CellDepot: A unified repository for scRNA-seq data and visual exploration

Authors: Dongdong Lin[1†], Yirui Chen[1†], Soumya Negi[1], Derrick Cheng[2], Zhengyu Ouyang[2], David Sexton[1], Kejie Li[1], Baohong Zhang[1*]

Affiliations:

[1]Research Department, Biogen, Inc., 225 Binney St, Cambridge, MA 02142, USA

[2]BioInfoRx, Inc., 510 Charmany Dr, Suite 275A, Madison, WI 53719, USA

Emails:

DL: dongdong.lin@biogen.com

YC: yirui.chen@biogen.com

SN: soumya.negi@biogen.com

DC: derrick@bioinforx.com

ZO: oyoung@bioinforx.com

DS: david.sexton@biogen.com

KL: kejie.li@biogen.com

BZ: baohong.zhang@biogen.com

† These authors contributed equally to this work

* Correspondence: Baohong Zhang, baohong.zhang@biogen.com

Abstract

CellDepot serves as an integrated web application to assist users in exploring single-cell RNA-seq (scRNA-seq) datasets and comparing the datasets among various studies through a user-friendly interface with advanced visualization and analytical tools. To begin with, it provides an efficient data management system that users can upload single cell datasets and query the database by multiple attributes such as species and cell types. In addition, the advanced query function incorporated in MySQL database system and its conditional filtering, allows users to quickly query and compare the expression of gene(s) across the datasets of interest. Moreover, by embedding the cellxgene VIP tool, CellDepot enables fast exploration of individual dataset in the manner of interactivity and scalability to gain more refined insights such as cell composition, gene expression profiles, and differentially expressed genes among cell types. In summary, the web portal allows large scale single cell data sharing, analysis and visualization for supporting decision-making, and encouraging scientists to contribute to the single-cell community in a tractable and collaborative way. Finally, CellDepot is released as open-source software to motivate crowd contribution, broad adoption, and local deployment for private data.

Keywords

Single cell, Single-cell RNAseq, cellxgene, data visualization, scRNA-seq portal

Introduction

Since the first print of using next-generation sequencing technology to analyze the single-cell transcriptome in 2009 [1], this technology has massively ignited the interest in obtaining high-resolution characterizations of each cell's transcriptome. Afterwards, an exponentially growing number of studies utilized this advanced scRNA-seq technology to investigate the expression changes among cells or groups from various sources such as cell lines, tissues, or species [2-5]. Meanwhile, substantial efforts have been invested in developing computational pipelines and tools to advance the analysis and visualization of large-scale scRNA-seq data [6-10]. With more scRNA-seq datasets have been published by community, it is urgent to launch some in-depth investigations of the associations of those identified biological variations with cellular heterogeneity. However, many datasets are from different studies or individual labs that are preprocessed in varied ways. Therefore, it is necessary to build a centralized space for managing, analyzing, and visualizing those datasets to yield comprehensive insights from this big data.

Several tools and databases have been built to store, analyze, and visualize scRNA-seq datasets, which allow scientists to compile and query the information at their hand. The 'Human Cell Atlas' consortium [11] leads an international collaboration to generates single-cell datasets from human body tissues, which are manually curated and processed by a uniform pipeline. 'JingleBells' provides single-cell data [12] focusing on immune cells. The 'conquer' database provides uniformly processed single-cell expression data to facilitate benchmarking of computational tools [13]. The 'PanglaoDB' database provides single-cell RNA-seq count matrices from public sequencing data in the National Center for Biotechnology Information Sequence Read Archive [14]. The 'EMBL-EBI Single-Cell Expression Atlas' provides uniformly processed data from submissions to 'ArrayExpress' [15]. The Broad Institute also offers a public 'Single-Cell Portal' (https://singlecell.broadinstitute.org/single_cell). Although those 'atlas' and portals have greatly facilitated the exploration of single-cell datasets from the database [16-20], they usually tackle certain aspects of managing, visualizing, or analyzing scRNA-seq datasets with limitations on both query and visualization functionalities. For comprehensive review of these portals, detailed feature comparison is outlined in supplementary Table S1.

We developed a database-backed single cell portal available at http://celldepot.bxgenomics.com with user-friendly interactive visualization and analysis capabilities, namely CellDepot to empower biologists and bioinformaticians to manage, explore, visualize, and compare scRNA-seq datasets in a comprehensive, flexible, and collaborative manner.

Results and Discussion

1. Overview of CellDepot

CellDepot is a user-friendly centralized database platform that integrates the database management system along with the search engine. The web-based platform enables the share of scRNA-seq data and efficient communication among the community, which in turn encourages crowd contributions to the online scRNA-seq data portal. Currently, there are more than two hundred scRNA-seq datasets from eight species hosted on the portal for public access. Notably, CellDepot integrates with advanced single-cell transcriptomic data explorer to conduct all analytical tasks on the webserver while presenting interactive results on the webpage through leveraging modern web development techniques (Figure 1a).

Highly unified space for managing, exploring, and analyzing comprehensive single-cell datasets is one of the significant strengths of CellDepot. Nowadays, while a few web-based portals for both visualizing and analyzing scRNA-seq datasets are available, they have limitations on scalability and capability in comparing among various datasets to meet increasing demands on examining single-cell RNA-seq datasets. CellDepot employs the MySQL relational database management system, a robust scRNA-seq data visualization tool cellxgene [21], and its versatile plugin cellxgene VIP [22] to allow project selection and filtering, scRNA-seq data visualization and analysis, and cross-project comparison of targeted genes as illustrated Figure 1b.

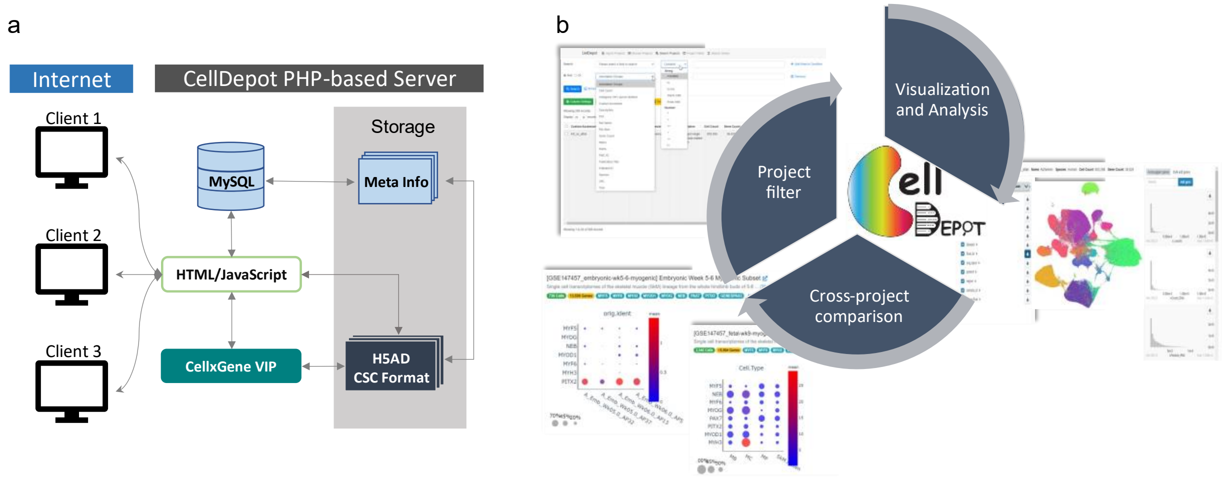


Figure 1. CellDepot portal overview. (a) The architecture of CellDepot. (b) Functional structure of CellDepot.

2. Features of CellDepot

There are mainly three functions built in the platform: project search/filter, visualization and analysis, and cross-project comparison, which will be discussed thoroughly in the following sections.

a. Data filtering

scRNA-seq datasets are stored as h5ad files in the compressed sparse column (CSC) format and managed efficiently in the CellDepot together with the metadata table of projects. For each dataset, some primary metadata fields are inputted by users and can be updated as needed. The main CellDepot application consists of five functionalities:

- **Importing projects** provides a user-friendly way to upload processed public or private datasets in h5ad format.
- **Browsing projects** displays a quick overview of all datasets in CellDepot with a quick search function and customized column setting such that users can easily find the projects of interest, personalize the look-n-feel of the project table and export the customized project spreadsheet.

- **Searching projects** enables advanced query of projects by joining multiple search conditions with the logical operators.
- **Project filters** refines the matched datasets by simply selecting ‘Year’ and/or ‘Species’. It is a user-friendly feature for a first-time user who is not familiar with the content of the database.
- **Searching genes** allows users to input any gene(s) of interest and generate the cross-dataset plots for those targeted genes.

Based on available meta information stored in the h5ad files, multiple dataset attributes are extracted for searching and filtering, such as species, cell type annotation, published year, etc. As shown in Figure S1, there are six datasets when searching by ‘Species is Human’ and ‘Annotation Group contains Neuron’.

It is desirable to include all possible abbreviations and synonyms of genes or cell types in the database to make search more flexible. In reality, CellDepot has not enumerated and standardized all possible terms used to name cell types in the query function beyond metadata collected from h5ad files. In the future, a comprehensive mapping table of terms based on cell ontology could be implemented to standardize keywords for searching. Alternatively, user can use advanced search function to query a cell type by multiple synonyms, e.g., searching by ‘Annotation Groups contains OPC’ or ‘Annotation Groups contains oligodendrocyte’ to identify datasets containing oligodendrocyte cell type.

b. Data visualization and analysis

Users can launch a cellxgene VIP instance (such as dataset GSE140231) [23] as shown in Figure 2 to explore a dataset interactively. Cellxgene VIP is based on the cellxgene platform where users can quickly overview meta information and cell embedding after dimensional reduction (e.g., TSNE or UMAP). In addition, users can use lasso selection from embedded cell maps to interactively select any cell cluster(s) as a group for refined analysis via a rich set of visualization and analysis functions provided in cellxgene VIP.

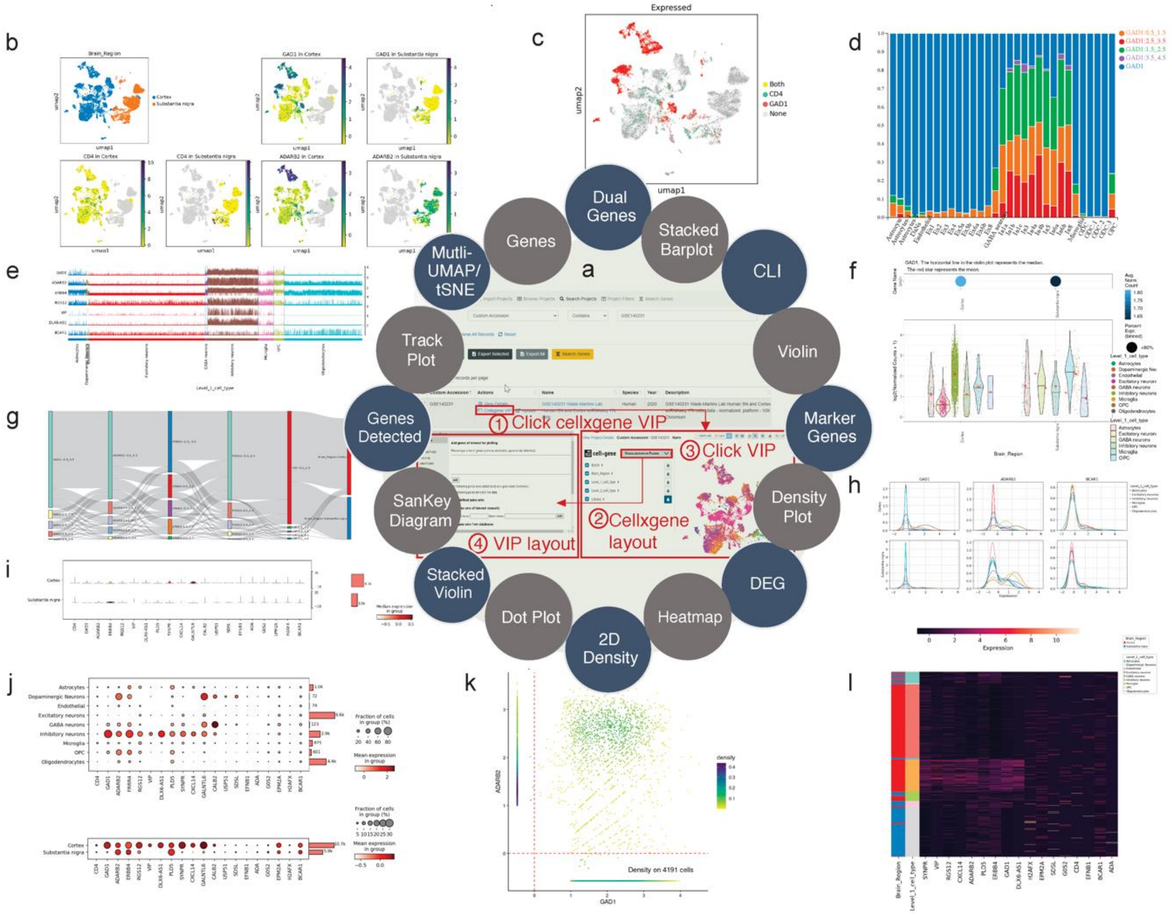


Figure 2. Exploration of visualization and analysis functions in cellxgene VIP on dataset GSE140231 [23]. (a) **Cellxgene VIP**, providing a plugin ecosystem of interactive data visualization, can be launched within four steps after querying ‘custom accession contains GSE140231’. (b) **Mutli-tSNE/UMAP plot** shows the differential expression of genes on split brain regions. (c) **Dual-gene plot** demonstrates the expression of CD4 and GAD1 on the selected embedding layout. (d) **Stacked barplot** indicates the fraction of cell distribution of the selected gene over different cell types. (e) **Trackplot** represents the distribution of gene expression across individual cells in annotated clusters. (f) **Sub-grouped violin plot** shows the GAD1 gene expression across the groups of cell types and subgroups of the brain regions. (g) **Sankey diagram** illustrates the inter-dependent relationship of annotated clusters based on the selected genes. (h) **Density plots** show expression of marker gene across varied cell types and split across the brain regions. (i) **Stacked violin plot** highlights the selected cell markers over cell types. (j) **Dot plots** highlights the gene markers over different annotated clusters. (k) **2D-density plot** illustrates the expression relationship of two genes (GAD1 and ADARB2). (l) **Heatmap** of selected marker genes in various cell types.

For example, user can go deep-dive to explore and visualize the expression of gene(s) across the cluster of cells under various conditions (Figure S2). As shown in Figure S2a, 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 S2a). 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 S2b). 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.

c. Cross-dataset query

Besides the function to visualize and explore individual datasets, CellDepot also allows users to query and compare the expression of genes of interest across multiple datasets to understand their cellular heterogeneity. As shown in Figure 3, users can explore the expression of skeletal muscle marker genes PAX3, PAX7, PITX2, MYF5, MYF6, MYOD1, MYOG, NEB, and MYH3 during human myogenic cells of development and differentiation from hx-protocol 4-6 weeks to fetal 9-18 weeks from datasets whose custom accession starting with GSE147457 [24]. Gene expression level in SkM cells highlighted by grey boxes in of Figure 3e decreases around 40% from week 4 to week 5 with hx-protocol until level-off at week 6. The expression level of genes in fetal stage over week 9 also shows the same trend, especially the gene expression level drops dramatically starting from week 12.

In summary, CellDepot is an easy-to-use and elaborative web portal for the exploration of the scRNA-seq datasets and data analysis results such that it allows biologists to easily access and reuse the rapid-increased scRNA-seq datasets in a highly scalable and interactive manner.

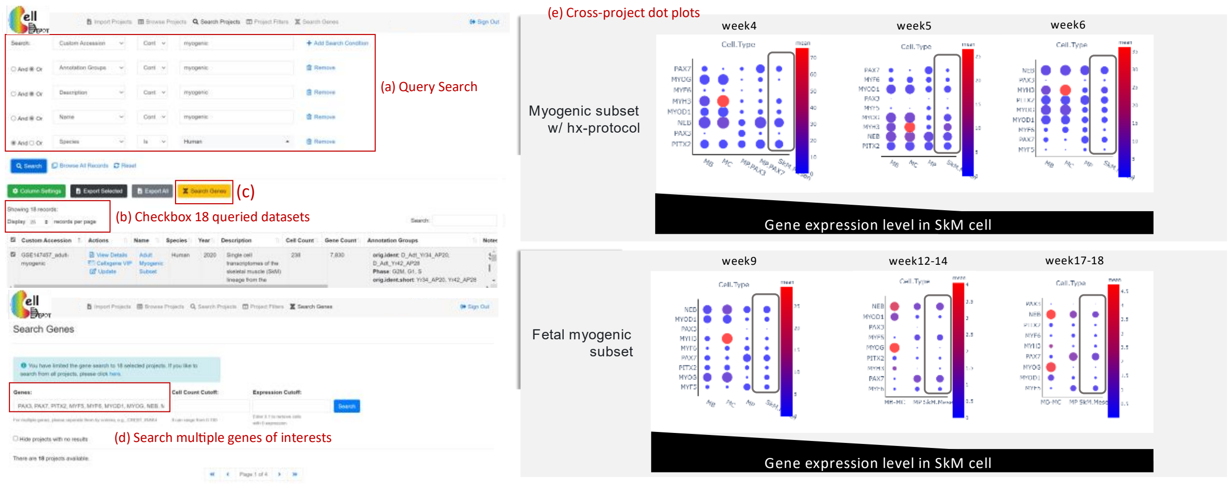


Figure 3. Cross-project view of the expression of genes among skeletal muscle development and differentiation times. (a) Search and (b) select the targeted datasets whose custom accession starting with GSE147457 [24]. Navigate to (c) 'Search Genes' page and (d) search genes of interest. (e) The summary of interactive plots from CellDepot that shows the cross-project comparison of gene expression level under varied conditions.

Materials and Methods

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 [6]. When importing the dataset to the system, user inputs additional metadata information as shown in Figure S5. Both metadata are collected and stored in a MySQL database table that is presented at http://celldepot.bxgenomics.com.

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 (Figure S3).

CellDepot platform and installation

The public version of CellDepot web portal is hosted at the web site, http://celldepot.bxgenomics.com and 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.

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 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 (Supplementary Tutorial Section 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.

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.

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.

Author Contributions

Dongdong Lin: Conceptualization, Methodology, Manuscript – writing & review

Yirui Chen: Testing, Manuscript – writing, review & editing

Kejie Li: Conceptualization, Testing, Manuscript – review & editing

Soumya Negi: Methodology, Testing, Manuscript – review & editing

Derrick Cheng: Web application development, Manuscript - review

Zhengyu Ouyang: Data curation, Manuscript - review

David Sexton: Conceptualization, Manuscript - review

Baohong Zhang: Conceptualization, Testing, Methodology, Manuscript – writing, review & editing

Declaration of competing interest

DL, KL, SN, DS and BZ hold Biogen stocks as Biogen employee.

Acknowledgements

The authors are grateful and indebted to BioInfoRx, Inc. for the management of server and storage to host the public web site.

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al., (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-82.

2. Katzenelenbogen Y, Sheban F, Yalin A, Yofe I, Svetlichnyy D, Jaitin DA, et al., (2020). Coupled scRNA-Seq and Intracellular Protein Activity Reveal an Immunosuppressive Role of TREM2 in Cancer. *Cell* **182**, 872-85 e19.

3. Madissoon E, Wilbrey-Clark A, Miragaia RJ, Saeb-Parsy K, Mahbubani KT, Georgakopoulos N, et al., (2019). scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol* **21**, 1.

4. Wang Z, Xu X. (2020). scRNA-seq Profiling of Human Testes Reveals the Presence of the ACE2 Receptor, A Target for SARS-CoV-2 Infection in Spermatogonia, Leydig and Sertoli Cells. *Cells* **9**.

5. Qi J, Zhou Y, Hua J, Zhang L, Bian J, Liu B, et al., (2021). The scRNA-seq Expression Profiling of the Receptor ACE2 and the Cellular Protease TMPRSS2 Reveals Human Organs Susceptible to SARS-CoV-2 Infection. *Int J Environ Res Public Health* **18**.

6. Wolf FA, Angerer P, Theis FJ. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15.

7. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502.

8. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-20.

9. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al., (2019). Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-902 e21.

10. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al., (2021). Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-87 e29.

11. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al., (2017). The Human Cell Atlas. *Elife* **6**.

12. Ner-Gaon H, Melchior A, Golan N, Ben-Haim Y, Shay T. (2017). JingleBells: A Repository of Immune-Related Single-Cell RNA-Sequencing Datasets. *J Immunol* **198**, 3375-9.

13. Soneson C, Robinson MD. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* **15**, 255-61.

14. Franzen O, Gan LM, Bjorkegren JLM. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* **2019**.

15. Papatheodorou I, Moreno P, Manning J, Fuentes AM, George N, Fexova S, et al., (2020). Expression Atlas update: from tissues to single cells. *Nucleic Acids Res* **48**, D77-D83.

16. Bernstein MN, Ni Z, Collins M, Burkard ME, Kendziorski C, Stewart R. (2020). CHARTS: A web application for characterizing and comparing tumor subpopulations in publicly available single-cell RNA-seq datasets. 2020.09.23.310441.

17. Orvis J, Gottfried B, Kancherla J, Adkins RS, Song Y, Dror AA, et al., (2021). gEAR: Gene Expression Analysis Resource portal for community-driven, multi-omic data exploration. *Nat Methods* **18**, 843-4.

18. Cai G, Xiao F. (2020). SCANNER: A Web Server for Annotation, Visualization and Sharing of Single Cell RNA-seq Data, *bioRxiv* 2020.01.25.919712 [Preprint], doi: https://doi.org/10.1101/2020.01.25.919712.

19. Darde TA, Lecluze E, Lardenois A, Stevant I, Alary N, Tuttelmann F, et al., (2019). The ReproGenomics Viewer: a multi-omics and cross-species resource compatible with single-cell studies for the reproductive science community. *Bioinformatics* **35**, 3133-9.

20. Cao Y, Zhu J, Han G, Jia P, Zhao Z. (2017). scRNASeqDB: a database for gene expression profiling in human single cell by RNA-seq. 104810.

21. Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, et al., (2021). cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices, *bioRxiv* 2021.04.05.438318 [Preprint], doi: https://doi.org/10.1101/2021.04.05.438318

22. Li K, Ouyang Z, Lin D, Mingueneau M, Chen W, Sexton D, et al., (2020). cellxgene VIP unleashes full power of interactive visualization, plotting and analysis of scRNA-seq data in the scale of millions of cells, *bioRxiv* 2020.08.28.270652 [Preprint], doi: https://doi.org/10.1101/2020.08.28.270652.

23. Agarwal D, Sandor C, Volpato V, Caffrey TM, Monzon-Sandoval J, Bowden R, et al., (2020). A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat Commun* **11**, 4183.

24. Xi H, Langerman J, Sabri S, Chien P, Young CS, Younesi S, et al., (2020). A Human Skeletal Muscle Atlas Identifies the Trajectories of Stem and Progenitor Cells across Development and from Human Pluripotent Stem Cells. *Cell Stem Cell* **27**, 181-5.

309 Supplementary Information

310 Supplementary Tables

311 Table S1. Comparison matrix of web portal tools

Web application repository	**CellDepot**	**Corpora Data Portal**	**gEAR**	**CHARTS**	**SCANNER**	**Single Cell Portal**	**Sfaria**	**Repro-Genomics**	**PanglaoDB**	**Expression Atlas**	**scRNA-SeqDB**	**conquer**	**Jingle-Bells**	**Human Cell ATLAS**
Year	2021	2021	2021	2020	2020	2020	2020	2019	2019	2019	2019	2018	2017	2017
Main function														
Database Explorer	Y	Y	Y	Y	NA	Y	Y	Y	Y	Y	Y	Y	Y	Y
Query Search	Advanced	NA	Basic I	NA	Intermediate	Advanced	Basic II	Intermediate	Advanced	Intermediate	Basic I	Basic I	Basic I	Advanced
Data Analysis Explorer	Advanced	Advanced	Advanced	Intermediate	Intermediate	Intermediate - Advanced	NA	Basic	Intermediate	Intermediate	Basic - Intermediate	Basic	NA	NA
Visualization features	Violin, dot, heatmap, bar, QC, scatter, density, embedding plots	Bar, scatter, embedding plots	Violin, dot, line, bar, QC, embedding plots, genome browser	Bar, embedding plots	Violin, dot, scatter, heatmap, embedding plots	Violin, scatter, heatmap, embedding plots	NA	Violin, density, scatter plots, genome browser	Bar, QC, scatter, embedding plots	Heatmap, scatter, embedding plots	Scatter, bar plots	Scatter, QC plots	Genome browser	NA
Data type supported														
Datasets	diverse	diverse	hearing/ brain	tumor	diverse	diverse	diverse	reproduction	diverse	diverse	diverse	diverse	immune -related	diverse
Datatypes	Single-cell	Single-cell	Single-cell, Epigenetics	Single-cell	Single-cell	Single-cell	Single-cell	Single-cell, Multi-omics	Single-cell	Single-cell, Protemics	Single-cell	Single-cell	Single-cell	Single-cell, Multi-omics
Links														
Source Code link	https://github.com/interactivereport/CellDepot	https://github.com/chanzuckerberg/corpora-data-portal	https://github.com/IGS/gEAR	https://github.com/stewart-lab/CHARTS	https://github.com/GuoshuaiCai/scanner	https://github.com/broadinstitute/single_cell_portal_core	https://github.com/theislab/sfaira-portal	https://github.com/fchalmel/RGV	https://github.com/oscar-franzen/PanglaoDB	https://github.com/ebi-gene-expression-group/atlas	NA	https://github.com/csoneson/conquer_comparison	https://drive.google.com/drive/folders/0BxSFjdiDhUI1amNoSks0SmpMdE0?resourcekey=0-s7Qw02gR8VhkS7hfdE9nPg	https://github.com/HumanCellAtlas/
Demo link	http://celldepot.bxgenomics.com/	https://cellxgene.cziscience.com/	umgear.org	https://charts.morgridge.org	https://www.thecailab.com/scanner/	https://singlecell.broadinstitute.org/single_cell	https://theislab.github.io/sfaira-portal/	https://rgv.genouest.org/	https://panglaodb.se	https://www.ebi.ac.uk/gxa/sc/home	https://bioinfo.uth.edu/scRNA-seqdb/	http://imlspenticton.uzh.ch:3838/conquer/	https://jinglebells.bgu.ac.il/	https://data.humancellatlas.org/

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

313

Table S2. Criterion for query search

Query Search	Keyword Search	Multiple Object Search	Category Filters
Basic	Y		
Intermediate	Y		Y
Advanced	Y	Y	Y

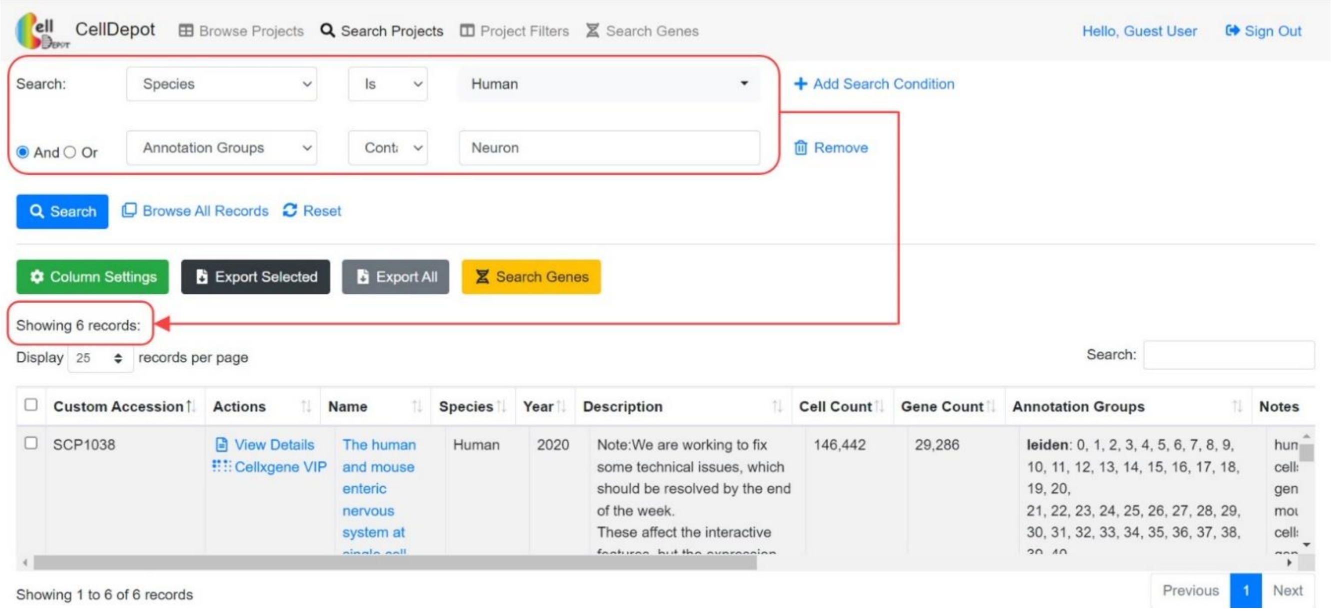
Table S3. Criterion for data analysis explorer

Data Analysis Explorer	Visualization and Analyze scRNA-seq Data	Visualization and Analyze scRNA-seq Data Analyze Gene Expression	Customize Displays
Basic	Y		
Intermediate	Y	Y	
Advanced	Y	Y	Y

Table S4. Project metadata captured in CellDepot

	General Category	**Variable type**	**Description**
Project Browser/Search	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	Digital 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	String	
	PubMed ID	Link	
	Species	String	
	URL	Link	
	Year	String	

Supplementary Figures



Search: Species Is Human

● And ○ Or Annotation Groups Conti Neuron

Showing 6 records:

Display 25 records per page

Search:

☐	Custom Accession	Actions	Name	Species	Year	Description	Cell Count	Gene Count	Annotation Groups	Notes
☐	SCP1038	View Details Cellxgene VIP	The human and mouse enteric nervous system at	Human	2020	Note:We are working to fix some technical issues, which should be resolved by the end of the week. These affect the interactive	146,442	29,286	**leiden**: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38,	hun cell gen mou cell

Showing 1 to 6 of 6 records

Figure S1. Data Filtering. Query search of ‘Species is Human’ and ‘Annotation Groups contains Neuron’ brings about nine datasets of interest.

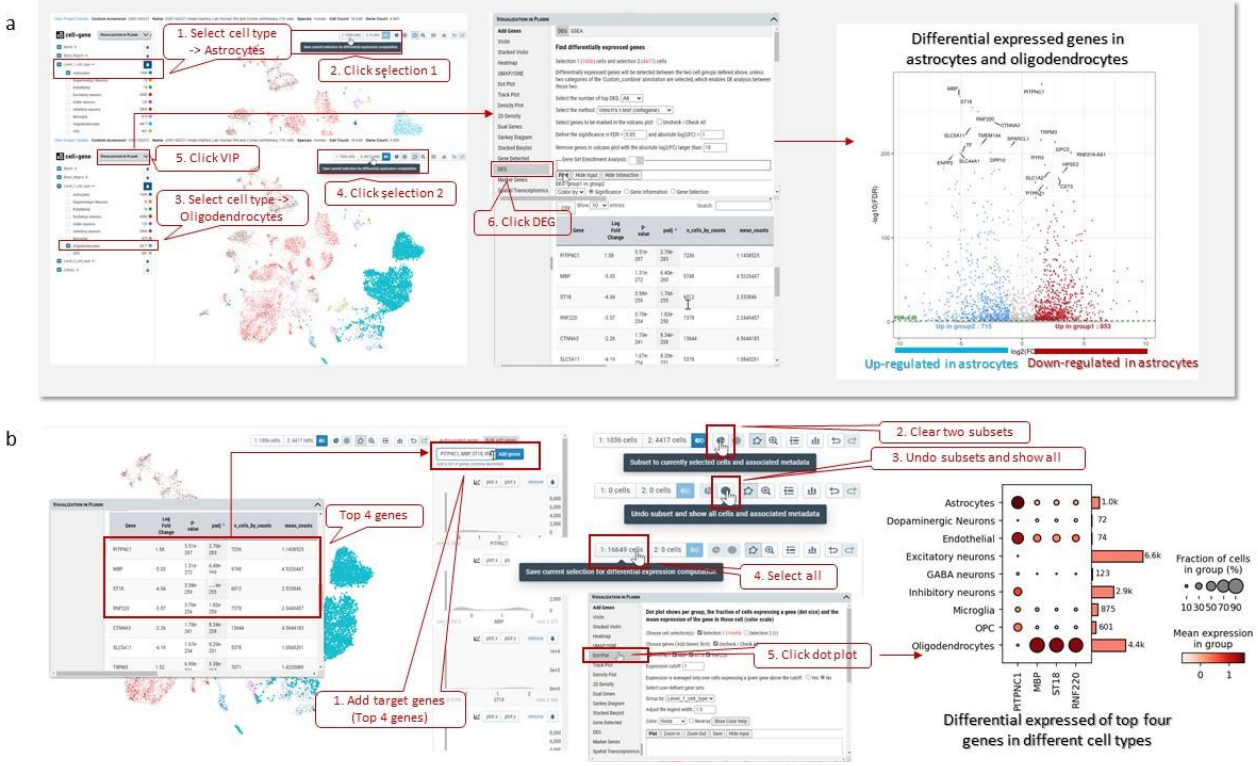
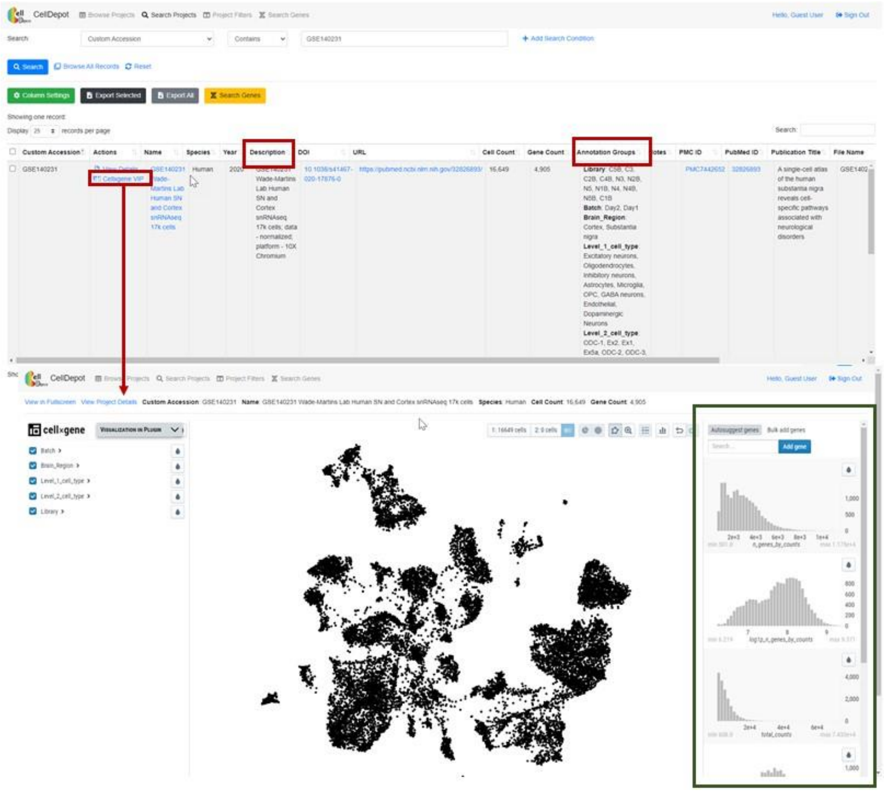


Figure S2. Exploration of differential expressed genes in dataset GSE140231 through cellxgene VIP [1]. (a) Differential expressed genes in astrocytes and oligodendrocytes. (b) The expression of top four genes in different cell types.



The annotation of the observation features from GSE140231.h5ad:

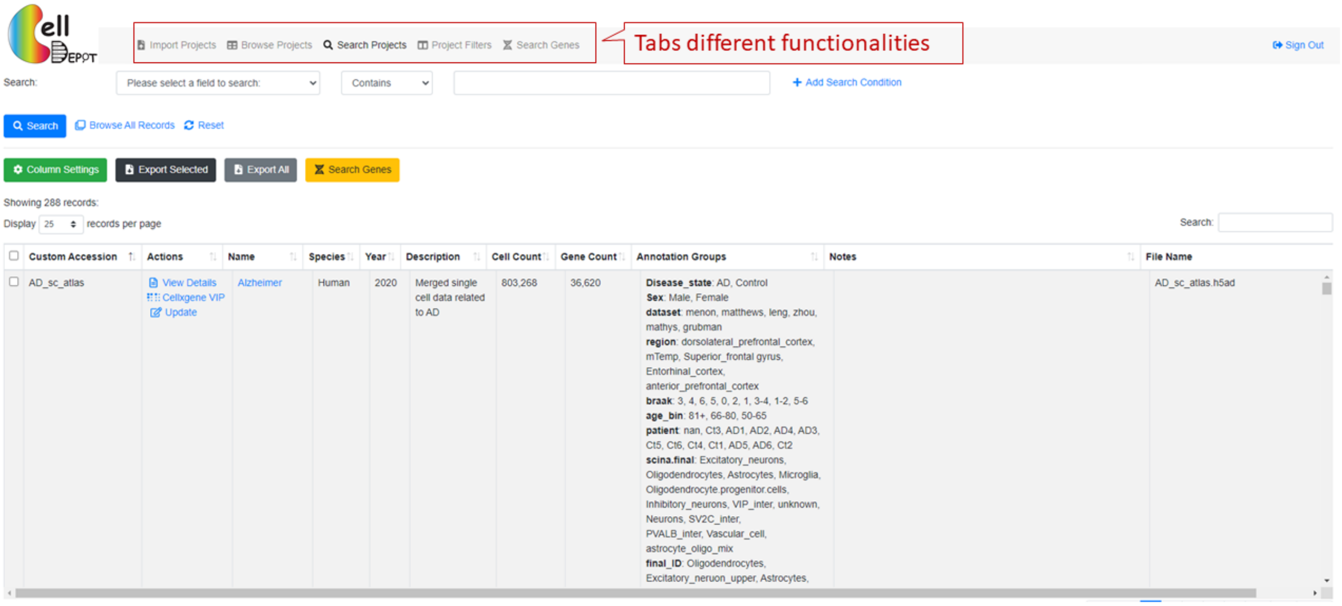
```
Index(['Library', 'Batch',
'Brain_Region', 'Level_1_cell_type',
'Level_2_cell_type', 'percent_mito',
'n_genes_by_counts',
'log1p_n_genes_by_counts',
'total_counts', 'log1p_total_counts',
'pct_counts_in_top_50_genes',
'pct_counts_in_top_100_genes',
'pct_counts_in_top_200_genes',
'pct_counts_in_top_500_genes',
'>Description'], dtype='object')
```

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.

Supplementary Tutorial

1. Introduction

CellDepot is database management system integrated with management system, query searching and data visualization tools [2, 3] for scRNA-seq datasets, which can be accessed by the link http://celldepot.bxgenomics.com. This is a supplemental tutorial provides a detailed guide.



Showing 288 records:

☐	Custom Accession	Actions	Name	Species	Year	Description	Cell Count	Gene Count	Annotation Groups	Notes	File Name
☐	AD_sc_atlas	View Details Cellxgene VIP Update	Alzheimer	Human	2020	Merged single cell data related to AD	803,268	36,620	**Disease_state**: AD, Control **Sex**: Male, Female **dataset**: menon, matthews, leng, zhou, mathys, grubman **region**: dorsolateral_prefrontal_cortex, mTemp, Superior_frontal gyrus, Entorhinal_cortex, anterior_prefrontal_cortex **braak**: 3, 4, 6, 5, 0, 2, 1, 3-4, 1-2, 5-6 **age_bin**: 81+, 66-80, 50-65 **patient**: nan, Ct3, AD1, AD2, AD4, AD3, Ct5, Ct6, Ct4, Ct1, AD5, AD6, Ct2 **scina.final**: Excitatory_neurons, Oligodendrocytes, Astrocytes, Microglia, Oligodendrocyte.progenitor.cells, Inhibitory_neurons, VIP_inter, unknown, Neurons, SV2C_inter, PVALB_inter, Vascular_cell, astrocyte_oligo_mix **final_ID**: Oligodendrocytes, Excitatory_neruon_upper, Astrocytes,		AD_sc_atlas.h5ad

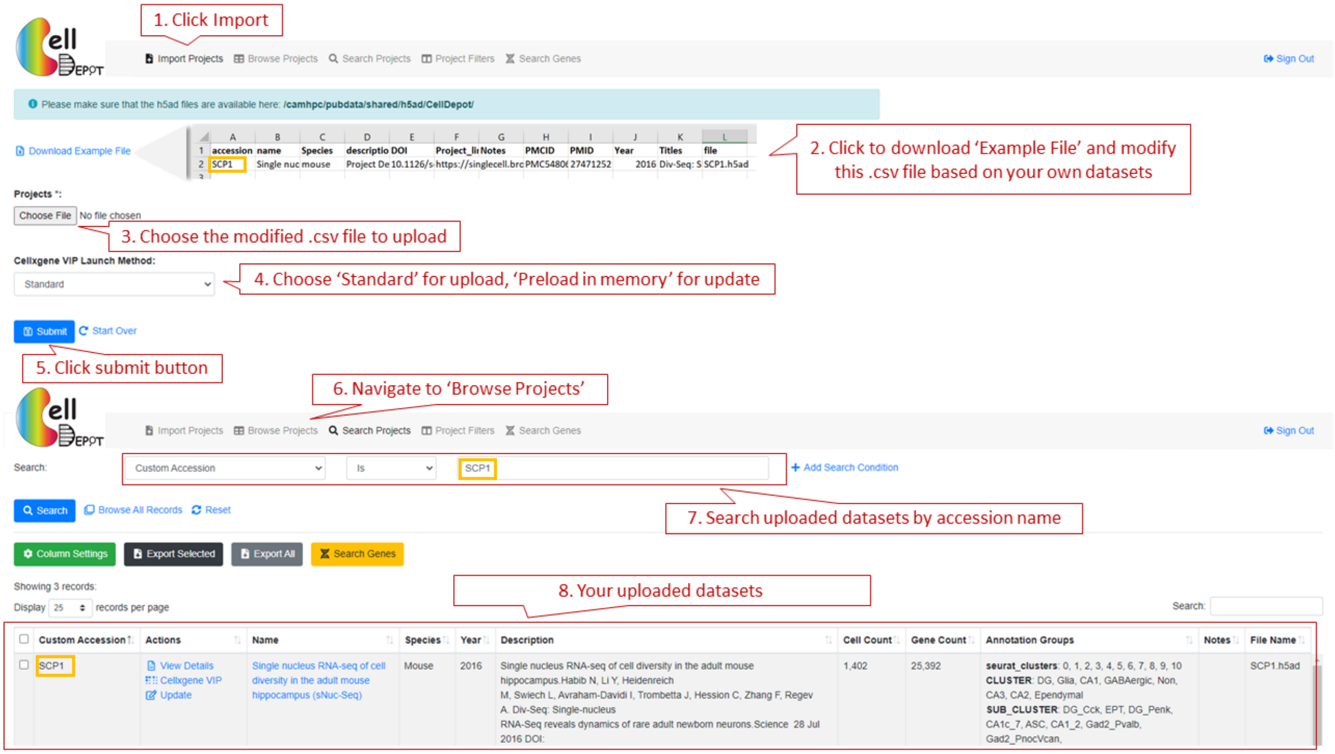
Figure S4. CellDepot website.

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.

2. 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.



1. Click Import

Please make sure that the h5ad files are available here: **/camhpc/pubdata/shared/h5ad/CellDepot/**

	A	B	C	D	E	F	G	H	I	J	K	L
1	accession	name	Species	descriptio	DOI	Project_li	Notes	PMCID	PMID	Year	Titles	file
2	SCP1	Single nuc	mouse	Project De	10.1126/s	https://singlecell.brc		PMC5480(	27471252	2016	Div-Seq: S	SCP1.h5ad

2. Click to download 'Example File' and modify this .csv file based on your own datasets

Projects *: No file chosen

3. Choose the modified .csv file to upload

Cellxgene VIP Launch Method: Standard

4. Choose 'Standard' for upload, 'Preload in memory' for update

5. Click submit button

6. Navigate to 'Browse Projects'

Search: Custom Accession Is SCP1

7. Search uploaded datasets by accession name

Showing 3 records:

Display 25 records per page

Search:

8. Your uploaded datasets

☐	Custom Accession	Actions	Name	Species	Year	Description	Cell Count	Gene Count	Annotation Groups	Notes	File Name
☐	SCP1	View Details Cellxgene VIP Update	Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus (sNuc-Seq)	Mouse	2016	Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus.Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta J, Hession C, Zhang F, Regev A. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons.Science 28 Jul 2016 DOI:	1,402	25,392	**seurat_clusters**: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 **CLUSTER**: DG, Glia, CA1, GABAergic, Non, CA3, CA2, Ependymal **SUB_CLUSTER**: DG_Cck, EPT, DG_Penk, CA1c_7, ASC, CA1_2, Gad2_Pvalb, Gad2_PnocVcan,		SCP1.h5ad

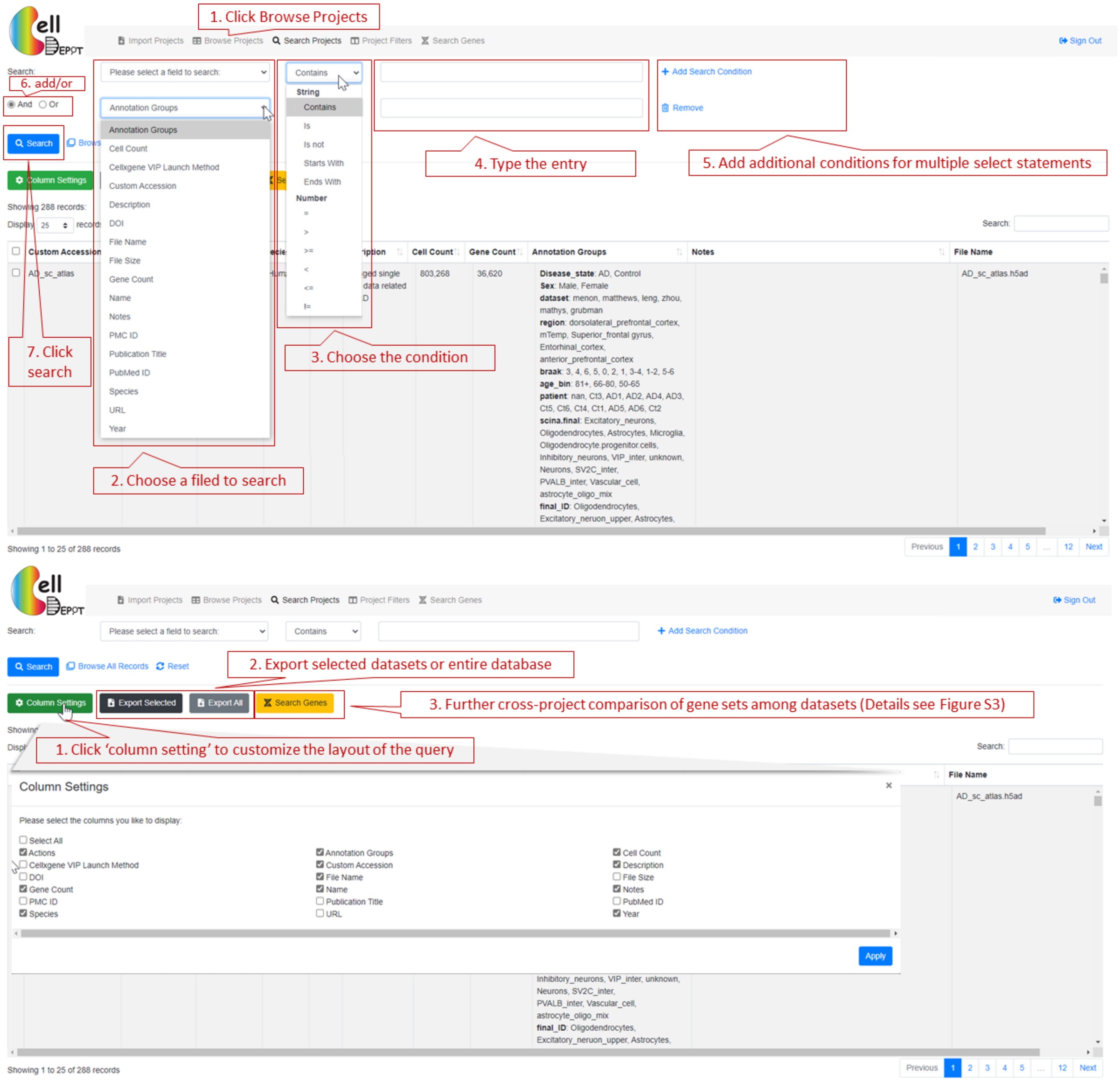
Figure S5. Workflow of how to import personal datasets.

3. Browse Projects

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.

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.


1. Click Browse Projects

2. Choose a filed to search

3. Choose the condition

4. Type the entry

5. Add additional conditions for multiple select statements

6. add/or

7. Click search

2. Export selected datasets or entire database

3. Further cross-project comparison of gene sets among datasets (Details see Figure S3)

1. Click 'column setting' to customize the layout of the query

Column Settings

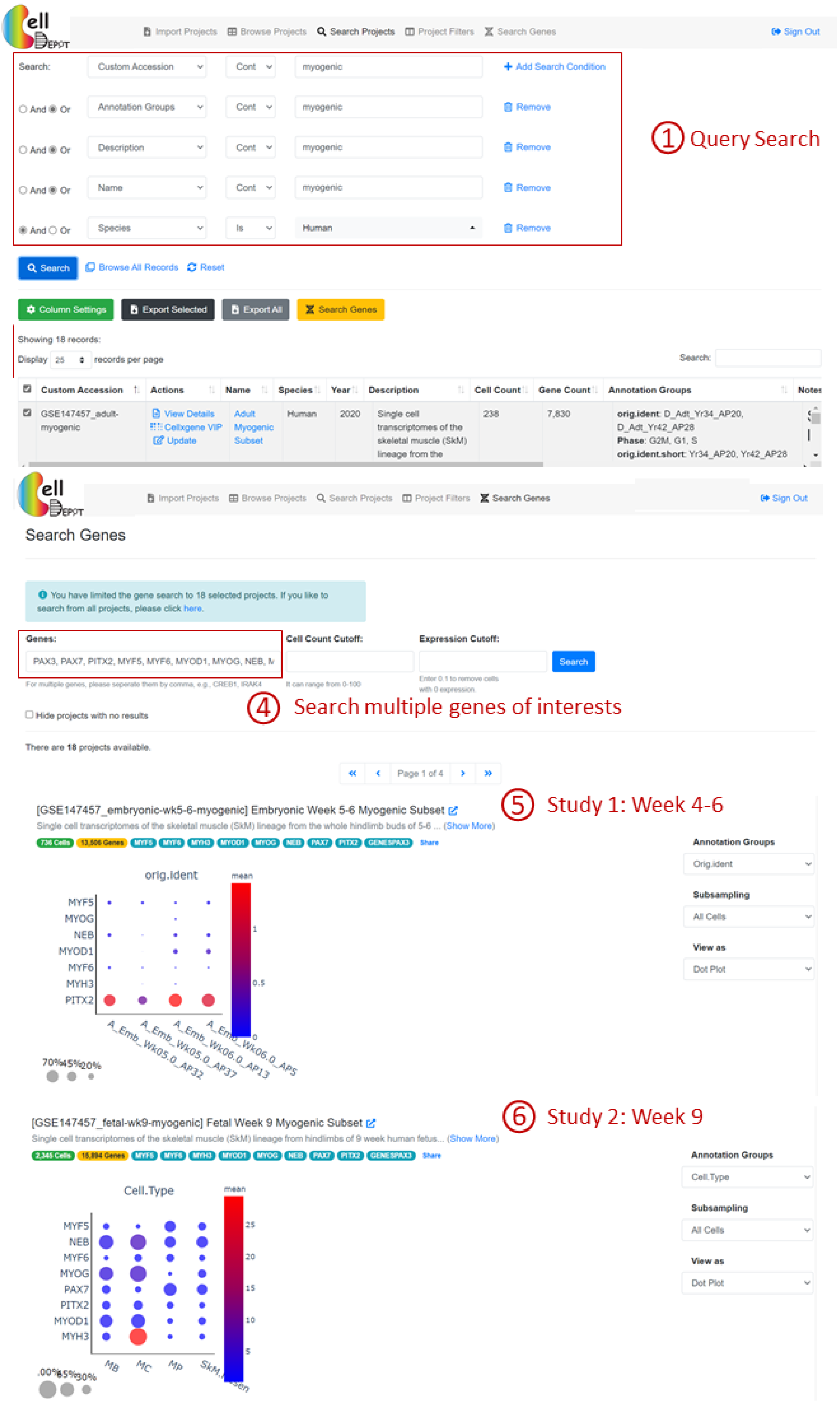
Please select the columns you like to display:

☐ Select All
☑ Actions
☐ Cellxgene VIP Launch Method
☐ DOI
☑ Gene Count
☐ PMC ID
☑ Species
☑ Annotation Groups
☑ Custom Accession
☑ File Name
☑ Name
☐ Publication Title
☐ URL
☑ Cell Count
☑ Description
☐ File Size
☑ Notes
☐ PubMed ID
☑ Year

Figure S6. Workflow of how to search query on 'Browse Projects' page.

2. Case Study 1

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.



① Query Search

Search: Custom Accession | Cont | myogenic

○ And ● Or | Annotation Groups | Cont | myogenic

○ And ● Or | Description | Cont | myogenic

○ And ● Or | Name | Cont | myogenic

● And ○ Or | Species | Is | Human

Showing 18 records:

Display 25 records per page

Search:

☑	Custom Accession	Actions	Name	Species	Year	Description	Cell Count	Gene Count	Annotation Groups	Notes
☑	GSE147457_adult-myogenic	View Details Cellxgene VIP Update	Adult Myogenic Subset	Human	2020	Single cell transcriptomes of the skeletal muscle (SkM) lineage from the	238	7,830	**orig.ident**: D_Adt_Yr34_AP20, D_Adt_Yr42_AP28 **Phase**: G2M, G1, S **orig.ident.short**: Yr34_AP20, Yr42_AP28	

Search Genes

You have limited the gene search to 18 selected projects. If you like to search from all projects, please click here.

Genes: PAX3, PAX7, PITX2, MYF5, MYF6, MYOD1, MYOG, NEB, M

For multiple genes, please seperate them by comma, e.g., CREB1, IRAK4

Cell Count Cutoff:

It can range from 0-100

Expression Cutoff:

Enter 0.1 to remove cells with 0 expression

④ Search multiple genes of interests

☐ Hide projects with no results

There are **18** projects available.

Page 1 of 4

⑤ Study 1: Week 4-6

[GSE147457_embryonic-wk5-6-myogenic] Embryonic Week 5-6 Myogenic Subset

Single cell transcriptomes of the skeletal muscle (SkM) lineage from the whole hindlimb buds of 5-6 ... (Show More)

736 Cells | 13,506 Genes | MYF5 | MYF6 | MYH3 | MYOD1 | MYOG | NEB | PAX7 | PITX2 | GENESPAX3

Annotation Groups: Orig.ident

Subsampling: All Cells

View as: Dot Plot

⑥ Study 2: Week 9

[GSE147457_fetal-wk9-myogenic] Fetal Week 9 Myogenic Subset

Single cell transcriptomes of the skeletal muscle (SkM) lineage from hindlimbs of 9 week human fetus... (Show More)

2,345 Cells | 18,894 Genes | MYF5 | MYF6 | MYH3 | MYOD1 | MYOG | NEB | PAX7 | PITX2 | GENESPAX3

Annotation Groups: Cell.Type

Subsampling: All Cells

View as: Dot Plot

Figure S7. Workflow of how to conduct the cross-project comparison of gene sets among the selected datasets.

3. 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 (Details see Figure S4).

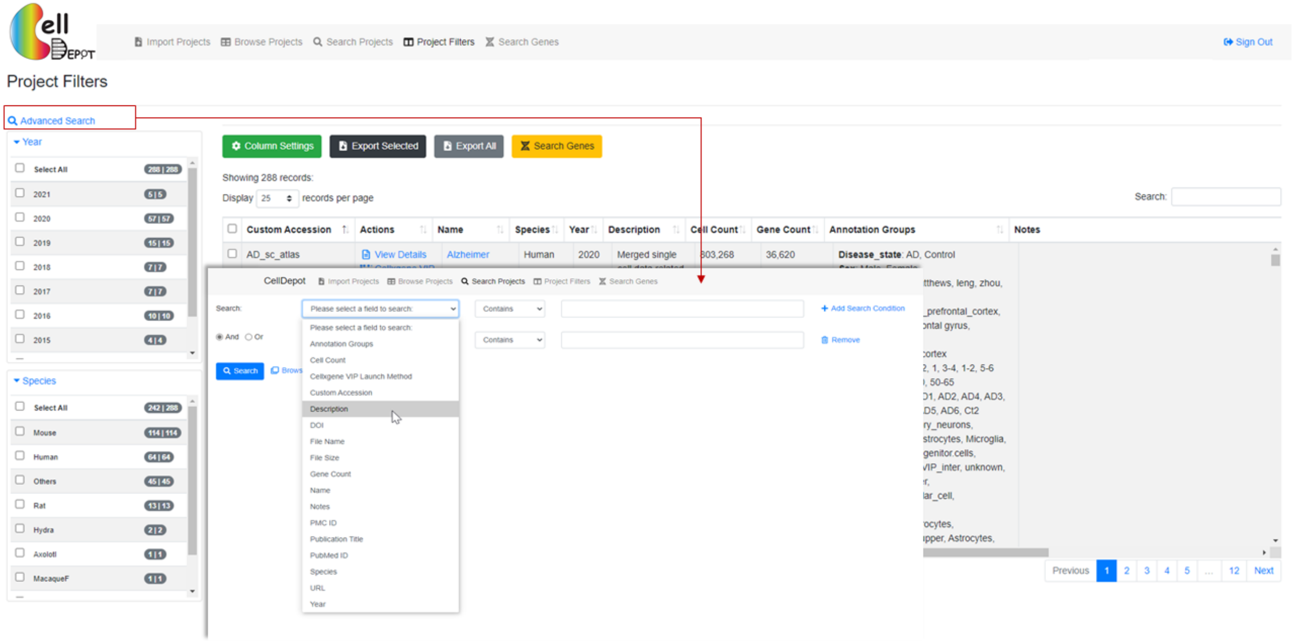
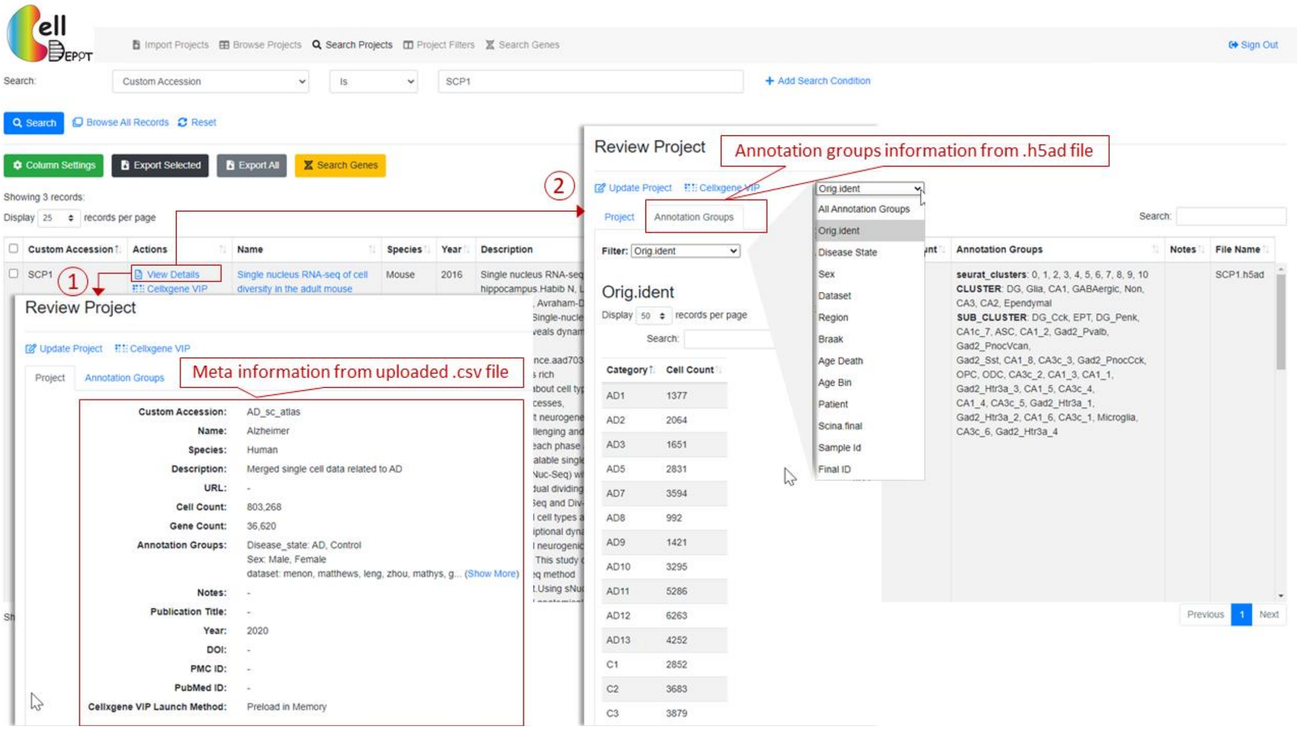


Figure S8. The layout of 'Project Filters' page.

4. Visualize Datasets

4.1 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.



Search: Custom Accession | Is | SCP1

Showing 3 records:

Display 25 records per page

Custom Accession	Actions	Name	Species	Year	Description
SCP1	View Details Cellxgene VIP	Single nucleus RNA-seq of cell diversity in the adult mouse	Mouse	2016	Single nucleus RNA-seq hippocampus.Habib N,

Review Project

Custom Accession: AD_sc_atlas
Name: Alzheimer
Species: Human
Description: Merged single cell data related to AD
URL: -
Cell Count: 803,268
Gene Count: 36,620
Annotation Groups: Disease_state: AD, Control
Sex: Male, Female
dataset: menon, matthews, leng, zhou, mathys, g... (Show More)
Notes: -
Publication Title: -
Year: 2020
DOI: -
PMC ID: -
PubMed ID: -
Cellxgene VIP Launch Method: Preload in Memory

Review Project

Filter: Orig.ident

Orig.ident

Display 50 records per page

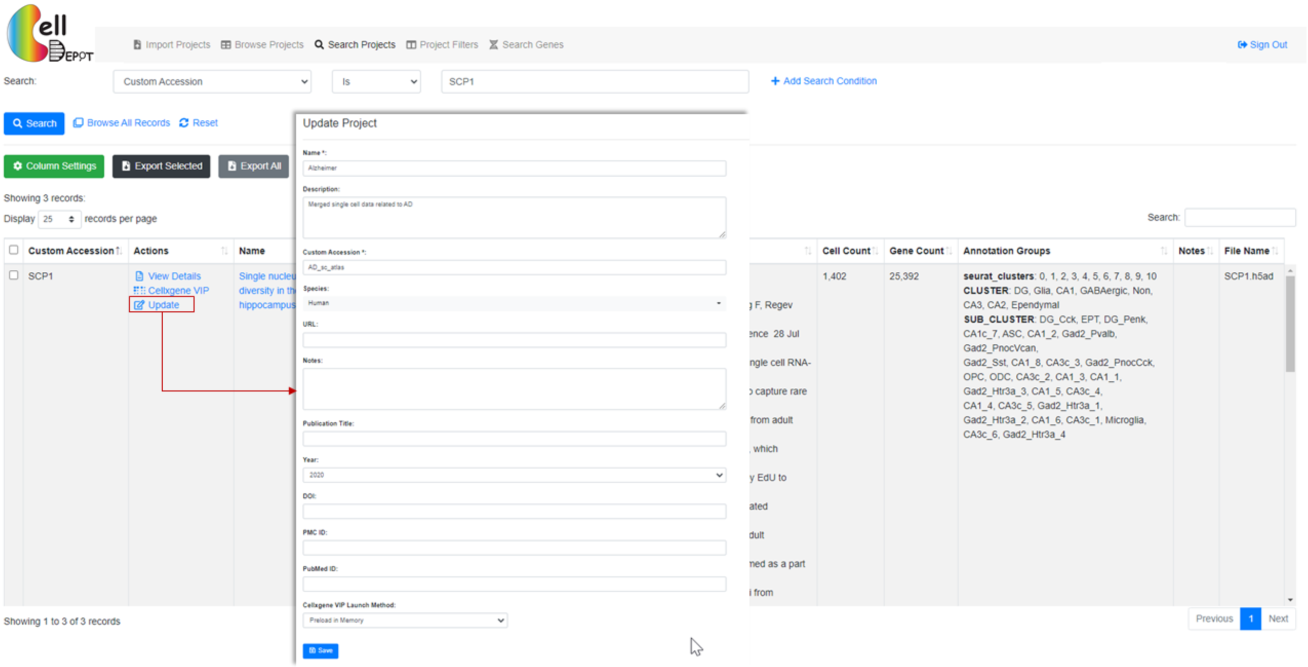
Category	Cell Count
AD1	1377
AD2	2064
AD3	1651
AD5	2831
AD7	3594
AD8	992
AD9	1421
AD10	3295
AD11	5286
AD12	6263
AD13	4252
C1	2852
C2	3683
C3	3879

Annotation Groups	Notes	File Name
seurat_clusters: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 **CLUSTER**: DG, Glia, CA1, GABAergic, Non, CA3, CA2, Ependymal **SUB_CLUSTER**: DG_Cck, EPT, DG_Penk, CA1c_7, ASC, CA1_2, Gad2_Pvalb, Gad2_PnocVcan, Gad2_Sst, CA1_8, CA3c_3, Gad2_PnocCck, OPC, ODC, CA3c_2, CA1_3, CA1_1, Gad2_Htr3a_3, CA1_5, CA3c_4, CA1_4, CA3c_5, Gad2_Htr3a_1, Gad2_Htr3a_2, CA1_6, CA3c_1, Microglia, CA3c_6, Gad2_Htr3a_4		SCP1.h5ad

Figure S9. Visualization of details of datasets.

4.2 Update

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



Search: Custom Accession | Is | SCP1

Showing 3 records:

Display 25 records per page

Custom Accession	Actions	Name	Cell Count	Gene Count	Annotation Groups	Notes	File Name
SCP1	View Details Cellxgene VIP Update	Single nucleus diversity in the hippocampus	1,402	25,392	**seurat_clusters**: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 **CLUSTER**: DG, Glia, CA1, GABAergic, Non, CA3, CA2, Ependymal **SUB_CLUSTER**: DG_Cck, EPT, DG_Penk, CA1c_7, ASC, CA1_2, Gad2_Pvalb, Gad2_PnocVcan, Gad2_Sst, CA1_8, CA3c_3, Gad2_PnocCck, OPC, ODC, CA3c_2, CA1_3, CA1_1, Gad2_Htr3a_3, CA1_5, CA3c_4, CA1_4, CA3c_5, Gad2_Htr3a_1, Gad2_Htr3a_2, CA1_6, CA3c_1, Microglia, CA3c_6, Gad2_Htr3a_4		SCP1.h5ad

Showing 1 to 3 of 3 records

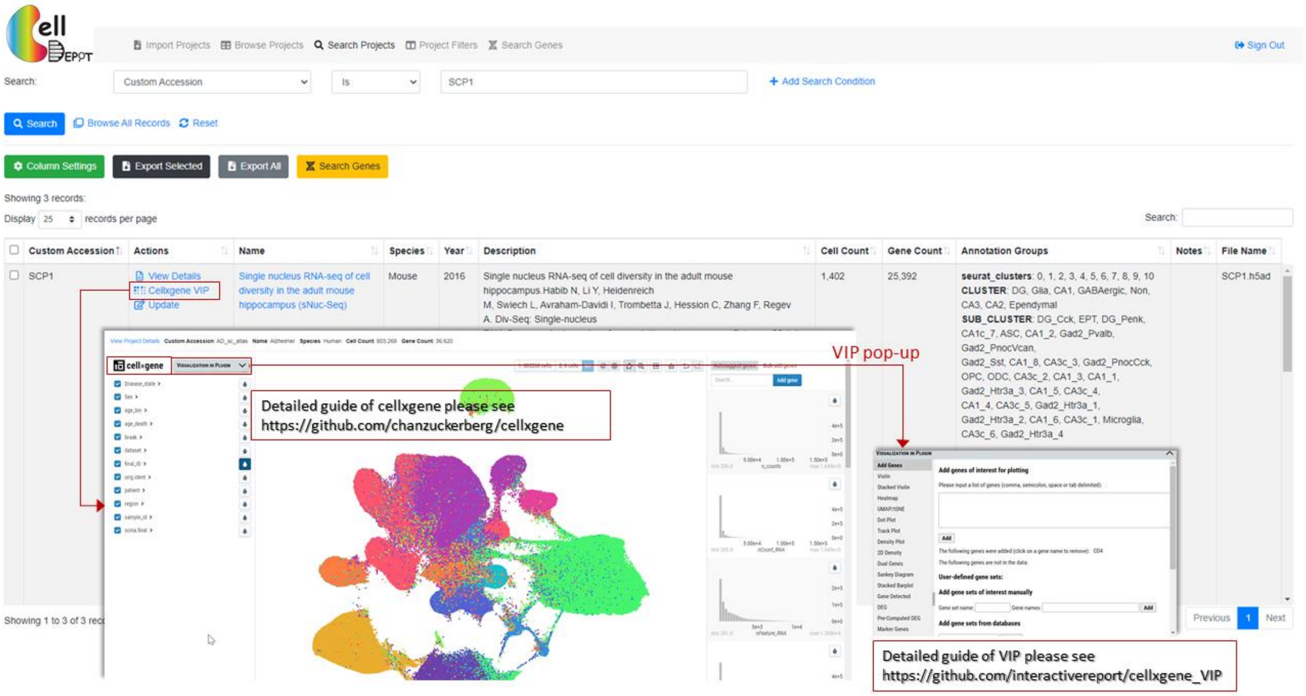
Update Project

Name *: Alzheimer
Description: Merged single cell data related to AD
Custom Accession *: AD_sc_atlas
Species: Human
URL:
Notes:
Publication Title:
Year: 2020
DOI:
PMC ID:
PubMed ID:
Cellxgene VIP Launch Method: Preload in Memory

Figure S10. Update project on ‘Browse Project’ page.

4.3 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://github.com/chanzuckerberg/cellxgene (cellxgene) and https://github.com/interactivereport/cellxgene_VIP (cellxgene_VIP).



Search: Custom Accession | Is | SCP1

Showing 3 records:

Display 25 records per page

Search:

☐	Custom Accession	Actions	Name	Species	Year	Description	Cell Count	Gene Count	Annotation Groups	Notes	File Name
☐	SCP1	View Details Cellxgene VIP Update	Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus (sNuc-Seq)	Mouse	2016	Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus.Habib N, Li Y, Heidenreich M, Swiech L, Avraham-Davidi I, Trombetta J, Hession C, Zhang F, Regev A. Div-Seq: Single-nucleus	1,402	25,392	**seurat_clusters**: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 **CLUSTER**: DG, Glia, CA1, GABAergic, Non, CA3, CA2, Ependymal **SUB_CLUSTER**: DG_Cck, EPT, DG_Penk, CA1c_7, ASC, CA1_2, Gad2_Pvalb, Gad2_PnocVcan, Gad2_Sst, CA1_8, CA3c_3, Gad2_PnocCck, OPC, ODC, CA3c_2, CA1_3, CA1_1, Gad2_Htr3a_3, CA1_5, CA3c_4, CA1_4, CA3c_5, Gad2_Htr3a_1, Gad2_Htr3a_2, CA1_6, CA3c_1, Microglia, CA3c_6, Gad2_Htr3a_4		SCP1.h5ad

Showing 1 to 3 of 3 rec

VIP pop-up

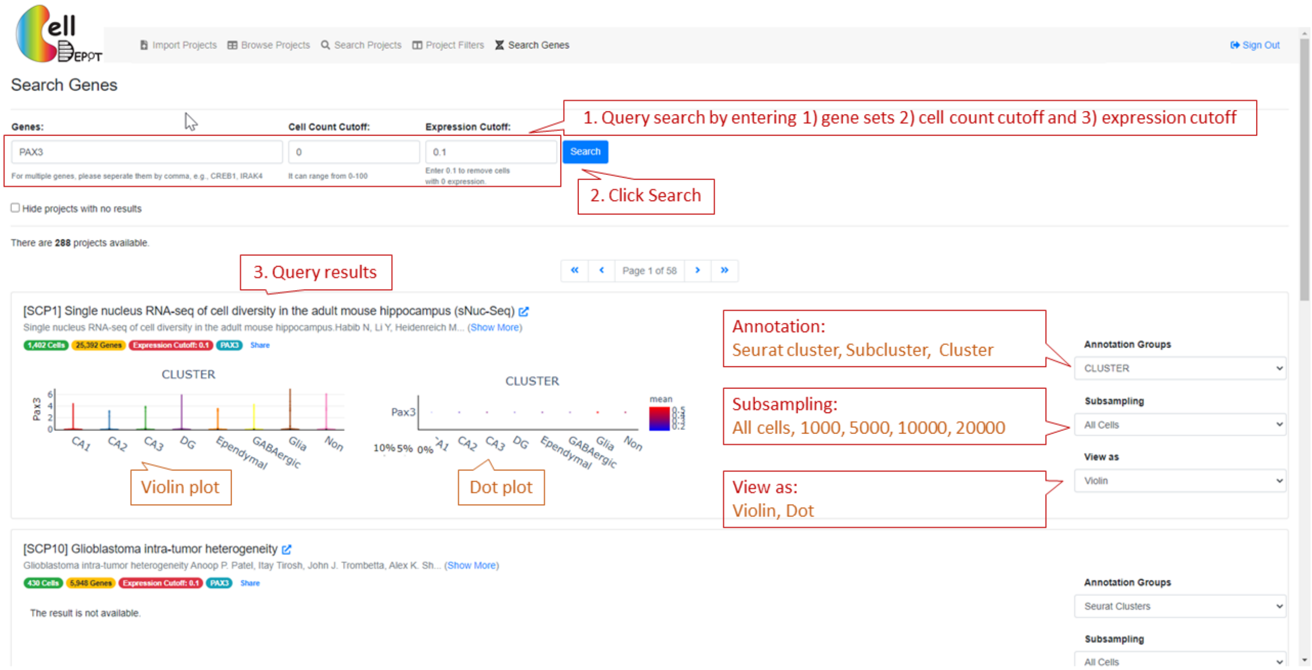
Detailed guide of cellxgene please see https://github.com/chanzuckerberg/cellxgene

Detailed guide of VIP please see https://github.com/interactivereport/cellxgene_VIP

Figure S11. See visualization of selected datasets.

5. 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.



Search Genes

Genes: PAX3

For multiple genes, please seperate them by comma, e.g., CREB1, IRAK4

Cell Count Cutoff: 0

It can range from 0-100

Expression Cutoff: 0.1

Enter 0.1 to remove cells with 0 expression.

1. Query search by entering 1) gene sets 2) cell count cutoff and 3) expression cutoff

2. Click Search

☐ Hide projects with no results

There are **288** projects available.

3. Query results

Page 1 of 58

[SCP1] Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus (sNuc-Seq)

Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus.Habib N, Li Y, Heidenreich M... (Show More)

1,402 Cells · 25,392 Genes · Expression Cutoff: 0.1 · PAX3

Violin plot

Dot plot

Annotation:
Seurat cluster, Subcluster, Cluster

Subsampling:
All cells, 1000, 5000, 10000, 20000

View as:
Violin, Dot

Annotation Groups: CLUSTER

Subsampling: All Cells

View as: Violin

[SCP10] Glioblastoma intra-tumor heterogeneity

Glioblastoma intra-tumor heterogeneity Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Sh... (Show More)

430 Cells · 5,948 Genes · Expression Cutoff: 0.1 · PAX3

The result is not available.

Annotation Groups: Seurat Clusters

Subsampling: All Cells

Figure S12. The layout of 'Search Genes'

6. 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.

Reference

[1] Agarwal D, Sandor C, Volpato V, Caffrey TM, Monzon-Sandoval J, Bowden R, et al. A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. Nat Commun. 2020;11:4183.

[2] Li K, Ouyang Z, Lin D, Mingueneau M, Chen W, Sexton D, et al. cellxgene VIP unleashes full power of interactive visualization, plotting and analysis of scRNA-seq data in the scale of millions of cells. 2020:2020.08.28.270652.

[3] Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, et al. cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. 2021:2021.04.05.438318.