Cytonaut™ Cloud Software
for Single-Cell RNA-seq Data Analysis

Catalog nb. 101-1000

USER GUIDE

For analysis of 3' single-cell RNA sequencing data
of samples prepared with Asteria™ kit

Revision nb. 1.8
Publication date: 09 November 2023

For Research Use Only. Not for use in clinical diagnostics procedures.

Manufacturer:
SCIPIO BIOSCIENCE,
Pépinière Paris Santé Cochin,
29 rue du Faubourg Saint Jacques,
75014 Paris, France
http://www.scipio.bio

Revision history:

Version nb	Publication Date	Description
1.8	09 Sep. 2023	Update of the user manual for the new Cytonaut™ product release v1.6. For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.7	04 Sep. 2023	Update of the user manual for the new Cytonaut™ product release v1.5. For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.6	27 Apr. 2023	Update of the user manual for the new Cytonaut™ product release v1.4. For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.5	02 Feb. 2023	Update of the user manual for the new Cytonaut™ product release v1.3 For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.4	12 Dec. 2022	Update of the user manual for the new Cytonaut™ product release v1.2.2. For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.3	02 Oct. 2022	Update of the user manual for the new Cytonaut™ product release v1.2. For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.2	08/08/2022	Update of the user manual for the new Cytonaut™ product release v1.1. For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.1	29/06/2022	Update of the user manual for the new Cytonaut™ product release v1.0.3. For more details, see the Release Notes available here: https://cytonaut-scipio.bio/release-notes
1.0	02/05/2022	First version of the user manual for Cytonaut™ product v1.0.2.

Important Licensing Information:
These products may be covered by one or more Limited Use Label Licenses. By use of these products, you accept the terms and conditions of all applicable Limited Use Label Licenses.

The information in this guide is subject to change without notice.

DISCLAIMER:
TO THE EXTENT ALLOWED BY LAW, SCIPIO BIOSCIENCE WILL NOT BE LIABLE FOR SPECIAL, INCIDENTAL, INDIRECT, PUNITIVE, MULTIPLE, OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH OR ARISING FROM THIS DOCUMENT, INCLUDING YOUR USE OF IT.

TRADEMARKS:
All trademarks are the property of SCIPIO BIOSCIENCE unless otherwise specified.

QUALITY AND CONTINUOUS IMPROVEMENT

We have implemented comprehensive quality control procedures to guarantee that Products delivered to our Customers match the highest standards of Quality.

In a continuous improvement effort, we encourage you to report any faults you observe by sending an email to us at *support@cytonaut-scipio.bio*.

We thank you in advance for supporting us to provide the best quality Products.

Table of Contents

I. PRODUCT INFORMATION

1. Product Description

Scipio bioscience Cytonaut™ is a cloud software solution for the analysis of 3' single-cell RNA sequencing data derived from samples prepared with Scipio bioscience's Asteria™ benchtop kit. Cytonaut™ is accessible at the website https://www.cytonaut-scipio.bio.

2. Intended Use

Cytonaut™ is a software-as-a-service application intended for researchers to perform analysis of 3' single-cell RNA sequencing data for Research Use Only (RUO). It is not for use in diagnostic procedures.

3. Compatibility Conditions

Cytonaut™ is compatible with any standard computer connected to the Internet. For optimal use, the minimum screen resolution is 1280x800 pixels.

Cytonaut™ supports all standard web browsers (e.g. Chrome, Firefox or Safari) and operating systems (e.g. Windows, Mac, Linux) in their latest version or in a version which is still supported by the manufacturer.

For raw sequencing data input, Cytonaut™ supports FASTQ files of samples which have been prepared with Asteria™ kit and sequenced by applying the instructions of use of Asteria™ kit available here, using one sample index per Asteria preparation: https://scipio.bio/resources/.

Besides, FASTQ files shall be provided in .fastq.gz format to Cytonaut™ application.

- To upload data in Cytonaut™, the user must previously agree with the following statement:

  "I certify that the sample data I am going to upload does not allow to identify a human individual and is not derived from human cell samples that have been collected for prevention, diagnosis, care or medico-social follow-up activities on behalf of natural or legal persons at the origin of the production or collection of this data or on behalf of the patient himself."

- Only the FASTQ files derived from samples prepared with Scipio Asteria™ kit technology are intended to be supported by Cytonaut™.

  However, other sample preparation technologies, library preparation or sequencing protocols may be compatible as far as the format of uploaded FASTQ files complies with the following acceptance criteria:

  - Length of read1 shall be either 25 or 26 or 30 or 31 bases;

  - If length of read1 is 30 or 31, then the global percentage of T bases in the last 4 bases of read1 shall be higher than 50% (accordingly to the polyT part of the barcode pattern used in Asteria™ kit);

  - Length of read2 shall be 50 bases or more.

  Besides, Cytonaut does not guarantee result accuracy for concatenated FASTQ files unless read order is randomized.

- Cytonaut™ also supports the upload of count matrix files without restriction of the used single-cell sample preparation technology, as far as the input format complies with one of these conditions:

  - (1) one non-compressed .tsv file including all count matrix information (genes in first column, then one column per cell barcode), similar to the .tsv count matrix files generated by Cytonaut

  - (2) one archive file (.zip, .tar, .gz, .tar.gz) containing 3 files which are compressed or not:

    - 1 .mtx file for the count matrix
    - 1 .tsv file for the list of cell barcodes with "barcode" or "cell" in its filename
    - 1 .tsv file for the list of genes with "gene" or "feature" in its filename.

    Note: for optimal use, the gene format shall start with the universal gene name.

4. User Workflow & Data Processing

The graphical interface of Cytonaut™ guides you step by step in a chronological order, where each step is presented from top to bottom in the left panel:

1. Upload Data / Define Samples
2. Create / Open Projects
3. Project Details
4. Run Pre-processing Analysis
5. Check Quality Indicators
6. Get Quality Reports & Count Matrices
7. Run Post-processing Analysis
8. Visualize Result Data
9. Download Results

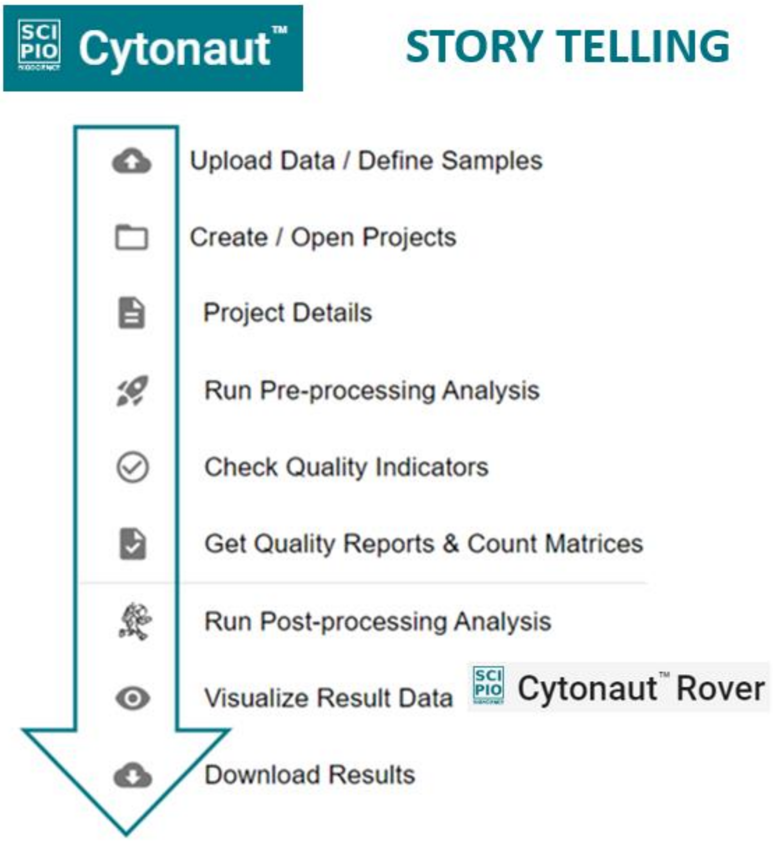


Figure 1. User Workflow in Cytonaut™

The data pre-processing phase (also called “primary analysis”) takes as input, for each sample to process, the couple of FASTQ files (R1 and R2) which constitute the raw sequencing data of the sample, and provides as output the sample quality indicators, as well as the sample **count matrix** which contains for each cell detected in the sample and each gene of the reference genome the number of detected transcripts expressed by the cell and assigned to the gene.

Note that the following conservative rules are applied for the pre-processing phase:

- Multimapping is not allowed (i.e. reads are considered as mapped only if they map to a maximum of 1 loci);
- After read alignment on the genome, only exonic reads are counted (i.e. intronic reads are not considered);
- Read assignment is done in the sense direction and multi-assignment is not allowed (i.e. if a read can be assigned to at least two distinct genes then the read is discarded).

When data pre-processing is completed, the data post-processing phase starts (also called “secondary analysis”) and takes as input the count matrix of each sample to process to perform the following successive steps:

1) **Filtering**: functions allowing some genes or cell to be excluded from the post processing analysis, by applying minimum or maximum threshold values to the number of cells expressing a gene and/or to number of genes or transcripts detected in a cell

2) **Selection of Highly Variable Genes (HVG) and normalizations**: functions which identify the top N genes of highest normalized variance (i.e. the genes that are the most representative of the gene expression variability in the cells), where N is given as input, and which normalize the expression of the genes in the detected cells

3) **Principal Component Analysis (PCA)**: function which applies a linear transformation on the normalized HVG signature of the cells to a new coordinate system, such that the linear combination of genes with highest variance lies on the 1st coordinate (called the 1st PCA dimension), the 2nd highest variance on the 2nd

coordinate (called the 2nd PCA dimension), etc., until a given number of PCA dimensions is reached, or until a given percentage of explained variance is reached

4) **Embedding**: function which applies a non-linear transformation of the PCA signature of the cells to project the cells as points onto a 2D plane (or a 3D volume) understandable by the human brain while preserving as much as possible the relative distances between the PCA signature of the cells

5) **Clustering**: function which associates each cell with a group of cells (called "cluster") by maximizing the distances between the PCA signatures of cells belonging to two different clusters and minimizing the distances between the PCA signatures of cells belonging to the same cluster, such that each cell cluster may be a relevant candidate for a biological cell population

6) **Differential Gene Expression (DGE)**: function which computes for each cell cluster and each HVG gene the statistical comparison of the normalized expression of the gene in the cells belonging to the cluster of interest compared to the cells belonging to the other clusters. The statistical comparison results include the z-score, the log fold change and the p-values (non-adjusted and adjusted), such that genes with adjusted p-value lower than 0.05 and positive (resp. negative) log fold change value can be considered as statistically over-expressed (resp. under-expressed) genes for the cluster of interest. These DGE results facilitates the manual annotation of cell populations.

Once data post-processing is completed, the Cytonaut™ Rover module enables the visualization of the cell clusters embedded in 2 or 3 dimensions as well as the interactive exploration of gene expression and cell attributes (e.g. ID of cell cluster, sample name, number of genes, number of transcripts, percentage of mitochondrial transcripts), allowing to access customized statistics (by applying specific filters) and to manually annotate cell populations with a high level of confidence.

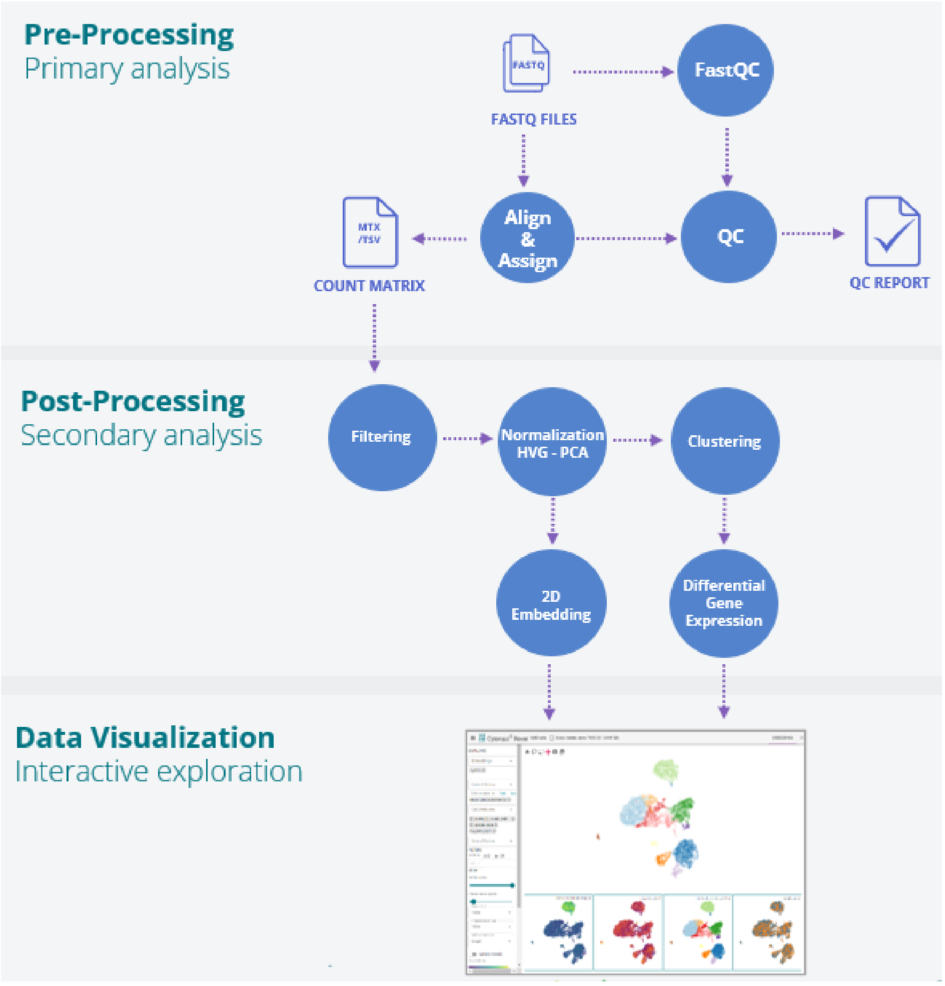


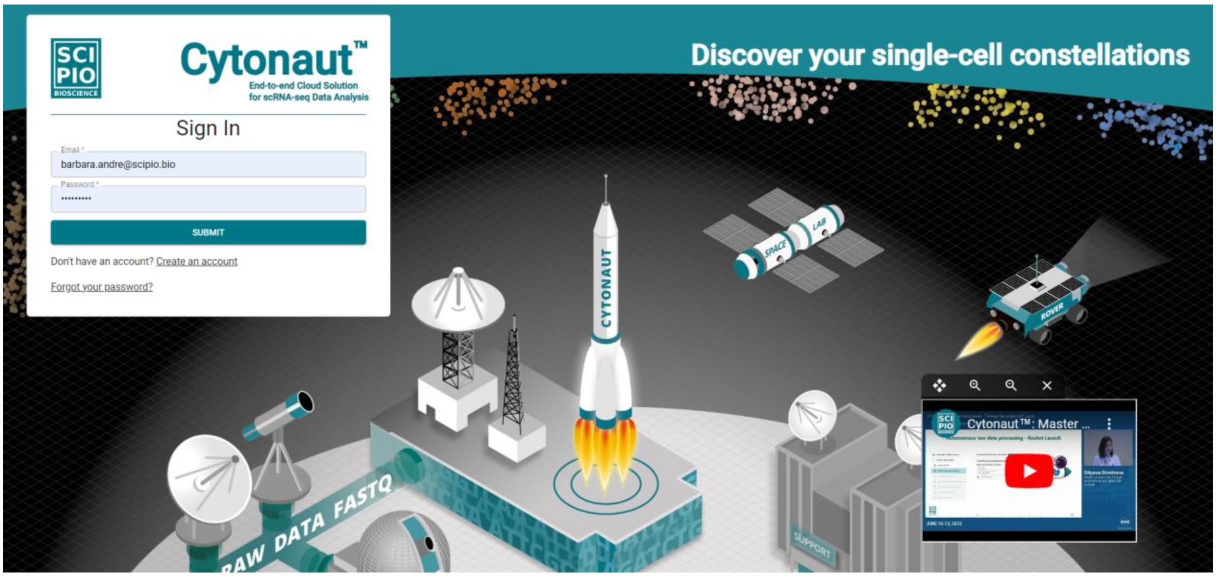
Figure 2. Flowchart of the end-to-end data analysis pipeline integrated in Cytonaut™.

II. CYTONAUT™ FUNCTIONALITIES - STEP BY STEP

1. Login to Cytonaut™ application

Using a typical computer connected to the external network, open your standard browser and go to https://www.cytonaut-scipio.bio.

- If you already have an account, sign in by simply entering your email and password:


Cytonaut™
End-to-end Cloud Solution for scRNA-seq Data Analysis
Discover your single-cell constellations
CYTONAUT
SPACE
LAB
ROVER
RAW DATA FASTQ
SUPPORT
Cytonaut™: Master ...

Sign In

Email *
barbara.andre@scipio.bio

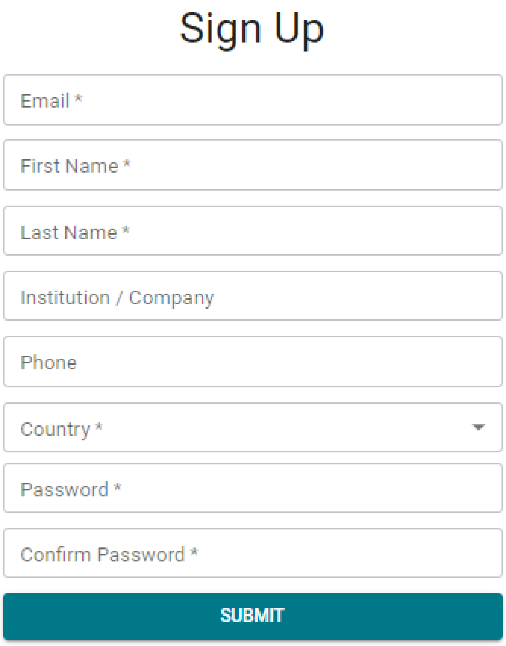
Password *
••••••••••

SUBMIT

Don't have an account? Create an account

Forgot your password?

- If you are not already registered, sign up by filling your email, first name, last name and password, then confirm your Cytonaut account by entering the code sent to your email. You will also be asked to enter your country in order to facilitate customer support.



Sign Up

Email *

First Name *

Last Name *

Institution / Company

Phone

Country *

Password *

Confirm Password *

SUBMIT

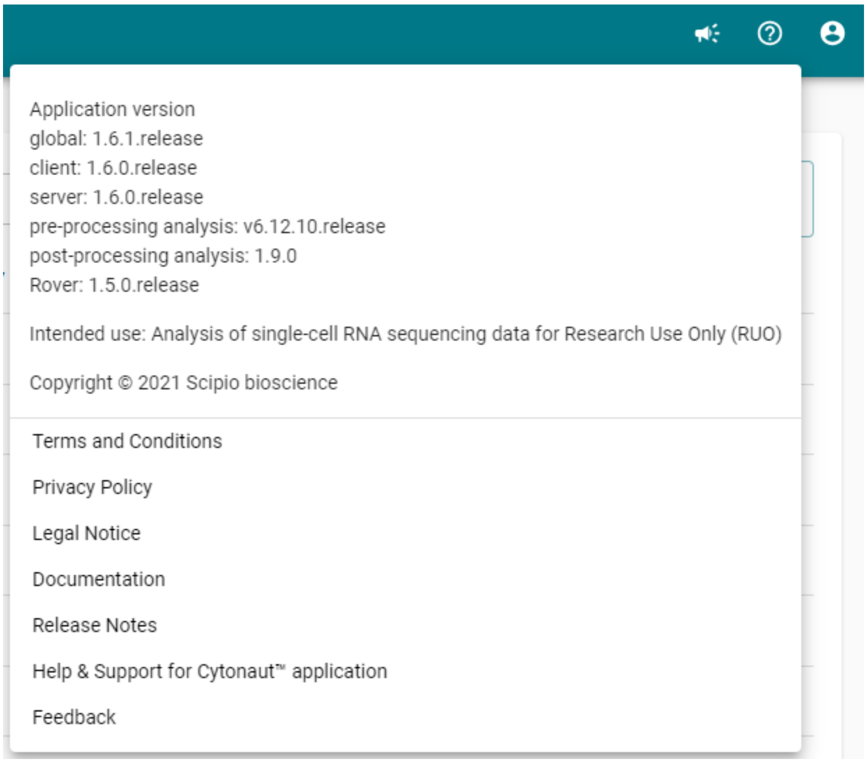
Notes:

- For security reasons, new user accounts are now automatically deleted if they are not confirmed by the user after 10 minutes, in which case a notification email inviting to register again is sent to the user.

- In order to use Cytonaut application, you will be asked to agree on the "Terms and Conditions" and on the "Privacy Policy" at first login, and whenever their content is updated.

- You can access the content of the Terms and Conditions or of the Privacy Policy anytime by clicking on the "question" icon ⓘ in the top banner

- By clicking on the "question" icon ⓘ in the top banner, you can also access the following: the Cytonaut™ version information (where "global" is the application version), the Legal Notice, the Release Notes, the Documentation (including the present User Manual), as well as information related to Help & Support and Feedback.

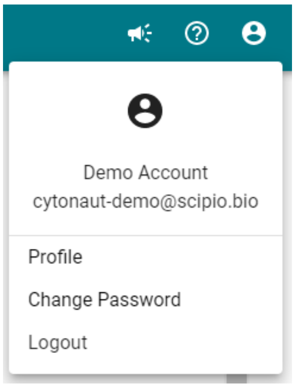


Application version
global: 1.6.1.release
client: 1.6.0.release
server: 1.6.0.release
pre-processing analysis: v6.12.10.release
post-processing analysis: 1.9.0
Rover: 1.5.0.release

Intended use: Analysis of single-cell RNA sequencing data for Research Use Only (RUO)

Copyright © 2021 Scipio bioscience

- You have the possibility to activate the Multi Factor Authentication (MFA) at each login by clicking on the "user" icon 👤 in the top banner and going to the "Profile" page, where you can fill your phone number, then enter the code sent to you by SMS, and finally check the box "Enable Multi Factor Authentication with your phone"



Demo Account
cytonaut-demo@scipio.bio

- You can change your password by clicking on the “user” icon in the top banner, then going to the “Change Password” page

- To logout, click on the “user” icon in the top banner, then click on “Logout”.

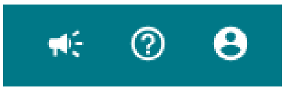
  Note that the login timeout is set to 24 hours.

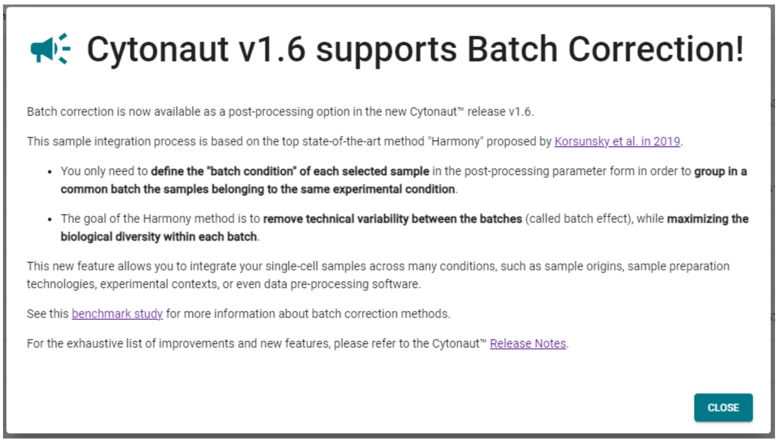
- In compliance with the GDPR law:

  - All retention logs are set to 15 days;

  - The Cytonaut application allows you to permanently delete your account, including you input data and output results, by typing “permanently delete” in the section “Delete my account” located at the bottom of the “Profile” page of the “user” menu .

- Whenever a new Cytonaut version is released, a notification popup window displayed at every new connection to Cytonaut, in order to provide a summary of the new features available. It is possible to deactivate this notification by clicking on “Don’t show again”, and to access it again on demand by clicking on the “loud speaker” icon in the top banner of Cytonaut.





Cytonaut v1.6 supports Batch Correction!

Batch correction is now available as a post-processing option in the new Cytonaut™ release v1.6.

This sample integration process is based on the top state-of-the-art method "Harmony" proposed by Korsunsky et al. in 2019.

- You only need to **define the "batch condition" of each selected sample** in the post-processing parameter form in order to **group in a common batch the samples belonging to the same experimental condition**.
- The goal of the Harmony method is to **remove technical variability between the batches** (called batch effect), while **maximizing the biological diversity within each batch**.

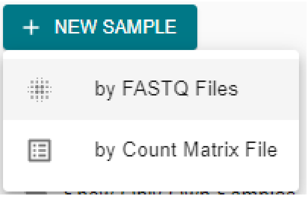
This new feature allows you to integrate your single-cell samples across many conditions, such as sample origins, sample preparation technologies, experimental contexts, or even data pre-processing software.

See this benchmark study for more information about batch correction methods.

For the exhaustive list of improvements and new features, please refer to the Cytonaut™ Release Notes.

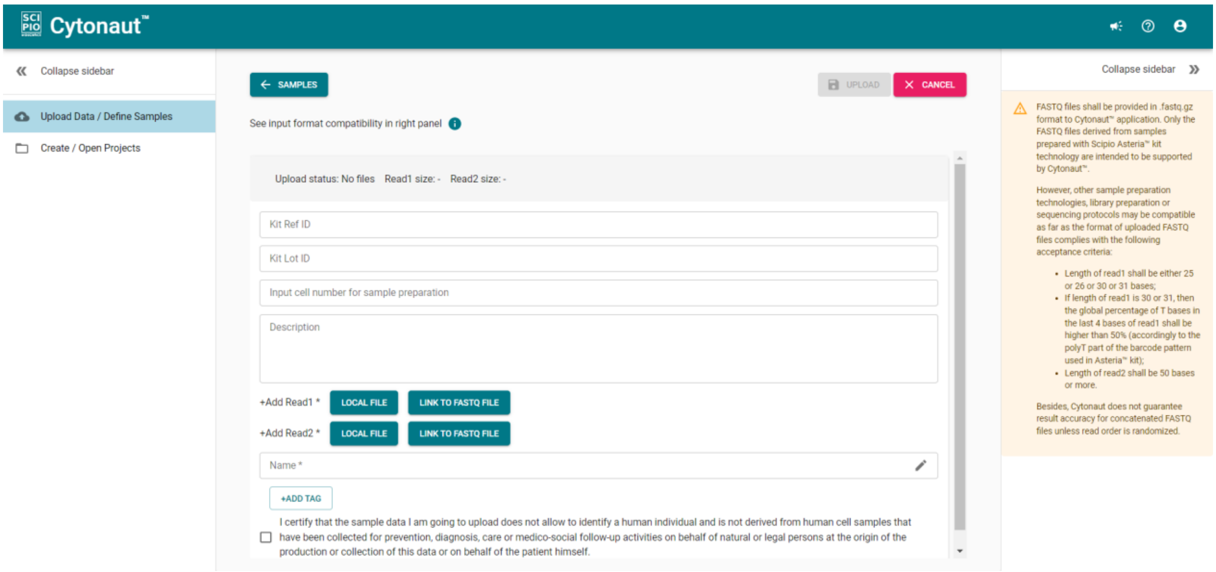
2. Upload FASTQ files to define your *Samples*

In the left panel, click on the menu Upload Data / Define Samples, then click on the button + NEW SAMPLE in the middle panel and select “by FASTQ Files” in the drop-down menu:



Note: it is also possible to upload a count matrix file to a define sample and directly perform post-processing analysis, as explained in the chapter “III. HOW TO UPLOAD COUNT MATRICES FOR DIRECT POST-PROCESSING”.

- Define a *Sample* by uploading the 2 FASTQ files provided as output of sequencing (R1 and R2 files in .fastq.gz format), either locally from your computer or via a link (a public HTTP or FTP link, or a S3 link provided at the discretion of Scipio bioscience).



See input format compatibility in right panel

Upload status: No files Read1 size: - Read2 size: -

□ I certify that the sample data I am going to upload does not allow to identify a human individual and is not derived from human cell samples that have been collected for prevention, diagnosis, care or medico-social follow-up activities on behalf of natural or legal persons at the origin of the production or collection of this data or on behalf of the patient himself.

FASTQ files shall be provided in .fastq.gz format to Cytonaut™ application. Only the FASTQ files derived from samples prepared with Scipio Asteria™ kit technology are intended to be supported by Cytonaut™.

However, other sample preparation technologies, library preparation or sequencing protocols may be compatible as far as the format of uploaded FASTQ files complies with the following acceptance criteria:

- Length of read1 shall be either 25 or 26 or 30 or 31 bases;
- If length of read1 is 30 or 31, then the global percentage of T bases in the last 4 bases of read1 shall be higher than 50% (accordingly to the polyT part of the barcode pattern used in Asteria™ kit);
- Length of read2 shall be 50 bases or more.

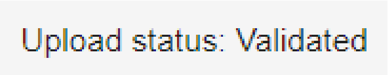
Besides, Cytonaut does not guarantee result accuracy for concatenated FASTQ files unless read order is randomized.

Once the FASTQ files are selected, a default name equal to the FASTQ filename is proposed for your sample. You may modify this default sample name, however it is better to keep it for traceability purposes.

- To upload these FASTQ files, you must previously check the box with the following statement:

  "I certify that the sample data I am going to upload does not allow to identify a human individual and is not derived from human cell samples that have been collected for prevention, diagnosis, care or medico-social follow-up activities on behalf of natural or legal persons at the origin of the production or collection of this data or on behalf of the patient himself."

- Click on UPLOAD and wait some seconds to minutes (depending on file size) until the end of the file upload and format verification process, when the sample upload status turns to "validated":



Upload status: Validated

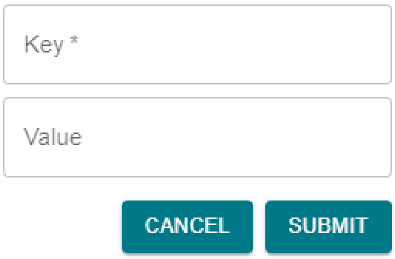
  Once your sample upload is done you will receive an email notification to the email address of your user account.

  During sample upload you have the possibility to create other samples or to perform other actions in parallel in Cytonaut application.

- Optional: you may enter some contextual information for your sample (e.g. "Kit Ref ID", "Kit Lot ID", "Input cell number for sample preparation", "Description" of sample, sample-level tags such as "ctrl:neg" for a negative control or "ctrl:pos" for positive control).

  This information is only indicative and not exploited for automated analysis; however it will be recorded in your exported data for traceability.

  To define a tag, enter a key (mandatory) and a value (optional):



Notes:

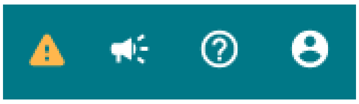
- Only the FASTQ files derived from samples prepared with the Asteria™ kit technology are intended to be supported by Cytonaut™.

  However, other sample preparation technologies, library preparation or sequencing protocols may be compatible as far as the format of uploaded FASTQ files complies with the following acceptance criteria:

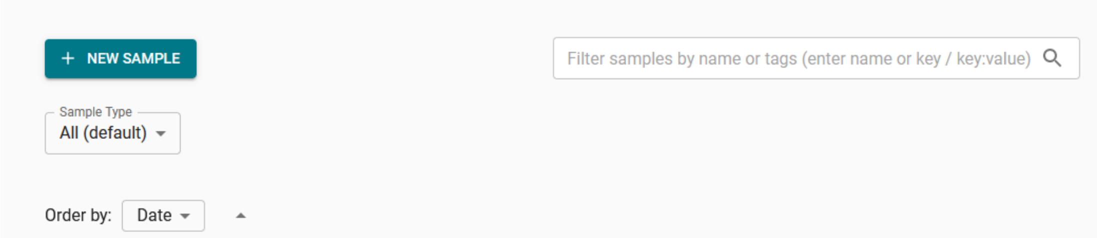
  - Length of read1 shall be either 25 or 26 or 30 or 31 bases;
  - If length of read1 is 30 or 31, then the global percentage of T bases in the last 4 bases of read1 shall be higher than 50% (accordingly to the polyT part of the barcode pattern used in Asteria™ kit);
  - Length of read2 shall be 50 bases or more.

- Cytonaut does not guarantee result accuracy for concatenated FASTQ files unless read order is randomized.

- If you are performing a local file upload from your computer, you shall not close your browser until the end of sample upload, as indicated by this warning icon in top banner:



- You can filter your existing samples by sample type (“All” by default, “FASTQ”, “Count Matrix”), or by sample name or sample tags in the search bar (by entering complete or partial name or tags). You can also sort the samples by the last update date of the sample (applied by default) or by the sample name.


+ NEW SAMPLE
Filter samples by name or tags (enter name or key / key:value)

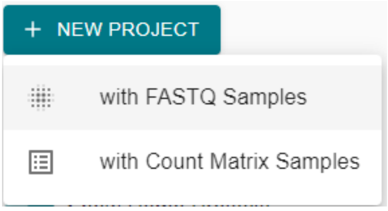
Sample Type: All (default)

Order by: Date

- The list of supported characters for sample name, project name, and post-processing run ID is the following: "0-9", "a-Z", "A-Z", "-", ".", "_"

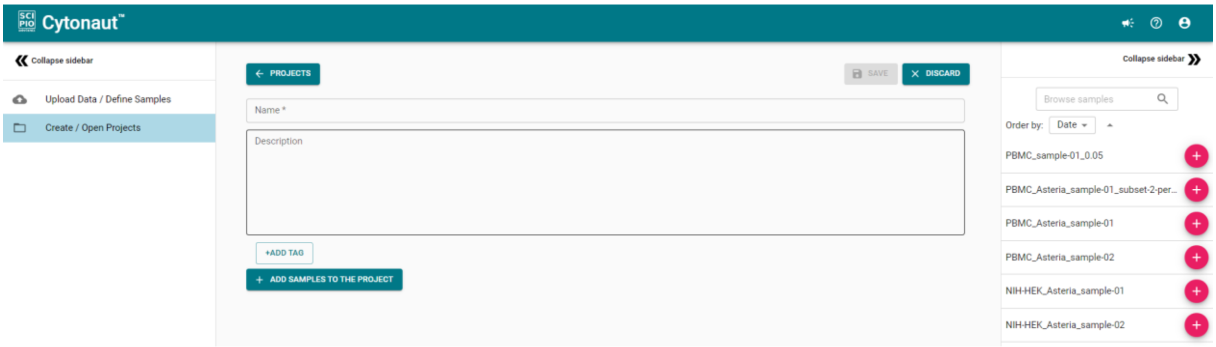
3. Create a *Project* of *Samples* to be analyzed

In the left panel, click on the menu Create / Open Projects, then click on the button + NEW PROJECT in the middle panel, and select "with FASTQ Samples" in the drop-down menu:



1) Define your *Project* as a set of *Samples* to be analyzed with same parameters (e.g. the same pre-processing parameters and by default the same post-processing parameters but with the freedom to integrate or not the samples in the same post-processing run):

   - Enter a unique name for your *Project*
   - Click on the button + ADD SAMPLES TO THE PROJECT, then search samples in the right panel and click on the icon + to add a *Sample* to your project



2) Optional: enter contextual information for your *Project* (e.g. project description, project-level tags such as "application:brain" for an application on the brain)

   This indicative information is not exploited for automated analysis, however it will be recorded in your exported data for traceability.

3) Click on the button SAVE to finalize the creation of your *Project*. This will automatically set the graphical interface in the referential of your project, with the name of the project written next to "Current project:" in the top banner.

Notes:

- A given *Sample* may be added to multiple *Projects* in order to be analyzed in different ways.

- *Projects* and *Samples* are searchable by names, dates and tags.

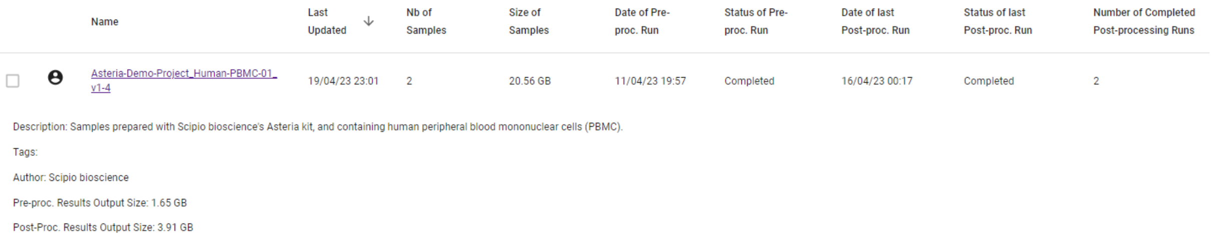
- An already existing *Project* can be selected to be opened in order to access results already generated for the project or to re-analyze all or parts of the project.

- *Demo Projects* are provided by Scipio bioscience in read-only mode: you can open them to explore or download their results in Cytonaut™ and Cytonaut™ Rover, however you cannot modify them neither their samples (i.e. no possibility to edit, re-analyze or delete). These demo projects as well as their demo samples are identifiable thanks to the icon 📖 displayed in front of their name.

- It is possible to sort projects according to each column of the project table shown in the menu "Create / Open Projects", where the sortable columns are displayed in this order from left to right for each project:

  - Name
  - Last Updated
  - Nb of Samples
  - Size of Samples
  - Date of Pre-proc. Run
  - Status of Pre-proc. Run (with completion percentage if running)
  - Date of last Post-proc. Run
  - Status of last Post-proc. Run (with completion percentage if running)
  - Nb of Completed Post-proc. Runs

- It is possible to see more details about a project without opening it by clicking on the icon on left side of project name:



		Name	Last Updated ↓	Nb of Samples	Size of Samples	Date of Pre-proc. Run	Status of Pre-proc. Run	Date of last Post-proc. Run	Status of last Post-proc. Run	Number of Completed Post-processing Runs
□	●	Asteria-Demo-Project_Human-PBMC-01_v1-4	19/04/23 23:01	2	20.56 GB	11/04/23 19:57	Completed	16/04/23 00:17	Completed	2

Description: Samples prepared with Scipio bioscience's Asteria kit, and containing human peripheral blood mononuclear cells (PBMC).

Tags:

Author: Scipio bioscience

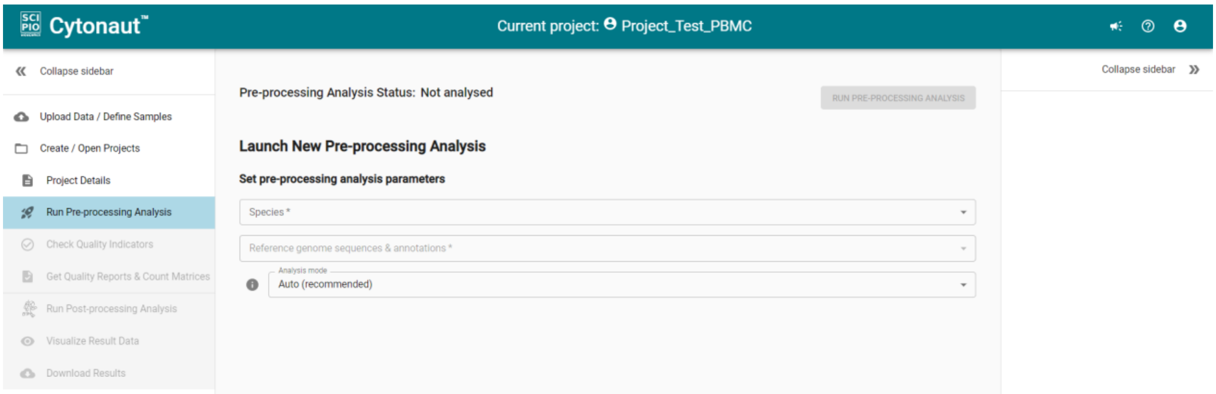
Pre-proc. Results Output Size: 1.65 GB

Post-Proc. Results Output Size: 3.91 GB

4. Launch Pre-processing Analysis on your Project

If the *Project* is not already opened: click on the menu Create / Open Projects in the left panel, then select the *Project* to be analyzed by clicking on its name.

1) In the left panel, click on the menu Run Pre-processing Analysis


Cytonaut™
Current project: Project_Test_PBMC
Collapse sidebar
Upload Data / Define Samples
Create / Open Projects
Project Details
Run Pre-processing Analysis
Check Quality Indicators
Get Quality Reports & Count Matrices
Run Post-processing Analysis
Visualize Result Data
Download Results
RUN PRE-PROCESSING ANALYSIS
Collapse sidebar

Pre-processing Analysis Status: Not analysed

Launch New Pre-processing Analysis

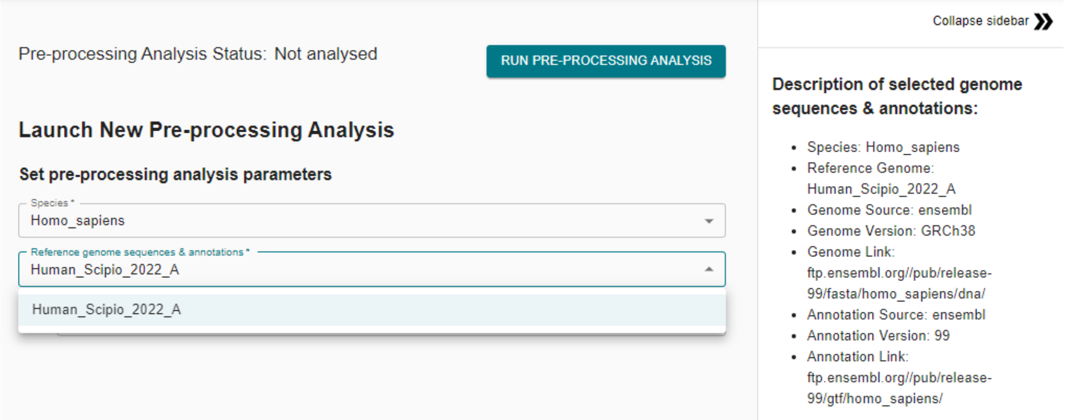
Set pre-processing analysis parameters

Species *

Reference genome sequences & annotations *

Analysis mode: Auto (recommended)

2) Select the “Species” and the genome applicable to the samples of your project in the drop-down menu, and check the details of the associated “Reference genome sequences & annotations” in the right panel:



Pre-processing Analysis Status: Not analysed

Launch New Pre-processing Analysis

Set pre-processing analysis parameters

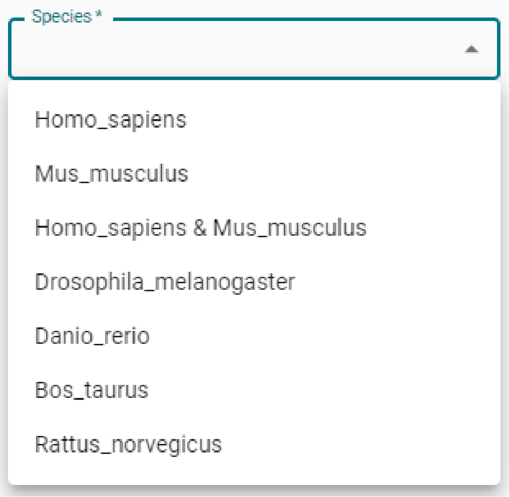
Species *: Homo_sapiens

Reference genome sequences & annotations *: Human_Scipio_2022_A

Description of selected genome sequences & annotations:

- Species: Homo_sapiens
- Reference Genome: Human_Scipio_2022_A
- Genome Source: ensembl
- Genome Version: GRCh38
- Genome Link: ftp.ensembl.org//pub/release-99/fasta/homo_sapiens/dna/
- Annotation Source: ensembl
- Annotation Version: 99
- Annotation Link: ftp.ensembl.org//pub/release-99/gtf/homo_sapiens/

The species and reference genomes currently supported by Cytonaut are listed below:



- **Homo sapiens (Human)**
  - Species: Homo_sapiens
  - Reference Genome: Human_Scipio_2022_A
  - Genome Source: ensembl
  - Genome Version: GRCh38
  - Genome Link: ftp.ensembl.org//pub/release-99/fasta/homo_sapiens/dna/
  - Annotation Source: ensembl
  - Annotation Version: 99
  - Annotation Link: ftp.ensembl.org//pub/release-99/gtf/homo_sapiens/
- **Mus musculus (Mouse)**
  - Species: Mus_musculus
  - Reference Genome: Mouse_Scipio_2022_A
  - Genome Source: ensembl
  - Genome Version: GRCm38
  - Genome Link: ftp.ensembl.org//pub/release-99/fasta/mus_musculus/dna/
  - Annotation Source: ensembl
  - Annotation Version: 99
  - Annotation Link: ftp.ensembl.org//pub/release-99/gtf/mus_musculus/
- **Humo sapiens & Mus musculus (Human & Mouse)**
  - Species: Homo_sapiens & Mus_musculus
  - Reference Genome: Human_Mouse_Scipio_2022_A
  - Genome Source: ensembl & ensembl
  - Genome Version: GRCh38 & GRCm38
  - Genome Link: ftp.ensembl.org//pub/release-99/fasta/homo_sapiens/dna/ & ftp.ensembl.org//pub/release-99/fasta/mus_musculus/dna/
  - Annotation Source: ensembl & ensembl
  - Annotation Version: 99 & 99
  - Annotation Link: ftp.ensembl.org//pub/release-99/gtf/homo_sapiens/ & ftp.ensembl.org//pub/release-99/gtf/mus_musculus/

- **Drosophila melanogaster**
  - Species: Drosophila_melanogaster
  - Reference Genome: Drosophila_Scipio_2022_A
  - Genome Source: ensembl
  - Genome Version: BDGP6.32
  - Genome Link: ftp.ensembl.org/pub/release-99/fasta/drosophila_melanogaster/dna/
  - Annotation Source: ensembl
  - Annotation Version: 106
  - Annotation Link: ftp.ensembl.org/pub/release-106/gtf/drosophila_melanogaster/

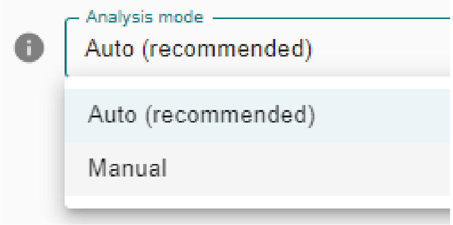
- **Danio rerio (Zebrafish)**
  - Species: Danio_rerio
  - Reference Genome: Zebrafish_Scipio_2022_A
  - Genome Source: ensembl
  - Genome Version: GRCz11
  - Genome Link: ftp.ensembl.org/pub/release-106/fasta/danio_rerio/dna/
  - Annotation Source: ensembl
  - Annotation Version: 106
  - Annotation Link: ftp.ensembl.org/pub/release-106/gtf/danio_rerio/

- **Bos taurus (Bovine)**
  - Species: Bos_taurus
  - Reference Genome: Bos_taurus_Scipio_2022_A
  - Genome Source: ncbi
  - Genome Version: ARS-UCD1.3
  - Genome Link: https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_002263795.2/
  - Annotation Source: ncbi
  - Version Annotation: 106
  - Annotation Link: https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_002263795.2/

- **Rattus norvegicus (Rat)**
  - Species: Rattus_norvegicus
  - Reference Genome: Rat_Scipio_2023_A
  - Genome Source: ensembl
  - Genome Version: Rnor_6.0
  - Genome Link: ftp.ensembl.org//pub/release-104/fasta/rattus_norvegicus/dna/
  - Annotation Source: ensembl
  - Version Annotation: 104
  - Annotation Link: ftp.ensembl.org//pub/release-104/gtf/rattus_norvegicus/

3) Keep the “Auto” analysis mode for the first time.

By default the method used to detect cell-associated barcodes in the sequencing data of each sample is set to “Auto” mode. It is highly recommended to keep the default "Auto" value for the pre-processing run, because this ensures objective detection by the automated method which retrieves the top N barcodes ranked in decreasing unique read counts until the first knee point.



However, if the pre-processing results which are then accessible in the menu Check Quality Indicators are not satisfying for one or several samples (in particular regarding the “Knee Plot” explained in chapter *“IV. HOW TO UNDERSTAND QUALITY INDICATORS”*), then it is possible to re-launch the pre-processing run in manual mode (see the chapter *“V. HOW TO RE-DO PRE-PROCESSING IN MANUAL MODE"*).

4) Launch the *Pre-processing Analysis* run by clicking on the button RUN PRE-PROCESSING ANALYSIS

5) Wait ~ 3 to 12 hours depending on sample sequencing depth and number of sample input cells (note that it is rather independent of the number of samples included in the project as their pre-processing analysis is parallelized).

A progress bar indicates the percentage of completed steps, where some steps can be much longer as others.

During the pre-processing you have the possibility to perform other actions in parallel, such as pre-processing runs on other projects.

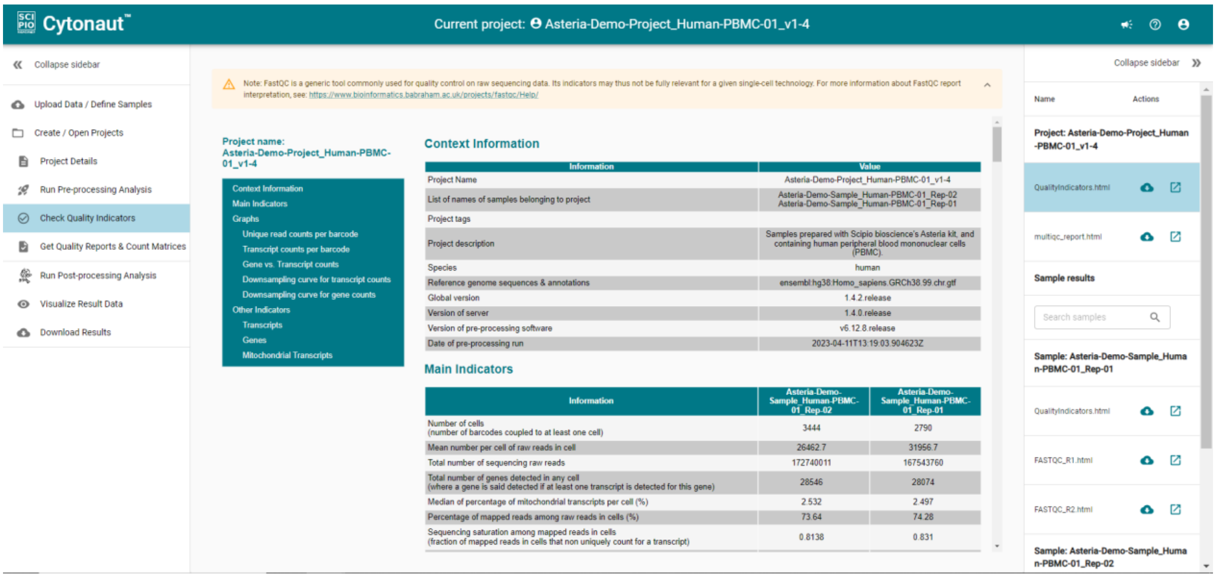
Once your project pre-processing is done you will receive an email notification to the email address of your user account, with a first link to directly access the computed quality indicators, and a second link to download them as well as sample count matrices.

Note: If needed, it is possible to cancel an ongoing pre-processing run by clicking on CANCEL ANALYSIS in the menu Run Pre-processing Analysis.

5. Check quality indicators of your pre-processed *Samples*

If the *Project* to be checked is not already opened: click on the menu Create / Open Projects in the left panel, then select the *Project* of interest by clicking on it.

1) In the left panel, click on the menu Check Quality Indicators, you will see in the middle panel the project-level Quality Indicator report "QualityIndicators.html" which aggregates the indicators of all the samples of the project:


Cytonaut™
Current project: Asteria-Demo-Project_Human-PBMC-01_v1-4
Collapse sidebar
Upload Data / Define Samples
Create / Open Projects
Project Details
Run Pre-processing Analysis
Check Quality Indicators
Get Quality Reports & Count Matrices
Run Post-processing Analysis
Visualize Result Data
Download Results
Project name: Asteria-Demo-Project_Human-PBMC-01_v1-4
Context Information
Main Indicators
Graphs
Unique read counts per barcode
Transcript counts per barcode
Gene vs. Transcript counts
Downsampling curve for transcript counts
Downsampling curve for gene counts
Other Indicators
Transcripts
Genes
Mitochondrial Transcripts
Collapse sidebar
Name
Actions
Project: Asteria-Demo-Project_Human-PBMC-01_v1-4
QualityIndicators.html
multiqc_report.html
Sample results
Search samples
Sample: Asteria-Demo-Sample_Human-PBMC-01_Rep-01
QualityIndicators.html
FASTQC_R1.html
FASTQC_R2.html
Sample: Asteria-Demo-Sample_Human-PBMC-01_Rep-02

Note: FastQC is a generic tool commonly used for quality control on raw sequencing data. Its indicators may thus not be fully relevant for a given single-cell technology. For more information about FastQC report interpretation, see: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

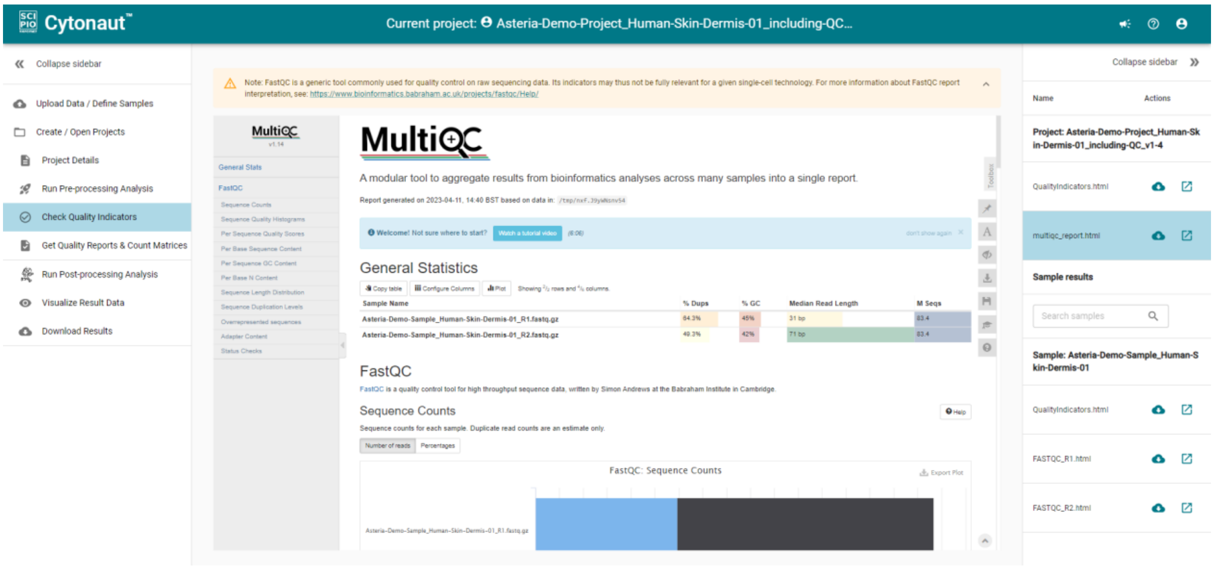
Context Information

Information	Value
Project Name	Asteria-Demo-Project_Human-PBMC-01_v1-4
List of names of samples belonging to project	Asteria-Demo-Sample_Human-PBMC-01_Rep-02 Asteria-Demo-Sample_Human-PBMC-01_Rep-01
Project tags	
Project description	Samples prepared with Scipio bioscience's Asteria kit, and containing human peripheral blood mononuclear cells (PBMC).
Species	human
Reference genome sequences & annotations	ensembl:hg38:Homo_sapiens.GRCh38.99.chr.gtf
Global version	1.4.2.release
Version of server	1.4.0.release
Version of pre-processing software	v6.12.8.release
Date of pre-processing run	2023-04-11T13:19:03.904623Z

Main Indicators

Information	Asteria-Demo-Sample_Human-PBMC-01_Rep-02	Asteria-Demo-Sample_Human-PBMC-01_Rep-01
Number of cells (number of barcodes coupled to at least one cell)	3444	2790
Mean number per cell of raw reads in cell	26462.7	31956.7
Total number of sequencing raw reads	172740011	167543760
Total number of genes detected in any cell (where a gene is said detected if at least one transcript is detected for this gene)	28546	28074
Median of percentage of mitochondrial transcripts per cell (%)	2.532	2.497
Percentage of mapped reads among raw reads in cells (%)	73.64	74.28
Sequencing saturation among mapped reads in cells (fraction of mapped reads in cells that non uniquely count for a transcript)	0.8138	0.831

You can also access the low-level Sequencing Quality report "multiqc_report.html" which aggregates the FastQC information of all the samples of the project:


Cytonaut™
Current project: Asteria-Demo-Project_Human-Skin-Dermis-01_including-QC...
Collapse sidebar
Upload Data / Define Samples
Create / Open Projects
Project Details
Run Pre-processing Analysis
Check Quality Indicators
Get Quality Reports & Count Matrices
Run Post-processing Analysis
Visualize Result Data
Download Results
MultiQC
v1.14
General Stats
FastQC
Sequence Counts
Sequence Quality Histograms
Per Sequence Quality Scores
Per Base Sequence Content
Per Sequence GC Content
Per Base N Content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
Status Checks
Toolbox
Copy table
Configure Columns
Plot
Number of reads
Percentages
FastQC: Sequence Counts
Export Plot
Asteria-Demo-Sample_Human-Skin-Dermis-01_R1.fastq.gz
Collapse sidebar
Name
Actions
Project: Asteria-Demo-Project_Human-Skin-Dermis-01_including-QC_v1-4
QualityIndicators.html
multiqc_report.html
Sample results
Search samples
Sample: Asteria-Demo-Sample_Human-Skin-Dermis-01
QualityIndicators.html
FASTQC_R1.html
FASTQC_R2.html

Note: FastQC is a generic tool commonly used for quality control on raw sequencing data. Its indicators may thus not be fully relevant for a given single-cell technology. For more information about FastQC report interpretation, see: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2023-04-11, 14:40 BST based on data in: `/tmp/nxF.J9yWNsnv54`

Welcome! Not sure where to start? Watch a tutorial video (6:06) don't show again

General Statistics

Showing 2/2 rows and 4/4 columns.

Sample Name	% Dups	% GC	Median Read Length	M Seqs
Asteria-Demo-Sample_Human-Skin-Dermis-01_R1.fastq.gz	84.3%	45%	31 bp	83.4
Asteria-Demo-Sample_Human-Skin-Dermis-01_R2.fastq.gz	49.3%	42%	71 bp	83.4

FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Counts

Sequence counts for each sample. Duplicate read counts are an estimate only.

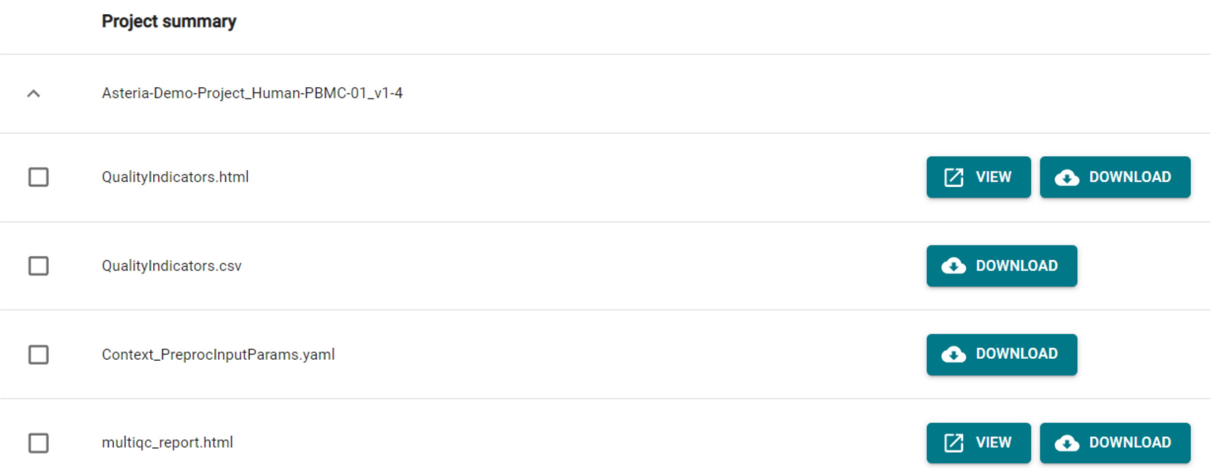
2) In the right panel you can select for each *Sample* of your *Project* the following files, in order to access the sample-level reports of each sample:

- the file “QualityIndicators.html” at the level of the sample, which recalls the context of the pre-processing analysis and provides the values of each quality indicator (see the chapter *“IV. HOW TO UNDERSTAND QUALITY INDICATORS”* for more details) as well as associated graph representations

- the files “FASTQC_R1.html” and “FASTQC_R2.html” which contain at the level of the sample the low-level quality indicators associated with the sequencing of R1 raw reads and R2 raw reads according to the FastQC tool

3) Optional: in order to download all or parts of the following files, click in the left panel on the menu Get Quality Reports & Count Matrices (click on “Select all” to download all of them at once) so you can export:

- At the project level:
  - the text file “Context_PreprocInputParams.yaml”, which contains all the contextual information, email of the author of the project, software version, UTC times of project creation and of pre-processing run start, and all pre-processing parameter values used for the project
  - the files “QualityIndicators.html” and “QualityIndicators.csv” at project level (table format with comma separated values)
  - the file “multiqc_report.html” which is an aggregation of the FastQC reports of all samples belonging to the project



Project summary

Asteria-Demo-Project_Human-PBMC-01_v1-4

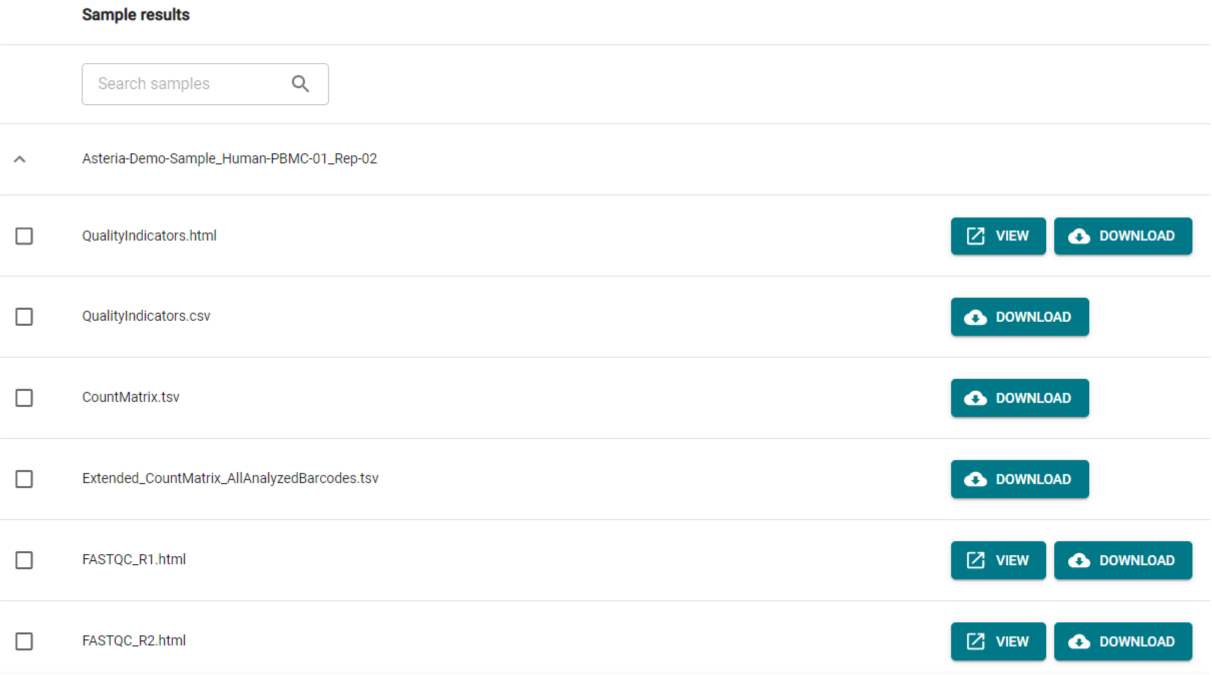
□	QualityIndicators.html	VIEW DOWNLOAD
□	QualityIndicators.csv	DOWNLOAD
□	Context_PreprocInputParams.yaml	DOWNLOAD
□	multiqc_report.html	VIEW DOWNLOAD

- At the sample level:
  - the files “QualityIndicators.html” and “QualityIndicators.csv” at sample level (table format with comma separated values)

- the files “FASTQC_R1.html” and “FASTQC_R2.html”

- the file “CountMatrix.tsv” (in .tsv format with tab separated values), which is the sample count matrix containing for each detected cell and each gene the number of detected transcripts expressed in the cell and assigned to the gene

- the file “Extended_CountMatrix_AllAnalyzedBarcodes.tsv” (in tsv format with tab separated values), which is the extension of the matrix “CountMatrix.tsv” to 4 times more aligned barcodes, including the detected cell-associated barcodes and the next barcodes in decreasing number of unique R1 reads.


Sample results

Search samples

Asteria-Demo-Sample_Human-PBMC-01_Rep-02

□ QualityIndicators.html

□ QualityIndicators.csv

□ CountMatrix.tsv

□ Extended_CountMatrix_AllAnalyzedBarcodes.tsv

□ FASTQC_R1.html

□ FASTQC_R2.html

Notes:

- For each file listed in the right panel, you may click on the icon ⬇ to download the file, or on the icon ↗ to open the file in a new tab (which is practical if you want to see multiple files simultaneously)

- In the count matrices generated by Cytonaut, each gene is named <GeneAlias>:<GeneID> (e.g. “MS4A1:ENSG00000156738”), where <GeneAlias> is the symbol of the gene and <GeneID> is the unique identifier of the gene in the reference genome annotations (e.g. its Ensembl identifier if the source of the reference genome is Ensembl).

- FastQC is a generic tool commonly used for quality control on raw sequencing data. Its indicators may thus not be fully relevant for a given single-cell technology. For more information about FastQC report interpretation, see: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

- The .yaml files can be opened using a simple text editor; the .csv files can be opened using Excel by specifying that separator is "," (comma) and floating value is "." (point).

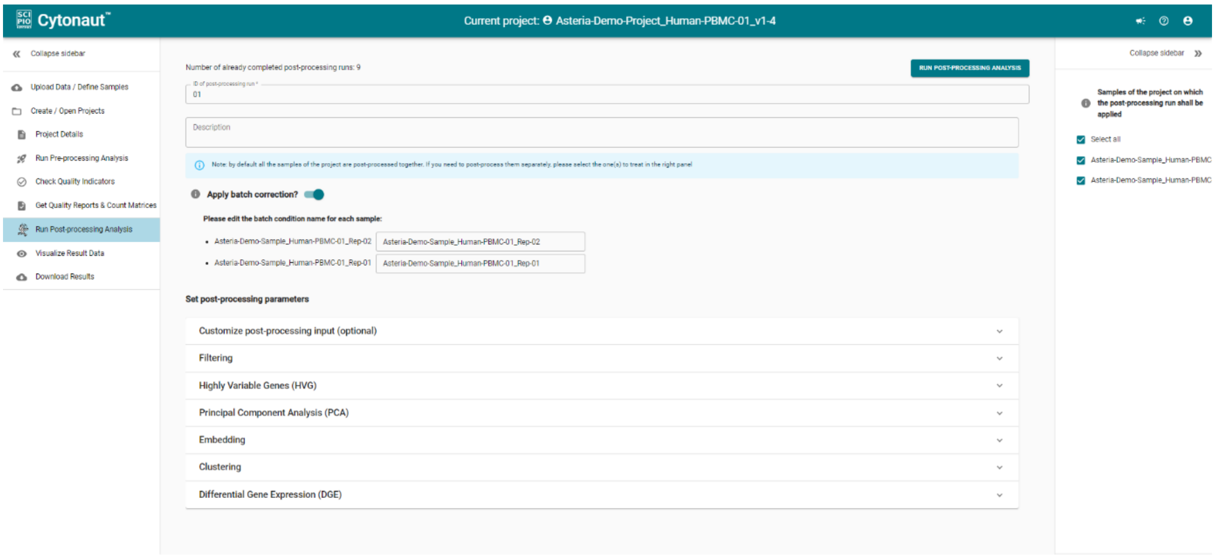
- Using the extended count matrix instead of the standard count matrix as input of post-processing analysis could allow to better detect cells of lower expression using more advanced filtering. To this purpose, select the option "Use extended count matrix" in the section "Customize post-processing (optional)".

- Note however that Cytonaut does not guarantee result accuracy if extended count matrices are used as input of post-processing analysis. Indeed, in most cases, most of the additional barcodes present in the extended count matrix are not cell-associated barcodes but barcodes of beads which have captured noise.

6. Launch a Post-processing Run on your Project

If the *Project* is not already opened: click on the menu Create / Open Projects in the left panel, then select the *Project* to be analyzed by clicking on it.

1) In the left panel, click on the menu Run Post-processing Analysis


Cytonaut™
Collapse sidebar
Upload Data / Define Samples
Create / Open Projects
Project Details
Run Pre-processing Analysis
Check Quality Indicators
Get Quality Reports & Count Matrices
Run Post-processing Analysis
Visualize Result Data
Download Results
RUN POST-PROCESSING ANALYSIS
Collapse sidebar
Customize post-processing input (optional)
Filtering
Highly Variable Genes (HVG)
Principal Component Analysis (PCA)
Embedding
Clustering
Differential Gene Expression (DGE)

Current project: Asteria-Demo-Project_Human-PBMC-01_v1-4

Number of already completed post-processing runs: 9

ID of post-processing run *: 01

Description

Note: by default all the samples of the project are post-processed together. If you need to post-process them separately, please select the one(s) to treat in the right panel

Apply batch correction? ●

Please edit the batch condition name for each sample:

- Asteria-Demo-Sample_Human-PBMC-01_Rep-02: Asteria-Demo-Sample_Human-PBMC-01_Rep-02
- Asteria-Demo-Sample_Human-PBMC-01_Rep-01: Asteria-Demo-Sample_Human-PBMC-01_Rep-01

Set post-processing parameters

Samples of the project on which the post-processing run shall be applied

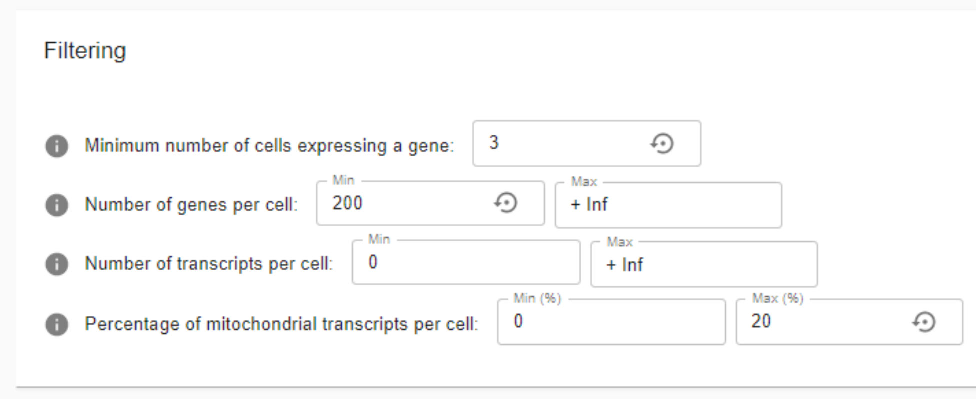
☑ Select all
☑ Asteria-Demo-Sample_Human-PBMC
☑ Asteria-Demo-Sample_Human-PBMC

1) Enter a unique identifier (Run ID) for the *Post-processing Run* to be launched on your *Project* (a default Run ID is proposed but you may change or extend it with any word to make it more explicit), and an optional description for your run

2) In the right panel, select the samples of the project that you want to integrate together in your post-processing run (by default all samples are checked); then, if you have at least two samples selected and want to apply batch correction (activated by default), enter the batch condition name of each selected sample in order to group in a common batch the samples with the same experimental conditions (see the chapter *"VII. HOW TO CONFIGURE BATCH CORRECTION"* for more details).

3) Set the parameters of your *Post-processing Run* for each expandable parameter category, by keeping their default values or changing them according to your application, in particular for the first parameters in the category "Filtering" (see the chapter *"VI. HOW TO SET POST-PROCESSING PARAMETERS"* for more details).

- Most default parameter values are relevant for a first post-processing run, except for the parameters belonging to the category "Filtering" because their optimal values depend on the application.
  It is this recommended to adapt the filtering parameter values, which are permissive by default, to the application of your project, for example as follows:



Filtering

Minimum number of cells expressing a gene: 3

Number of genes per cell: Min 200 Max + Inf

Number of transcripts per cell: Min 0 Max + Inf

Percentage of mitochondrial transcripts per cell: Min (%) 0 Max (%) 20

- To know more about the definition a given post-processing parameter and how to set its value, click on the "information" icon ⓘ next to the parameter and see the contextual help displayed in the right panel.

- Optionally, you can do more advanced parametrization by expanding the category "Customize post-processing input (optional)" which allows the following actions: UPLOAD CONFIGURATION and UPLOAD SUBSET OF CELL BARCODES (see the chapter *"VI. HOW TO SET POST-PROCESSING PARAMETERS"* for more details).

4) Launch the *Post-processing Run* by clicking on RUN POST-PROCESSING ANALYSIS

5) Wait ~ 5 min to 30 minutes depending on the number of *Samples* of the *Project* that are integrated in the post-processing run (a progress bar indicates the percentage of completed steps).

During the post-processing you have the possibility to perform other actions in parallel.

Once your post-processing run is done you will receive an email notification to the email address of your user account, with a first link to visually explore the post-processed results with Cytonaut™ Rover, and a second link to download the results.

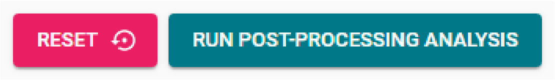
Notes:

- Several *Post-processing Runs* can be sequentially launched on a given *Project* in order to test different parameter values, to integrate different subsets of samples, or to post-process each sample individually (i.e. with one post-processing run per sample)

- If multiple samples of your project are selected to be integrated in the same post-processing run, their count matrices will be merged before running the post-processing, so these samples will not be analyzed independently. However, once the post-processing run is completed, it is possible in the

Cytonaut™ Rover module to identify which cell belongs to which samples and to filter cells by “sample_name”, which may be useful in particular to detect outlier samples or batch effect (in which case it is recommended to post-process the samples separately or to apply batch correction)

- Whenever the default value of at least one post-processing parameter is manually modified, a global RESET button is displayed in magenta color on top of the page, in order to give the possibility to reset all parameter values to their default values, and a small revert icon is displayed next to each modified value to allow resetting the specific parameter value:



- Explicit error messages are displayed in case where a post-processing run has failed, for example if too strict post-processing parameter values result in empty count matrix after filtering.

- If needed, it is possible to cancel an ongoing post-processing run by clicking on STOP ANALYSIS in the menu Run Post-processing Analysis.

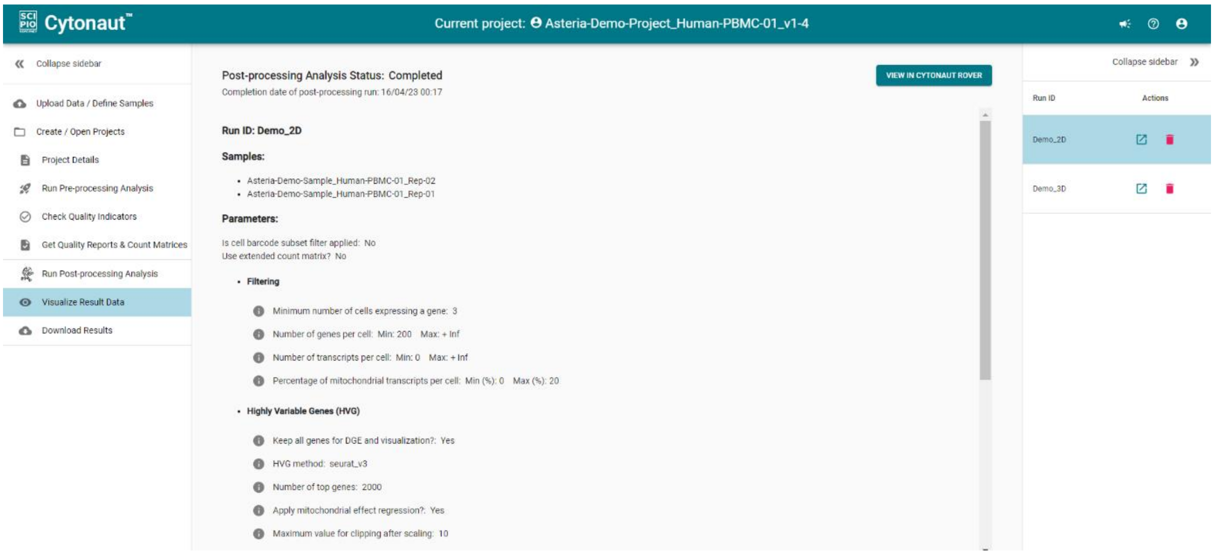
7. Visualize and annotate result data of your *Project Run* with Cytonaut™ Rover

If the *Project* is not already opened: click on the menu Create / Open Projects in the left panel, then select the *Project* to be checked by clicking on it.

1) In the left panel, click on the menu Visualize Result Data

2) In the right panel, select the *Run ID* to be displayed in the middle panel (by default the last run is displayed)

All parameter values which have been set for the run are recalled in the middle panel, as well as the samples to which the run has been applied.



Post-processing Analysis Status: Completed
Completion date of post-processing run: 16/04/23 00:17

Run ID: Demo_2D

Samples:

- Asteria-Demo-Sample_Human-PBMC-01_Rep-02
- Asteria-Demo-Sample_Human-PBMC-01_Rep-01

Parameters:

Is cell barcode subset filter applied: No
Use extended count matrix? No

- Filtering
  - Minimum number of cells expressing a gene: 3
  - Number of genes per cell: Min: 200 Max: + Inf
  - Number of transcripts per cell: Min: 0 Max: + Inf
  - Percentage of mitochondrial transcripts per cell: Min (%): 0 Max (%): 20
- Highly Variable Genes (HVG)
  - Keep all genes for DGE and visualization?: Yes
  - HVG method: seurat_v3
  - Number of top genes: 2000
  - Apply mitochondrial effect regression?: Yes
  - Maximum value for clipping after scaling: 10

3) Click on VIEW IN CYTONAUT ROVER in the middle panel (or on the “open in new tab” icon next to the Run ID in the right panel) to open a new tab with the Cytonaut™ Rover module applied to the post-processed results of your run.

4) The Cytonaut™ Rover page displays on the top banner the name of the project as well as the Run ID in parenthesis.

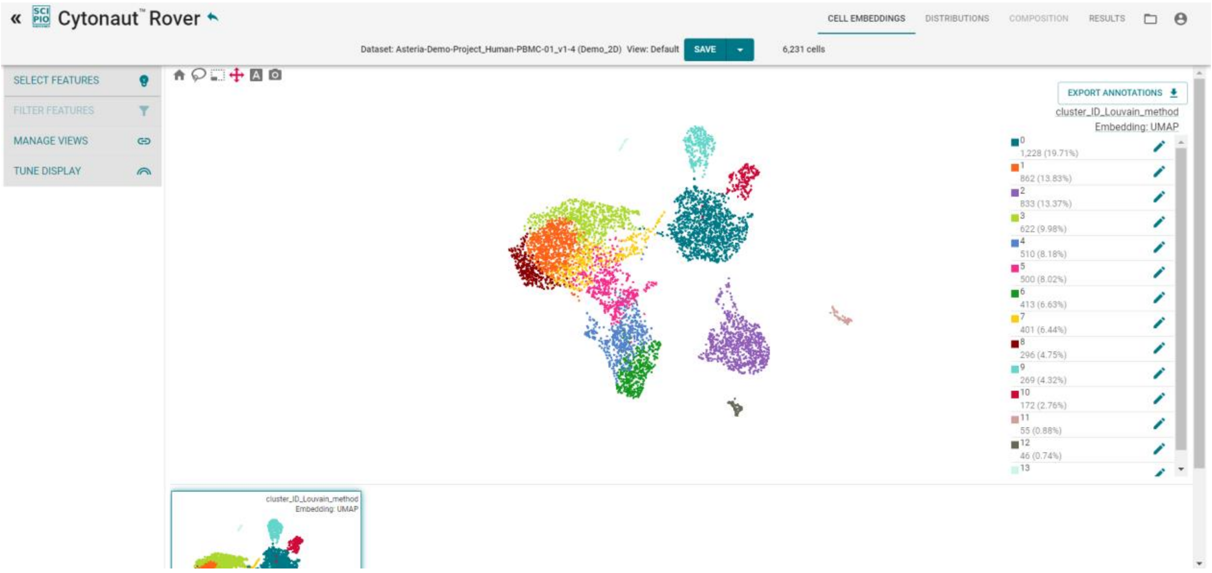
The graphical interface of Cytonaut™ Rover is organized by menus in the top banner:

- CELL EMBEDDINGS
- DISTRIBUTIONS
- COMPOSITION
- RESULTS

and by interactive sections in the left panel:

- SELECT FEATURES
- FILTER FEATURES
- MANAGE VIEWS
- TUNE DISPLAY

By default the page is always opened in the CELL EMBEDDINGS menu and the cluster ID map is selected in the “Cell Attributes” widget of the SELECT FEATURES section, so that each cell is represented as a point with the color of its cluster in the middle part of the screen (where the first cluster ID always starts from 0):



Dataset: Asteria-Demo-Project_Human-PBMC-01_v1-4 (Demo_2D) View: Default

6,231 cells

cluster_ID_Louvain_method
Embedding: UMAP

0	1,228 (19.71%)
1	862 (13.83%)
2	833 (13.37%)
3	622 (9.98%)
4	510 (8.18%)
5	500 (8.02%)
6	413 (6.63%)
7	401 (6.44%)
8	296 (4.75%)
9	269 (4.32%)
10	172 (2.76%)
11	55 (0.88%)
12	46 (0.74%)
13	

You can visually explore each cell embedding follows:

- zoom / unzoom using your mouse wheel
- pan the cell using your left (resp. right) mouse button for 2D embeddings (resp. 3D embeddings)
- in case of UMAP 3D embedding: rotate 3D using your left mouse button

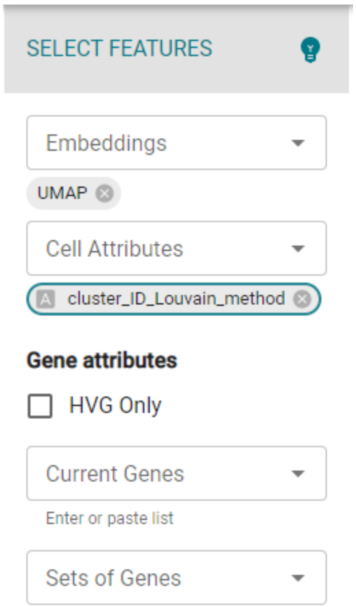
Note that the genes explorable in Cytonaut Rover are:

- restricted to the top N Highly Variable Genes (HVG) which have been identified by the post-processing run (e.g. the top N genes which have been automatically identified as highly variable genes during post-processing, where the post-processing parameter N is set to 2000 by default but can be changed by the user);

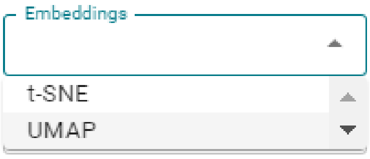
- named with their symbol followed by their unique gene identifier (e.g. MS4A1:ENSG00000156738) if coming from count matrices generated by Cytonaut;

- associated with a gene expression level which is equal to ln2(10000*X+1), where X is the ratio of transcripts detected in the cell for the gene of interest, “ln2()” is the logarithm in base 2.

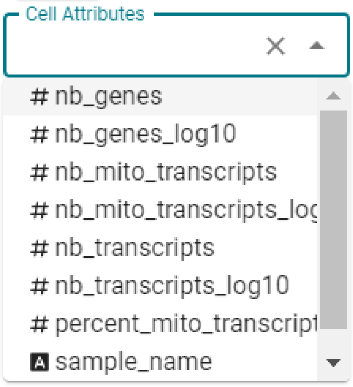
5) You can explore the data by expanding the SELECT FEATURES section in the left panel and playing with its 4 exploration widgets, such that new cell embedding(s) are dynamically added (resp. removed) in the middle part each time new data is selected (resp. unselected by clicking on the “cross” icon):



- (a) The “Embeddings” widget: select / unselect the embedding mode(s) applied by the post-processing run (e.g. UMAP 2D or 3D, t-SNE 2D) to embed the cells accordingly in the middle panel (note that the applied mode is then displayed just below the title of each cell embedding)

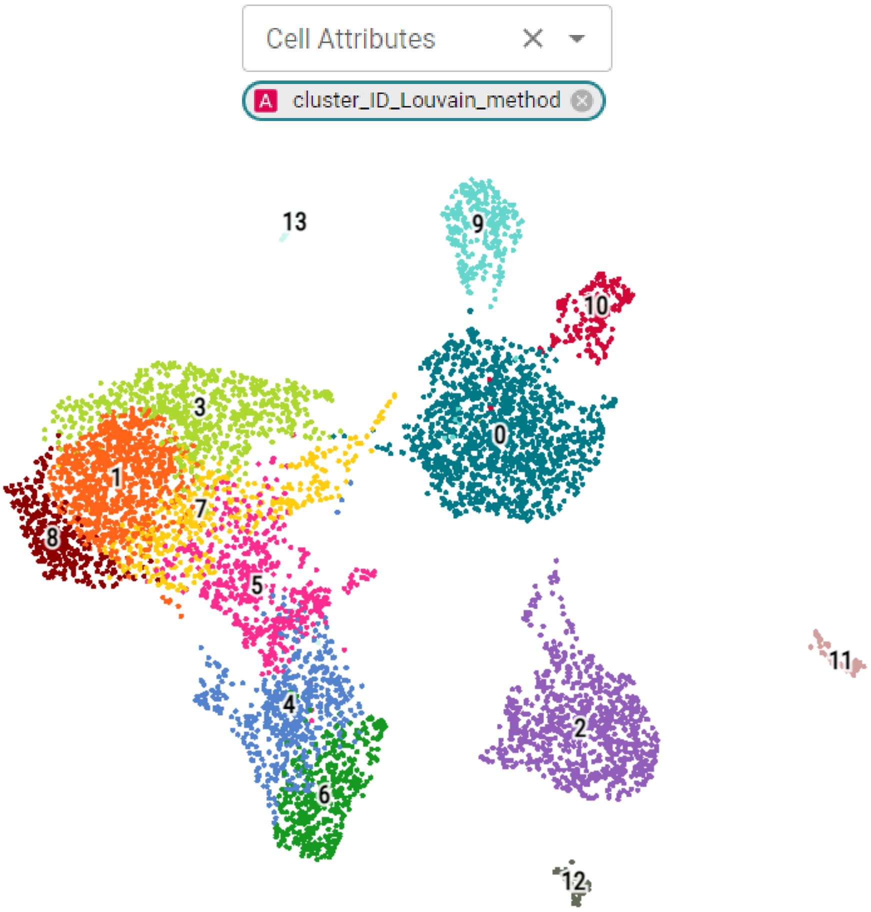


- (b) The “Cell Attributes” widget: select / unselect the cell attribute(s) to visualize the value of each cell attribute in the embedded cells (e.g. number of genes, number of transcripts, percentage of mitochondrial transcripts, cluster ID, sample name).

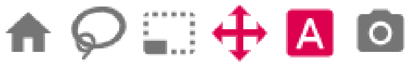


The cell attribute selected by default is the ID of the cluster which the cell belongs to and is named “cluster_ID_Louvain_method” if Louvain clustering has been applied, or “cluster_ID_Leiden_method” if Leiden clustering has been applied.

Note that it is possible to overlap in the “clustering” cell embedding the name of each cell cluster (cluster ID potentially followed by a manual annotation) by clicking on the “A” icon on the left side of “cluster_ID*” once selected in the left panel, which will set the “A” icon in magenta color (clicking again on it will set it back to gray and disable the overlap):



This “A” action is replicated in the top part of the CELL EMBEDDINGS menu to directly activate / deactivate all overlap of cell group attributes (e.g. cluster names, sample names, species) in the “clustering” cell embedding:



The cell attribute “sample_name” indicates the name of the sample which the cell belongs to, which enables to check the presence of batch effect if at least 2 samples have been analyzed in the post-processing run.

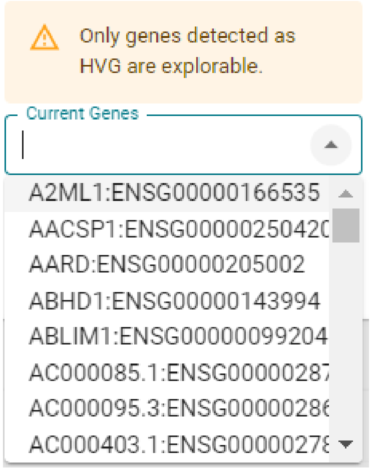
Note that:

- “nb_hvg_transcripts” is the number of HVG transcripts detected in the cell;
- “percent_hvg_transcripts” is percentage of transcripts detected in the cell which are HVG transcripts
- the “_log10” suffix on attribute X (where X is nb_genes, nb_transcripts or nb_mito_transcripts) means that logarithm in base 10 has been applied to X+0.001.

- (c) The “Current Genes” widget: type or select / unselect gene(s) of interest among the list of genes, in order to visualize the level of expression of each selected gene in the embedded cells:

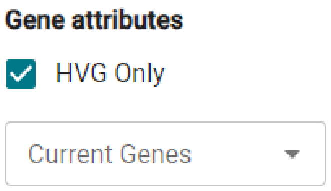
Note that:

- If the option “Keep all genes for DGE and visualization?” has not been activated for the post-processing run (which is the case by default), then the explorable genes are restricted to the N top genes detected as high variable genes (HVG) during post-processing (by default N = 2000):



⚠ Only genes detected as HVG are explorable.

- If the option “Keep all genes for DGE and visualization?” has been activated for the post-processing run, then all genes are explorable in Rover and a checkbox “HVG only” allows to restrict or not the selection to the highly variable genes:

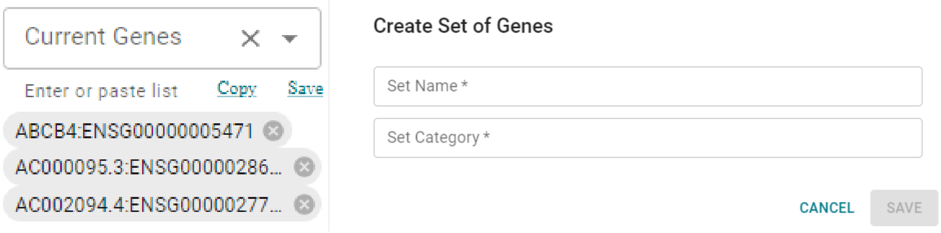


To enter a list of pre-defined genes written in a file (with one gene per line), you can simply drag and drop the file in the field.

You can also use Ctrl+V to copy paste one gene or a list of genes (separated by comma "," or semicolon ";" or tab character or end of line) in the following fields of Cytonaut Rover: "Current Genes", "Cell Attributes", "Sets of Genes", "Positive Genes" and "Negative Genes", knowing that these 2 rules are applied:

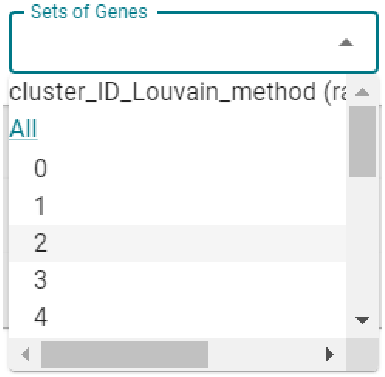
- if only one gene is pasted (without any separator): if it matches one and only one gene in the HVG genes, then this gene is selected, else all possible matching genes are proposed in a drop-down;
- if a list of several genes is pasted: only the HVG genes that match one and only one of the input genes are selected, and the others are ignored (note however that exact matching gene is automatically selected even if there are other partial matches)

Note that you have the possibility to save the list of your currently selected genes as a specific "set of genes" by clicking on the "Save" link below the widget, and then naming your custom set and its category, which will automatically register it in the "Set of Genes" widget (explained below):

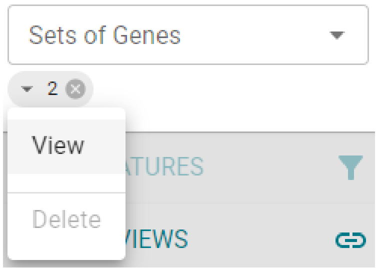


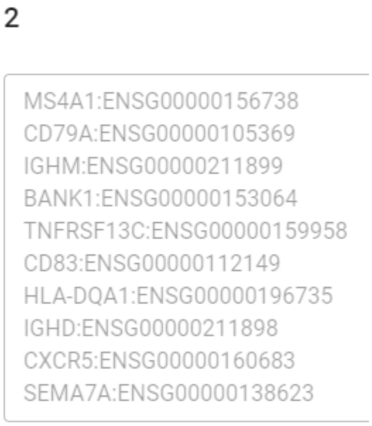
- ○ (d) The "Set of Genes" widget: select / unselect a set of genes in a given category, where the default category is the clustering organization (e.g. "cluster_ID_Louvain_method") which automatically presents for each cluster ID (0, 1, 2, etc.) the set of the top 10 differentially expressed genes "DEG" (i.e. those having the highest z-score's absolute value) in the differential gene expression of the cluster of interest, may they be over-expressed genes (z-score > 0) or under-expressed genes (z-score < 0).

  Note that selecting "All" will select all the top 10 DEG of all the clusters.



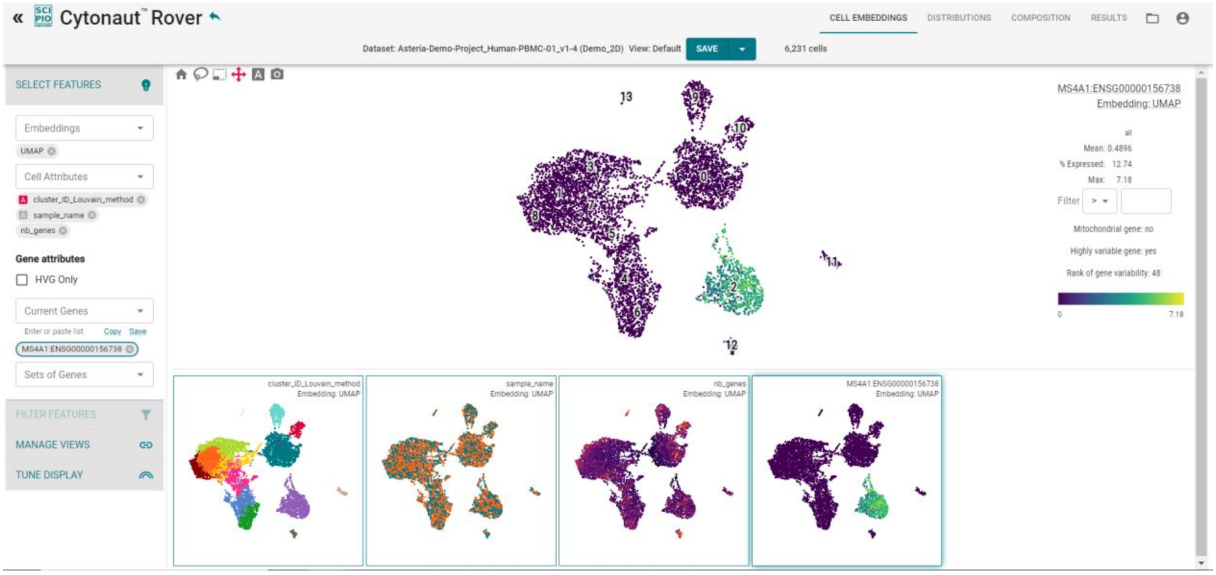
Once a set of genes is selected (e.g. the set of the top 10 DEG of the cluster ID “0”), you can see the list of its genes by clicking on its expander icon and then on “View” (and you have the possibility to copy all or some of them):




2

MS4A1:ENSG00000156738
CD79A:ENSG00000105369
IGHM:ENSG00000211899
BANK1:ENSG00000153064
TNFRSF13C:ENSG00000159958
CD83:ENSG00000112149
HLA-DQA1:ENSG00000196735
IGHD:ENSG00000211898
CXCR5:ENSG00000160683
SEMA7A:ENSG00000138623

Here is an example of what you can get by applying such selections on a given post-processing run, and the current display is associated with a given gene:

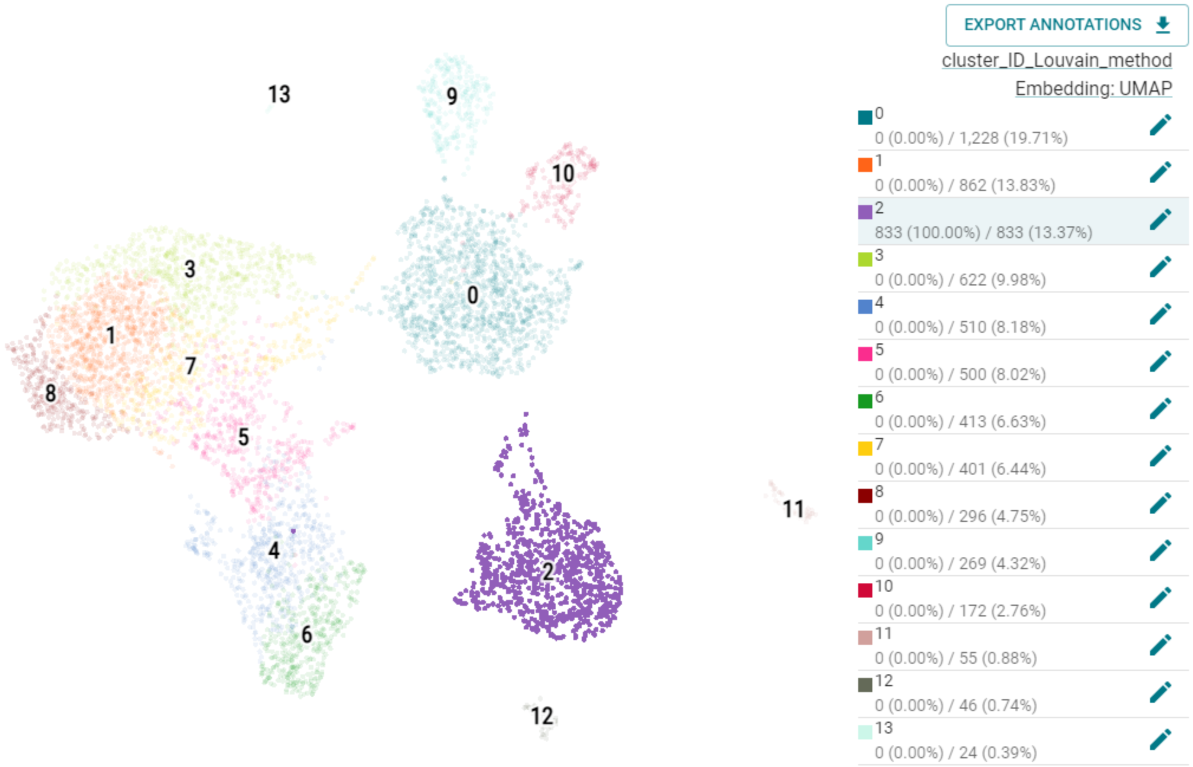


6) You may get the value of an individual cell for a given cell attribute (e.g. expression level of a given gene, number of transcripts) by putting your mouse over the cell and look at the "cell attribute value" displayed on top left corner of the screen.

7) You may apply filter(s) to the cells by applying one or several of these filtering strategies, which will automatically enable the FILTER FEATURES section in the left panel:

- Click on "cluster_ID*" (e.g. "cluster_ID_Louvain_method") to display the "clustering" cell embedding in the middle part):
  - In the right part you will see the absolute number as well as the percentage of cells belonging to each cluster
  - You may select in the right part one or several cluster ID(s) to only keep their cells, by clicking on the cluster ID to select it (Ctrl+click and Shift+click allows to select multiple cluster IDs at once; clicking again on a cluster allows to unselect it).

    Here is an example when selecting the cluster "2":



cluster_ID_Louvain_method
Embedding: UMAP

0
0 (0.00%) / 1,228 (19.71%)

1
0 (0.00%) / 862 (13.83%)

2
833 (100.00%) / 833 (13.37%)

3
0 (0.00%) / 622 (9.98%)

4
0 (0.00%) / 510 (8.18%)

5
0 (0.00%) / 500 (8.02%)

6
0 (0.00%) / 413 (6.63%)

7
0 (0.00%) / 401 (6.44%)

8
0 (0.00%) / 296 (4.75%)

9
0 (0.00%) / 269 (4.32%)

10
0 (0.00%) / 172 (2.76%)

11
0 (0.00%) / 55 (0.88%)

12
0 (0.00%) / 46 (0.74%)

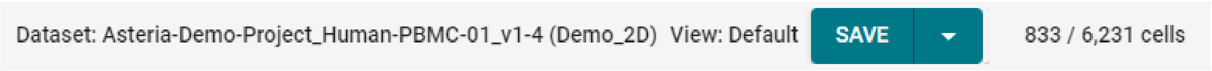
13
0 (0.00%) / 24 (0.39%)

- Click on "sample_name" to display the "sample name" cell embedding in the middle part, then select in the right part one or several sample(s) to only keep their cells

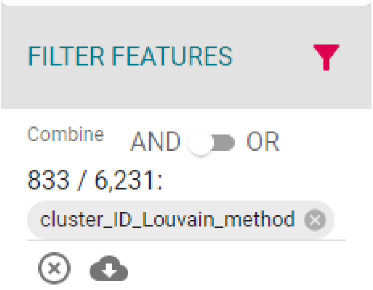
- Click on the "lasso" icon on top of the middle panel to manually select the cells (left click then drag and drop) to be kept inside the lasso (in this case a filter named "selection" will be applied). You can use "Ctrl" key to make multiple lasso selections.

- Click on the cell embedding representing a gene expression or a cell attribute in the middle part, then select in the right part the filter you want to apply (>, <, >=, <=, =, !=) to only keep the cells verifying this constraint

Once a filter is applied, the number of kept cells verifying the current filter(s) is displayed on the top banner, and the new filter name is displayed in the FILTER FEATURES section of the left panel, with the possibility to add / remove / combine filters (using the AND / OR switch):


Dataset: Asteria-Demo-Project_Human-PBMC-01_v1-4 (Demo_2D) View: Default

833 / 6,231 cells



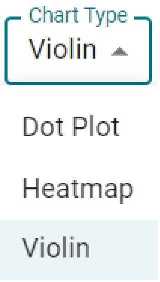
The FILTER FEATURE section is automatically expanded and activated with the filter icon set to magenta color whenever at least one filter has been applied by the user.

Note that it is possible to export the list of barcode sequences of the cells in a "_selection.txt" file, by clicking on the "download" icon ☁ just below the FILTER FEATURES section. Such export is useful to launch another post-processing run on a subset of cell barcodes, for example in order to refine the clustering and the cell annotation (for more details see the explanations of the action "UPLOAD SUBSET OF CELL BARCODES" in the section *"V. HOW TO SET POST-PROCESSING PARAMETERS"*).

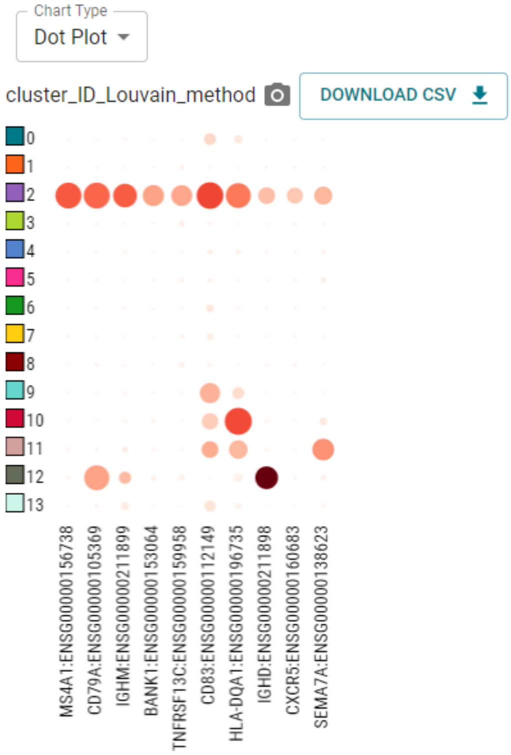
8) You may click on the DISTRIBUTIONS menu on the top to access the customized distribution statistics of:

- the genes currently selected in the SELECT FEATURES section, where gene statistics are displayed in the top part named "Features";
- the cell attributes currently selected in the SELECT FEATURES section, where cell attribute statistics are displayed in the bottom part named "Observations".

Multiple chart types are available ("Dot Plot", "Heatmap", "Violin"):

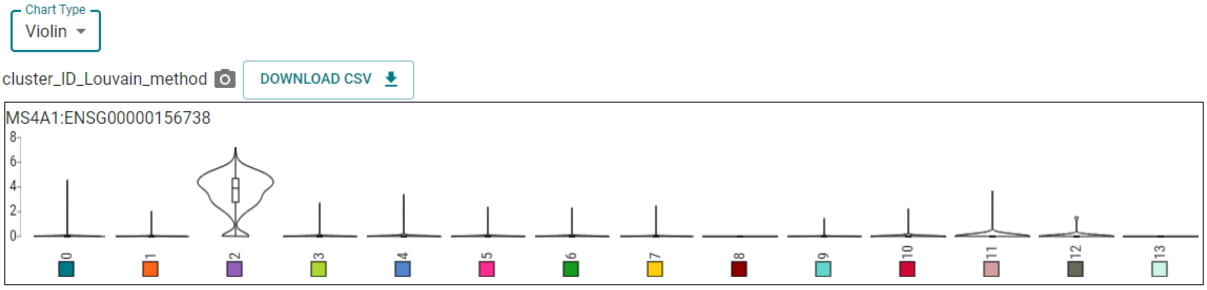


- The "Dot Plot" chart displays a disk for each currently selected gene and each cell cluster (or more generally each cell group), where disk color represents the mean value of gene expression and disk size represents the percentage of cells expressing the gene (i.e. percentage of cells in the group for which gene expression is not zero).



- ○ The “Heatmap” chart displays a color value for each currently selected gene and each cell cluster (or more generally each cell group), where the color represents the mean value of gene expression.

- ○ By selecting the “Violin” chart type, you can access more complete statistics including data dispersion, which are provided at mouseover to evaluate the heterogeneity of data distribution:
  - Mean
  - Median
  - Q1 (first quartile)
  - Q3 (third quartile)
  - LAV (lower adjacent value, defined as the lowest observation that is higher than Q1 - 1.5 * (Q3 - Q1))
  - UAV (upper adjacent value, defined as the highest observation that is lower than Q3 + 1.5 * (Q3 - Q1))
  - Min
  - Max
  - % Expressed (percentage of cells in the group for which gene expression is not zero)



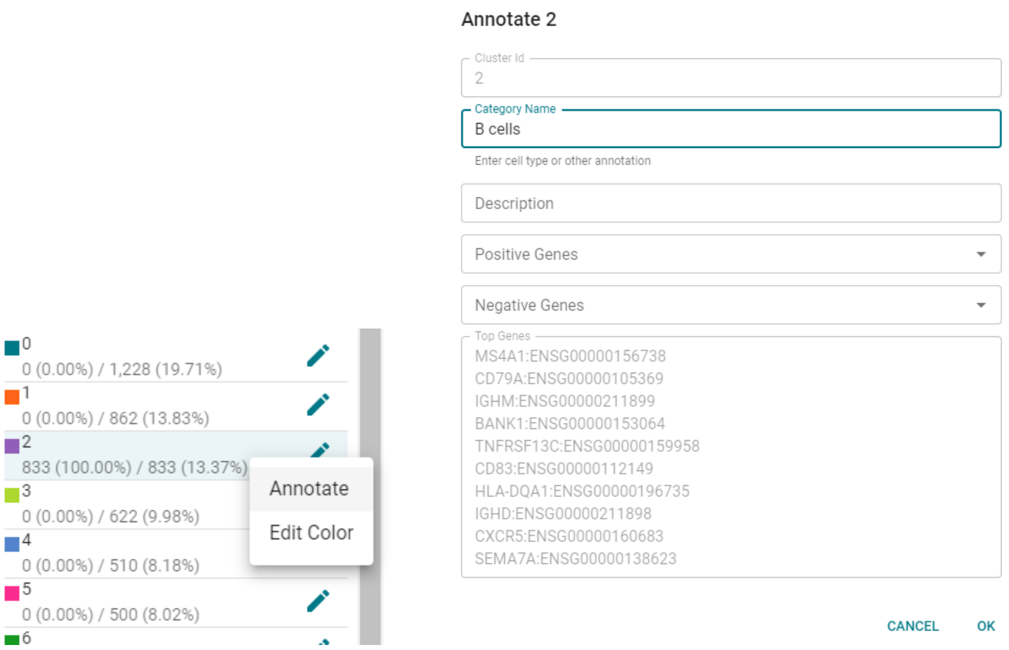
In case at least one filter has been applied in the FILTER FEATURES section, every statistical graph displayed in the menus DISTRIBUTIONS and COMPOSITION is duplicated just below as a new graph with the keyword “(selection)” its title and which restricts the statistics to the cells selected by the currently applied filter(s).

9) If you select at least two cell attributes which are grouping cells by categories (e.g. “cluster_ID*”, “sample_name”, “species”), then you can go to the COMPOSITION menu to visualize how the cells in each group of the first selected category (e.g. “sample_name”) are distributed across the groups of the second selected category (e.g. “cluster_ID*”). Note that it is possible to re-order such cell attributes using drag & drop in the left panel.

10) You can download each table, graph or cell embeddings displayed in Cytonaut™ Rover in .csv / .png / .svg format when applicable by clicking on the associated download or capture icon.

11) You can manually annotate a cell cluster in the “clustering” cell embedding (only for owner of the project) by clicking in the right part of the window on the pencil icon ✎ which is located next to the cluster ID, then by clicking on “Annotate”, so that you can:

- Enter in the field “Category Name” the name of the cluster, typically the cell type of the cells belonging to this cluster (this cluster name will be concatenated to the cluster ID in most of the displayed figures and tables)
- Optional: enter a “Description” for the cluster
- Optional: manually select “Positive Genes” for the cluster (e.g. significantly over-expressed genes)
- Optional: manually select “Negative Genes” for the cluster (e.g. significantly under-expressed genes)
- Optional: check the list of the top 10 DEG (differentially expressed genes) automatically detected for the cluster based on the absolute value of the z-score in the differential gene expression results when compared to all other clusters



0
0 (0.00%) / 1,228 (19.71%)
1
0 (0.00%) / 862 (13.83%)
2
833 (100.00%) / 833 (13.37%)
3
0 (0.00%) / 622 (9.98%)
4
0 (0.00%) / 510 (8.18%)
5
0 (0.00%) / 500 (8.02%)
6

Annotate 2

Cluster Id: 2

Category Name: B cells

Enter cell type or other annotation

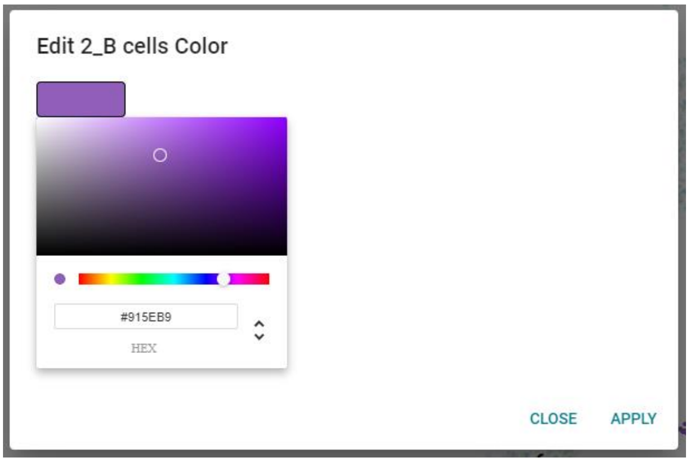
Description

Positive Genes

Negative Genes

Top Genes:
MS4A1:ENSG00000156738
CD79A:ENSG00000105369
IGHM:ENSG00000211899
BANK1:ENSG00000153064
TNFRSF13C:ENSG00000159958
CD83:ENSG00000112149
HLA-DQA1:ENSG00000196735
IGHD:ENSG00000211898
CXCR5:ENSG00000160683
SEMA7A:ENSG00000138623

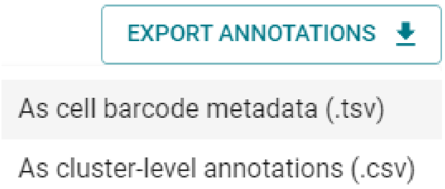
It is also possible to edit the color of a cluster by clicking on the pencil icon and then on “Edit Color” in RGB, HSL or HEX format (after editing the color, click on APPLY and then on CLOSE to apply the new color value):



Edit 2_B cells Color

HEX: #915EB9

You can export your annotations anytime by clicking on the button EXPORT ANNOTATIONS located just above the list of cell groups displayed on the right part, and which proposes the 2 following options:


EXPORT ANNOTATIONS
As cell barcode metadata (.tsv)
As cluster-level annotations (.csv)

- Option (a): Export cell-level annotation as cell barcode metadata (in .tsv format)

  The cell barcode metadata are exported in a file having the following .tsv format (tab separated values, compatible with Excel), with one line per cell barcode:

  - 1st row: barcode_sequence, barcode_annotation
  - 1st column from 2nd row: name of the sample which the cell barcode belongs to, followed with the sequence of the cell barcode (e.g. Sample1_ACCAATCTACTG)
  - 2nd column from 2nd row: annotation of the cluster which the cell barcode belongs to (e.g. "B cells"), or the ID of this cluster in case it has not been annotated

- Option (b): Export cluster-level annotations (in .csv format)

  The cluster metadata are exported in a file in .csv format (comma separated values, compatible with Excel) with the following information for each cluster:

  - cluster ID,
  - cluster name (i.e. cluster annotation, typically lits cell type),
  - number of cells in the cluster before Rover filtering,
  - number of cells in the cluster after Rover filtering,
  - cluster description,
  - cluster color (in HEX format),
  - list of “positive genes” manually entered for the cluster,
  - list of “negative genes” manually entered for the cluster,
  - list of the top 10 DEG (differentially expressed genes) detected for the cluster based on the ansolute value of the z-score value in the differential gene expression results when compared to all other clusters.

If a filter is currently applied in Rover, then only the cell barcodes currently selected in this filter will be exported.

Notes:

- If the project is a read-only project, such as demo projects or shared projects, then it is not possible to edit cluster annotations but it is possible to access them in read-only mode by clicking on the “eye” icon 👁

- They are two ways of accessing the 10 top DEG (differentially expressed genes) in the CELL EMBEDDINGS menu when visualizing a cell attribute of type automated clustering (“cluster_ID_*”) or custom DGE (“custom_groups”):

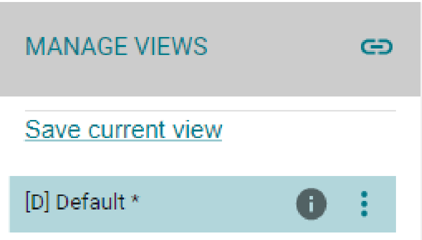
    - Option (a): Click on “Pencil icon + Annotate” (or on the “Eye icon” in read-only mode) next to a the cell population of interest in the top right list, to see the list of top 10 DEG at the bottom of the window;

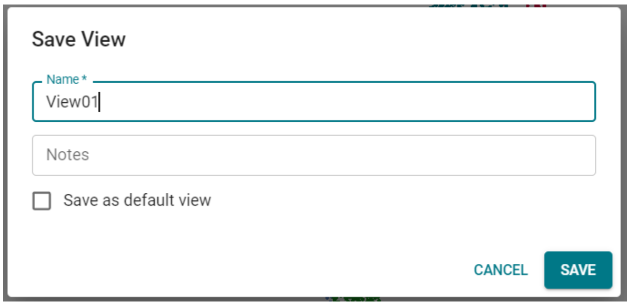
    - Option (b): Select the cell population of interest in “Set of Genes” in left panel, then click on “View” to see the list of top 10 DEG.

  Besides, to access more than 10 top DEG and/or only “over-expressed genes” (resp. “under-expressed genes”) with z-score > 0 (resp. z-score < 0), the RESULTS menu of Cytonaut Rover still allows to apply such dynamic filters or to download the entire DGE results in CSV format so that custom filters can be applied using Excel for example.

- It is possible to perform multiple cluster annotations in the same Rover page (i.e. same project and same post-processing run) by saving / loading each performed annotation in a different view in the section “MANAGE VIEWS” (see below)

12) You can save anytime all the cluster annotations you have performed, as well as all the selections and/or filters that you have applied on the Rover page of a given post-processing run, by expanding the MANAGE VIEWS section in the left panel, and then clicking on “Save current view” (only for owner of the project). This action will invite you to name your custom view and add a potential description, leading to the creation of a new view. You will be later able to access this view again by simply clicking on its name.




Save View

Name *: View01

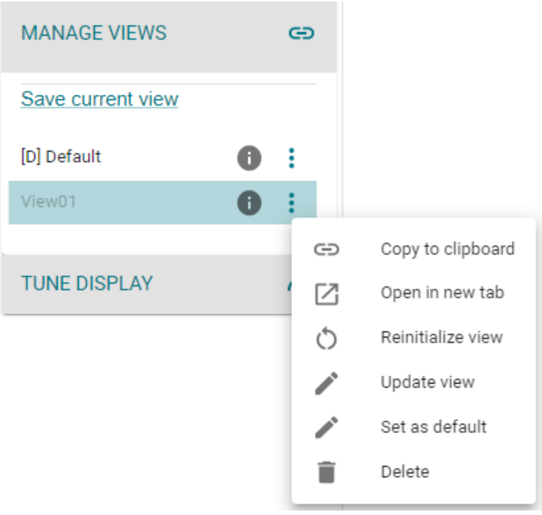
Notes

□ Save as default view

Once a view is created, you can click on the “more actions” icon ⋮ displayed next to this view in order to do one of the following actions:

- copy the link of the view
- open it in a new tab
- reinitialize the view to come back to the first default view which is generated at the creation of the Rover page (only for owner of the project)
- edit the name or the notes of the view (only for owner of the project)

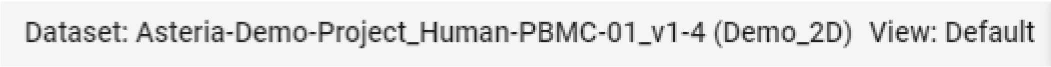
- set the view as default view to be displayed when opening the Rover page (only for owner of the project)
- delete the view (only for owner of the project)



Multiple views can thus be created and saved for a given post-processing run.

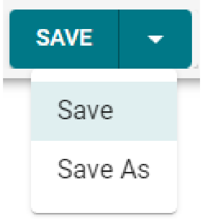
In all cases, key information is displayed in the top banner of Cytonaut Rover:

- The name of the currently loaded view (if not the default one) is displayed next to project name and run name



Dataset: Asteria-Demo-Project_Human-PBMC-01_v1-4 (Demo_2D) View: Default

- A SAVE button, next to the view name, allows to save all current settings (including cluster colors and annotations in particular) in a view:
    - either in the already selected view (via the option “Save”, also via Ctrl+S)
    - or in a new view (via the option “Save As”, also via Ctrl+Shift+S)

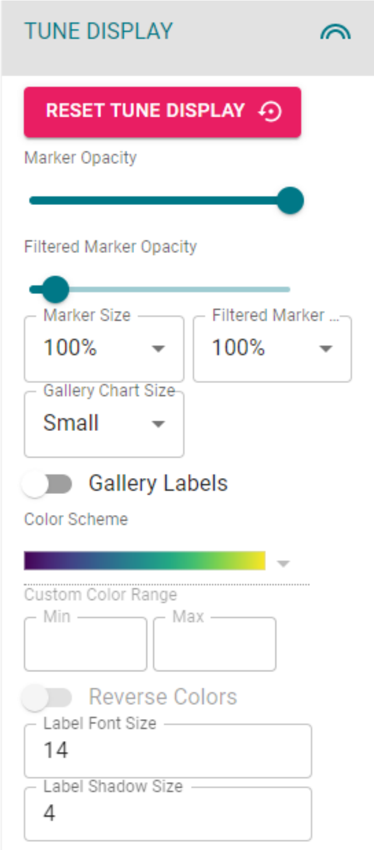


- A “back” icon, located next to “Cytonaut Rover”, allows to come back to the Cytonaut menu “Visualize Result Data” where the project of interest and its post-processing run of interest are selected (this facilitates iterative parameter tuning)

Cytonaut™ Rover

- In case of page refresh or page closing, a warning message inviting to save the current view is displayed if at least one modification has been done without saving.

Finally, you can use the TUNE DISPLAY section in the left panel to modify data rendering such as dot size or color scheme, where such rendering customization is also recorded in saved views:



TUNE DISPLAY

RESET TUNE DISPLAY

Marker Opacity

Filtered Marker Opacity

Marker Size: 100%

Filtered Marker ...: 100%

Gallery Chart Size: Small

Gallery Labels

Color Scheme

Custom Color Range

Min

Max

Reverse Colors

Label Font Size: 14

Label Shadow Size: 4

8. Get Differential Gene Expression Results & Download your processed data

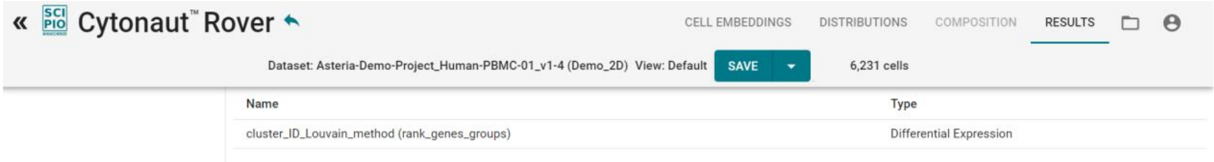
Important preliminary note:

If at least two samples have been selected in the post-processing run, then the standard Differential Gene Expression (DGE) methodology will be applied on the union of the selected samples based on the automatically defined clusters, and there is no way to separate the statistical contributions of each sample.

The only way to obtain DGE results for each sample individually is to launch one post-processing run per sample, by only checking one sample for each successive run, in which case it is recommended to enter the sample name in the Run ID field for traceability purposes.

Besides, if you want to customize the way you want to apply a DGE, based on user-defined groups of cells instead of automatically defined clusters, please see chapter *"VIII. HOW TO RUN CUSTOM DIFFERENTIAL GENE EXPRESSION"*).

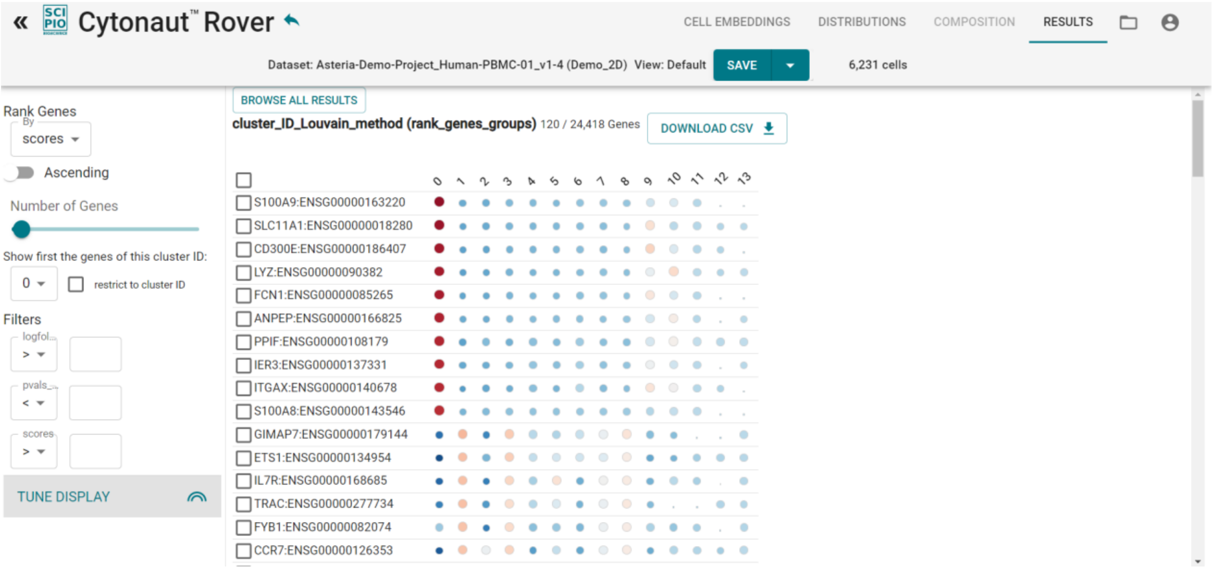
1) From the Cytonaut™ Rover page associated with the *Project Run*, click on the RESULTS menu on the top, then click on the line starting with "cluster_ID" and you will access the DGE results:



Dataset: Asteria-Demo-Project_Human-PBMC-01_v1-4 (Demo_2D) View: Default

6,231 cells

Name	Type
cluster_ID_Louvain_method (rank_genes_groups)	Differential Expression



Dataset: Asteria-Demo-Project_Human-PBMC-01_v1-4 (Demo_2D) View: Default

6,231 cells

Rank Genes By: scores

Ascending

Number of Genes

Show first the genes of this cluster ID: 0 □ restrict to cluster ID

Filters

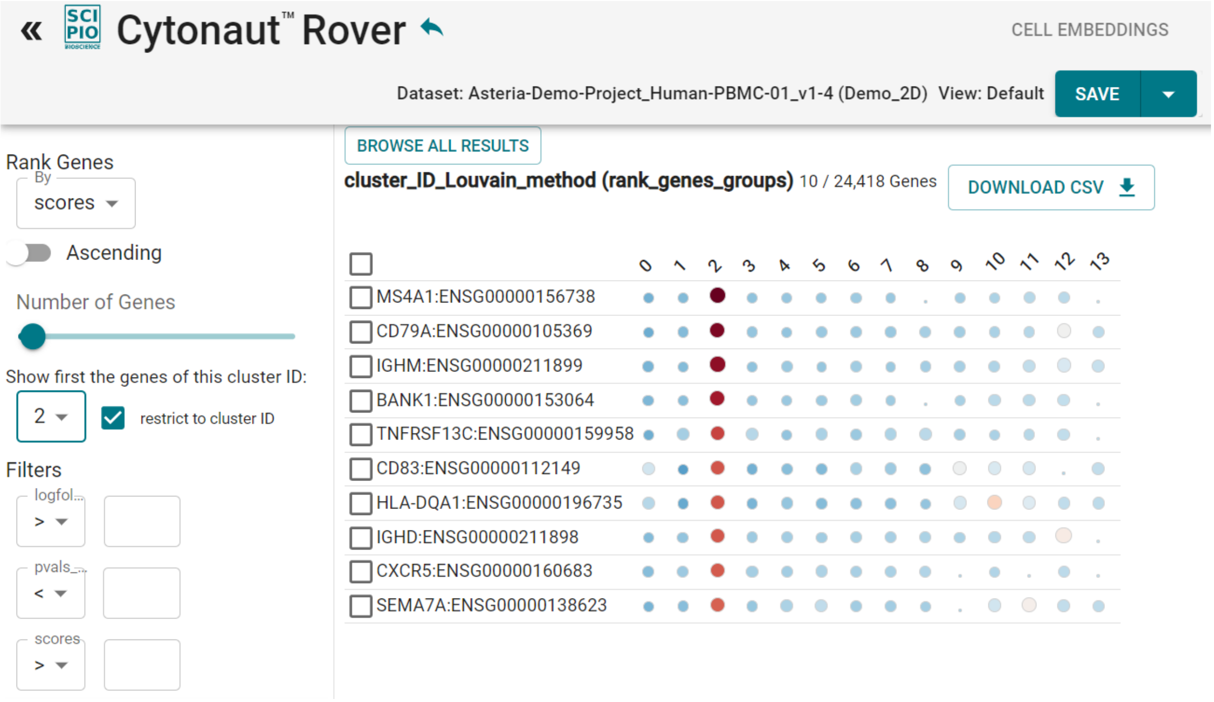
logfol... >

pvals_... <

scores >

cluster_ID_Louvain_method (rank_genes_groups) 120 / 24,418 Genes

In left panel, the checkbox “restrict to cluster ID” allows to quickly identify in the top lines of the DGE matrix the most differentially expressed genes in a specific cluster ID, for example the cluster 2 as shown below:



Dataset: Asteria-Demo-Project_Human-PBMC-01_v1-4 (Demo_2D) View: Default

Rank Genes By: scores

Ascending

Show first the genes of this cluster ID: 2

☑ restrict to cluster ID

cluster_ID_Louvain_method (rank_genes_groups) 10 / 24,418 Genes

2) You can customize the DGE table by selecting in the left panel z-scores (“scores”), adjusted p-values (pvals_adj) or log fold change (logfoldchanges) for the following actions:

- Ranking the genes (ranking parameter set to “scores” by default; choice of ascending / descending order; choice of N for the top N genes; choice of first cluster ID to be considered for gene ordering);
- Filtering the genes (min / max values);
- Setting the color of the disk for the gene-cluster point (set to “scores” by default);
- Setting the size of the circle for the gene-cluster point (set to “logfoldchanges” by default).

3) You can access the values associated with each gene-cluster point at mouseover: they will be displayed on top of the DGE table

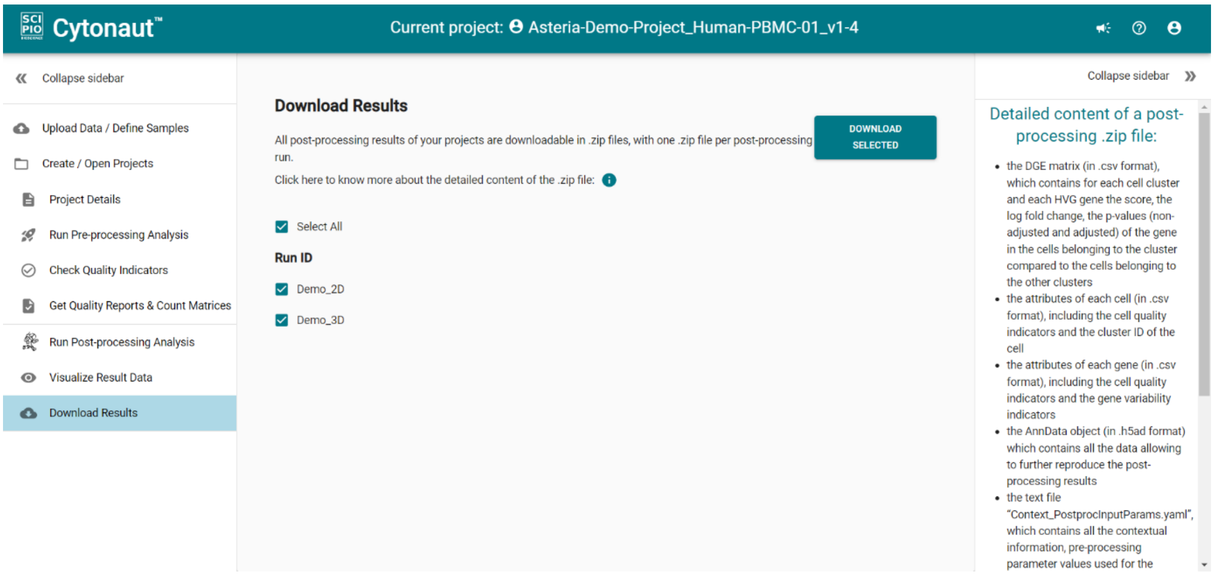
4) You can select one or several gene(s) shown in the DGE table by checking the box next to the gene, leading to the selection of the gene(s) in the “Current Genes” widget.

5) Finally, you can download the whole DGE matrix in .csv format (comma separated values) by clicking on the DOWNLOAD CSV button.

Note that this same DGE matrix can also be downloaded from the Download Results menu of the main Cytonaut application (as explained in the next section)

9. Download your processed data

From the main Cytonaut™ application, you can download all post-processing results in a single .zip file per run of the project, by clicking on the menu Download Results. The downloaded results include the following files:


Cytonaut™
Current project: Asteria-Demo-Project_Human-PBMC-01_v1-4
Collapse sidebar
Upload Data / Define Samples
Create / Open Projects
Project Details
Run Pre-processing Analysis
Check Quality Indicators
Get Quality Reports & Count Matrices
Run Post-processing Analysis
Visualize Result Data
Download Results
DOWNLOAD SELECTED
Collapse sidebar

Download Results

All post-processing results of your projects are downloadable in .zip files, with one .zip file per post-processing run.

Click here to know more about the detailed content of the .zip file:

☑ Select All

Run ID

☑ Demo_2D

☑ Demo_3D

Detailed content of a post-processing .zip file:

- the DGE matrix (in .csv format), which contains for each cell cluster and each HVG gene the score, the log fold change, the p-values (non-adjusted and adjusted) of the gene in the cells belonging to the cluster compared to the cells belonging to the other clusters
- the attributes of each cell (in .csv format), including the cell quality indicators and the cluster ID of the cell
- the attributes of each gene (in .csv format), including the cell quality indicators and the gene variability indicators
- the AnnData object (in .h5ad format) which contains all the data allowing to further reproduce the post-processing results
- the text file "Context_PostprocInputParams.yaml", which contains all the contextual information, pre-processing parameter values used for the

- The text file “Context_PostprocInputParams.yaml”, which contains all the contextual information, email of the author of the project, software version, UTC times of project creation and of pre-processing run start, pre-processing parameter values used for the project, UTC time of post-processing run start, and all post-processing parameter values used for the given run of the project.

  Note that you may apply the set of custom post-processing parameter values that you used for a past run to a new run (may it be on the same project or not) by downloading the file “Context_PostprocInputParams.yaml” associated with the past run and later uploading it before launching the new run by simply clicking on the button UPLOAD CONFIGURATION in the optional “Customize post-processing input” section of the menu Run Post-processing Analysis.

- The DGE matrix (in .csv format), which contains for each cell cluster “cluster_id” and each HVG gene “gene_name”:
  - “scores”: the z-score;
  - “logfoldchanges”: the log fold change;
  - “pvals”: the p-value of the gene in the cells belonging to the cluster compared to the cells belonging to the other clusters;
  - “pvals_adj”: ”: the adjusted p-value of the gene in the cells belonging to the cluster compared to the cells belonging to the other clusters;
  - “pct_nz_group”: the ratio of cells in the cluster for which the gene is detected (i.e. at least one transcript is detected);
  - “pct_nz_reference”: the ratio of cells in all the other clusters for which the gene is detected (i.e. at least one transcript is detected).

- The attributes of each cell (in .csv format), including the cell quality indicators and the cluster ID of the cell.

- The attributes of each gene (in .csv format), including the cell quality indicators and the gene variability indicators.

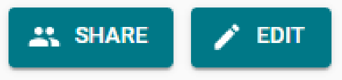
- The PCA information (in .csv format), including the results of the Principal Component Analysis step, allowing particular to check how many PCA dimensions have been used for the next post-processing steps (see the suffix "_used" in column names), and for each PCA dimension the percentage of cumulative explained variance among the first 100 dimensions.

- The AnnData object (in .h5ad format) which contains all the data allowing to further reproduce the post-processing results.

10. Share a project with another Cytonaut™ user

If the *Project* is not already opened: click on the menu Create / Open Projects in the left panel, then select the *Project* to be analyzed by clicking on it.

- In the menu Project Details of your project, click on the button SHARE



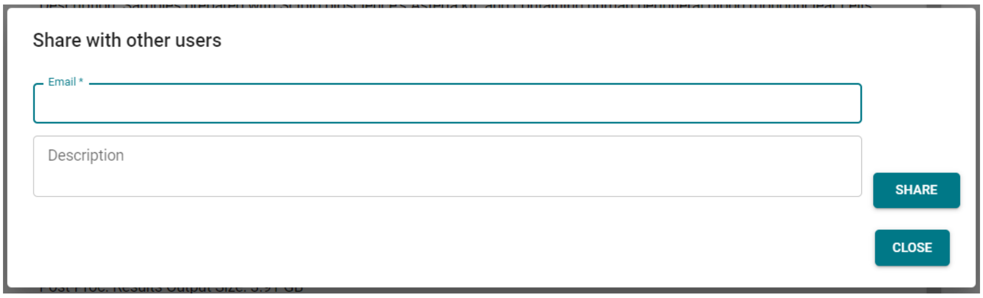
- In the pop-up window, enter the email of the Cytonaut user (this must be the email used for the Cytonaut account of this user) and an optional “Description” message to be included in the email, then click on “SHARE” and “CLOSE”

- As a result, the Cytonaut user will receive a notification email with a direct link to the shared project in Cytonaut application, allowing to access in read-only mode all the past and future content of the project (including Cytonaut Rover’s pages and saved views): data can be explored but not modified, and all results and figures can be downloaded.

- A shared project is equivalent to a demo project for the Cytonaut user who receives it, except that the shared projects are represented with the icon 👥 instead of the icon 📖 used for the demo projects.

- If a project has been shared with at least one Cytonaut user, then a shared icon ⋖ is displayed before the project name.

You may share a project with multiple Cytonaut users, and you may remove the sharing with a user by clicking on the “cross” icon:

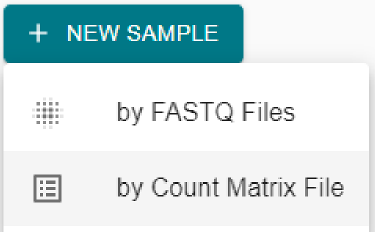


III. HOW TO UPLOAD COUNT MATRICES FOR DIRECT POST-PROCESSING

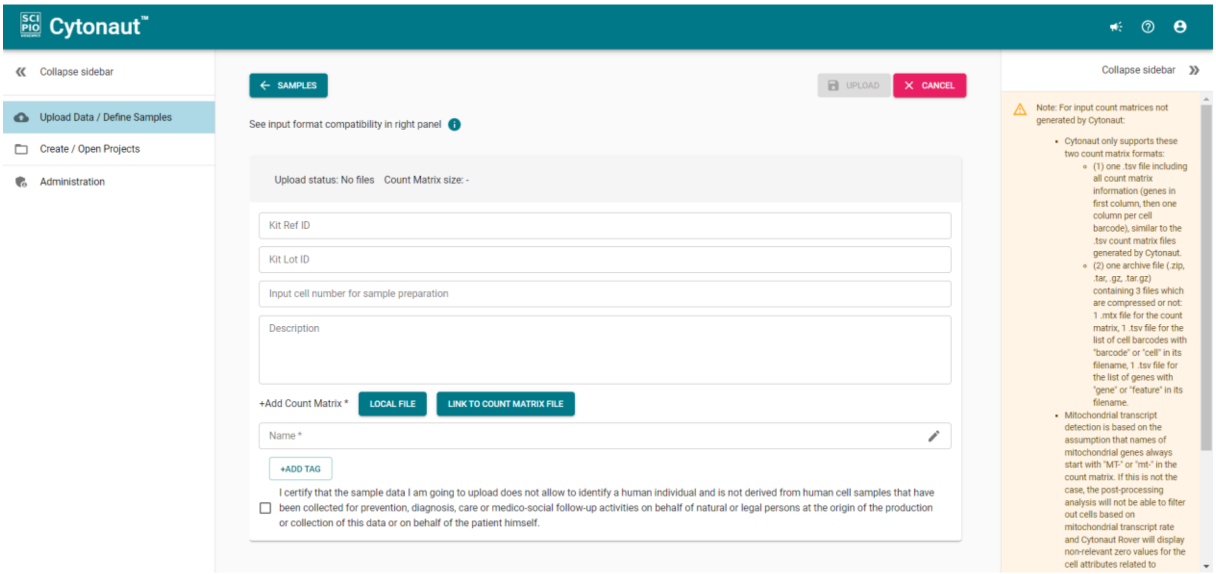
It is possible to directly input a count matrix file to define a sample in order to directly perform post-processing analysis, without restriction of the used single-cell sample preparation technology as far as input count matrix format is supported by Cytonaut.

This can be particularly useful to process count matrices provided as public data of already published paper.

- In the left panel, click on the menu Upload Data / Define Samples, then click on the button + NEW SAMPLE in the middle panel and select then “by Count Matrix File” in the drop-down menu:



- Cytonaut only supports these two count matrix formats:
  - (1) one non-compressed .tsv file including all count matrix information (genes in first column, then one column per cell barcode), similar to the .tsv count matrix files generated by Cytonaut.
  - (2) one archive file (.zip, .tar, .gz, tar.gz) containing 3 files which are compressed or not:
    - 1 .mtx file for the count matrix,
    - 1 .tsv file for the list of cell barcodes with "barcode" or "cell" in its filename,
    - 1 .tsv file for the list of genes with "gene" or "feature" in its filename.



See input format compatibility in right panel

Upload status: No files Count Matrix size: -

□ I certify that the sample data I am going to upload does not allow to identify a human individual and is not derived from human cell samples that have been collected for prevention, diagnosis, care or medico-social follow-up activities on behalf of natural or legal persons at the origin of the production or collection of this data or on behalf of the patient himself.

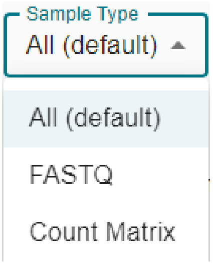
Note: For input count matrices not generated by Cytonaut:

- Cytonaut only supports these two count matrix formats:
  - (1) one .tsv file including all count matrix information (genes in first column, then one column per cell barcode), similar to the .tsv count matrix files generated by Cytonaut.
  - (2) one archive file (.zip, .tar, .gz, .tar.gz) containing 3 files which are compressed or not: 1 .mtx file for the count matrix, 1 .tsv file for the list of cell barcodes with "barcode" or "cell" in its filename, 1 .tsv file for the list of genes with "gene" or "feature" in its filename.
- Mitochondrial transcript detection is based on the assumption that names of mitochondrial genes always start with "MT-" or "mt-" in the count matrix. If this is not the case, the post-processing analysis will not be able to filter out cells based on mitochondrial transcript rate and Cytonaut Rover will display non-relevant zero values for the cell attributes related to

- For input count matrices not generated by Cytonaut:
    - Cytonaut is not responsible for input data quality and thus does not guarantee result accuracy.
    - Mitochondrial transcript detection is based on the assumption that names of mitochondrial genes always start with "MT-" or "mt-" in the count matrix. If this is not the case, the post-processing analysis will not be able to filter out cells based on mitochondrial transcript rate and Cytonaut Rover will display non-relevant zero values for the cell attributes related to mitochondrial transcripts.
    - The transcript count values in the count matrix are expected to be strictly lower than the value 65536 to be compatible with the 16 bytes data type supported by post-processing analysis.

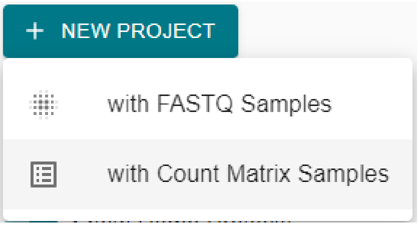
- Samples defined by uploading count matrices are categorized with the type “count matrix”, while other samples are set with type “FASTQ”. Similarly, projects including samples of type “count matrix” are categorized with the type “count matrix”, while other projects are set with the type “FASTQ”.

- A filter allows to search existing samples (resp. projects) by sample type (resp. project type):

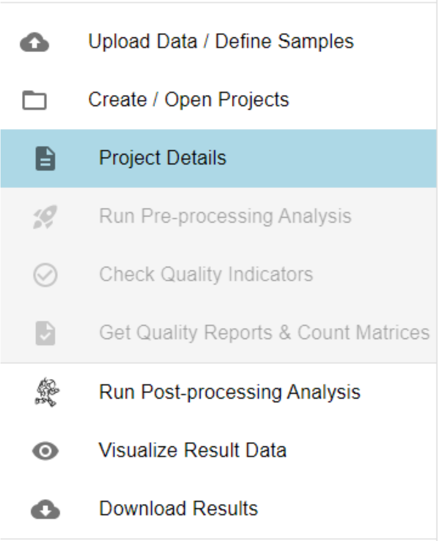


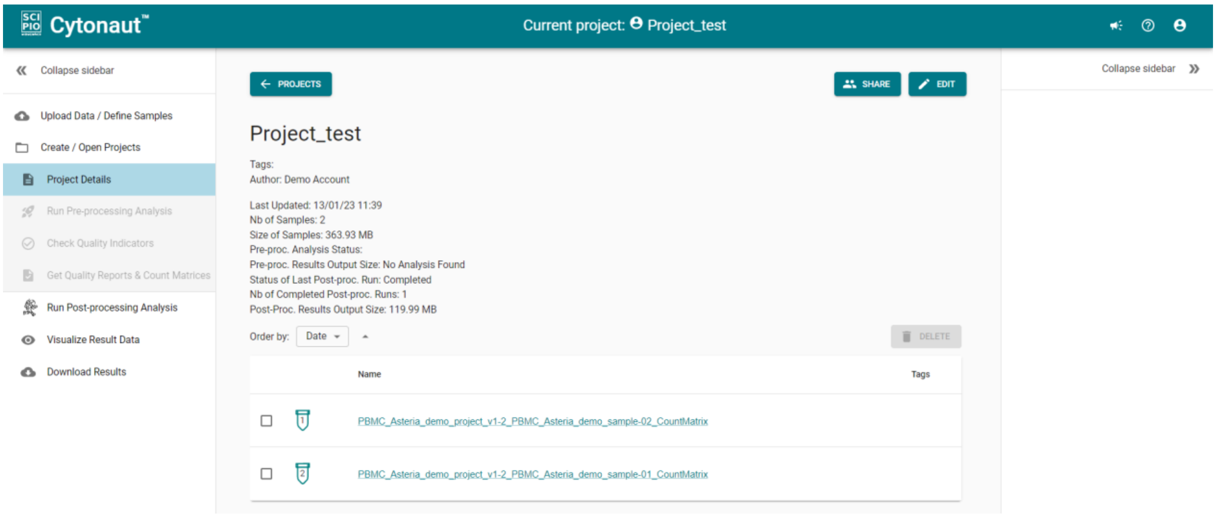
- Samples of type “count matrix” and samples of type “FASTQ” cannot be added in the same project.

- To create a project of type “count matrix”, select “with Count Matrix Samples” in the drop-down menu:



- Once a project of type “count matrix” is created with at least one “count matrix” sample, only the last menus “Run Post-processing Analysis”, “Visualize Result Data” and “Download Results” are enabled, allowing to directly perform these next steps:


Upload Data / Define Samples
Create / Open Projects
Project Details
Run Pre-processing Analysis
Check Quality Indicators
Get Quality Reports & Count Matrices
Run Post-processing Analysis
Visualize Result Data
Download Results



Project_test

Tags:
Author: Demo Account

Last Updated: 13/01/23 11:39
Nb of Samples: 2
Size of Samples: 363.93 MB
Pre-proc. Analysis Status:
Pre-proc. Results Output Size: No Analysis Found
Status of Last Post-proc. Run: Completed
Nb of Completed Post-proc. Runs: 1
Post-Proc. Results Output Size: 119.99 MB

Order by: Date

		Name	Tags
□	1	PBMC_Asteria_demo_project_v1-2_PBMC_Asteria_demo_sample-02_CountMatrix	
□	2	PBMC_Asteria_demo_project_v1-2_PBMC_Asteria_demo_sample-01_CountMatrix	

IV. HOW TO UNDERSTAND QUALITY INDICATORS

1. General quality indicators

Category	Criterion	Definition / Explanations
Main indicators	Number of cells	Number of barcodes coupled to at least one cell
Main indicators	Mean number per cell of raw reads in cell	
Main indicators	Total number of sequencing raw reads	
Main indicators	Median number of transcripts per cell	
Main indicators	Median number of genes per cell	
Main indicators	Total number of genes detected in any cell	Where a gene is said detected if at least one transcript is detected for this gene
Main indicators	Median of percentage of mitochondrial transcripts per cell (%)	
Main indicators	Percentage of mapped reads among raw reads in cells (%)	
Main indicators	Sequencing saturation among mapped reads in cells	Fraction of mapped reads in cells that non uniquely count for a transcript
Other indicators	Statistics about the distribution of the number of transcripts per cell	Distribution statistics: ○ Mean ○ Std: standard deviation ○ CV: coefficient of variation = std/mean ○ Min ○ Max ○ Median ○ Q1: first quartile ○ Q3: third quartile ○ Heterogeneity: relative inter-quartile range = (Q3-Q1)/median
Other indicators	Statistics about the distribution of the number of genes per cell	
Other indicators	Statistics about the distribution of the percentage of mitochondrial transcripts per cell	

2. Additional quality indicators for the two-species case (human & mouse)

Category	Indicator (multi-species case of human & mouse)	Definition / Explanations
Main indicators	Median purity per cell	Where cell purity is the probability that a transcript captured by the cell barcode has been expressed by the main cell coupled to the barcode, i.e. is not an ambient transcript
Main indicators	Percentage of species-defined cells (%)	Percentage of cells with purity > 0.95, knowing that cells with purity <= 0.95 are called undefined cells
Main indicators	Percentage of observed hetero-species cell multiplets (%)	Percentage of cells with purity < 2/3 ~ 0.67
Main indicators	Percentage of cell multiplets (%)	Percentage of barcodes coupled to at least two cells, including both hetero-species and homo-species cell multiplets, extrapolated from number of observed hetero-species cell multiplets
Main indicators	Number of human cells, i.e. defined for human species	Number of cells with cell purity > 0.95 and having a majority of human transcripts
Main indicators	Number of mouse cells, i.e. defined for mouse species	Number of cells with cell purity > 0.95 and having a majority of mouse transcripts
Main indicators	Median number of transcripts per cell for human species	Median number of human transcripts per human cell
Main indicators	Median number of transcripts per cell for mouse species	Median number of mouse transcripts per mouse cell
Main indicators	Median number of genes per cell for human species	Median number of human genes per human cell
Main indicators	Median number of genes per cell for mouse species	Median number of mouse genes per mouse cell
Main indicators	Total number of detected genes in any cell for human species	Where a gene is said detected if at least one transcript is detected for this gene
Main indicators	Total number of detected genes in any cell for mouse species	Where a gene is said detected if at least one transcript is detected for this gene
Other indicators	Statistics about the distribution of the purity per human cell	Distribution statistics: ○ Mean ○ Std: standard deviation ○ CV: coefficient of variation = std/mean ○ Min
Other indicators	Statistics about the distribution of the purity per mouse cell	
Other indicators	Statistics about the distribution of the number of transcripts per human cell	

Other indicators	Statistics about the distribution of the number of transcripts per mouse cell	○ Max ○ Median ○ Q1: first quartile ○ Q3: third quartile ○ Heterogeneity: relative inter-quartile range = (Q3-Q1)/median
Other indicators	Statistics about the distribution of the number of genes per human cell	
Other indicators	Statistics about the distribution of the number of genes per mouse cell	

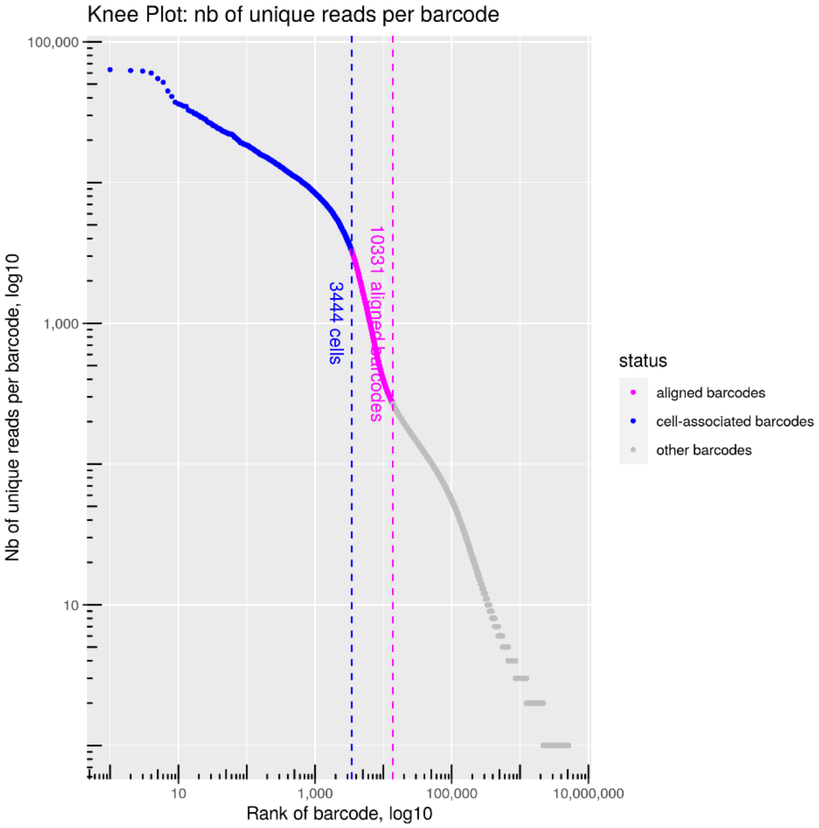
3. Graphs included in Quality Indicators reports

2. **Unique read counts per barcode**, also called “Knee Plot”

This graph shows in log-log scale the “Knee Plot” curve, which displays the number of unique reads (i.e. unique read1 sequences including barcode sequence and random sequence) for each barcode present in the R1 FASTQ file, where the barcodes are ranked in decreasing order from highest to lowest number of unique reads.

- The blue vertical dotted line is the threshold indicating the number of cell-associated barcodes detected in the sequencing data during pre-processing analysis (in automated mode it is intended to detect the first knee point of the “Knee Plot” curve): all barcodes before this blue threshold are defined as cell-associated barcodes (i.e. barcodes couples to at least one cell)

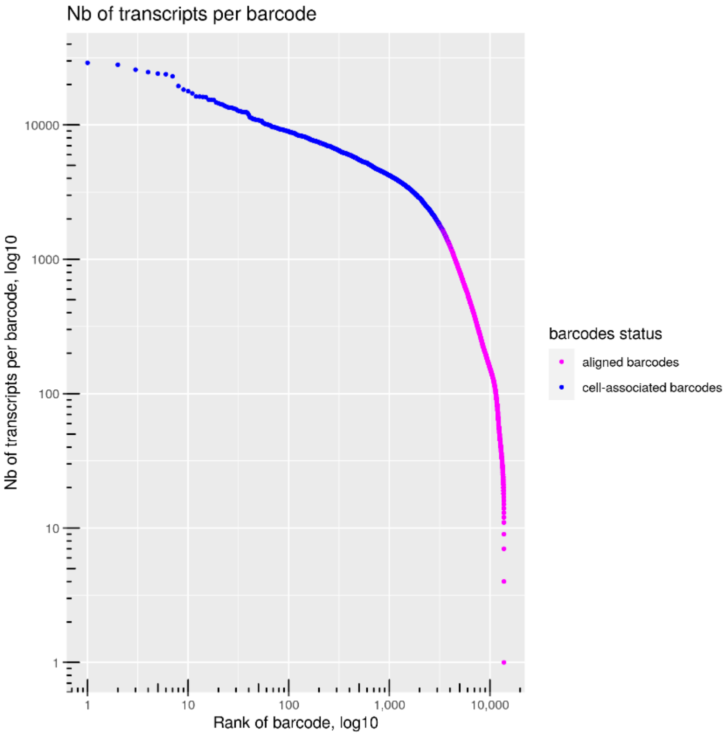
- The magenta vertical dotted line is the threshold indicating the number of “analyzed barcodes” and corresponds to 4 times the number of cell-associated barcodes, and all barcode before this magenta threshold are aligned to the genome of interest during pre-processing analysis (to get more quality control information).



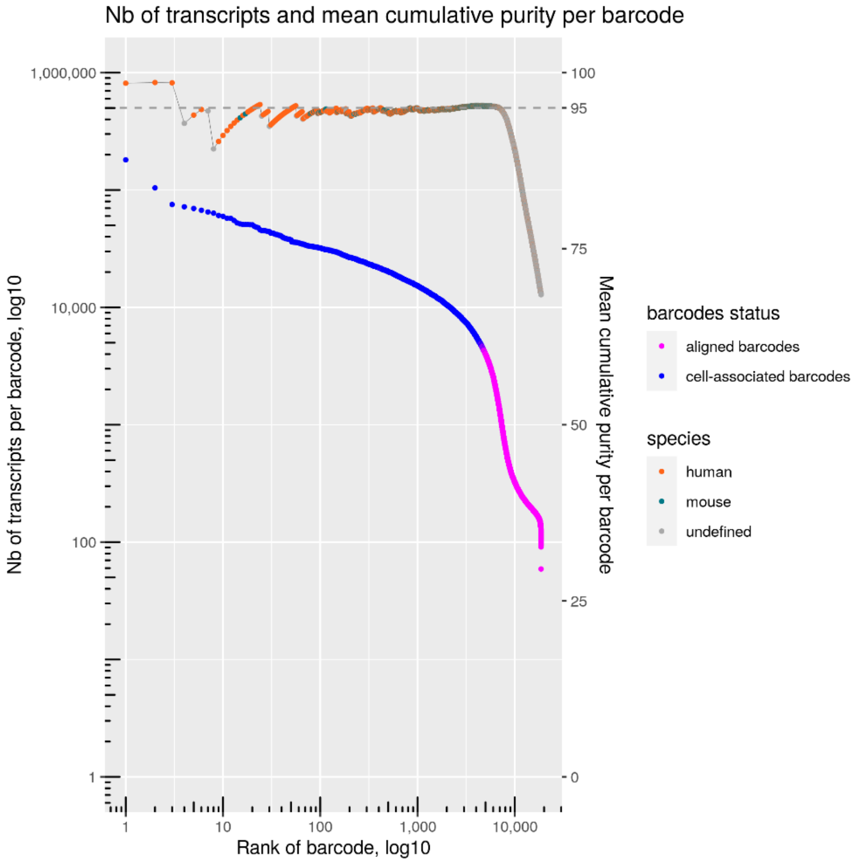
3. Transcript counts per barcode

This graph shows in log-log scale the number of transcripts detected in each "analyzed barcode", where the barcodes are ranked in decreasing order from highest to lowest number of transcripts.

- The blue points represent the cell-associated barcodes detected in the sequencing data during pre-processing analysis, while the magenta points represent the barcodes which have not been detected as cell-associated barcodes but which have also been aligned to the genome of interest (to get more quality control information).

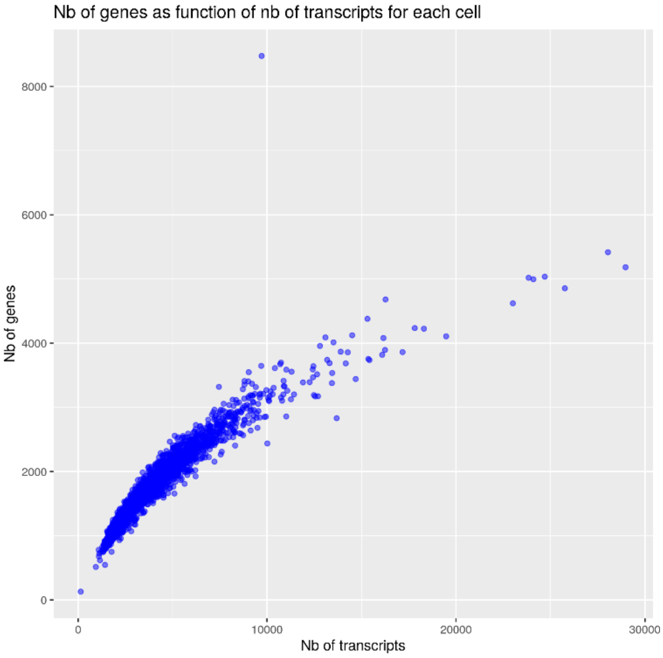


- In case of a mix species sample (e.g. human & mouse species), an additional curve is displayed and shows the mean cumulative purity per barcode:



4. Gene vs. Transcript counts

This graph shows for each cell-associated barcode the number of genes detected in the barcode (in y-axis) and the number of transcripts detected in the barcode (in x-axis).

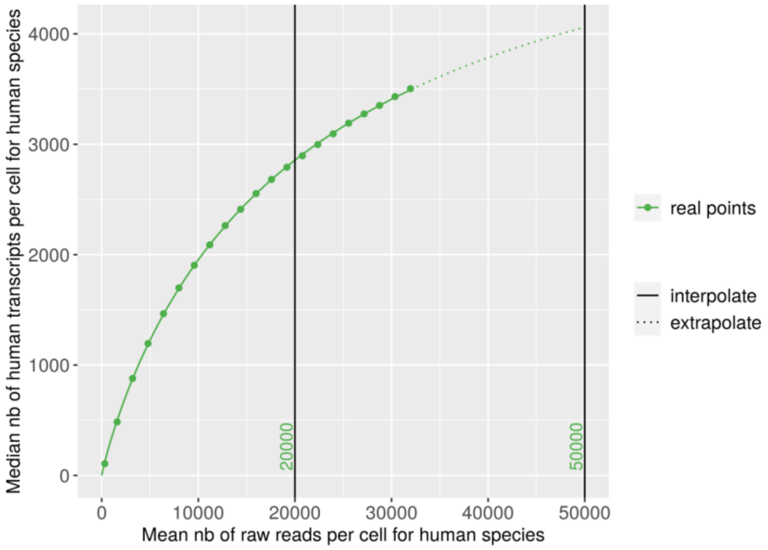


In case of a mix species sample (e.g. human & mouse species), such graph is provided with the restriction to the human (resp. mouse) genes and transcripts of the human species (resp. mouse species) cell-associated barcodes.

5. Downsampling curve for transcript counts

This graph shows the median number of detected transcripts per cell-associated barcode as a function of the sequencing depth defined as mean number of raw reads belonging to cell-associated barcodes.

- Lower sequencing depths have been automatically simulated by subsampling the raw reads of the FASTQ files.
- The interpolation of the resulting downsampled points allows to compare multiple samples at a common sequencing depth.
- The fitted transcript downsampling curve is then extrapolated using non-linear least squares to predict results for higher sequencing depths (prediction only, not truth).



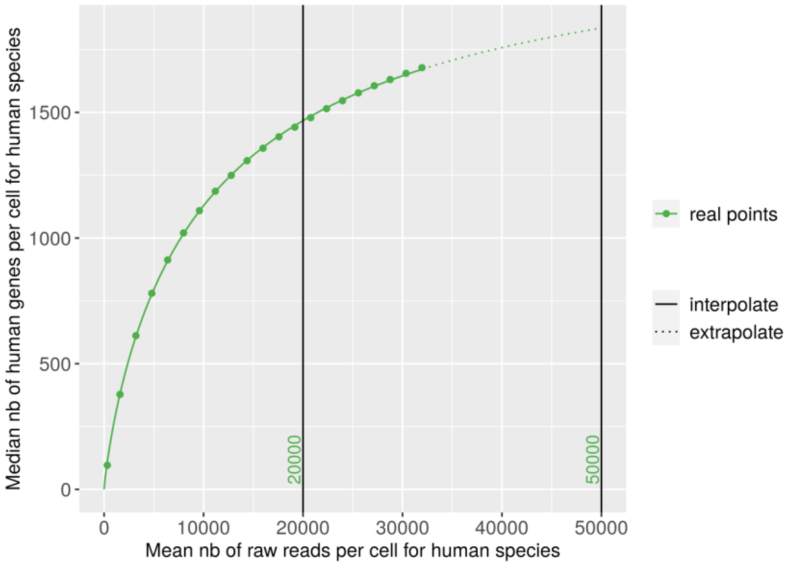
In case of a mix species sample (e.g. human & mouse species), such downsampling curve is provided with the restriction to the human (resp. mouse) transcripts of the human species (resp. mouse species) cell-associated barcodes.

6. Downsampling curve for gene counts

This graph shows the median number of detected genes per cell-associated barcode as a function of the sequencing depth defined as mean number of raw reads belonging to cell-associated barcodes.

- Lower sequencing depths have been automatically simulated by subsampling the raw reads of the FASTQ files.
- The interpolation of the resulting downsampled points allows to compare multiple samples at a common sequencing depth.

- The fitted gene downsampling curve is then extrapolated using non-linear least squares to predict results for higher sequencing depths (prediction only, not truth).



In case of a mix species sample (e.g. human & mouse species), such downsampling curve is provided with the restriction to the human (resp. mouse) genes of the human species (resp. mouse species) cell-associated barcodes.

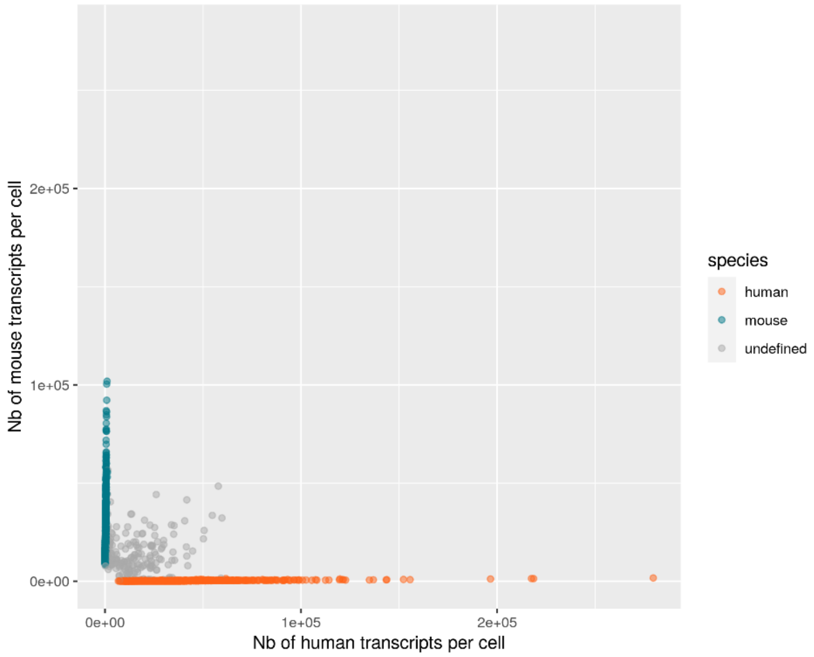
7. **Barnyard Plot** (only for mix species samples, e.g. human & mouse species)

In case of a mix species sample (e.g. human & mouse species), this graph shows for each cell-associated barcode the number of detected mouse transcripts (in y-axis) and the number of detected human transcripts (in x-axis).

- Orange points indicate human species barcodes (i.e. barcodes having a majority of human transcripts and a purity higher than 95%)

- Green points indicate human species barcodes (i.e. barcodes having a majority of mouse transcripts and a purity higher than 95%)

- Gray points are barcodes having a purity lower than 95% (they includes cell multiplets but also singlets with ambient noise)


Nb of mouse transcripts per cell
2e+05
1e+05
0e+00
0e+00
1e+05
2e+05
Nb of human transcripts per cell
species
human
mouse
undefined

V. HOW TO RE-DO PRE-PROCESSING IN MANUAL MODE

We assume that a pre-processing run has first been completed in “Auto” mode on the project of interest (highly recommended), and that the automated method has failed in accurately retrieving the cell-associated barcodes for at least one sample of the project.

Note that there is a systematic interruption of the pre-processing run if the number of detected cell-associated barcodes is larger than 15,000 (which corresponds to the maximum input cell number for one Asteria sample preparation). If this happens it means that the automated method of Cytonaut has failed to detect cell-associated barcodes for at least one sample of the project, for which it is then recommended to re-launch pre-processing analysis in “manual” mode.

As the pre-processing run has failed, the output data could not be completely generated, so post-processing analysis cannot be performed. The non-completion of pre-processing may be caused by bad quality of raw sequencing data and/or a highly overestimated number of cell-associated barcodes at least one sample of the project.

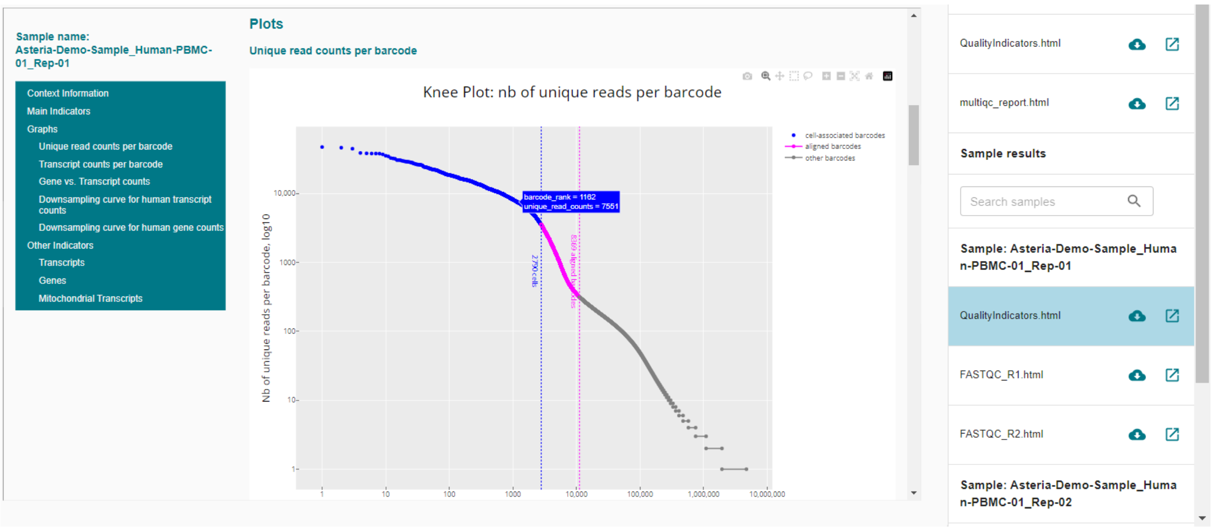
1) Go to the menu Check Quality Indicators and check the quality of sequencing data in the HTML files “multiqc_report” and “FATSQC” for R1 and R2.

   In case of poor sequencing quality for at least one sample of the project, you may check with your sequencing platform that all input FASTQ files are demultiplexed (not "undetermined") files of sufficient quality.

2) If the file “QualityIndicators.html” is not available for any sample, check the quality of cell-associated barcode detection in the file "Unique-raw-reads_KneePlot" which is likely generated for at least one of the samples.

   If the file “QualityIndicators.html” is available for at least one sample in the section “Sample results” in the right panel, check in this report each of the 2 interactive knee plot curves, named “Unique read counts per barcode” and “Transcript counts per barcode” in the “Graphs” section, to identify the position of the first knee point of the curve (e.g. point of maximum curvature) by starting on left side of the curve.

   - Note that the curve “Unique read counts per barcode” always contains one last knee plot on the right side of the curve, which corresponds to the number of beads put in the sample (typically 100,000 beads for a sample of 10,000 input cells). This last knee point must be ignored as the objective is to find cell-associated barcodes with significantly more unique raw reads than artefact barcodes.



Sample name: Asteria-Demo-Sample_Human-PBMC-01_Rep-01

Plots

Unique read counts per barcode

Knee Plot: nb of unique reads per barcode

barcode_rank = 1162
unique_read_counts = 7551

- ○ In most cases, the number N of cell-associated barcodes is provided by the x-coordinate of this knee point, which is displayed at mouseover on the knee point. If the automated method has detected much less or much more barcodes than this number N, then it is recommended to ask for N cells to analyze in “Manual” mode for the sample of interest.

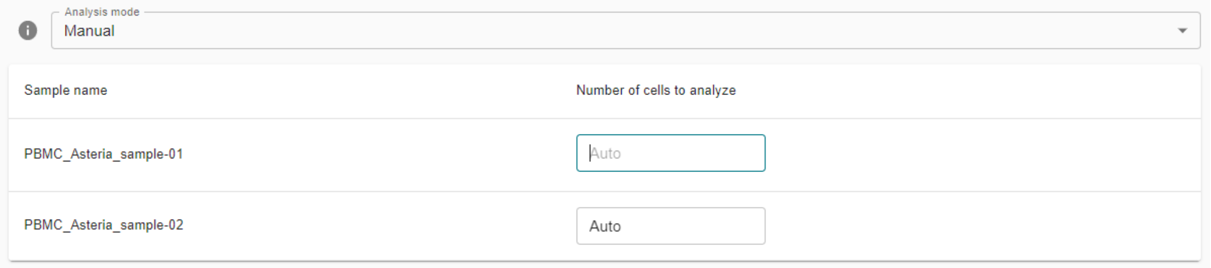
- ○ In some challenging cases, the first knee point is not obvious or a second knee point is appearing, due to the presence in the sample of some cells having much lower gene expression than the other cells (e.g. neutrophils among white cells). In this situation it is recommended to ask for a higher number N of cells to be analyzed (yet not more than the number of input cells) in order to be sure to detect all real cells coupled to beads.

3) Once the value N of cells has been estimated, there are 2 options (a) or (b) to re-do the pre-processing analysis in “Manual” mode for the current project:

   - ○ (a) Leave your current project as it is, to keep the history of the pre-processing run already completed in “Auto” mode, and create another project with the same samples

   - ○ (b) Download the QualityIndicator.html files in case they may be needed later, then delete the existing pre-processing results of the current project by clicking on the button DELETE PRE-PROCESSING RESULTS in the menu Run Pre-processing Analysis

4) Then, fill again the pre-processing parameters, set the field “Analysis Mode” to "Manual" and enter for each sample of interest the top N barcodes to be selected in decreasing number of unique raw reads (note that you can keep the “Auto” mode for parts of the samples):



Analysis mode: Manual

Sample name	Number of cells to analyze
PBMC_Asteria_sample-01	Auto
PBMC_Asteria_sample-02	Auto

5) Finally, launch the *Pre-processing Analysis* run by clicking on the button RUN PRE-PROCESSING ANALYSIS.

   For traceability purposes, the mode of each pre-processing run as well as the number of entered for each sample analyzed in “Manual” mode are automatically recorded in the *_Context_PostprocInputParams.yaml file which is downloadable in the menu Download Results.

VI. HOW TO SET POST-PROCESSING PARAMETERS

The data post-processing pipeline of Cytonaut™ uses the open-source Scanpy toolkit and follows most of the recommendations of this tutorial.

Most of default values are commonly used value for single-cell data post-processing, however it is recommended to adjust some value, in particular the filtering parameter values, depending on your application.

Category	Parameter	Default value	How to set it ?
Sample management	Apply batch correction?	Yes	This feature allows you to apply batch correction between multiple samples selected for the post-processing run. This sample integration process is based on the state-of-the-art method "Harmony" proposed by Korsunsky et al. in 2019. You only need to define as input the "batch condition" of each selected sample (by entering a condition name) in order to group in a common batch the samples belonging to the same experimental condition. The goal of the Harmony method is to remove technical variability between the batches (called batch effect), while maximizing the biological diversity within each batch. By default the sample condition is set to the sample name, meaning that each sample is considered as a distinct batch with a risk of over-correction, so it is recommended to edit the default sample condition name. If batch correction is de-activated, then all the selected samples will be pooled together and processed as one big sample, as if all samples were belonging to the same batch (meaning that no batch correction is applied). See this benchmark study for more information about batch correction methods.
Customize post-processing input (optional)	UPLOAD CONFIGURATION	Not applied	This feature allows you to apply for your current post-processing run the set of custom post-processing parameter values that you used for a past post-processing run (may it be on the same project or not) by first downloading the .zip file associated with the past run in the "Download Results" menu, then extracting from the .zip file the .yaml file ending with “Context_PostprocInputParams.yaml”, and finally selecting this .yaml file by clicking on the button "UPLOAD CONFIGURATION" in the "Run Post-processing Analysis" menu.

Category	Parameter	Default value	How to set it ?
Customize post-processing input (optional)	UPLOAD SUBSET OF CELL BARCODES	Not applied	This advanced feature enables you to apply your current post-processing run on a subset of cell barcodes that you have previously identified by filtering in Cytonaut Rover the results of a past post-processing run performed on the same sample(s). This allows for example to exclude outlier barcodes or to refine the clustering and the differential gene expression on a subset of cell barcodes. You first need to download in Cytonaut Rover the .txt file containing the subset of currently filtered cell barcodes (by clicking on the download icon), and then to upload this .txt file in Cytonaut by clicking on the button "UPLOAD SUBSET OF CELL BARCODES" in the "Run Post-processing Analysis" menu.
Customize post-processing input (optional)	UPLOAD BARCODE METADATA	Not applied	This feature allows you to upload cell barcode metadata for the input sample(s) to be analyzed by the post-processing run, in order to visualize and manage these metadata as “Cell Attributes” in the output Rover page. This requires to upload a file in .tsv format including one header line including the metadata name(s) from the second value, and at least 2 columns where the first column contains each cell barcode sequence prefixed with its sample name (e.g. "Sample1_AAAAATAACACT"), and the second column contains the metadata value of each cell barcode. If the file contains more than 2 columns, then the additional columns are interpreted as additional cell barcode metadata to be included as new cell attributes. A given metadata column contains either only numerical values or only string values (i.e. for categorical metadata). Empty values are interpreted as "NaN" values. If a cell barcode of one input sample is not found in the uploaded .tsv file, then its metadata values will be set to "NaN" (not applicable).
Customize post-processing input (optional)	UPLOAD GROUPS OF CELLS FOR CUSTOM DGE		This feature enables you to perform a custom DGE (Differential Gene Expression) based on at least 2 custom groups of cells previously defined in the Rover page of a previous post-processing run (named “parent run”).

Category	Parameter	Default value	How to set it ?
			The expected input file is an archive file (.zip, .tar, .gz, .tar.gz) which must include a set of .txt files (at least 2 .txt files). Each of these .txt files can be downloaded from the Rover menu “FILTER FEATURES” managing the subset of currently filtered cell barcodes, and must then be renamed with a name representing its custom group of cells (e.g. GroupA.txt). Such .txt file contains one column of barcode sequences prefixed with their belonging sample name (e.g. "Sample1_AAAAATAACACT"). Each .txt file must contain at least one barcode belonging to the past reference post-processing run. As a result, the post-processing run will provide these new results in addition to the classical clusters and DGE: (1) a custom clustering named “custom_groups” (explorable as a new cell attribute in Rover) made of the input groups of cells, and (2) the associated custom DGE in one versus all mode, named “custom_one-vs-all”, as well as in pairwise mode for each pair of groups (group XXX as test versus group YYY as reference), named “custom_pw_XXX-vs-YYY” (explorable as a new DGE result in Rover).
Customize post-processing input (optional)	Use extended count matrix?	No	If set to "yes", then the post-processing run will use as input the count matrix extended to all aligned barcodes (e.g. 4 times more barcodes than the number of detected cell-associated barcodes, ranked by decreasing number of unique reads). This could allow to detect more cells of lower gene expression by playing with post-processing or data visualization filtering parameters. Note that Cytonaut does not guarantee result accuracy if extended count matrices are used as input of post-processing analysis. Indeed, in most cases, most of the additional barcodes present in the extended count matrix are not cell-associated barcodes but barcodes of beads which have captured noise.
Filtering	Minimum number of cells expressing a gene	3	Genes which are strictly below this threshold will be excluded for the post-processing run. The default value is the most permissive value, however it is recommended to change it depending on your application (a commonly used value is 3).

Category	Parameter	Default value	How to set it ?
Filtering	Minimum value for number of genes per cell	0	Cells which are strictly below this threshold will be excluded for the post-processing run. The default value is the most permissive value, however it is recommended to change it depending on your application (a commonly used value is 200). Note that a barcode with very few genes may represent a low-quality cell or a barcode not coupled to any cell.
Filtering	Maximum value for number of genes per cell	+Inf	Cells which are strictly above this threshold will be excluded for the post-processing run. The default value is the most permissive value, however it is recommended to change it depending on your application (a commonly used value is the permissive value +Inf). Note that a barcode with significantly more genes may represent a cell multiplet, and that cell multiplets are not automatically excluded by the pre-processing pipeline.
Filtering	Minimum value for number of transcripts per cell	0	Cells which are strictly below this threshold will be excluded for the post-processing run. The default value is the most permissive value, however it is recommended to change it depending on your application (a commonly used value is the permissive value 0). Note that a barcode with very few transcripts may represent a low-quality cell or a barcode not coupled to any cell.
Filtering	Maximum value for number of transcripts per cell	+Inf	Cells which are strictly above this threshold will be excluded for the post-processing run. The default value is the most permissive value, however it is recommended to change it depending on your application (a commonly used value is the permissive value +Inf). Note that a barcode with significantly more transcripts may represent a cell multiplet, and that cell multiplets are not automatically excluded by the pre-processing pipeline.
Filtering	Minimum value for percentage of mitochondrial transcripts per cell (%)	0	Cells which are strictly below this threshold will be excluded for the post-processing run. The default value is the most permissive value and is applicable to most applications (a commonly used value is the permissive value 0).

Category	Parameter	Default value	How to set it ?
Filtering	Maximum value for percentage of mitochondrial transcripts per cell (%)	+Inf	Cells which are strictly above this threshold will be excluded for the post-processing run. The default value is the most permissive value, however it is recommended to change it depending on your application (commonly used values range from 5 to 20, meaning 5% to 20%). Note that a cell with very high percentage of mitochondrial transcripts may represent a poor-quality cell or a dying cell, but some samples such as tumor samples may have a rather high mitochondrial gene expression.
Highly Variable Genes (HVG)	Keep all genes for DGE and visualization?	No	If set to "yes", then the post-processing run will still restrict to the HVG all the next steps, but the DGE results will be computed for all genes (HVG and non-HVG) and it will be possible to visualize and explore all genes (HVG and non-HVG) in the Rover page, though at the cost of increased runtime.
Highly Variable Genes (HVG)	HVG method	seurat_v3	Method applied for HVG detection. The "seurat_v3" method (selected by default) reproduces the R-implementations of Seurat v3 which takes original transcript counts as input (see more details here: https://scanpy.readthedocs.io/en/stable/generated/scanpy.pp.highly_variable_genes.html) with all default parameter values. The "seurat_v3" method applies HVG before data log-normalization, whereas the "Cytonaut_v3.1" method takes log-normalized counts as input. It is recommended to use "seurat_v3" method to be closer to Seurat's state-of-the-art method. Cell data normalization and log-transformation are applied using the formula ln2(10000*X+1)), where: X is the ratio of transcripts detected in the cell for the gene of interest and ln2() is the logarithm in base 2. Note that this same normalization and log-transformation formula is used for the visualization of gene expression level in Cytonaut Rover.
Highly Variable Genes (HVG)	Number of top genes	2000	Number of most highly variable genes to keep as input for the next steps covering PCA, Embedding, Clustering and DGE steps. Note that genes which

Category	Parameter	Default value	How to set it ?
			are not selected as highly variable will not be explorable in Rover. The default value set to 2000 is a commonly used value, but it may be changed depending on your application: a higher (resp. lower) gene diversity in the expected cell populations may require a higher (resp. lower) value.
Highly Variable Genes (HVG)	Apply mitochondrial effect regression?	Yes	Applied by default. If set to "Yes": regress out effects of percentage of mitochondrial transcripts detected (note that this function tends to overcorrect in certain circumstances as described here: https://github.com/theislab/scanpy/issues/526). In all cases, we regress out the effects of number of detected transcripts per cell.
Highly Variable Genes (HVG)	Maximum value for clipping after scaling	10	Maximum value to which scaled gene expression is clipped, set to 10 by default as commonly used. Note that setting this value to +Inf is equivalent to not clipping at all after scaling. Setting this value to 10 or a numerical value close to 10 can help reduce the effects of genes that are only expressed in a very small number of cells. Just before this clipping step, the scaling function is applied to the normalized and log-transformed gene expression values by performing mean centering (i.e. subtracting the mean across all cells) and then unit variance scaling (i.e. dividing by the standard deviation across all cells). Note that this scaling function is used to prepare the PCA step but is not applied for the visualization of gene expression in Cytonaut Rover.
Principal Component Analysis (PCA)	Minimum PCA variability (%) for the next steps (if applied as exclusive choice)	95 (but not applied)	Not applied by default. Minimum value for the percentage of cumulative explained variance of the PCA, allowing to automatically set the number of PCA dimensions from an initial number of 100 dimensions. The resulting number of PCA dimensions is used as input for cell embedding as well as for cell clustering. The default value set to 95 (for 95%) is a commonly used value allowing to automate the number of PCA dimensions, but it may be changed depending on your application: a higher (resp. lower) number of expected cell populations may require a higher (resp. lower) value.

Category	Parameter	Default value	How to set it ?
Principal Component Analysis (PCA)	Number of PCA dimensions for the next steps (if applied as exclusive choice)	10 (applied)	Applied by default. Explicit value for the number of PCA dimensions to be used as input for cell embedding as well as for cell clustering (if not applied, it is automatically determined from the minimum PCA variability parameter value). The default value set to 10 is a commonly used value in case you need to explicitly set the number of PCA dimensions (up to 100), but it may be changed depending on your application: a higher (resp. lower) number of expected cell populations may require a higher (resp. lower) value. Note that setting a too high value (e.g. 100) may cause non-relevant small cell clusters to be generated, which may then cause false cell populations to be manually annotated.
Embedding	Number of neighbors	10	Number of nearest neighbors used for cell embedding as well as for cell clustering (where the inter-cell distance used is the Euclidean distance in the PCA coordinate system with the previously generated number of PCA dimensions). The default value set to 10 is a commonly used value, but it may be changed (from 5 to 100) depending on your application: this parameter controls the balance between local structure (lower values) and global structure (higher values).
Embedding	min_dist parameter of UMAP method (if UMAP applied)	0.5 (applied)	Minimum distance value used by the UMAP algorithm for cell embedding. The default value set to 0.5 is a commonly used value, but it may be changed (from 0 to 1) depending on your application: this parameter represents the minimum distance between points in the low-dimensional space (2D or 3D if applied), with higher (resp. lower) values leading to more loosely (resp. more tightly) packed embeddings. It is recommended to apply at least UMAP embedding (and optionally t-SNE embedding) as UMAP method is faster, less stochastic, and better preserves both local and global distances between cells. However, UMAP methods tends to product more compact results than t-SNE method.
Embedding	2D or 3D exclusive choice	2D (applied)	2D is applied by default.

Category	Parameter	Default value	How to set it ?
	for UMAP method (if UMAP applied)		If 3D is selected then the UMAP embedding projects the cells in a 3D dot plot which can be visualized using Cytonaut™ Rover. For a more reliable cell cluster annotation it is recommended to keep UMAP embedding in 2D, while 3D may be used for 3D exploration purposes or attractive rendering.
Embedding	Perplexity parameter of t-SNE method (only in 2D) (if t-SNE applied)	30 (not applied)	Perplexity value used by the t-SNE algorithm for cell embedding. The default value set to 30 is a commonly used value, but it may be changed (up to 50 for denser or larger data) depending on your application: this parameter controls the balance between local structure (lower values) and global structure (higher values). The parameter is, in a sense, a guess about the number of close neighbors each point has.
Clustering	Resolution parameter of Louvain method (if Louvain applied as exclusive choice)	0.8 (applied)	Resolution value used by the Louvain method which is applied by default as it is the most commonly used cell clustering algorithm. The default value set to 0.8 is a commonly used value, but it may be changed depending on your application: a higher (resp. lower) resolution value will lead to a higher (resp. lower) number of clusters and to smaller (resp. larger) cluster sizes. Note that setting a too high (resp. too low) value may cause over-segmentation (resp. under-segmentation) of cells, which may then cause false positives (resp. false negatives) in cell population detection.
Clustering	Resolution parameter of Leiden method (if Louvain applied as exclusive choice)	0.8 (not applied)	Resolution value used by the Leiden method which is not applied by default as it is less commonly used than Louvain method for cell clustering. The default value set to 0.8 is a commonly used value, but it may be changed depending on your application: a higher (resp. lower) resolution value will lead to a higher (resp. lower) number of clusters. Note that setting a too high (resp. too low) value may cause over-segmentation (resp. under-segmentation) of cells, which may then cause false positives (resp. false negatives) in cell population detection.

Category	Parameter	Default value	How to set it ?
Differential Gene Expression (DGE)	Test method (exclusive choice: Mann-Whitney-Wilcoxon, t-test, t-test_overestim_var)	Mann-Whitney-Wilcoxon	Set to Mann–Whitney–Wilcoxon method by default. Statistical test method used for Differential Gene Expression (DGE), to detect for each cluster which genes have a statistically different expression in this cluster compared to all other clusters. Note that “Mann–Whitney–Wilcoxon” uses the Mann–Whitney U test, also called Wilcoxon rank-sum test (most commonly used method, recommended as it is a non-parametrical test). Besides, “t-test_overestim_var” overestimates variance of each group and is recommended for small sample size.

VII. HOW TO CONFIGURE BATCH CORRECTION

What we call “batch effect” between multiple samples is a non-biological variability between these samples which manifests itself by misaligned 2D-projected cells belonging to distinct samples but common cell type (e.g. non-overlapping points in the UMAP view) and by non-relevant automated clusters capturing technical variability in addition to biological variability.

If no batch correction is applied and if the samples selected in the post-processing run are not a simple set of technical replicate samples, then there is a risk of “batch effect” (which can be check in Cytonaut Rover with the “sample_name” cell attribute) resulting in wrong DGE results and compromised biological interpretation.

The goal of batch correction (also called sample integration) is to group samples into batches according to their experimental conditions and to remove non-biological variability between the batches (called batch effect), while maximizing the biological diversity within each batch.

- You only need to define as input the "batch condition" of each selected sample (by entering a condition name) in order to group in a common batch the samples belonging to the same experimental condition.
- By default the sample condition is set to the sample name, meaning that each sample is considered as a distinct batch with a risk of over-correction, so it is recommended to edit the default sample condition name.
- If batch correction is de-activated, then all the selected samples will be pooled together and processed as one big sample, as if all samples were belonging to the same batch (meaning that no batch correction is applied).

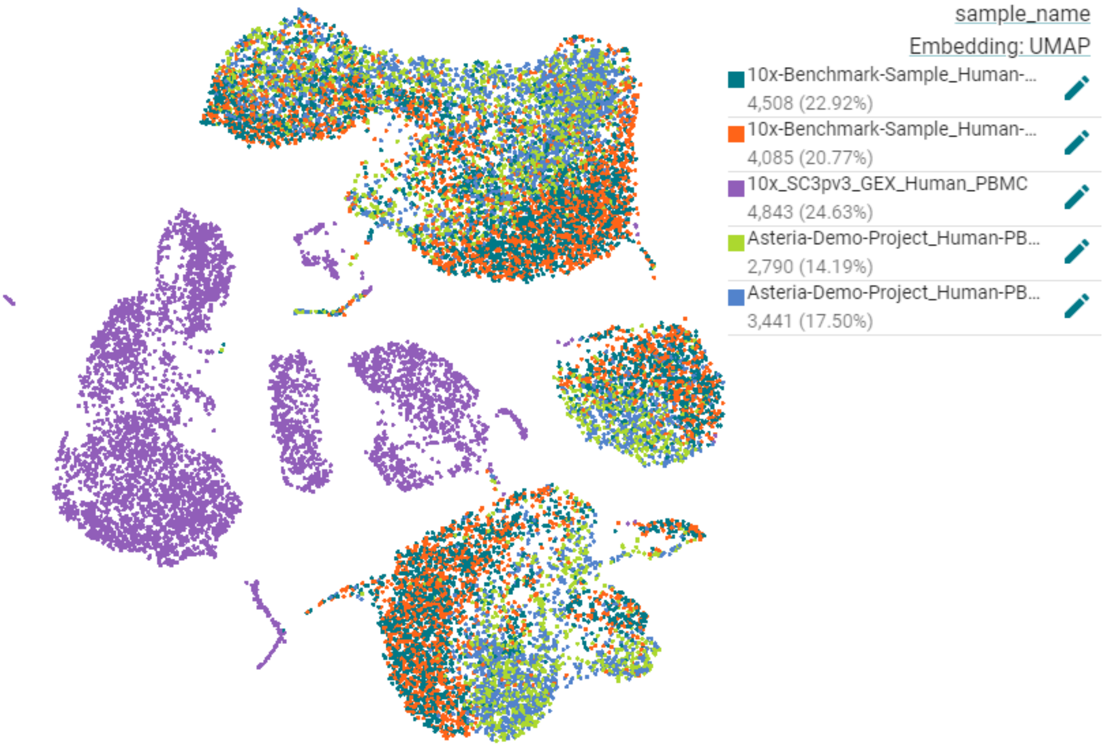
Note that the batch correction algorithm changes the way to perform post-processing analysis just after the HVG step (as highly variable genes are still identified in each sample separately), by transforming the PCA space and thus impacting the clustering step (based on PCA results), the DGE step (based on clustering results), and 2D embedding step (based on PCA results as well). The DGE results are impacted because they are based on batch-corrected clusters but they still takes as input the original transcriptomic signatures of the cells.

The batch correction method offered by Cytonaut is based on the state-of-the-art method "Harmony" proposed by Korsunsky et al. in 2019.

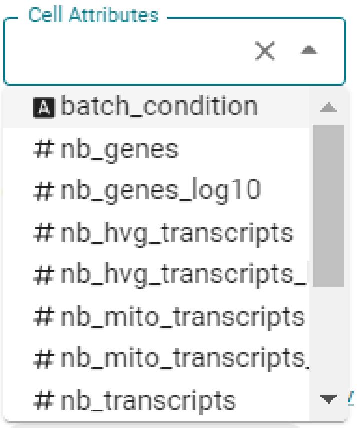
See this benchmark study for more information about batch correction methods.

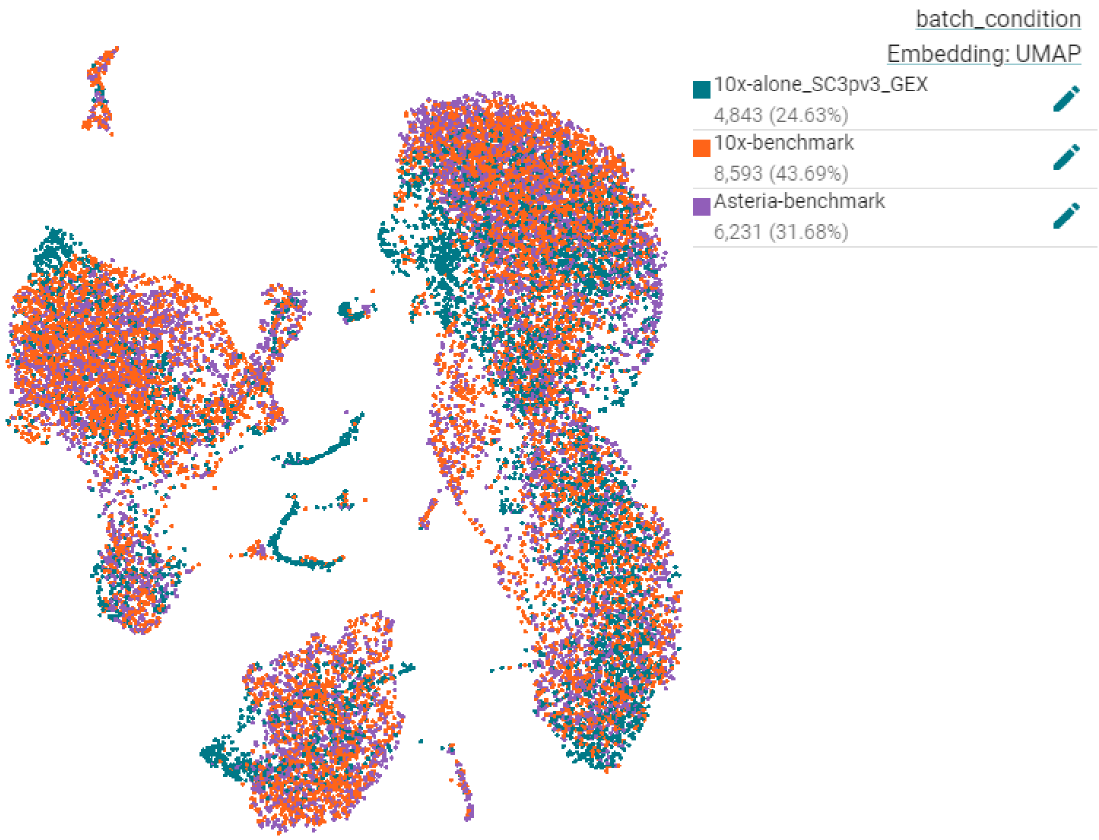
Let’s take for example 5 PBMC samples, coming from different donors, prepared with different sample preparation technologies (Asteria technology and 10x Genomics technology) and pre-processed with different software tools (Cytonaut and Cell Ranger).

- If we don’t apply batch correction for the post-processing run, then we see an obvious batch effect in the generated Rover page when we select the cell attribute “sample_name” (where only the technical replicate samples are aligned together):



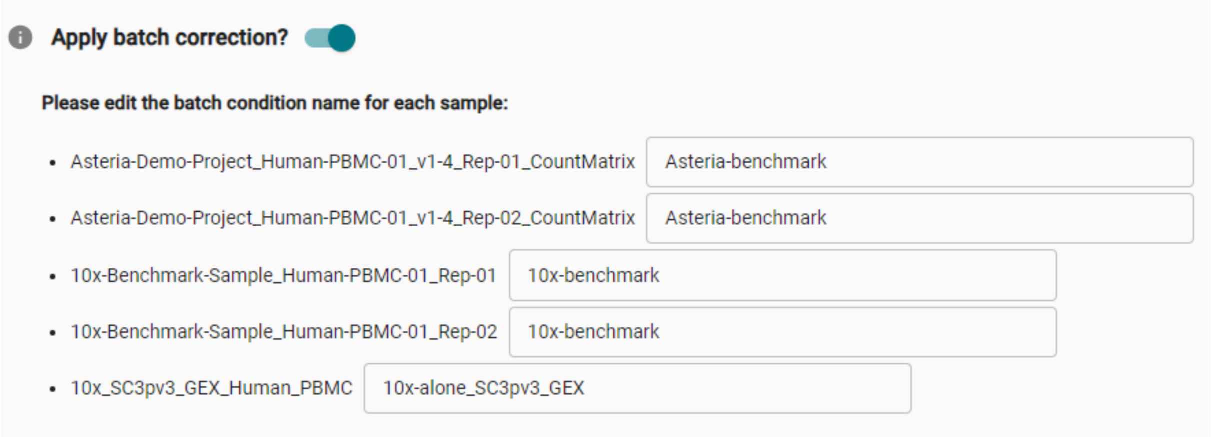
- If we apply batch correction for the post-processing run, by grouping the samples by condition, then we remove the batch effect, as shown in the generated Rover page when we select the new available cell attribute “batch_condition”:





We have grouped the 5 samples into 3 batches named as follows:

- “Asteria-benchmark” (two technical replicate samples from a donor A, prepared with Asteria technology, and pre-processed with Cytonaut v1.4);
- “10x-benchmark” (two technical replicate samples from the same donor A, prepared with 10x technology, and pre-processed with Cytonaut v1.4);
- “10x-alone_SC3pv3_GEX” (one sample coming from a distinct set of donors, prepared with 10x technology, and pre-processing with Cell Ranger v7.0).

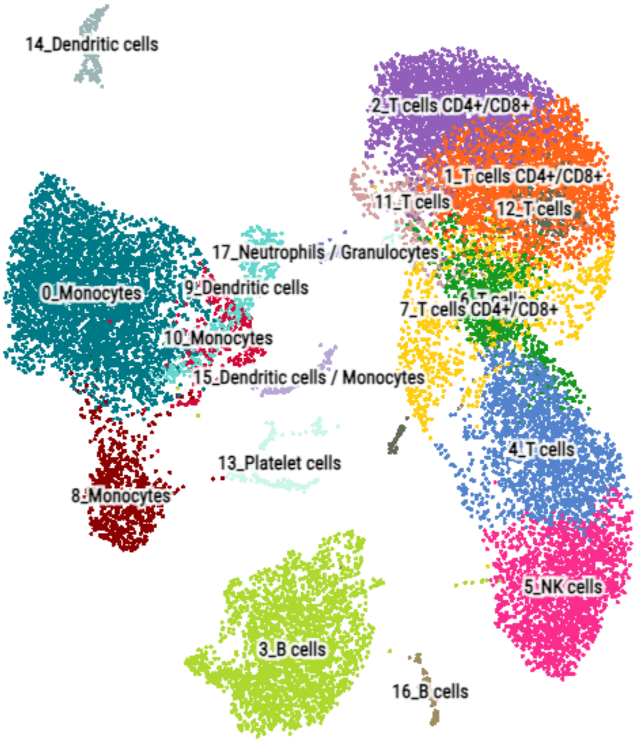


Apply batch correction? ●

Please edit the batch condition name for each sample:

Asteria-Demo-Project_Human-PBMC-01_v1-4_Rep-01_CountMatrix	Asteria-benchmark
Asteria-Demo-Project_Human-PBMC-01_v1-4_Rep-02_CountMatrix	Asteria-benchmark
10x-Benchmark-Sample_Human-PBMC-01_Rep-01	10x-benchmark
10x-Benchmark-Sample_Human-PBMC-01_Rep-02	10x-benchmark
10x_SC3pv3_GEX_Human_PBMC	10x-alone_SC3pv3_GEX

Thanks to batch correction, the samples are nicely aligned together with respect to their common PBMC cell types, resulting in relevant automated clusters and DGE results:



Note that selecting the new categorical cell attribute “batch_condition” will dynamically display the color of each of the batch condition in the CELL EMBEDDINGS menu, but also adapt the displayed statistics per batch condition in the menus DISTRIBUTIONS and COMPOSITION, exactly as if the categorical cell attribute “batch_condition” were a set of custom clusters.

VIII. HOW TO RUN CUSTOM DIFFERENTIAL GENE EXPRESSION

"Standard DGE" (Differential Gene Expression) compares the expression of the genes in each cluster with respect to all other clusters, where each standard cluster of cells has been automatically computed by running the post-processing step on one input sample or on a set of multiple input samples.

However, "Custom DGE" aims at comparing the expression of the genes between groups of cells which can be customized by the user, for example by grouping cells according to their sample conditions in order to detect statistical differences between test and control samples.

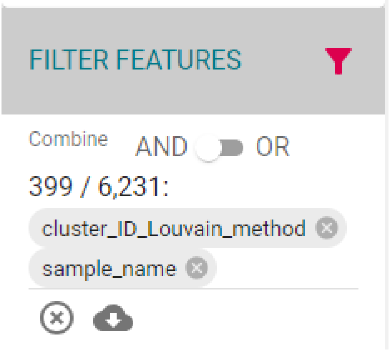
Here is how to run a Custom DGE on the sample(s) of interest of a given Cytonaut project:

1. It is first required to have performed at least one post-processing run on the project, where such run, named for example "RunID", includes the sample(s) of interest;

2. Then, open the Rover page which has been generated by the run "RunID", and define & export each group of cells to be compared for your Custom DGE, as follows:

   - Define your group of cells in Cytonaut Rover by filtering them in the menu CELL EMBEDDINGS, for example by clicking on one or several sample(s) or standard cluster(s), and/or by manually applying lasso selection or custom filter on gene expression and/or cell attribute values;

   - Your group of cells is now selected in the section FILTER FEATURES in the left panel of Cytonaut Rover. Click on the download icon to export it as a .txt file and rename this .txt file with an explicit filename which refers to your custom group of cells:

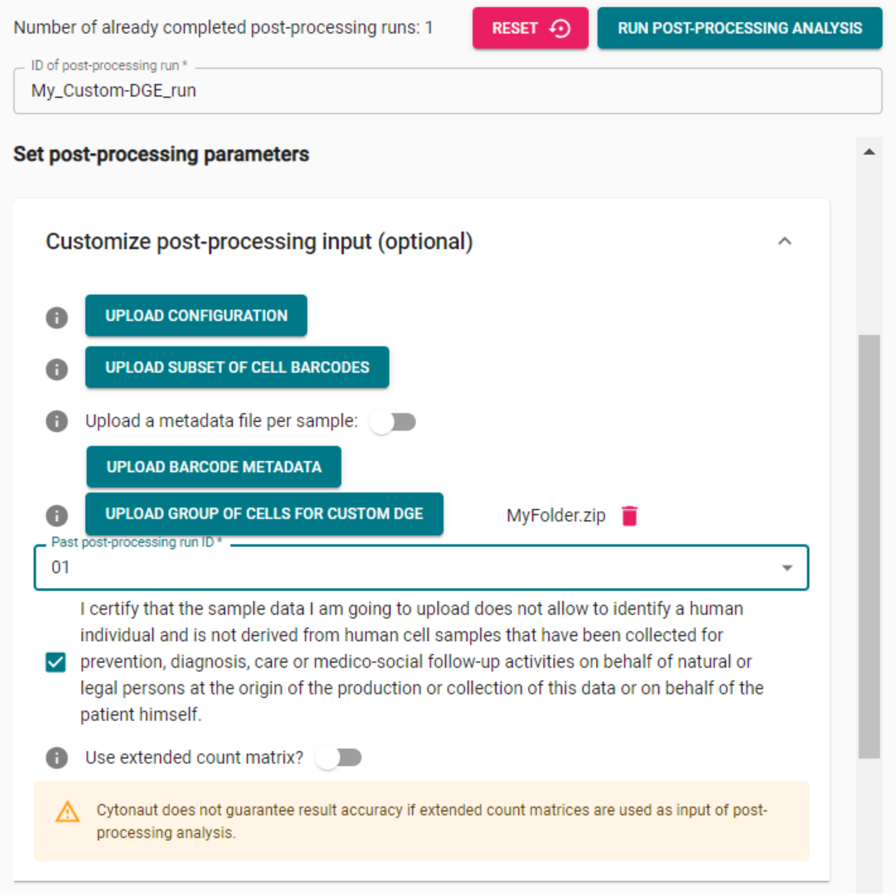


3. Once you have defined and exported all the group of cells you want to compare, you should have in your local computer one .txt file for each group of cells. Create a folder <MyFolder> (you can name it as you want) in your local computer and move all these .txt files in this folder, then compress the folder as a .zip file <MyFolder>.zip.

4. Come back to the Cytonaut menu Run Post-processing Analysis for your Cytonaut project to define a new custom post-processing run as follows:

   - Enter a unique name for the ID of this post-processing run;
   - In the right panel, check that the sample(s) of interest on which you want to apply the Custom DGE are checked;
   - Click in the first parameter section Customize post-processing input to expand it, then click on the button UPLOAD GROUP OF CELLS FOR CUSTOM DGE and select your compressed folder <MyFolder>.zip.

- You need then to select the Run ID of the previous post-processing run considered as the "parent run", and to check the checkbox which certifies that input data does not include sensitive data, as shown in the snapshot below:


Number of already completed post-processing runs: 1

RESET
RUN POST-PROCESSING ANALYSIS

ID of post-processing run *: My_Custom-DGE_run

Set post-processing parameters

Customize post-processing input (optional)

UPLOAD CONFIGURATION
UPLOAD SUBSET OF CELL BARCODES

Upload a metadata file per sample:

UPLOAD BARCODE METADATA
UPLOAD GROUP OF CELLS FOR CUSTOM DGE

MyFolder.zip

Past post-processing run ID *: 01

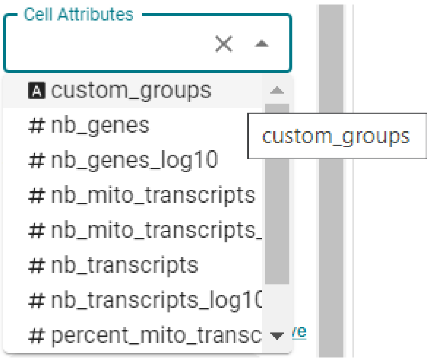
☑ I certify that the sample data I am going to upload does not allow to identify a human individual and is not derived from human cell samples that have been collected for prevention, diagnosis, care or medico-social follow-up activities on behalf of natural or legal persons at the origin of the production or collection of this data or on behalf of the patient himself.

Use extended count matrix?

Cytonaut does not guarantee result accuracy if extended count matrices are used as input of post-processing analysis.

5. As the Custom DGE will be computed from the results of the selected "parent run", there is no need to set other analysis parameters: you just need to click on the top button RUN POST-PROCESSING ANALYSIS in order to launch your Custom DGE.

6. When your custom post-processing run is completed, click on the button VIEW IN CYTONAUT ROVER to explore its generated Rover page. In the section SELECT FEATURES in the left panel of Cytonaut Rover, select the cell attribute "custom_groups" in Cell Attributes: this will dynamically display the color of each of your input custom groups of cells in the CELL EMBEDDINGS menu, and adapt the displayed statistics per custom group in the menus DISTRIBUTIONS and COMPOSITION, exactly as if the categorical cell attribute "custom_groups" were a set of custom clusters:



7. Go to the RESULTS menu of Rover, then you will see all the custom DGE results in addition the standard DGE results:

    - the custom DGE results in one versus all mode, named “custom_one-vs-all”, where each custom group is compared with respect to all the other custom groups;

    - for each combination of couples of custom groups: the custom DGE results in pairwise mode, named “custom_pw_XXX-vs-YYY”, where the test group XXX is compared with respect to the reference group YYY (note that the ratio of cells in the reference group is not provided in the pairwise custom DGE results because it corresponds to the ratio of cells in the test group when comparing YYY with respect to XXX).

    If you need to go back to the list of DGE results, click on the button BROWSE ALL RESULTS:

    Finally, you can export all your custom DGE results in .csv format the same way you export the standard DGE results, either from the RESULTS menu of Cytonaut Rover, or from the Download Results menu of Cytonaut.

IX. DOCUMENTATION AND SUPPORT

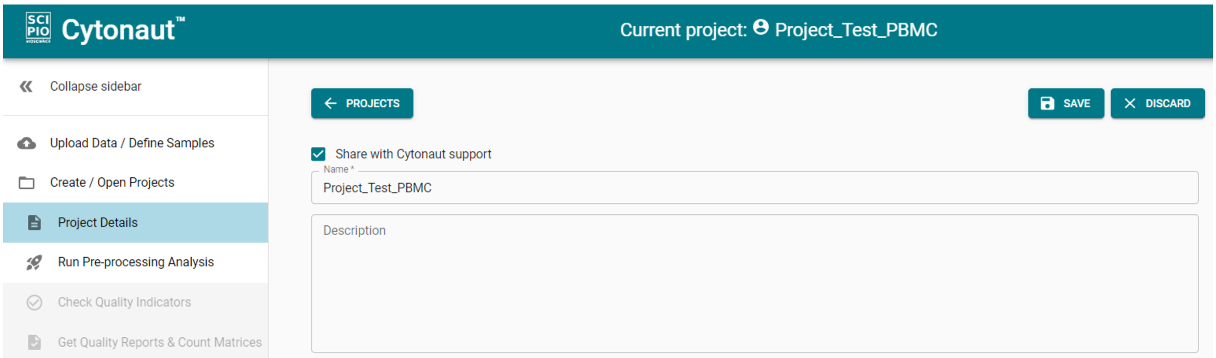
Customer Technical Support

Visit www.scipio.bio to access for the latest service and support information from Scipio bioscience or our nearest representative for your country.

For any support related to using Cytonaut™ software, contact us at:

support@cytonaut-scipio.bio

- If you need help on a specific Cytonaut project, you can authorize the technical support team of Cytonaut to access the project in read-only mode, by checking the option "Share with Cytonaut support" in the project edition page (Project Details menu > EDIT button) and then click on SAVE. Such explicit authorization can facilitate customer support (note that it is possible to unshare using the same checkbox):



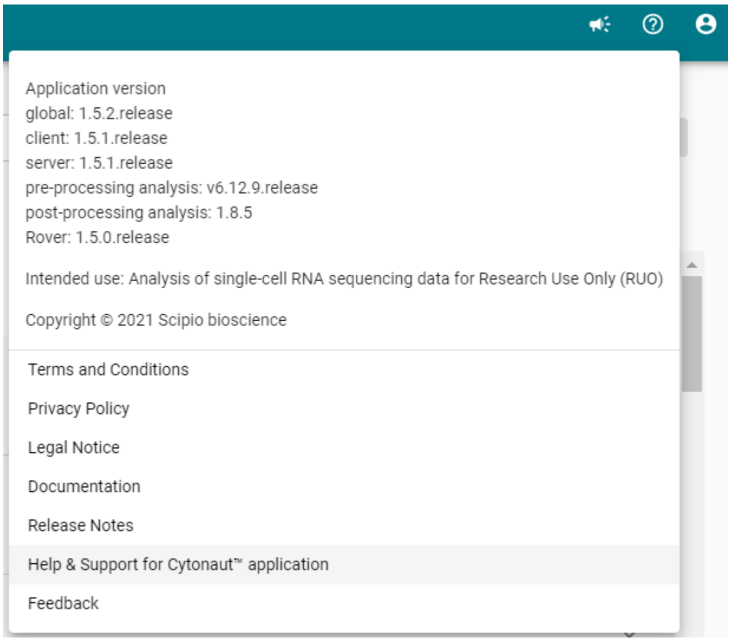
Cytonaut™
Current project: Project_Test_PBMC
Collapse sidebar
Upload Data / Define Samples
Create / Open Projects
Project Details
Run Pre-processing Analysis
Check Quality Indicators
Get Quality Reports & Count Matrices
PROJECTS
SAVE
DISCARD

☑ Share with Cytonaut support

Name * Project_Test_PBMC

Description

- Besides, you can fill a customized support form by clicking on "Help & Support for Cytonaut application" from the help icon in the top banner of Cytonaut:

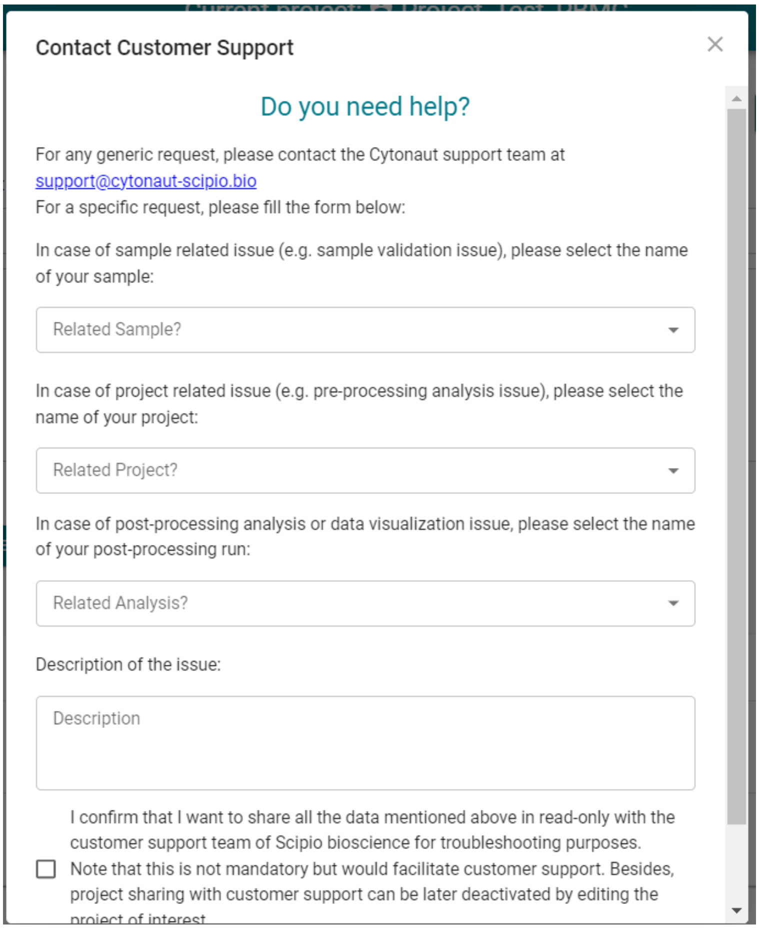


Application version
global: 1.5.2.release
client: 1.5.1.release
server: 1.5.1.release
pre-processing analysis: v6.12.9.release
post-processing analysis: 1.8.5
Rover: 1.5.0.release

Intended use: Analysis of single-cell RNA sequencing data for Research Use Only (RUO)

Copyright © 2021 Scipio bioscience

This form allows you to ask for help on a specific sample and/or project and/or pre-processing or postprocessing run. Besides, an optional agreement check box allows to share the related project in read-only mode with Cytonaut support team to facilitate troubleshooting (not mandatory and can be later disabled by editing the project of interest):



Contact Customer Support

Do you need help?

For any generic request, please contact the Cytonaut support team at support@cytonaut-scipio.bio
For a specific request, please fill the form below:

In case of sample related issue (e.g. sample validation issue), please select the name of your sample:

Related Sample?

In case of project related issue (e.g. pre-processing analysis issue), please select the name of your project:

Related Project?

In case of post-processing analysis or data visualization issue, please select the name of your post-processing run:

Related Analysis?

Description of the issue:

Description

□ I confirm that I want to share all the data mentioned above in read-only with the customer support team of Scipio bioscience for troubleshooting purposes. Note that this is not mandatory but would facilitate customer support. Besides, project sharing with customer support can be later deactivated by editing the project of interest

Sending this form will automatically send you a confirmation email with a unique request ID, as well as a notification email including this same unique request ID to Cytonaut support team.

Limited Product Warranty

SCIPIO BIOSCIENCE warrants its products as set forth in SCIPIO BIOSCIENCE's General Terms and Conditions of Sale at https://scipio.bio/legal/#general-terms-and-contitions-of-sale.

If you have any other questions, please contact us at support@scipio.bio.