Takara Bio USA

Cogent™ NGS Discovery Software v1.5 User Manual

(052323)

Takara Bio USA, Inc.
2560 Orchard Parkway, San Jose, CA 95131, USA
U.S. Technical Support: technical_support@takarabio.com

United States/Canada	Asia Pacific	Europe	Japan
800.662.2566	+1.650.919.7300	+33.(0)1.3904.6880	+81.(0)77.565.6999

Table of Contents

Table of Figures

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 the ICELL8® cx Single-Cell system and plate-based NGS workflows.

II. Before You Begin

A. Supported Operating Systems

The CogentDS is designed to be installed on a user workstation 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 7, Windows 10
- macOS Mojave (10.14) or higher
- Linux CentOS 6.9 or higher

B. Hardware Requirements

CogentDS with its dependencies is a lightweight program. It should work on any basic workstation (desktop or laptop) with >2 GB of free disk space and a minimum of 8 GB RAM (recommended 16 GB).

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 dependences (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.

- **devtools**

  devtools is a free, open-source R tool that enhances the development and installation of R packages; it is used to install Cogent NGS Discovery Software. More information on obtaining and installing RStudio devtools is available in Section IV.D.

- **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 package, 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.

- **Pandoc (optional)**

  Pandoc is another R utility that is installed natively with RStudio Desktop. For advanced users that wish to forgo RStudio and run R from the command line, the Pandoc Software Package must be downloaded, installed, and placed in the computer's `PATH`.

E. Required Input Files

CogentDS requires one of the two following file options as input. The listed files are output from Cogent NGS Analysis Pipeline v2.0.1 (CogentAP):

- `CogentDS.analysis.rda`, 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 transcript-count based, gene-fusion, and immune analysis, if these optional analyses have been run first in CogentAP. An added advantage to this input is that it opens the option to use precalculated settings for the quality control, clustering, and differential expression modules, resulting in faster upload speeds.

- 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 1.0 and its predecessor, mappa™ v1.0). More information on this process can be found in Section V.B.

III. Software Overview

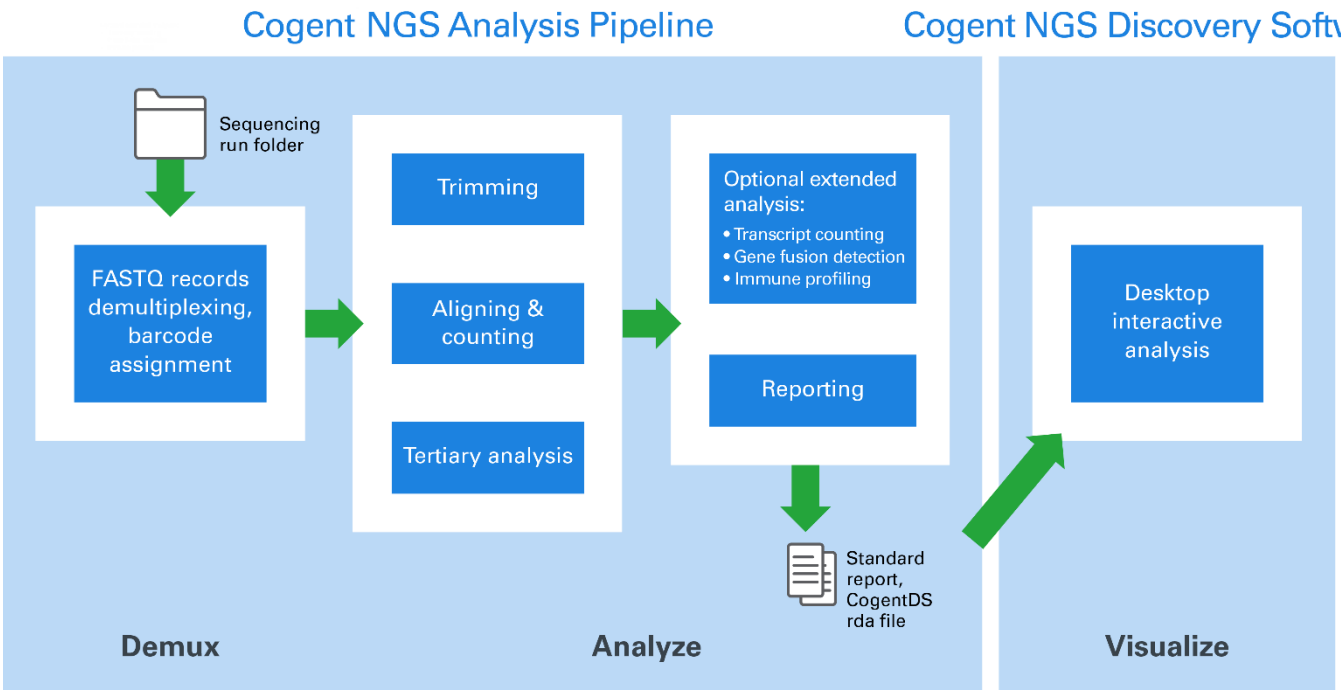


Figure 1. High-level analysis workflow of Cogent NGS Analysis Pipeline (CogentAP) and Cogent NGS Discovery Software (CogentDS).

Figure 1 depicts the high-level workflow of the analysis provided by CogentAP and how its output is carried over to CogentDS. For more information about Cogent NGS Analysis Pipeline, see the Cogent NGS Analysis Pipeline v2.0.1 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, skip to Section IV.F.

A. Install R

R and many of the contributed packages are available on the Comprehensive R Archive Network (CRAN). If R is not installed on your system, please download and install R version 4.0.0 or higher from https://www.r-project.org, by first choosing a CRAN mirror of your choice.

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 third-party software be installed prior to installing RStudio.

1. Windows

On Windows, R requires Rtools to build and install packages from a source file. Download the most recent version of Rtools compatible with R 4.0.x (e.g., rtools40) 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 which 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.

2. macOS

R version 4.0.x requires `clang-7.0.0.pkg`. This can be downloaded from https://cran.r-project.org/bin/macosx/tools/.

C. 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.

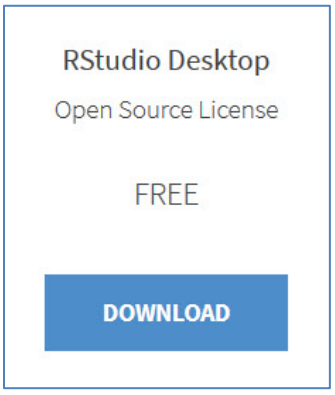


Figure 2. rstudio.com screenshot of the package to download.

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).

D. Install Devtools

CogentDS requires devtools version 2.4.2 or higher be installed on the computer prior to its installation. Refer to https://cran.r-project.org/web/packages/devtools/readme/README.html for detailed information about this package.

1. If devtools is already installed on the computer, verify that the version is 2.4.2 or later. Start Rstudio and run the following command from the Console prompt:

   ```
   packageVersion("devtools")
   ```

   ```
   > packageVersion("devtools")
   [1] '2.4.2'
   ```

   Figure 3. Example devtools version check in RStudio.

2. To install or update the devtools version, enter the following command into the Console window of RStudio:

   ```
   install.packages("devtools")
   ```

E. (Optional) Install Pandoc

Instructions for downloading and installing Pandoc can be found at pandoc.org.

F. Uninstall Previous Instances of CogentDS

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

If an earlier version of CogentDS or hanta™ Software was installed on the server, it should be uninstalled prior to installing Cogent NGS Discovery Software using the procedure in Section IV.G.

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

G. Install Cogent NGS Discovery Software v1.5

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

1. Download the CogentDS ZIP file (`Cogent_NGS_Discovery_Software_v1.5.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. From the devtools command line, run:

   ```
   devtools::install_local("<PATH>/Cogent_NGS_Discovery_Software_v1.5.zip")
   ```

   where **<PATH>** will be the folder location where the ZIP file was saved during Step 1.

   Example:

   If the `Cogent_NGS_Discovery_Software_v1.5.zip` file was downloaded to `C:\temp` on a Windows server, the command would resemble:

   ```
   devtools::install_local("C:/temp/Cogent_NGS_Discovery_Software_v1.5.zip")
   ```

```
> devtools::install_local("C:/temp/Cogent_NGS_Discovery_Software_v1.5.zip")
Installing 121 packages: matrixStats, S4Arrays, IRanges, S4Vectors, MatrixGenerics,
```

Figure 4. Example of the text displayed when the Cogent NGS Discovery Software installation starts.

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 the user to accept downloading and installing certain packages from source. Answer 'yes' to any such prompts.

During the installation, the script may notice updates to the R dependencies installed. If that occurs, it is recommended to select whatever the 'All' value is (1 in the Figure 5 example, below) to upgrade them.

```
These packages have more recent versions available.
It is recommended to update all of them.
Which would you like to update?

 1: All
 2: CRAN packages only
 3: None
 4: xfun          (0.26   -> 0.27  ) [CRAN]
 5: rlang         (0.4.11 -> 0.4.12) [CRAN]
 6: crayon        (1.4.1  -> 1.4.2 ) [CRAN]
 7: cli           (3.0.1  -> 3.1.0 ) [CRAN]
 8: pillar        (1.6.3  -> 1.6.4 ) [CRAN]
 9: generics      (0.1.0  -> 0.1.1 ) [CRAN]
10: future        (1.22.1 -> 1.23.0) [CRAN]
11: igraph        (1.2.6  -> 1.2.7 ) [CRAN]
12: deldir        (1.0-5  -> 1.0-6 ) [CRAN]
13: spatstat.... (2.3-0  -> 2.3-1 ) [CRAN]
14: rvest         (1.0.1  -> 1.0.2 ) [CRAN]
15: Seurat        (4.0.4  -> 4.0.5 ) [CRAN]

Enter one or more numbers, or an empty line to skip updates: 1
```

Figure 5. Example of CogentDS installation process detecting newer R dependency packages.

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

H. How to Uninstall Cogent DS

To uninstall Cogent NGS Discovery Software, 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. Launching Cogent NGS Discovery Software

A. Getting Started

Once installation is complete, CogentDS can be launched with the following command at the Console prompt of RStudio:

```
CogentDS::launch()
```

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).

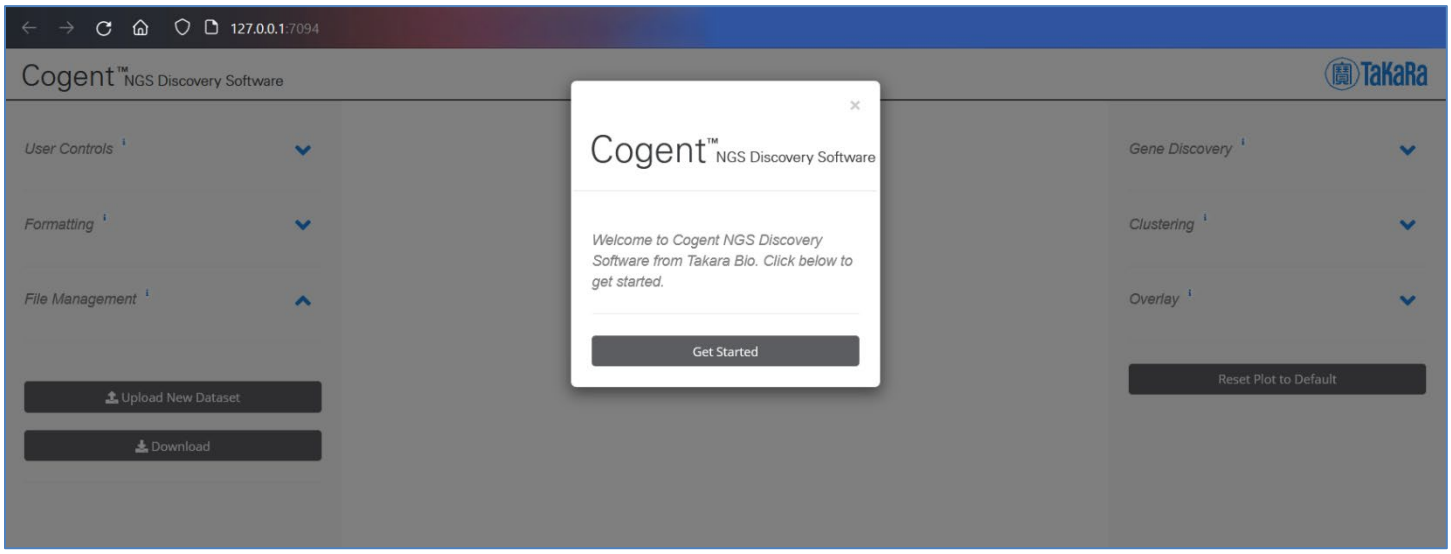


Figure 6. Initial screen of CogentDS in the web browser.

B. Upload Data

Click [Get Started] to start the process. The *Select Input Data* window will pop up. Choose the input data type from the dropdown menu, which displays 'CogentDS Rda file' by default.

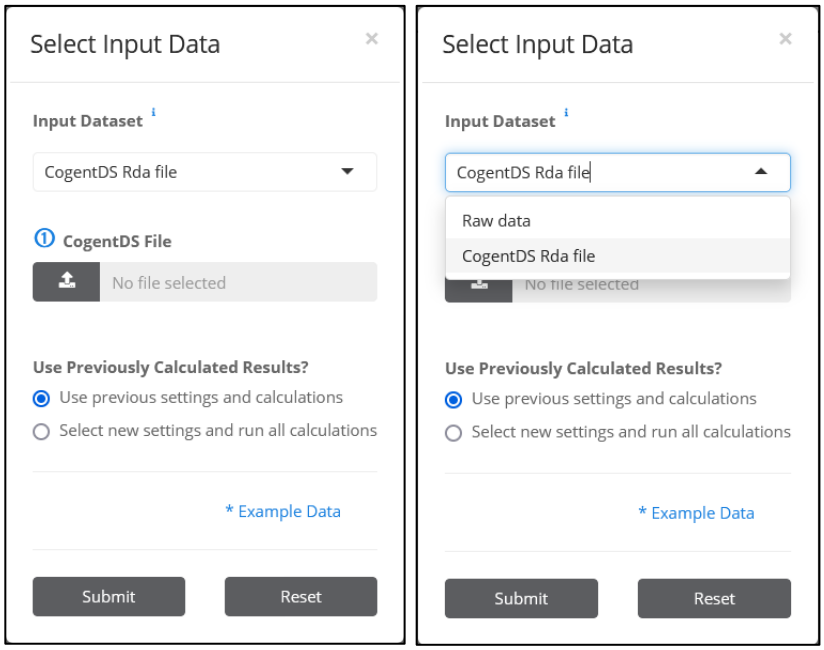


Figure 7. The *Select Input Data* browser pop-up window. **(Left)** Default view upon launch. **(Right)** Dropdown menu for input options.

NOTE: The *Select Input Data* menu can also be accessed with the [Upload New Dataset] button from the **File Management** menu in an established CogentDS session. See Figure 57 in Section V.F for where to locate the button.

The *Select Input Data* window allows the user to enter data in one of two ways.

1. Through a `CogentDS.analysis.rda` file, an R object file which is output from CogentAP.

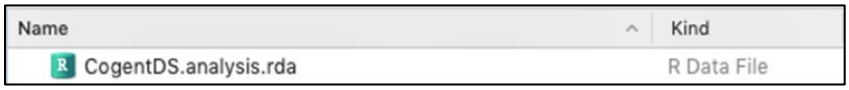


Figure 8. Typical R data file generated by CogentAP, that is the recommended input into CogentDS.

This option is recommended for the following two advantages:

- It enables full analysis capabilities in CogentDS, which includes using either gene- or transcript-count data for clustering and the ability to overlay gene fusion and immune analysis data on the clustering. When using our full-length chemistry, selecting transcript counts for analysis can lead to discovery of specific transcript variants that drive clustering, especially in single cells (refer to Section V.E for more details).

- It allows you to choose between 'Use previous settings and calculations' and 'Select new settings and run all calculations.'

    - The first option takes advantage of existing data and settings that have already been run through the quality control, clustering, and differential expression modules in CogentAP, thus allowing quick visualization of the data.

    - The second option allows you to customize the quality control (QC) and analysis settings.

2. Through raw gene matrix and stats/metadata files, also output from CogentAP.

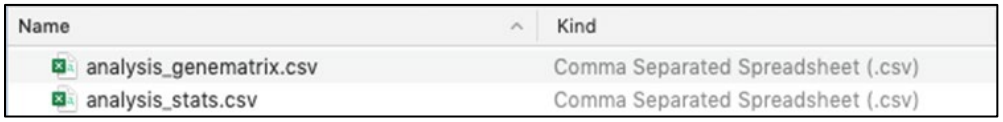


Figure 9. Typical CSV Data files generated by CogentAP, that can be input into CogentDS.

Refer to Section V.C (next section) for more information about these files. This option enables analysis using files generated by previous versions of CogentAP (v1.0 and mappa v1.0) and can only be used to perform gene-count based clustering.

NOTE: For optimal performance, compress (.zip) the gene matrix file before uploading.

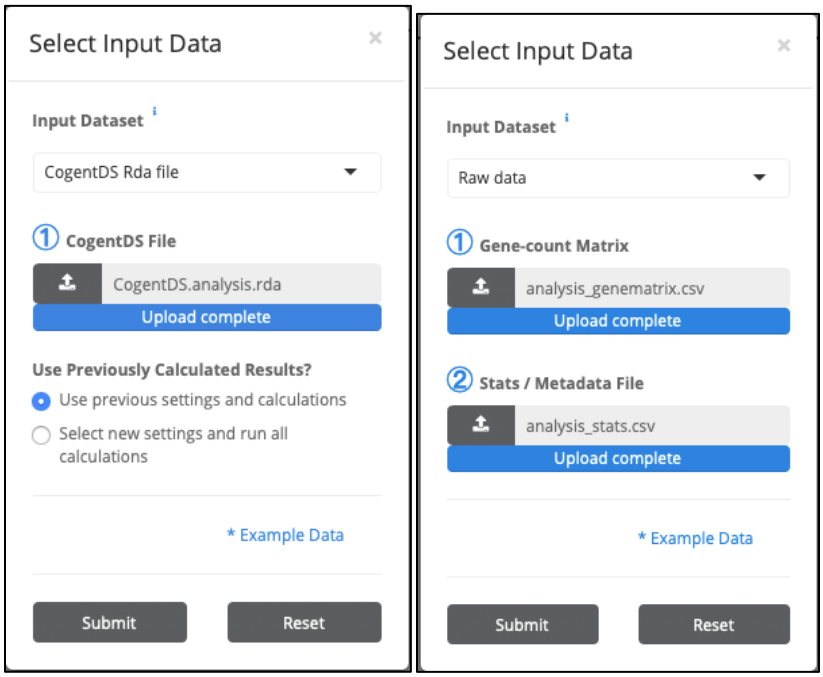


Figure 10. Selecting the example Cogent NGS Analysis Pipeline output for input into CogentDS. (Left) The CogentDS.analysis.rda file, **(Right)** The CSV gene matrix and stats files for input.

3. Once the input data source is selected, hit [Submit] to continue.

NOTE: The *Example Data link in Figure 10 can be used to download a demo dataset (named `analysis_demo.rda`) to test the software post installation. This is a small dataset and should not be used for biological inference.

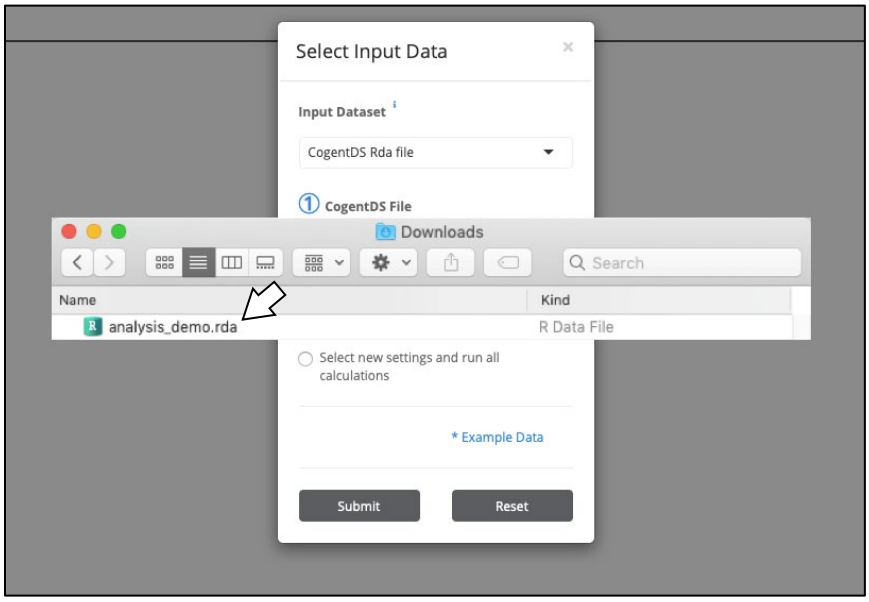


Figure 11. The prompt shown to download the `analysis_demo.rda` file from the *Example Data link.

C. Run Quality Control and Clustering Modules

If you choose 'Select new settings and run all calculations' when using a CogentDS Rda file or raw data files as your input, you will be sent to the *Quality Control* (QC) module. The QC module and following menus come populated with standard values for filters to help novice users get started with their analysis. It is recommended that users vary these values and repeat the analysis to arrive at output best suited for the dataset in question.

NOTE: If you're using the raw data CSV files or the CogentDS rda file does not contain transcript analysis data, skip to Step 2.

1. If you're using a CogentDS rda file as input that includes transcript analysis from CogentAP (such as the `analysis_demo.rda` file from the Example Data), you will first be asked whether to use gene counts or transcript counts for clustering. Select whichever option you prefer to do the analysis and click [Accept].

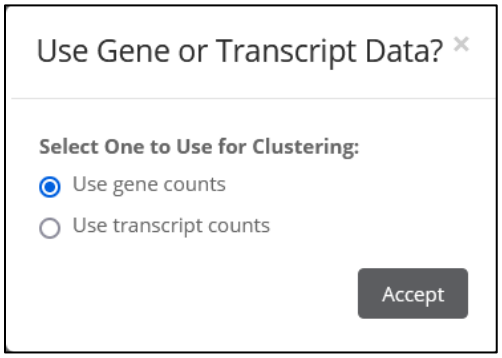


Figure 12. Initial window prompt prior to entering the *Quality Control* window when the data includes transcript analysis information.

2. The next screen is the *Quality Control* window for QC filter options. By default, the "Advanced QC Options" is checked (selected), and you will see the options below it (Figure 13).

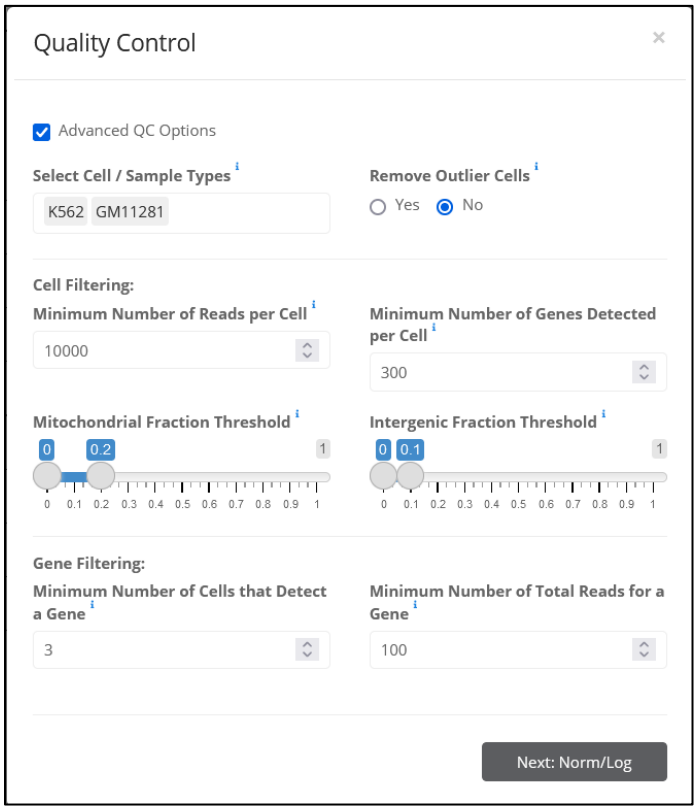


Figure 13. Default *Quality Control* menu for QC filter options.

Select how you would like to filter out uninformative cells and genes from the gene- or transcript-count data. Make changes to the QC menu and click the [Next: Norm/Log] button when finished.

3. The *Normalization* window is for normalization and transformation options.

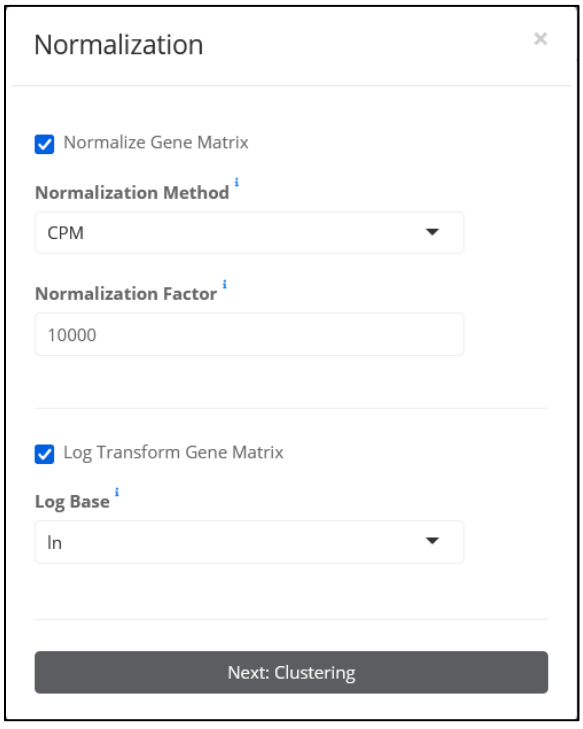


Figure 14. The *Normalization* window of the quality control module, for normalization and transformation options.

a. **Normalize Gene Matrix**

The available normalization methods include Counts Per Million (CPM), Transcripts Per Kilobase Million (TPM), and Reads Per Kilobase Million (RPKM).

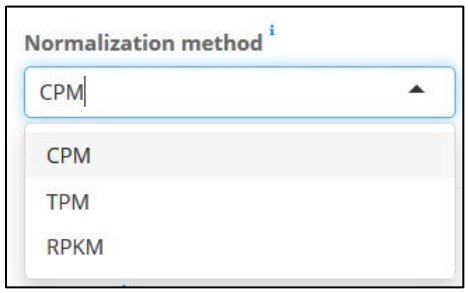


Figure 15. Expanded "Normalization method" drop-down menu.

To normalize by 'median cell coverage', select 'CPM' from the "Normalization method" drop-down menu and type 'median' into the "Normalization factor" input box (Figure 16).

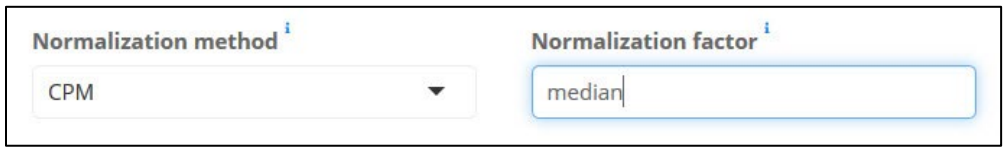


Figure 16. Parameters to normalize by median cell coverage.

b. **Log transform gene matrix?**

To log transform the data, the available options are natural log, Base 2, and Base 10.

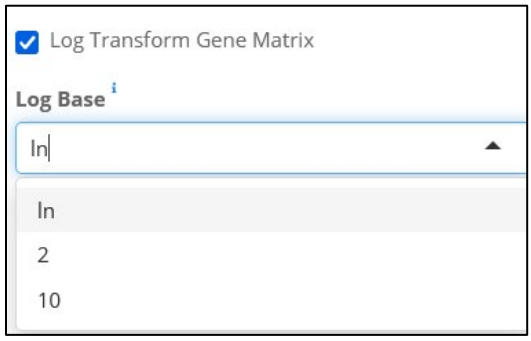


Figure 17. Expanded "Log base" transform gene matrix drop-down menu.

When all desired parameters are populated, click [Next: Clustering].

4. The next window will prompt you to enter options for the *Cluster Analysis* module. By default, CogentDS performs clustering analysis based on the 2,000 most variable genes (or transcripts, based on selection made during Step 1). These genes are used as input for a principal component analysis, and the top principal components are selected for further reduction and visualization using either the Uniform Manifold Approximation and Projection (UMAP, recommended) or t-Distributed Stochastic Neighbor Embedding (t-SNE) method.

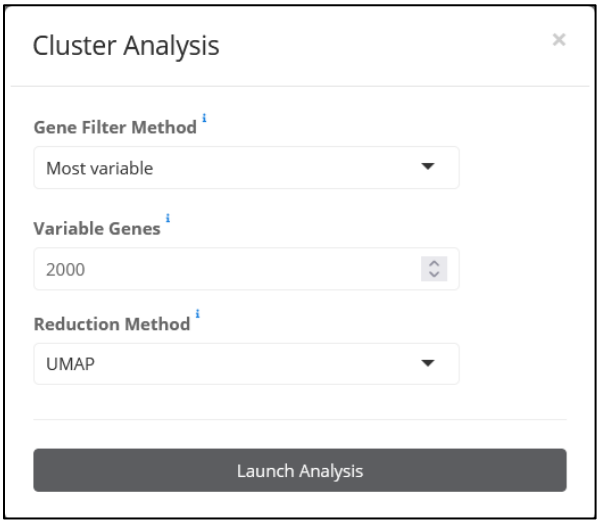


Figure 18. *Cluster Analysis* menu.

Input your parameters and click [Launch Analysis]. Data transformation will begin, and status pop-ups similar to those in Figure 19 will display on the bottom right-hand corner of the browser window.

Figure 19. Example stages of the status pop-up while the cluster analysis is running.

After running the cluster analysis, the plot will be rendered in the center of the window, with menus in the right and left sidebars and a QC table below the plot.

D. Discover

The baseline plot displays the results of graph-based clustering; however, this can be modified with the **User Controls** and **Formatting** menu options in the sidebar panel to the left of the screen.

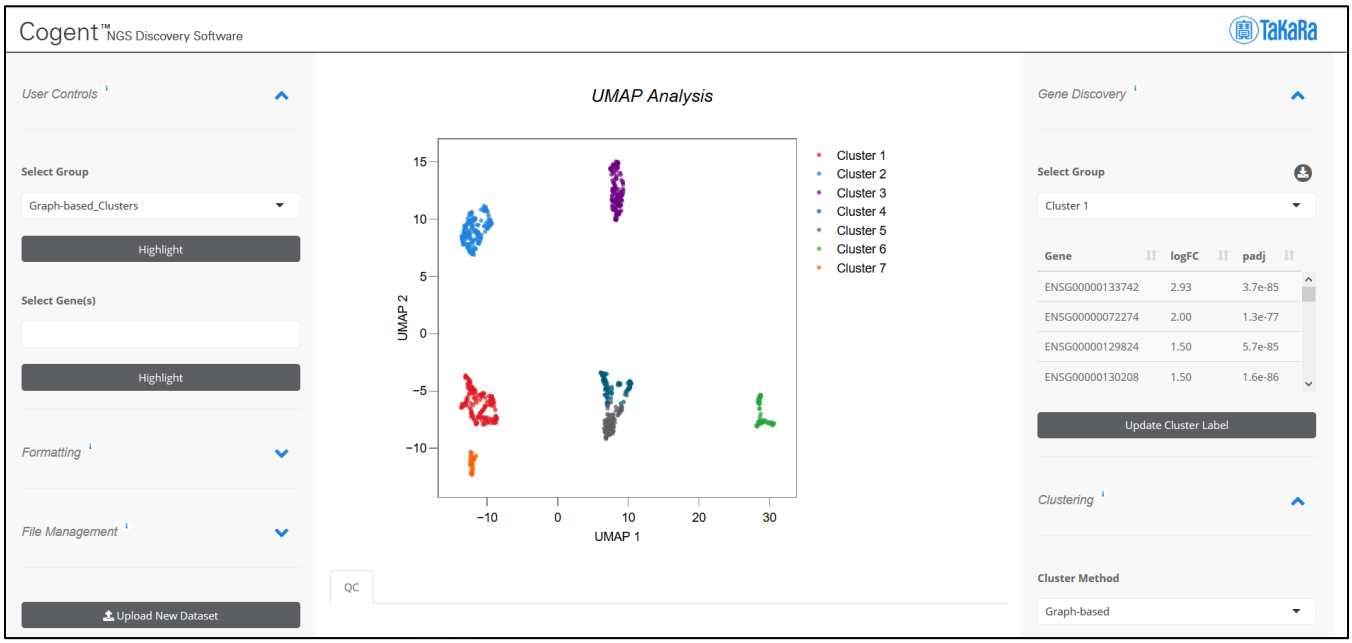
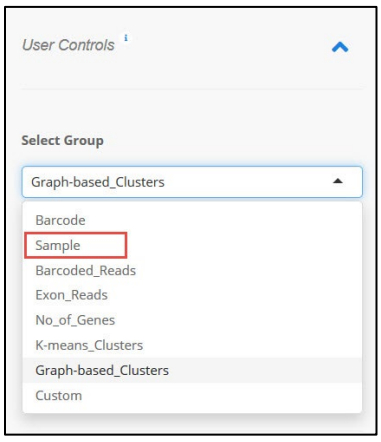


Figure 20. Baseline UMAP analysis plot.

1. User Controls

1. To highlight cells by cell type, go to the **User Controls** accordion menu in the left sidebar and select an option from the "Select Group(s)" drop-down box. This field is pre-populated with all column headers from the metadata file. Selecting different metadata features allows the user to highlight the cells by any desired method.



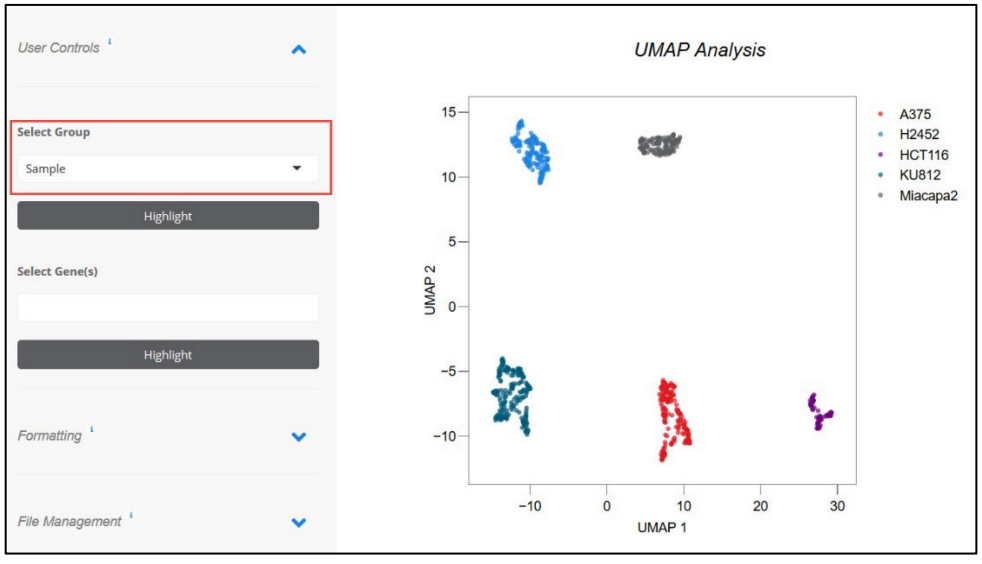


Figure 21. Example "Select Group(s)" drop-down menu and the resulting UMAP analysis plot with cells highlighted by cell sample type.

2. Another method to highlight cells is by expression levels for genes (or transcripts based on selection shown in Figure 12). Entering one or multiple genes into the "Select Gene(s)" field

plots the average expression across the panel for each cell and renders the expression into the plot. In Figure 22, the plot highlights *ENSG00000261857*, a single marker for the A375 cell type.

Figure 22. Example "Select Gene(s)" drop-down menu and the resulting UMAP analysis plot with cells highlighted for gene *ENSG00000261857*.

3. If you choose the t-SNE method, the "Perplexity" parameter can also be configured from within the **Clustering** menu in the right sidebar (see Figure 23). A feature of the t-SNE calculation that broadly serves as an estimate of the cluster size(s) within the data, high perplexity parameters will define large, global structures within the dataset, while smaller perplexities will identify small, local structures.

NOTE: Perplexity defaults have been optimized for general-use cases of the ICELL8 cx Single-Cell and ICELL8 Single-Cell Systems. These values are different from the standard defaults in the Rtsne package and may need to be reoptimized for unique applications.

For more information, please refer to https://distill.pub/2016/misread-tsne/ or https://cran.r-project.org/web/packages/Rtsne/Rtsne.pdf.

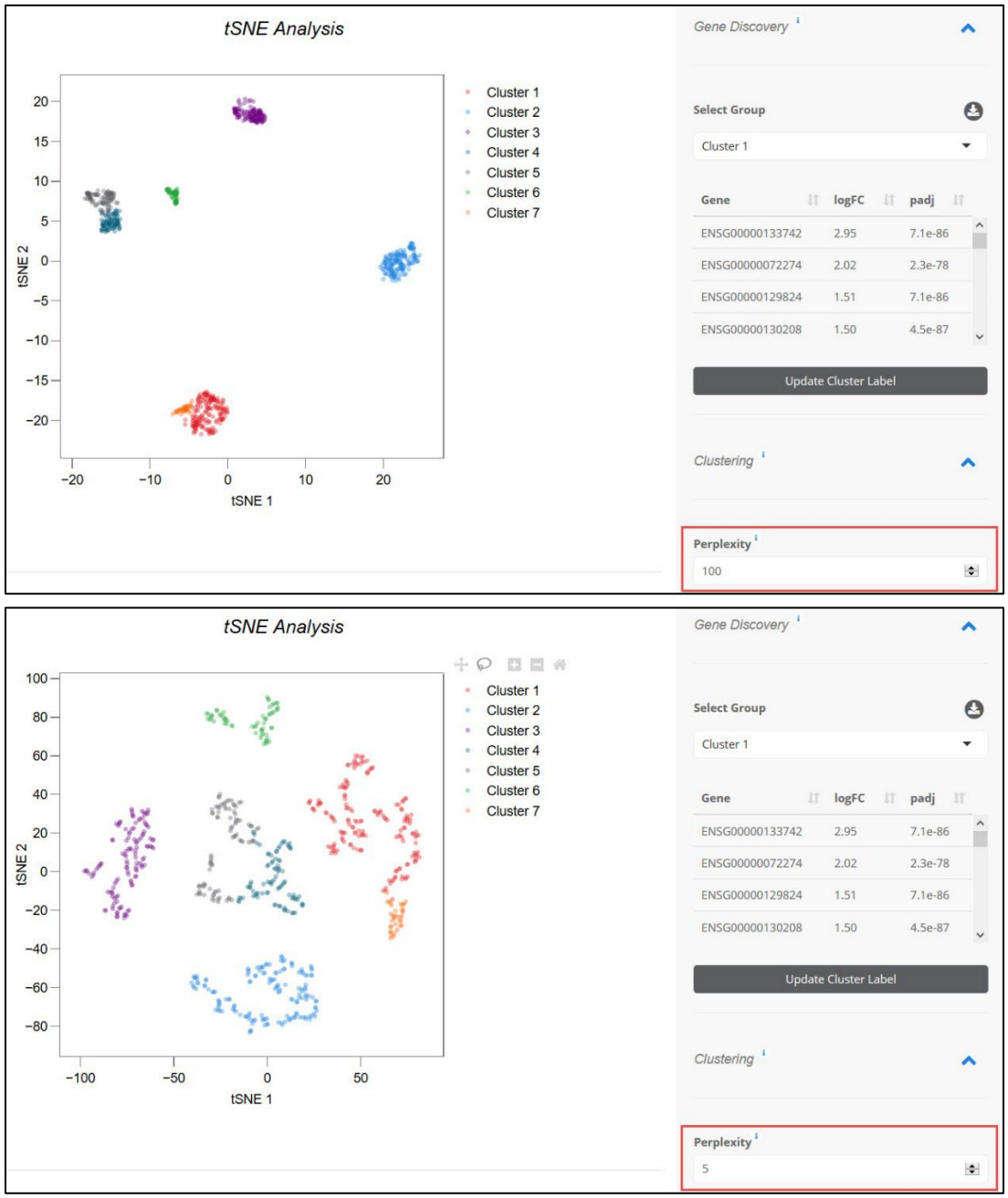
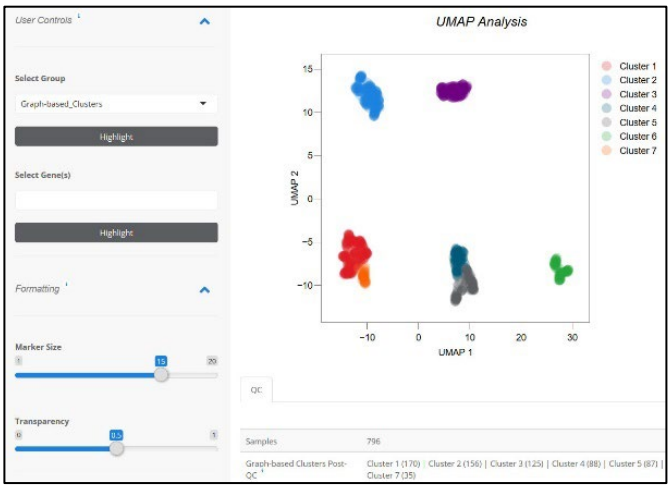


Figure 23. t-SNE analysis plot with cells clustered by high and low perplexity. **Top:** High perplexity (100). **Bottom:** Low perplexity (5).

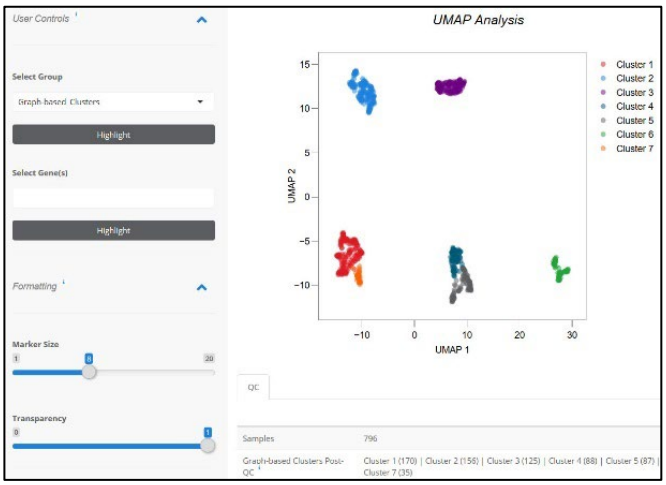
2. Formatting

In the **Formatting** tab in the left sidebar panel, the marker size and transparency can be changed. These can be used to visualize the data to the user's preferences but are also useful for identifying individual cells within larger clusters.

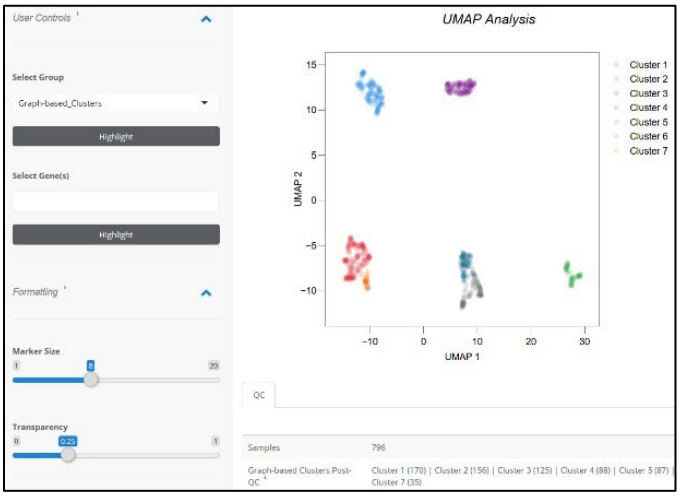


A.

B.



C.



D.

Figure 24. Illustration of different marker sizes and transparency selections applied to the same data plot. Top: (Panel A) larger and **(Panel B)** smaller marker size. **Bottom: (Panel C)** less and **(Panel D)** more transparency.

3. Floating menu

To the right of the chart, there is a menu of icons that only displays when hovered over with the mouse cursor.

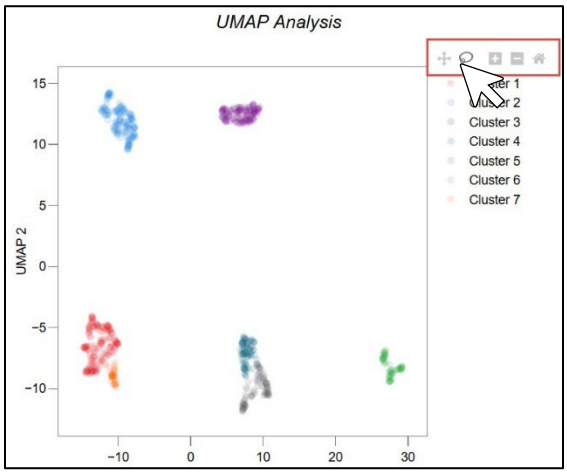


Figure 25. Location of the floating menu icons to the right of the UMAP chart.

1. **Pan** and **Reset** axes

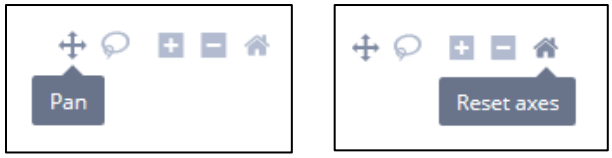


Figure 26. Identification of the Pan and Reset axes icons in the floating menu.

a. 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.

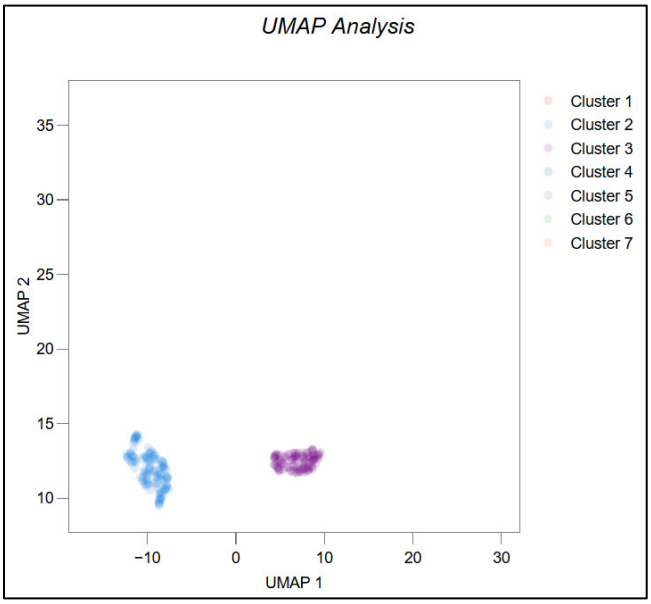


Figure 27. Example after using the Pan function to move the plots down the page. The values of the Y-axis are increased compared to the default in Figure 25.

b. [Reset axes] will return the plot to the default view (Figure 25) after using the [Pan], [Zoom in], and/or [Zoom out] functions.

2. **Zoom in** and **Zoom out**

Figure 28. 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).

3. **Lasso Select**

Figure 29. Identification of the Lasso Select icon in the floating menu.

The [Lasso Select] feature can be used to select, group, and label cells in a custom manner.

a. Click the [Lasso Select] icon.

b. 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.

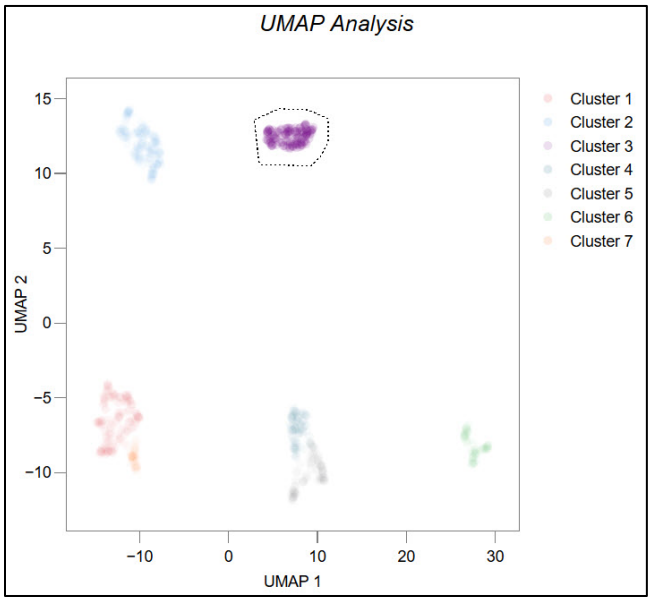


Figure 30. Lassoing a cell cluster of interest.

c. Stop pressing on the mouse button, and the *Custom Selection* window will pop up.

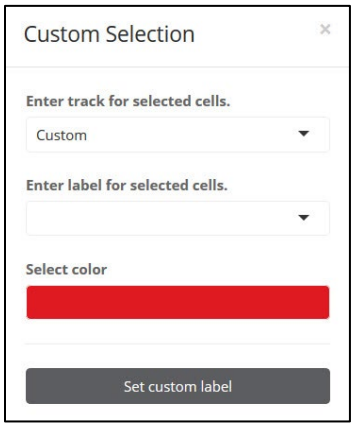


Figure 31. Default *Custom Selection* pop-up window.

i. **Enter label for selected cells**: (Optional) Type in text that will identify the cluster in the legend on the right-hand side of the chart.

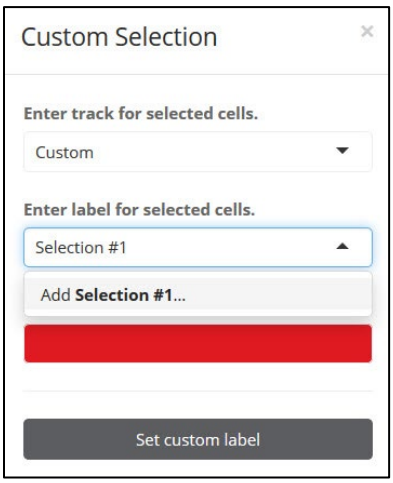


Figure 32. Typing in a custom label name for a lassoed cluster on the chart.

ii. **Select color**: (Optional) Click on the color bar to expand out to a color selector gradient. Macro changes can be made on the vertical rainbow bar, while finer gradients can be selected by moving the dot around the larger square of color shades on the left.

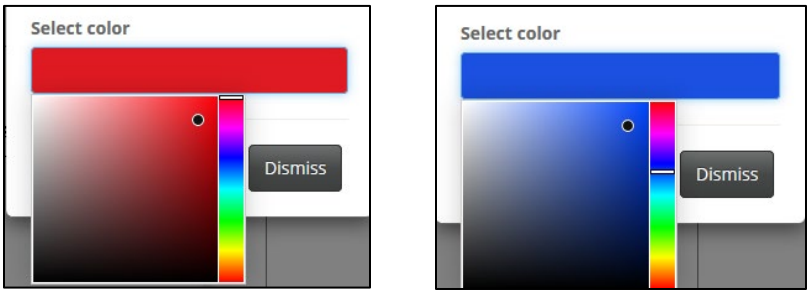


Figure 33. Examples of two different color selections using the macro and finer gradients fields.

d. Once the options are selected, hit the [Set custom label] button to apply them. To quit without applying the customization, press the [Dismiss] button.

e. If the customizations are set, the *Custom Selection* pop-up will disappear, and the chart will reflect the changes made.

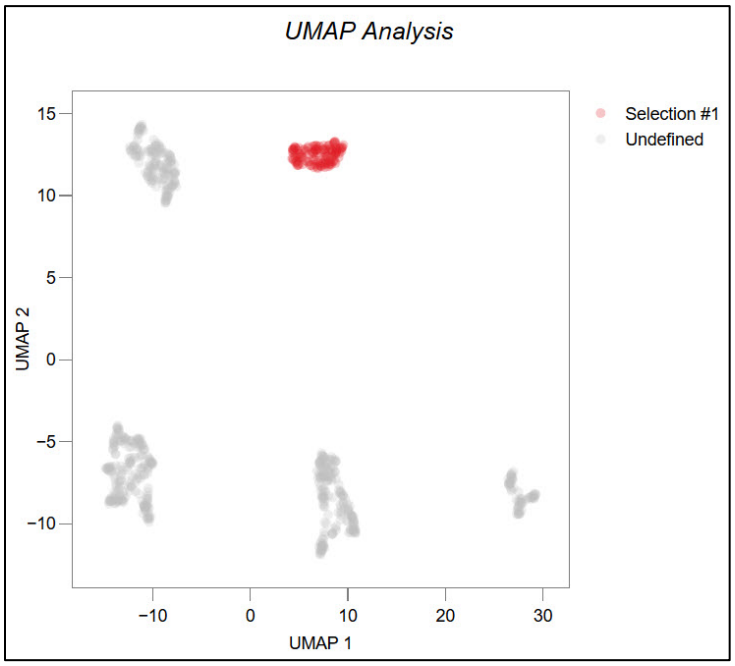


Figure 34. Illustrating the application of cluster display customization.

f. Repeat this process, if desired, for other clusters (as shown in the figure below). To reset back to the default, click the [Clear custom selections] button, located in the **User Controls** menu in the left sidebar.

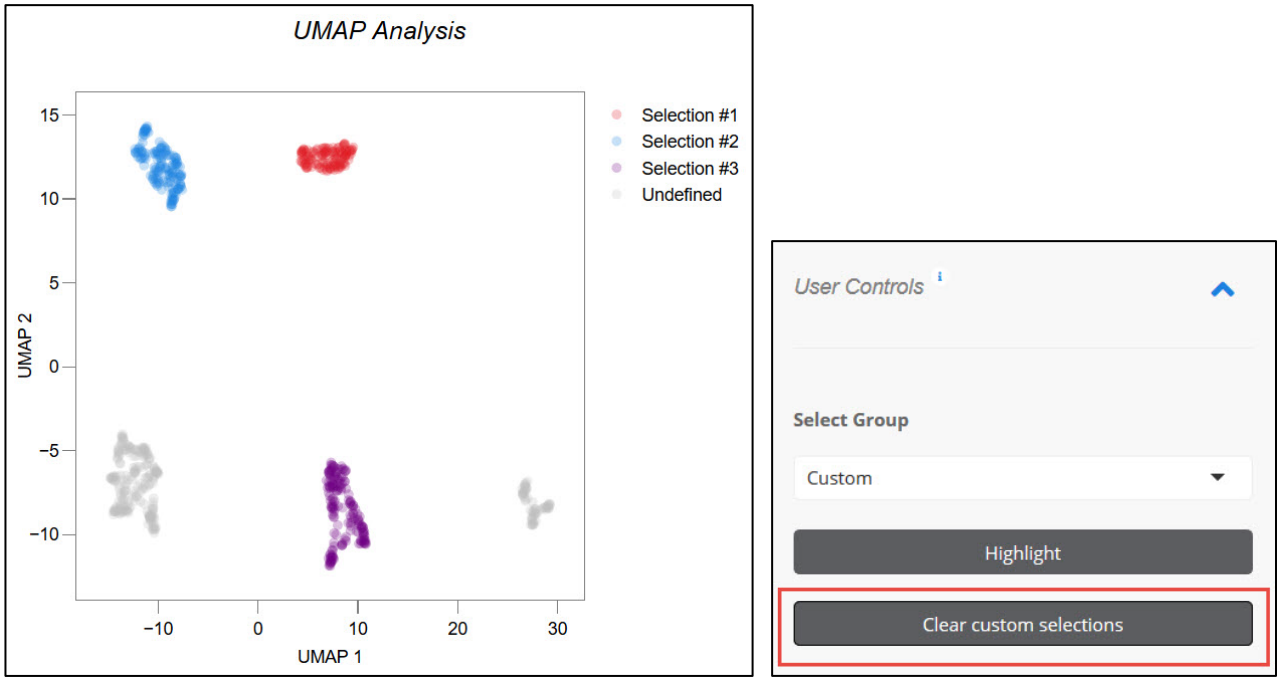


Figure 35. Multiple custom labeled and tinted clusters and the [Clear custom selections] button.

4. QC table

The Quality Control table under the main figure provides data on QC-filtering and experimental overview statistics.

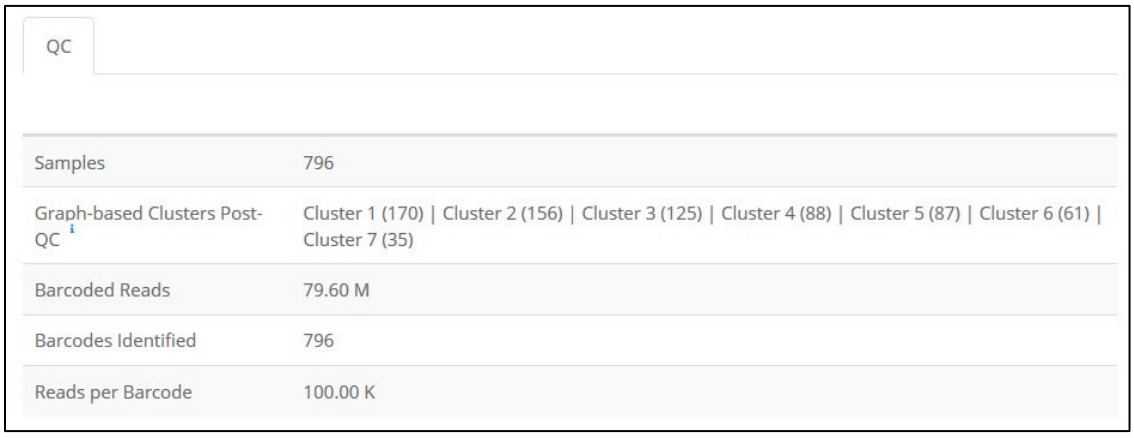


Figure 36. Example *QC* table, located in the center frame, below the main analysis figure.

- Samples: the total number of cells/samples identified for the current analysis
- Graph-based Clusters Post-QC : the number of cells/samples that remain after QC-filtering.

The remaining rows contain experimental overview statistics.

- Barcoded Reads: total number of barcoded reads found for the experiment
- Barcodes Identified: total number of barcodes identified
- Reads per Barcode: average reads per barcode

By clicking on the info icon in the first column of the second row (see Figure 37), the user can view a detailed overview of the QC methods that were selected for the current analysis (Figure 38).

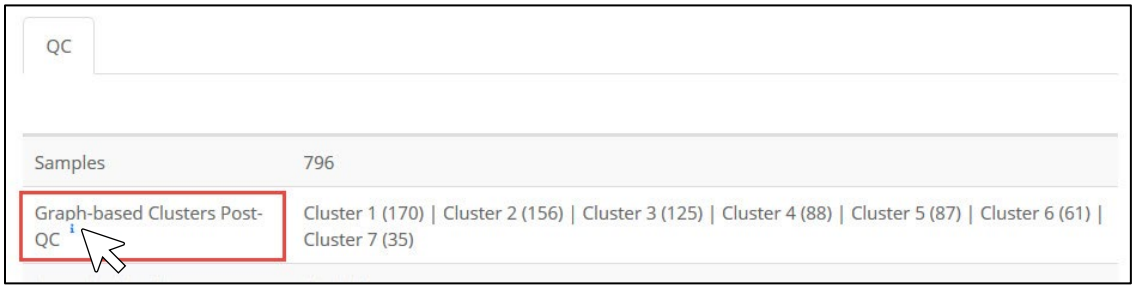


Figure 37. Info icon in the first column of the second row of the *QC* table.

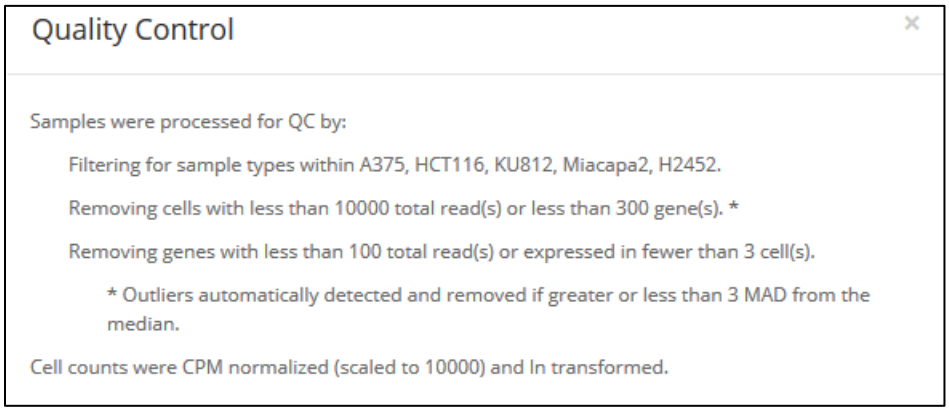


Figure 38. Detailed overview of QC methods currently in use, accessed from the QC table info icon.

5. Gene Discovery menu

The **Gene Discovery** menu in the upper-right sidebar contains information and user controls to find genes or transcripts that distinguish individual clusters.

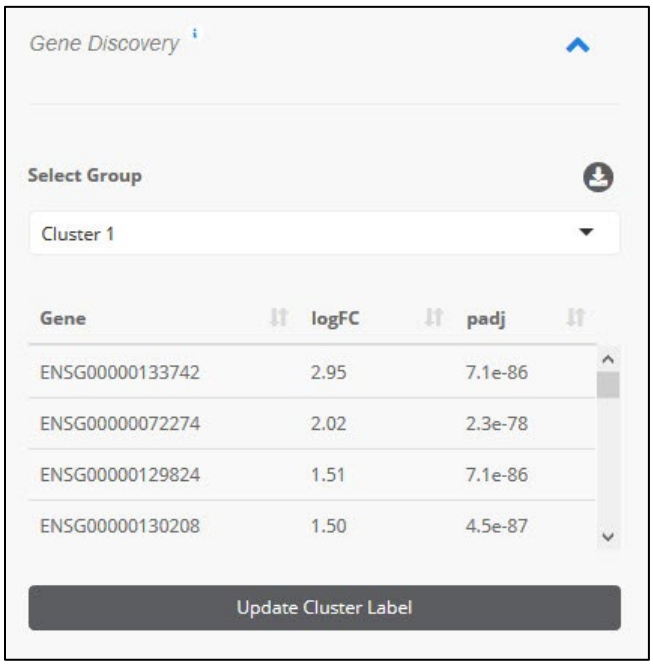


Figure 39. Gene Discovery menu.

The data in this menu is produced from the differential expression (DE) module. The module calculates DE using the Wilcoxon Rank Sum test with a Benjamini-Hochberg correction for multiple comparisons. The test is performed for each gene in each group within the selected track. The within-group expression is tested against the background, skipping genes with either <0.5 natural logFC or expressed in <10% of both groups.

The results are displayed in the **Gene Discovery** table, with the top 500 most differentially expressed genes displayed for the selected group. Scroll through the table to view genes of interest.

Clicking on a given gene will display the expression on the UMAP/t-SNE plot.

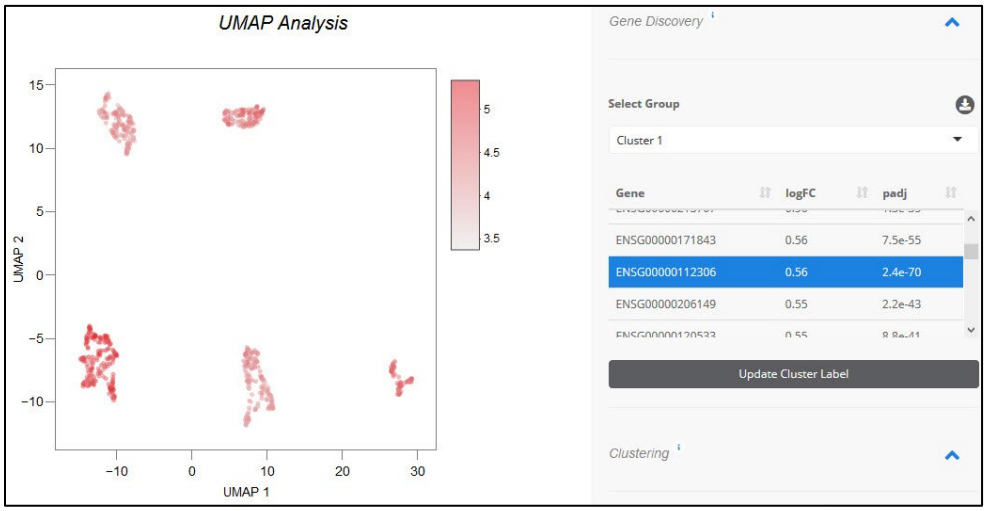


Figure 40. Display of expression of specific gene (left) selected in the Gene Discovery menu (highlighted in blue, right).

- For a full list of the differentially expressed genes, click the download link at the top right of the table.

Figure 41. Download icon for the full list of differentially expressed genes. The icon is highlighted by the red box. The downloaded list file is in CSV format.

- For graph-based clusters, K-means clusters, and any custom track selections, an [Update Cluster Label] button will appear below the **Gene Discovery** table to allow the user to update the selected cluster label.

Figure 42. Button for updating/customizing cluster labels.

This may be useful for updating generic names such as 'Cluster 1' to informative names once a cell type has been classified. For example, a cluster (e.g., 'Cluster 1') that has high expression of the gene *MS4A1* may be identified as a B-cell cluster. Click to update the generic 'Cluster 1' to the desired label.

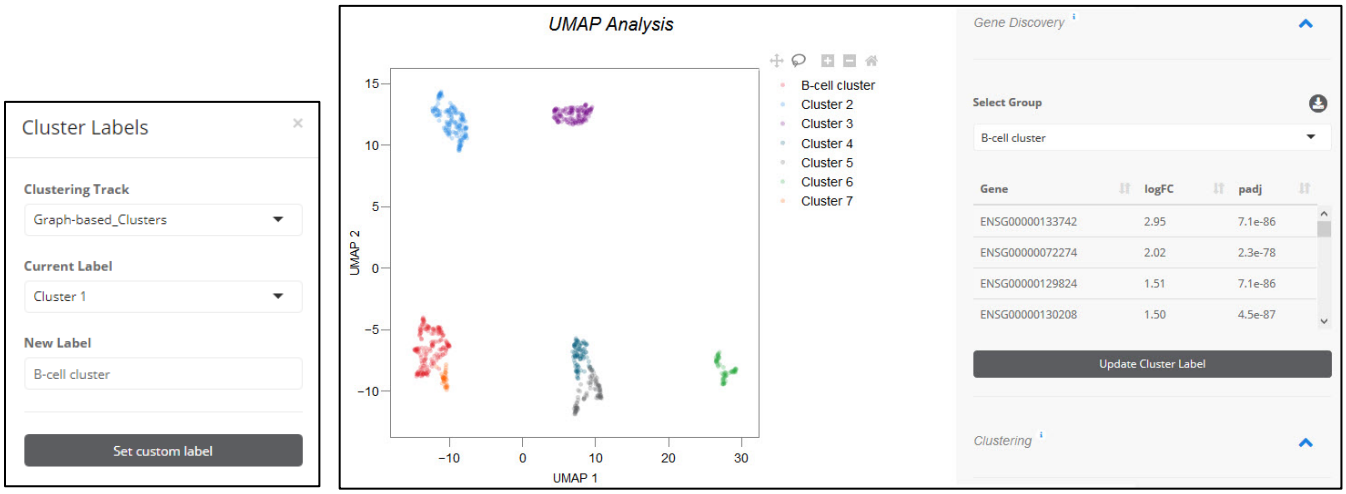


Figure 43. Example of 'Cluster 1' label updated to 'B-cell cluster'. This change is reflected both in **(Left)** the plot and **(Right)** the Gene Discovery table.

6. Clustering menu

The **Clustering** menu contains information and user controls on both dimensionality reduction and cluster identification approaches.

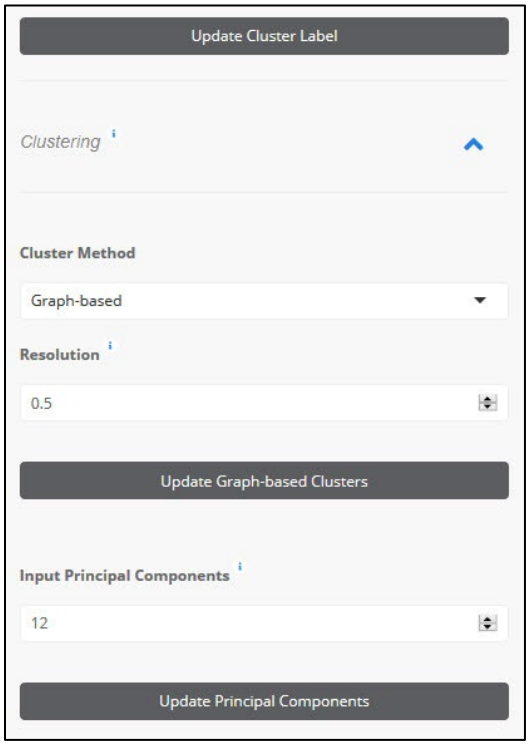


Figure 44. Clustering menu in the right sidebar. The **Clustering** menu contains information and user controls on both dimensionality reduction (PCA and UMAP/t-SNE) and cluster identification approaches (graph-based and K-means clustering).

The data in this menu is produced from the clustering module, which takes the top genes/features identified from the quality controlled, log-normalized data and uses them to perform a Principal Component Analysis (PCA). The top components from this analysis are then determined using a heuristic to find the elbow on the screen plot (with a default minimum of 8 principal components).

The top components are used as the input for clustering and visualization (UMAP/t-SNE) and can be manually set by updating the "Input Principal Components" section of the menu.

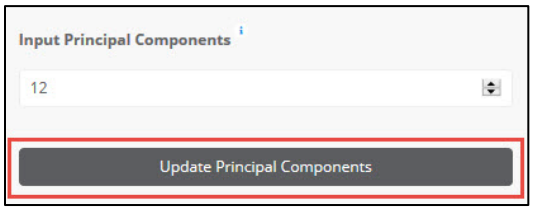
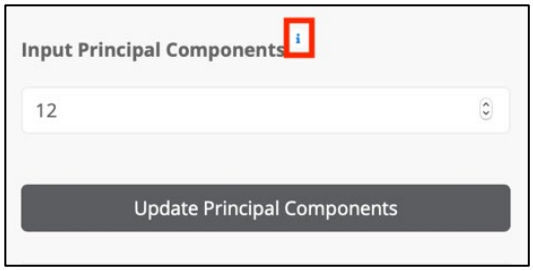


Figure 45. "Input Principal Components" section of the Clustering menu.

Clicking the [i] (info) button (Figure 46, top) for this field provides more information and plots to visualize the top principal components (PCs; Figure 46, bottom).



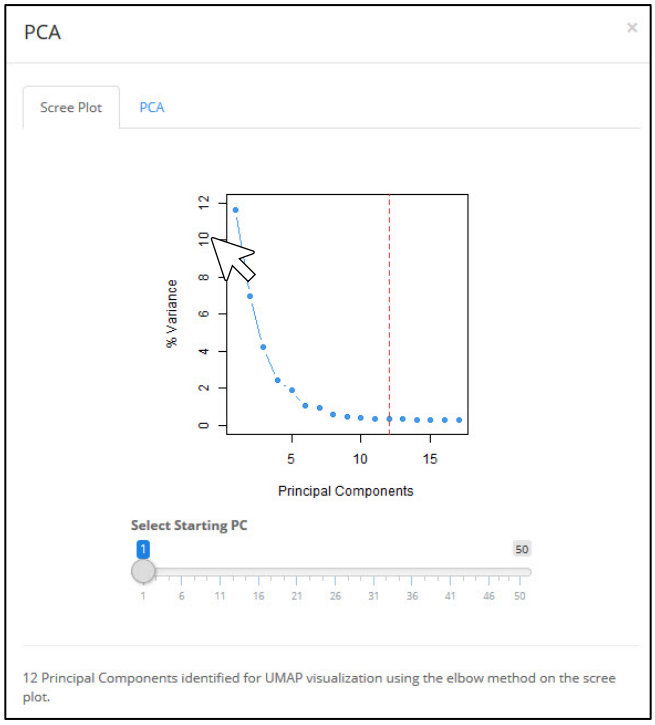


Figure 46. Accessing (top) information and plots (bottom) to visualize the top principal components.

Clustering is performed on the top PCs using either a graph-based approach (adapted from Seurat) or K-means clustering. To switch between these methods, make the desired selection from the "Cluster Method" input within the menu.

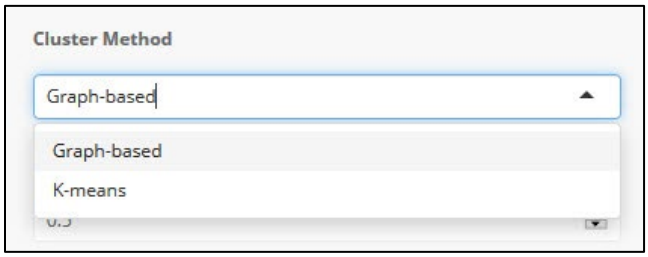


Figure 47. "Cluster Method" section of the Clustering menu.

Graph-based clustering is the default method used for identifying clusters. The granularity of the clusters can be changed by setting the "Resolution" parameter within the menu. Values closer to 0 will find broader structure in the data, while values closer to 1 and above will reveal finer structure.

Figure 48. Two UMAP images of the same data with different "Resolution" values in the Clustering menu. (Top) Resolution of 0.2. **(Bottom)** Resolution of 0.8. The two examples illustrate how different values affect the display.

K-means clusters are calculated for values of K = 2–15 by default. The optimal number of clusters from this set is determined using the Average Silhouette Score Method. Users may also select custom values of K to calculate and observe a desired number of clusters.

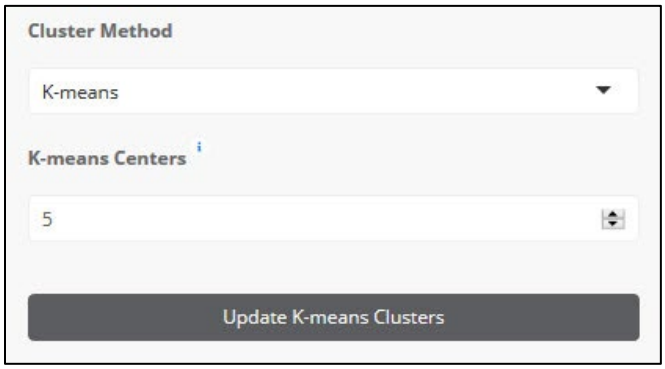
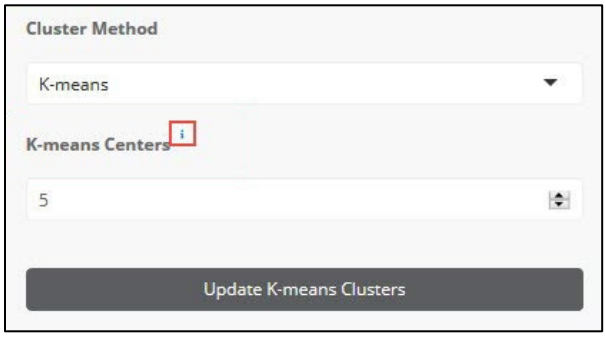


Figure 49. K-means Cluster menu.

For visualization of this approach, click the [i] (info) button (Figure 50, top) within the menu.



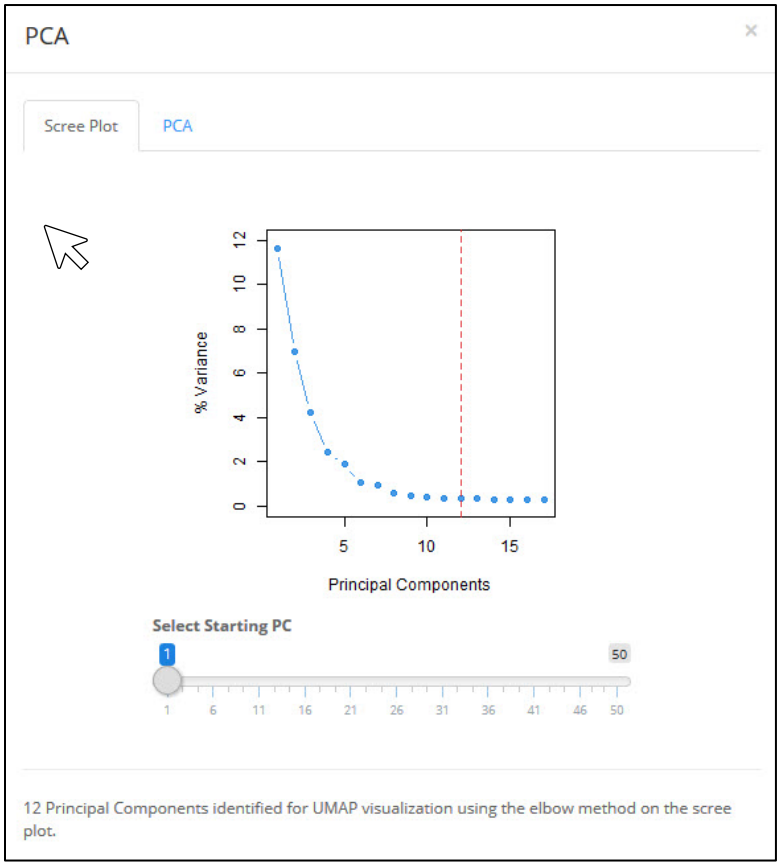


Figure 50. Accessing information (top) and plots (bottom) to visualize the K-means cluster approach.

E. Leveraging Full-length Data in Analysis

Using the SMART-Seq® Pro application provides an advantage to 3'-end chemistries by assaying fragments along the full-length of the transcriptome which can then be leveraged for transcriptome-based expression analysis, gene-fusion analysis, etc. The full-length data can also provide significantly increased biological insights when used in conjunction with 3'-end based data.

CogentAP and CogentDS enable the analyses listed below towards this goal. These analyses are first run in CogentAP as extensions of gene-based analysis; the results get stored in the same rda output file where the gene-based counts are stored and used as input into CogentDS (Section II.E).

7. Transcript-count Analysis

This feature requires that the transcript analysis option be performed in CogentAP to generate transcript counts (Cogent NGS Analysis Pipeline v2.0.1 User Manual, Section V.D.1). When the data is uploaded to CogentDS, transcript counts can be chosen for the basis of analysis, as shown in Figure 12 (Section V.B).

Specific differences in the expression of transcript variants of genes can be used to separate the clusters—information that is lost in gene-based analysis. After the plot is rendered, you can:

- Search for known transcripts and color the clusters by the transcript expression (Figure 51).

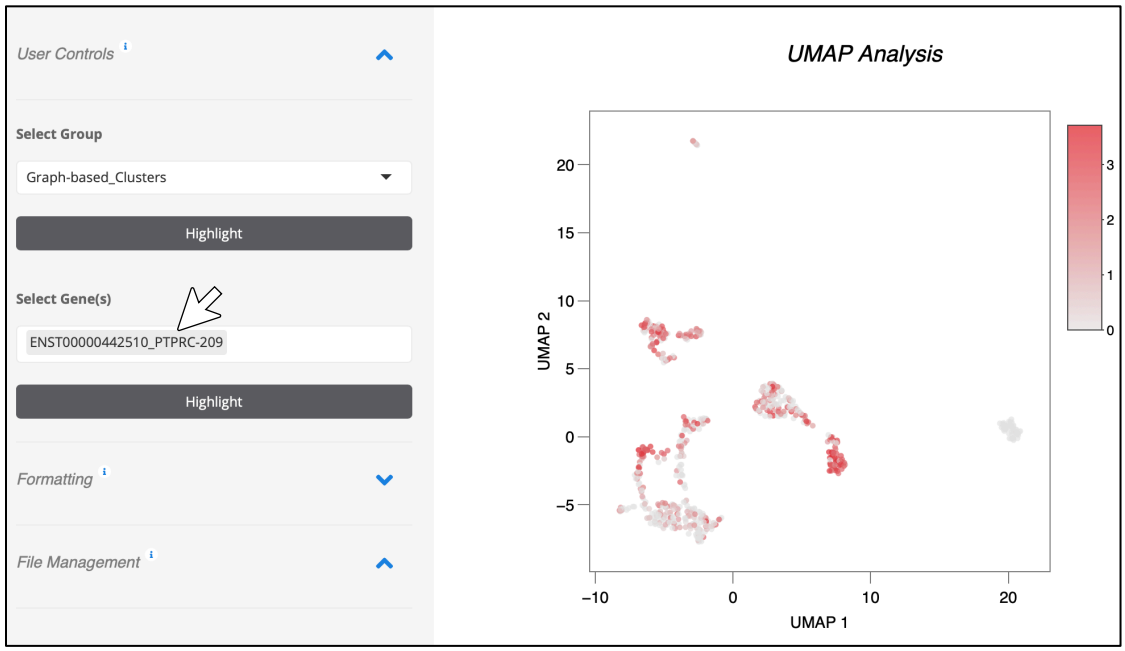


Figure 51. Selection of a given transcript and clicking 'Highlight' colors the clustering by its expression.

- Discover transcripts that are differentially expressed in a certain cluster as compared to the other clusters, using the **Gene Discovery** sidebar (Figure 52).

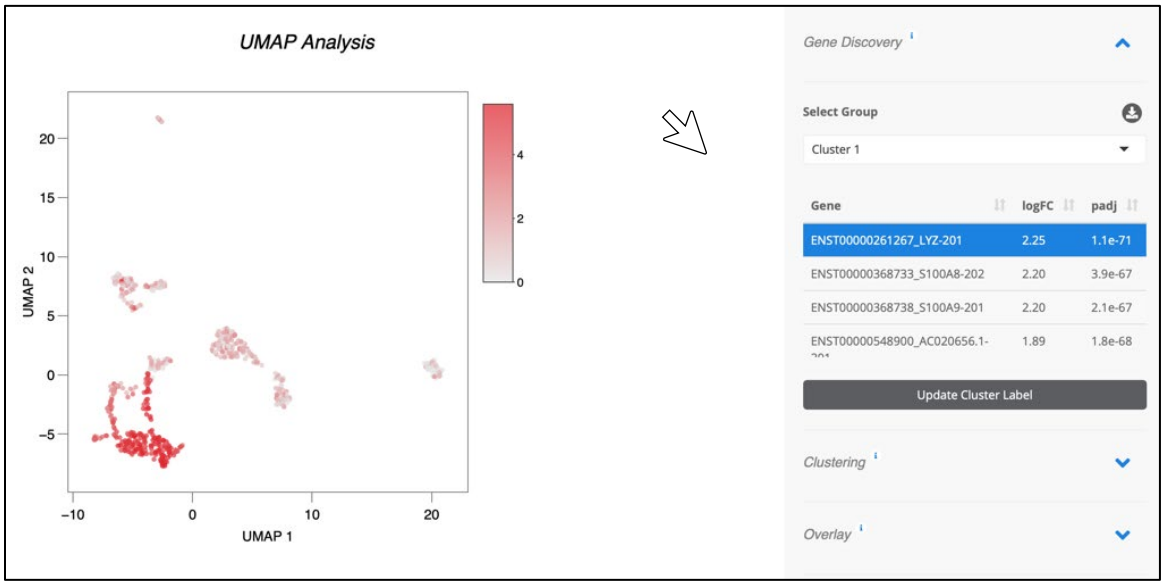


Figure 52. Selection of a given cluster in the Gene Discovery module. The sidebar will list the differentially expressed transcripts in that cluster. Selecting a transcript shows its expression.

8. Gene-fusion Overlay

When fusion analysis is performed in CogentAP (Cogent NGS Analysis Pipeline v2.0.1 User Manual, Section V.D.2), the results can be overlayed on the gene- or transcript-based clustering plot by using the **Overlay** module.

The top fusions that were detected during the analysis are displayed in a pull-down menu; select a fusion of interest and click the [Add Overlays] button to mark the cells in which the fusion was detected.

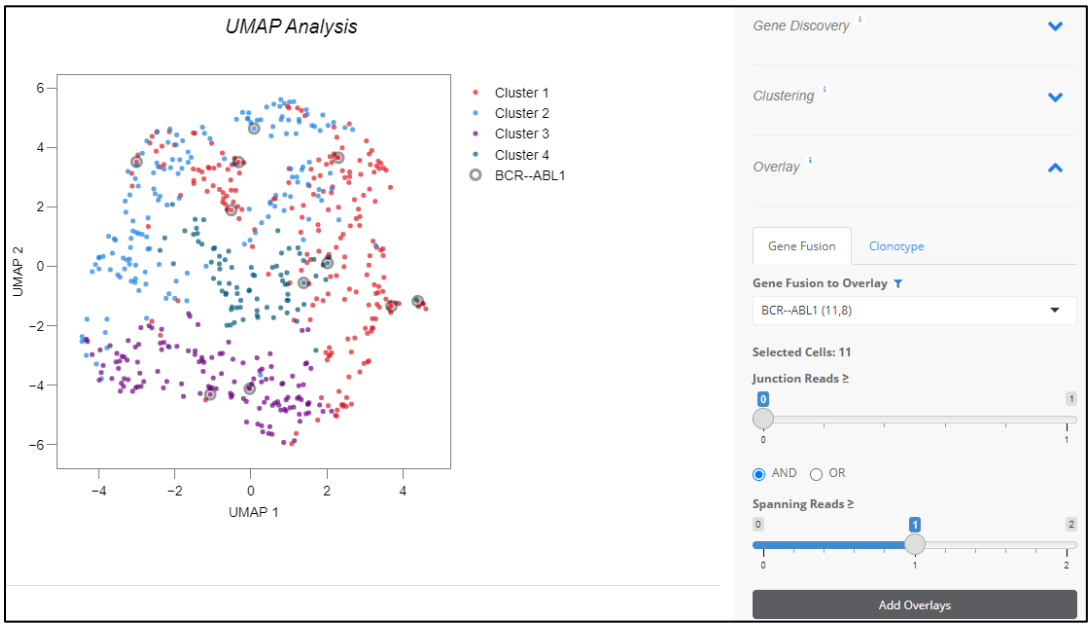


Figure 53. Overlay module for gene fusions. The **Overlay** menu option shown in the right sidebar can be opened to display a drop-down menu of gene-fusions that are detected in the analyzed dataset. An individual fusion can be selected to overlay on the cluster plot, depicted by gray circles drawn on the cells in which the selected fusion is detected. Multiple fusions can be overlaid simultaneously, if desired.

The number of fusions to display in the menu can be changed by clicking on the filter icon and changing the value in the pop-up window. The sliders can be used to filter the detection based on the minimum number of junction and spanning reads you decide are required to identify a fusion. A close-up view of this module is shown in Figure 54.

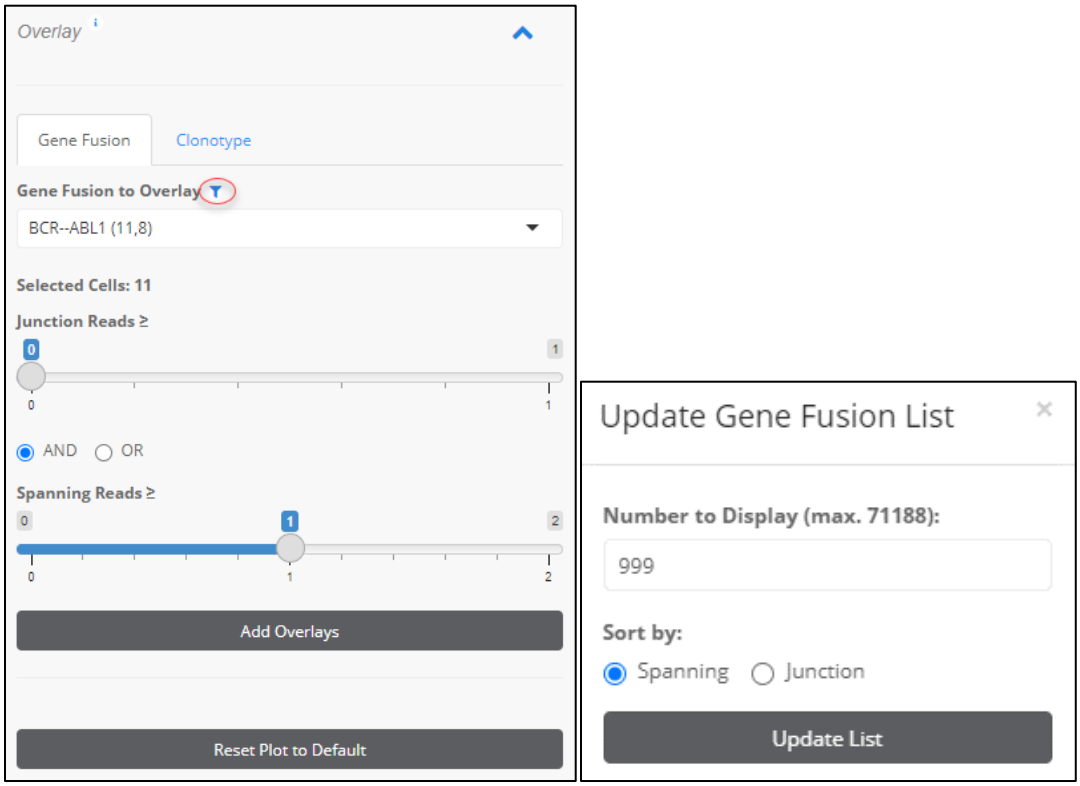


Figure 54. Close-up of the Overlay options for gene-fusion. **(Left, top):** The red circle highlights the filter icon for the gene fusion overlay; clicking on it opens the pop-up window **(Right)**, where you can choose how many fusions to display in the overlay and whether to sort on junction or spanning reads. **(Left, bottom):** The sliders can be used to set the number of reads required to define that a fusion is detected. Clicking the [Reset Plot to Default] button will clear all overlays and show the original cluster plot.

9. Immune Clonotype Overlay

When immune analysis is performed in CogentAP (Cogent NGS Analysis Pipeline v2.0.1 User Manual, Section V.D.3), the results can overlayed on the gene or transcript based clustering plot by using the **Overlay** module as shown in Figure 55.

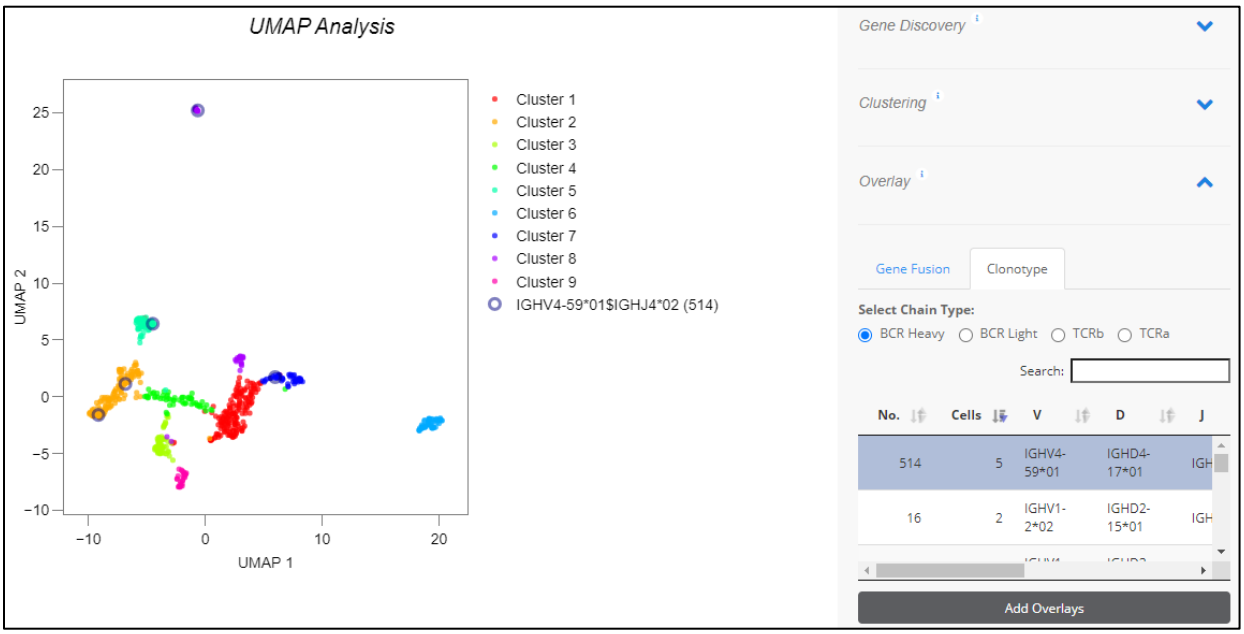


Figure 55. Overlay module for immune profiling. The **Overlay** menu option shown in the right sidebar can be opened to display a list of clonotypes detected in this dataset. After selecting the chain type of interest, an individual clonotype can be selected to overlay on the cluster plot, depicted by circles on the cells in which the selected clonotype is detected. Multiple clonotypes can be overlaid simultaneously, if desired.

The clonotypes identified from the immune analysis in CogentAP are listed under four categories:

- B-cell receptor (BCR) Heavy
- BCR Light
- T-cell receptor (TCR) alpha
- TCR beta

After selecting one of the categories, the clonotypes detected of that type are listed as a table with the following information:

- The number of cells in which the clonotype is detected
- V, D, and J genes and constant region (display based on the category type)
- CDR3 amino acid (CDR3aa) sequence information

Each of the columns can be sorted; for example, sort by the number of cells in which the clonotype is detected to study the top clonotypes. The "Search" box can be used to find a specific clonotype. A close-up view of this module is shown in Figure 56.

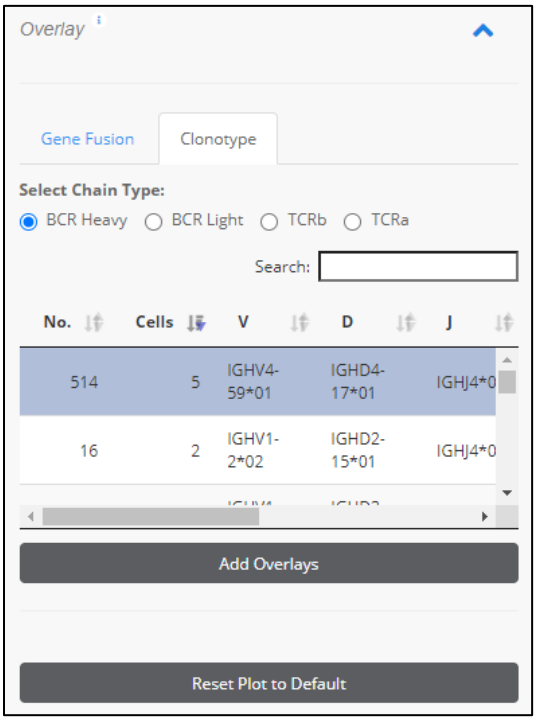


Figure 56. Close-up of the Overlay module for immune analysis. The four clonotype categories are shown at the top of the tab display, while the table of individual clonotypes in that category is displayed below them. Typically, the table is reverse sorted by the "Cells" column to show clonotypes that are detected in most cells. Clicking the [Reset Plot to Default] button will clear all overlays and show the original cluster plot.

F. Export the Data

After applying any data manipulations from Section V.D or V.E, it can be saved in its edited form.

1. Expand the **File Management** option in the left sidebar menu and click the [Download] button.

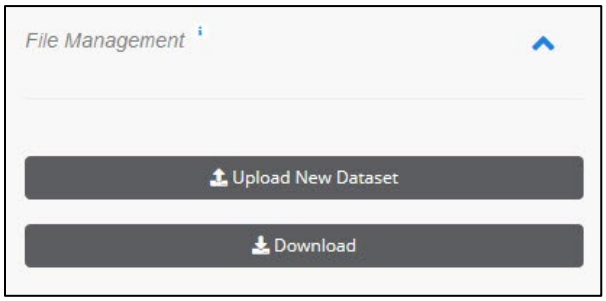


Figure 57. The File Management submenu.

2. Choose the type of file you wish to download from the dropdown menu. Available options are:
   - High-resolution UMAP/t-SNE plot
   - Updated stats/metadata CSV file
   - Quality-controlled log-normalized gene matrix
   - Comprehensive HTML report of the analysis. This HTML report is similar to the one generated from the Cogent NGS Analysis Pipeline but is customized to any new specifications made in CogentDS.

- CogentDS rda file to save changes to the analysis as a new rda file. If full-length data was initially imported into CogentDS, the new rda file will also save fusions or clonotypes that are overlayed on the clustering (Section V.E). The rda file saved here can be reloaded into CogentDS at a later time to continue analysis.

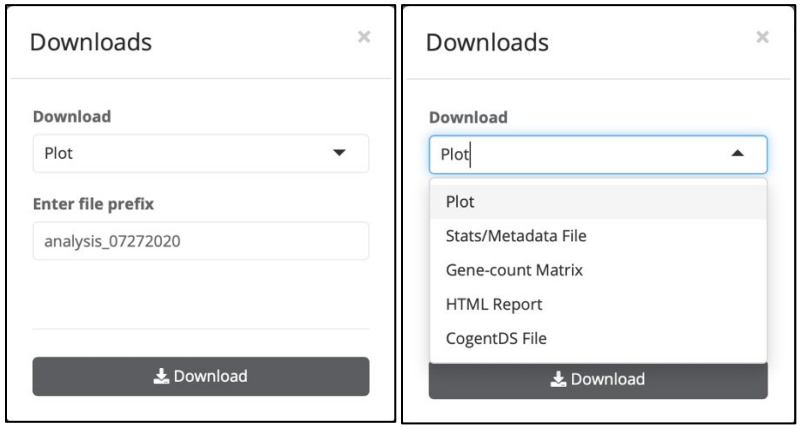


Figure 58. The *Downloads* pop-up menu and dropdown list of available file types.

3. If desired, enter a custom file prefix.
4. If downloading an HTML report, two additional fields are available (Figure 59):

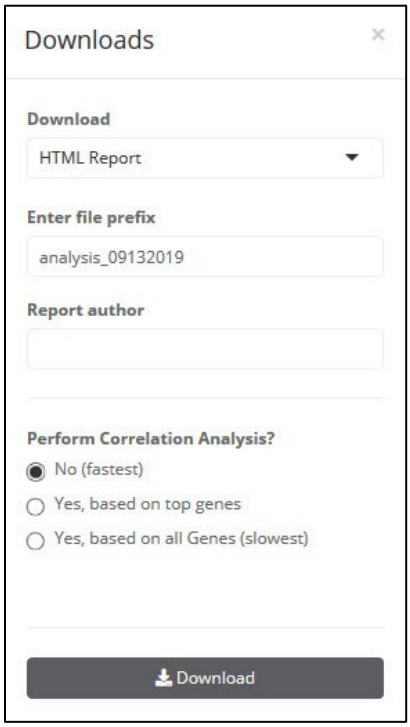


Figure 59. The *Downloads* pop-up menu for HTML reports.

- Enter a report author (Optional): This will add the text input into this field to the top of the HTML display (Figure 60).

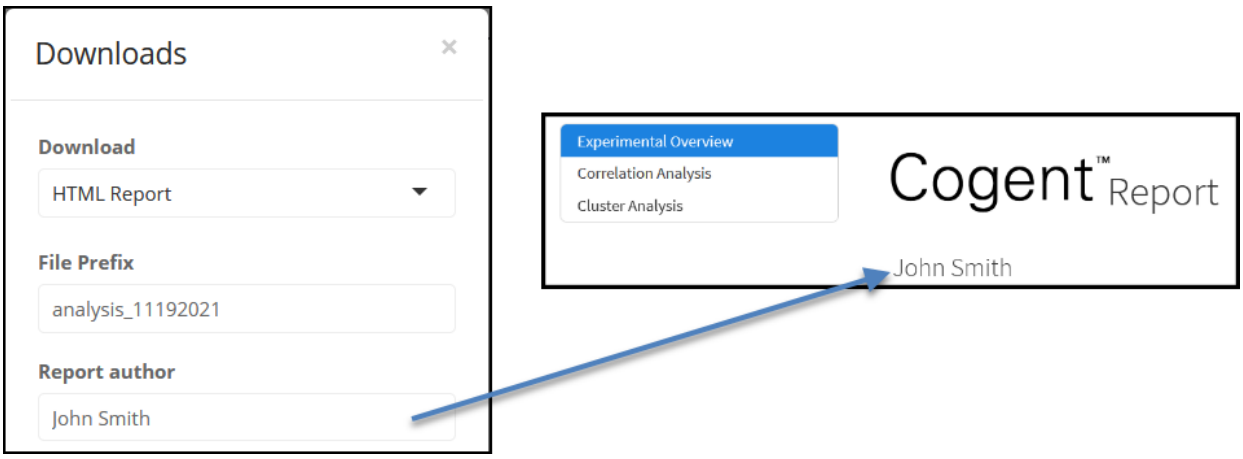


Figure 60. The effect of the "Report author" field on HTML reports.

- Change selection in order to perform correlation analysis. (Optional)
  - No (fastest): **(DEFAULT)** This will skip the correlation analysis and is the recommended selection. This selection does not affect the rest of the analysis steps.

NOTE: Correlation analysis can take a significant amount of time. Please be aware of this if selecting either of the following 'Yes' options.

  - Yes, based on top genes: Perform the analysis on only the top genes that are used for clustering.
  - Yes, based on all Genes (slowest): Perform the analysis using all the genes in the input data.

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.

© 2023 Takara Bio Inc. All Rights Reserved.

All trademarks are the property of Takara Bio Inc. or its affiliate(s) in the U.S. and/or other countries or their respective owners. Certain trademarks may not be registered in all jurisdictions. Additional product, intellectual property, and restricted use information is available at takarabio.com.

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