omnibusx
FROM DATA TO DISCOVERY

Single-cell analysis

User guide for OmnibusX application

omnibusx.com

Table of contents

Introduction 8

Installation 9

1. System requirements 9

1.1. Operating systems 9

1.2. Network 9

1.3. Hard drive 9

1.4. Memory 9

1.5. Notice 10

2. Software installation 10

1.1. Windows 10

1.2. Ubuntu/Debian 10

1.3. MacOS 10

Importing data 12

1. Genome reference 12

2. Data format 13

2.1. Scanpy object 14

2.2. Seurat object 15

2.3. HDF5 output from Cell Ranger (10X) 17

2.4. MTX files (10X) 18

2.5. TSV/CSV/TXT matrix 19

3. Pre-processing pipeline 20

3.1. Data conversion 20

3.2. Gene annotation 21

3.3. Quality control metrics 22

4. Processing pipeline 23

4.1. Filtering cells 23

4.2. Doublet detection 24

4.3. Demultiplexing with HTOs 26

4.4. Normalization 27

4.4. Initial analysis 28

Data exploration 30

1. Visualization tools 31

1.1. Zoom 31

1.2. Navigate 32

1.3. Reset 32

1.4. Settings 33

1.5. Export 33

2. Expression tools 34

2.1. Selecting omics 34

2.2. Query genes by name 35

2.3. Color modes for expression visualization 37

2.4. Gene sets 41

2.5. Gene set enrichment 43

2.6. Finding markers 44

3. Computational methods 47

3.1. Clustering 47

3.2. t-SNE/UMAP 48

3.3. Cell type prediction 50

4. Annotation management 51

4.1. Find a metadata 51

4.2. Categorical metadata 53

4.3. Numerical metadata 55

4.4. Combine metadata 56

4.5. Metadata group 56

4.6. Edit a metadata 58

4.7. Restore deleted metadata 60

5. Sub-cluster management 61

5.1. Create sub-clusters 61

5.2. Multiple levels of sub-clustering 62

5.3. Remove all sub-clusters 63

6. Data management 64

6.1. Import additional data 64

6.2. Scatter management 65

6.3. Embedding management 68

6.4. Quality control 72

6.5. Default color palette 73

6.6. Update dataset information 74

Cell annotation 76

1. Cell selection 76

1.1. Select with existed metadata 76

1.2. Select with expression values 77

1.3. Manually draw to select 80

2. Manual annotation 82

3. Programmatic annotation 84

Downstream analysis 86

1. Composition analysis 87

2. Basic charts 92

2.1. Violin plot 96

2.2. Box plot 99

2.3. Bar plot 102

2.4. Bubble plot 105

3. Differential expression genes (DEGs) analysis 109

3.1. Select groups for comparison 110

3.2. Automatic group selection 112

3.3. Parameters for DEGs analysis 114

3.4. Manage DEGs results 116

3.5. Volcano plot 118

3.6. DEGs table 121

3.7. Differential enrichment analysis 125

3.8. Heatmap 130

4. Enrichment analysis 134

4.1. Create new enrichment analysis 134

4.2. Manage enrichment results 136

4.3. Analyze enrichment result 137

4.4. Meant plot 139

4.5. Scatter 141

4.6. Heatmap 142

5. Heatmap analysis 144

5.1. Create new heatmap analysis 145

5.2. Manage heatmap results 147

5.3. Analyze heatmap result 149

Dataset management 154

2. Dataset folders 155

3. Delete and restore 157

Your libraries 160

1. Gene set library 160

1.1. Group management 161

1.2. Gene set management 163

1.3. Import and export 166

2. Color palette library 167

2.1. Create new color palette 168

2.2. Default color palette 169

Task management 170

1. Tasks 170

2. Resource monitor 174

Thank you! 175

Introduction

In the world of biological data science, navigating vast and complex datasets can be a daunting task. Our software toolkit is designed to empower you by accelerating analysis, promoting accessibility, and providing comprehensive visualization tools to help you make sense of these intricate biological datasets.

Our platform is built with data accessibility at its core. With just a few clicks, you can gain seamless access to extensive, high-dimensional biological datasets, including bulk RNA-seq, scRNA-seq, CITE-seq, ATAC-seq, Visium, GeoMx, and more. This allows you to focus less on data wrangling and more on analysis and discovery.

Our integrated standard processing pipeline includes the latest tools for standard data processing. From normalization and batch correction to PCA, t-SNE, UMAP, clustering, and beyond, our pipeline is designed to streamline your workflow and provide you with reliable, reproducible results.

But our toolkit doesn't stop at data processing. We also offer comprehensive analysis and visualization capabilities. Whether you need to query genes, draw basic charts, create complex heatmaps, find marker genes, perform differential expression genes, gene set enrichment analysis, pseudotime, AUCell, cell type prediction, or more, our platform has you covered.

In this user guide, we'll walk you through each step of using our Single-Cell Application, from accessing your data to performing complex analyses and visualizing your results. We're excited to help you unlock the potential of your biological data!

Installation

1. System requirements

1.1. Operating systems

- Windows 11 (64-bit only).
- macOS: macOS Ventura (version 13) or higher.
- Ubuntu 20.04 or higher.
- Debian 10 or higher.

1.2. Network

- Ethernet connection (LAN) or a wireless adapter (Wi-Fi).

1.3. Hard drive

- Minimum 10 GB.
- Recommend 50 GB or above.

1.4. Memory

- Minimum 8 GB.
- Recommend 16 GB or above.

1.5. Notice

For an optimal experience with large data containing more than 50,000 cells/samples, please use a computer with the recommended configuration.

If you experience difficulty running the software after fulfilling all these requirements, please contact support@omnibusx.com.

2. Software installation

1.1. Windows

- Download the installer .exe file on our website and run it. The installation process will start automatically.
- Administrator privilege is not required.
- If your computer has more than one account using the software, each account can only access its data.

1.2. Ubuntu/Debian

- Download the installer .deb on our website.
- Make a folder to store the extracted files.
- Install the application using the command below.
- Run the application using the command below.

```
mkdir <path-to-store-extracted-files>
dpkg -x omnibusx_amd64.deb <path-to-store-extracted-files>
./<path-to-store-extracted-files>/usr/bin/omnibusx
```

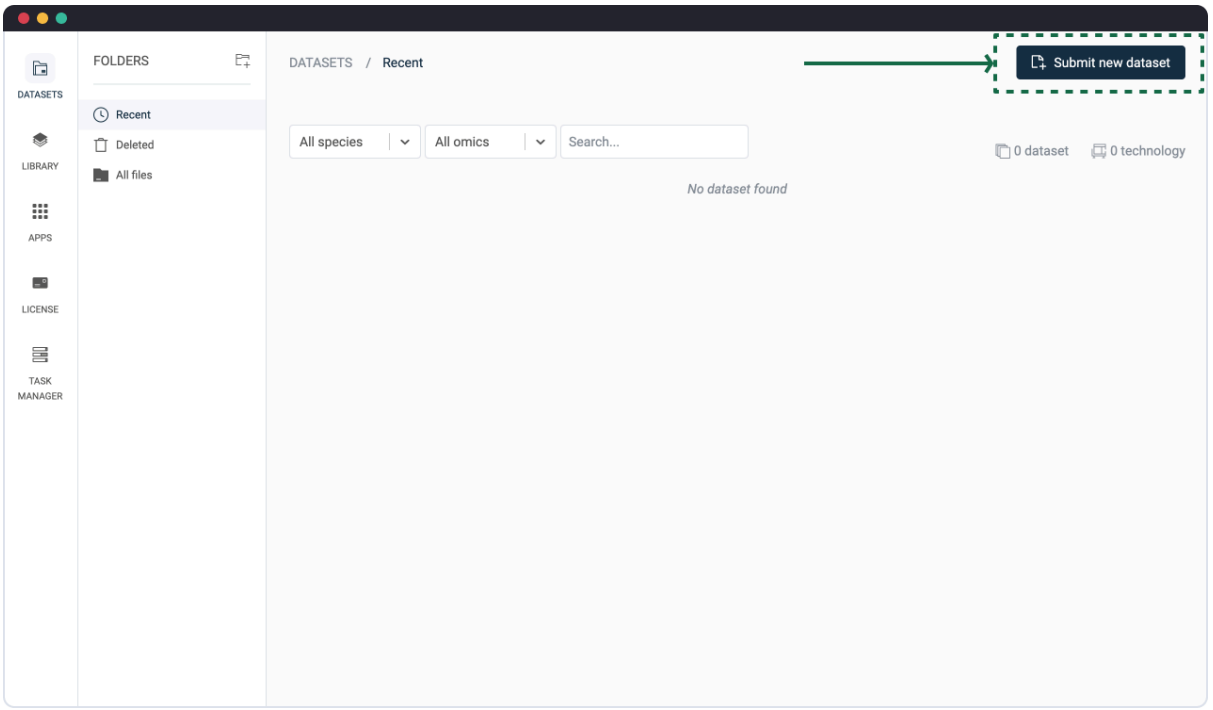
1.3. MacOS

- Download OmnibusX for macOS.
- Open the browser's download list and locate the downloaded archive.

- Extract the archive contents. Use double-click for some browsers or select the 'magnifying glass' icon with Safari.
- Drag OmnibusX.app to the Applications folder, making it available in the macOS Launchpad.
- Open OmnibusX from the Applications folder, by double clicking the icon.
- Add OmnibusX to your Dock by right-clicking on the icon, located in the Dock, to bring up the context menu and choosing Options, Keep in Dock.

Importing data

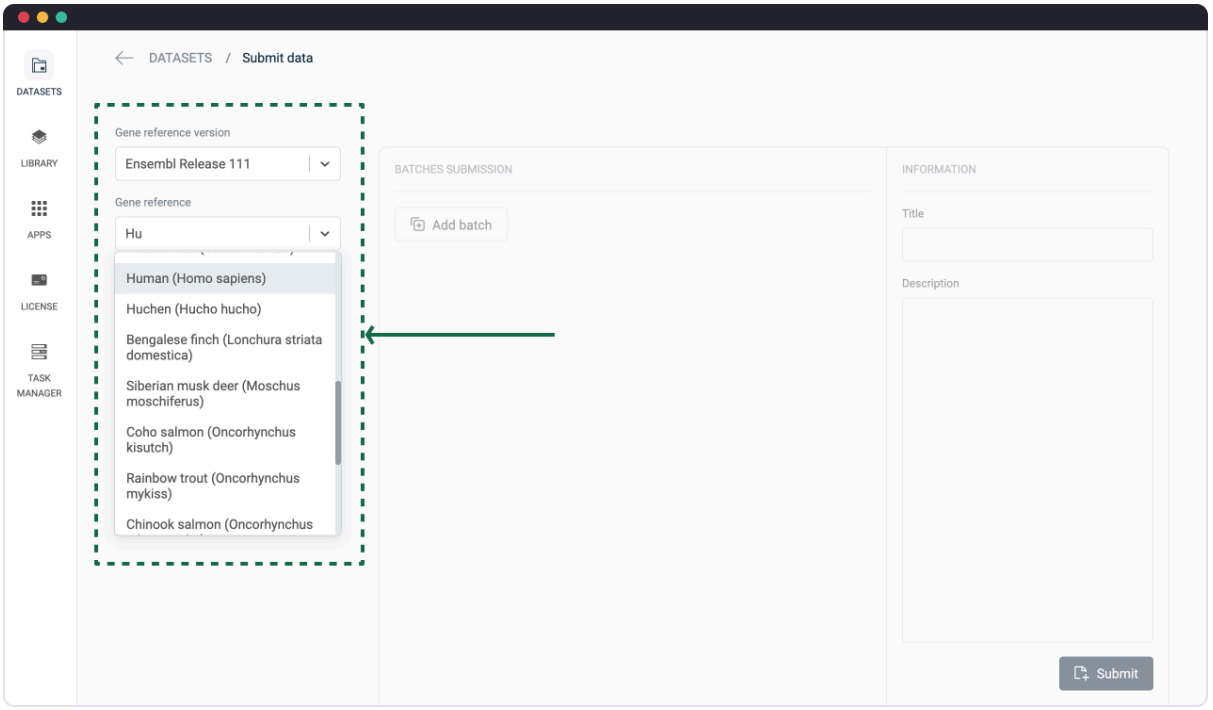
To start importing your data into OmnibusX, click on the `Submit new dataset` button located at the top right corner of the page. Currently, OmnibusX accepts expression matrices as input, which stores the expression values processed from alignment results.



1. Genome reference

Single-cell data can be processed using various alignment tools and different gene annotation versions. Therefore, OmnibusX incorporates gene nomenclature conversion as a crucial first step. This step ensures that gene IDs and gene aliases are converted to currently accepted gene names, curated by Ensembl.

OmnibusX has downloaded and versioned indexes for all available species from Ensembl. You simply need to select the corresponding species that matches your input data and the appropriate version to ensure reproducibility.



2. Data format

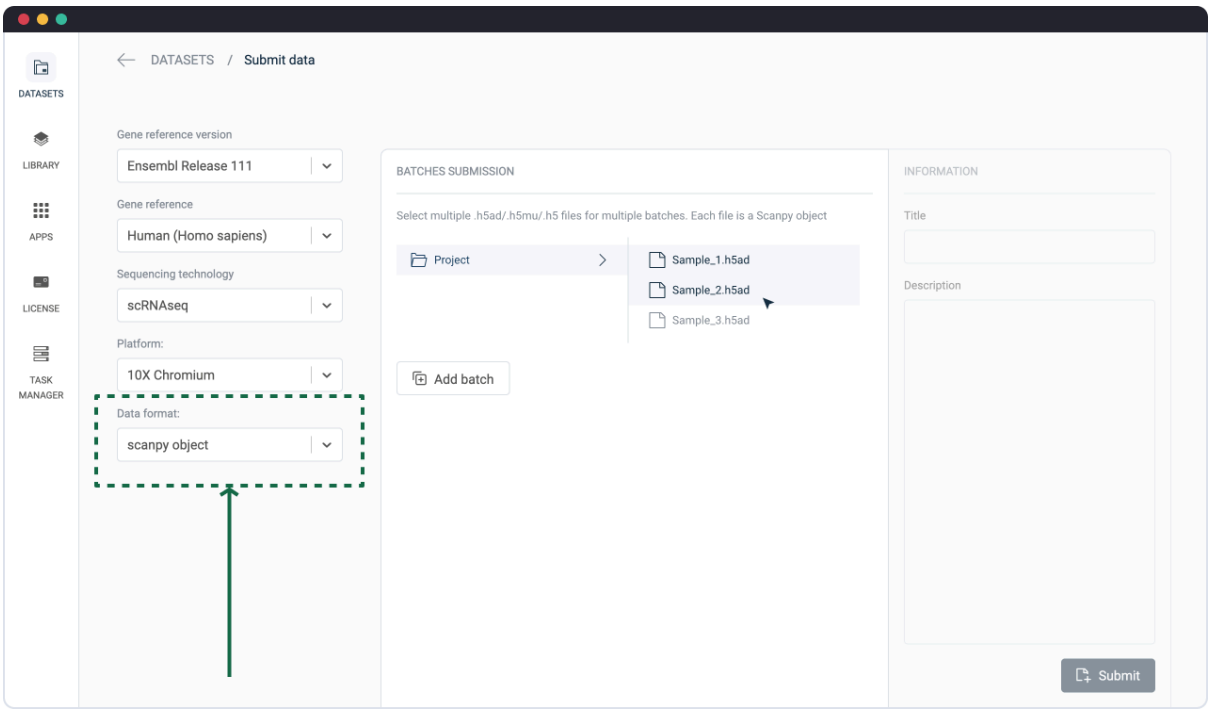
OmnibusX supports a variety of data formats to accommodate the diverse nature of publicly available datasets, which come from different laboratories and use various pipelines for data processing. The supported formats include:

- `scanpy object`: A widely-used format for single-cell RNA sequencing data analysis in Python.
- `Seurat object`: A popular format for single-cell RNA sequencing data analysis in R.
- `HDF5`: A format used by 10X Genomics for storing single-cell gene expression data.
- `MTX sparse matrix:` A Matrix Market Exchange format used for sparse matrices.

- `TSV/CSV/TXT full matrix`: Standard text file formats for storing full matrices.

Handling these different data formats typically requires significant effort to customize your processing pipeline and substantial time to learn the specifics of each format. OmnibusX simplifies this process by providing a unified platform that can handle various data formats. This allows you to focus more on data analysis and exploration, rather than spending time learning how to process different data formats.

2.1. Scanpy object



You can provide a single scanpy object that contains at least an expression matrix (raw or normalized) saved as `adata.X` or `adata.raw.X` or both. For single scanpy object submission, all available analysis results will be imported, including:

- `adata.raw.X` as raw count.
- `adata.X` as normalized data (if adata.raw.X is absent, adata.X will be treated as raw count).
- `adata.obs` as metadata.

- `adata.obsm` as dimensional reduction results.

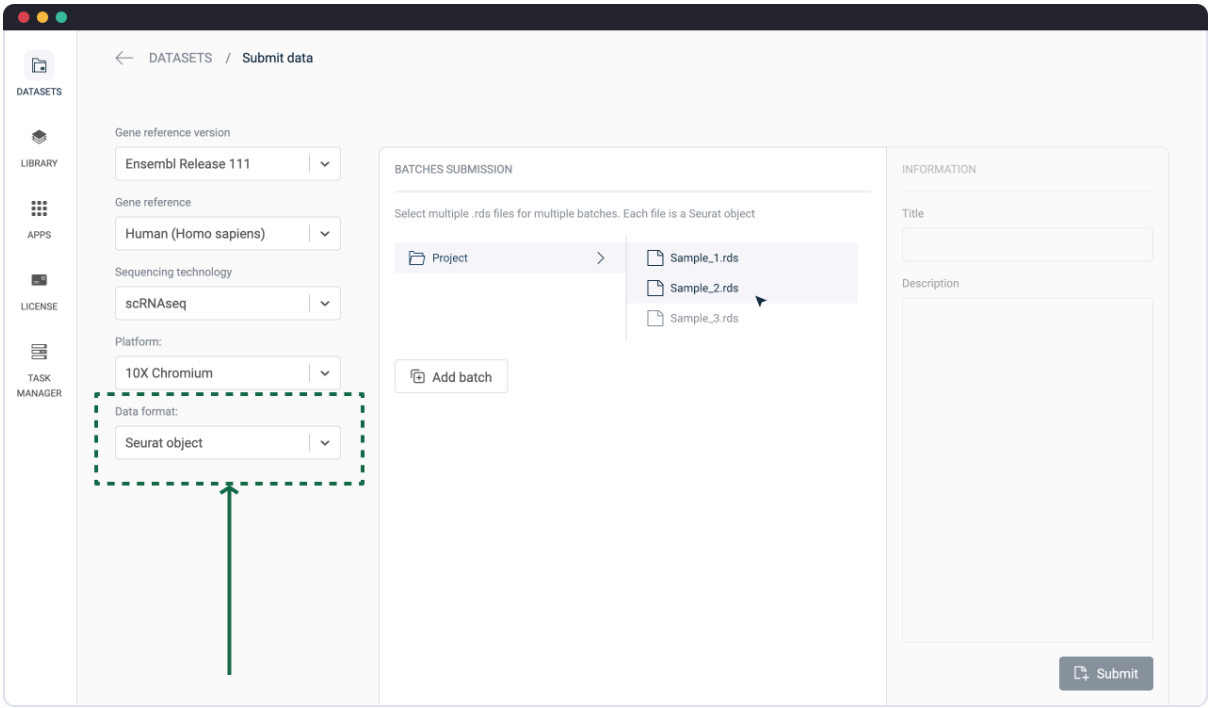
If all necessary analysis results are available in the object, you can skip the processing pipeline in later steps.

You can also provide multiple scanpy objects in a single submission. In this case, each object will be treated as an experiment batch. All processed results from each object (such as PCA, t-SNE, UMAP, etc.) will be ignored, as the data needs to be combined and re-processed. The prioritization for expression data is as follows:

- If `adata.raw.X` is available, it will be used, and `adata.X` will be ignored.
- If `adata.raw.X` is absent, `adata.X` will be used and treated as raw count. If adata.X is already normalized and you cannot access the raw count data, you can choose to skip normalization in a later step.

The metadata of all batches are also merged and imported.

2.2. Seurat object



You can provide a single Seurat object, which must contain the `RNA` assay. For single Seurat object submission, all available assays and analysis results will be imported, including:

- `assay.count` as raw count.
- `assay.data` as normalized data.
- `meta.data` as metadata.
- `reductions` as dimensional reduction results.

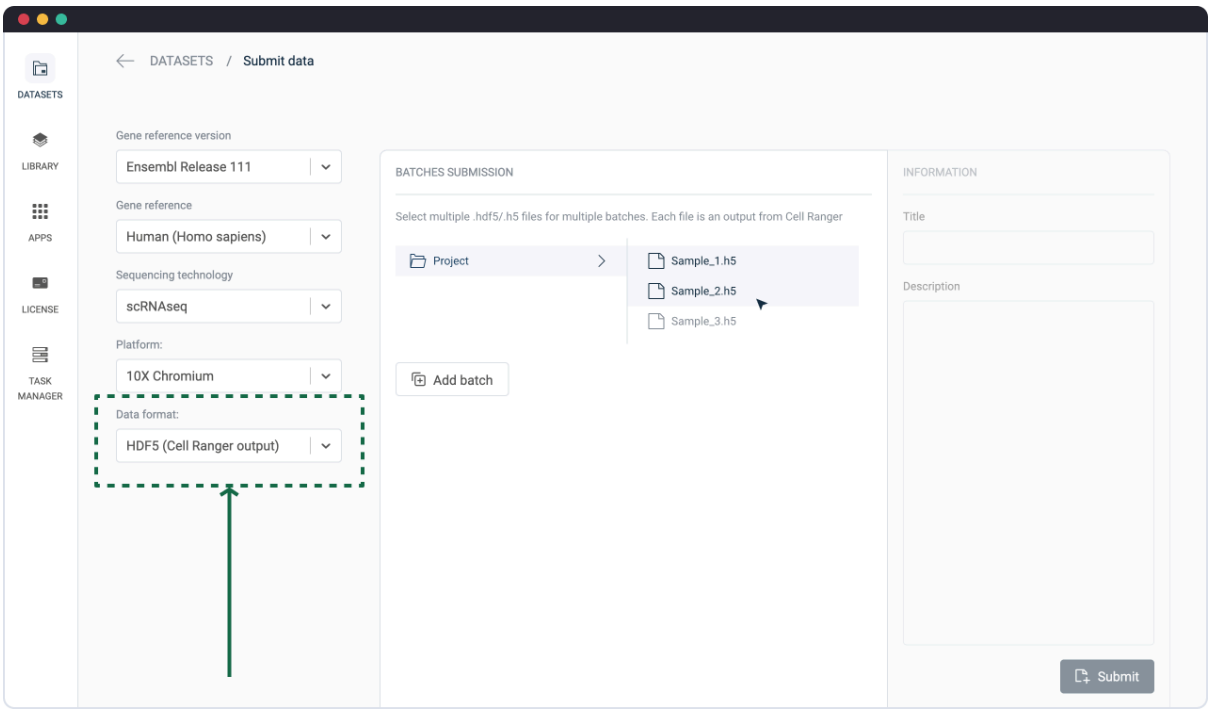
If all necessary analysis results are available in the object, you can skip the processing pipeline in later steps.

You can also provide multiple Seurat objects in a single submission. In this case, each object will be treated as an experiment batch. All processed results from each object (such as PCA, t-SNE, UMAP, etc.) will be ignored, as the data needs to be combined and re-processed. The prioritization for expression data is as follows:

- If `assay.count` is available, it will be used, and `assay.data` will be ignored.
- If `assay.count` is absent, `assay.data` will be used. In this case, we recommend you to skip normalization in a later step to avoid double normalization.

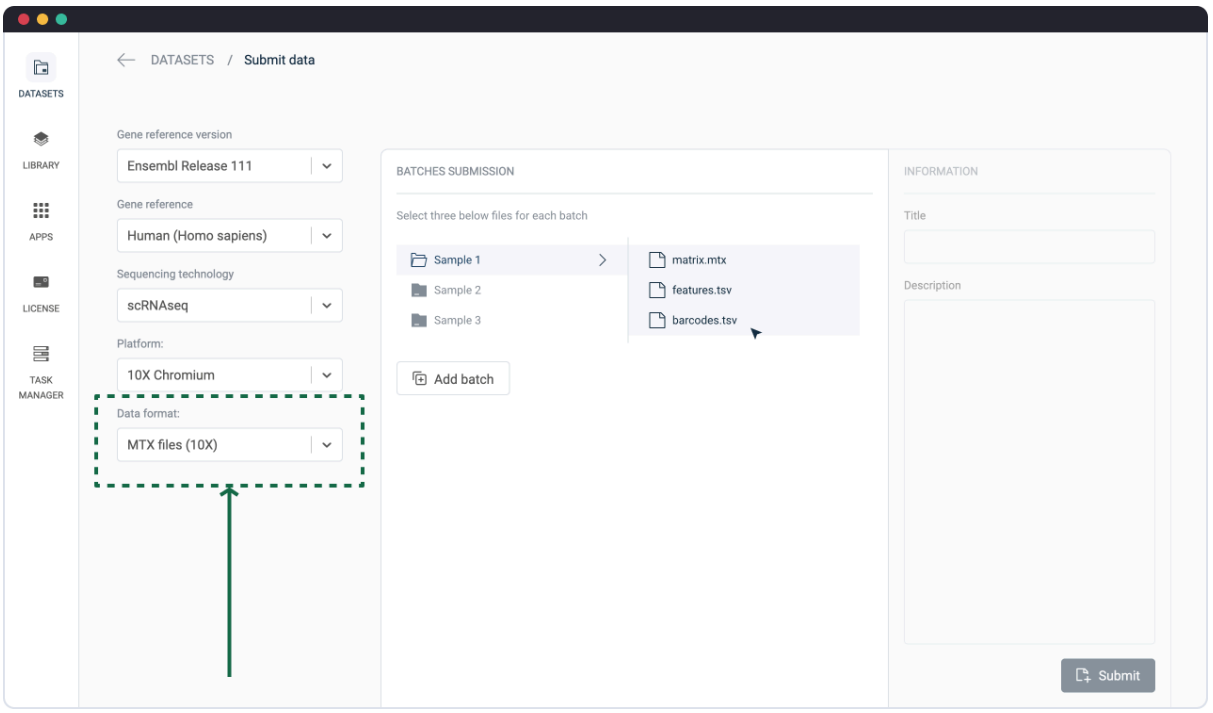
The metadata of all batches will be merged and imported. Since Seurat objects can contain multi-modal data, we will retain assay keys that are shared between all batches and process each assay independently.

2.3. HDF5 output from Cell Ranger (10X)



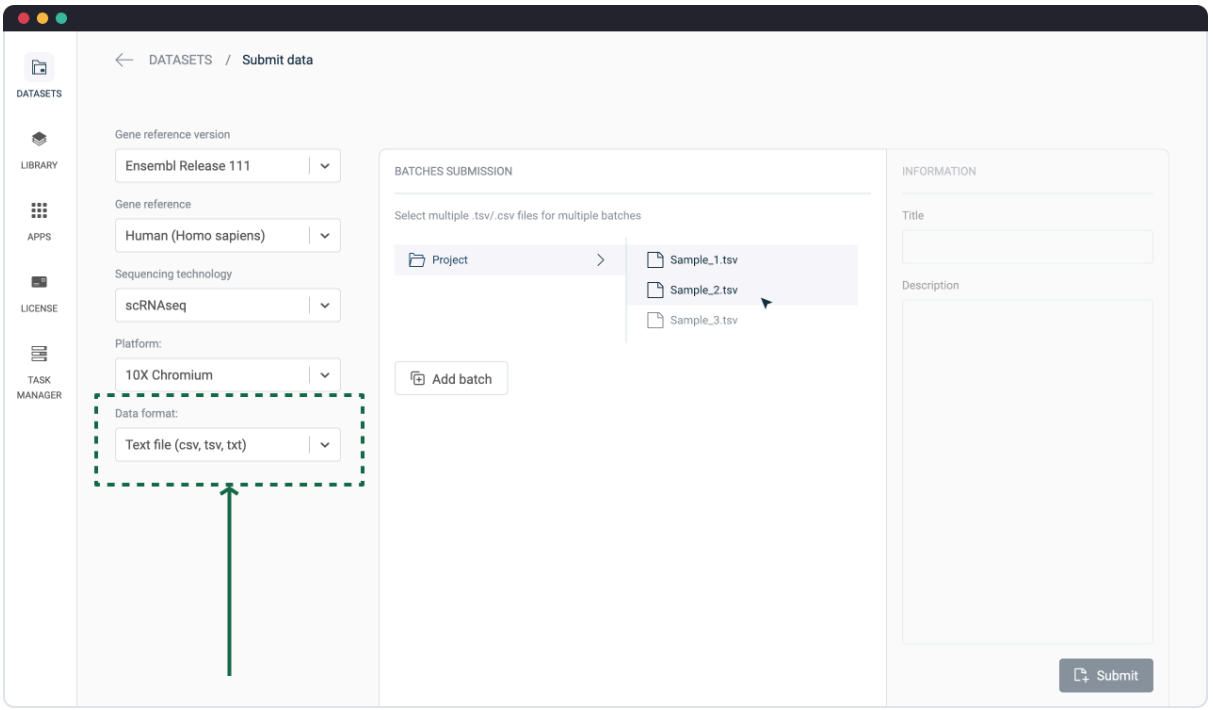
You can submit one or multiple HDF5 files generated by Cell Ranger, each corresponding to an experiment batch. The detailed file structure for these HDF5 files is described by 10X Genomics here. OmnibusX will automatically process and import the data from these files.

2.4. MTX files (10X)



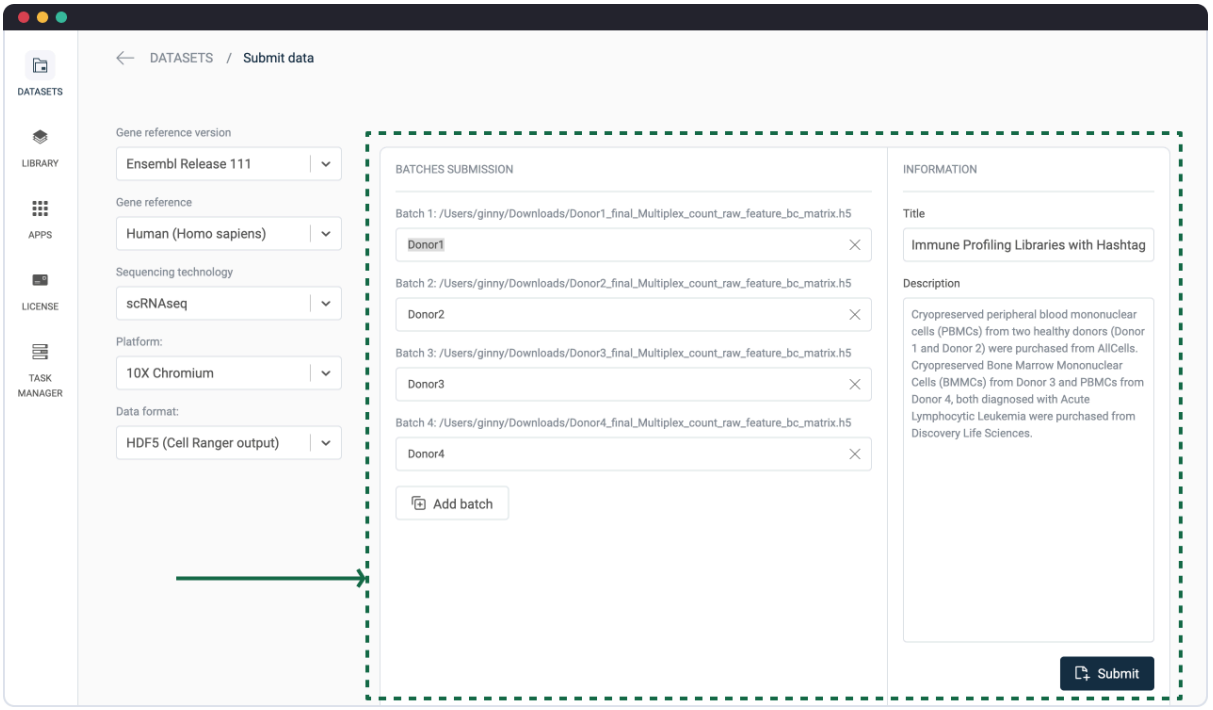
MTX files, generated by Cell Ranger, require three associated files for each sample: `matrix.mtx`, `barcodes.tsv`, and `genes.tsv`/`features.tsv`. The detailed file structure is described by 10X Genomics here. When submitting MTX files, you need to select the three files for each batch at a time.

2.5. TSV/CSV/TXT matrix



You can submit full matrices in TSV, CSV, or TXT format. The matrix should have headers for column names, with the first column representing row names. The matrix can be formatted as either `cells x genes` or `genes x cells`. OmnibusX will detect the genes according to your selected reference. You can input one or multiple files, each corresponding to an experimental batch.

3. Pre-processing pipeline



After selecting files from your local directory, you will need to provide submission information such as title, description, and batch names. Then, click on the `Submit` button located at the bottom right corner to start the pre-processing pipeline. This pipeline handles data conversion, gene annotation, and data quality control.

3.1. Data conversion

OmnibusX automatically detects the expression matrix, metadata, and other processed results (PCA, t-SNE, UMAP, etc.) based on your input file format. The expression matrix is then split into distinct omics as follows:

- `Seurat object`: Each input assay corresponds to an omic. Based on the omic name, the `RNA` assay is assigned to transcriptomics, `ADT` to proteomics, and `HTO` to group hashing. These omics will follow specific processing pipelines later. Other omics are imported as they are.
- `HDF5` and `MTX`: Omic types are split based on the feature type provided by 10X Genomics: `Gene Expression` as transcriptomics, and `Antibody`

`Capture` as proteomics. The proteomics data is further split into proteomics and group hashtags based on the protein name (whether it contains the prefix `hashtag` or not).

- `scanpy object`, `TSV/CSV/TXT`: As these formats do not specifically support multi-modal data, OmnibusX auto-detects based on the genes/features names. If the name contains the prefix `adt`, it is split as proteomics; if it contains `hto` or `hashtag`, it is split as a hashtag. The remaining data is considered transcriptomics.

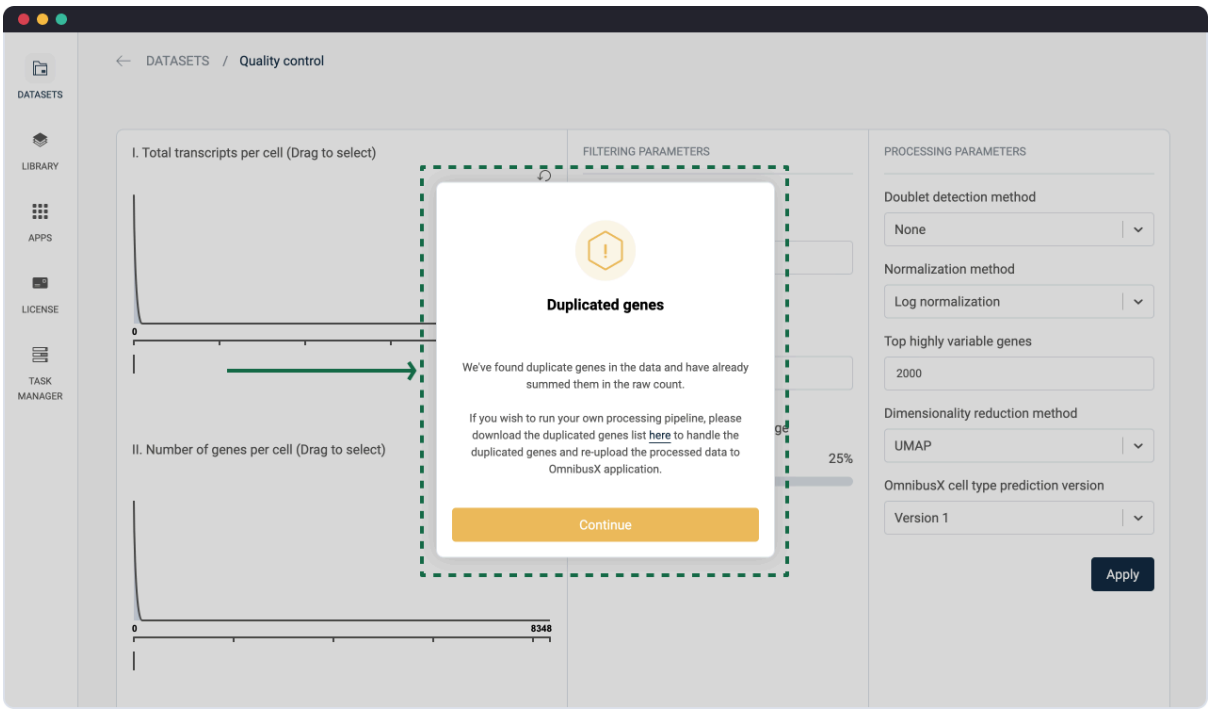
When multiple batches are input, OmnibusX prioritizes selecting the raw count matrix and metadata, while ignoring all other results. If raw count data is absent, normalized data is used as raw. OmnibusX then merges multiple batches into one. To avoid duplicate barcodes between batches, OmnibusX adds the batch name (input in the previous step) as the barcode prefix for each batch.

3.2. Gene annotation

The transcriptomic matrix is then processed with your selected reference. Gene annotations from your input are converted into gene names using Ensembl annotation. Mapped genes are retained as transcriptomics, while unmapped ones, which may include other numeric annotations (number of genes, group IDs, etc.), are split into another omic named `Others`.

Data with duplicated gene names, which occurs when multiple gene IDs are mapped to the same gene name or multiple gene aliases from different alignment references, will be further processed as follows:

- If raw count data is detected from the input, OmnibusX will sum the raw counts of duplicate genes and renormalize the data. The normalized matrix, if detected, will be ignored as it leads to different results.
- In the case of missing raw count data, OmnibusX sums the normalized values to continue processing, but it is recommended to proceed only if raw count data is inaccessible.



This approach is based on the rationale that the same gene names encode for the same product, regardless of their location on the chromosome. Therefore, the final production is contributed by the total transcript of all duplicate genes. By summing the raw counts, we can obtain a more accurate representation of gene expression levels.

On the other hand, keeping duplicate genes separate can lead to confusing situations during data analysis. For example, if a researcher queries for gene A but the alignment tool used a different reference and named all of gene A transcripts as gene B, the researcher might conclude that gene A is not expressed. However, this would be a false conclusion, as the transcripts for gene A were simply named differently. By summing the raw counts of duplicate genes, we can avoid such confusion and ensure that researchers obtain accurate results when querying for specific genes.

3.3. Quality control metrics

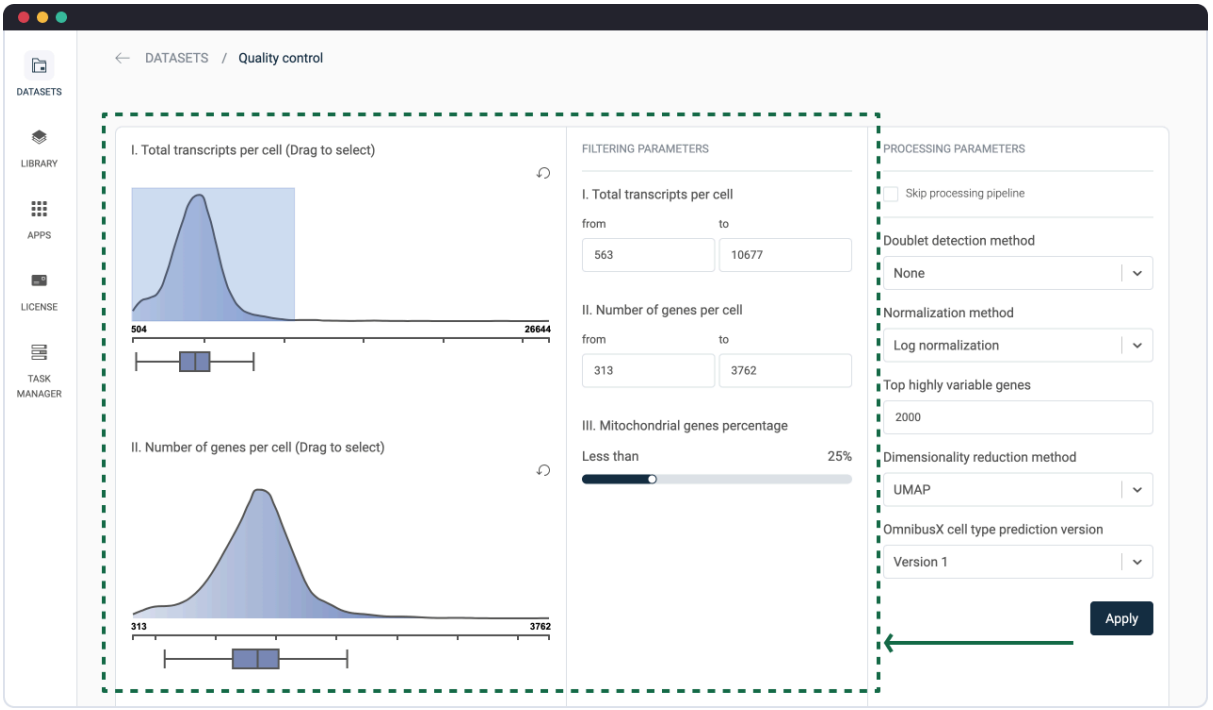
OmnibusX then calculates necessary data for quality control steps, including the number of transcripts per cell, the number of genes (mapped to the selected

reference) per cell, and the percentage of mitochondrial genes. Mitochondrial genes are detected using Ensembl annotation, mapped to the MT chromosome in the annotation file.

4. Processing pipeline

The OmnibusX processing pipeline encompasses doublet detection, filtering, normalization, dimensional reduction, and cell type prediction. The primary objective is to filter out low-quality cells, ensuring accurate and reliable downstream analysis. This step can be repeated multiple times with different sets of filtering parameters on the pre-processed results until the desired data quality is achieved. Additionally, other analyses, such as dimensional reduction and cell type prediction, can be easily updated and customized in later stages of downstream analysis.

4.1. Filtering cells



Low-quality cells or dead cells are typically characterized by a low number of detected transcripts and genes, which also results in an abnormally high fraction of mitochondrial gene expression. However, since the number of detected genes/transcripts varies depending on the technology used, there is no fixed threshold for this filtering step. Therefore, OmnibusX visualizes the distribution of transcript counts and gene counts per cell, allowing you to make informed decisions.

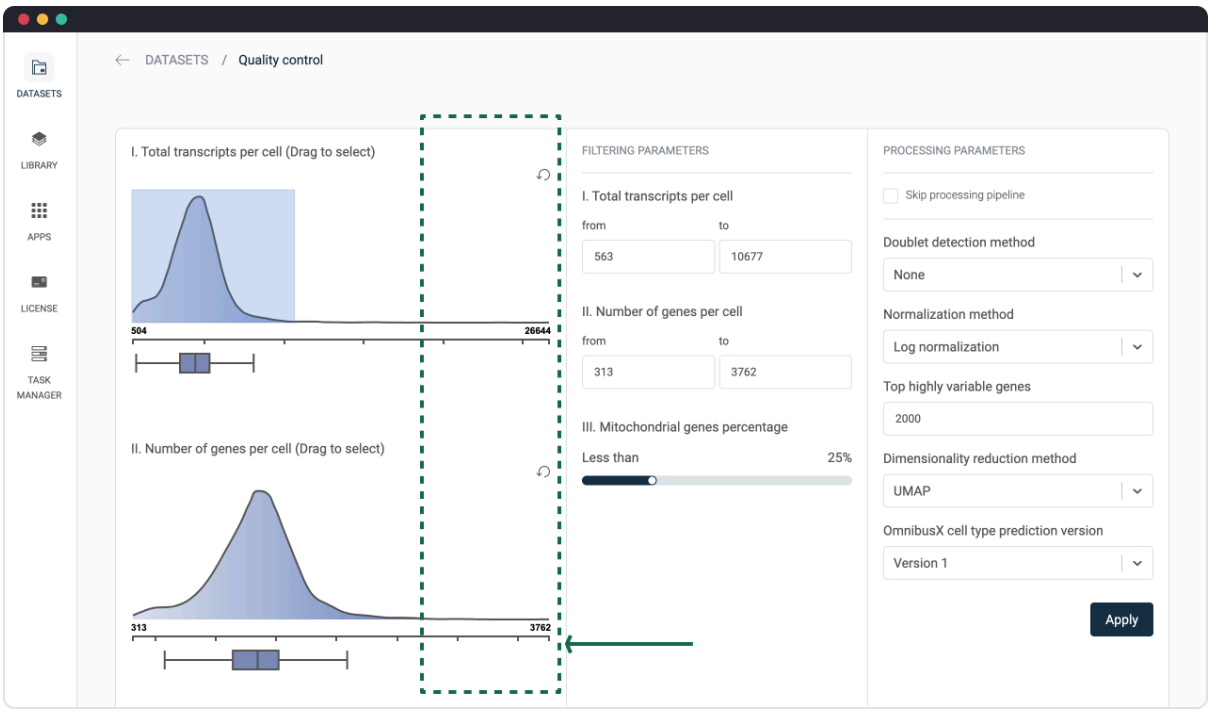
You can drag to select the region of interest from the visualization to filter the cells. Additionally, you can refine the filtering criteria by manually editing the filtering numbers for reproducibility.

Cells with an abnormally high number of transcripts or genes may also indicate the presence of doublets, where multiple cells are counted as one. These should also be filtered out to ensure the integrity of the data.

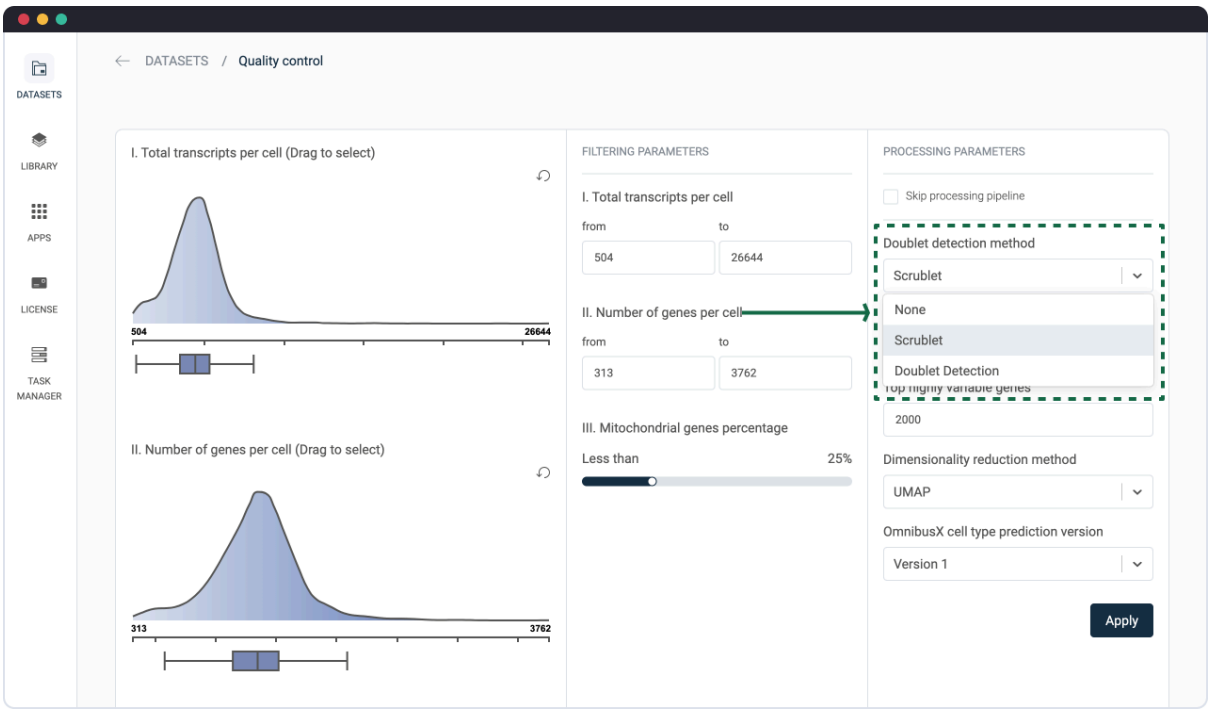
4.2. Doublet detection

Doublets can be identified and removed using multiple approaches:

- High Transcript/Gene Counts: Doublets may exhibit abnormally high numbers of transcripts or genes compared to the rest of the population. These can be filtered out by manually adjusting the filtering thresholds as described in the previous section.



- Computational Methods: OmnibusX integrates commonly used methods for doublet detection:
    - `Scrublet` (*Wolock et al., 2019*): Using the implement from the Scanpy team: scanpy.pp.scrublet.
    - `Doublet detection` (*Gayoso et al., 2020*): Using the implement from their GitHub repository.



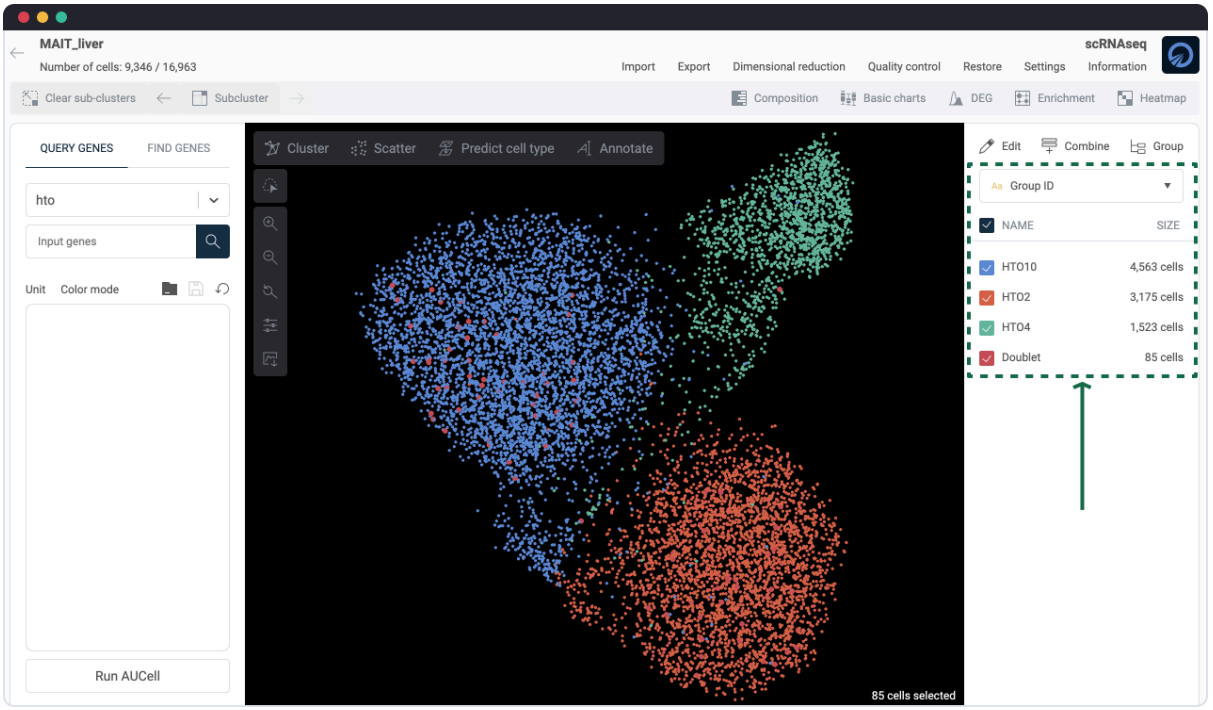
- Demultiplexing with Hashtag Oligos (HTOs): If HTO information is available from the input data, cells with more than one HTO tag after the demultiplexing process will be annotated as doublets. Further details on this process are described in the section on HTOs.

4.3. Demultiplexing with HTOs

If HTO data is provided, OmnibusX will automatically apply a demultiplexing process to assign cells to their respective experimental groups. The resulting group IDs and doublet types will be added as new metadata fields called `Group ID` and `Doublet Type` under the metadata group `OmnibusX Demultiplexing`. This process follows the method developed by the Seurat team (detailed here). In summary, the demultiplexing process involves the following steps:

- `HTO filtering`: Identifies HTOs expressed in at least 10 cells to ensure relevance.
- `K-Means Clustering`: Groups cells based on their HTO expression patterns.
- `Threshold determination`: Sets thresholds to classify cells as positive or negative for each HTO based on the expression value of the cluster with the lowest average expression.

- `Label assignment`: Classifies cells and identifies doublets (cells positive for more than one HTO). Cells with no HTO assignment are labeled as negative.



4.4. Normalization

For transcriptomics, the normalization process includes two steps:

- `Scale by total counts`: For each cell, the total gene count is scaled to 10,000 using the formula: `(gene_count / total_count) * 10,000`. This step uses the Scanpy team's method: scanpy.pp.normalize_total.
- `Logarithmize the data matrix` (log1p): This step applies log normalization to the scaled data, implemented using the Scanpy team's method: scanpy.pp.log1p.

If you have already normalized the data using your own method, you can choose to skip the normalization step. In this case, OmnibusX will treat the input matrix as normalized.

For proteomics and HTOs data, OmnibusX automatically applies `Central Log Ratio` (CLR) normalization. However, the original CLR formula results in negative values, which can be confusing when analyzing expression data in downstream analysis. To address this, OmnibusX modifies the CLR normalization process by shifting the entire value range to the right, ensuring all values are nonzero, while preserving original raw zeros as zeros. Below is the Python implementation of the CLR normalization method used by OmnibusX:

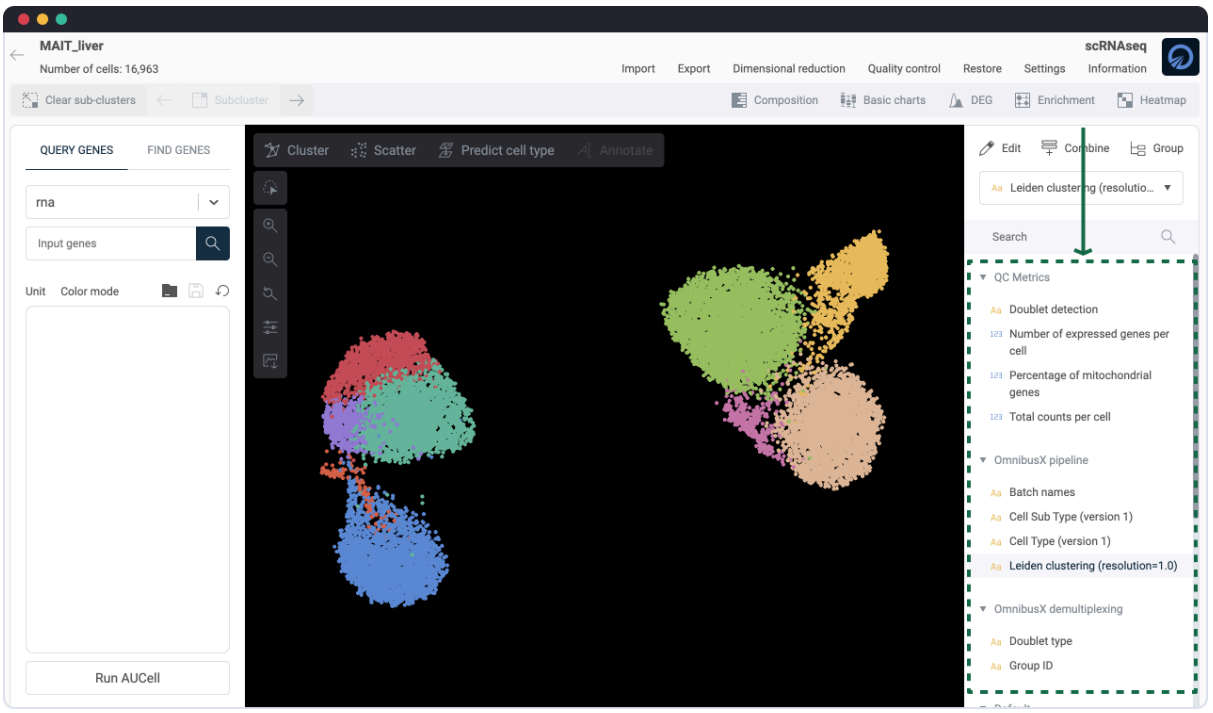
```
def _clr_normalize(data: np.ndarray) -> np.ndarray:
    """Central log ratio normalize the expression matrix."""
    geometric_mean = np.exp(np.mean(np.log(data + 1), axis=0))

    data = data / geometric_mean
    idx = data > 0
    data[idx] = np.log(data[idx])
    data += np.abs(data.min()) + 1
    data[~idx] = 0

    return data
```

4.4. Initial analysis

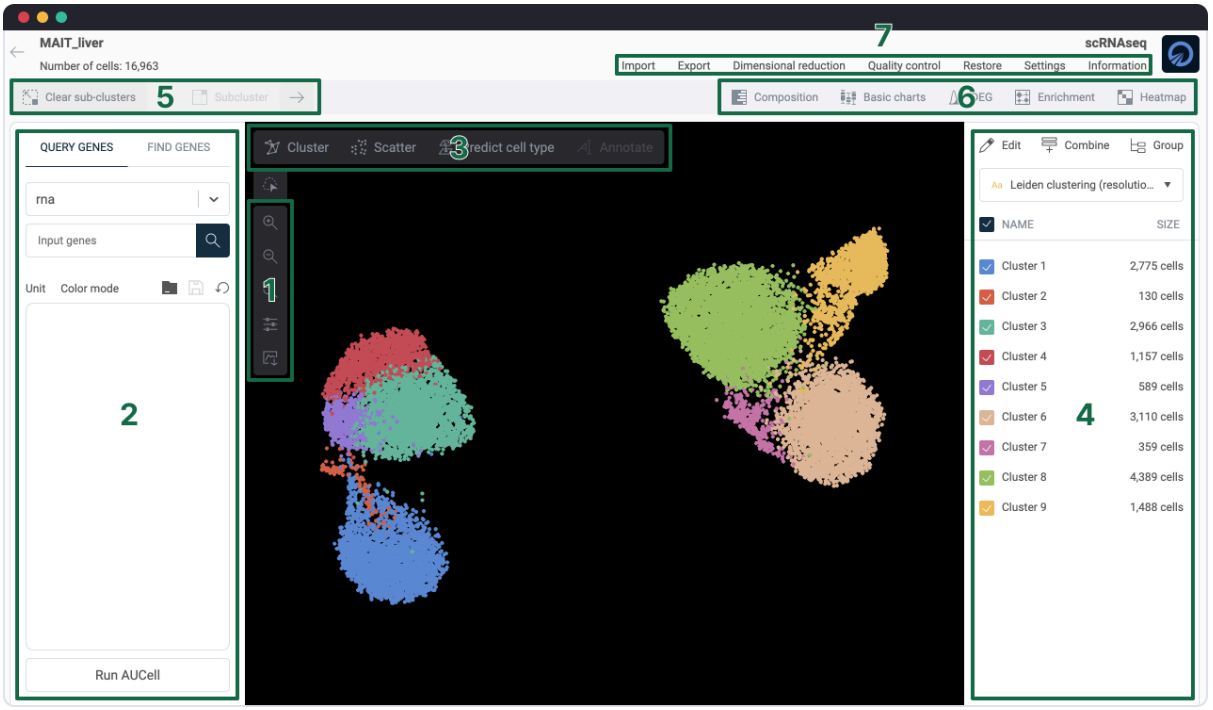
You can initiate a dimensional reduction analysis by specifying the number of highly variable genes to use for running PCA, t-SNE, or UMAP for visualization. Additionally, by default, OmnibusX also performs clustering (using the Leiden algorithm with a resolution of 1) and cell type prediction (using the latest version available for the specified species). These analyses can be adjusted at any time in later steps, providing flexibility to refine your results as needed. Detailed instructions for each analysis method will be provided in the following sections.



If your input Scanpy or Seurat object already contains normalized data and dimensional reduction results, you can choose to skip this initial analysis step.

Data exploration

After successfully submitting your dataset, you can start exploring the data using our provided functionalities. These are organized into seven groups for easy navigation, allowing you to effectively visualize and annotate your data.



1. `Visualization tools`: Adjust scatter settings, zoom in and out, and navigate through cell cluster patterns.
2. `Expression tools`: Visualize gene, gene sets, and feature expression from distinct omics, extract cluster markers, and explore gene set enrichment patterns..
3. `Computational methods`: Generate cell labels using clustering algorithms, perform automatic cell typing with cell type prediction, or gene set

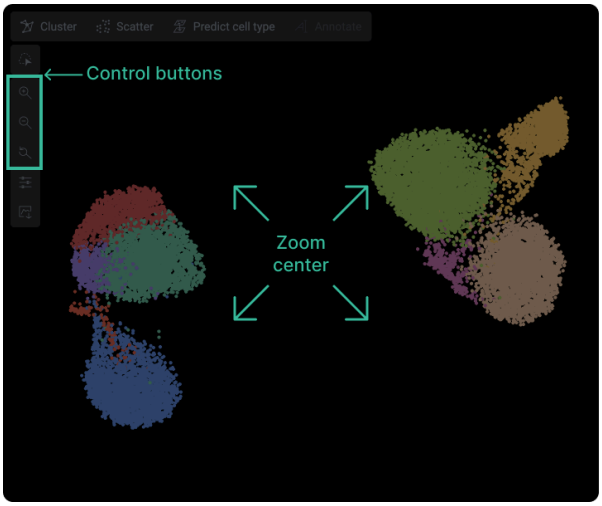
enrichment scores. Create t-SNE and UMAP visualizations to refine and uncover hidden patterns in your data..

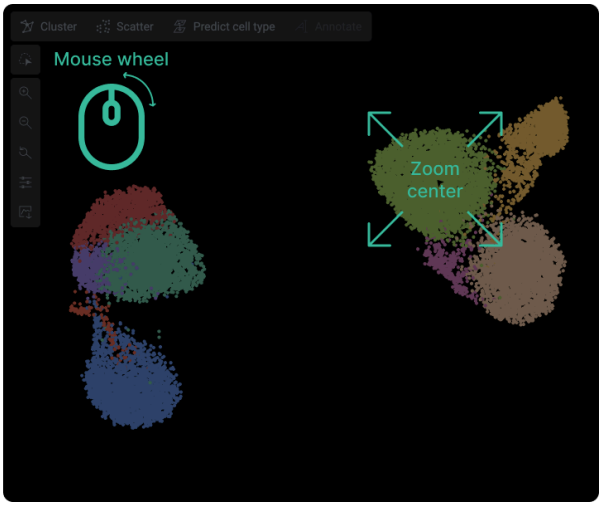
4. `Annotation management`: Add, remove, modify, and organize your cell labels.
5. `Sub-cluster management`: Create and manage sub-clusters within your main clusters. This is useful for breaking down the data further and analyzing smaller, more specific groups of cells..
6. `Downstream analysis`: This section includes tools for various downstream analyses such as differential expression, enrichment analysis, heatmap, etc, which will be explained in a separate section.
7. `Data management`: This group provides tools for modifying and managing your data, including importing additional data, creating new analysis results, and re-generating data from the quality control step, etc.

1. Visualization tools

1.1. Zoom

There are two ways to control the zoom level of the cell scatter plot:

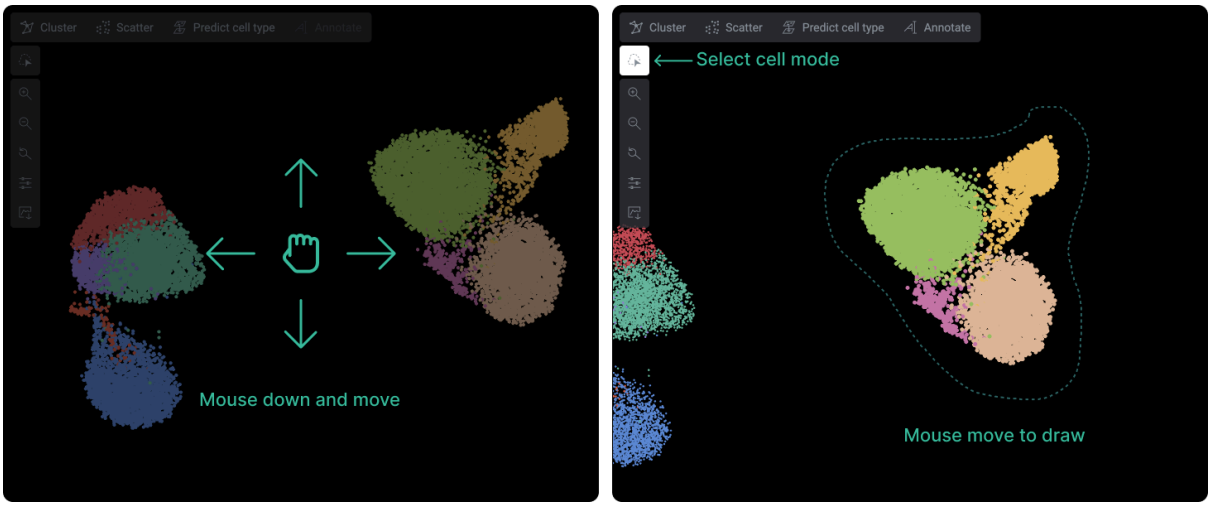




- **`Control buttons:`** Located at the top right corner of the scatter plot, these buttons include `Zoom In` and `Zoom Out`. When using these buttons, the zoom level will adjust at the center of the scatter plot.
- **`Mouse wheel:`** Scrolling the mouse wheel up or down will zoom in or out of the scatter plot. When using the mouse wheel, the zoom center will be at your mouse pointer, allowing you to focus on specific areas of interest.

1.2. Navigate

You can navigate through the scatter plot by dragging it with your mouse. This feature is disabled when you are in `Select cell` mode to avoid confusion between moving the scatter plot and drawing a selection area.

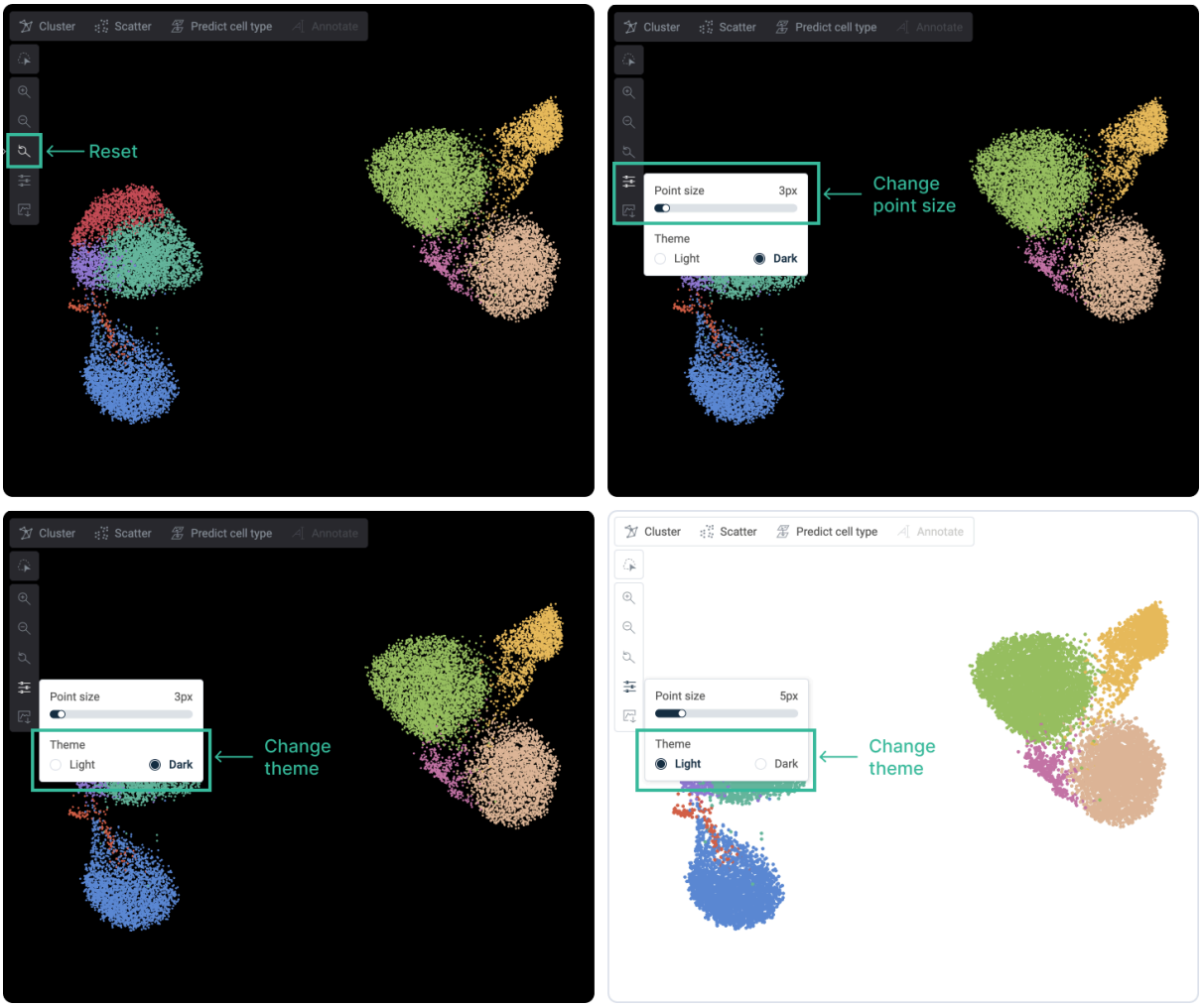


1.3. Reset

Clicking the `Reset` button will restore the scatter plot to its default zoom level and position.

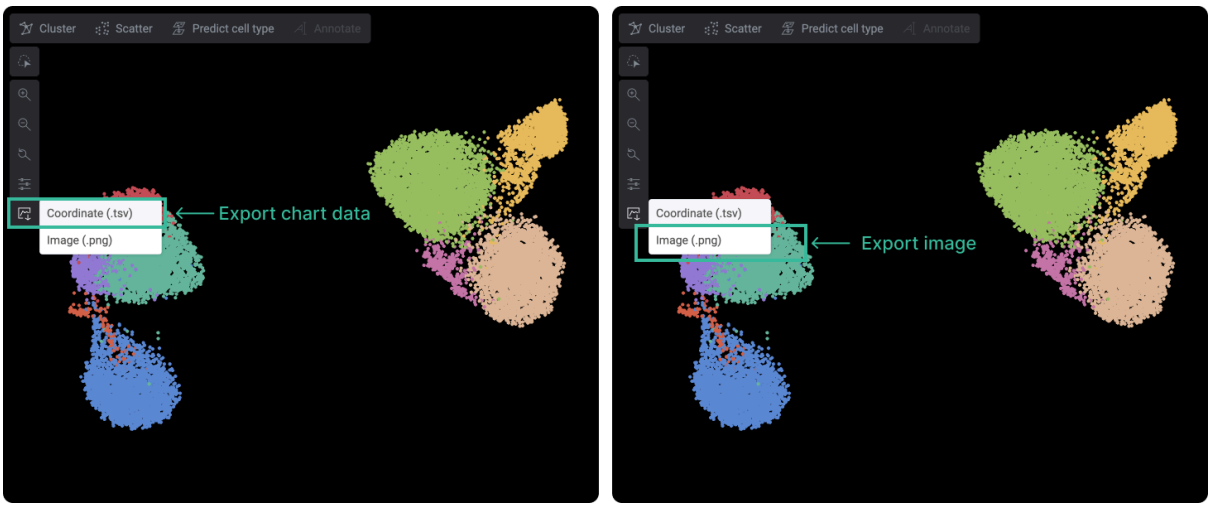
1.4. Settings

Adjust the point size by moving the slider or switch between light and dark themes to suit your preference.



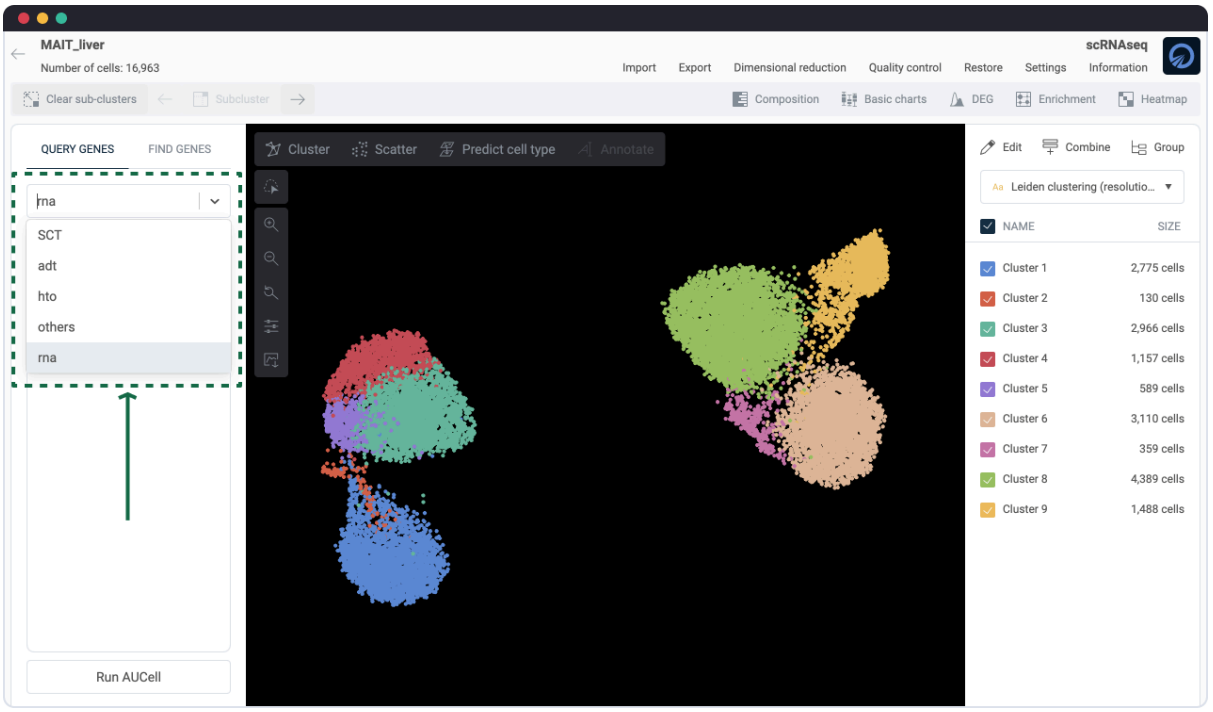
1.5. Export

You can export the current scatter plot view, including the current zoom level, scatter position, and cell colors, as a PNG image. Additionally, you can export the chart data as a TSV file.



2. Expression tools

2.1. Selecting omics

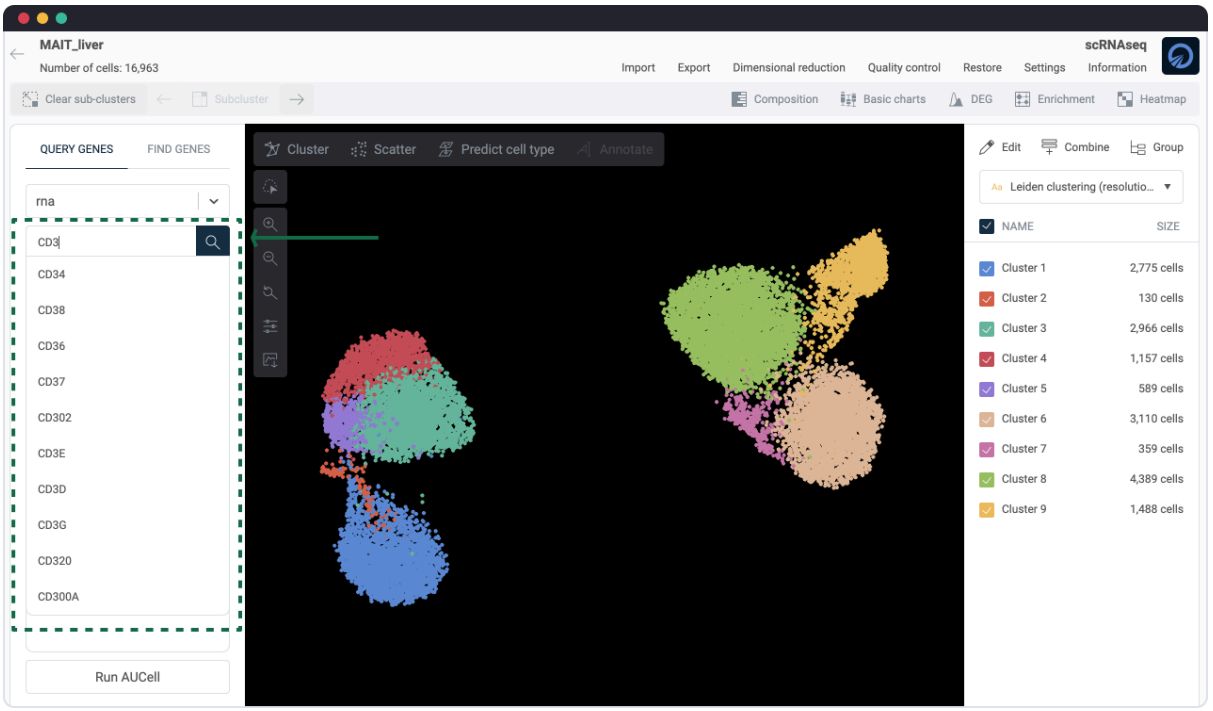


From the `Query genes` panel, all the available omics from your input data will be listed in the selection box. OmnibusX processes each omic separately, so you must first select your omic of interest. This applies to most analyses, except for cell type

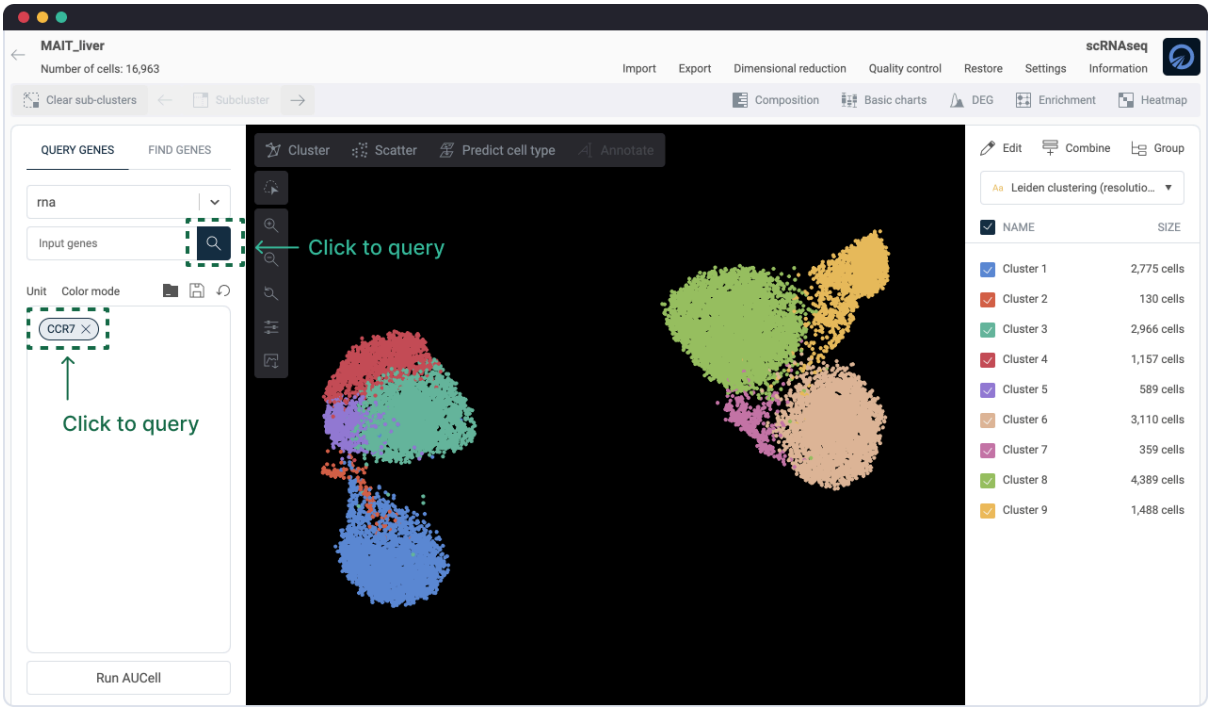
prediction, gene set analysis, and enrichment analysis, which can only be applied to transcriptomics due to their dependence on known annotated references/databases.

2.2. Query genes by name

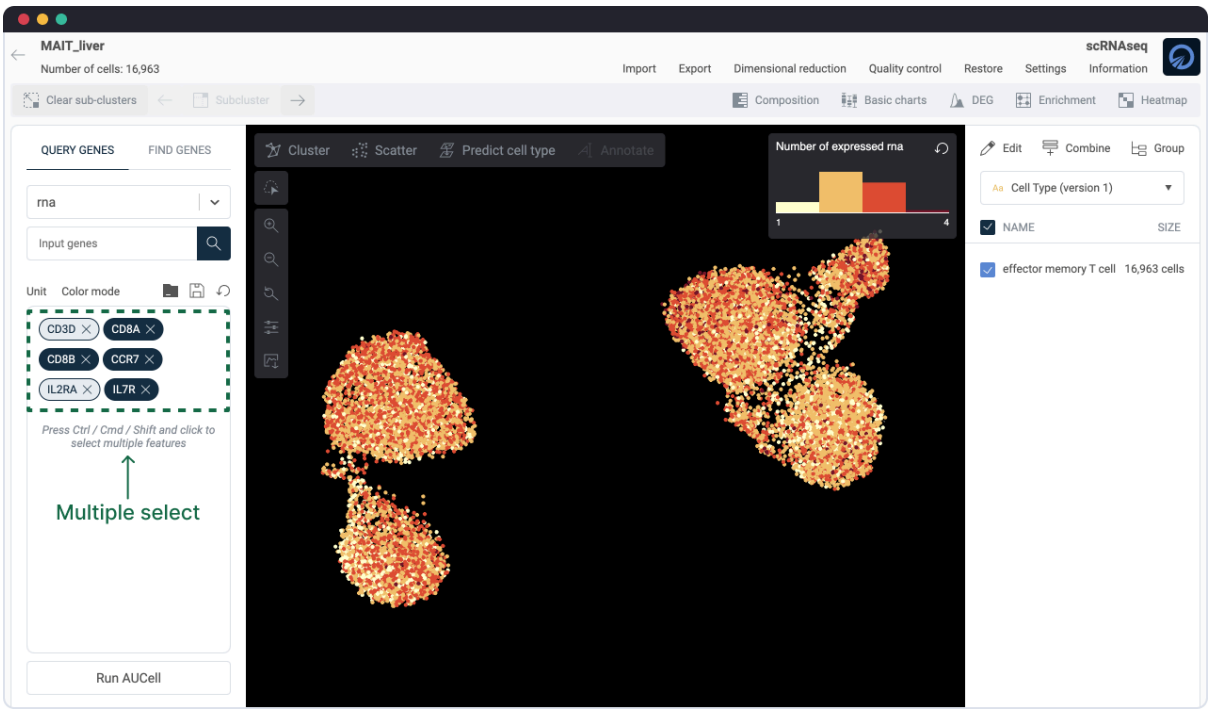
After selecting your omic of interest, you can explore target genes by typing their names into the input box. OmnibusX will suggest related gene names. If your target gene does not appear, and you are working with transcriptomics (`rna` omic), try switching to the `Others` omic where unmapped genes are stored. Alternatively, check for aliases via Google and try those. If no results are found, your target may not be present in the original data, which can apply to other omics (proteomics, genomics, etc.).



Once you find your target name, select the desired option from the suggestions by clicking on it or using the arrow keys (`↑Up` / `↓Down` and `↵Enter`). Press `↵Enter` or click the search icon to query the expression.



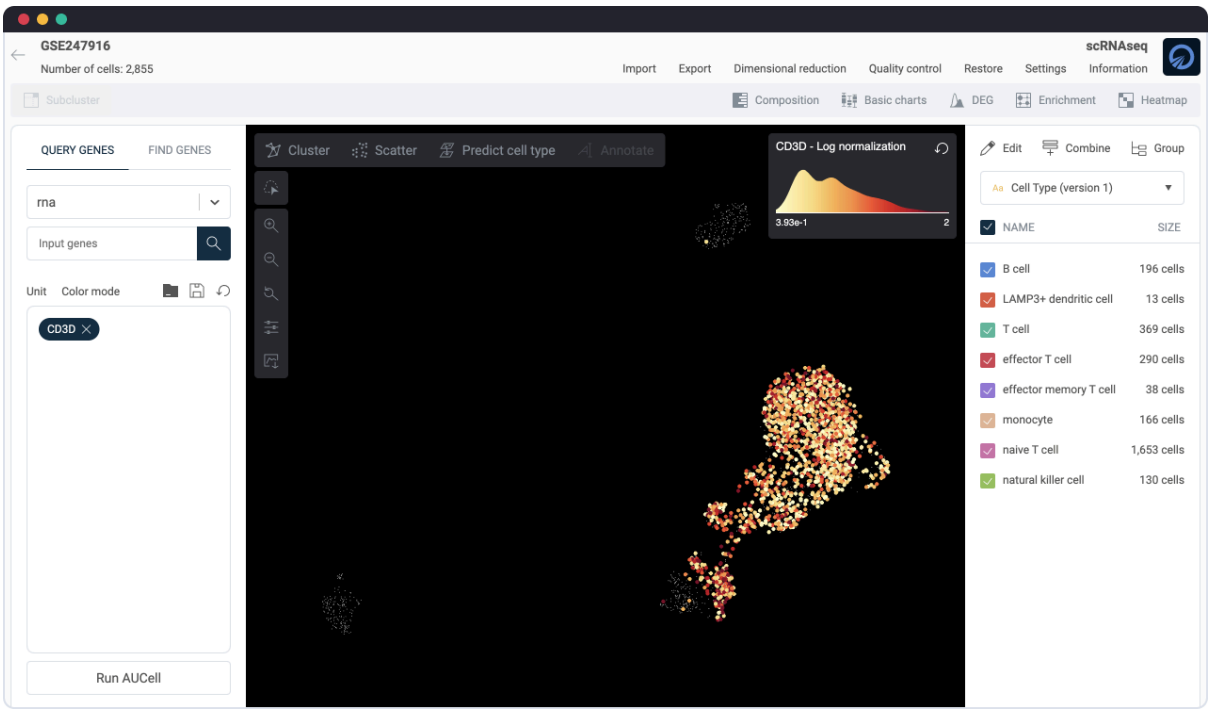
You can select multiple genes and switch between their visualizations by clicking on each gene name tag below. Hold `Shift`, `Ctrl`, or `Cmd` while clicking to select multiple genes from the list. Press the search icon again to select all input genes. The color mode changes accordingly as the number of selected genes for visualization changes.



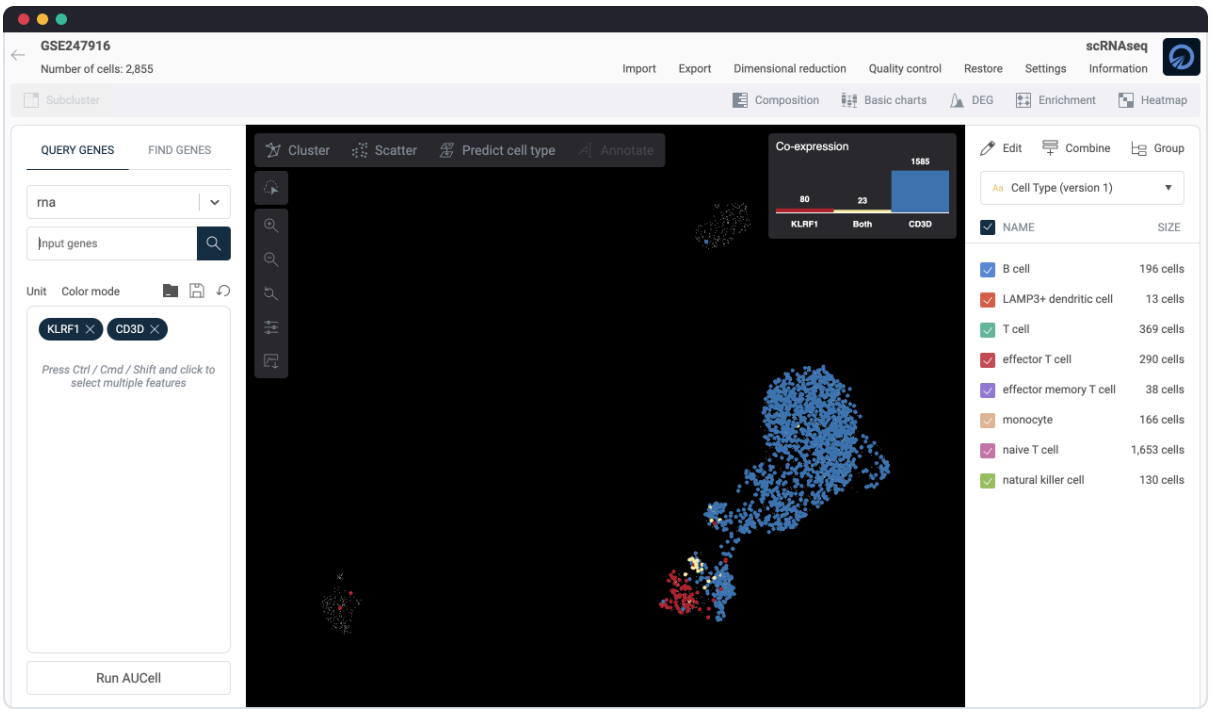
2.3. Color modes for expression visualization

There are four color modes to visualize gene/gene set expression:

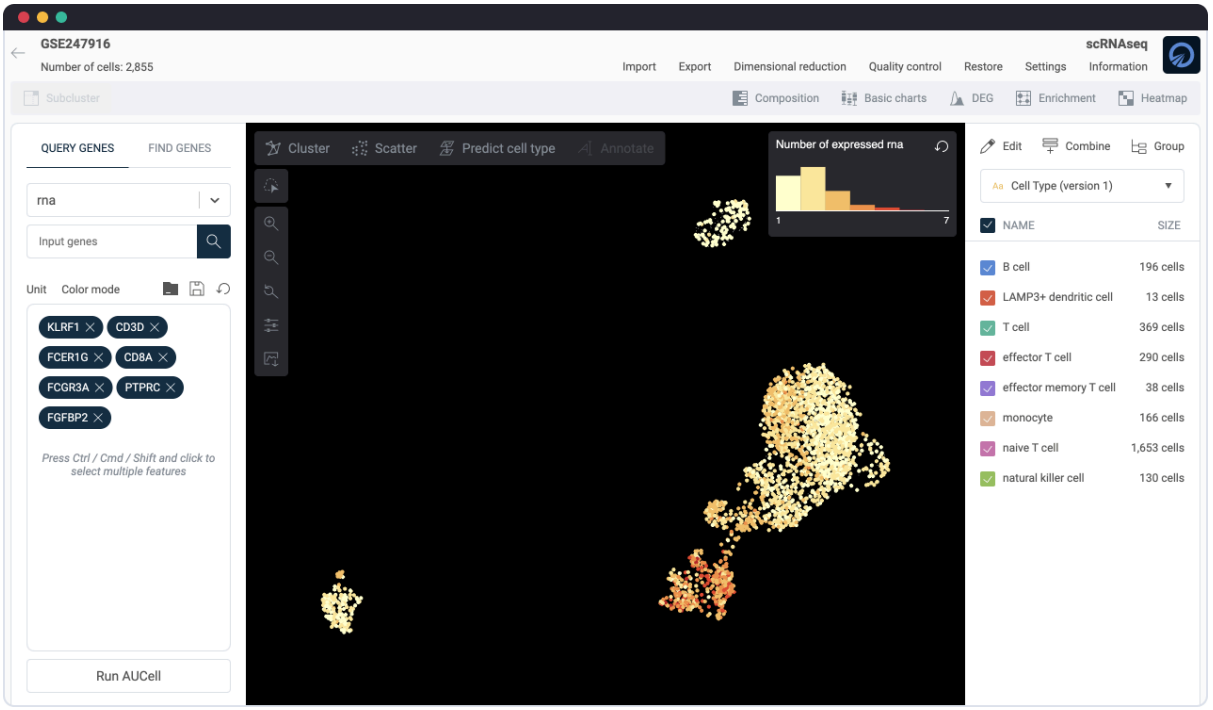
- `Feature`: Colors are scaled from the `5th` to `95th` percentile of non-zero values to avoid outliers. Zero values are colored **gray**.



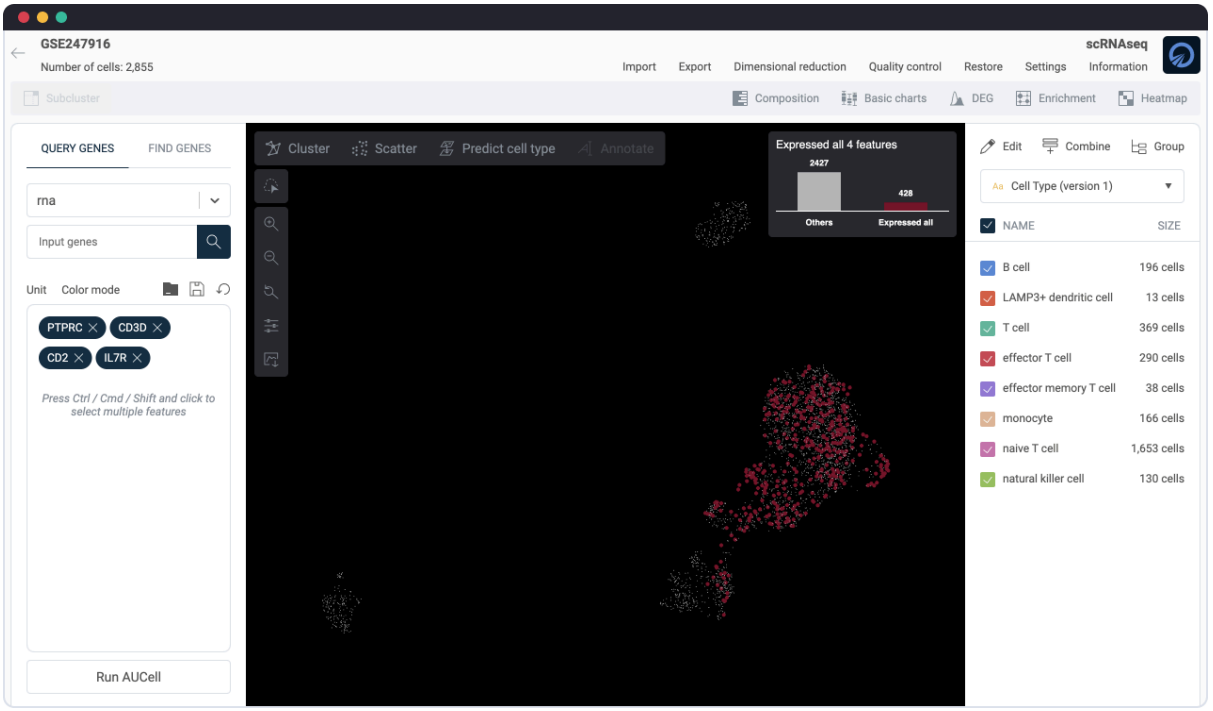
- `Tri-color`: Cells are categorized into four colors based on expression of two genes: **gray** for no expression, **red** for first gene only, **blue** for second gene only, and **yellow** for both genes.



- `Count`: Colors the scatter plot based on the number of genes expressed by each cell.



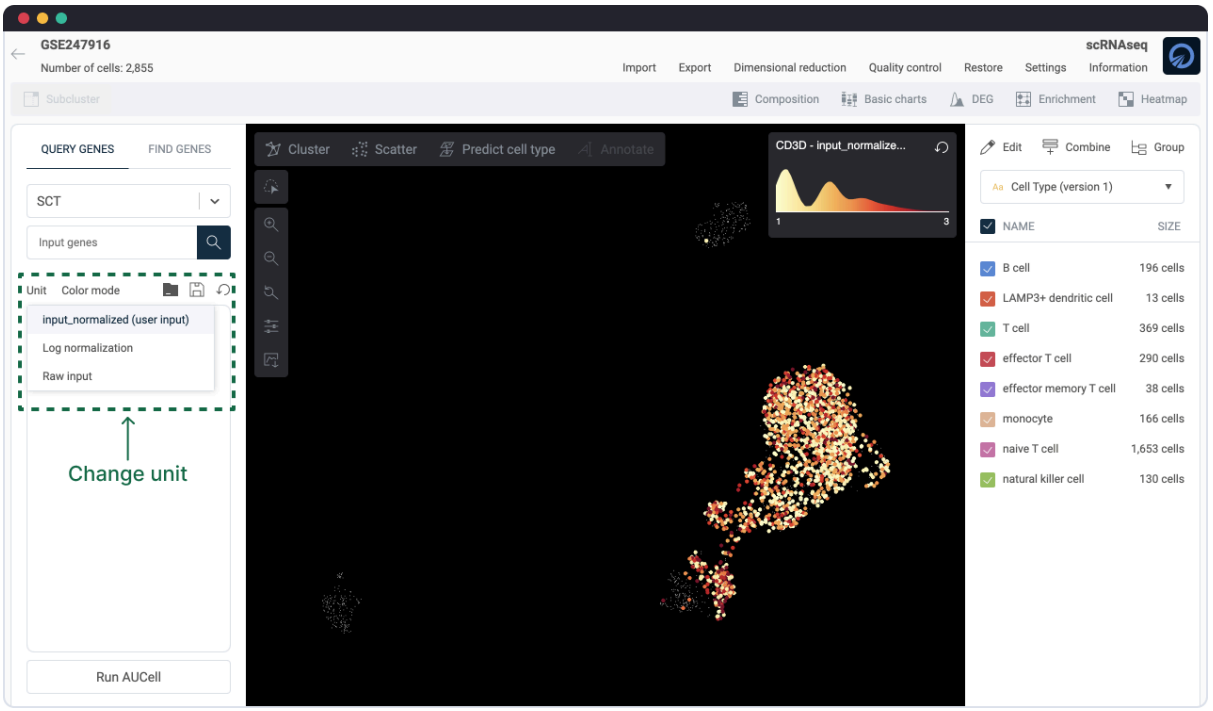
- `Binary`: Uses two colors: **red** (cells expressing all selected genes) and **gray** (remaining cells).

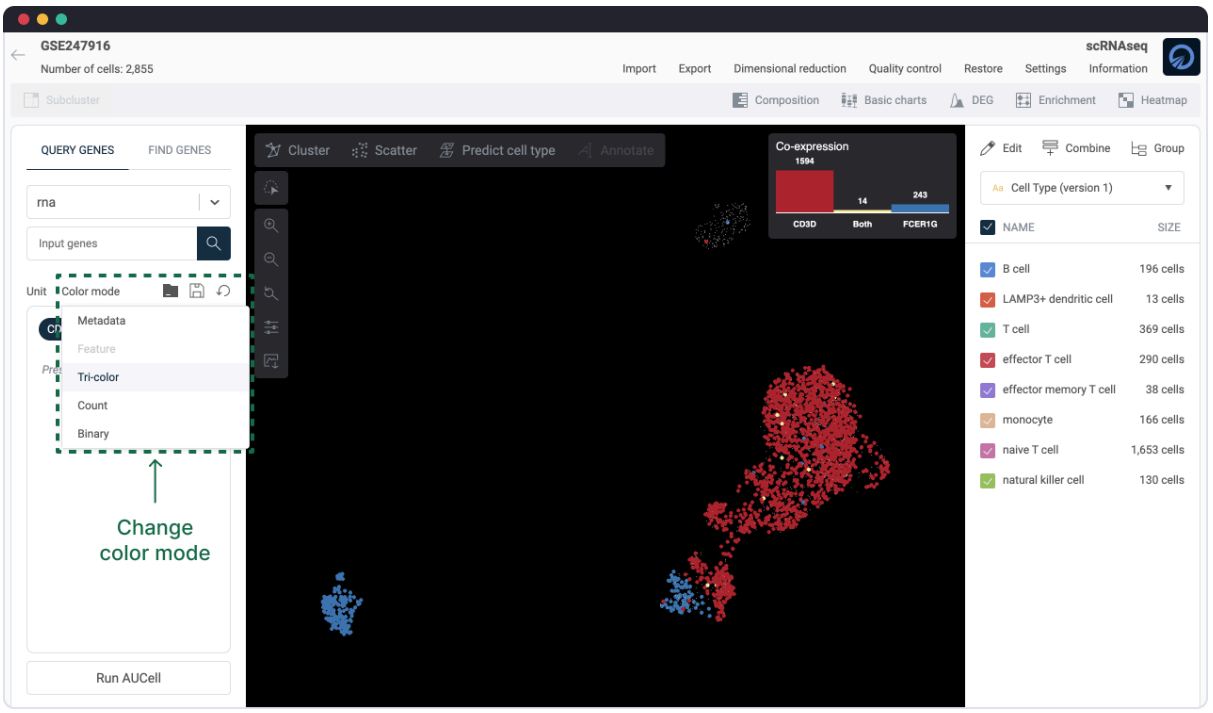


By default, OmnibusX applies the following color modes based on the number of selected genes:

- One gene: `Feature` mode. You can switch between available units to see how normalization methods affect the expression pattern.
- Two genes: `Tri-color` mode.
- More than two genes: `Count` mode.

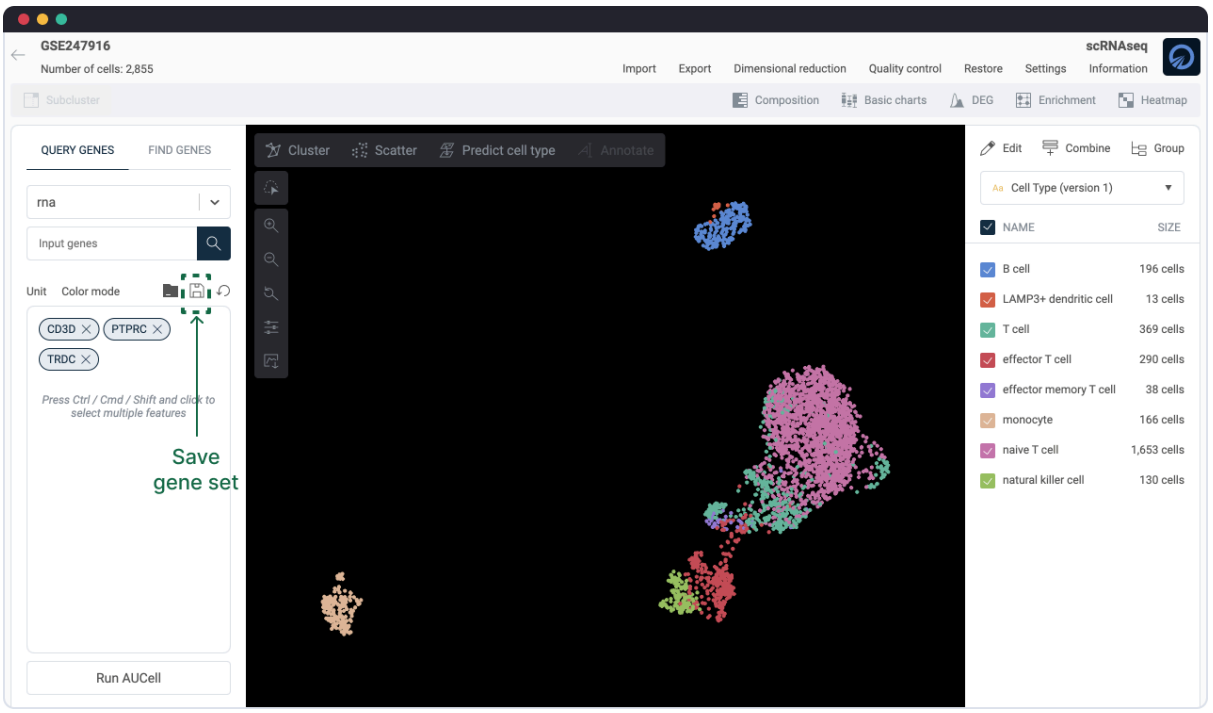
You can then switch to any color mode according to your preference.

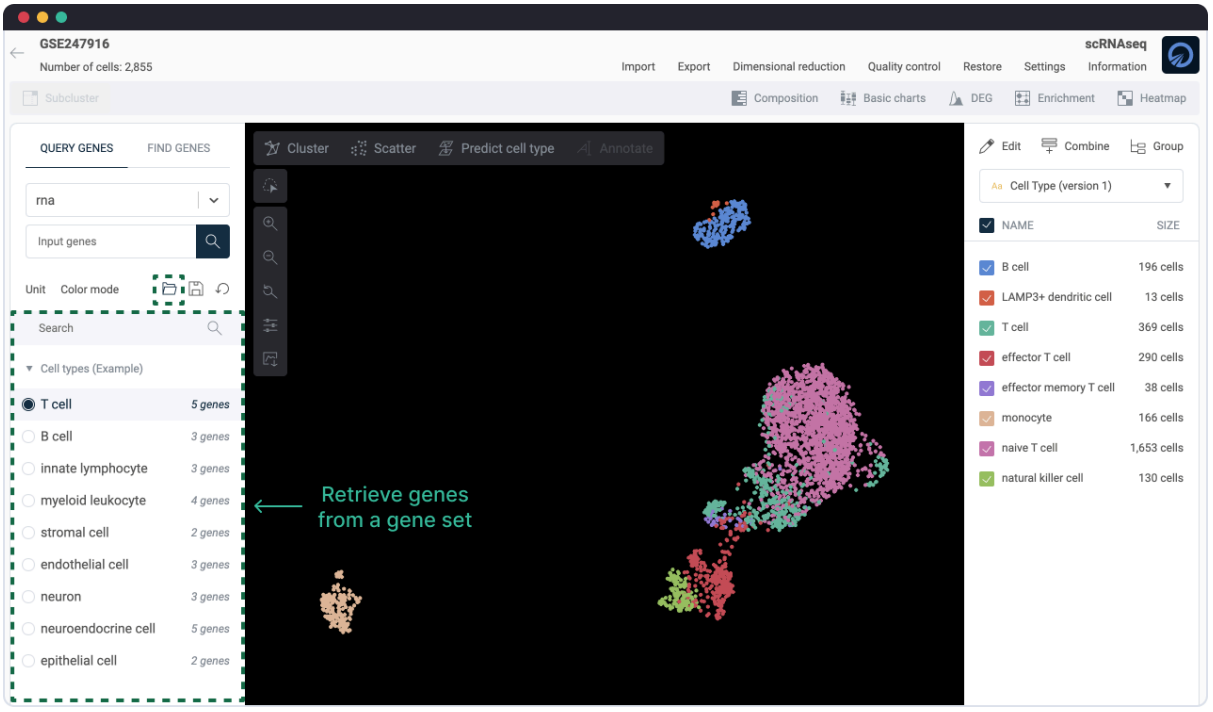




2.4. Gene sets

You can save your current input genes to your gene set library by clicking on the save icon located at the top of the gene box. This function is applicable for the rna omic only, as the gene set library is designed to be used across different studies. Therefore, it needs to utilize common shareable names, which OmnibusX standardized using Ensembl gene annotation.



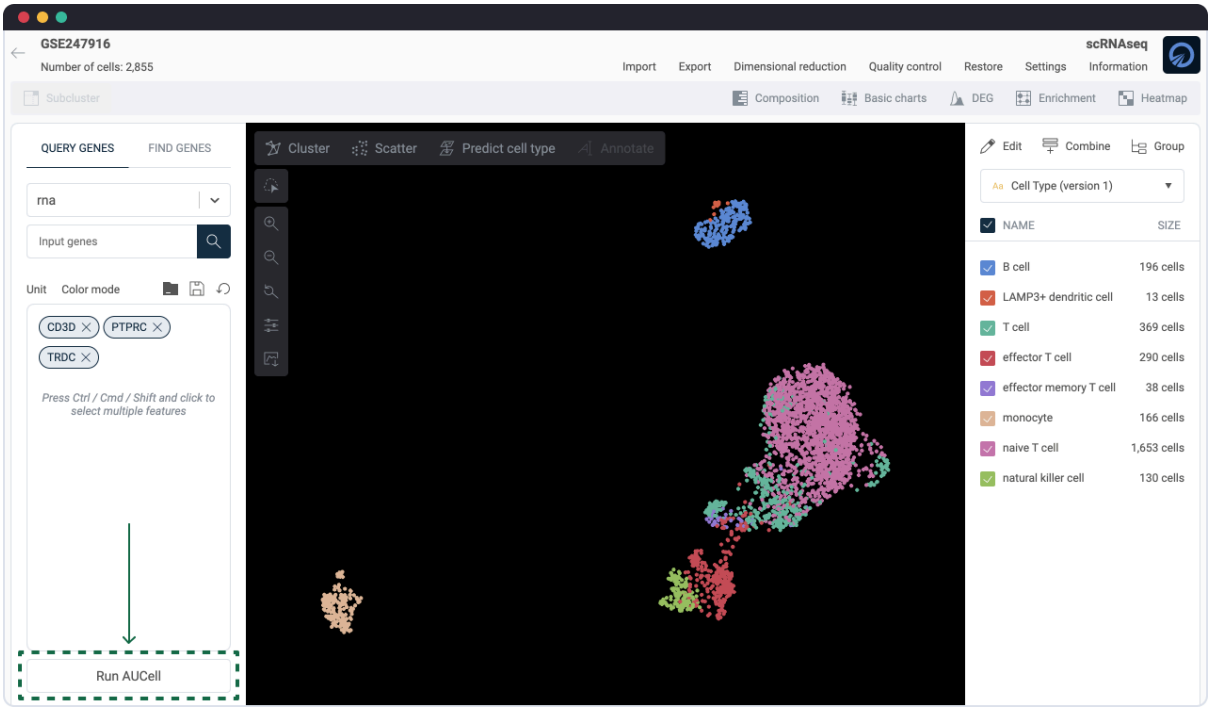


Once you have saved your gene set, you can query genes using your saved gene sets by clicking on the gene set folder icon next to the save icon. This allows you to retrieve a list of target genes from a specific gene set and explore their expression within your dataset.

Detailed instructions on how to create and manage your gene set library will be provided in the Your Libraries section later.

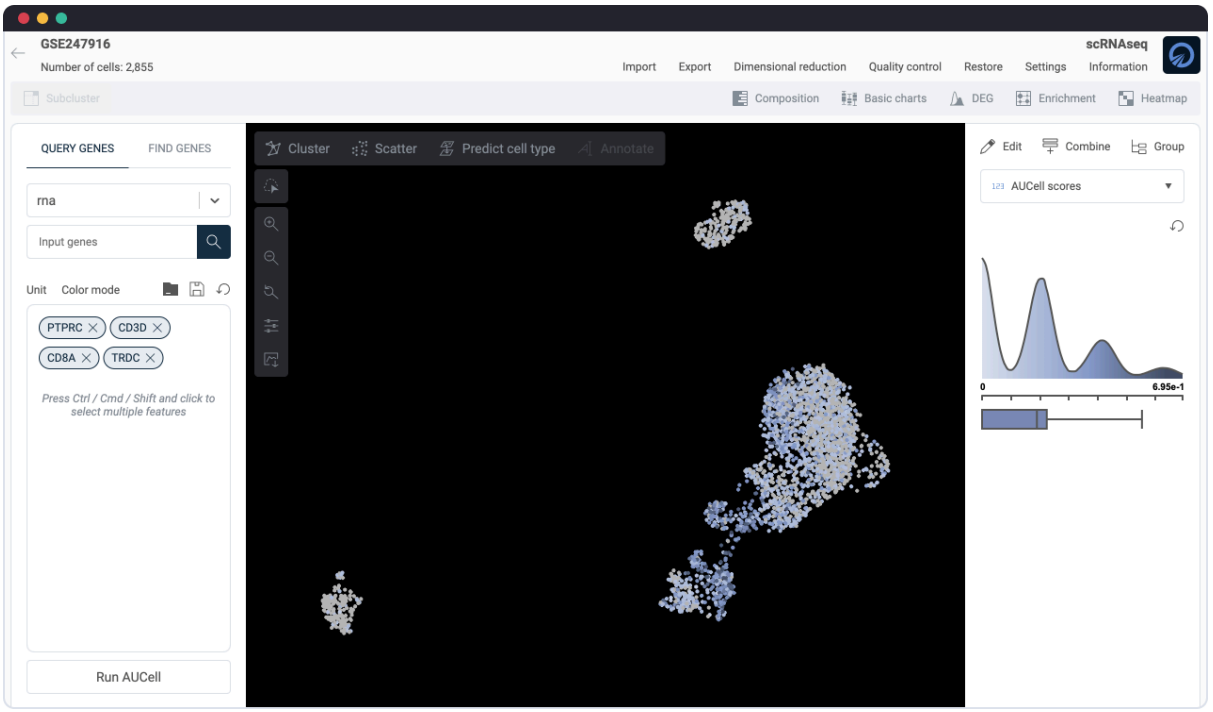
2.5. Gene set enrichment

For analyzing a list of genes, in addition to visualizing the number of expressed genes in each cell and observing their patterns across cell clusters, OmnibusX also allows you to analyze a list of genes using a computational method called AUCell (*Aerts et al., 2016*).



AUCell enables you to identify cells with active gene sets, such as signatures or gene modules, in single-cell RNA-seq data. AUCell uses the "Area Under the Curve" (AUC) to calculate whether a critical subset of the input gene set is enriched within the expressed genes for each cell. This scoring method is ranking-based and independent of the gene expression units and normalization procedure. For more details on the AUCell algorithm, please refer to the Bioconductor package documentation.

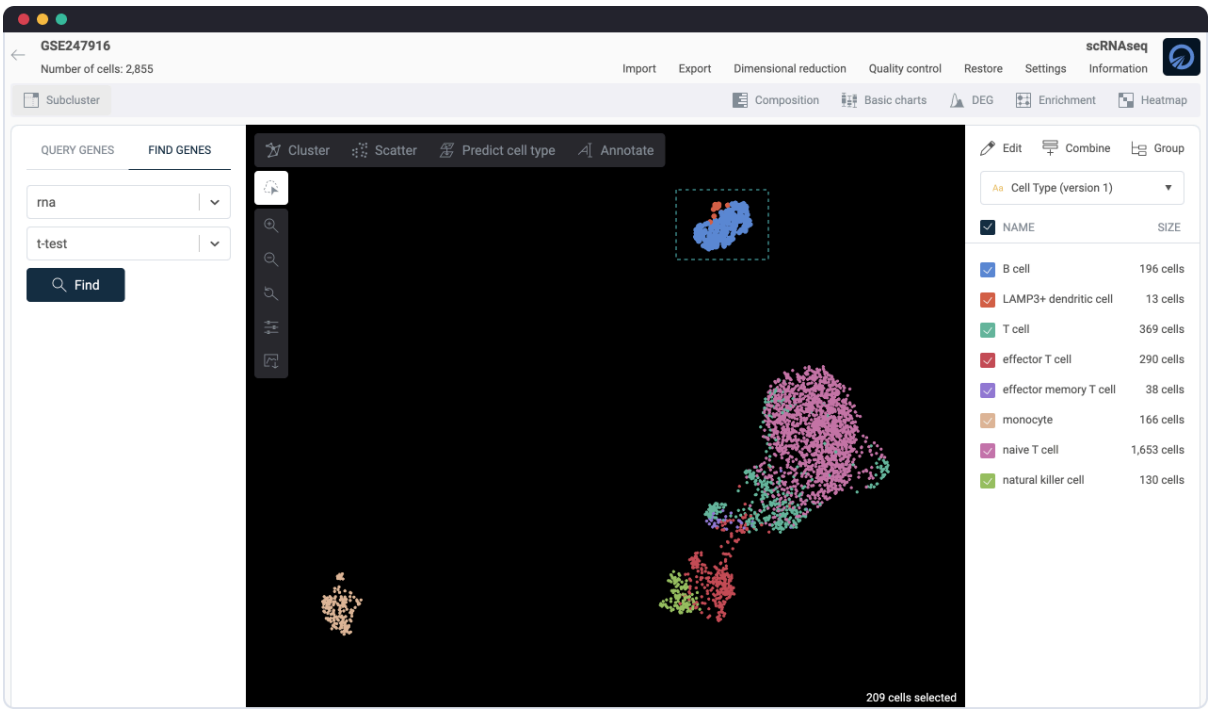
To perform AUCell analysis, simply select the gene set of interest and choose AUCell as the analysis method. The resulting enrichment score will be saved as a new annotation for later access. This allows you to explore the activity of your gene set across different cell clusters and identify potential cell types or states associated with its activity.



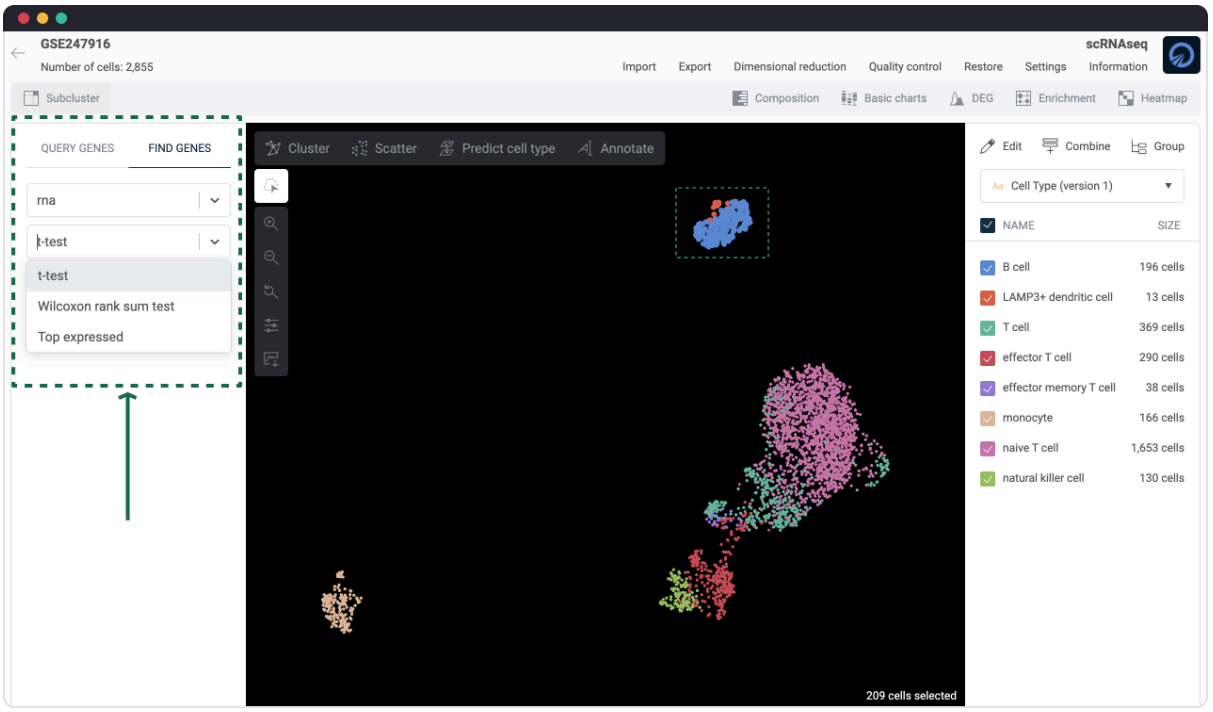
2.6. Finding markers

Besides exploring the expression of known targets in your dataset, another way to study your data is by profiling the markers of specific clusters of interest to find new targets. In OmnibusX, you can find genes associated with a specific cell cluster by following these steps:

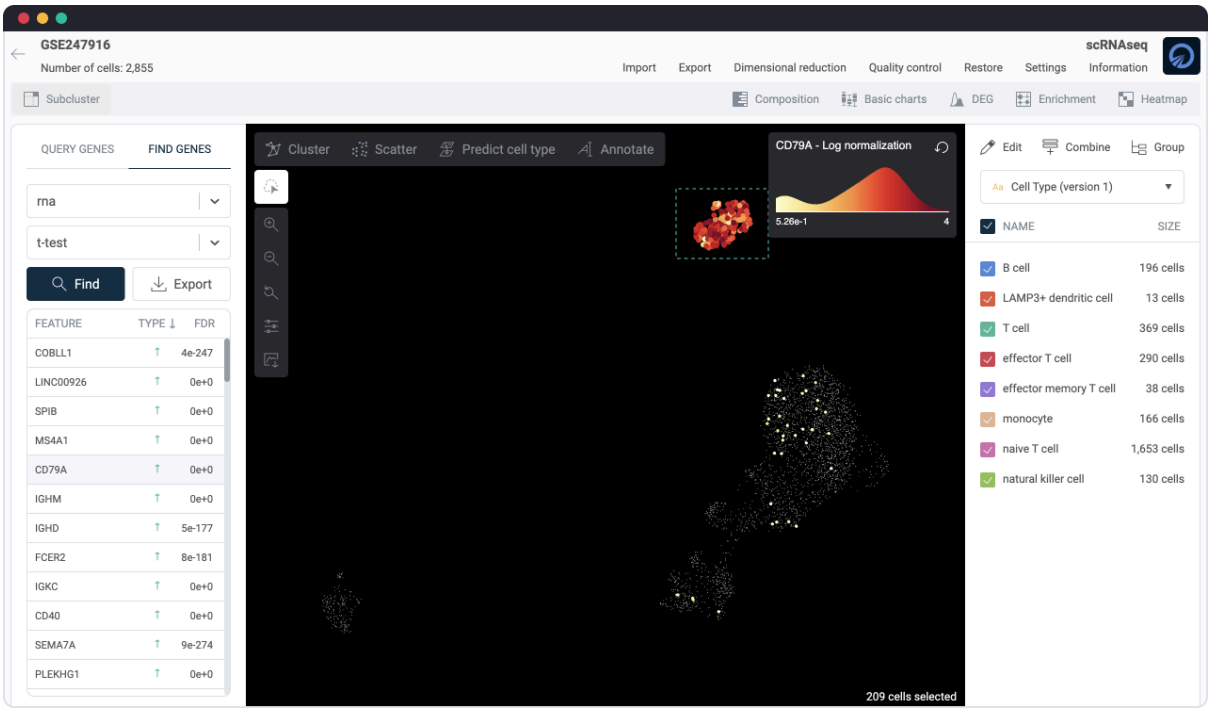
1. `Select the population of interest`: First, select the cell population you are interested in. There are multiple ways to select a set of cells, which will be described in detail in the Cell selection section.



2. `Select omic and method`: Once you have selected your population, switch to the `Find genes` panel. Select the omic of interest and choose a method for finding markers. The available methods include:



- `t-test`: Perform a t-test across all features of your selected omic between your selected cell population and the remaining cells on the screen (which can be a specific sub-cluster). Features with expression coverage lower than 30% in both groups are removed. The results are filtered with a threshold of p-value < 0.01, and the top 100 results with the lowest p-values are returned. The t-test implementation is from the SciPy team: scipy.stats.ttest_ind.
- `Wilcoxon rank sum test`: Similar to the t-test procedure, but using the Wilcoxon rank sum test. This method also filters features with expression coverage lower than 30% in both groups and returns the top 100 results with the lowest p-values. The implementation is from the SciPy team: scipy.stats.mannwhitneyu.
- `Top expressed`: This method finds the top 100 features with the highest total expression values in the currently selected population, regardless of how they are expressed in the remaining cells.

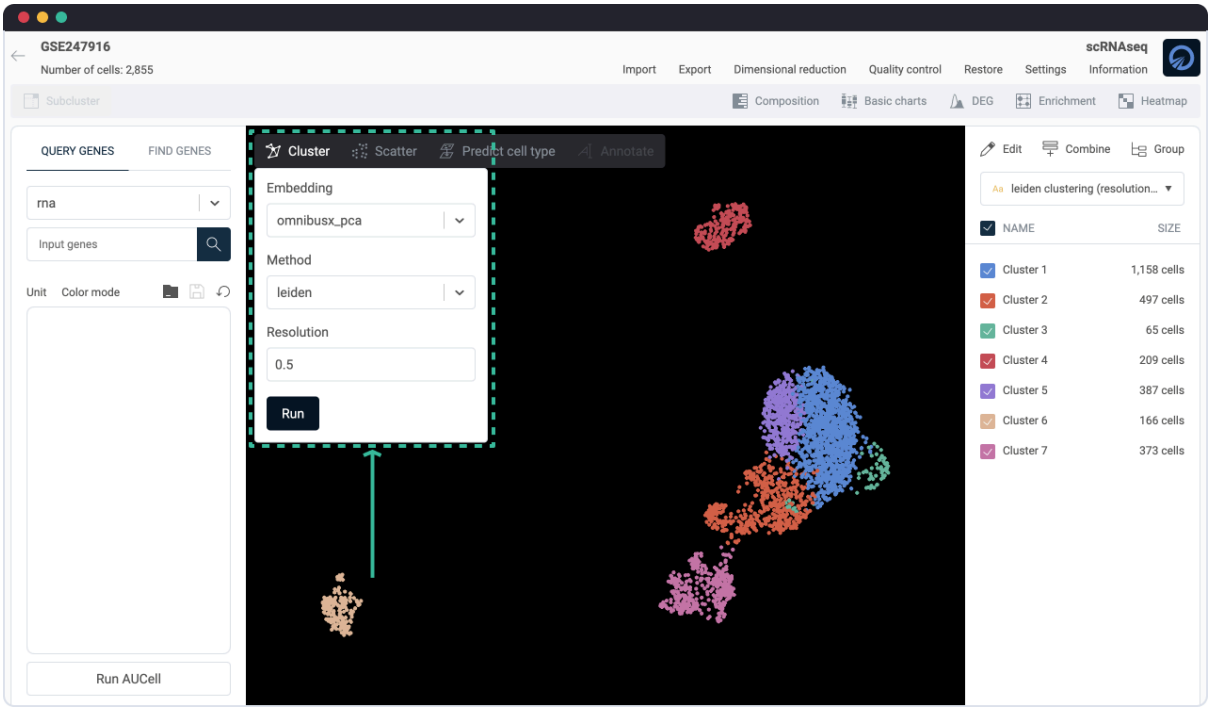


3. Computational methods

3.1. Clustering

OmnibusX provides two clustering methods to quickly generate cell clusters with different resolutions:

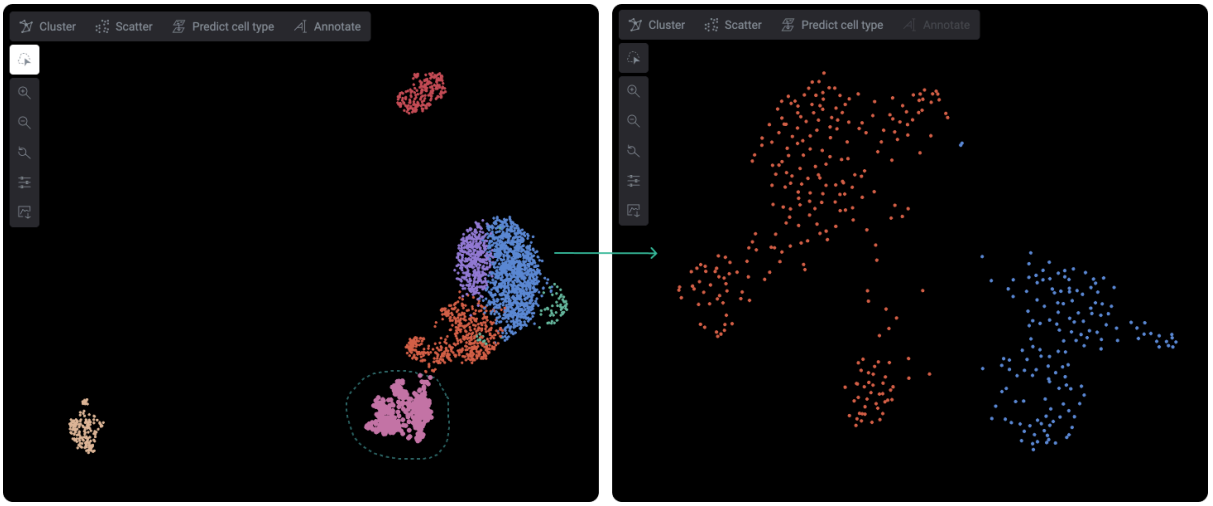
- `Louvain clustering` (*Blondel et al., 2008*; *Levine et al., 2015*; *Traag, 2015*): Clusters cells using the Louvain algorithm. OmnibusX uses the implementation from the Scanpy team: scanpy.tl.louvain.
- `Leiden clustering` (Traag et al., 2019): An improved version of the Louvain algorithm. OmnibusX uses the implementation from the Scanpy team: scanpy.tl.leiden.



These clustering methods can be applied to different omics, allowing for the simultaneous analysis of multimodal datasets. The clustering process requires a cell embedding as input. Initially, OmnibusX automatically generates a PCA embedding from `rna` omics data as part of the submission pipeline. Users also have the flexibility to switch to other pre-existing embeddings or create new ones

for different omics, which will be covered more comprehensively in the Cell embedding management section.

When triggering the clustering function, only the cells currently displayed on the screen are considered. This feature facilitates the clustering of specific sub-populations. Clustering a subset of cells helps in distinguishing closely similar cell populations that may be challenging to separate on a global scale. In such cases, increasing the clustering resolution might lead to overly segmented clusters or unexpected splits, potentially obscuring true biological differences. More on generating sub-clusterings will be discussed in the Sub-cluster management section.

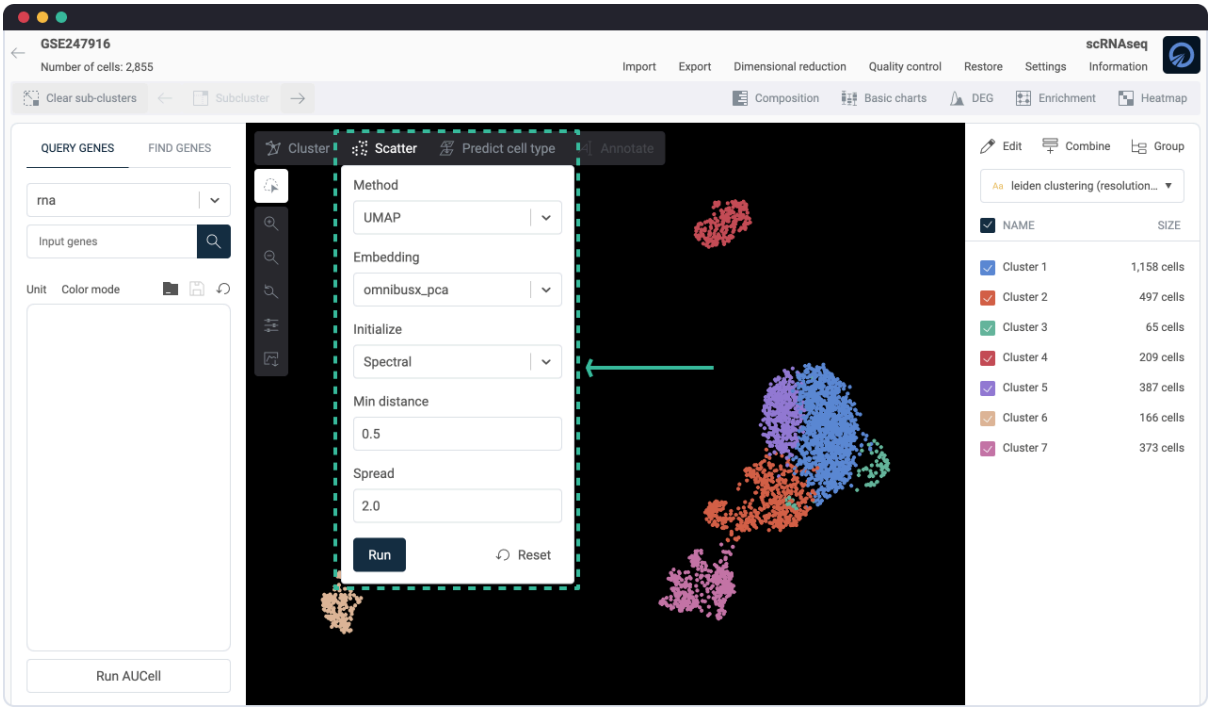


3.2. t-SNE/UMAP

OmnibusX provides two popular dimensionality reduction techniques for quickly adjusting cell scatter visualizations:

- `t-SNE` (*Amir et al., 2013*; *Pedregosa et al., 2011*; *van der Maaten and Hinton, 2008*): This method optimizes the distribution of nearest-neighbor distances in the embedding to closely match those in the high-dimensional space. OmnibusX employs the t-SNE implementation provided by the Scipy team: scanpy.tl.tsne.

- UMAP (*McInnes et al., 2018*): Besides faster than t-SNE, UMAP optimizes the embedding to best reflect the data's topology. The implementation used by OmnibusX is also sourced from the Scipy team: scanpy.tl.umap.

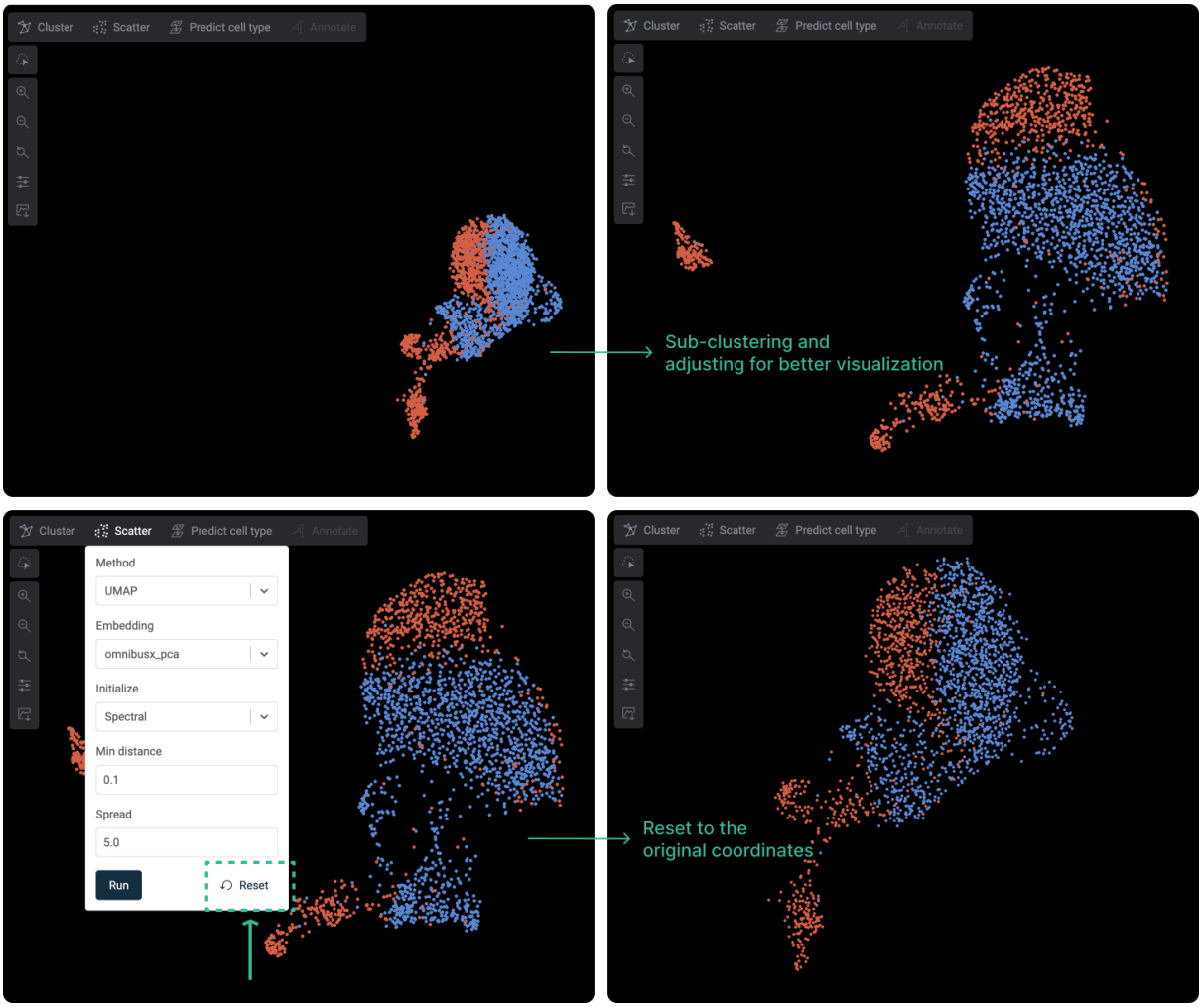


Similar to the clustering process, t-SNE and UMAP require a cell embedding as input. OmnibusX automatically generates a PCA embedding from rna omics data as part of the initial submission pipeline. Users can also choose to utilize other existing embeddings or create new ones for different omics, which will be further explained in the Cell embedding management section.

When the t-SNE/UMAP function is triggered, only the cells currently displayed on the screen are considered. This selective visualization aids in the analysis of specific sub-populations, helping to distinguish closely similar cell populations that may be difficult to separate on a global scale. Details on generating sub-clusterings will be further discussed in the Sub-cluster management section.

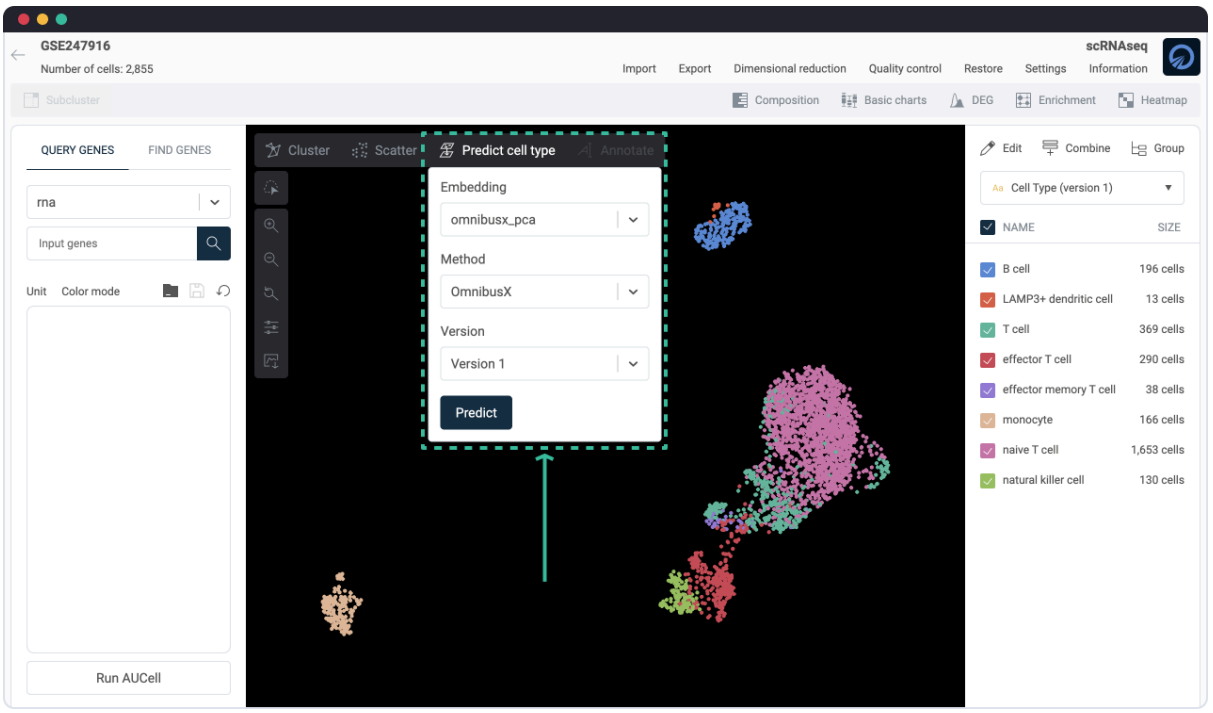
When a specific cell population is selected and moved to sub-clusters for deeper analysis, the default cell scatter subset is extracted from the current scatter coordinates. Any subsequent adjustments will only apply to this localized scope.

You can revert to the original scatter coordinates by clicking the `Reset` button located at the bottom right corner.



3.3. Cell type prediction

With OmnibusXLab advanced cell type prediction feature, you can precisely label your data at the single-cell level. Our methodology, published in Biorxiv, uses a customized algorithm that builds on the AUCell method to achieve accurate cell labeling. By manually curating cell type names and markers from 280 publications, we have unified the nomenclature to consistently identify 166 cell types and subtypes. Our algorithm surpasses the performance of reference-based tools like Azimuth, especially in distinguishing closely related subtypes.



In addition to our pre-built cell type library, OmnibusX also allows you to label your dataset using your own curated gene set. This enables you to detect your own new cell subtypes precisely at the single-cell level, even with a very small population. The Gene set library section provides more detail on how to create and manage your own gene set.

4. Annotation management

OmnibusX stores and manages all dataset annotations as metadata and metadata groups, which can be accessed and updated from the right panel.

4.1. Find a metadata

To find a specific metadata, click on the metadata select box in the metadata panel to open the dropdown list of all metadata groups and metadata. You can organize metadata into groups to make it easier to navigate. Click on the group name to expand or collapse the metadata list.

For efficient navigation, simply type the desired metadata name in the search box and click on it to retrieve the corresponding annotation data.

OmnibusX categorizes metadata into two types:

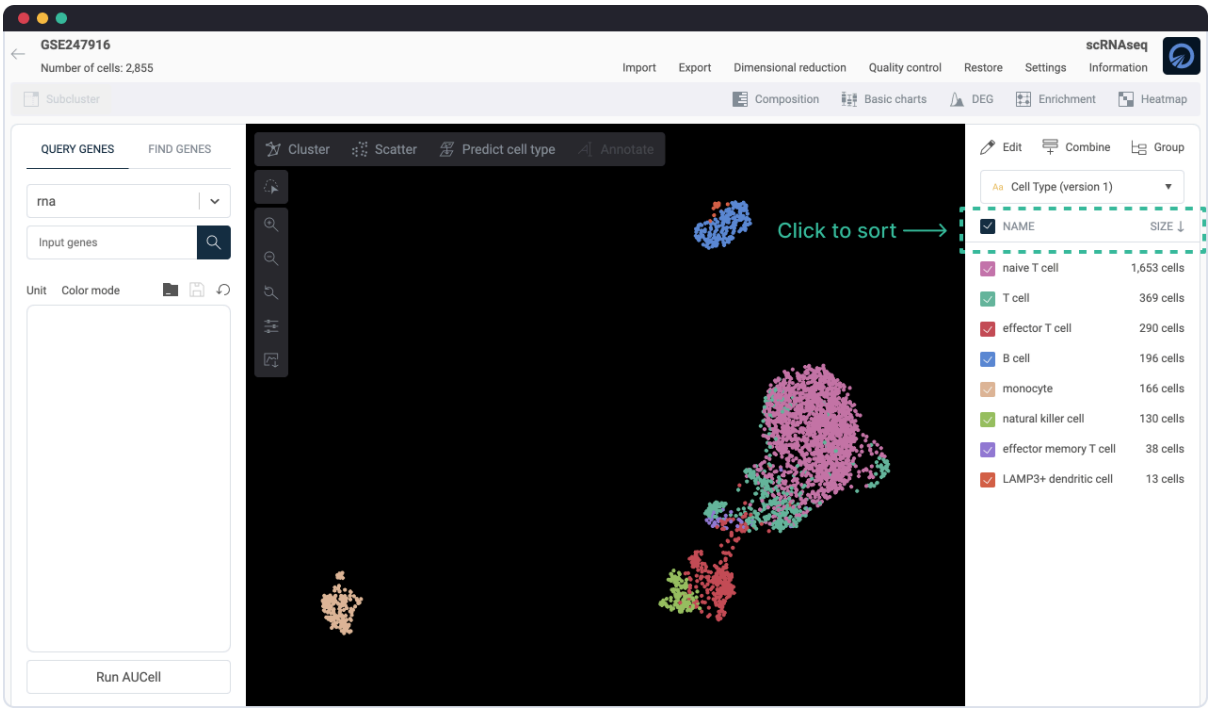
- `Categorical metadata`: `Aa` represents distinct labels.

- `Numerical metadata`: 123 represents continuous labels.

Each type is represented by a small icon before each metadata name. Changing the metadata will also change the cell scatter color accordingly.

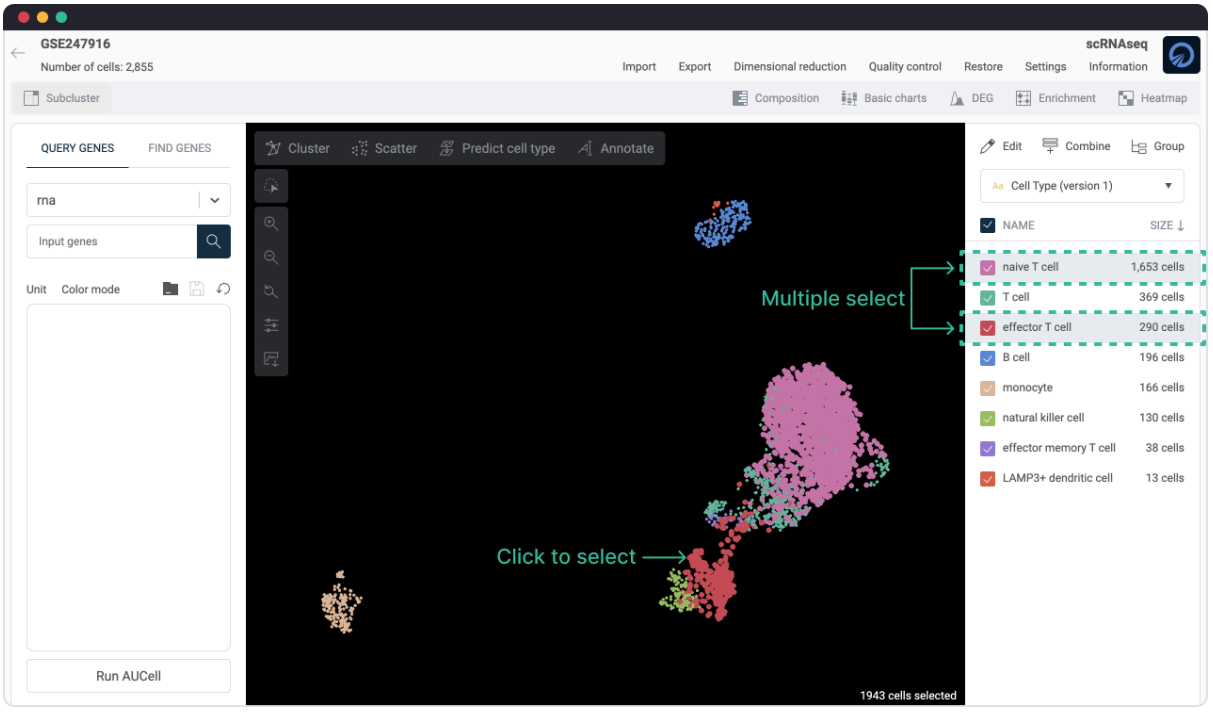
4.2. Categorical metadata

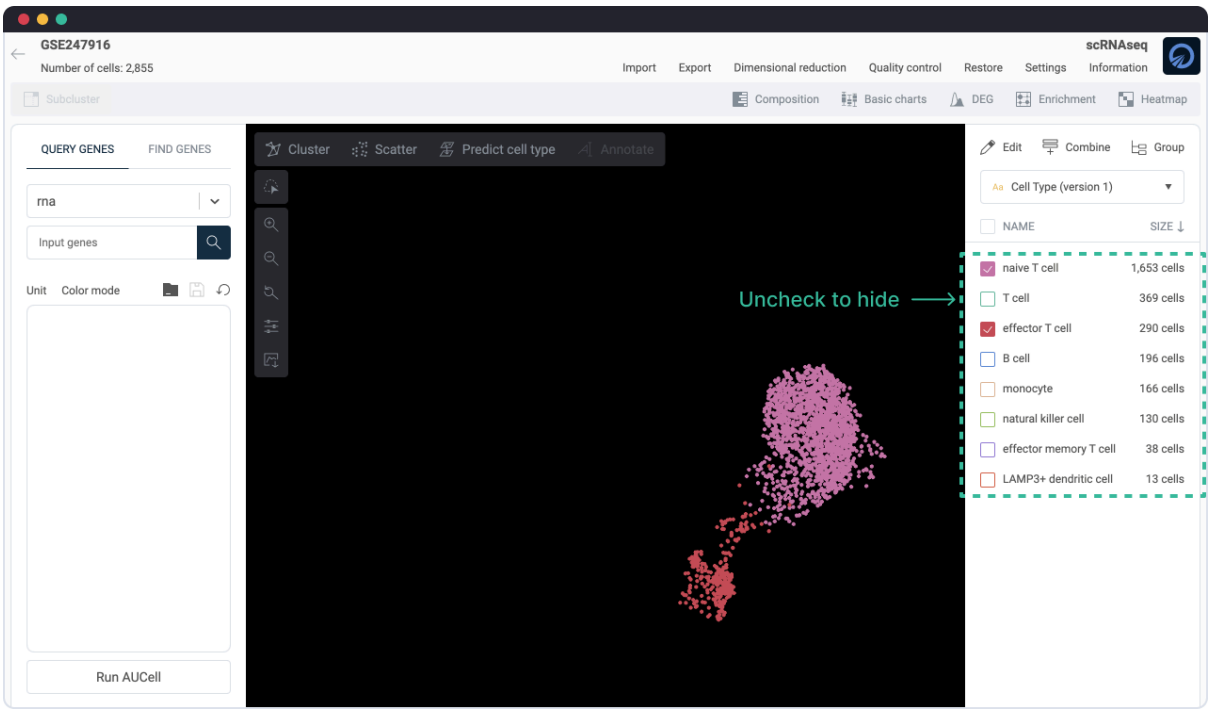
Each label in categorical metadata is represented by a distinct color for easy distinguishing. If the number of labels exceeds the number of colors in your selected color palette, the palette will repeat. You can sort the labels by name or number of cells by clicking on the table header. This only changes the order of label display on the screen and does not affect the actual order of labels in storage. To change the order of labels, you can use the edit metadata function described in the Edit and update section.



Clicking on a label name from the metadata table selects the corresponding cell population, which is necessary for other analyses. You can also select cell populations by clicking on the corresponding cell population on the cell scatter. Multiple labels can be selected by holding the `ctrl`/`cmd` buttons while clicking on

each metadata label or cell population on the scatter. Holding the `Shift` button selects multiple labels between the first and second clicks.



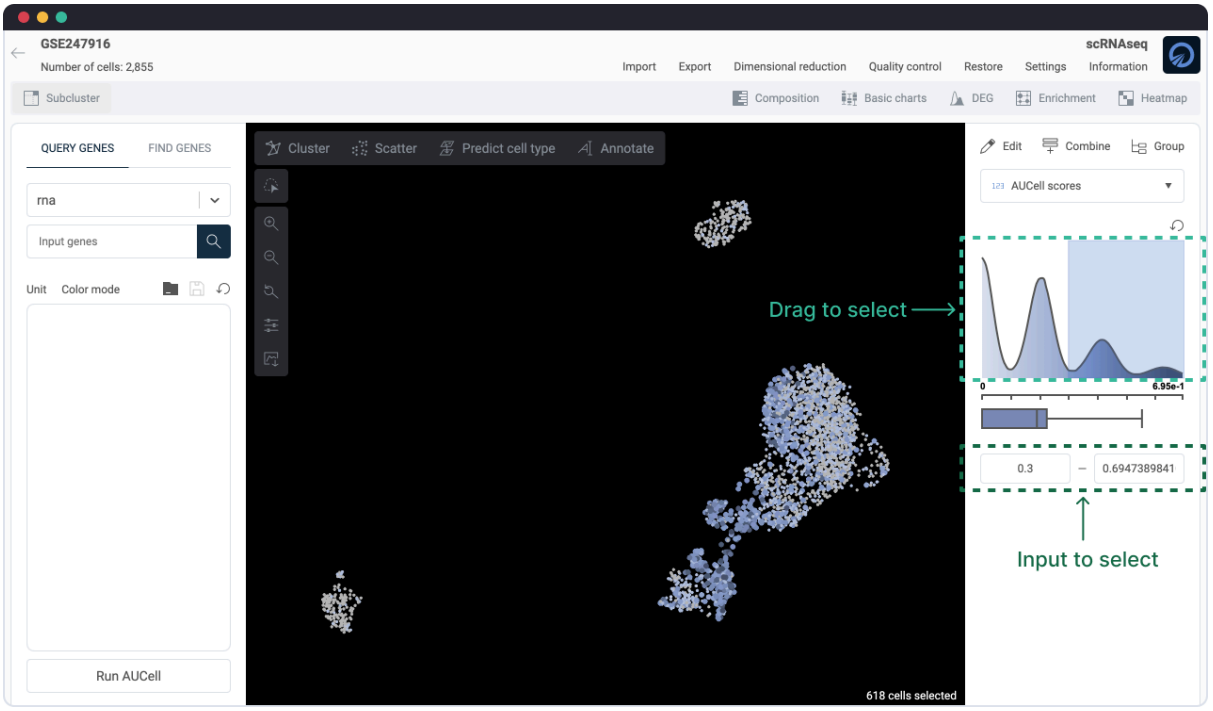


If cell clusters overlap, you can hide some cell populations for a clearer view by clicking on the checkbox before each metadata label. This only hides the cell

population on the screen and does not create sub-clusters. For sub-clustering, see the Create sub-clusters section.

4.3. Numerical metadata

In OmnibusX, cells can be assigned continuous values such as the number of genes, AUCell score, etc., which are represented as numerical metadata. The distribution of these values is visualized using a distribution plot, smoothed through Kernel Density Estimation (KDE). OmnibusX uses the implementation from the Scipy team: scipy.stats.gaussian_kde.



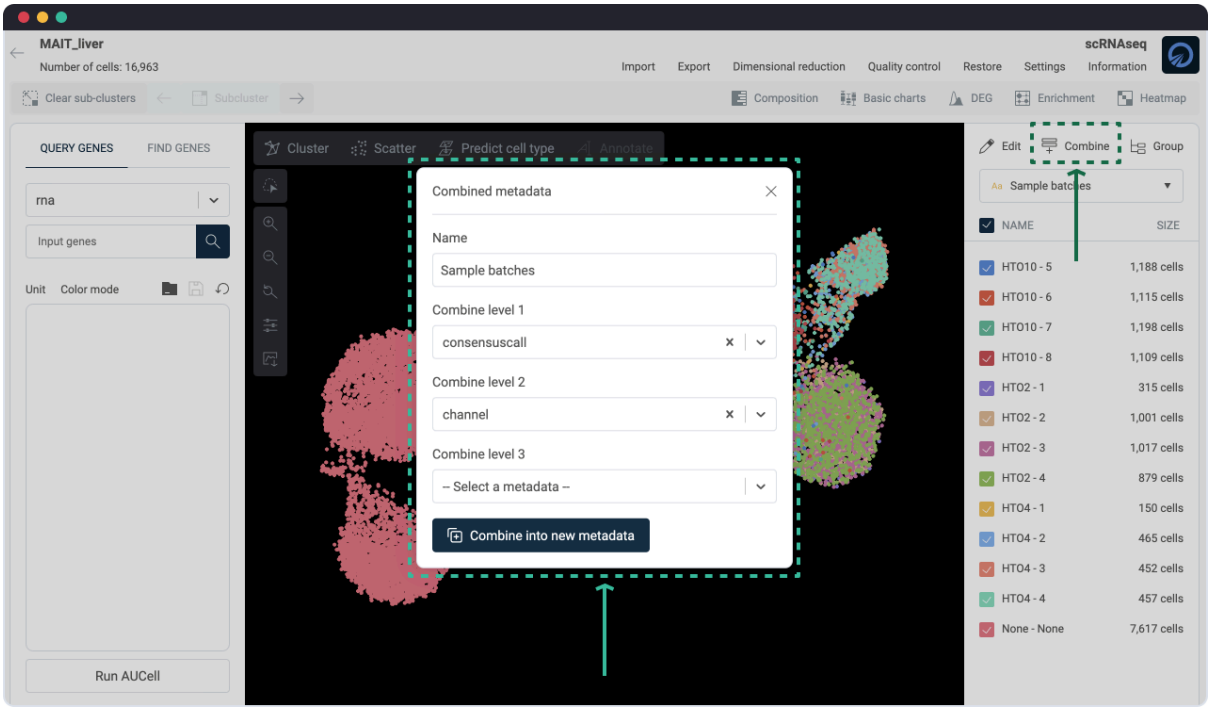
To select a specific cell population within a range of values, you have two options:

- `Interactive selection`: Click and drag directly on the distribution plot to highlight and select the desired region.
- `Precise input`: For more reproducible results, you can enter exact start and end values in the provided input boxes.

4.4. Combine metadata

In certain research scenarios, it may be necessary to synthesize information from multiple annotations, such as combining `Condition` and `Treatment` data. You can perform this function by following below steps:

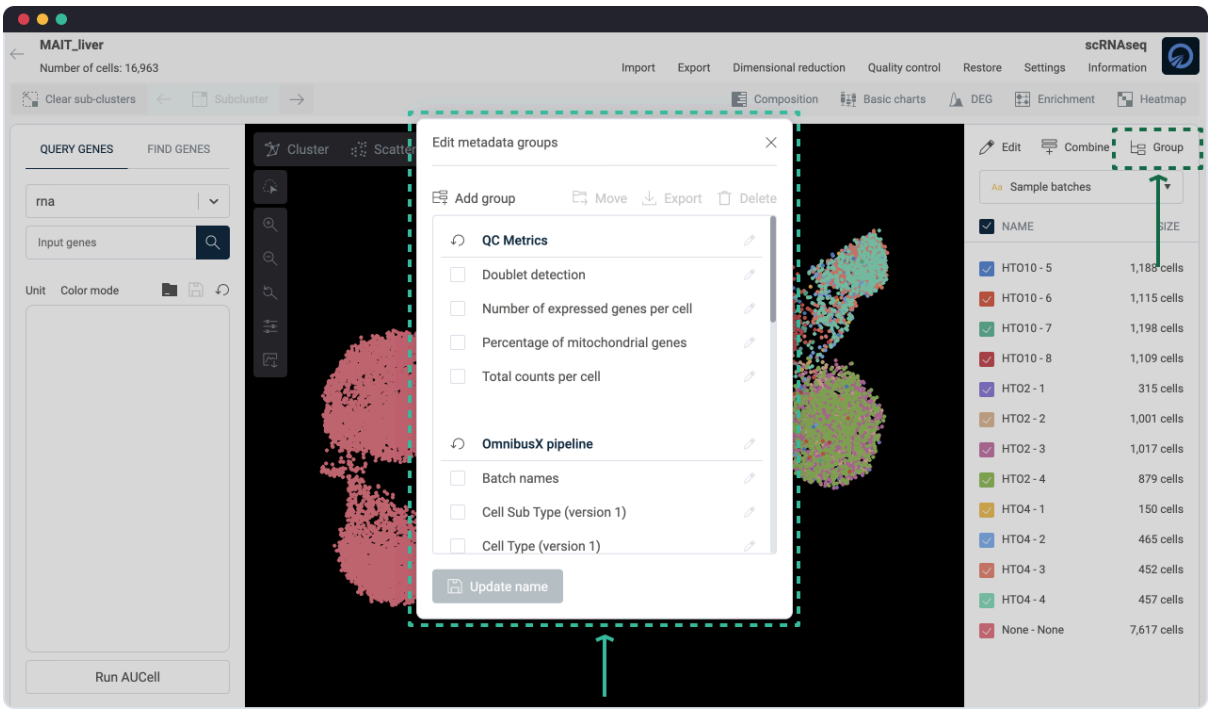
1. Click on the `Combine` button located at the top of the metadata panel.
2. Enter a name for the new combined metadata.
3. Select up to three different metadata that you want to combine together.



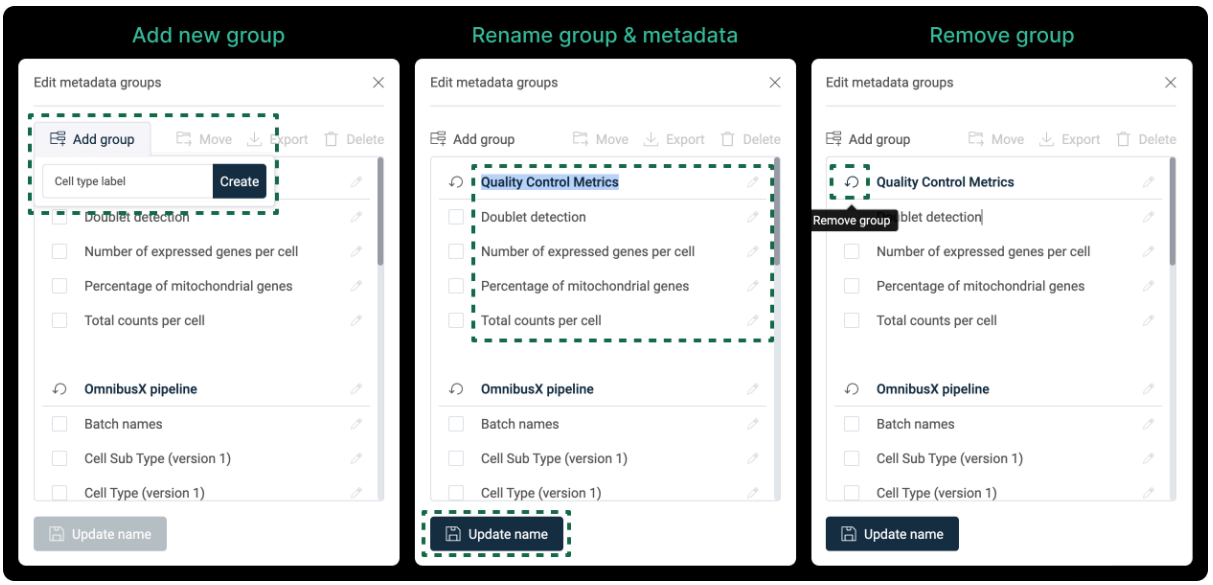
The new metadata, formed by integrating the selected categories, will be saved under the name you provide and appears as a new entry in the metadata panel.

4.5. Metadata group

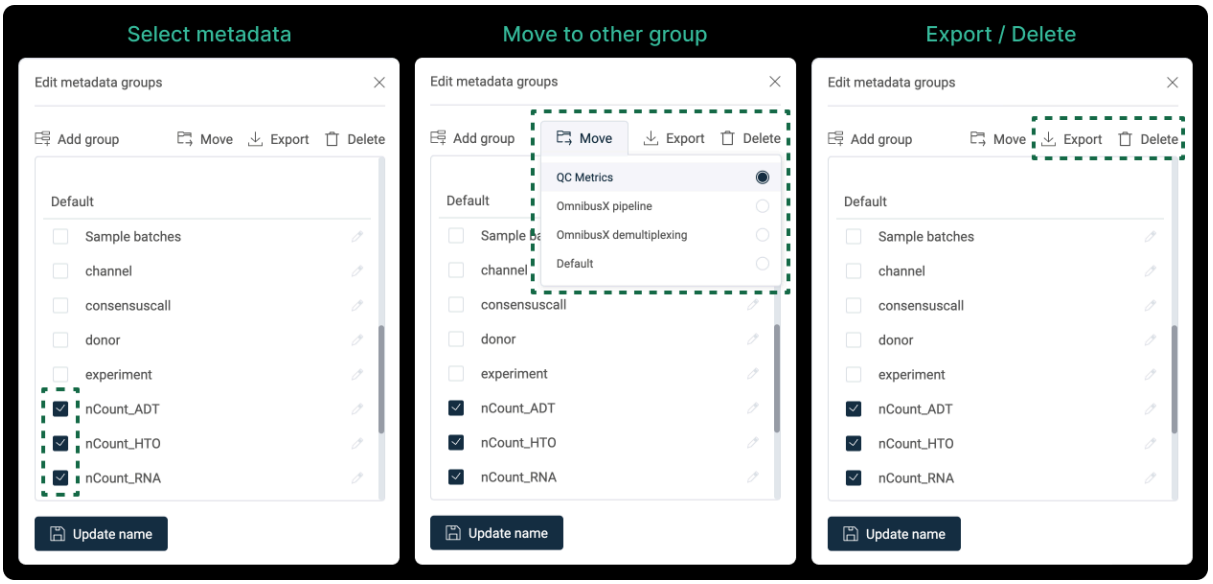
Grouping metadata into groups can help you better organize your annotations in OmnibusX. To manage your metadata groups, click on the `Group` button at the top of the metadata panel.



To create a new metadata group, click on the `Add group` button, input a group name, and then click `Create`. You can update the group names as well as metadata names by entering the new name in the corresponding input field and then clicking on the `Update name` button to save all changes. To remove a metadata group, click on the `Clear` icon at the beginning of each group name. All metadata in that group will be moved back to the `Default` group, which cannot be deleted.



You can select multiple metadata by clicking on the checkbox at the beginning of each metadata name. Once selected, you can update their group by clicking on the `Move` button and selecting a new group to move them to. Additionally, you can export the selected metadata as a `.tsv` file by clicking on the `Export` button or delete the selected metadata by clicking on the `Delete` button. Deleted metadata can be restored within 30 days in case of accidental removal of important annotations.



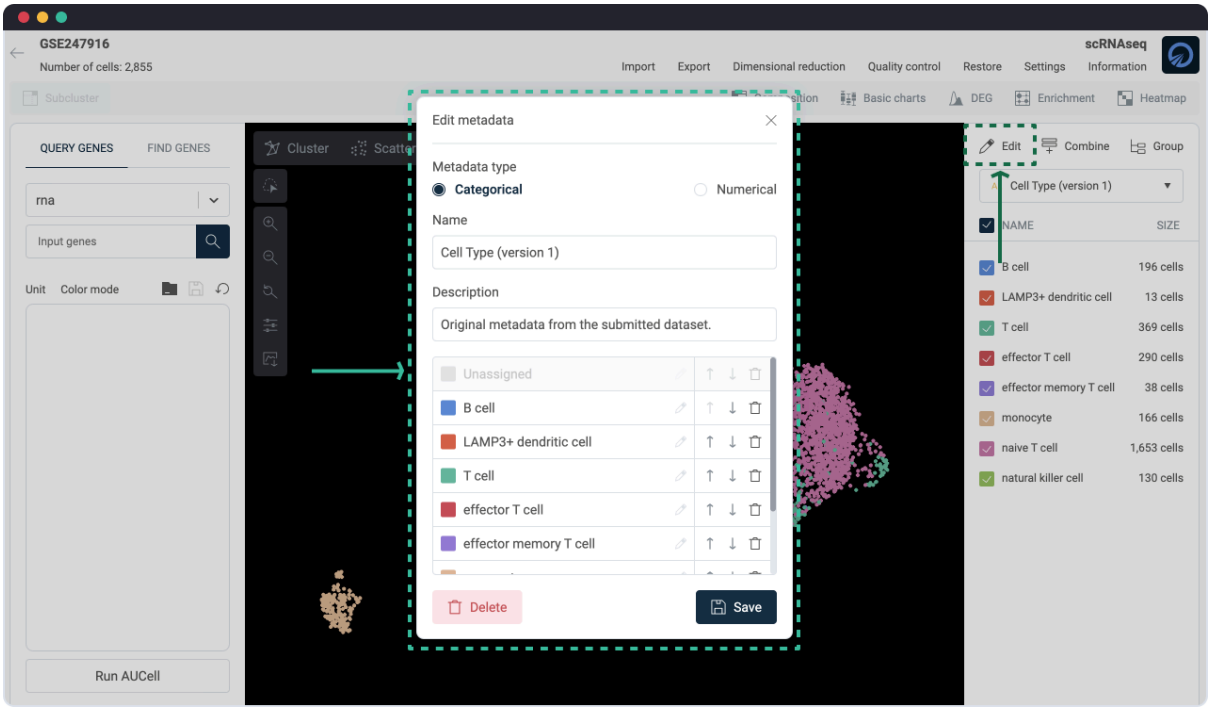
4.6. Edit a metadata

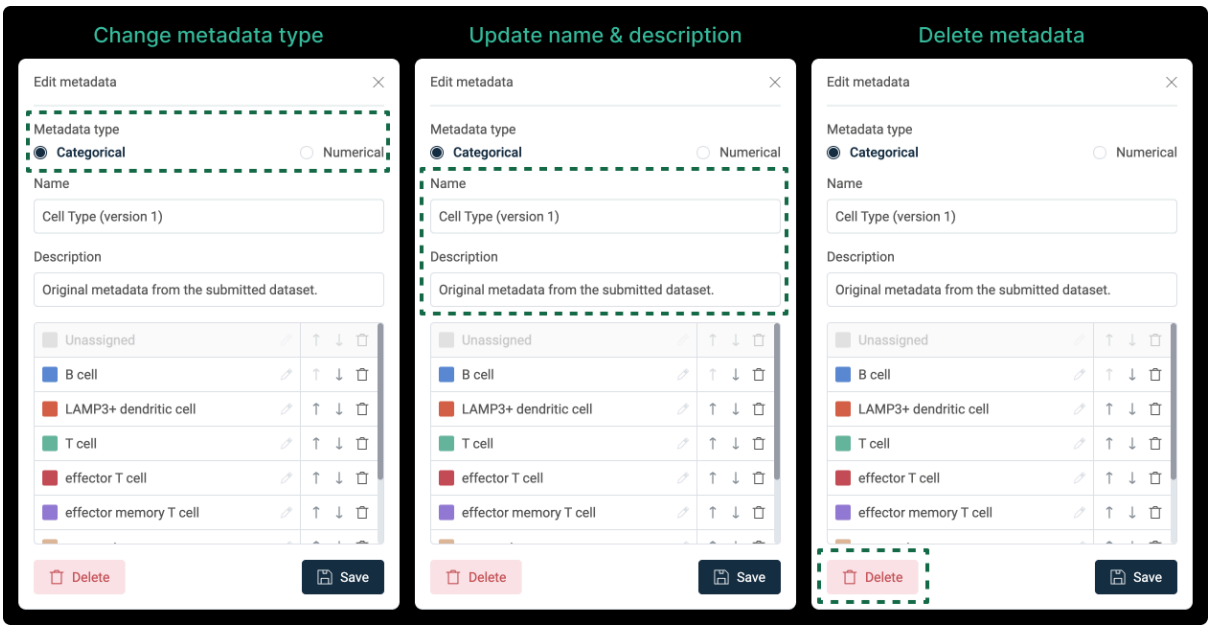
OmnibusX allows you to modify your current metadata by clicking on the `Edit` button at the top of the metadata panel.

Some metadata, such as age or height, can be represented either as distinct categories or as numerical values. OmnibusX enables you to switch between these types as needed. However, when converting numerical values to categories, ensure that the number of unique values does not exceed `1000` to avoid complications with handling a large number of categories.

You can also rename the metadata and update its description, which serves as an extra note to keep track of necessary information. To delete the current metadata,

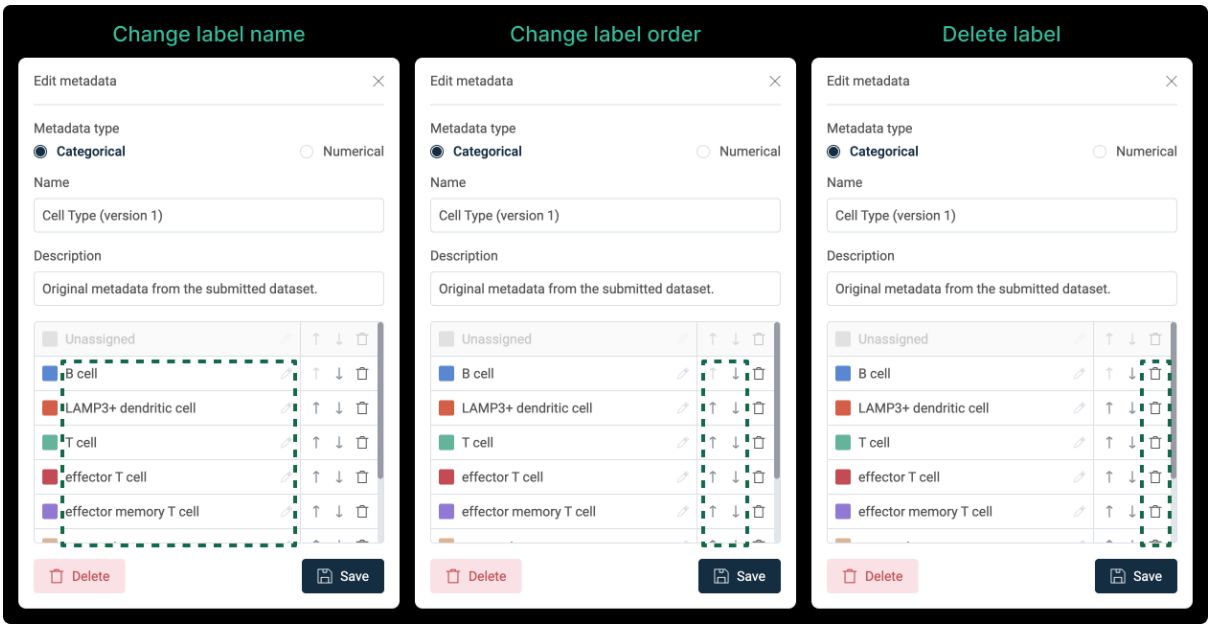
click on the `Delete` button at the bottom left corner. Deleted metadata can be restored within 30 days in case of accidental removal of important annotations.





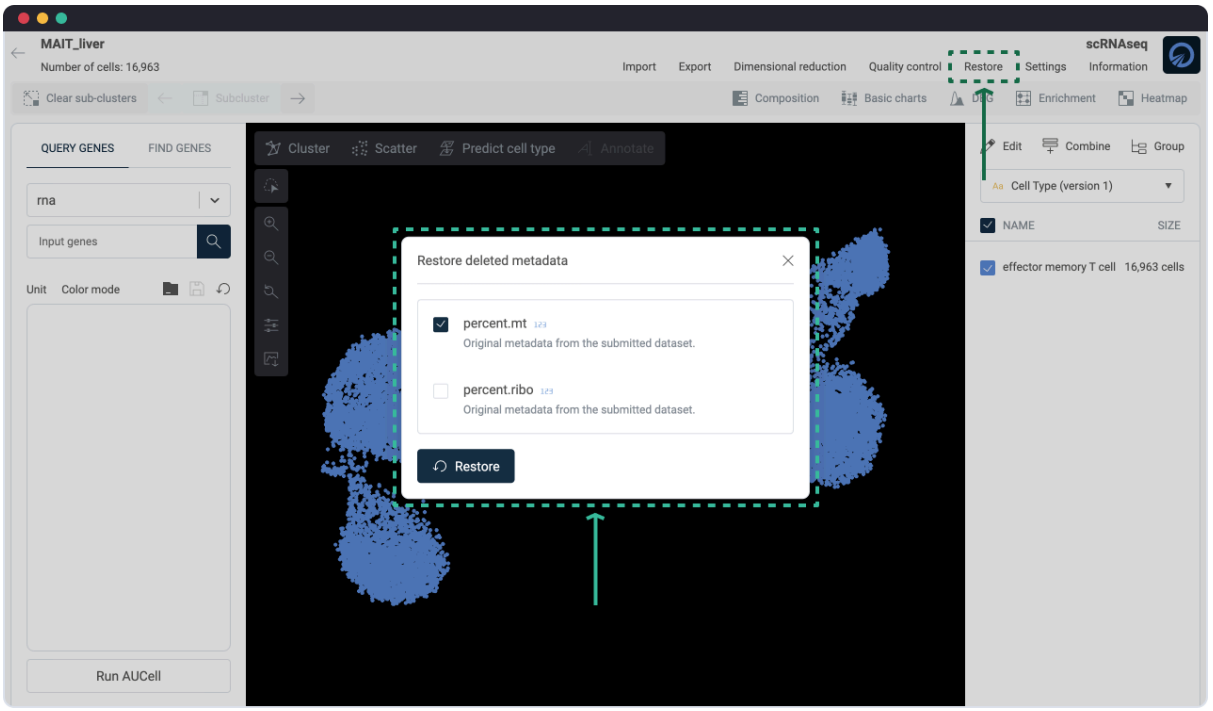
For categorical metadata, you can update label names by entering the new name in the corresponding input field. Adjust the order of labels by using the arrow buttons to move them up or down. This feature is particularly useful for grouping

related labels together for clearer comparative analysis in downstream plots. Remove unnecessary labels by clicking on the delete icon at the end of each label.



Finally, click on the Save button to save all your changes.

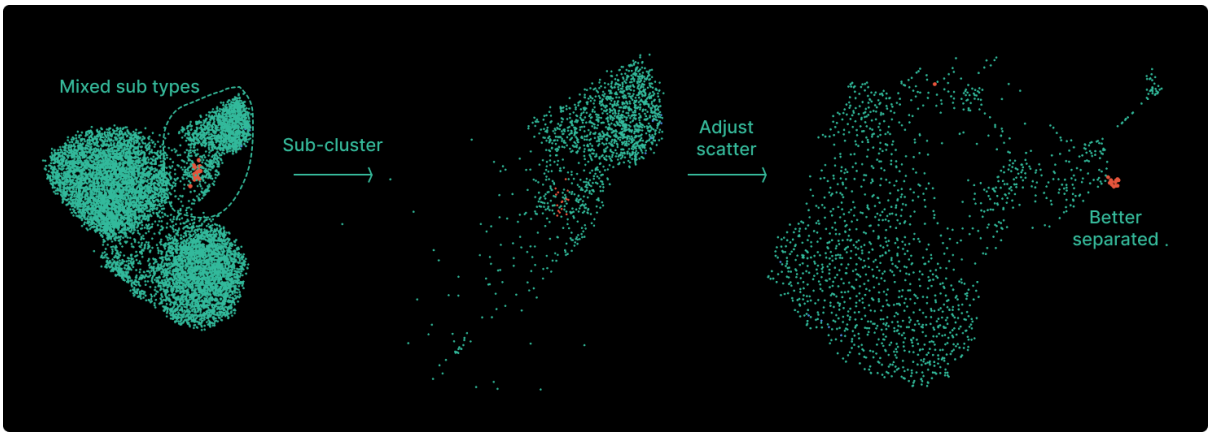
4.7. Restore deleted metadata



If you have accidentally deleted any metadata, you can restore it within 30 days from the date of deletion. To restore deleted metadata, click on the `Restore` button located on the top menu. A list of all deleted metadata will appear, and you can select the metadata you want to restore by clicking on the checkbox before the metadata name. Once you have selected all the metadata you want to restore, click on the `Restore` button to retrieve them. The restored metadata will be added back to your metadata panel, under the `Default` group.

5. Sub-cluster management

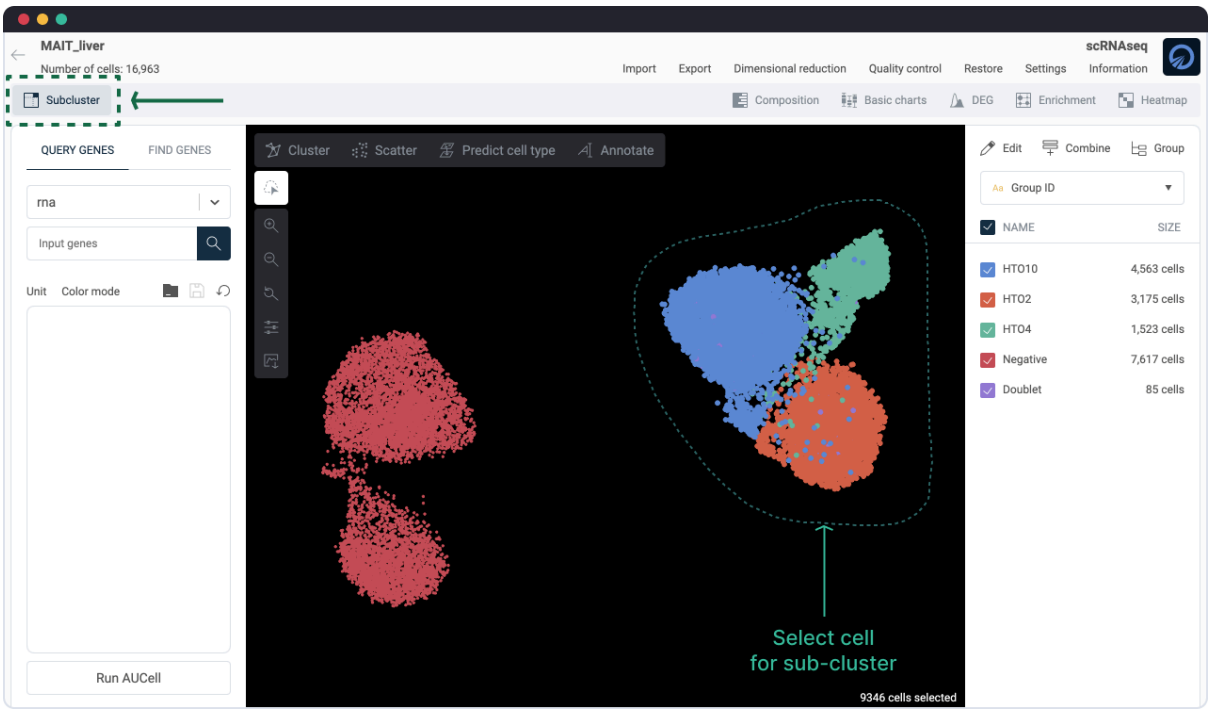
Sub-clustering is essential to explore the heterogeneity of each cell population and discover new cell subtypes, potentially rare with a small number of cells, which might otherwise be missed due to their proximity to closely related cell types in the global scope.



5.1. Create sub-clusters

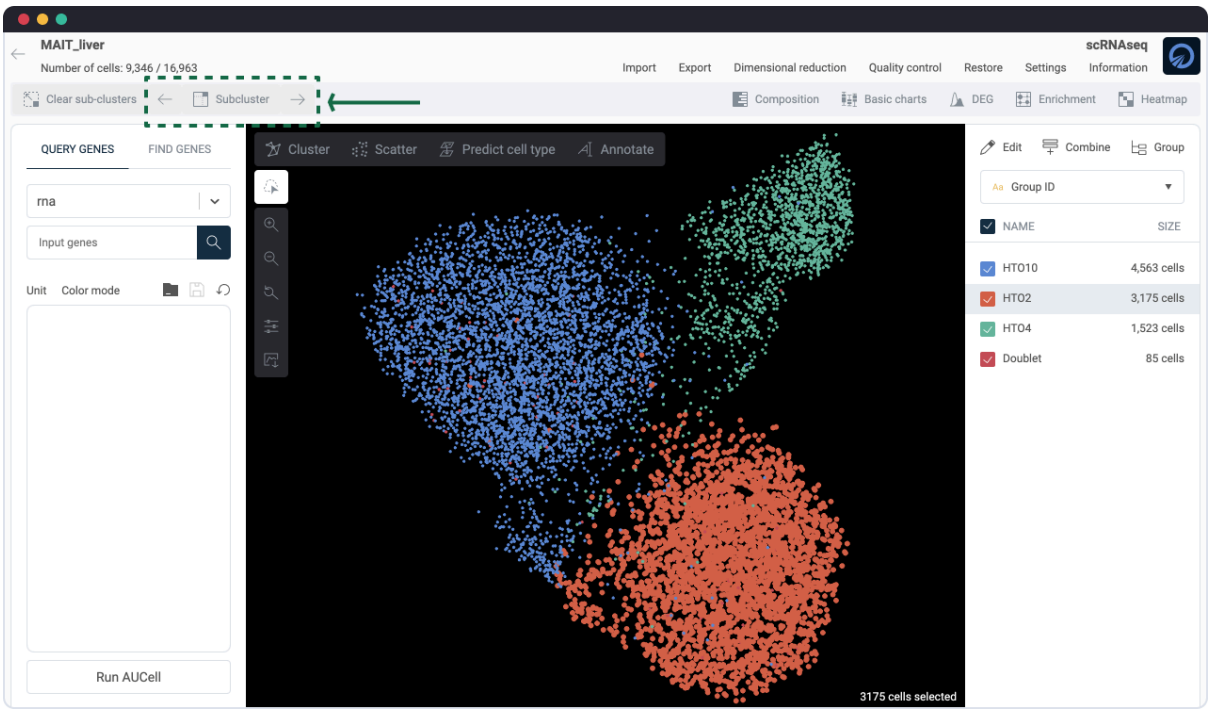
To initiate sub-clustering, first select a cell population of interest for further analysis. There are many ways for selecting a cell population, which are detailed in the Cell selection section. Once your population is selected, click on the `Sub-cluster` button located in the top right corner of the interface to subset the currently selected cells.

By default, the existing coordinates of the selected cells are used as the initial visualization scatter for the new sub-cluster. You can then adjust the scatter (as described in the t-SNE/UMAP section) to enhance separation between cell sub-populations. Other analyses, including clustering, 2D dimensional reduction, differential expression genes, composition, basic chart, heatmap, etc., will be applied only to this subset of cells.



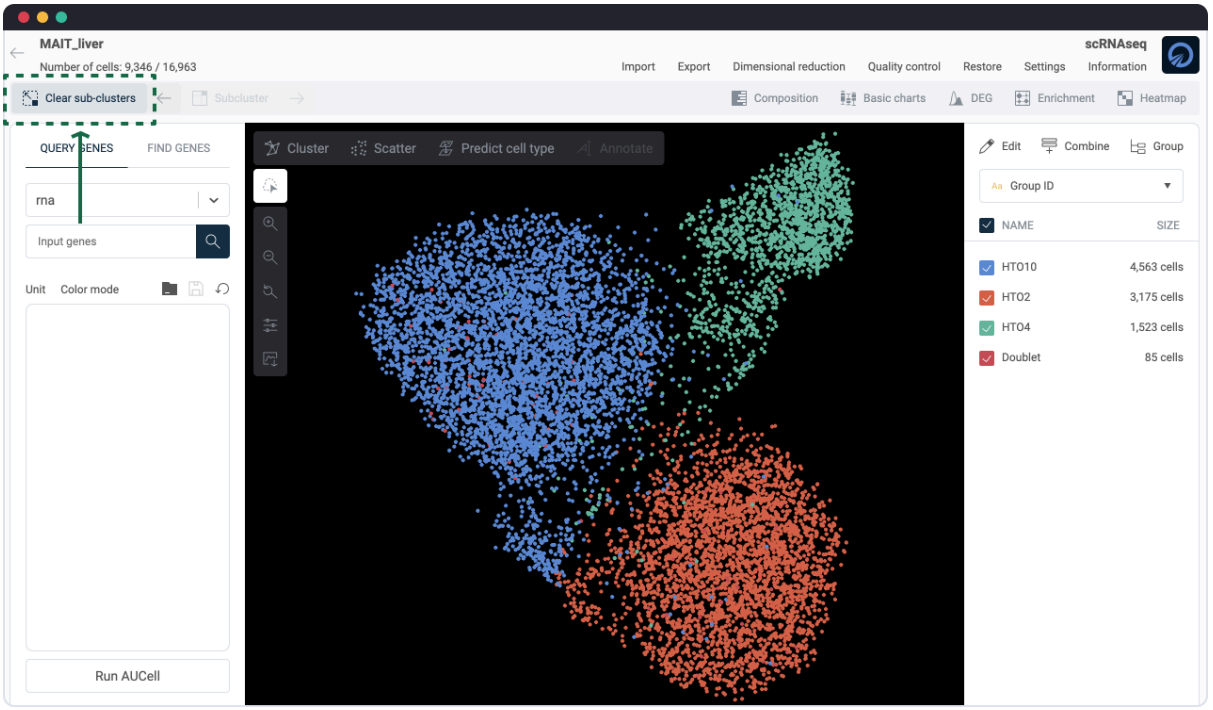
5.2. Multiple levels of sub-clustering

Sub-clustering can be performed multiple times, allowing you to dive deeper into subsets of cells based on your analytical needs. You can navigate between sub-clusters by using the arrow buttons next to the `Sub-cluster` button. Creating a new sub-cluster from an existing one replaces all downstream sub-clusters with the new subset. The current working sub-cluster is preserved even when the application is closed and reopened, enabling you to continue your analysis without needing to reselect sub-clusters.



5.3. Remove all sub-clusters

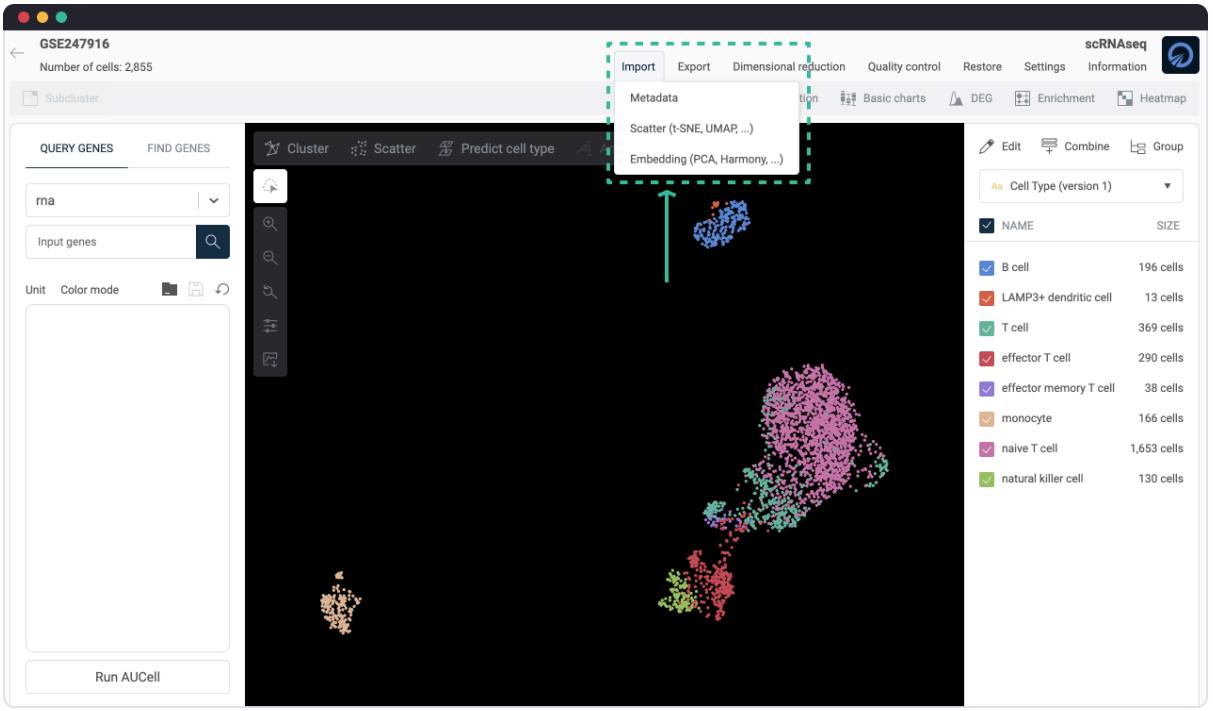
To reset your analysis and return to the full dataset, click on the `Clear sub-cluster` button located next to the `Sub-cluster` button.



6. Data management

6.1. Import additional data

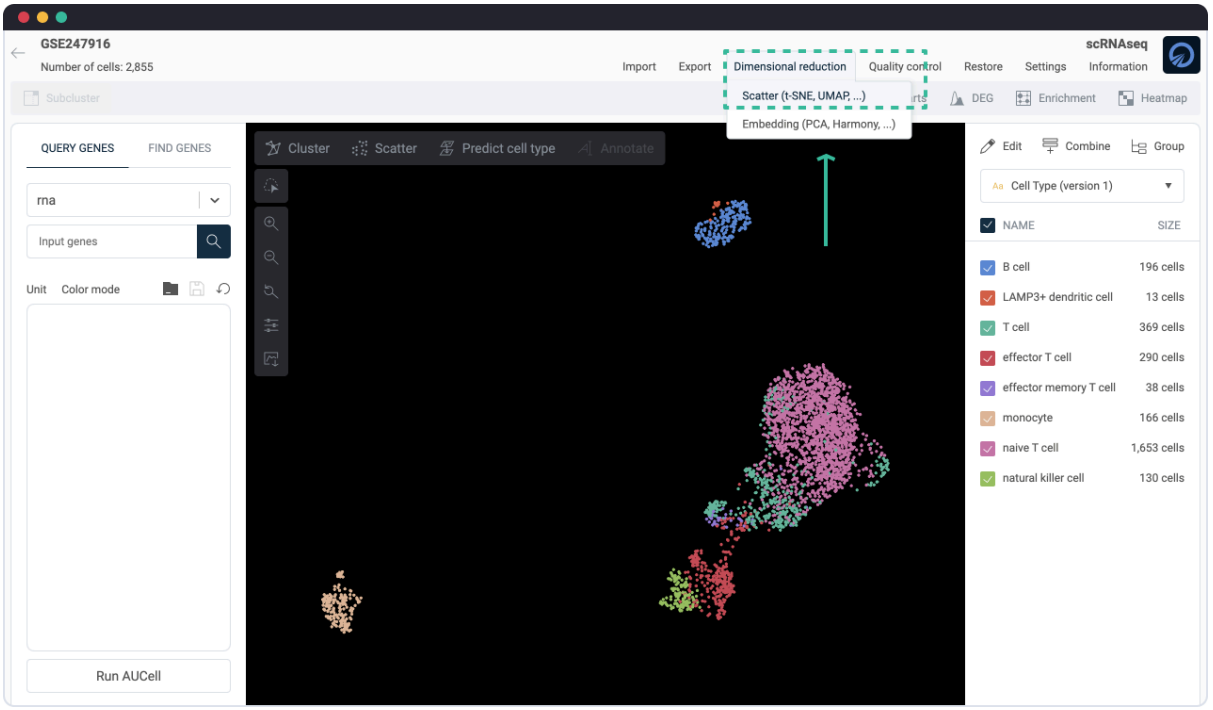
OmnibusX supports the importation of additional data to enrich your analysis:



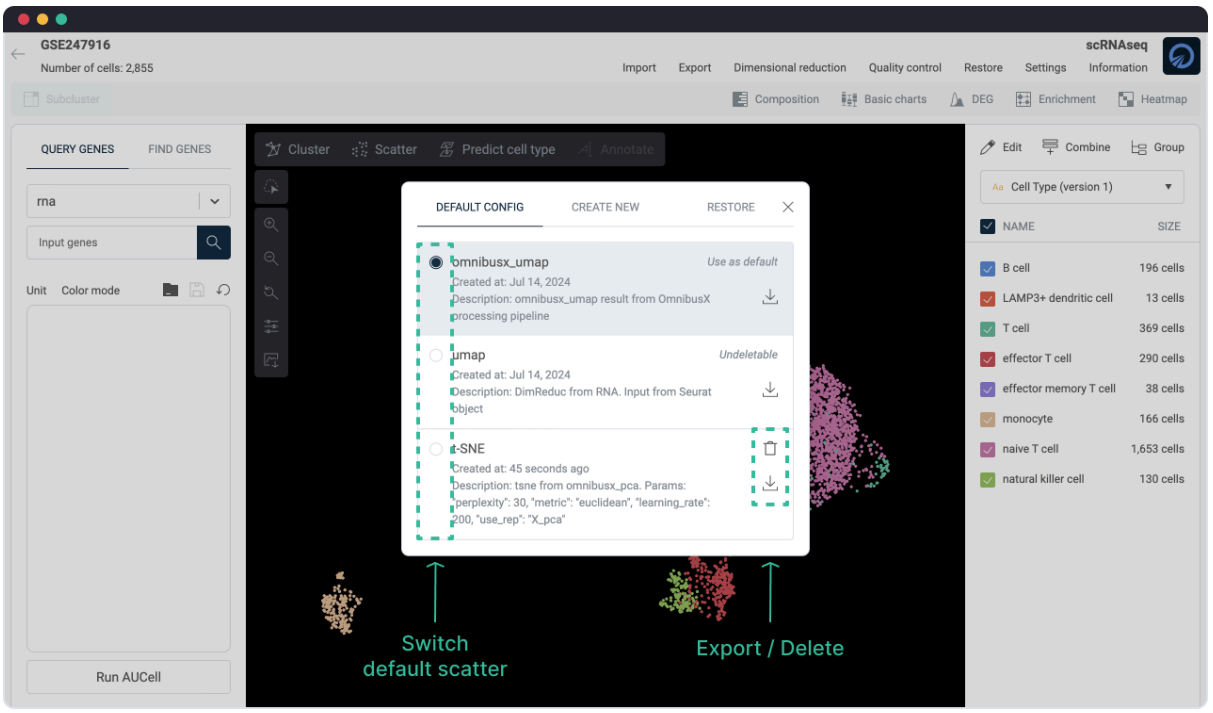
- `Metadata`: Accepts tabular files (`.tsv`, `.csv`, `.txt`) with the first column presenting cell barcodes and the first row containing metadata names. For datasets submitted from multiple batches, cell barcodes are modified to ensure no duplicates. To align your import correctly, you can export the current metadata from OmnibusX to retrieve the exact cell barcodes format required.
- `Scatter (t-SNE, UMAP, …)`: OmnibusX accepts tabular files (`.tsv`, `.csv`, `.txt`) that must contain three columns: the first column for cell barcodes and the next two columns for 2D scatter coordinates of each cell.
- `Embedding (PCA, Harmony, …)`: Similar to the above, OmnibusX accepts tabular files (`.tsv`, `.csv`, `.txt`) with the first column for cell barcodes and the remaining columns representing dimensions from dimensional reduction results.

6.2. Scatter management

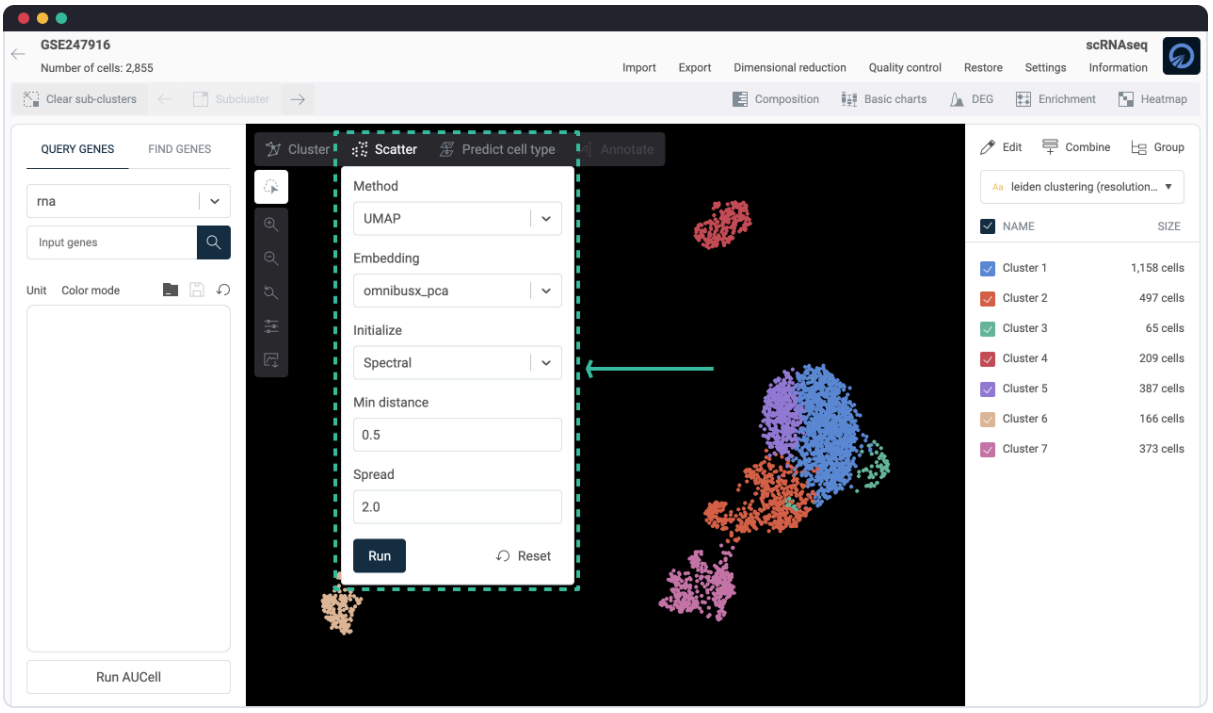
Each dataset can have multiple cell scatters for visualization, which can be sourced from either the submitted object (Scanpy, Seurat), imported as described above, or newly generated while analyzing your data with OmnibusX. To manage these results, click on the `Dimensional reduction` button on the top menu and select `Scatter (t-SNE, UMAP, ...)`.



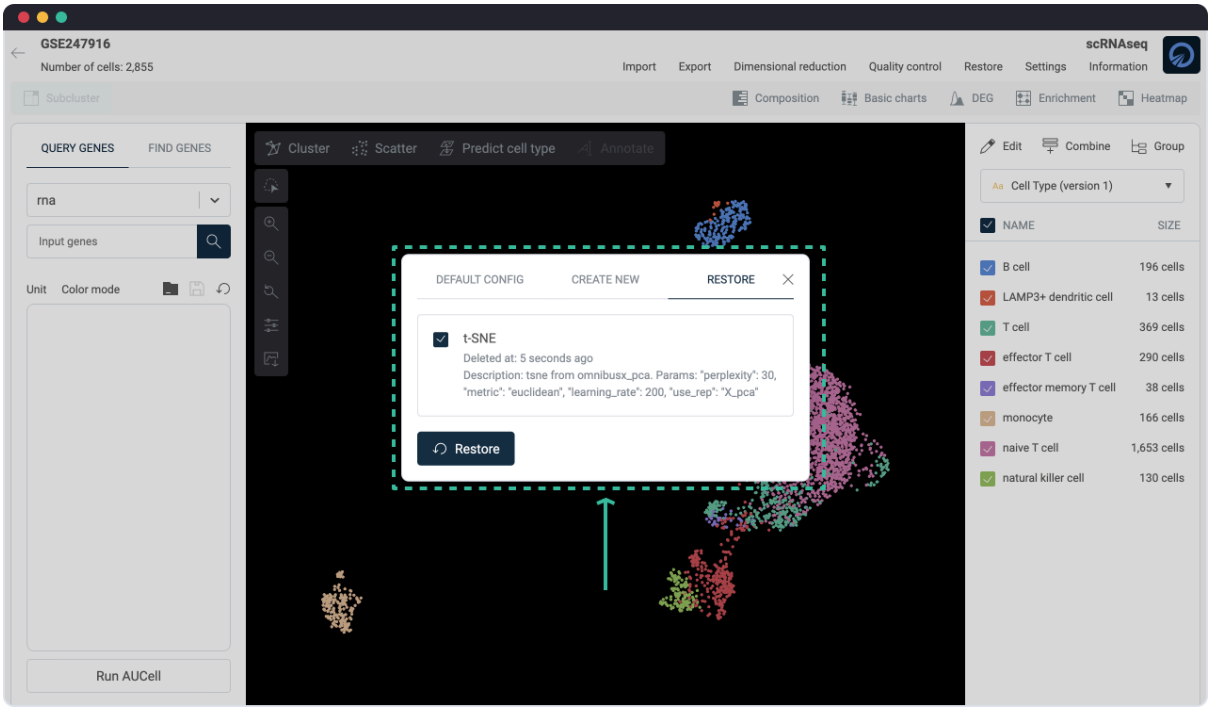
All available scatter results are listed in the first panel `Default config`. You can click on the radio button before each scatter result to set it as the default display.



This default scatter will be used as the primary visualization for the dataset, which affects the visualization of the whole dataset only. When you are in a sub-cluster, it does not modify your current sub-cluster scatter. Only when you click on the `Reset` button in the `Scatter` panel (as described in the t-SNE/UMAP section), the default scatter coordinate will be used instead.

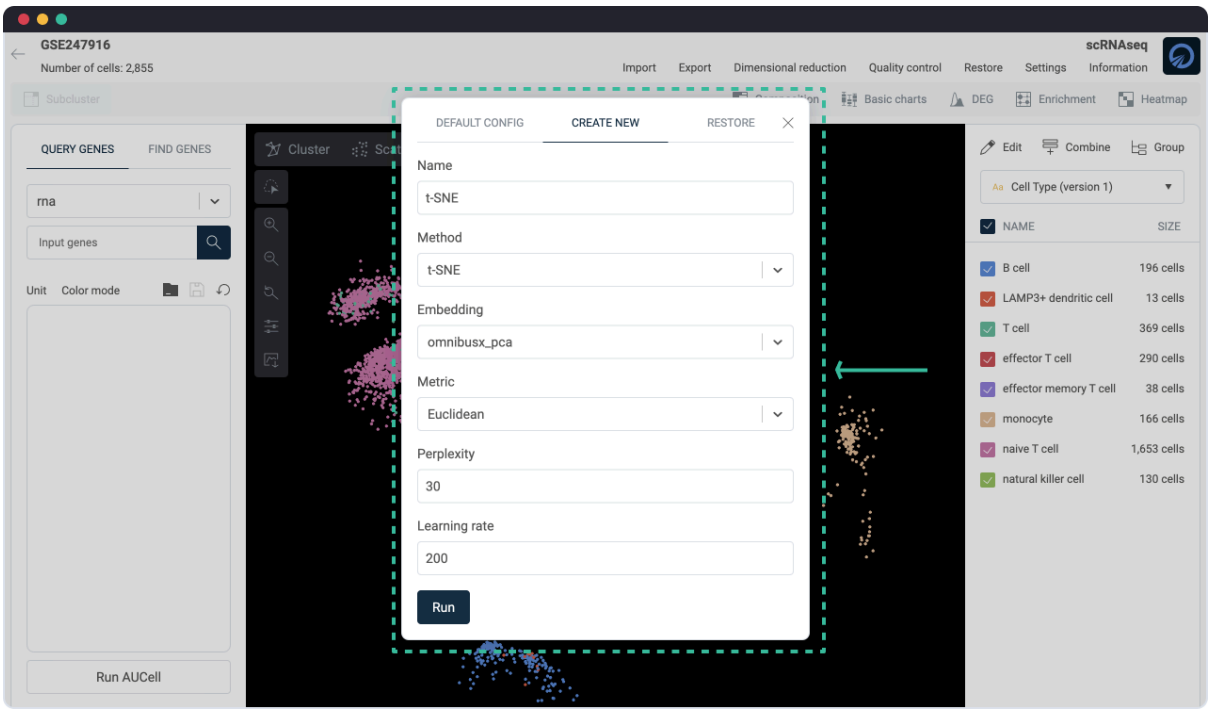


To remove a saved scatter result, click on the delete icon on each result item. Deleted results can be restored within 30 days in case of accidental removal. To restore deleted results, switch to the `Restore` panel, click on the checkbox before the needed result, and then click on the `Restore` button.



Additionally, you can generate new scatter coordinates directly within OmnibusX, which support two methods:

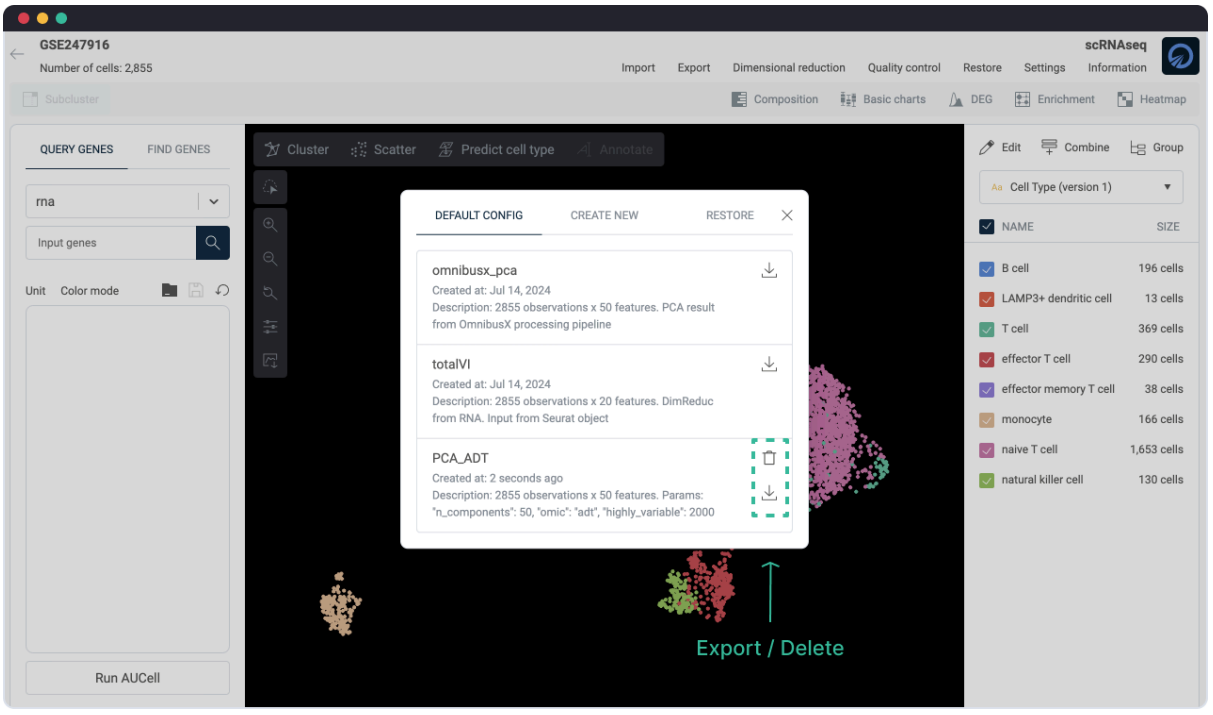
- `t-SNE` (*Amir et al., 2013*; *Pedregosa et al., 2011*; *van der Maaten and Hinton, 2008*): This method optimizes the distribution of nearest-neighbor distances in the embedding to closely match those in the high-dimensional space. OmnibusX employs the t-SNE implementation provided by the Scipy team: scanpy.tl.tsne.
- `UMAP` (*McInnes et al., 2018*): Besides faster than t-SNE, UMAP optimizes the embedding to best reflect the data's topology. The implementation used by OmnibusX is also sourced from the Scipy team: scanpy.tl.umap.



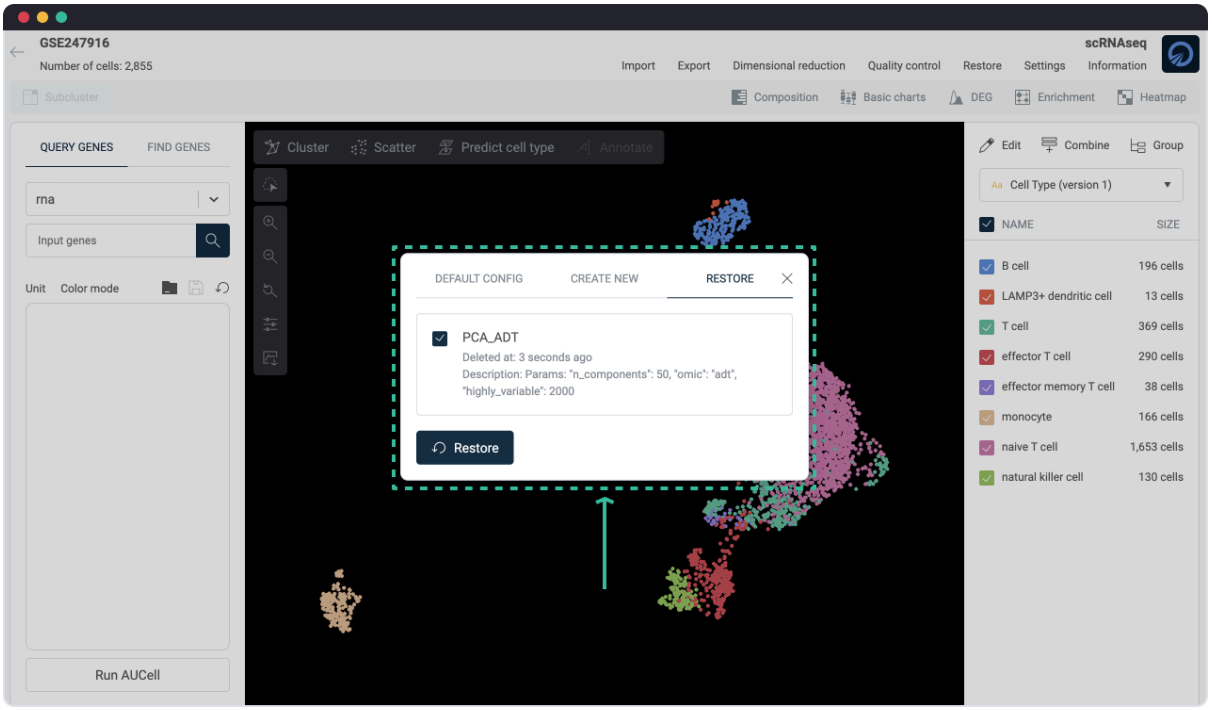
6.3. Embedding management

Each dataset in OmnibusX can have multiple cell embeddings (e.g., PCA, Harmony). To manage all of these results, click on the `Dimensional reduction` button on the top menu and select `Embedding (PCA, Harmony, etc.)`.

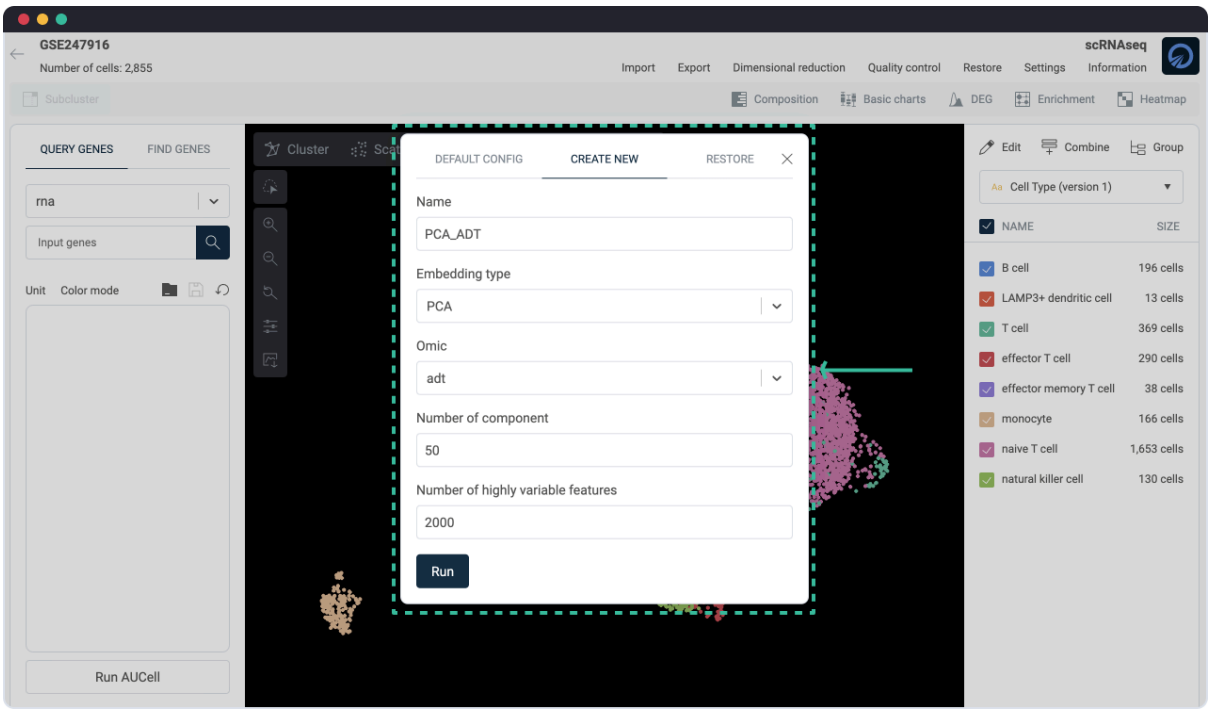
In the first panel, `Default config`, all available results are listed. These results can be used for other analyses, including clustering, generating scatter plots, cell type prediction, and more. To remove a saved embedding result, click on the delete icon on each result item.



Deleted results can be restored within 30 days in case of accidental removal. To restore deleted results, switch to the `Restore` panel, click on the checkbox before the needed result, and then click on the `Restore` button.



Besides importing existing embeddings, OmnibusX allows you to generate new embeddings using three methods:



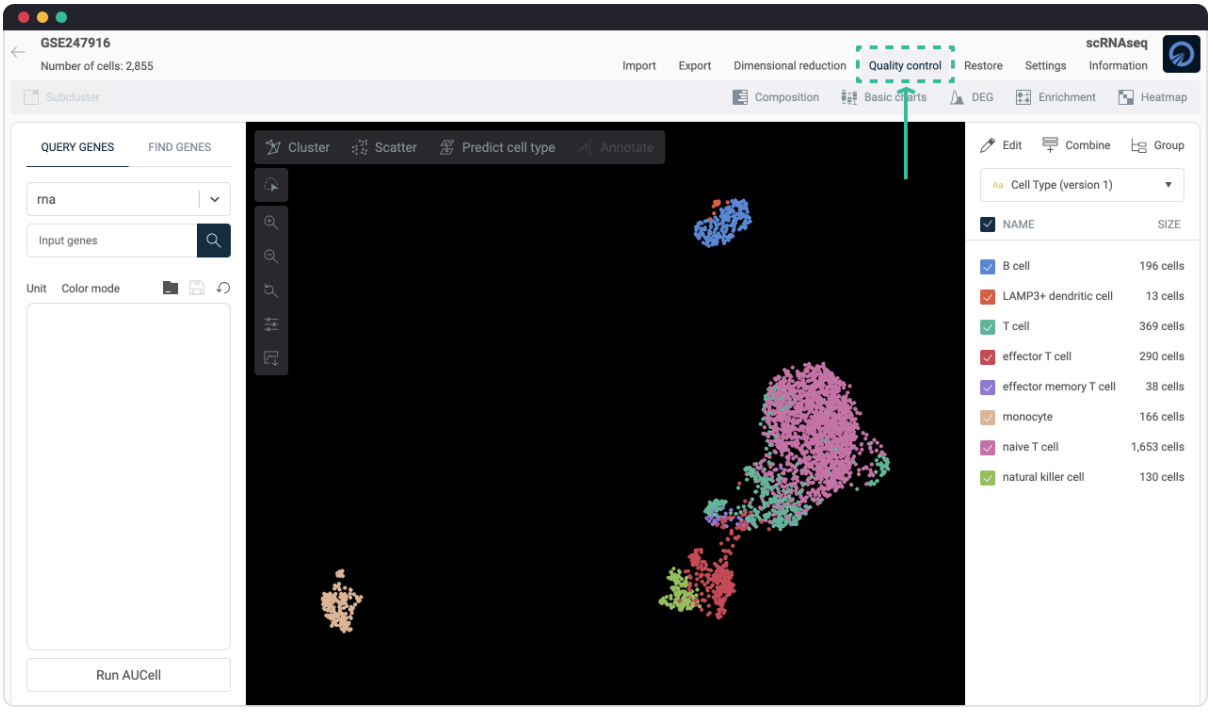
- `PCA` (Principal Component Analysis): PCA reduces data dimensionality by identifying principal components that capture the greatest variance in the data. OmnibusX uses the PCA implementation from the Scanpy team: scanpy.pp.pca
- `LDA` (Linear Discriminant Analysis): This method finds a linear combination of features that characterizes or separates two or more classes of objects or events. The LDA implementation is sourced from scikit-learn: LinearDiscriminantAnalysis, suitable for visualizing datasets with known labels.
- `Batch effect removal`: Allows you to remove batch effects based on an existing metadata (each label presents a batch). OmnibusX supports two methods for batch correction:

    1. `Combat`: Adjusts for batch effects using linear models, improving statistical power through an empirical Bayes framework that leverages information across genes. Combat is implemented by the Scanpy team: scanpy.pp.combat.
    2. `Harmony`: Designed to integrate multi-sample datasets, Harmony adjusts data to minimize discrepancies between batches. OmnibusX

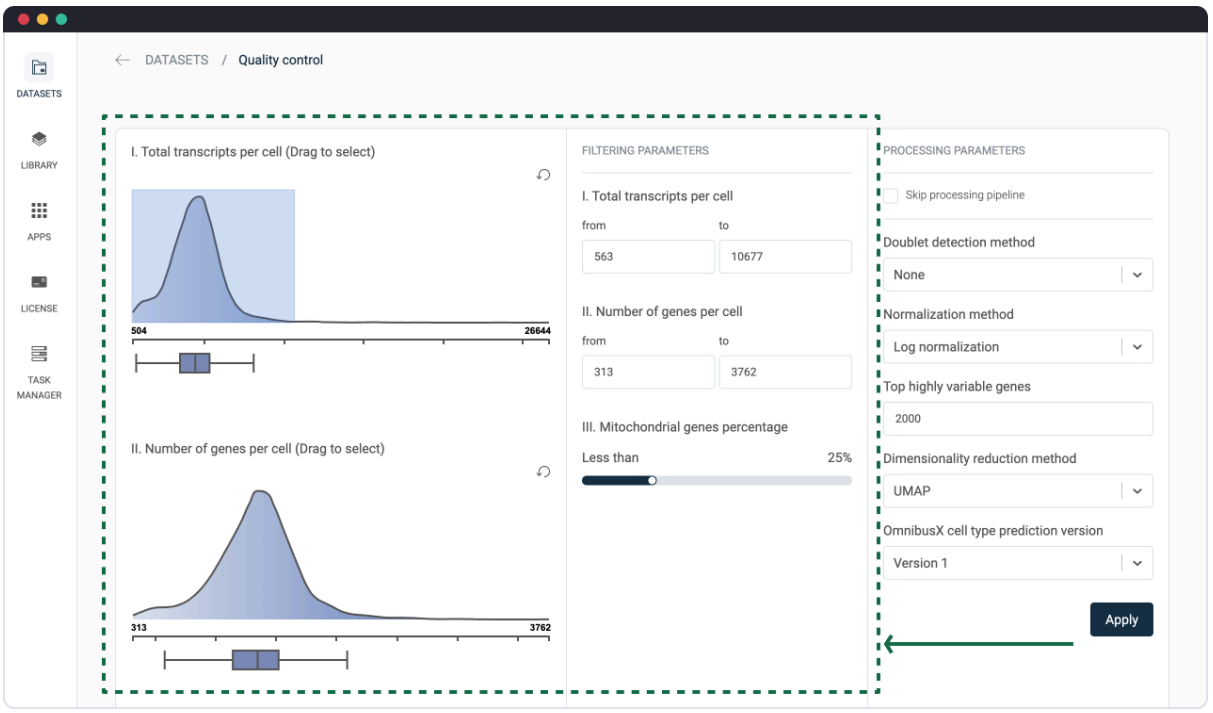
uses the Python implementation by Kamil Slowikowski, available here.

You can generate different embeddings from various available omics in your dataset, facilitating the simultaneous analysis of multimodal data.

6.4. Quality control



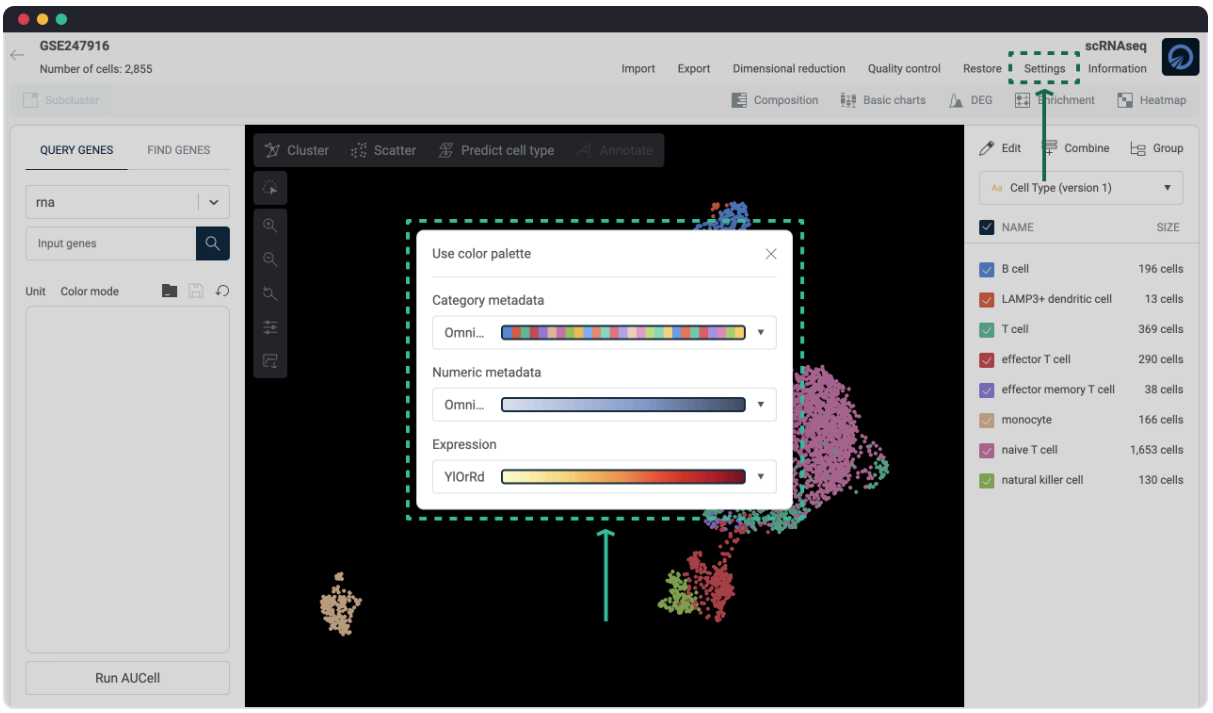
Maintaining high-quality data is critical for reliable analyses. If you find that the cell filtering parameters applied during a previous round of quality control did not yield satisfactory results—either over-filtering or under-filtering—you have the option to adjust these settings. To re-perform the quality control step, click on the `Quality control` button located in the top menu. This action will reopen the quality control panel for re-submitting.



It's important to note that submitting a new round of quality control will invalidate all current analysis results, including new annotations, differential expression genes (DEGs), heatmaps, and enrichment analyses. These results will be removed because they depend on the composition and characteristics of the filtered cell population.

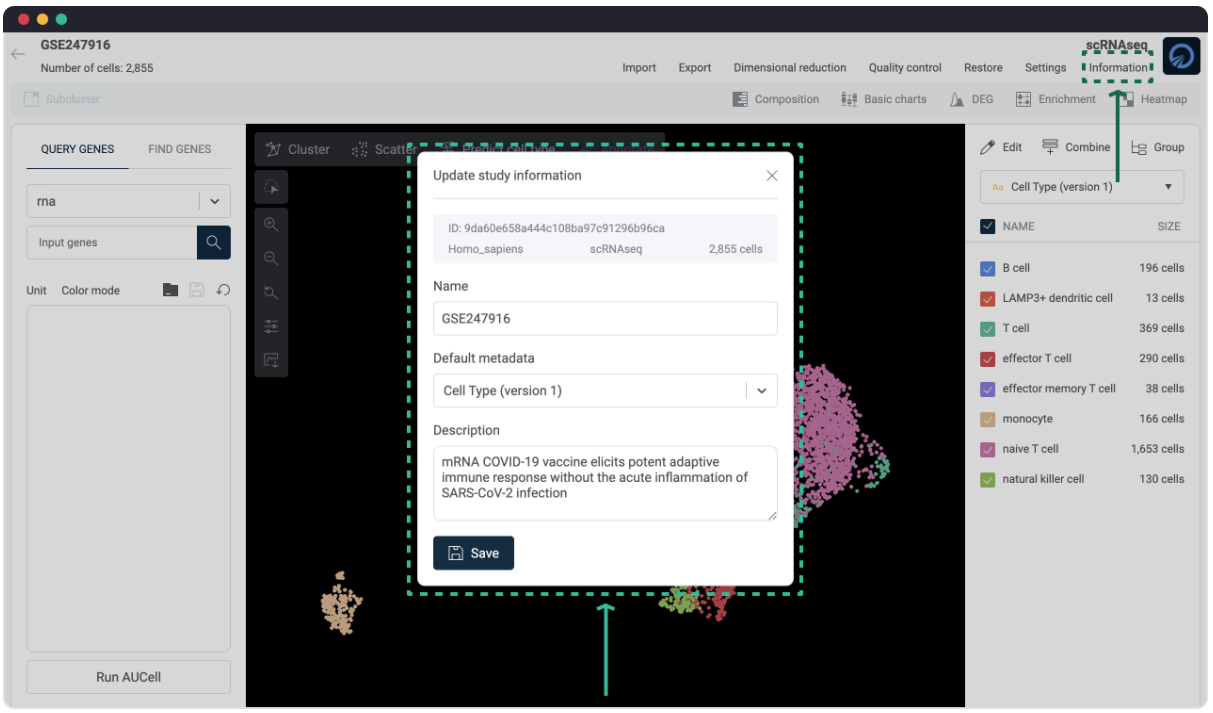
6.5. Default color palette

OmnibusX uses three default color palettes to visually represent different types of data: categorical metadata, numerical metadata, and expression values. To change these palettes to your preferred ones, simply click on the `Settings` button located on the top menu. OmnibusX offers a wide variety of color palettes for you to choose from. Additionally, you have the option to add your own custom color palette, which will be explained in detail in the Color palette library section.



6.6. Update dataset information

To modify your dataset's basic information, click on the 'Information' button located on the top menu. This will allow you to update the dataset title, description, and default metadata. The default metadata is used to initially color the cells when opening a dataset.

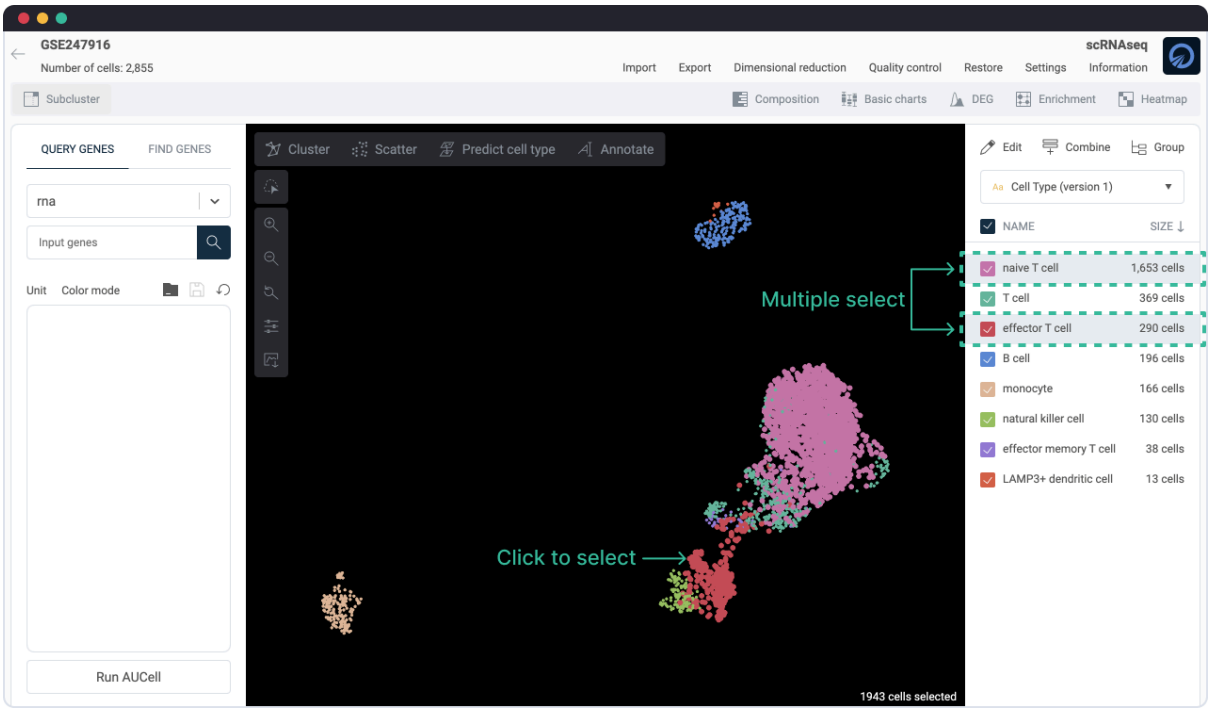

GSE247916
Number of cells: 2,855
Import
Export
Dimensional reduction
Quality control
Restore
Settings
scRNAseq
Information
Subcluster
Composition
Basic charts
DEG
Enrichment
Heatmap
QUERY GENES
FIND GENES
rna
Input genes
Unit
Color mode
Run AUCell
Cluster
Update study information
ID: 9da60e658a444c108ba97c91296b96ca
Homo_sapiens
scRNAseq
2,855 cells
Name
GSE247916
Default metadata
Cell Type (version 1)
Description
mRNA COVID-19 vaccine elicits potent adaptive immune response without the acute inflammation of SARS-CoV-2 infection
Save
Edit
Combine
Group
Cell Type (version 1)
NAME
SIZE
B cell 196 cells
LAMP3+ dendritic cell 13 cells
T cell 369 cells
effector T cell 290 cells
effector memory T cell 38 cells
monocyte 166 cells
naive T cell 1,653 cells
natural killer cell 130 cells

Cell annotation

1. Cell selection

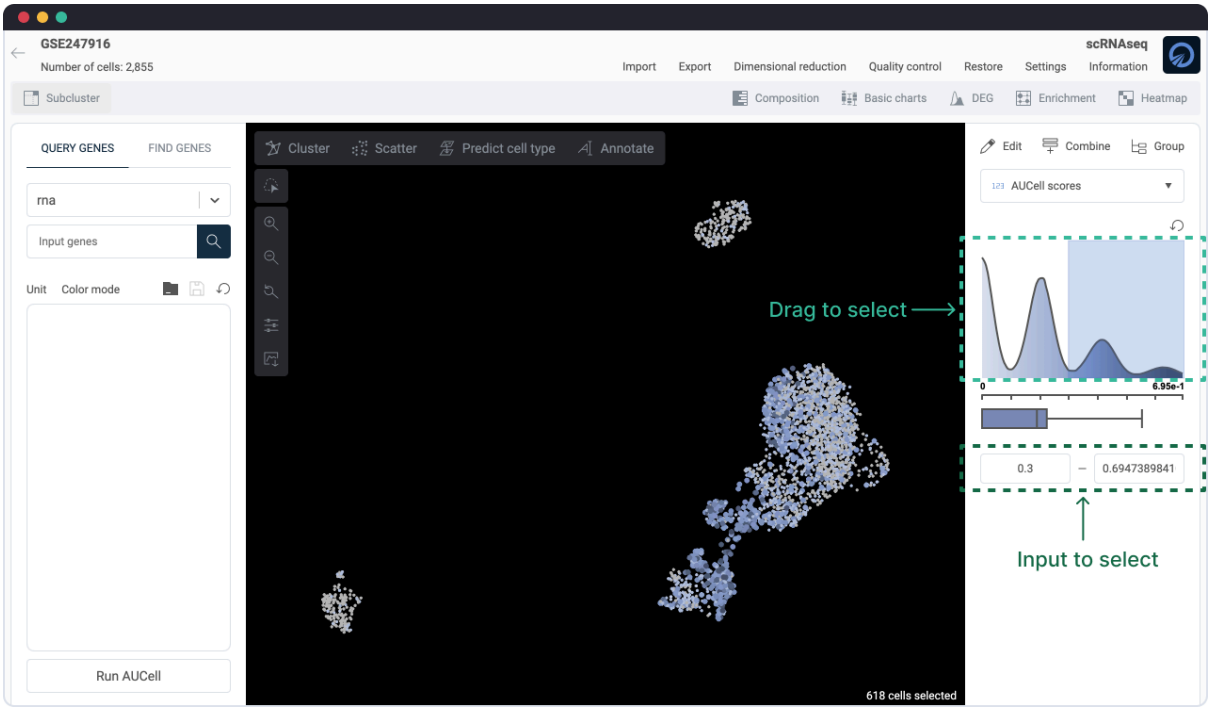
1.1. Select with existed metadata

For categorical metadata, clicking on a label name from the metadata table to select the corresponding cell population. You can also select cell populations by clicking on the corresponding cell population on the cell scatter. Multiple labels can be selected by holding the `ctrl`/`cmd` buttons while clicking on each metadata label or cell population on the scatter. Holding the `Shift` button selects multiple labels between the first and second clicks.



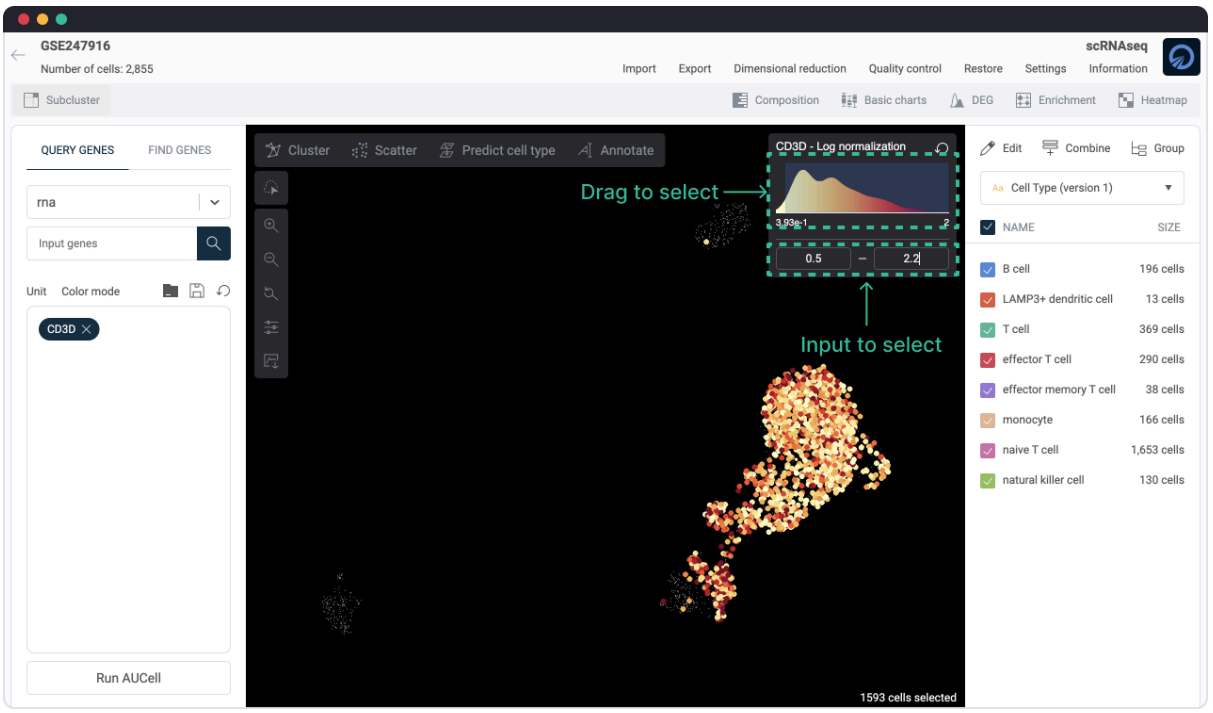
For numerical metadata, to select a specific cell population within a range of values, you have two options:

- `Interactive selection`: Click and drag directly on the distribution plot to highlight and select the desired region.
- `Precise input`: For more reproducible results, you can enter exact start and end values in the provided input boxes.

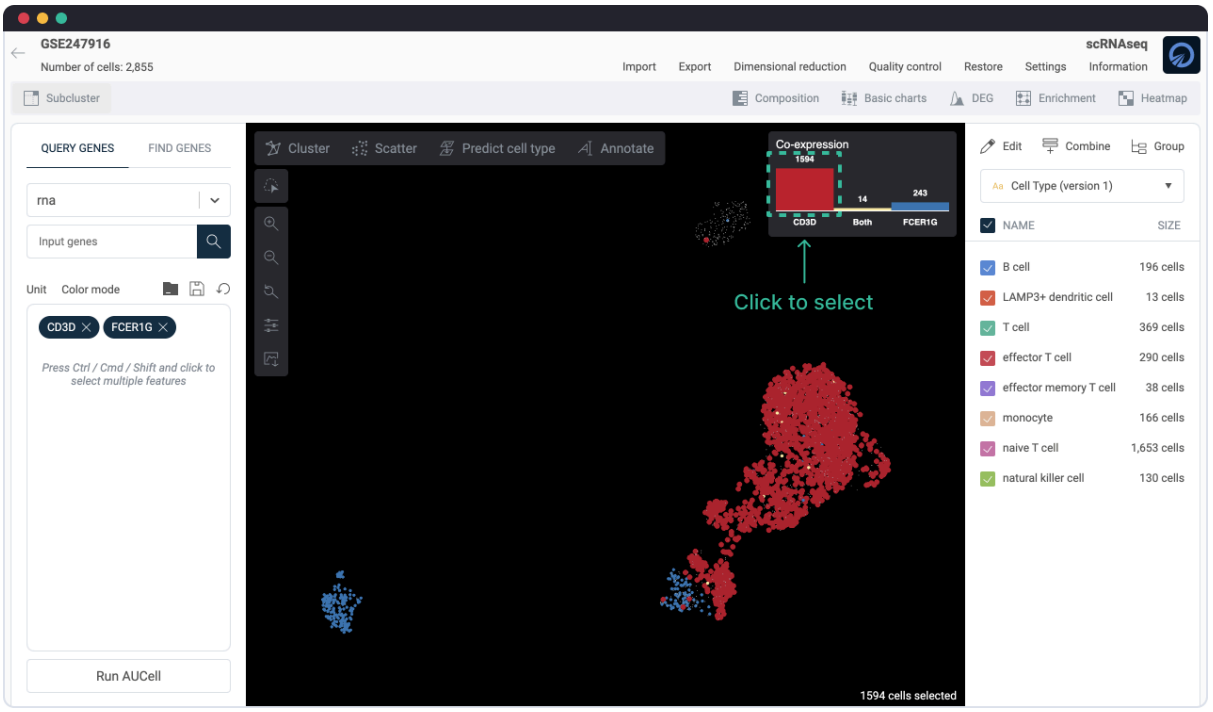


1.2. Select with expression values

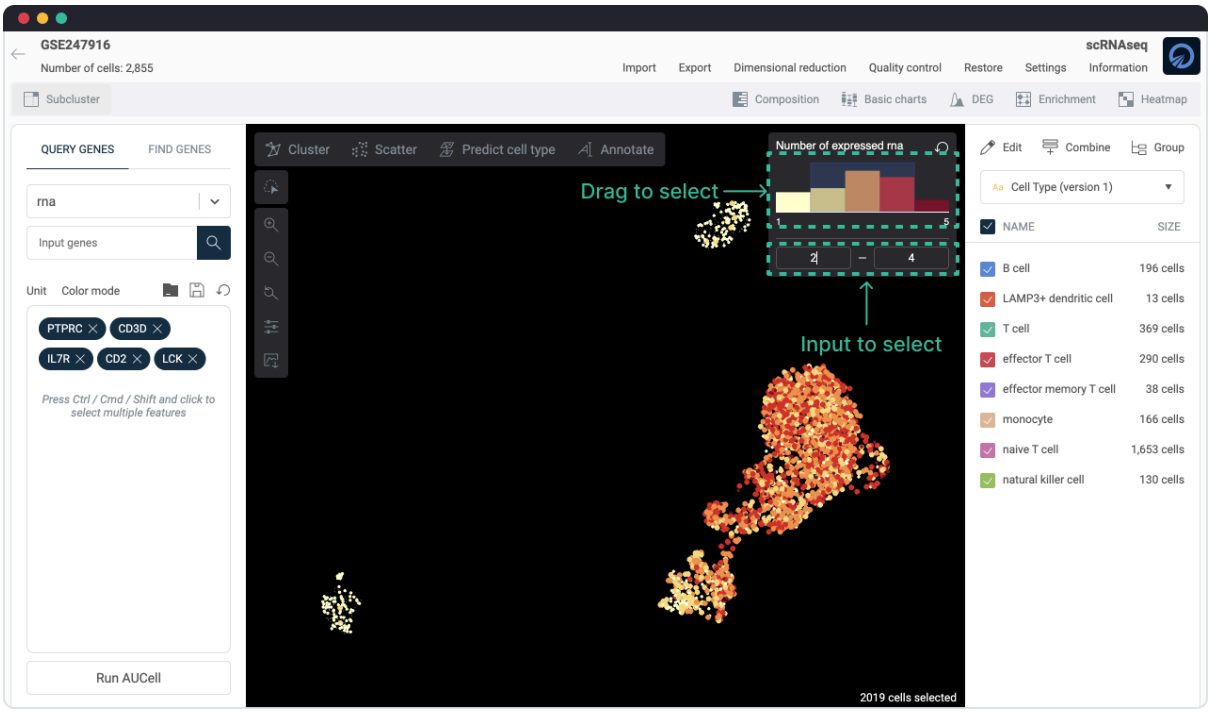
When querying the expression of a single gene, the distribution of expression values is visualized in the top right panel using a distribution plot, smoothed via Kernel density estimation (KDE), implementation from the Scipy team: [scipy.stats.gaussian_kde](). You can select cells within a specific range of expression values by dragging across the desired region on the distribution plot or by entering precise values into an input box for exact selection, enhancing reproducibility.



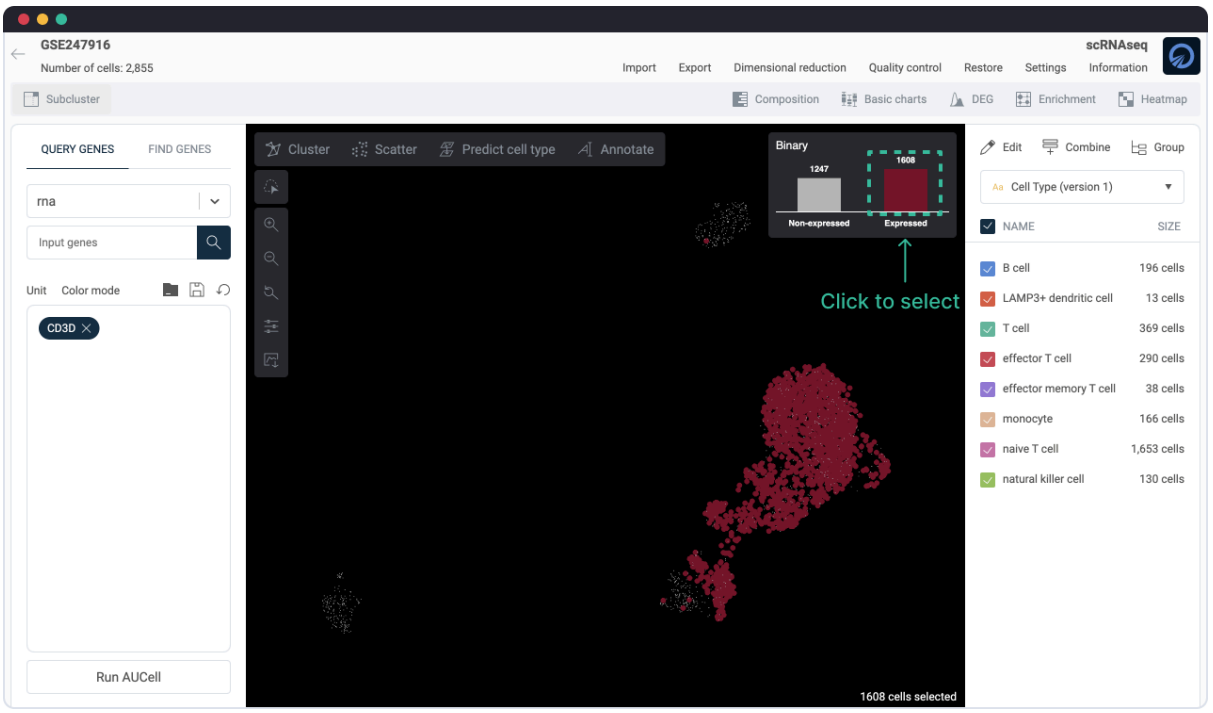
For queries involving two genes, the expression values are visualized as distinct categories and summarized as a bar plot. You can click on each bar to select the corresponding cell population.



When querying more than two genes, expression values are displayed as distinct bars, each representing the number of cells expressing from one up to all queried genes. You can select the cell population expressing a certain range of the number of genes by either dragging to select a region on the bar plot or inputting a precise number in the input box.



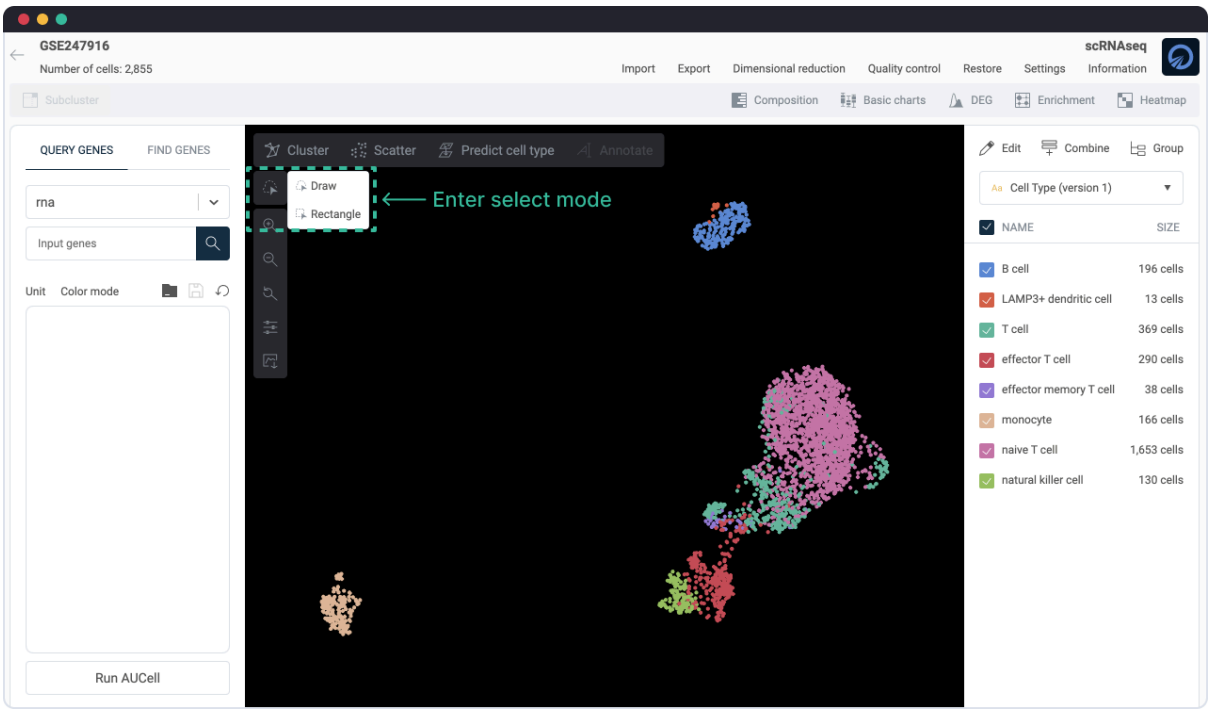
When using binary color mode for multiple genes, the summarized plot will show two bars, one representing cells that express all queried genes and another for the remainder. You can click on each bar to select the corresponding cell population.



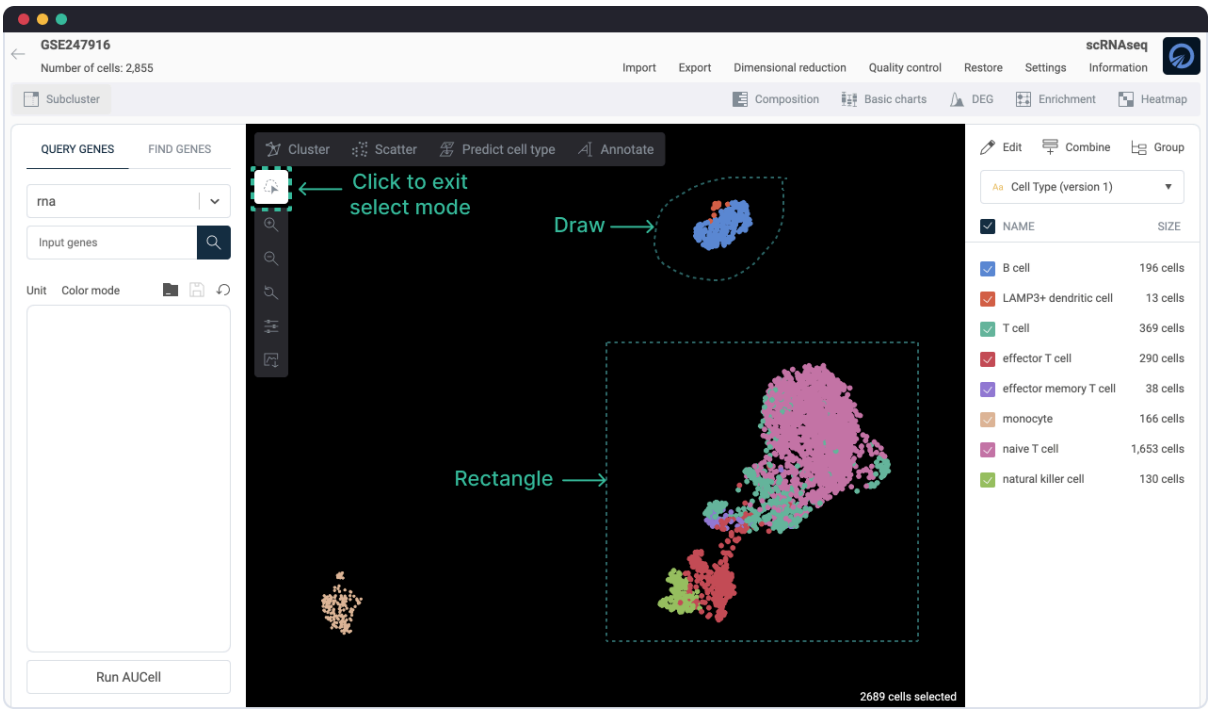
1.3. Manually draw to select

In some cases, you may need to manually draw to select a specific cell population. To start the select mode, click on the 'select' icon in the left menu. OmnibusX supports two methods for manual selection:

- `Draw`: Freely draw a path around the desired cell population. All cells enclosed within this path will be selected.
- `Rectangle`: Draw a rectangular area to select all cells within this shape.



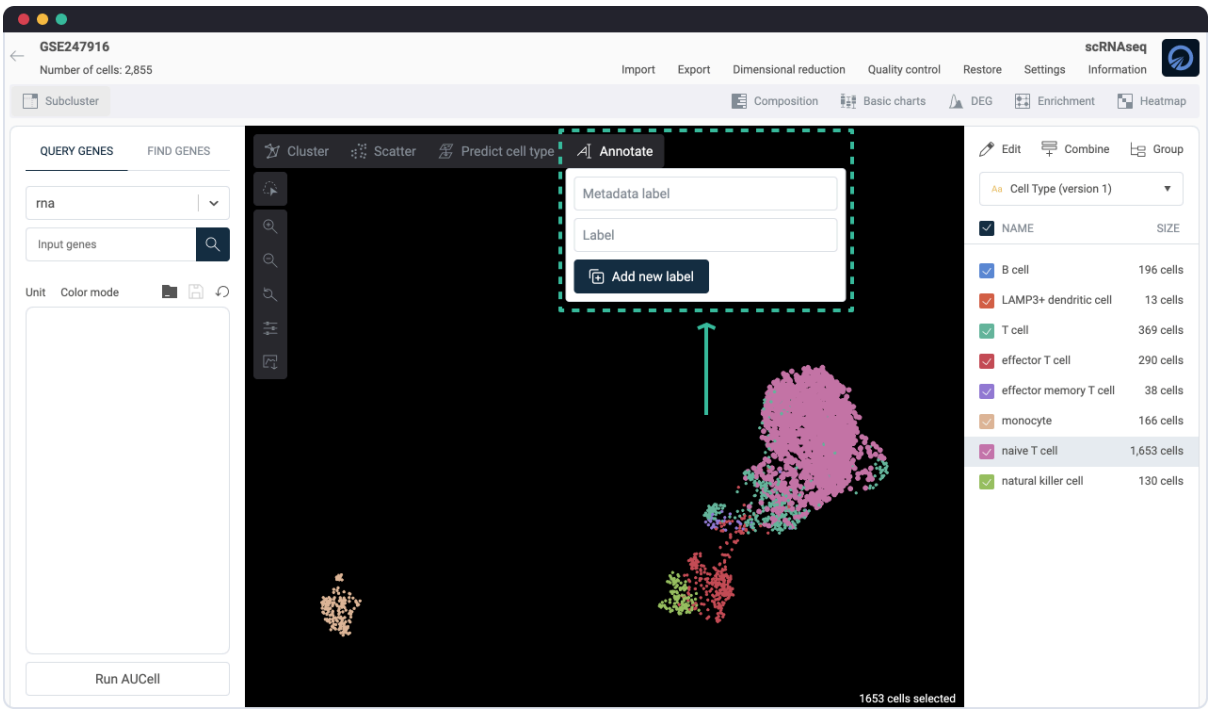
During the drawing process, you can hold the `ctrl`, `cmd`, or `Shift` keys to continue selecting multiple populations without resetting the previous selections. To simplify the selection process and minimize accidental scatter plot adjustments, zooming and panning features are disabled when in selection mode. This ensures that your drawing actions are precise and not confused with other interactions. Once you have completed your selections, click on the `Select` icon again to exit the selection mode. This action re-enables normal interaction with the scatter plot, including zooming and moving.



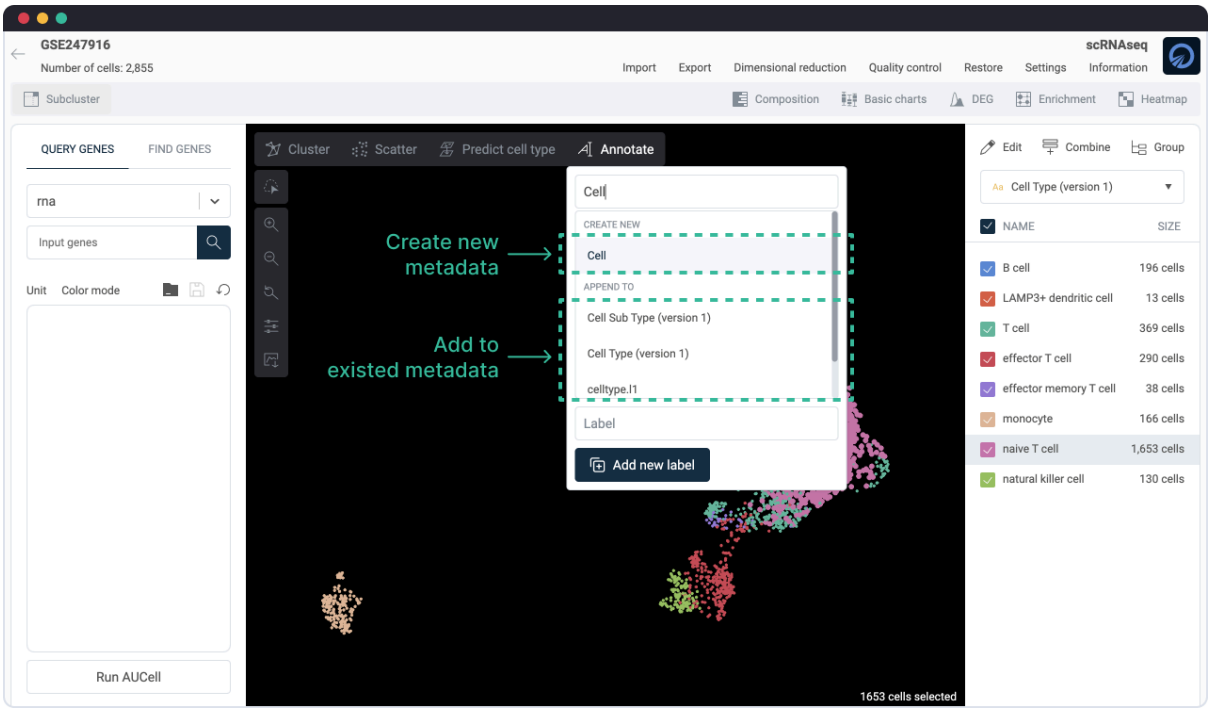
2. Manual annotation

Manual annotation is an essential step in single-cell data analysis, allowing you to categorize cells based on specific characteristics. Here's how you can manually label selected cell populations in OmnibusX:

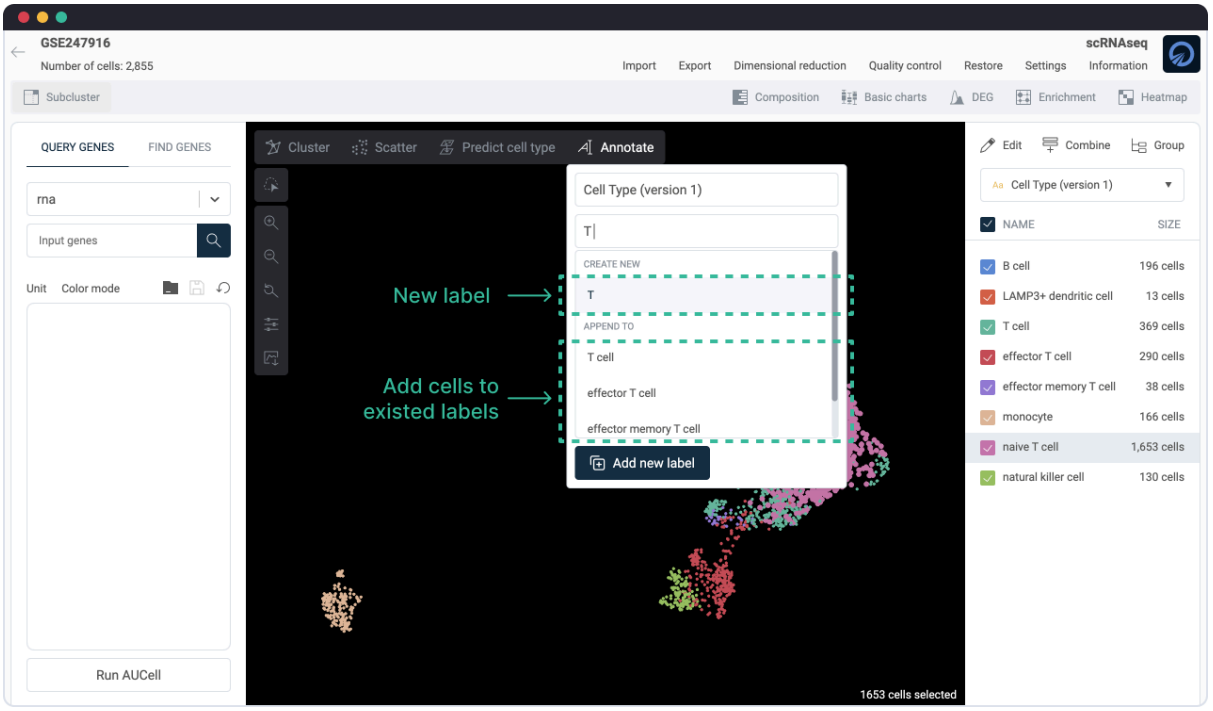
- Click on the `Annotation` button located at the top of the scatter plot after selecting your desired cell population.



- In the metadata name field, type a name for your new annotation. As you type, related existing metadata will be suggested for you. You can choose to create a new metadata with a unique name or append your label to an existing metadata.



- If you select to create a new metadata, input a label for the current selected population. If you choose to append cells to an existing metadata, you have two options:
    1. `Create a new label`: Input a new label for the selected population.
    2. `Append to existing label`: Add the selected population to an already existing label within the metadata.



If any cells in the selected population already have a label within this metadata, the old label will be overwritten with the newly selected label.

3. Programmatic annotation

In addition to manually curating cells based on observed characteristics within the dataset, OmnibusX supports automated methods for labeling cells using established knowledge:

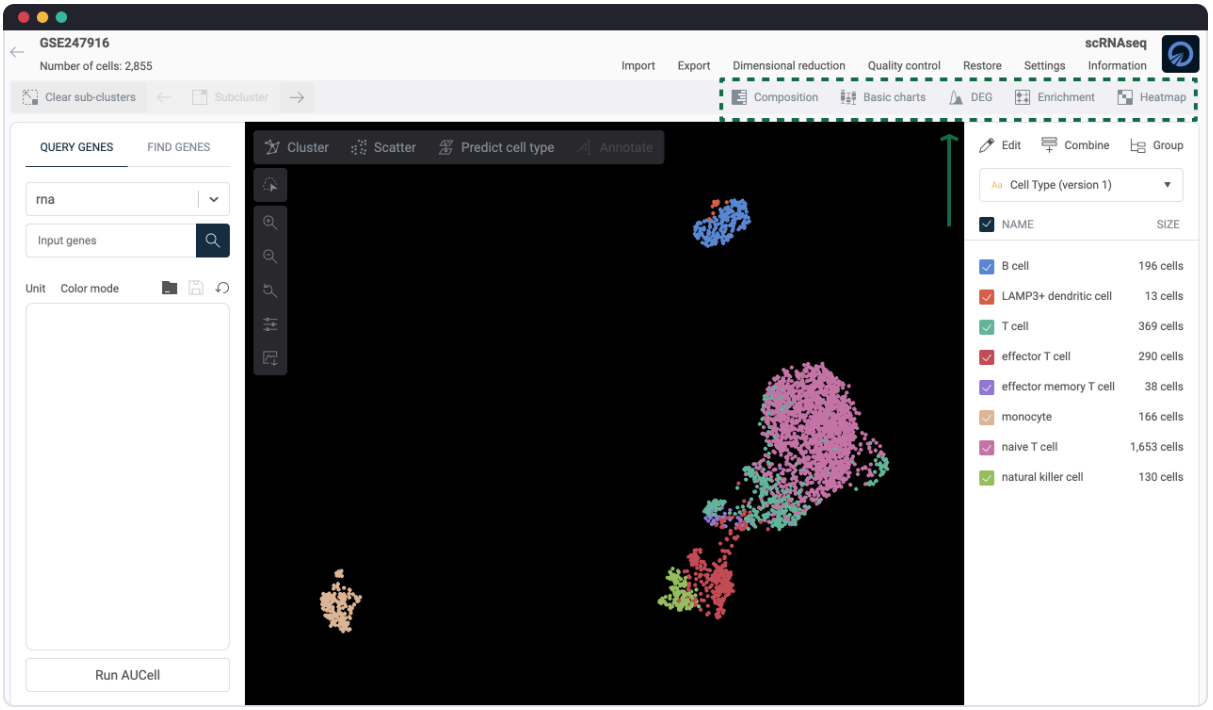
1. `OmnibusX cell type prediction`: By manually curating cell type names and markers from 280 publications, we have unified the nomenclature to consistently identify 166 cell types and subtypes. Our methodology,

published in Biorxiv, offers a fast and precise way to label your cell types (detail in Cell type prediction section).

2. `Automatically classify using your own gene set`: If you have developed a custom set of gene sets, you can apply OmnibusX's methodology (similar to the cell type prediction method) to label cell populations based on the most relevant enriched gene set (detail in Cell type prediction section).
3. `Clustering for annotation`: Clustering is another effective method for grouping cells into similar populations based on their expression profiles or other cellular characteristics. While clustering provides a useful initial grouping, it may not always align perfectly with true biological cell populations, especially for closely related subtypes. To enhance accuracy, we recommend combining clustering with the sub-cluster function (detail in Clustering section and Sub-clusters management section). This allows for a multi-resolution examination of cell populations.

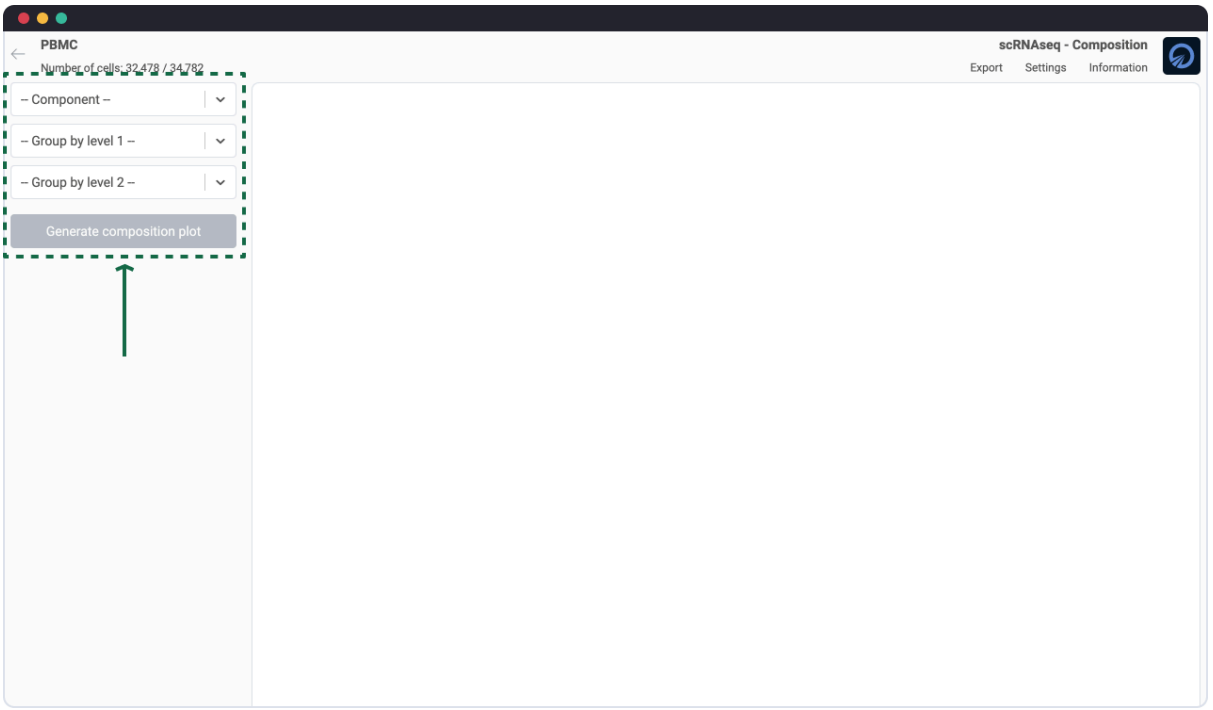
Downstream analysis

OmnibusX supports a range of functions that allow for further exploration of the dataset based on expression data and curated annotations stored as metadata. These analyses can be applied to the whole dataset, to a subset if you are working within a specific subcluster, or across different available omics. Supported functions include `Composition` analysis, `Basic charts` for reports and presentations, `Differential expression gene/feature` analysis, `Enrichment` analysis, and `Heatmap`. You can initiate any of these analyses by selecting the corresponding icon from the top menu.

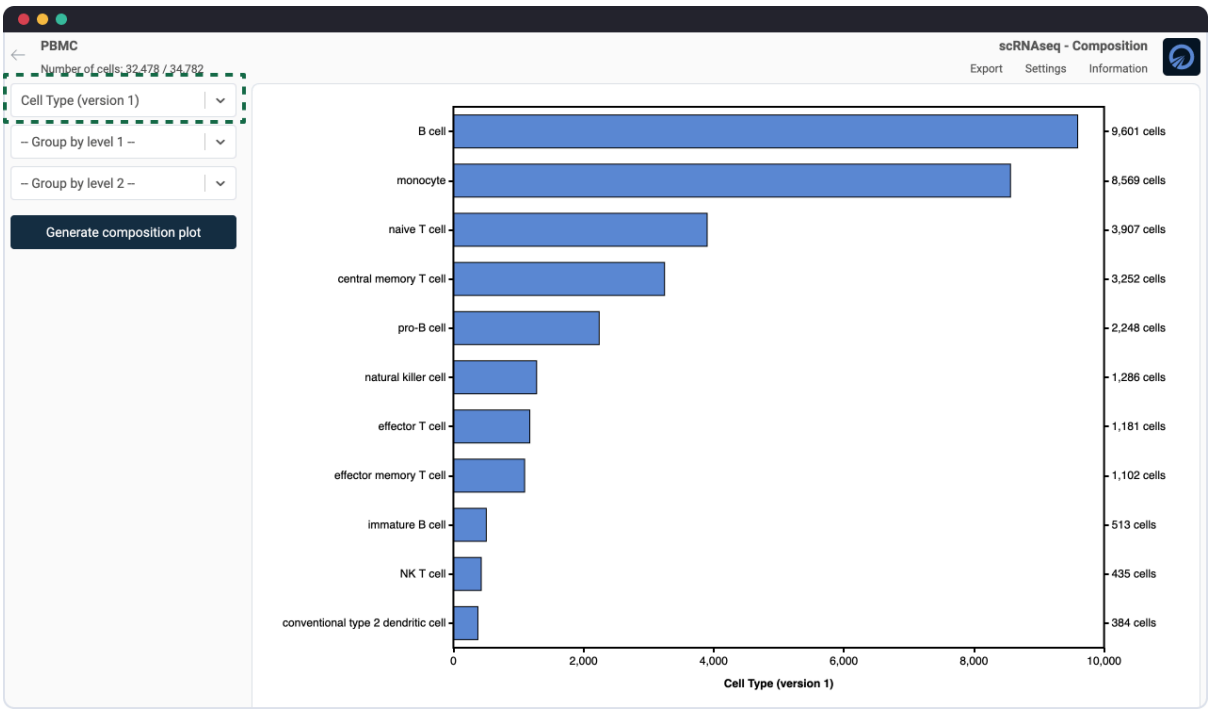


1. Composition analysis

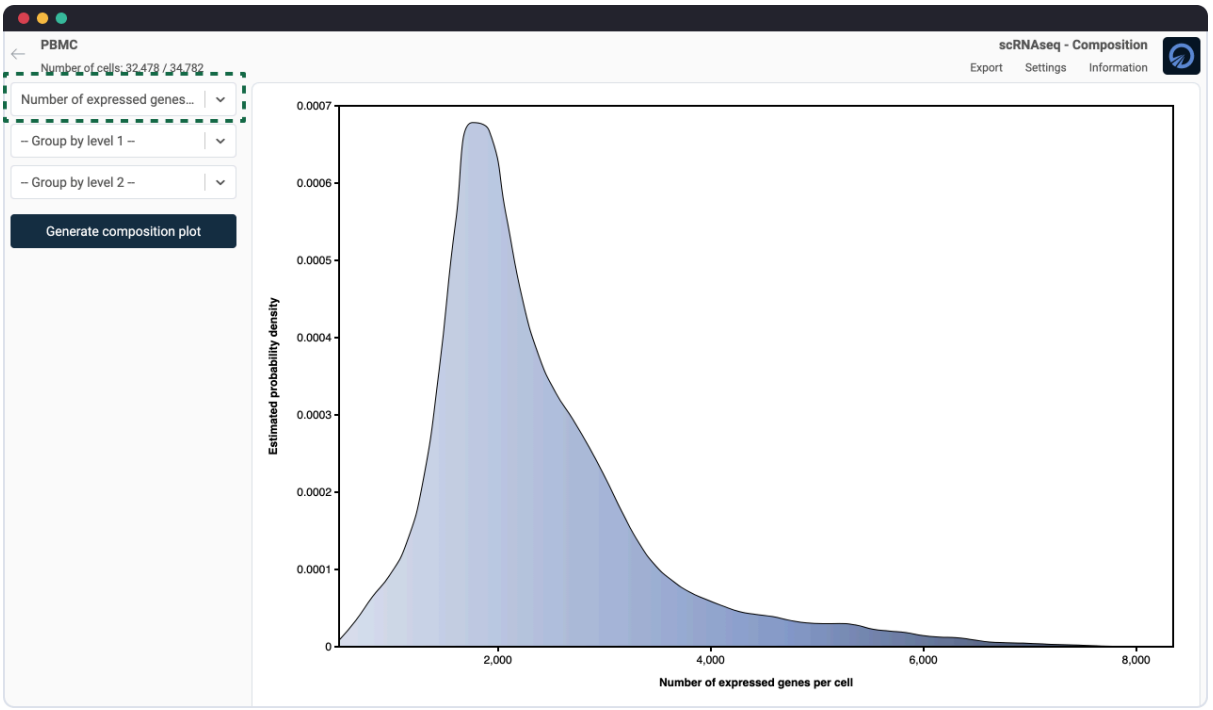
Composition analysis provides insights into the characteristics of your data derived from metadata information. To start, select a metadata of interest from the `Component` select box and click on the `Generate composition plot` button.



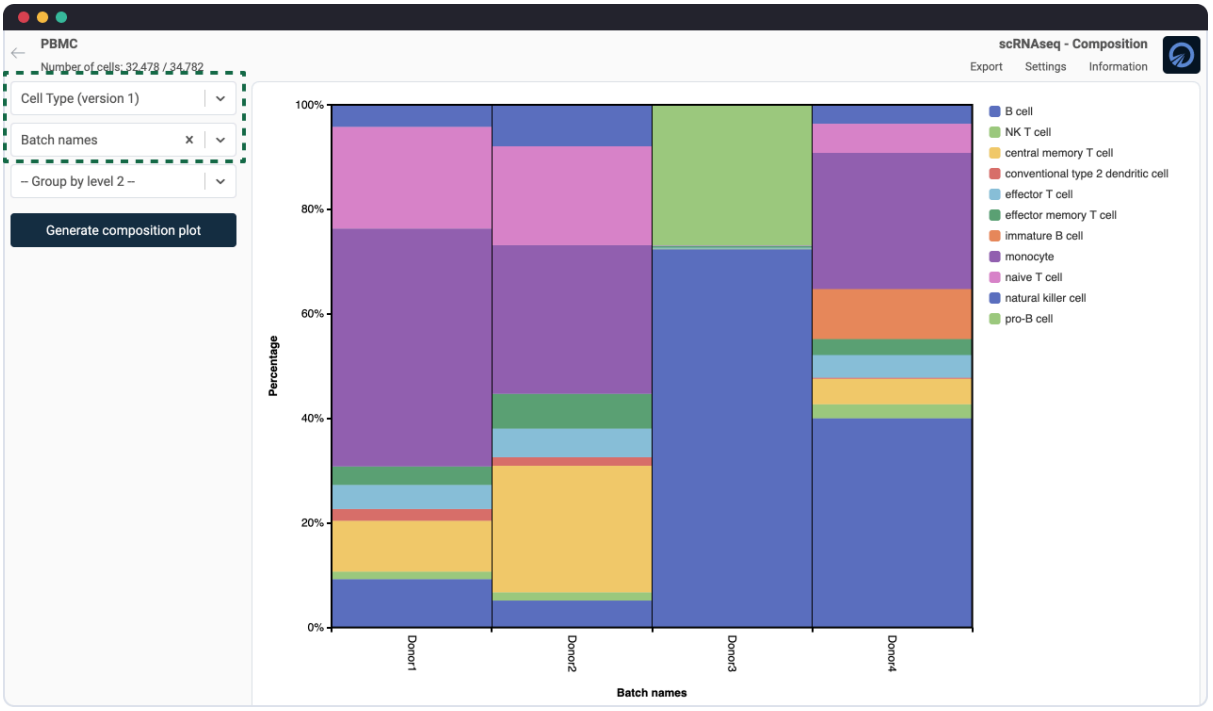
For categorical metadata, OmnibusX generates a bar chart summarizing all available components within the selected metadata and the number of cells in each component.



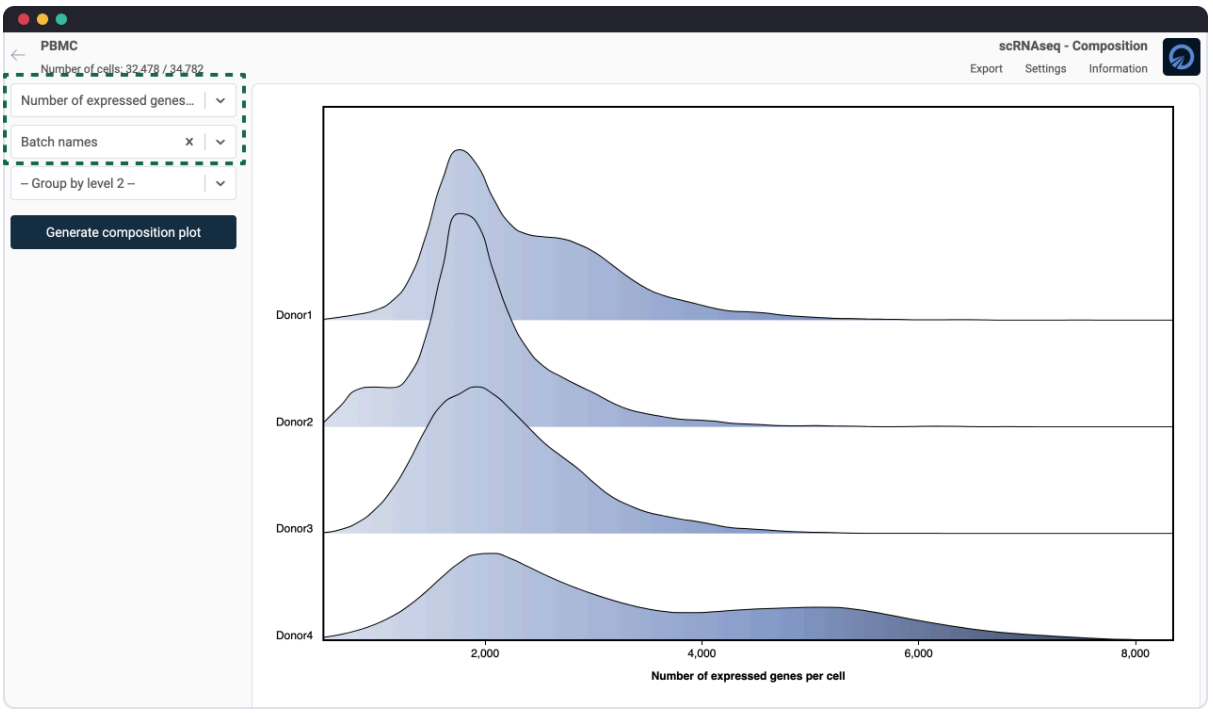
For numerical metadata, OmnibusX creates a distribution plot that summarizes the value distribution of this metadata, smoothed using Kernel Density Estimation (KDE) with implementation sourced from the Scipy team: scipy.stats.gaussian_kde.



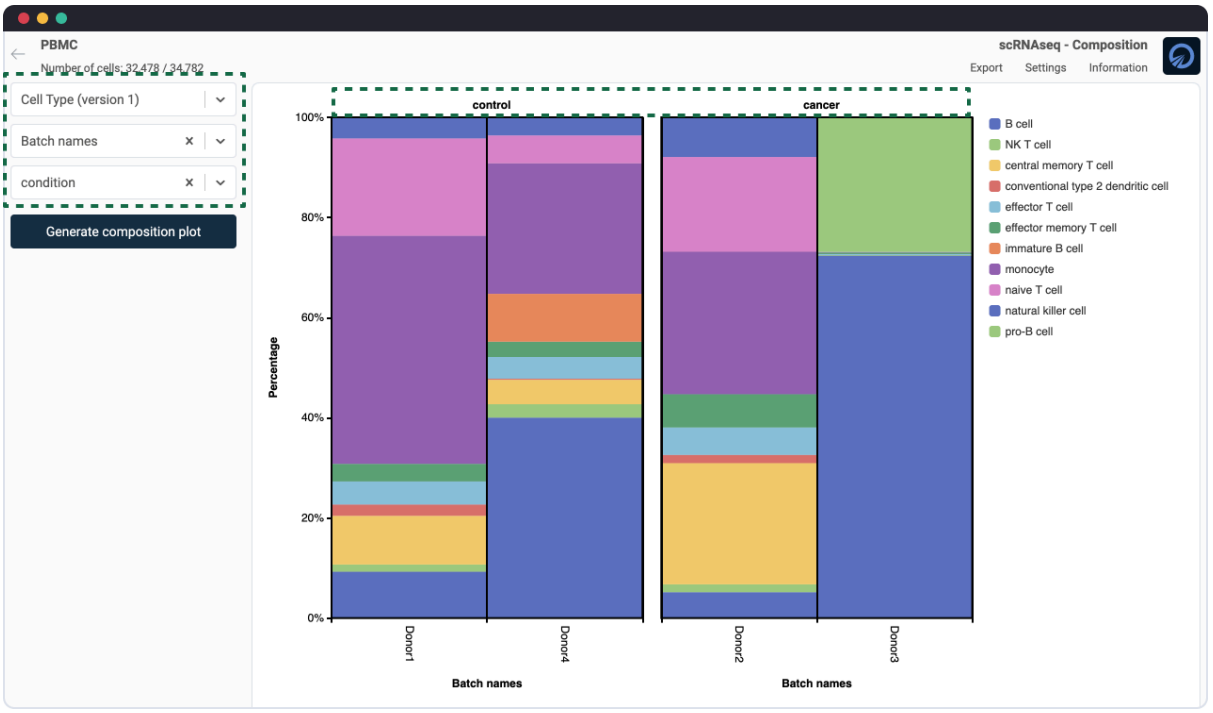
You can increase the complexity of the analysis by adding one more level of grouping. Select a metadata from the `Group by level 1` dropdown to divide the current cell population into corresponding groups. For categorical data, OmnibusX generates a composition plot displaying the composition of component metadata within each group.



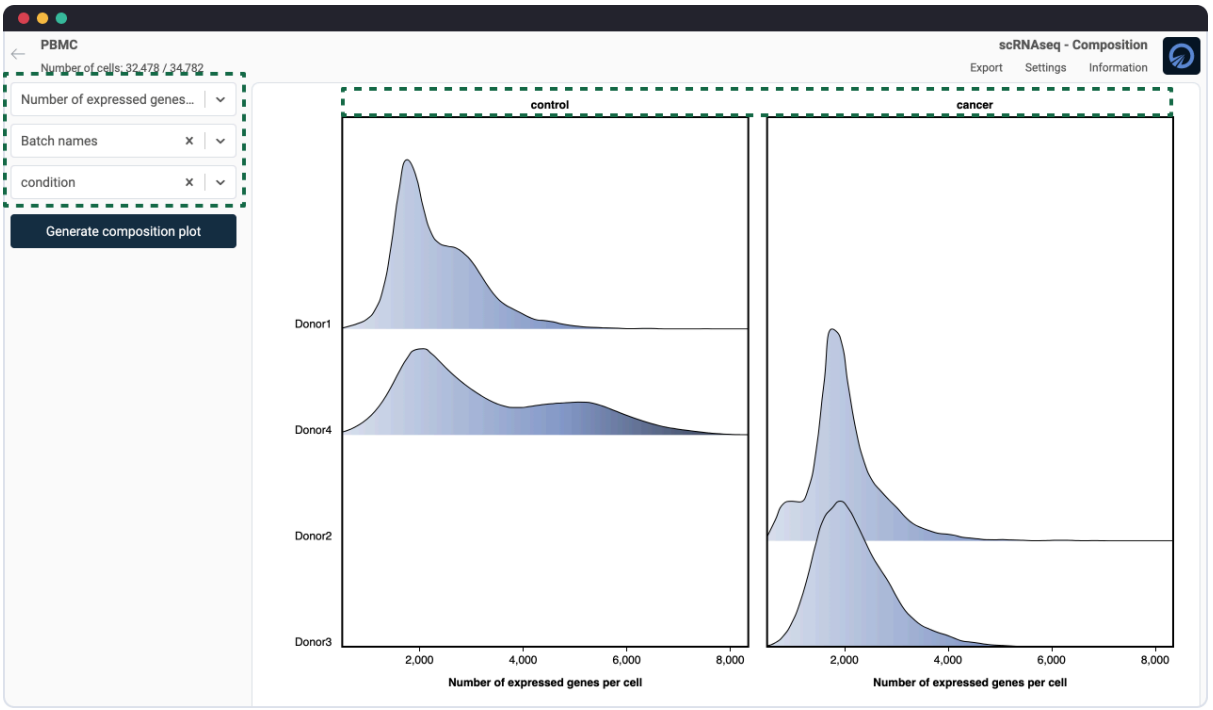
For numerical data, OmnibusX produces stacked distribution plots, with each plot summarizing the values in each group.



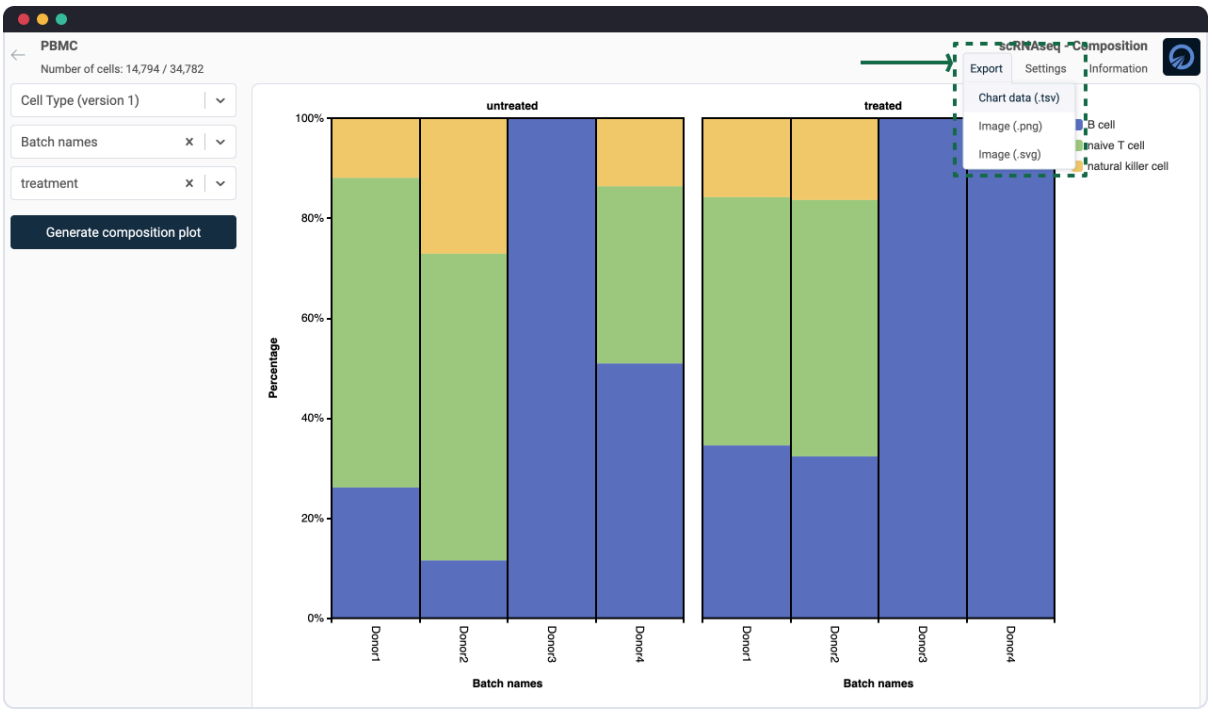
OmnibusX allows for up to two levels of grouping. Adding a second level divides the current cell populations according to labels from the `Group by level 2` metadata. Then, for categorical metadata, OmnibusX draws multiple composition plots from component metadata and level 1 metadata for each group.



For numerical metadata, multiple stacked distribution plots are created from component metadata and level 1 metadata for each group.



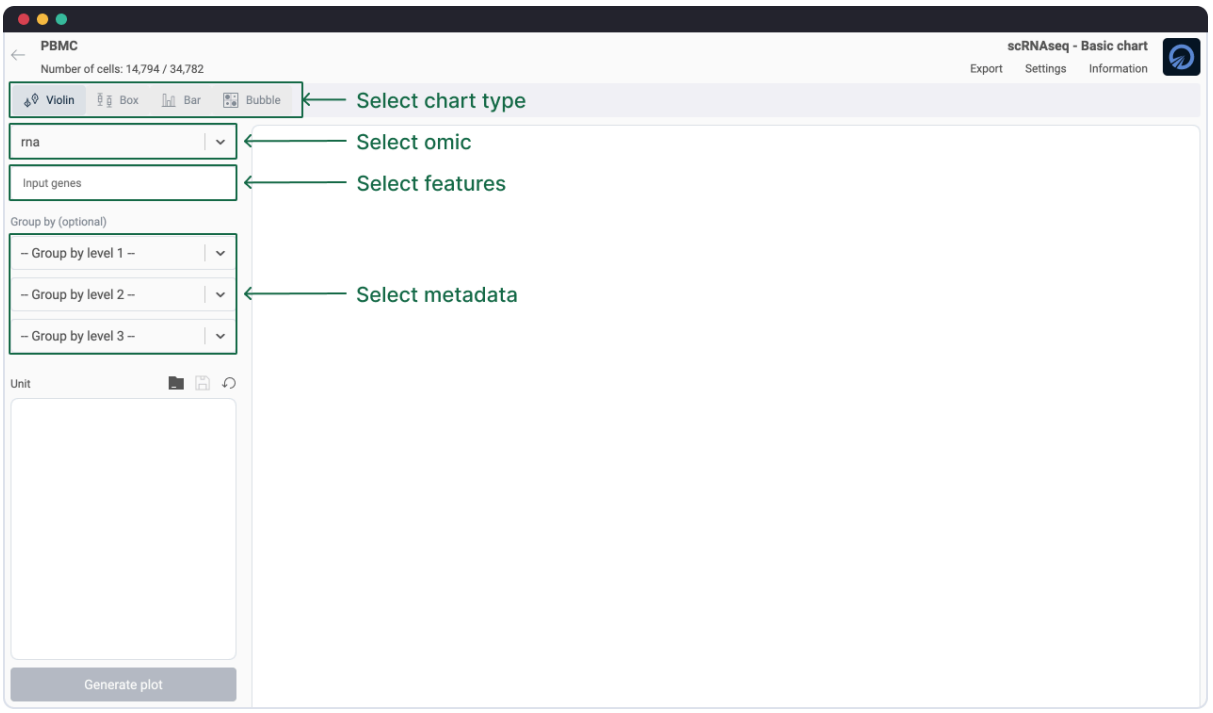
For each generated plot, you can export the chart data as a .tsv file, the image as a PNG, and the image as an SVG by clicking on the export button located on the top menu.



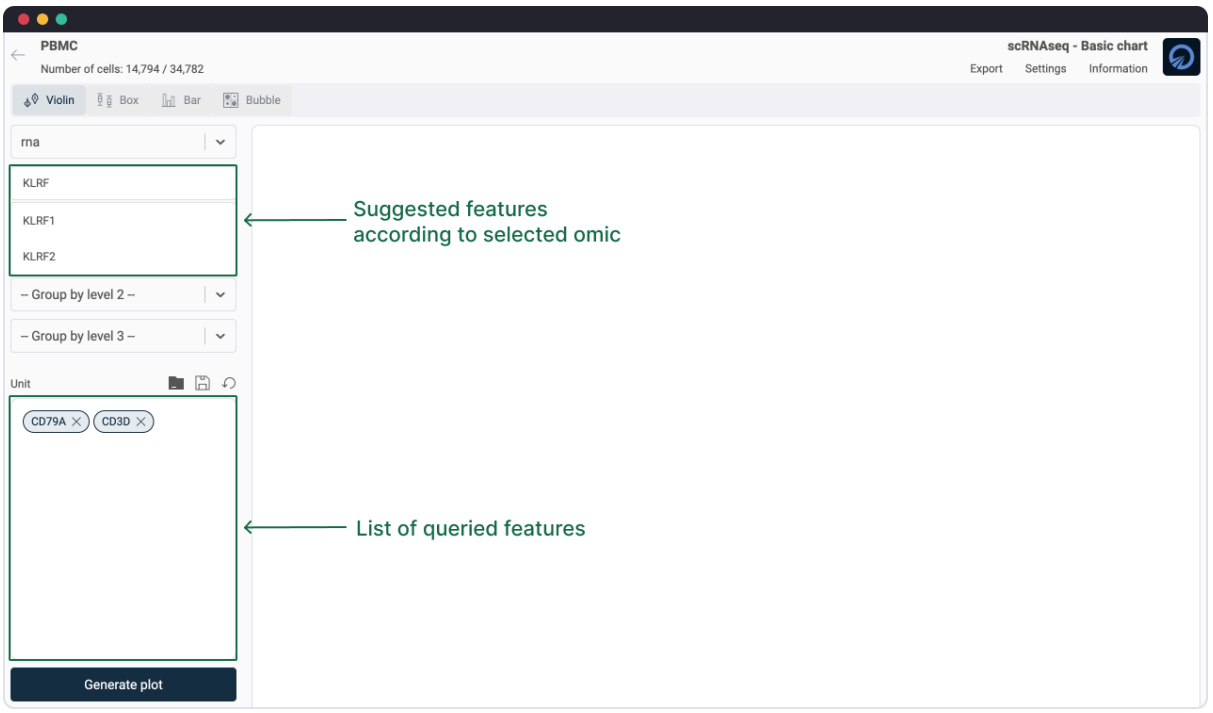
2. Basic charts

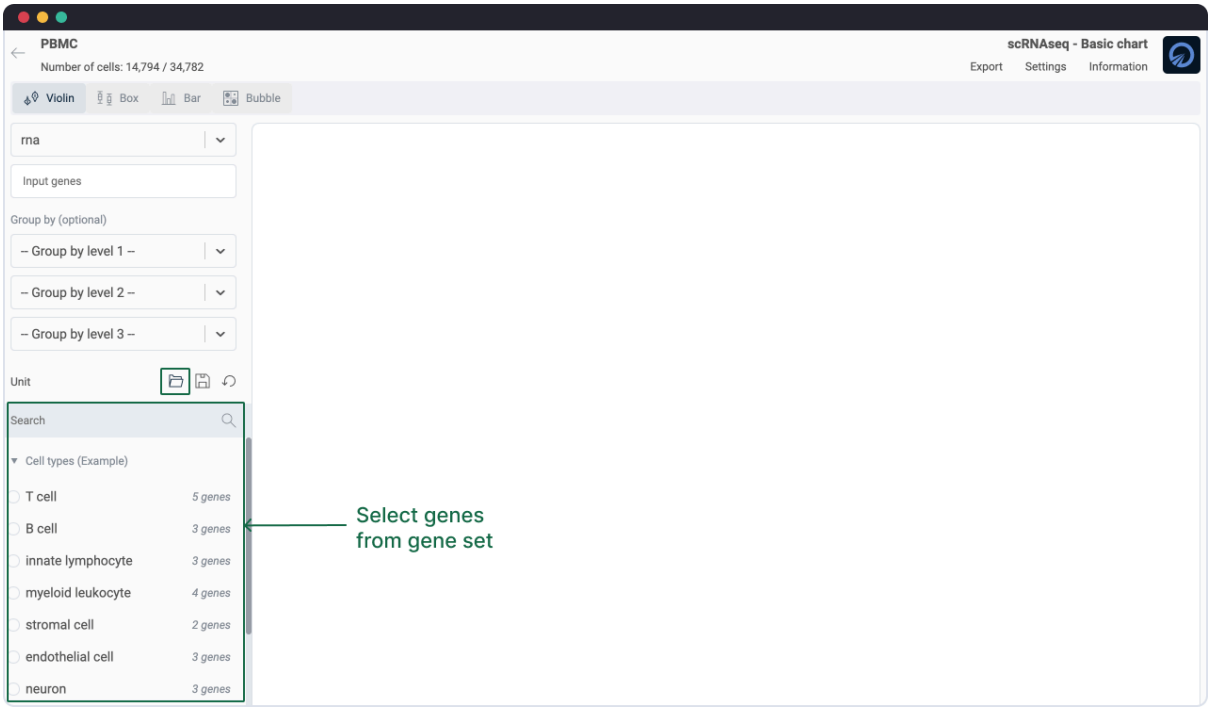
OmnibusX supports four distinct types of plots for visualizing expression values across different cell populations: `Violin` plot, `Box` plot, `Bar` plot, and `Bubble` plot. These visualization tools can be applied to different omics available in your dataset. You can follow the below steps to start plotting:

1. `Select chart type`: select the chart type of your choice from the top menu bar.
2. `Select omic`: From the left control panel, select an omic of interest from the dropdown box.



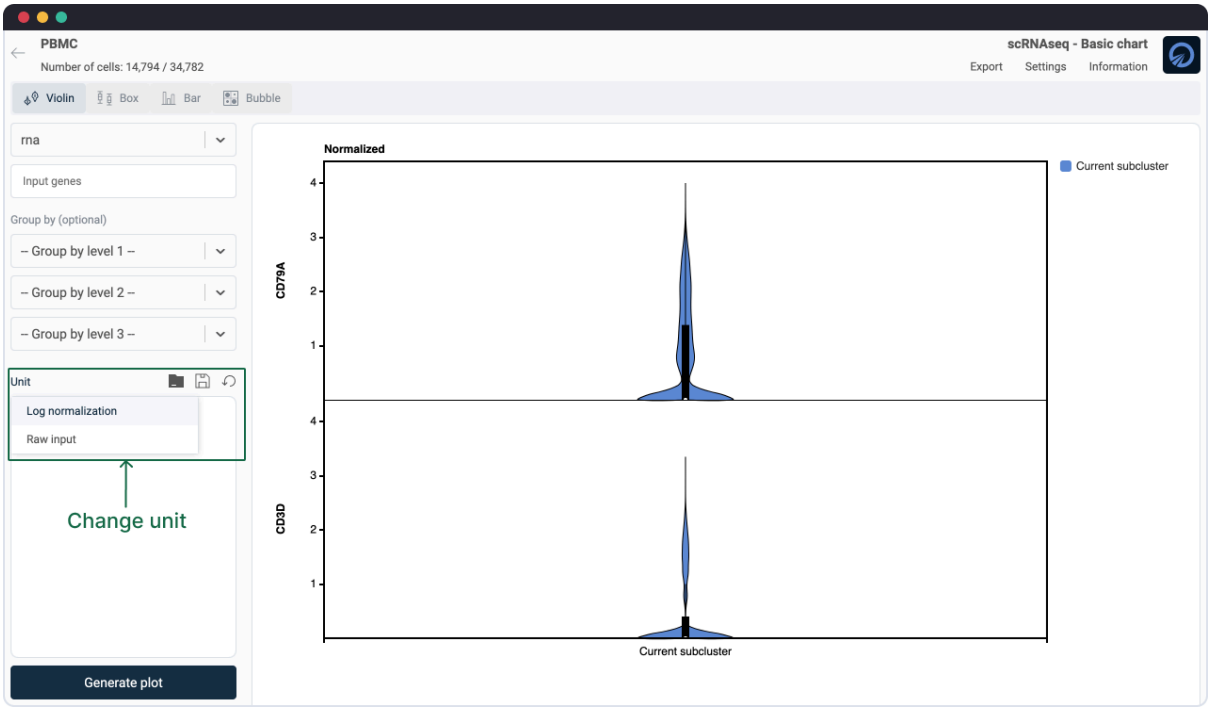
3. `Select features`: Below the omic selection, type the name of your target feature into the input box. As you type, OmnibusX will suggest related and available features from your dataset. Click on the suggested feature name to select it. For `rna` omic, you can quickly incorporate a list of genes from your gene set library by clicking on the folder icon above the query box. A list of available gene sets will appear, allowing you to select one and automatically add its genes to your query list.



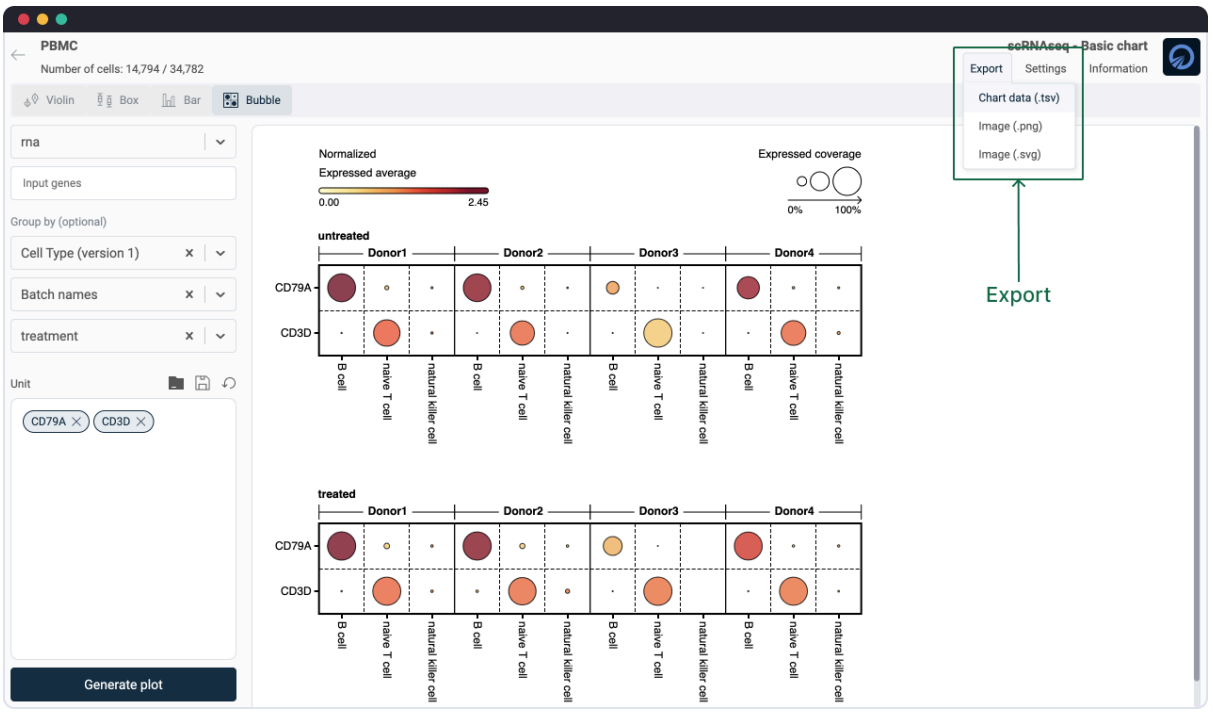


4. `Select metadata`: By default, if no metadata is selected for grouping, OmnibusX treats all cells as one group. However, you can choose to group cells using up to three levels of metadata from the select boxes for more detailed analysis.

After selecting the features and group by metadata, click on the `Generate plot` button at the bottom to get the corresponding plot. Once the plot is generated, you can change the unit by clicking on the unit button above the query box and selecting your preferred unit.



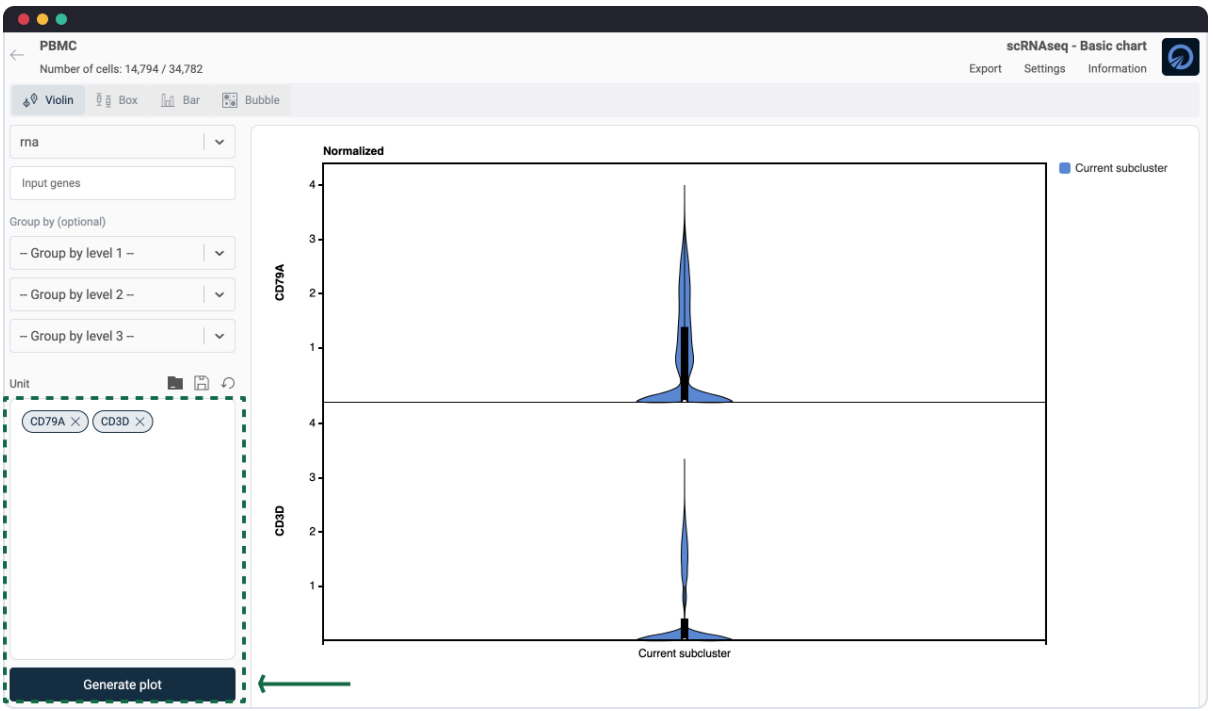
Moreover, you can export the data by clicking on the `Export` button in the top menu. OmnibusX supports exporting the chart data as a .tsv file, or the image as PNG or SVG files for use in reports and presentations



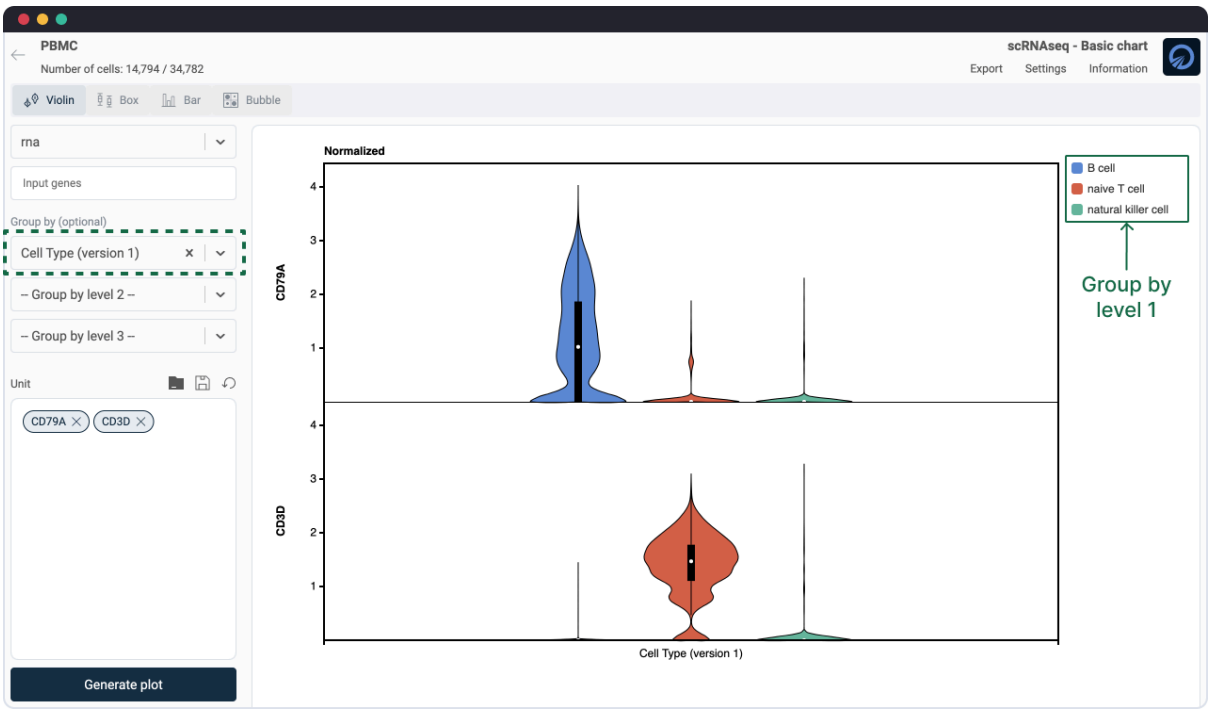
2.1. Violin plot

Violin plot is generated from expression data and cell annotations as below:

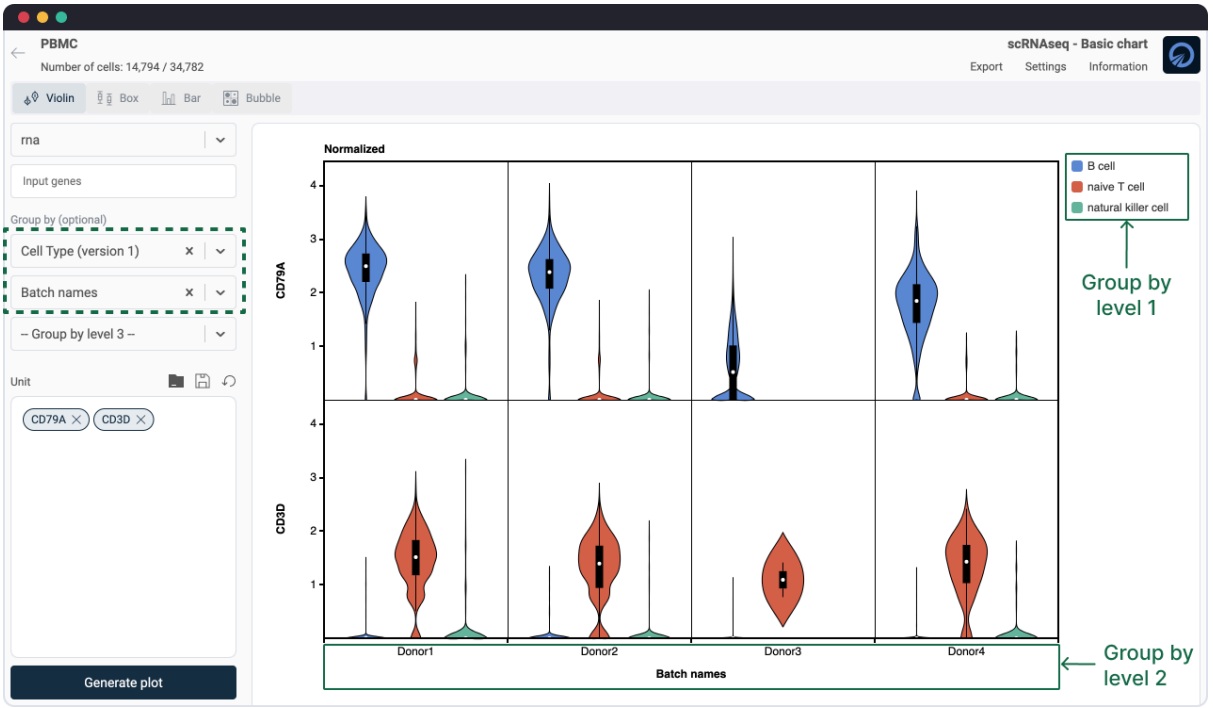
1. `Data preparation and grouping`: OmnibusX retrieves the expression values of the queried features based on the selected omic and unit. The data is then divided into groups according to the selected metadata annotations.
2. `Data for violin`: The distribution of each feature within each group is smoothed using Kernel Density Estimation (KDE), implemented by the Scipy team: [scipy.stats.gaussian_kde], to create the violin plots.
3. `Plotting`: Depending on the number of metadata levels selected, different plots are generated:
    - `None`: all cells are treated as one group, and OmnibusX draws one violin for each queried feature.



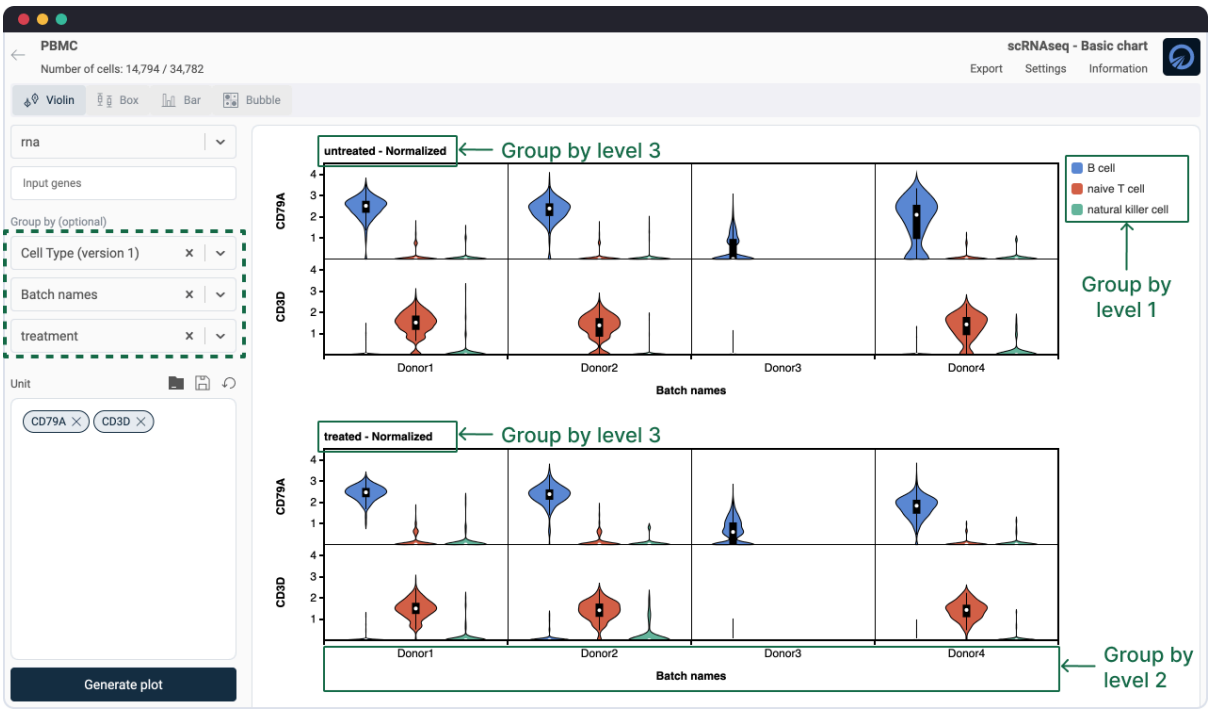
- `Group by level 1`: Cells are divided into groups using the annotation from the input metadata. Each group is visualized with a distinct color.



- `Group by level 2`: Cells are divided into groups using the combined information from both metadata. Groups from the same annotation of metadata level 1 will have the same color. Groups from metadata level 2 are displayed as distinct xAxis values.



- `Group by level 3`: Cells are first divided by the third level of metadata, with each group presented as a separate plot using combined information from the first two metadata levels.



All plots are drawn using the same scale for comparative analysis.

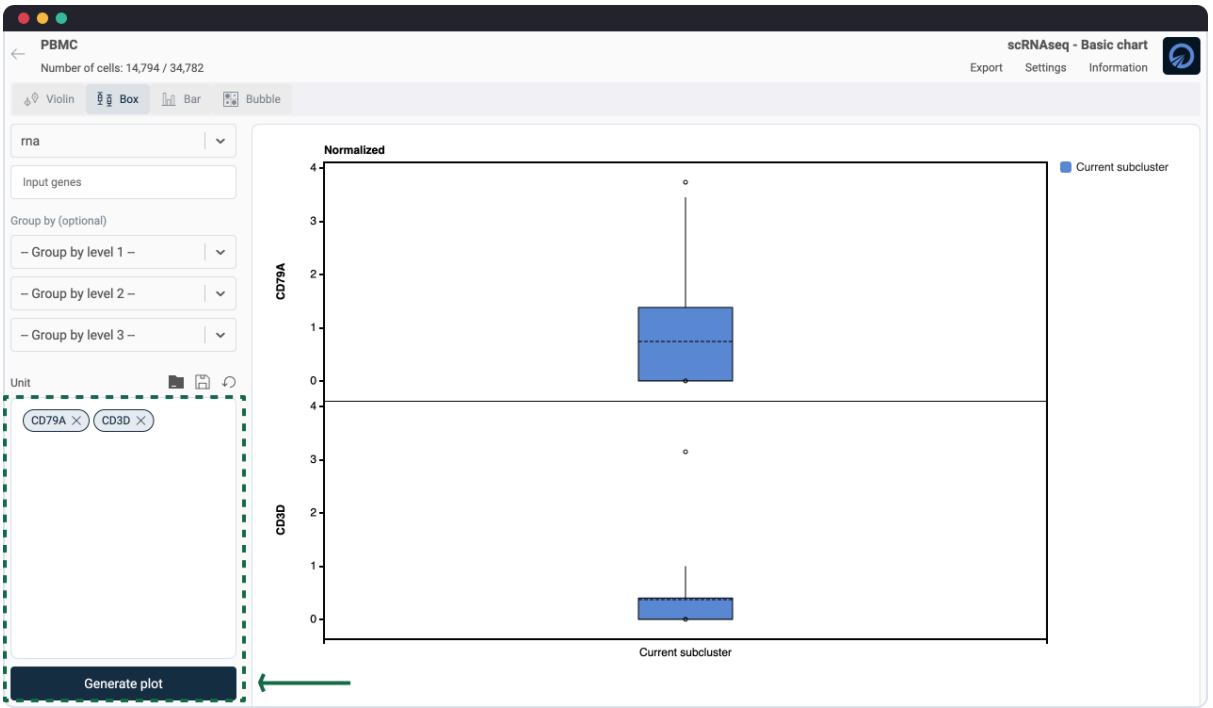
2.2. Box plot

Box plot is generated from expression data and cell annotations as below:

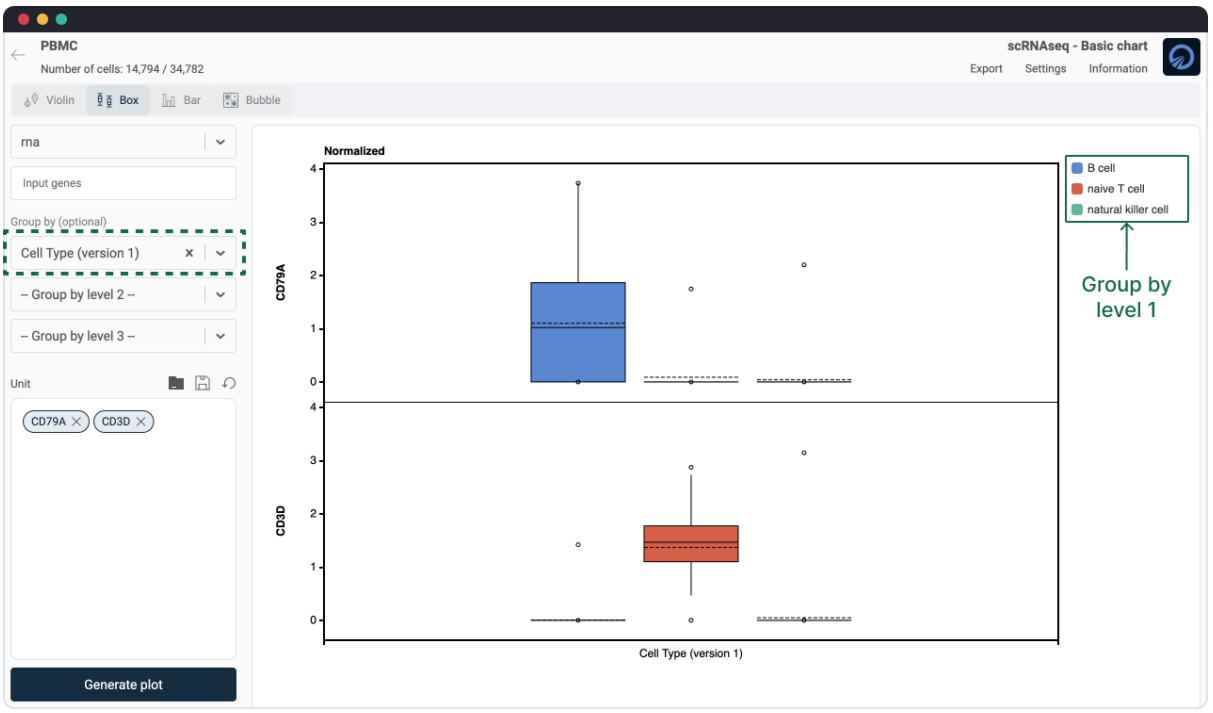
1. `Data preparation and grouping`: OmnibusX retrieves the expression values of the queried features based on the selected omic and unit. The data is then divided into groups according to the selected metadata annotations.
2. `Data for box`: For each group, the following statistical measures are computed:
    - `Minimum`: The lowest expression value of the data.
    - `Maximum`: The highest expression value of the data.
    - `Mean`: The average expression value of the data.
    - `Median`: The middle value separating the higher half from the lower half of the data.
    - `Q1` (First Quartile): The 25th percentile of the data.
    - `Q3` (Third Quartile): The 75th percentile of the data.

- `Lower whisker`: Calculated as `Q1 - 1.5IQR` (interquartile range, which is the distance between Q1 and Q3).
- `Upper whisker`: Calculated as `Q3 + 1.5IQR`.

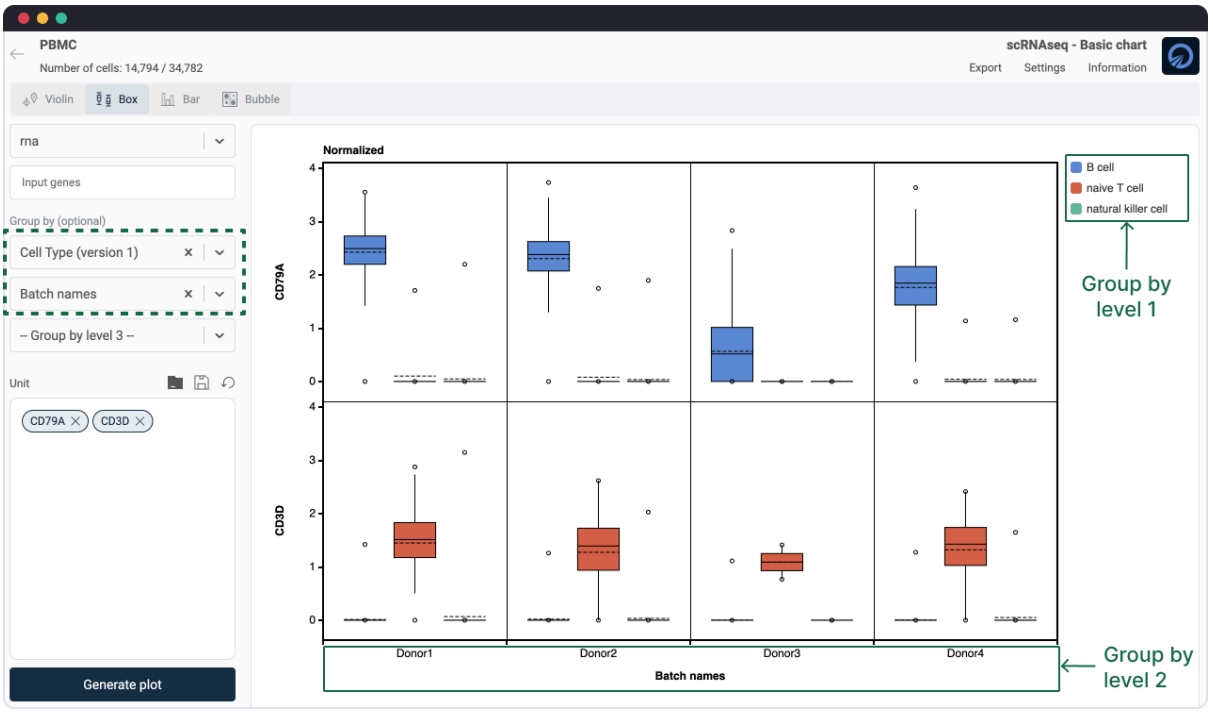
3. `Plotting`: Depending on the number of metadata levels selected, different plots are generated:
    - `None`: all cells are treated as one group, and OmnibusX draws one box for each queried feature.



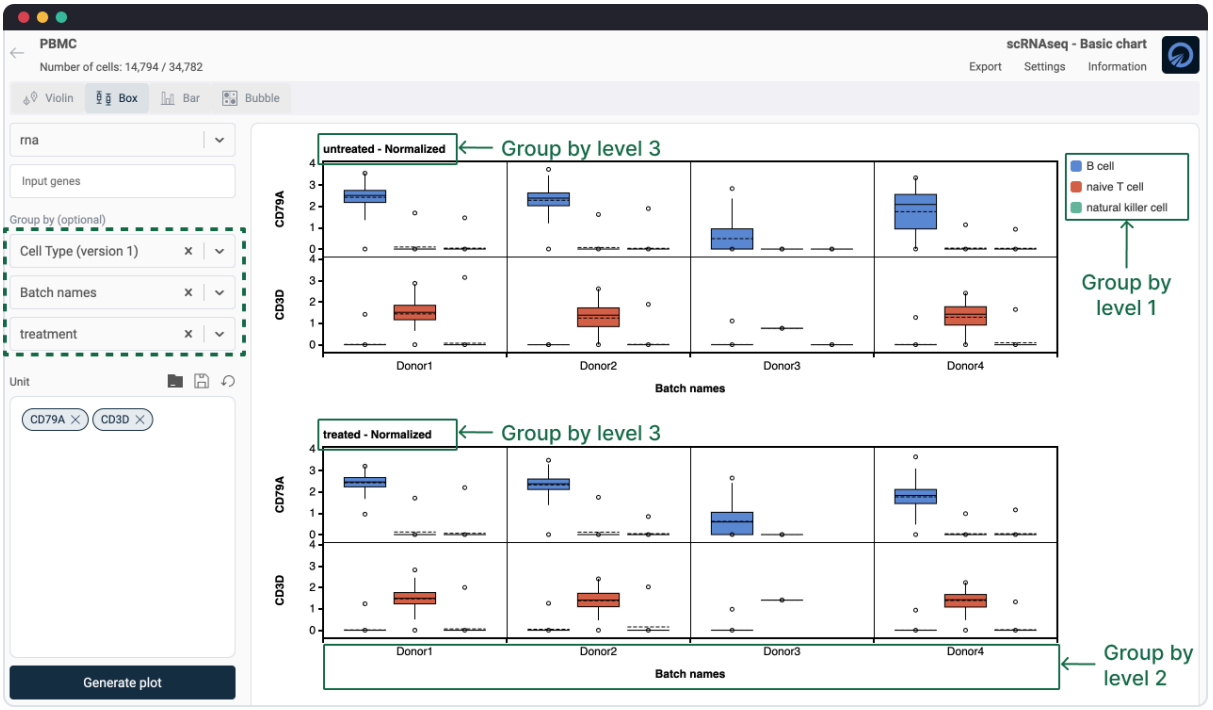
- `Group by level 1`: Cells are divided into groups using the annotation from the input metadata. Each group is visualized with a distinct color.



- `Group by level 2`: Cells are divided into groups using the combined information from both metadata. Groups from the same annotation of metadata level 1 will have the same color. Groups from metadata level 2 are displayed as distinct xAxis values.



- `Group by level 3`: Cells are first divided by the third level of metadata, with each group presented as a separate plot using combined information from the first two metadata levels.



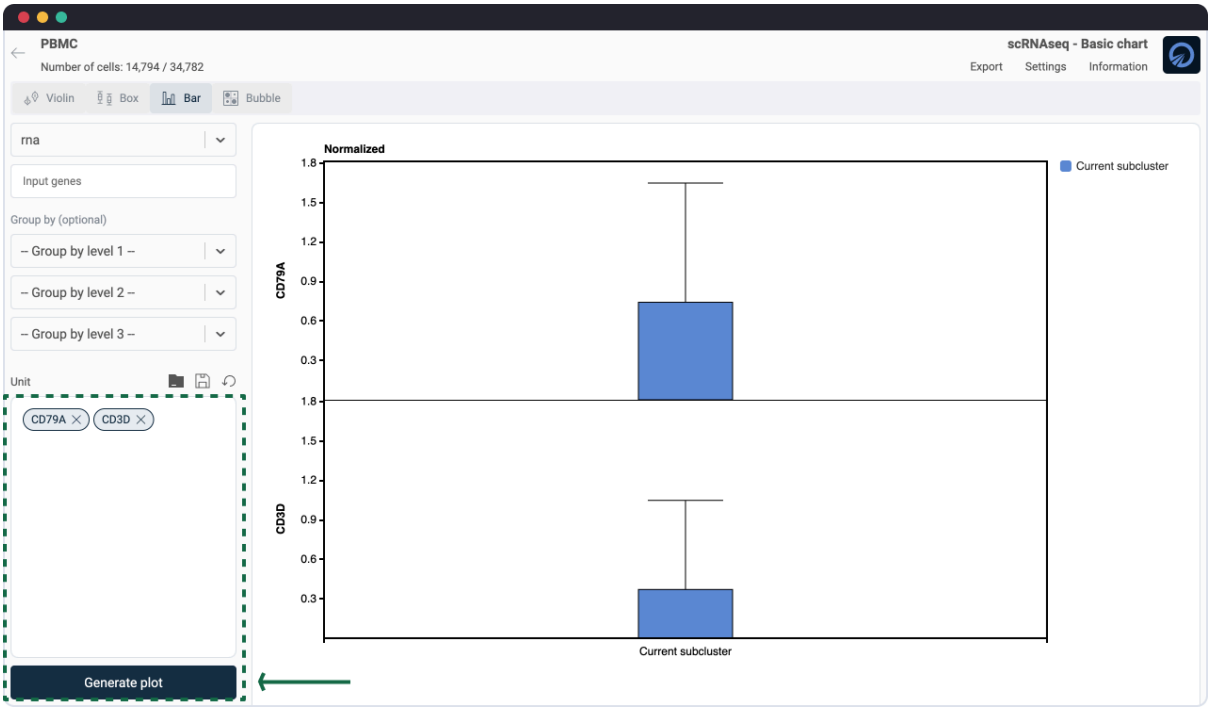
All plots are drawn using the same scale for comparative analysis.

2.3. Bar plot

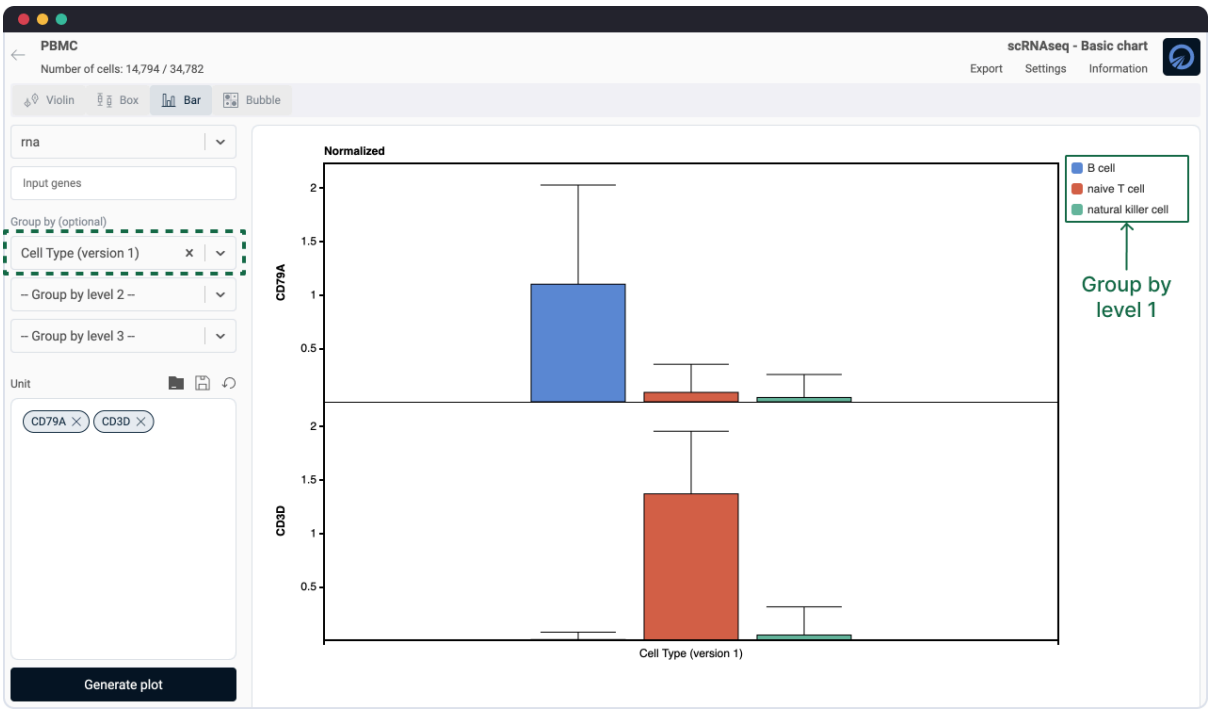
Bar plot is generated from expression data and cell annotations as below:

1. `Data preparation and grouping`: OmnibusX retrieves the expression values of the queried features based on the selected omic and unit. The data is then divided into groups according to the selected metadata annotations.
2. `Data for bar`: For each group, the following statistical measures are computed:
    - `Mean`: The average expression value of the data.
    - `Standard deviation` (Std): Measures the amount of variation, which is calculated for drawing the error bars.

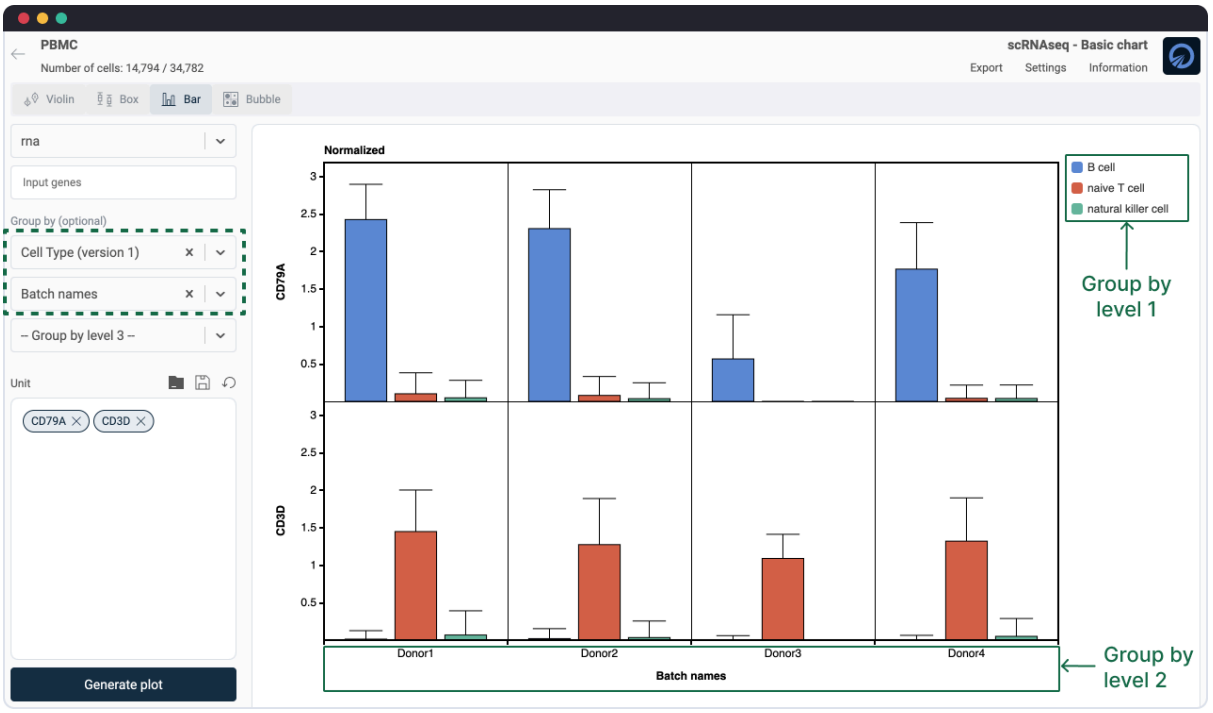
3. `Plotting`: Depending on the number of metadata levels selected, different plots are generated:
    - `None`: all cells are treated as one group, and OmnibusX draws one bar for each queried feature.



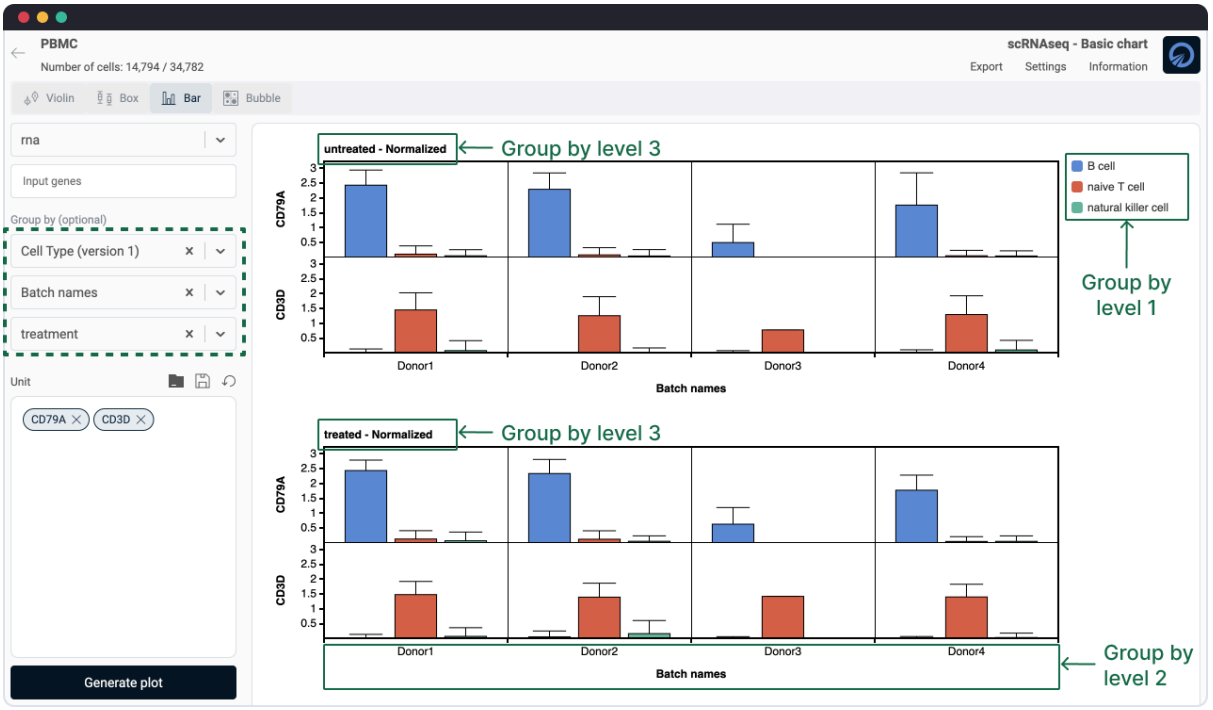
    - `Group by level 1`: Cells are divided into groups using the annotation from the input metadata. Each group is visualized with a distinct color.



- `Group by level 2`: Cells are divided into groups using the combined information from both metadata. Groups from the same annotation of metadata level 1 will have the same color. Groups from metadata level 2 are displayed as distinct xAxis values.



- `Group by level 3`: Cells are first divided by the third level of metadata, with each group presented as a separate plot using combined information from the first two metadata levels.



All plots are drawn using the same scale for comparative analysis.

2.4. Bubble plot

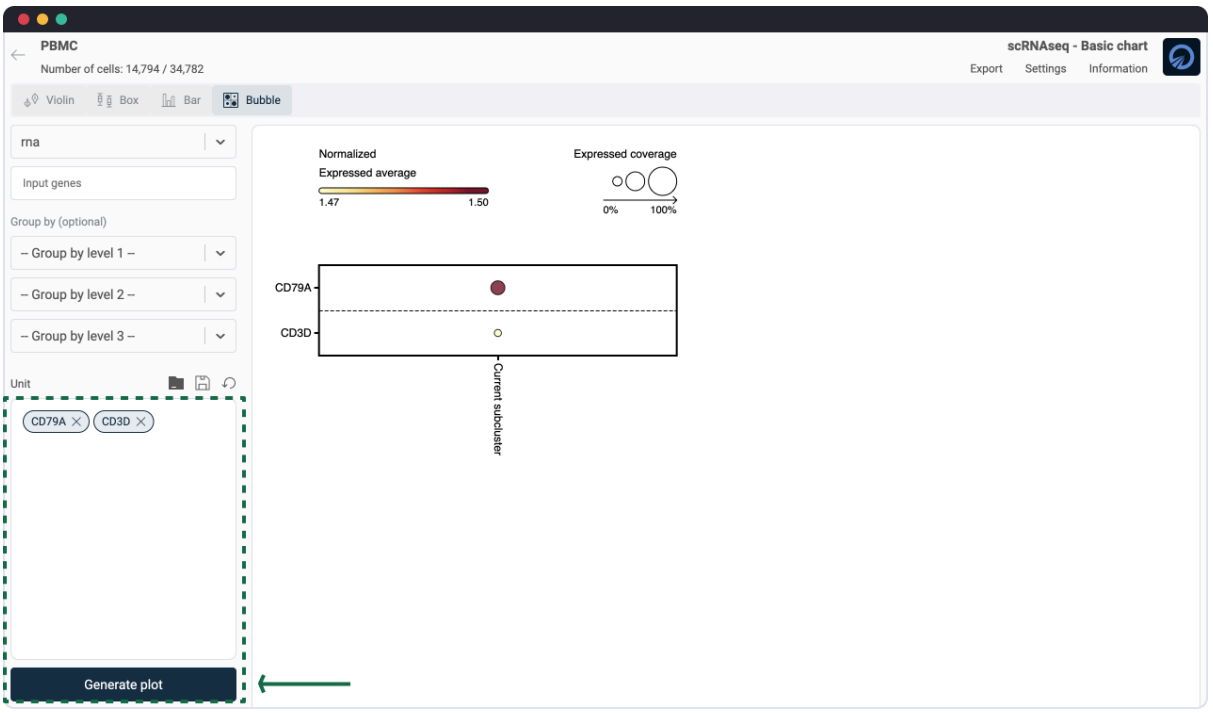
Bubble plots are generated from expression data and cell annotations as follows:

1. `Data preparation and grouping`: OmnibusX retrieves the expression values of the queried features based on the selected omic and unit. The data is then divided into groups according to the selected metadata annotations.
2. `Data for bubble`: Bubble: Each feature of each group is visualized by a bubble, which is calculated as follows:
    - `Color`: The color represents the mean expression of non-zero cells only.
    - `Size`: The bubble size represents the percentage of cells in this group that express the feature.

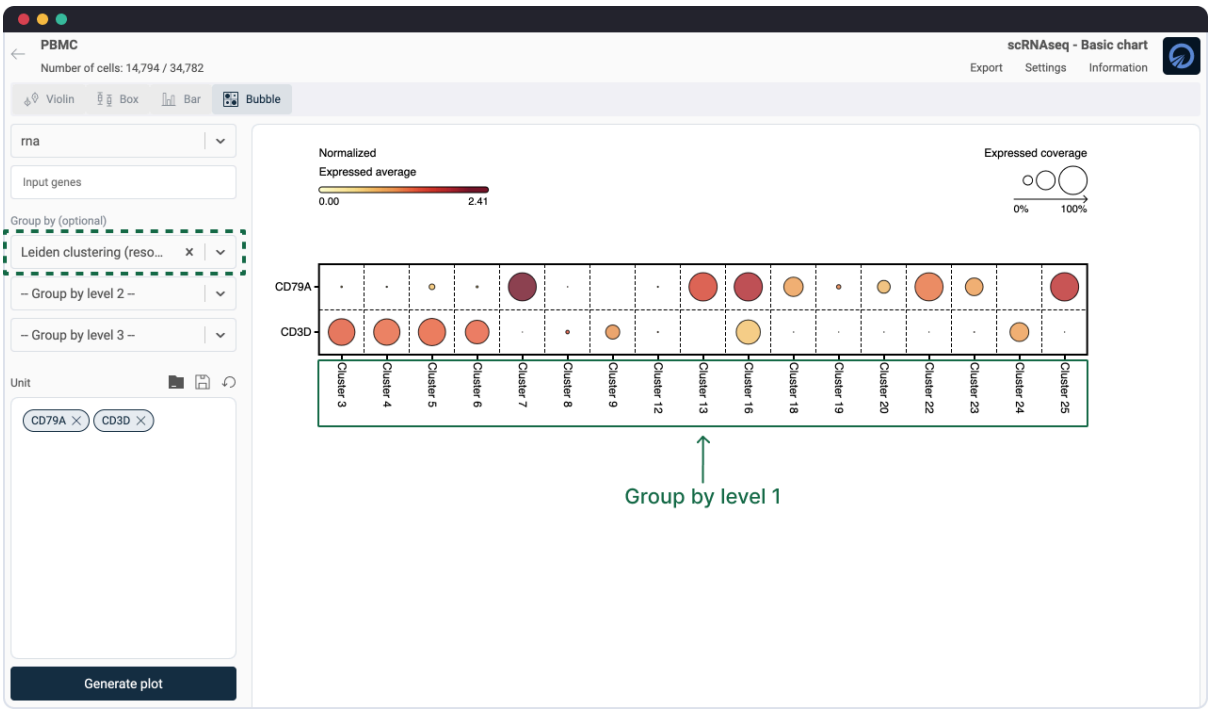
As drop-out events happen in single-cell data, where zero values come from the limit of the technology to capture the true signal, combining information from both mean expression of expressed cells only with the coverage of expressed cells can provide a more effective way to analyze the data. Four scenarios can be observed:

- `Large, highly colored`: Indicates that a majority of the cells express the feature at high levels, suggesting the true population signal and effective capture rate.
- `Large, lightly colored`: Suggests that most cells express the feature but at low levels.
- `Small, highly colored`: Represents a small fraction of the population expressing the feature at high levels, which may indicate a significant sub-population or low capture rate for an actively expressed gene..
- `Small, lightly colored`: potentially indicative of random noise or alignment errors.

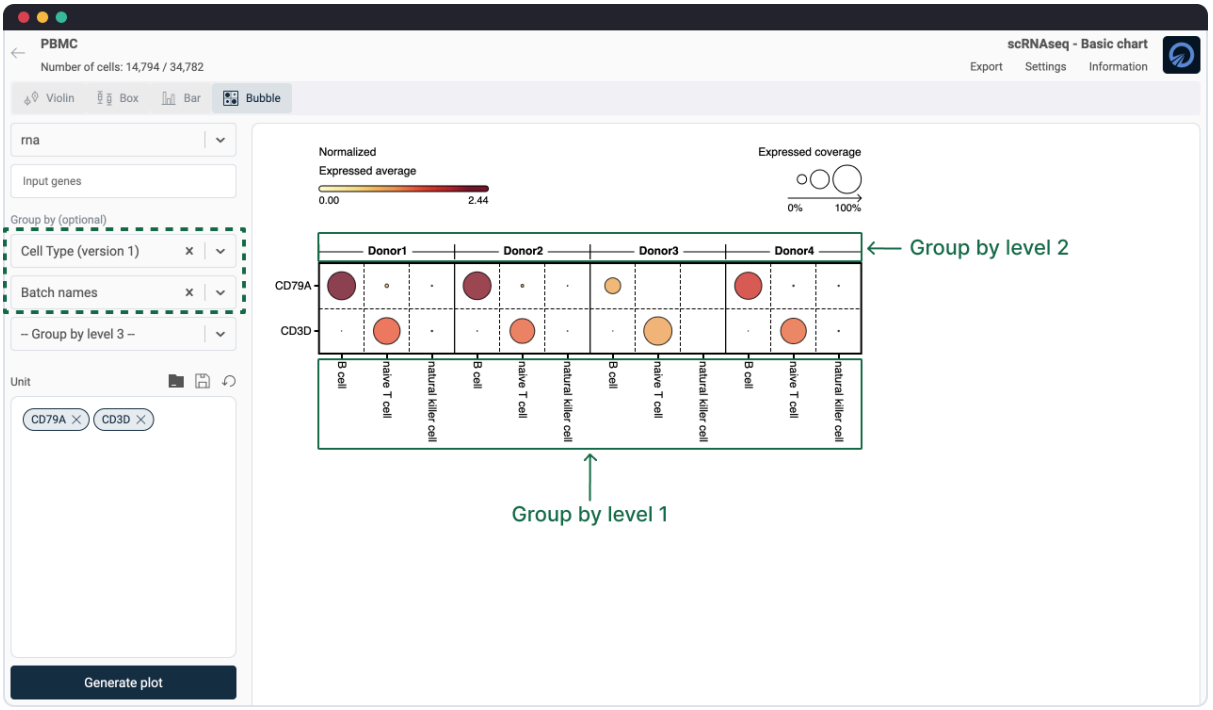
3. `Plotting`: Depending on the number of metadata levels selected, different plots are generated:
    - `None`: If no metadata is selected for grouping, all cells are treated as a single group, with each feature represented by a bubble.



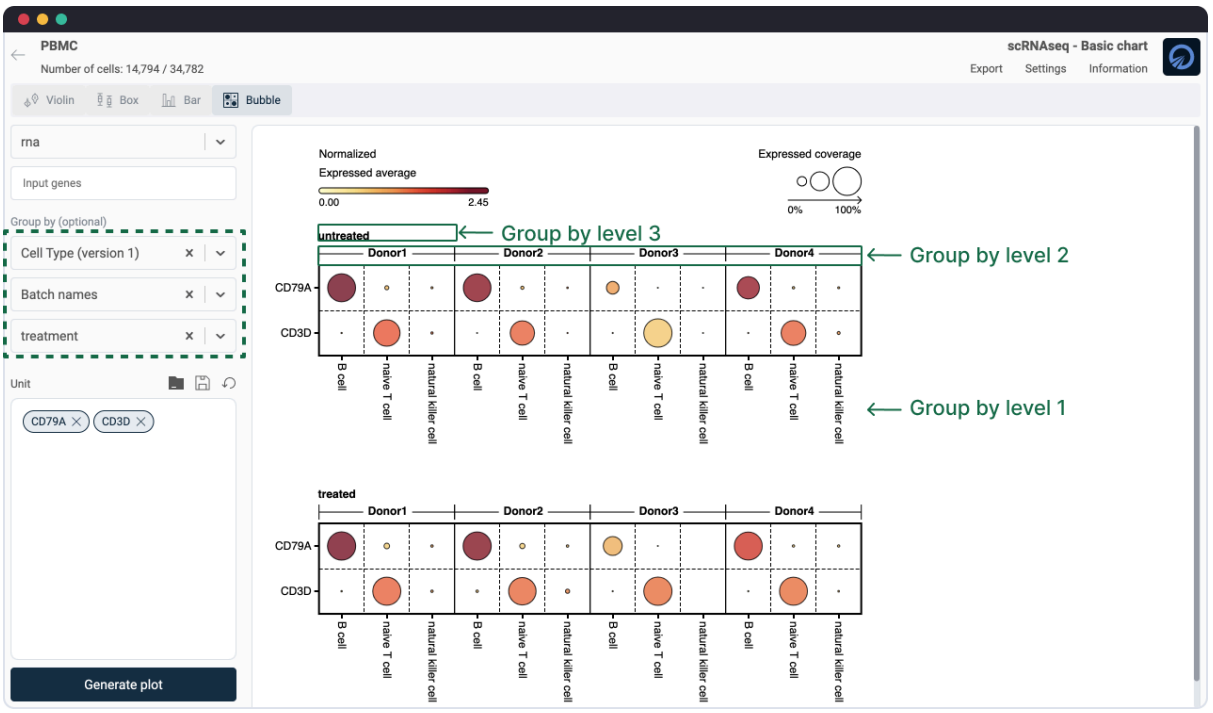
- `Group by Level 1`: Cells are divided into groups based on the first level of metadata, with each feature within these groups represented by individual bubbles.



- `Group by Level 2`: Cells are categorized by combining information from both the first and second metadata levels. Groups sharing the same level 2 metadata are presented adjacent to each other.



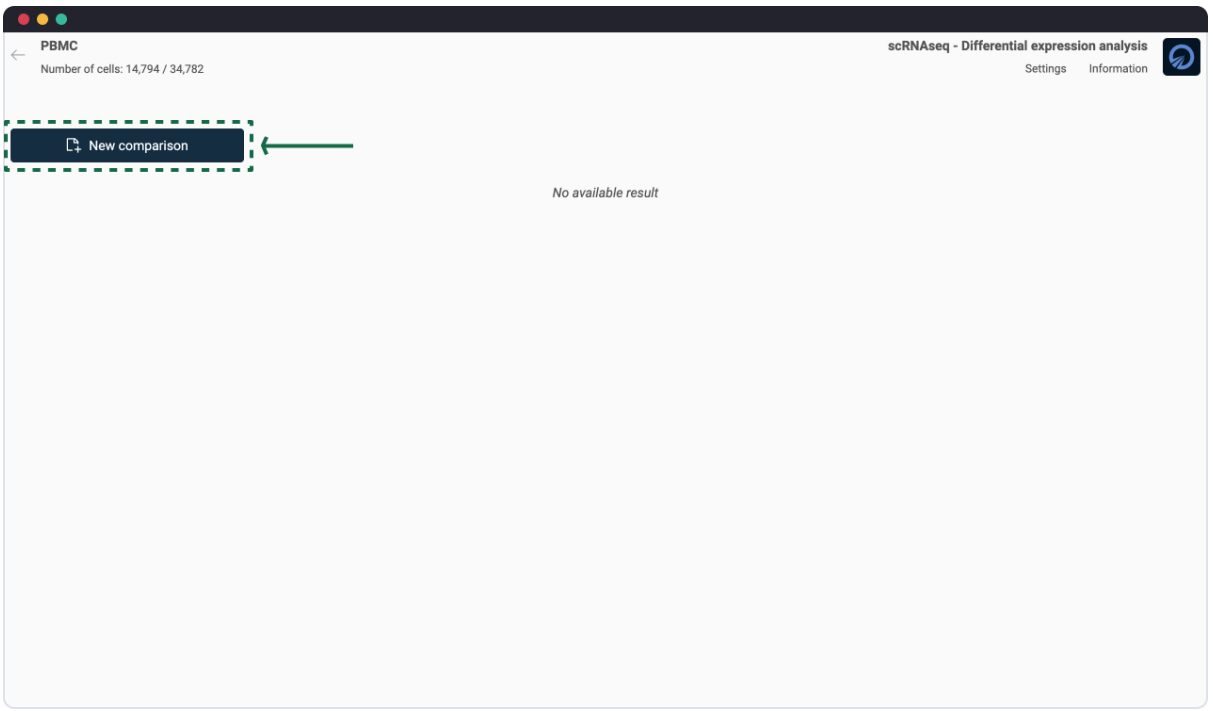
- `Group by Level 3`: Cells are first divided using the third level of metadata. Each such group is then visualized as a separate plot, incorporating features from the first two metadata levels.



All plots are drawn using the same scale for comparative analysis.

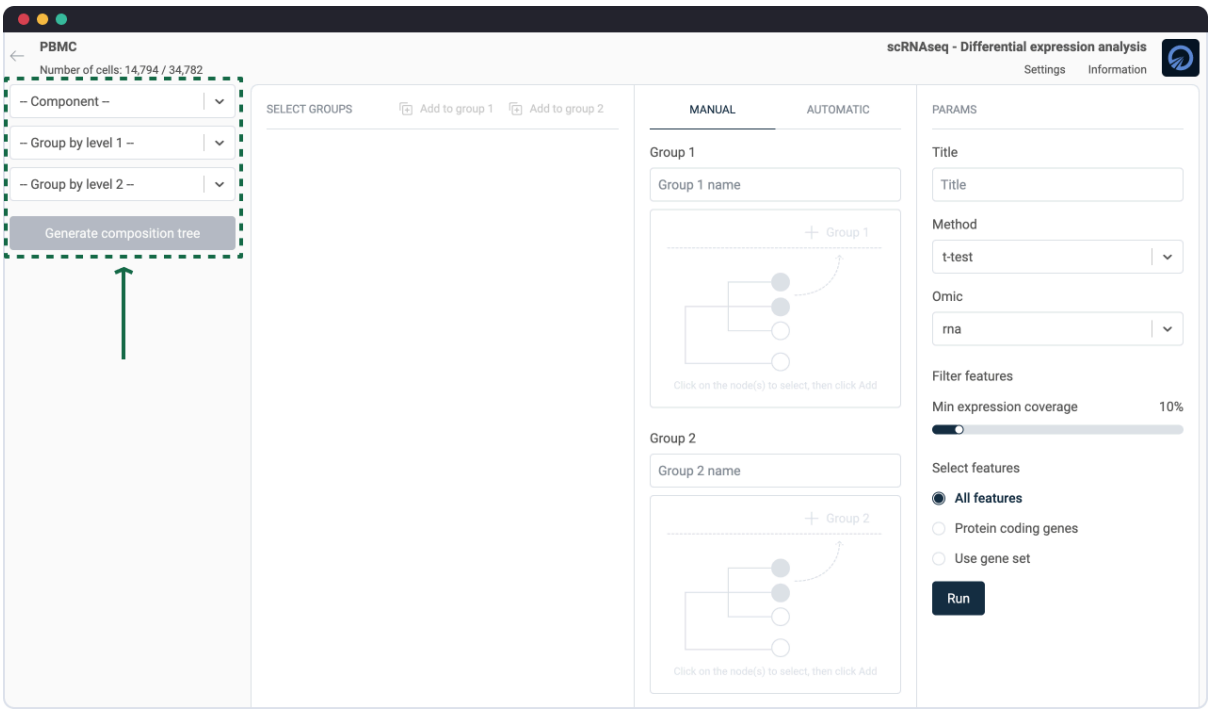
3. Differential expression genes (DEGs) analysis

To create a new DEGs analysis, click on the `New comparison` button on the top:

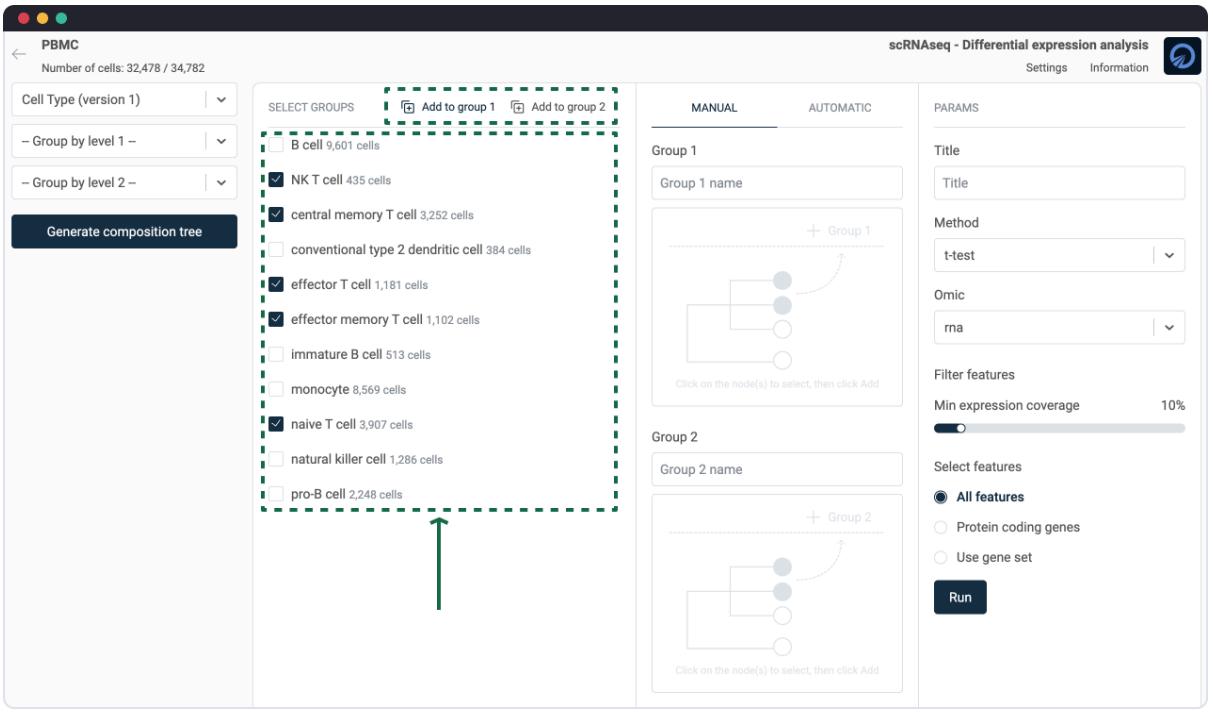


3.1. Select groups for comparison

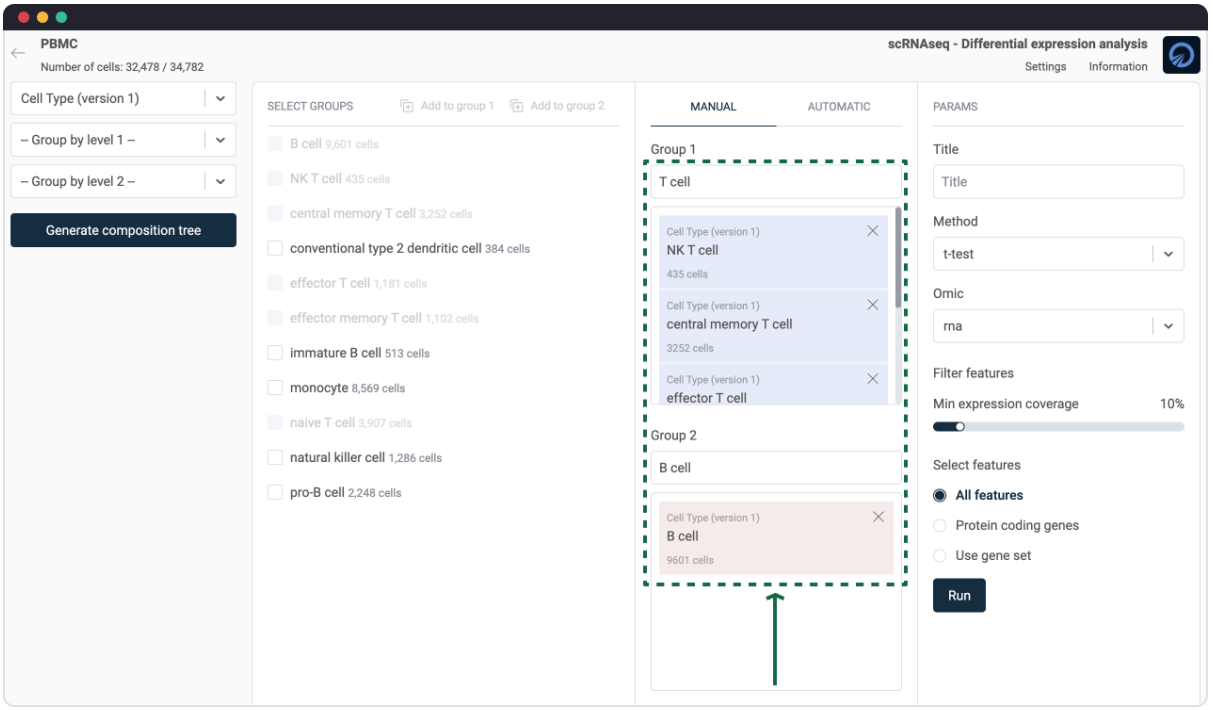
The first step in performing DEGs analysis is to define two groups for comparison. You can use the provided selection boxes to choose up to three metadata levels, then click on the generate composition tree button to generate a tree of all possible groups formed from the combined metadata information.



From the generated composition tree, select the desired nodes by checking the boxes next to each node name. You can add selected nodes to one of the comparison groups using the `Add to group 1` or `Add to group 2` buttons. These groups will then appear in the right panel.



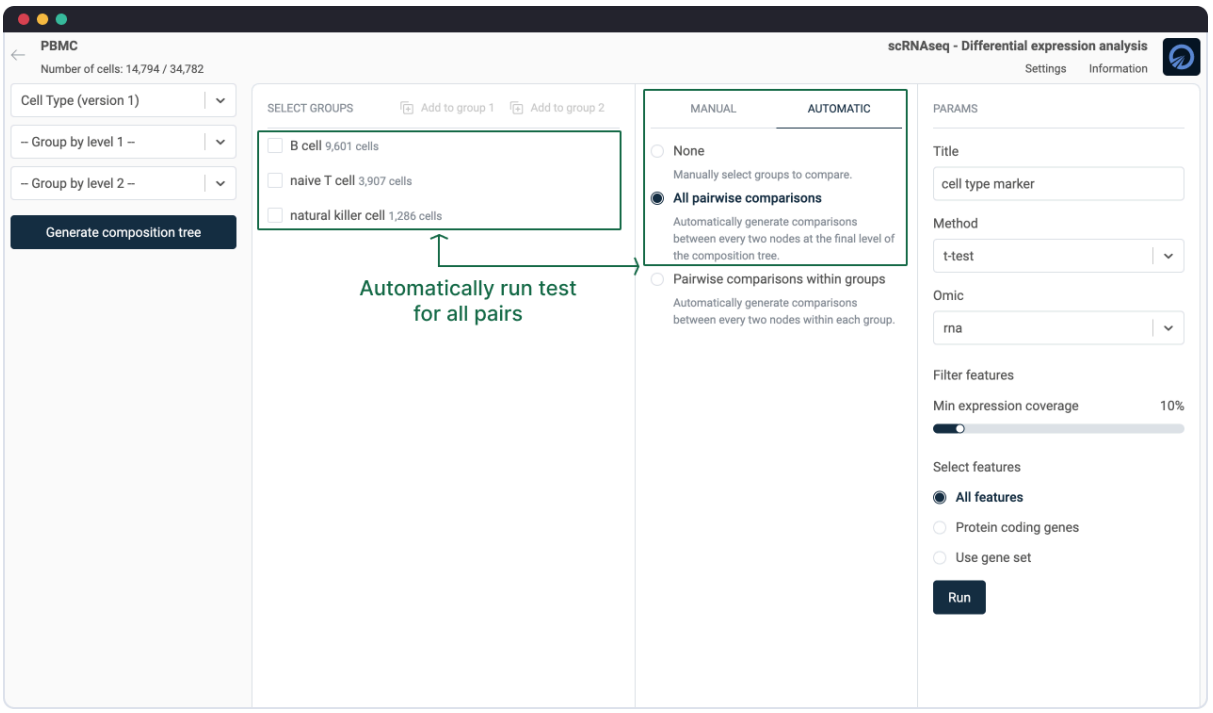
You can review your selected nodes in the right panel and remove any mistakenly selected ones by clicking the remove icon. Finally, input group names for both groups in the input box. Ensure each group contains at least 30 cells to maintain statistical significance.

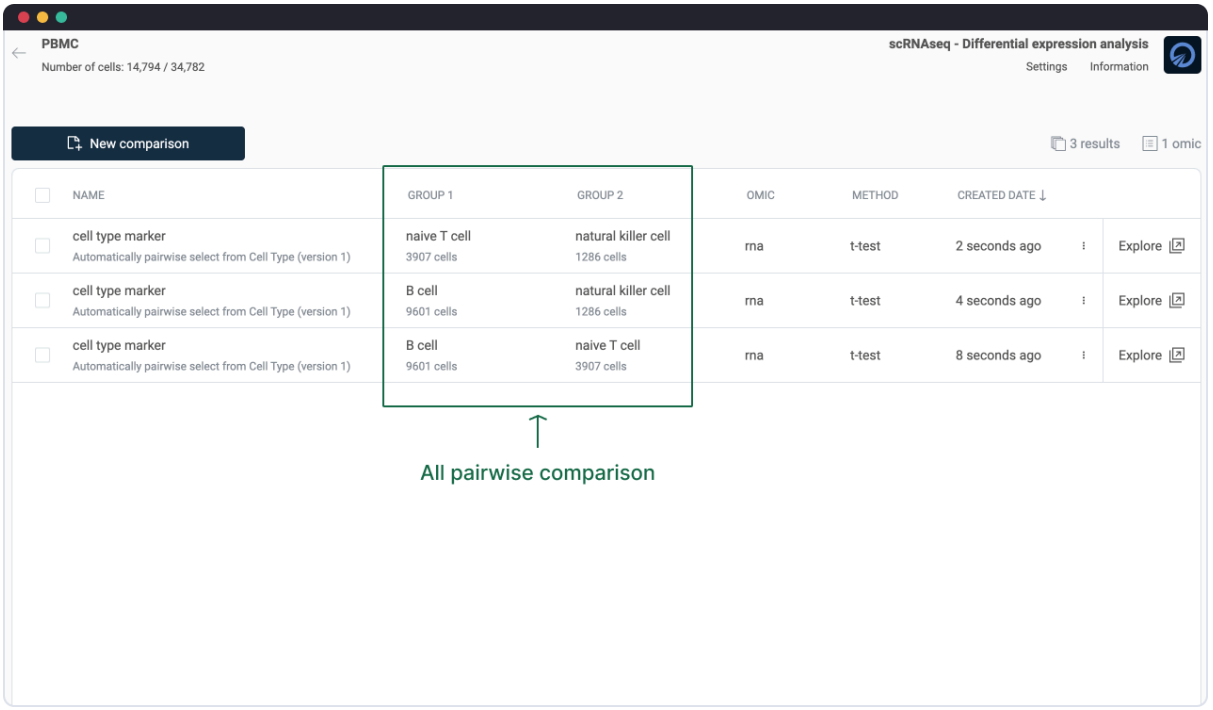


3.2. Automatic group selection

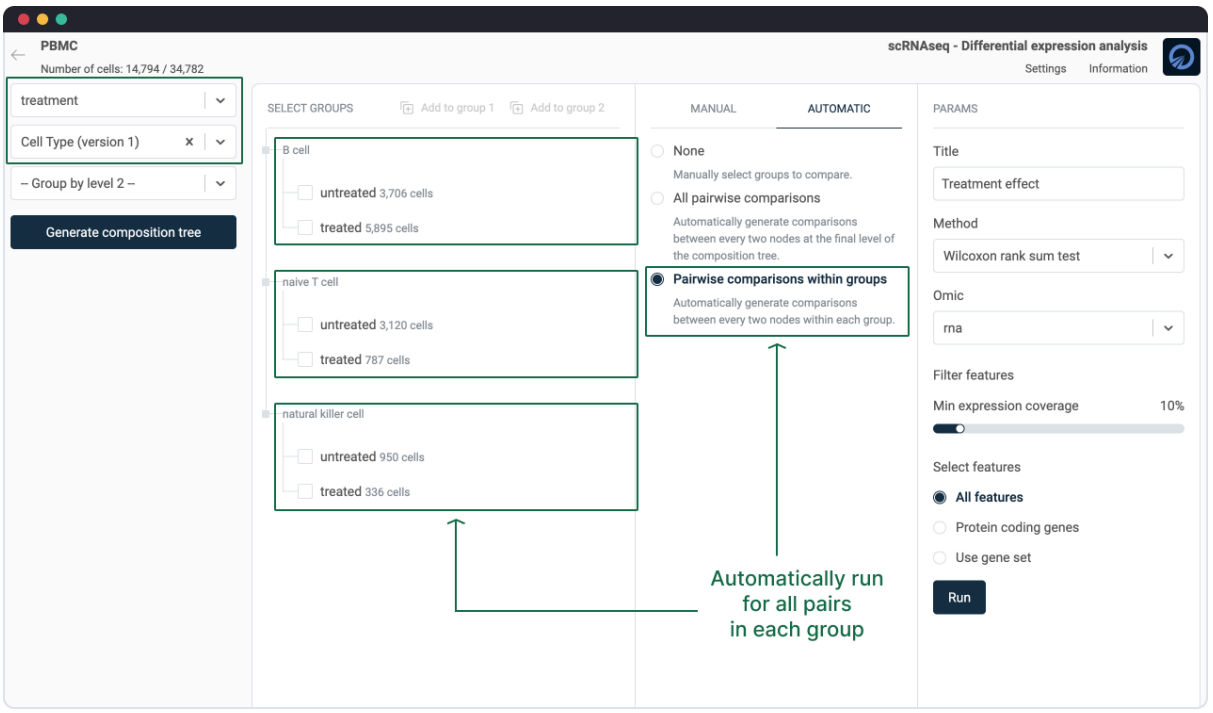
Beside manual selecting groups and performing each comparison, OmnibusX supports automated approaches to selecting and comparing groups, which is particularly useful for large-scale studies where manual selection would be time-consuming and less feasible. This functionality allows for systematic comparisons across all potential groupings derived from your specified metadata, which includes two below types:

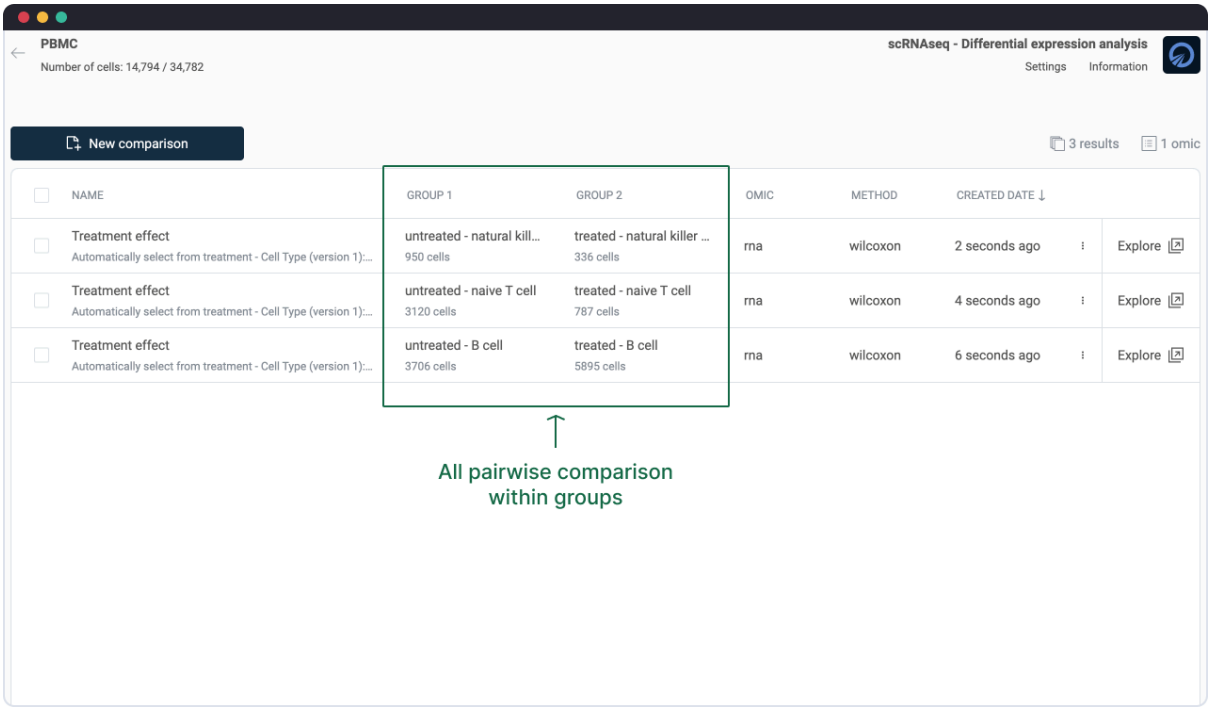
- `All pairwise comparison`: This option automatically generates comparisons between every two nodes at the final level of your composition tree. For example, if using cell type as a metadata, comparisons will be made between each pair of cell types. Groups with fewer than 30 cells will be skipped.





- `Pairwise comparisons within groups`: Comparisons are generated between nodes within each group based on the composition tree hierarchy. For instance, comparisons between different treatment types within the same cell type (e.g., treated vs. untreated B cells).

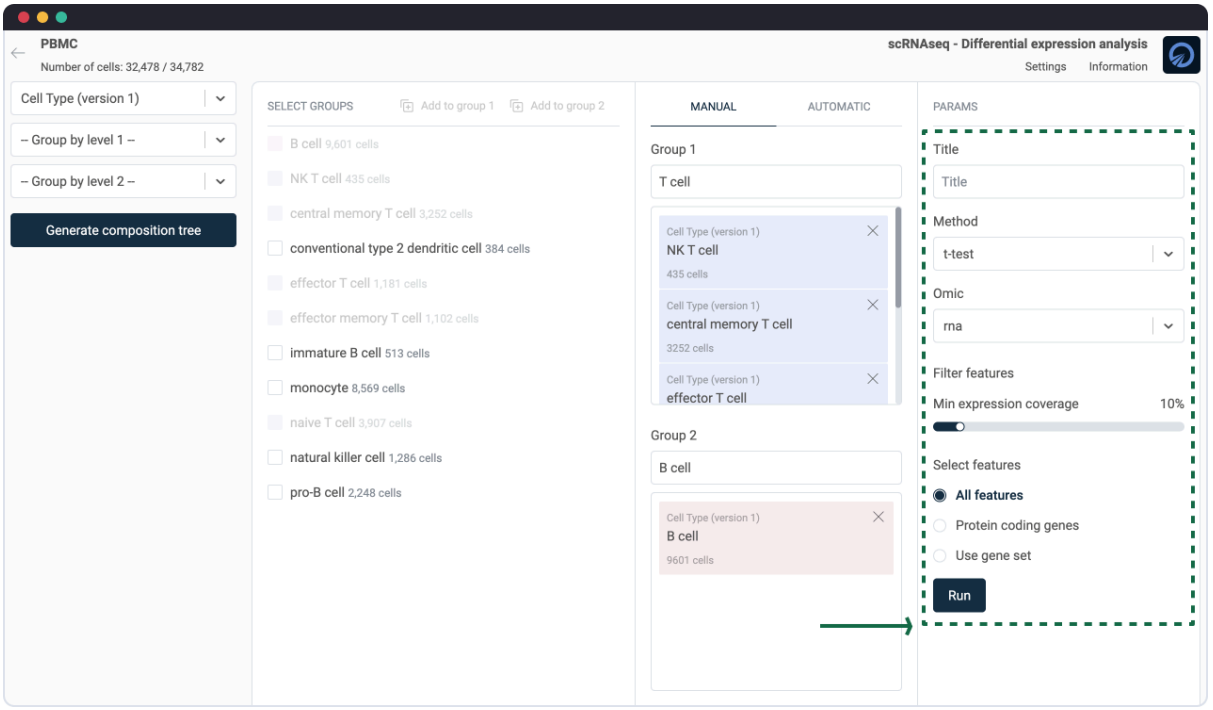




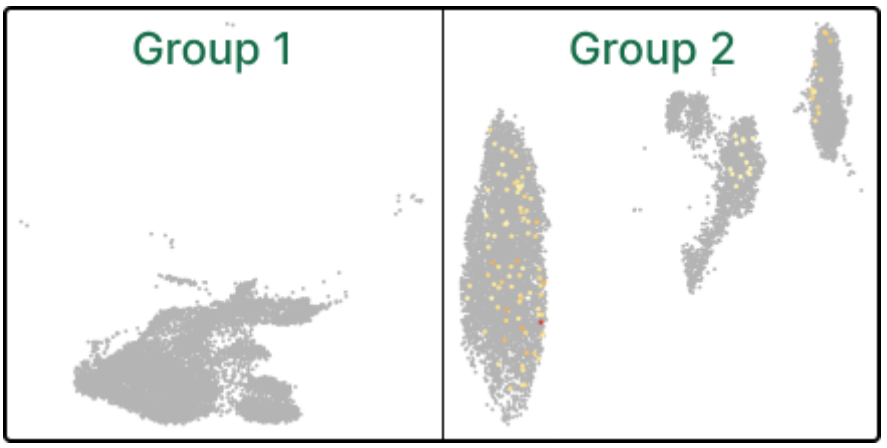
3.3. Parameters for DEGs analysis

OmnibusX allows you to customize several parameters to tailor the analysis to your specific research needs, including:

- `Method`: OmnibusX supports two statistical methods to determine differential expression:
    - `t-test`: Suitable for comparing two groups under the assumption that the feature expression data follows a normal distribution. The t-test implementation is from the SciPy team: scipy.stats.ttest_ind.
    - `Wilcoxon rank sum test`: A non-parametric test that does not assume a normal distribution of data, making it more robust against non-normal data distributions and outliers. The implementation is from the SciPy team: scipy.stats.mannwhitneyu.



- `Omic`: Select an omic you want to analyze from available omics in your dataset.
- `Min expression coverage:` You can set the minimum expression coverage to filter out features expressed by only a small percentage of cells, reducing noise in the results.

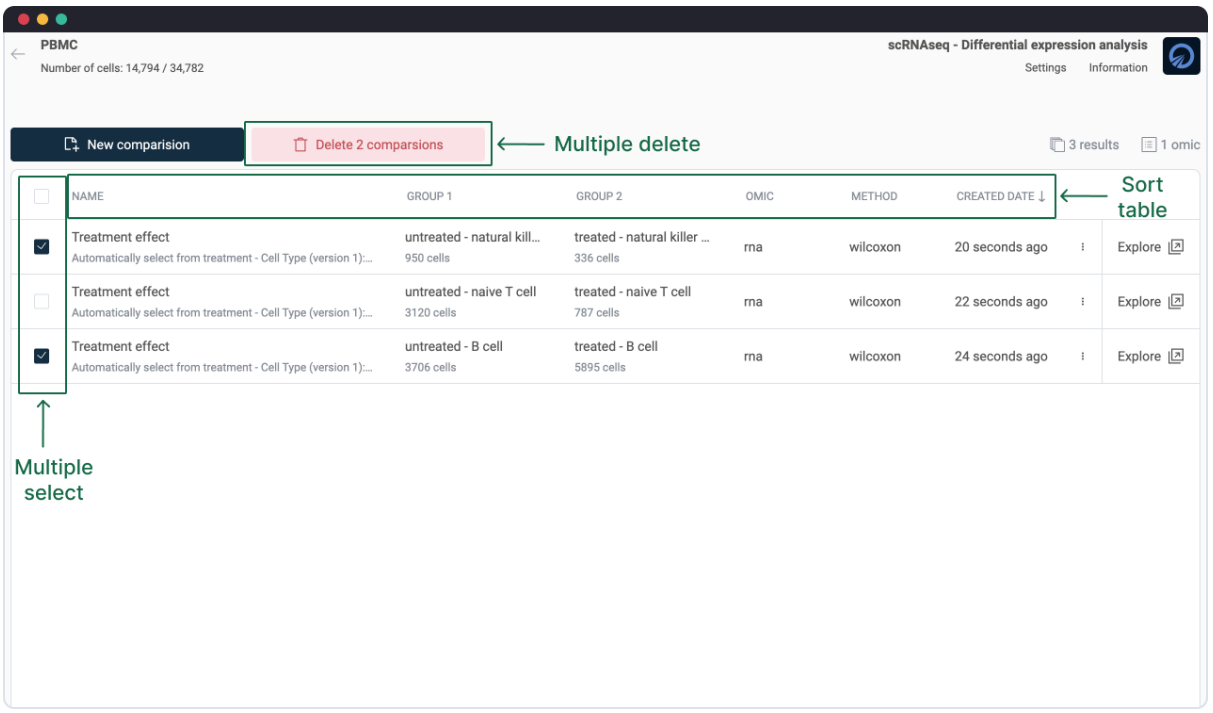


- `Select feature:` for `rna` omic, you can choose to include `All features` in the analysis, or focus on `Protein coding genes`, which are often of more biological interest. The list of protein-coding genes is sourced from the Ensemble annotation you selected during dataset submission. If you have defined specific gene sets in your gene set library (e.g., genes involved in a particular pathway), you can choose to analyze only these genes. This is particularly useful for targeted analyses where you are interested in specific biological processes or pathways.

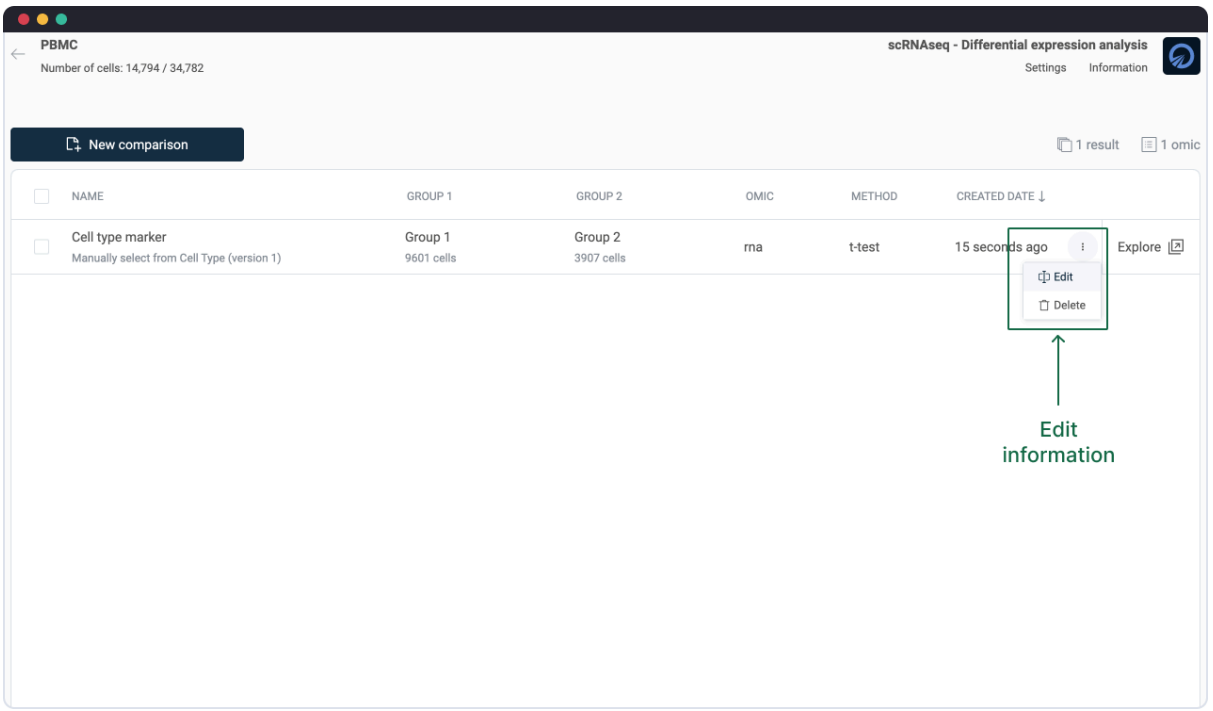
After setting all parameters, clicking the `Run` button will initiate the analysis. Depending on the complexity and size of the dataset, the analysis might take some time.

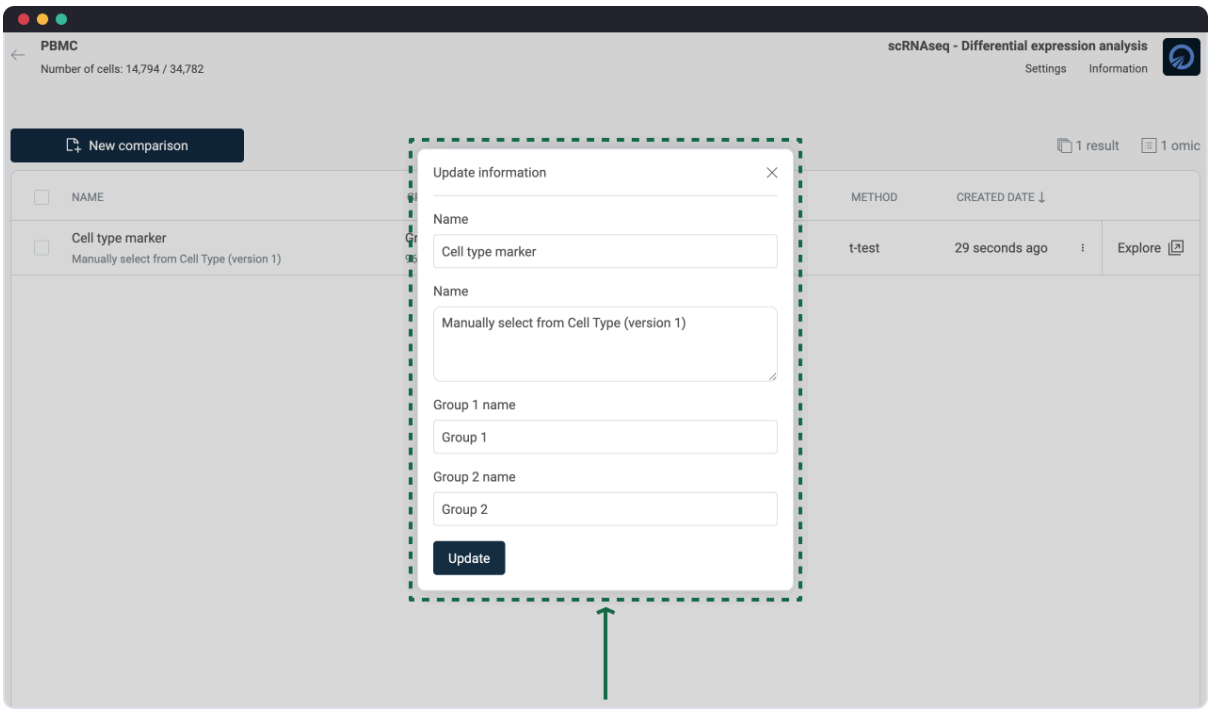
3.4. Manage DEGs results

All DEG analyses are stored and displayed in a table within the DEGs panel. You can click on the table headers to sort the results. You can also select multiple results using the corresponding checkboxes and perform bulk actions like deleting.



You can edit the details of a result by hovering over the dot icon next to each result and selecting `Edit`. Update the result title, group names, or description for better management.





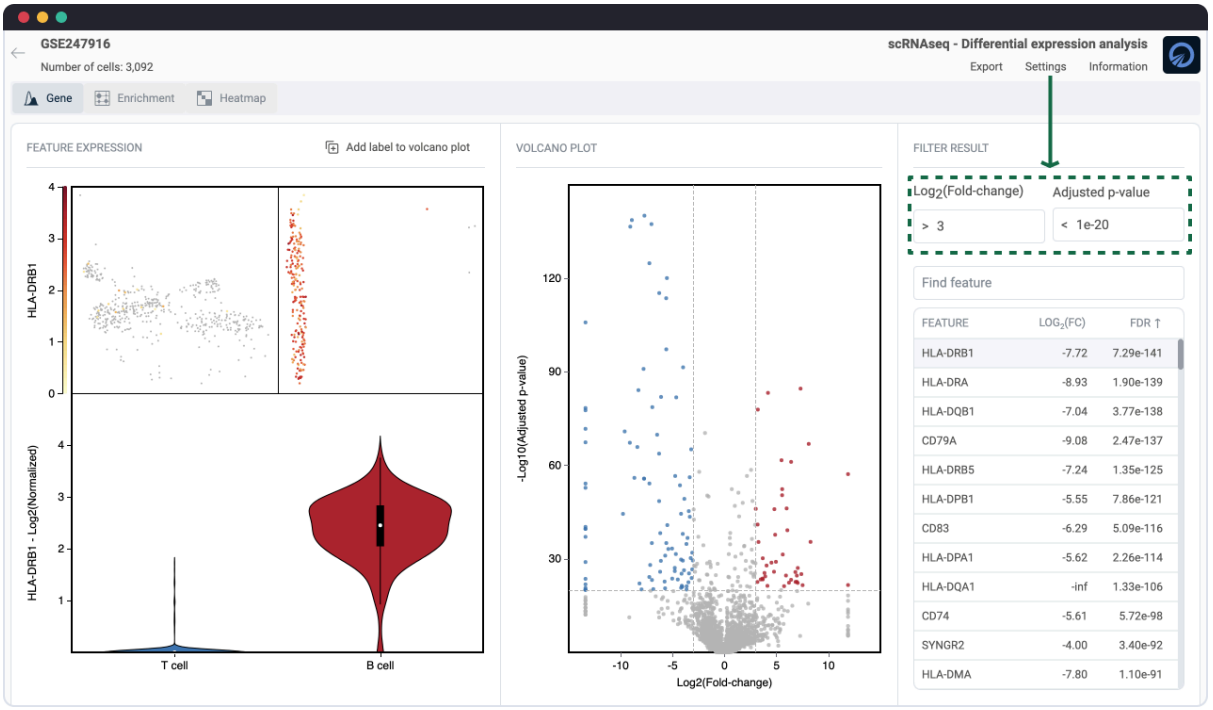
Clicking the `Explore` button to start analyzing a specific result.

3.5. Volcano plot

OmnibusX visualizes DEGs results using a volcano plot as described below:

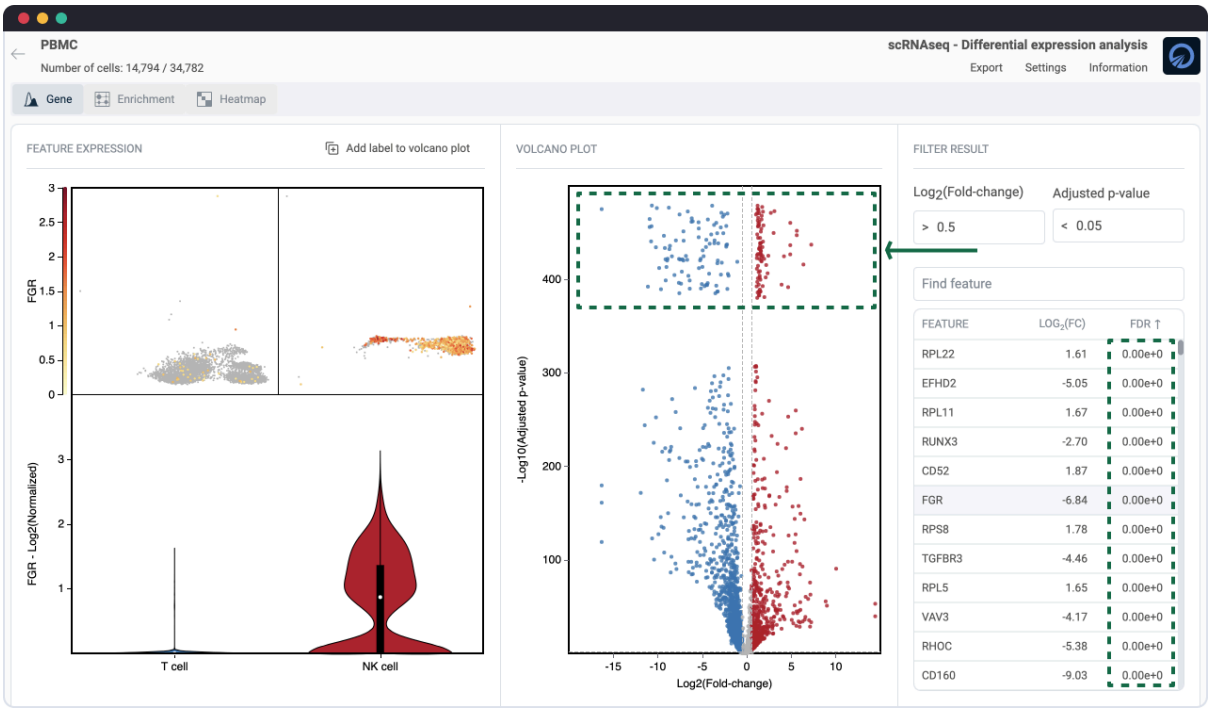
- `X-axis`: Represents the `log2(Fold-change)` of each feature between the two groups being compared. Positive values indicate higher expression in the first group, while negative values indicate higher expression in the second group.
- `Y-axis`: Represents the `-log10(Adjusted p-value)` of each test result, indicating the statistical significance of the expression differences.
- `Dots`: Each dot on the plot corresponds to a test result for a specific feature. By default, dots are colored as follows:
    - `red dots`: Indicate features with an `adjusted p-value` smaller than `0.05` and an `log2(Fold-change)` greater than `0.5`, indicating significant upregulation in the first group.
    - `blue dots`: Indicate features with an `adjusted p-value` smaller than `0.05` and an `log2(Fold-change)` smaller than `-0.5`.
    - `gray dots`: Features that do not meet the `adjusted p-value` and `log2(Fold-change)` thresholds for significant differential expression

You can customize the `adjusted p-value` and `log2(Fold-change)` thresholds for coloring the dots using the Filter result panel on the right.

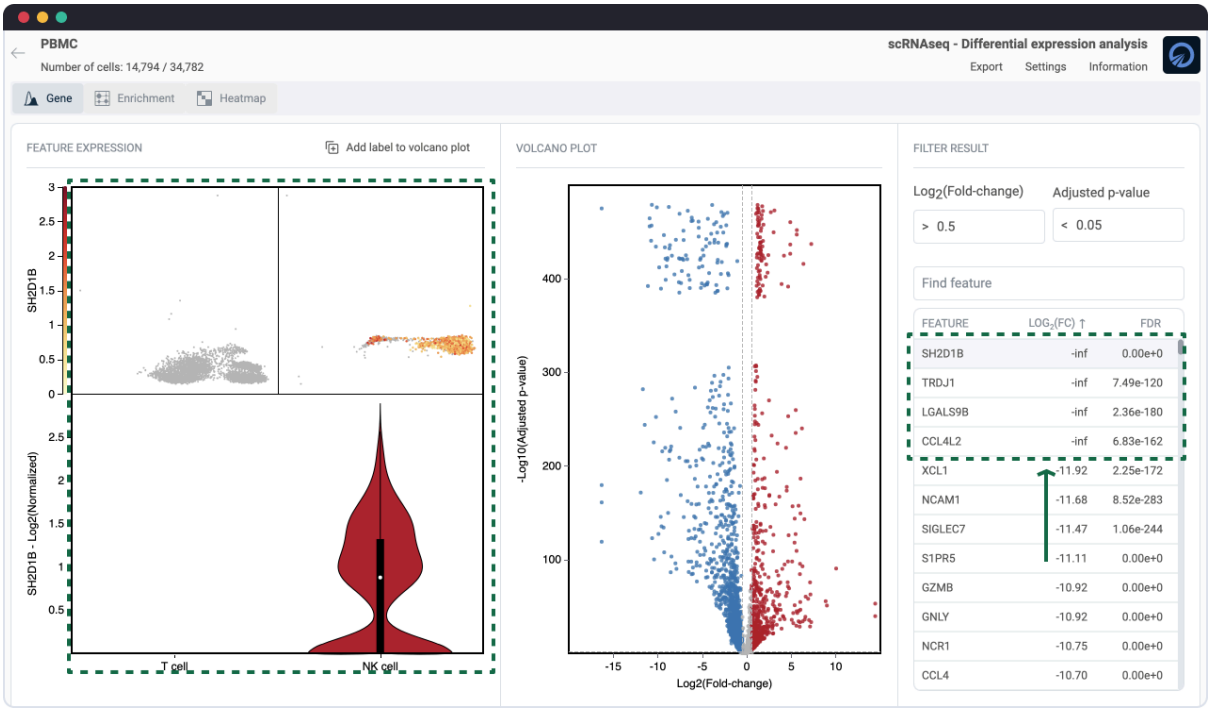


Given the vast number of features tested in a typical DEGs analysis, OmnibusX applies the Benjamini-Hochberg procedure to adjust raw p-values for multiple comparisons. This method, implemented by the Scipy team: statsmodels.stats.multitest.multipletests, controls the false discovery rate.

In single-cell data analyses, the abundance of data points can lead to extremely small p-values, some approaching the lower limits of computational precision. These small values can complicate the visualization of results in the volcano plot; as a p-value approaches zero, the value of `-log10(p-value)` approaches infinity. This extreme value can disrupt the scale and readability of the volcano plot, making it difficult to distinguish between significant results. To manage this, OmnibusX adjusts the plot for p-values smaller than `2.2250738585072014e-308` (the smallest positive number representable in the float64 data type), which have a `-log10` value of approximately `307.954`. For visualization purposes, any p-value below this threshold is plotted at `308 ± a small random variance`. Adding a small variance to these extremely significant points helps to jitter overlapping points.

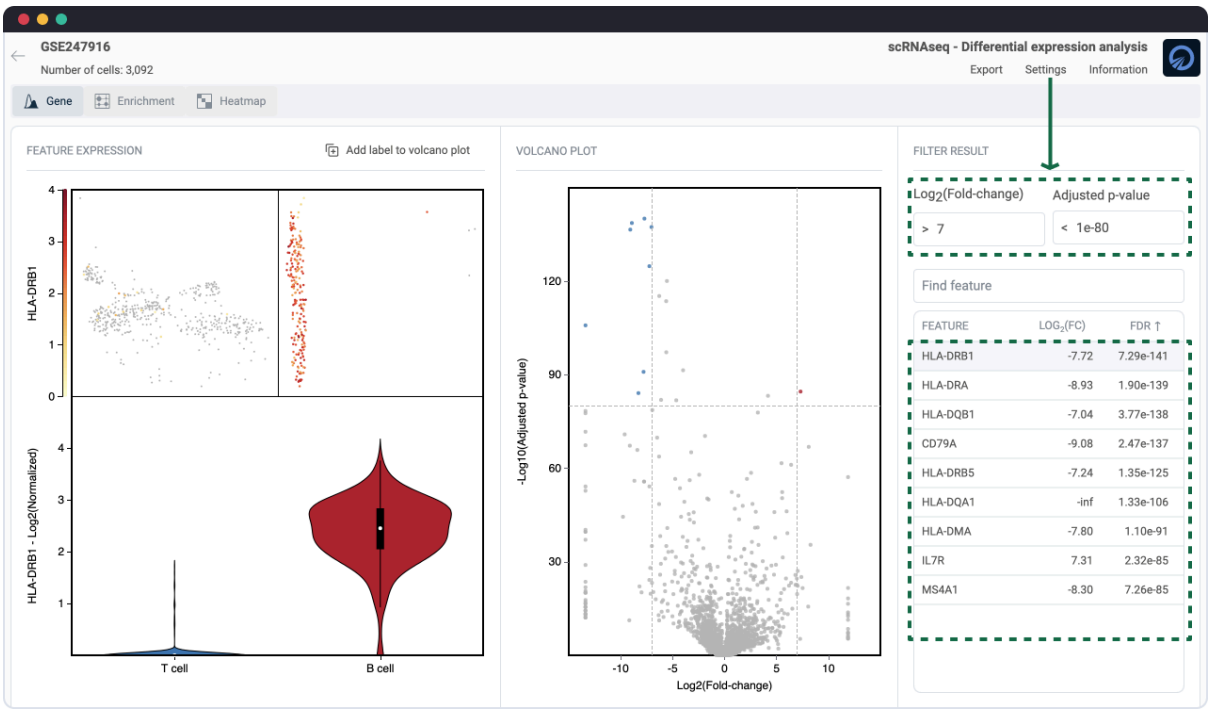


There are instances where the fold change calculation `log2(mean_1 / mean_2)` becomes problematic, such as when a feature is not expressed in one of the groups, resulting in a mean of zero. This situation leads to a division by zero error. As `mean_2` approaches zero, the ratio `mean_1 / mean_2` approaches infinity. OmnibusX handles this by assigning a value of `Infinity` to the `log2(Fold-change)` when mean_2 is zero, and `-Infinity` when mean_1 is zero. Since `Infinity` is not feasible to display on a standard plot, OmnibusX presents these values as the maximum `log2(Fold-change)` observed in the data plus a small delta.

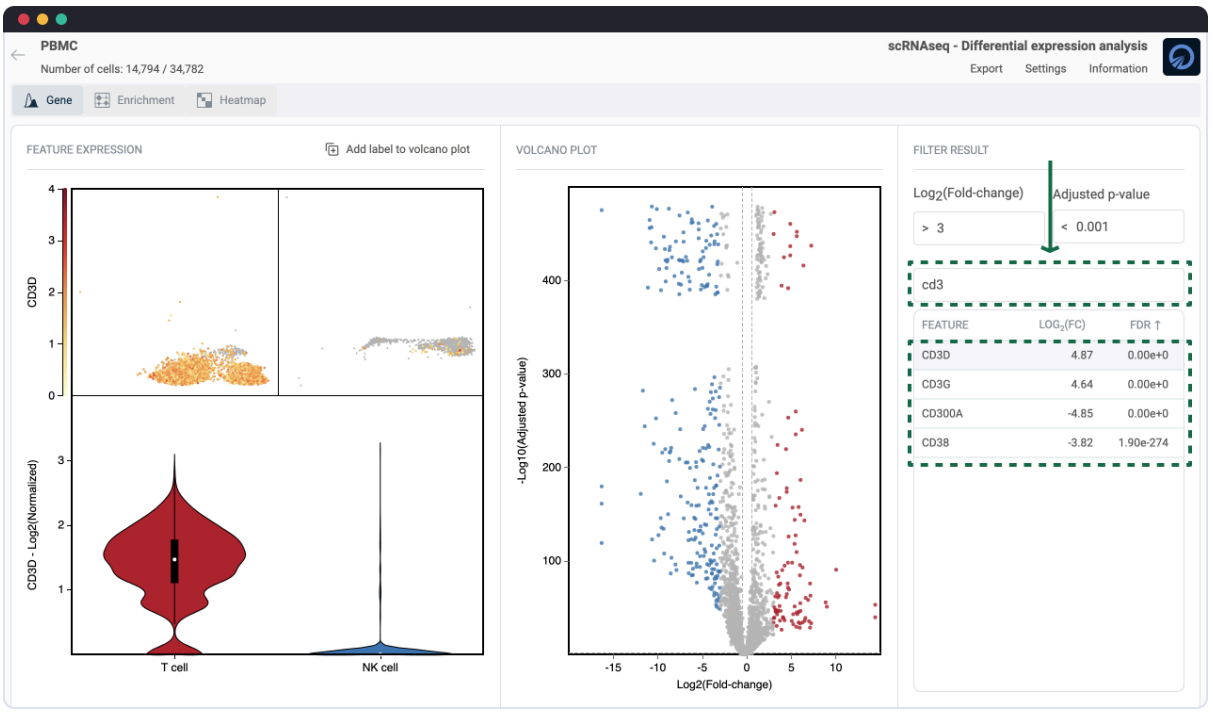


3.6. DEGs table

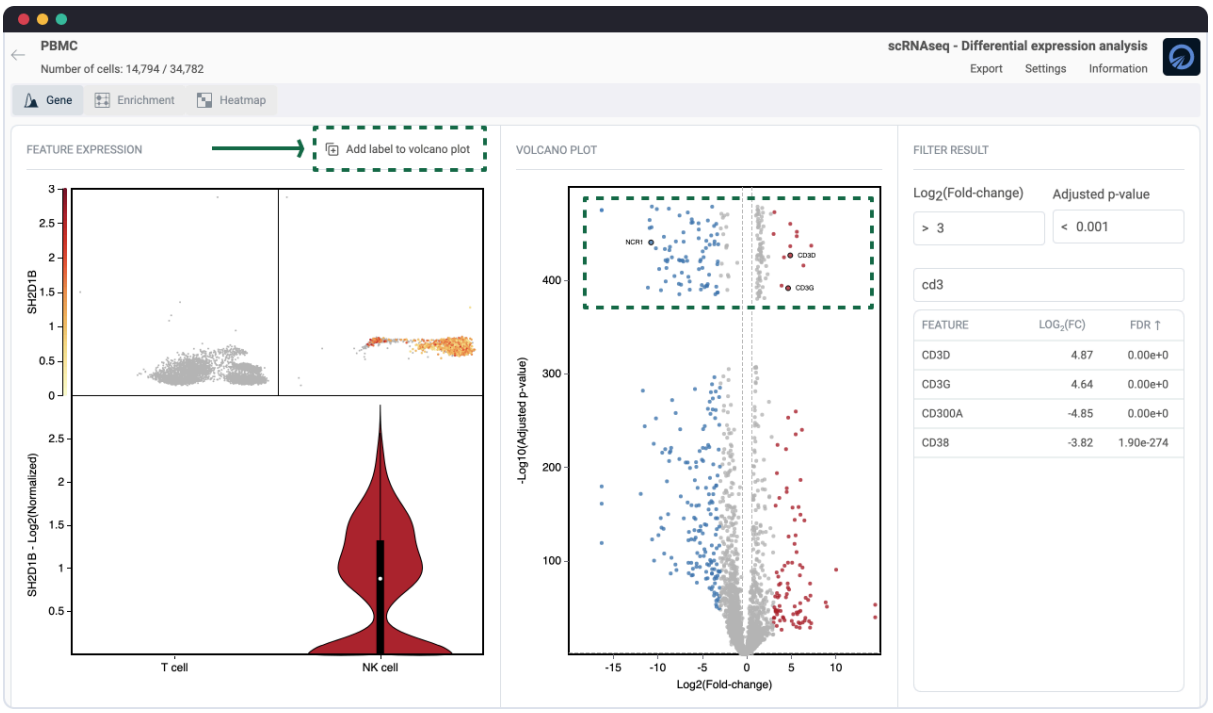
In addition to visualizing differential expression results via the volcano plot, OmnibusX provides a detailed DEGs table for further exploration and analysis. You can sort the table by clicking on any column header. Use the input boxes above the table to filter the results by `log2(Fold-change)` and `Adjusted p-value` thresholds. The `log2(Fold-change)` filter uses the absolute value.



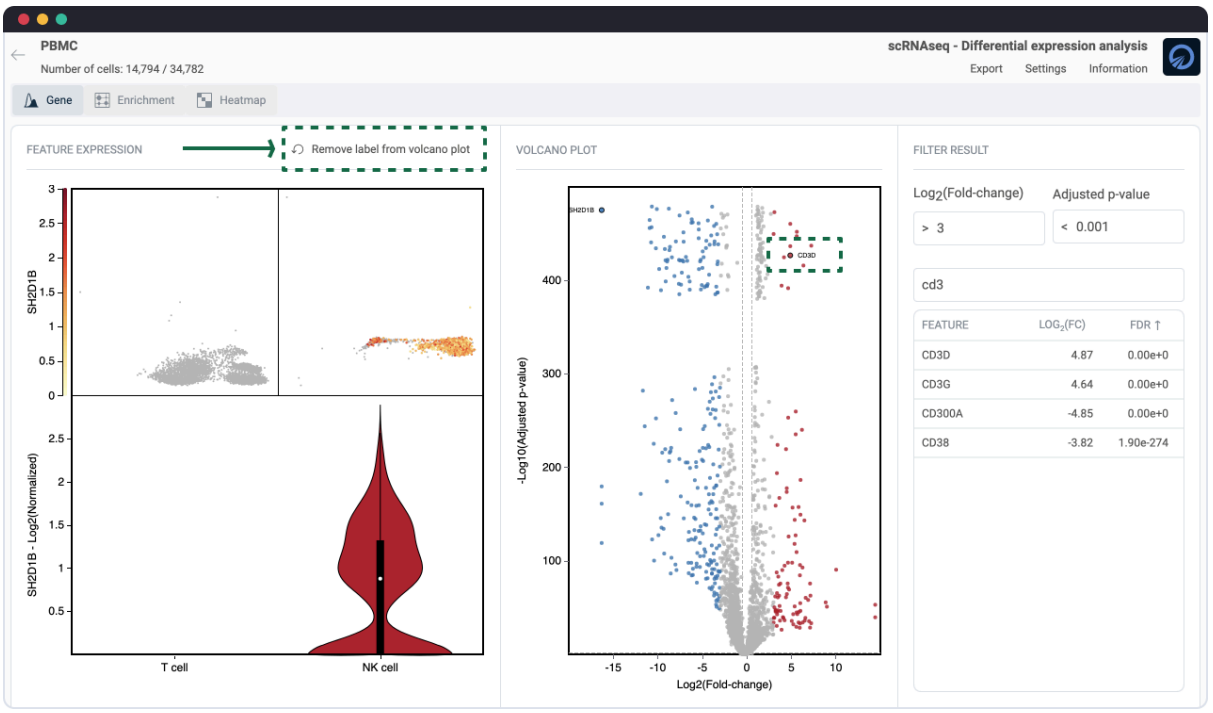
You can also find your target gene by typing its name in the search box and pressing `Enter`. This filters the table to show only entries related to the searched gene, making it easy to locate and analyze specific data quickly.



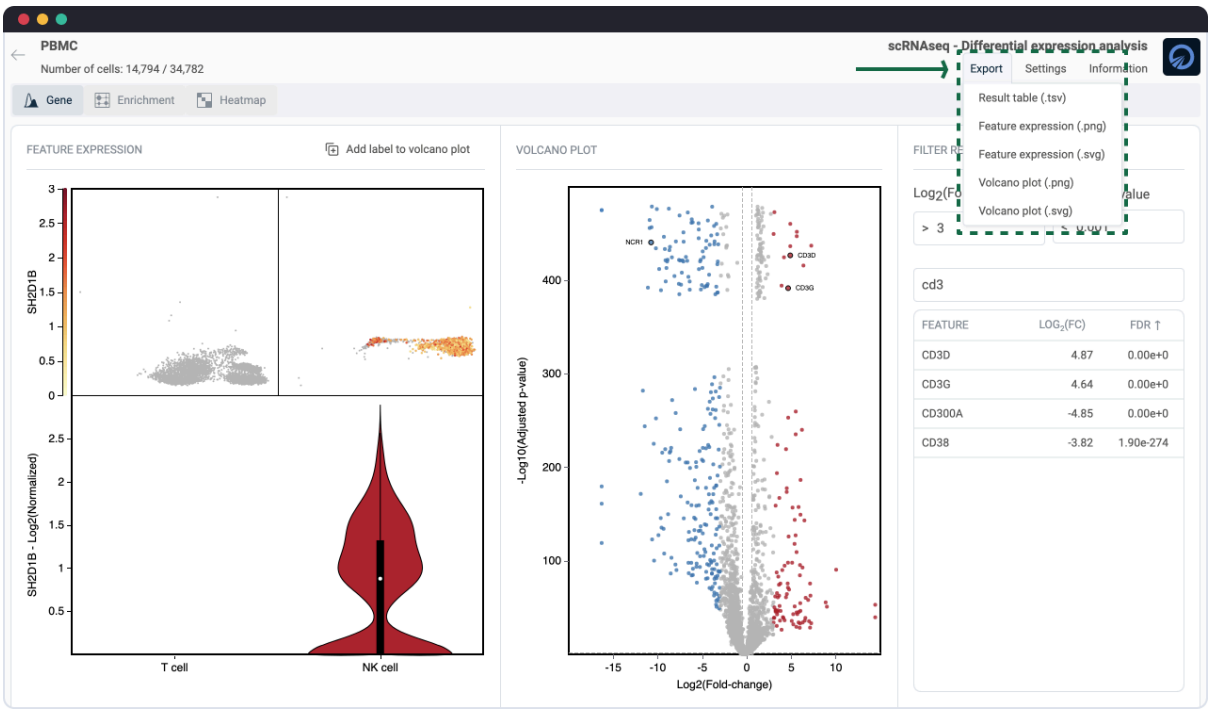
By clicking on a feature within the table, you can view detailed expression patterns for that particular feature across the two comparison groups. These patterns are visualized through scatter plots and violin plots displayed in the left panel. To highlight a specific feature on the volcano plot, click the `Add label to the volcano` plot button. This action places a label on the corresponding dot on the volcano plot.



Labels can be removed by either clicking again on the labeled dot in the volcano plot or by pressing the `Remove label from volcano plot` button.



The DEGs results can be exported as a `.tsv` file by clicking the `Export` button located on the top menu. All plots can be exported as PNG or SVG files.

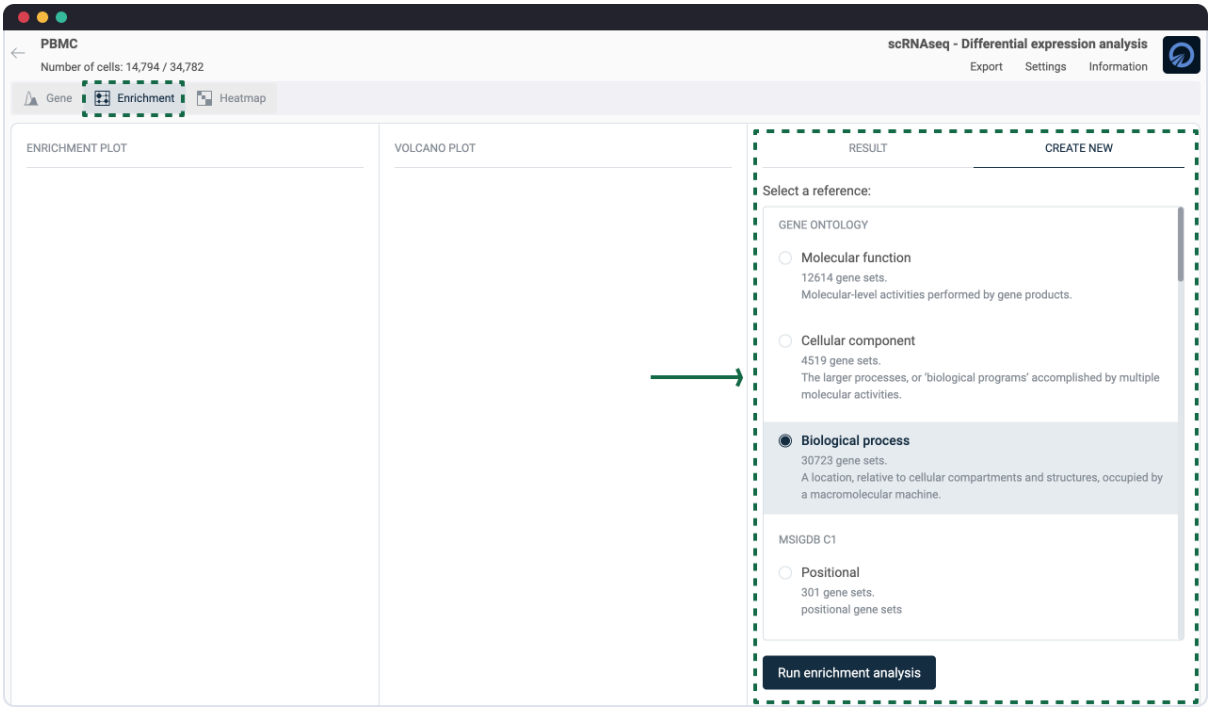


3.7. Differential enrichment analysis

OmnibusX supports comprehensive analysis for identifying significant differences in gene set enrichment scores between two groups. This analysis provides insights beyond mean expression differences by focusing on the behavior of predefined gene sets. OmnibusX integrates gene sets from two primary sources:

- MsigDB: Molecular Signatures Database
- Gene ontology: Gene Ontology Consortium

Choose a collection that aligns with your research to start analysis, which includes the below steps:



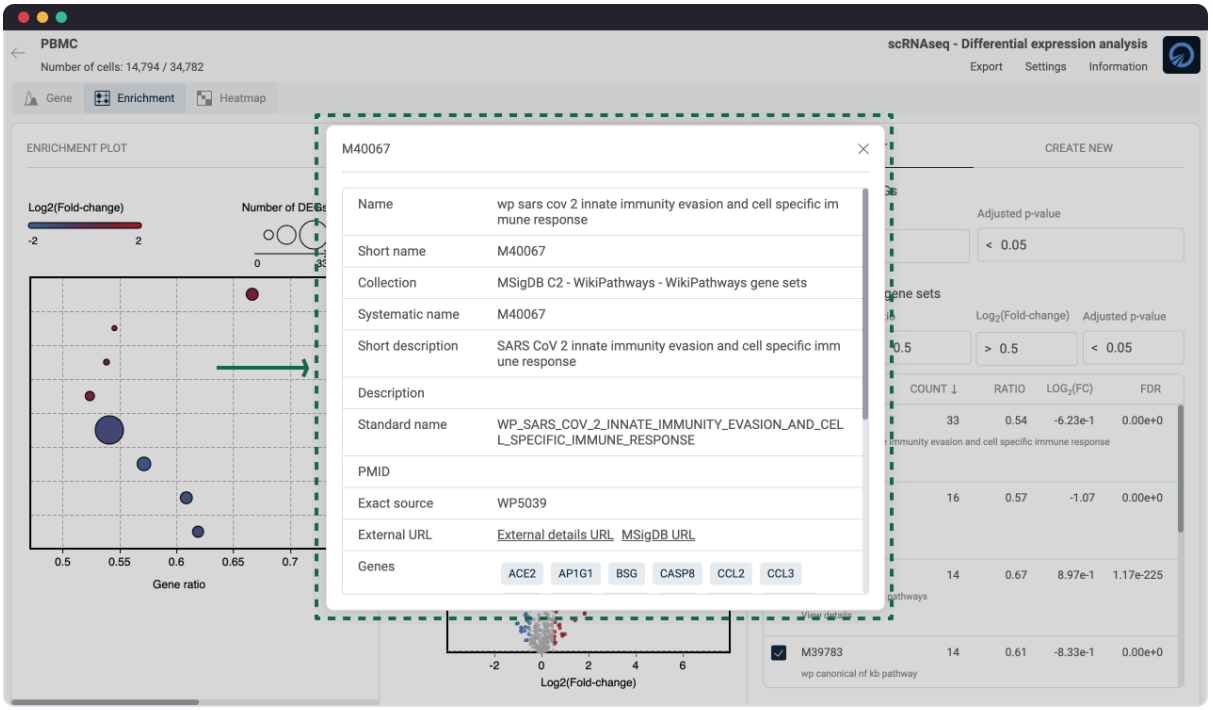
- `Enrichment score calculation`: Once a collection is selected, OmnibusX calculates the enrichment scores for each cell in both groups against all gene sets in the collection. The enrichment scores are calculated using AUCell algorithm (Bioconductor package documentation).
- `Statistical testing`: A t-test (scipy.stats.ttest_ind) is performed to compare the enrichment scores between the two groups for each gene set. `p-values` generated from these tests are corrected for multiple

comparisons using the Benjamini-Hochberg procedure to control the false discovery rate.

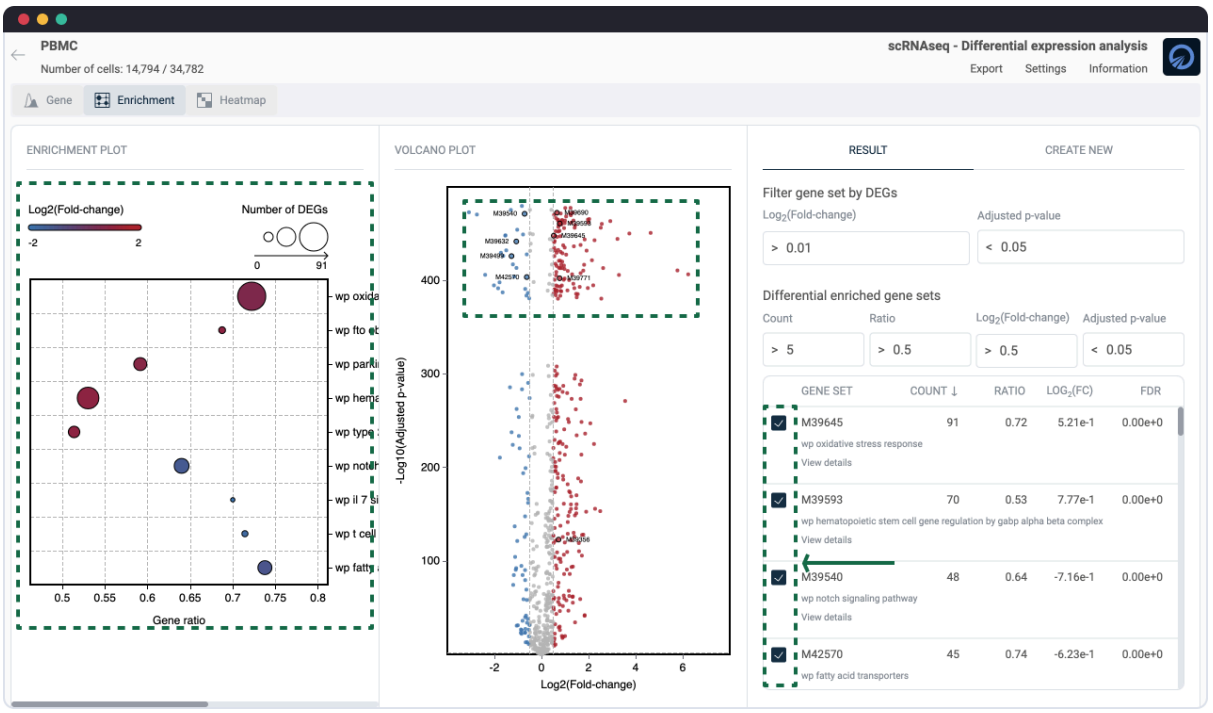
- `DEGs integration`: OmnibusX uses the results of DEGs analysis between the two groups to calculate the number of genes in each gene set that are differentially expressed (`Count`) and the fraction of these genes relative to the total number of genes in the gene set (`Ratio`).

The statistical testing results are visualized using a volcano plot, similar to DEGs analysis, where each dot represents a gene set. All analysis results are presented in a table displaying four key values for each gene set: `Count`, `Ratio`, `Log2(Fold-change)`, and `Adjusted p-value`. You can sort the results by clicking on the table headers or filter them based on thresholds set via the input above the table.

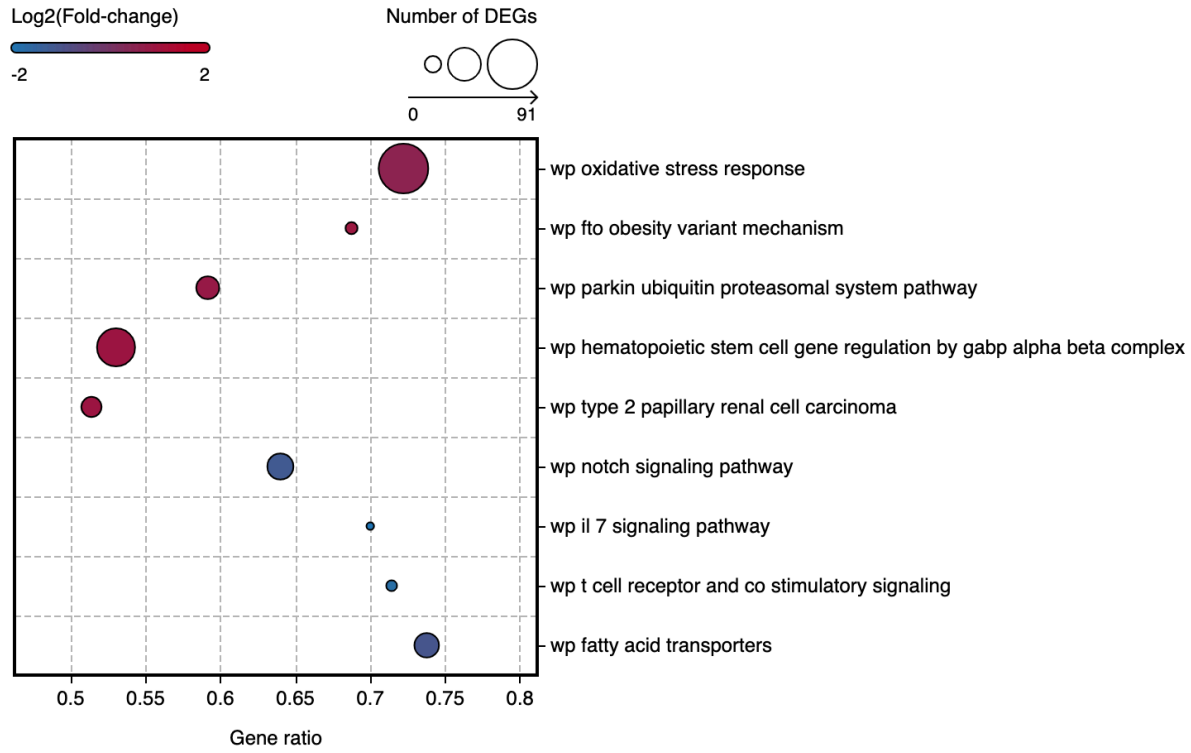
Click on `view detail` in each row to access detailed descriptions about each gene set.



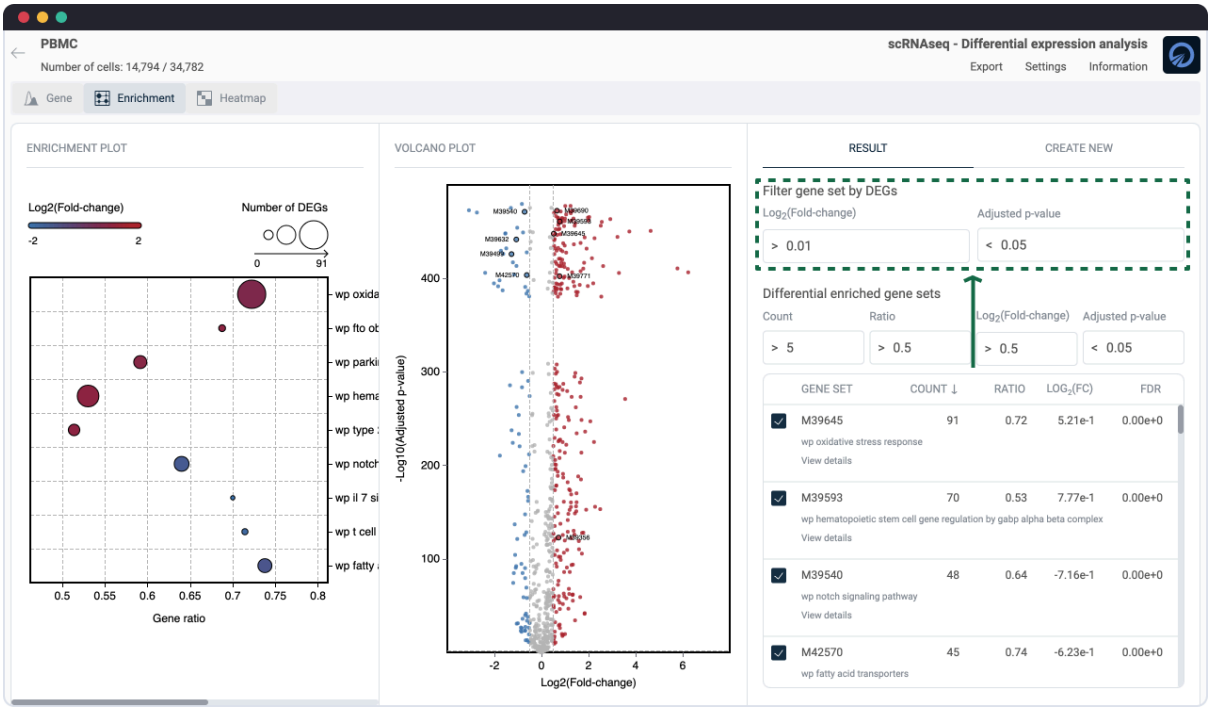
Once you find an interesting gene set, you can click on the checkbox to select it. The selected gene sets are labeled on the volcano plot and visualized using a scatter plot for further exploration.



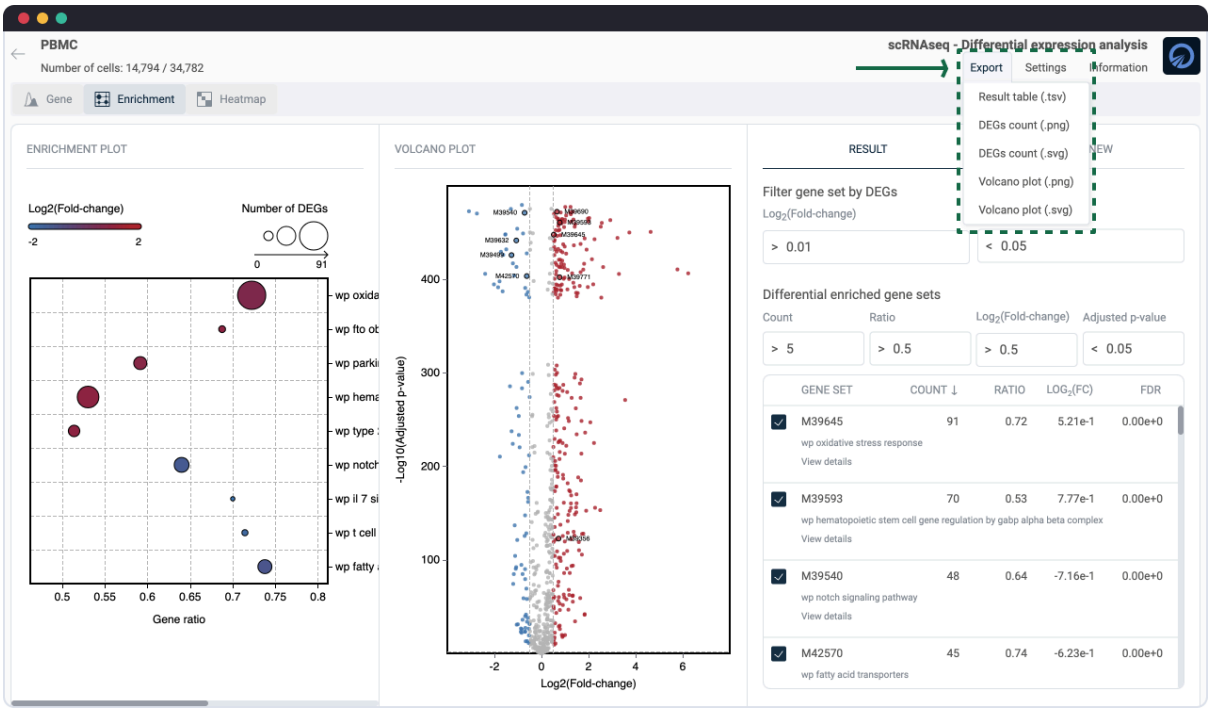
On the scatter plot, each dot represents a gene set enrichment result. The x-axis presents the `Ratio` of differentially expressed genes versus total gene set genes. The color represents the `log2(fold-change)` of enrichment score between the two groups for that gene set. The size of the dot represents the number of differentially expressed genes (`Count`) between the two groups that belong to that gene set.



You can adjust the `Count` and `Ratio` values by modifying the input list of DEGs using the DEGs `log2(fold-change)` and `adjusted p-value` thresholds for filtering.



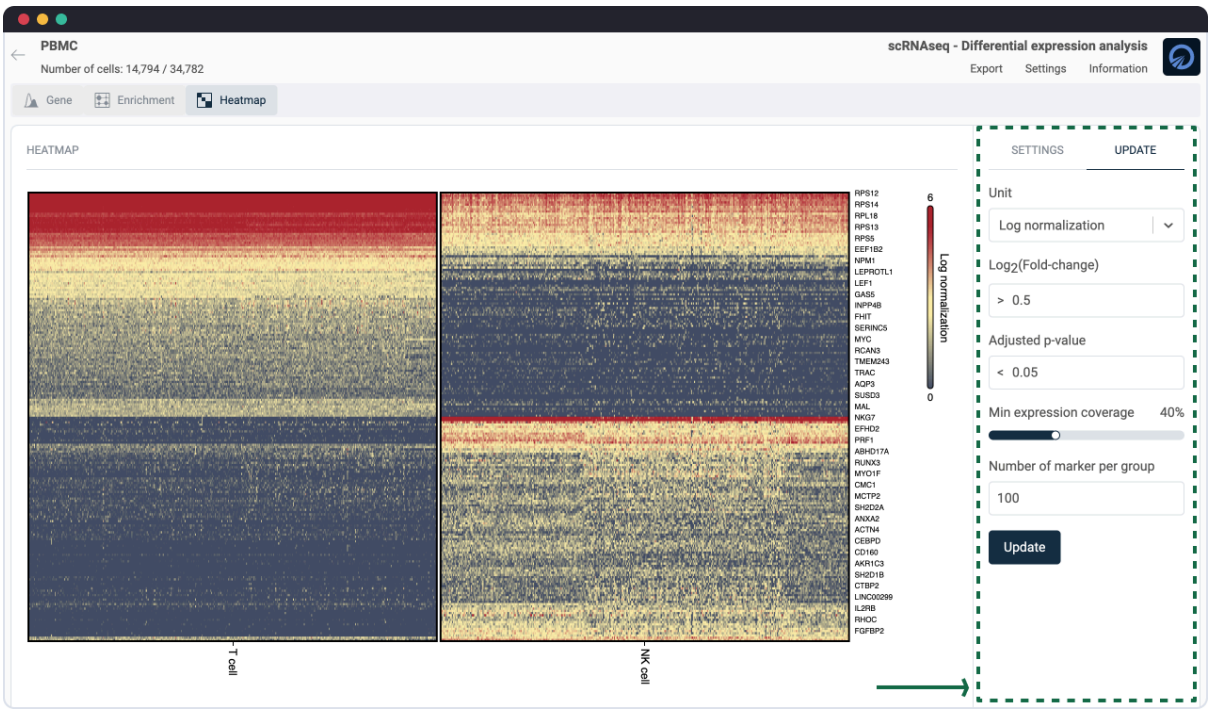
Results can be exported as a .tsv file using the export button located at the top menu. All plots can be exported as PNG or SVG files.



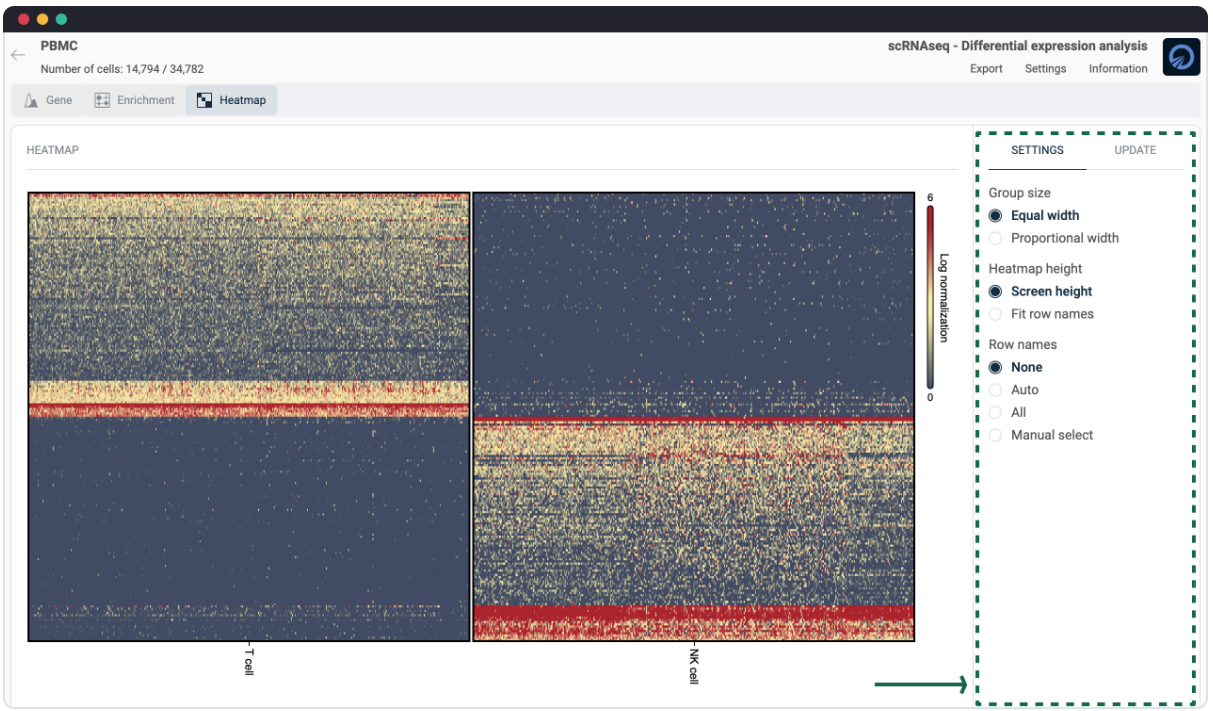
3.8. Heatmap

Heatmaps are an effective way to assess the overall expression patterns of significantly up-regulated and down-regulated features from DEGs results. OmnibusX supports you to create a heatmap starting by filtering the DEGs results according to a `p-value` threshold. Once filtered, sort the features by `log2(fold-change)` value to identify the top highest (up-regulated) and lowest (down-regulated) features. Then, draw a heatmap with these features using expression values of a selected unit.

You can customize the number of top results to display, adjust the filtering threshold, or change the unit of expression used for visualization. These settings are accessible in the `Update` panel.

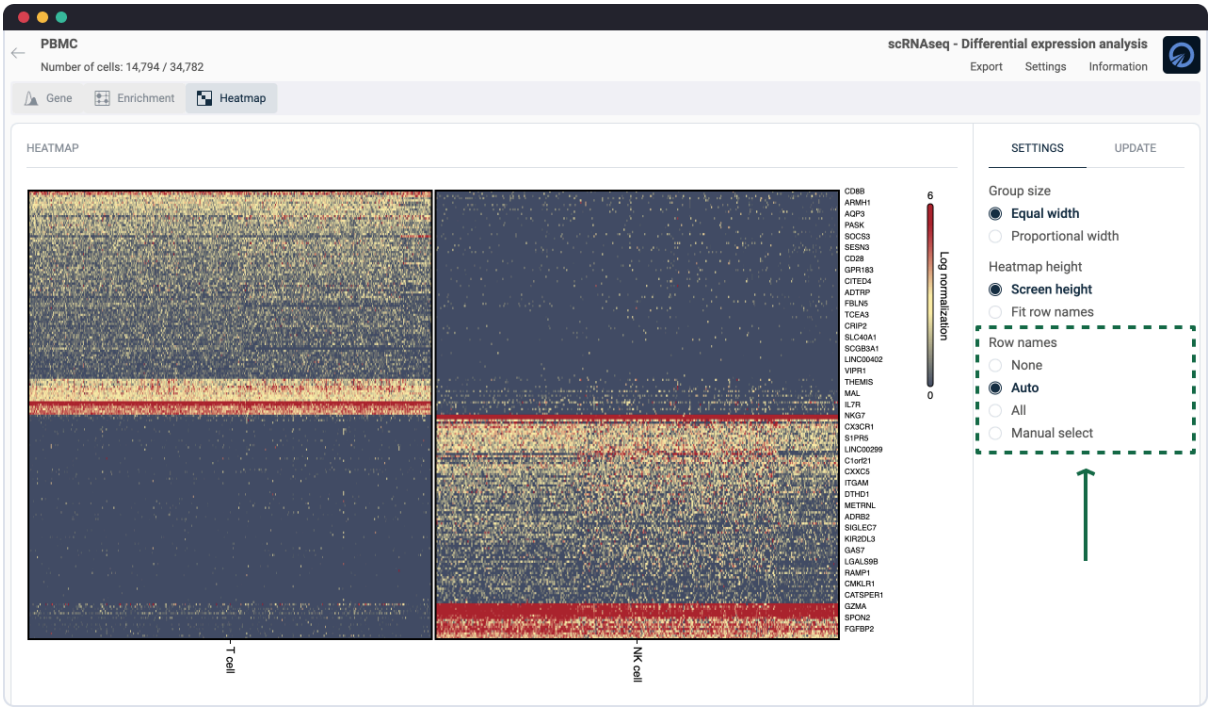


Once the heatmap is created, you can modify its appearance through the `Settings` panel.

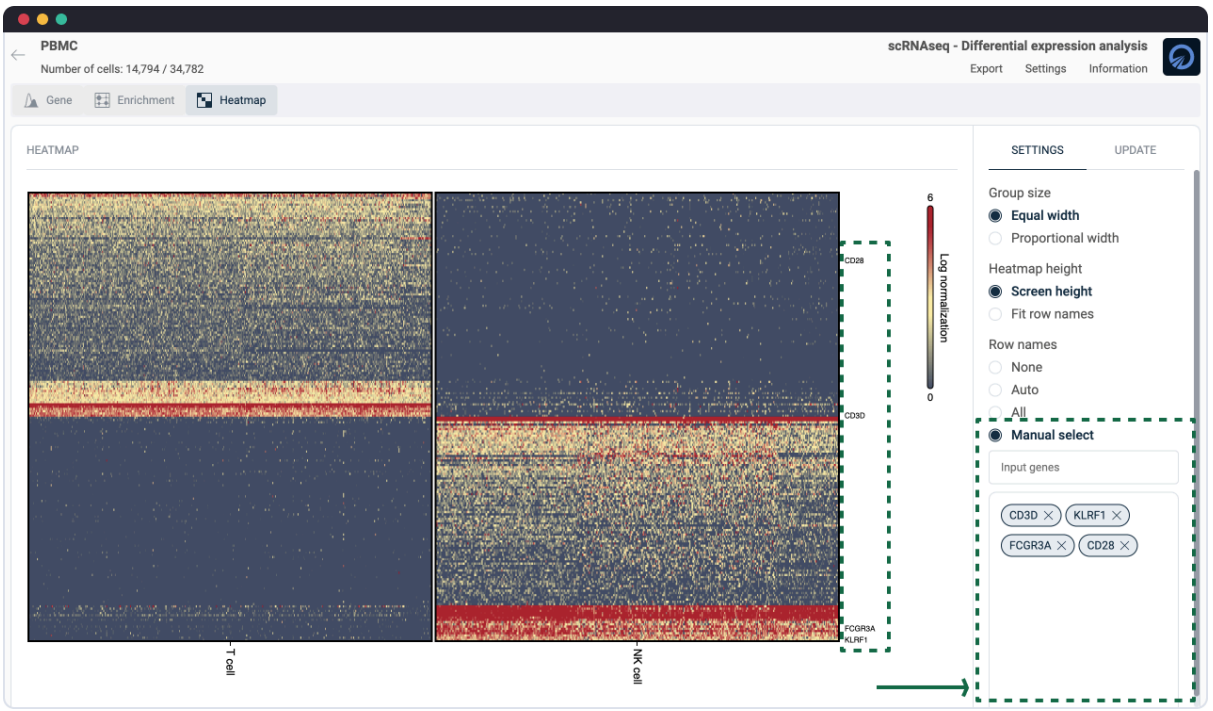


Here, you can adjust various visual aspects of the heatmap such as: width, height, or feature name display options, which includes:

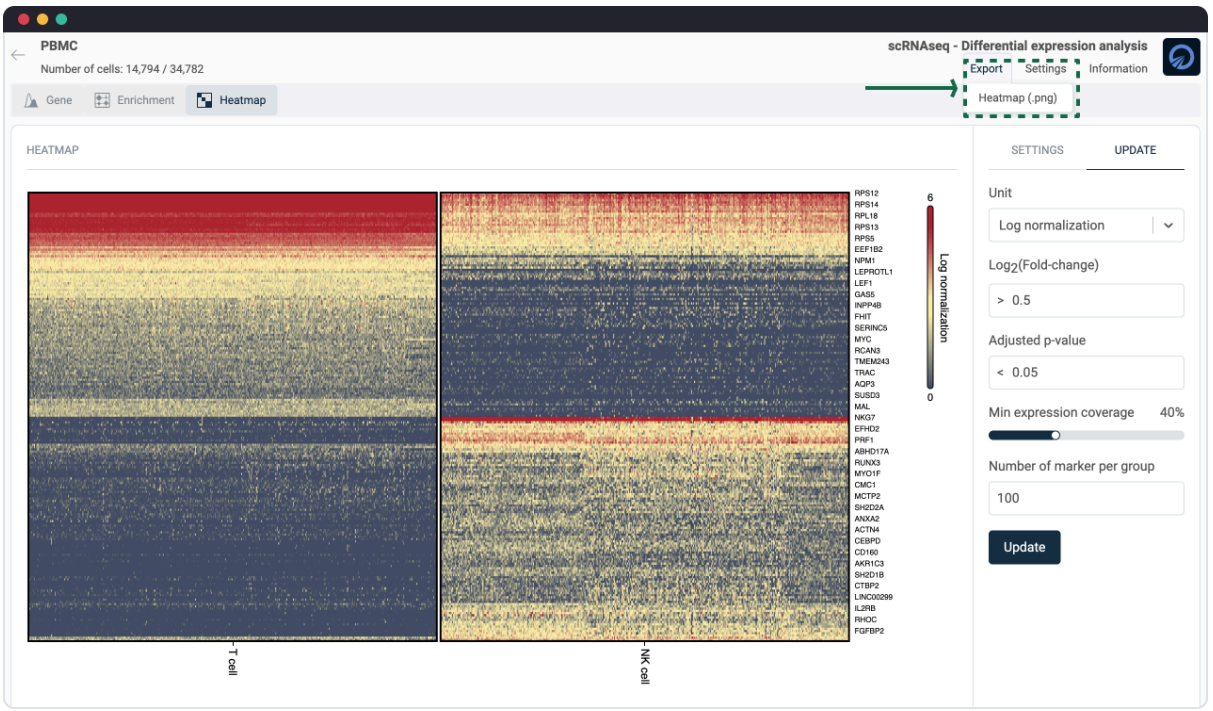
- `None`: No feature names are displayed.
- `Auto`: Automatically displays non-overlapping feature names, adjusting to fit within the heatmap's height.



- `All`: All feature names are shown. This may require adjusting the heatmap's height to prevent overlapping labels.
- `Manual select`: Manually input a list of features to highlight on the heatmap.



Heatmap can be exported as a png file by clicking on the `Export` button on the top menu.



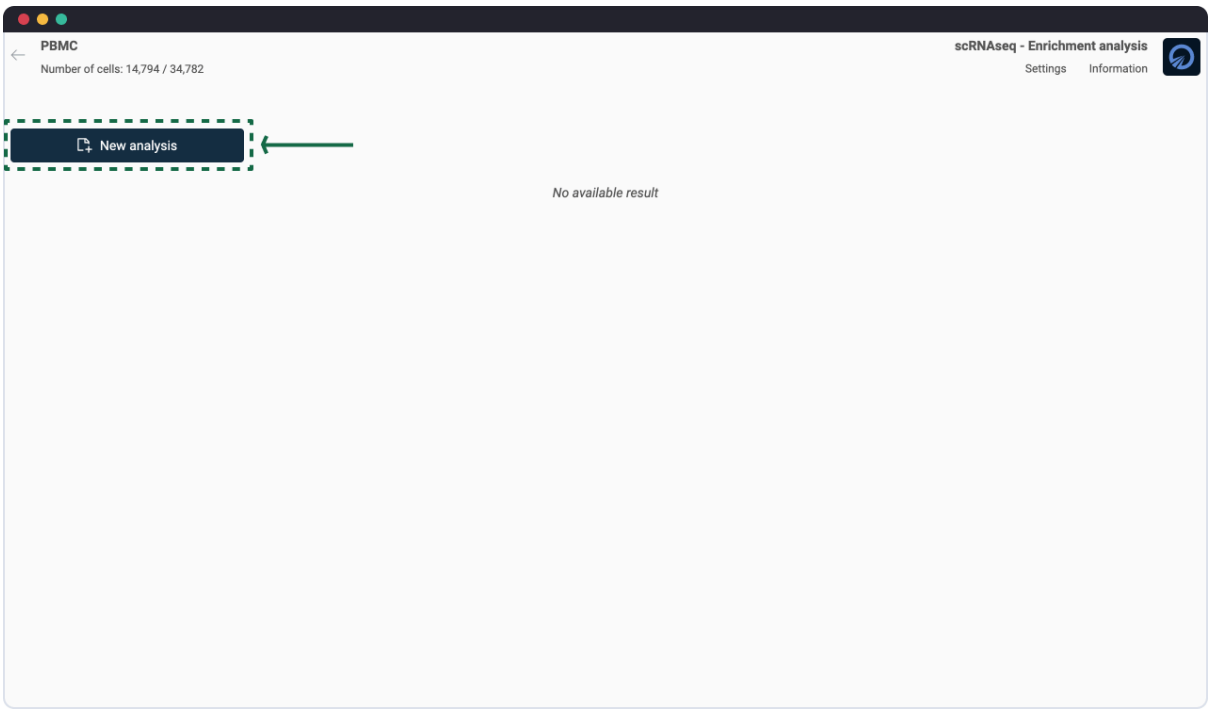
4. Enrichment analysis

Enrichment analysis in OmnibusX is designed to statistically evaluate how sets of genes are enriched in different cell populations under various conditions. This process involves several computational steps that combine bioinformatics algorithms with statistical tests to provide insights into the biological significance of the gene sets. Here's a step-by-step breakdown of how enrichment analysis is calculated:

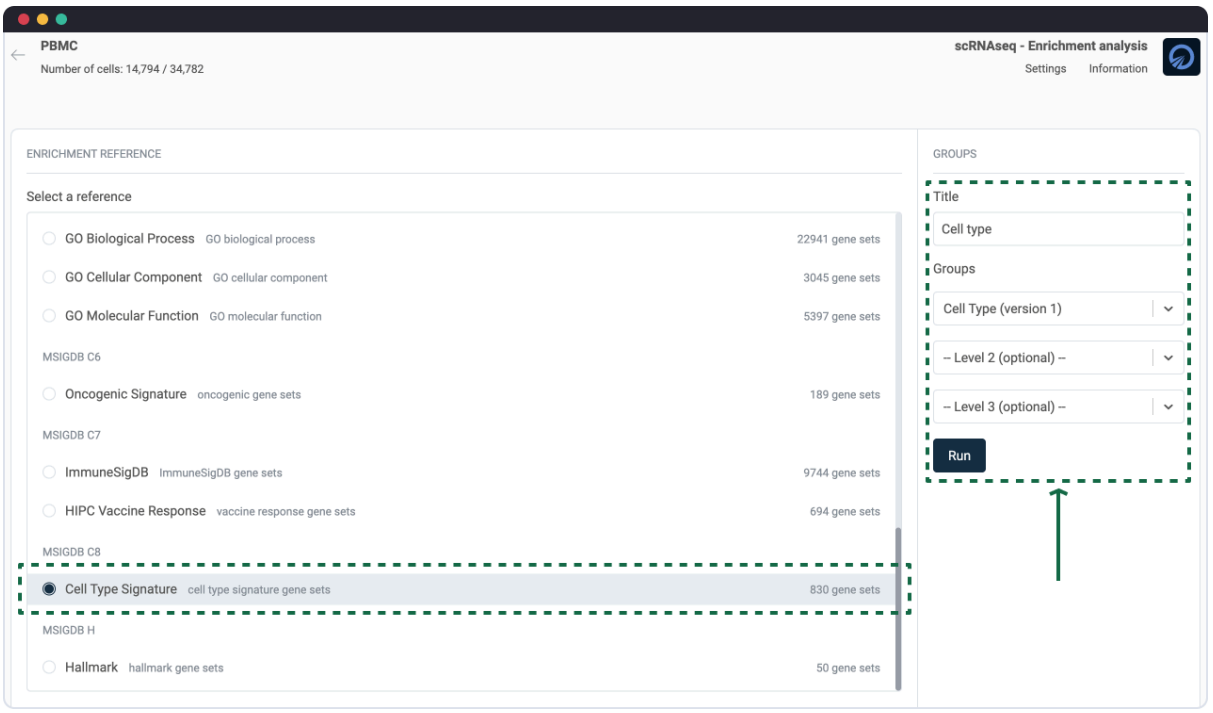
1. `Select gene set collection`: OmnibusX integrates gene sets from two primary sources:OmnibusX integrates gene sets from two primary sources: MsigDB and Gene ontology
2. `Calculate enrichment scores`: For each cell, an enrichment score for each gene set in the selected collection is calculated using the AUCell algorithm (Bioconductor package documentation).
3. `ANOVA testing`: Cells are grouped based on metadata annotations, which can include factors like treatment, cell type, or any other experimental or biological condition. Once cells are grouped, an Analysis of Variance (ANOVA) is performed to compare the enrichment scores across these groups. ANOVA helps determine whether there are statistically significant differences in the mean enrichment scores between two or more groups.
4. `p-value adjustment`: Since multiple gene sets are tested simultaneously, the `p-values` obtained from the ANOVA tests are adjusted using the Benjamini-Hochberg procedure.

4.1. Create new enrichment analysis

To create an enrichment analysis, click on the `New analysis` button.

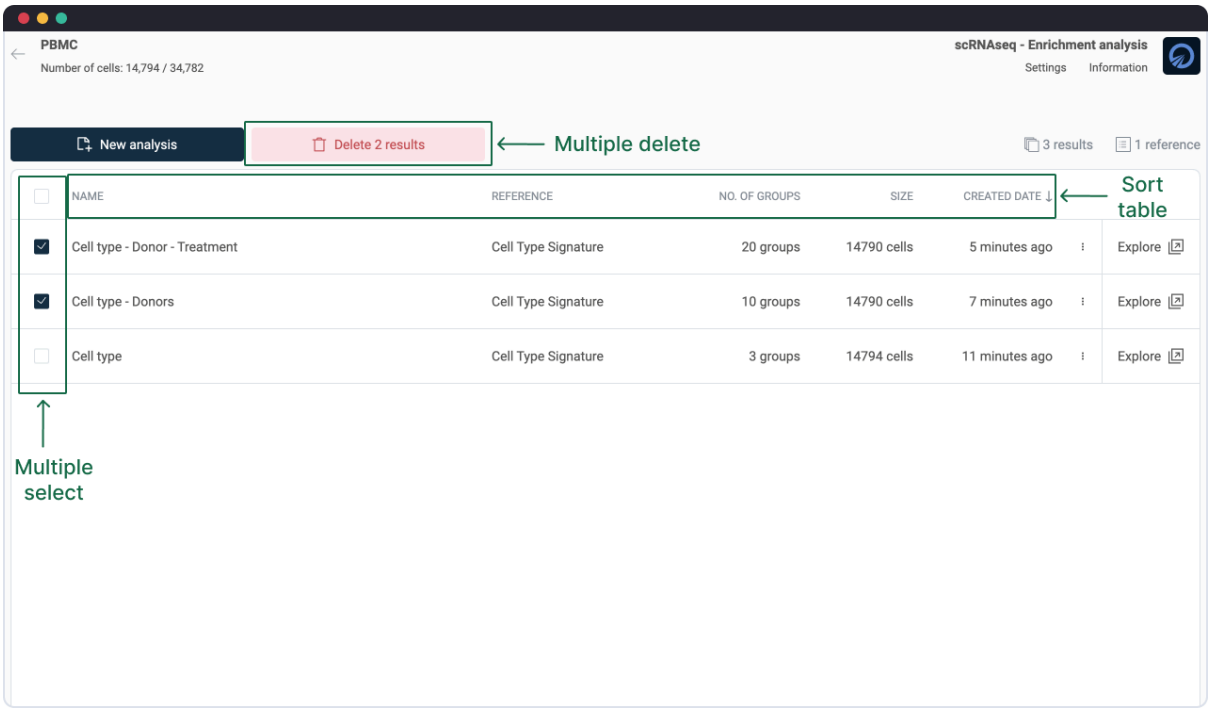


Select a collection from the `Enrichment reference` table. Then, choose up to three metadata levels for grouping cells in your analysis. Enter a name for your analysis and click the Run button to initiate the process.

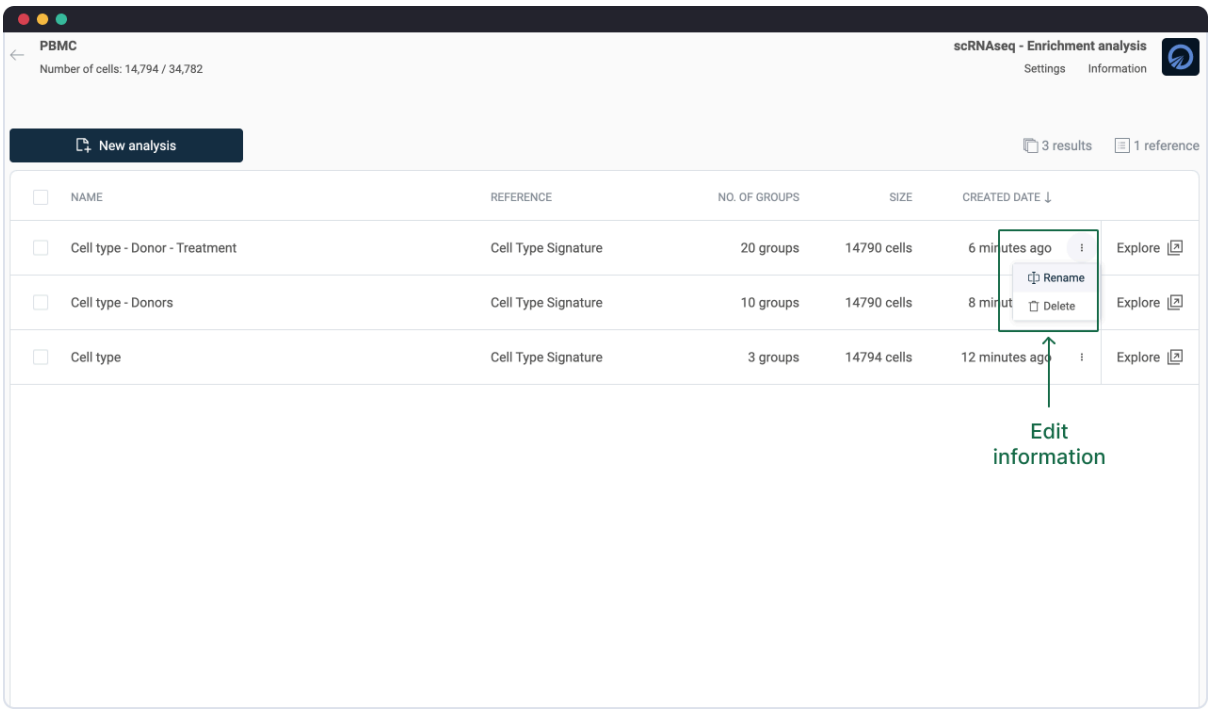


4.2. Manage enrichment results

All results are listed in an interactive table on the enrichment analysis page. You can sort these results by clicking on the table headers. To delete multiple results, select them using the checkboxes and hit the `Delete` button to perform a bulk deletion.



Update the name of an analysis by hovering over the three-dot icon on the right of a result and selecting `Rename`.



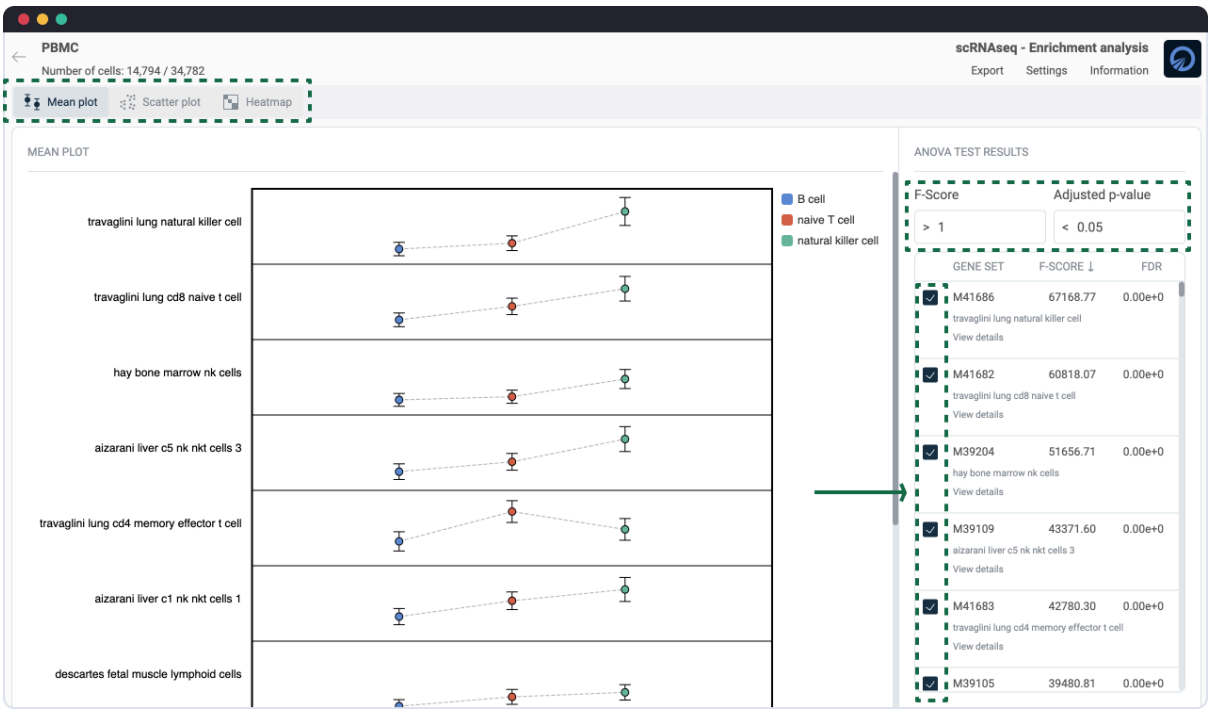
Clicking the `Explore` button to start analyzing a specific result.

4.3. Analyze enrichment result

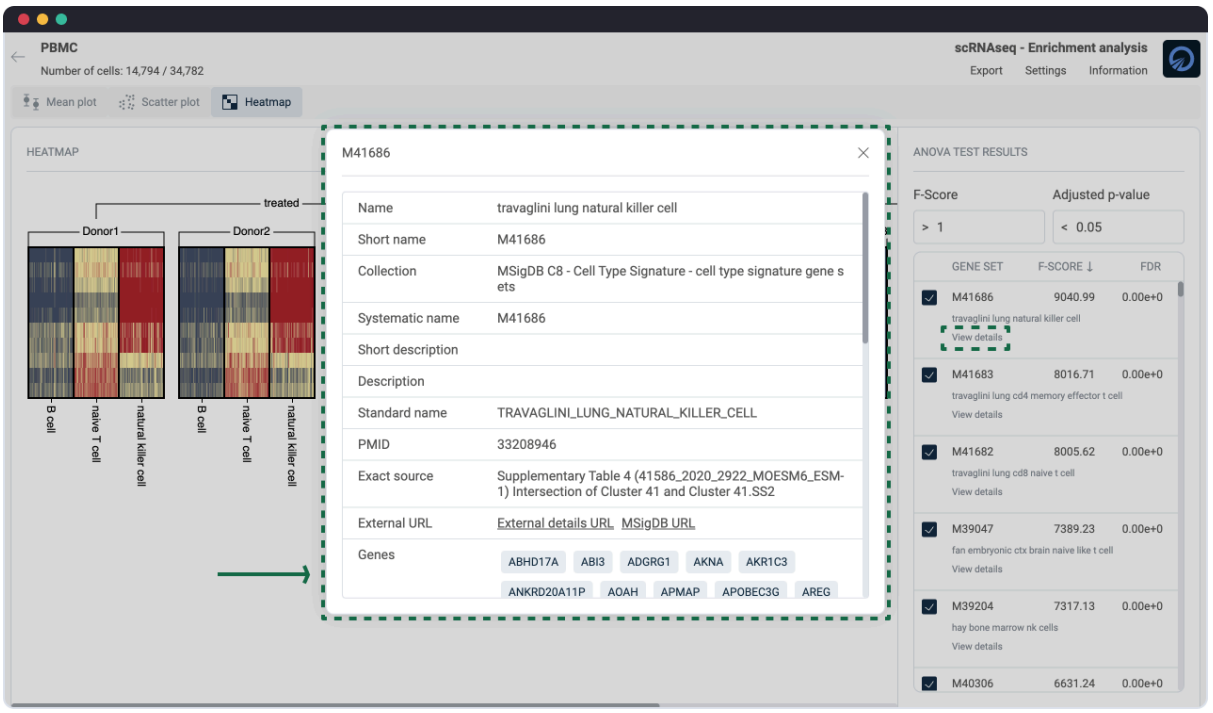
OmnibusX offers three types of visualizations for analyzing enrichment results:

- `Mean plot`: Visualizes the average enrichment score across groups.
- `Scatter plot`: Visualizes the pattern of differential enrichment of each gene set across groups.
- `Heatmap`: Visualizes the enrichment scores pattern of each gene set across groups.

Enrichment test results, including `F-scores` and `adjusted p-values`, are displayed in a table on the right panel. You can click on the table header to sort or filter results using the inputs above the table. Check the boxes next to gene sets in the table to select them for visualization in your chosen format.



Click on `view detail` in each row to access detailed descriptions about each gene set.

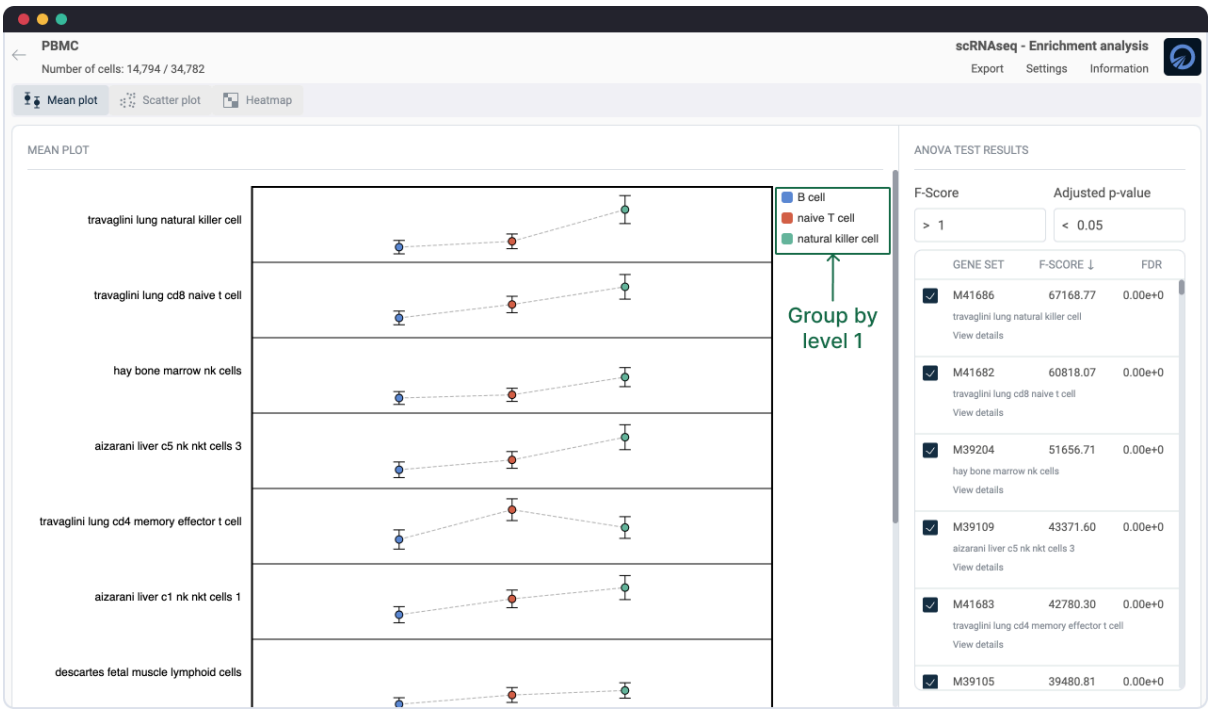


Results can be exported as a .tsv file using the export button located at the top menu. All plots can be exported as PNG or SVG files.

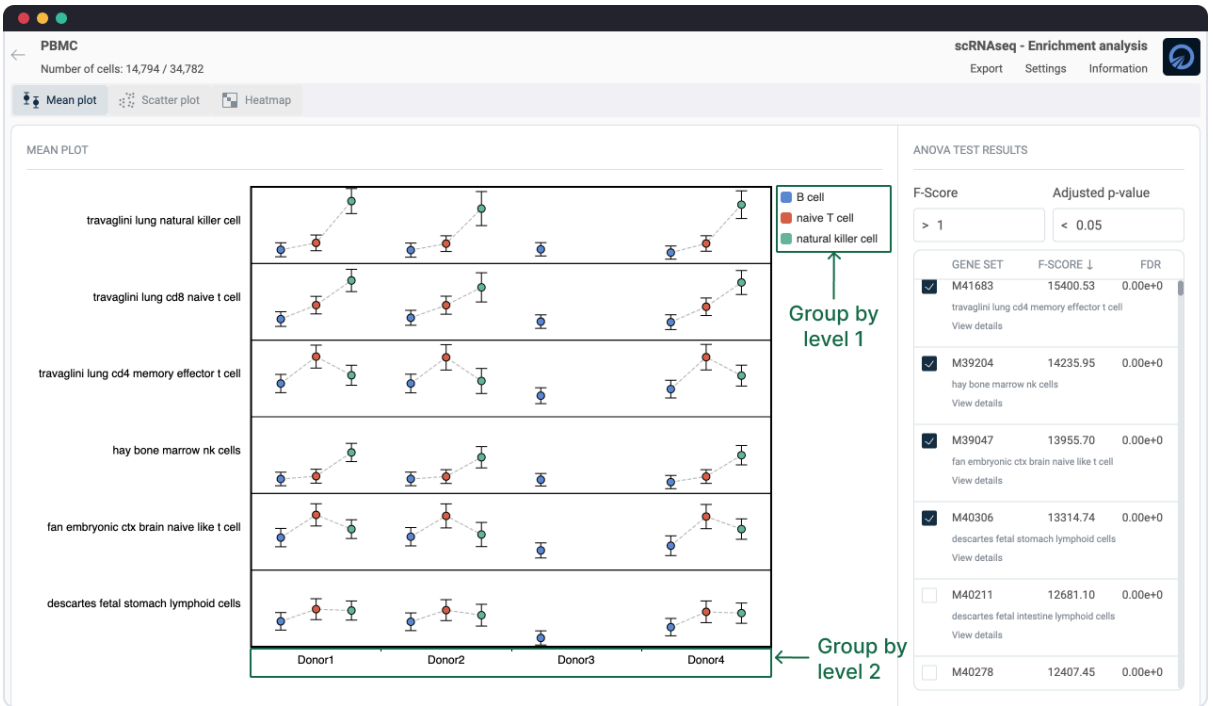
4.4. Meant plot

The mean plot in OmnibusX visualizes the means of enrichment scores of a gene set across different groups defined by metadata levels. Each selected gene set is represented as a separate row within the plot. Each dot in a row represents the mean enrichment score for a gene set within a specific group. Attached to each dot, error bars show the variance of the enrichment scores within that group.

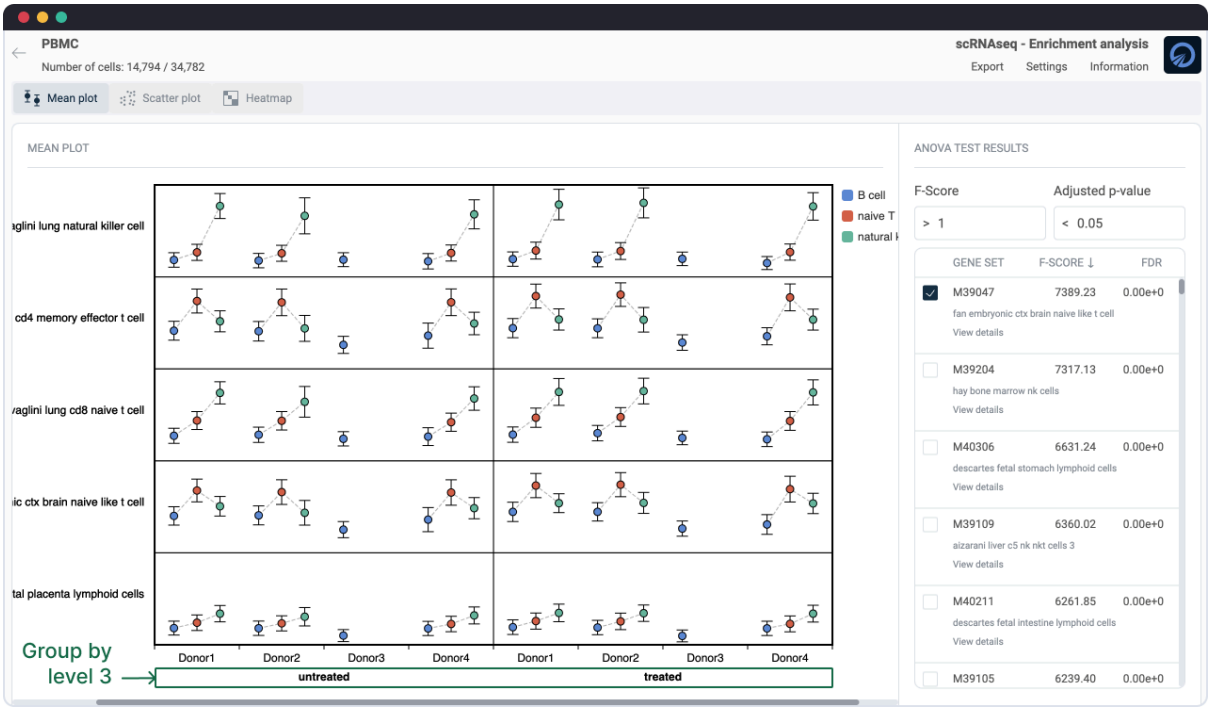
When inputting one metadata, cells are divided into distinct groups using annotation from that metadata. Each group is assigned a unique color for easy identification.



If two metadata are selected, cells are divided into groups using the combined information from both metadata. Groups sharing the same annotation from the first metadata level are colored identically. Those sharing the same annotation from the second metadata level are aligned along the same X-axis value.



If three metadata are selected, cells are first segmented based on the third metadata level. Each of these segments is presented as a distinct subplot, which is drawn using the combined information from the first and second metadata.



4.5. Scatter

OmnibusX also provides a scatter plot to visualize the enrichment results. This plot helps identify clusters or patterns in gene set expression across different groups:

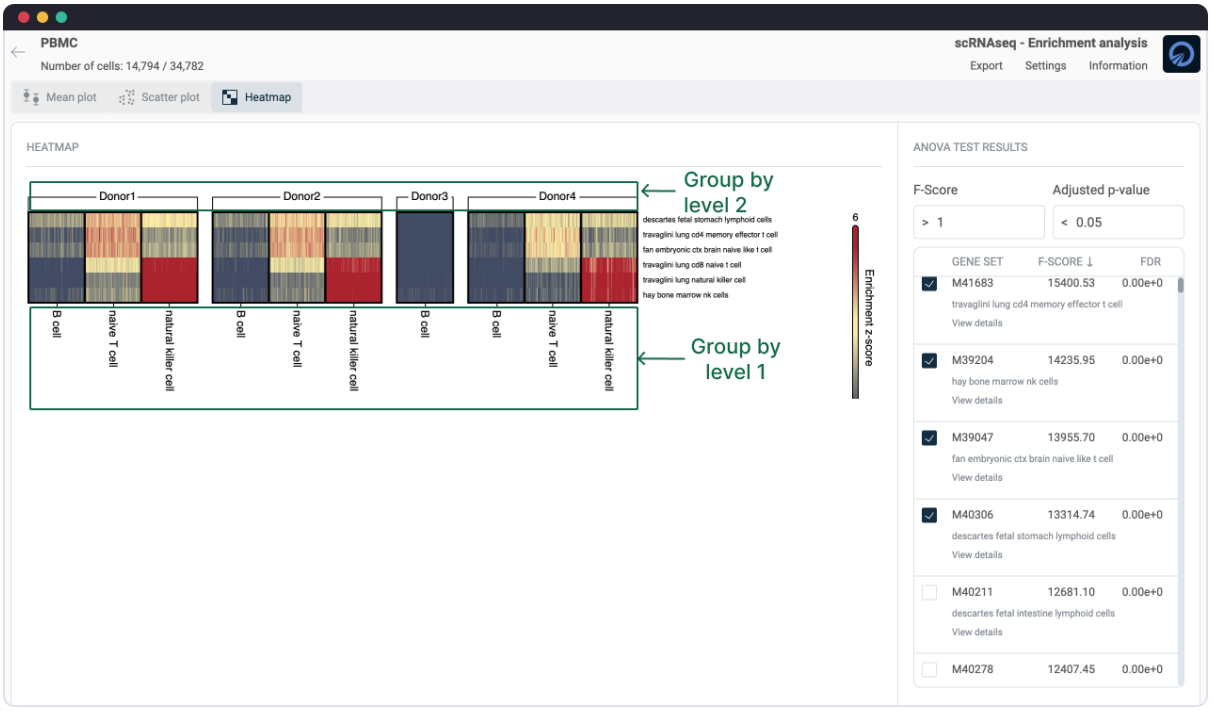
- Using the mean enrichment score of each gene set of each group, OmnibusX performs `PCA` and then `t-SNE` to visualize all gene sets as a 2D scatter plot. This can reveal relevant groups of gene sets based on the enrichment pattern across groups.
- Gene sets selected from the results table are highlighted in red, making them stand out against others, which are shown in gray. This distinction helps focus on specific gene sets of interest during analysis.
- The point size presents the `-Log10(Adjusted p-value)`. This means that larger points indicate more statistical significance associated with each gene set's differential enrichment.

4.6. Heatmap

The heatmap in OmnibusX is another way to observe differential enrichment patterns of gene sets across various groups. It is particularly useful for identifying consistent trends and anomalies in gene set enrichment across different cellular conditions or treatments.

The values displayed in the heatmap are the enrichment scores of selected gene sets for all cells. These scores are grouped according to the combined information from input metadata. To ensure consistency and comparability across different gene sets within the heatmap, OmnibusX applies `row-wise normalization`. Each row in the heatmap, which corresponds to a specific gene set, is normalized individually into z-scores by subtracting the mean of the row (`row_mean`) from each score in that row and then dividing the result by the standard deviation of the row (`row_std`): `(score - row_mean) / row_std`.

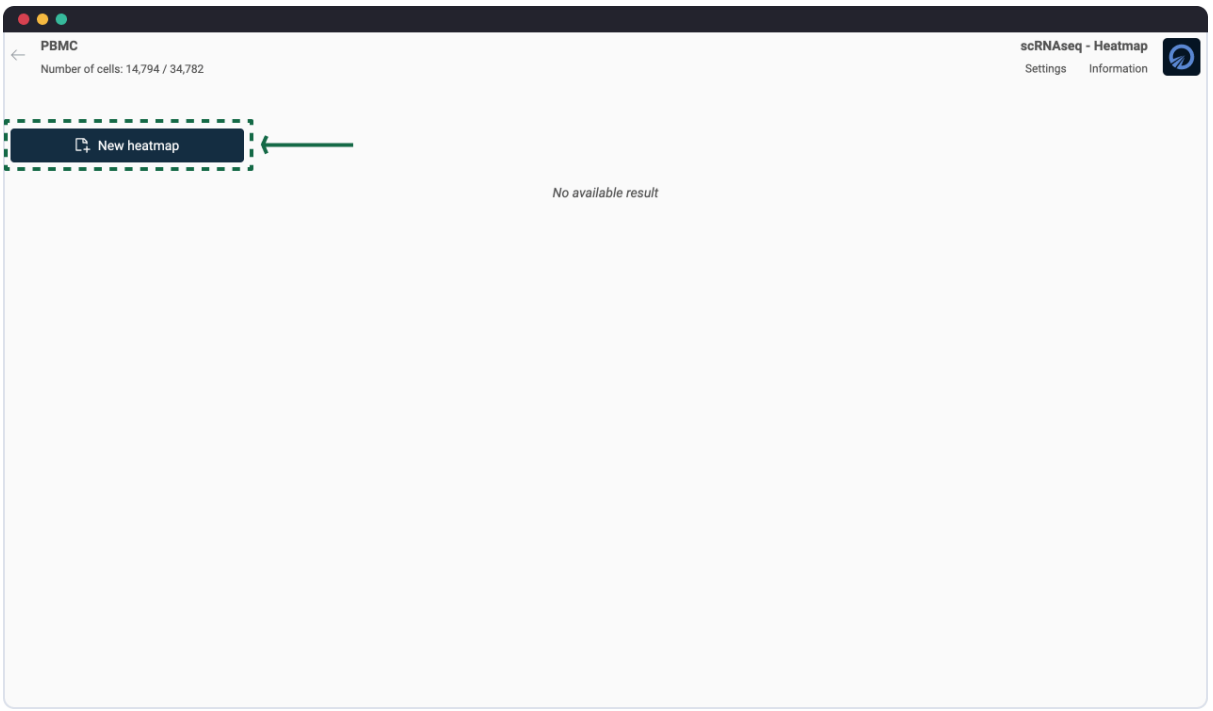

PBMC
Number of cells: 14,794 / 34,782
scRNAseq - Enrichment analysis
Export Settings Information
Mean plot Scatter plot Heatmap
HEATMAP
descartes fetal stomach myeloid cells
he lim sun fetal lung c4 cd16 pos nk cell
travaglini lung natural killer t cell
travaglini lung natural killer cell
hay bone marrow nk cells
aizarani liver c12 nk nkt cells 4
durante adult olfactory neuroepithelium nk cells
aizarani liver c5 nk nkt cells 3
travaglini lung cd8 naive t cell
fan embryonic ctx brain effector t cell
travaglini lung cd8 memory effector t cell
durante adult olfactory neuroepithelium cd4 t cells
travaglini lung cd4 memory effector t cell
fan embryonic ctx brain naive like t cell
hay bone marrow naive t cell
aizarani liver c3 nk nkt cells 2
aizarani liver c1 nk nkt cells 1
cui developing heart c9 b t cell
durante adult olfactory neuroepithelium cd8 t cells
descartes fetal muscle lymphoid cells
descartes fetal stomach lymphoid cells
descartes fetal intestine lymphoid cells
descartes fetal placenta lymphoid cells
7
-4
Enrichment z-score
B cell
naive T cell
natural killer cell
Group by level 1
ANOVA TEST RESULTS
F-Score
Adjusted p-value
> 1
< 0.05
GENE SET F-SCORE ↓ FDR
M40278 18341.71 0.00e+0
descartes fetal placenta lymphoid cells
View details
M39202 17746.32 0.00e+0
hay bone marrow naive t cell
View details
M41685 16513.62 0.00e+0
travaglini lung natural killer t cell
View details
M40302 13097.68 0.00e+0
descartes fetal stomach myeloid cells
View details
M39107 12970.78 0.00e+0
aizarani liver c3 nk nkt cells 2
View details
M40144 12425.97 0.00e+0


PBMC
Number of cells: 14,794 / 34,782
scRNAseq - Enrichment analysis
Export Settings Information
Mean plot Scatter plot Heatmap
HEATMAP
Donor1
Donor2
Donor3
Donor4
Group by level 2
descartes fetal stomach lymphoid cells
travaglini lung cd4 memory effector t cell
fan embryonic ctx brain naive like t cell
travaglini lung cd8 naive t cell
travaglini lung natural killer cell
hay bone marrow nk cells
B cell
naive T cell
natural killer cell
Group by level 1
6
Enrichment z-score
ANOVA TEST RESULTS
F-Score
Adjusted p-value
> 1
< 0.05
GENE SET F-SCORE ↓ FDR
M41683 15400.53 0.00e+0
travaglini lung cd4 memory effector t cell
View details
M39204 14235.95 0.00e+0
hay bone marrow nk cells
View details
M39047 13955.70 0.00e+0
fan embryonic ctx brain naive like t cell
View details
M40306 13314.74 0.00e+0
descartes fetal stomach lymphoid cells
View details
M40211 12681.10 0.00e+0
descartes fetal intestine lymphoid cells
View details
M40278 12407.45 0.00e+0

You can select which gene sets to display on the heatmap from the table located on the right panel. To optimize disk storage, OmnibusX calculates the heatmap in real-time rather than pre-calculating and storing enrichment scores for all gene sets across all cells. Therefore, it might take time for large datasets. It is recommended to pre-select gene sets of interest using other visualization tools provided by OmnibusX, such as mean plots or scatter plots. Once you have identified key gene sets through preliminary visualizations, switch to the heatmap view to generate a new heatmap that visually represents the enrichment scores of these selected gene sets.

5. Heatmap analysis

The heatmap is used to illustrate the expression patterns of genes/features across groups of cells. To create a new heatmap, click on the `New heatmap` button.

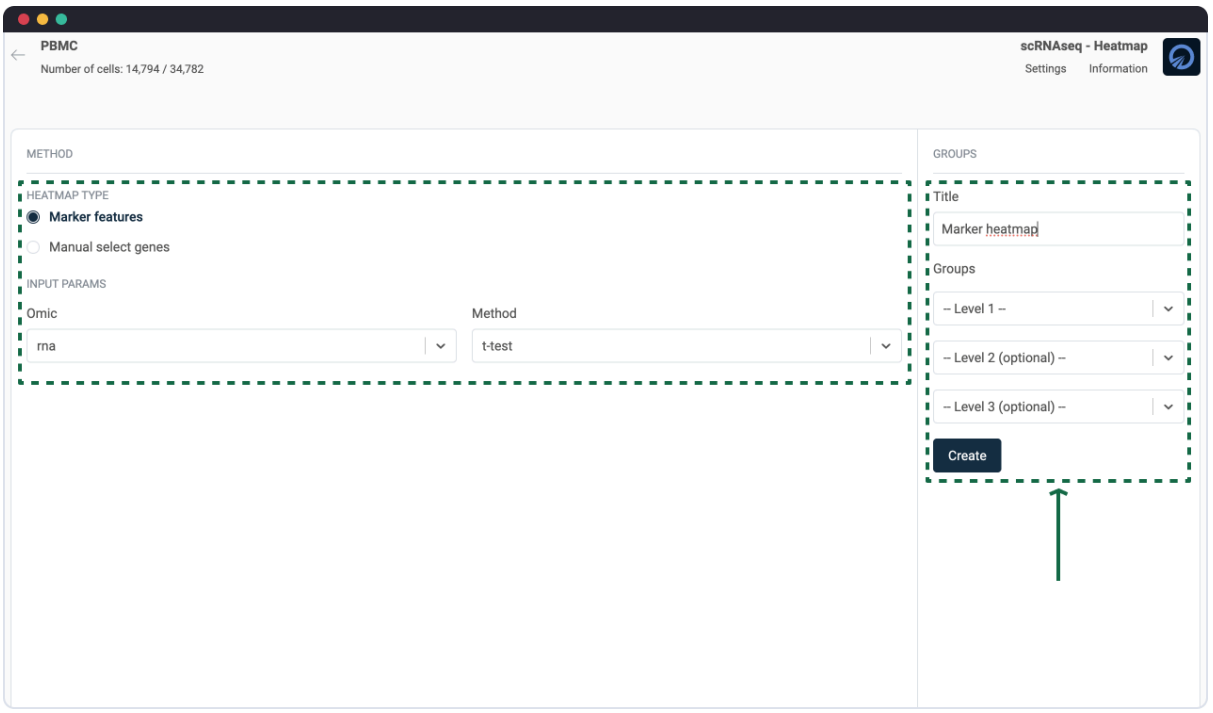


5.1. Create new heatmap analysis

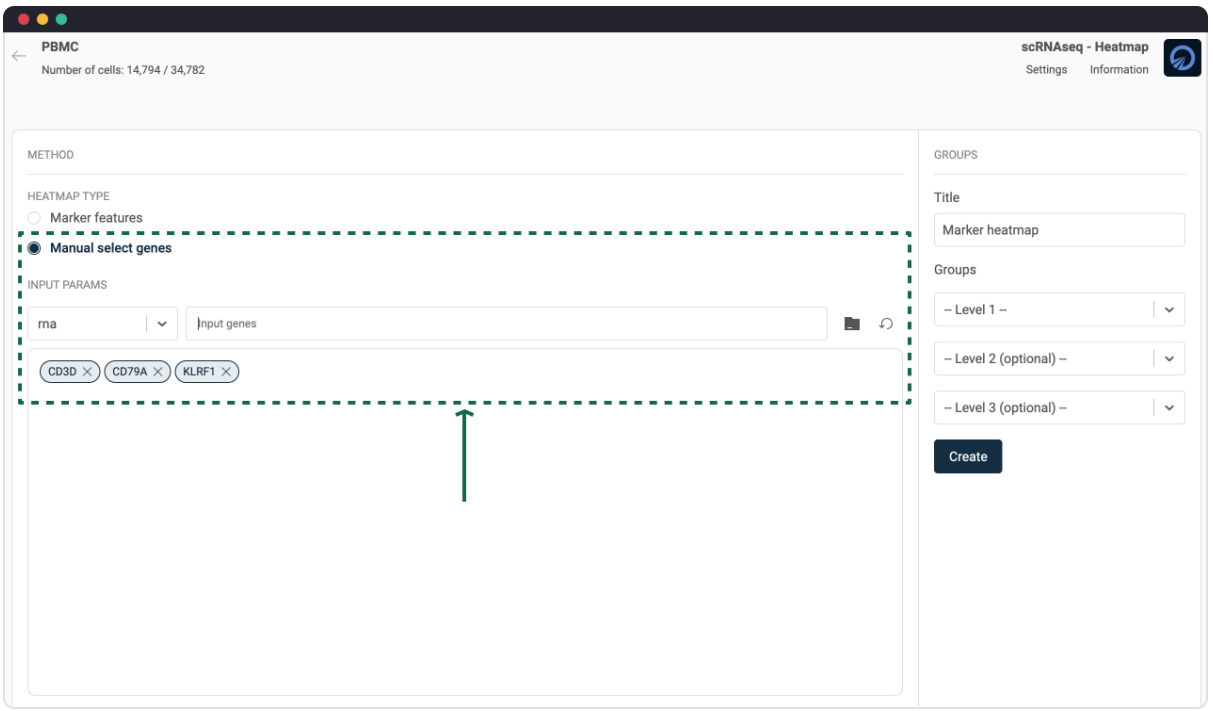
Select up to three metadata levels to group cells. Grouping cells helps in comparing expression patterns across defined categories or conditions.

Choose one of the two ways to select a list of features:

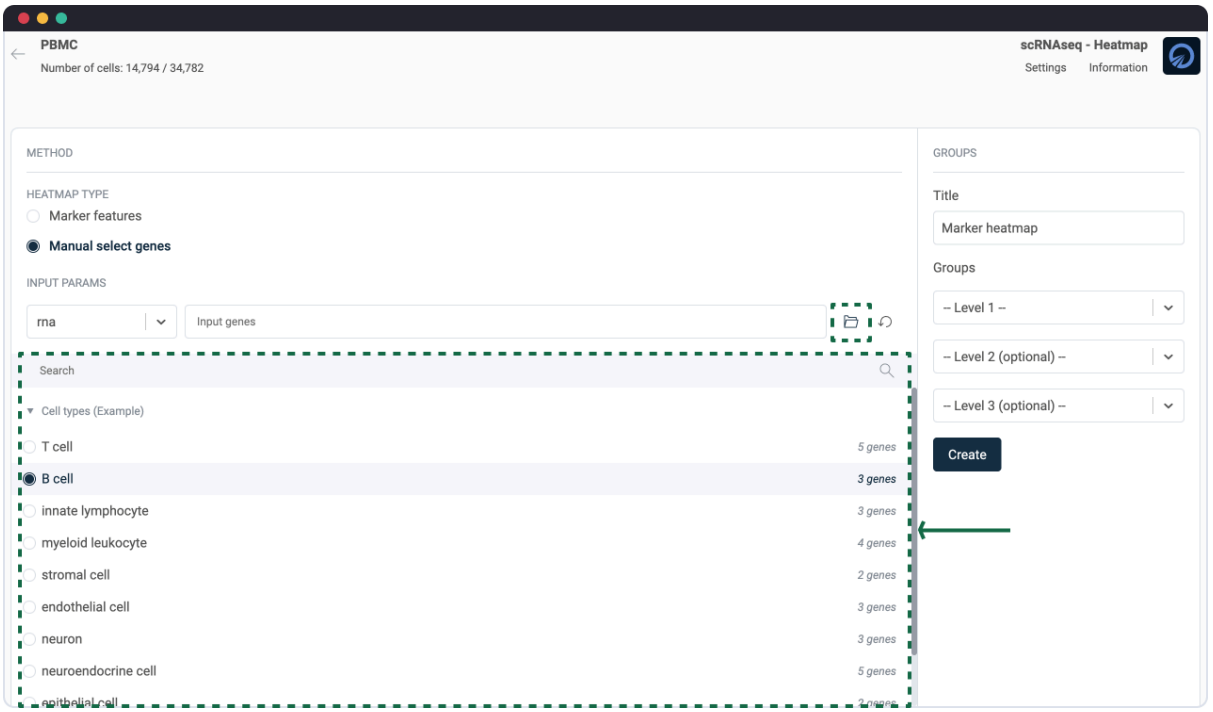
1. `Marker features`: OmnibusX automatically identifies marker features for each input group compared to the remaining cells. You can choose an omic from available omics in the dataset and the statistical method (`t-test`, `Wilcoxon rank sum test`) for calculating these marker features.



2. `Manual select features`: Directly input the names of target genes/features in the provided input box. As you type, OmnibusX will suggest available features based on the omic you have selected.



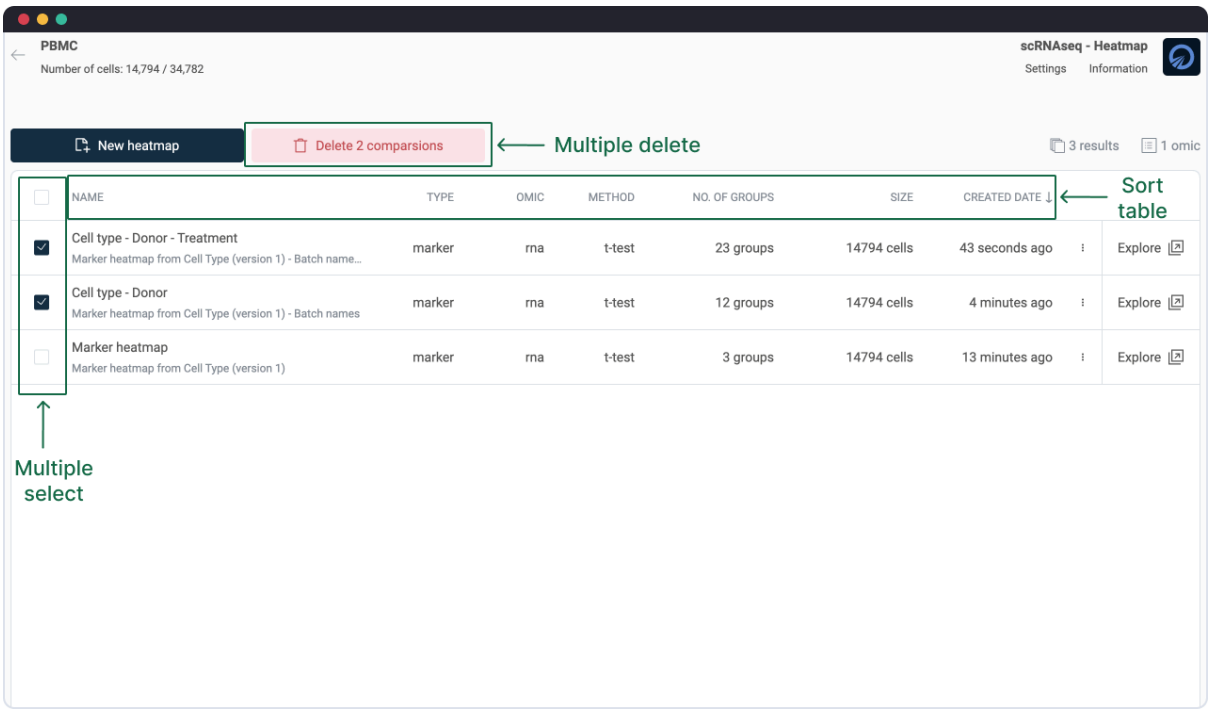
Alternatively, you can quickly select genes related to a gene set in your library by clicking on the folder icon and selecting a gene set from the dropdown list.



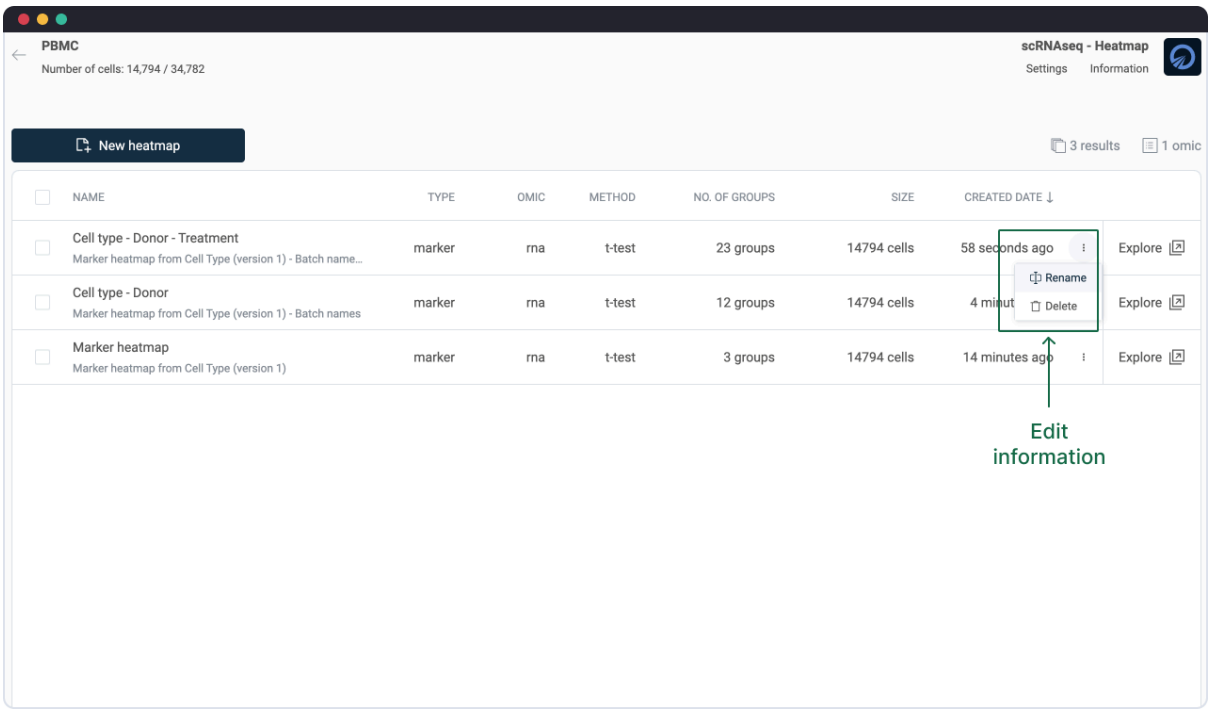
Click the `Create` button to generate the heatmap.

5.2. Manage heatmap results

All results are listed in an interactive table on the heatmap analysis page. You can sort these results by clicking on the table headers. To delete multiple results, select them using the checkboxes and hit the `Delete` button to perform a bulk deletion.



Update the name of an analysis by hovering over the three-dot icon on the right of a result and selecting `Rename`.

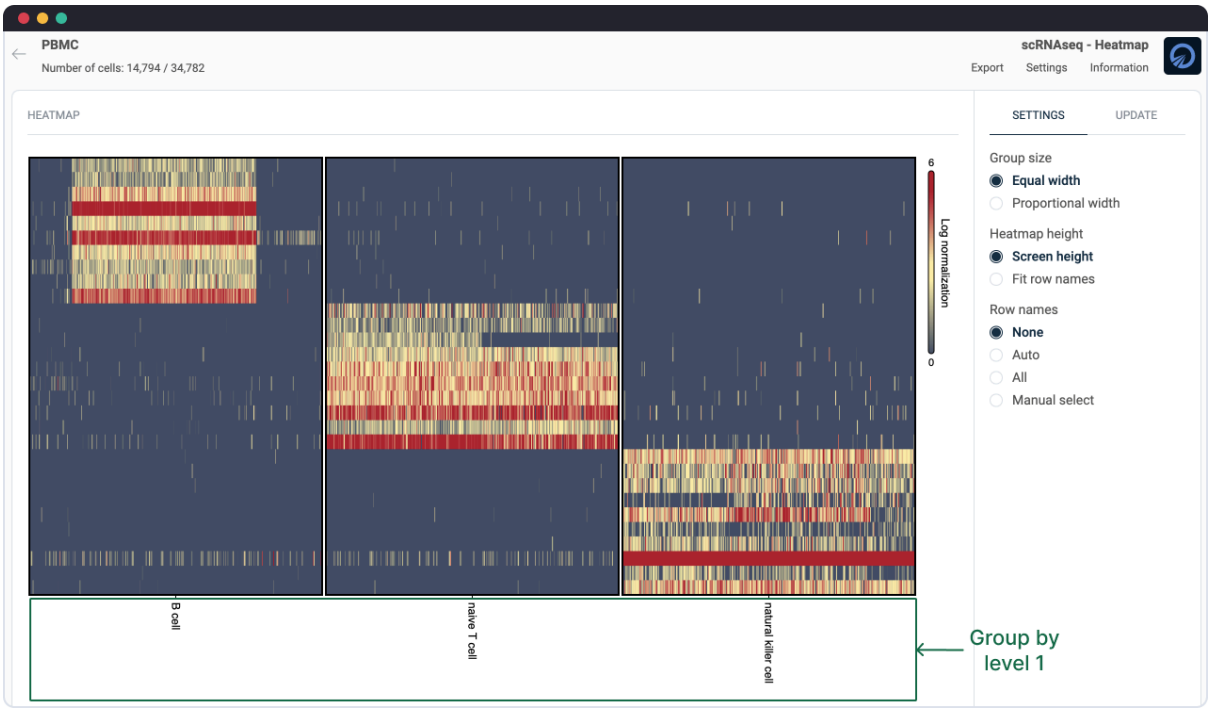


Clicking the `Explore` button to start analyzing a specific result.

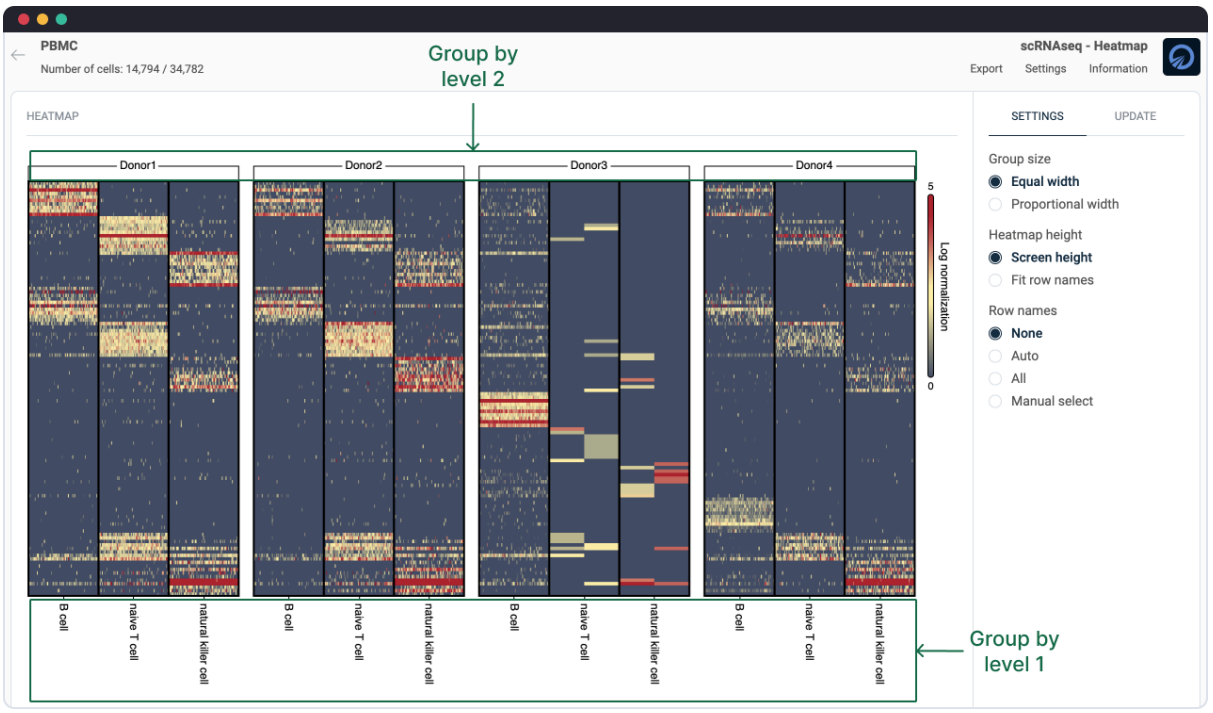
5.3. Analyze heatmap result

Depending on the level of metadata used for grouping, OmnibusX generates different types of heatmaps:

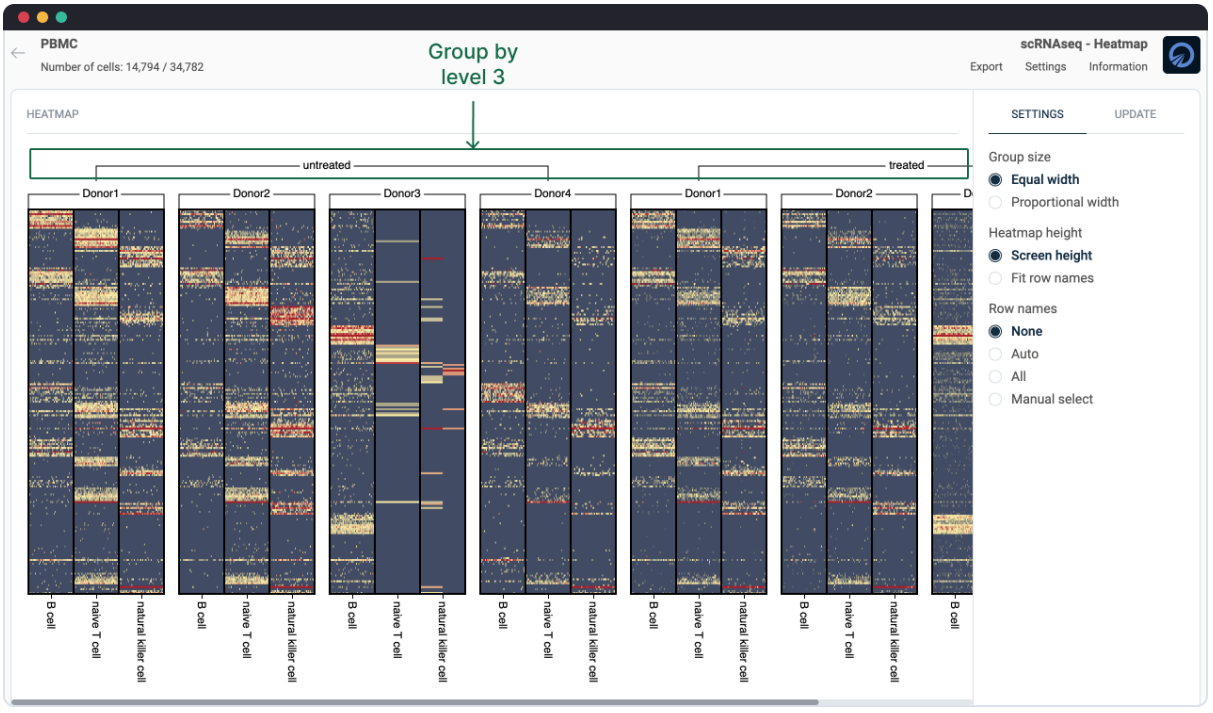
1. One metadata level



2. Two metadata levels:



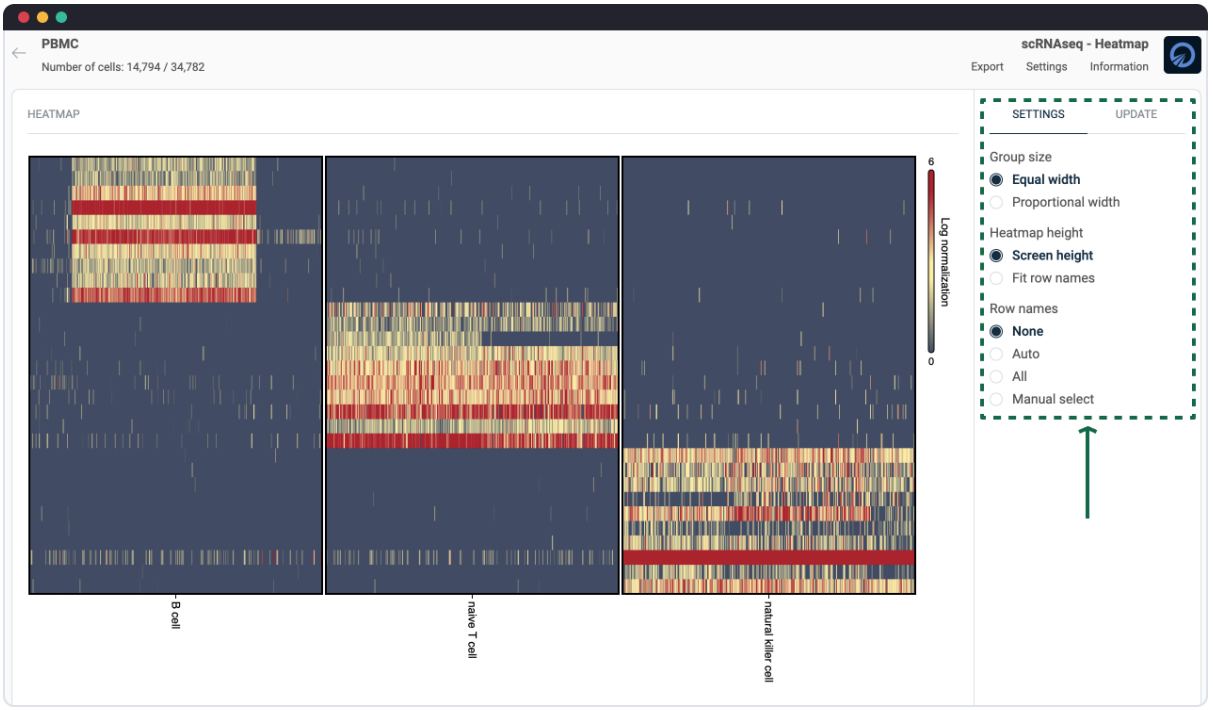
3. Three metadata levels:



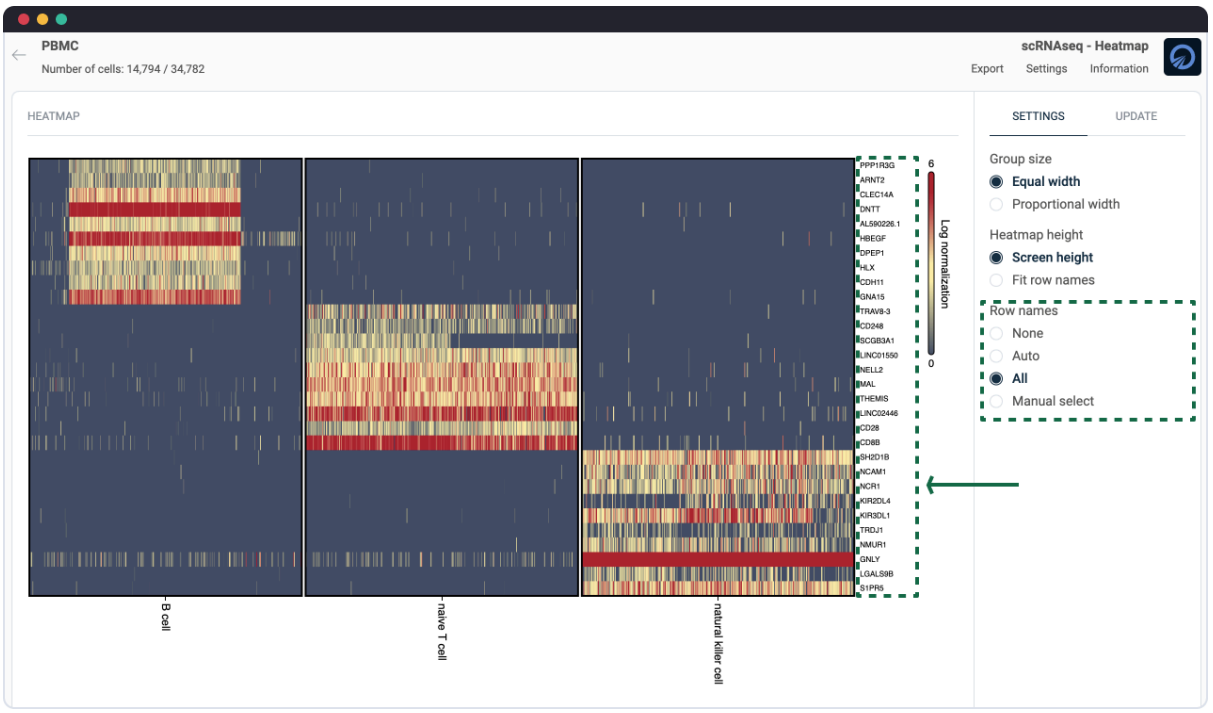
Once you have created a heatmap in OmnibusX, you have multiple options to tailor and manage the visualization. From the Setting panel on the right, you can

adjust the heatmap's width and height. You can change the feature name display as below options:

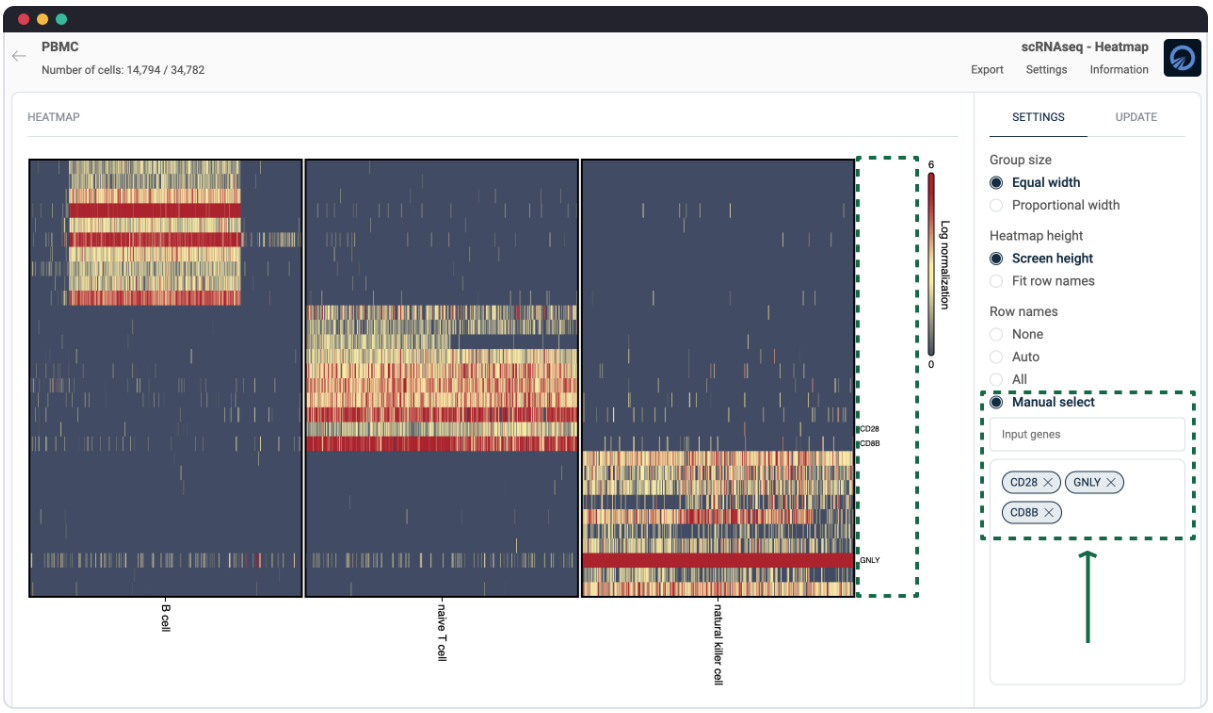
- None: No feature names are displayed.

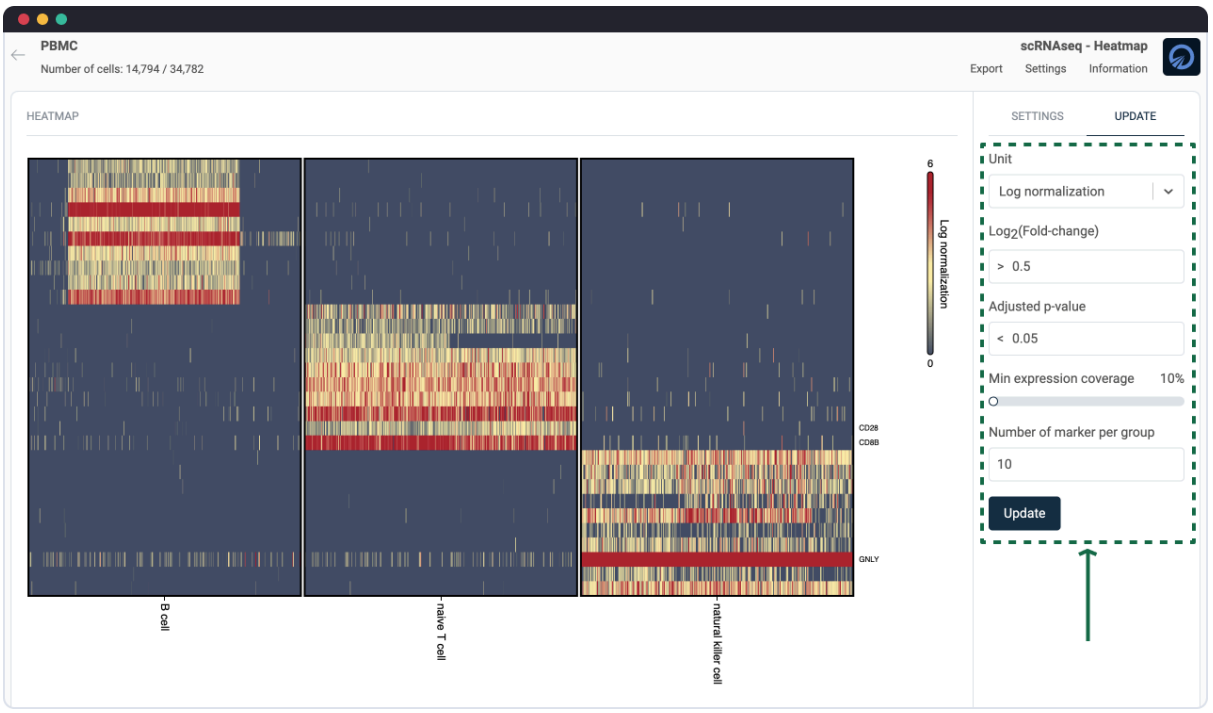


- Auto: Automatically displays non-overlapping feature names, adjusting to fit within the heatmap's height.
- All: All feature names are shown. This may require adjusting the heatmap's height to prevent overlapping labels.

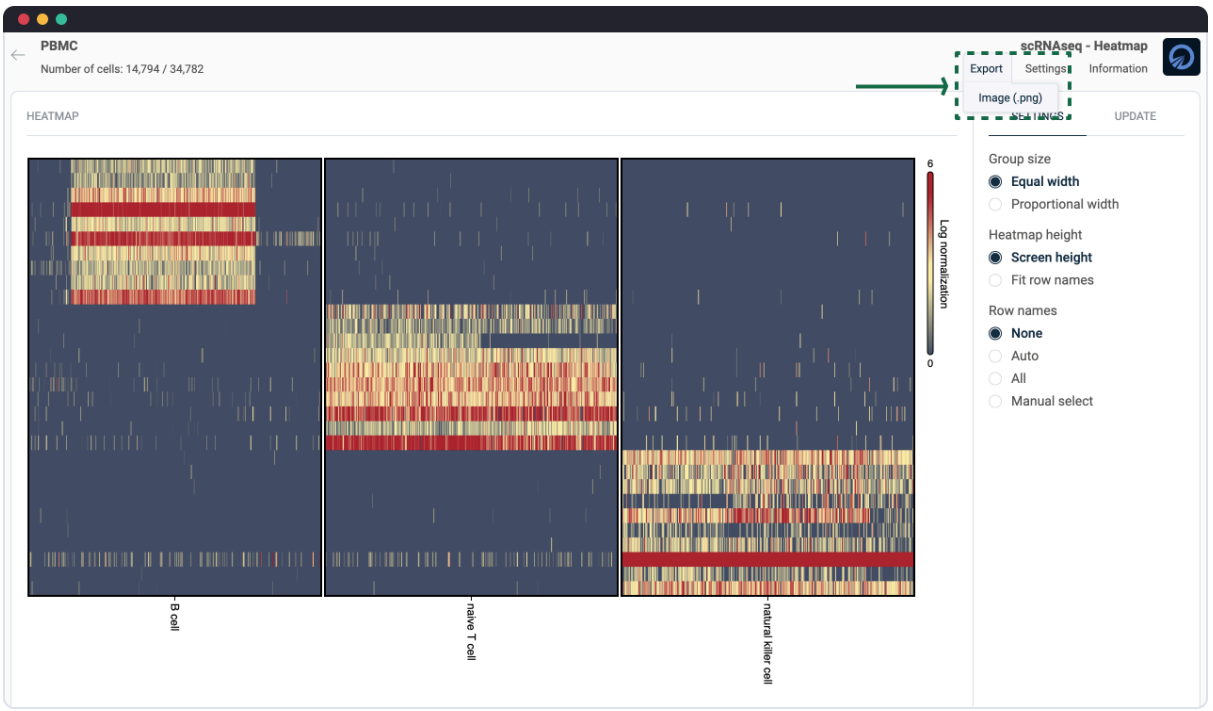


- `Manual select`: Manually input a list of features to highlight on the heatmap.



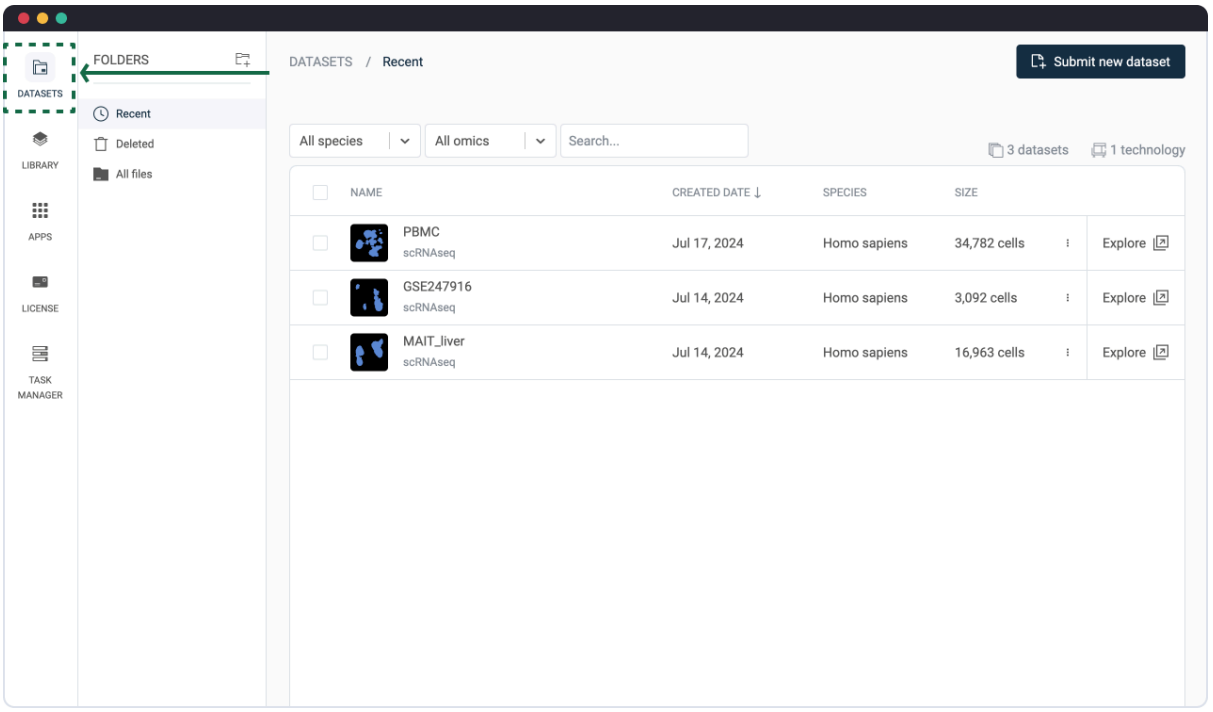


Heatmap can be exported as a png file by clicking on the `Export` button on the top menu.

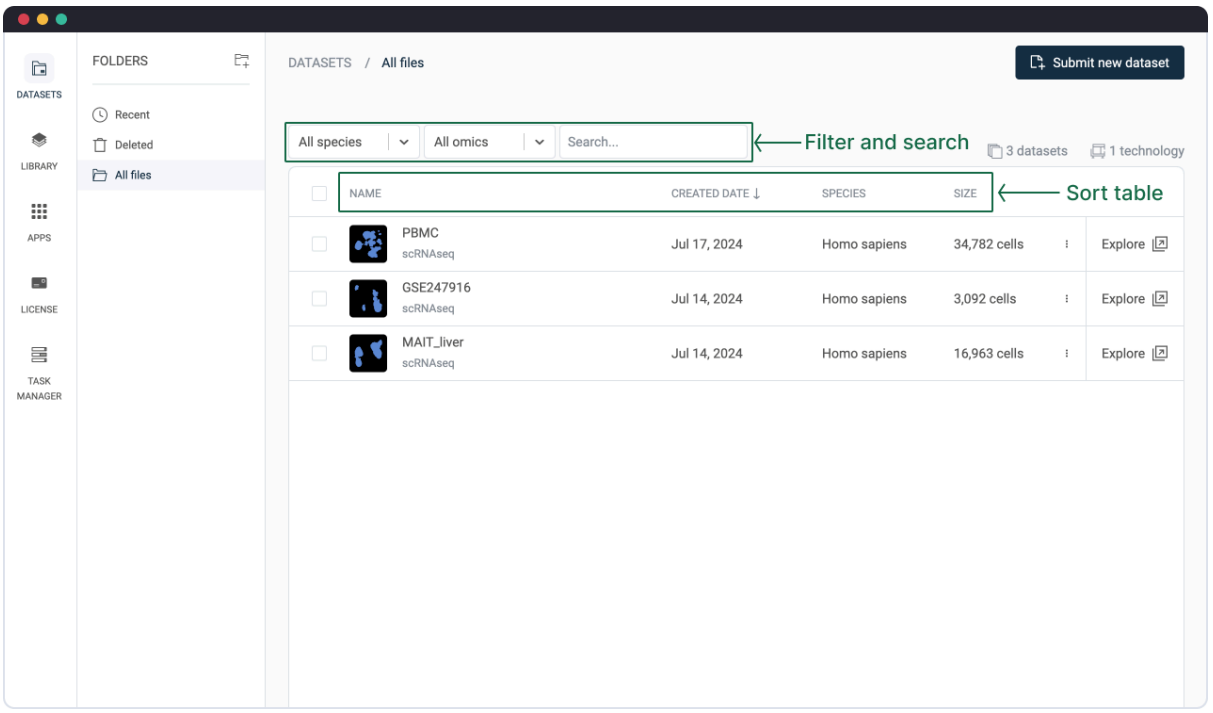


Dataset management

To view all your submitted datasets, click on the Dataset button located on the left navigator.

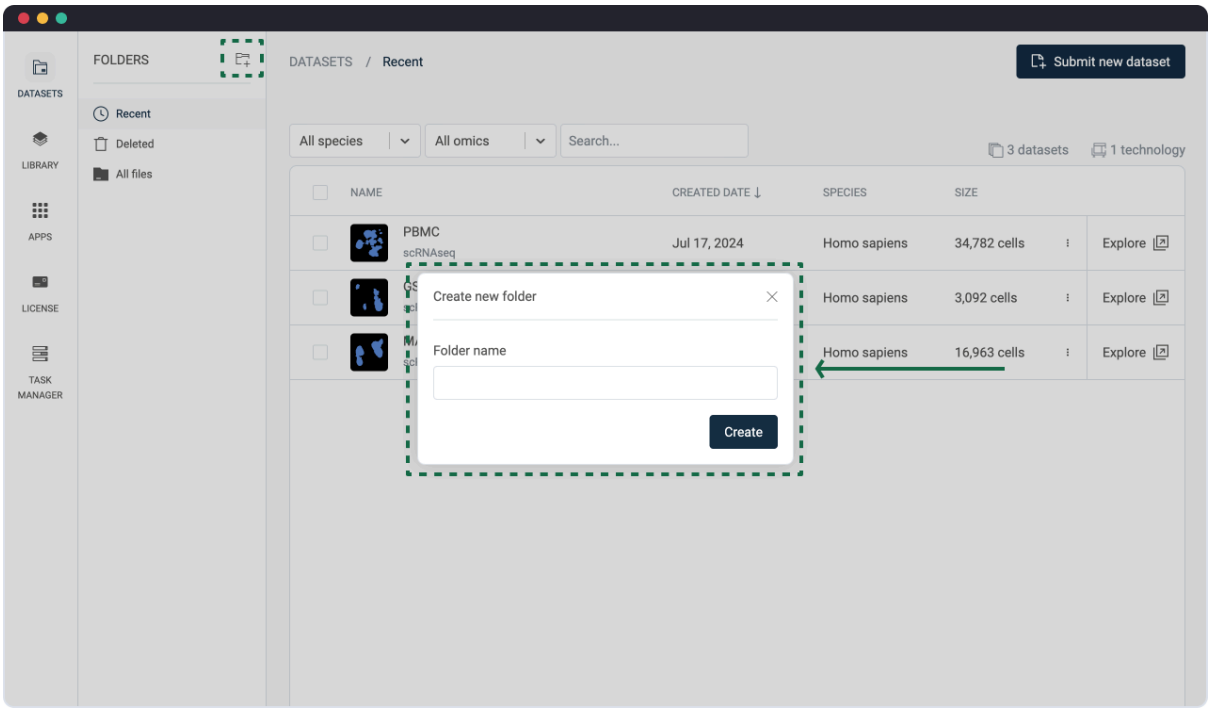


By default, all submitted datasets are stored under the `All files` folder and presented in a tabular format. This table can be sorted by clicking on any table header. You can use the filter options above the table to quickly find specific datasets based on your criteria. To explore a dataset in more detail, click on the Explore button associated with the dataset.



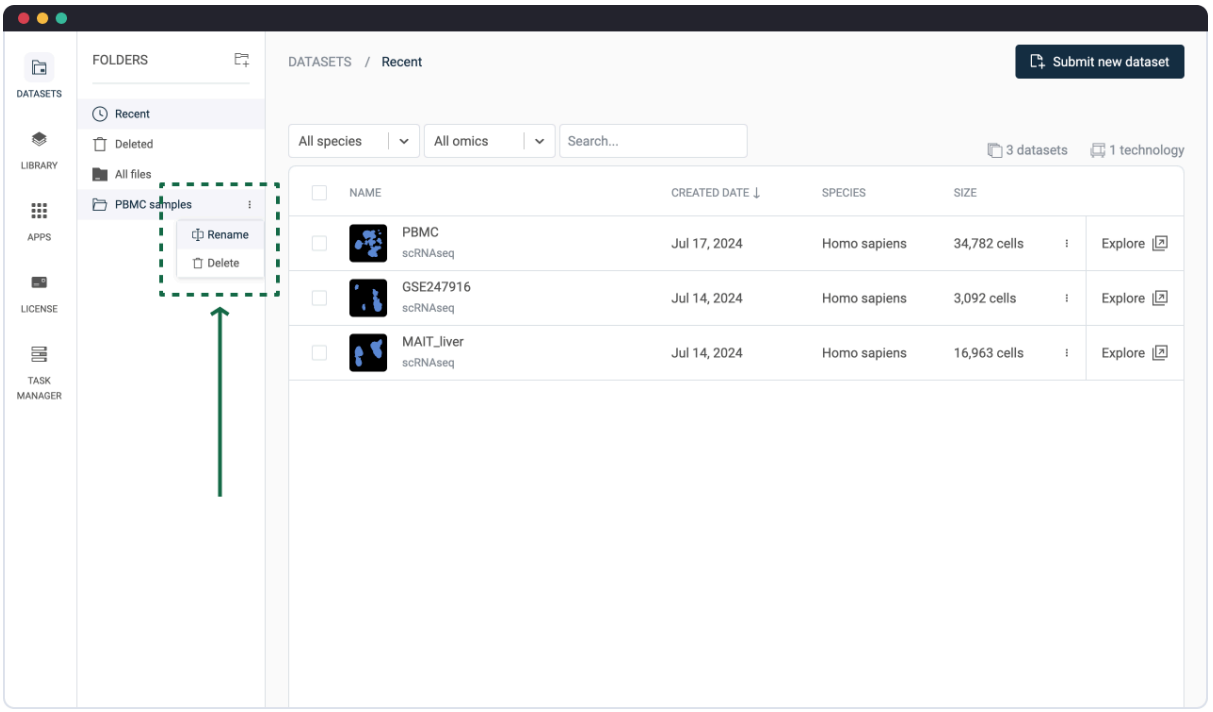
2. Dataset folders

To better organize your datasets, you can group them into folders. You can create a new folder by clicking on the folder icon at the top of the `Folders` panel.

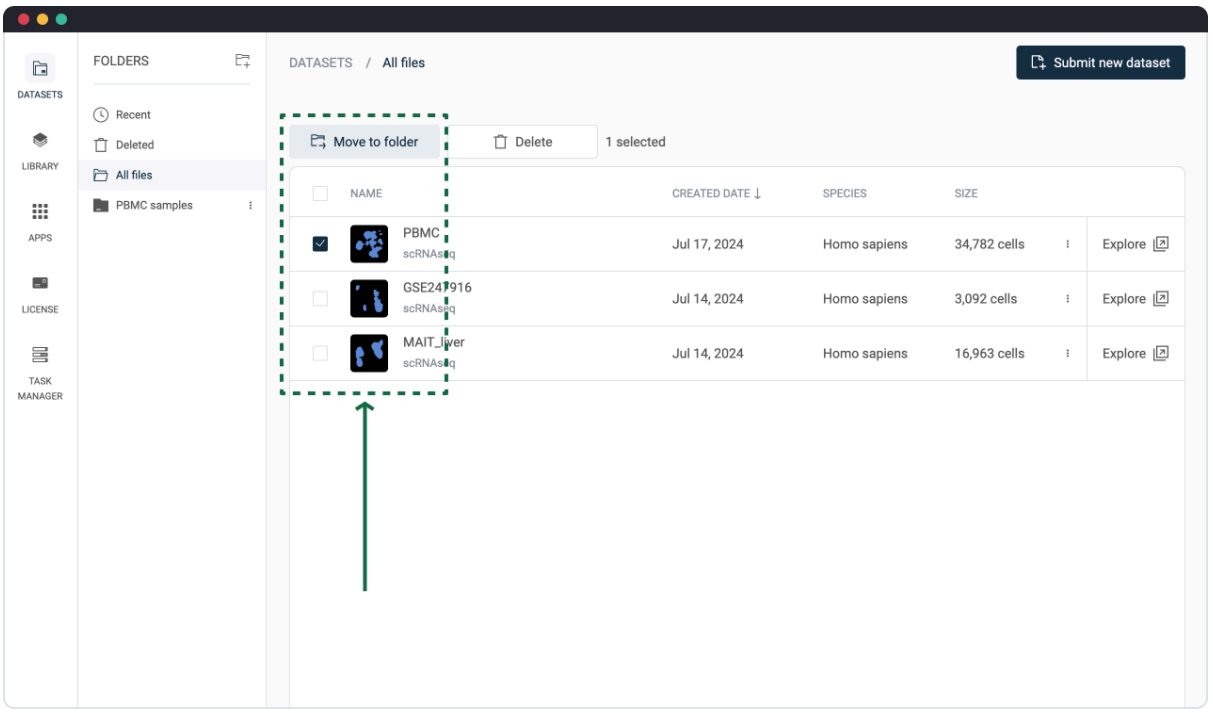


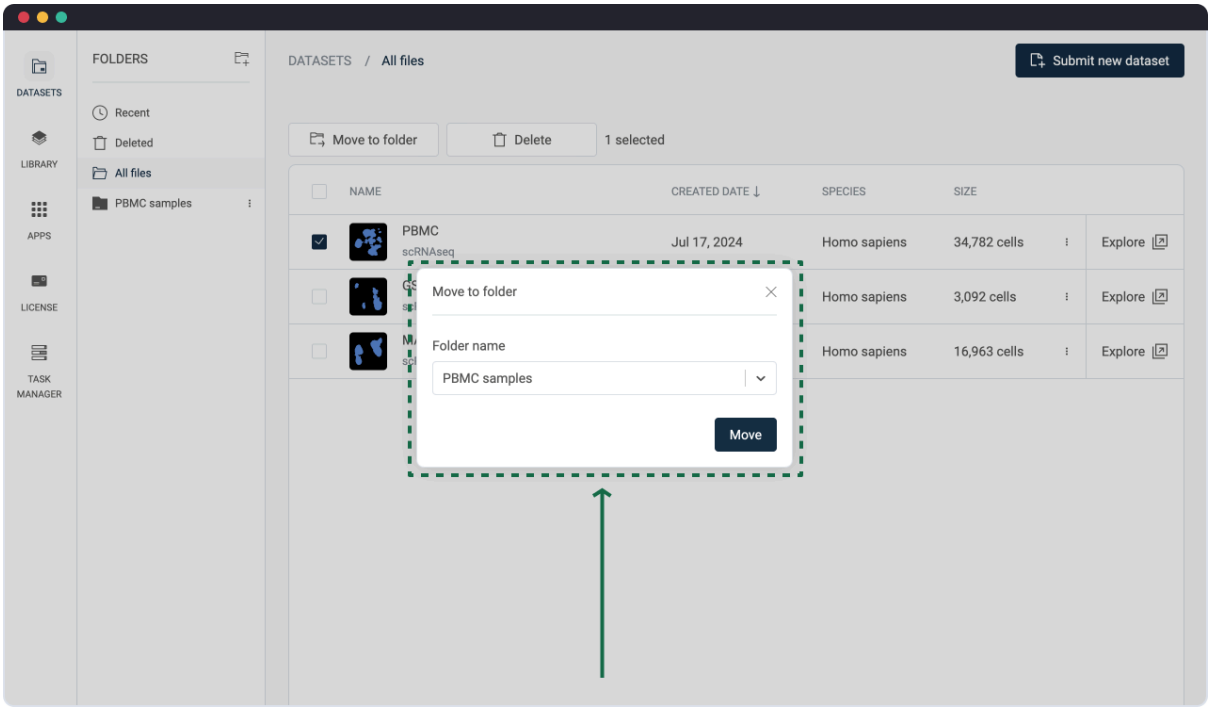
All created folders are listed in the 'Folders' panel. You can manage these folders by hovering over the three dots icon next to each folder name:

- `Rename`: Update the folder name.
- `Delete`: Remove the folder. All datasets within a deleted folder will be moved back to the `All files` folder.



To organize your datasets into different folders, from the dataset table, select the datasets you wish to move by checking the box next to each dataset. Click on the `Move to folder` button located above the table, and in the pop-up modal, select your destination folder.

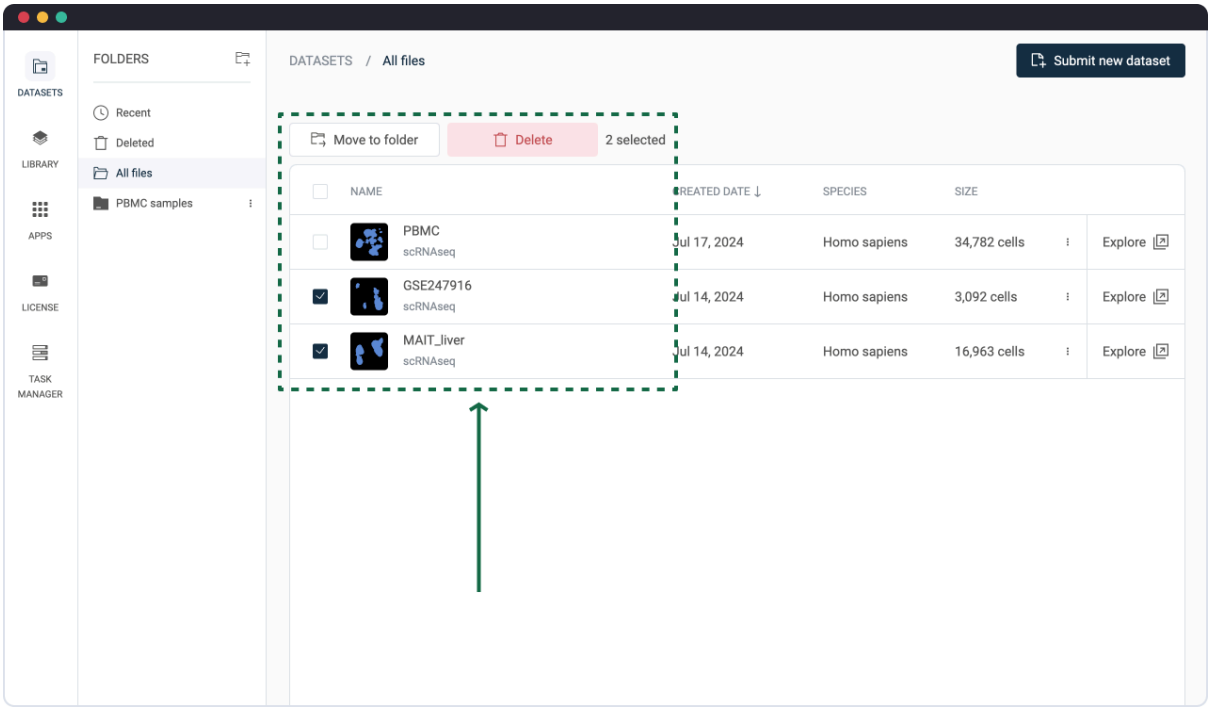




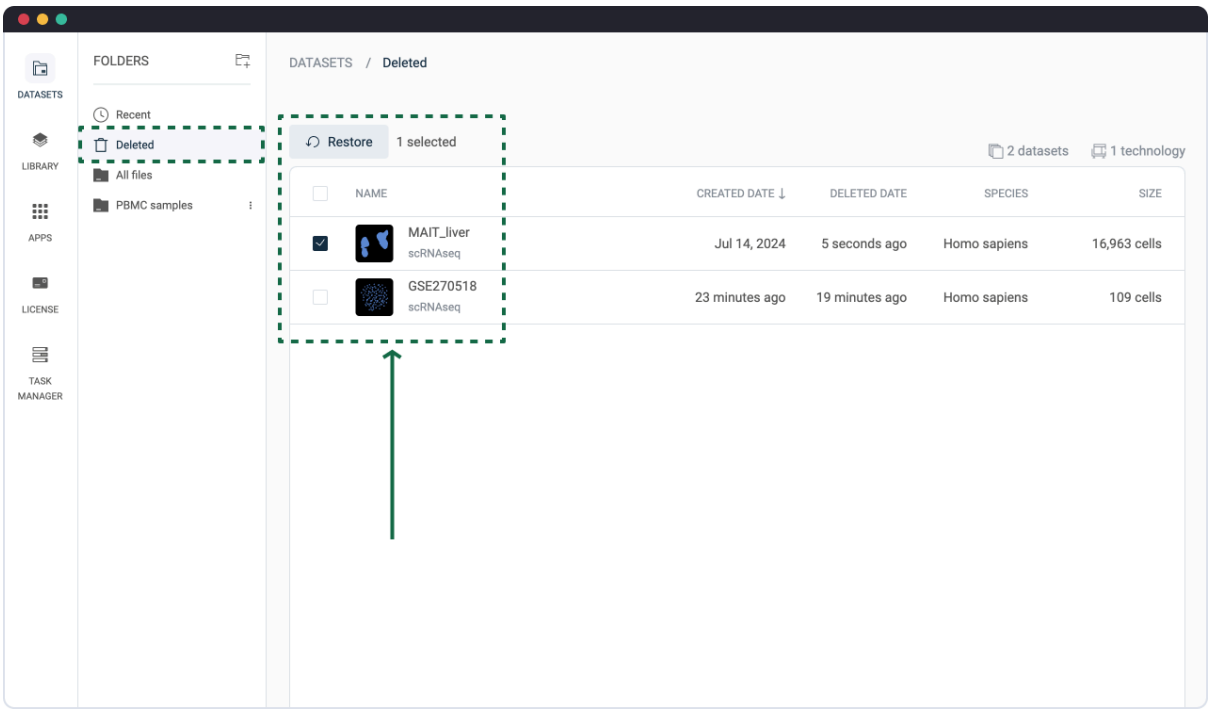
3. Delete and restore

To remove unneeded datasets, you can select datasets to delete by checking their corresponding boxes, then click the `Delete` button above the table. Deleted

datasets are moved to the `Deleted` folder and are held there for 30 days before permanent deletion. This temporary holding period helps prevent accidental data loss.

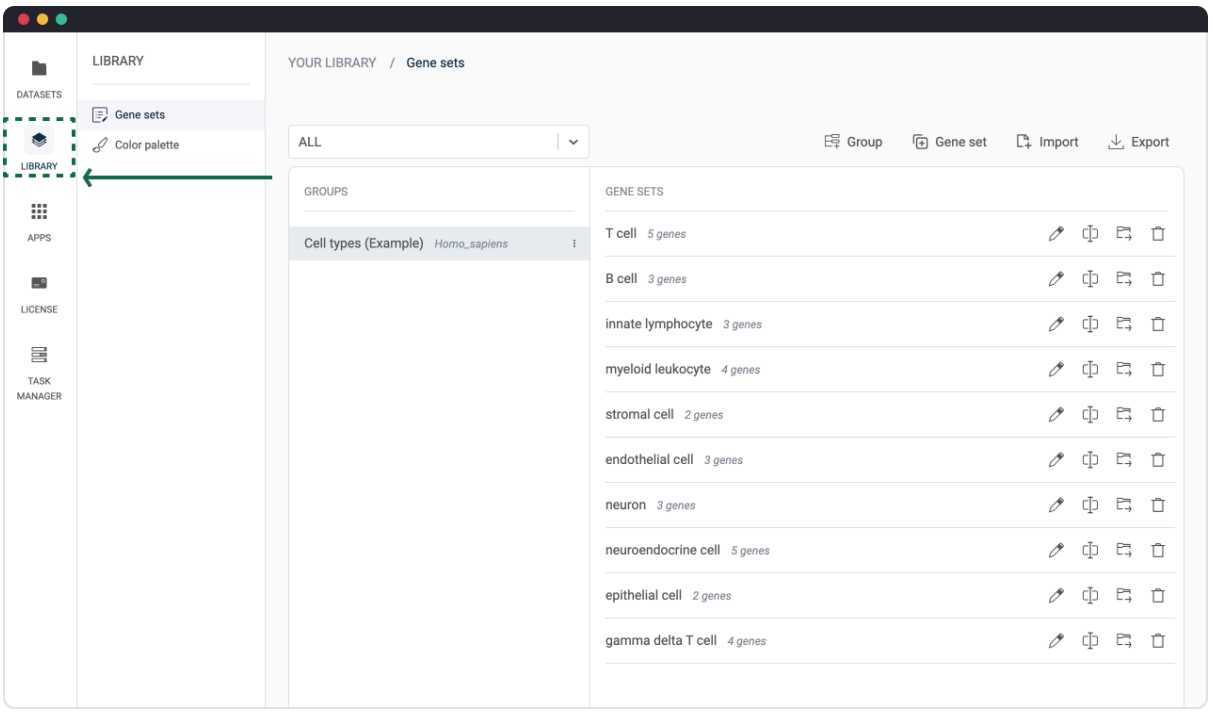


To restore a deleted dataset, navigate to the `Deleted` folder, select the datasets you wish to restore by checking their boxes, and then click the `Restore` button located above the table.


DATASETS
LIBRARY
APPS
LICENSE
TASK MANAGER
FOLDERS
Recent
Deleted
All files
PBMC samples
DATASETS / Deleted
Restore
1 selected
2 datasets
1 technology
NAME
CREATED DATE
DELETED DATE
SPECIES
SIZE
MAIT_liver
scRNAseq
Jul 14, 2024
5 seconds ago
Homo sapiens
16,963 cells
GSE270518
scRNAseq
23 minutes ago
19 minutes ago
Homo sapiens
109 cells

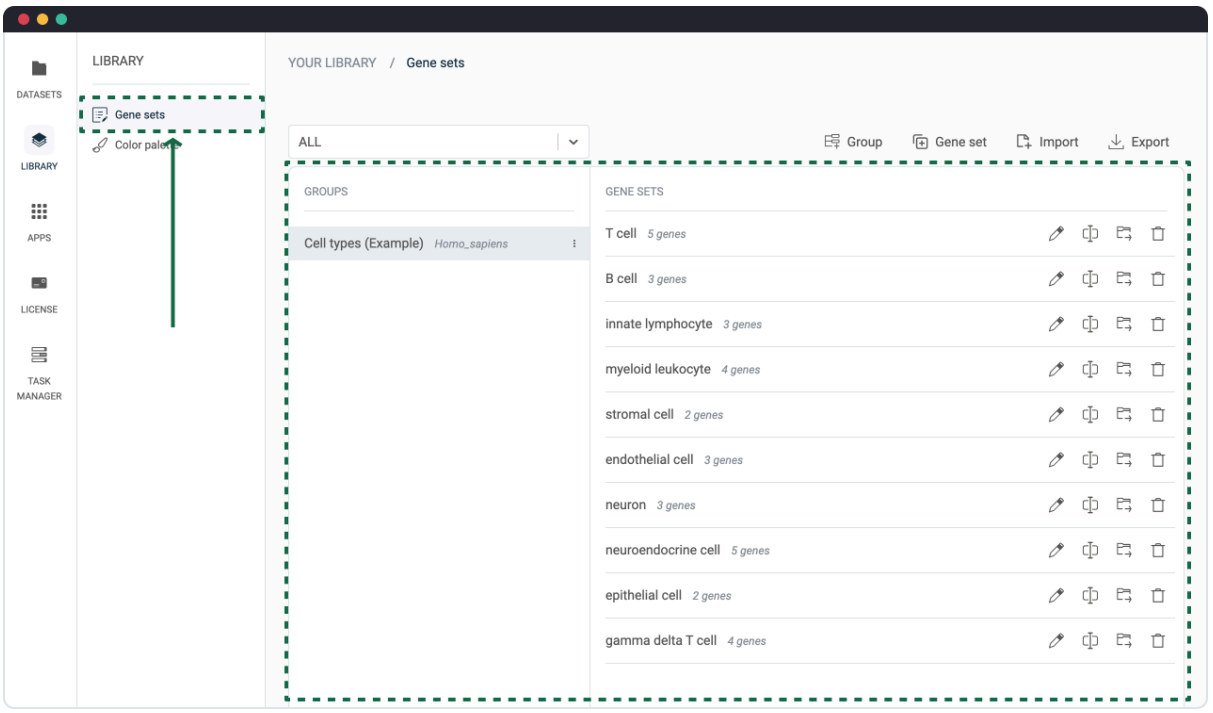
Your libraries

OmnibusX allows you to manage your own libraries for customizing your analysis. You can access and manage your libraries by clicking on the `Library` button on the left navigator.



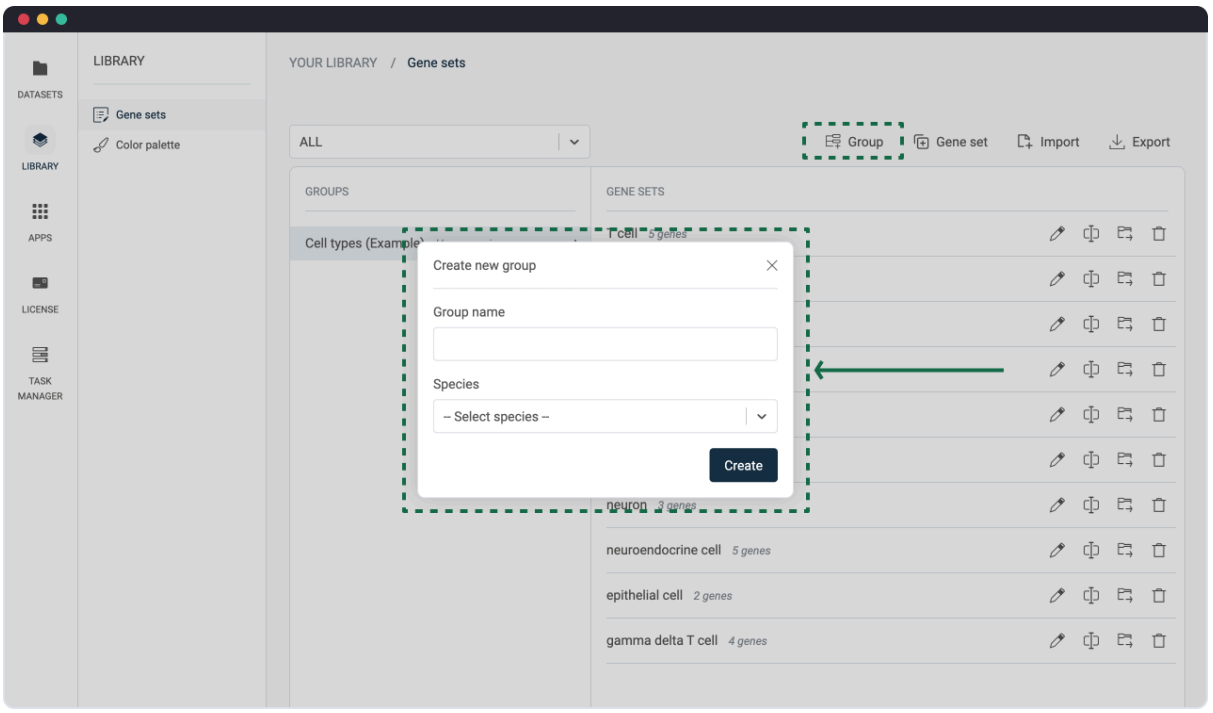
1. Gene set library

From the `Library` panel, select `Gene set` to view all your groups of gene sets. OmnibusX allows you to organize related gene sets into groups for easier management and to facilitate automatic labeling of cell populations based on these sets.



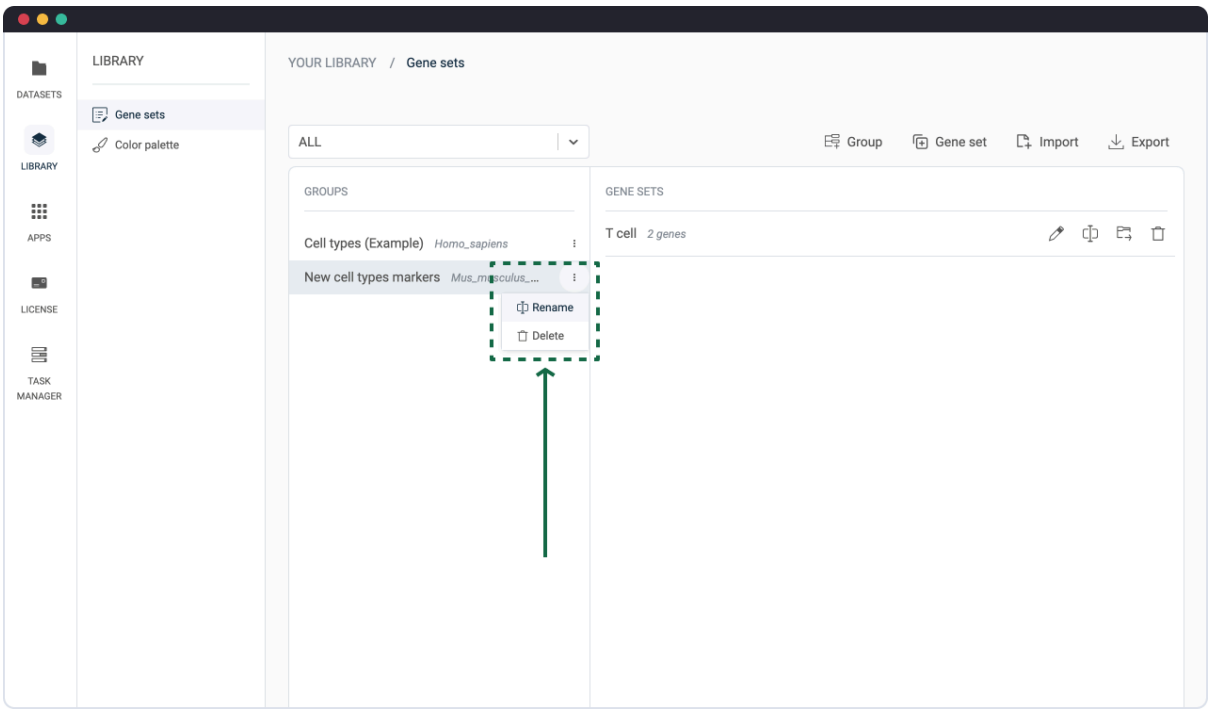
1.1. Group management

To create a group for your gene sets, click on the `Group` button at the top of the panel. You will need to provide a group name and specify the desired species, as gene sets are used across different datasets and require mapping to a common gene annotation. Based on the selected species, OmnibusX retrieves the corresponding gene reference from Ensembl to ensure that your gene names are accurately mapped to a standardized annotation.



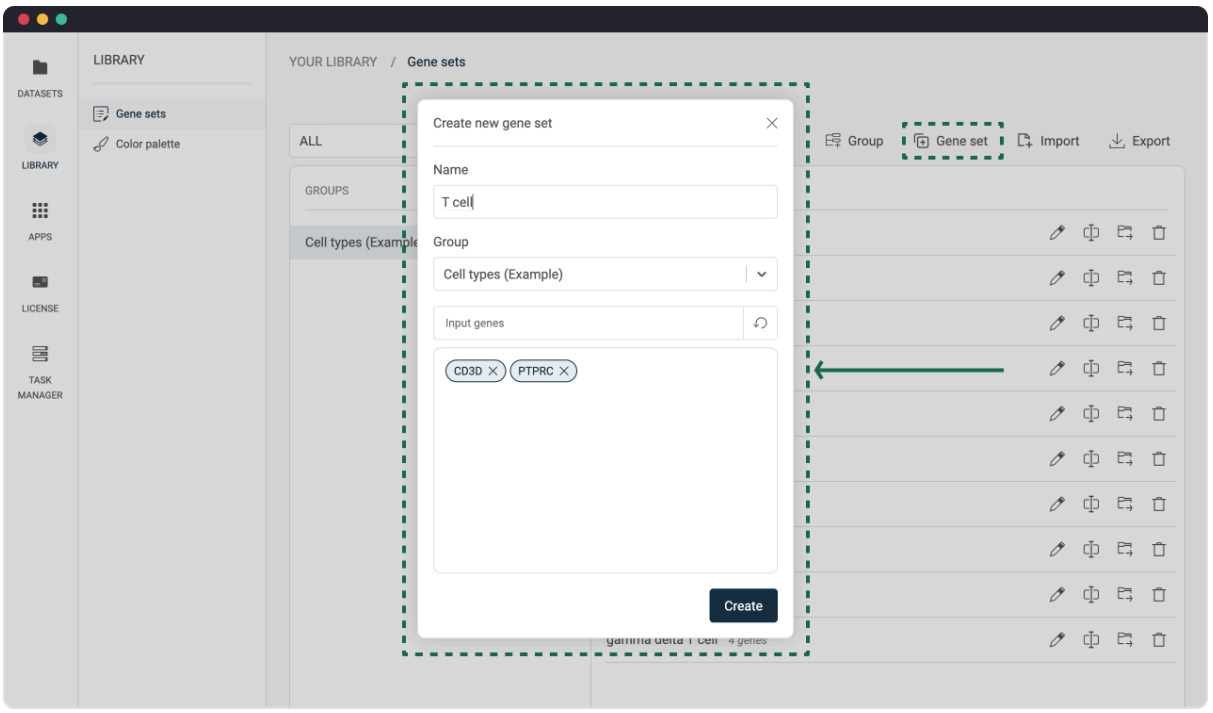
All created groups are listed in the left panel of the table. You can manage these groups by hovering over the three-dot icon next to each group name. Options include:

- `Rename`: Update the name of the group.
- `Delete`: Remove the group and all associated gene sets within that group.



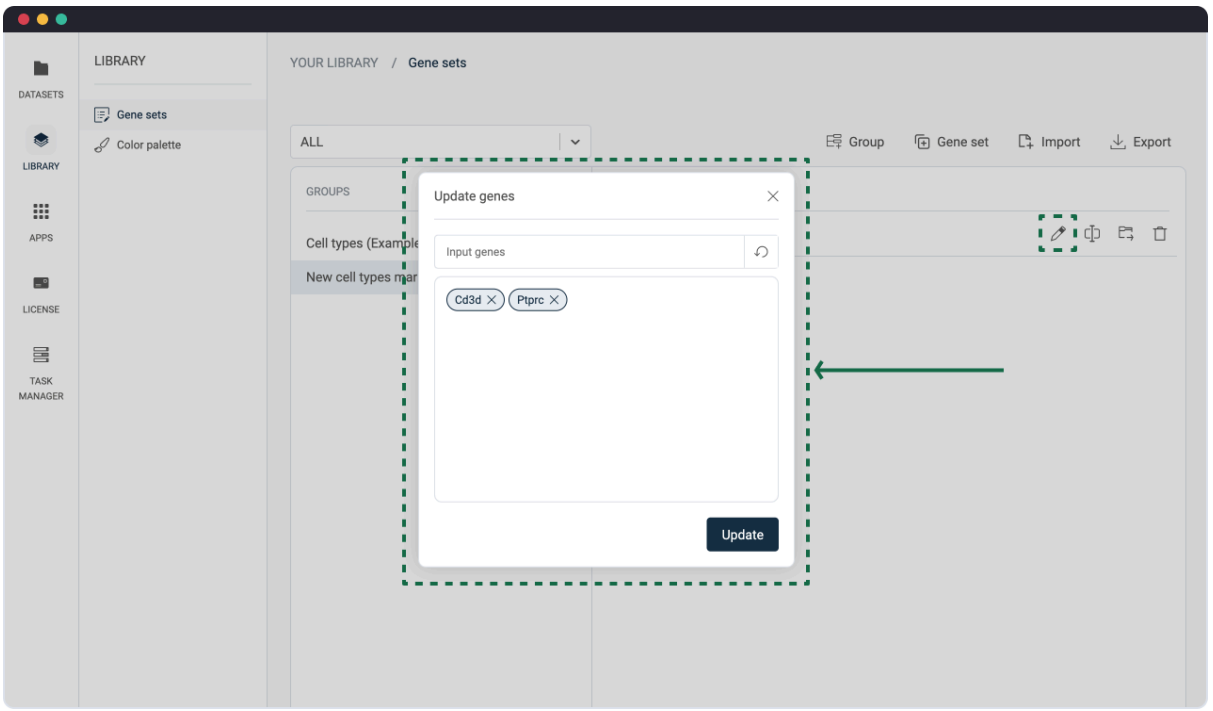
1.2. Gene set management

To add a gene set to your library, click on the `Gene set` button above the table. Enter the gene set name and select the target group. Based on the selected group species, OmnibusX will retrieve the corresponding gene annotation. You can manually input genes into the provided input box and select matches from OmnibusX's suggestions, or you can paste a list of genes at once. The pasted list will automatically be mapped against the gene annotation reference, standardizing names and removing unmapped genes. Click `Create` to save your gene set to the library.

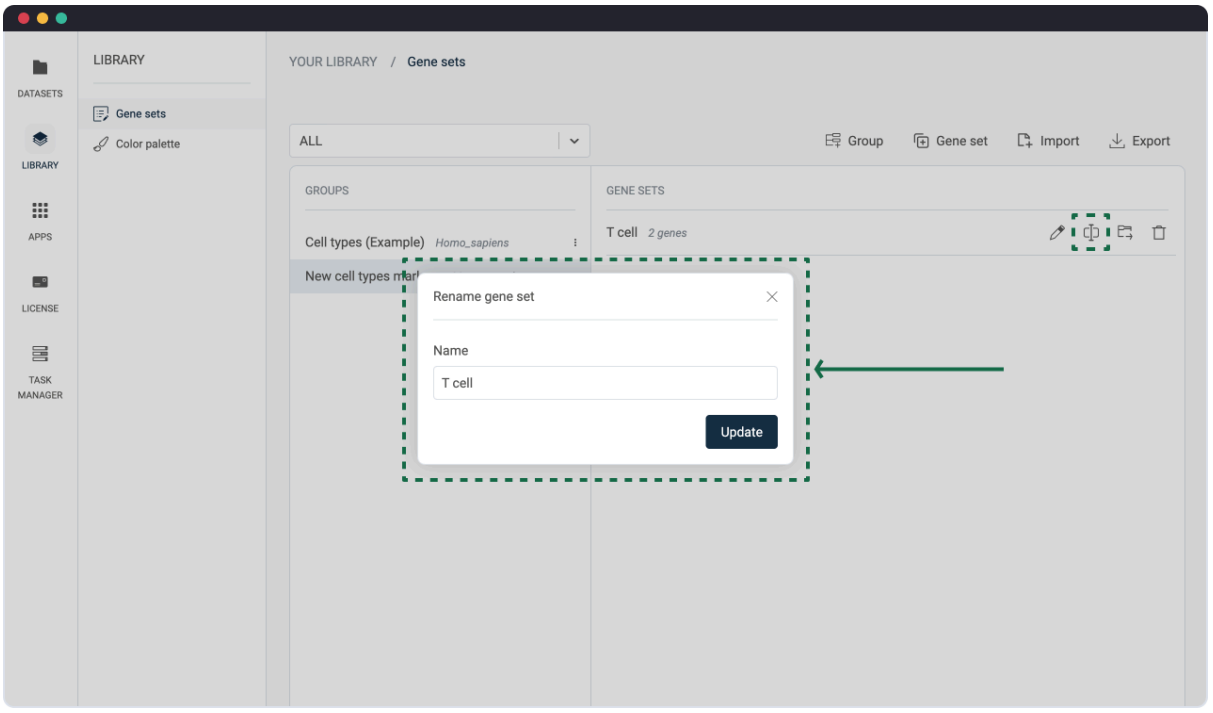


Click on any group name in the left table to view its gene sets in the right panel. OmnibusX offers several functions for managing your gene sets:

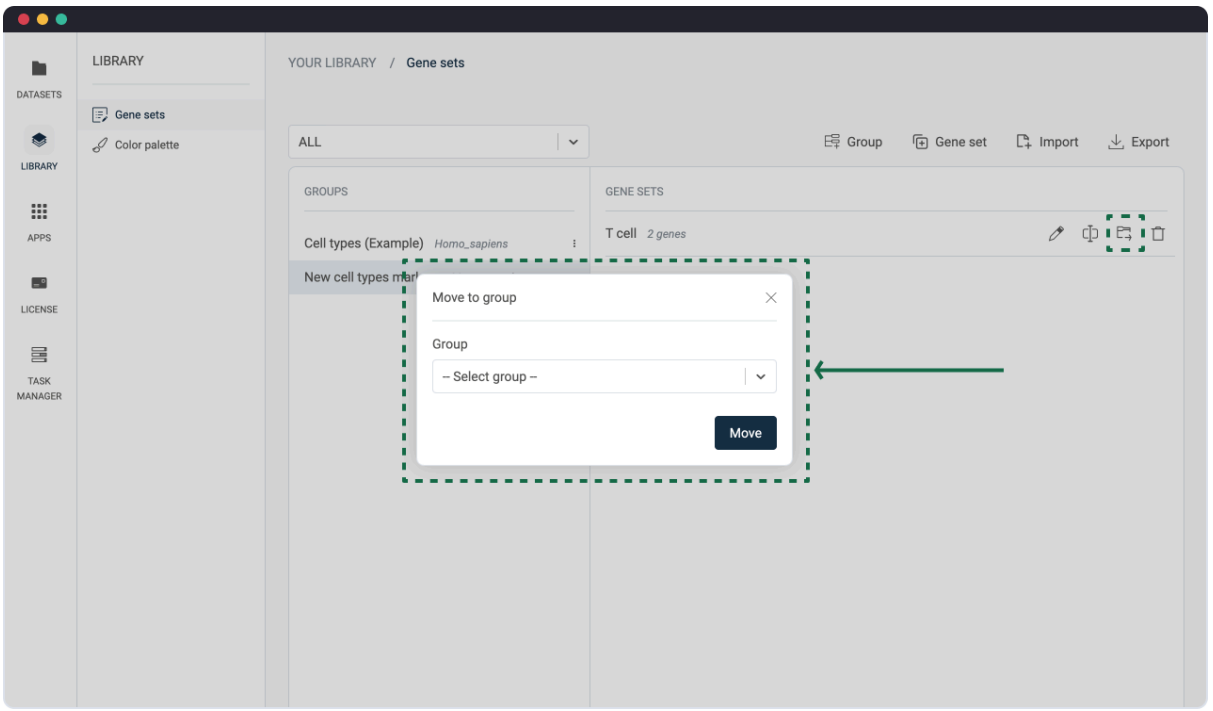
- `Update`: Modify the list of genes within a gene set.



- `Rename`: Update the gene set name.



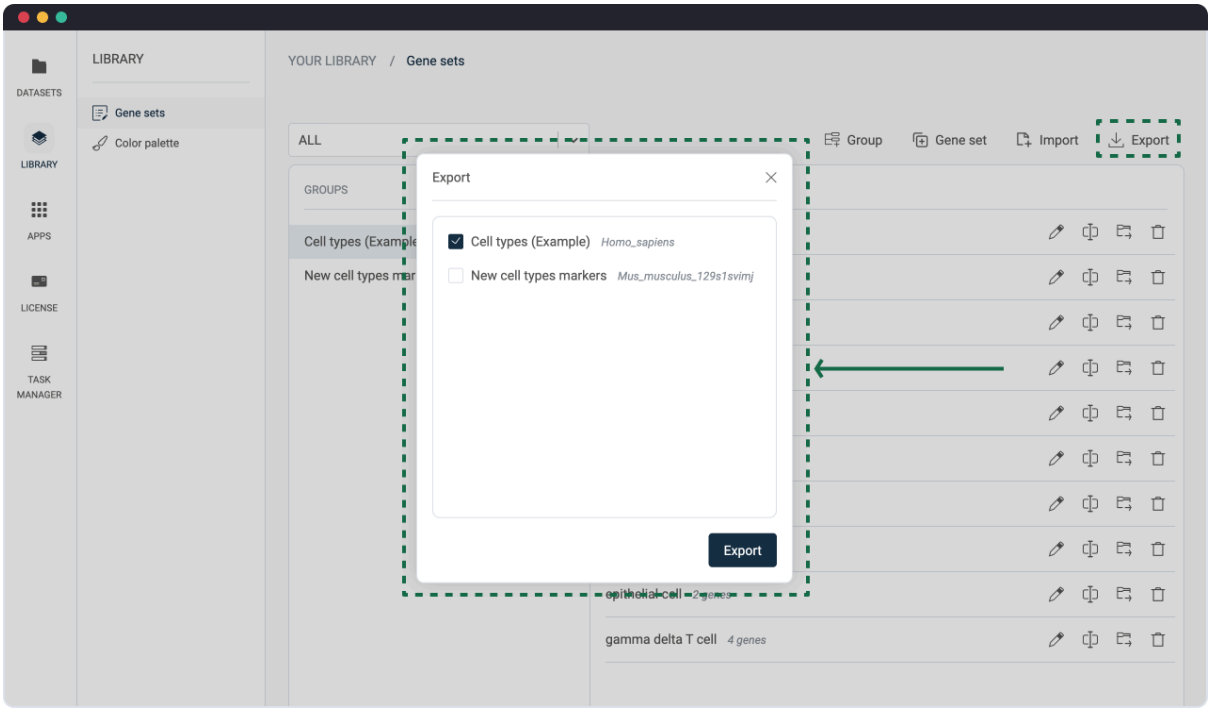
- `Move to group`: Change the group of the gene set. You can select new target groups from a dropdown, filtered by species compatibility.



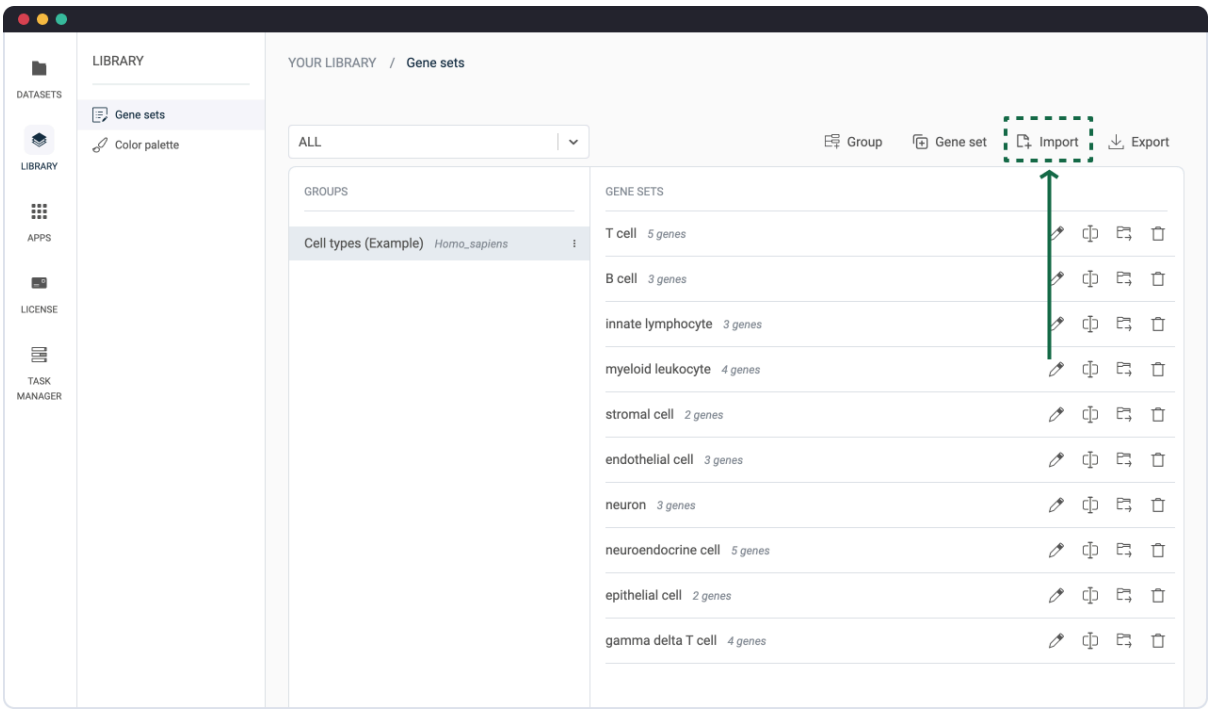
- `Delete`: Remove the gene set from the library.

1.3. Import and export

To share your gene set libraries with colleagues, click on the `Export` button above the table, select the groups of gene sets you wish to share, and then click `Export`.

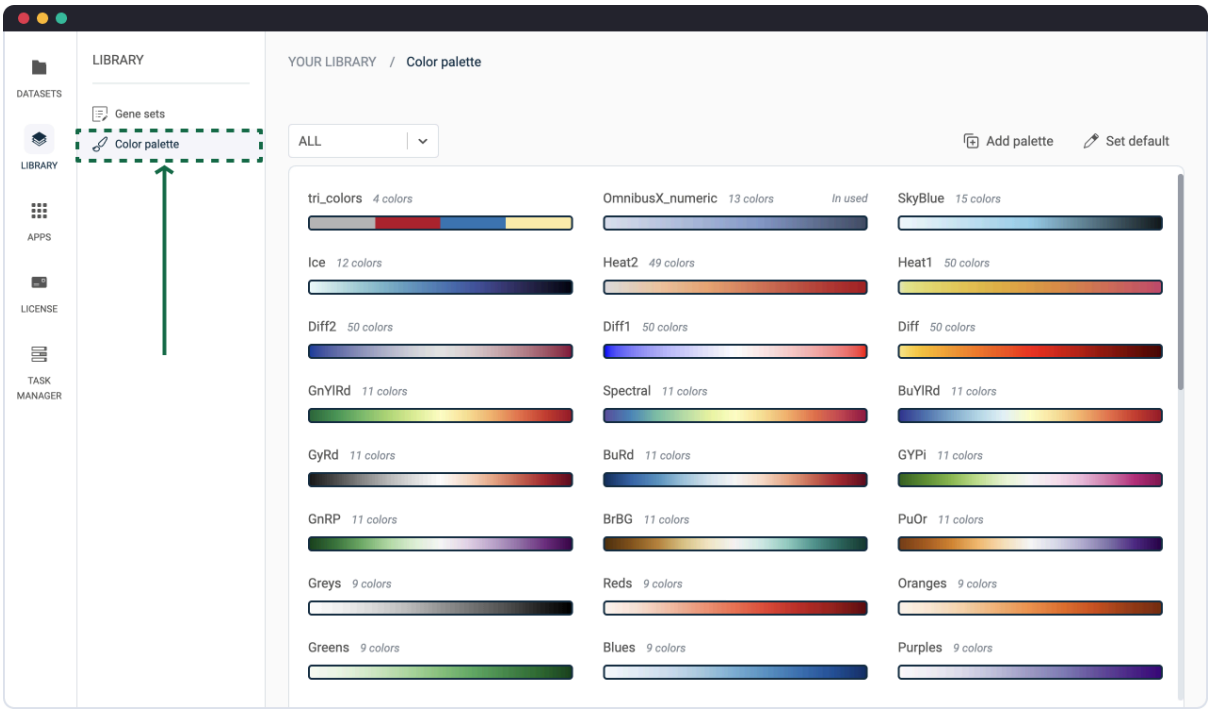


Your colleagues can import these gene sets by clicking on the `Import` button and selecting the provided file.



2. Color palette library

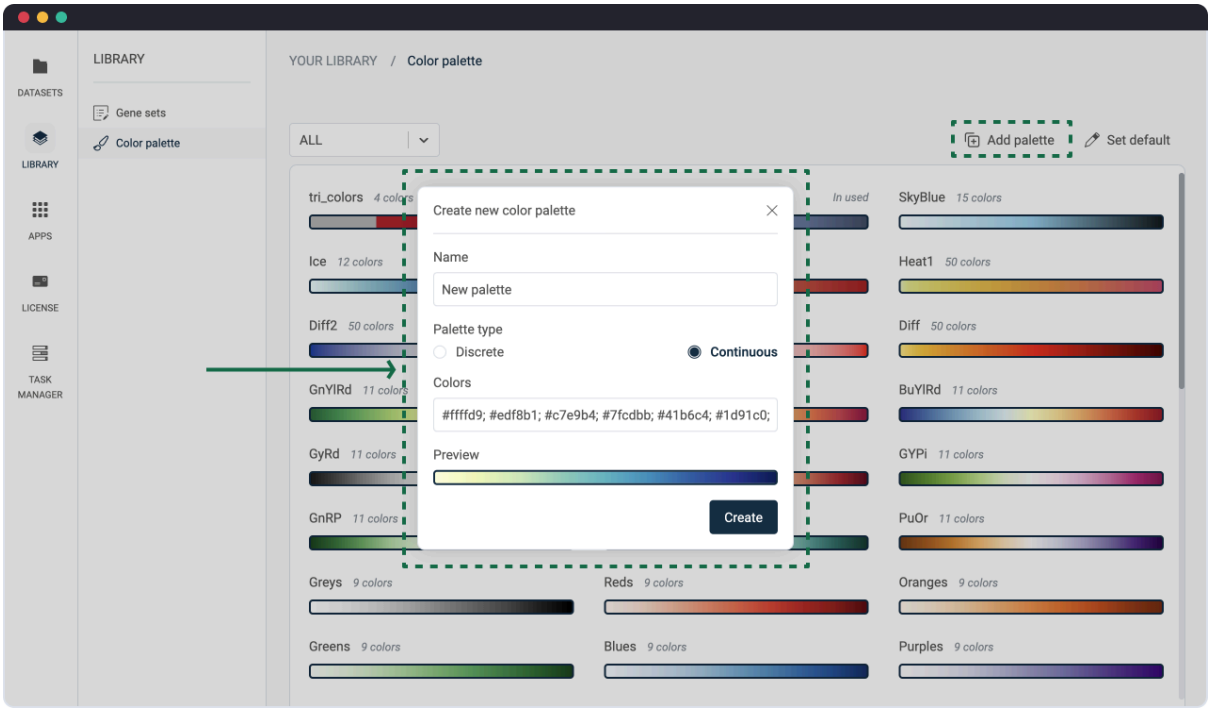
To begin creating a new color palette, navigate to the `Library` panel and select `Color palette`.



2.1. Create new color palette

By default, OmnibusX provides various palettes, you can add your own palette by clicking on the `Add palette` button located at the top of the panel. A form will appear where you can define the details of your new color palette.

- `Name`: Input a name for your color palette to easily identify it later.
- `Palette type`: Specify whether the palette is meant for discrete (categorical) data or continuous data. Discrete palettes are typically used for categorical labels, while continuous palettes are better suited for representing numerical values where a gradient might be necessary.
- `Colors`: Input the colors that will make up your palette. Colors can be specified in hex code (e.g., #FF5733) or RGB code (e.g., rgb(255,87,51)), separated by semicolons `;`.

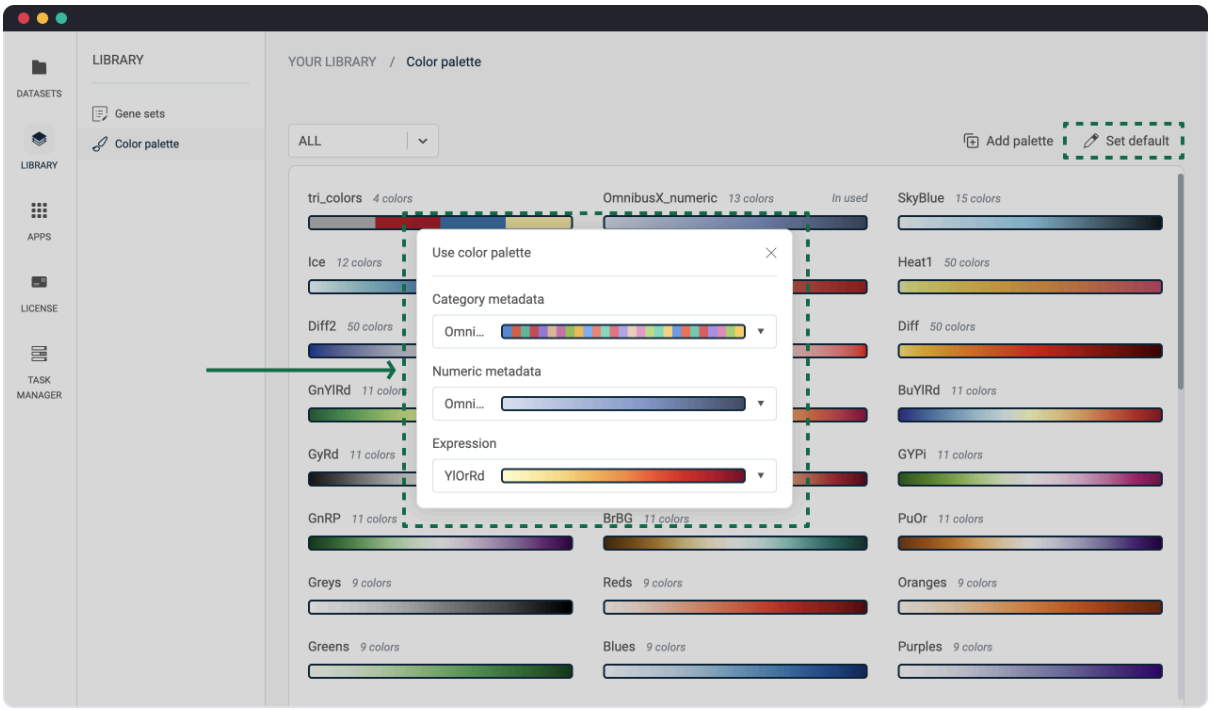


After entering all the details, click `Create` to add your new color palette to your library.

2.2. Default color palette

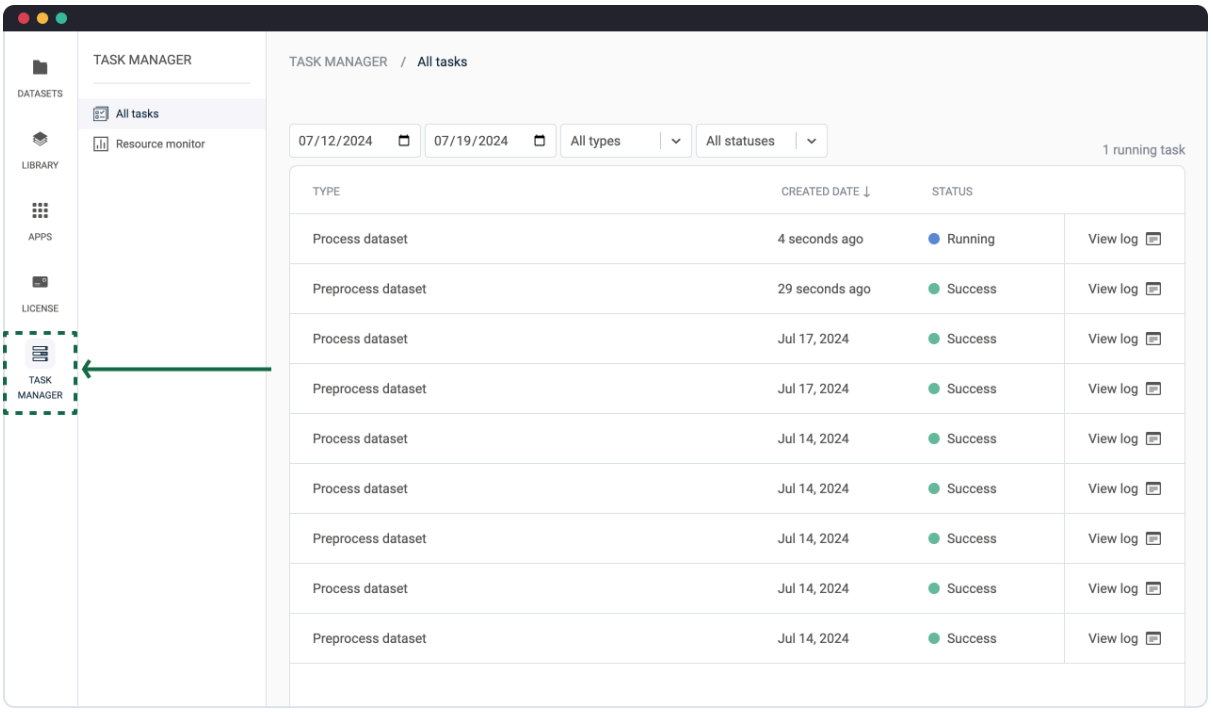
OmnibusX enables you to set default color palettes for different types of data, ensuring consistency across all your analyses. You need to configure default palettes for the following categories:

- `Category metadata`: Choose a palette that will be used for coloring plots based on categorical metadata.
- `Numeric metadata`: Select a palette for numeric metadata which typically requires a gradient to effectively represent the range and distribution of values.
- `Expression`: Define the palette used for visualizing gene expression data.



Task management

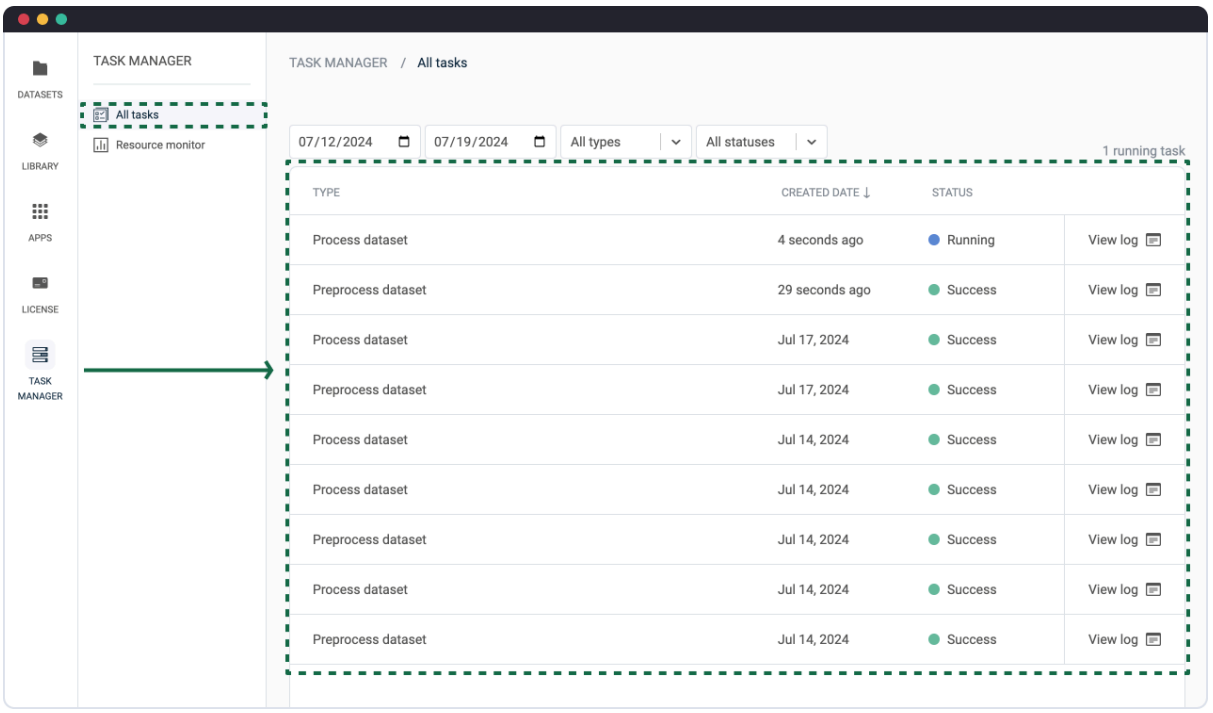
In OmnibusX, the Task Manager is an essential feature that allows you to efficiently manage and monitor tasks that are computationally intensive and time-consuming. By running these tasks in the background, OmnibusX enables you to continue working on other datasets without disruption, optimizing your workflow and resource utilization. To view and manage all ongoing and completed tasks, click on the `Task manager` button on the left navigator



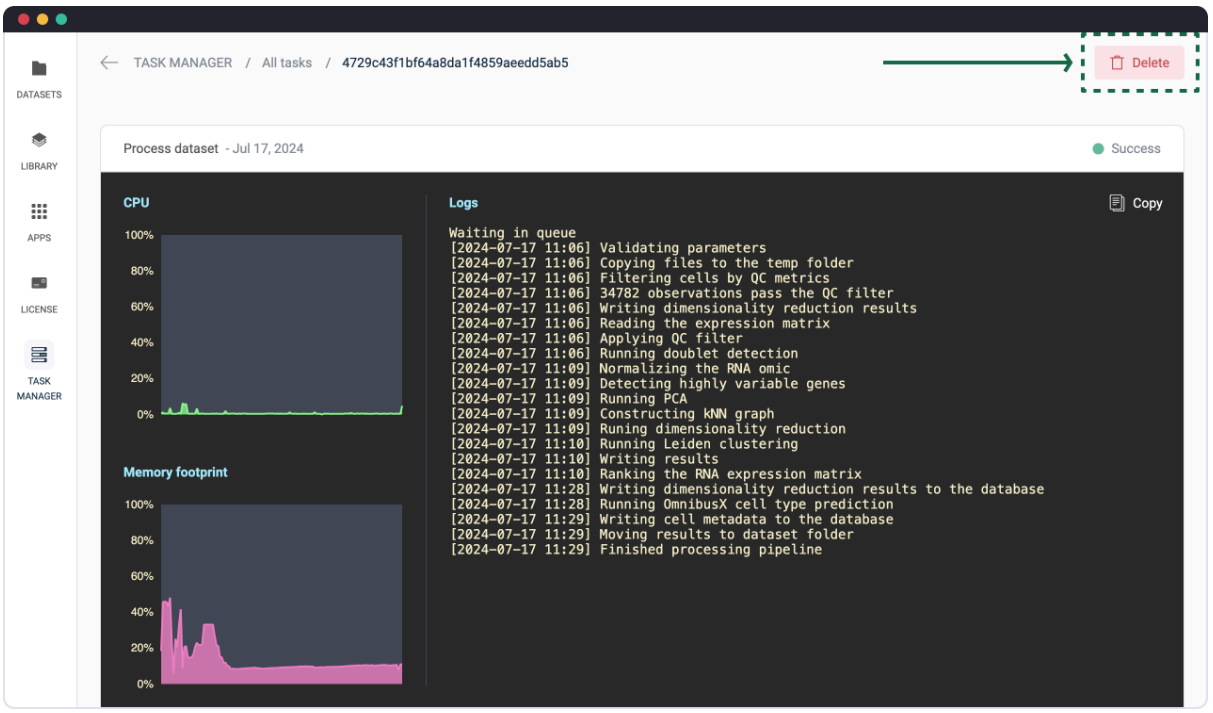
1. Tasks

In the `Task manager` panel, select `All tasks`. All submitted tasks are listed in a table in the right panel, each with its current status:

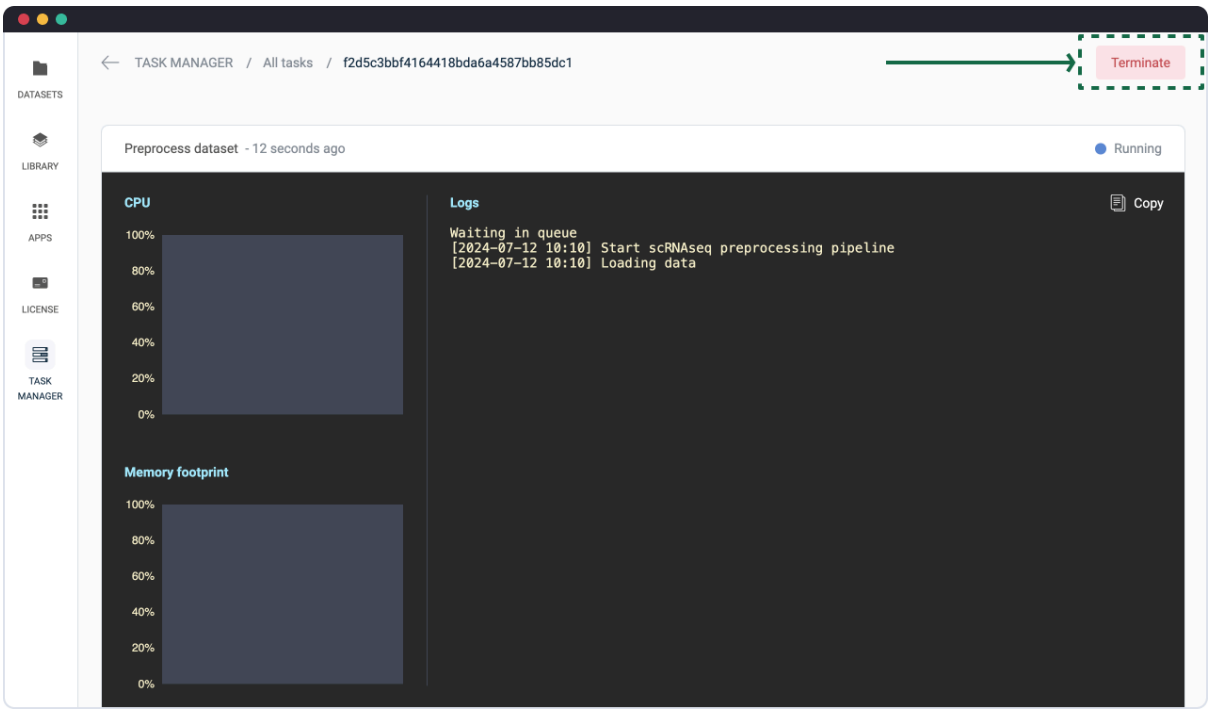
- `Pending`: The task is queued and waiting for processing. Since multiple tasks can be submitted simultaneously, OmnibusX processes them sequentially to prevent exhaustion of computer resources.
- `Running`: The task is currently being processed.
- **`Success`**: The task has been completed successfully.
- **`Error`**: The task has been corrupted because of errors during execution.



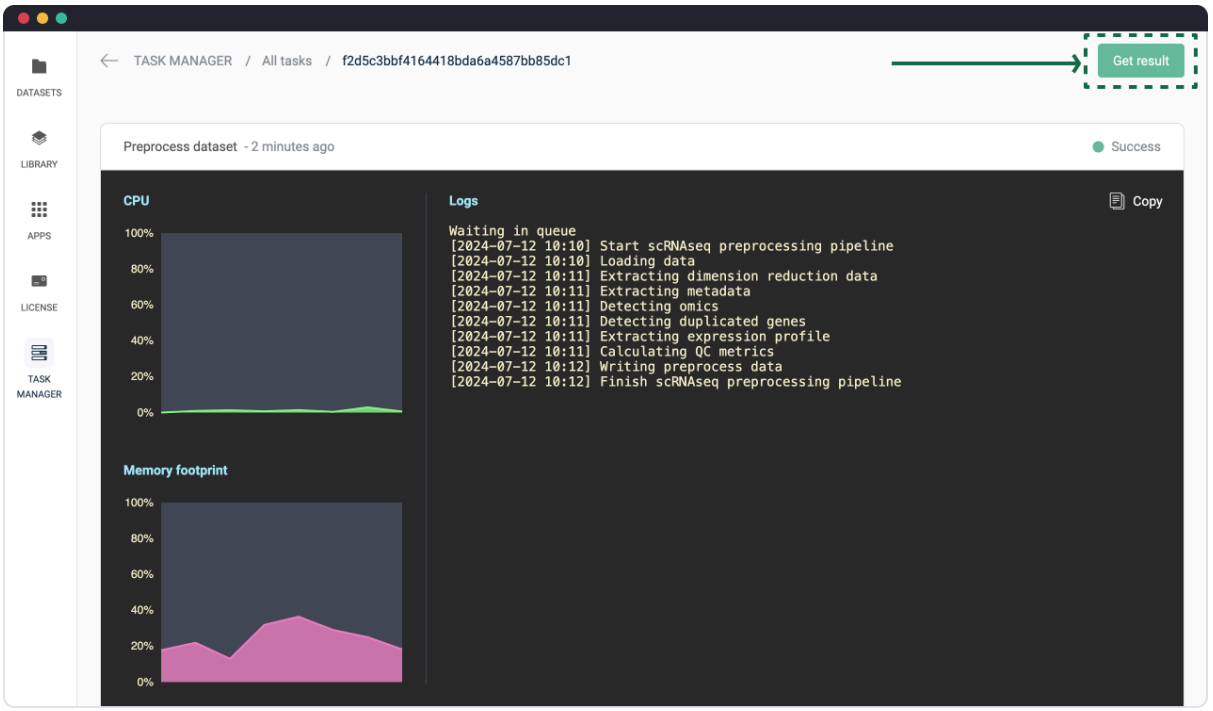
Click on the `View log` button beside each task to access detailed process information. For completed tasks, logs can be removed by clicking the `Delete` button at the top of the detail page.



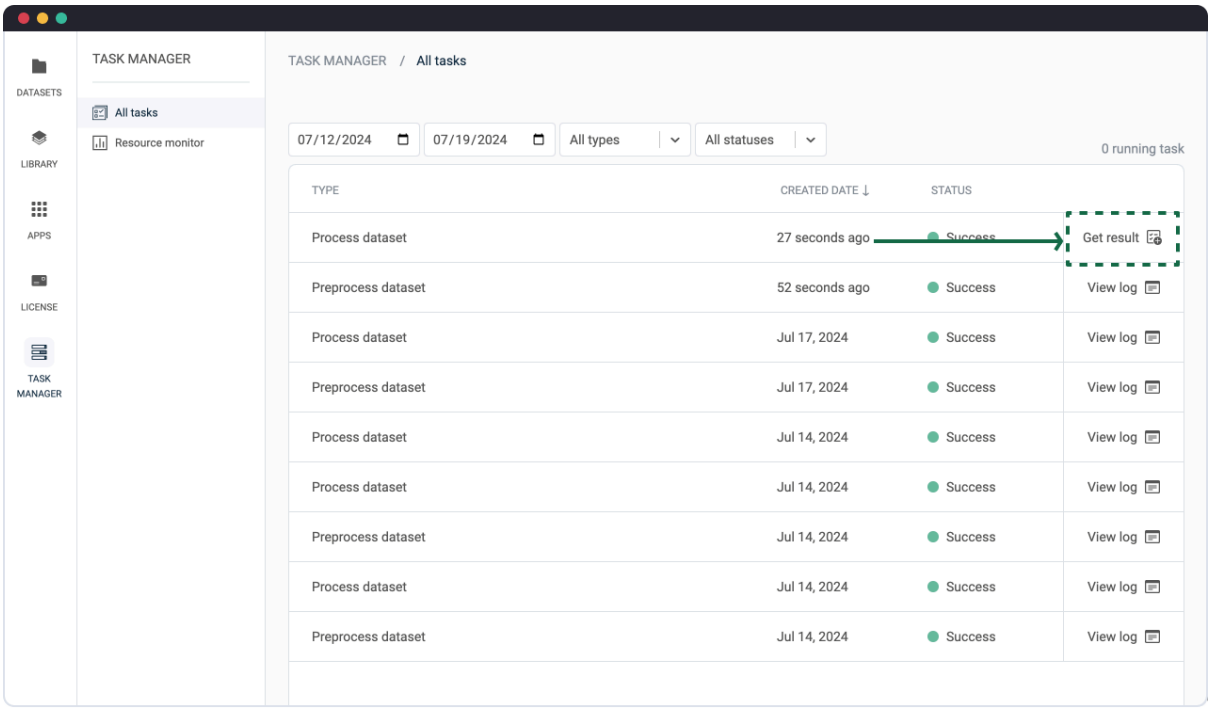
For running tasks, the log updates in real-time. If you wish to stop a running task, you can terminate it by clicking the `Terminate` button.



If a task completes while you are viewing the log, the results are automatically updated in your application.

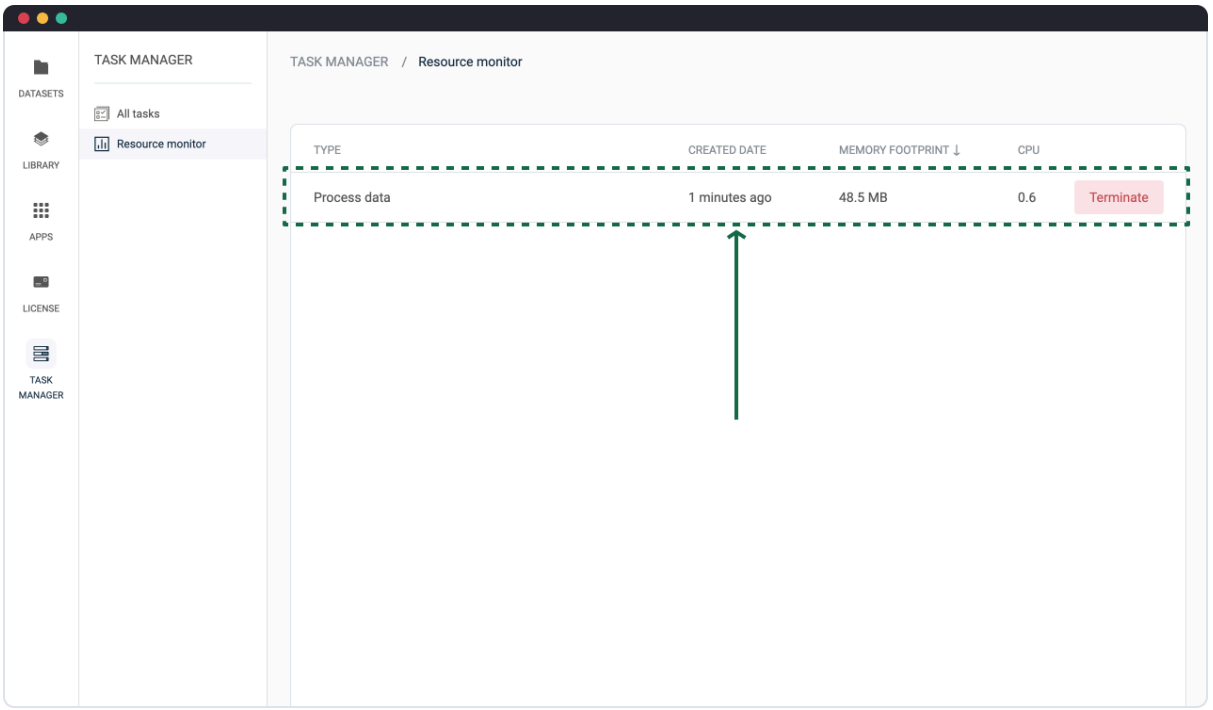


If you are in another section when the task finishes, the result remains on the task server. Retrieve it by navigating back to the `Task manager` page and clicking the `Get result` button next to the relevant task.



2. Resource monitor

The `Resource monitor` section allows you to track the resource usage of currently running tasks. If a task is using an excessive amount of resources, or if system limits are being approached, you can preemptively terminate the task from the `Resource monitor` to prevent any potential issues.



Thank you!

We sincerely thank you for choosing OmnibusX for your research and data analysis needs. We hope this user guide has been helpful in assisting you to effectively navigate and utilize the various features within OmnibusX. Our goal is to support your scientific endeavors by providing a powerful and intuitive platform that enhances your ability to analyze, visualize, and interpret complex datasets.

We continuously strive to improve OmnibusX and tailor it to the needs of our diverse user base. Your feedback is invaluable to us as it helps us enhance the functionality and user experience of our platform. Please do not hesitate to contact us at support@omnibusx.com with any feedback, questions, or suggestions you may have. Your insights are crucial for us to deliver a better and more efficient user experience.

Thank you once again for your trust in OmnibusX. We are excited to be a part of your research journey and look forward to supporting your groundbreaking work in the years to come.