Takara Bio USA

Cogent™ NGS Discovery Software User Manual

(073025)

Table of Contents

Table of Figures

Table of Tables

I. Introduction

Cogent NGS Discovery Software (referred to as CogentDS in this manual) is bioinformatics software for user-friendly analysis of sequencing data derived from Takara Bio applications, such as kits designed for the Shasta™ Single Cell System and plate-based NGS workflows.

A. What's New

Unless otherwise noted, the software contains all features included in previous versions.

- **Cogent NGS Discovery Software v2.2**
  - scDNA app: added new SNV modules
  - Bulk RNA app: added a batch correction step
  - Barcode Rank Plot module updated

NOTE: Find release notes for prior versions on the Cogent NGS Discovery Software product page.

II. Before You Begin

A. Supported Operating Systems

The CogentDS is designed to be installed on a user workstation (laptop or desktop) and should work on any system that supports R (see Section II.D). Installation and functionality have been tested and supported for the following OSs:

- Windows 11
- macOS Mojave (10.14) or higher
- Linux Ubuntu

B. Hardware Requirements

CogentDS, with its dependencies, is a lightweight program. It should work on any basic workstation (desktop or laptop) with ≥20 GB of free disk space and a minimum of 8 GB RAM (recommended 16 GB).

For optimal performance when running Shasta Total RNA-Seq kit data, a minimum of 64 GB of RAM and ≥60 GB of disk space is required. If your system does not meet this requirement, you may experience potential issues including application crashes and incomplete processing of some modules.

C. User Account Requirements

By default, administrative privileges are not required to install or run CogentDS. However, if working in an environment where R is installed with IT restrictions, an administrator may need to install the necessary software dependencies (Section II.D) and the Cogent NGS Discovery Software.

D. Additional Hardware and Software Dependencies and Recommendations

- **Internet connectivity on the computer/server**
- **R**

  R is a free, open-source software for statistical computing that provides support across a variety of operating systems. CogentDS is designed to work within an R environment. More information on obtaining and installing R is available in Section IV.A.

- **RStudio (IDE for R)**

  RStudio is a free, open-source program that provides graphical user interface (GUI) access to R. More information on obtaining and installing RStudio is available in Section IV.C.

- **An open network port on the install machine**

  As the CogentDS interface is accessed through a web GUI, a network port needs to be available on the computer it will be installed on. The port number is selected at random by the RStudio software, shiny, checking for open ports on the install computer or server until one is found. For more information about this assignment process, please see https://shiny.rstudio.com/reference/shiny/1.0.1/runApp.html.

  If running in an environment where the TCP/IP ports are locked down, please check with your local IT to ensure a port is available on the computer for CogentDS to use.

E. Required Input Files

The CogentDS apps require specific input files for processing, depending on the app or mode. Use the list below to determine what files are required for your desired outcome:

1. scRNA app > Analysis Mode (Section VI.A) & Bulk RNA app (Section VII)

Single-cell and bulk RNA analysis requires one of the two following file options as input. The listed files are output from Cogent NGS Analysis Pipeline (CogentAP):

- `CogentDS.analysis.rds`, an R-object file. This is the recommended input option since it enables full analysis capabilities, especially those added for leveraging our full-length chemistries. These capabilities include isoform-count as a standard, and gene-fusion and immune analysis if these optional analyses have been run first in CogentAP.
- Raw gene-count matrix and stats files. This input option is mainly for users only interested in gene-based analysis; it also provides backwards-compatibility for files generated using previous versions of CogentAP software (Version 2.0).

2. scRNA app > Discovery Mode (Section VI.B)

The single-cell RNA discovery mode requires a `CogentDS.analysis.rds` file saved after performing CogentDS analysis using the scRNA Analysis Mode (Section VI.A.17, "Download CogentDS Processed .rds Data").

3. scRNA app > Barcode Rank Plot (Section VI.C)

The single-cell RNA barcode rank plot mode requires as input a `demultiplexed_fastqs_counts_all.estimated.csv` file, generated by CogentAP using the `--dry_run` argument.

An example of the dry_run command syntax is shown in Section V.B.1.b) of the Cogent NGS Analysis Pipeline User Manual, "RNA Demux Dry Run (For Shasta Total RNA-Seq Kit Data)". The dry_run output from any single-cell RNA-seq kit type can be used for this CogentDS mode.

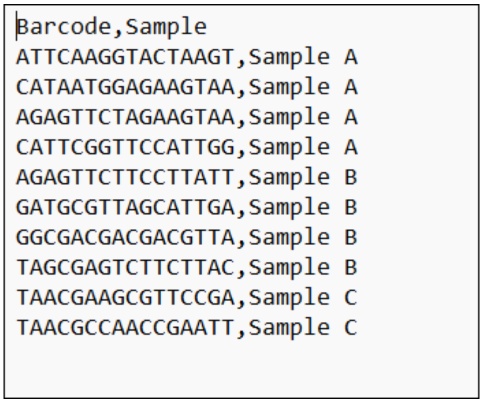
4. scDNA app (Section VIII)

Single-cell DNA analysis requires a `CogentDS.analysis.rds` file from CogentAP (see above).

F. Optional Input File

For the scRNA app Analysis Mode and Bulk RNA app, a `metadata.csv` file can additionally be uploaded along with the required input data file.

This text-based file of comma-separated values (CSV) contains additional identifying information associated with the sequencing data barcodes, such as sample name, biotype, biological experiment, etc. Figure 1 is an example of a very simple metadata file that includes sample names to associate with barcodes. Users can select the column of interest from the uploaded metadata file for downstream analysis in CogentDS.



```
Barcode,Sample
ATTCAAGGTACTAAGT,Sample A
CATAATGGAGAAGTAA,Sample A
AGAGTTCTAGAAGTAA,Sample A
CATTCGGTTCCATTGG,Sample A
AGAGTTCTTCCTTATT,Sample B
GATGCGTTAGCATTGA,Sample B
GGCGACGACGACGTTA,Sample B
TAGCGAGTCTTCTTAC,Sample B
TAACGAAGCGTTCCGA,Sample C
TAACGCCAACCGAATT,Sample C
```

Figure 1. Example metadata file contents. This example, displayed in a text editor, is provided to illustrate the type of information contained in the file and how the file is constructed.

A metadata file is not required but, when used, will make available additional parameters that can be applied for data visualization in the process of the downstream analysis.

III. Cogent NGS Analysis and Discovery Software Overview

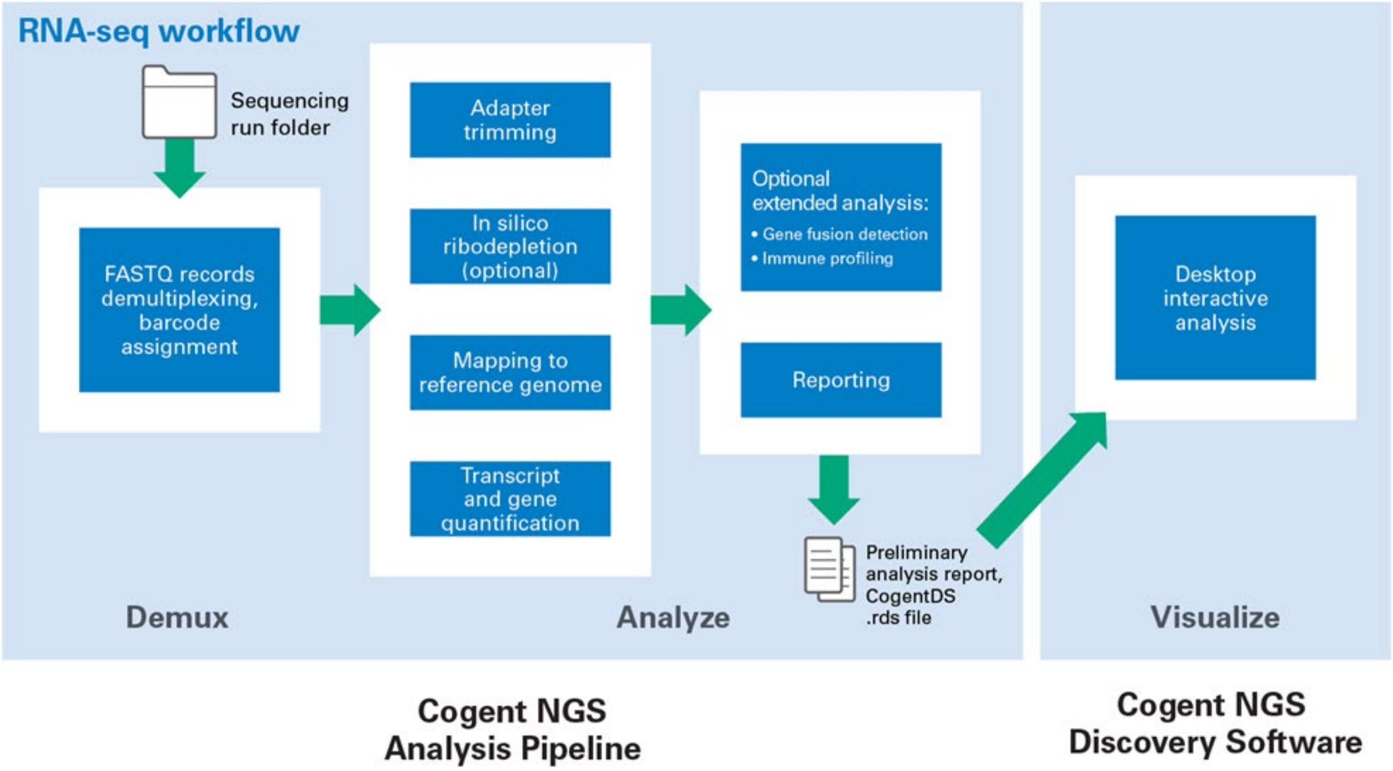

RNA-seq workflow
Sequencing run folder
FASTQ records demultiplexing, barcode assignment
Adapter trimming
In silico ribodepletion (optional)
Mapping to reference genome
Transcript and gene quantification
Optional extended analysis:
• Gene fusion detection
• Immune profiling
Reporting
Preliminary analysis report, CogentDS .rds file
Desktop interactive analysis
Demux
Analyze
Visualize
Cogent NGS Analysis Pipeline
Cogent NGS Discovery Software

Figure 2. High-level overview of the RNA-seq analysis workflow of Cogent NGS Analysis Pipeline and Cogent NGS Discovery Software.

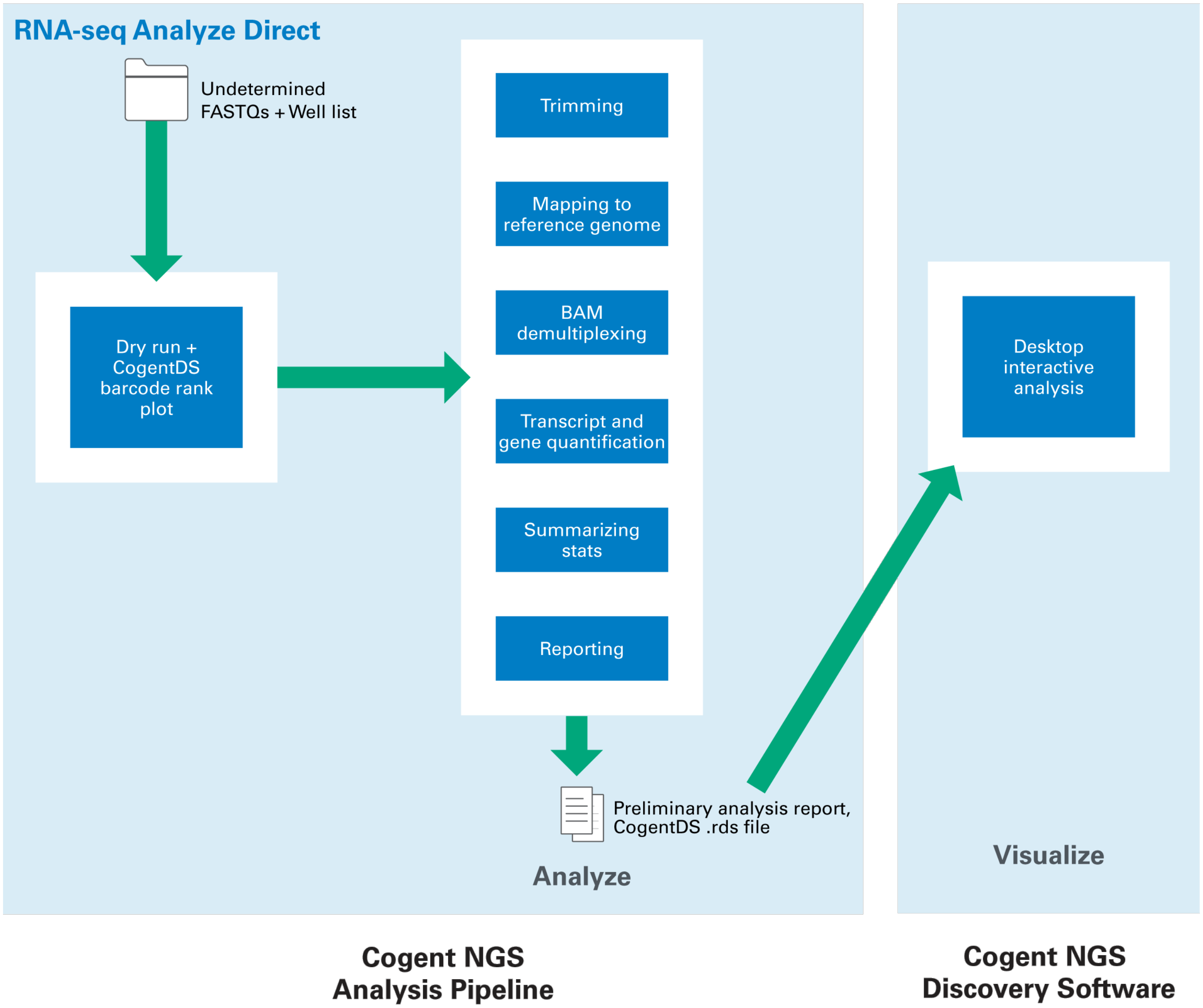

RNA-seq Analyze Direct
Undetermined FASTQs + Well list
Dry run + CogentDS barcode rank plot
Trimming
Mapping to reference genome
BAM demultiplexing
Transcript and gene quantification
Summarizing stats
Reporting
Preliminary analysis report, CogentDS .rds file
Analyze
Desktop interactive analysis
Visualize
Cogent NGS Analysis Pipeline
Cogent NGS Discovery Software

Figure 3. High-level RNA-seq Analyze Direct workflow of Cogent NGS Analysis Pipeline and Cogent NGS Discovery Software.

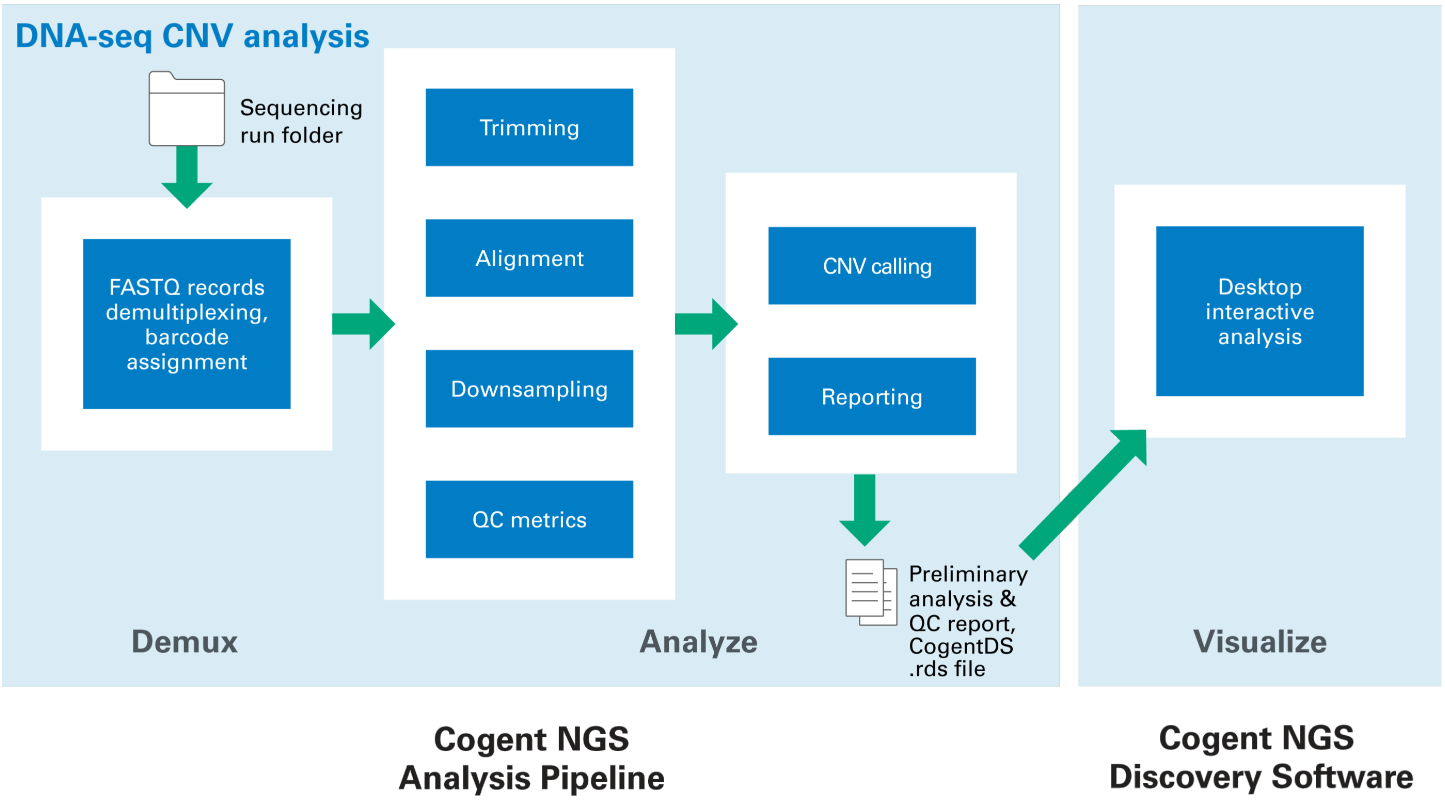

DNA-seq CNV analysis
Sequencing run folder
FASTQ records demultiplexing, barcode assignment
Trimming
Alignment
Downsampling
QC metrics
CNV calling
Reporting
Preliminary analysis & QC report, CogentDS .rds file
Desktop interactive analysis
Demux
Analyze
Visualize
Cogent NGS Analysis Pipeline
Cogent NGS Discovery Software

Figure 4. High-level scDNA-seq (CNV analysis) workflow of Cogent NGS Analysis Pipeline and Cogent NGS Discovery Software.

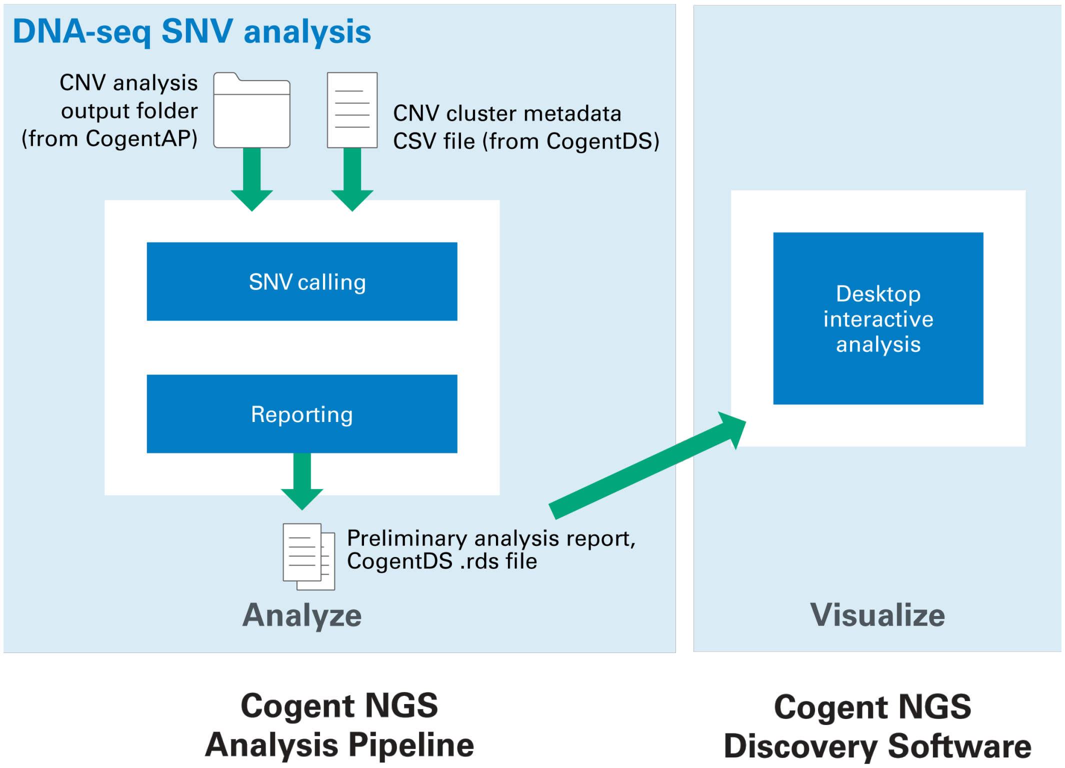

DNA-seq SNV analysis
CNV analysis output folder (from CogentAP)
CNV cluster metadata CSV file (from CogentDS)
SNV calling
Reporting
Preliminary analysis report, CogentDS .rds file
Analyze
Desktop interactive analysis
Visualize
Cogent NGS Analysis Pipeline
Cogent NGS Discovery Software

Figure 5. High-level scDNA-seq (SNV analysis) workflow of Cogent NGS Analysis Pipeline and Cogent NGS Discovery Software.

Figure 2 & 3 (RNA-seq) and 4 & 5 (DNA-seq) depict the high-level workflows of the analysis provided by CogentAP and how their output is carried over to CogentDS. For more information about Cogent NGS Analysis Pipeline, see the Cogent NGS Analysis Pipeline User Manual.

Once CogentDS and required dependencies are installed, analysis can be launched in an interactive RStudio session.

IV. Installation & Configuration Options

Run through the steps in this section to set up the R environment and install the CogentDS software.

NOTE: If you're upgrading from an older version of CogentDS, go to Section IV.D to uninstall it first.

A. Install R

R and many of the contributing packages are available on the Comprehensive R Archive Network (CRAN). If R is not installed on your system, please download and install R version 4.4 or 4.5 from https://www.r-project.org, by first choosing a CRAN mirror of your choice.

NOTES:

– Ensure you download the installation package specific to the platform on which it is to be installed. I.e., download the Windows installation executable for a computer running Windows.

– R version 4.6 and later is not currently supported.

For more information on installing R, see the tutorial at datacamp.com.

B. Install Platform-specific Tools

Installation of CogentDS on Windows or Macintosh workstations requires additional software to be installed prior to installing RStudio.

1. Windows 11

On Windows, R requires Rtools to build and install packages from a source file. Download the most recent version of Rtools compatible with R 4.4.x (e.g., rtools44) from https://cran.r-project.org/bin/windows/Rtools/. During installation, ensure Rtools is included in the system PATH. For more information on installing Rtools, see the instructions on the download page.

NOTE: Rtools must be installed in a file path with directory names that do not include spaces (i.e., it cannot be installed in `C:\Program Files\` but could be installed in `C:\Program\`). Installing it in a file path with spaces in the directory names will cause the Cogent NGS Discovery Software installation to fail.

If Rtools is installed in such a location on the target computer, please uninstall Rtools and re-install in a folder with a path that conforms to these requirements.

To improve performance, you can increase your system's virtual memory. For detailed instructions on how to adjust virtual memory settings, please refer to the following link: https://helpdeskgeek.com/windows-11/how-to-increase-virtual-memory-in-windows-11/.

2. macOS Mojave (10.14) or higher

For macOS: Homebrew; after R installation, please install `xcode` tools and run `brew install libomp` and `brew install gcc`.

In addition to these, ensure XQuartz is installed for successful execution of the single-cell DNA-seq application.

3. Ubuntu (24.04.2)

For Linux (Ubuntu): use apt or apt-get to install libomp-dev, gcc, zlib1g-dev, libcurl-dev, libssl-dev, libxml2-dev, libpng-dev, libicu-dev, libcairo2-dev, perl, libfreetype6-dev, libjpeg-dev, libtiff5-dev, libfontconfig-dev, libfribidi-dev, libharfbuzz-dev, libblas-dev, liblapack-dev, gfortran, cmake, gsfonts, libglpk-dev, libmagick++-dev, make, libnode-dev, pandoc, libcurl4-openssl-dev, libX11-dev and r-base (≥v4.4.0).

NOTE: To identify Linux OS-level dependencies of R packages required by CogentDS, one can use the pak R package (https://pak.r-lib.org/). Please note that there might be some OS-level dependencies not listed here, which may need to be installed for a successful installation of the CogentDS on a Linux OS.

C. Set Up RStudio

1. Install RStudio

If RStudio is not installed on your system, please download and install the RStudio Desktop (Open Source License) version for your Operating System from rstudio.com. Please install the latest version to make sure it is compatible with the version of R (v4.4 or higher) installed on your system.

For more information on installing RStudio on Windows, refer to the same tutorial at datacamp.org as in Section IV.A for installing R (scroll down towards the bottom of that page).

Once it's installed and running, verify that RStudio is configured to use R 4.4.0 or higher by default. This can be checked in the upper right-hand corner of RStudio.

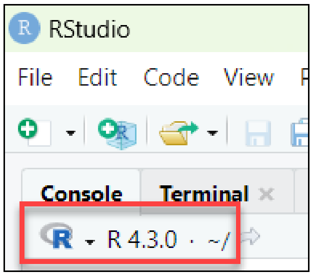


Figure 6. RStudio version environment. The screenshot shows that the R version is 4.3.0.

If it shows an older version, such as in Figure 6, follow the instructions in the next section. Otherwise, proceed to Section D.

2. Setting R Version in RStudio Environment

If the environment version of R needs to be changed in your install of RStudio, do the following to change it.

1. Navigate to **Tools > Global Options…**

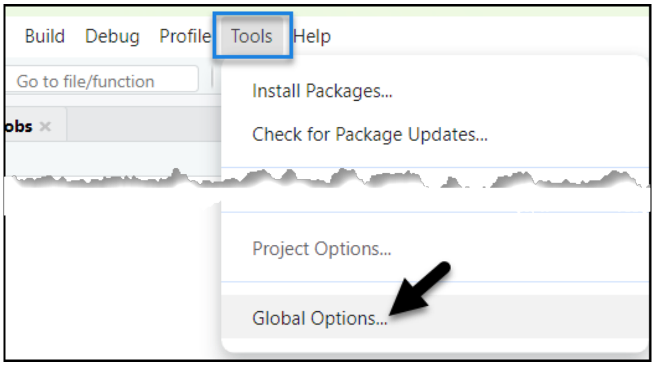
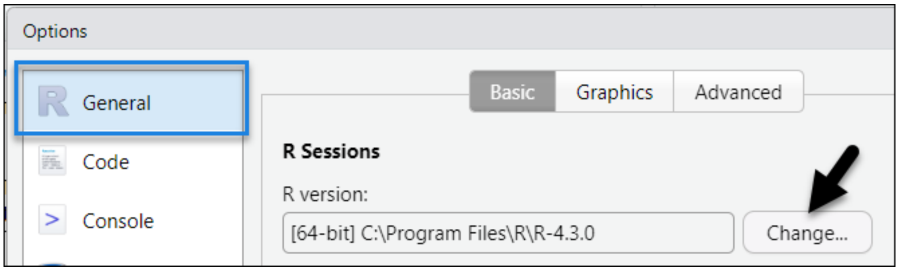


Figure 7. RStudio Tools > Global Options… menu item.

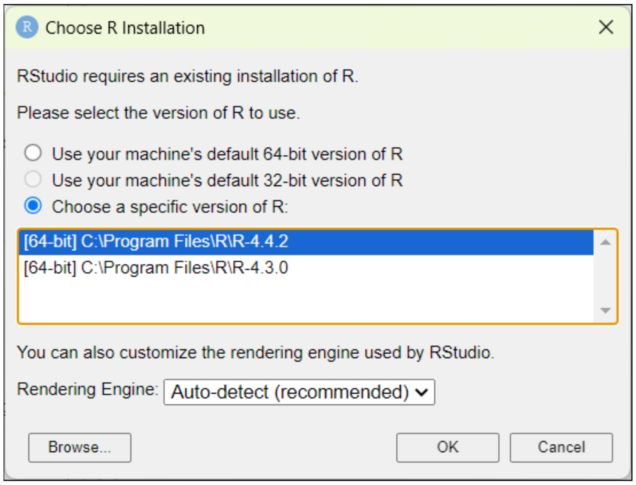
2. In the resulting dialog window, under **General**, click the [Change…] button.



R version:
[64-bit] C:\Program Files\R\R-4.3.0

Figure 8. Where to change the version of R being used by RStudio.

3. From the next dialog window, select:
    - 'Use your machine's default 64-bit version of R', assuming that R 4.4.0 or higher is installed,
    - Or if you're unsure or the default version is earlier than 4.4.0, 'Choose a specific version of R'. This will allow you to select the desired version of R from a list of versions installed on your system.



Choose R Installation

RStudio requires an existing installation of R.

Please select the version of R to use.

○ Use your machine's default 64-bit version of R
○ Use your machine's default 32-bit version of R
● Choose a specific version of R:

[64-bit] C:\Program Files\R\R-4.4.2
[64-bit] C:\Program Files\R\R-4.3.0

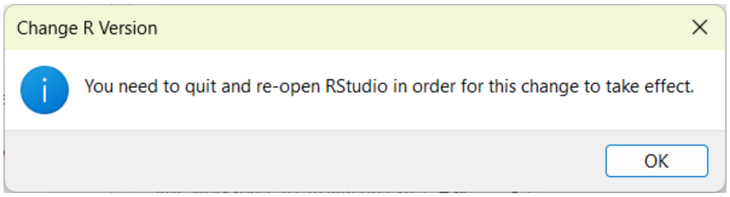
You can also customize the rendering engine used by RStudio.

Rendering Engine: Auto-detect (recommended)

Figure 9. Choosing a specific version of R for the RStudio environment.

Click [OK] to implement the change.

4. A message will pop up notifying you that RStudio has to be restarted after the change. Click [OK], then [OK] again to quit out of the Options, and shut down RStudio.



Change R Version

You need to quit and re-open RStudio in order for this change to take effect.

Figure 10. *Change R Version* restart prompt after changing the RStudio version environment.

5. Run RStudio again. The version should show as updated in the upper right corner (refer to Figure 6).

D. Uninstall Previous Instances of CogentDS

NOTE: If CogentDS has never been installed on the server, skip to the next section (Section IV.E).

If CogentDS v1.5 or hanta™ Software was installed on the server previously, it should be uninstalled prior to installing Cogent NGS Discovery Software using the procedure in Section IV.F.

Follow the uninstall directions in Section IV.F ("How to Uninstall CogentDS").

E. Install Cogent NGS Discovery Software Dependencies

CogentDS is available for download as a compressed file from the CogentDS product page.

1. Download the CogentDS ZIP file (`Cogent_NGS_Discovery_Software_v2.2.zip`), following the directions (a) on the page seen after submitting the sign-up form on the CogentDS product page or (b) in the email sent to the email address submitted in the form.
2. Unzip the CogentDS zip file.
3. In RStudio type the following commands:

   ```
   setwd("<PATH>")
   ```

   ```
   source ("setup_CogentDS.R")
   ```

   where **`<PATH>`** is replaced by the full path where the unzipped CogentDS file is stored.

Example:

If the `Cogent_NGS_Discovery_Software_v2.2.zip` file was downloaded to `C:\temp` on a Windows server, the command would resemble:

```
setwd("C:/temp/Cogent_NGS_Discovery_Software_v2.2")
```

```
source("setup_CogentDS.R")
```

For first-time users, the installation process may take 10–20 minutes, as many dependencies are automatically downloaded and installed. The installation may also prompt you to accept downloading and installing certain packages from the source. Answer 'yes' (or 'y', case insensitive) to any such prompts.

If an error is displayed indicating RStudio could not remove a prior package installation, please refer to Cogent NGS Discovery Software notices for a potential fix.

If you encounter issues downloading and installing the package from GitHub during the installation (possibly caused by GitHub SSH keys), please follow the steps below to resolve the issue:

- Check in RStudio for a GITHUB token using: `Sys.getenv("GITHUB_PAT")`
- In some cases: Personal access tokens (PAT) may need to be created and set using: `gitcreds::gitcreds_set()`
- Use https://github.com/settings/tokens (classic) or https://github.com/settings/personal-access-tokens (fine-grained), run `gitcreds::gitcreds_set()` in R or RStudio
- Replace these credentials, and then paste the PAT when prompted
- Optional resource: https://github.com/orgs/community/discussions/140956#discussioncomment-10890333
- Once these steps have been performed, re-run the installation command.

If you continue to encounter a problem, please reach out to our technical support team for additional assistance.

F. How to Uninstall CogentDS

To uninstall older versions of Cogent NGS Discovery Software (v1.5), run the following command at the Rstudio prompt:

```
remove.packages("CogentDS")
```

To uninstall a previous installation of the original hanta Software, run:

```
remove.packages("hanta")
```

V. Cogent NGS Discovery Software: Getting Started

A. Run CogentDS

1. Run RStudio.
2. From the RStudio console, run the command:

```
setwd('<PATH>')
```

where **<PATH>** is replaced by the full path where the unzipped CogentDS file is stored.

NOTE: This needs to be done after every time RStudio is started.

3. Start CogentDS with the following command from the RStudio console:

```
source('launch_CogentDS.R')
```

This command will launch the **default browser** on your computer and create a new instance of the CogentDS user interface (GUI), running through the localhost of your computer (IP address 127.0.0.1) and a randomly assigned, available TCP/IP port (see Section II.D, for more information about the selection of the TCP/IP port).

B. Software Interface Overview

The initial screen will look like Figure 11.

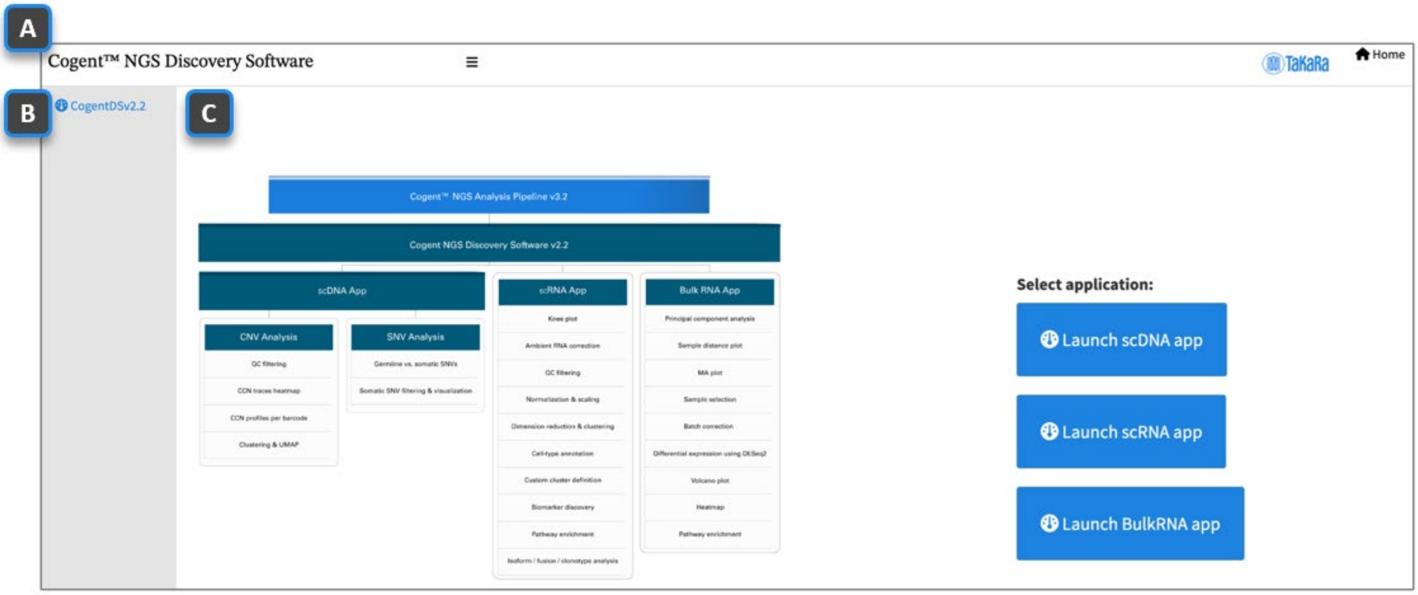


Figure 11. Initial screen of CogentDS in the web browser.

The parts of the screen, labeled by the boxes, are described below:

[A] Window title bar—this includes the product name, Cogent NGS Discovery Software

[B] Side bar—the name of the module displaying in the main window ([C]) is shown here

[C] Main window—where the application will run and display

The individual elements of the title bar are:

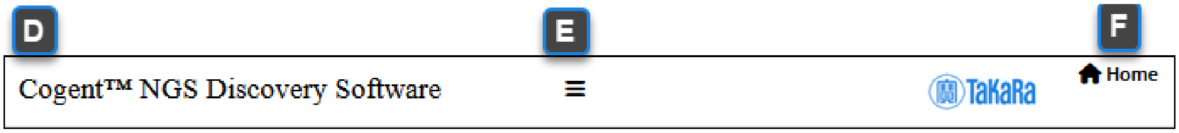


Figure 12. CogentDS title bar in the web browser.

[D] The full name of the CogentDS software

[E] The hamburger menu icon—toggles the view of the side bar ([B], above) to make it visible or hidden

[F] Home—use this to reload the CogentDS view page, rather than the refresh button of the browser

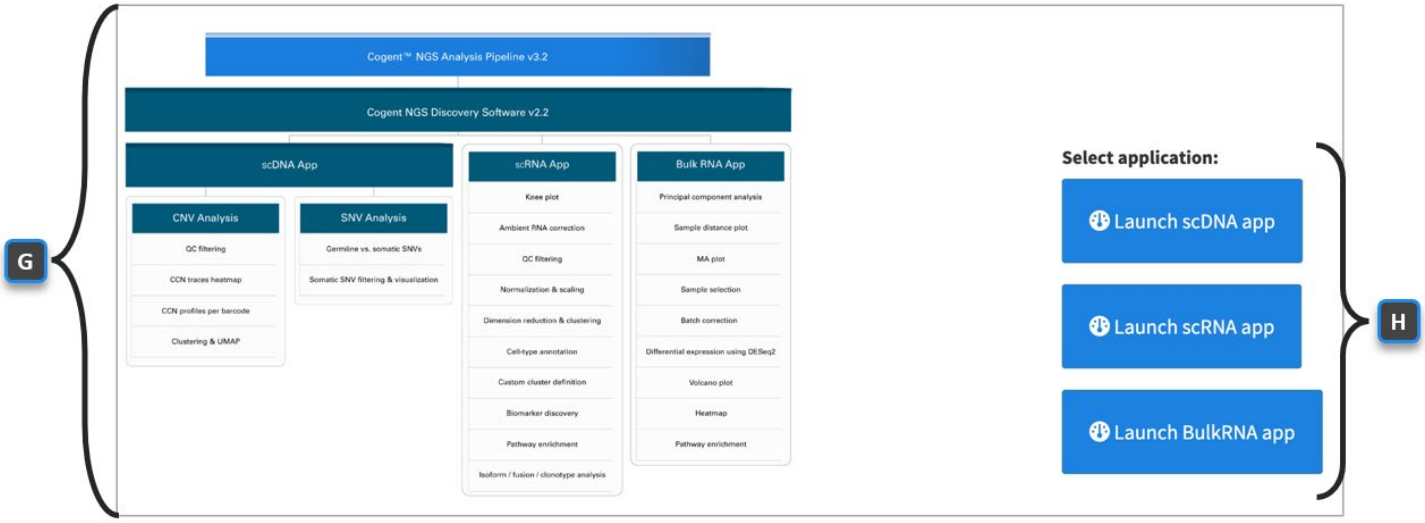


Figure 13. CogentDS initial main window view.

The elements on the initial main window are:

[G] Workflow graphic—the graphic shows all current applications available from the module displaying in the window. On the initial view, this will be all apps (scDNA, scRNA, and Bulk RNA)

[H] Application selection—the list of available applications from the screen. On the initial view, click on any of the three applications to start it.

C. CogentDS Applications

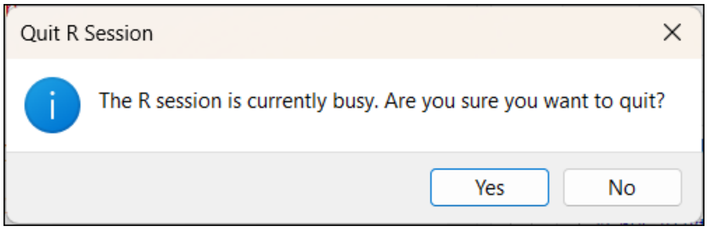
There are three application options for further exploring the data from CogentAP. The chosen option depends on the sequencing data you are attempting to analyze.

- scRNA app—for analysis of single-cell RNA-seq data. This app offers three modes: Analysis Mode, Discovery Mode, and Barcode Rank Plot.
- BulkRNA app—for analysis of bulk RNA-seq data
- scDNA app—for analysis of single-cell DNA-seq data

D. Closing CogentDS and RStudio

Once you have completed your analysis using CogentDS, follow the steps below to shut down CogentDS and RStudio.

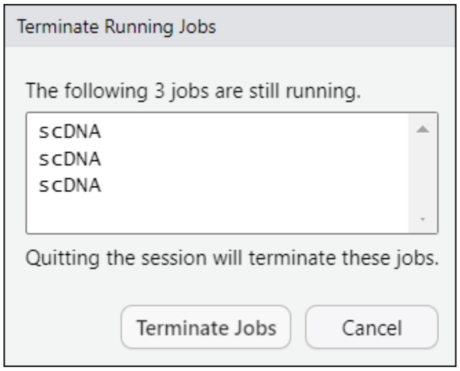
1. Close either all browser tabs opened by CogentDS or shut down the browser program entirely.
2. Quit the RStudio program. You will see a message similar to Figure 14, notifying you that there are processes running and requiring you to confirm termination of the jobs to quit. Click [Yes] to continue.



The R session is currently busy. Are you sure you want to quit?

Figure 14. ***Quit R Session*** **confirmation window in RStudio.**

4. A second pop-up window will display requesting confirmation to terminate the running jobs (Figure 15). Click [Terminate Jobs] to confirm.



The following 3 jobs are still running.

scDNA
scDNA
scDNA

Quitting the session will terminate these jobs.

Figure 15. ***Terminate Running Jobs*** **confirmation window in RStudio.**

RStudio will then shut down.

VI. Application: scRNA Analysis

To start a single-cell RNA-seq analysis, click [Launch scRNA app]. The app will open in a new browser tab with the scRNA app main window. The screen will appear as in Figure 16.

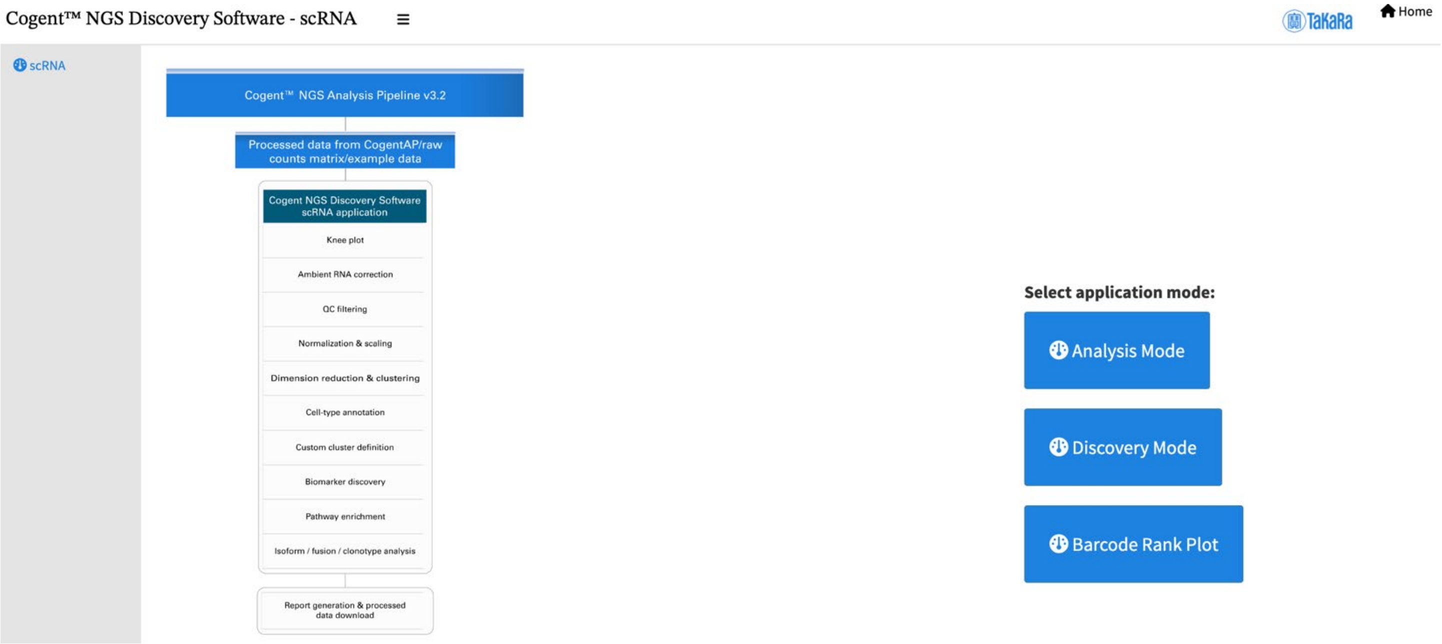

Cogent™ NGS Discovery Software - scRNA
TaKaRa
Home
scRNA
Cogent™ NGS Analysis Pipeline v3.2
Processed data from CogentAP/raw counts matrix/example data
Cogent NGS Discovery Software scRNA application
Knee plot
Ambient RNA correction
QC filtering
Normalization & scaling
Dimension reduction & clustering
Cell-type annotation
Custom cluster definition
Biomarker discovery
Pathway enrichment
Isoform / fusion / clonotype analysis
Report generation & processed data download
Select application mode:
Analysis Mode
Discovery Mode
Barcode Rank Plot

Figure 16. scRNA application main window.

Elements of the scRNA app main window include:

[A] Workflow graphic—shows the flow of modules within the scRNA app. This image is not clickable.

[B] Application mode selection—the available application modes within the scRNA app. Click on any of the three modes to start it.

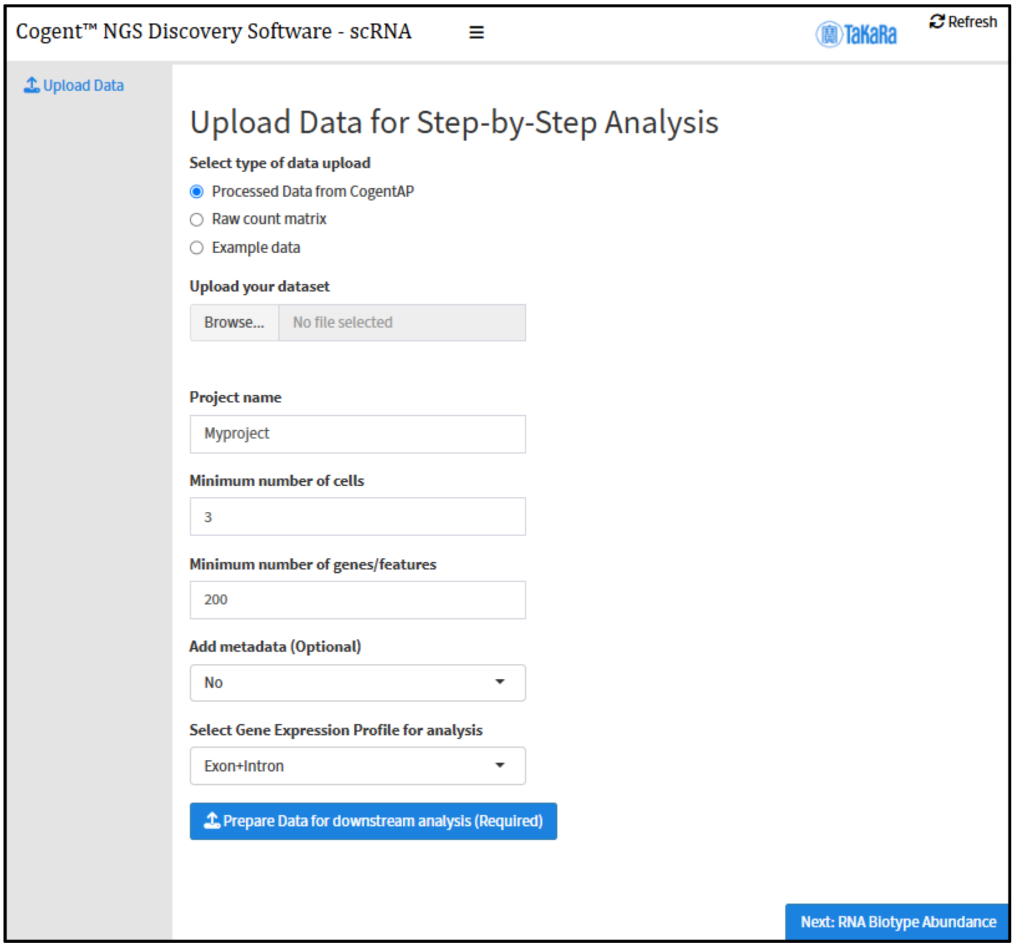
- Analysis Mode
- Discovery Mode
- Barcode Rank Plot

A. Analysis Mode

The Analysis Mode within the scRNA app analyzes scRNA-seq data and takes you through each step of the scRNA-seq analysis process. To start a scRNA-seq analysis, click [Analysis Mode].

1. Upload Data

Upon entering Analysis Mode, a new browser tab will open with the *Upload Data for Step-by-Step Analysis* window (Figure 17).


Cogent™ NGS Discovery Software - scRNA
TaKaRa
Refresh
Upload Data
Browse...
Prepare Data for downstream analysis (Required)
Next: RNA Biotype Abundance

Upload Data for Step-by-Step Analysis

Select type of data upload

● Processed Data from CogentAP

○ Raw count matrix

○ Example data

Upload your dataset: No file selected

Project name: Myproject

Minimum number of cells: 3

Minimum number of genes/features: 200

Add metadata (Optional): No

Select Gene Expression Profile for analysis: Exon+Intron

Figure 17. *Upload Data for Step-by-Step Analysis* window within the scRNA app.

The *Upload Data for Step-by-Step Analysis* window allows you to input data files generated using CogentAP for scRNA-seq analysis. There are three data upload options to choose from:

- Processed Data from CogentAP—allows for upload of a `CogentDS.analysis.rds` file

  NOTE: Upload of a `Cogent.analysis.rds` file from CogentAP v3.2 is recommended to utilize the full analysis capabilities of the scRNA app.

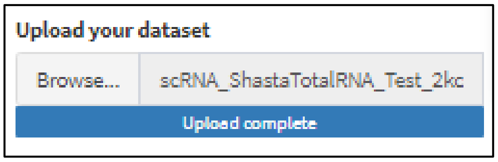
- Raw count matrix—allows for upload of raw gene-count matrix and stats/metadata files in .csv or .csv.gz format. These files can be found in the `counts_matrix\` folder in CogentAP when the number of cells is ≤5,000. For more information on this folder, please reference the Cogent NGS Analysis Pipeline User Manual. Please note that the gene info file is required for this option to be used.

- Example data—allows you to go through all the steps of the scRNA app analysis with an example dataset consisting of sequencing data from 1,640 PBMCs.

To begin the step-by-step analysis, use the radio button under "Select type of data to upload" to choose the desired data input option, following the directions below.

If you have selected "Processed Data from CogentAP" or "Raw count matrix" for data upload:

Click on [Browse] under **Upload your dataset** to select a file for upload. Once a file is selected, wait for the file upload to complete (Figure 18)



Upload your dataset

Browse... scRNA_ShastaTotalRNA_Test_2kc

Upload complete

Figure 18. "Upload complete" under Upload your dataset step.

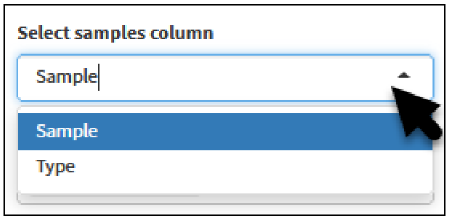
Specify the following parameters:

- Project name—allows users to tag their Seurat object with a unique identifier
- Minimum number of cells—sets a threshold for filtering out features/genes expressed in fewer than a specified number of cells. The default setting for this parameter is 3. For the default setting, only features/genes expressed in at least 3 cells will be retained for analysis.
- Minimum number of genes/features—sets a threshold for the minimum number of features/genes that a cell must express to be included in the analysis. The default setting for this parameter is 200. For the default setting, only cells that express at least 200 genes/features will be retained for analysis.
- Add metadata (Optional)—provides an option to upload a metadata file (Section II.F)

  To associate the metadata file with the selected dataset, first select 'Yes' from the drop-down menu. The following options will appear on the screen:

  - Upload metadata file—click [Browse] to select the file for upload. Once the file is selected, wait for the file upload to complete.
  - Select samples column—(optional) select a sample column from the uploaded file as configuration for the next option ("Select samples to analyze"). The options listed in the menu are dynamic and correspond to the column headers in the metadata file, omitting the "Barcodes" column.

    Only one value can be selected from this dropdown. The default value is the name of the first column header after "Barcodes".



Select samples column

Sample

Figure 19. Example of the "Select samples column" drop-down menu for a `metadata.csv` file.

  - Select samples to analyze—(optional) based on the sample column of interest from the previous option ("Select samples column"), you can restrict downstream analysis to barcodes identified by a specific value or set of values. The options available to select are dynamic and defined as all unique text entries present for the column of interest.

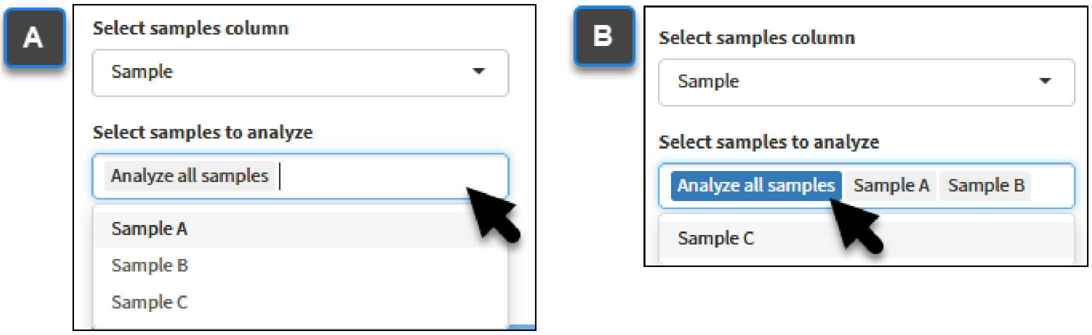
    Multiple values can be selected by clicking on the list value; the default value is 'Analyze all samples'.

NOTE: If custom values (a subset of 'all') are selected to restrict analysis, delete the 'Analyze all samples' entry from the input box.

Example:

Figure 20 illustrates the behavior of this option and is based on the metadata file shown in Figure 1. **Panel A** is the default view; the metadata file column of interest is 'Sample' (shown in the "Select samples column" box), and the options in the dropdown list below it are the values for 'Sample' in the metadata file.

Panel B shows what it would look like after selecting 'Sample A' and 'Sample B' to restrict downstream analysis. The default 'Analyze all samples' option is still listed; it should be removed by clicking on the name to highlight it (the white text on a blue background) and then clicking the **[Backspace]** key to remove it.



Panel A:

Select samples column: Sample

Select samples to analyze: Analyze all samples

Panel B:

Select samples column: Sample

Select samples to analyze: Analyze all samples, Sample A, Sample B

Figure 20. Example of the "Select samples to analyze" drop-down menu for a `metadata.csv` file. (Panel A) 'Analyze all samples' is selected, but the drop-down menu shows all the values contained within the "Sample" column (value of "Select samples column" option) in the metadata file. **(Panel B)** Example of 'Sample A' and 'Sample B' being added to analyze and selecting the third option ('Analyze all samples') in order to remove it from the input box.

- Select Gene Expression Profile for analysis (if using the "Processed Data from CogentAP" option)—allows you to choose between Exon+Intron or Exon-Only analysis
- Upload gene info file (Required) (if using the "Raw count matrix" option)—allows for upload of a gene info file from CogentAP

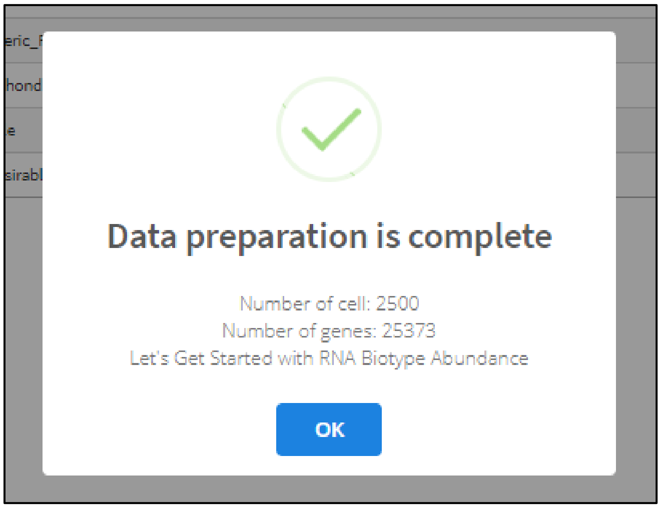
NOTE: If you have selected "Raw count matrix" and uploaded a .csv or .csv.gz file, you must upload the gene info file (info.csv) generated from CogentAP that contains gene annotations.

Click [Prepare Data for downstream analysis (Required)] to start the data preparation process. This action triggers the app to organize the uploaded data into a format suitable for downstream analysis.

<u>If you have selected "Example data":</u>

Click [Prepare Data for downstream analysis (Required)] to start the data preparation process. This action triggers the app to organize the uploaded data into a format suitable for downstream analysis.

When data preparation is complete, a *Data preparation is complete* popup window will appear (Figure 21). Click [OK].

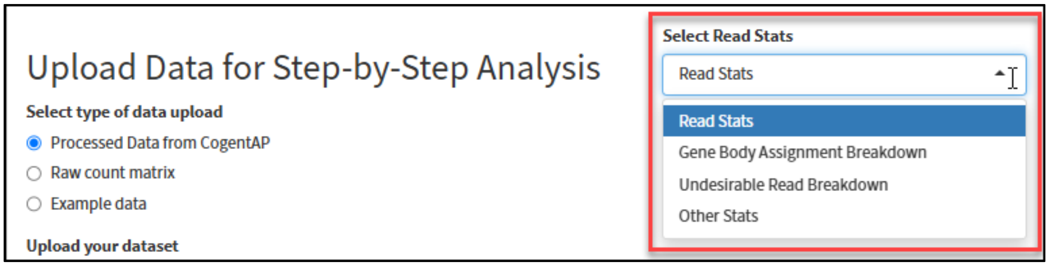


Data preparation is complete

Number of cell: 2500
Number of genes: 25373
Let's Get Started with RNA Biotype Abundance

Figure 21. *Data preparation is complete* popup window.

After data preparation, the "Select Read Stats" drop-down menu will appear to the left of the input fields (Figure 22).



Upload Data for Step-by-Step Analysis

Select type of data upload

● Processed Data from CogentAP
○ Raw count matrix
○ Example data

Upload your dataset

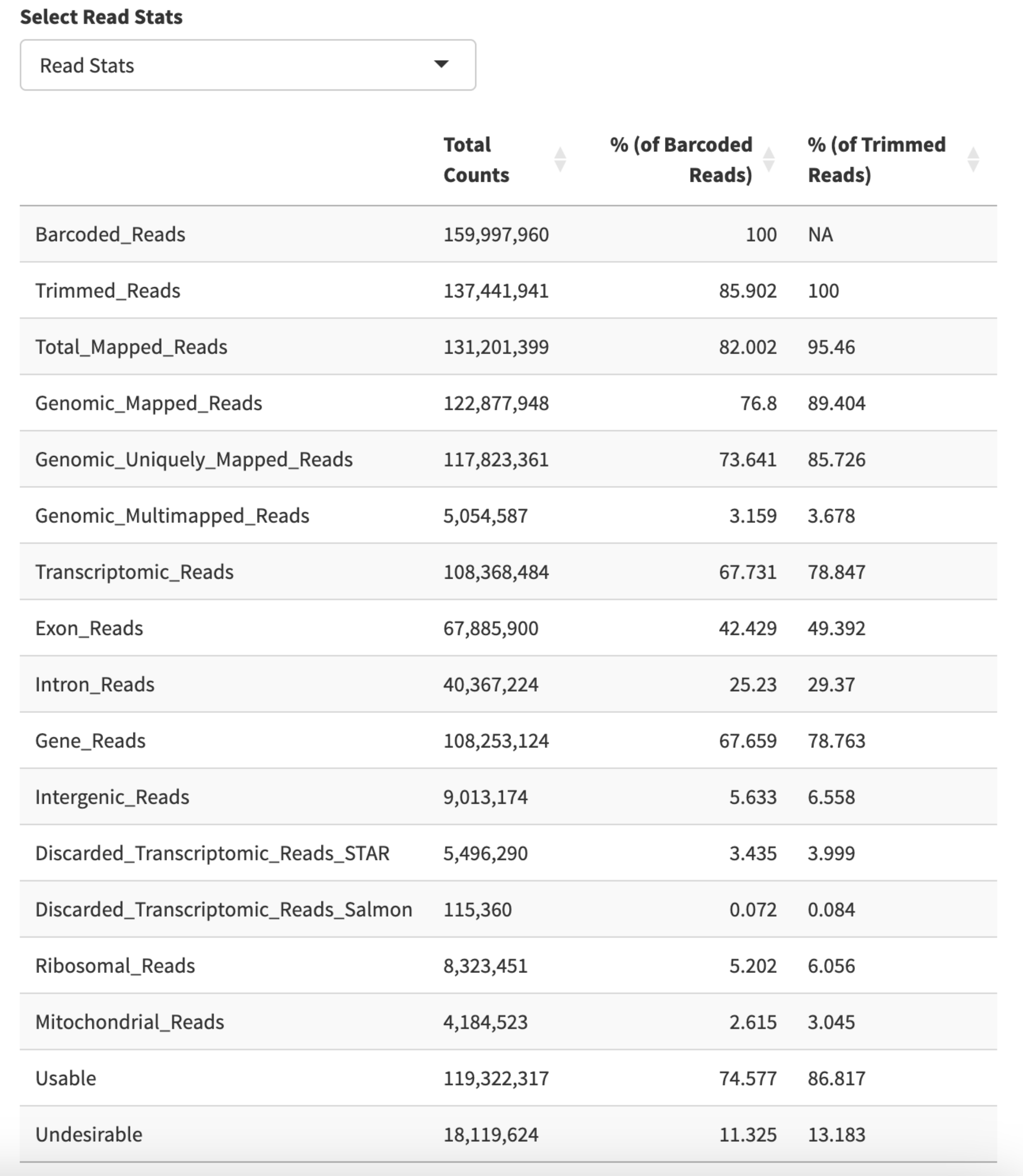
Select Read Stats: Read Stats

Figure 22. "Select Read Stats" drop-down menu.

Choosing an option from the drop-down menu will bring up a table containing those read statistics for your data. These tables will appear only if the necessary data is present.

There are four possible choices:

- Read Stats—displays metrics related to various types of reads, which may include barcoded reads, trimmed reads, mitochondrial reads, ribosomal reads (if SortmeRNA is run during the analysis step in CogentAP), mapped reads, unmapped reads, uniquely mapped reads, multimapped reads, usable reads, and undesirable reads. For each type, the total counts, along with the percentages of barcoded and trimmed reads, are shown (Figure 23).



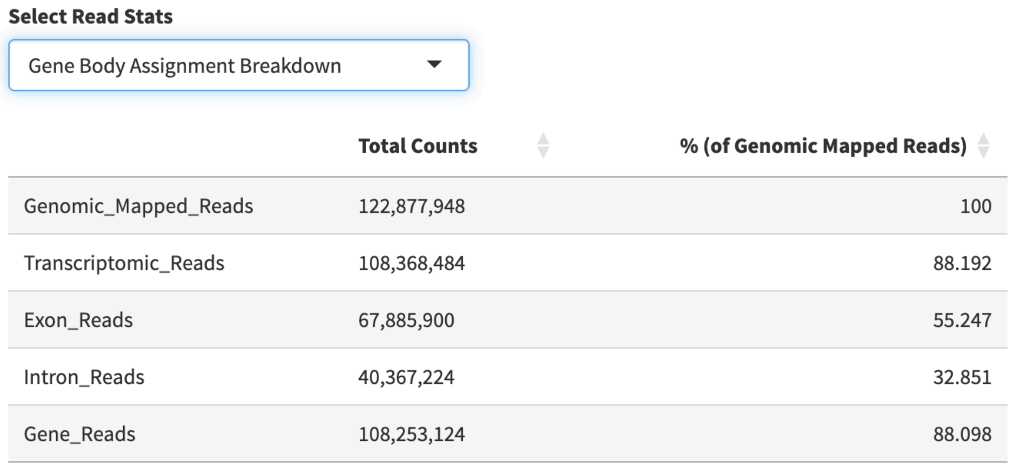
Select Read Stats

Read Stats ▾

	Total Counts	% (of Barcoded Reads)	% (of Trimmed Reads)
Barcoded_Reads	159,997,960	100	NA
Trimmed_Reads	137,441,941	85.902	100
Total_Mapped_Reads	131,201,399	82.002	95.46
Genomic_Mapped_Reads	122,877,948	76.8	89.404
Genomic_Uniquely_Mapped_Reads	117,823,361	73.641	85.726
Genomic_Multimapped_Reads	5,054,587	3.159	3.678
Transcriptomic_Reads	108,368,484	67.731	78.847
Exon_Reads	67,885,900	42.429	49.392
Intron_Reads	40,367,224	25.23	29.37
Gene_Reads	108,253,124	67.659	78.763
Intergenic_Reads	9,013,174	5.633	6.558
Discarded_Transcriptomic_Reads_STAR	5,496,290	3.435	3.999
Discarded_Transcriptomic_Reads_Salmon	115,360	0.072	0.084
Ribosomal_Reads	8,323,451	5.202	6.056
Mitochondrial_Reads	4,184,523	2.615	3.045
Usable	119,322,317	74.577	86.817
Undesirable	18,119,624	11.325	13.183

Figure 23. Read Stats table.

- Gene Body Assignment Breakdown—provides a breakdown of gene-body assignment, listing mapped reads, exon reads, intron reads, gene reads (exon + intron reads), and intergenic reads. For each type, the total counts and the percentage of mapped reads are shown (Figure 24).



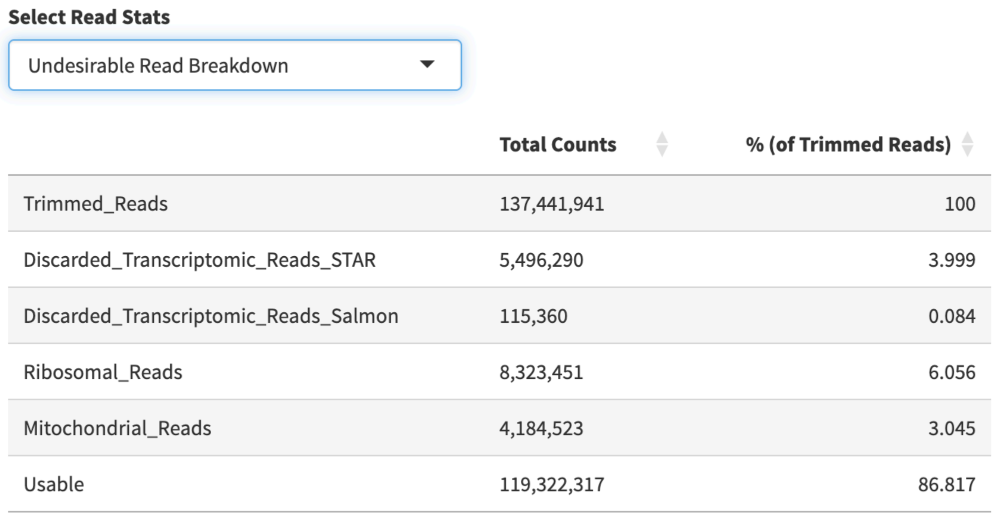
Select Read Stats

Gene Body Assignment Breakdown

	Total Counts	% (of Genomic Mapped Reads)
Genomic_Mapped_Reads	122,877,948	100
Transcriptomic_Reads	108,368,484	88.192
Exon_Reads	67,885,900	55.247
Intron_Reads	40,367,224	32.851
Gene_Reads	108,253,124	88.098

Figure 24. Gene Body Assignment Breakdown table.

- Undesirable Read Breakdown—categorizes various undesirable reads, including mapped reads, mitochondrial reads, ribosomal reads, and usable reads. Each type is accompanied by total counts and its percentage of mapped reads (Figure 25).



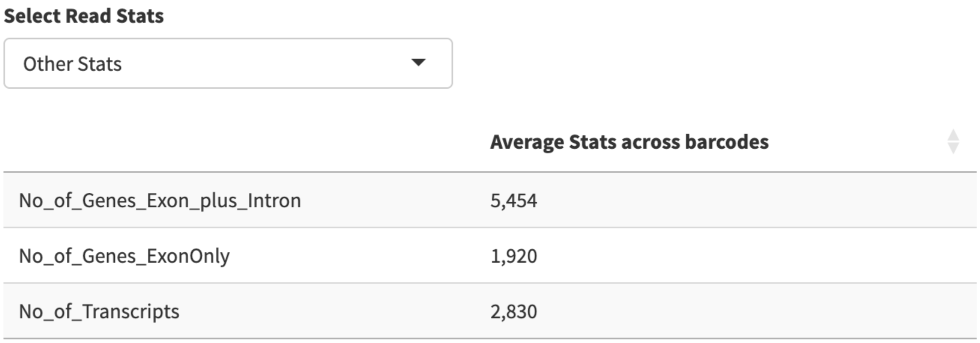
Select Read Stats

Undesirable Read Breakdown

	Total Counts	% (of Trimmed Reads)
Trimmed_Reads	137,441,941	100
Discarded_Transcriptomic_Reads_STAR	5,496,290	3.999
Discarded_Transcriptomic_Reads_Salmon	115,360	0.084
Ribosomal_Reads	8,323,451	6.056
Mitochondrial_Reads	4,184,523	3.045
Usable	119,322,317	86.817

Figure 25. Undesirable Read Breakdown table.

- Other Stats—includes additional stats such as the number of genes or transcripts and strand specificity, along with average statistics across barcodes (Figure 26).



Select Read Stats

Other Stats ▾

	Average Stats across barcodes
No_of_Genes_Exon_plus_Intron	5,454
No_of_Genes_ExonOnly	1,920
No_of_Transcripts	2,830

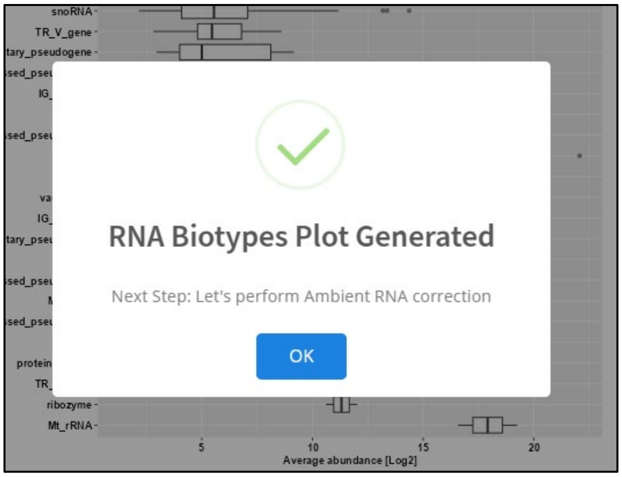
Figure 26. Other Stats table.

Click [Next: RNA Biotype Abundance] on the bottom right-hand corner of the screen to move to the next step of the scRNA-seq analysis. If you can't see the button, try scrolling down the browser page until it is visible.

2. RNA Biotype

The RNA Biotype step allows for visualization of the average abundance of a given RNA biotype across all cells in a dataset. It generates a boxplot that shows the distribution of the average abundance of each RNA biotype, using Ensembl annotations to represent each biotype (https://useast.ensembl.org/info/genome/genebuild/biotypes.html), and allows for comparisons of abundances between biotypes. If a metadata file is uploaded and >1 samples ares selected for analysis, then RNA-Biotype plot is split by sample for a more granular view.

Upon clicking [Next: RNA Biotype Abundance], the *RNA Biotypes Plot Generated* popup window will appear (Figure 27).

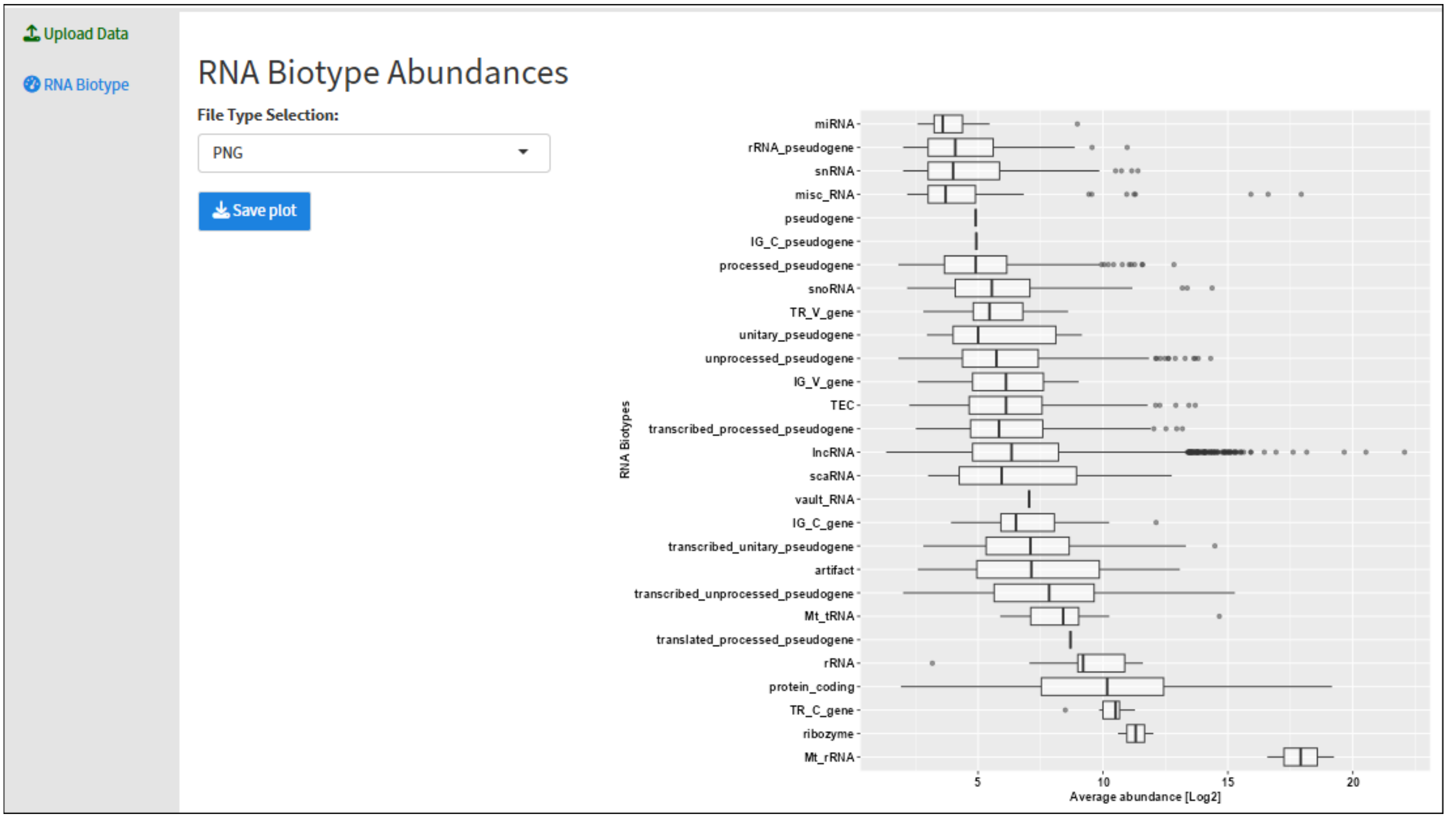


RNA Biotypes Plot Generated

Next Step: Let's perform Ambient RNA correction

Figure 27. *RNA Biotypes Plot Generated* popup window.

Click [OK] to remove the popup and view the RNA biotypes boxplot in the *RNA Biotype Abundances* window (Figure 28).



Upload Data

RNA Biotype

RNA Biotype Abundances

File Type Selection:

PNG

Figure 28. *RNA Biotype Abundances* window.

- The RNA biotypes plot can be downloaded as a .png, .pdf, .svg, or .jpeg file by choosing the desired file type from the "File Type Selection" drop-down menu and clicking the [Save plot] button (Figure 28, left side). The file will be saved to the "Downloads" folder associated with your browser.

- The left-nav sidebar has been updated with the new/current step. This happens at each step of the workflow; the sidebar list of modules can be used to navigate back to previous steps.

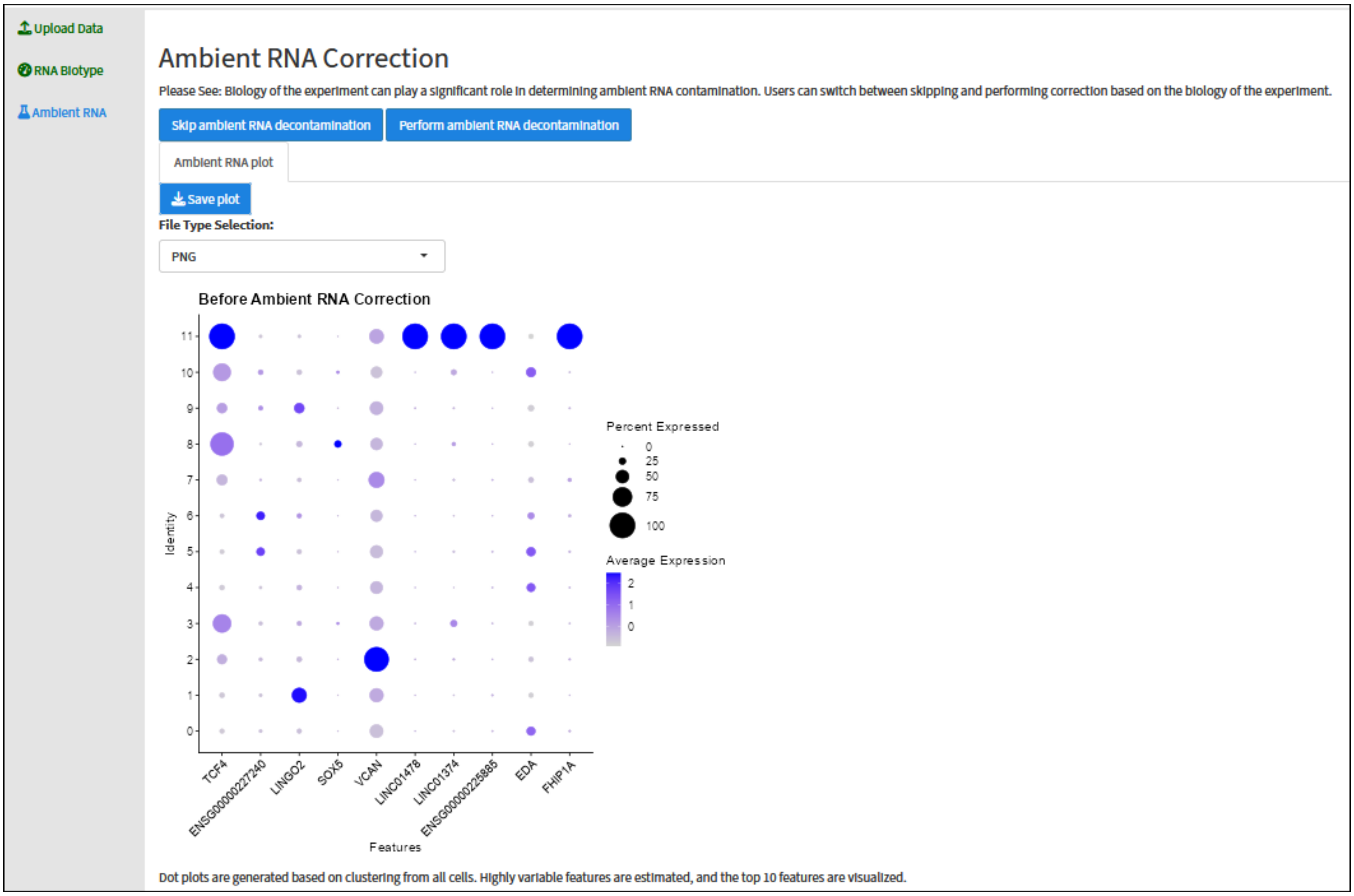
  NOTE: If you do navigate back to a previous step, change parameters, and run the analysis step on that page/module, you cannot jump forward and skip steps. You will need to proceed through each step again to generate the data with the new parameters applied, overwriting the previous calculations.

Click [Next: Perform Ambient RNA correction] to proceed to the next analysis step.

3. Ambient RNA

This step performs ambient RNA decontamination, which corrects for RNA contamination in individual cells.

The initial view in the *Ambient RNA Correction* window is shown in Figure 29, below. The dot plot displayed in the initial view is generated from the clustering of all cells, and the top 10 most highly variable features are highlighted.



Ambient RNA Correction

Please See: Biology of the experiment can play a significant role in determining ambient RNA contamination. Users can switch between skipping and performing correction based on the biology of the experiment.

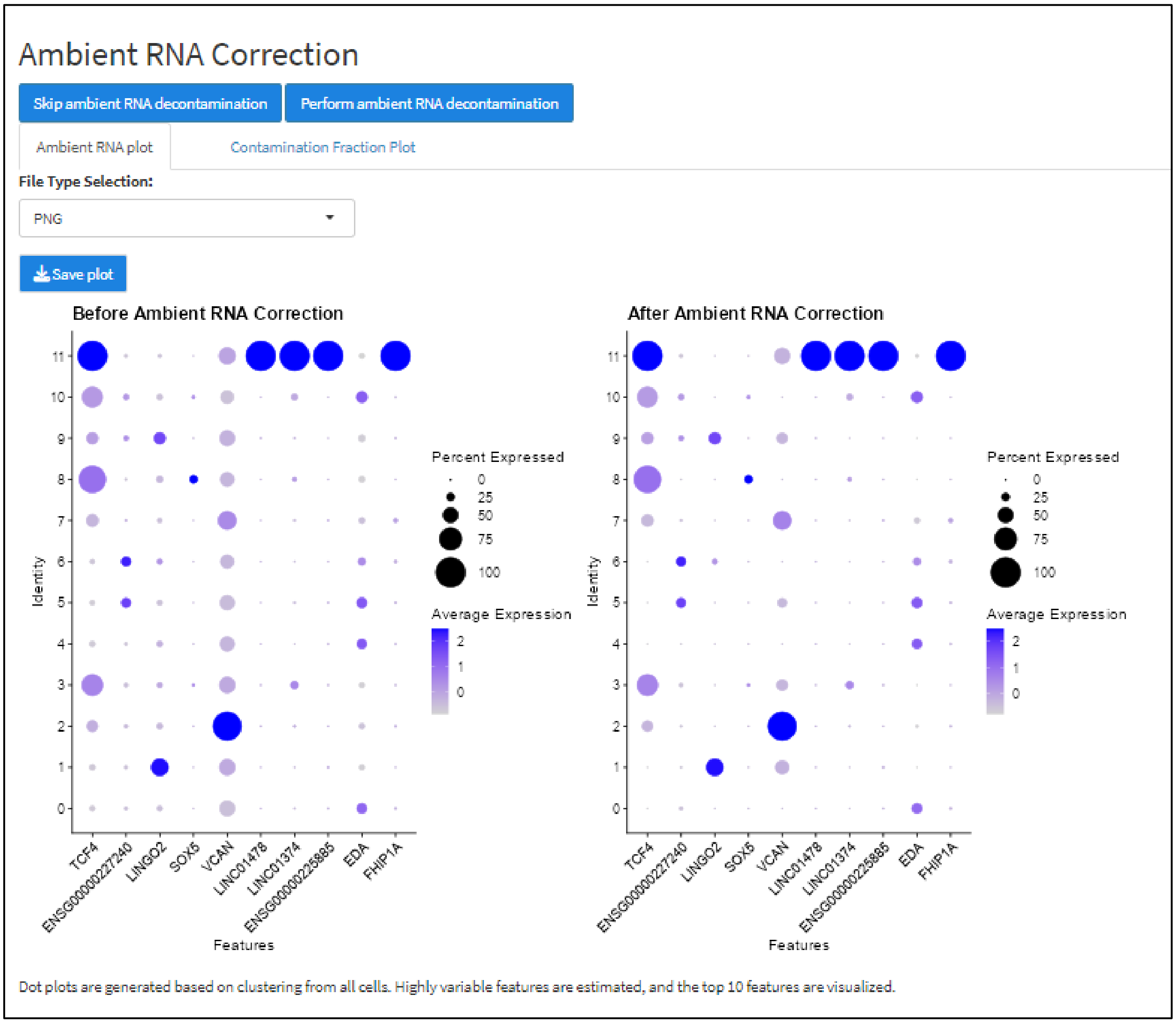
File Type Selection:

PNG

Dot plots are generated based on clustering from all cells. Highly variable features are estimated, and the top 10 features are visualized.

Figure 29. The initial view in the *Ambient RNA Correction* window.

From the initial view, ambient RNA correction can be performed by clicking [Perform ambient RNA decontamination]. After the decontamination process is complete, dot plots using data from before and after RNA decontamination are shown to allow you to see the impact of ambient RNA contamination on your data and how effective the decontamination process was in mitigating these effects (Figure 30).



Ambient RNA Correction

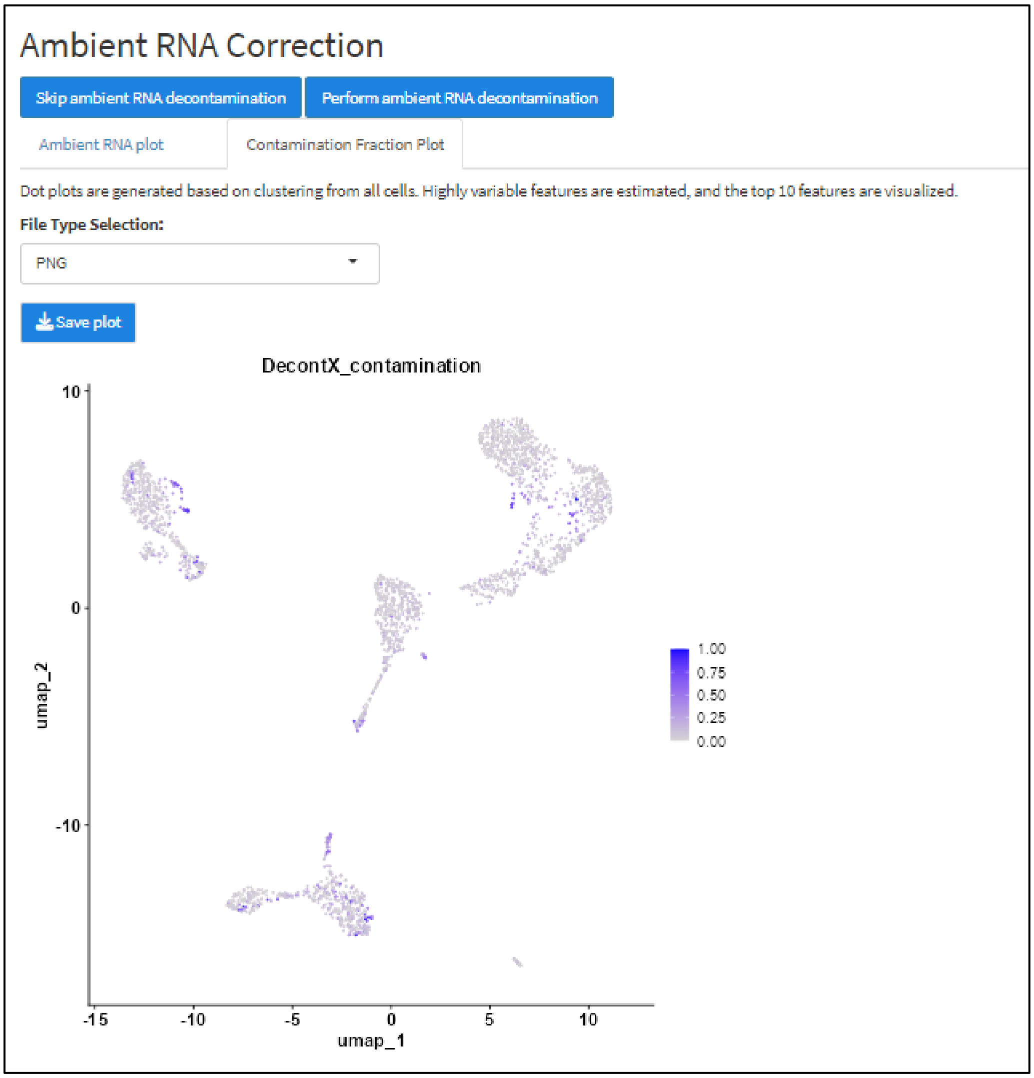
File Type Selection: PNG

Dot plots are generated based on clustering from all cells. Highly variable features are estimated, and the top 10 features are visualized.

Figure 30. The *Ambient RNA Correction* window after ambient RNA decontamination. The gene expression of several genes was affected by ambient RNA contamination, most notably LINGO2 and VCAN.

In addition to viewing the before and after ambient RNA correction dot plots, choosing the "Contamination Fraction Plot" option displays a UMAP plot showing the percentage of contamination in each cell, helping you identify which clusters may have high levels of ambient RNA contamination (Figure 31).

NOTE: The biology of the experiment can play a significant role in determining ambient RNA contamination. Users can switch between skipping and performing correction based on the biology of the experiment.



Ambient RNA Correction
Skip ambient RNA decontamination
Perform ambient RNA decontamination
Ambient RNA plot
Contamination Fraction Plot

Dot plots are generated based on clustering from all cells. Highly variable features are estimated, and the top 10 features are visualized.

File Type Selection: PNG

Figure 31. The contamination fraction plot after ambient RNA decontamination.

Ultimately, the plots available after ambient RNA decontamination are meant to help you make an informed decision on whether to perform ambient RNA decontamination.

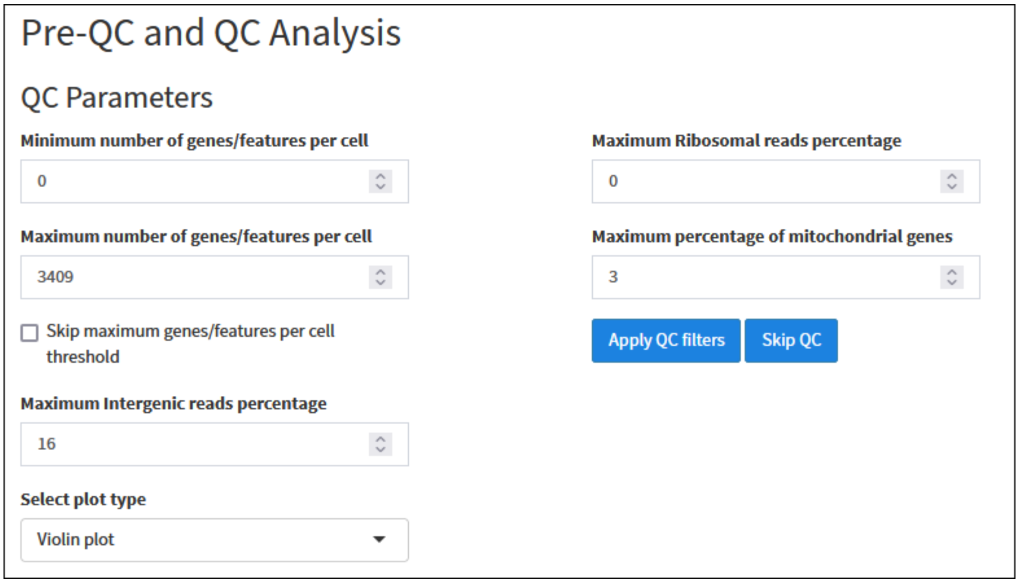
If you have performed ambient RNA decontamination but want to continue scRNA-seq analysis with uncorrected data, click [Skip Ambient RNA Decontamination]. The *Ambient RNA Correction Window* will revert to the initial view (Figure 29).

The dot plots and contamination fraction UMAP plot can be downloaded as a .png, .pdf, .svg, or .jpeg file by choosing the desired file type from the "File Type Selection" drop-down menu and clicking the [Save plot] button. The file will be saved respecting the configuration for downloads via your browser.

Click [Next: Perform QC] to proceed to the next analysis step

4. QC (Quality Control)

The QC step filters out cells from the dataset based on specified quality metrics. The initial view in the *Pre-QC and QC Analysis* window has two distinct sections: QC Parameters (Figure 32) and Pre-QC Plots (Figure 33).



Pre-QC and QC Analysis

QC Parameters

Minimum number of genes/features per cell: 0

Maximum number of genes/features per cell: 3409

□ Skip maximum genes/features per cell threshold

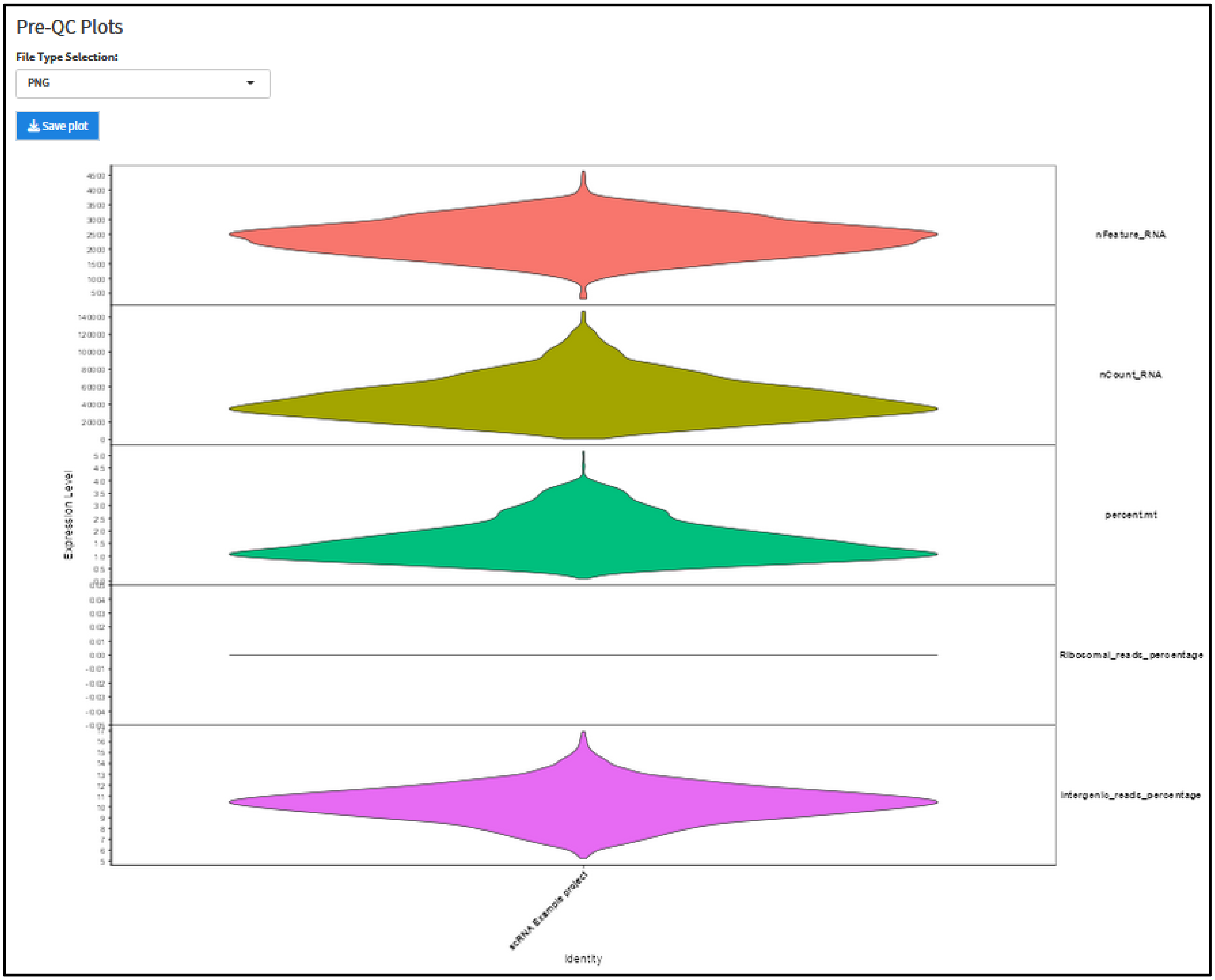
Maximum Intergenic reads percentage: 16

Select plot type: Violin plot

Maximum Ribosomal reads percentage: 0

Maximum percentage of mitochondrial genes: 3

Figure 32. The QC Parameters section in the initial view of the *Pre-QC and QC Analysis* window.



Pre-QC Plots

File Type Selection: PNG

Figure 33. The Pre-QC Plots section of the *Pre-QC and QC Analysis* window. 'Violin plot' is chosen under "Select plot type" in the QC Parameters section.

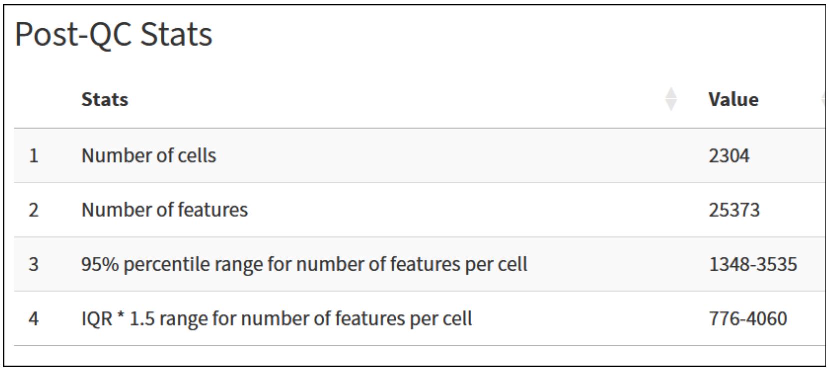
The desired QC parameters can be set or selected in the QC Parameters section (Figure 32). The options/ functionalities available include:

- Minimum number of genes/features per cell—filters cells based on a minimum threshold for number of genes or features each cell must have to be included in the analysis, as cells with very few features might be of low quality.
- Maximum number of genes/features per cell—filters cells based on a maximum threshold for the number of genes or features in each cell. The default value is set to the 95th percentile of the number of features across all cells in the dataset. This setting helps exclude cells with an unusually high number of features, which might indicate doublets or multiplets.
- Skip maximum genes/features per cell threshold—checkbox to bypass setting a maximum threshold. Selecting this option allows you to proceed without excluding any cells based on the number of features identified.

- Maximum Ribosomal reads percentage—filters cells based on the maximum percentage of reads that map to ribosomal RNA. Cells with high ribosomal content could be low-quality or stressed cells. The default value is calculated using the median absolute deviation (MAD) method to ensure appropriate thresholds are set for the dataset.
- Maximum percentage of mitochondrial genes—filters cells based on the maximum percentage of reads that map to mitochondrial genes. Cells with high mitochondrial contamination are often considered to be of low quality or dying. The default threshold is set to the 90th percentile of mitochondrial gene percentage across all cells in the dataset.
- Maximum Intergenic reads percentage—filters cells based on the chosen maximum percentage of reads that map to intergenic regions. This filter helps to keep cells in the analysis where reads mostly map to RNA and not in the intergenic reads, which can increase the quality of cells in the analysis.
- Select plot type—allows for selection of the plot type to display, either violin or scatter plot.
    - When 'Violin plot' is selected, all violin plots for the key metrics are displayed:
        - nFeature_RNA—number of genes detected per cell
        - nCount_RNA—number of mapped gene reads detected per cell
        - percent.mt—percentage of mitochondrial gene counts
        - Ribosomal_reads_percentage—percentage of reads mapping to ribosomal RNA
        - Intergenic_reads_percentage—percentage of reads mapping to intergenic RNA
    - 'Scatter plot' shows two charts, with nCount_RNA on the X axis of both and percent.mt and nFeature_RNA on the Y axis, respectively. Pearson correlation values are displayed for each chart as well.

After adjusting all the desired changes, click the [Apply QC filters] button to have them take effect and reflect in the data presented. The module generates:

- A table with statistics:
    - The number of cells and features
    - The 95th percentile range for the number of features per cell
    - The interquartile range (IQR) * 1.5 range for the number of features per cell to aid in identifying outliers



Post-QC Stats

	Stats	Value
1	Number of cells	2304
2	Number of features	25373
3	95% percentile range for number of features per cell	1348-3535
4	IQR * 1.5 range for number of features per cell	776-4060

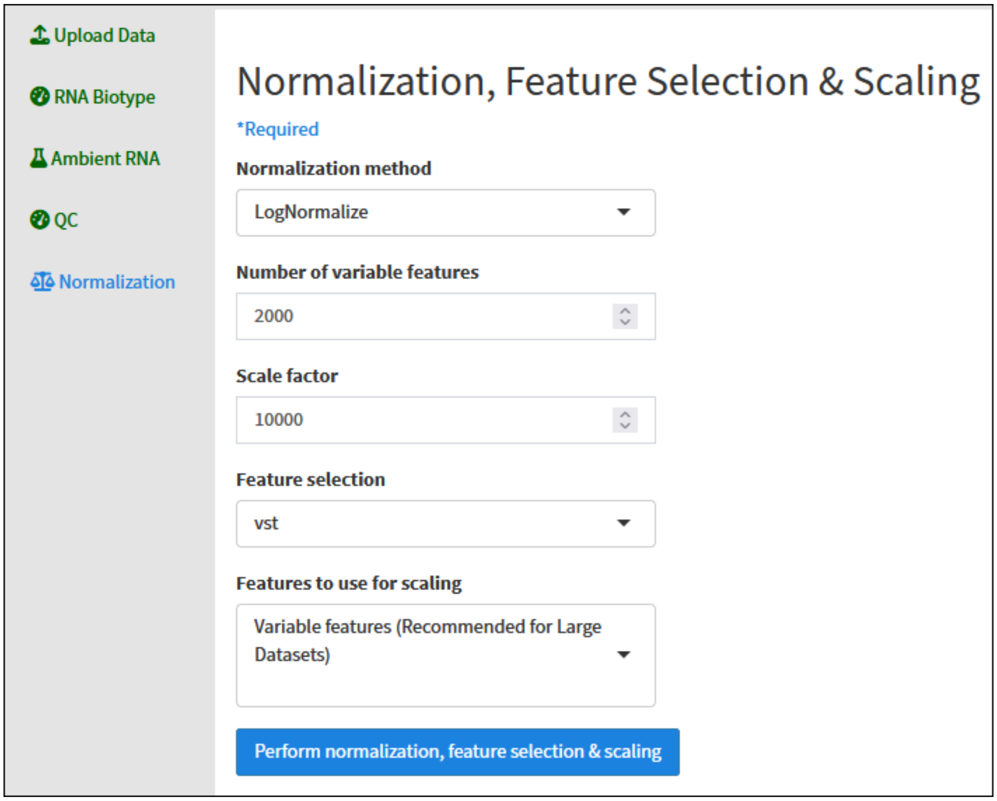
Figure 34. Example Post-QC Stats table.

- A set of post-QC plots of the type selected in the "Select Plot Type" option. Like the pre-QC plots, the post-QC plots can be downloaded in PNG, PDF, SVG, or JPEG format.

Click [Next: Perform Normalization, Feature Selection & Scaling] to proceed to the next analysis step.

5. Normalization, Feature Selection & Scaling

IMPORTANT: Before navigating away from the module, you will need to click the [Perform normalization, feature selection & scaling] button.


Normalization, Feature Selection & Scaling

*Required

Normalization method: LogNormalize

Number of variable features: 2000

Scale factor: 10000

Feature selection: vst

Features to use for scaling: Variable features (Recommended for Large Datasets)

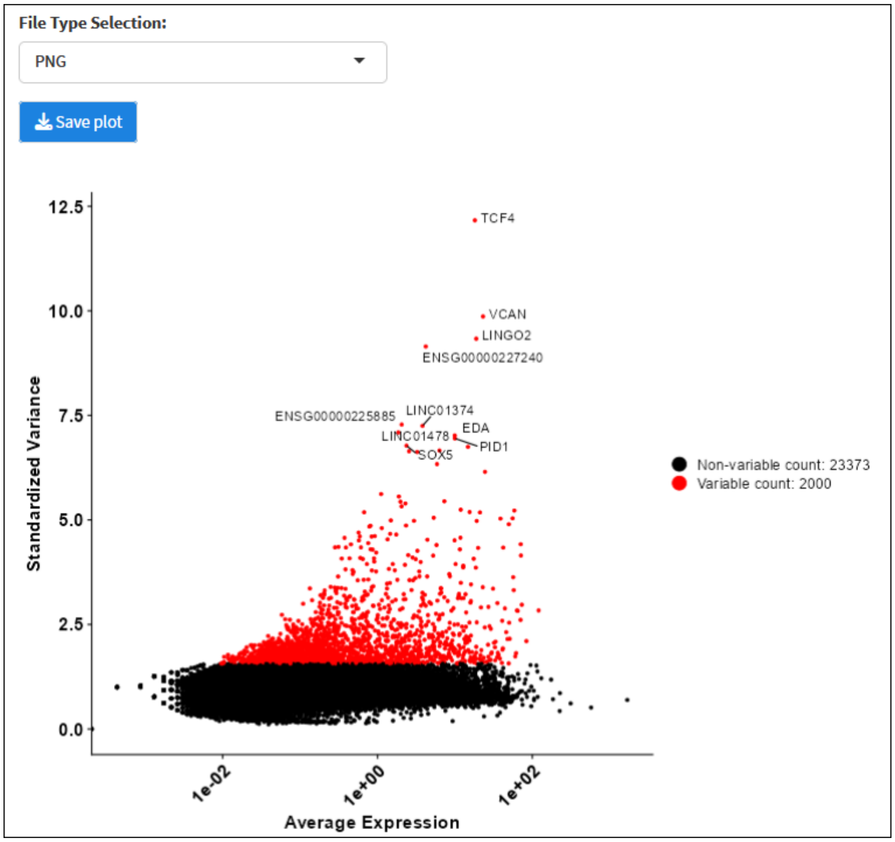
Figure 35. The initial normalization (*Normalization, Feature Selection, & Scaling*) view.

This module helps to normalize data, select features using the vst (variance-stabilizing transformation) method, and apply scaling. Below are the options available; for more information on these methods and parameters, please refer to the Seurat documentation (https://satijalab.org/seurat/).

- Select Normalization Method from the drop-down menu. Options include:
  - LogNormalize—normalizes the feature counts for each cell by dividing by the total counts for that cell and multiplying by scale factor. The default value for scale factor is 10,000. The resulting values are then transformed by natural log.
  - SCT: (SCTransform)—SCTransform is a statistical approach for normalizing single-cell datasets. It uses a generalized linear model (GLM) to show the relationship between sequencing depth and gene expression. This method normalizes data while stabilizing variance. CogentDS uses SCT v2, which is default in Seurat v5.

- Number of variable features—specifies the number of features you wish to include.
- Scale factor—the desired scale factor for normalization.
- Feature selection—the only option is 'vst'.
- Features to use for scaling—chose whether to scale variable features only or all features. For large datasets, it is recommended to choose the 'scale variable features' option as it is faster. Scale all features is computationally expensive, requiring more memory usage and time to complete.

Perform the analysis by clicking the button [Perform Normalization, feature selection & scaling] with the selected settings. After the calculations are complete, a plot will display on the page with the results.



File Type Selection:

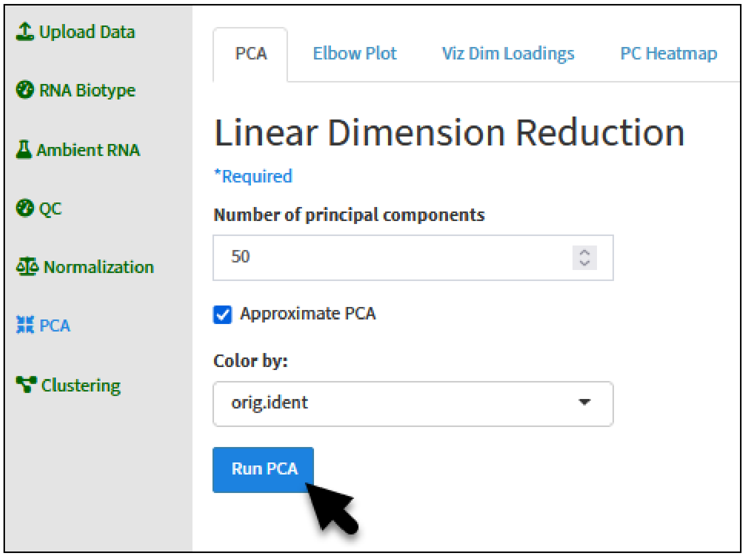
PNG

Figure 36. Normalized plot results after applying feature selections and scaling parameters.

Click [Next: Perform PCA Analysis] to proceed to the next analysis step.

6. Principal Component Analysis (PCA)

IMPORTANT: Before navigating away from the default *PCA* tab to one of the other tabs, you will need to click the [Run PCA] button.



Linear Dimension Reduction

*Required

Number of principal components: 50

☑ Approximate PCA

Color by: orig.ident

Figure 37. Default *PCA* analysis (Linear Dimension Reduction) window and tab. Before proceeding, click the [Run PCA] button, as indicated in the image.

This module allows you to perform Principal Component Analysis. This module includes four tabs: PCA, Elbow Plot, Viz Dim Loadings (Dimensional Loadings), and PC Heatmap.

Modify the parameters (described below), if desired, but click [Run PCA] before navigating to any of the other tabs.

Refer to the following sections for high-level information about each tab. For more in-depth details on PCA, the associated visualization, and parameters, please refer to the Seurat (https://satijalab.org/seurat/) documentation.

When ready to proceed to the next section, click [Next: Perform Clustering] at the bottom right corner of the screen.

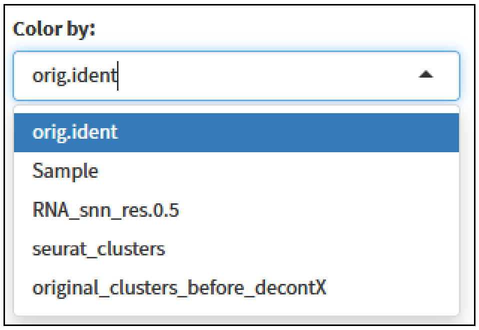
a) PCA

This tab allows users to configure principal component analysis settings and generate and display a PCA plot.

The available parameters include:

- Number of principal components—specify the number of principal components to be calculated. The default number of principal components is 50, but this can be adjusted based on specific analysis needs.
- Approximate PCA—option to determine whether to use an approximate algorithm for PCA. By default, this parameter is checked to be enabled.

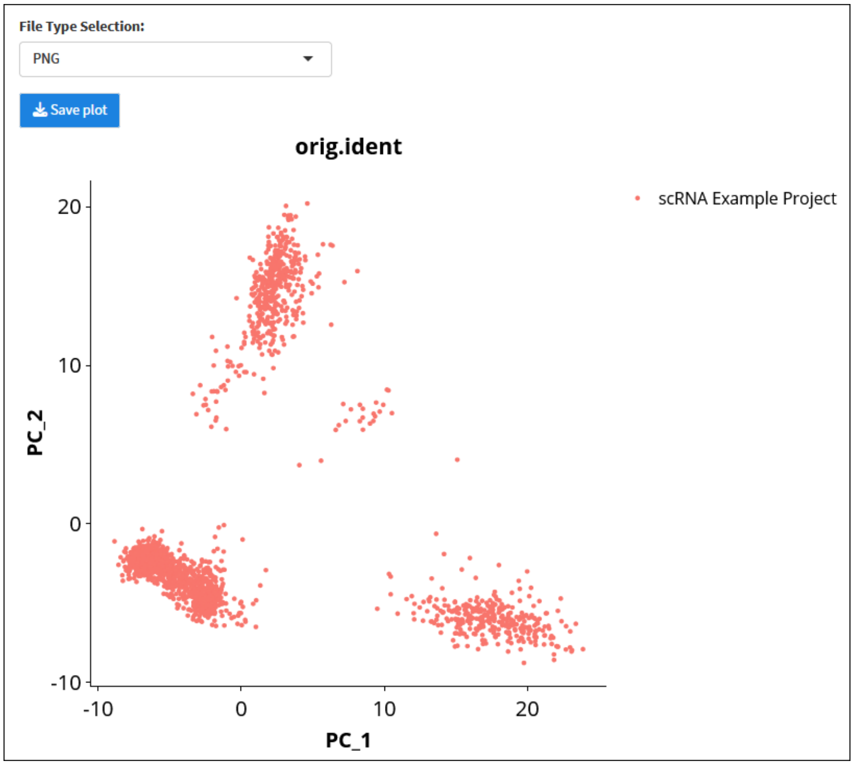
- Color by—select a metadata attribute by which the PCA plot points will be colored.



Color by: orig.ident

Figure 38. The "Color by" menu options on the *PCA* tab. The options listed in the drop-down menu will depend on the column headers of your metadata information.

After making your selections and clicking [Run PCA] or making any change to the "Color by" option, a scatter plot is generated showing the first principal component (PC1) vs. the second principal component (PC2).



File Type Selection: PNG

Figure 39. Default *PCA* plot.

This plot allows some customization through the floating menu visible only when mousing over the upper right corner of the plot (Figure 40). For more details on the options available in the floating menu, see the Appendix.

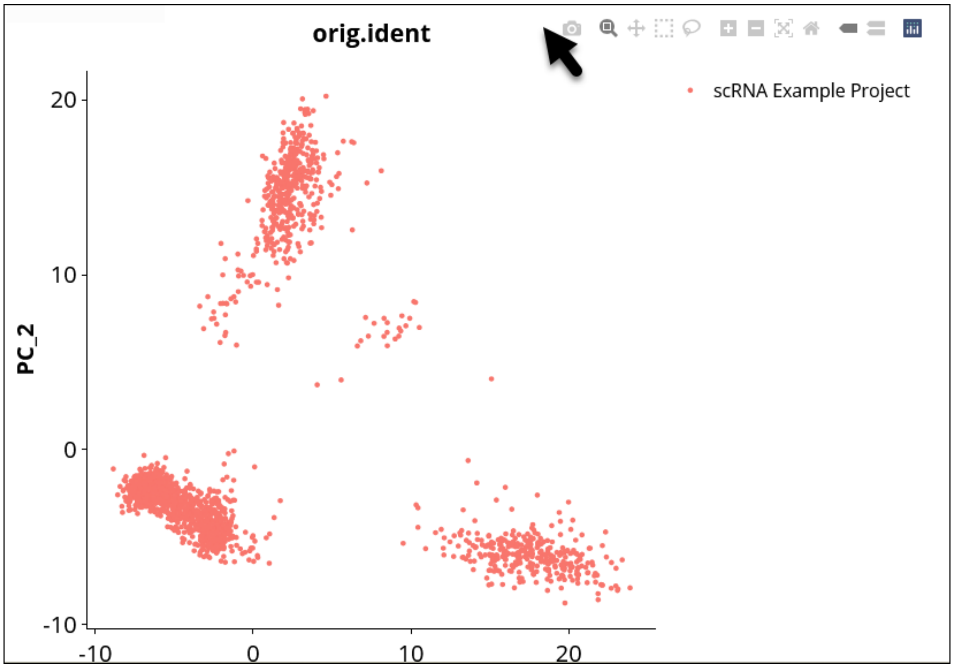


Figure 40. Chart modification floating menu location. The black arrow is demonstrating how to place the mouse cursor in order for the menu to become visible.

b) Elbow Plot

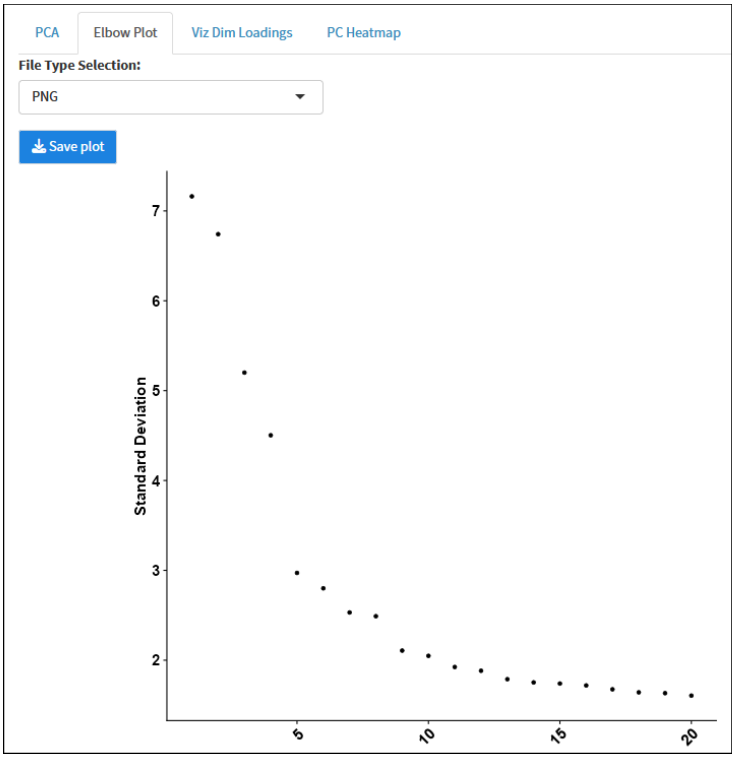


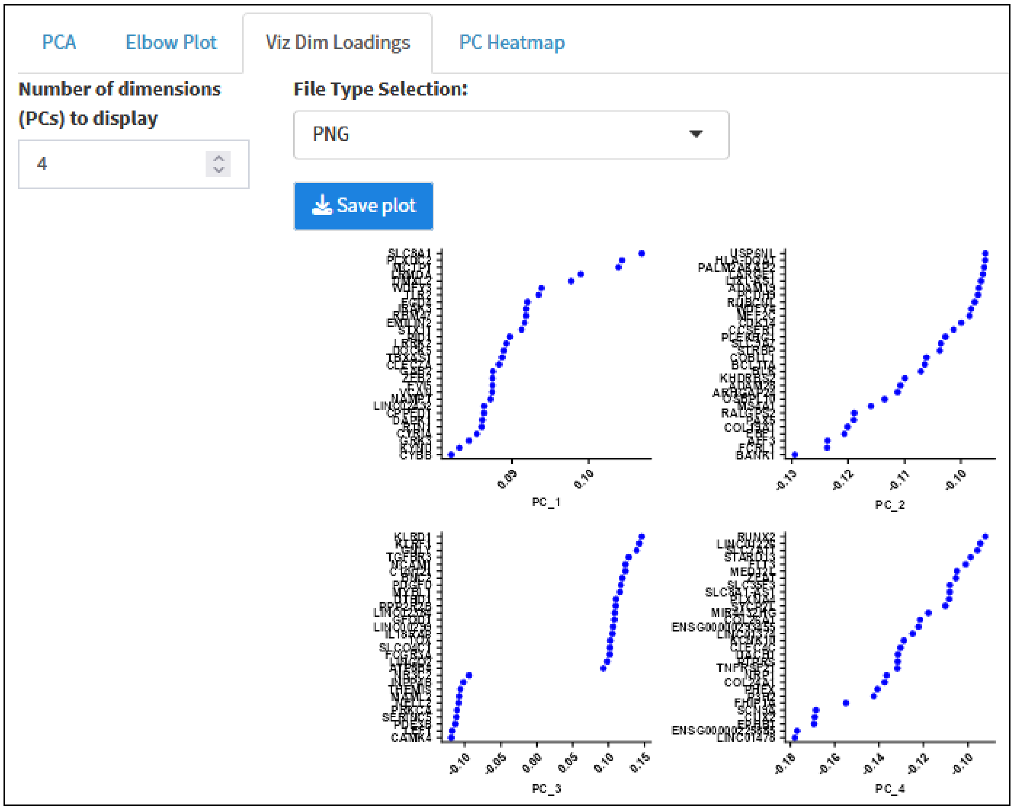
Figure 41. Default *Elbow Plot* tab chart and view.

The elbow plot helps to visualize the variance (standard deviations) calculated for each principal component, assisting in the identification of the "elbow point" where the explained variance (slope of the curve) decreases sharply. This point is used to determine the number of principal components to retain for further analysis.

c) *Viz Dim Loadings (Dimensional Loadings)*

This tab displays the loadings of features/genes, which can be helpful in understanding the contribution of each feature/gene to the principal component.

The "Number of Dimensions (PCs) to display" parameter specifies how many principal components (PCs) are visualized in the loading plots. These plots show the top genes with the most significant positive and negative loadings for each selected PC, highlighting their contributions to the observed variance in the data.



Number of dimensions (PCs) to display: 4

File Type Selection: PNG

Figure 42. Example *Viz Dim Loadings* tab and charts. The default number of dimensions (charts) that display is '2'; a customized value of '4' is shown to demonstrate how the value impacts the display.

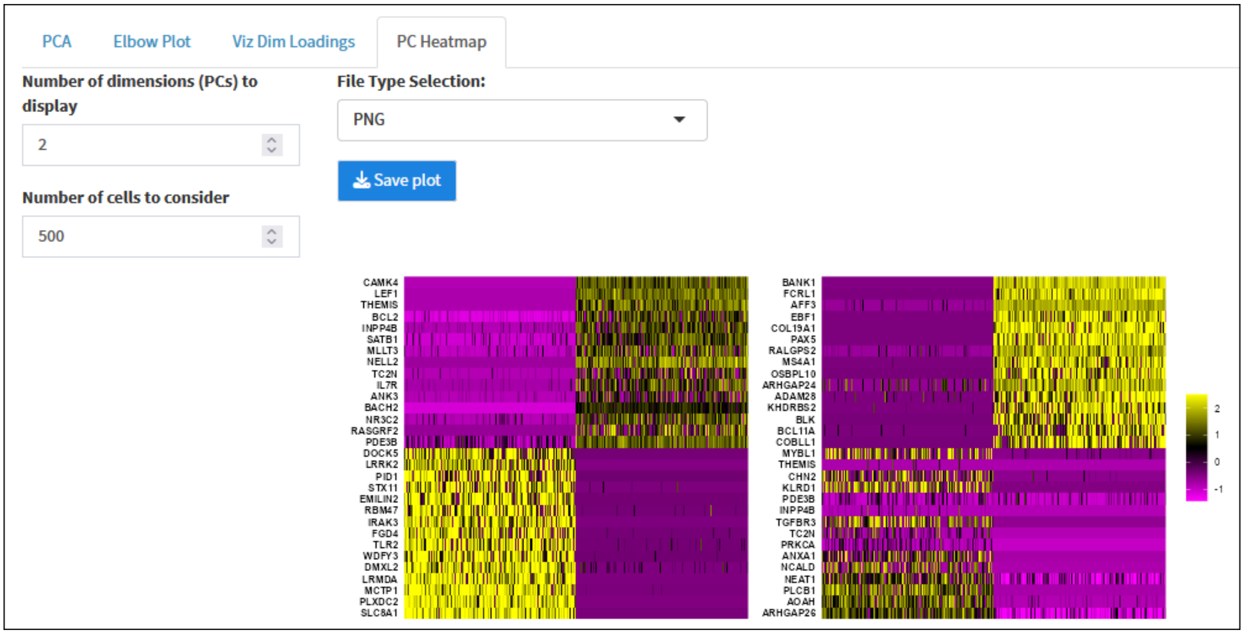
d) *PC (Principal Component) Heatmap*

This tab displays a heatmap focusing on principal components. Cells and features are ordered based on their PC scores, allowing for easy investigation into sources of heterogeneity in the dataset.

Parameters:

- Number of dimensions (PCs) to Display—how many of the PCs (dimensions) to be displayed as a heatmap. For example, setting it to '2' will display a heatmap for each of the first two principal components in order (not based on importance or ranking). Increasing the value will increase the number of heatmaps generated.

- Number of cells to consider—the number of cells to include in the heatmap. The default is set to 500, meaning that the top 500 cells based on their principal component scores are plotted. Adjust this value based on the need of your analysis.



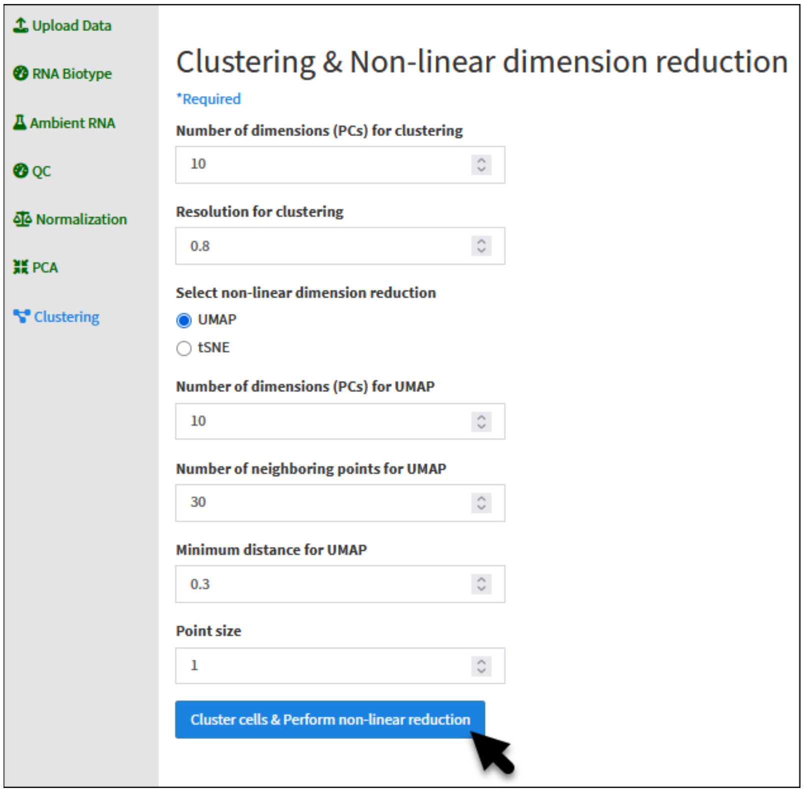
Number of dimensions (PCs) to display: 2

File Type Selection: PNG

Number of cells to consider: 500

Figure 43. Example *PC Heatmap* tab and charts.

7. Clustering & Non-Linear Dimension Reduction



Clustering & Non-linear dimension reduction

*Required

Number of dimensions (PCs) for clustering: 10

Resolution for clustering: 0.8

Select non-linear dimension reduction

● UMAP

○ tSNE

Number of dimensions (PCs) for UMAP: 10

Number of neighboring points for UMAP: 30

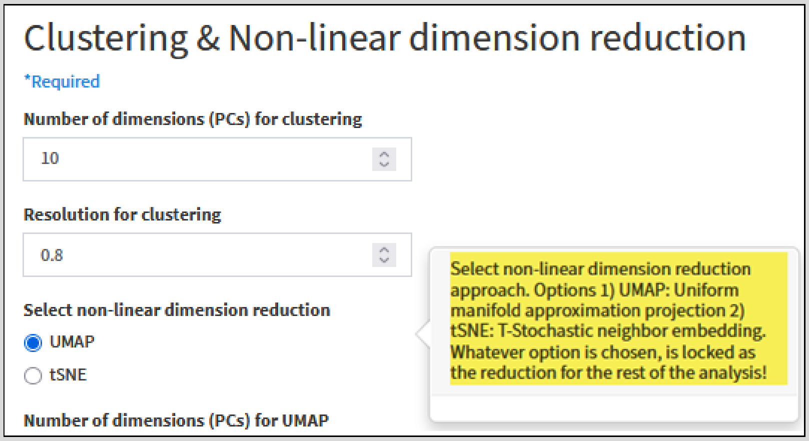
Minimum distance for UMAP: 0.3

Point size: 1

Figure 44. Default *Clustering & Non-linear dimension reduction* display and options. 'UMAP' is selected by default; the options below that subsection are specific to the UMAP chart-type. Select 'tSNE' to view the t-SNE-specific options.

This module facilitates clustering and non-linear dimension reduction to return UMAP (Uniform Manifold Approximation and Projection) or t-SNE (t-Distributed Stochastic Neighbor Embedding) charts based on the specified parameters. Pages downstream from this step will adapt based on the dimension reduction method selected here. For information on functions and parameters, please refer to the Seurat documentation.

NOTE: Once the cell clustering and non-linear reduction button is pushed, the option to select UMAP or t-SNE is locked out. To select the other option, restart the app.



Clustering & Non-linear dimension reduction

*Required

Number of dimensions (PCs) for clustering: 10

Resolution for clustering: 0.8

Select non-linear dimension reduction

● UMAP

○ tSNE

Number of dimensions (PCs) for UMAP

Select non-linear dimension reduction approach. Options 1) UMAP: Uniform manifold approximation projection 2) tSNE: T-Stochastic neighbor embedding. Whatever option is chosen, is locked as the reduction for the rest of the analysis!

Figure 45. Pop-up warning message on the *Clustering and Non-linear dimension reduction* page for the reduction method selection.

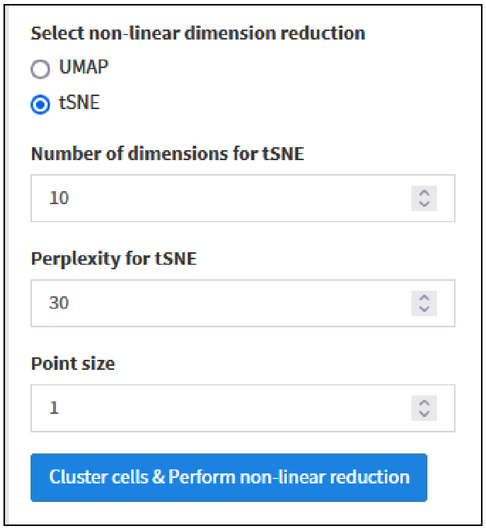
The options above and below the radio button option can be grouped as follows:

- Clustering parameters
  - Number of Dimensions (PCs) for Clustering—this parameter sets the number of principal components to use for clustering. It is suggested to select the number of dimensions based on the elbow plot obtained from the previous PCA module. The default value is '10'.
  - Resolution for Clustering—this parameter allows users to control resolution of clustering. A higher value results in more communities, while a lower value results in fewer communities. The default value is '0.8'.
- Non-linear dimension reduction parameters

  The choice of dimension reduction (UMAP or t-SNE) will determine the available parameters and the downstream visualization.

  - When UMAP is selected, the following parameters are available:
    - Number of dimensions (PCs) for UMAP—the number of dimensions to use for UMAP. The default value is '10'.
    - Number of neighboring points for UMAP— the number of neighboring points used in local approximations. The default value is '30'.

    - Minimum distance for UMAP—controls how closely points are packed together in the UMAP embedding. Larger values spread points more evenly, while smaller values allow the algorithm to focus more on local structure. The default value is '0.3'.
    - Point size—adjust the size of points in the UMAP plot. The default value is '1'.
  - When t-SNE is selected, the following parameters are available:
    - Number of dimensions to use for t-SNE—defines the number of dimensions to be used for t-SNE.
    - Perplexity for tSNE—Perplexity is a key tunable parameter for t-SNE, which influences the balance between local and global structure in the data.
    - Point Size—adjusts the size of points in the t-SNE plot



Select non-linear dimension reduction

○ UMAP

● tSNE

Number of dimensions for tSNE: 10

Perplexity for tSNE: 30

Point size: 1

Figure 46. Non-linear dimension reduction parameters for t-SNE plots.

After setting the desired parameters, click [Cluster cells & Perform non-linear reduction] to perform clustering and visualize dimension reduction plots overlaid with clusters based on selected configurations.

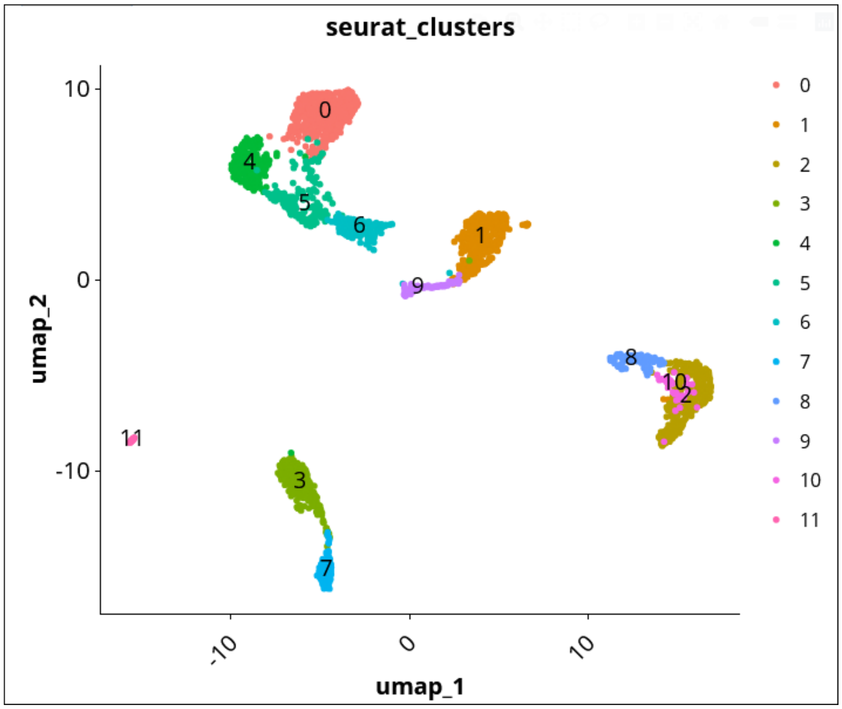
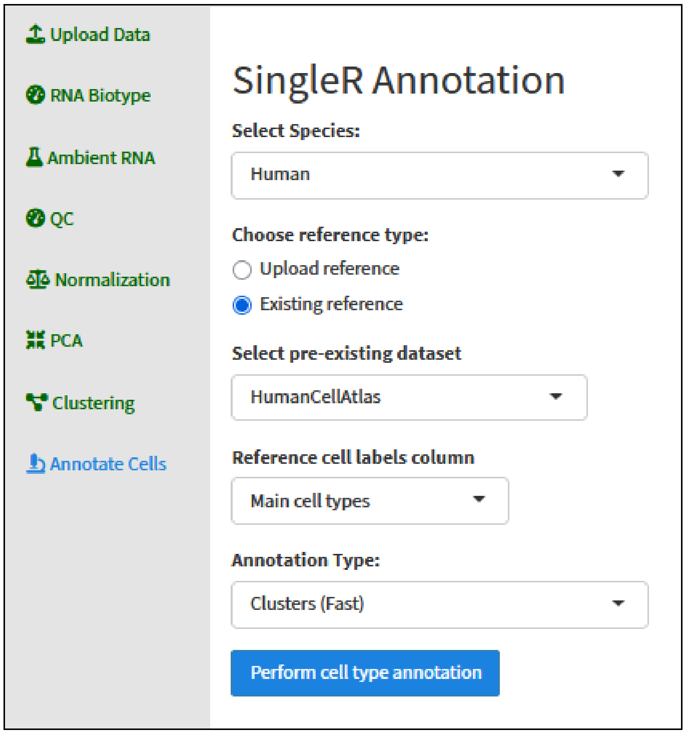


Figure 47. Example UMAP output after performing cell clustering and non-linear reduction.

The resulting plots may also be interacted with via a floating menu visible only when hovering the mouse over the top right corner of the chart. Please see the Appendix for more information about the menu options.

Click [Next: Perform Cell Type Annotation] to proceed to the next analysis step.

8. Annotate Cells



SingleR Annotation

Select Species: Human

Choose reference type:
○ Upload reference
● Existing reference

Select pre-existing dataset: HumanCellAtlas

Reference cell labels column: Main cell types

Annotation Type: Clusters (Fast)

Figure 48. Default *Annotate Cells* display and options. 'Existing reference' is selected by default; the options below that subsection are specific to that option. Select 'Upload reference' to view parameters specific to using your own reference file.

Users have the option to annotate their cell types, either by uploading their own reference data or selecting a pre-existing reference from the celldex R package (https://bioconductor.org/packages/release/data/experiment/html/celldex.html). The tool can work with both Seurat and SingleCellExperiment (SCE) objects as reference. For more detailed information, users can refer to the documentation for the SingleR and celldex R packages.

- Select Species—choose the appropriate species based on your data. The available options are human and mouse
- Choose Reference Type
  - Upload Reference Data
    - Upload your own reference data, which can be in the form of a Seurat object or an SCE object
    - Reference Cell Labels Column—use the dropdown to select the metadata column name from the uploaded reference that contains cell labels
  - If Pre-existing Reference is selected:

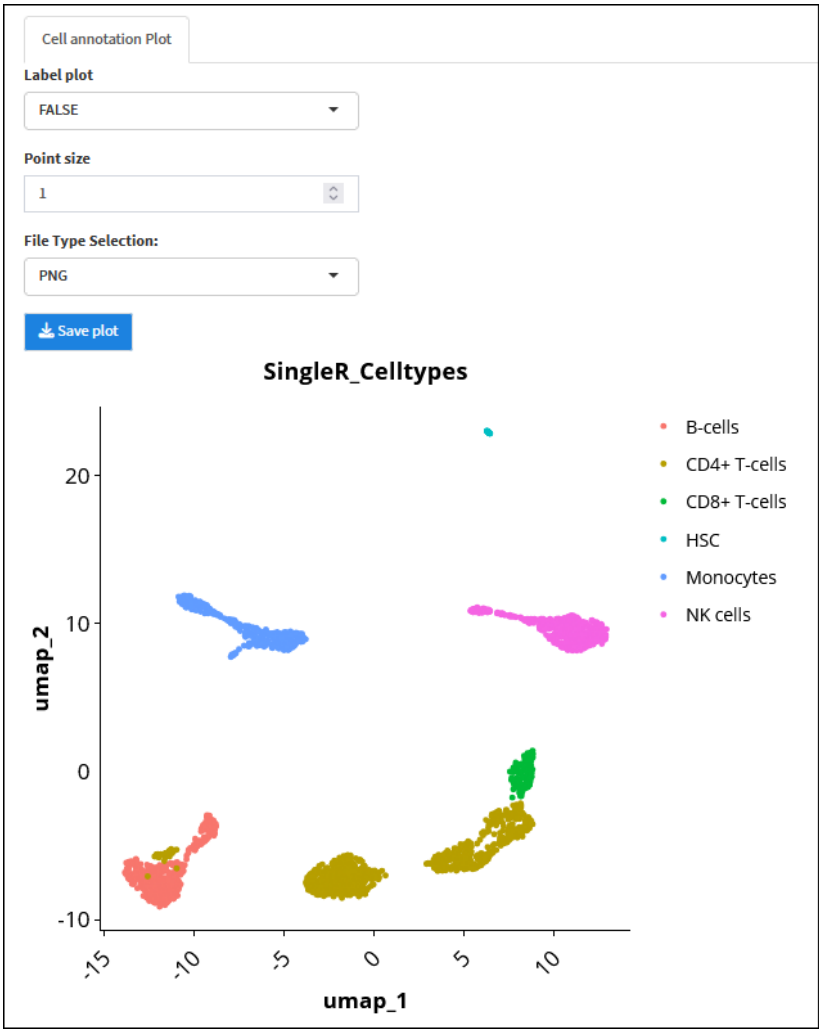
    This module uses the celldex R package to fetch pre-existing reference datasets based on the selected species (Human or Mouse).
    - Human Reference Datasets

      When Human is selected, the following reference datasets are available in the drop-down menu:
      - HumanCellAtlas—Human Primary Cell Atlas Data, which represents a broad range of human primary cells.
      - BluePrintEncode—Blueprint Encode Data, which represents bulk RNA-seq data from Blueprint and ENCODE. It consists of stromal and immune cells.
      - ImmuneDatabase—obtains bulk RNA-seq data of immune cell population from the Database of Immune Cell Expression (DICE)
      - Hematopoietic Data—retrieves bulk microarray expression data for sorted hematopoietic cells
      - ImmuneMonaco—bulk RNA-seq dataset for immune cells.
    - Mouse Reference Datasets

      When Mouse is selected, the following reference datasets are available:
      - MouseRNAseqData—bulk RNA-seq expression data of sorted cell population in mice.
      - ImmGenData—normalized microarray expression data from pure populations of murine cells, provided by the Immunologic Genome Project.
- Reference Cell Labels Column—use the drop-down menu to select the column name from the chosen pre-existing reference that contains cell labels.
- Annotation Type—the app offers two annotation methods:

- Cluster-Based Annotation—fast, completing the annotation of aggregated cluster profiles in about 2–3 minutes.
- Cell Type Annotation—per cell annotation that can take 10–15 min for large datasets, such as those with 100,000 cells.

Click on [Perform cell type annotation] button to initiate the cell type annotation using the selected reference data. A dimension reduction plot will be displayed, overlaid with the predicted cell types.



Label plot

FALSE

Point size

1

File Type Selection:

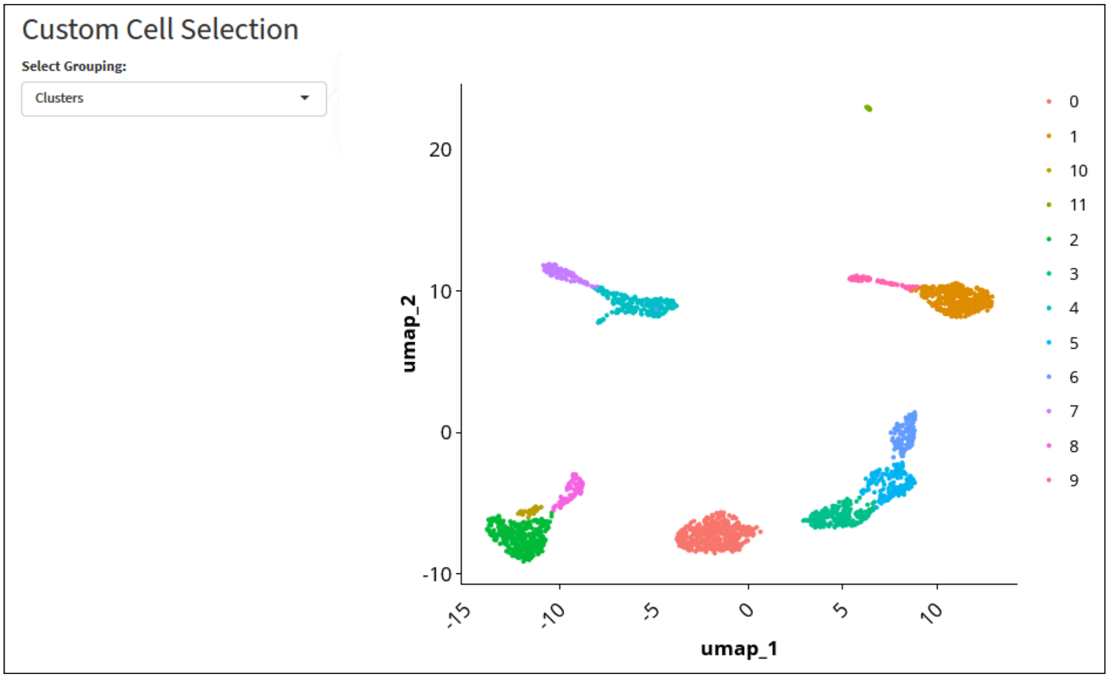
PNG

Figure 49. Chart resulting from a cell annotation example.

The resulting plots may also be interacted with via a floating menu visible only when hovering the mouse over the top right corner of the chart. Please see the Appendix for more information about the menu options.

Click [Next: Perform Custom Lasso Selection] to proceed to the next analysis step.

9. Custom Lasso Selection



Custom Cell Selection

Select Grouping:

Clusters

Figure 50. Example Custom Cell Selection initial page. The chart shown when you first access this step is carried over from previous analysis steps.

The custom selection module allows users to interactively select groups of cells based on clustering or cell type annotations. These selections can then be named and saved for subsequent differential gene expression analysis.

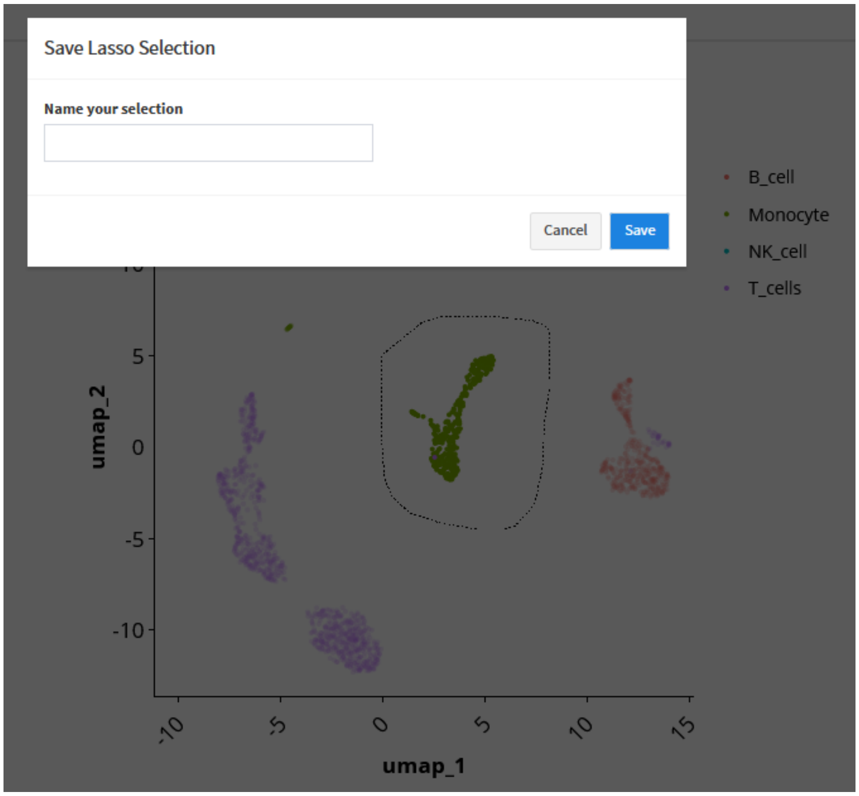
Details on this module are provided below:

a) *Select Groupings:*

- Choose between Clusters or Cell Types from the drop-down menu.
- By default, a Dimension reduction plot is displayed overlaid with Clusters.
- If the Cell Type Annotation module has been run and Cell Types is selected, the plot will be overlaid with cell types.

b) *Performing Lasso Selection*

- Use the lasso option of the floating menu (Appendix, Section C) to make selections on the plot. Multiple selections are allowed.
- Each selection can be named for easy identification.
- As each selection is made, a *Save Lasso Selection* dialog window will pop up with a text input field, "Name your Selection". Input the name you want to apply to the designated/selected cluster and click [Save] to apply your change, or [Cancel] to quit the selection with no changes applied.

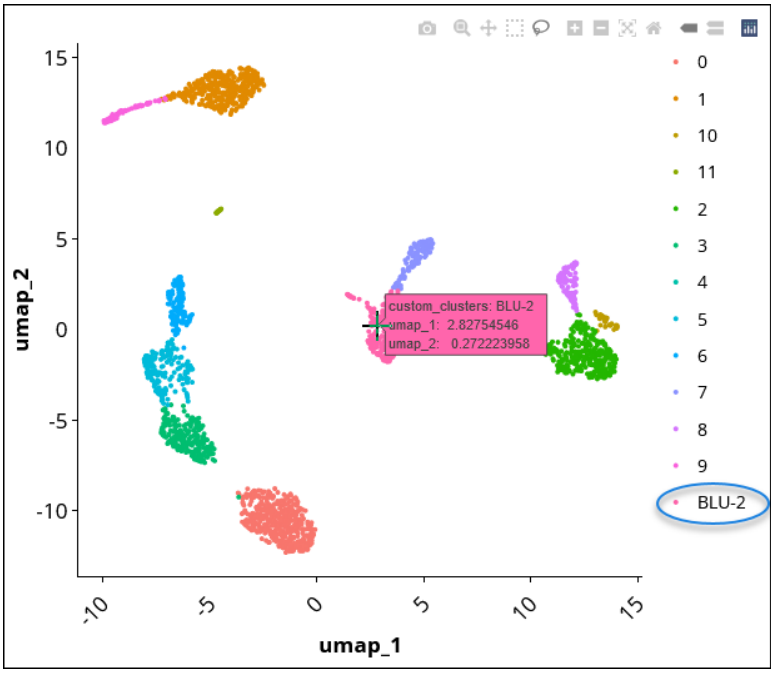


Save Lasso Selection

Name your selection

Figure 51. *Save Lasso Selection* dialog window after using the lasso selection tool.

- Enter the name for your selection in the text input field.
- Click the [Save] button to store your selection



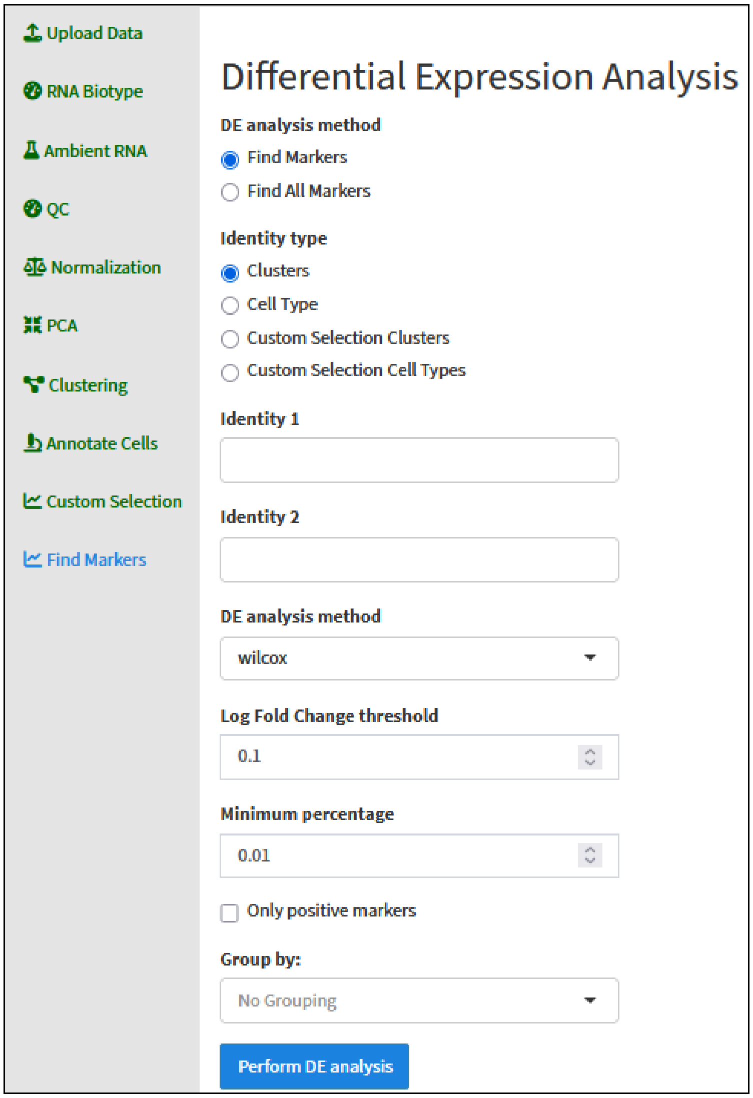
custom_clusters: BLU-2
umap_1: 2.82754546
umap_2: 0.272223958

Figure 52. Example results of custom naming a cluster selected with the chart [Lasso] function. The name given to the custom cluster in Figure 51, BLU-2, is now listed on the legend. When hovering the mouse cursor over the cluster (center), the pop-up box now shows it is a "custom_clusters" with the name 'BLU-2'.

Once selections are named and saved, they can be used in the subsequent Differential Gene Expression Analysis module, Find Markers. More details on the Find Markers module are provided in the next section.

Click [Next: Perform Differential Gene Expression Analysis] to proceed to the next analysis step.

10. Differential Expression Analysis (Find Markers)



Differential Expression Analysis

DE analysis method

● Find Markers

○ Find All Markers

Identity type

● Clusters

○ Cell Type

○ Custom Selection Clusters

○ Custom Selection Cell Types

Identity 1

Identity 2

DE analysis method: wilcox

Log Fold Change threshold: 0.1

Minimum percentage: 0.01

□ Only positive markers

Group by: No Grouping

Figure 53. *Differential Expression Analysis* default page.

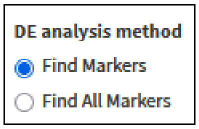
The Find Markers module (*Differential Expression Analysis*) allows users to identify differentially expressed genes between specified groups of cells. This module offers various options for DE analysis methods, parameters, and visualization.

The options on this page have both shared and analysis-specific settings. To simplify the documentation, the options have been grouped following that divide rather than in the order they're listed on the page.

If you have any questions about using the page, please reach out to technical support.

a) *Workflow*

1. Begin by selecting which DE analysis method you wish to use. The two options are described below.



DE analysis method

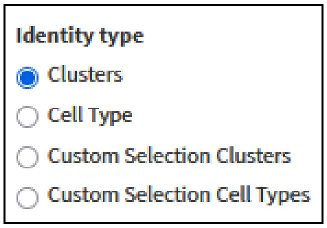
● Find Markers

○ Find All Markers

Figure 54. The "DE analysis method" selection of the *Differential Expression Analysis* default page.

- Find Markers—identifies markers between specified identities. This option allows users to compare two groups of identities (clusters or cell types) and find genes that are differentially expressed between them. For the additional parameters specific to this option type, refer to Section b), Find Markers Parameters.
- Find All Markers—identifies markers for all identities in the dataset. It compares each identity against all others, providing overview of differentially expressed genes across all identities. For the additional parameters specific to this option type, refer to Section c), Find All Markers Parameters.

2. Choose Identity Type—selects the type of identity for the analysis.



Identity type

● Clusters

○ Cell Type

○ Custom Selection Clusters

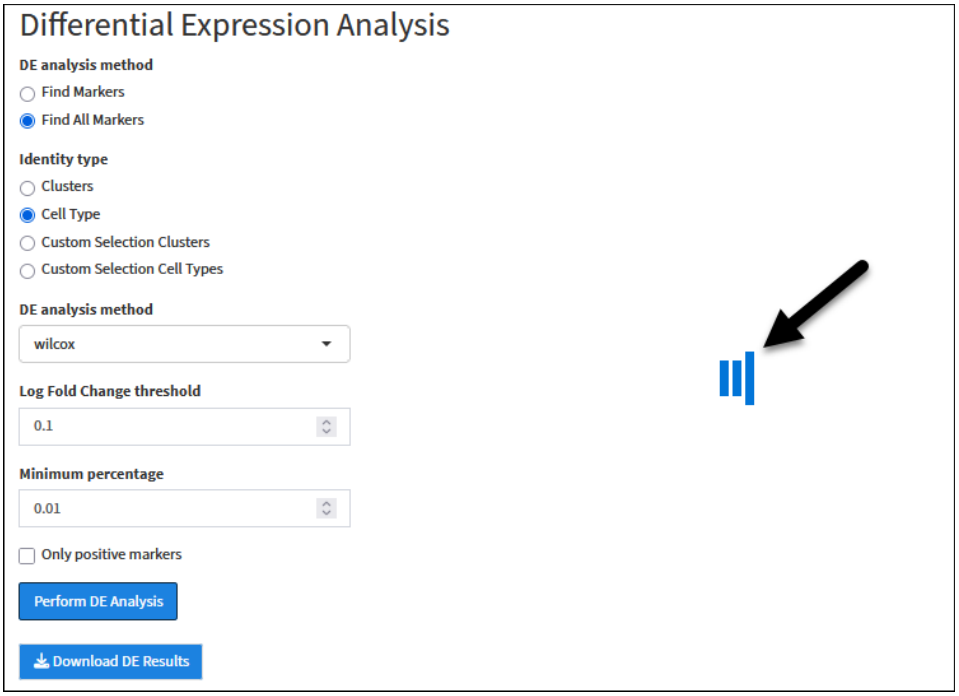
○ Custom Selection Cell Types

Figure 55. The "Identity type" selection of the *Differential Expression Analysis* default page.

- Clusters—uses predefined clusters in your Seurat object. This option leverages the clustering information from your Seurat object for the analysis.
- Cell Type—uses annotated cell types from the cell type annotation module. If the cell type annotation module has been run, the Seurat object will contain these annotated cell types. Selecting this option will perform differential expression analysis based on the annotated cell type.
- Custom Selection Clusters—if you used the previous module (Section VI.A.9, "Custom Lasso Selection") to custom name any clusters on the UMAP, use this option to have access to the custom cluster name in the Identity 1 or Identity 2 drop-down menus alongside the original, predefined clusters.
- Custom Selection Cell Types—similar to 'Custom Selection Clusters', use this option to have access to any custom cell types defined in the previous module in the Identity drop-down menus.

3. Choose DE Analysis Method— the method for differential gene expression analysis:
    - Wilcox—uses the Wilcoxon ran sum test method to find differential gene expression. This is the default method.
    - bimod—uses a Likelihood-ratio test for single-cell gene expression datasets.
    - roc—uses ROC analysis for identifying markers.
    - LR—logistic regression to identify differentially expressed genes.
    - t—uses a Student's t-test to find differentially expressed genes between two groups of cells.
    - negbinom—uses a negative generalized linear model to identify differentially expressed genes between two groups of cells. It is suggested to use this option for UMI-based datasets only.
    - poisson—uses poisson generalized linear model for finding differentially expressed genes between two groups of cells. It is suggested to use this option for UMI-based datasets only.
    - MAST—identifies differentially expressed genes in single-cell RNA-seq data using the hurdle model.
    - Deseq2—uses negative binomial distribution to determine differentially expressed genes.
4. Set the Log Fold Change Threshold—restricts the testing of genes or features that show at least a specified log fold change between two cell groups. By default, this value is 0.1.
5. Set the Minimum Percentage—sets a threshold for the minimum fraction of cells in either group that must express a gene for it to be tested. By default, this value is 0.01.
6. Check or uncheck the Only Positive Markers—checking this box considers only genes that are positively differentially expressed in the analysis. By default, it is FALSE.
7. For information on the remaining options, refer to the appropriate section below.
8. Click on the [Perform DE Analysis] button, and a section of postanalysis plot parameters specific to the analysis type will display after the button. Again, refer to the analysis-specific section information to fill these fields out.

    NOTE: While the software is waiting for the postanalysis plot parameters to be filled out, an animated icon (indicated in still format in Figure 56, with the arrow pointing to the animation caught in progress). This will persist until the next step.



Differential Expression Analysis

DE analysis method

○ Find Markers

● Find All Markers

Identity type

○ Clusters

● Cell Type

○ Custom Selection Clusters

○ Custom Selection Cell Types

DE analysis method: wilcox

Log Fold Change threshold: 0.1

Minimum percentage: 0.01

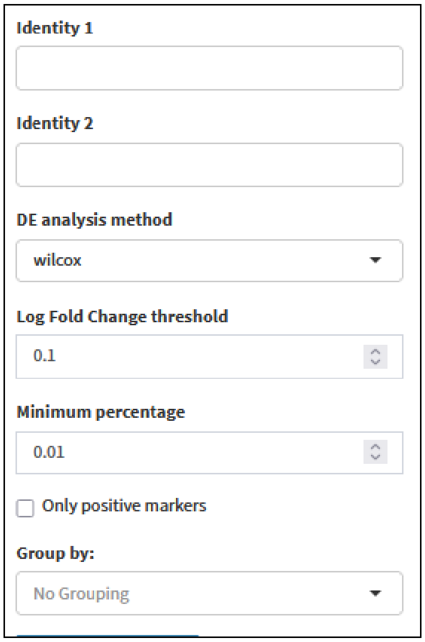
□ Only positive markers

Figure 56. In-progress icon on the *Differential Expression Analysis* page midway through the analysis steps.

9. After filling out the postanalysis plot parameters, click the [Generate plots] button. This will replace the animated in-progress icon with the respective chart.
10. View or download the chart in the desired format, specified by the drop-down menu.
11. Click [Next: Perform Pathway Analysis] to proceed to the next module.

b) *Find Markers Parameters*

(1) Preanalysis Parameters



Identity 1

Identity 2

DE analysis method: wilcox

Log Fold Change threshold: 0.1

Minimum percentage: 0.01

□ Only positive markers

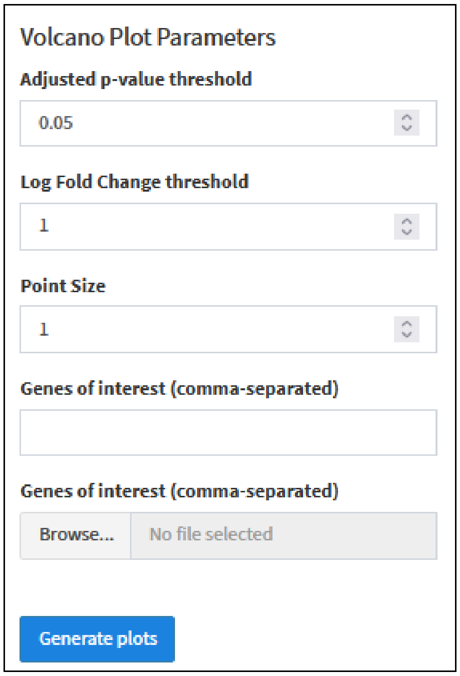
Group by: No Grouping

Figure 57. Find Markers parameters. The section includes parameters in common with "Find All Markers" and some specific to the "Find Markers" option. The options are described below.

When Find Markers is selected as the DE analysis method, along with any identity type, specify the following parameters:

- Identity 1—(required) allows for the selection of one or more identities (clusters or cell types) for comparison. You can choose multiple clusters or cell types to be included in this first group (For example, Cluster 1 and Cluster 2).
- Identity 2—(required) allows for the selection of one or more identities (clusters or cell types) to compare against Identity 1. Similarly, multiple clusters or cell types can be included in this group (For example, Cluster 3 and Cluster 4)
- Group by Option—allows users to specify a metadata column in Seurat Object that defines how cells are grouped before differential gene expression analysis

(2) Postanalysis Plot Parameters: Volcano Plot



Volcano Plot Parameters

Adjusted p-value threshold: 0.05

Log Fold Change threshold: 1

Point Size: 1

Genes of interest (comma-separated):

Genes of interest (comma-separated): No file selected

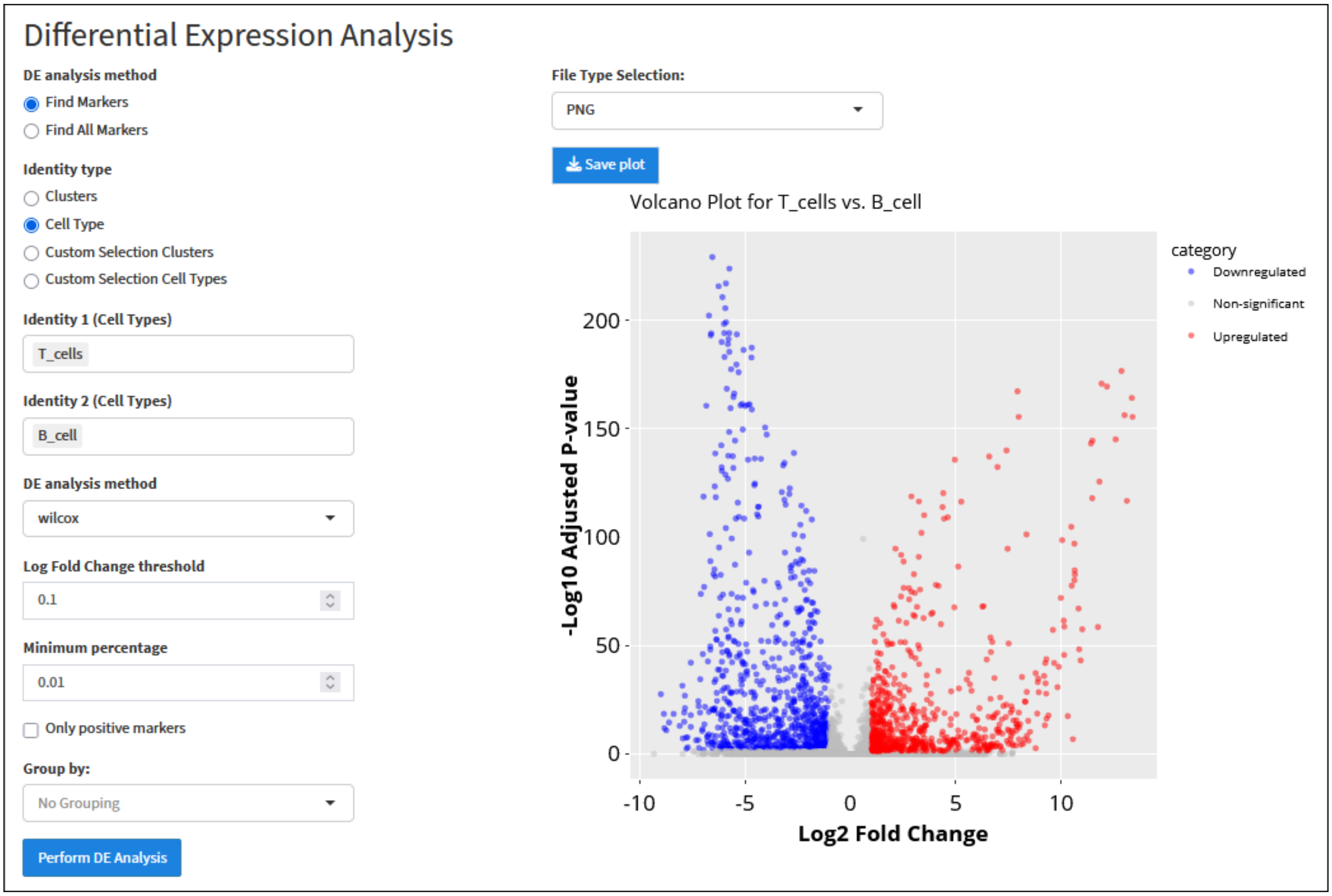
Figure 58. Find Markers>Volcano Plot Parameters.

If Find Markers is selected, additional parameters for generating the Volcano plot will appear:

- Adjusted P-value Threshold—sets the cutoff for adjusted p-values, helping to identify statistically significant genes. The default value is '0.05'.
- Log Fold Change Threshold—defines the minimum absolute value for the log fold change required for a gene to be considered significant. The default value is '1'.

- Point Size—allows you to set the size of the points in the volcano plot.
- Genes of interest—allows you to enter specific genes of interest separated by commas or upload a CSV file with the list of genes.

Click the [Generate plots] button to generate the volcano plot based on specified parameters.



Differential Expression Analysis

DE analysis method
● Find Markers
○ Find All Markers

Identity type
○ Clusters
● Cell Type
○ Custom Selection Clusters
○ Custom Selection Cell Types

Identity 1 (Cell Types): T_cells

Identity 2 (Cell Types): B_cell

DE analysis method: wilcox

Log Fold Change threshold: 0.1

Minimum percentage: 0.01

□ Only positive markers

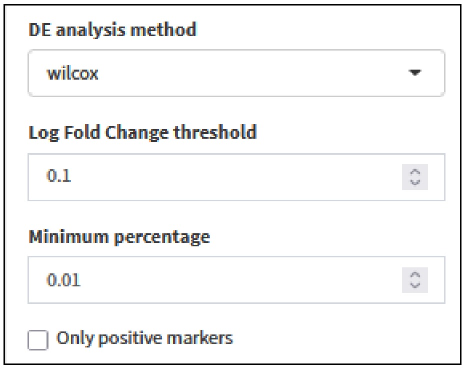
Group by: No Grouping

File Type Selection: PNG

Figure 59. Find Markers example volcano plot.

c) Find All Markers Parameters

(1) Preanalysis Parameters



DE analysis method: wilcox

Log Fold Change threshold: 0.1

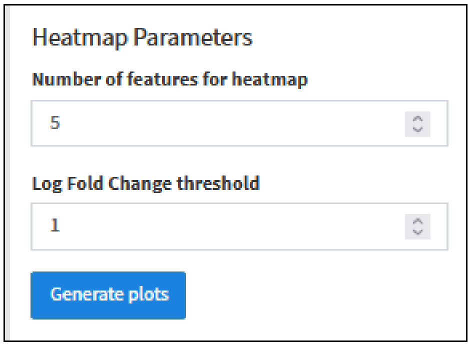
Minimum percentage: 0.01

□ Only positive markers

Figure 60. Find All Markers Parameters.

All of these parameters are in common with the "Find Markers" option. Refer to the Workflow for more information.

(2) Postanalysis Plot Parameters—Heatmap



Heatmap Parameters

Number of features for heatmap: 5

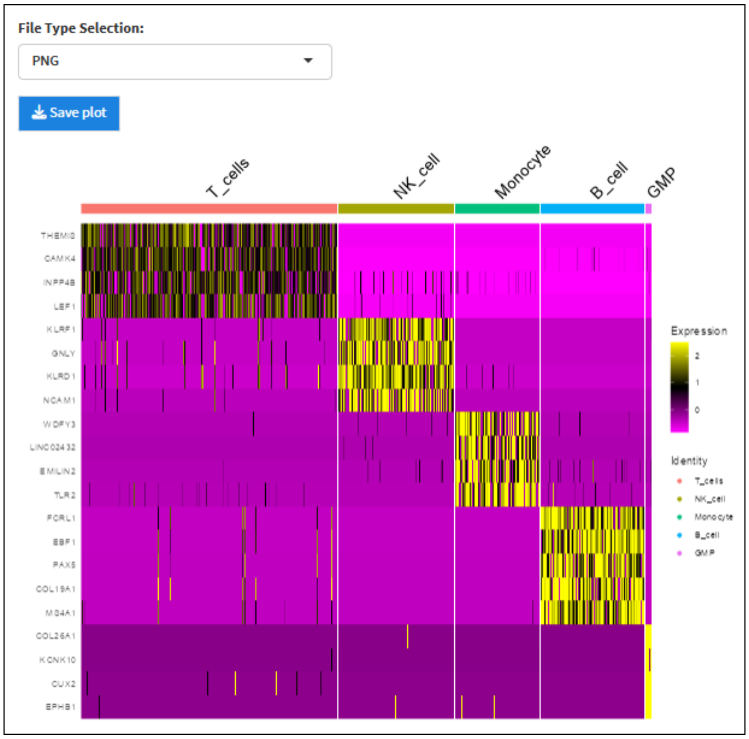
Log Fold Change threshold: 1

Figure 61. Find All Markers>Heatmap Parameters.

When Find All Markers is selected, users can generate a Heatmap with the following parameters:

- Number of Features for Heatmap—specifies the number of top features/genes to display in the heatmap.
- Log Fold Change Threshold—defines the minimum absolute value for the log fold change required for a gene to be considered significant. The default value is 1.

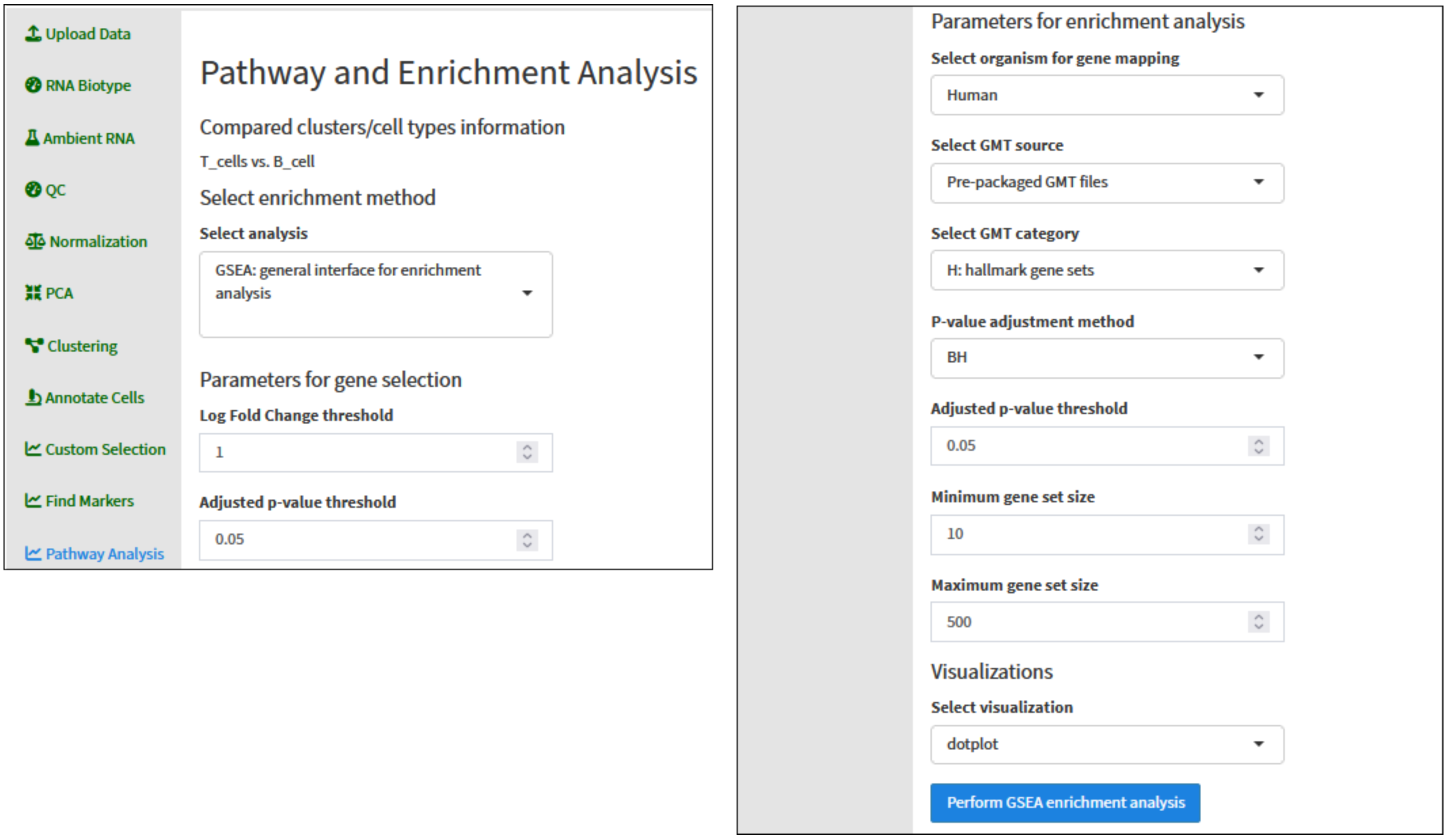
Click the [Generate plots] button to generate the heatmap based on the specified parameters.



File Type Selection: PNG

Figure 62. Find All Markers example heatmap.

11. Pathway and Enrichment Analysis



Pathway and Enrichment Analysis

Compared clusters/cell types information

T_cells vs. B_cell

Select enrichment method

Select analysis: GSEA: general interface for enrichment analysis

Parameters for gene selection

Log Fold Change threshold: 1

Adjusted p-value threshold: 0.05

Parameters for enrichment analysis

Select organism for gene mapping: Human

Select GMT source: Pre-packaged GMT files

Select GMT category: H: hallmark gene sets

P-value adjustment method: BH

Adjusted p-value threshold: 0.05

Minimum gene set size: 10

Maximum gene set size: 500

Visualizations

Select visualization: dotplot

Figure 63. *Pathway and Enrichment Analysis* page. The list of options is long, so has been split into two columns in the screenshot above.

The pathway analysis module helps you perform analysis using different enrichment methods. After selecting the method, define the gene selection parameters and visualization methods, and then download the results. The module supports analysis with both human and mouse datasets, utilizing clusterProfiler and ReactomePA R packages.

a) Workflow

1. Select Enrichment Method

   There is only one parameter in this subsection, "Select analysis", which is a drop-down menu. The options in the menu are described below.

   - GSEA: general interface for enrichment analysis—provides a general interface for conducting gene set enrichment analysis (GSEA) using the clusterProfiler R package. It enables flexible enrichment analysis as it allows various gene sets to be utilized. For the additional parameters specific to this option type, refer to Section b), "General Interface for Enrichment Analysis Options".
   - GSEA GO Enrichment Analysis—performs gene ontology (GO) enrichment analysis using GSEA leveraging clusterProfiler R package. For the additional parameters specific to this option type, refer to Section c), "GO Enrichment Analysis Options".

- GSEA WikiPathways Enrichment analysis—GSEA-based WikiPathways enrichment analysis, utilizing the clusterProfiler package. For the additional parameters specific to this option type, refer to Section d), "WikiPathways and Reactome Pathway Enrichment Analysis".
- Reactome pathway enrichment analysis—performs GSEA on Reactome pathways using the ReactomePA R package. For the additional parameters specific to this option type, refer to Section d), "WikiPathways and Reactome Pathway Enrichment Analysis".

2. Parameters for Gene Selection

   The two parameters in this section help determine which genes to include in the analysis based on differential expression results from the previous module.

   - Log Fold Change Threshold—defines the minimum absolute value for the log fold change required for a gene to be considered significant. The default value is '1'.
   - Adjusted p-value Threshold—sets the cutoff for adjusted p-values, helping to identify statistically significant genes. The default value is '0.05'.

3. Parameters for Enrichment Analysis

   The bulk of the parameter options on this page are under this subsection. Like the previous page (Differential Expression Analysis), the options here will vary depending on the initial selection made. The fields in common are summarized below, while the subsequent sections describe the analysis type-specific options.

   - Select organism for gene mapping—decides which gene sets to use for the analysis and also converts gene symbols to ENTREZ IDs. Ensure the correct organism species is selected for accurate gene mapping and GMT (Gene Matrix Transposed) file usage.
   - P-value Adjustment Method—choose a method to adjust the p values for multiple testing from the dropdown. Options include: BH (Benjamini & Hochberg—the default), Holm, Hochberg, Hommel, Bonferroni, BY (Benjamini Yekutieli), FDR, and None. If 'None' is selected, the p-value will equal the "Adjusted p-value Threshold" value (next bullet). For details about these option types, refer to the p.adjust module documentation.

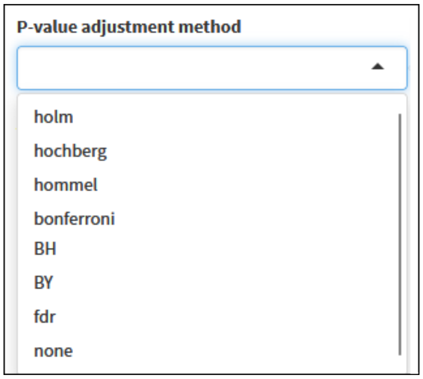


Figure 64. "P-value adjustment method" option drop-down menu on the *Pathway and Enrichment Analysis* page.

- Adjusted p-value Threshold—the p-value cutoff. The default value is '0.05'.
- Minimum Gene Set Size—represents the minimum size of the gene set. The default value is '10'.
- Maximum Gene Set Size—the maximum gene set size for analysis. The default value is '50'.

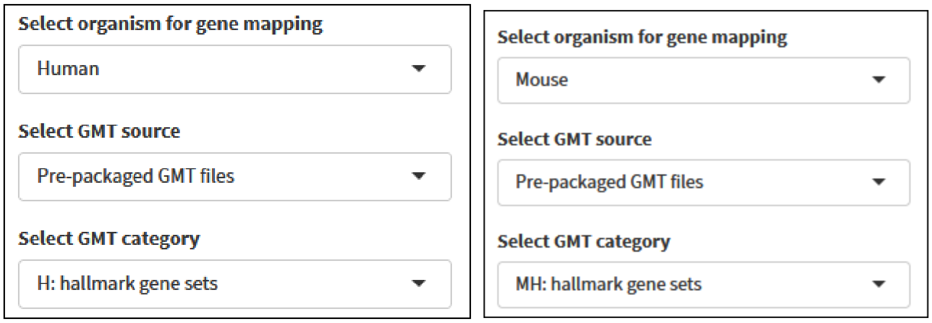
4. Visualizations
    - Select visualization—'dotplot' is the only available option. The plot helps to visualize significant pathways.
5. After filling out or selecting all the options, click the [Perform GSEA enrichment analysis] button to generate a chart and table of results, which will display to the right of the options.
6. Save the plot or download the results, if desired.
7. Click [Next: Generate Additional Plots] to proceed to the next module.

b) General Interface for Enrichment Analysis Options

Select GMT source—there are two options in this drop-down menu, 'Pre-packaged GMT files' and 'Upload GMT File'. Based on which of these are selected, a second option will display.

If 'Pre-packaged GMT files' is selected:

1. Select GMT category—allows for selection of a gene set collection from the drop-down menu; options are outlined in Table 1 (next page). Additional information on these collections can be obtained from the MSigDB website at https://www.gsea-msigdb.org/gsea/msigdb.



Select organism for gene mapping: Human
Select GMT source: Pre-packaged GMT files
Select GMT category: H: hallmark gene sets

Select organism for gene mapping: Mouse
Select GMT source: Pre-packaged GMT files
Select GMT category: MH: hallmark gene sets

Figure 65. GSEA general interface options for prepackaged GMT files in the enrichment analysis parameters subsection of the *Pathway analysis* page. Note that the initial "Select GMT category" option in the drop-down menu changes depending on the organism selected for gene mapping **(Left)** Human, **(Right)** Mouse.

Table 1. "Select GMT category" drop-down menu options, based on selected organism.

Human gene set collections
H: Hallmark
C1: Positional
C2: Curated
C3: Regulatory target
C4: Computational
C5: Ontology
C6: Oncogenic signature
C7: Oncogenic
C8: Cell type signature

Mouse gene set collections
MH: Hallmark
M1: Positional
M2: Curated
M3: Regulatory target
M5: Ontology
M8: Cell type signature

If 'Upload GMT File' is selected:

- Upload GMT file— uploads your preferred GMT file using this option to include custom gene sets tailored to your specific research focus or experimental conditions within human or mouse data. This allows you to incorporate specialized pathways or updated gene sets for a more personalized analysis.
- Select Term ID—the column in your file that contains term IDs.
- Select Gene ID—the column containing genes associated with each term ID.

NOTE: GMT files contain gene sets where each line represents gene set/pathway with a name, description, and list of genes corresponding to that gene set/pathway. For information on the file format, please refer to https://docs.gsea-msigdb.org/#GSEA/Data_Formats/#gmt-gene-matrix-transposed-file-format-gmt.

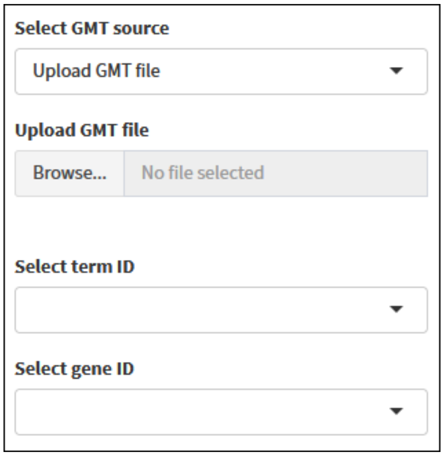
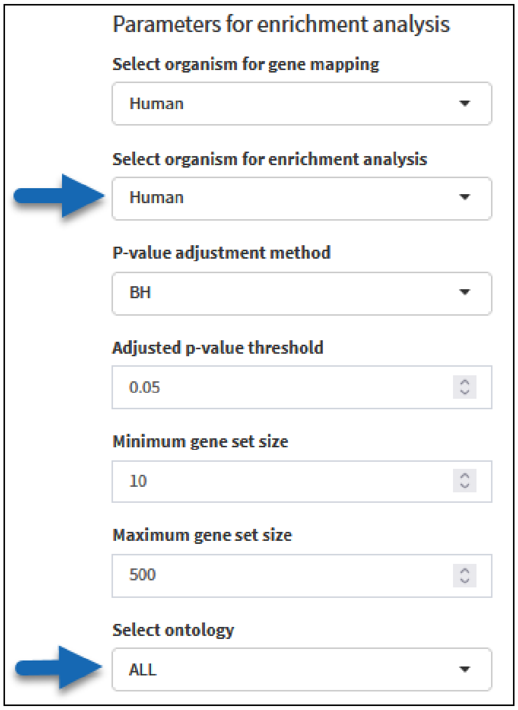


Figure 66. GSEA general interface options for uploading a GMT file in the enrichment analysis parameters subsection of the *Pathway analysis* page.

c) GO Enrichment Analysis Options



Parameters for enrichment analysis

Select organism for gene mapping: Human

Select organism for enrichment analysis: Human

P-value adjustment method: BH

Adjusted p-value threshold: 0.05

Minimum gene set size: 10

Maximum gene set size: 500

Select ontology: ALL

Figure 67. GSEA GO enrichment options in the enrichment analysis parameters subsection of the *Pathway analysis* page. The two options specific to this analysis type are indicated by the blue arrows.

2. Select organism for enrichment analysis—in addition to the "Select organism for gene mapping" option, 'Human' or 'Mouse' needs to be selected in this one as well.

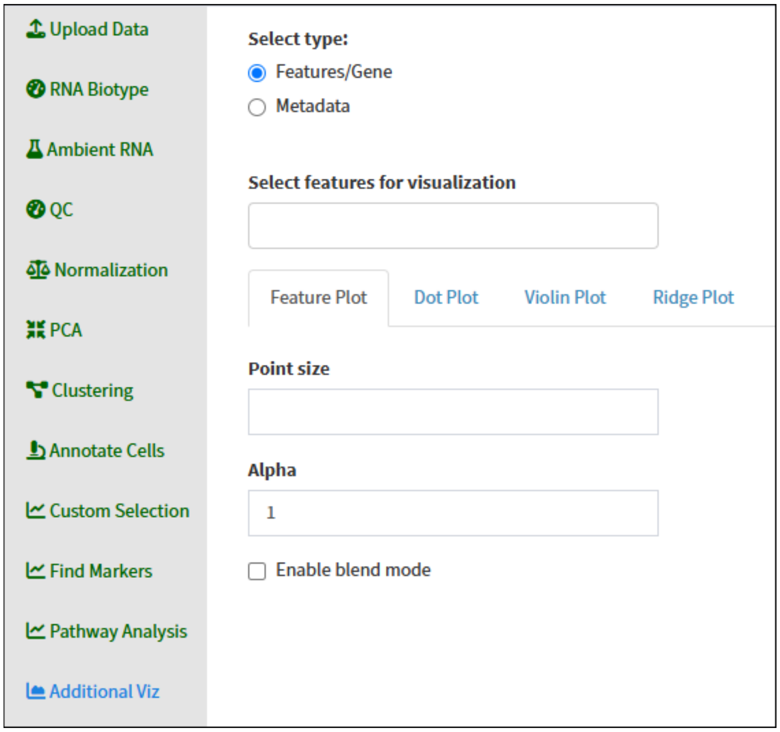
- Select ontology—choose the ontology type for analysis from the drop-down menu. The four options and the meaning of the abbreviations present in the menu are listed below:
  - BP—Biological Process
  - MF—Molecular Function
  - CC—Cellular Component
  - ALL—include all three ontologies (default)

  More information about these types can be found at the website https://advaitabio.com/faq-items/understanding-gene-ontology/.

d) WikiPathways and Reactome Pathway Enrichment Analysis

3. Select organism for enrichment analysis—in addition to the "Select organism for gene mapping" option, 'Human' or 'Mouse' needs to be selected in this one as well.

12. Generate Additional Plots (Additional Viz)



Select type:

● Features/Gene

○ Metadata

Select features for visualization

Point size

Alpha

1

□ Enable blend mode

Figure 68. The *Generate Additional Plots* default page. This screenshot also shows the tabs available for the 'Features/Genes' type option.

The additional visualization modules provide plotting options for exploring gene expression data and metadata. This module offers a number of visualization choices to help users gain insights into their data.

Options in this module are:

1. Select the type of visualization—choose one of the radio buttons for the following options:
   - Features/Gene—visualizes genes or features
   - Metadata—visualizes numerical metadata associated with the dataset
2. Select features for visualization—use this option to choose which features or metadata to visualize in the plot

   NOTE: There is a functional limit to the number of features displayed by default in the selection drop-down menu, listed in alphanumeric order.

   If you do not see the feature of interest listed:

   a. Click on the box (Figure 69, left)

   b. Delete the current contents if present (Figure 69, middle)

   c. Manually type in the whole or partial text match for your feature of interest (Figure 69, right)

   If no match returns, the given feature may not be found in the imported dataset.

The tabs available below these two options depend on the type of visualization selected. See Sections a) and b), below, for more information.

3. Once analysis on this page is completed, click [Next: Isoform Analysis] to proceed to the next module.

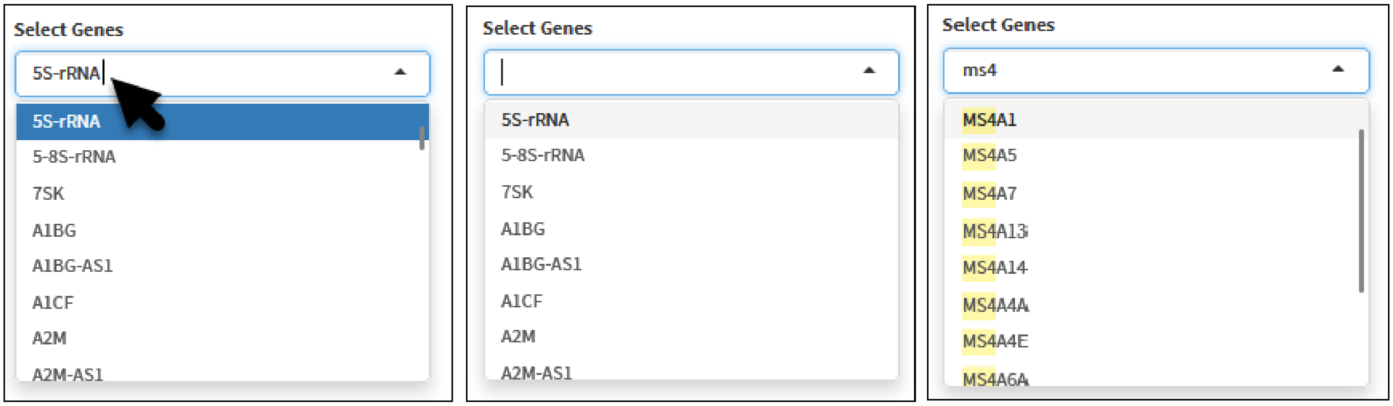
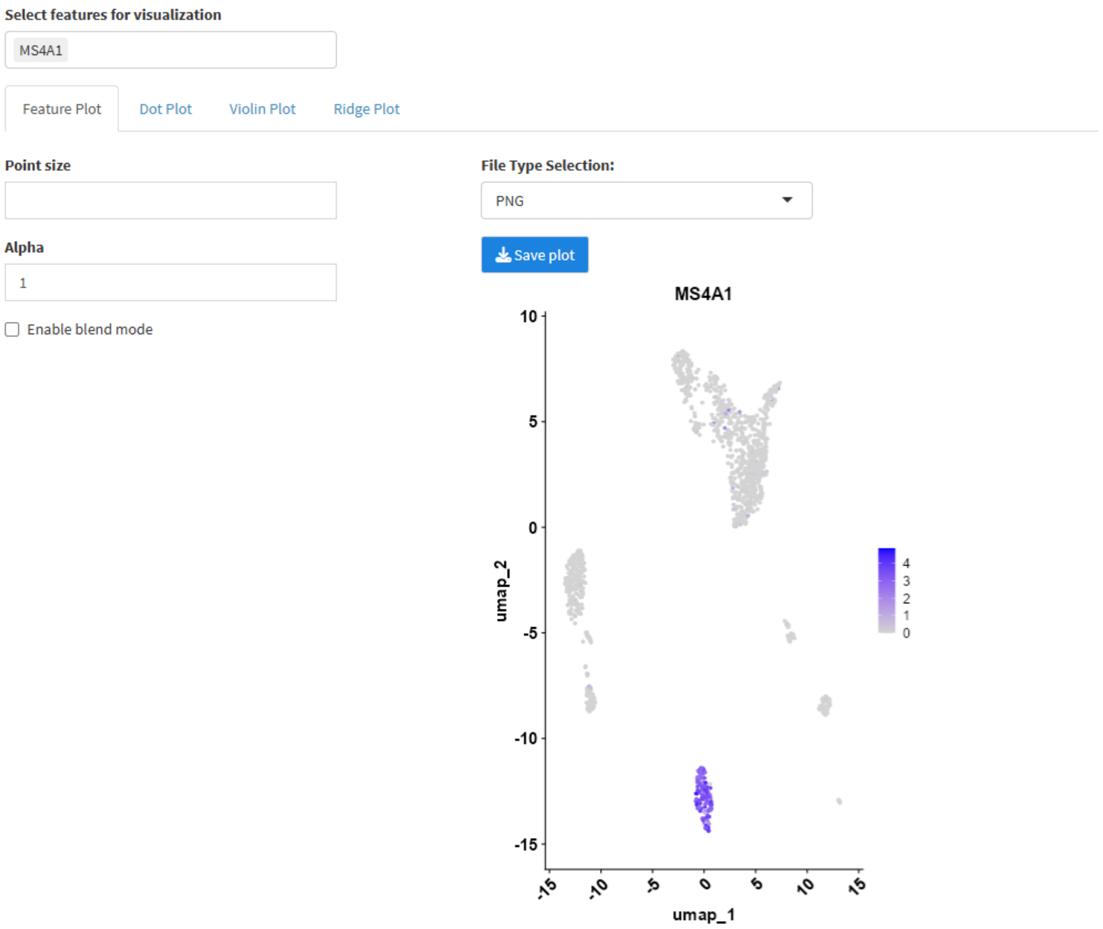


Figure 69. Selection of features or genes using manual text matching.

a) Features/Genes


Select features for visualization: MS4A1

Point size:

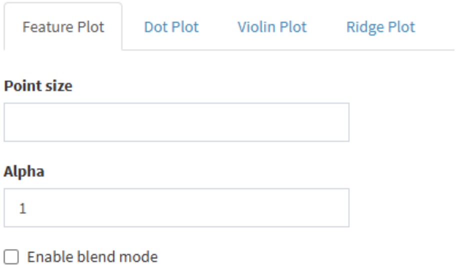
Alpha: 1

□ Enable blend mode

File Type Selection: PNG

Figure 70. Example *Feature Plot* view on a specific gene (*MS4A1*).

- Feature Plot—visualizes the expression of features or genes on dimension reduction plot (e.g., t-SNE or UMAP). It colors single cells according to the expression levels of selected features/genes.



Point size

Alpha

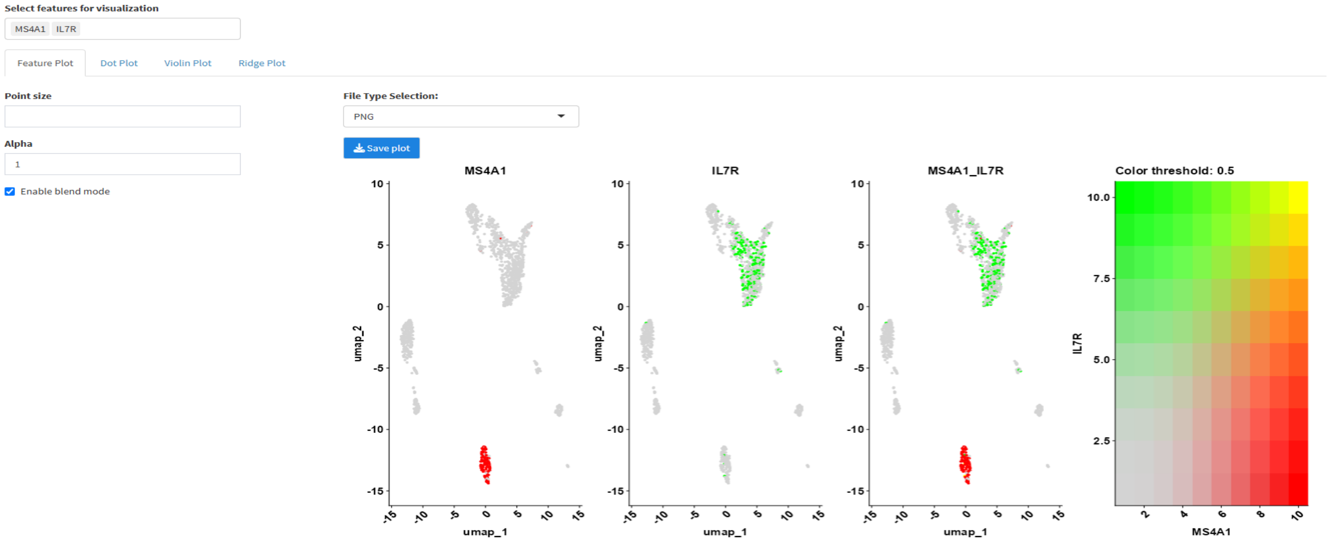
1

□ Enable blend mode

Figure 71. *Feature Plot* options in the Generate Additional Plots module of scRNA analysis.

Table 2. Description of *Feature plot* options in the Generate Additional Plots module of scRNA analysis.

Parameter	Default	Description
Point Size	(blank)	Controls the size of the points on the dimension reduction plot
Alpha	1	Sets the transparency of the points
Enable blend mode	Disabled	if the box is checked, this enables visualizing co-expression of two features. It only works when two features are specified in "Select features for visualization"



Select features for visualization

MS4A1 IL7R

Point size

Alpha

1

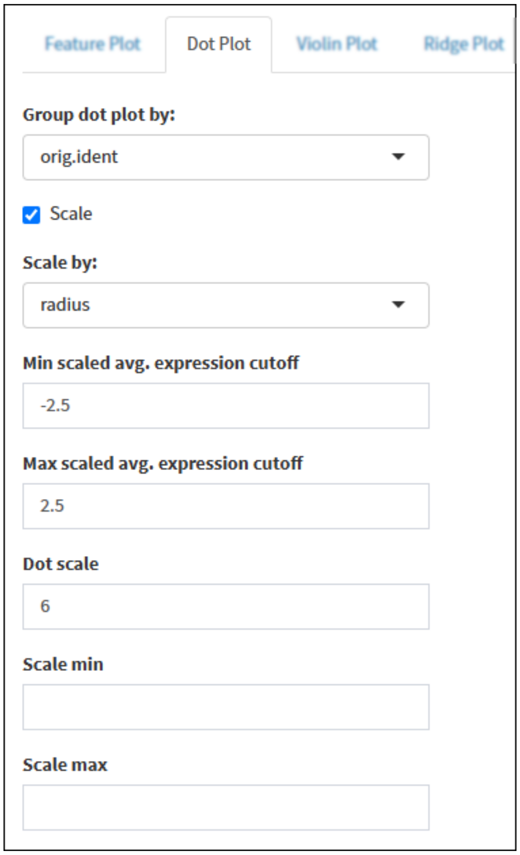
☑ Enable blend mode

File Type Selection:

PNG

Figure 72. Example *Feature Plot* showing the expression of *MS4A1* and *IL7R* with blend mode enabled.

- Dot Plot—displays the expression of selected features across different categories or groups within the dataset. The size and color of each dot represent the percentage and average expression of a given feature across all cells within each group, respectively.



Group dot plot by:
orig.ident

☑ Scale

Scale by:
radius

Min scaled avg. expression cutoff
-2.5

Max scaled avg. expression cutoff
2.5

Dot scale
6

Scale min

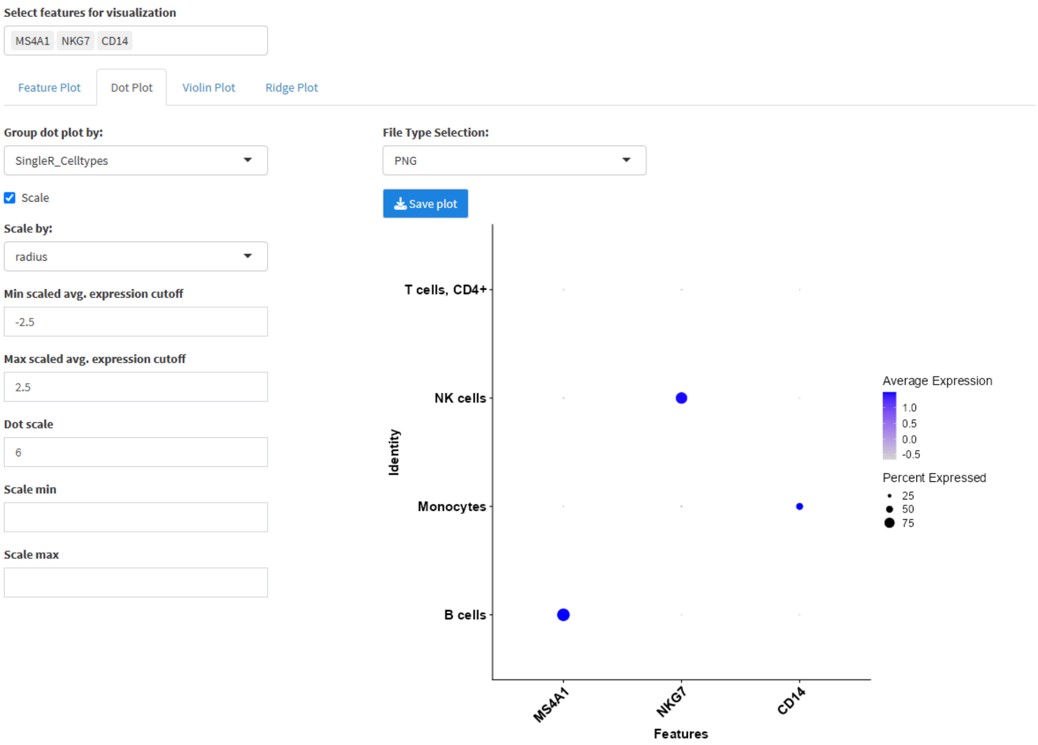
Scale max

Figure 73. *Dot plot* options in the Generate Additional Plots module of scRNA analysis.

Table 3. Description of the *Dot plot* options in the Generate Additional Plots module of scRNA analysis.

Parameter	Default	Description
Group dot plot by	orig.ident	Select how to group the data (e.g., by clusters, cell types, etc.) from the dropdown menu.
Scale	Enabled	When enabled, the data will be scaled.
Scale by	radius	Controls how dot size is scaled. Options are "radius" or "size".
Min scaled avg. expression cutoff	–2.5	The minimum scaled average expression cutoff threshold. Values below this will be clipped.
Max scaled avg. expression cutoff	2.5	The maximum scaled average expression cutoff threshold. Values above this will be clipped.
Dot scale	6	Adjusts the diameter of the dots in the plot.
Scale min	(blank)	Sets a lower limit for dot size scaling.

Parameter	Default	Description
Scale max	(blank)	Sets an upper limit for dot size scaling.



Select features for visualization: MS4A1 NKG7 CD14

Group dot plot by: SingleR_Celltypes

☑ Scale

Scale by: radius

Min scaled avg. expression cutoff: -2.5

Max scaled avg. expression cutoff: 2.5

Dot scale: 6

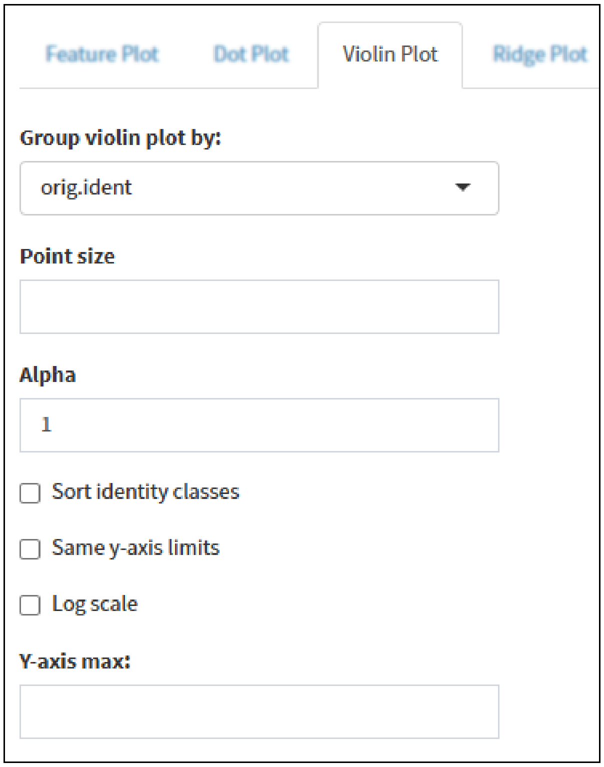
Scale min:

Scale max:

File Type Selection: PNG

Figure 74. Example *Dot Plot* showing expression of *MS4A1*, *NKG7*, and *CD14* across cell types.

- Violin Plot—illustrates the distribution of feature expression values across different groups.



Group violin plot by:

orig.ident

Point size

Alpha

1

□ Sort identity classes

□ Same y-axis limits

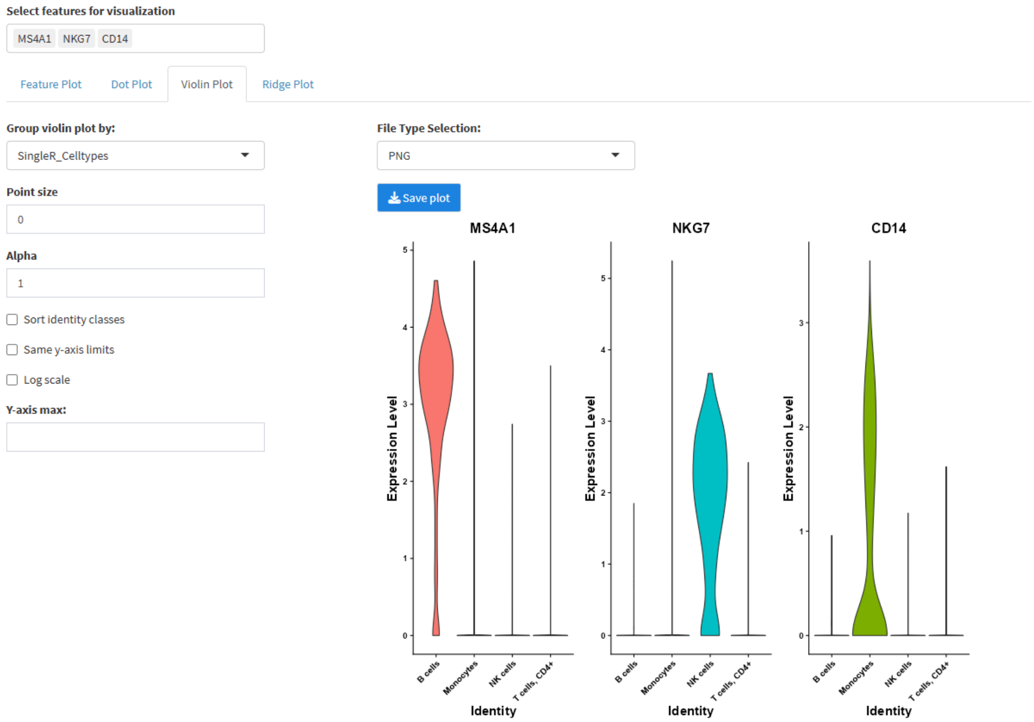
□ Log scale

Y-axis max:

Figure 75. *Violin plot* options in the Generate Additional Plots module of scRNA analysis.

Table 4. Description of the *Violin plot* options in the Generate Additional Plots module of scRNA analysis.

Parameter	Default	Description
Group violin plot by	orig.ident	Select how to group the data from the dropdown menu.
Point size	(blank)	Controls the size of the points on the plot. Set to 0 to hide the dots completely.
Alpha	1	Sets the transparency of the points. Values can range from 0–1.
Sort identity classes	Disabled	Sort identity classes on the x axis based on their average expression values.
Same y-axis limits	Disabled	Ensure all violin plots share the same y-axis scale.
Log scale	Disabled	Apply a log transformation to the y-axis.
Y-axis max	(blank)	Set a maximum y-axis limit for the violin plot.



Select features for visualization

MS4A1 NKG7 CD14

Group violin plot by: SingleR_Celltypes

Point size: 0

Alpha: 1

□ Sort identity classes

□ Same y-axis limits

□ Log scale

Y-axis max:

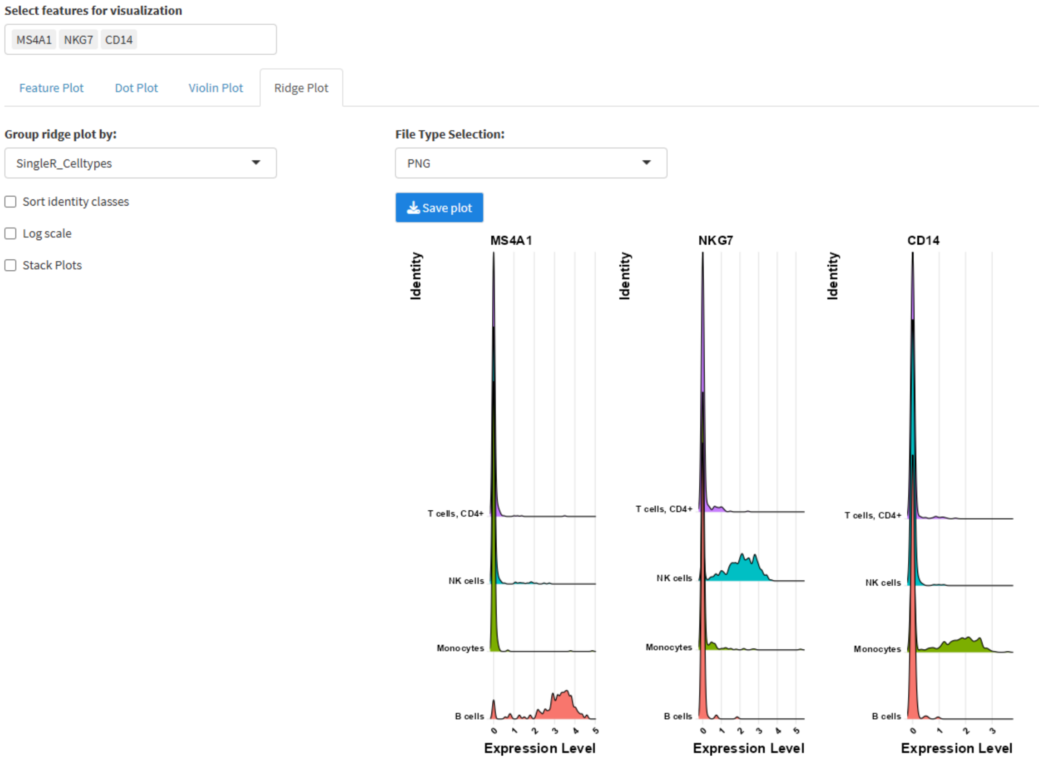
File Type Selection: PNG

Figure 76. Example *Violin Plot* showing expression of *MS4A1*, *NKG7*, and *CD14* across cell types.

- Ridge Plot—displays the distribution of feature expression values across different groups. Users can select the grouping variable for the plot from the "Group by" drop-down menu.

Table 5. *Ridge plot* options in the Generate Additional Plots module of scRNA analysis.

Parameter	Default	Description
Group ridge plot by	orig.ident	Select how to group the data from the dropdown menu.
Sort identity classes	Disabled	Sort identity classes on the x axis based on their average expression values.
Log scale	Disabled	Apply a log transformation to feature axis.
Stack Plots	Disabled	For each feature, stack plots horizontally.



Select features for visualization: MS4A1 NKG7 CD14

Group ridge plot by: SingleR_Celltypes

File Type Selection: PNG

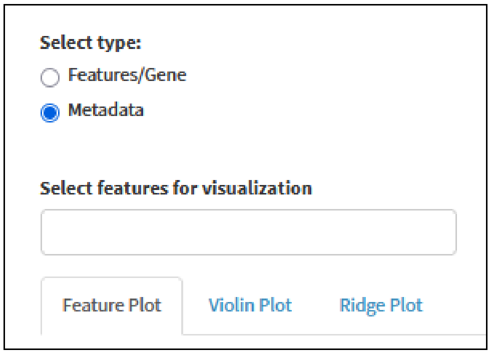
☐ Sort identity classes

☐ Log scale

☐ Stack Plots

Figure 77. Example *Ridge Plot* showing expression of *MS4A1*, *NKG7*, and *CD14* across cell types.

b) *Metadata*



Select type:

○ Features/Gene

● Metadata

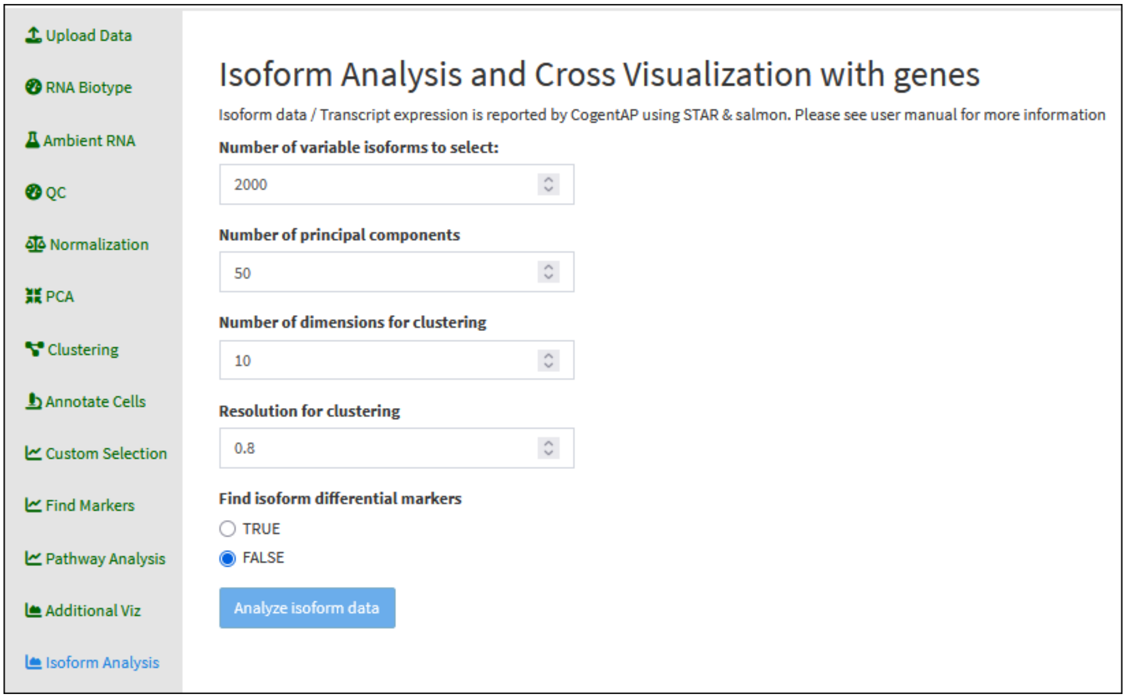
Select features for visualization

Figure 78. *Generate Additional Plots* options visible for the 'Metadata' type.

NOTE: Parameters for metadata are similar to Features/Genes (previous section). However, a dot plot is not available for metadata.

- Feature Plot—visualizes the numerical metadata on a dimension reduction plot, coloring cells based on metadata values to gain insights into data patterns.
- Violin Plot—illustrates the distribution of metadata values across different groups.
- Ridge Plot—shows the distribution of metadata values across different groups.

13. Isoform Analysis and Cross Visualization with Genes



Isoform Analysis and Cross Visualization with genes

Isoform data / Transcript expression is reported by CogentAP using STAR & salmon. Please see user manual for more information

Number of variable isoforms to select: 2000

Number of principal components: 50

Number of dimensions for clustering: 10

Resolution for clustering: 0.8

Find isoform differential markers

○ TRUE

● FALSE

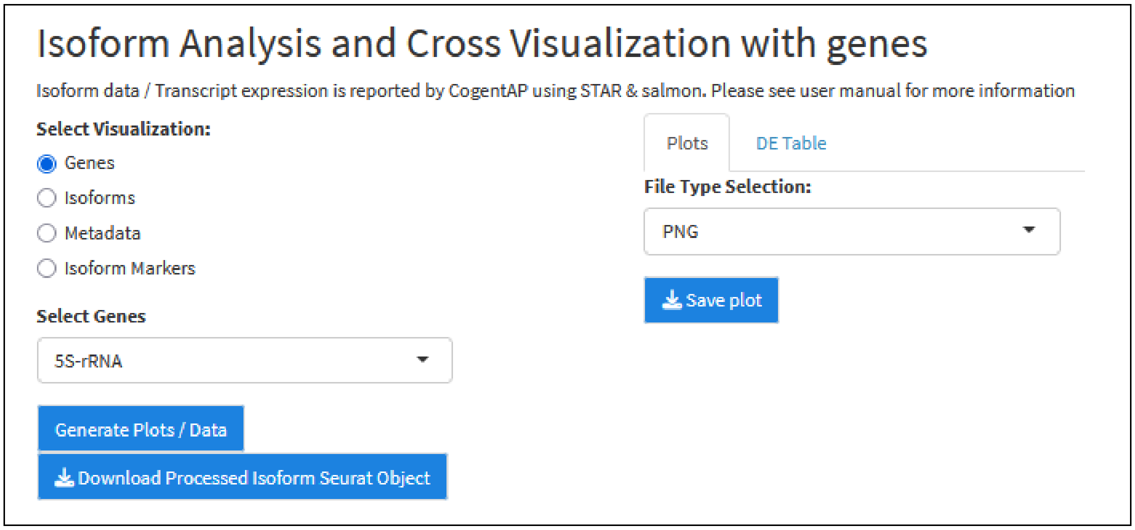
Figure 79. Default *Isoform Analysis and Cross Visualization with genes* page.

This module allows users to perform isoform analysis, which includes cross-visualization of transcript expression on gene expression-based UMAP/t-SNE plots, and isoform differential marker analysis.

Key components of this module include:

- Number of variable isoforms to select—determines the number of variable features to be selected for downstream analyses such as PCA, clustering, and visualization (e.g., UMAP or t-SNE). The default value is '2,000' (features), which is enough to capture sufficient variability in most datasets. This number can be adjusted depending on the complexity of the dataset and the specific goals of the analysis.
- Number of principal components—used for clustering. The default value is '50'.
- Number of dimensions for clustering—the default is '10'.
- Resolution for clustering—determines the number of clusters. The default is '0.8'.
- Find isoform differential markers—select the 'TRUE' radio button to obtain differential isoform markers per cluster or leave it as 'FALSE' (the default) to decline the option.

Click [Analyze isoform data] to initiate the data analysis process. Once the analysis is completed, the module will display a second set of options.



Isoform Analysis and Cross Visualization with genes

Isoform data / Transcript expression is reported by CogentAP using STAR & salmon. Please see user manual for more information

Select Visualization:

● Genes
○ Isoforms
○ Metadata
○ Isoform Markers

Select Genes: 5S-rRNA

File Type Selection: PNG

Figure 80. ***Isoform Analysis and Cross Visualization with genes*** **page after initial analysis of the isoform data.**

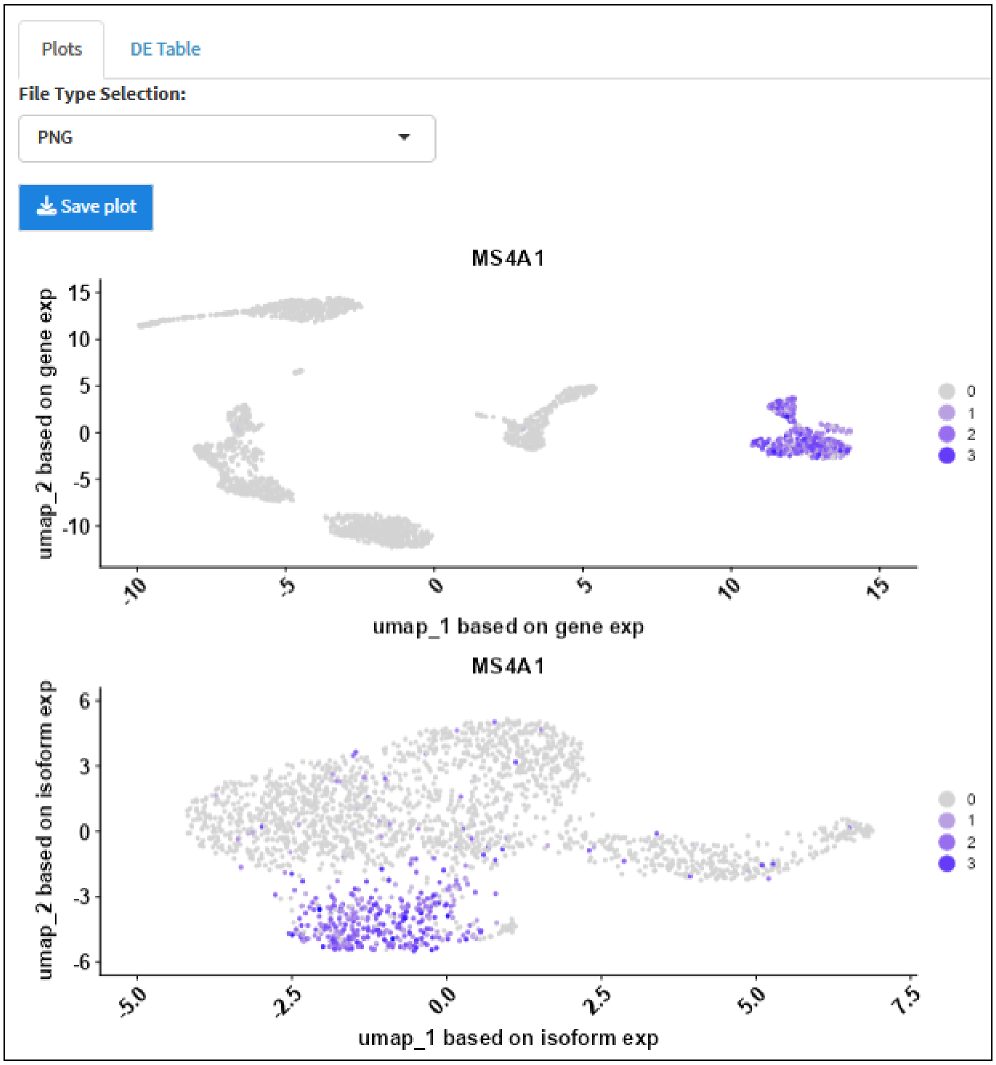
Four radio buttons for "Select Visualization" will appear for further exploration:

- Genes—this provides an option to select a gene from the drop-down menu (see Figure 80).
- Isoforms—Similar to 'Genes', select an isoform to visualize gene expression and isoform expression-based dimension reduction (UMAP/t-SNE).
- Metadata—select a "Color by" option (similar to the option of the same name in Section VI.A.6, "Principal Component Analysis (PCA)") from the drop-down menu to color code single cells in a UMAP/t-SNE chart for gene expression and isoform expression.
- Isoform Markers—displays isoform markers if the "Find isoform differential markers" option was selected as 'TRUE'. The subselection for this option is to visualize the data either as a heatmap or a DE isoforms table.

  NOTE: In order to change the "Find isoform differential markers" option to 'TRUE' at this point, you will need to restart the app.

Select the visualization you wish to produce then click on [Generate Plots/ Data] to create the UMAPs/t-SNE. This will display two UMAP/t-SNE charts corresponding to gene expression and isoform expression, respectively.

NOTE: This module requires >30 cells to perform analysis.



Plots | DE Table

File Type Selection:

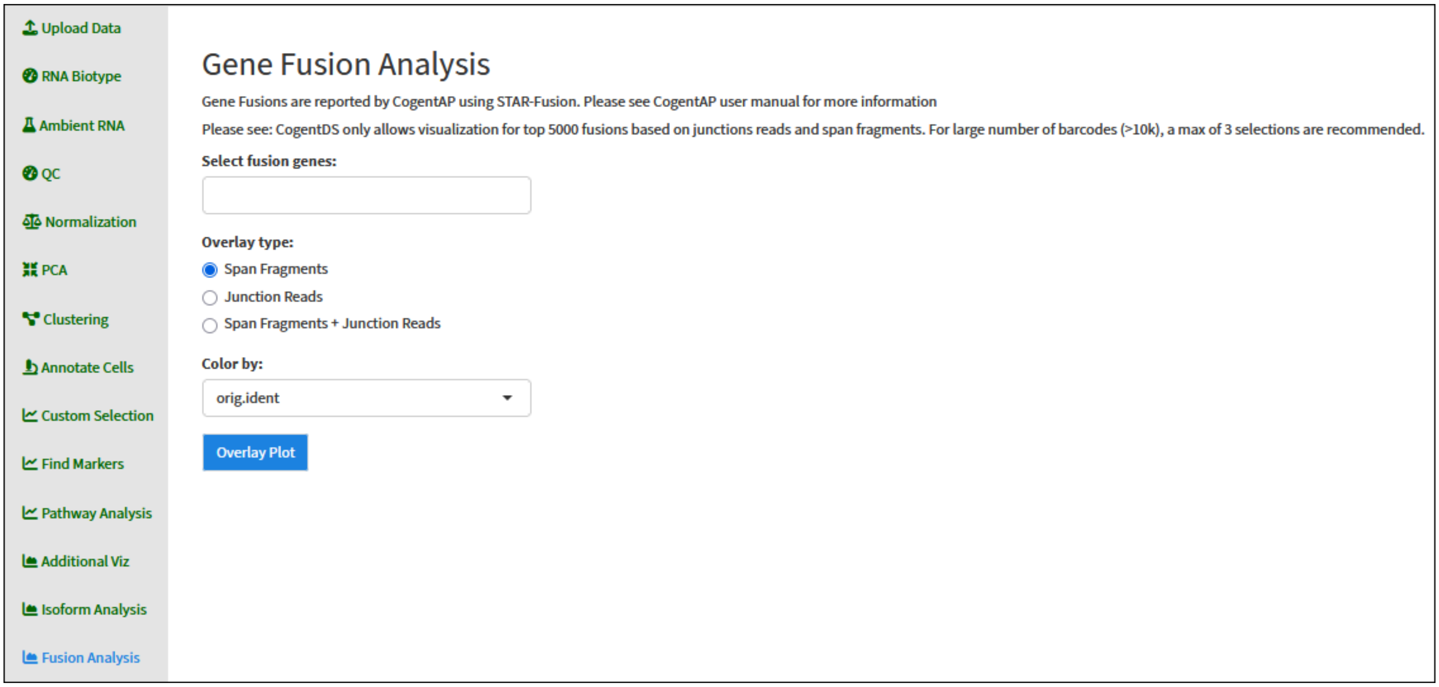
PNG

Save plot

Figure 81. *Isoform Analysis and Cross Visualization with genes* page after plot generation using 'Genes' visualization. The gene used in the example is 'MS4A1'.

Save the plot or download the results, if desired, then click [Next: Gene Fusion Analysis] to proceed to the next module.

14. Gene Fusion Analysis



Gene Fusion Analysis

Gene Fusions are reported by CogentAP using STAR-Fusion. Please see CogentAP user manual for more information

Please see: CogentDS only allows visualization for top 5000 fusions based on junctions reads and span fragments. For large number of barcodes (>10k), a max of 3 selections are recommended.

Select fusion genes:

Overlay type:

● Span Fragments
○ Junction Reads
○ Span Fragments + Junction Reads

Color by: orig.ident

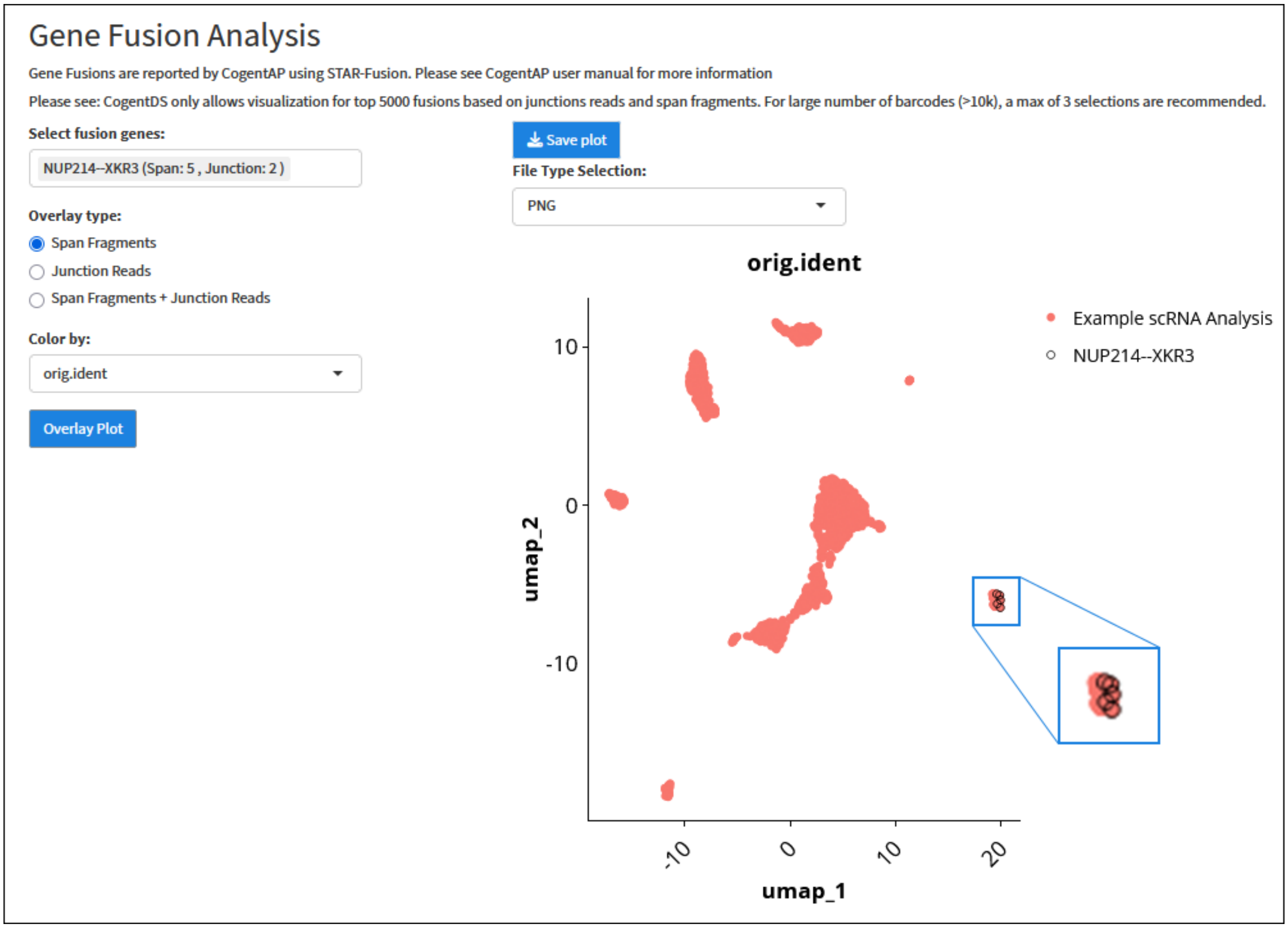
Figure 82. The *Gene Fusion Analysis* page when fusions are detected in the imported data.

If fusion analysis is performed in CogentAP (Cogent NGS Analysis Pipeline User Manual, Section V.B.2, "Optional Extended Analysis"), this module detects the fusion results and activates the options on the page at this step, as illustrated in Figure 82; it allows the fusion analysis results to be overlayed on the gene-based UMAP/t-SNE plot.

The fusions detected during the analysis are available to select in the "Select fusion genes" drop-down menu. Choose the fusion of interest, then the overlay type, which includes span fragments, junction reads, or both ('Span Fragments + Junction reads'). Multiple gene fusions can be overlayed on the dimension reduction plot for comparative analysis. Gene fusions listed in the selection box can be removed by clicking on the fusion name and hitting the **[backspace]** key on your keyboard.

The "Color by" option allows you to assign a color to the single cells in the resulting plot based on the metadata column.

Once the options are selected, click the [Overlay Plot] button to apply the effects to the cells in which the fusion was detected.



Gene Fusion Analysis

Gene Fusions are reported by CogentAP using STAR-Fusion. Please see CogentAP user manual for more information

Please see: CogentDS only allows visualization for top 5000 fusions based on junctions reads and span fragments. For large number of barcodes (>10k), a max of 3 selections are recommended.

Select fusion genes: NUP214--XKR3 (Span: 5 , Junction: 2)

Overlay type:

● Span Fragments
○ Junction Reads
○ Span Fragments + Junction Reads

Color by: orig.ident

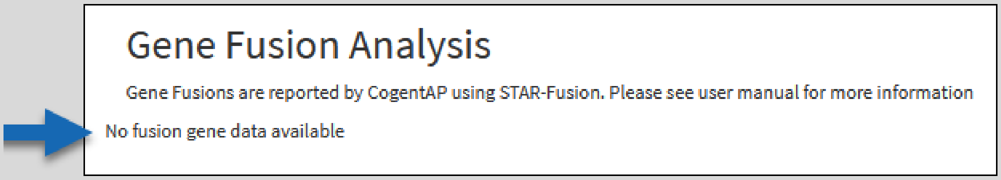
Figure 83. Example Gene Fusion Analysis results with fusion overlay. The cluster outlined by the blue square demonstrates the overlay of the NUP214--XKR3 fusion; the inset shows an enlargement of the original to better display the circled cells within the cluster in this figure.

NOTES:

– CogentDS only allows overlay of the top 5,000 gene fusions obtained, based on selection of span fragments, junction reads, or span fragments + junction reads. For large datasets where the number of barcodes >10,000, a maximum of three selections for overlay is recommended.

– When analyzing large datasets, you may notice a delay (lag) in the population of the drop-down menus after switching between overlay types.

– If no gene fusions are detected within the data file being analyzed, the message "No fusion gene data available" will display instead of the options (Figure 84).



Gene Fusion Analysis

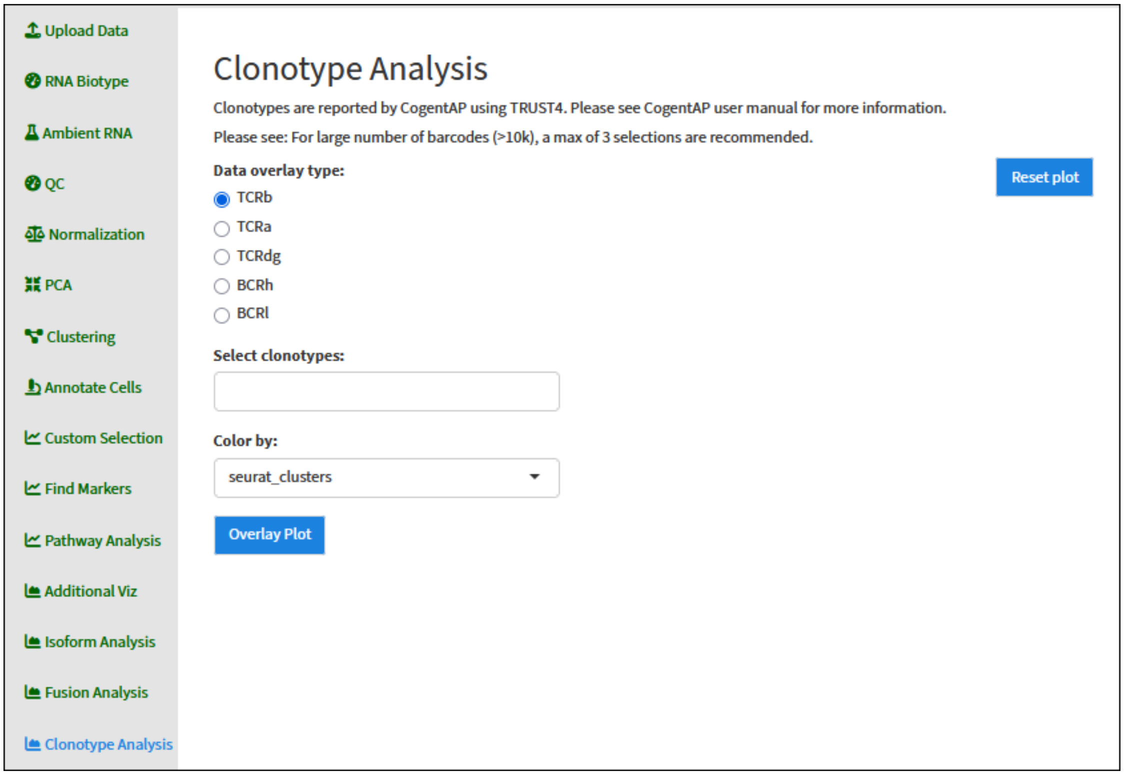
Gene Fusions are reported by CogentAP using STAR-Fusion. Please see user manual for more information

No fusion gene data available

Figure 84. *Gene fusion analysis* page, if no fusion gene data is found in the imported data.

Save the plot or download the results, if desired, then click [Next: Clonotype Analysis] to proceed to the next module.

15. Clonotype Analysis



Clonotype Analysis

Clonotypes are reported by CogentAP using TRUST4. Please see CogentAP user manual for more information.

Please see: For large number of barcodes (>10k), a max of 3 selections are recommended.

Data overlay type:

- ● TCRb
- ○ TCRa
- ○ TCRdg
- ○ BCRh
- ○ BCRl

Select clonotypes:

Color by: seurat_clusters

Figure 85. *Clonotype Analysis* page.

When immune analysis is performed in CogentAP (Cogent NGS Analysis Pipeline User Manual, Section V.B.2, "Optional Extended Analysis"), the results can be overlayed on the gene-based clustering plot by using the options on the *Clonotype Analysis* page.

Click [Next: Generate HTML Report] to proceed to the next module.

a) Before You Begin

Please keep the following information in mind before exploring this module:

- For large datasets where the number of barcodes >10,000, a maximum of three selections for overlay is recommended.
- When analyzing large datasets, you may notice a delay (lag) in the population of the "Select clonotypes" drop-down menus after switching between data overlay types.
- When analyzing >50,000 barcodes, you may also experience slowness (lag) in the display of the list under "Select clonotypes" during the selection process.
- If no clonotype information is contained in the imported data, the message "No Clonotypes data available for the selected assay type." will display instead of the second set of options, and the initial radio buttons will make no changes to the page.

b) *Overlay Options*

Select the options for the overlays from the list.

- Data overlay type—select one of the clonotype options using the radio buttons.
  - TCRb—T-cell receptor (TCR) β
  - TCRa—TCRα
  - TCRdg—TCRδ, TCRγ
  - BCRh—B-cell receptor (BCR) heavy chain components
  - BCRl—BCR light chain components
- Selecting clonotypes—after choosing an overlay type, you can select from the clonotypes detected by CogentAP (in the imported data) within that type.

  Multiple clonotypes can be overlayed on the dimension reduction plot for comparative analysis. Figure 86 shows how the list of detected clonotypes display as a drop-down menu when the input box is clicked on (Panel A); it also demonstrates how multiple clonotypes can be selected, with Panel B showing two clonotypes selected with the menu open to add a third. Clonotypes listed in the selection box can be removed by clicking on the clonotype name and hitting the **[Backspace]** key on your keyboard.

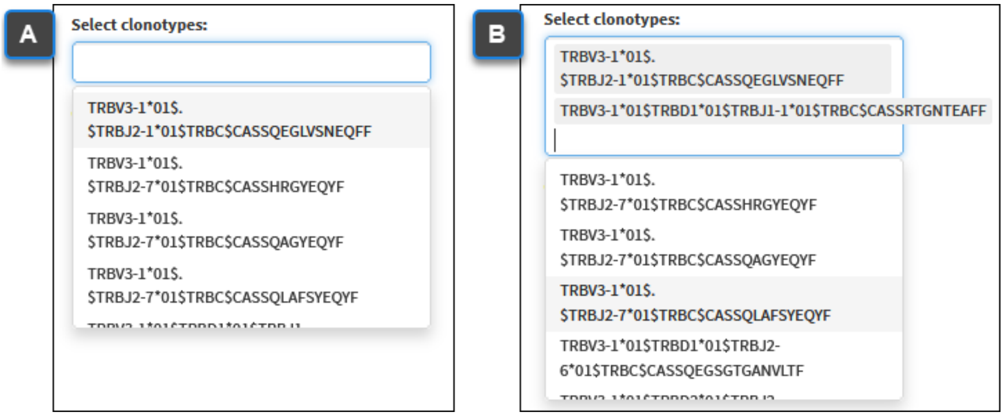
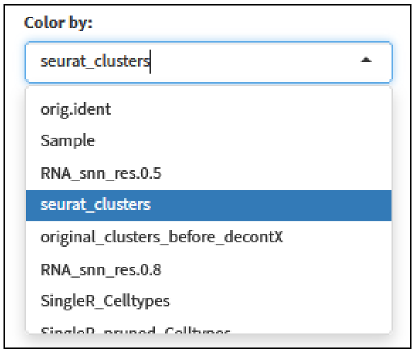

A
Select clonotypes:
TRBV3-1*01$.
$TRBJ2-1*01$TRBC$CASSQEGLVSNEQFF
TRBV3-1*01$.
$TRBJ2-7*01$TRBC$CASSHRGYEQYF
TRBV3-1*01$.
$TRBJ2-7*01$TRBC$CASSQAGYEQYF
TRBV3-1*01$.
$TRBJ2-7*01$TRBC$CASSQLAFSYEQYF
B
Select clonotypes:
TRBV3-1*01$.
$TRBJ2-1*01$TRBC$CASSQEGLVSNEQFF
TRBV3-1*01$TRBD1*01$TRBJ1-1*01$TRBC$CASSRTGNTEAFF
TRBV3-1*01$.
$TRBJ2-7*01$TRBC$CASSHRGYEQYF
TRBV3-1*01$.
$TRBJ2-7*01$TRBC$CASSQAGYEQYF
TRBV3-1*01$.
$TRBJ2-7*01$TRBC$CASSQLAFSYEQYF
TRBV3-1*01$TRBD1*01$TRBJ2-
6*01$TRBC$CASSQEGSGTGANVLTF

Figure 86. Selecting specific clonotypes to overlay on the *Clonotype Analysis* page. (**Panel A**) The initial selection drop-down menu. (**Panel B**) The selection drop-down menu after addition of two clonotypes to overlay.

- Color by—as in other modules, this option can be used to assign a color to the single cells in the resulting plot based on the metadata column.



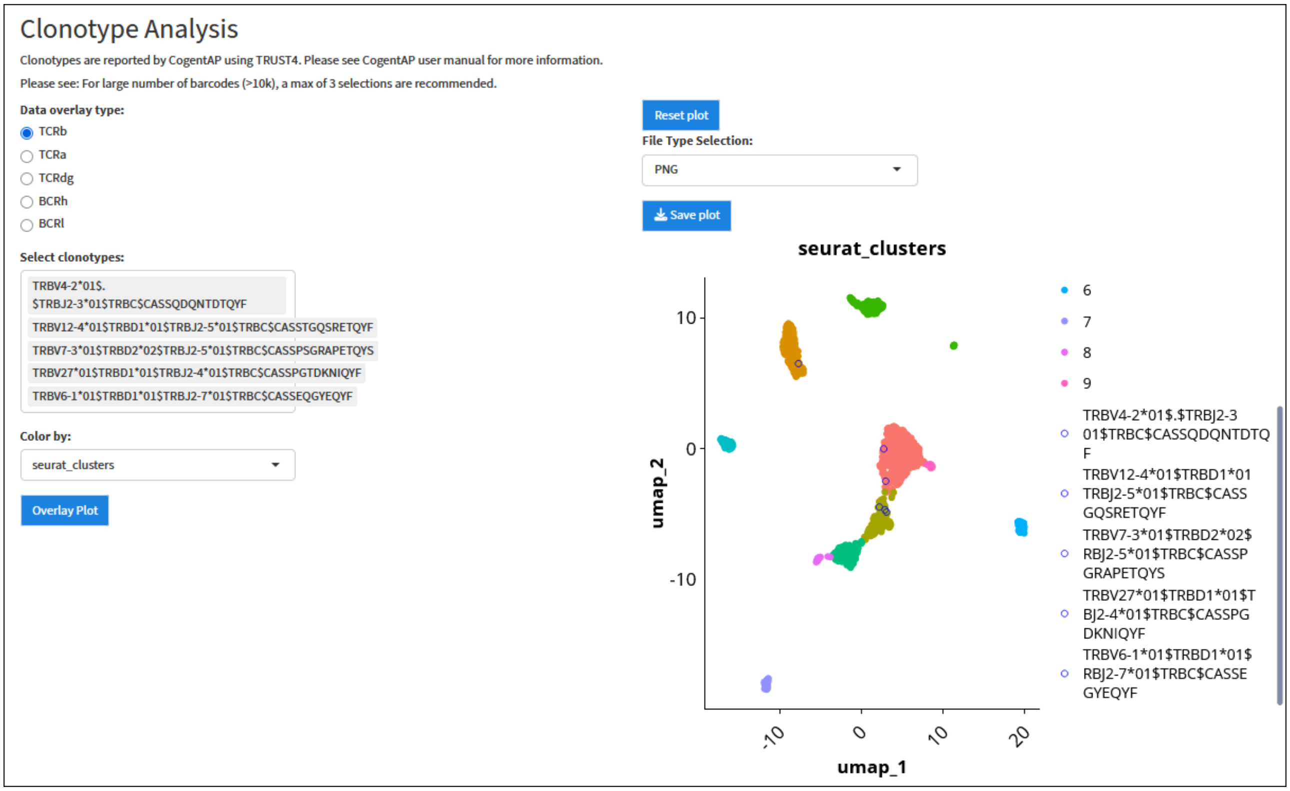
Color by:

seurat_clusters

Figure 87. "Color by" drop-down menu options on the *Clonotype Analysis* page.

c) Generate Overlay Plot

After making your selections, click on [Overlay Plot] to generate the chart, which will display to the right of the option list (Figure 88).



Clonotype Analysis

Clonotypes are reported by CogentAP using TRUST4. Please see CogentAP user manual for more information.

Please see: For large number of barcodes (>10k), a max of 3 selections are recommended.

Data overlay type:

● TCRb
○ TCRa
○ TCRdg
○ BCRh
○ BCRl

Select clonotypes:

TRBV4-2*01$.$TRBJ2-3*01$TRBC$CASSQDQNTDTQYF
TRBV12-4*01$TRBD1*01$TRBJ2-5*01$TRBC$CASSTGQSRETQYF
TRBV7-3*01$TRBD2*02$TRBJ2-5*01$TRBC$CASSPSGRAPETQYS
TRBV27*01$TRBD1*01$TRBJ2-4*01$TRBC$CASSPGTDKNIQYF
TRBV6-1*01$TRBD1*01$TRBJ2-7*01$TRBC$CASSEQGYEQYF

Color by:

seurat_clusters

Overlay Plot
Reset plot
File Type Selection:
PNG
Save plot
seurat_clusters
umap_2
10
0
-10
umap_1
-10
0
10
20
6
7
8
9
TRBV4-2*01$.$TRBJ2-3 01$TRBC$CASSQDQNTDTQ F
TRBV12-4*01$TRBD1*01 TRBJ2-5*01$TRBC$CASS GQSRETQYF
TRBV7-3*01$TRBD2*02$ RBJ2-5*01$TRBC$CASSP GRAPETQYS
TRBV27*01$TRBD1*01$T BJ2-4*01$TRBC$CASSPG DKNIQYF
TRBV6-1*01$TRBD1*01$ RBJ2-7*01$TRBC$CASSE GYEQYF

Figure 88. Example UMAP chart with selected clonotypes overlaid on the *Clonotype Analysis* page.

d) Modify the Clonotype Analysis Overlay Options

If an overlay plot has been generated (Section c) but you want to modify any of the selected options (Section b), follow the procedure below for the respective option you want to change.

(1) Data Overlay Type

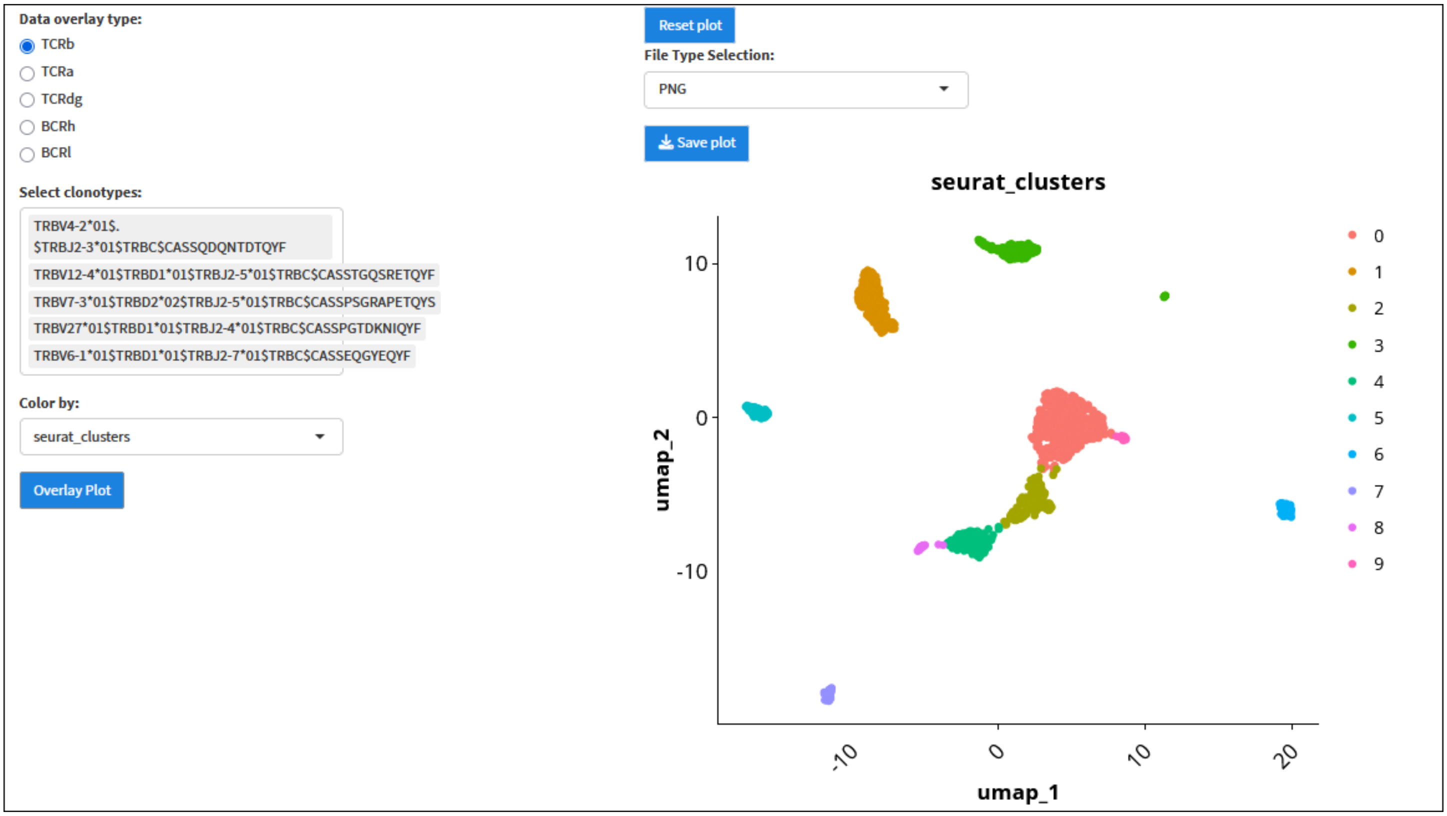
Select a new radio button to select the new data type. The chart will regenerate automatically after the selection.

(2) Select Clonotypes

1. Click the [Reset plot] button.

   NOTE: No changes are made to the display after this action.

2. Click on the [Overlay Plot] button. This will clear the overlay on the chart and the chart legend but will not reset the selected options (Figure 89).



Data overlay type:

● TCRb
○ TCRa
○ TCRdg
○ BCRh
○ BCRl

Select clonotypes:

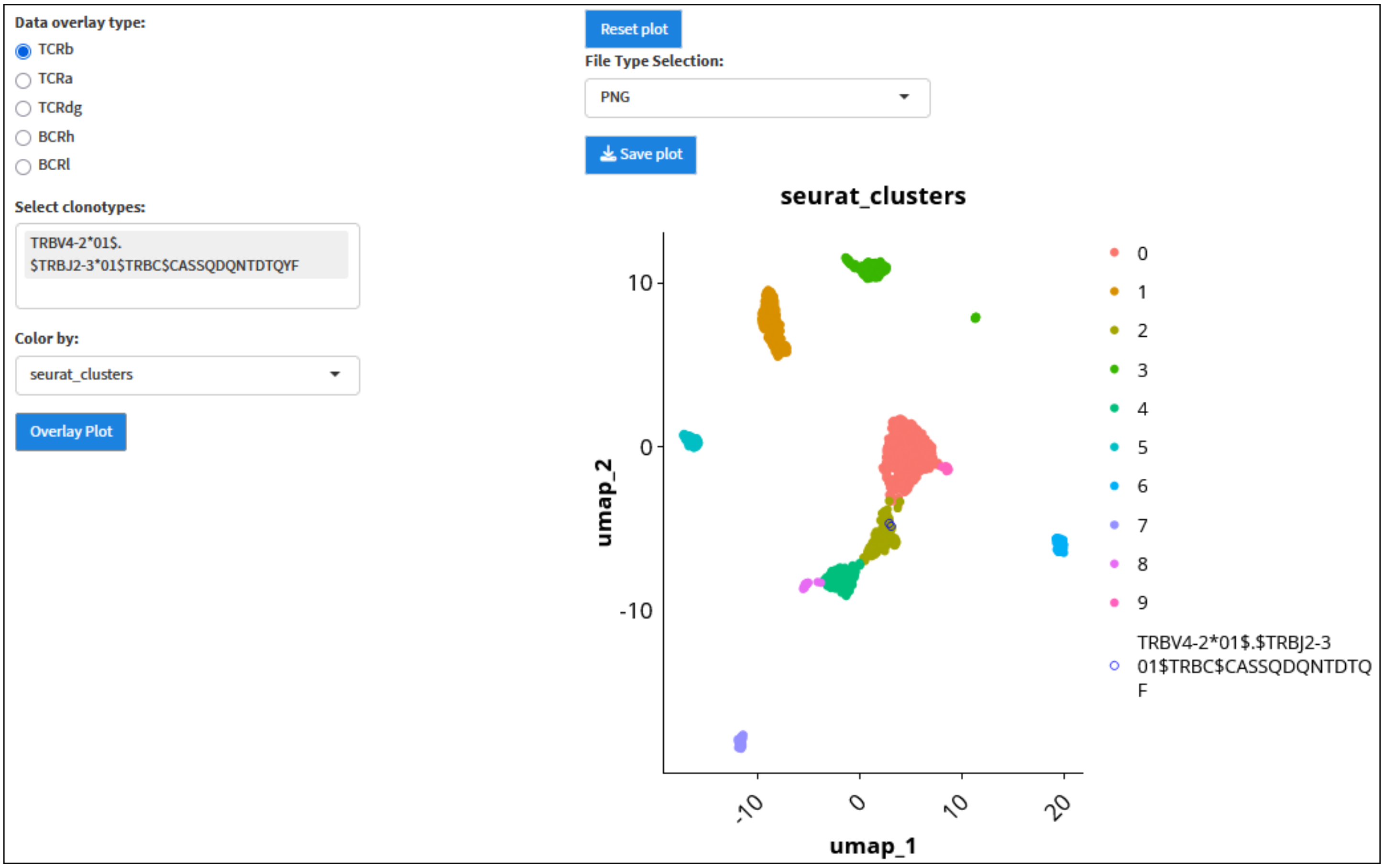
TRBV4-2*01$.
$TRBJ2-3*01$TRBC$CASSQDQNTDTQYF
TRBV12-4*01$TRBD1*01$TRBJ2-5*01$TRBC$CASSTGQSRETQYF
TRBV7-3*01$TRBD2*02$TRBJ2-5*01$TRBC$CASSPSGRAPETQYS
TRBV27*01$TRBD1*01$TRBJ2-4*01$TRBC$CASSPGTDKNIQYF
TRBV6-1*01$TRBD1*01$TRBJ2-7*01$TRBC$CASSEQGYEQYF

Color by: seurat_clusters

File Type Selection: PNG

Figure 89. *Clonotype Analysis* page, after plot reset. The parameters shown are the results of doing [Reset Plot] > [Overlay Plot] on the chart shown in Figure 88 Note that the selected clonotypes are still displayed in the drop-down menu box but are absent in the chart legend.

3. Add or remove a clonotype from the "Select clonotypes" list—a change MUST occur to regenerate the plot.
4. Click the [Overlay Plot] button again to visualize the new chart (Figure 90).



Data overlay type:

● TCRb
○ TCRa
○ TCRdg
○ BCRh
○ BCRl

Select clonotypes:

TRBV4-2*01$.
$TRBJ2-3*01$TRBC$CASSQDQNTDTQYF

Color by: seurat_clusters

File Type Selection: PNG

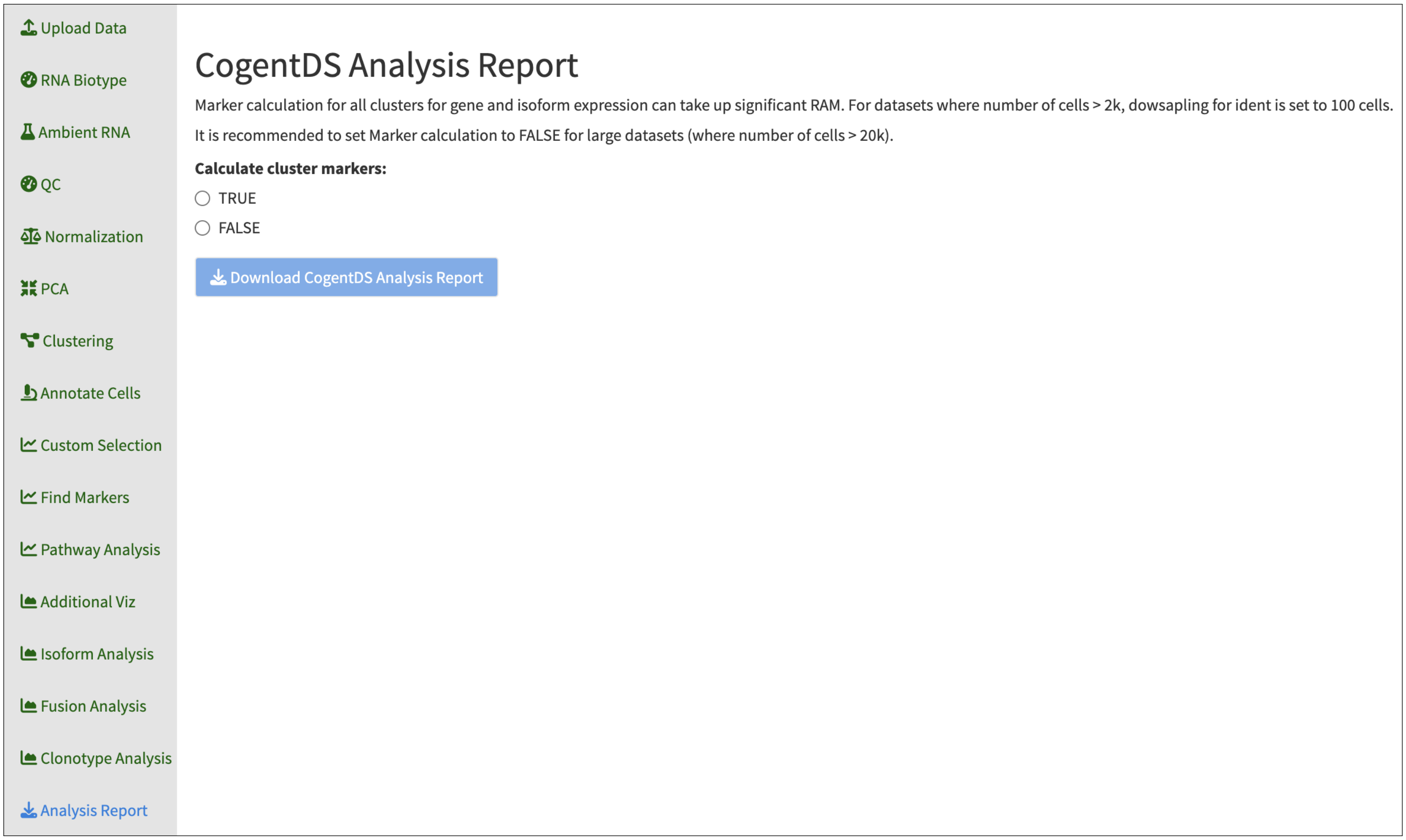
Figure 90. *Clonotype Analysis* page, after new clonotype selection and overlay execution. This chart started with the one depicted in Figure 89; the selected clonotypes of that figure were removed except the one listed in the screenshot.

(3) Color By

NOTE: Unlike the "Select clonotypes", you do not need to reset the plot prior to performing the steps below.

1. Select a new option from the "Color by" drop-down menu.
2. Click [Overlay Plot].

16. CogentDS Analysis Report



CogentDS Analysis Report

Marker calculation for all clusters for gene and isoform expression can take up significant RAM. For datasets where number of cells > 2k, dowsapling for ident is set to 100 cells. It is recommended to set Marker calculation to FALSE for large datasets (where number of cells > 20k).

Calculate cluster markers:

○ TRUE

○ FALSE

Figure 91. *CogentDS Analysis Report* page.

This page lets you download a comprehensive HTML report that reflects all the parameters and refinements applied during the analysis workflow. This HTML report is similar to the preliminary one generated from the Cogent NGS Analysis Pipeline but is customized to any new specifications made in CogentDS. While both the reports share similarities, they may display differences due to the following:

- The CogentAP report may show different results in PCA, Clustering, and Differential Expression Analysis because ambient RNA and QC modules are not performed.
- The CogentDS report also includes additional and customized results.

The "Calculate cluster markers" option, when marked 'TRUE', calculates markers for clusters and includes the markers in the report, while 'FALSE' omits this from the report.

Clicking [Download CogentDS Analysis Report] will prepare the report to pass to your browser as an HTML file. Depending on the settings of your browser, it may open the HTML file in the browser itself, follow the browser configuration to save the file, or prompt you to save or open the file.

Once it's saved (if desired), click [Next: Download Processed Data] to proceed to the final module.

17. Download CogentDS Processed .rds Data

In addition to downloading the modified HTML-version report, a Seurat-processed R data (RDS) file can be downloaded that includes all the additional analysis applied by CogentDS during the course of these steps

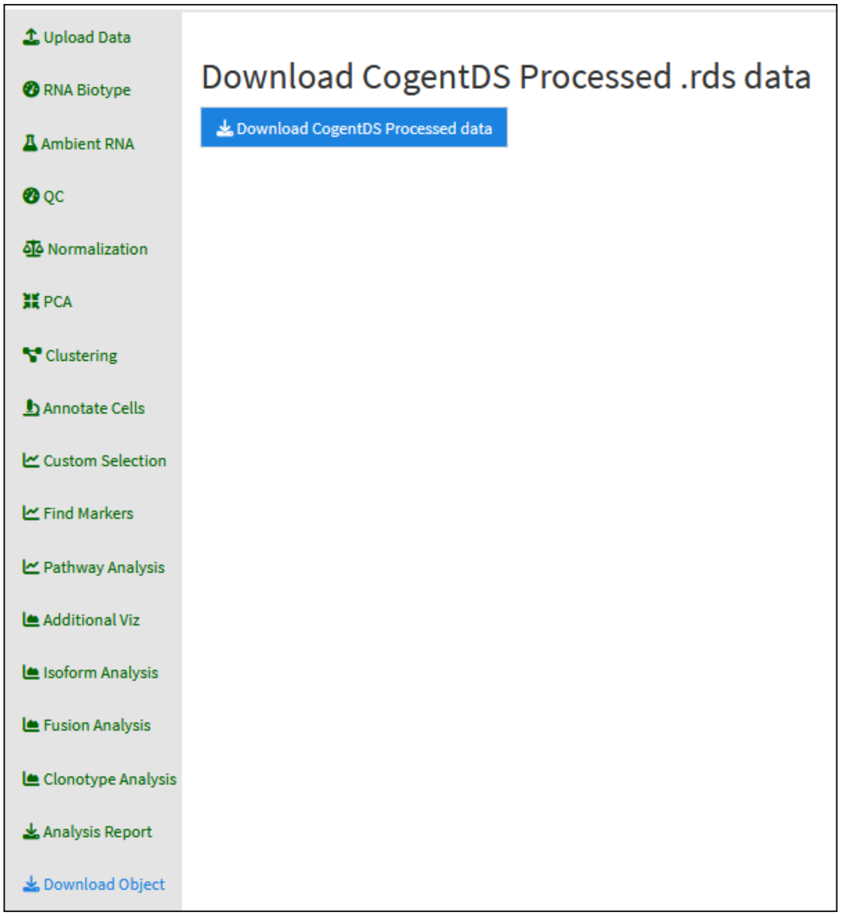


Figure 92. *Download CogentDS Processed .rds data* page.

Clicking the [Download CogentDS Processed data] button will initiate a data clean-up and synthesis of the new file, before honoring the browser download file configuration (generally either saving to the `Downloads/` folder or prompting to ask where it should be saved).

Once the file is saved (if desired), do one of the following:

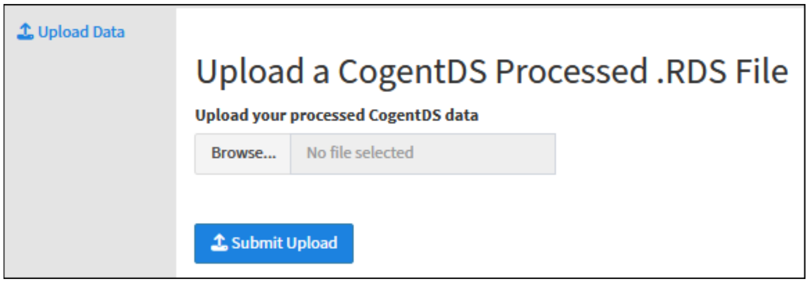
- Click the [Go to Main Page] button on the bottom right corner, which returns to the initial scRNA Analysis page
- Click the [Home] icon in the top right corner, in the title bar, which will also return to the scRNA Analysis page, or
- Close the browser tab or window. If the tab is closed, the main CogentDS window with the three main CogentDS applications (Section V.B) will still be open and available for use.

B. Discovery Mode

CogentDS v2.2 offers the Discovery Mode application, which allows users to upload the resulting RDS file saved after processing in Analysis Mode (Section VI.A.17, "Download CogentDS Processed .rds Data"). This application offers intuitive data visualization through dimension reduction techniques such as UMAP, t-SNE, and PCA, depending on the availability of these reductions in the imported RDS (Seurat object) file.

The modules of this application are described in the subsections below.

1. Upload Data

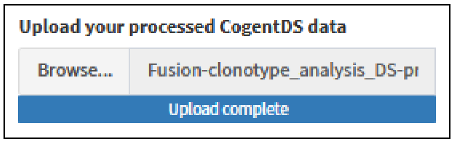


Upload a CogentDS Processed .RDS File

Upload your processed CogentDS data

Figure 93. *Upload a CogentDS Processed .RDS File* page of the scRNA Discovery application.

Click on [Browse…] to select your processed CogentDS data file. After selecting the file, click [Submit Upload]. Wait for the confirmation message that your data has been successfully uploaded.



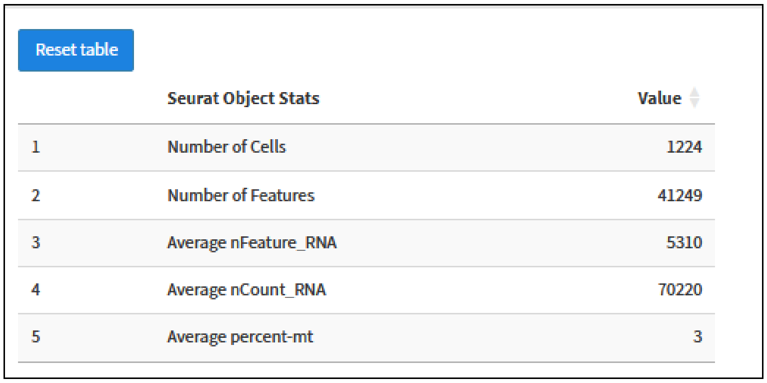
Upload your processed CogentDS data

Fusion-clonotype_analysis_DS-pı

Upload complete

Figure 94. The "Upload complete" message after importing an RDS file.

A table will display on the right side of the screen summarizing detected features found in your imported file. The table can be sorted by clicking on the Value column for ascending or descending order; [Reset table] will reset the display to the default, undoing the sort.

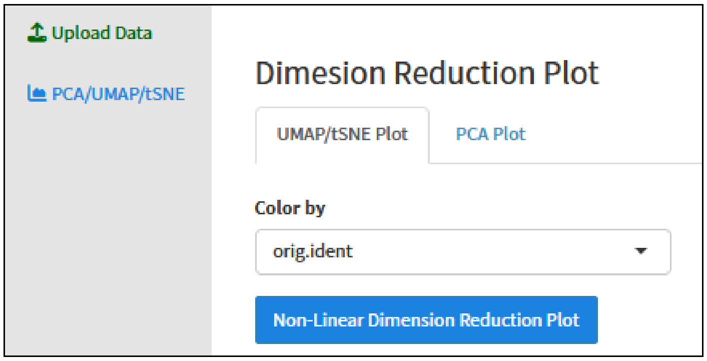


	Seurat Object Stats	Value
1	Number of Cells	1224
2	Number of Features	41249
3	Average nFeature_RNA	5310
4	Average nCount_RNA	70220
5	Average percent-mt	3

Figure 95. Seurat object stats table displayed after importing an RDS file into scRNA Discovery Mode.

Click on the [Next: Dimension Reduction Visualization] button in the bottom right corner of the screen to proceed to the next module. If the button is not visible, you may need to scroll down the page to view it.

2. Dimension Reduction Visualization (PCA/UMAP/t-SNE)



Dimesion Reduction Plot

Color by

orig.ident

Figure 96. *Dimension reduction plot* page in scRNA Discovery Mode.

This section contains two tabs: *UMAP/tSNE Plot* and *PCA Plot*.

- *UMAP/tSNE Plot*

  This tab displays the dimensionality reduction plot based on the available embeddings in the Seurat object:

  - UMAP—if UMAP embeddings are present, this will be shown.
  - tSNE—if t-SNE embeddings are available, this will be displayed.

  Use the drop-down menu of the "Color by" option to select a metadata column for coloring the single cell. This feature allows users to visually distinguish cells based on specific metadata information.

  When the setting is as desired, click the [Non-Linear Dimension Reduction Plot] button to generate the new chart.

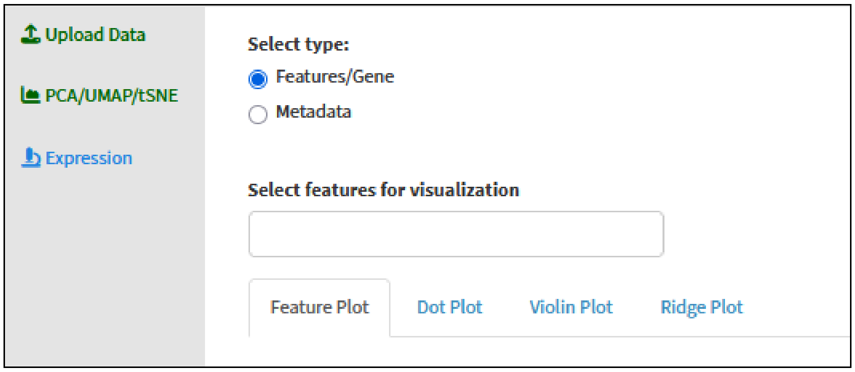
- *PCA Plot*

  This tab provides a visualization of PCA for the data.

  Use the drop-down menu of the "Color PC by" option to select a metadata column for coloring the single cell (similar to "Color by", above). When the setting is as desired, click the [PCA Plot] button to generate the new chart.

Save the plot(s), if desired, then click the [Next: Data Visualization] to proceed to the next module.

3. Data Visualization (Expression)



Select type:

● Features/Gene

○ Metadata

Select features for visualization

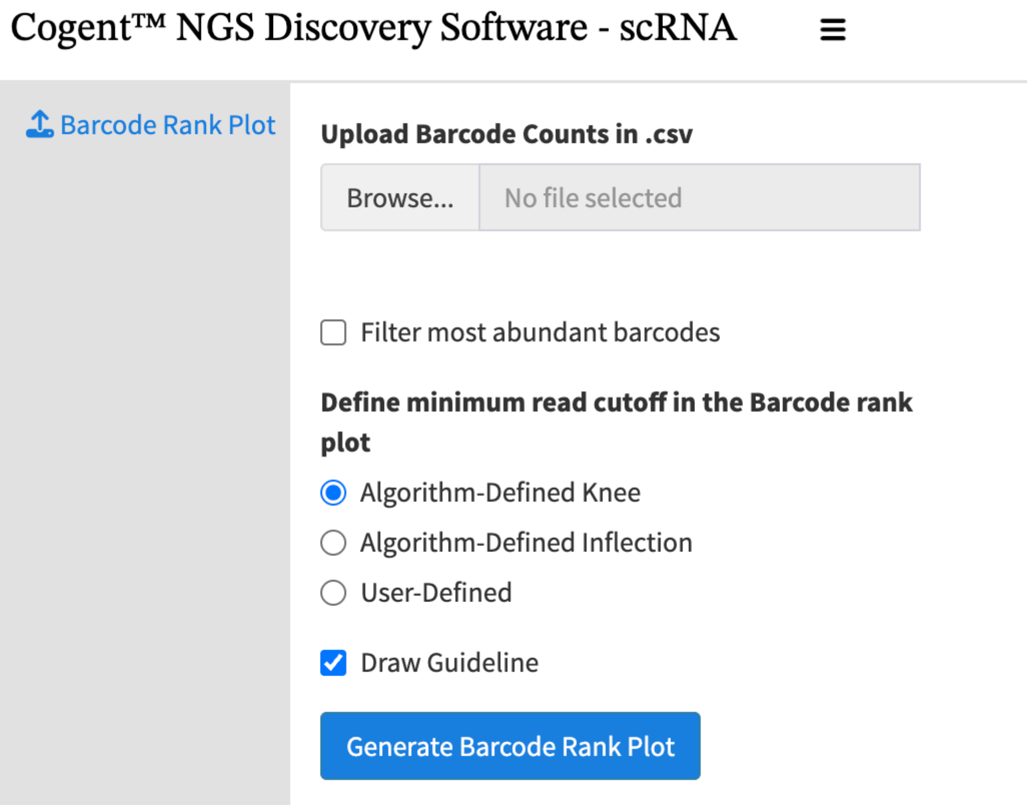
Figure 97. *Data visualization* page in scRNA Discovery Mode.

This module offers functionalities similar to the *Generate additional plots* module in Analysis Mode and is used to visualize gene expression patterns in your single-cell dataset. For more details, refer to Section VI.A.12, "Generate Additional Plots (Additional Viz)".

This is the final module of this application. To quit out of the application, do one of the following:

- Click the [Home] icon in the top right corner, in the title bar, which will also return to the scRNA Analysis page
- Close the browser tab or window. If the tab is closed, the main CogentDS window with the three main CogentDS applications (Section V.B) will still be open and available for use.

C. Barcode Rank Plot



Cogent™ NGS Discovery Software - scRNA ≡

Barcode Rank Plot

Upload Barcode Counts in .csv

Browse... No file selected

□ Filter most abundant barcodes

Define minimum read cutoff in the Barcode rank plot

● Algorithm-Defined Knee

○ Algorithm-Defined Inflection

○ User-Defined

☑ Draw Guideline

Generate Barcode Rank Plot

Figure 98. *Barcode rank plot* application page in scRNA Analysis.

This module generates a chart that visualizes the distribution of total reads across barcodes with the barcode rank on the X axis and total reads per cell on the Y axis. This module computes the rank of total reads and determines the inflection and "knee" points. The intersection point on the barcode rank plot displays the number of barcodes and corresponding total reads per barcode.

To generate this plot, the module uses an input file, `demultiplexed_fastqs_counts_all.estimated.csv`, a demux counts file generated from the CogentAP pipeline (see Section V.B of the Cogent NGS Analysis Pipeline User Manual), which contains barcodes along with their associated total read counts.

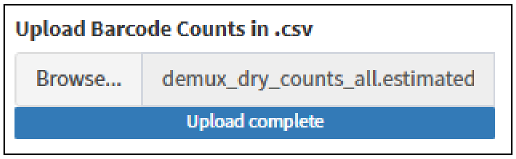
1. Run the CogentAP demux command (Section V.B.1, "RNA-Seq Analysis/Primary Analysis Commands") on your data of interest with the `--dry_run` argument.

   NOTE: An example of the dry_run command syntax is shown in subsection b) of the section referenced above, "RNA Demux Dry Run (For Analysis of Shasta Total RNA-Seq Kit Data)".

   However, the barcode rank plot can be generated in CogentDS for Shasta Total RNA-Seq kit data.

2. On **CogentDS>scRNA app>Barcode Rank Plot** page, click on [Browse…] to select the `demultiplexed_fastqs_counts_all.estimated.csv` file resulting from Step 1.

   Wait for the confirmation message that your data has been successfully uploaded.



Upload Barcode Counts in .csv

demux_dry_counts_all.estimated

Figure 99. Upload complete confirmation: dry run count output from CogentAP successfully uploaded to CogentDS.

The rest of the page has the following options to select from.

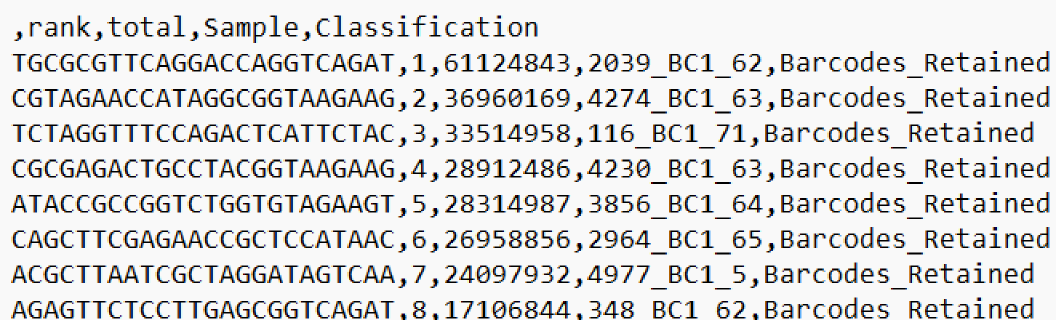
- Filter overly abundant barcodes: (Default: disabled) Enable this option to filter out barcodes that are unusually abundant, which may represent multiplets.
- Define minimum read cutoff in the Barcode rank plot: Choose from the options below for how to define the cutoff point that separates real cell barcodes from background noise:
    - Algorithm-Defined Knee: (Default option) The knee point is defined as the point where the signed curvature (second derivative) is minimized. To reduce noise sensitivity, a smooth spline is fitted to log total counts vs. log rank, and derivatives are calculated from this fit.
    - Algorithm-Defined Inflection: The inflection point is computed as the point where the first derivative of the curve is minimized.
    - User-Defined: Allows manual specification of the knee threshold in cases where the automatic (algorithm-defined) thresholds are too stringent or not suitable for your dataset. When this option is selected, an input field labeled "Minimum knee read cutoff" will appear, allowing you to enter a custom read count value.

  NOTE: For more information on how the knee and inflection points are defined, please refer to DropletUtils R package.

- Draw Guideline: (Default: enabled) The plot will display visual guidelines (e.g., cutoff lines) to aid interpretation of knee and inflection points. Unchecking this option will remove the lines from the chart.

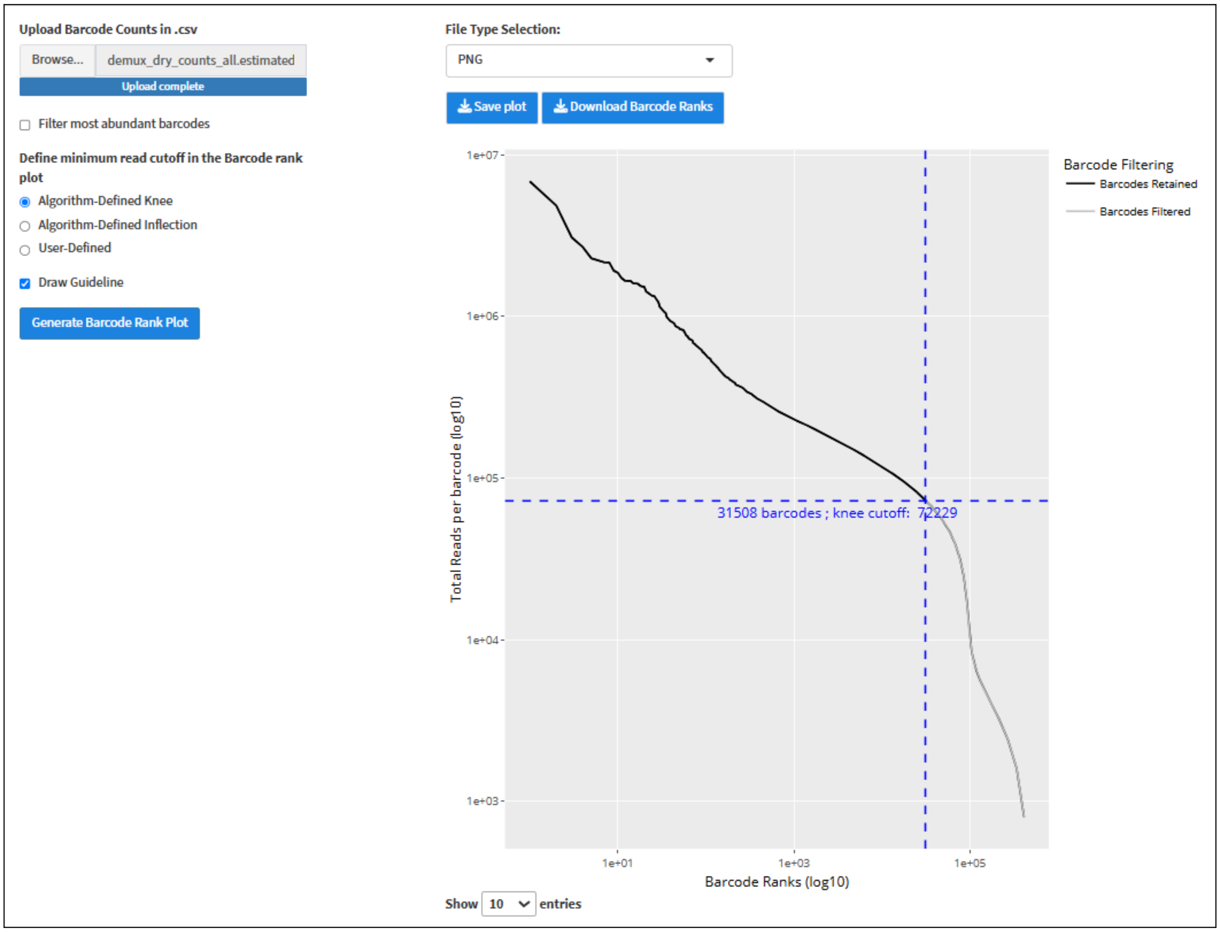
3. Once all parameters are set, click [Generate Barcode Rank Plot]. A chart will be displayed on the right side of the screen.

After the Barcode rank plot is generated, download the barcode ranks metadata by clicking the [Download Barcode Ranks] button. The file includes barcodes along with their rank, total count, fitted value (if applicable), sample information (TSO plate layout based) and classification, which indicates whether the barcode was retained. This metadata file is required as an input for the `analyze_direct` command in CogentAP v3.2 for analyzing Shasta Total RNA datasets.



```
,rank,total,Sample,Classification
TGCGCGTTCAGGACCAGGTCAGAT,1,61124843,2039_BC1_62,Barcodes_Retained
CGTAGAACCATAGGCGGTAAGAAG,2,36960169,4274_BC1_63,Barcodes_Retained
TCTAGGTTTCCAGACTCATTCTAC,3,33514958,116_BC1_71,Barcodes_Retained
CGCGAGACTGCCTACGGTAAGAAG,4,28912486,4230_BC1_63,Barcodes_Retained
ATACCGCCGGTCTGGTGTAGAAGT,5,28314987,3856_BC1_64,Barcodes_Retained
CAGCTTCGAGAACCGCTCCATAAC,6,26958856,2964_BC1_65,Barcodes_Retained
ACGCTTAATCGCTAGGATAGTCAA,7,24097932,4977_BC1_5,Barcodes_Retained
AGAGTTCTCCTTGAGCGGTCAGAT,8,17106844,348_BC1_62,Barcodes_Retained
```

Figure 100. Example of downloaded barcode ranks metadata.



Upload Barcode Counts in .csv

Browse... demux_dry_counts_all.estimated

Upload complete

□ Filter most abundant barcodes

Define minimum read cutoff in the Barcode rank plot

● Algorithm-Defined Knee

○ Algorithm-Defined Inflection

○ User-Defined

☑ Draw Guideline

File Type Selection: PNG

Show 10 entries

Figure 101. Example barcode rank plot in scRNA Analysis.

The barcode rank plot shows the barcode rank on the X axis based on the total reads per barcode on the Y axis. The inflection point is calculated, and based on this, you can select the number of barcodes to keep in the demultiplexing analysis in CogentAP.

This is the only module of this application. To quit out of the application, do one of the following:

- Click the [Home] icon in the top right corner, in the title bar, which will also return to the scRNA Analysis page
- Close the browser tab or window. If the tab is closed, the main CogentDS window with the three main CogentDS applications (Section V.B) will still be open and available for use.

VII. Application: Bulk RNA Analysis

From the initial CogentDS screen, click [Launch BulkRNA app], which will bring up the Bulk RNA application in a second browser window. From there, click on [Analysis Mode] to begin analysis of a bulk RNA-seq dataset.

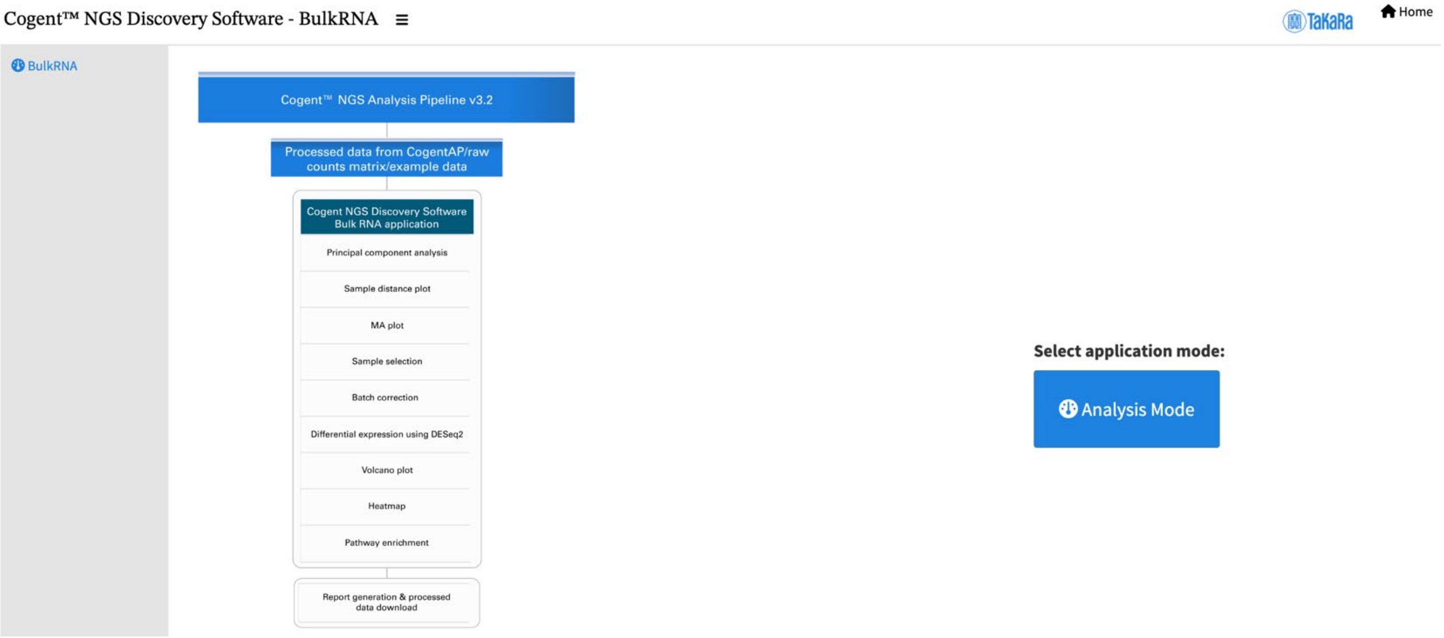
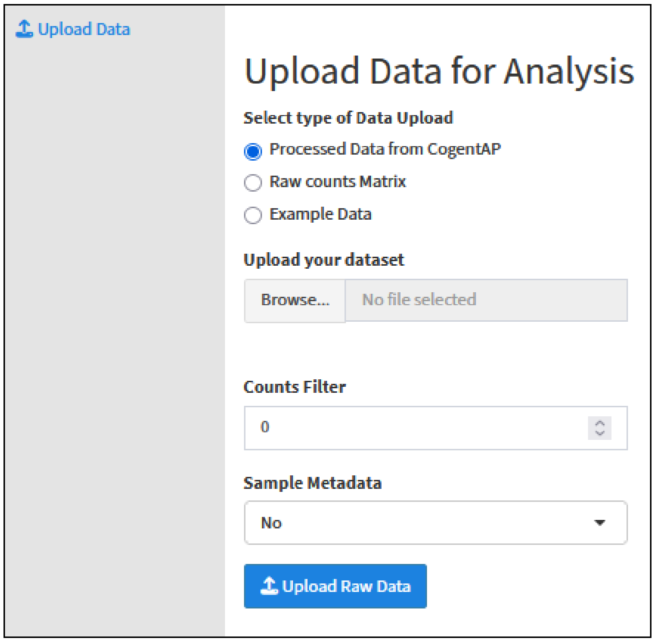

Cogent™ NGS Discovery Software - BulkRNA ≡
TaKaRa
Home
BulkRNA
Cogent™ NGS Analysis Pipeline v3.2
Processed data from CogentAP/raw counts matrix/example data
Cogent NGS Discovery Software Bulk RNA application
Principal component analysis
Sample distance plot
MA plot
Sample selection
Batch correction
Differential expression using DESeq2
Volcano plot
Heatmap
Pathway enrichment
Report generation & processed data download
Select application mode:
Analysis Mode

Figure 102. Initiating analysis mode in the Bulk RNA application. From the list of applications on the initial CogentDS page, click [Launch BulkRNA app] to bring up the *Bulk RNA app* window. Click [Analysis Mode] to begin the workflow.

A. Upload Data for Analysis



Upload Data for Analysis

Select type of Data Upload

● Processed Data from CogentAP

○ Raw counts Matrix

○ Example Data

Upload your dataset

Browse... No file selected

Counts Filter

0

Sample Metadata

No

Figure 103. ***Upload data for analysis*** **page in the Bulk RNA application.**

This module facilitates the uploading and preparation of bulk RNA sequencing data for analysis.

1. Choose the appropriate input data type using the radio buttons. Currently, CogentDS accepts either processed data from CogentAP (an RDS file) or a raw counts matrix (in CSV format). A small example dataset is also integrated into CogentDS to run if you would like to see how the application works.

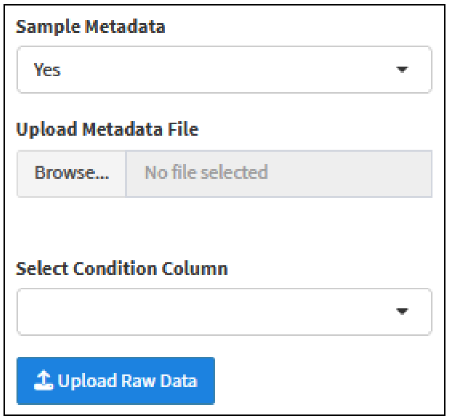
2. For either 'Processed Data from CogentAP' or the 'Raw counts Matrix' options, click [Browse…] to locate the appropriate input file. After selecting the file and clicking [Open], the file will automatically be uploaded to the tool.
3. After the upload complete message displays, you can configure the last two options on the page.
    - Counts Filter—allows for specification of the threshold for filtering of genes based on counts across a group of samples.
    - Sample Metadata—provides an option to add a metadata file. This is a simple 'Yes/No' drop-down menu that defaults to 'No'. If there is no sample metadata file, proceed to the next step.

      If 'Yes' is selected, the page display will change to what's shown in Figure 104.

      a. Click [Browse…] to locate and upload the metadata file.

      b. Use the drop-down menu of the "Select Condition Column" to specify the condition for samples like control and treatment from the uploaded metadata. This is required for further analysis, especially MA plot and DE Analysis.

      Select the Condition column from the uploaded metadata. This is required for further analysis, especially for MA plot and DE Analysis.



Sample Metadata: Yes

Upload Metadata File: Browse... No file selected

Select Condition Column

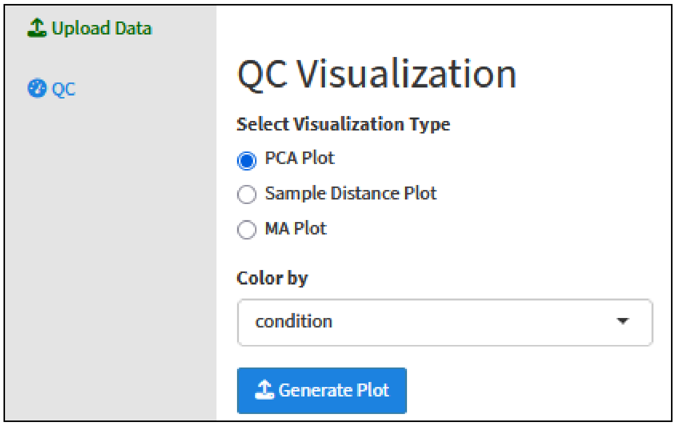
Figure 104. Uploading sample metadata in the Bulk RNA application.

3. Click [Upload Raw Data].

Analysis of the input or example data will proceed, and a table will display on the right side of the window listing statistics about the data that could provide insights into its quality.

Click [Next: Check Data Quality] to proceed to the next module.

B. QC Visualization (Check Data Quality)



Upload Data

QC

QC Visualization

Select Visualization Type

● PCA Plot

○ Sample Distance Plot

○ MA Plot

Color by

condition ▾

Generate Plot

Figure 105. The *QC visualization* page in the Bulk RNA application.

The QC module allows you to assess the quality of your data using several visualization types. Select the type of chart you would like to generate and values for any additional option related to the chart type, then click [Generate Plot] to visualize the data.

The options and example plot charts, visualized from the 'Example data' option in the previous section, are listed below.

- PCA Plot—displays the variance covered by the first principal component (PC1) vs. the second principal component (PC2) across your samples, providing an overview of concordance among samples and their groups.

  Use the "Color by" dropdown to select columns from the metadata file to color code the samples in the plot.

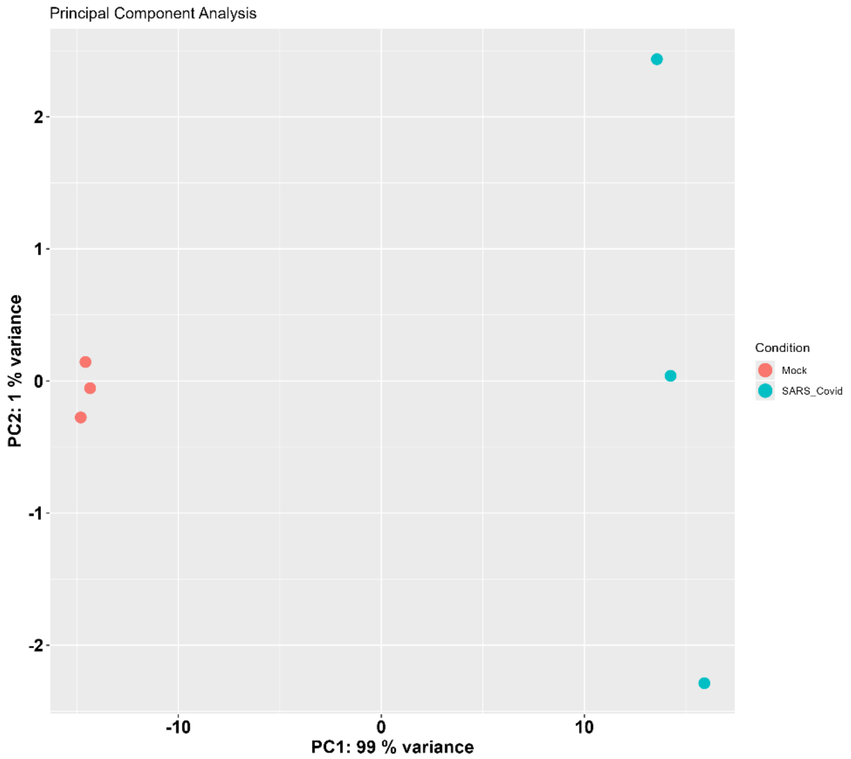


Figure 106. Example PCA plot under the *QC visualization* module of the Bulk RNA application.

- Sample Distance Plot—drawn as a heatmap, this chart illustrates the distances/correlations among the samples, based on the mapped read counts across genes.

  Use the "Color by" drop-down menu to select a column from the metadata, which will add annotations for the column in the heatmap. This annotation might help users in identifying the specific groupings for the samples.

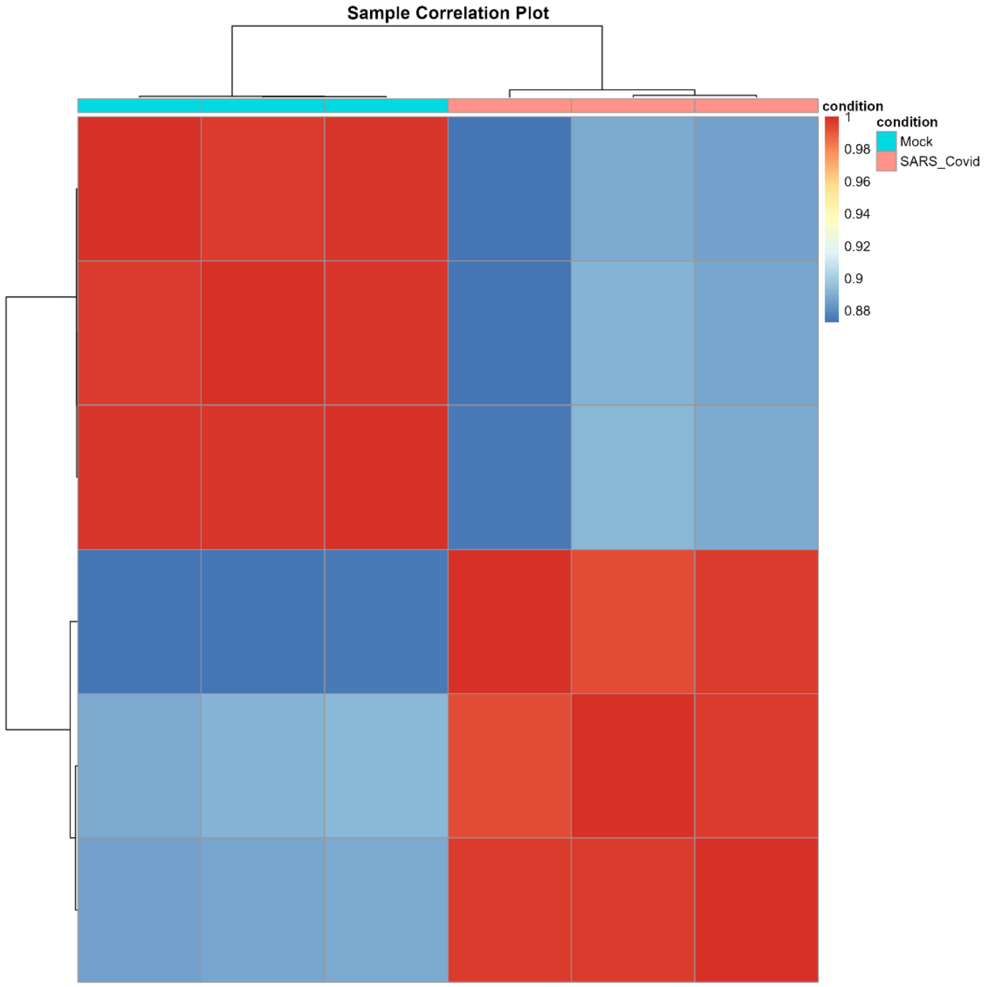


Figure 107. Example sample distance heatmap plot under the *QC visualization* module of the Bulk RNA application.

- MA Plot—visualizes the differences between the groups of samples, as specified by the conditions set in the *Upload Data* module.

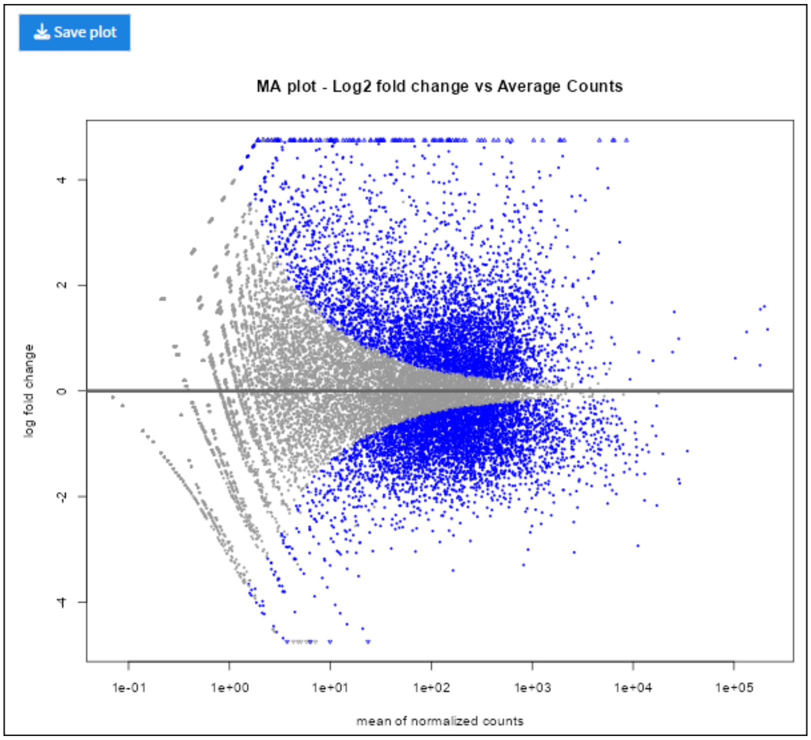
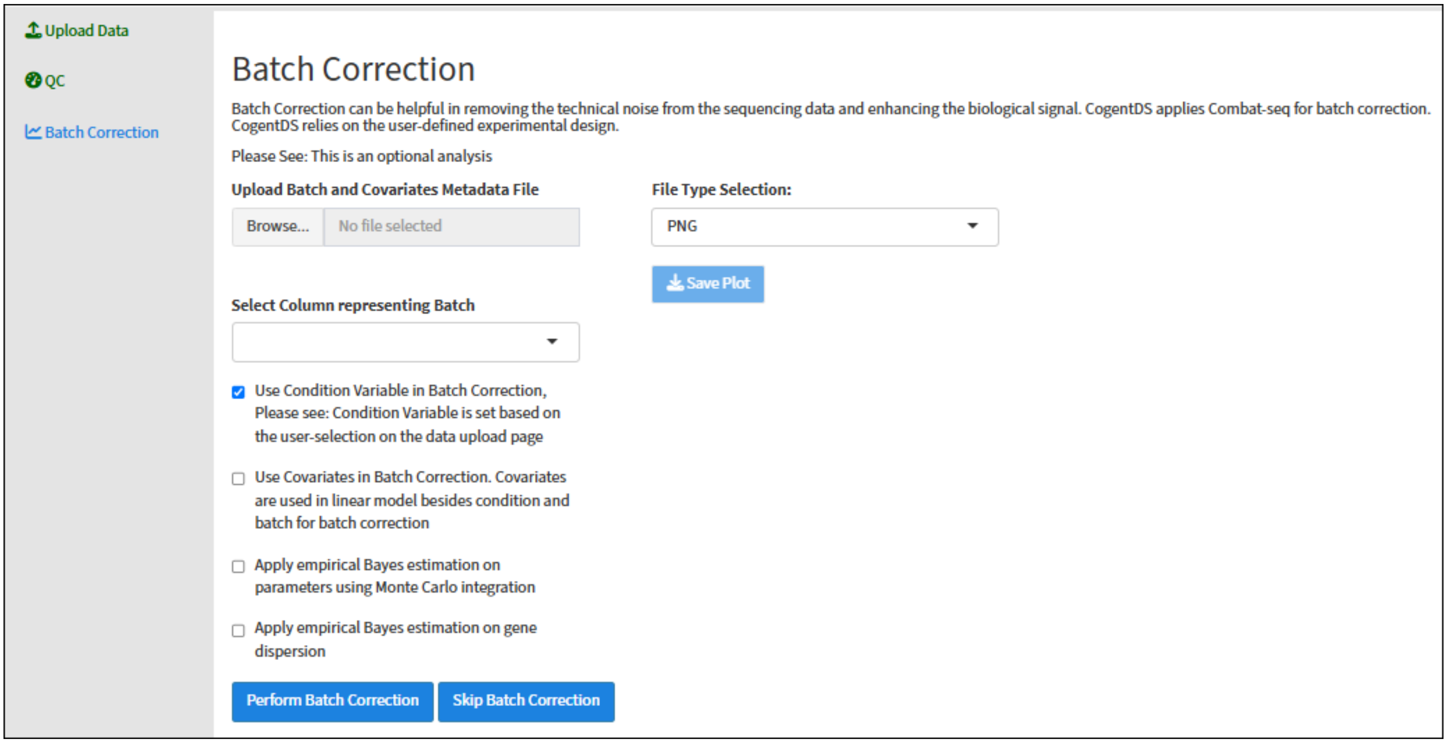


Figure 108. Example MA plot under the *QC visualization* module of the Bulk RNA application.

The resulting plots can be downloaded to your local computer using the [Save plot] button, if desired, and then click [Next: Perform Differential Expression] to proceed to the next module.

C. Batch Correction



Batch Correction

Batch Correction can be helpful in removing the technical noise from the sequencing data and enhancing the biological signal. CogentDS applies Combat-seq for batch correction. CogentDS relies on the user-defined experimental design.

Please See: This is an optional analysis

Upload Batch and Covariates Metadata File

Browse... No file selected

File Type Selection: PNG

Select Column representing Batch

☑ Use Condition Variable in Batch Correction, Please see: Condition Variable is set based on the user-selection on the data upload page

□ Use Covariates in Batch Correction. Covariates are used in linear model besides condition and batch for batch correction

□ Apply empirical Bayes estimation on parameters using Monte Carlo integration

□ Apply empirical Bayes estimation on gene dispersion

Figure 109. The *Batch Correction* page in the Bulk RNA application.

The Batch Correction module leverages Combat-seq to correct for batch effects in bulk RNA-seq data. This helps minimize technical variation and highlight true biological differences across samples.

NOTES:

- This is an optional step in the process. It can be skipped by clicking the [Next: Perform Differential Expression] button on the bottom right corner of the page.
- For more information about the application of batch correction to your data, please see https://academic.oup.com/nargab/article/2/3/lqaa078/5909519

To implement batch correction, perform the following steps.

1. Upload Batch and Covariates Metadata File: (Required) upload a metadata file (CSV) containing information about sample batches and any covariates you wish to include in the correction model.
2. Select Column representing Batch: (Required) after uploading the file, select the column representing the batch value from the dropdown. The input will be a numerical column.

sample	condition	batch
Sample1	condition1	1
Sample2	condition1	1
Sample3	condition1	2
Sample4	condition1	2
Sample5	condition2	2
Sample6	condition2	3

Figure 110. Example metadata file for batch correction. This example is provided to illustrate the type of information contained in the metadata file. Please make sure that the batch column is numerical and the file is saved as .csv.

NOTE: This must be a column in the metadata file not selected as the condition column during the upload (Section VII.A).

3. Select from the following options to perform batch correction:
    - Use Condition Variable in Batch Correction: (Default: enabled)

      The condition variable defined during data upload is included in the model.
    - Use Covariates in Batch Correction: (Default: disabled)

      If enabled, additional variables besides batch and condition from your metadata may be included in the model.

      When enabled, a selection box is displayed.
    - Apply empirical Bayes estimation on parameters using Monte Carlo integration: (Default: disabled)

      Enable this option to apply empirical Bayes shrinkage to parameter estimates using Monte Carlo integration.

      If this is enabled, an additional parameter, "Number of Genes for Estimation", becomes available (Default: 1000) to specify how many genes to use for empirical Bayes estimation.
    - Apply empirical Bayes estimation on gene dispersion: (Default: disabled)

      Enable to apply shrinkage on dispersion.

NOTE: Empirical Bayes options may help when data contains outliers, but can increase run time.

4. Once all desired parameters are configured, click [Perform Batch Correction] to apply the selected settings.

   The following plots are generated to visualize the effect of batch variation on your data.

   - PCA plot, before (No batch correction)
   - Sample Correlation Heatmap, before (No batch correction)

   The module will generate:

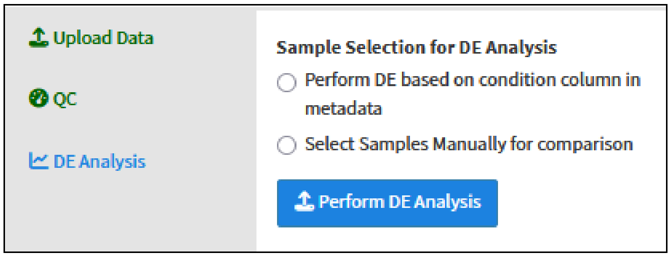
   - PCA plot after batch correction
   - Sample Correlation Heatmap after batch correction.

   The before and after charts are displayed side-by-side to better allow for direct comparison between the charts.

If by comparing the data, you determine that batch correction is not needed to proceed with your analysis, you may click [Skip Batch Correction] to continue with the original data.

After either option, click [Next: Perform Differential Expression] to continue to the next module.

D. DE Analysis (Perform Differential Expression)



Sample Selection for DE Analysis

○ Perform DE based on condition column in metadata

○ Select Samples Manually for comparison

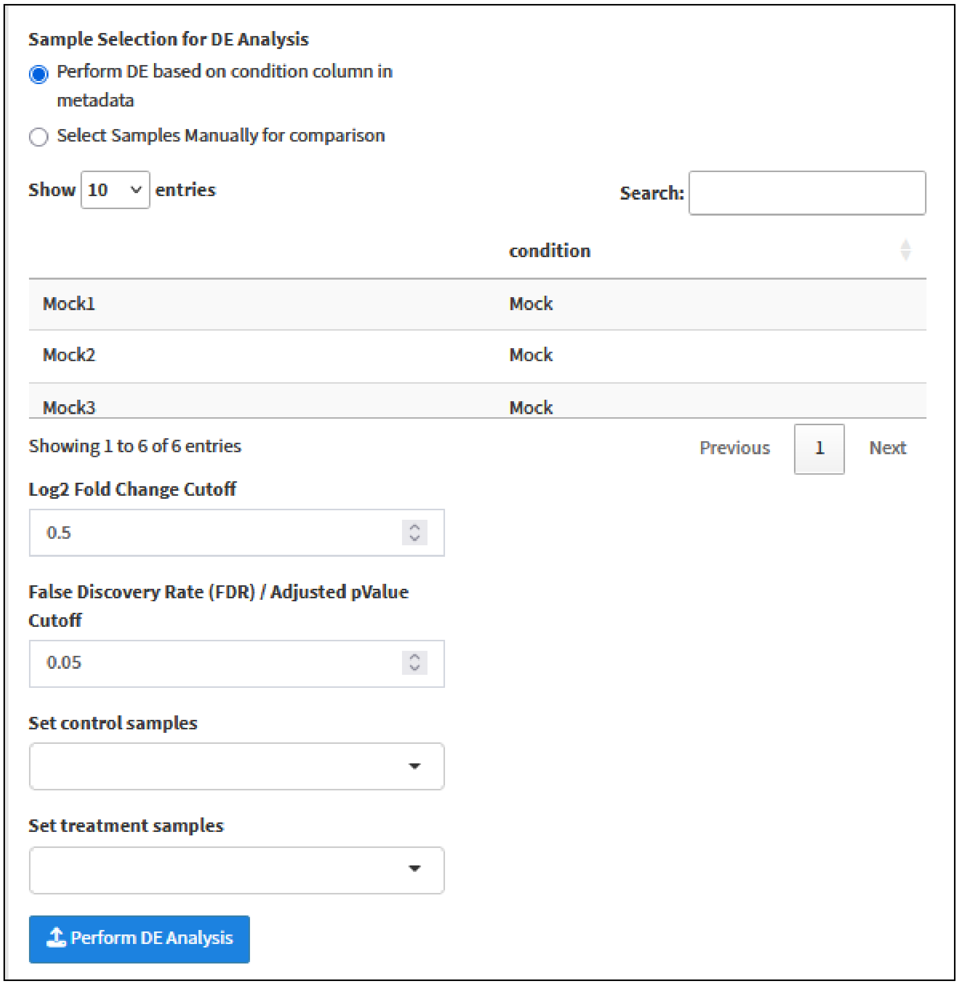
Figure 111. The *DE Analysis* page in the Bulk RNA application.

This module performs differential expression (DE) analysis on your data. This can be done in two ways: either based on the condition column in the metadata or by manually selecting samples for comparison.

The screenshots in this section are based on the Example Data provided with CogentDS. Your display will vary slightly, based on your input data.

The resulting table can be downloaded to your local computer using the [Save Data] button, if desired, and then click [Next: Visualization] to proceed to the next module.

1. Perform DE Based on the Condition Column



Sample Selection for DE Analysis

● Perform DE based on condition column in metadata

○ Select Samples Manually for comparison

Show 10 entries

Search:

	condition
Mock1	Mock
Mock2	Mock
Mock3	Mock

Showing 1 to 6 of 6 entries

Log2 Fold Change Cutoff: 0.5

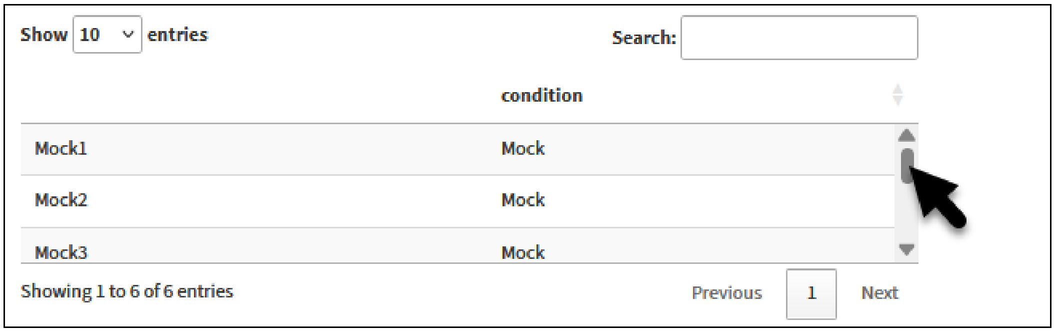
False Discovery Rate (FDR) / Adjusted pValue Cutoff: 0.05

Set control samples

Set treatment samples

Figure 112. Bulk RNA *DE Analysis* page where the 'Perform DE based on condition column in metadata' option is selected.

If this option is selected, the first information shown is the condition table, which lists the available conditions from the selected metadata. The drop-down menu for "Show entries" can be used to increase or decrease the number of results (from 10–100) shown per 'page' of the table, and the "Search" box can be used to filter by specific text strings. The viewing scrollbar becomes active when hovering your mouse cursor over the table contents (Figure 100).



Show 10 entries

Search:

	condition
Mock1	Mock
Mock2	Mock
Mock3	Mock

Showing 1 to 6 of 6 entries

Figure 113. The condition table display in the Bulk RNA *DE analysis* page. The scrollbar for the table is shown as activated when the mouse cursor hovers over it.

After the condition table, the following configuration options display:

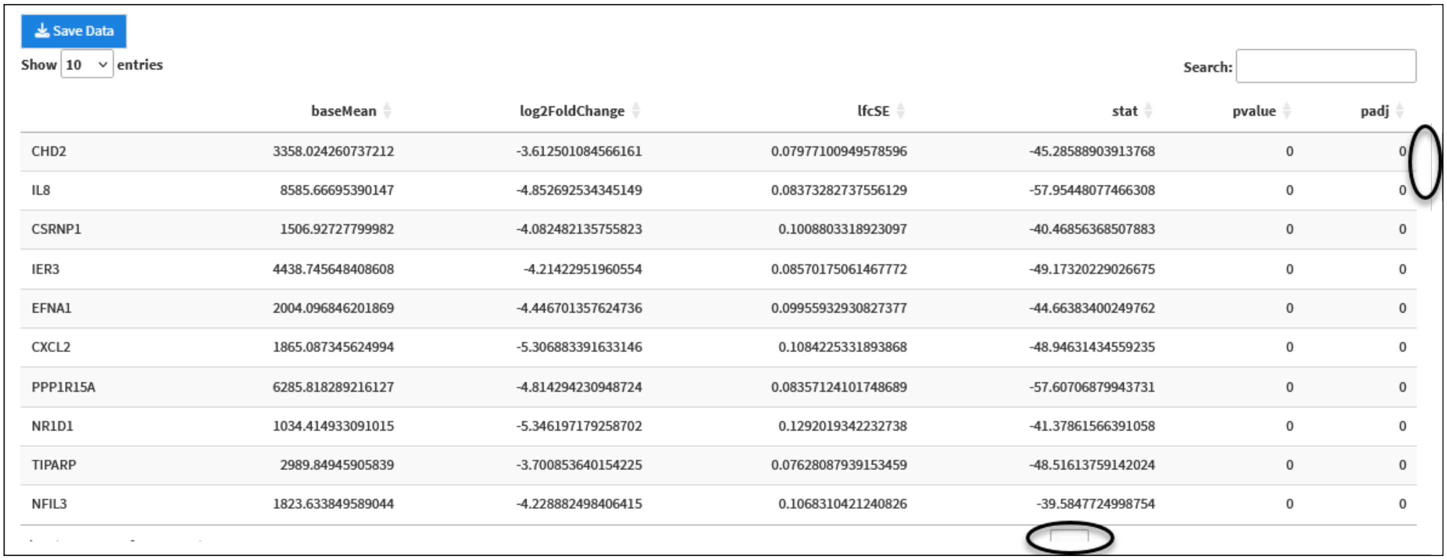
- Log2 Fold Change Cutoff—defines the minimum absolute value for the log fold change required for a gene to be considered significant. The default value is '0.5'.
- False Discover Rate (FDR)/Adjusted pValue Cutoff—sets the cutoff for adjusted p-values, helping to identify statistically significant genes. The default value is '0.05'.
- Set control samples/Set treatment samples—the drop-down menus for each of these parameters allow you to define the conditions corresponding to control and treatment groups, respectively, for DE analysis. These are required to be different values.

After setting the parameters and sample groups, click the [Perform DE Analysis] button to execute the analysis using a Deseq2-based approach. An option, "Select Coefficients", will display under the button showing only information about the two conditions for which the DE analysis will be performed. The results of the analysis will be displayed in a table format with the columns shown in Table 6.

Table 6. Columns of the bulk RNA DE analysis results table performed based on the condition column.

Column name	Description
baseMean	Average of normalized counts across all samples.
log2FoldChange	The change in expression of a gene between two groups being compared, measured on a logarithmic scale with base-2.
lfcSE	Standard error estimates of the log2 fold change.
stat	Corresponds to Wald Statistics.
pvalue	Wald test p-values.
padj	Represents the p-value adjusted for multiple testing. This is the most important column of the results.

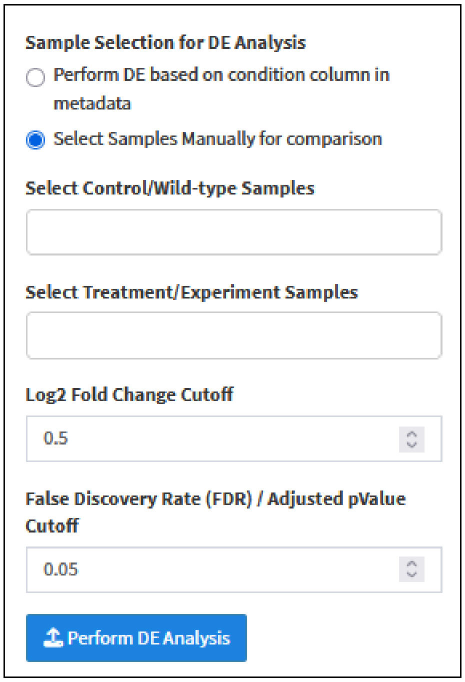
NOTE: For a detailed interpretation of Deseq2 output, it is recommended to refer to the official Deseq2 documentation.



	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
CHD2	3358.024260737212	-3.612501084566161	0.07977100949578596	-45.28588903913768	0	0
IL8	8585.66695390147	-4.852692534345149	0.08373282737556129	-57.95448077466308	0	0
CSRNP1	1506.92727799982	-4.082482135755823	0.1008803318923097	-40.46856368507883	0	0
IER3	4438.745648408608	-4.21422951960554	0.08570175061467772	-49.17320229026675	0	0
EFNA1	2004.096846201869	-4.446701357624736	0.09955932930827377	-44.66383400249762	0	0
CXCL2	1865.087345624994	-5.306883391633146	0.1084225331893868	-48.94631434559235	0	0
PPP1R15A	6285.818289216127	-4.814294230948724	0.08357124101748689	-57.60706879943731	0	0
NR1D1	1034.414933091015	-5.346197179258702	0.1292019342232738	-41.37861566391058	0	0
TIPARP	2989.84945905839	-3.700853640154225	0.07628087939153459	-48.51613759142024	0	0
NFIL3	1823.633849589044	-4.228882498406415	0.1068310421240826	-39.5847724998754	0	0

Figure 114. DE Analysis result table. Scrollbars (rough location circled in the screenshot) allow both horizontal and vertical scrolling of the data.

2. Select Samples Manually for Comparison



Sample Selection for DE Analysis

○ Perform DE based on condition column in metadata

● Select Samples Manually for comparison

Select Control/Wild-type Samples

Select Treatment/Experiment Samples

Log2 Fold Change Cutoff: 0.5

False Discovery Rate (FDR) / Adjusted pValue Cutoff: 0.05

Figure 115. Bulk RNA *DE Analysis* page where the 'Select Samples Manually for comparison' option is selected.

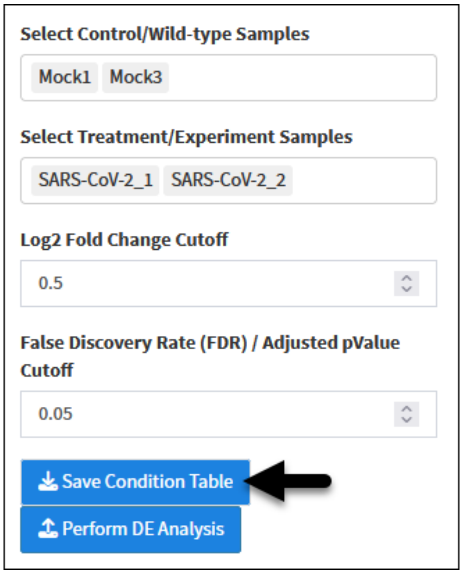
If Select samples manually for comparison is selected as an option, the following options/parameters will be displayed:

- "Select Control/Wild-type Samples" and "Select Treatment/Experiment Samples"—define the control and treatment groups that will be used in DE analysis.

  NOTE: More than one sample should be selected for both control and treatment groups for valid comparison

- Log2 Fold Change Cutoff—defines the minimum absolute value for the log fold change required for a gene to be considered significant. The default value is '0.5'.
- False Discovery Rate (FDR)/Adjusted pValue Cutoff—sets the cutoff for adjusted p-values, which aids in identifying statistically significant genes. The default value is '0.05'.

After the control and treatment sample types are selected from the drop-down menus, a second button, [Save Condition Table], displays after the FDR option (Figure 116). If desired, click on the button to save the manual parameter definitions as a CSV file for future reference.



Select Control/Wild-type Samples: Mock1 Mock3

Select Treatment/Experiment Samples: SARS-CoV-2_1 SARS-CoV-2_2

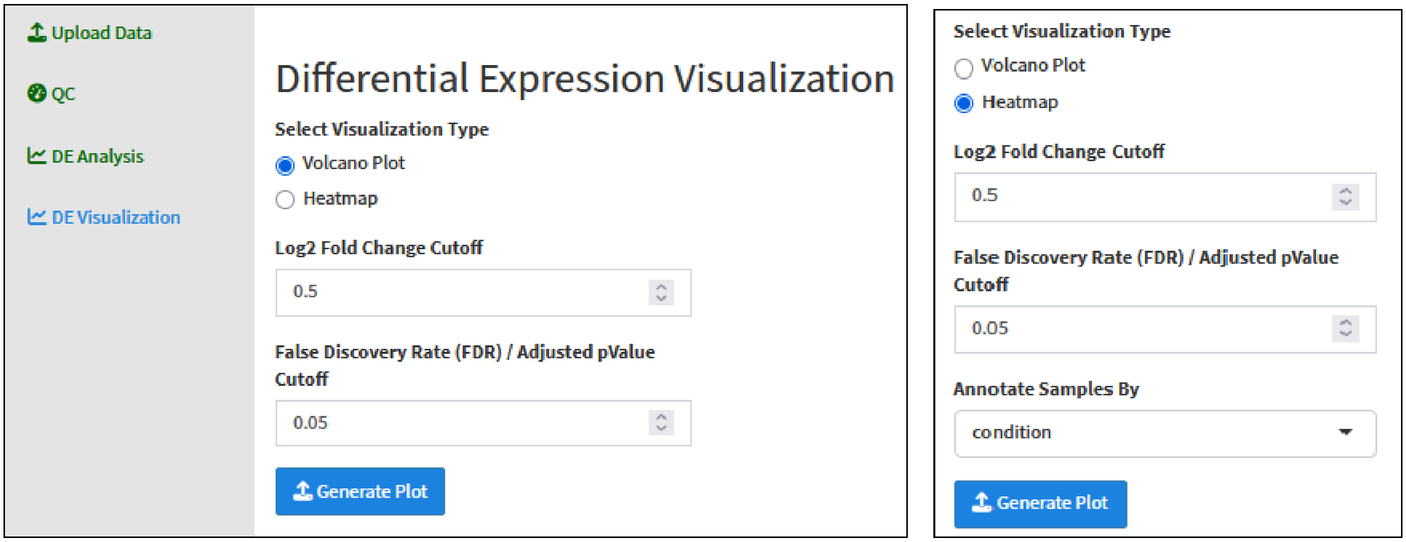
Log2 Fold Change Cutoff: 0.5

False Discovery Rate (FDR) / Adjusted pValue Cutoff: 0.05

Figure 116. [Save Condition Table] button, after control and experiment sample types are selected manually for DE analysis.

After setting the parameters and sample groups, click [Perform DE Analysis] to execute the analysis using a Deseq2-based approach. The results will be displayed in a table format, similar to the previous section (Figure 114).

E. Differential Expression Visualization (DE Visualization)



Differential Expression Visualization

Select Visualization Type
● Volcano Plot
○ Heatmap

Log2 Fold Change Cutoff: 0.5

False Discovery Rate (FDR) / Adjusted pValue Cutoff: 0.05

Select Visualization Type
○ Volcano Plot
● Heatmap

Log2 Fold Change Cutoff: 0.5

False Discovery Rate (FDR) / Adjusted pValue Cutoff: 0.05

Annotate Samples By: condition

Figure 117. The *Differential Expression Visualization* page in the Bulk RNA application. **(Left)** The default view with 'Volcano Plot' selected. **(Right)** With 'Heatmap' selected.

This module provides two visualization options for helping users to interpret the results of DE analysis: volcano plot and heatmap.

- Volcano Plot (default)

  When Volcano Plot is selected, the following parameters will be displayed:

  - ○ Log2 Fold Change Cutoff—defines the minimum absolute value for the log fold change required for a gene to be considered significant. The default value is '0.5'.
  - ○ False Discovery Rate (FDR)/Adjusted pValue Cutoff—sets the cutoff for adjusted p-values, which helps to identify statistically significant genes. The default value is '0.05'.

- Heatmap

  When you select this option, the same parameters as the volcano plot (Log2 Fold Change, FDR) will be displayed, with the addition of:

  - ○ Annotate Samples By—select a metadata column using the drop-down menu to add annotations to the heatmap.

Once you've reviewed and/or set your parameters, click [Generate Plot] to synthesize the charts. Download the plots to your local computer using the [Save plot] button, if desired, and then click [Next: Pathway Enrichment] to proceed to the next module.

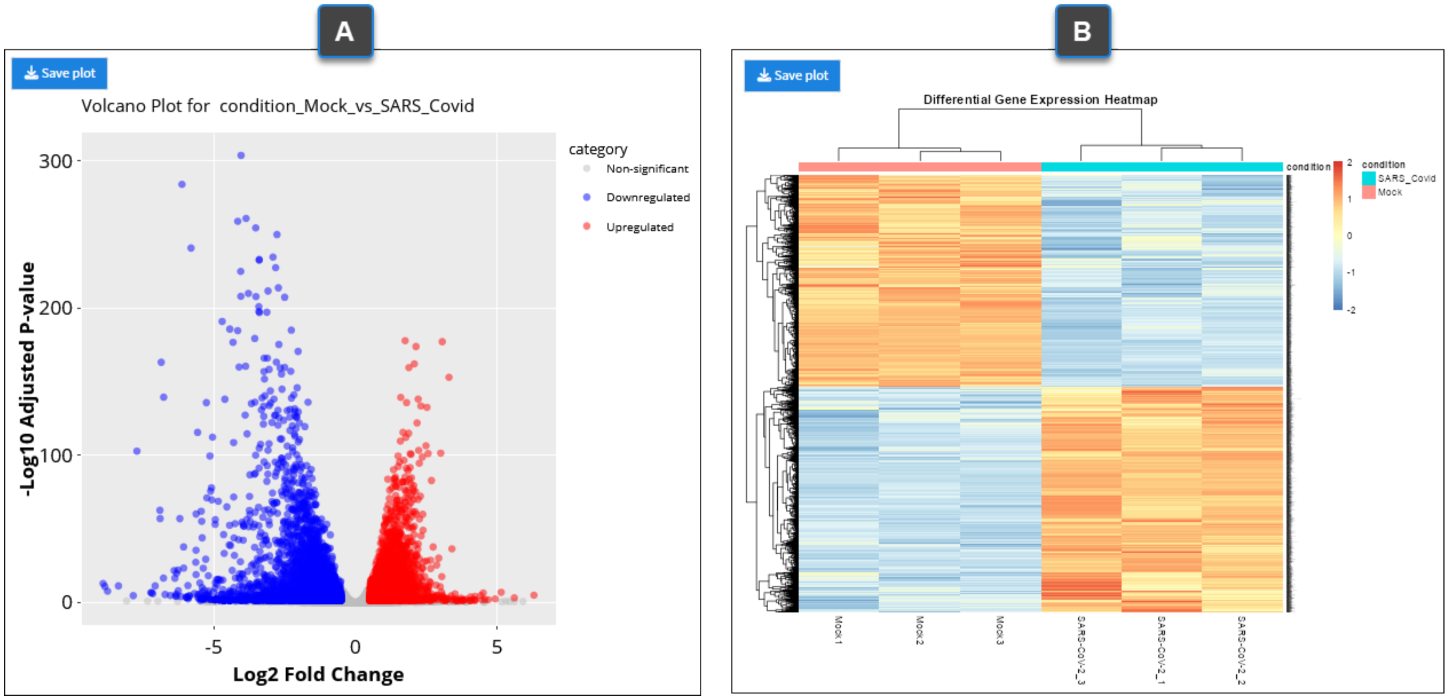


Figure 118. *Differential Expression Visualization* result plots in the Bulk RNA application. **(Left)** Volcano plot from the included example data with default parameter configuration. **(Right)** Heatmap with default parameters.

F. Pathway and Enrichment Analysis (Pathway Enrichment)

The pathway and enrichment module helps perform enrichment analysis based on the DE genes obtained in DE Module (Section D).

The functionality is identical to the same module in the scRNA app>Analysis Mode. For more information, including available options and parameters, refer to Section VI.A.11, "Pathway and Enrichment Analysis".

After applying the desired parameters and generating the results, save the plot or download the results, if desired, and click [Go to Analysis Report] to proceed to the next module.

G. CogentDS Analysis Report & Download CogentDS Processed .rds Data

The last two pages of the Bulk RNA application are identical to the final modules of the scRNA app>Analysis Mode workflow.

1. For more information on the analysis report page, refer to Section VI.A.16, "CogentDS Analysis Report". Click [Go to Download Page] to proceed to the final module.
2. For more information on the data download page, refer to Section VI.A.17, "Download CogentDS Processed .rds Data".

Once the data file is saved (if desired), do one of the following:

- Click the [Go to Main Page] button on the bottom right corner, which returns to the initial BulkRNA Analysis page
- Click the [Home] icon in the top right corner, in the title bar, which will also return to the BulkRNA Analysis page, or
- Close the browser tab or window. If the tab is closed, the main CogentDS window with the three main CogentDS applications (Section V.B) will still be open and available for use.

VIII. Application: scDNA Analysis

From the initial CogentDS screen, click [Launch scDNA app], which will bring up the single-cell DNA application in a second browser window. From there, click on [CNV Analysis Mode] or [SNV Analysis Mode] to begin.

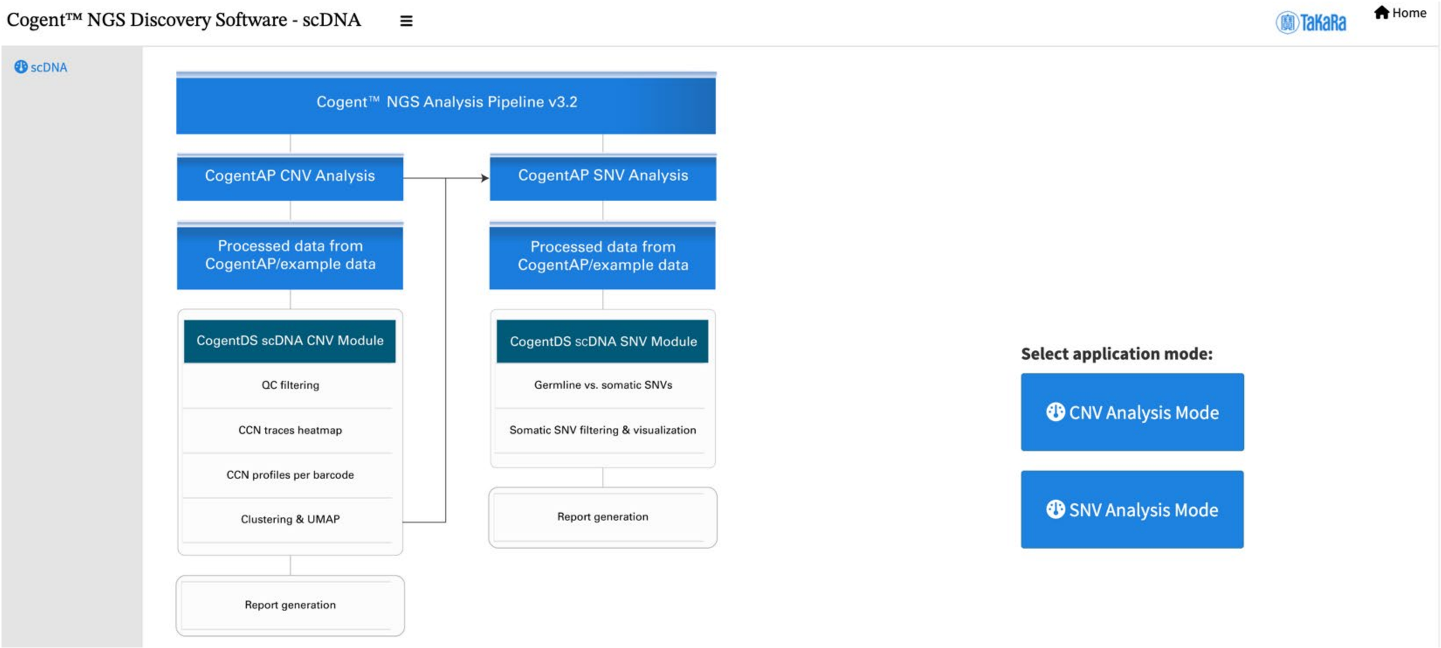
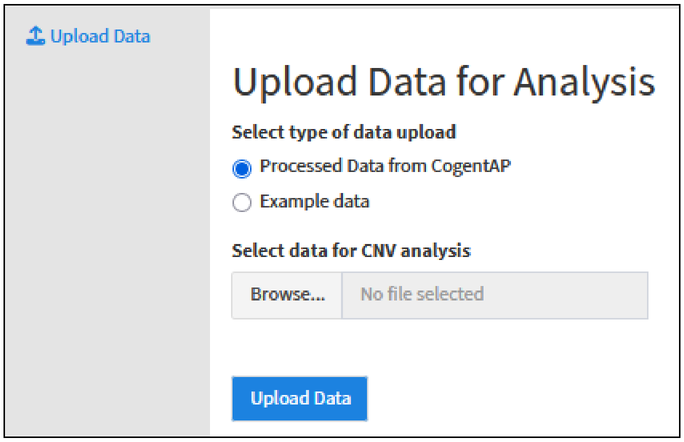

Cogent™ NGS Discovery Software - scDNA
≡
TaKaRa
Home
scDNA
Cogent™ NGS Analysis Pipeline v3.2
CogentAP CNV Analysis
CogentAP SNV Analysis
Processed data from CogentAP/example data
Processed data from CogentAP/example data
CogentDS scDNA CNV Module
QC filtering
CCN traces heatmap
CCN profiles per barcode
Clustering & UMAP
Report generation
CogentDS scDNA SNV Module
Germline vs. somatic SNVs
Somatic SNV filtering & visualization
Report generation
Select application mode:
CNV Analysis Mode
SNV Analysis Mode

Figure 119. Initiating analysis mode in the scDNA application.

A. CNV Analysis Mode

1. Upload Data for Analysis



Upload Data for Analysis

Select type of data upload

● Processed Data from CogentAP

○ Example data

Select data for CNV analysis

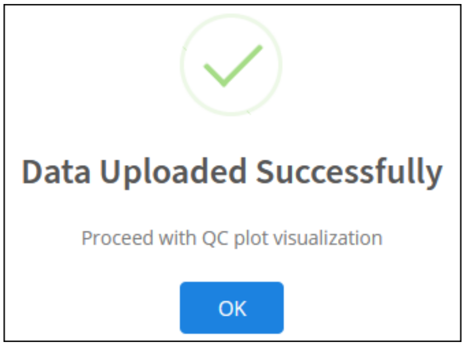
Browse... No file selected

Figure 120. *Upload data for analysis* page in the scDNA application.

This module enables the uploading and preparation of single-cell DNA sequencing CNV data for downstream analysis.

1. Choose the appropriate input data type using the radio buttons. Currently, CogentDS accepts either processed data (an RDS file) from CogentAP CNV analysis or you can test the workflow with a small example dataset integrated into CogentDS.

   For the 'Processed Data from CogentAP' option, click [Browse…] to locate the appropriate input file. After selecting the file and clicking [Open], the file will be selected for upload.
2. Once the option is selected and the data file is selected, if necessary, click [Upload Data]. A pop-up window will confirm completion.



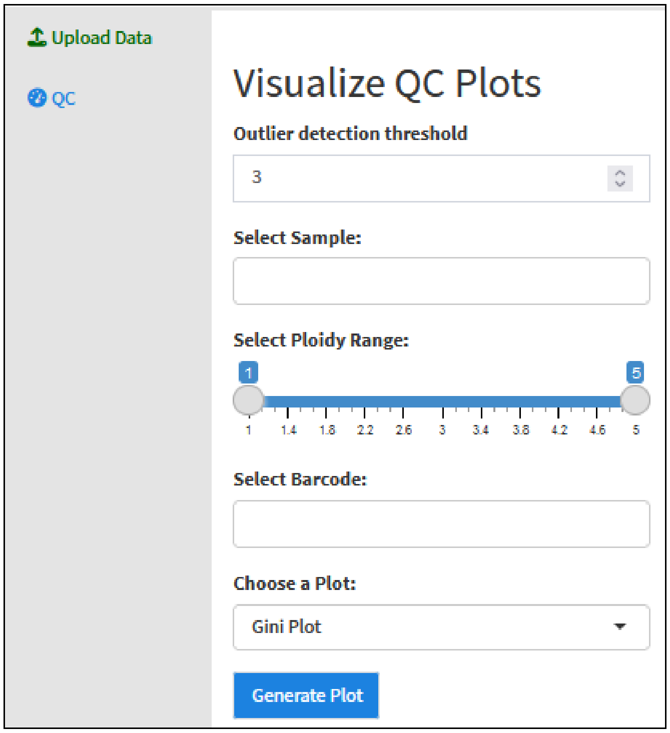
Data Uploaded Successfully

Proceed with QC plot visualization

Figure 121. Successful data upload in the scDNA application.

Click the [OK] button, then [Next: Visualize QC plots] to proceed to the next module.

2. Visualize QC Plots



Visualize QC Plots

Outlier detection threshold

3

Select Sample:

Select Ploidy Range:

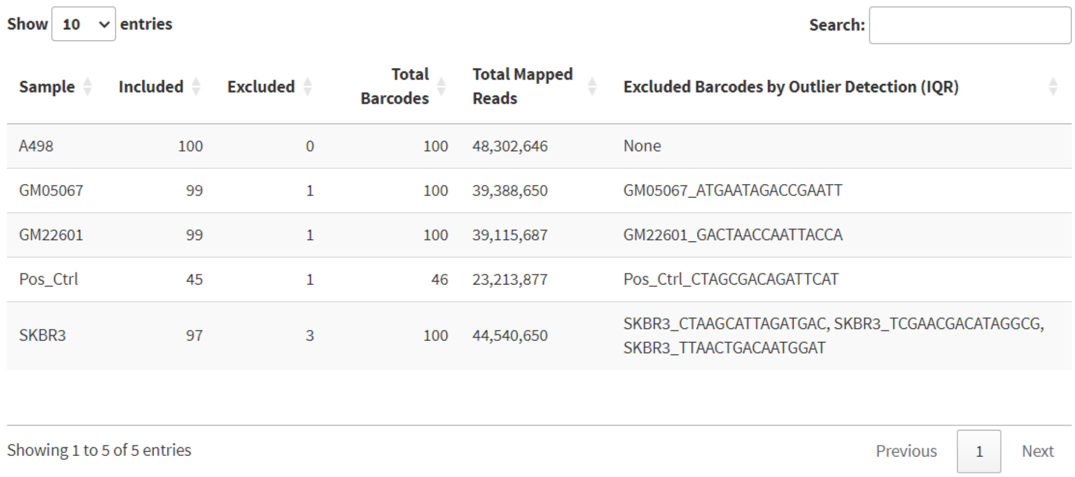
Select Barcode:

Choose a Plot:

Gini Plot

Figure 122. The *Visualize QC Plots* page in the scDNA CNV analysis application.

This module provides users with multiple filters and visualization options for quality control (QC) plots to access data quality. Outlier detection is based on Gini scores according to a user input threshold. A table summarizing the number of barcodes retained and excluded per sample after outlier detection is also provided in this module, along with total mapped reads and a list of excluded barcodes (Figure 123).



Sample	Included	Excluded	Total Barcodes	Total Mapped Reads	Excluded Barcodes by Outlier Detection (IQR)
A498	100	0	100	48,302,646	None
GM05067	99	1	100	39,388,650	GM05067_ATGAATAGACCGAATT
GM22601	99	1	100	39,115,687	GM22601_GACTAACCAATTACCA
Pos_Ctrl	45	1	46	23,213,877	Pos_Ctrl_CTAGCGACAGATTCAT
SKBR3	97	3	100	44,540,650	SKBR3_CTAAGCATTAGATGAC, SKBR3_TCGAACGACATAGGCG, SKBR3_TTAACTGACAATGGAT

Showing 1 to 5 of 5 entries

Figure 123. Summary table showing the number of retained and excluded barcodes per sample based on Gini score outlier detection. The table also includes the total mapped reads and a list of excluded barcodes.

The types of QC plots available in this module include Lorenz plot, Gini plot, Loess plot, and MAD plot. See the detailed description for each type of plot below, and select the type of chart you would like to generate and values for any additional option related to the chart type, then click [Generate Plot] to visualize the data. The resulting plots can be downloaded to your local computer using the [Save plot] button, if desired, then click [Next: Visualize CCN Heatmap] to proceed to the next module.

a) *Filter Options*

- Outlier detection threshold—this threshold value is used to identify outlier cells which are then removed from downstream analysis. The default value is '3', which is three times the difference between the first and the third quartile (interquartile range) of the Gini scores. A smaller value will correspond to a more stringent threshold to exclude more cells with a Gini score different from the median value.
- Select sample—(Optional) select a specific sample from the drop-down menu to analyze. The values of the menu are dependent on the identified samples in your input data, but multiple samples can be selected for this field. Samples listed in the selection box can be removed by clicking on the sample name and hitting the **[Backspace]** key on your keyboard.

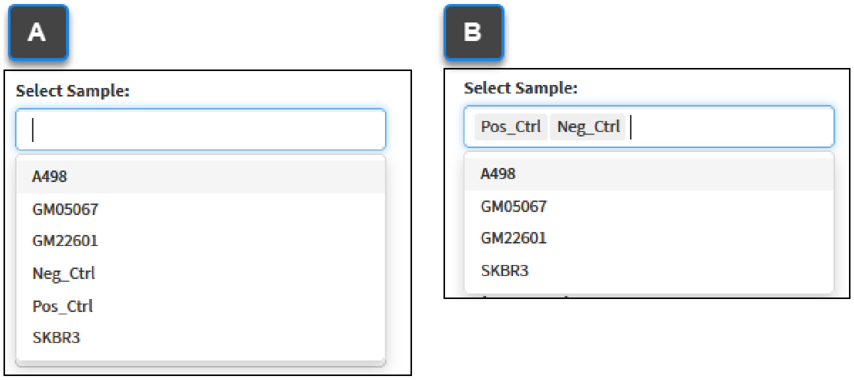


Figure 124. Selecting multiple samples on the scDNA>*Visualize QC plots* page. (**Panel A**) The initial selection drop-down menu. (**Panel B**) The selection drop-down menu after addition of two samples.

- Ploidy—(Optional) sets the minimum and maximum values for the amount of DNA levels, to focus on a desired range. Selection is done by clicking with your mouse on a gray circle (on the left and right in Figure 125, below) then, while continuing to click, dragging it left or right to the desired value. To aid with selecting the exact value, the number of the setting is displayed in the blue box above the circle.

  The circle on the left denotes the minimum value while the circle on the right end of the blue line (the range) is for the maximum value.

  NOTE: The maximum or minimum value cannot be set to exceed its counterpart. I.e., in the example depicted in Figure 125, the maximum end of the slider cannot be moved lower than 2.8, the value of the minimum.

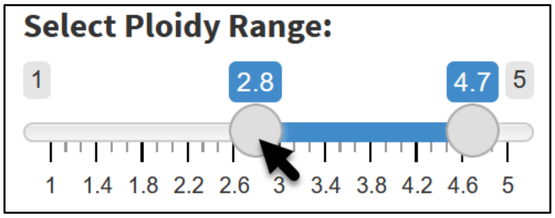


Figure 125. Setting the minimum and maximum ploidy values via the slider on the scDNA>*Visualize QC plots* page.

- Barcode—(Optional) exclude specific barcodes of interest from the dataset. Similar to the sample selection, multiple barcodes can be selected from the drop-down menu.

b) *Choose a Plot*

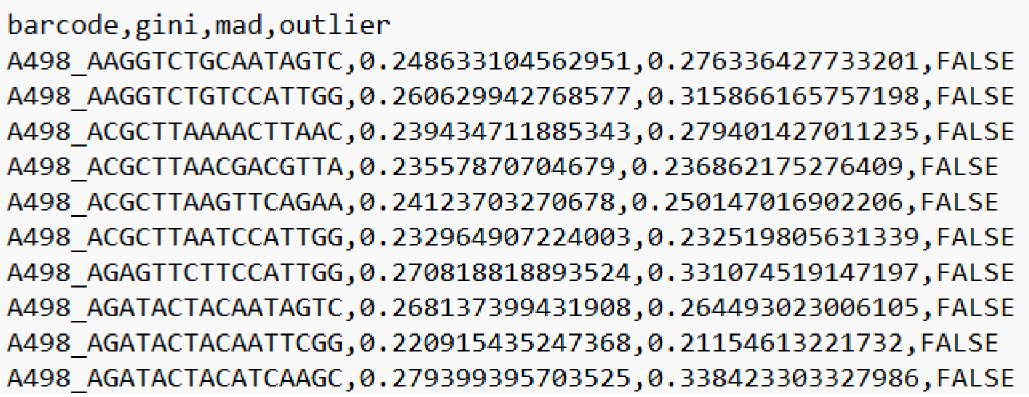
Select a plot type from the drop-down menu for how you would like to visualize the data.

- Lorenz Plot—assesses coverage uniformity across the dataset. A Lorenz curve is plotted by the cumulative fraction of the genome versus the cumulative fraction of total reads. The perfect coverage uniformity is represented by the diagonal dotted line. Coverage uniformity decreases as the curve gets further away from the diagonal dotted line.
- Gini Plot—the distribution of the Gini coefficient across samples, helping to identify samples with high- or low-coverage uniformity. The Gini coefficient is a value between 0 and 1 summarizing the degree of nonuniformity of the Lorenz curve. A

Gini coefficient of '0' indicates perfect uniformity (every site has equal coverage), whereas a Gini coefficient of '1' indicates perfect nonuniformity (one site has all the reads and all the other sites have no reads).

- Loess Plot—shows the LOESS fit of GC content against log-normalized bin counts across all barcodes, thereby visualizing GC-bias.
- MAD Plot (By Samples)—displays a sample-level boxplot of median absolute deviation (MAD) scores across bins, illustrating the bin-by-bin variability within each sample.
- MAD Plot—similar to a sample-based MAD plot, but without sample grouping (i.e., the entire dataset).

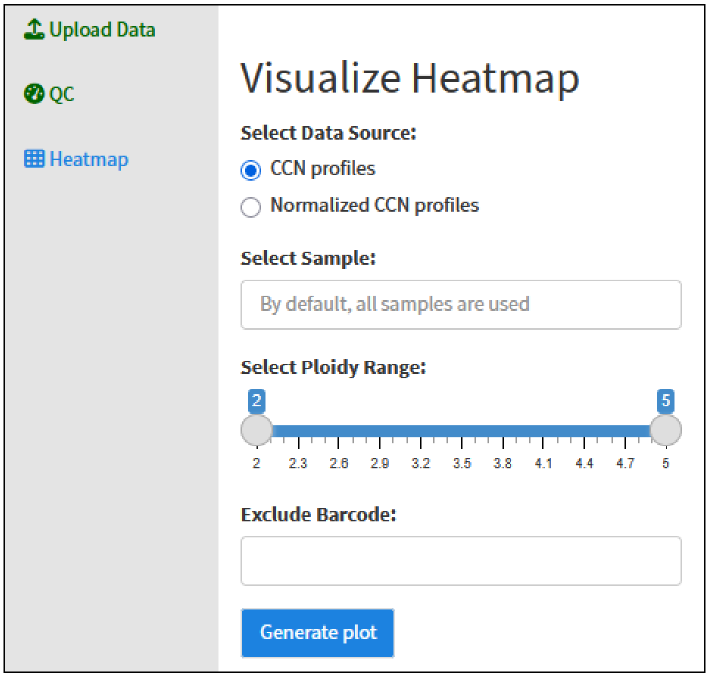
Once the parameters are set, click [Generate Plot]. The metadata becomes available for download using the [Download Gini & MAD Scores] button. The CSV file contains the computed Gini scores and MAD scores for each barcode along with the outlier status.



```
barcode,gini,mad,outlier
A498_AAGGTCTGCAATAGTC,0.248633104562951,0.276336427733201,FALSE
A498_AAGGTCTGTCCATTGG,0.260629942768577,0.315866165757198,FALSE
A498_ACGCTTAAAACTTAAC,0.239434711885343,0.279401427011235,FALSE
A498_ACGCTTAACGACGTTA,0.23557870704679,0.236862175276409,FALSE
A498_ACGCTTAAGTTCAGAA,0.24123703270678,0.250147016902206,FALSE
A498_ACGCTTAATCCATTGG,0.232964907224003,0.232519805631339,FALSE
A498_AGAGTTCTTCCATTGG,0.270818818893524,0.331074519147197,FALSE
A498_AGATACTACAATAGTC,0.268137399431908,0.264493023006105,FALSE
A498_AGATACTACAATTCGG,0.220915435247368,0.21154613221732,FALSE
A498_AGATACTACATCAAGC,0.279399395703525,0.338423303327986,FALSE
```

Figure 126. Screenshot showing the metadata table with computed Gini scores, MAD scores, and outlier status available for download.

3. Visualize Heatmap (Heatmap of CCN Traces)



Visualize Heatmap

Select Data Source:

● CCN profiles

○ Normalized CCN profiles

Select Sample:

By default, all samples are used

Select Ploidy Range:

Exclude Barcode:

Figure 127. The *Visualize heatmap* page in the scDNA CNV analysis application.

This module generates a heatmap of chromosomal copy number (CCN) traces, allowing users to visualize CNV data patterns. It utilizes the data processed in the QC module and includes the same filtering options (Section 2) but does not perform outlier detection. For further details, refer to Section VIII.A.2.a, "Filter Options".

After selecting or setting any desired custom options, click [Generate plot] to synthesize the visualization, which will display within the box under "Original heatmap".

The heatmap can be resized by clicking and dragging the bottom right-hand corner of the box (indicated by the arrow in Figure 128) either vertically, horizontally, or diagonally.

NOTE: The menu items below the heatmap (brush, save file, and resize) should not be used.

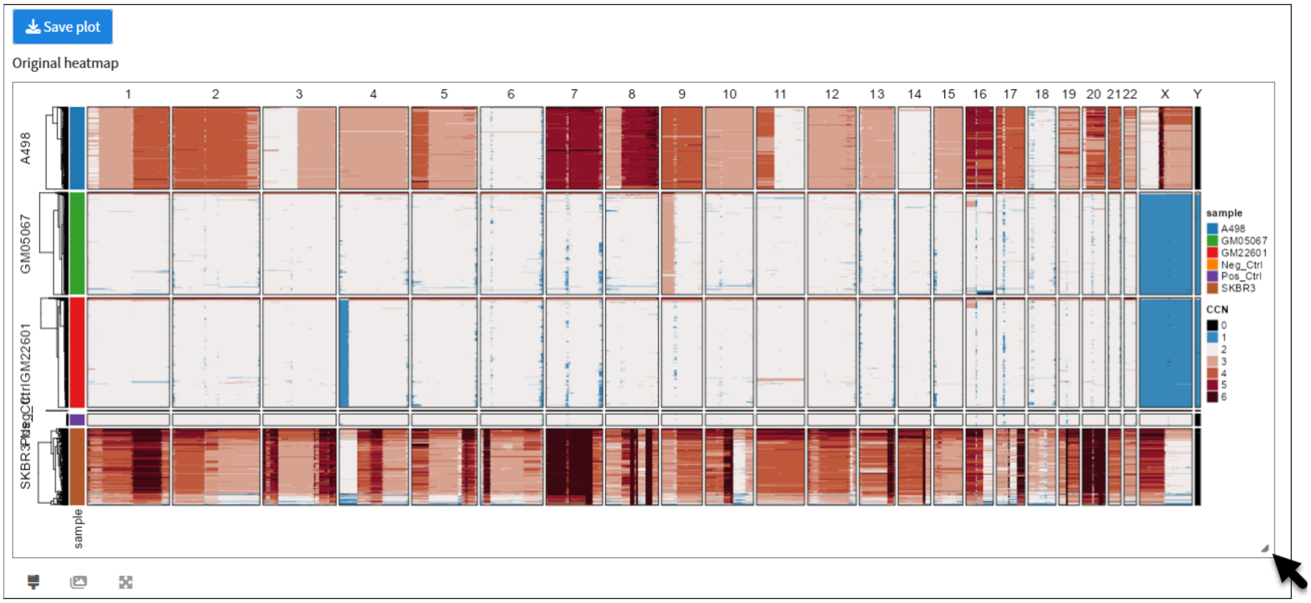


Figure 128. The CCN profiles heatmap generated in the scDNA application.

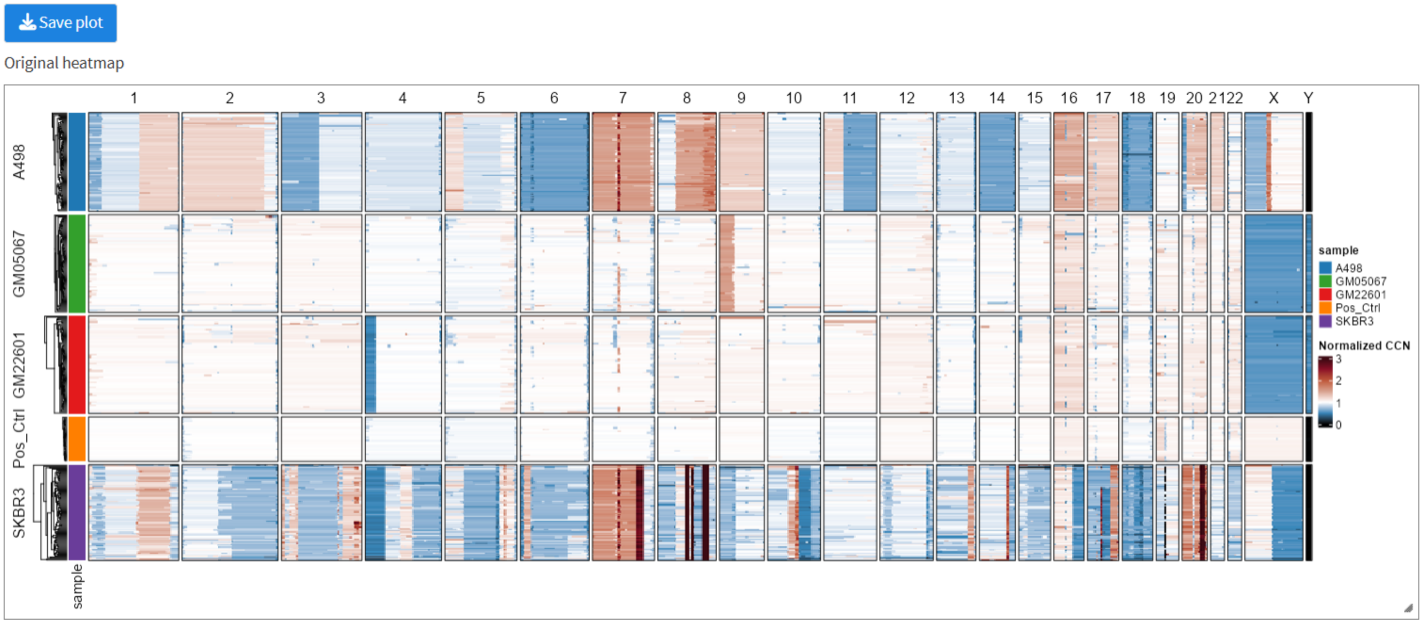


Figure 129. The Normalized CCN profiles heatmap generated in the scDNA application.

Selecting Data Source: choose between CCN profiles or Normalized CCN profiles for generating the heatmap:

- CCN profiles: Represents the integer copy number of profiles with ploidy estimation. Ginkgo infers ploidy via a numerical optimization based on sum-of-squares (SoS) error to find the copy-number multiplier that causes the normalized segmented bin counts to best align with integer copy-number values.
- Normalized CCN profiles: Represents the normalized copy number ratios, which accounts for GC bias and mappability, but without ploidy estimation.

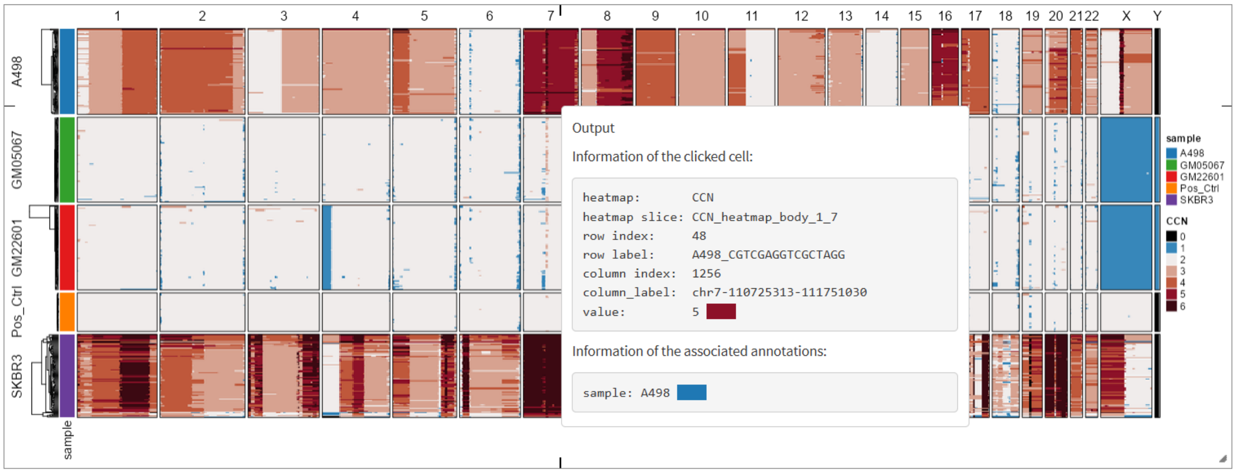
NOTE: If [Normalized CCN profiles] is selected, ploidy filtering will not be available. Normalized data doesn't have ploidy information.

The heatmap provides an overview of CNV patterns of all the chromosomes, enabling identification of copy number changes across samples. Each horizontal line represents the CCN profile of a single cell, and the cells clustered within each sample based on the Euclidean distance of their CCN profile patterns.

The heatmap plots shows color-coded integer CCN levels from 0 to 6.

- CCN of 0 is color-coded in black; e.g., in Figure 130, chromosome Y of the A498 cells.
- CCN of 1 is color-coded in moderate blue; e.g., in Figure 130, the segmental loss at chromosome 4 of the GM22601 cells.
- CCN of 2 is color-coded in light gray; e.g., in Figure 130, all the autosomes of the control genomic DNA.
- CCN of 3 is color-coded in orange; e.g., in Figure 130, the segmental gain at chromosome 9 of the GM05067 cells.
- CCN of 4–6 is color-coded from moderate to dark red.

Clicking on the heatmap provides additional information about selected data points, including sample details, chromosome number, location, and ploidy.



Output

Information of the clicked cell:

```
heatmap:        CCN
heatmap slice:  CCN_heatmap_body_1_7
row index:      48
row label:      A498_CGTCGAGGTCGCTAGG
column index:   1256
column_label:   chr7-110725313-111751030
value:          5
```

Information of the associated annotations:

```
sample: A498
```

Figure 130. The heatmap with additional information generated in the scDNA application.

The resulting heatmap can be downloaded as an image to your local computer using the [Save plot] button, if desired. Click [Next: Visualize CN profile plot] to proceed to the next module.

4. CN Profile

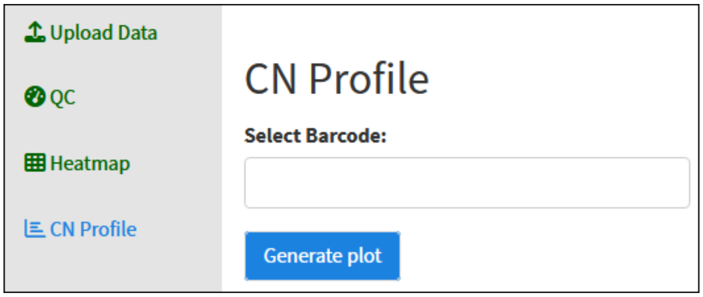


Figure 131. *CN profile* page in the scDNA application.

This module enables generation of the plot for copy number (CN) profiles for individual barcodes using the "Select Barcode:" option. Choose at minimum one and up to three barcodes from the drop-down menu, and then click the [Generate plot] button to visualize CN profiles for each barcode selected.

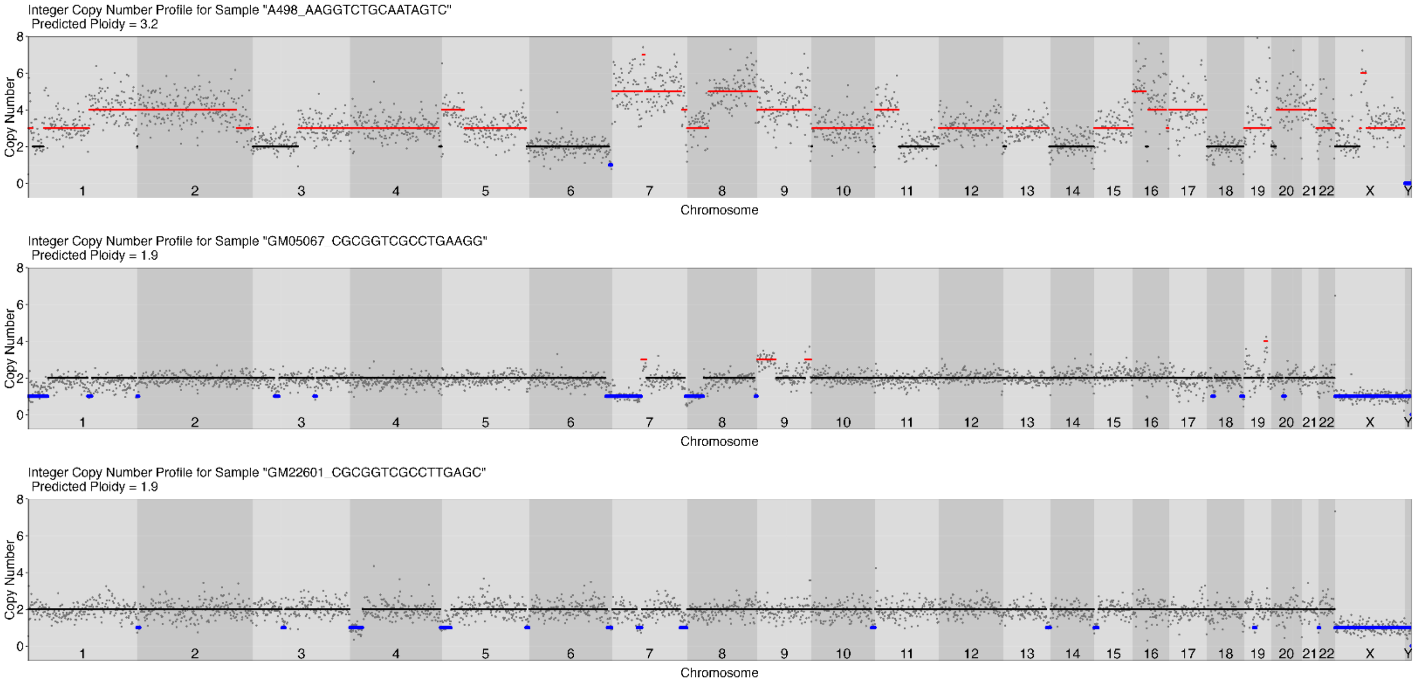
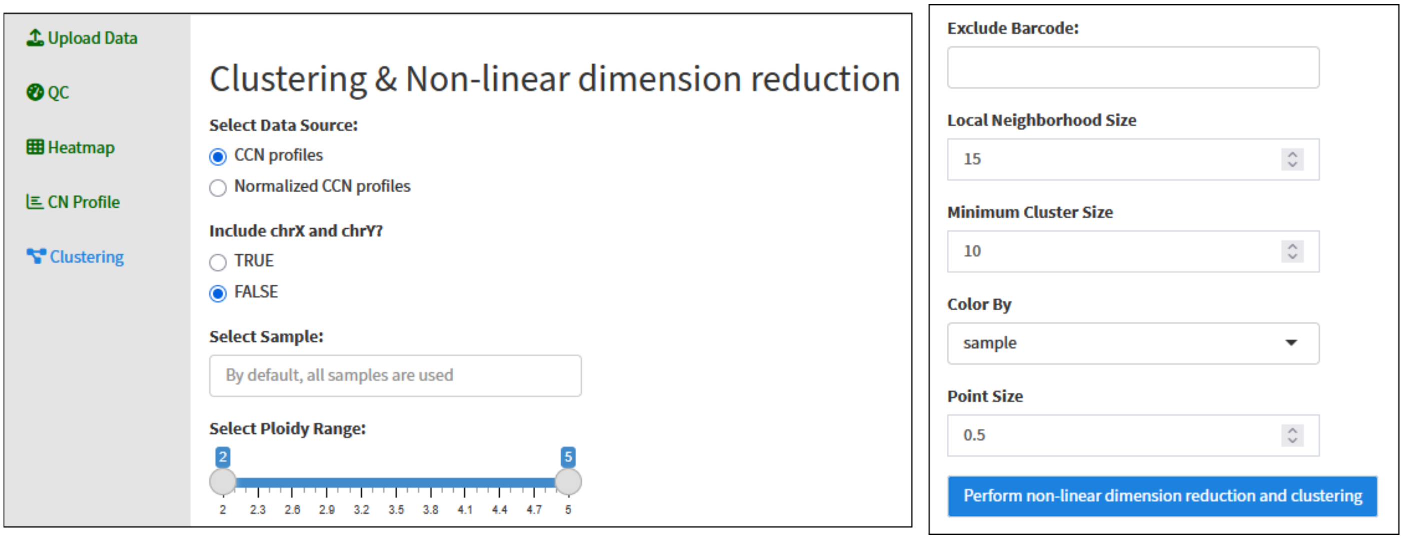


Figure 132. Example plots on the *CN profile* page in the scDNA application.

The resulting plots can be downloaded as a single image to your local computer using the [Save plot] button, if desired. Click [Next: Perform Clustering] to proceed to the next module.

5. Clustering & Non-Linear Dimension Reduction



Clustering & Non-linear dimension reduction

Select Data Source:

● CCN profiles
○ Normalized CCN profiles

Include chrX and chrY?

○ TRUE
● FALSE

Select Sample:

By default, all samples are used

Select Ploidy Range: 2 – 5

Exclude Barcode:

Local Neighborhood Size: 15

Minimum Cluster Size: 10

Color By: sample

Point Size: 0.5

Figure 133. The *Clustering* page in the scDNA application. The page is shown split horizontally for space purposes in this manual but appears vertically on the web page.

This module performs nonlinear dimensionality reduction using UMAP and applies density-based clustering (DBSCAN) on the UMAP coordinates. It utilizes the data processed in the QC module and includes the same filtering options available (Section 2) but doesn't perform outlier detection. For further details, refer to Section VIII.A.2.a, "Filter Options".

The data source selection for UMAP follows the same logic as Section 3, the "Visualize Heatmap (Heatmap of CCN Traces)" module.

- Choose between CCN profiles and Normalized CCN profiles.

> **NOTE:** If [Normalized CCN profiles] is selected, ploidy filtering will not be available. Normalized data doesn't have ploidy information.

Additional options for this module:

- Include chrX and chrY—(Default: FALSE) This indicates whether the X and Y sex chromosomes are included or excluded from the data. Excluding them helps to reduce sex-related bias, while marking TRUE will include the sex chromosomes in the resulting data output.
- Local neighborhood size—sets the number of neighboring points used in local approximations. The default value is '15'. (n_neighbors)
- Minimum cluster size—specifies the minimum number of cells required to form a cluster. The default value is '10'. (minPts)
- Color by—select either 'sample' or 'cluster' by which the plot points will be colored.
- Point size—determines how big or small the points appear on the plot. A larger number makes the point size bigger and a smaller number makes points smaller. The default value is '0.5'.

After setting the desired parameters, click [Perform non-linear dimension reduction and clustering] to run the analysis and generate the UMAP plot. Hovering the mouse cursor over the plot will cause a pop-up window to display that shows the UMAP coordinates, sample name, and barcode for the point in the cursor location.

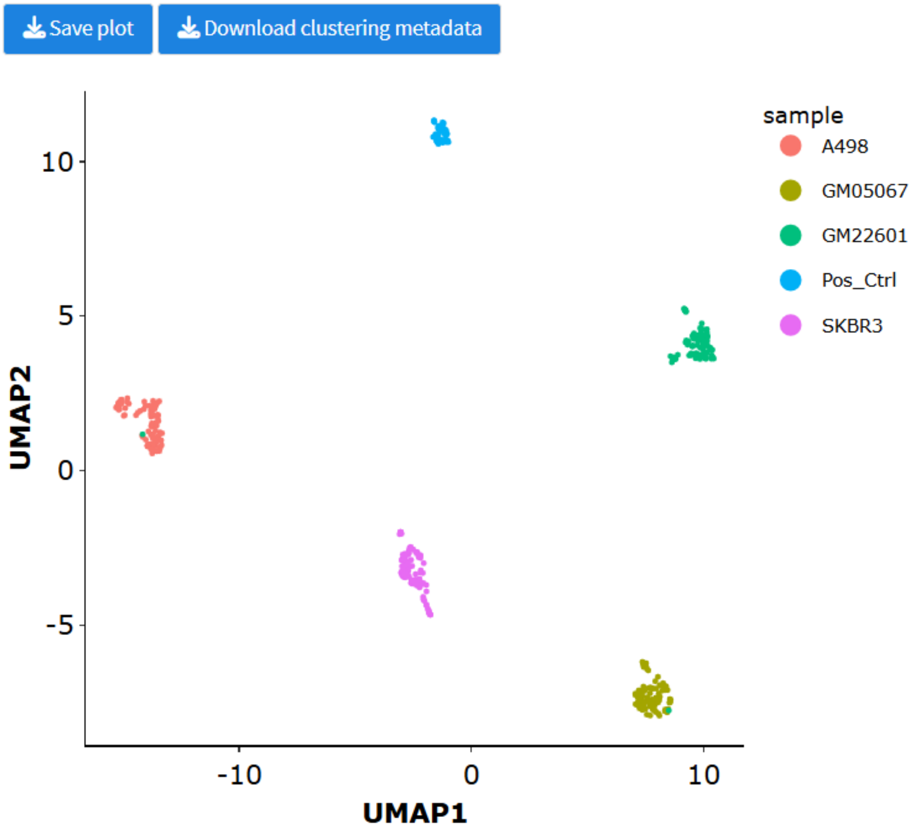


Figure 134. The UMAP generated in the clustering module of the scDNA-seq CNV application. Downloadable clustering metadata is available for CogentAP and CogentDS SNV analysis.

The resulting plot can be downloaded to your local computer by clicking [Save plot] or cluster information downloaded as a CSV file with the [Download clustering metadata] button.

> **NOTE:** Clustering metadata is required to perform SNV analysis in both CogentAP and CogentDS after CNV analysis

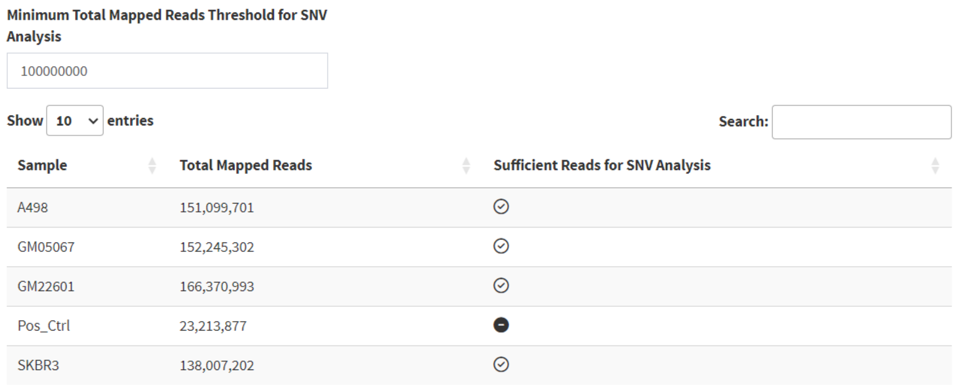
Once the plot is generated, an additional filter option will appear: Minimum Total Mapped Reads Threshold for SNV Analysis.

This allows users to specify the minimum number of mapped reads required for a sample or cluster to be considered suitable for SNV analysis. By default, this threshold is set to 100 million reads.

A corresponding filter table will also be displayed. This table provides:

- A list of samples or clusters (depending on the “Color by” option selected)
- Their respective total mapped reads, and

- An indication of whether they meet the specified threshold for inclusion in the SNV analysis within CogentAP



Minimum Total Mapped Reads Threshold for SNV Analysis

100000000

Show 10 entries

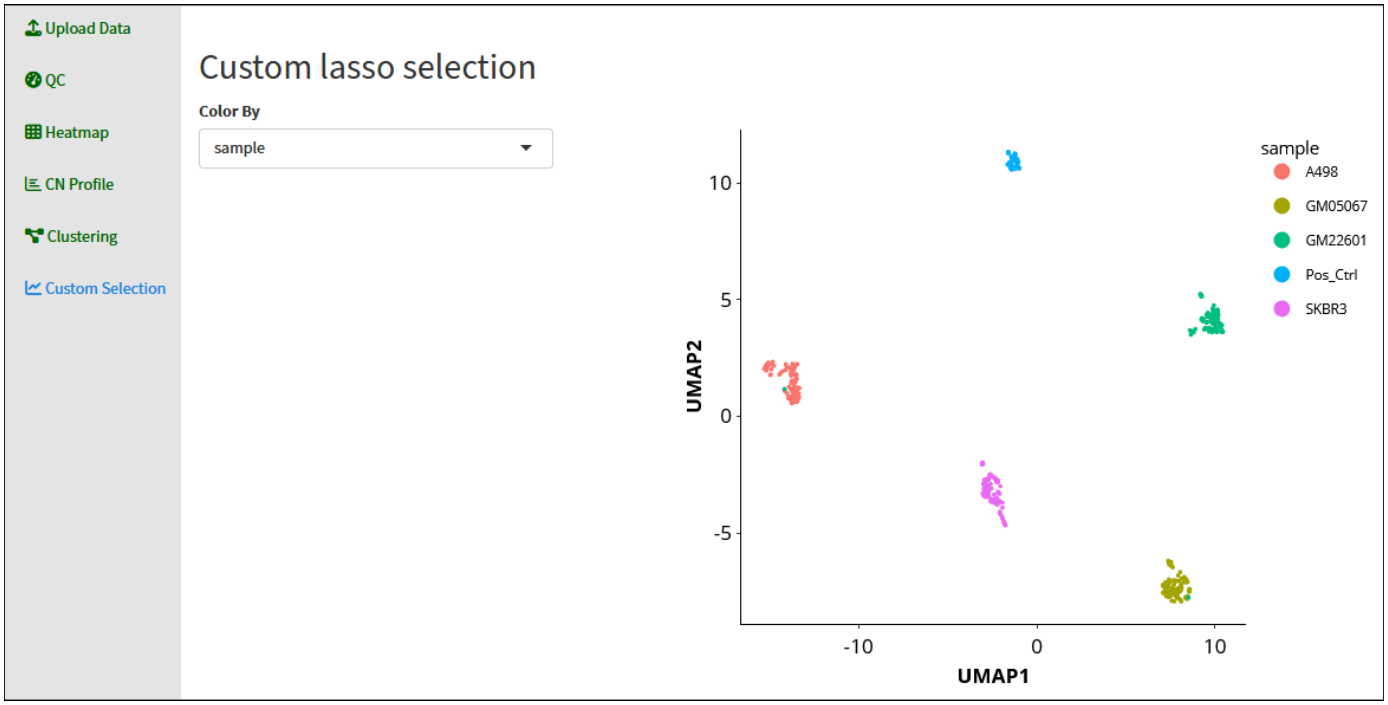
Search:

Sample	Total Mapped Reads	Sufficient Reads for SNV Analysis
A498	151,099,701	✓
GM05067	152,245,302	✓
GM22601	166,370,993	✓
Pos_Ctrl	23,213,877	⊖
SKBR3	138,007,202	✓

Figure 135. Screenshot showing the filtering option that appears after UMAP clustering: Minimum Total Mapped Reads Threshold for SNV Analysis.

Once clustering analysis is performed, click [Next: Perform Custom Lasso Selection] to proceed to the next module.

6. Custom Lasso Selection



Custom lasso selection

Color By: sample

Figure 136. The *Custom lasso selection* page in the scDNA application.

This module enables interactive selection of a group of cells based on samples or clusters. The selections can then be named and saved for further analysis.

Additional options:

- Color by—select from the drop-down menu to overlay either by 'sample' or 'cluster' onto the UMAP. By default, the overlay type is 'sample'.

Lasso selection here is similar to lasso selection in the scRNA analysis module. For more details, refer to Section VI.A.9.b, "Performing Lasso Selection".

Once any customized lasso is defined, a [Download metadata] button displays under the Color by dropdown box.

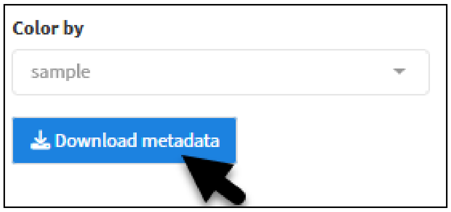


Figure 137. The [Download metadata] button on the *Custom lasso selection* page in the scDNA application. The screenshot also shows the "Color by" drop-down menu text as grayed out (inactive).

Click the button to download a CSV file that includes barcodes, UMAP coordinates, original samples, and clusters, along with either custom CNV samples or custom CNV clusters, depending on previous selection and additional information such as Gini scores and ploidy.

Once saved, click [Generate HTML Report] to proceed to the next module.

7. CogentDS Analysis Report

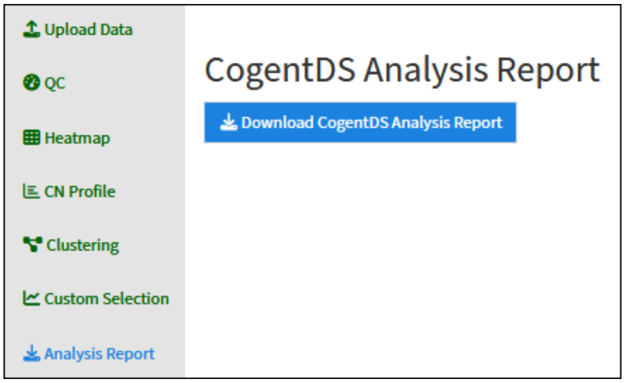


Figure 138. The *CogentDS Analysis Report* page in the scDNA application.

Click the [Download CogentDS Analysis Report] to save an HTML file containing reports based on the customizations applied during the course of the analysis workflow. This report will include a summary table showing the number of barcodes included and excluded for each sample, based on the outlier detection threshold. The table summarizes how many barcodes passed the filter and how many were flagged as outliers and lists the barcodes excluded for each sample. Alongside this table, the report will show QC plots, the CCN Profile heatmap, and the normalized CCN profile heatmap, as well as the UMAP plots for both CCN and normalized CCN profile plots.

After saving, if desired, click [Next: Download Processed Data] to proceed to the final module.

8. Download CogentDS .rds Data

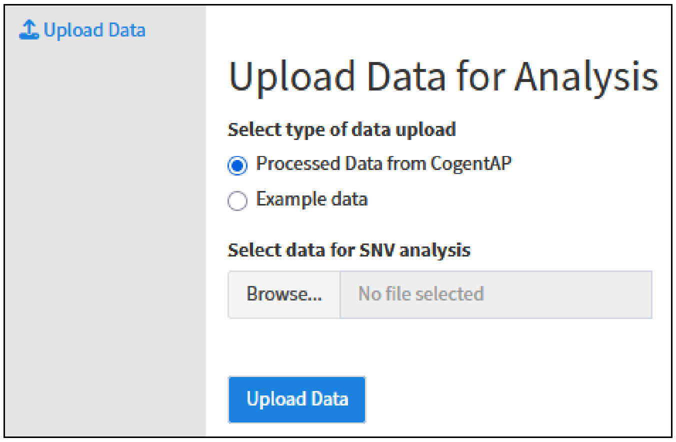
The last page of the scDNA application is identical to the final module of the scRNA app>Analysis Mode workflow. For more information on the data download page, refer to Section VI.A.17, "Download CogentDS Processed .rds Data".

Once the data file is saved (if desired), do one of the following:

- Click the [Go to Main Page] button in the lower right corner
- Click the [Home] icon in the top right corner, in the title bar, which will also return to the scDNA Analysis page, or
- Close the browser tab or window. If the tab is closed, the main CogentDS window with the three main CogentDS applications (Section V.B) will still be open and available for use.

B. SNV Analysis Mode

1. Upload Data for Analysis



Upload Data

Upload Data for Analysis

Select type of data upload

● Processed Data from CogentAP

○ Example data

Select data for SNV analysis

Figure 139. *Upload data for analysis* page in the scDNA SNV analysis application.

This module enables the uploading and preparation of single-cell DNA sequencing SNV data for downstream analysis.

1. Choose the input data type using the radio buttons. CogentDS accepts either processed data from CogentAP (i.e., CogentDS_SNV_analysis.rds from the snv_report/ folder), or you can test the workflow with a small example dataset integrated into CogentDS.
2. For the 'Processed Data from CogentAP' option, click [Browse…] to locate the appropriate input file. After selecting the file and clicking [Open], the file will automatically be uploaded to the tool.
3. For either option, click [Upload Data]. After some initial processing, a pop-up window will confirm completion.

Click the [OK] button, then [Next: Visualize variants barplot] to proceed to the next module.

2. Visualize Variants Barplot

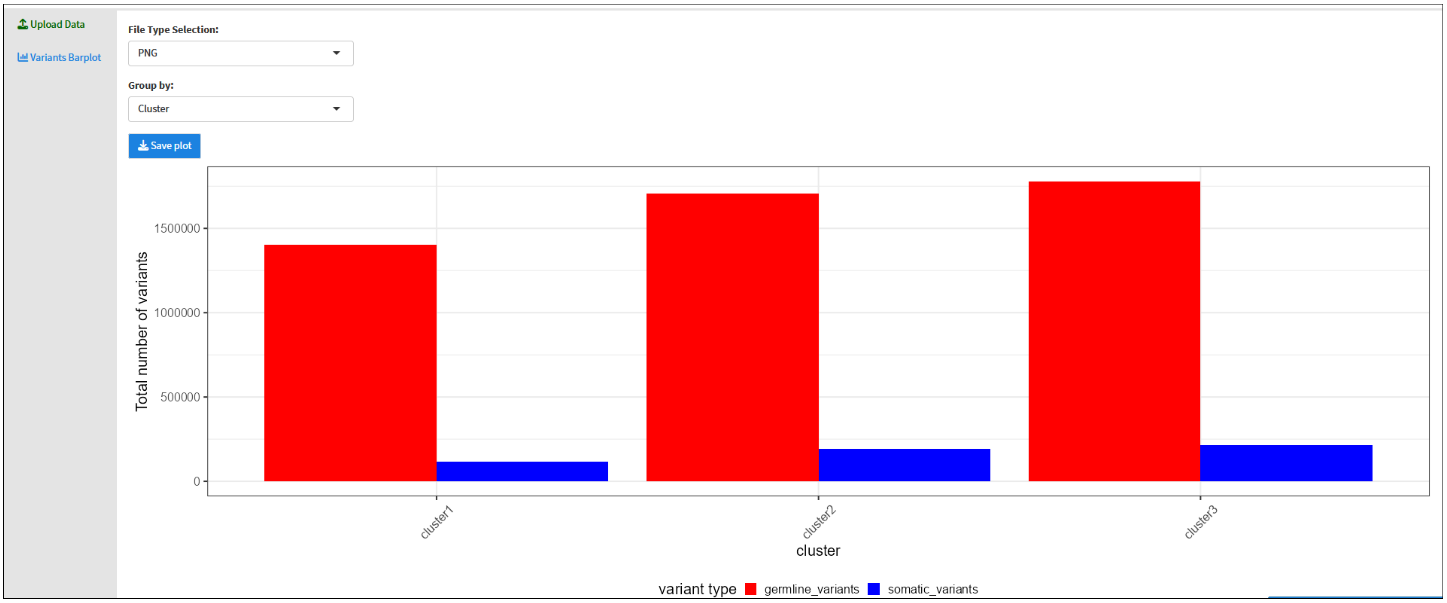


Figure 140. The *Visualize QC Plots* page in the scDNA SNV analysis application.

This module allows users to visualize the distribution of somatic and germline variants across clusters and chromosomes. You can choose to display variant counts grouped by those parameters within each cluster. A specific chromosome can also be selected for focused analysis. Barplots can be downloaded in various formats.

a) Choosing How to Display the Data

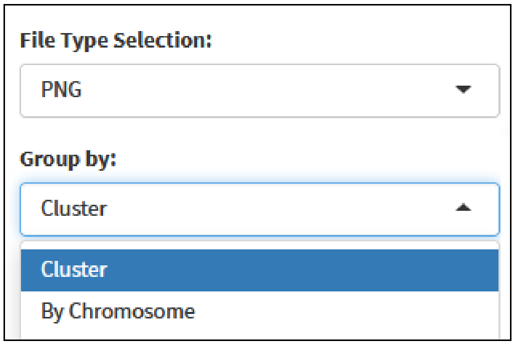
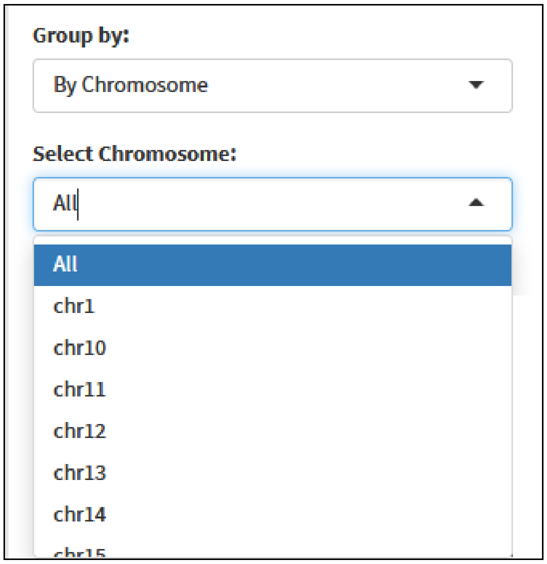


Figure 141. The "Group by" dropdown in the *Variants Barplot* page of the scDNA SNV analysis application.

Use the “Group by” dropdown to select how the variants should be displayed:

- Cluster (Default option): Display somatic and germline variant counts across different clusters.
- By Chromosome: Organizes the data by chromosome, displaying the somatic and germline variant counts across clusters.

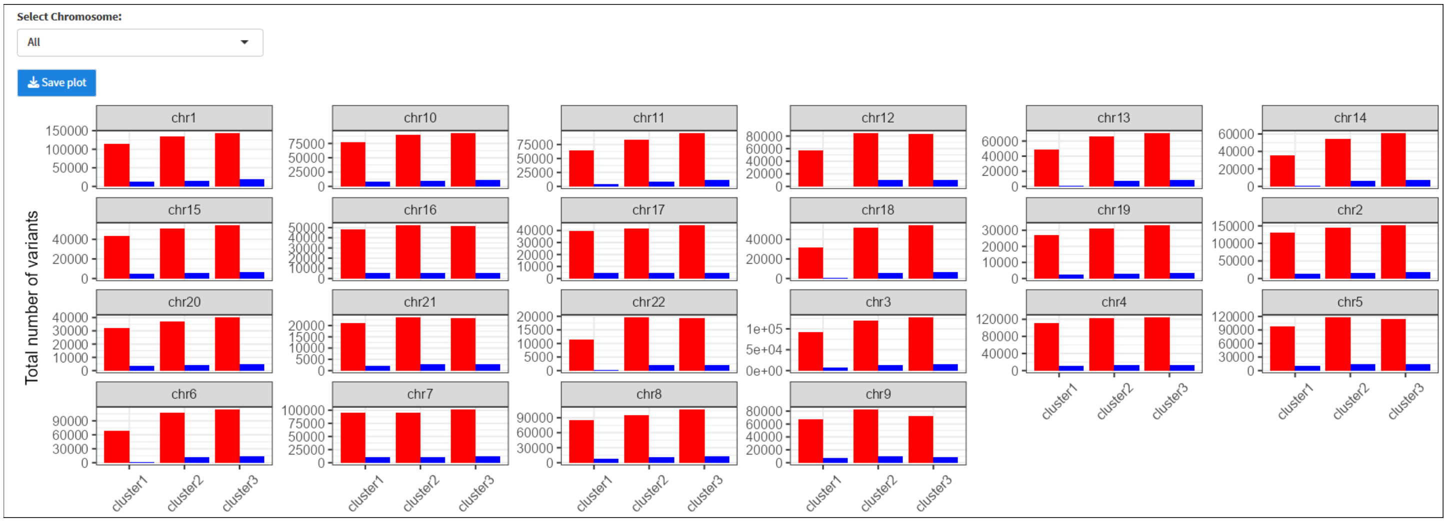


Group by: By Chromosome

Select Chromosome: All

Figure 142. The "Select Chromosome" drop-down menu in the *Variants Barplot* page of the scDNA SNV analysis application.

If the 'By Chromosome' option is selected, a dropdown labeled "Select Chromosome" appears to filter the plot. Keep 'All' selected to display the variant counts for all chromosomes across the clusters or choose a specific chromosome to focus only on its mutation counts across clusters. By default, 'All' chromosomes are shown in the barplot.

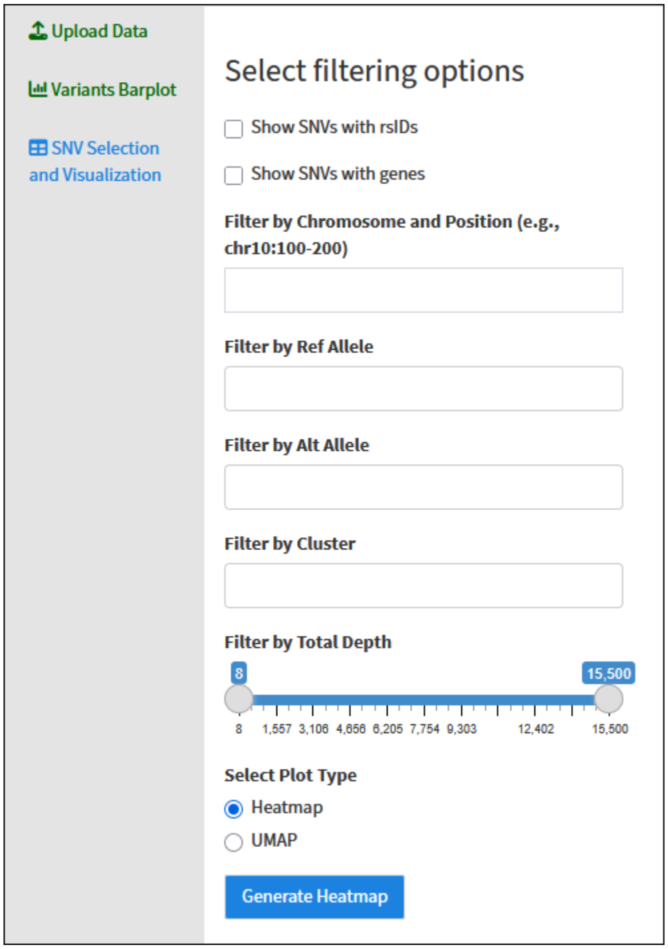


Select Chromosome: All

Figure 143. Example of the chromosome barplots on the *Variants Barplot* page of the scDNA SNV analysis application.

The resulting bar plots can be downloaded to your local computer using the [Save plot] button, if desired. Click [Next: SNV selection and visualization] to proceed to the next module.

3. SNV Selection and Visualization



Select filtering options

□ Show SNVs with rsIDs

□ Show SNVs with genes

Filter by Chromosome and Position (e.g., chr10:100-200)

Filter by Ref Allele

Filter by Alt Allele

Filter by Cluster

Filter by Total Depth: 8 – 15,500

Select Plot Type

● Heatmap

○ UMAP

Figure 144. The *SNV Selection and Visualization* page in the scDNA SNV analysis application.

This module allows users to filter and explore SNV data based on multiple criteria. Refine the displayed SNVs using filters such as rsIDs, gene annotations, chromosome position, reference/alternative allele, clusters, and sequencing depth. Additionally, you can generate heatmap and UMAP overlay plots for selected SNVs and download both filtered table and plots. Annotations are based on dbSNP Build 155 and the Grch38 reference genome.

a) *Filtering SNV Data*

The SNV Selection and Visualization module provides multiple options to refine the SNV data displayed in the table:

- Show SNVs with rsIDs: Displays only variants that have an associated dbSNP rsID.
- Show SNVs with Genes: Displays only variants mapped to known genes.
- Filter by Chromosome and Position:
    - Enter a UCSC style genomic range (ch10:100–200) to filter SNVs within that region.
    - If the input format is incorrect or range is invalid (e.g., ***start*** >***end*** value), an error message will display.

- Filter by Ref Allele/Filter by Alt Allele: Select one or more nucleotides (A, T, G, C) to filter the SNVs table based on the reference or alternative allele, respectively.

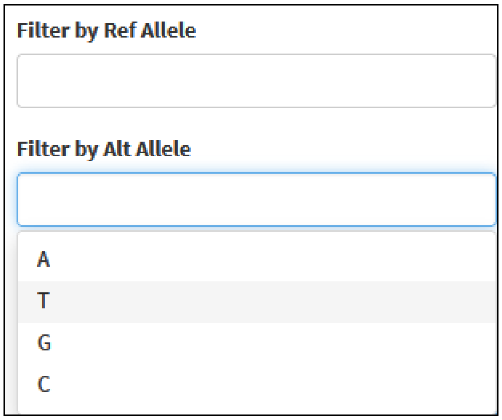


Figure 145. "Filter by Alt Allele" drop-down menu in the *SNV Selection and Visualization* page. The "Filter by Ref Allele" dropdown menu is identical.

- Filter by Cluster: Allows users to filter SNVs based on cluster derived from CNV analysis. These clustering definitions are generated in the clustering module of the CNV Analysis workflow. After clustering, users can download the resulting clustering metadata, which is then used as input when running SNV analysis in CogentAP. Upon completion of the analysis, the clustering metadata is embedded into the resulting .rds file. Choose one or more clusters to display SNVs within the selected clusters.

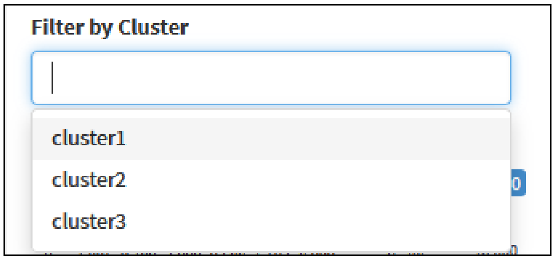


Figure 146. Example "Filter by Cluster" drop-down menu in the *SNV Selection and Visualization* page.

- Filter by Total Depth: Adjust the slider to filter the SNVs based on total sequencing depth. The range dynamically adjusts based on the data.

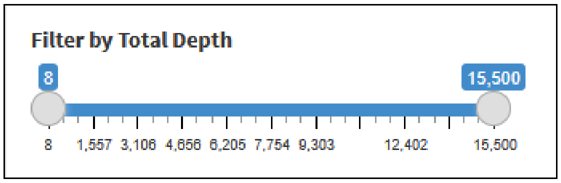
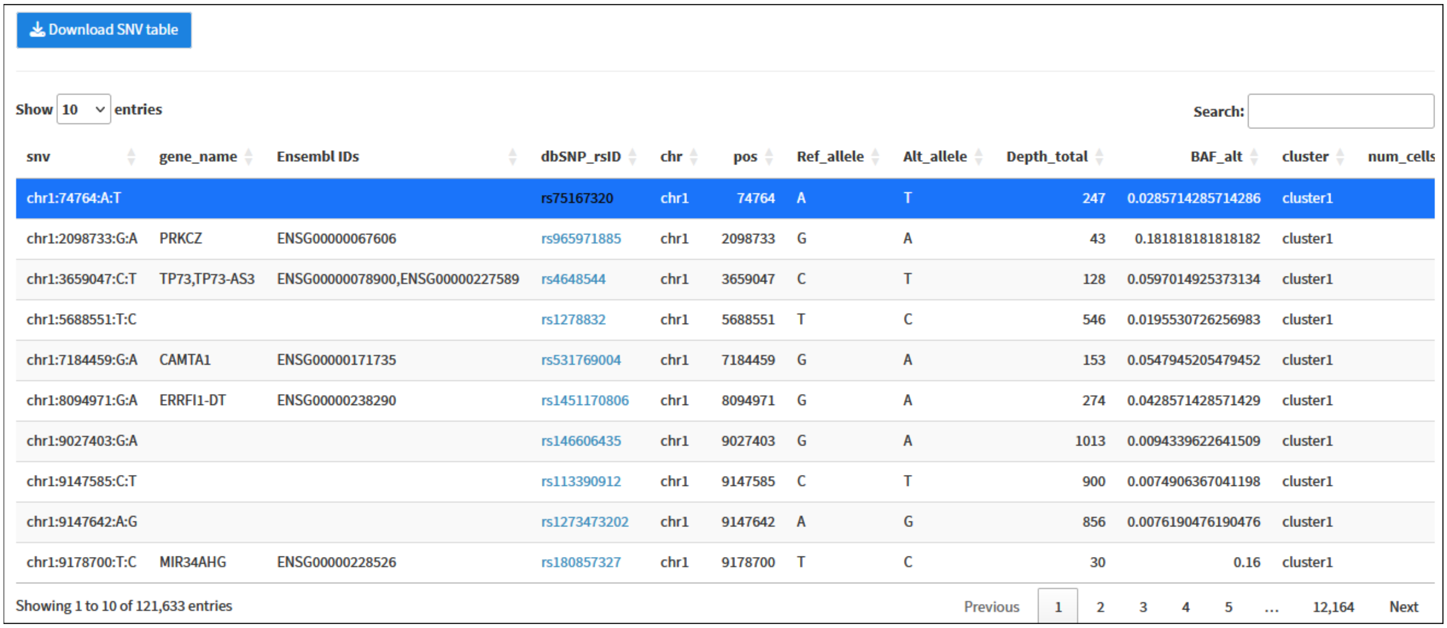


Figure 147. Example "Filter by Total Depth" slider in the *SNV Selection and Visualization* page.

The SNV summary table updates automatically as filters are applied, displaying only SNVs that match the criteria.

b) *Viewing and Downloading the Filtered SNV table*

Once the SNV data is filtered, the results are displayed in a table.



snv	gene_name	Ensembl IDs	dbSNP_rsID	chr	pos	Ref_allele	Alt_allele	Depth_total	BAF_alt	cluster	num_cells
chr1:74764:A:T			rs75167320	chr1	74764	A	T	247	0.0285714285714286	cluster1	
chr1:2098733:G:A	PRKCZ	ENSG00000067606	rs965971885	chr1	2098733	G	A	43	0.181818181818182	cluster1	
chr1:3659047:C:T	TP73,TP73-AS3	ENSG00000078900,ENSG00000227589	rs4648544	chr1	3659047	C	T	128	0.0597014925373134	cluster1	
chr1:5688551:T:C			rs1278832	chr1	5688551	T	C	546	0.0195530726256983	cluster1	
chr1:7184459:G:A	CAMTA1	ENSG00000171735	rs531769004	chr1	7184459	G	A	153	0.0547945205479452	cluster1	
chr1:8094971:G:A	ERRFI1-DT	ENSG00000238290	rs1451170806	chr1	8094971	G	A	274	0.0428571428571429	cluster1	
chr1:9027403:G:A			rs146606435	chr1	9027403	G	A	1013	0.0094339622641509	cluster1	
chr1:9147585:C:T			rs113390912	chr1	9147585	C	T	900	0.0074906367041198	cluster1	
chr1:9147642:A:G			rs1273473202	chr1	9147642	A	G	856	0.0076190476190476	cluster1	
chr1:9178700:T:C	MIR34AHG	ENSG00000228526	rs180857327	chr1	9178700	T	C	30	0.16	cluster1	

Showing 1 to 10 of 121,633 entries

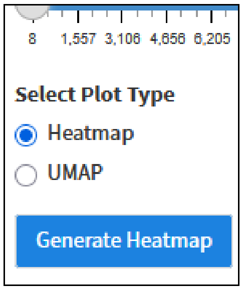
Figure 148. SNV table on the *SNV Selection and Visualization* page.

- Specific SNVs can be manually selected for visualization from the table
- The filtered SNV table can be downloaded as a TSV file by clicking the [Download SNV Table]

c) *Selecting and Generating Plots*

After filtering SNVs, individual SNV results (rows) can be visualized by selecting a plot type (heatmap or UMAP).

- Heatmap: The heatmap visualizes the occurrence of selected SNVs across clusters by displaying the percentage of cells within each cluster that have given SNV.
  - Select Heatmap as the plot type.



Select Plot Type

● Heatmap

○ UMAP

Figure 149. "Select Plot Type" options in the *SNV Selection and Visualization* page, with 'Heatmap' selected.

  - Select up to 10 SNVs from the table by clicking on the desired rows.
  - Click [Generate Heatmap] to create a plot.

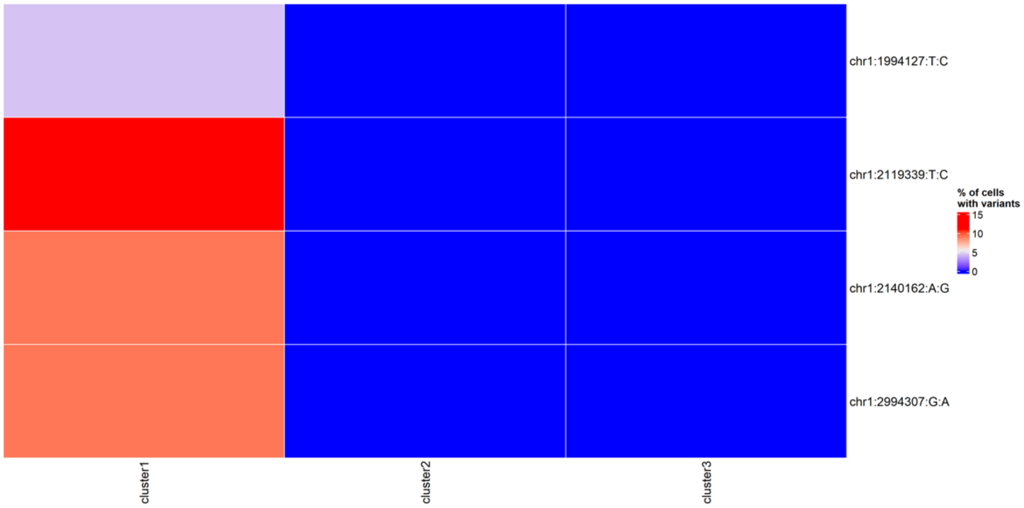
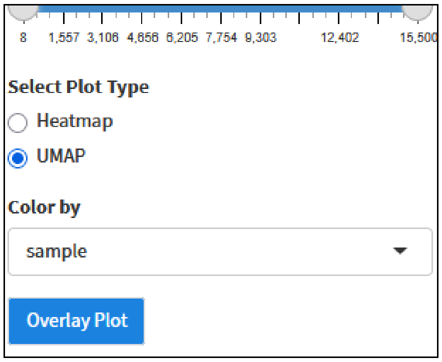


Figure 150. SNV heatmap on the *SNV Selection and Visualization* page.

- UMAP: The UMAP overlay plot allows users to overlay selected SNVs on a CNV based UMAP.
  - Select UMAP as the plot type.



Select Plot Type

○ Heatmap

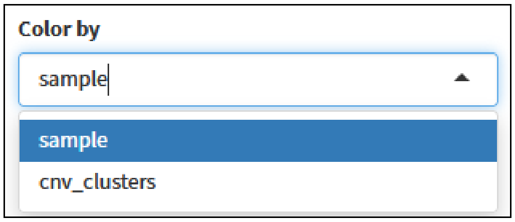
● UMAP

Color by

sample

Figure 151. "Select Plot Type" options in the *SNV Selection and Visualization* page, with 'UMAP' selected.

  - Select up to 10 SNVs from the table
  - Choose a "Color by" option: The UMAP can be color-coded by samples, cnv clusters or cnv custom clusters



Color by

sample

Figure 152. "Color by" dropdown in the *SNV Selection and Visualization* page, with 'sample' selected.

  - Click [Overlay Plot] to generate the UMAP with the distribution of selected SNVs overlayed on the UMAP. All the selected SNVs will be overlayed on the UMAP; click on the specific SNVs to show or hide them on the UMAP.

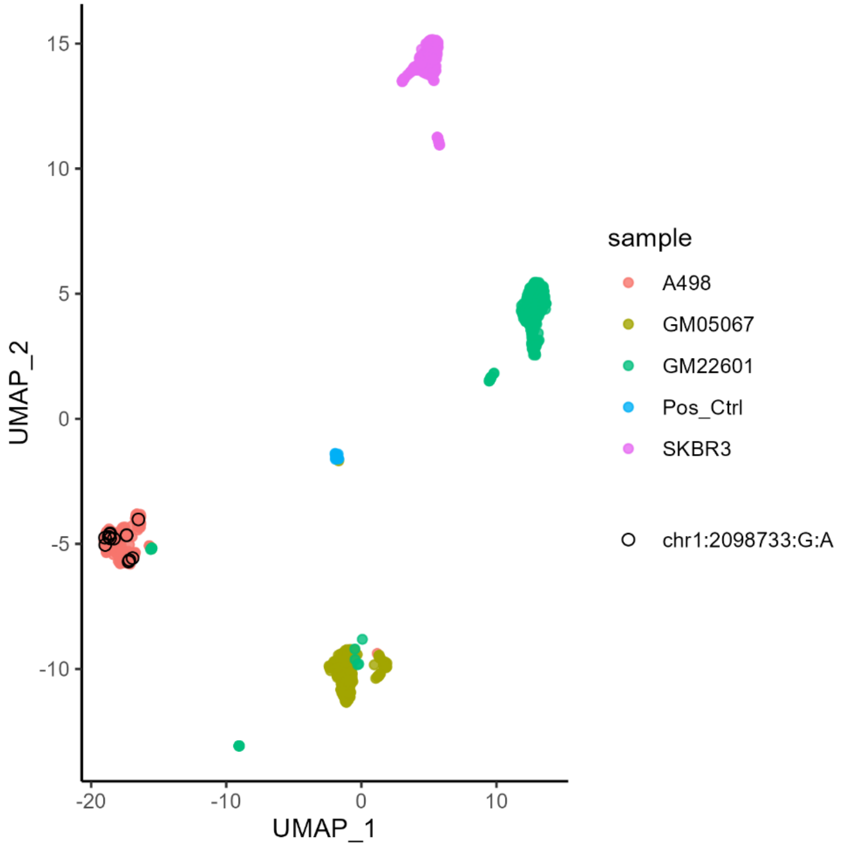


Figure 153. SNV UMAP overlaid with SNVs on the *SNV Selection and Visualization* page.

NOTE: In the CNV analysis mode, users can define custom groups using the Custom Lasso Selection module. This allows selection based on samples or clusters via a drop-down menu. Once all lasso selections are completed, a metadata file can be downloaded, which includes a "custom_cnv_clusters" column.

This metadata file serves as an input to CogentAP SNV analysis. CogentAP will use custom grouping for SNV analysis.

If custom samples or custom clusters were used as input for CogentAP SNV analysis, then the "custom cnv clusters" option will automatically appear in the drop-down list for UMAP coloring in CogentDS scDNA SNV analysis mode.

The resulting plot can be downloaded to your local computer by clicking [Save plot], if desired. Click [Next: Generate HTML Report] to proceed to the next module.

4. CogentDS Analysis Report

Click the [Download CogentDS Analysis Report] to save an HTML file that includes a barplot displaying the number of germline and somatic variants across the clusters. If the heatmap and UMAP module have been run, the report will include a heatmap, UMAP overlaid with selected SNVs, and a table of selected SNVs. The table provides detailed variant information, including rsID and gene annotations, chromosomal location, reference (ref) and alternative (Alt) alleles, total depth, number of cells with the variant, and the total number of cells.

Note on the default behavior for UMAP and Heatmap Modules

When generating the visualization, the SNV module handles missing parameters using the following defaults:

- UMAP defaults:
  - If no SNVs are explicitly selected for UMAP, it will default to the SNVs selected for the Heatmap
  - Coloring defaults to "sample" unless specified otherwise
- Heatmap defaults:
  - If no SNVs are explicitly selected for the heatmap, it will default to the SNVs selected for UMAP
- Fallback handling:
  - If neither UMAP nor Heatmap SNV selections are available, the module will notify the user and prompt them to run the corresponding module first
- SNV table content:
  - If both UMAP and heatmap SNVs are available, the resulting table will include merged SNVs from both
  - If only one is available, that set is used
  - If neither is available, a notification is displayed to inform the user

Appendix. UMAP Plot Floating Menu

In the top right corner of many charts, there is a menu of icons that only displays when hovered over with the mouse cursor.

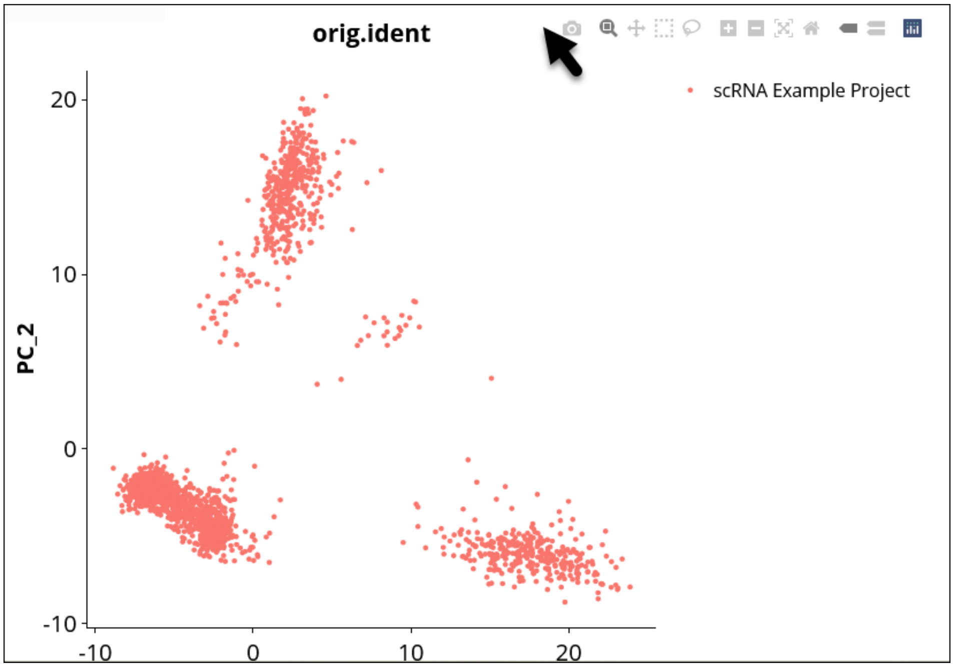


Figure 154. Location of the floating menu icons at the top right of a UMAP chart. The black arrow is demonstrating how to place the mouse cursor for the menu to become visible.

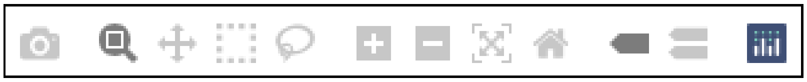


Figure 155. Chart modification floating menu (enlarged).

Table 7. Breakdown of the UMAP plots floating menu options.

Icon	Icon name	Description
📷	Download plot as png	Plotly method of downloading the UMAP plot as a PNG file. This is separate from the [Save plot] button of CogentDS, which allows for download in four file types (Section D).
🔍	Zoom	Do not use.
✥	Pan	The [Pan] function can be used to move the scatter plot within the frame of the chart axes (Section AI).
⬚	Box select	Used to group cells within a rectangular area.
◯	Lasso select	Used to group cells within a contiguous but non-polygonal area (i.e., freeform). This is only used in the scRNA Analysis Custom Cell Selection module (Section VI.A.9).
⊞ / ⊟	Zoom in/Zoom out	Can be used to zoom in for a more granular view of a smaller portion of the chart or zoom out for a less granular view. (Section B)
⤢	Autoscale	Auto-fits the plot within the boundaries of the window. Similar to the Reset Axes button, this option can be useful when you have zoomed, panned or modified the view and want to quickly reset the plot to show the full dataset again.
⌂	Reset axes	[Reset axes] will return the plot to the default view (Section A).
◀	Show closest data	Do not use.
☰	Compare data	If multiple points are close to same x-coordinate, enabling this function and hovering your mouse over a point displays the details for all the points near that x value.
📊	Plotly logo	Provides a link to the Plotly homepage for more information on the third-party module.

A. Pan and Reset Axes

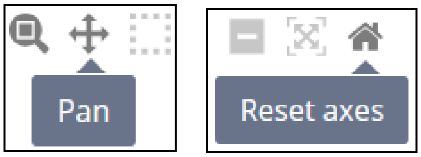


Figure 156. Identification of the Pan and Reset axes icons in the floating menu.

- The [Pan] function can be used to move the scatter plot within the frame of the chart axes, changing not only what plots are visible, but also the range values on the X and Y axes.

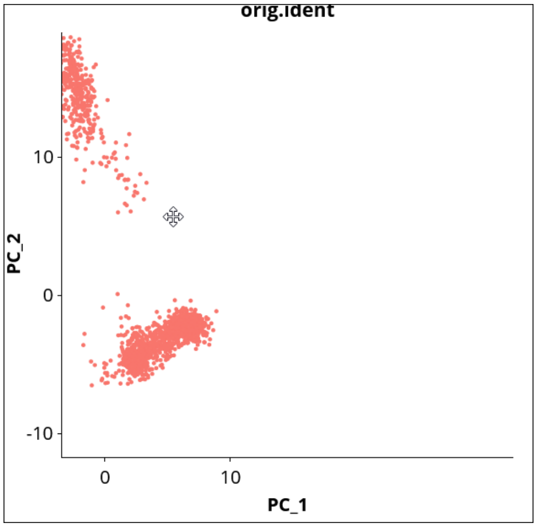


Figure 157. Example after using the Pan function to move the plots down the page. The values of the Y axis are decreased compared to the default in Figure 154.

- [Reset axes] will return the plot to the default view (Figure 154) after using the [Pan], [Zoom in], and/or [Zoom out] functions.

B. Zoom in and Zoom out

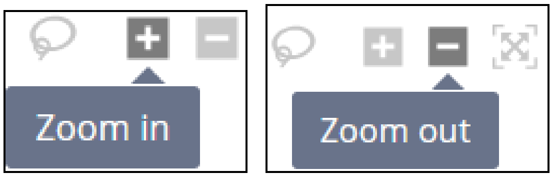


Figure 158. Identification of the Zoom in and Zoom out icons in the floating menu.

The [Zoom in] and [Zoom out] buttons can be used to either enlarge or shrink the plots within the chart, decreasing or increasing the scale of the axes (respectively).

C. Lasso Select

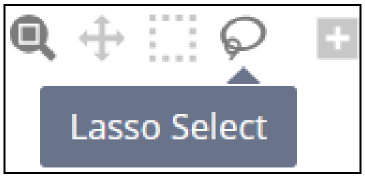


Figure 159. Identification of the Lasso Select icon in the floating menu.

The [Lasso Select] feature can be used to group cells.

1. Click the [Lasso Select] icon.

2. Click in the plot area and, while holding the mouse button down, use the mouse cursor to draw around the cells of interest. The line will automatically adjust its shape based on the movement of the mouse cursor.

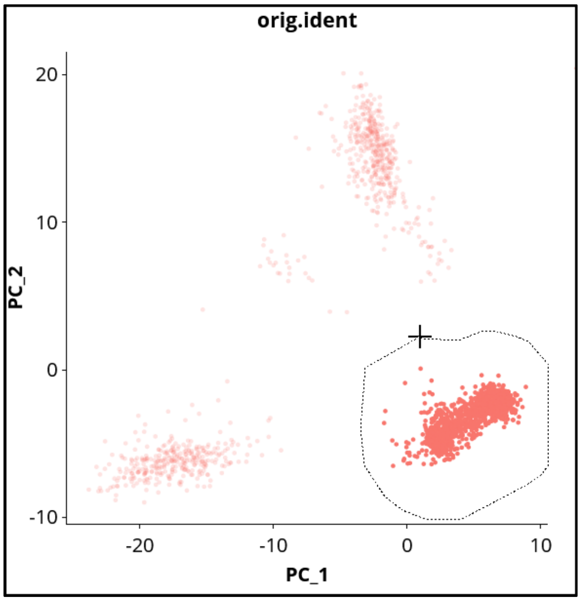


Figure 160. Lassoing a cell cluster of interest.

3. Stop pressing down on the left mouse button to complete the "select" action.

D. Download Plot as PNG

The [Download Plot as PNG] option of the floating menu is the Plotly method of downloading the UMAP plot; this is separate from the [Save plot] button of CogentDS.

There are two primary differences between the two methods:

- When using [Save plot], you can select from up to four different file types (PNG, PDF, SVG, or JPEG). The [Download plot as png] button only allows download as PNG file.
- When using [Save plot], the original plot view (with all clusters) will be saved in a large format (e.g., 3,000 x 3,000 px). The [Download plot as png] option will respect any modifications you made through the floating menu, such as zooming in or out, or panning, but as a much smaller file (700 x 500 px).

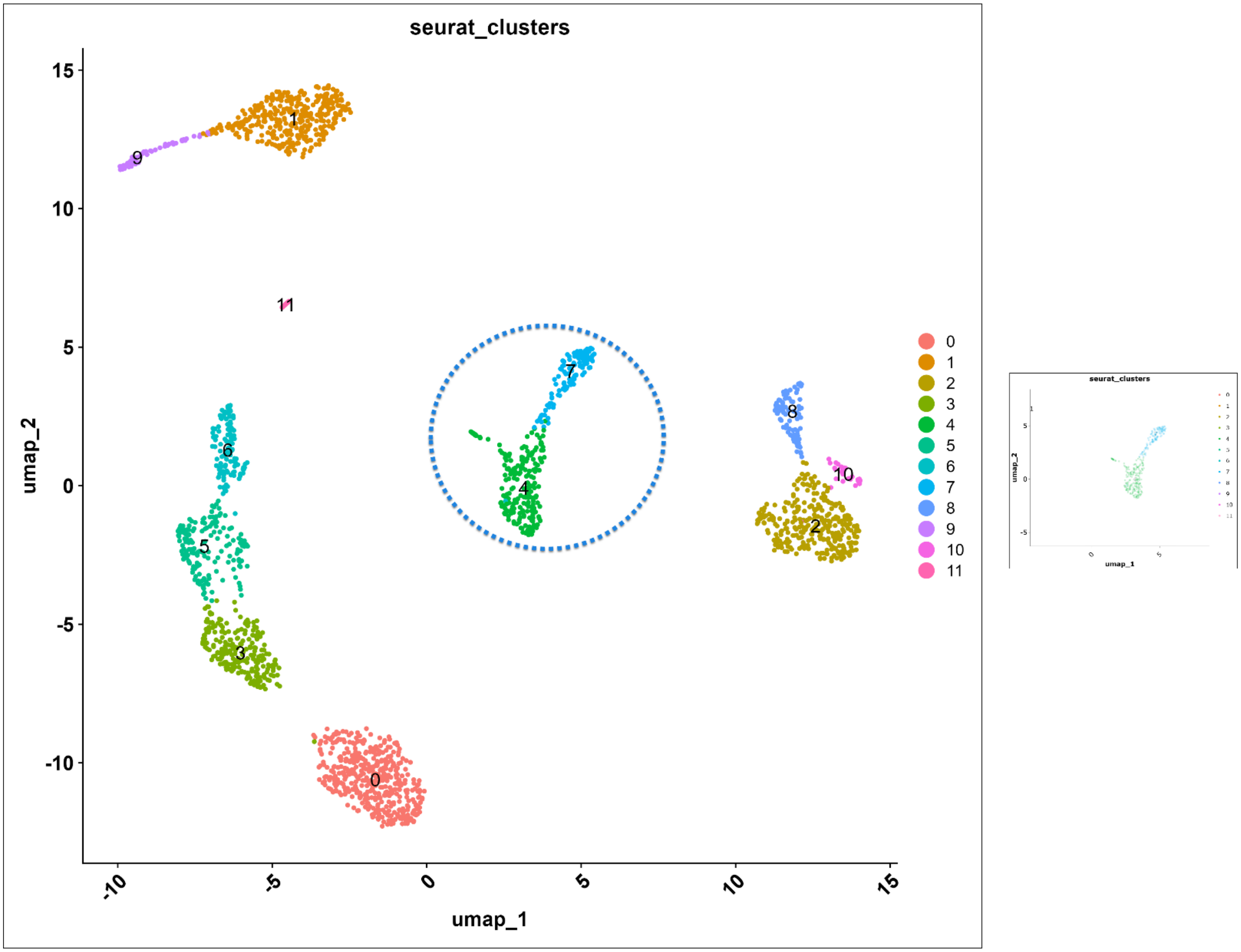


Figure 161. Comparison of a UMAP plot from [Save plot] versus [Download plot as png]. Before saving the images, the [Zoom in] button was used to focus on the middle cluster, circled in the larger image. Images are to-scale compared to each other. (**Left**) A UMAP plot saved with [Save plot] displays all the plots, ignoring the zoomed in view. (**Right**) The plot saved with the [Download plot as png] button. It honors the zoomed in view, showing only the middle cluster, but saves with a much lower resolution.

Contact Us	
Customer Service/Ordering	**Technical Support**
tel: 800.662.2566 (toll-free)	tel: 800.662.2566 (toll-free)
fax: 800.424.1350 (toll-free)	fax: 800.424.1350 (toll-free)
web: takarabio.com/service	web: takarabio.com/support
e-mail: ordersUS@takarabio.com	e-mail: technical_support@takarabio.com

Notice to Purchaser

Our products are to be used for **Research Use Only**. They may not be used for any other purpose, including, but not limited to, use in humans, therapeutic or diagnostic use, or commercial use of any kind. Our products may not be transferred to third parties, resold, modified for resale, or used to manufacture commercial products or to provide a service to third parties without our prior written approval.

Your use of this product is also subject to compliance with any applicable licensing requirements described on the product's web page at takarabio.com. It is your responsibility to review, understand and adhere to any restrictions imposed by such statements.

This document has been reviewed and approved by the Quality Department.