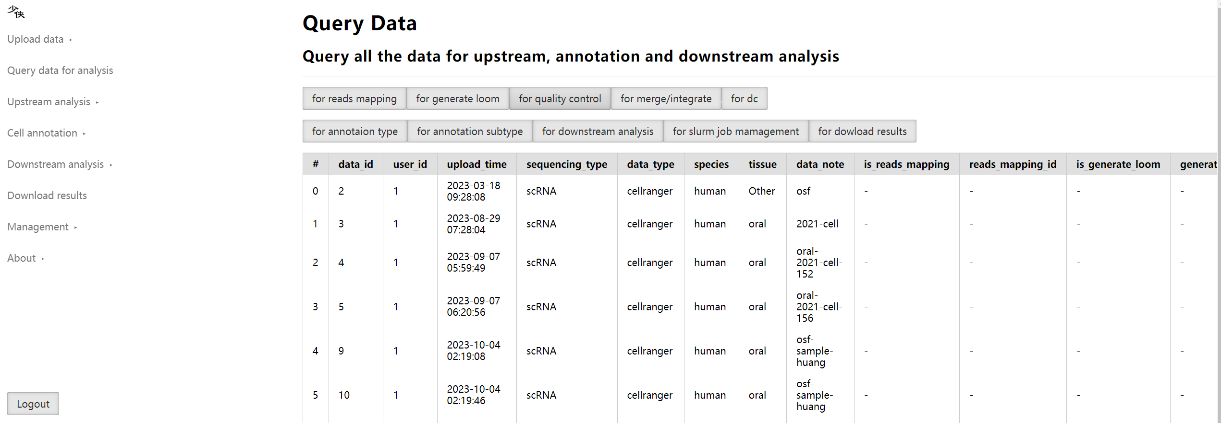


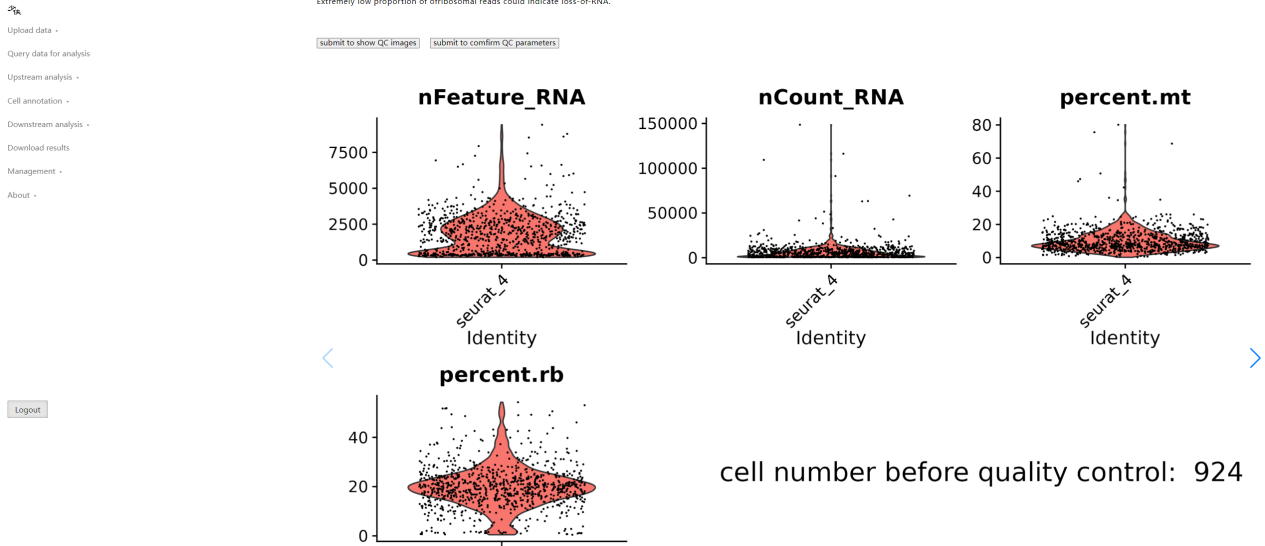
2.2) 生成 loom文件：该步骤为下游分析RNA速率分析打基础，如果不做该步骤则不能进行RNA速率分析，反之，则不做。依次点击“Query data for analysis -> for generate loom”，从表中得到数据的序号，然后依次点击 “Upstream analysis -> generate loom” 进入分析页面，输入数据序号，最后点击上传，开始分析。

图九、查询上传的FASTQ数据并进行generate loom 分析



2.3) 细胞质控：该分析是实时交互式的，即用户可以尝试不同的质控参数，查看不同质控参数产生的不同结果，决定最优参数，然后上传。依次点击“Query data for analysis -> for quality control”，从表中得到数据的序号，然后依次点击 “Upstream analysis -> quality control” 进入分析页面，输入数据序号以及质控参数，点击“sumbit to show QC images”，查看质控结果，重复输入质控参数并点击“sumbit to show QC images”来确定最优参数，最后点击 “sumbit to confirm QC parameters” 上传。



图十、查询质控数据的序号并进行质控分析

2.3) 多样本数据融合/整合 (merge/integrate)：本平台支持两种形式的多样本数据组合，一般融合用于组织类型一样的样本数据，整合用于组织类型不一样的样本类型。同样的，点击“Query data for analysis -> for merge/integrate”查询数据的序号，然后点击“Upstream analysis -> multi-sample integration” 或者 “Upstream analysis -> multi-sample merge”进行数据融合/整合操作。Data ID 输入框输入数据序号，Group name 输入框输入自定义的组别名，点击 “Add” 添加一个数据，点击“Delete”删除当前数据，点击“submit”进行数据融合/整合操作（如图十一所示）。

2.4) 数据降维以及细胞聚类：该操作同样是实时交互式的，即用户可以尝试不同的参数，查看不同参数产生的不同聚类结果，决定最优参数，然后上传。依次点击“Query data for analysis -> for dc”，从表中得到数据的序号，然后依次点击 “Upstream analysis -> dimension reduction & clutering” 进入分析页面，输入数据序号以及参数，点击“sumbit to show DC images”，查看质控结果，重复输入质控参数并点击“sumbit to show QC images”来确定最优参数，最后点击 “sumbit to confirm DC parameters” 上传。

## 细胞身份注释

细胞聚类之后，就可以通过对细胞类群表达的基因情况 (cell marker) 从而对其身份进行注释。该部分分为两个步骤：1) 细胞主要类群身份注释，2) 细胞亚类群身份注释 (图十三)。前者是进行下游分析之前的必须步骤，后者这是可选的步骤，但是下游分析中的个别分析建议先进行细胞亚类群身份注释。细胞身份注释也是实时交互式的。

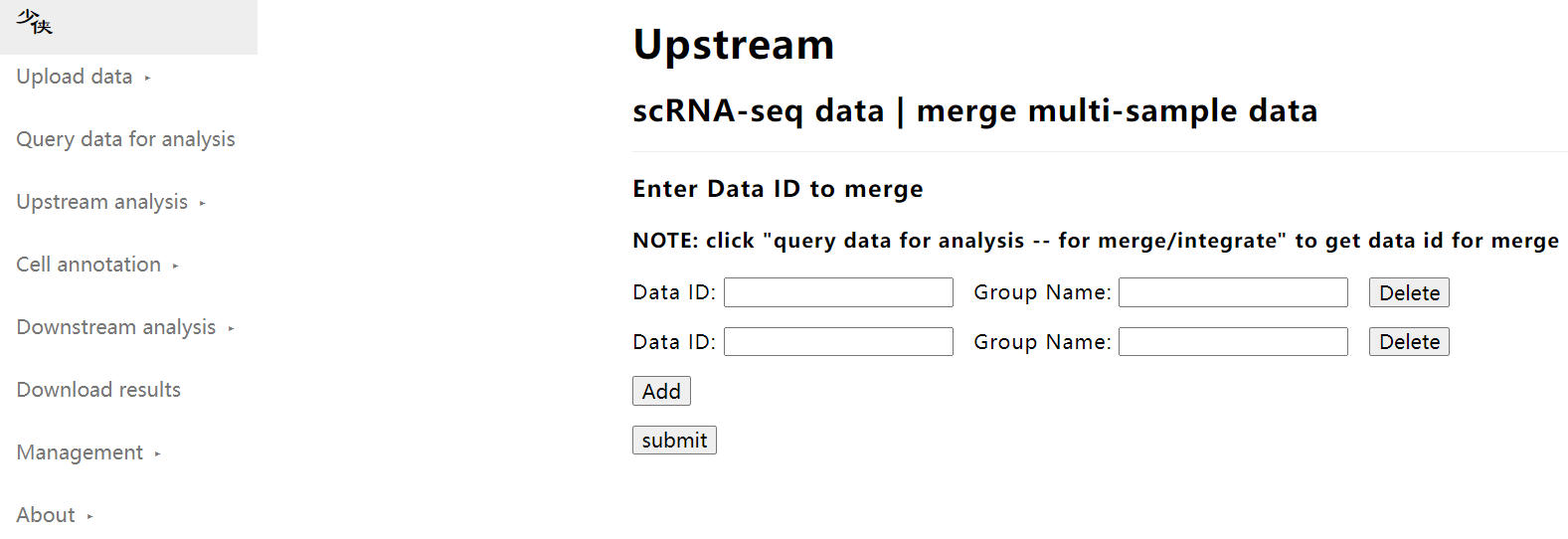
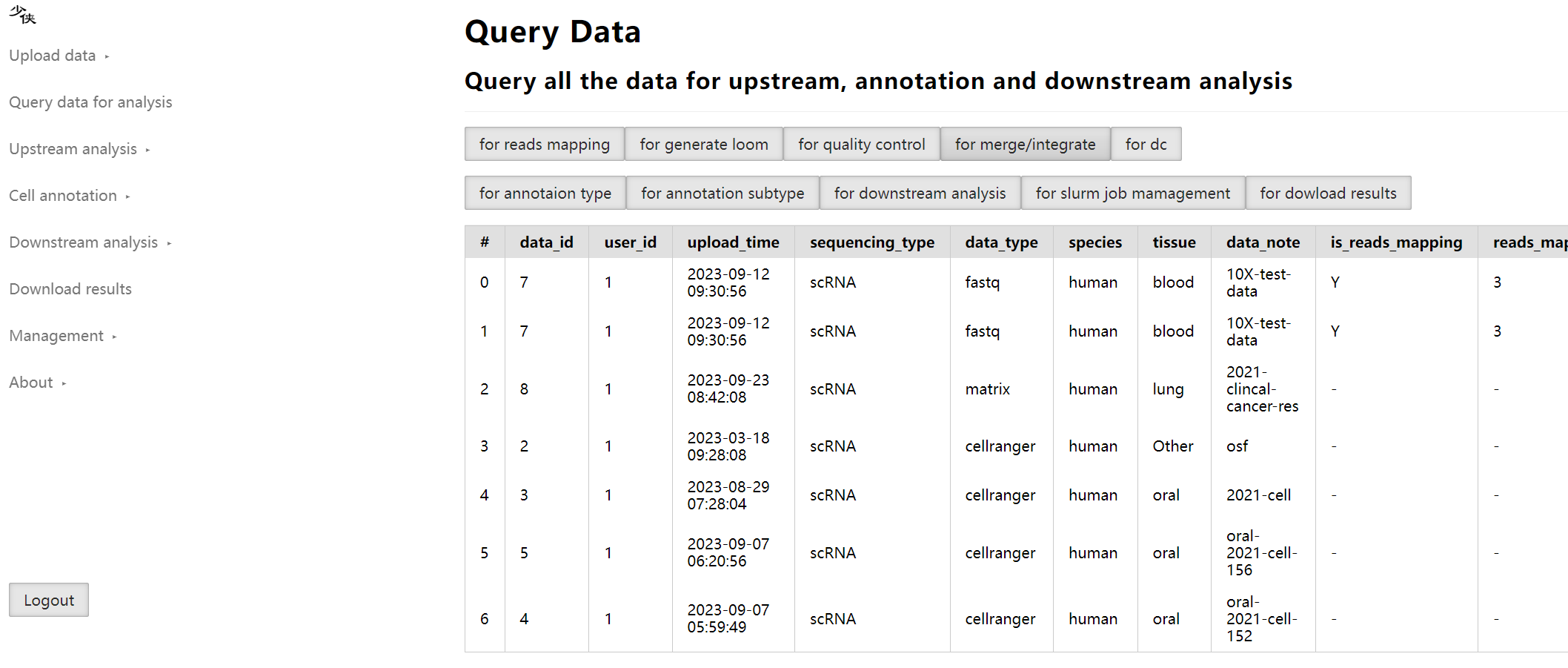
3.1) 细胞主要类群身份注释：该分析步骤基于细胞聚类的结果。同样地，首先依次点击 “Query data for analysis -> for annotation type” 查询用于细胞身份注释数据，得到相应数据的序号，而后依次点击“annotation -> cell-type annotation”，进入页面，输入数据序号，点击“ok”开始细胞身份注释。点击“download all marker file”下载所有细胞类群表达的基因(marker)，输入想要查看的基因，然后点击输入框后的“ok”，查看该基因在所有细胞类群表达的情况，根据基因在不同细胞类群中的表达情况，就可以对细胞类群进行注释了。在分析页面的右边，对应每一个类群(cluster)，在相应的输入框中输入细胞类群名。在页面底部，在相应的输入框中输入挑选好的基因名序列(cell marker)以及细胞类群名序列，例如：

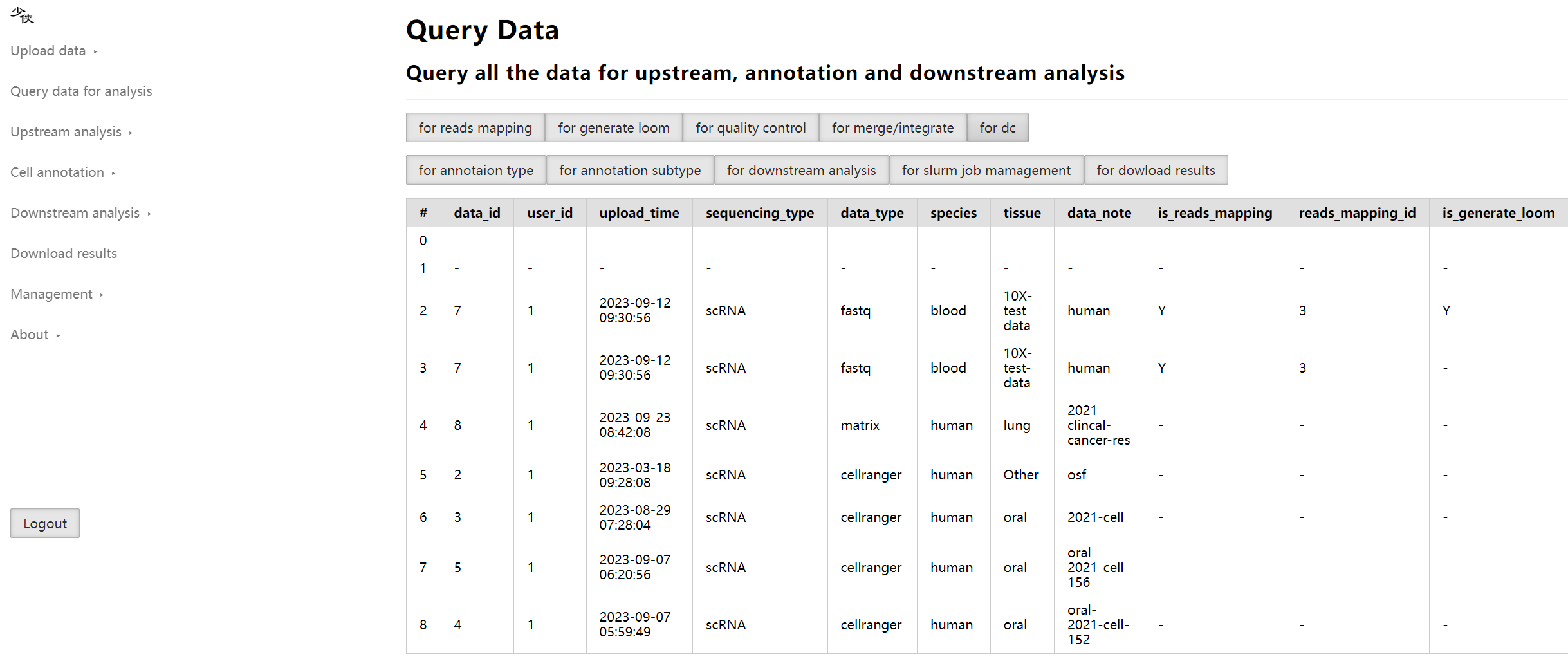
cluster name order -- epithelial:endothelial:immune:fibroblast:mesenchymal

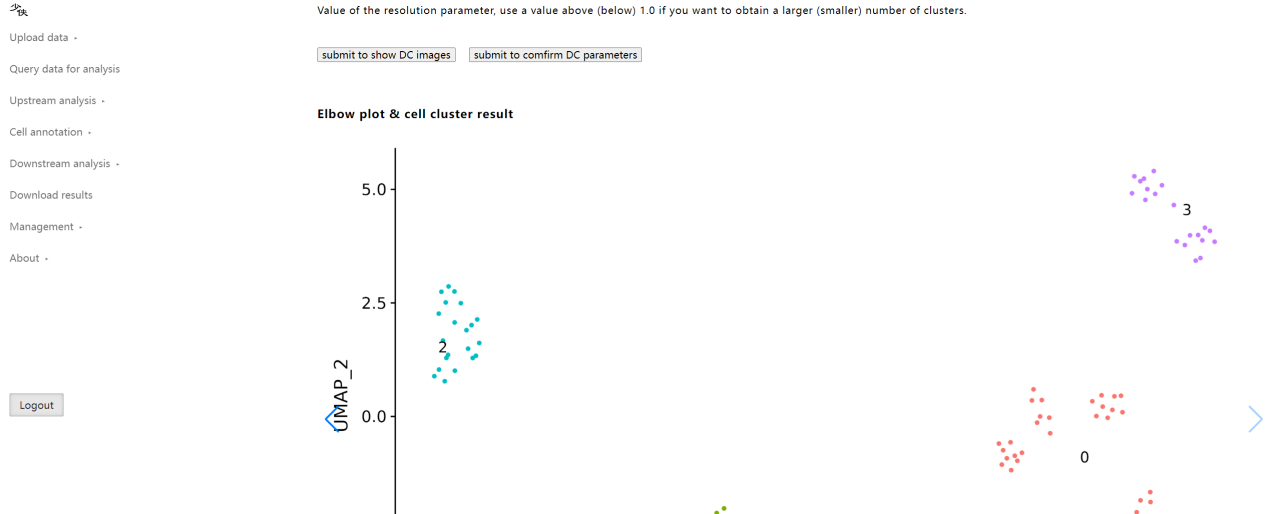
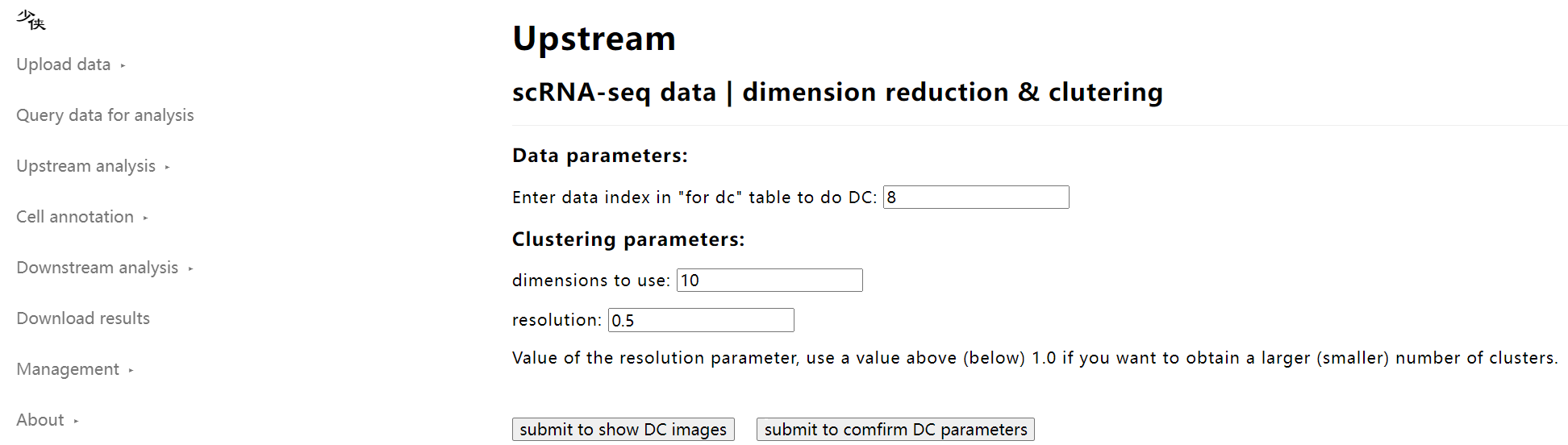
marker gene order -- KRT5:KRT14:SELE:VWF:CD86:CD36:COL3A1:COL4A3:CDH3:KIT

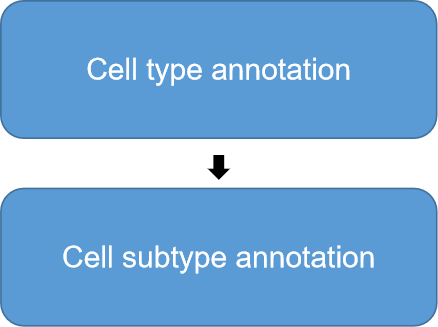
这两个信息是用来产生细胞身份注释结果图的，如图十五所示。

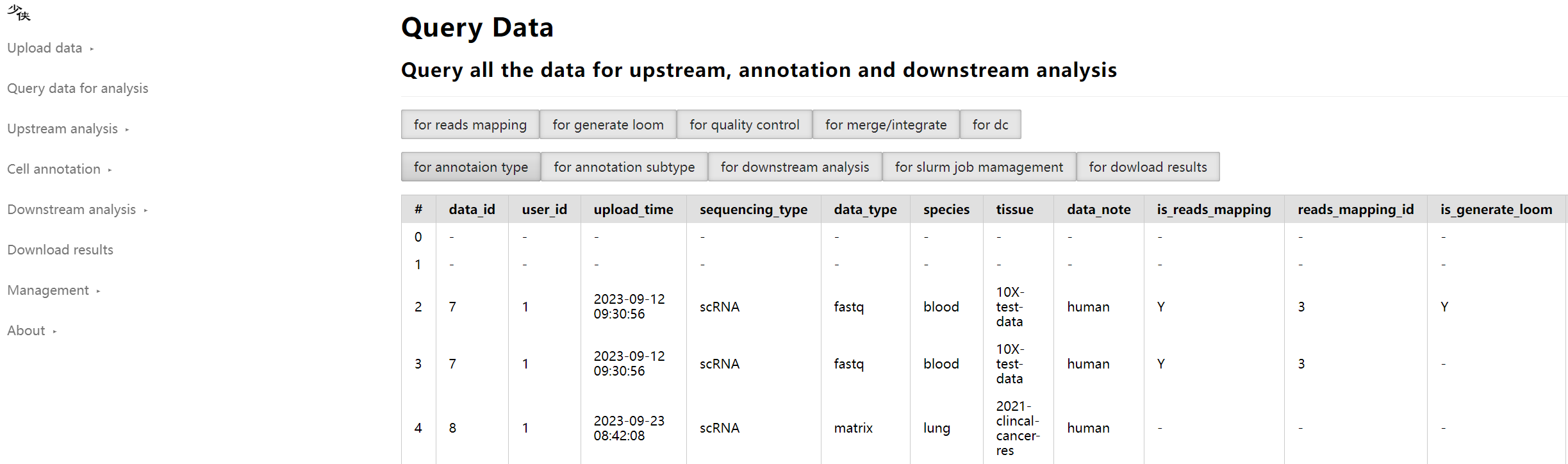
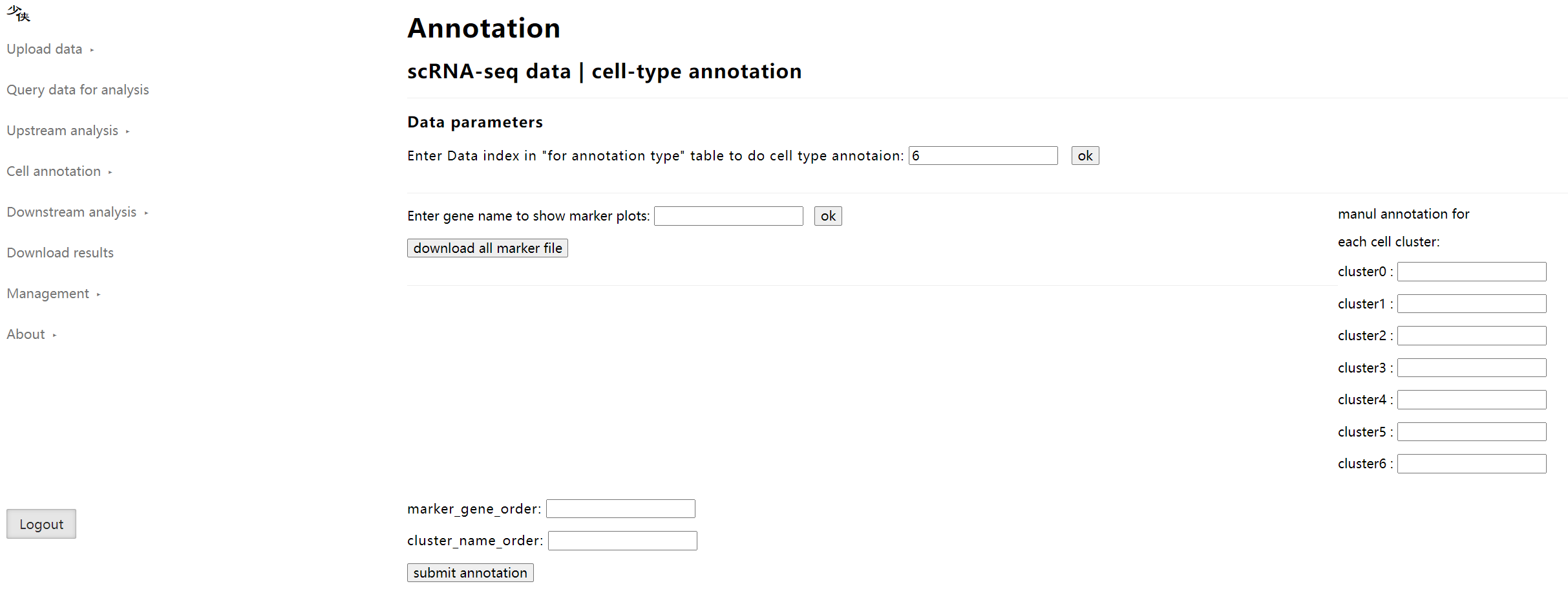
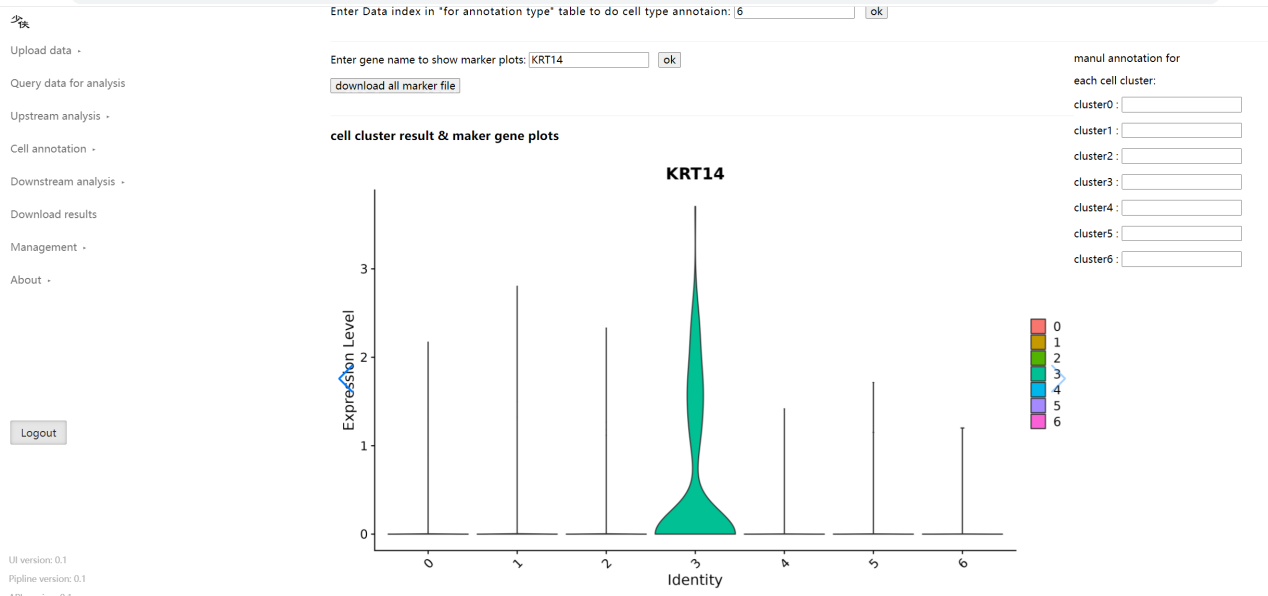
图十一、查询数据序号并进行多个样本数据融合/整合

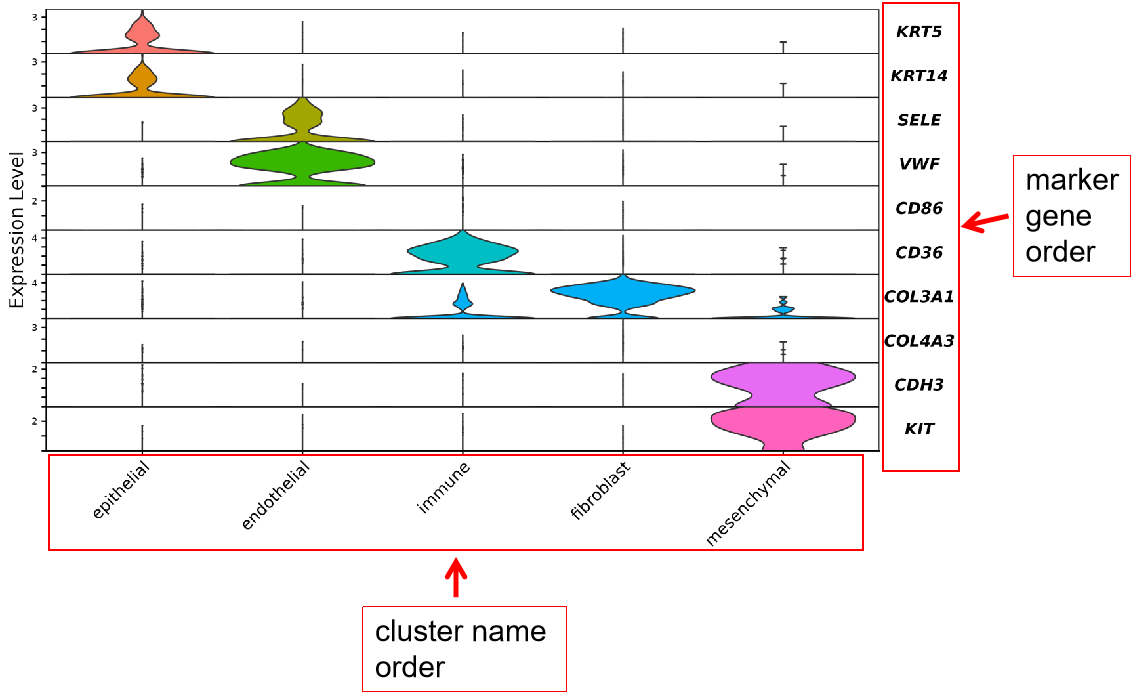


图十二、查询数据并进行降维及细胞聚类分析



图十三、细胞身份注释

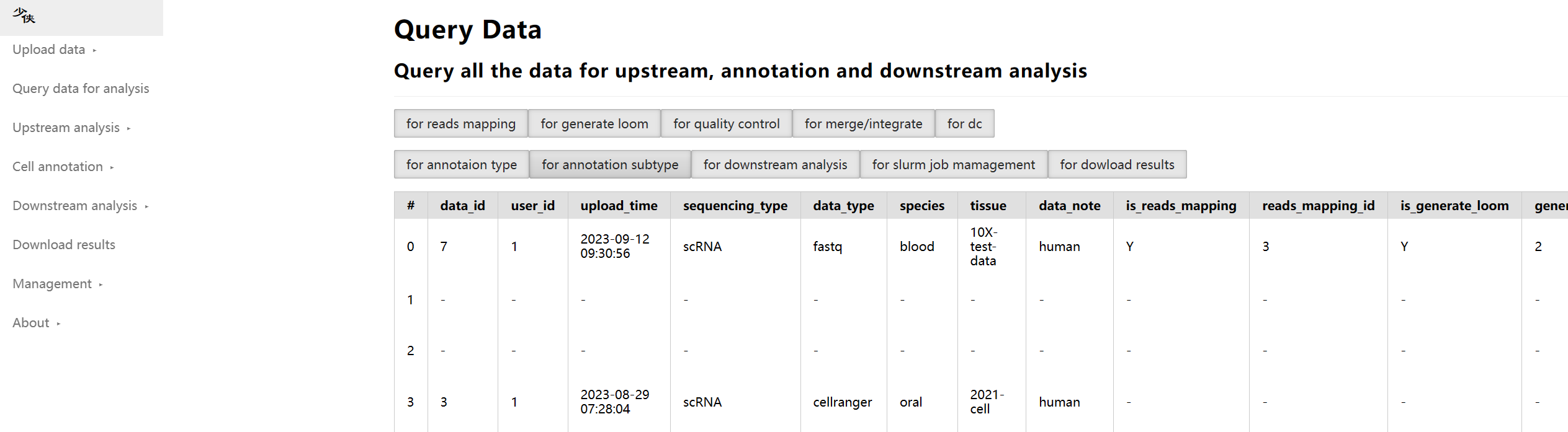
图十四、查询数据并进行细胞身份注释



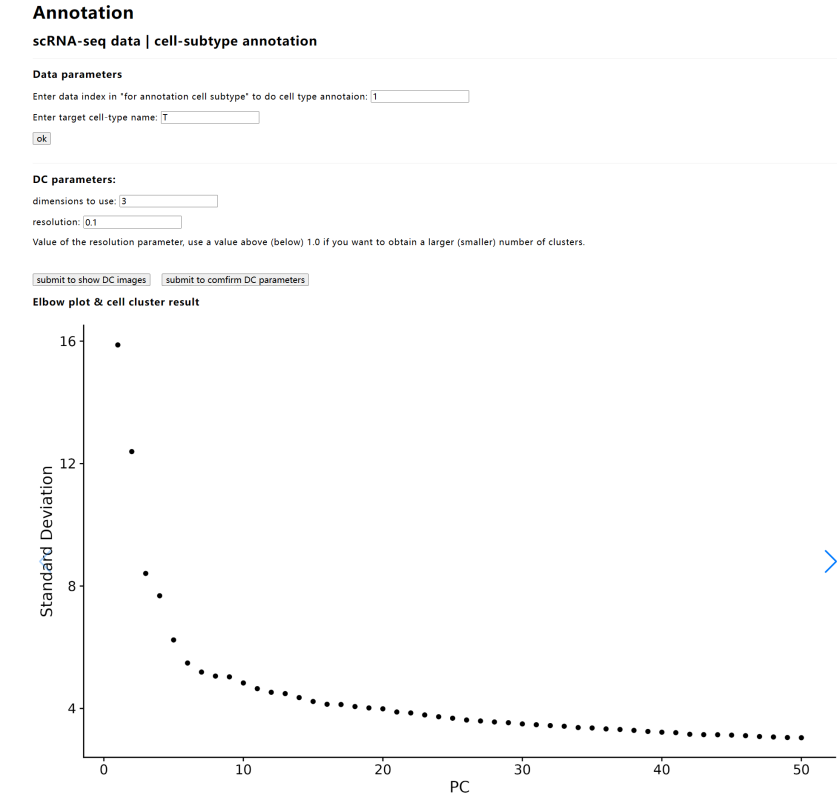
图十五、细胞身份注释结果

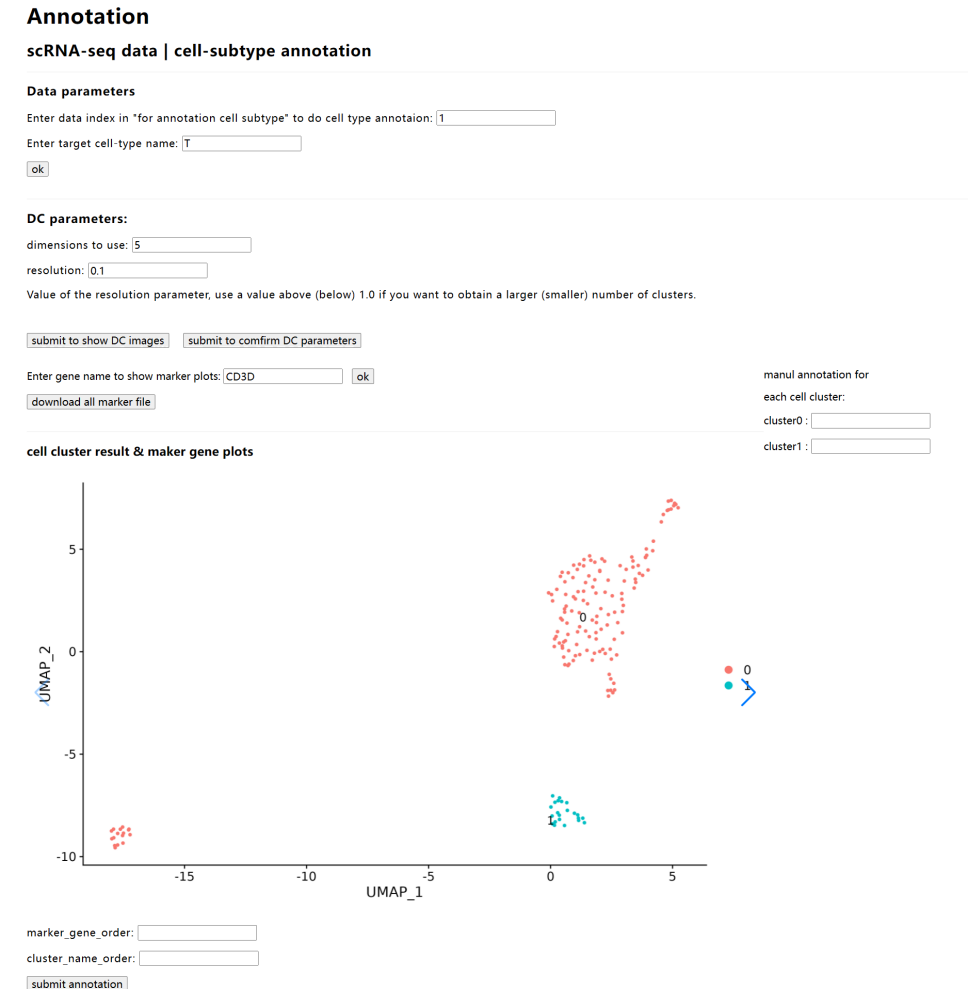
3.2) 细胞亚型注释：该分析基于细胞主要类群细胞身份注释结果，并且该分析是可选的。例如，如果你想对一些免疫细胞进行亚型鉴定，就可以做该分析。反之，则不需要。点击

“Query data for analysis -> for annotation subtype”查询数据的序号(图十六)，然后点击“Cell annotation -> cell-subtype annotation”进入页面，输入数据序号开始分析。细胞亚型分析有两个步骤—— DC (dimension reduction & cell clustering) (图十七) 以及 annotation (图十八)。操作与上游分析的数据降维及细胞聚类一样和细胞主要类群身份注释操作一样



图十六、查询用于细胞亚型鉴定的数据

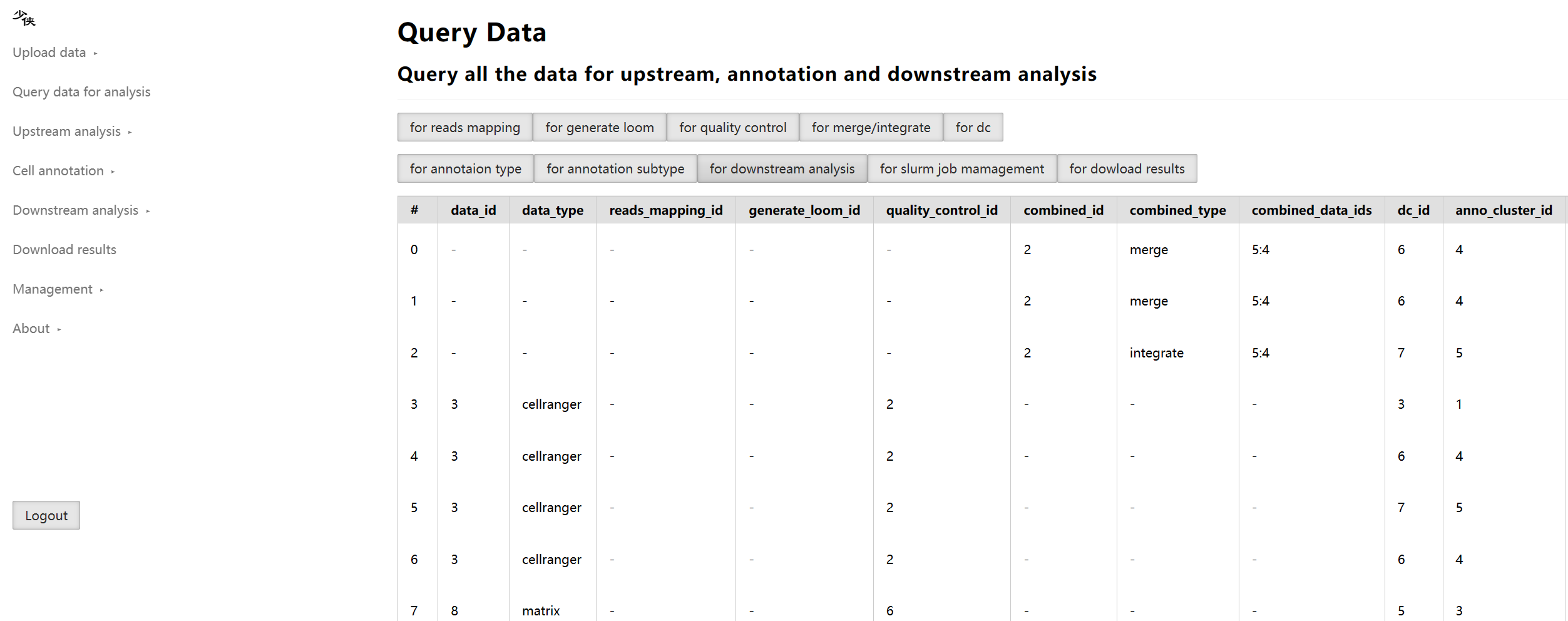
图十七、细胞亚型鉴定中的数据降维及细胞重新聚类

图十八、细胞亚型鉴定

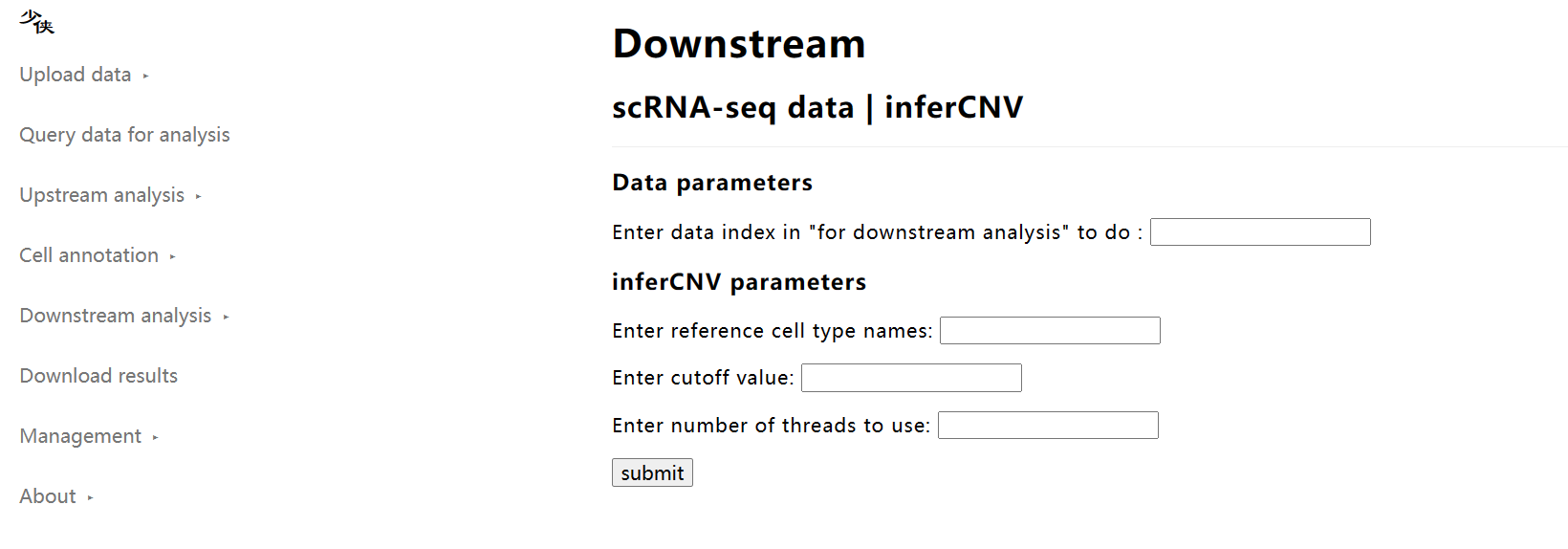
## 下游分析

在进行细胞身份注释之后，就可以进行下游的11个分析了。遵循同样的操作逻辑，要进行下游分析，首先要查询下游分析所需的数据，然后输入数据的序号以及分析所需的参数，点击上传，完成分析。所有下游分析均使用SLURM进行分析任务调度，使得分析平台具有较高的伸缩性。

4.1) 拷贝数变异分析 (inferCNV)：该分析应用于肿瘤样本，用于识别肿瘤细胞。不是肿瘤样本不应作拷贝数变异分析。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> inferCNV”，进入分析页面，输入数据序号以及分析所需参数，然后点击上传，即可提交分析任务到SLURM任务队列(图二十)。

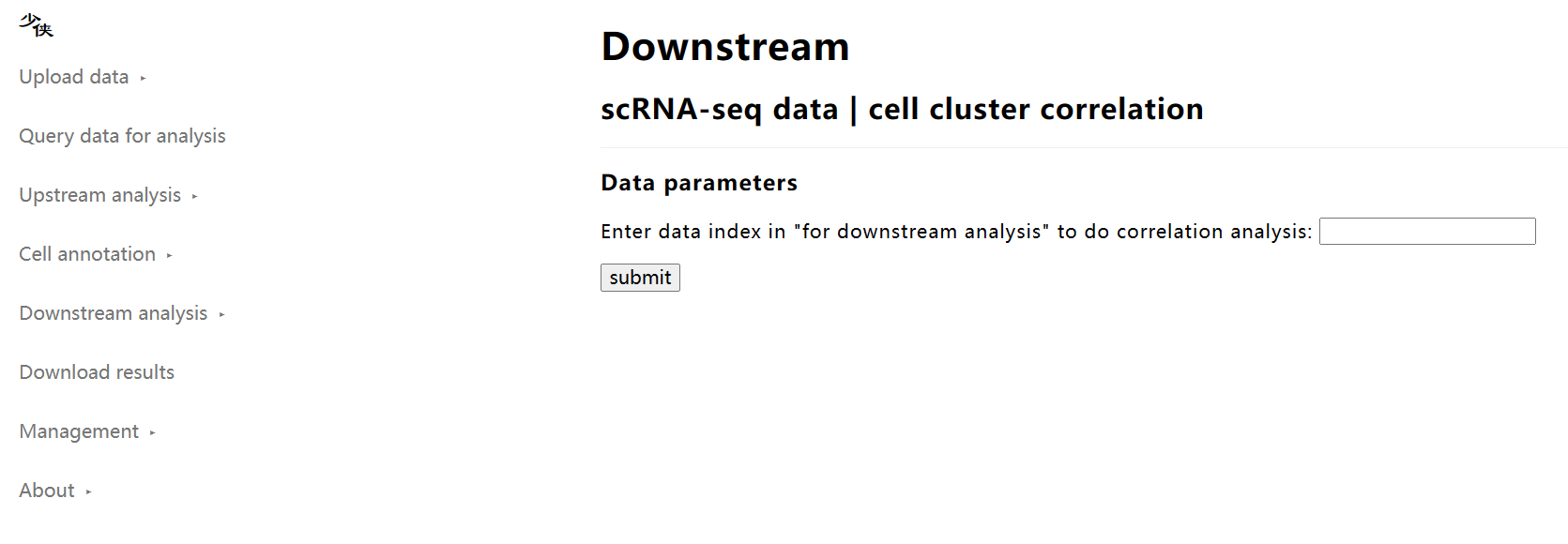


图十九、查询用于下游分析的数据



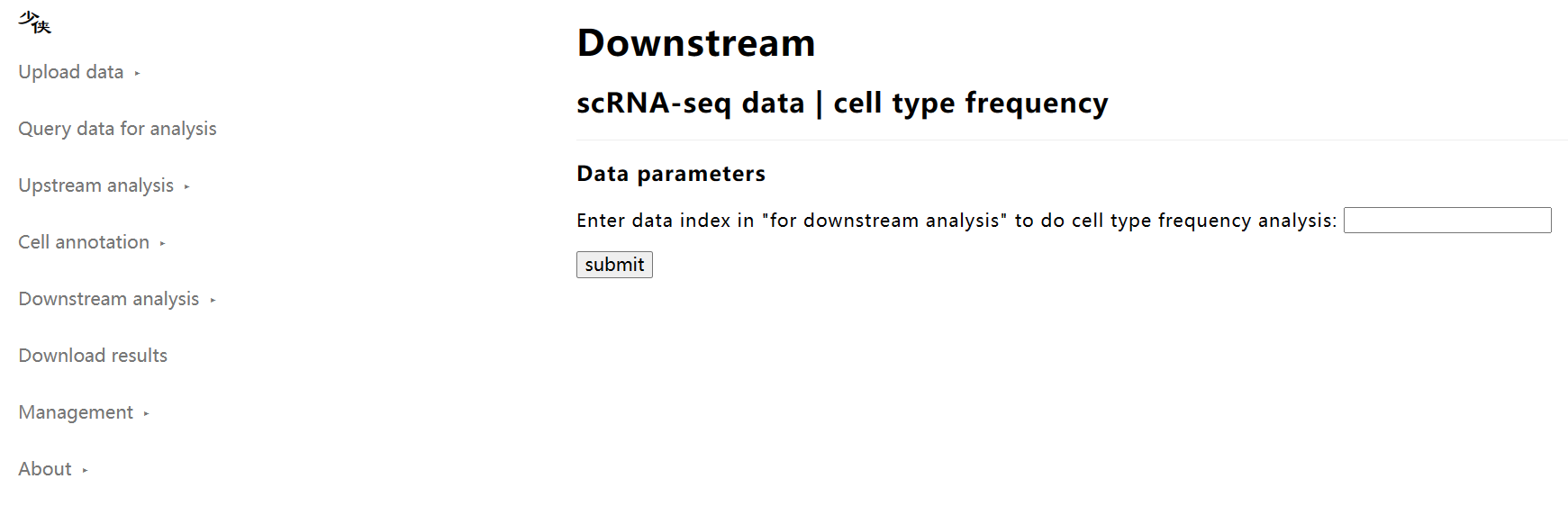
图二十、拷贝数变异分析

4.2) 细胞类群相关性分析(cell cluster correlation)：该分析适用于所有的样本类型，计算注释出来的细胞类型之间的相关性。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> cell cluster correlation”，进入分析页面，输入数据序号，然后点击上传，即可提交分析任务到SLURM任务队列(图二十一)。



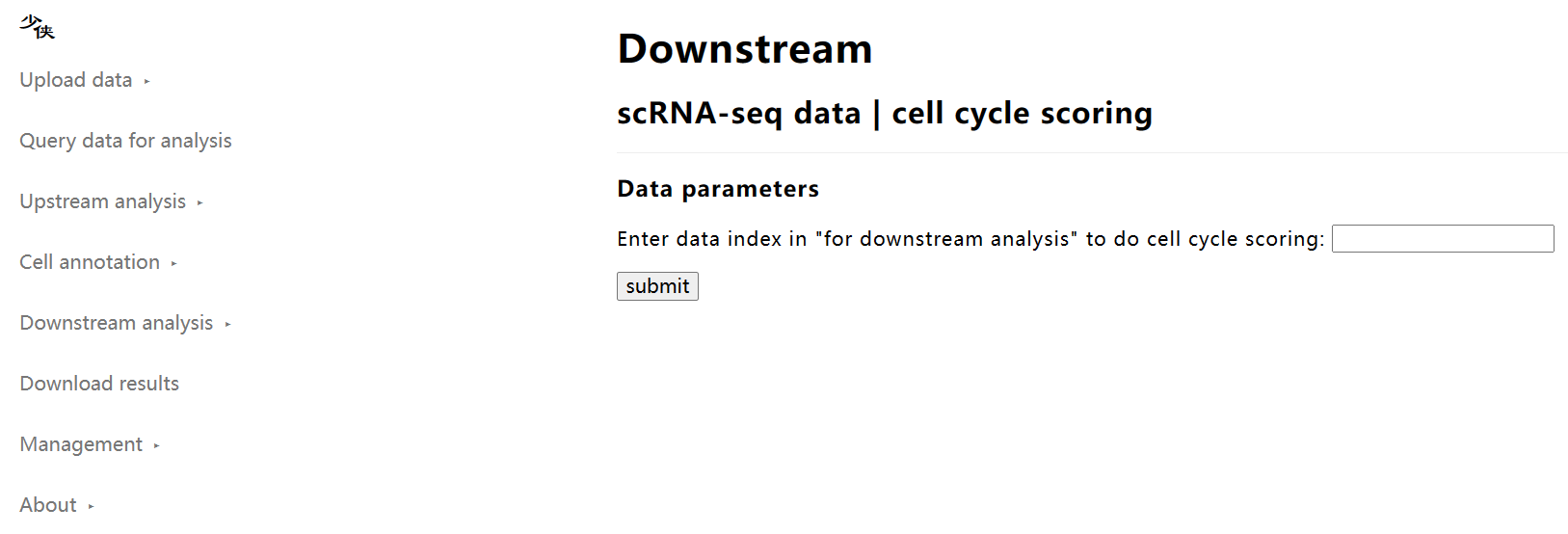
图二十一、细胞类群相关性分析

4.3) 细胞类型组份分析(cell type frequency)：该分析适用于所有的样本类型，计算各个细胞类型的占比。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> cell type frequency”，进入分析页面，输入数据序号，然后点击上传，即可提交分析任务到SLURM任务队列(图二十二)。



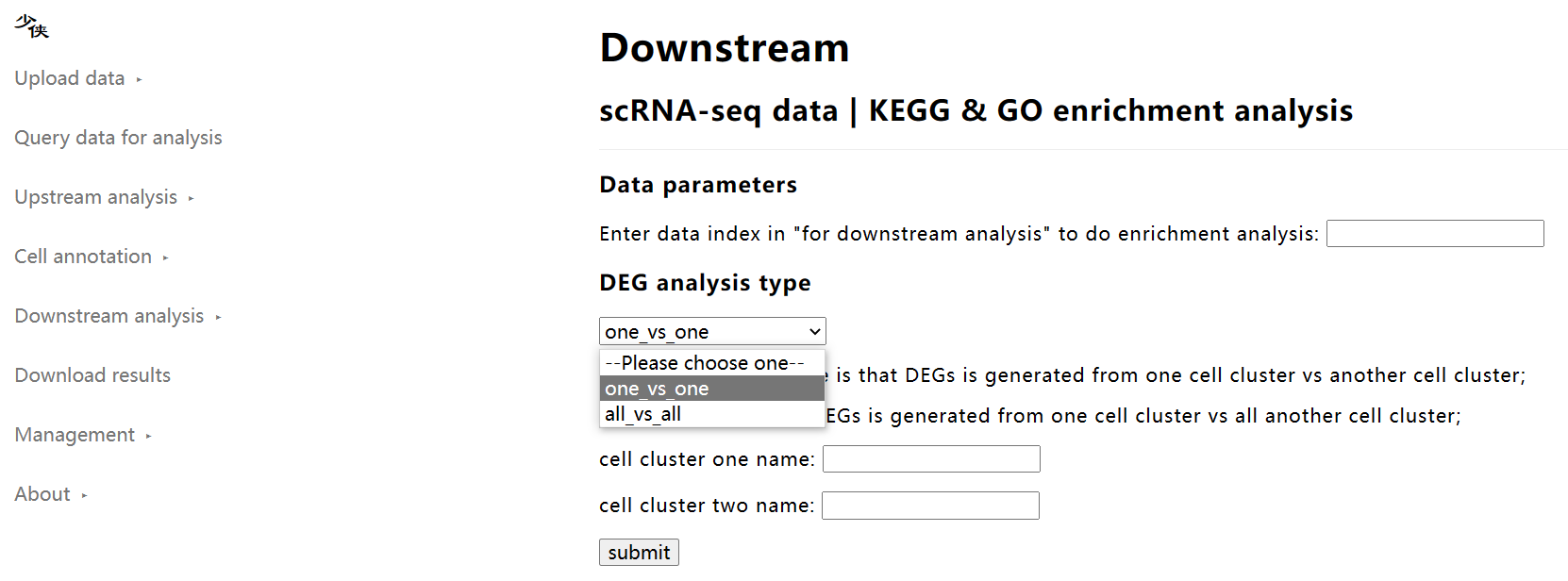
图二十二、细胞类型组份分析

4.4) 细胞周期评分 (cell cycle analysis)：根据细胞周期相关基因的表达量，计算每一个细胞所在的周期，适用于所有样本。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> cell cycle analysis”，进入分析页面，输入数据序号，然后点击上传，即可提交分析任务到SLURM任务队列(图二十三)。



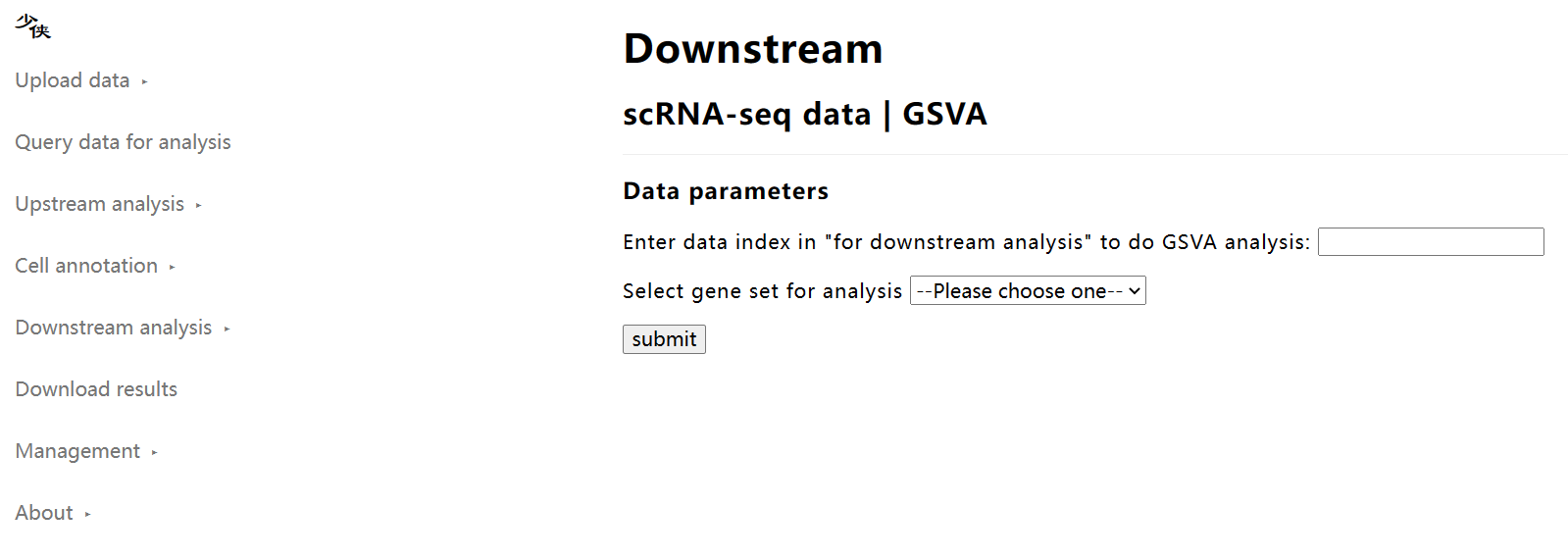
图二十三、细胞周期分析

4.5) KEGG & GO 富集分析：适用于所有数据类型，计算两个细胞类群的差异表达基因，然后用这些基因进行KEGG以及GO的富集分析。它包含两种分析方式，第一种是所有细胞类群中两两比较(all\_vs\_all)，另外一种是选择两个细胞类群进行比较(one\_vs\_one)。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> ”，进入分析页面，输入数据序号，然后点击上传，即可提交分析任务到SLURM任务队列(图二十四)。



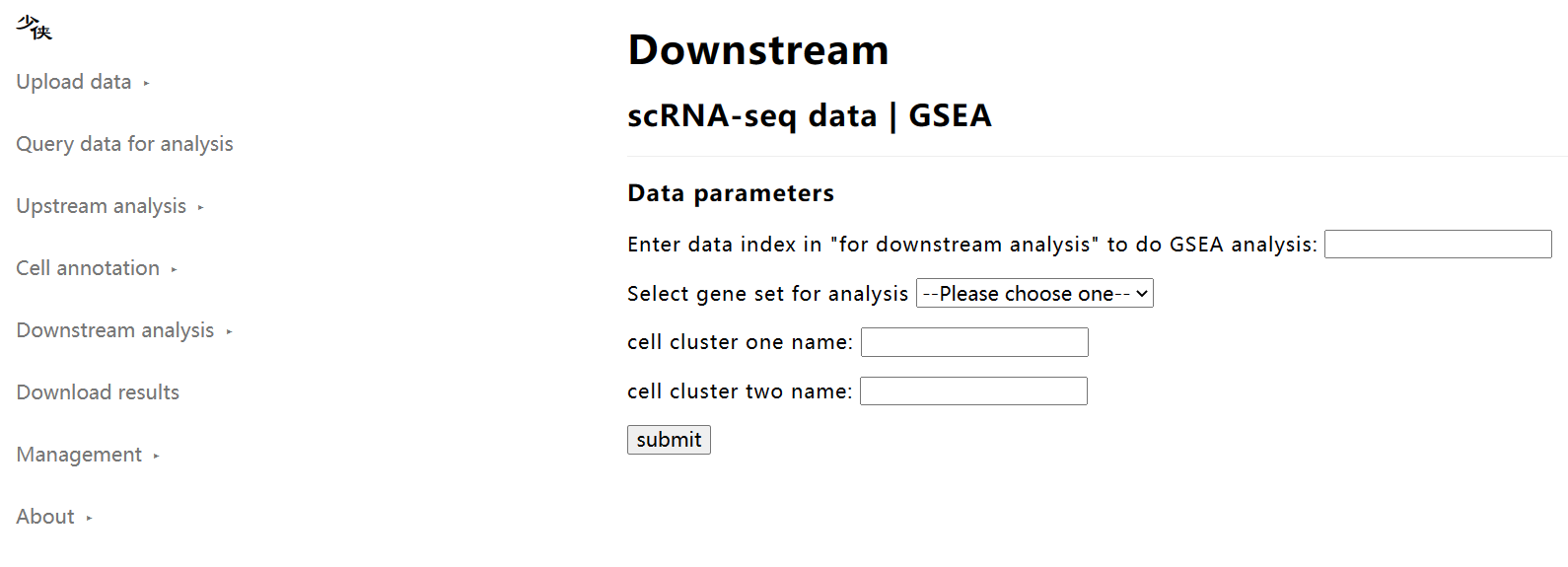
图二十四、KEGG & GO 富集分析

4.6) GSVA分析：基因集变异分析，不进行差异表达基因的计算，也针对两个细胞类群的比较，直接将基因的表达富集到不同基因集，分析其样本整体结构与功能。适用于所有数据类型。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> GSVA”，进入分析页面，输入数据序号，选择基因集，然后点击上传，即可提交分析任务到SLURM任务队列(图二十五)。



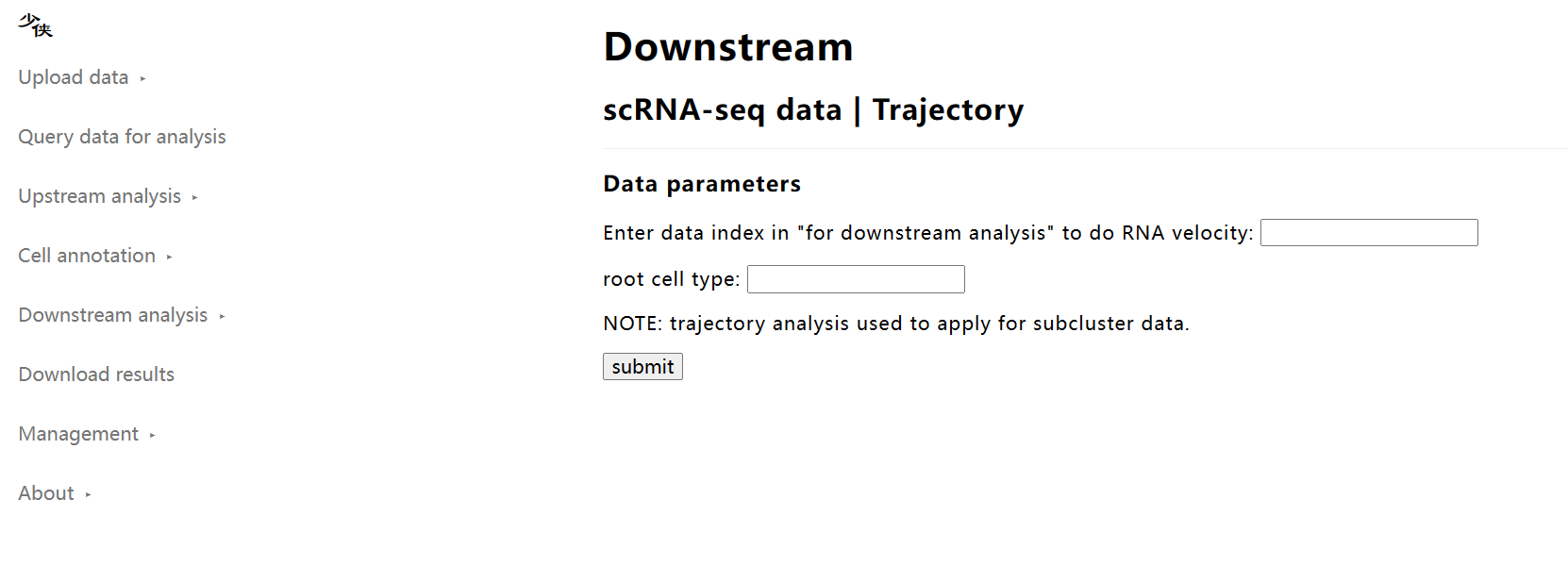
图二十五、GSVA分析

4.7) GSEA分析：不计算两个细胞类群的差异基因，直接比较两个细胞类群的基因表达在某个基因集的差异，从而分析两者之间的功能差异。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> GSVA”，进入分析页面，输入数据序号，选择基因集，输入要比较的两个细胞类群，然后点击上传，即可提交分析任务到SLURM任务队列(图二十六)。



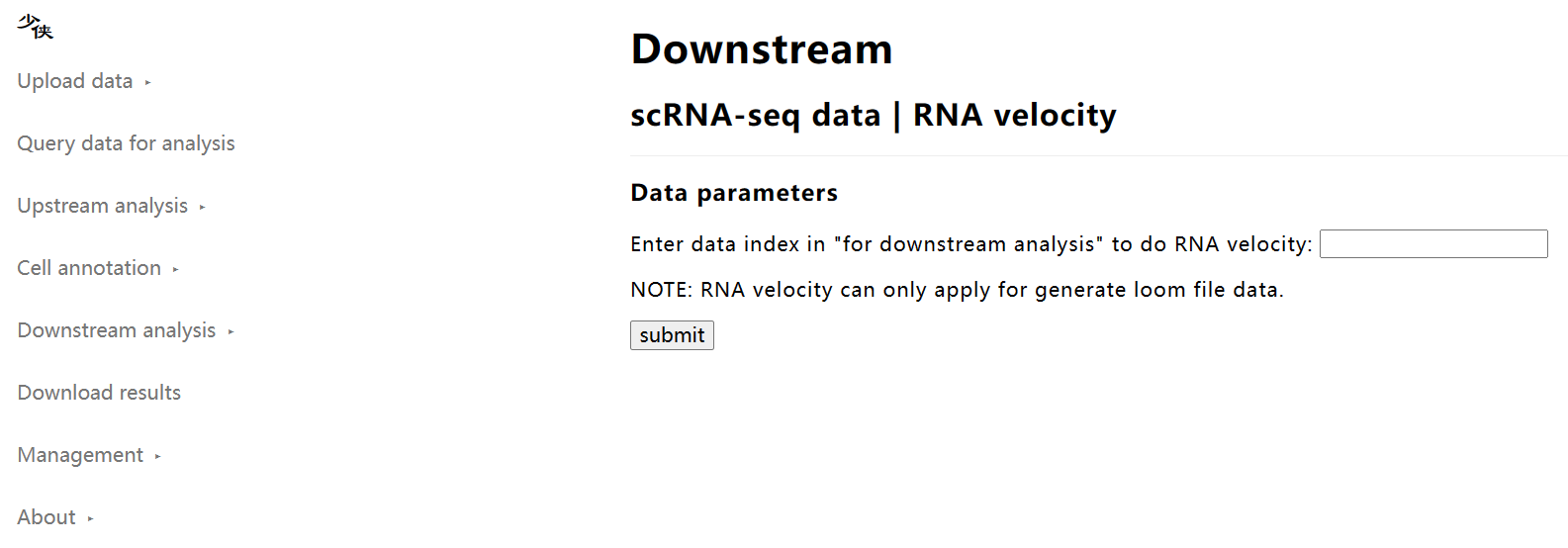
图二十六、GSEA分析

4.8) 拟时分析：一般用于细胞亚群注释结果，用于解读细胞亚型类型之间的变化。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> trajectory”，进入分析页面，输入数据序号以及相应的参数，然后点击上传，即可提交分析任务到SLURM任务队列(图二十七)。



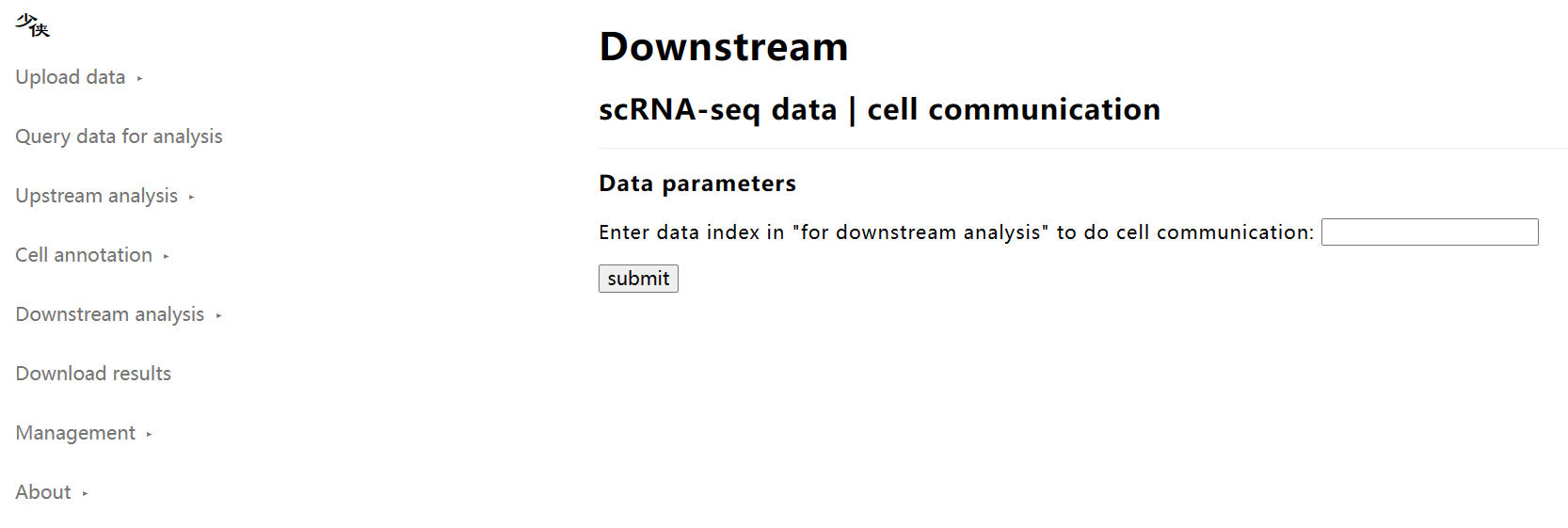
图二十七、拟时分析

4.9) RNA速率分析：只能应用于做过上游分析中的生成loom文件分析的数据。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> RNA velocity”，进入分析页面，输入数据序号，然后点击上传，即可提交分析任务到SLURM任务队列(图二十八)。



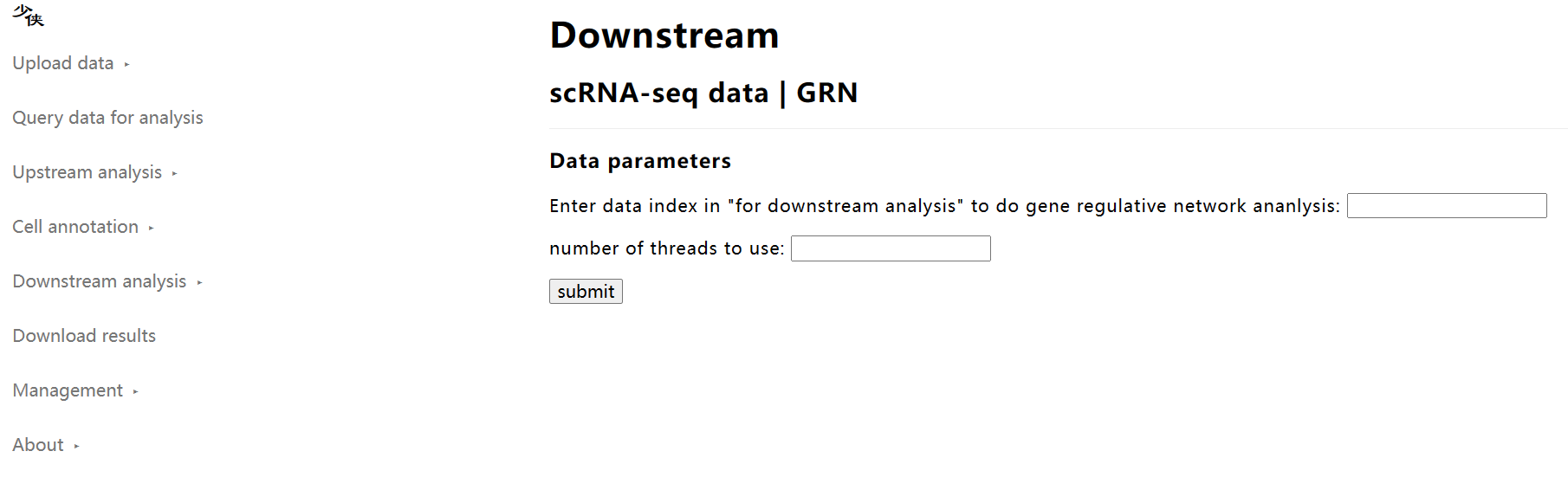
图二十八、RNA 速率分析

4.10) 细胞通讯分析：用于分析细胞类群之间的通讯分析，适用于所有数据类型。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> cell communication”，进入分析页面，输入数据序号，然后点击上传，即可提交分析任务到SLURM任务队列(图二十九)。



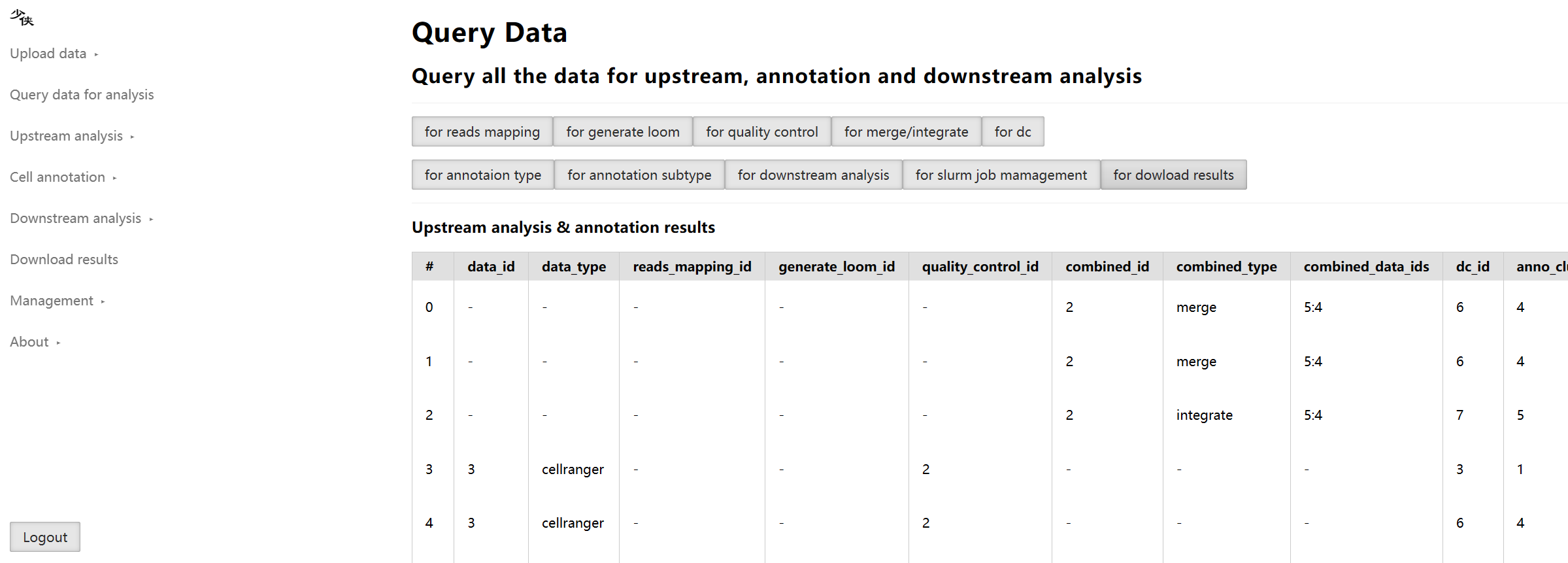
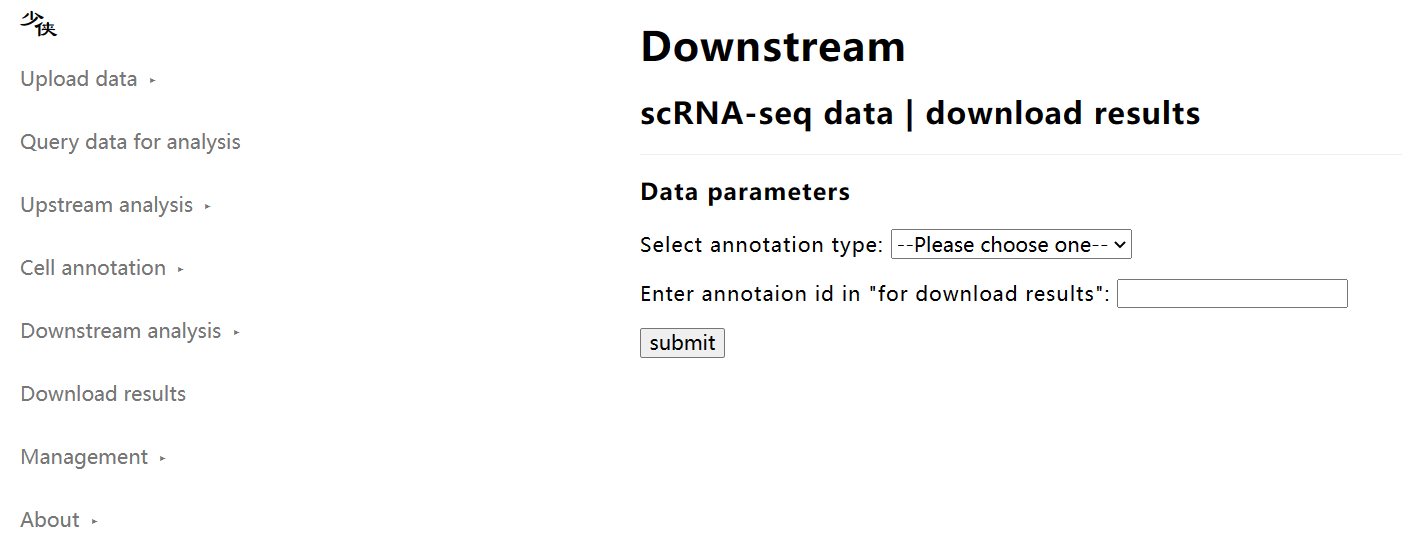
图二十九、细胞通讯分析

4.11) 基因调控网络分析：适用于所有数据类型。首先点击“Query data for analysis -> for downstream analysis”得到数据序号 (图十九)，然后点击“Downstream analysis -> regulateNet”，进入分析页面，输入数据序号，然后点击上传，即可提交分析任务到SLURM任务队列(图三十)。

图三十、基因调控网络分析

## 下载分析结果：

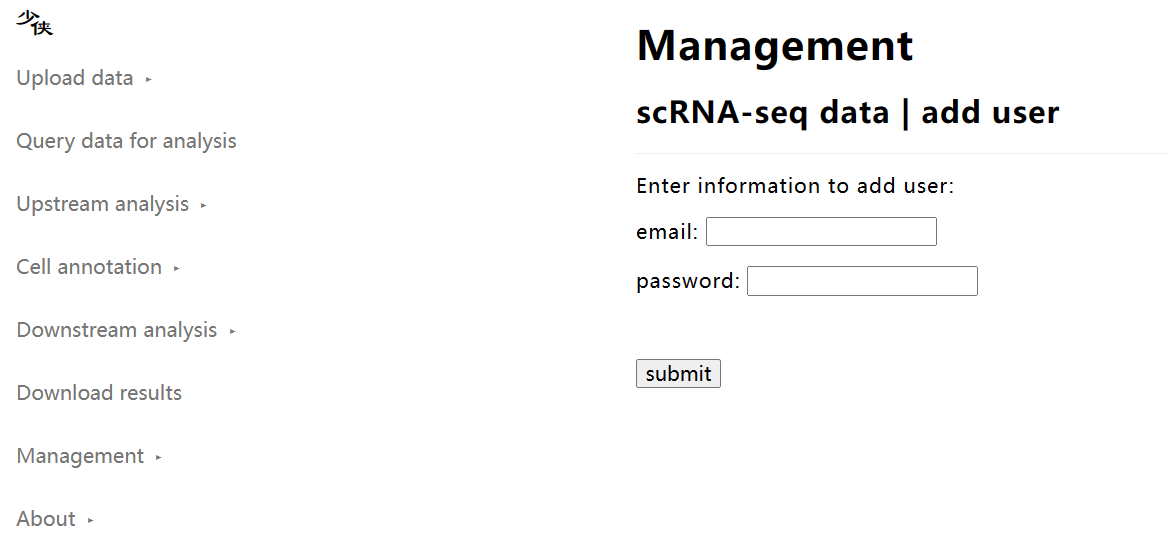
只要进行了至少一项下游分析的样本数据，就可以下载分析结果。首先点击“Query data for analysis -> for download results”得到样本的注释ID，然后点击“Download results”，进入下载页面，选择注释类型，输入相应的注释ID号，然后点击上传，即可下载分析结果。

图三十一、下载数据分析结果

## 平台管理

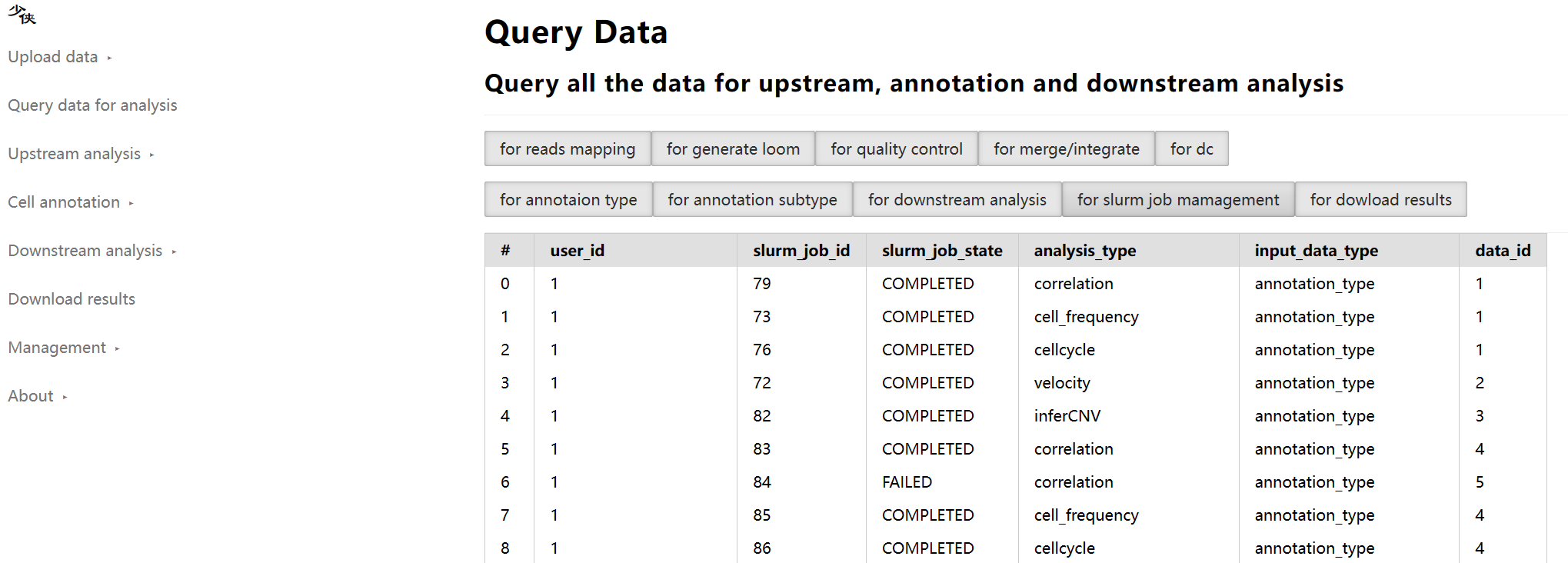
平台的管理有两个方面：1) 用户添加，管理员可以添加普通用户；2) 任务管理，通过SLURM，可以取消、暂停、启动分析任务。

6.1) 添加普通用户：管理员可以添加普通平台用户，点击“Management -> add user”，进入添加用户页面，输入相应信息，点击上传，即可。

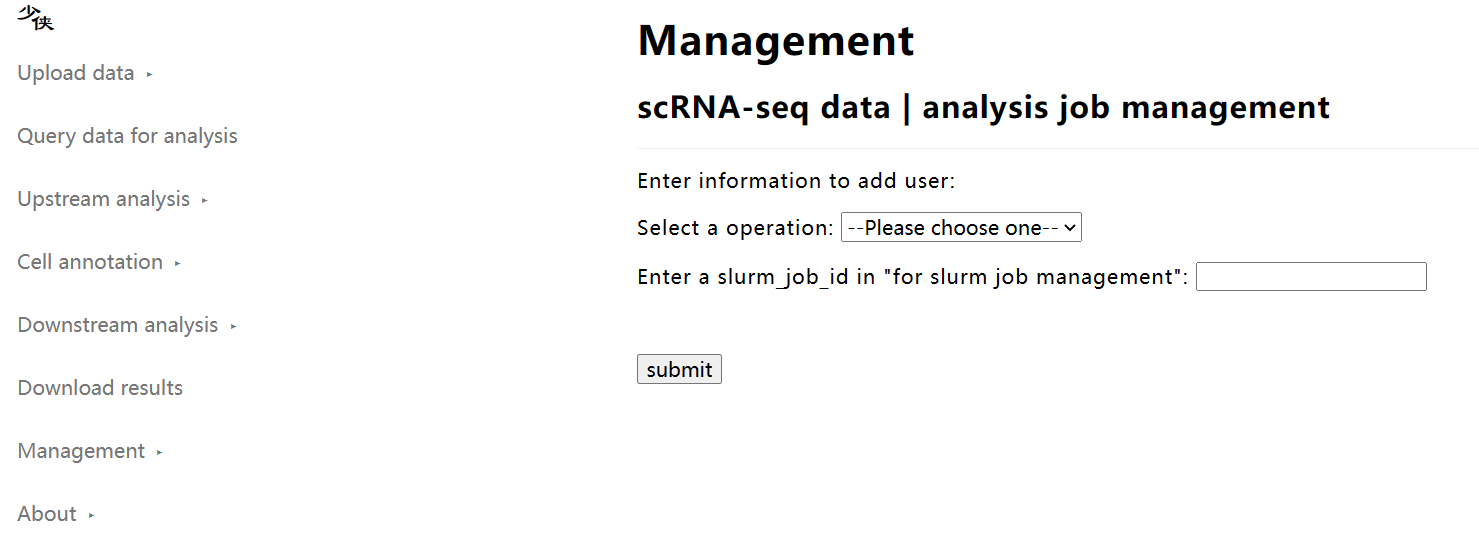


图三十二、添加用户

6.2) 任务管理：上游分析中的reads mapping和generate loom，以及所有下游分析任务都是通过SLURM进行任务调度的，也就可以进行相应的任务管理。点击“Query data for analysis -> for slurm job management”查询SLURM 任务ID号(图三十三)，然后点击“Management -> manage analysis jobs” 进入管理页面，输入任务ID，选择想进行的操作，点击上传(图三十四)。



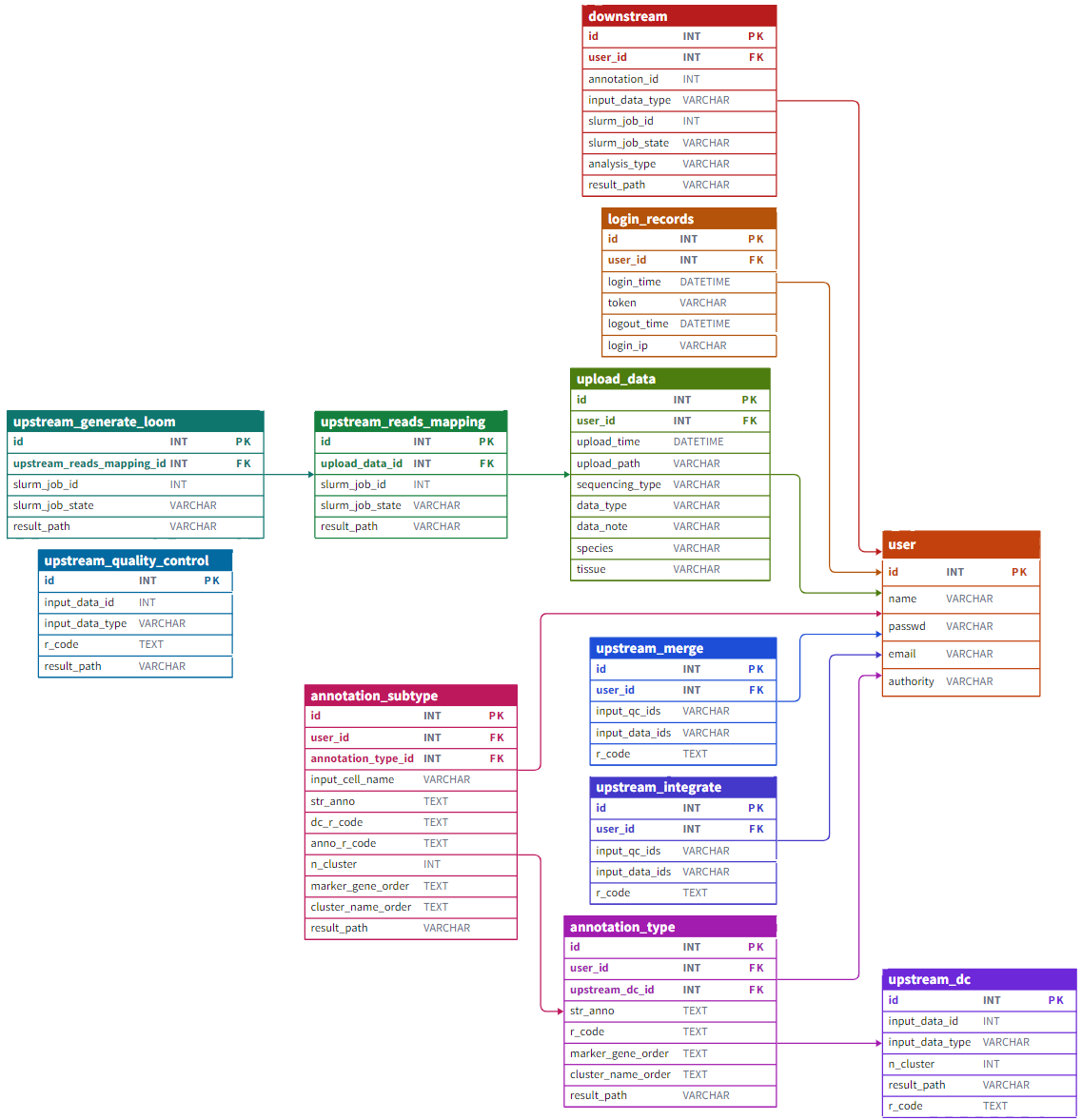
图三十三、查询SLURM任务ID号



图三十四、分析任务管理

# 数据库设计

数据库设计如下：

图三十五、数据库框架设计