Supplemental Tutorial - CellDepot: A unified repository for scRNAseq data and visual exploration

Dongdong Lin Yirui Chen Soumya Negi Derrick Cheng Zhengyu Ouyang
David Sexton Kejie Li Baohong Zhang*

2021-11-29

Contents

Chapter 1

Preface

This is a **supplemental tutorial** written in Markdown, which provides the detailed guide for CellDepot web portal.

Chapter 2

Getting start with CellDepot

2.1 Sources of annotation and metadata

The original metadata information of each single cell RNA-seq dataset is retrieved from h5ad file, which is a preferred way of sharing and storing an on-disk representation of anndata object. When importing the dataset to the system, user inputs additional metadata information as shown in (4.4). Both metadata are collected and stored in a MySQL database table that is presented at http://celldepot.bxgenomics.com and Biogen internal link http://go.biogen.com/CellDepot.

2.2 Data format, availability and preparation

CellDepot requires scRNA-seq data in h5ad file where the expression matrix is stored in CSC (compressed sparse column) instead of CSR (compressed sparse row) format to improve the speed of data retrieving. For example, designating genes as columns in the h5ad file creates the interactive plot five times faster than as rows. Just in case, we provide sample scripts to help users generate h5ad files. Having gene expression matrix, metadata, and layout files, users can easily combine and convert their data to h5ad file by following this R script on https://github.com/interactivereport/CellDepot/blob/main/toH5ad.R. In the case of lacking layout file, users can also create h5ad file by following the Jupyter notebook https://github.com/interactivereport/CellDepot/blob/main/raw2h5ad.ipynb with custom python script tailored to their own data. Categorical features extracted from a h5ad file are shown in the 'annotation groups' column of the table on CellDepot home page, while the numerical features are shown as the distributions in the rightmost panel on cellxgene VIP

2.3 CellDepot platform and installation

The public version of CellDepot web portal is hosted at the web site, http://celldepot.bxgenomics.com and Biogen internal link http://go.biogen.com/CellDepot. It is implemented with MySQL database, an advanced search engine, and powerful interactive visualizing tools that allow users to explore attributes of datasets as well as scRNA-seq analysis results. Also, users can intentionally select single-cell RNA-seq datasets on the web interface by simply browsing the online dataset table or applying advanced search to perform the cross-dataset comparison. Moreover, CellDepot also provides comprehensive data analysis tools via an embedded interactive visualization plugin. To host private datasets, local instance of CellDepot on Unix server can be installed by following the guide here, https://celldepot.bxgenomics.com/celldepot_manual/install_environment.php.

2.4 Data import on user's server

The prepared h5ad files are required to copy to a folder defined in the configuration file, e.g., /data/celldepot/all_h5ad_files/. Afterwards, users can navigate to the CellDepot home page, click 'Import Project' at the top menu, then 'Download

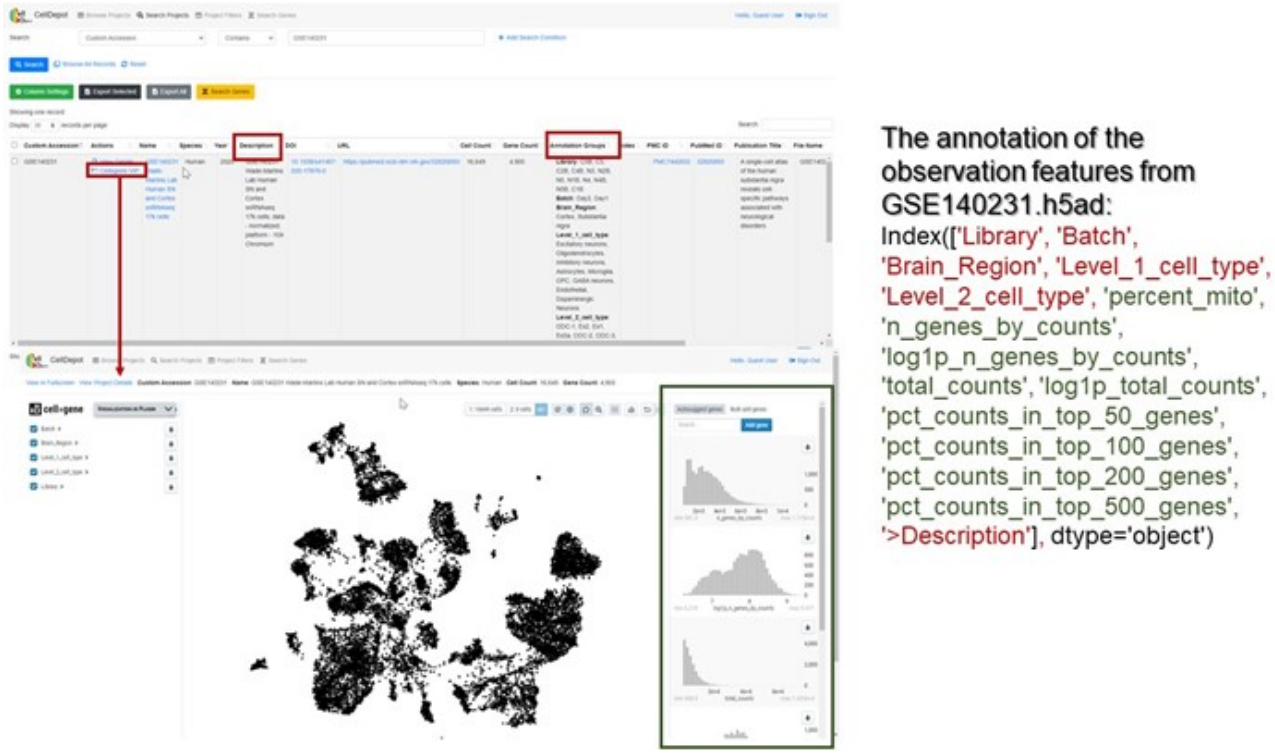

Figure 2.1: Figure S3. The exploration of observation features of dataset GSE140231. Red-marked categorical features are shown on CellDepot Project page (highlighted by red framed box), the numerical features marked in green color can be visualized as distribution plots on the rightmost panel in cellxgene VIP tool, which is highlighted by green box.

Example File' to fill in meta information of datasets into the downloaded template for submission. After the metadata file is uploaded, CellDepot will automatically convert the dataset to CSC format if needed through a cron job (4.5). To explore the detail of imported datasets, users can enter 'Browse Projects' page and then search these datasets by user assigned accessions in the metadata file.

2.5 CellDepot API (Application Programming Interface)

The CellDepot API web service provides a direct way to generate figures for users to share or embed in web page. For example, the following URL will generate a gene expression violin plot across cell clusters for IRAK4 gene for the data set with ID equaling one, https://celldepot.bxgenomics.com/celldepot/app/core/api_gene_plot.php?ID=1&Genes=IRAK4&Plot_Type=violin&Subsampling=0&n=0&g=0&Project_Group=CLUSTER. The complete format of the URL and explanation of parameters are detailed in the web page, https://celldepot.bxgenomics.com/celldepot_manual/api_gene_plot.php.

2.6 Code availability

The source code, local installation guide and complete tutorial of visualization and analysis tool are provided at https://github.com/interactivereport/CellDepot. With broad adoption and contribution in mind, CellDepot is released under the MIT License.

Chapter 3

Supplemental Tables

Table S1 - Comparsion matrix of web portal tools

Note: The criteria for query search and data analysis explorer please see Table S2 and S3.

Table S2 - Criterion for query search

Table S3 - Criterion for data analysis explorer

Table S4 - Project metadata captured in CellDepot

Table S1: Comparsion matrix of web portal tools

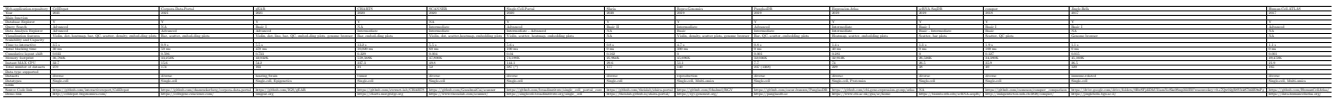

Table S2: Criterion for query search

Query.Search	Keyword.Search	Multiple.Object.Search	Category.Filters
Basic S2	Y		
Basic II			Y
Intermediate	Y		Y
Advanced	Y	Y	Y

Table S3: Criterion for data analysis explorer

Data.Analysis.Explorer	Analyze.scRNAseq.Data	Anaylze.Gene.Expression	Customize.Displays
Basic	Y		
Intermediate	Y	Y	
Advanced	Y	Y	Y

Table S4: Project metadata captured in Celldepot

General.Category	Expected.Variable.Type	Description
Annotation Groups	String	categorical features from h5ad file
Cell Count	Integer	numbers of cell in study
Actions	Link	three options: 1) Study summary information; 2) Data visualization and analysis; 3) Update project information
Custom Accession	String	Customized accession name for individual project
Description	String	Additional information
DOI	Link	Digtial Object Identifier
File Name	String	h5ad file name
File Size	Integer	size of h5ad file
Gene Count	Integer	numbers of gene in study
Name	Link	project name
Notes	String	study notes
PMC ID	Link	
Publication Title		
PubMed ID	Link	
Speices	String	
URL	Link	
Year	String	

Chapter 4

Supplemental Tutorial

CellDepot is database management system integrated with management system, query searching and data visualization tools for scRNA-seq datasets, which can be accessed by the link http://celldepot.bxgenomics.com and Biogen internal link http://go.biogen.com/CellDepot. This is a supplemental tutorial providing a detailed guide.

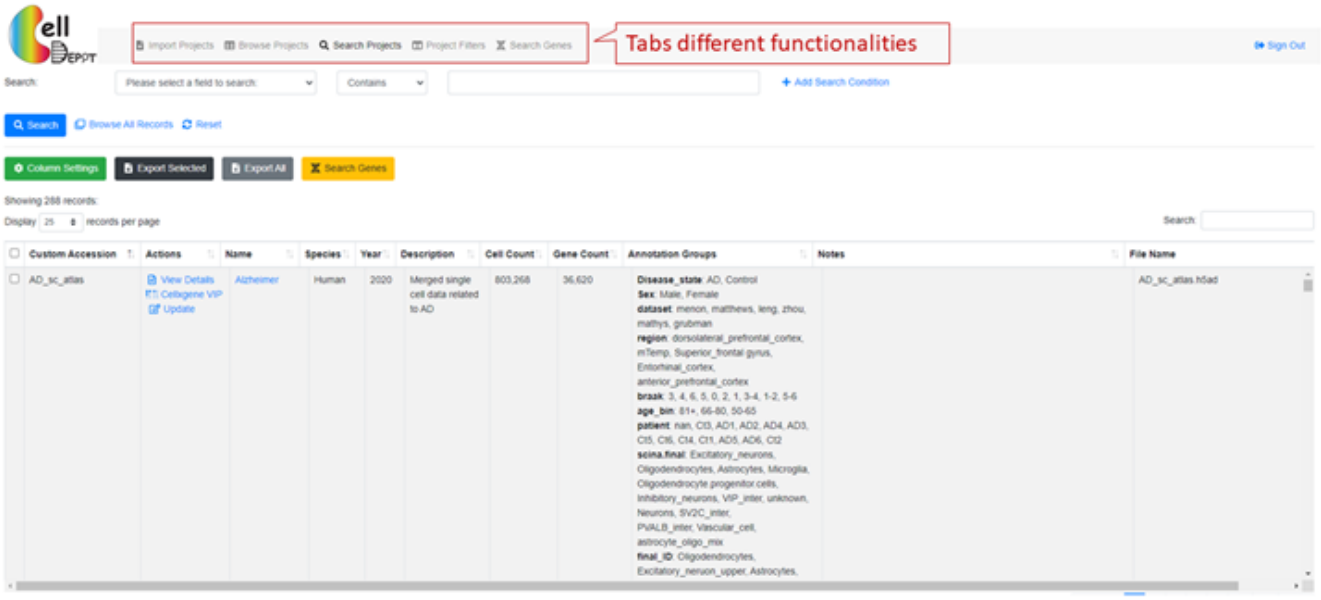

Figure 4.1: Figure S1. CellDepot website

The interface contains multiple tabs, corresponding to import and/or select objects in CellDepot scRNA-seq database, that can be accessed on top panel of the webpage. Users can upload their own dataset or explore the existing datasets for visualization and analysis.

4.1 Browse Projects

4.1.1 Search Projects

This function allows the user to search any targeted interests, which also can be accessed through search projects on the top panel of the webpage. Users is allowed to select the projects by 17 attributes: annotation groups, cell count, cellxgene VIP launch method, Custom accession, description, DOI, file name, file size, gene count, name, notes, PMC ID, Publication Title, PubMed ID, Species, URL, Year. These 17 fields can also be (partially) displayed on the webpage through 'column setting' on the webpage. Users can also search projects by the keywords via the

search function on the right of the webpage. In addition, by 'column setting', users can set up the customized layout of targeted projects; thereby exporting to csv format.

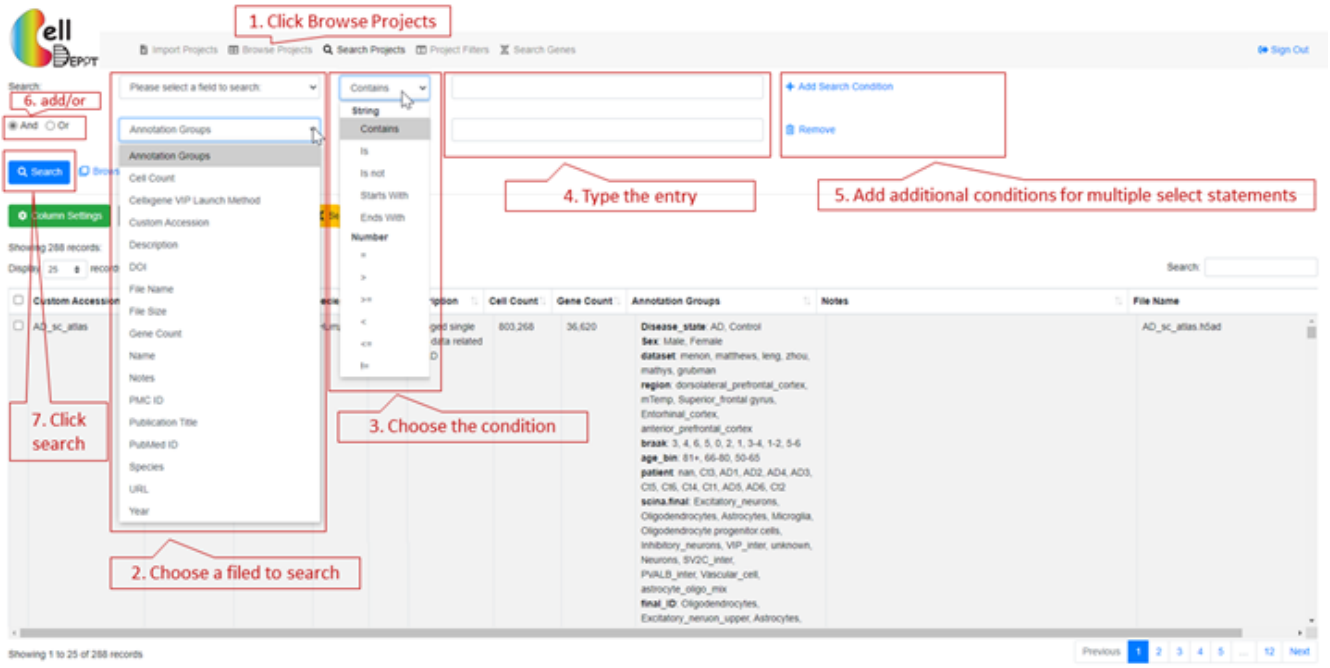

Figure 4.2: Figure S3. Workflow of how to search query on 'Browse Projects' page.

4.1.1.1 Case study 1

Six datasets are filtered when searching by 'Species is Human' and 'Annotation Group contains Neuron'.

4.1.1.2 Case Study 2

Cross-project comparison of skeletal muscle marker genes PAX3, PAX7, PITX2, MYF5, MYF6, MYOD1, MYOG, NEB, and MYH3 among the datasets whose species is human and cell type is myogenic.

For each project, users can view the datasets information, visualize data analysis, and conduct update through clicking on "View Details", "cellxgene VIP", and "Update" links, respectively.

4.1.2 Project Filters

This page provides the matched datasets by simply clicking the categories. It is a first-time user-friendly functionality as users may not be familiar with the content of the database. The advance search function is the same as that on the 'Browse Projects' page (4.1.1).

4.2 Visaulize Datasets

4.2.1 View Details

The datasets information consists of project summary and annotation groups. The project summary is provided by each user when uploading projects. The information of annotation groups is retrieved from uploaded .h5ad file.

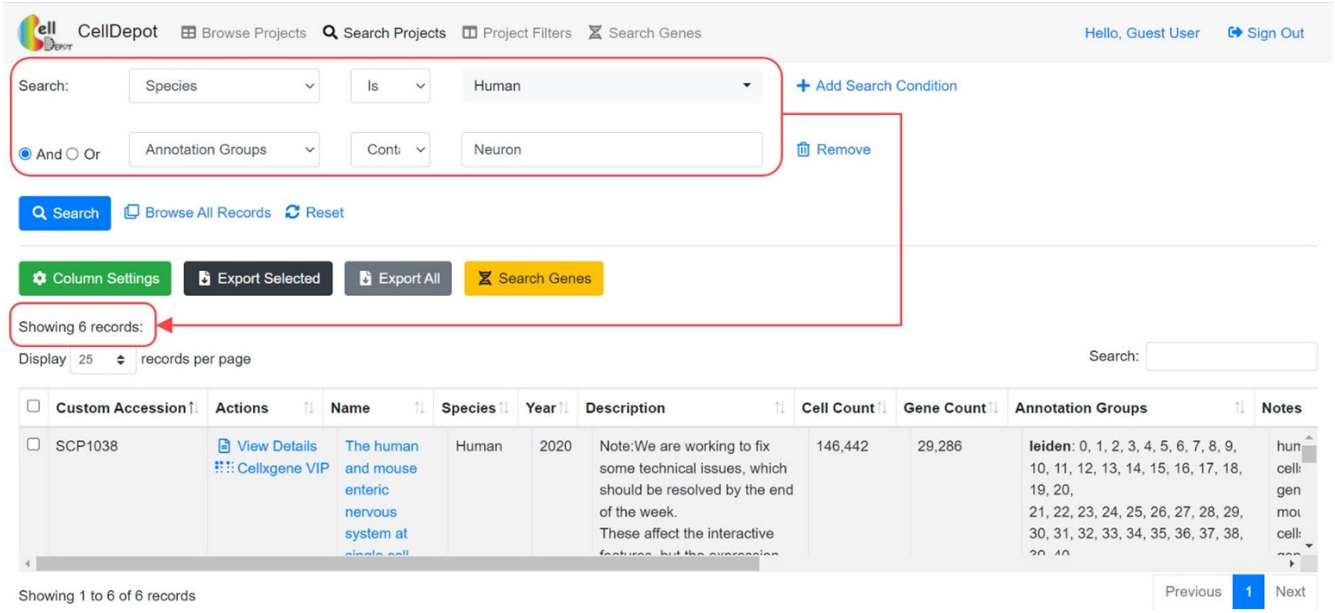

Figure 4.3: Figure S4. Data Filtering. Query search of 'Species is Human' and 'Annotation Groups contains Neuron' brings about nine datasets of interest.

4.2.2 Update

Project summary information can be updated on 'Browse Project' page with 'Preload in Memory' cellxgene VIP launch method via click 'Update' hyperlink.

4.2.3 Data Visualization and Analysis

CellDepot is not only a database management system, but also a web portal to visualize the scRNA-seq dataset. Here, we embed cellxgene and cellxgene VIP in CellDepot. By clicking 'Cellxgene VIP', data analysis results can be visualized. Detailed guides of cellxgene and cellxgene VIP, please go to https://interactivereport.github.io/cellxgene_VIP/tutorial/docs/.

4.2.3.1 Case study 1

Exploration and visualization of the expression of gene(s) across the cluster of cells under various conditions.

As shown in Figure S10a, two cell groups from Astrocytes (1036 cells) and Oligodendrocytes (4417 cells) are selected. By running differential analysis with one of the built-in statistical methods such as Welch's t-test, we detected 1578 differential expressed genes (DEGs), including 715 up-regulated and 853 down-regulated genes in astrocytes compared to oligodendrocytes (Figure S10a). The expression of the top four DEGs among the cell types indicates that gene MBP, ST18 and RNF220 are expressed explicitly in oligodendrocytes, while gene PITPNC3 is expressed mainly in astrocytes and endothelial cells (Figure S10b). In the future, we plan to add other multi-omics data modalities, which can be incorporated and integrated with scRNA-seq, such as spatial transcriptomics and scATAC-seq data.

4.3 Search Genes

This tab allows searching on targeted genes with cell count cutoff and expression cutoff. The search outcome provides users every project contains the targeted genes. Each project displays a link to project page and a figure plot if applicable. This plot can be either violin plot or dot plot shows the gene expression level in each annotation groups.

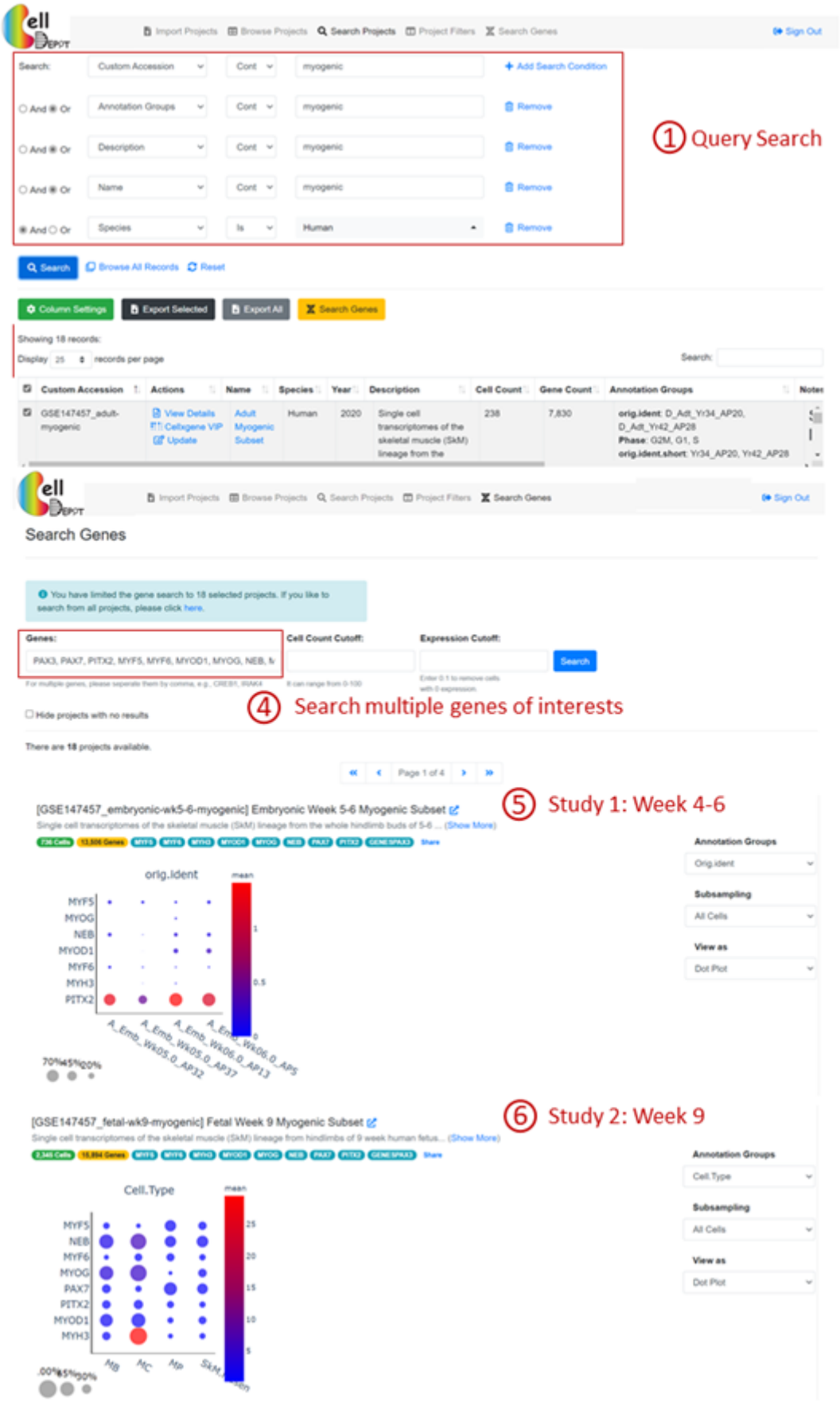

Figure 4.4: Figure S5. Workflow of how to conduct the cross-project comparison of gene sets among the selected datasets.

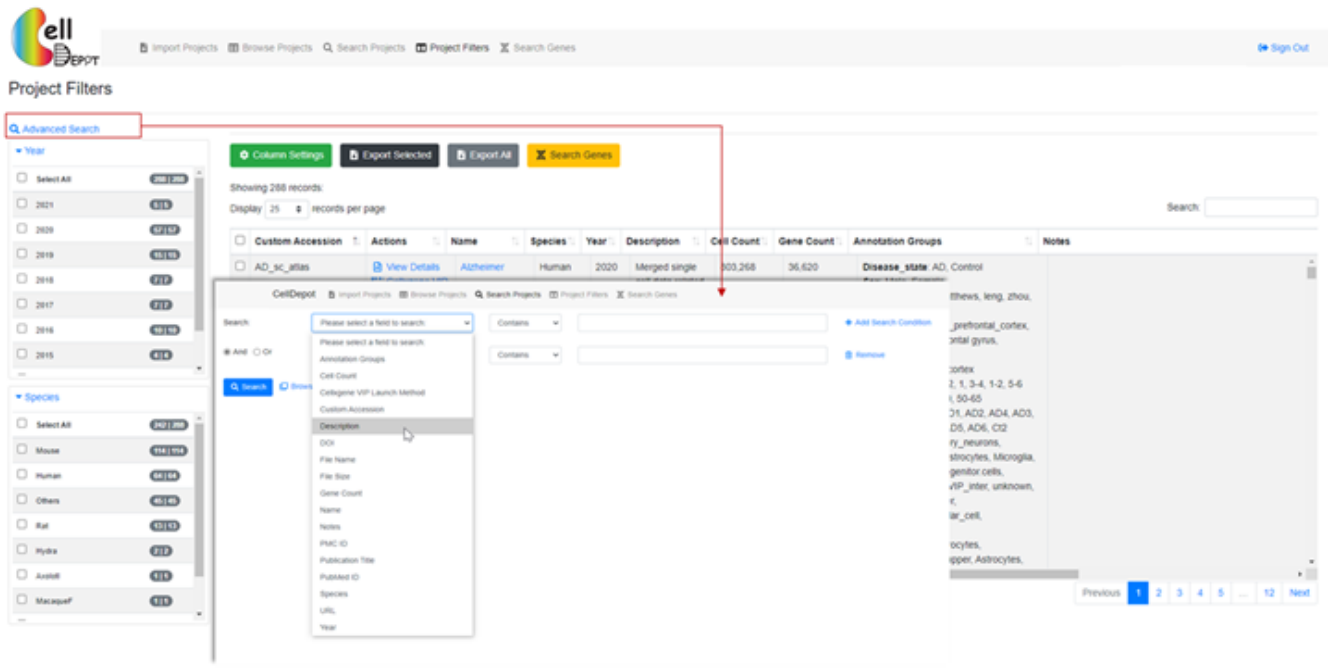

Figure 4.5: Figure S6. The layout of 'Project Filters' page.

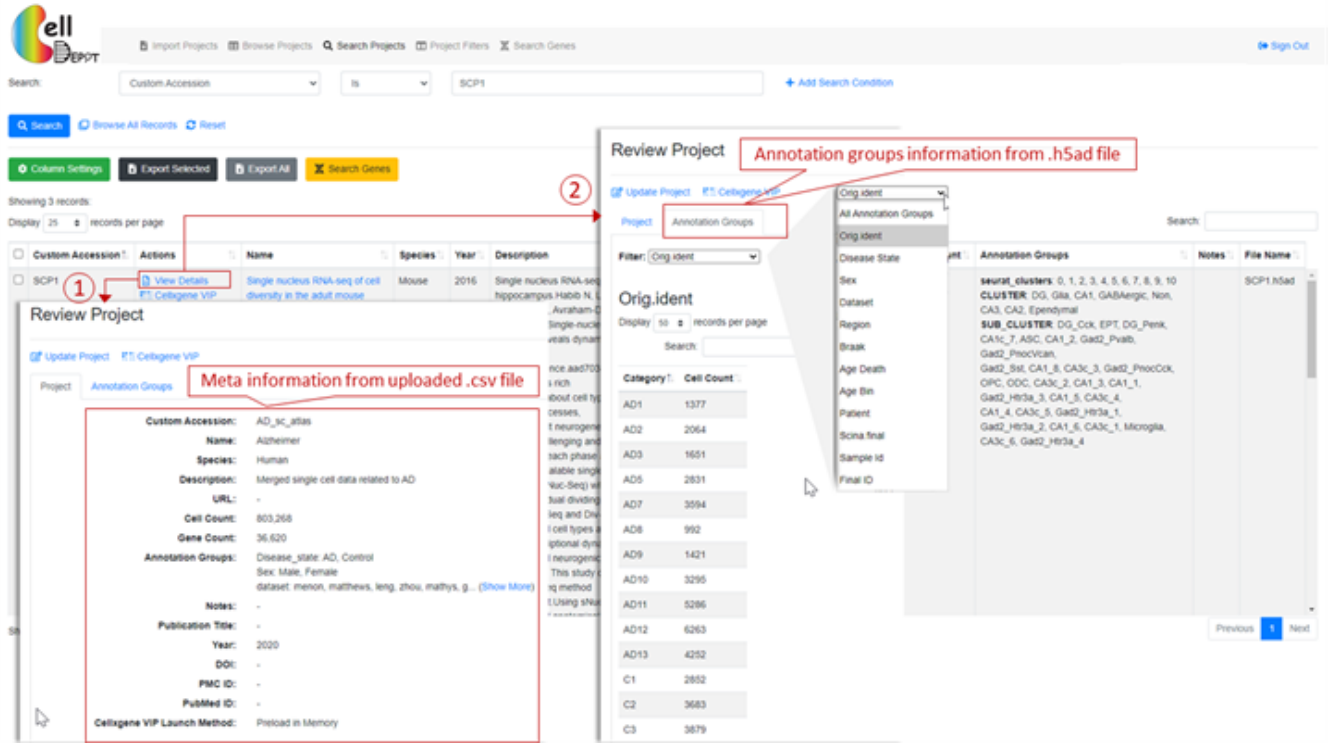

Figure 4.6: Figure S7. Visualization of details of datasets.

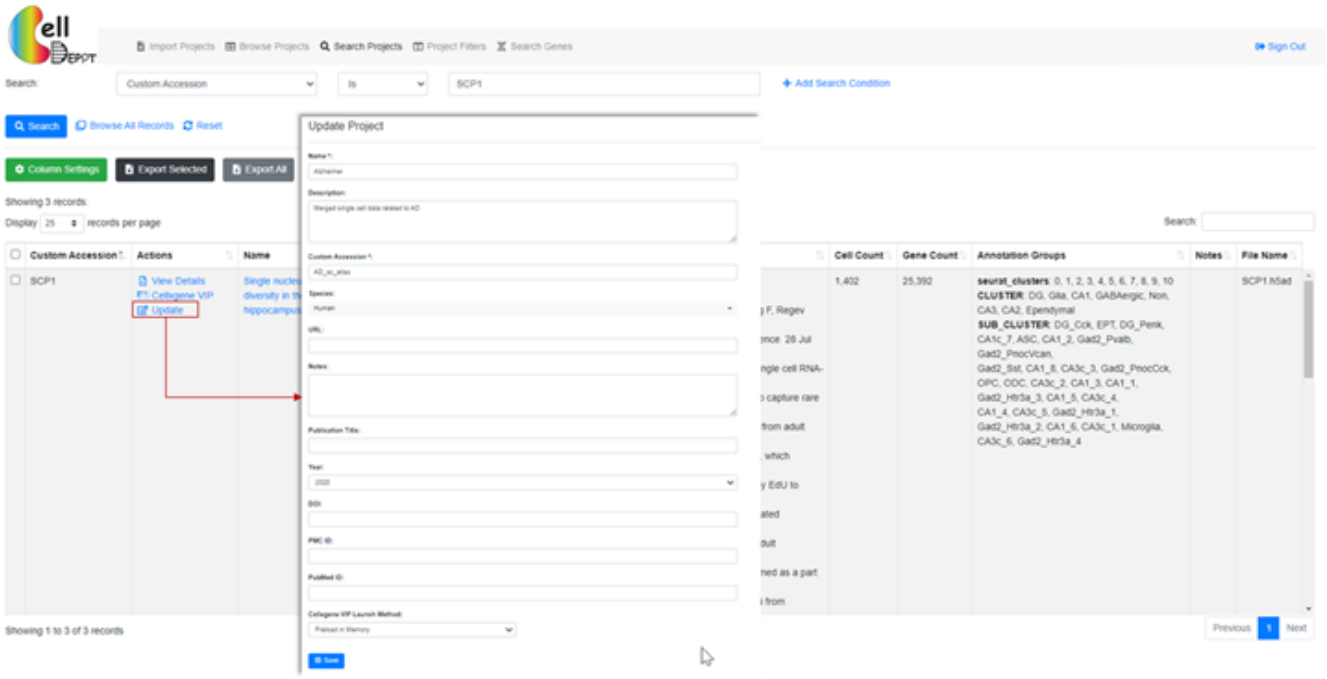

Figure 4.7: Figure S8. Update project on 'Browse Project' page.

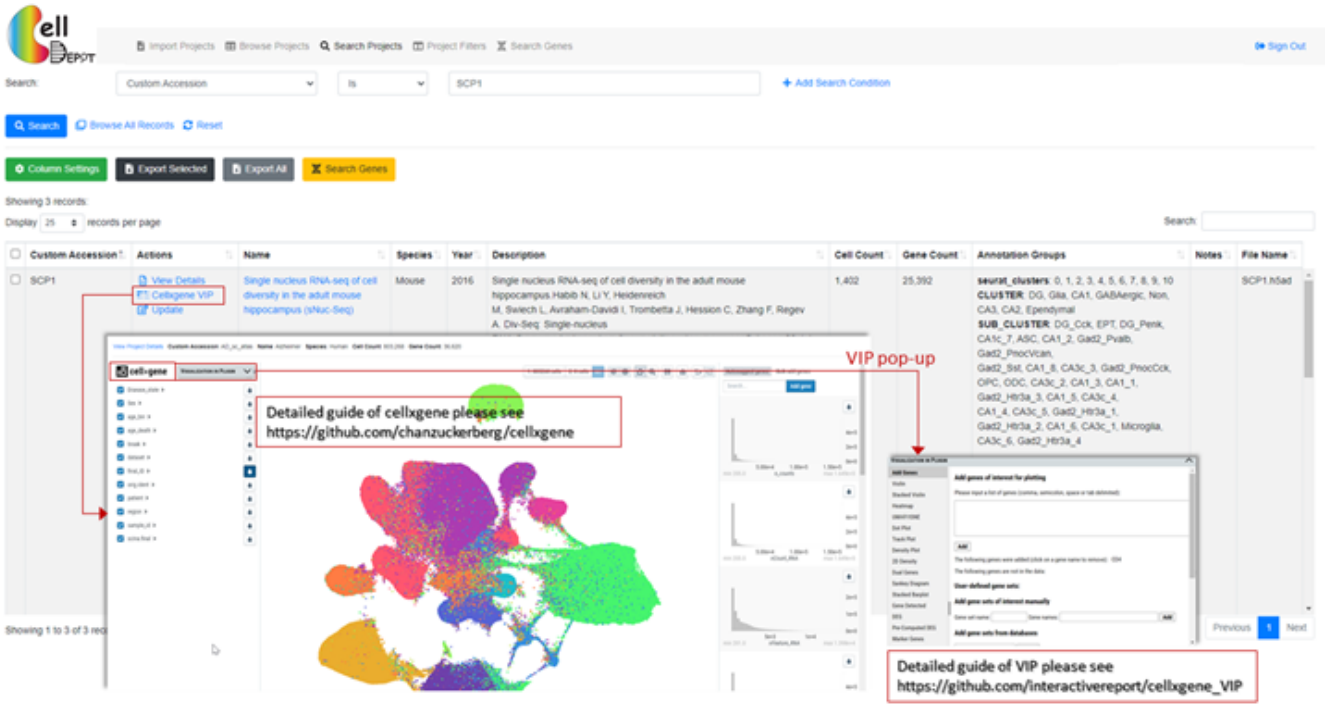

Figure 4.8: Figure S9. See visualization of selected datasets.

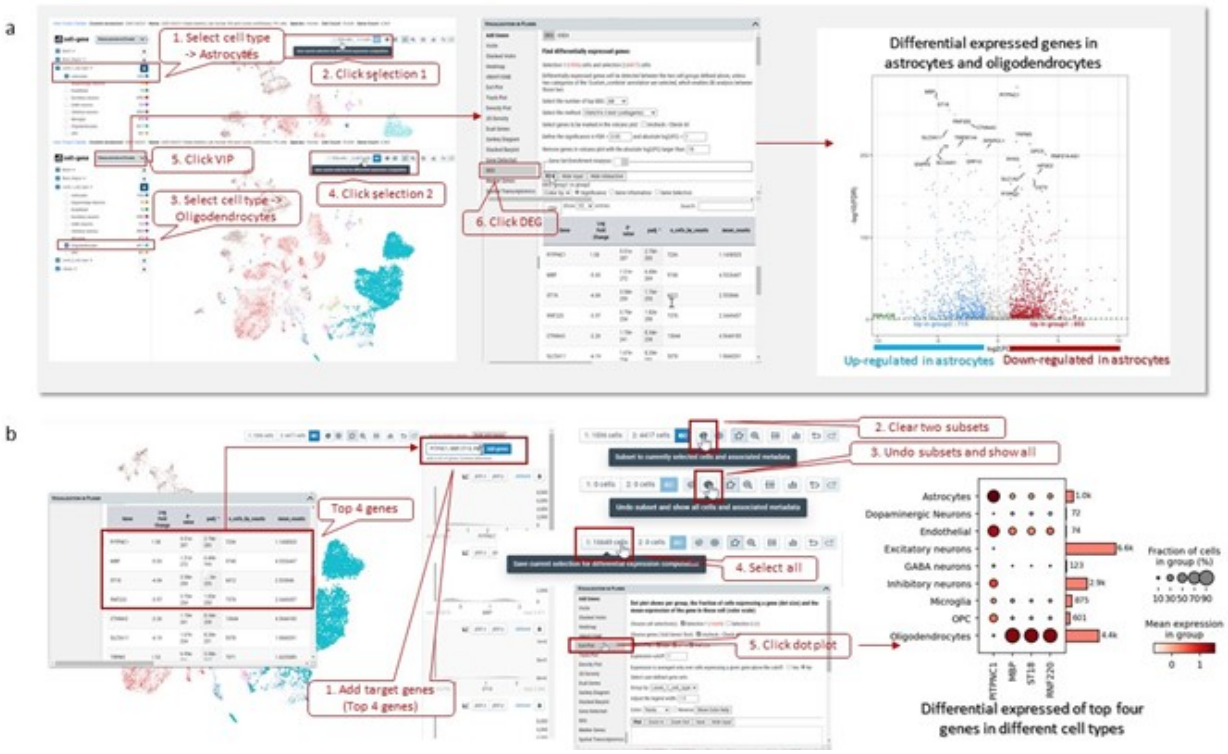

Figure 4.9: Figure S10. Exploration of differential expressed genes in dataset GSE140231 through cellxgene VIP. (a) Differential expressed genes in astrocytes and oligodendrocytes. (b) The expression of top four genes in different cell types.

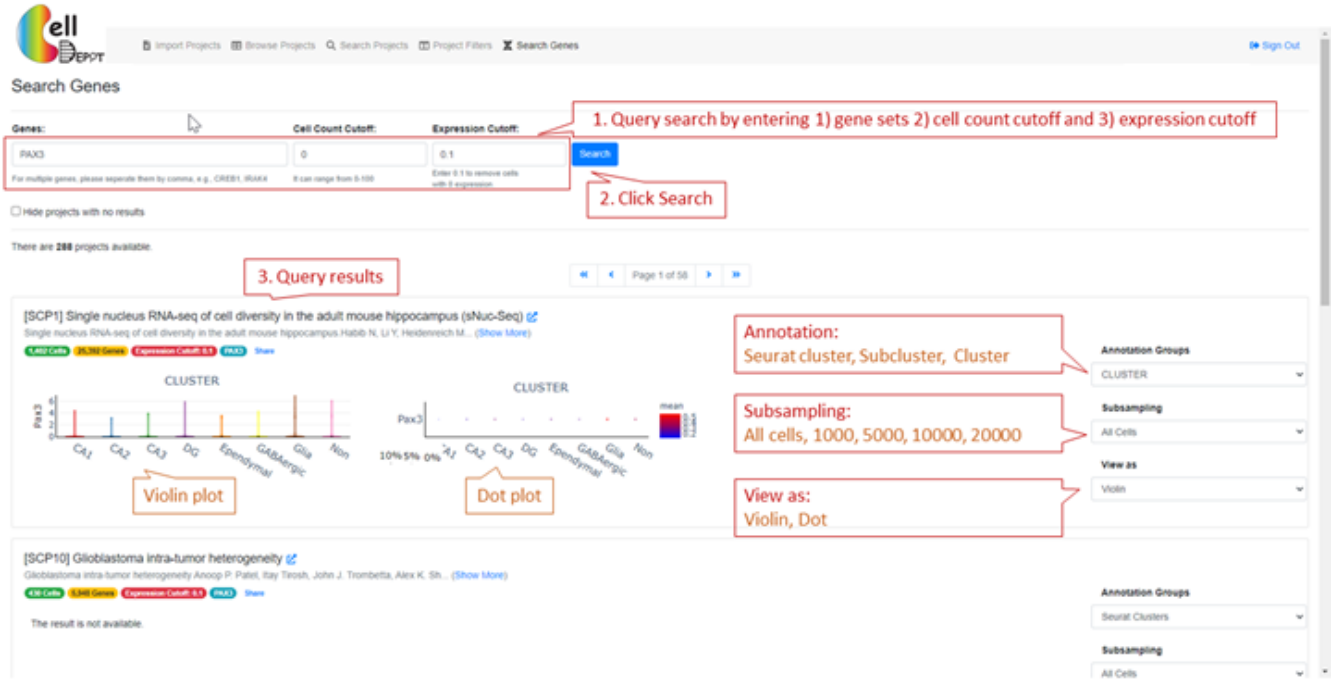

Figure 4.10: Figure S11. The layout of 'Search Genes'

4.4 Upload Projects

To upload new projects in CellDepot database, two files are required: 1) .h5ad files and 2) project information in csv format. Detailed formatting guidance can be found by 'Download Example File' hyperlink on webpage. In addition, two cellxgene VIP launch methods are provided: standard and preload in memory. Standard mode is for the first-time imported datasets, while preload in memory should be chosen when users update the meta information of datasets. After the projects are submitted, CellDepot will automatically analyze the datasets. To explore the detail of uploaded datasets, users can navigate to 'browse projects' page and then search the imported datasets by the customized accession number.

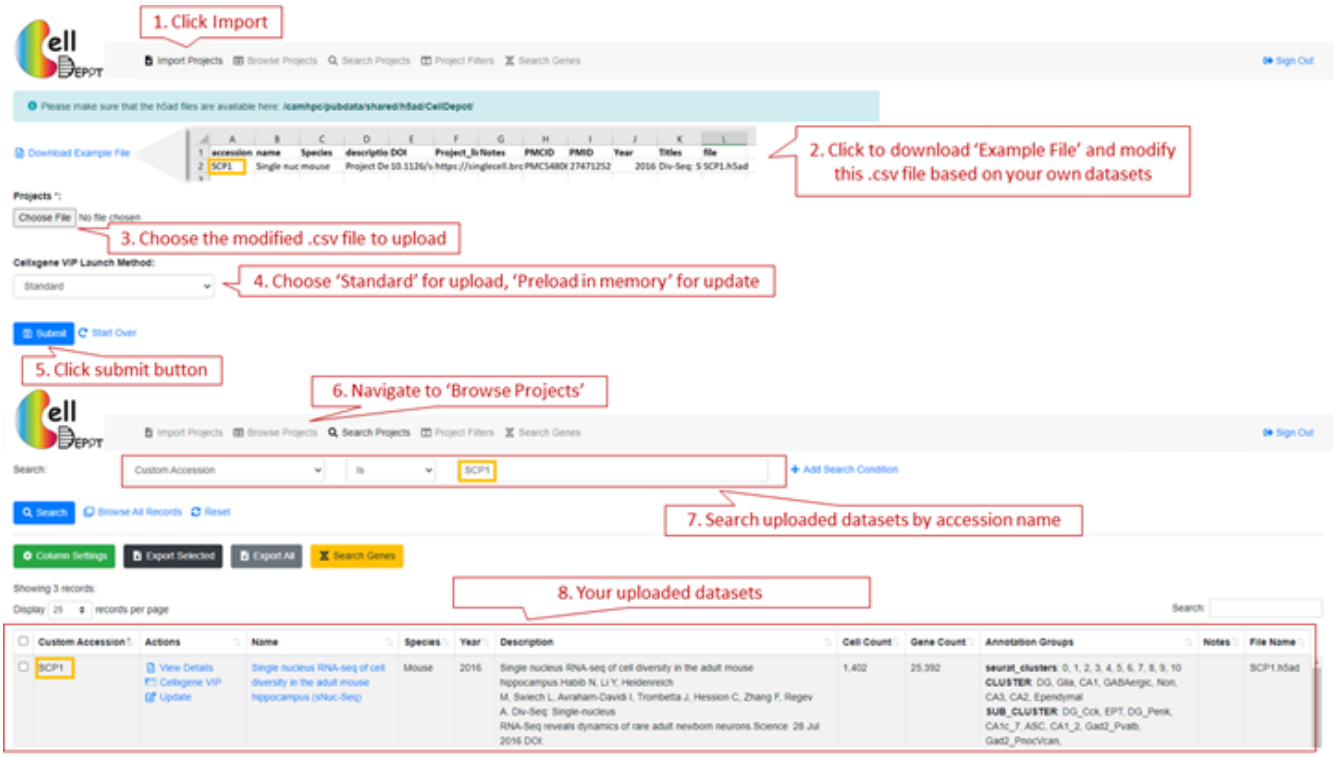

Figure 4.11: Figure S2. Workflow of how to import personal datasets.

4.5 How to set up cron job?

The following cron job entry is needed to convert h5ad file to CSC format on the background,

```
@hourly <user-name> cd /var/www/html/celldepot/app/core; php ./api_toCSCh5ad.php
```

Please make sure that the user has the permission to write in the data directory.