Workshop:

Single Cell RNA-seq analysis using Cellenics

https://www.biomage.net

Oliver@biomage.net

Vicky@biomage.net

peter_kharchenko@hms.harvard.edu

GOAL

Q: How many of you have a single cell RNA-seq dataset that you are trying to / want to analyse?

-> Write ‘yes’ in the zoom chat

We will help you to get your analysis done!

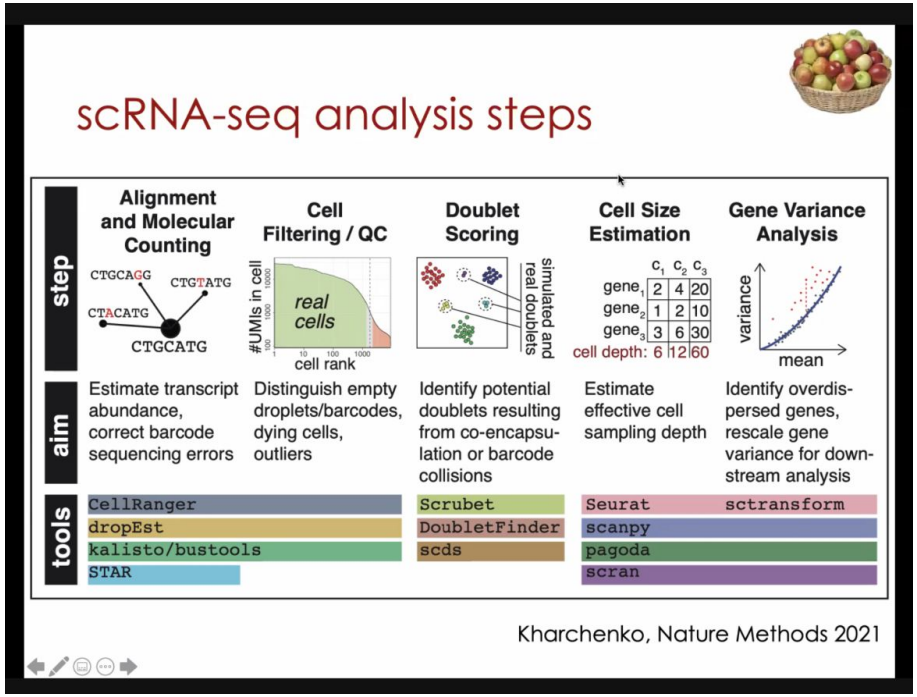
Book a free 45-minute meeting with Vicky (biologist) and Oliver (bioinformatician) to discuss your analysis needs:

https://calendly.com/vicky-morrison/cellenics-onboarding

INTRODUCTION

By Peter Kharchenko, Associate Professor of Biomedical Informatics at Harvard Medical School

https://www.nature.com/articles/s41592-021-01171-x



Review Article | Published: 21 June 2021

The triumphs and limitations of computational methods for scRNA-seq

Peter V. Kharchenko

Nature Methods **18**, 723–732 (2021) | Cite this article

15k Accesses | **1** Citations | **241** Altmetric | Metrics

Cellenics Overview

In this workshop you will be analysing a scRNA-seq dataset using Cellenics.

- Cellenics is cloud-based (AWS) - there is no need for you to install anything on your computer.
- Free for academic use
- Platform was released 16 weeks ago

We want to hear your feedback!

-> use the ‘Feedback’ button at the top

You may encounter issues/bugs

-> tell us and we will fix them!

If you need help now or at any point during the tutorial exercise, post a message in the zoom chat and one of the Biomage team will help.

Logging in

- We have created an account for each of you and we have already uploaded the example dataset in order to save time.

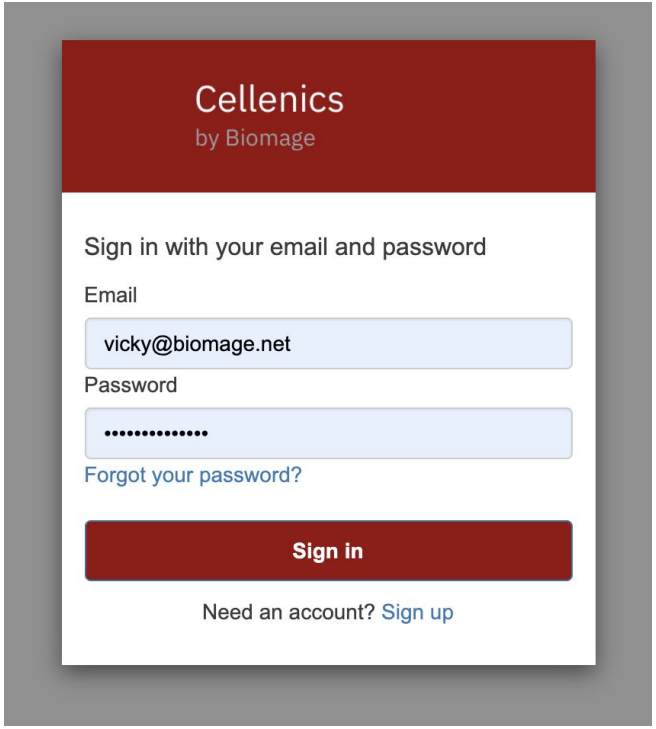
- Let's log in to Cellenics now using Chrome: https://scp.biomage.net/

You should have received your login details by email ~1 hour ago.

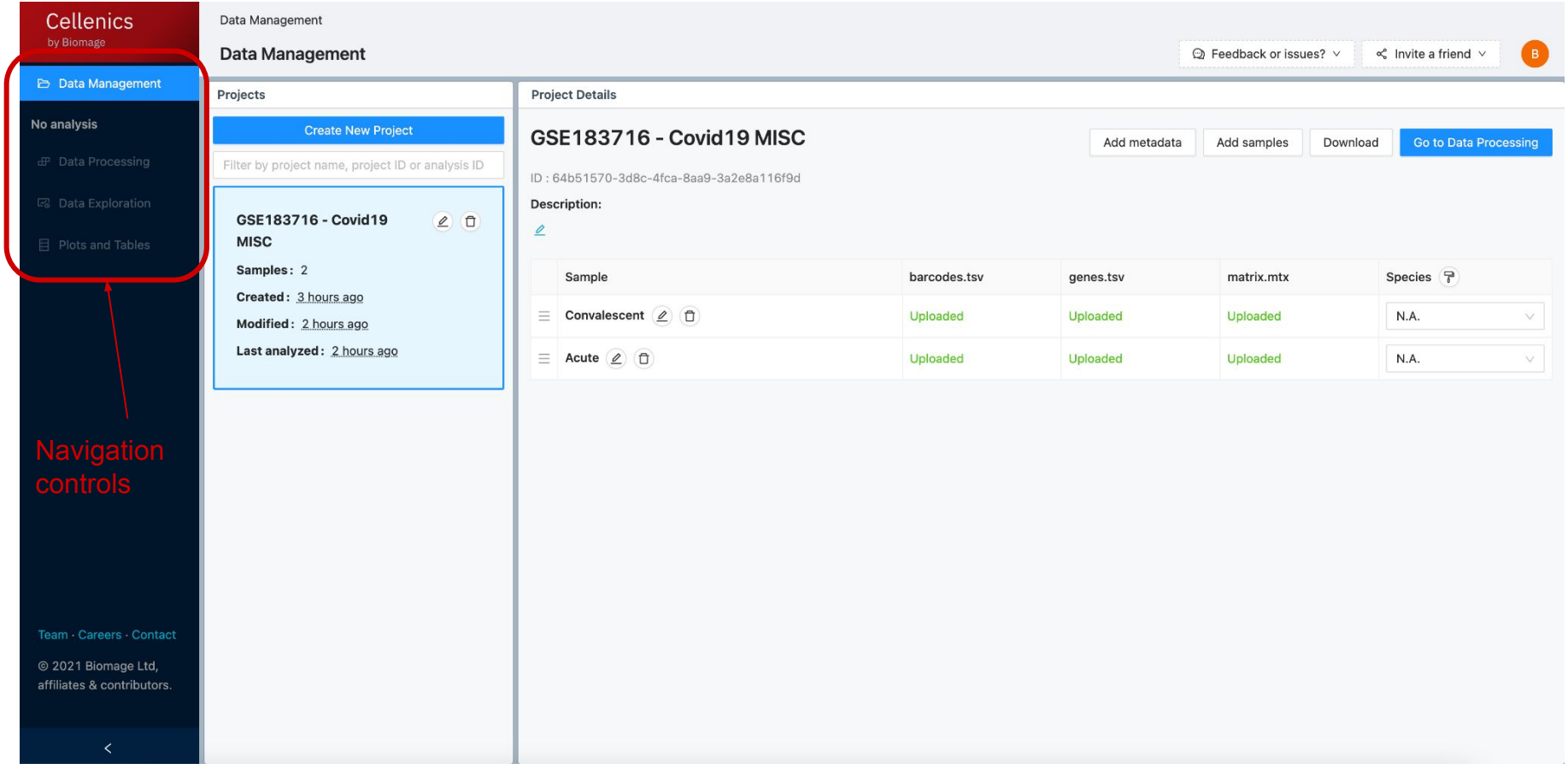
Do check your junk folder.

If you can't find your login details or have issues, let us know in the zoom chat.

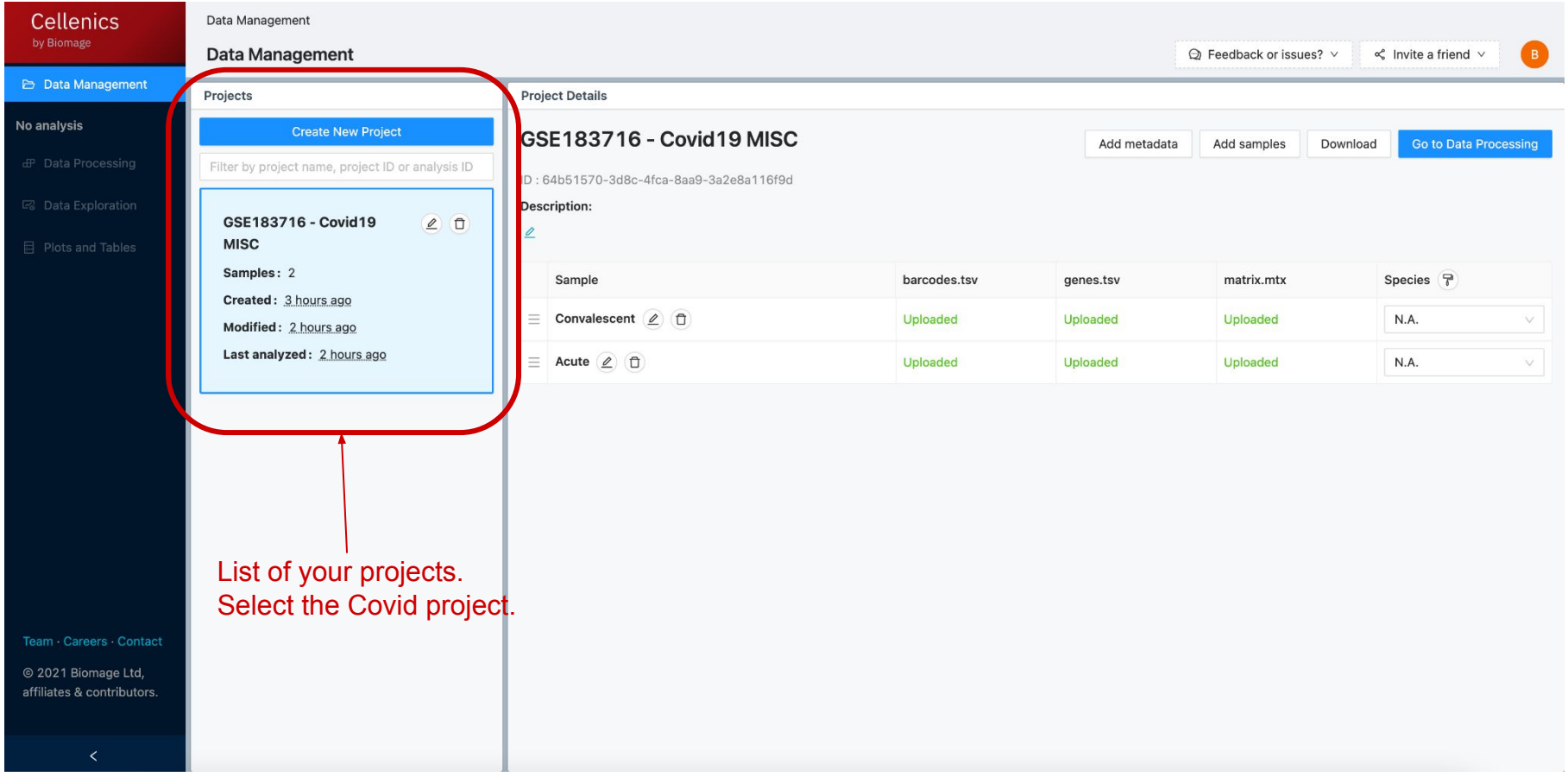
Once you have logged in, type "I'm in" in the zoom chat!



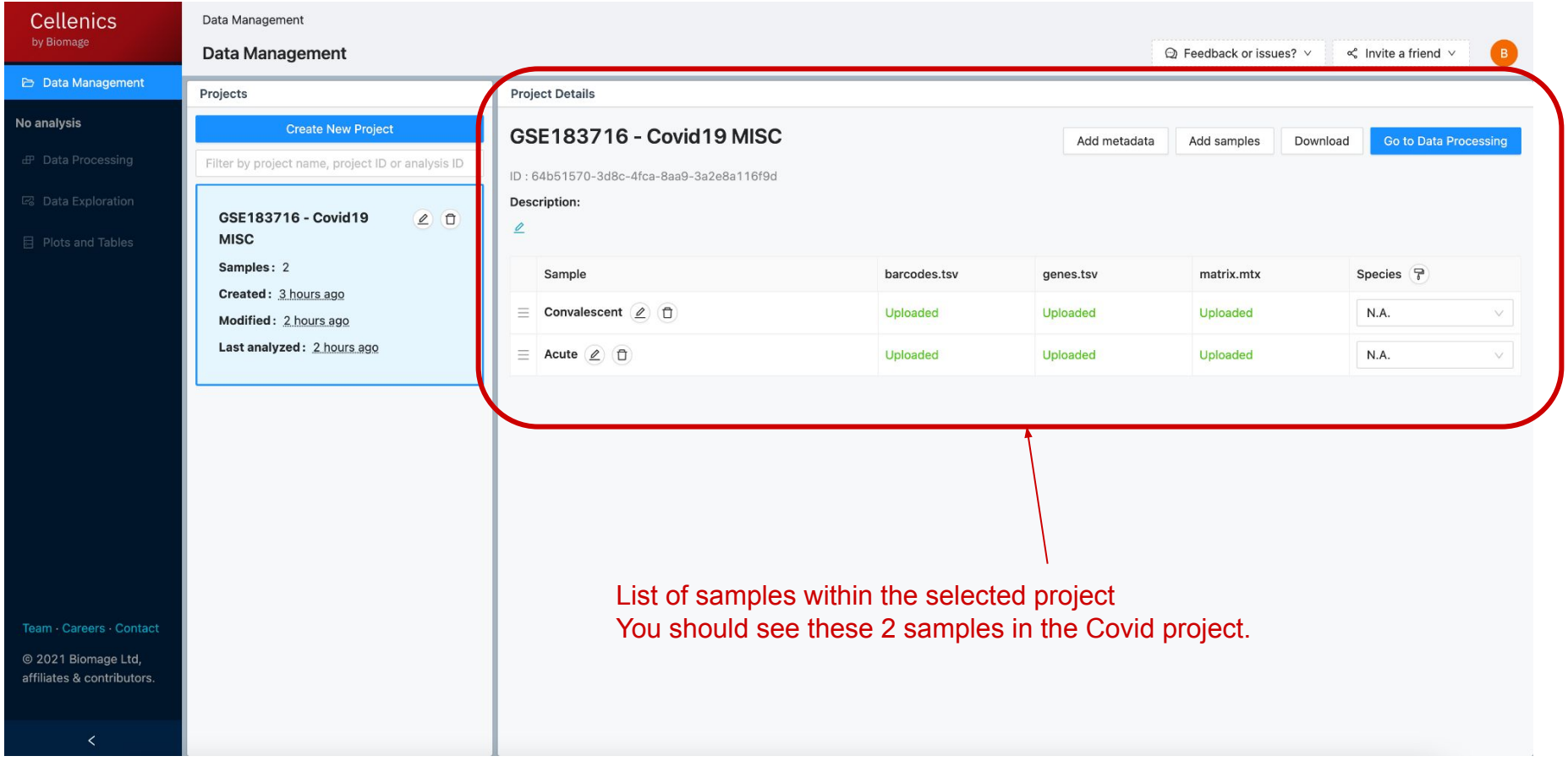
Data Management module



Data Management module

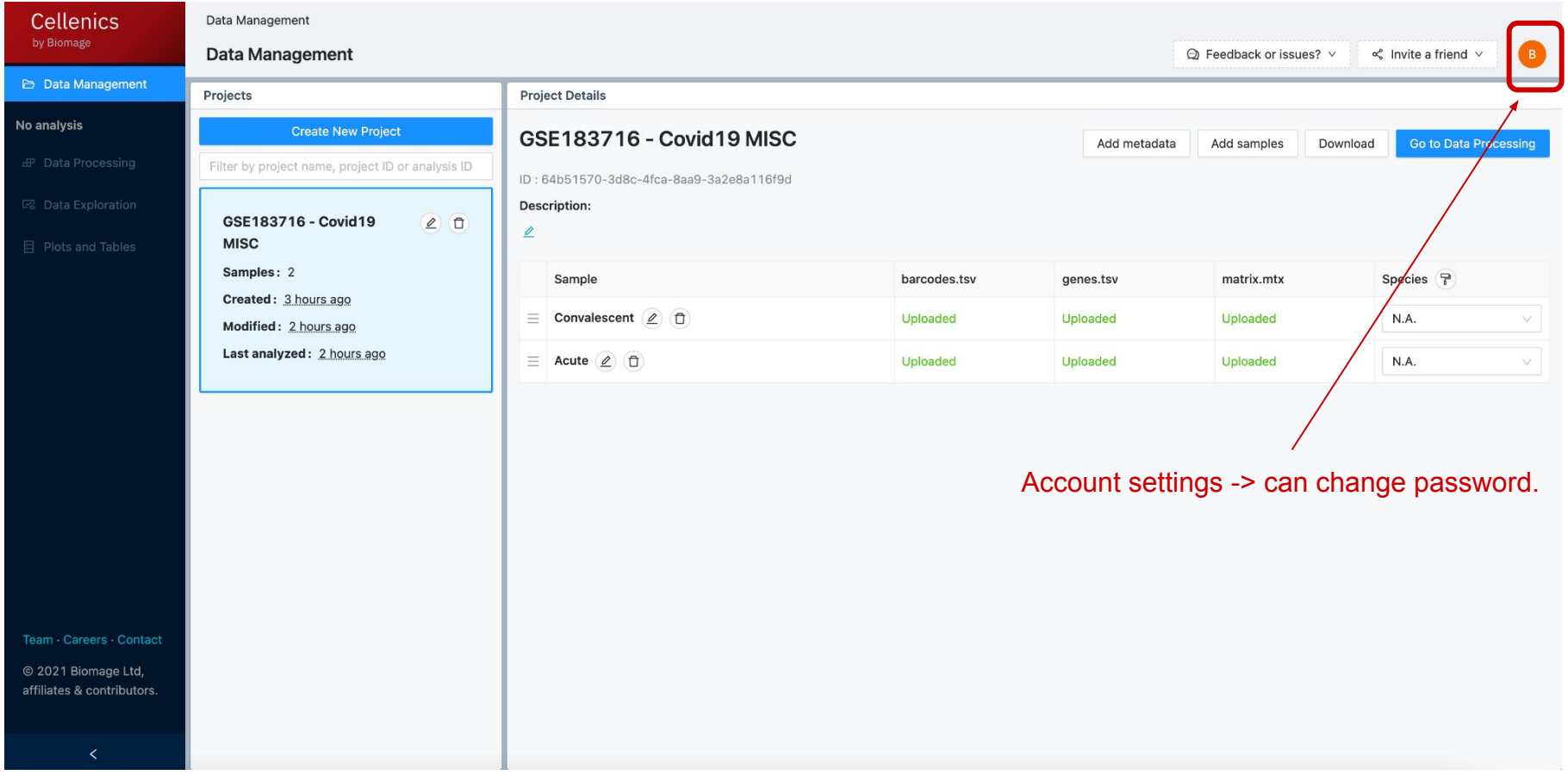


Data Management module



List of samples within the selected project
You should see these 2 samples in the Covid project.

Data Management module



Tutorial Exercise Dataset

Reference:

The innate and adaptive immune landscape of SARS-CoV-2-associated multisystem inflammatory syndrome in children (MIS-C) from acute disease to recovery.

GEO accession GSE183716

https://www.medrxiv.org/content/10.1101/2020.08.06.20164848v2.full.pdf

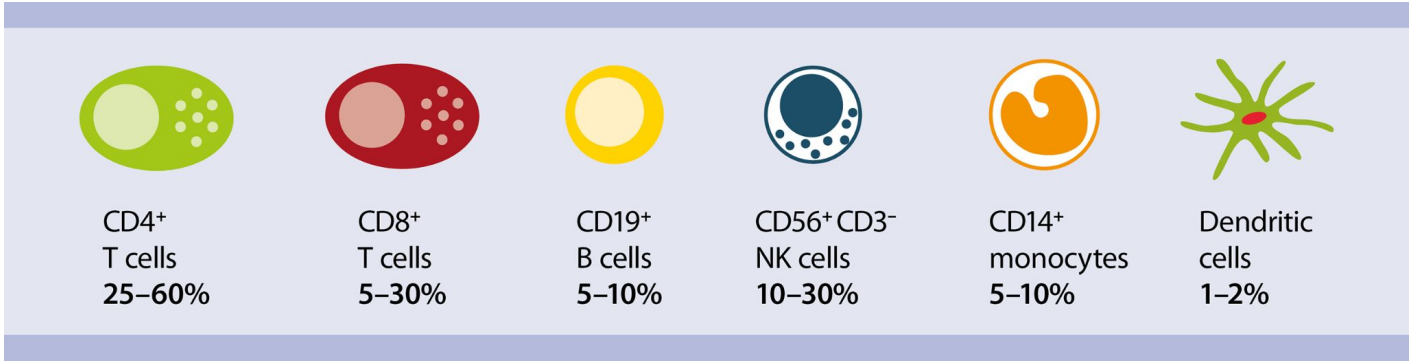
About the dataset:

Publicly available dataset, published on GEO on Sep 10, 2021
Two PBMC samples isolated from humans with COVID-19 associated MIS-C:

1. Acute patient
2. Convalescent post-treatment patient

PBMCs

- Critical components of the immune system
- Easily isolated from peripheral blood
- Widely used in areas relating to immunology, infectious disease, vaccine development, and others.



Common Cell Types in PBMCs: Lymphocytes, Monocytes and Dendritic cells.

https://www.miltenyibiotec.com/US-en/resources/macs-handbook/human-cells-and-organs/human-cell-sources/blood-human.html#gref

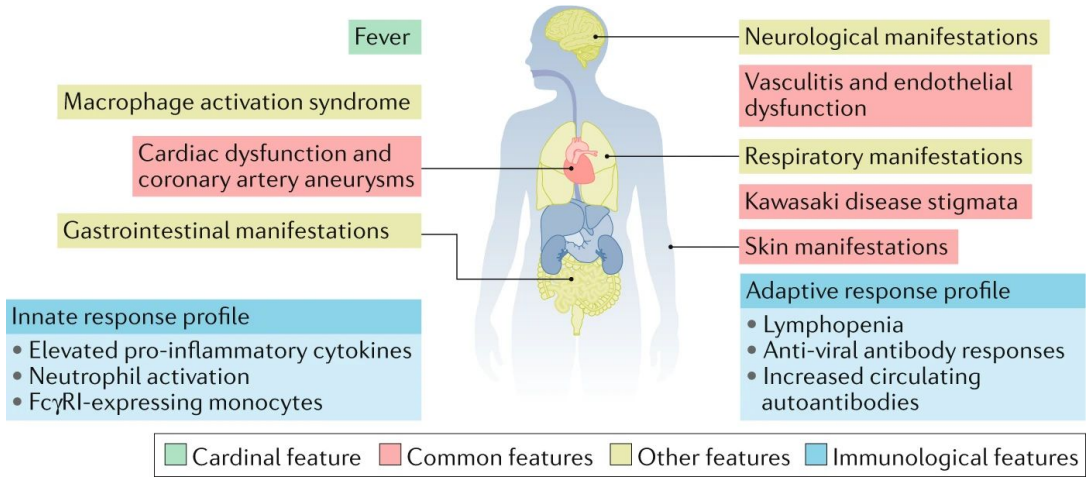
Multisystem inflammatory syndrome in children (MIS-C)

- Associated with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Usually develops 4-6 weeks after SARS-CoV-2 infection.

- Characterized by a marked inflammatory state in various organs.

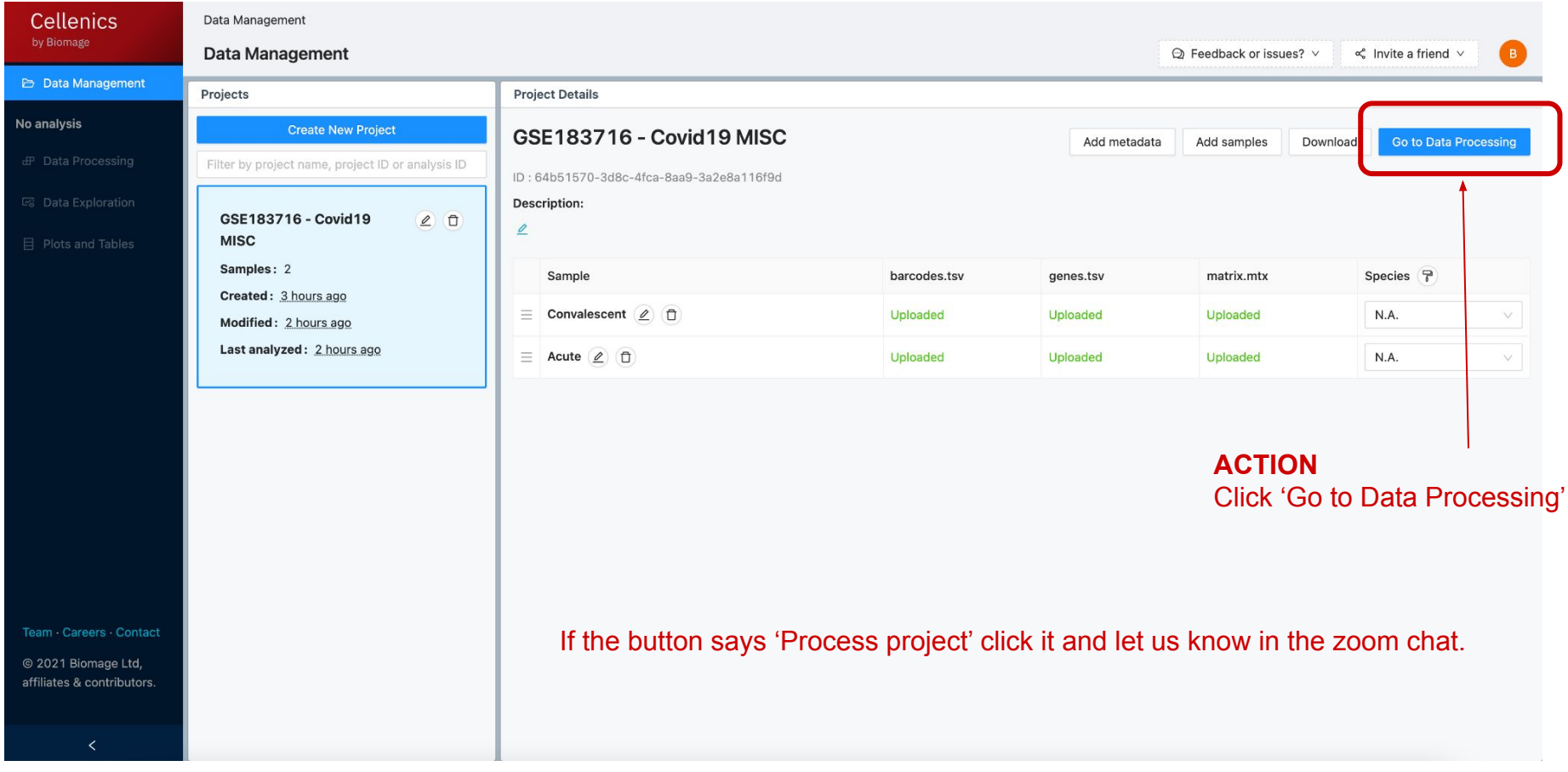
- Can be serious and potentially fatal: cardiac symptoms are common.

- Our understanding of this disease remains poor.



https://www.nature.com/articles/s41584-020-00566-y

Data Management module



ACTION
Click ‘Go to Data Processing’

If the button says ‘Process project’ click it and let us know in the zoom chat.

Data Processing module

Data processing pipeline is already applied to the dataset:

Steps 1-5: Filtering
-> to remove empty droplets, dead cells, doublets, etc.

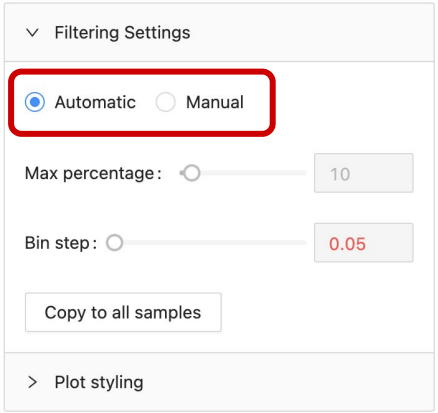
↓

Step 6: Data Integration & PCA
-> to control for batch effect & reduce complexity

↓

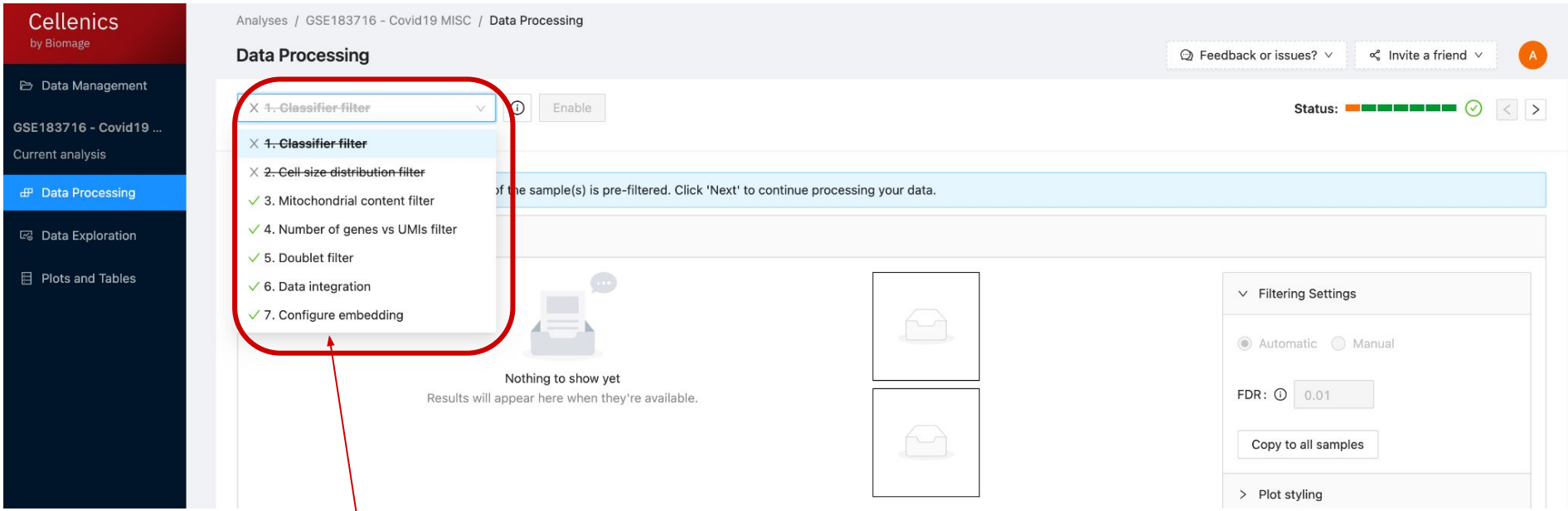
Step 7: Configure embedding
-> UMAP or t-SNE; clustering settings

Manual control over the settings:



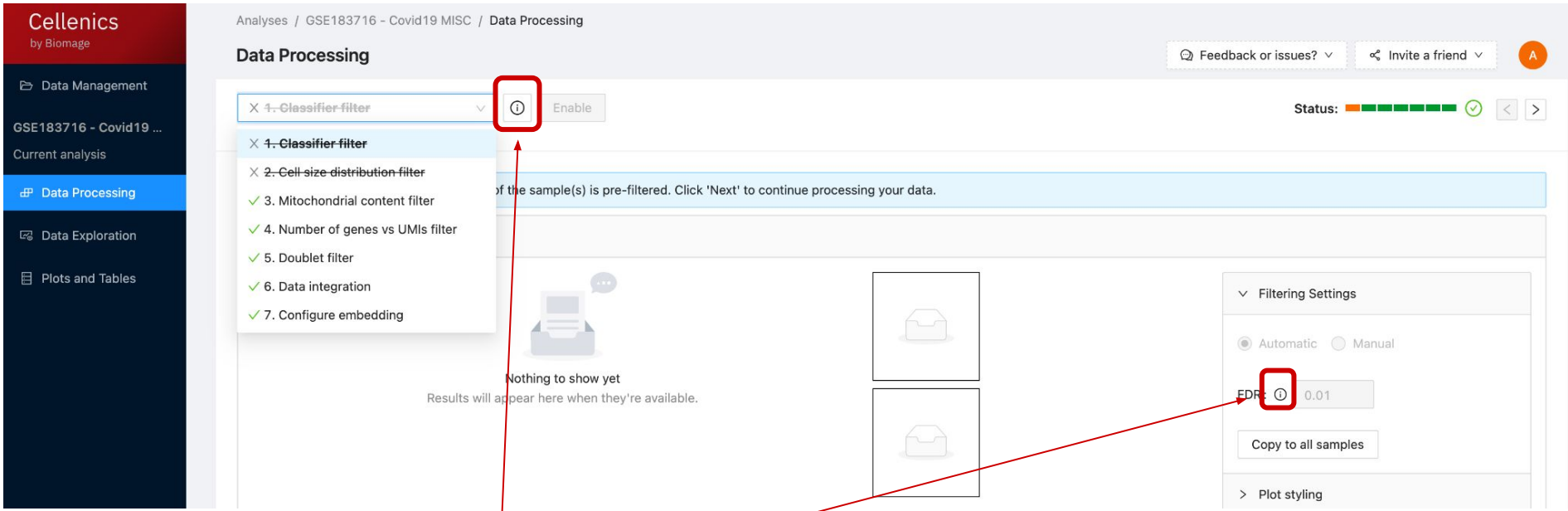
Please DO NOT make and apply (‘Run’) changes to the Data Processing module right now

Data Processing module



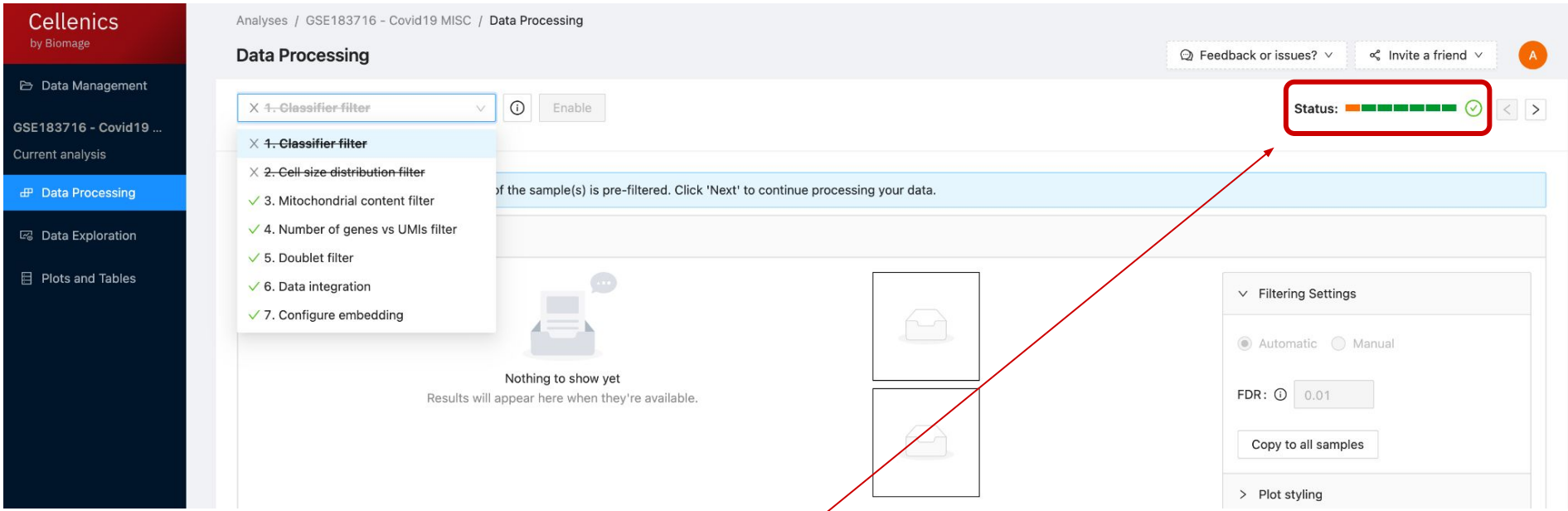
Dropdown menu shows the 7 steps

Data Processing module



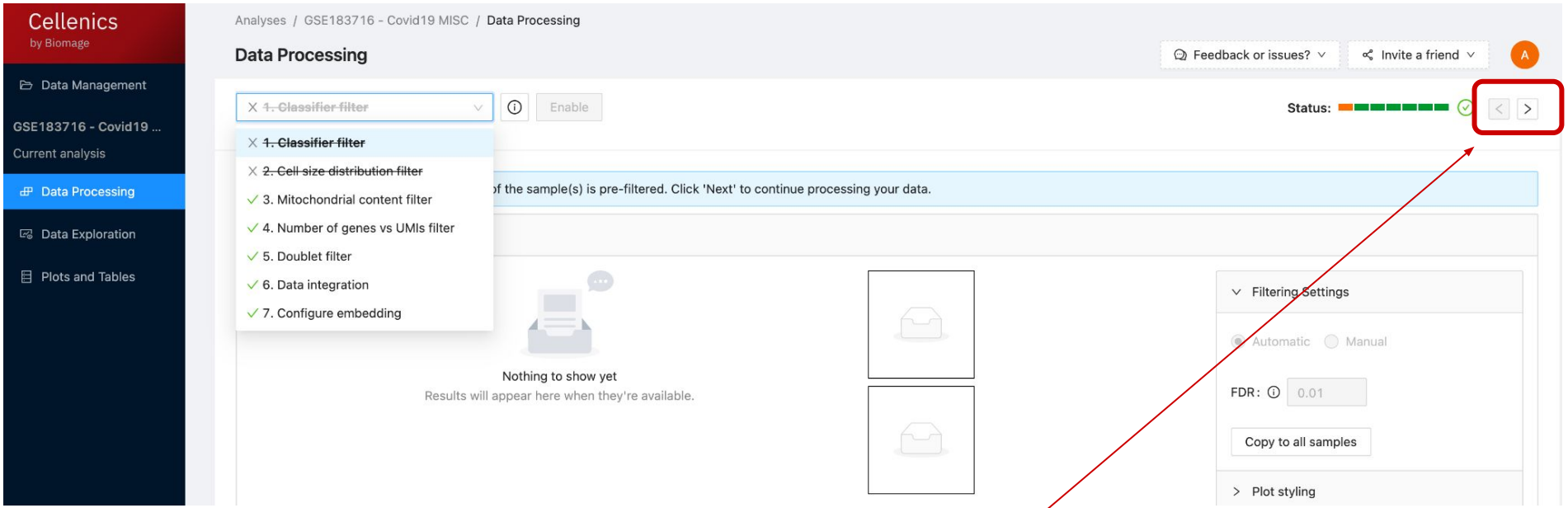
Explanation tooltips throughout

Data Processing module



Pipeline status indicator

Data Processing module



Navigate between data processing steps

Exercise 1 - Assessment of sample quality [8mins]

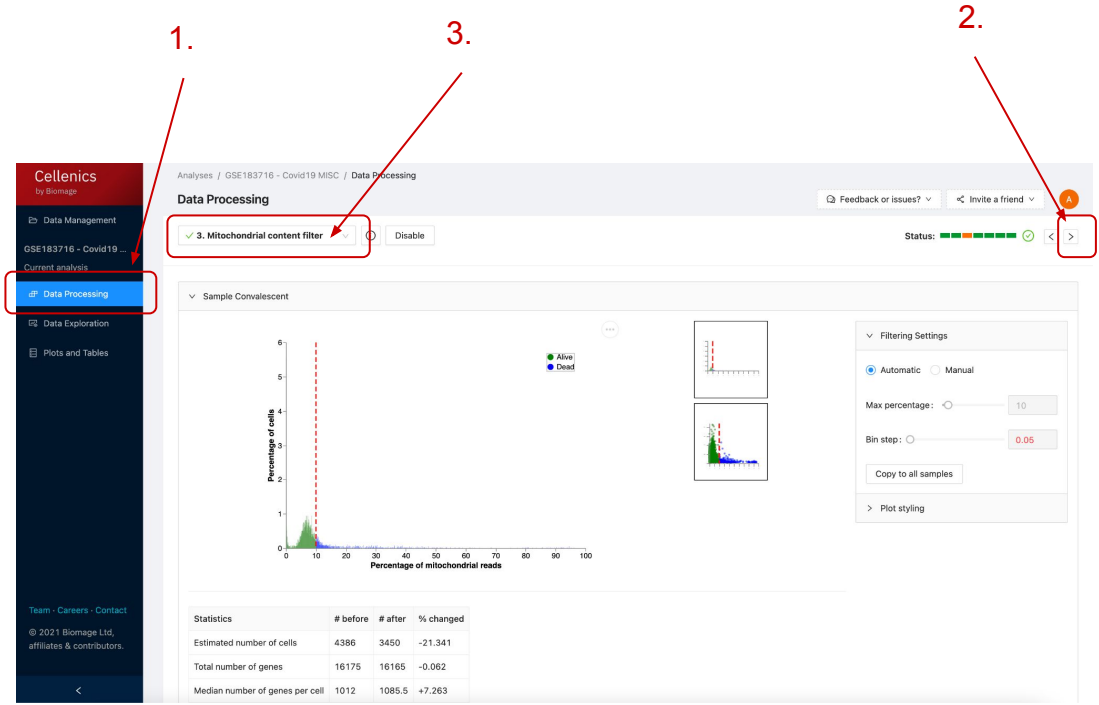
Determine whether the two samples differ significantly in quality.

a) Do they contain similar proportions of dead cells?
b) Do they contain similar proportions of doublets?
c) Are the 2 samples integrated well?

INSTRUCTIONS:

1. Go to Data Processing module
2. Use the ‘next’ button to navigate through the 7 steps
3. Look in detail at the mitochondrial content filter (step 3), doublet filter (step 5) and integration (step 6)

When you have completed it, write “Done 1” in chat.

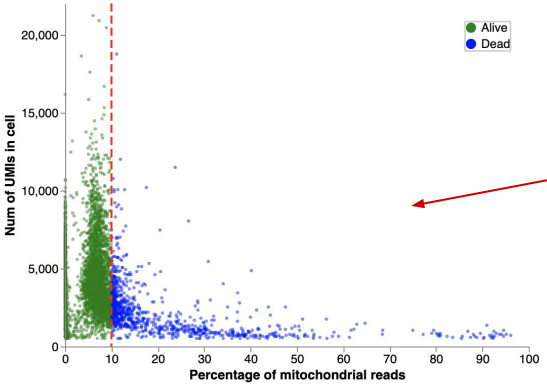


Answer 1 - Assessment of sample quality

Determine whether the two samples differ significantly in quality.

a) Do they contain similar proportions of dead cells?

convalescent

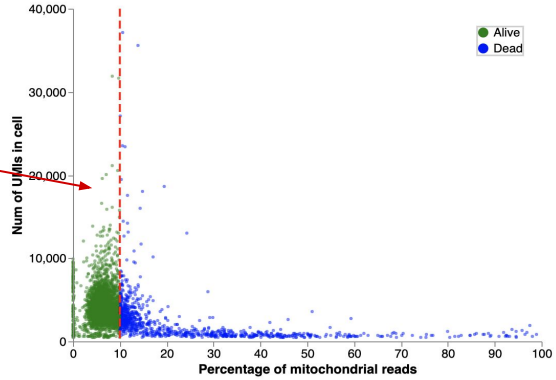


Statistics	# before	# after	% changed
Estimated number of cells	4479	3553	-20.674
Total number of genes	16031	16021	-0.062
Median number of genes per cell	1049	1094	+4.290
Median UMI counts per cell	3551	3877	+9.181

Data spread is similar

% of cells filtered out is similar

acute



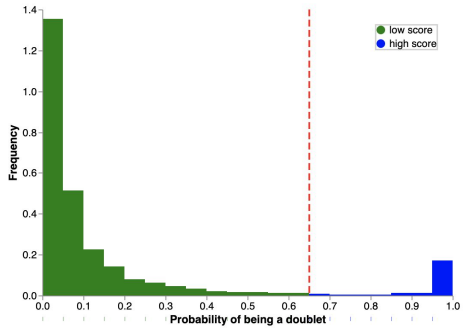
Statistics	# before	# after	% changed
Estimated number of cells	4386	3450	-21.341
Total number of genes	16175	16165	-0.062
Median number of genes per cell	1012	1085.5	+7.263
Median UMI counts per cell	3489	3954.5	+13.342

Answer 1 - Assessment of sample quality

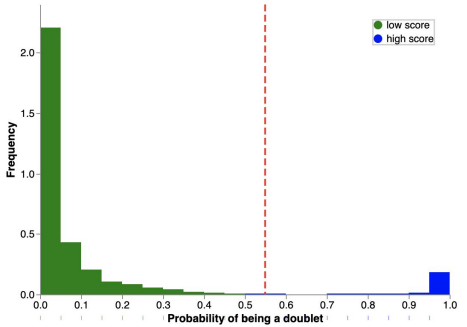
Determine whether the two samples differ significantly in quality.

b) Do they contain similar proportions of doublets?

convalescent



acute



Data spread is similar

Statistics	# before	# after	% changed
Estimated number of cells	2754	2536	-7.916
Total number of genes	16149	16118	-0.192
Median number of genes per cell	1196.5	1173.5	-1.922
Median UMI counts per cell	3998	3897	-2.526

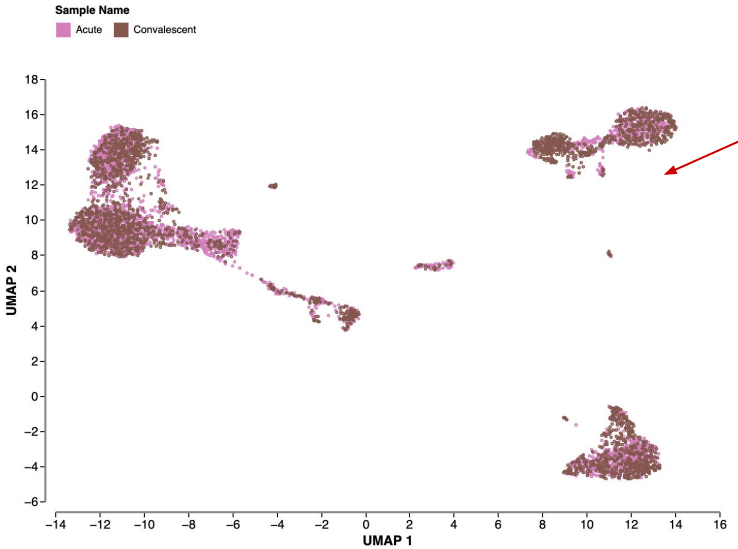
% of cells filtered out is similar

Statistics	# before	# after	% changed
Estimated number of cells	3406	3171	-6.900
Total number of genes	16009	15989	-0.125
Median number of genes per cell	1105	1089	-1.448
Median UMI counts per cell	3870.5	3747	-3.191

Answer 1 - Assessment of sample quality

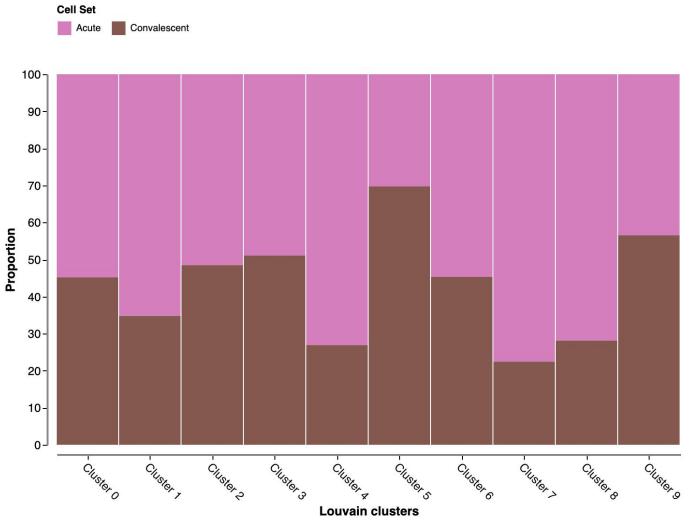
Determine whether the two samples differ significantly in quality.

c) Is the data well integrated?



The two samples show similar distribution on the UMAP embedding.

The frequency plot shows that all clusters contain cells from both samples.



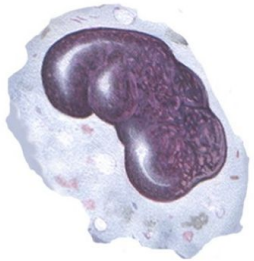
Data Exploration module

- fast, interactive data visualisation
- cluster annotation
- view gene expression in just a few clicks
- calculate differential expression between clusters or groups

-> Define what cell types are present -> annotate clusters

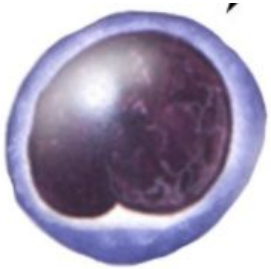
-> Find interesting things from the dataset

Exercise 2 - Dataset annotation

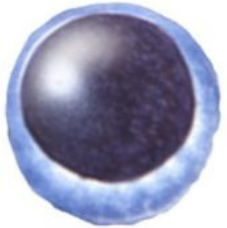


Monocytes

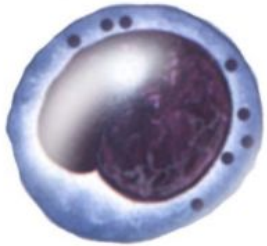
CD14+
CD16+



B cells

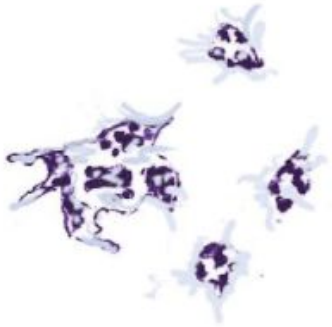


T cells



NK cells

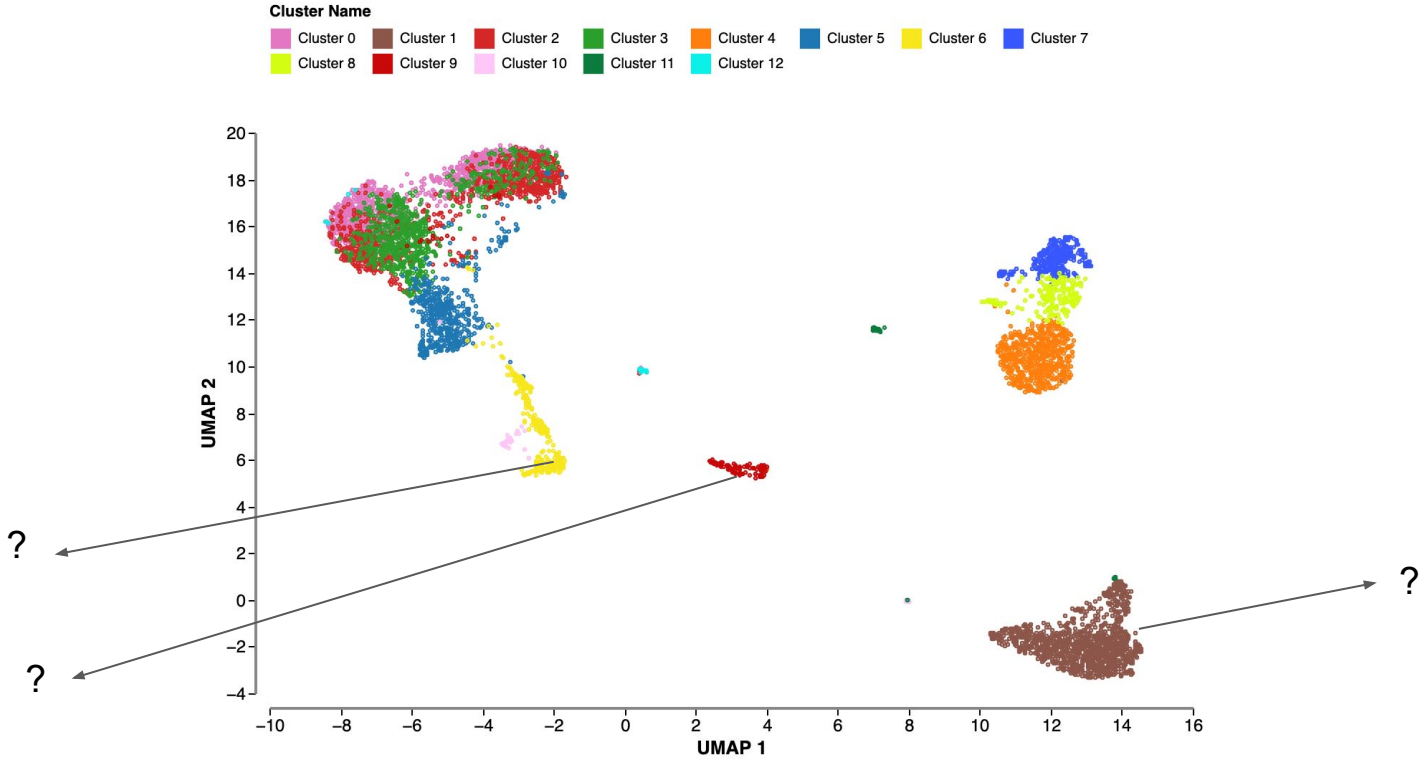
Erythrocytes



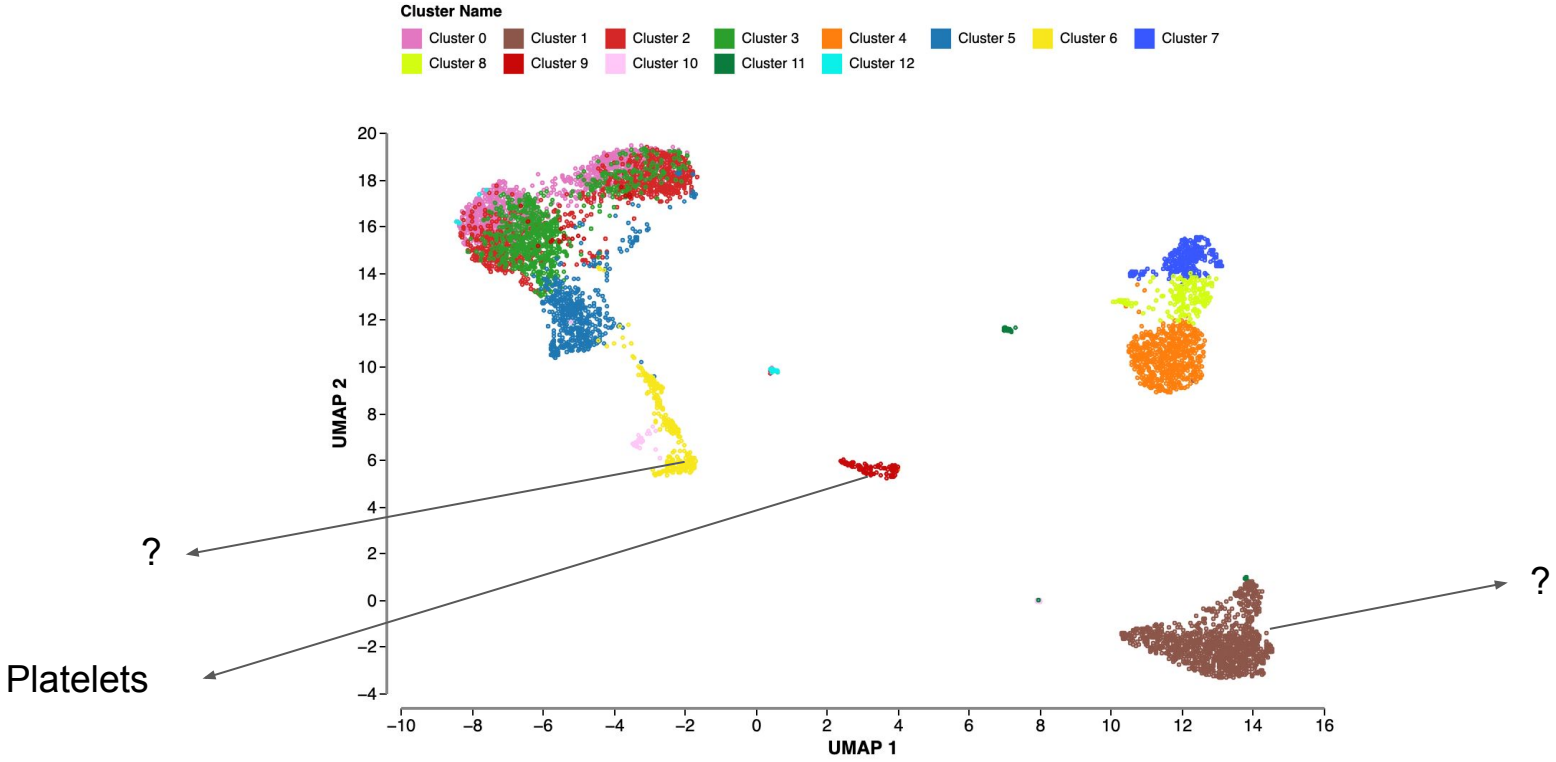
Platelets

https://www.cancer.gov/publications/dictionaries/cancer-terms/def/t-cell

Exercise 2 - Dataset annotation



Exercise 2 - Annotate these two unknown clusters [10mins]



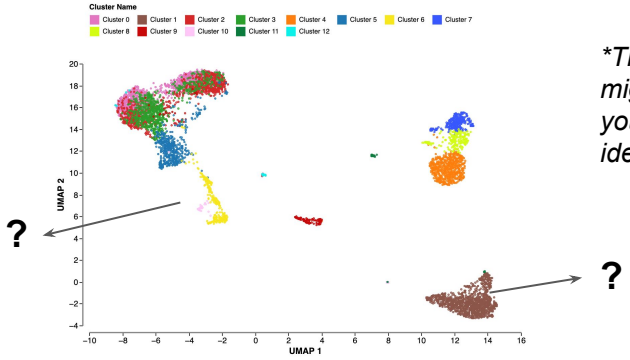
Exercise 2 - Annotate these 2 unknown clusters [10mins]

INSTRUCTIONS:

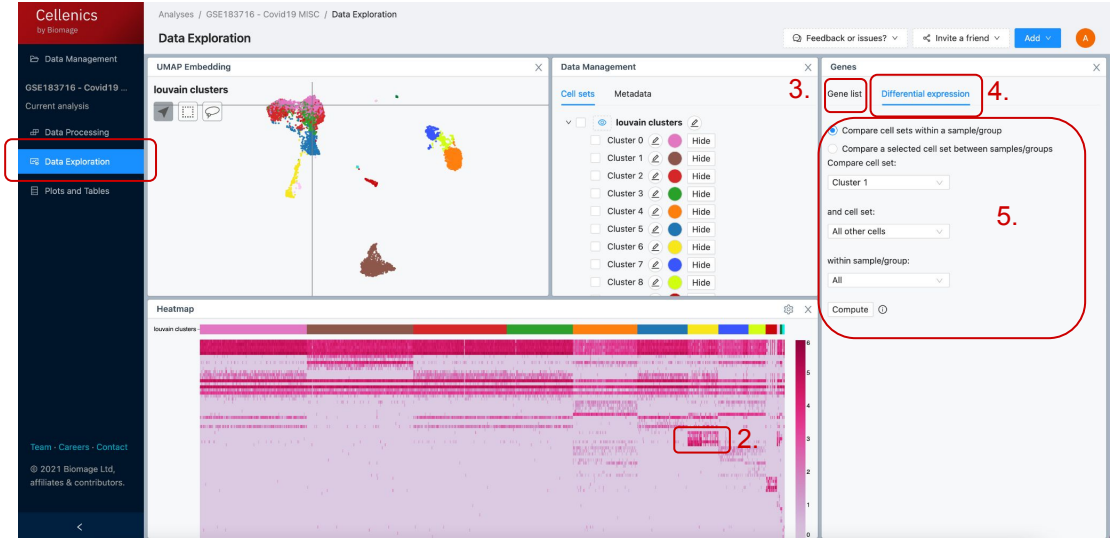
1. Go to Data Exploration module
2. For the **yellow cluster***: hover over the marker genes of this cluster in the heatmap.
3. Search for the relevant genes in the gene list.
4. For the **brown cluster***: Select the Differential Expression tab from the ‘Genes’ tile.
5. Select the relevant cluster; compare to all other cells; using all samples; click ‘Compute’. Look at the results (next slide).

When you have completed it, write “Done 2” in chat.

If you have time left, annotate another cluster of your choice.



**The cluster colors might be different but you will be able to identify the shapes*



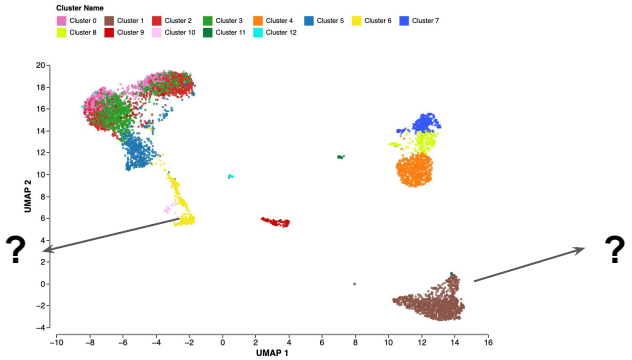
Exercise 2 - Annotate these 2 unknown clusters [10mins]

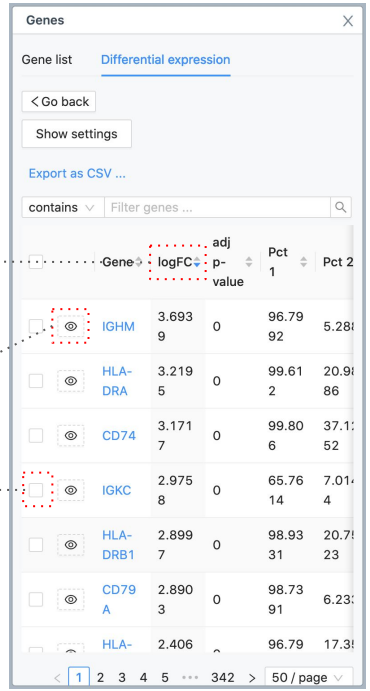
INSTRUCTIONS:

1. Go to Data Exploration module
2. For the **yellow cluster***: hover over the marker genes of this cluster in the heatmap.
3. Search for the relevant genes in the gene list.
4. For the **brown cluster***: Select the Differential Expression tab from the ‘Genes’ tile.
5. Select the relevant cluster; compare to all other cells; using all samples; click ‘Compute’. Look at the results (next slide).

When you have completed it, write “Done 2” in chat.

If you have time left, annotate another cluster of your choice.



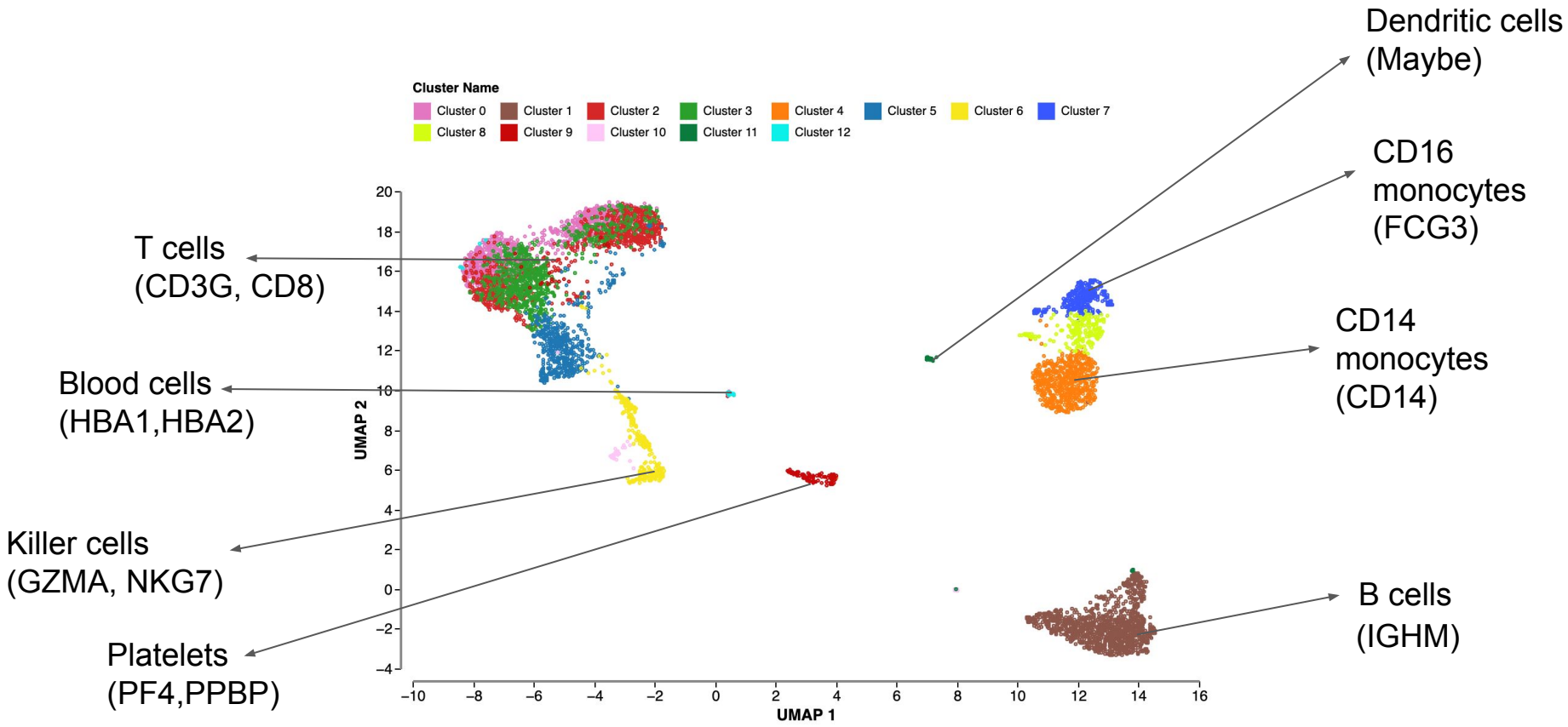


Click twice to sort descending by logFC

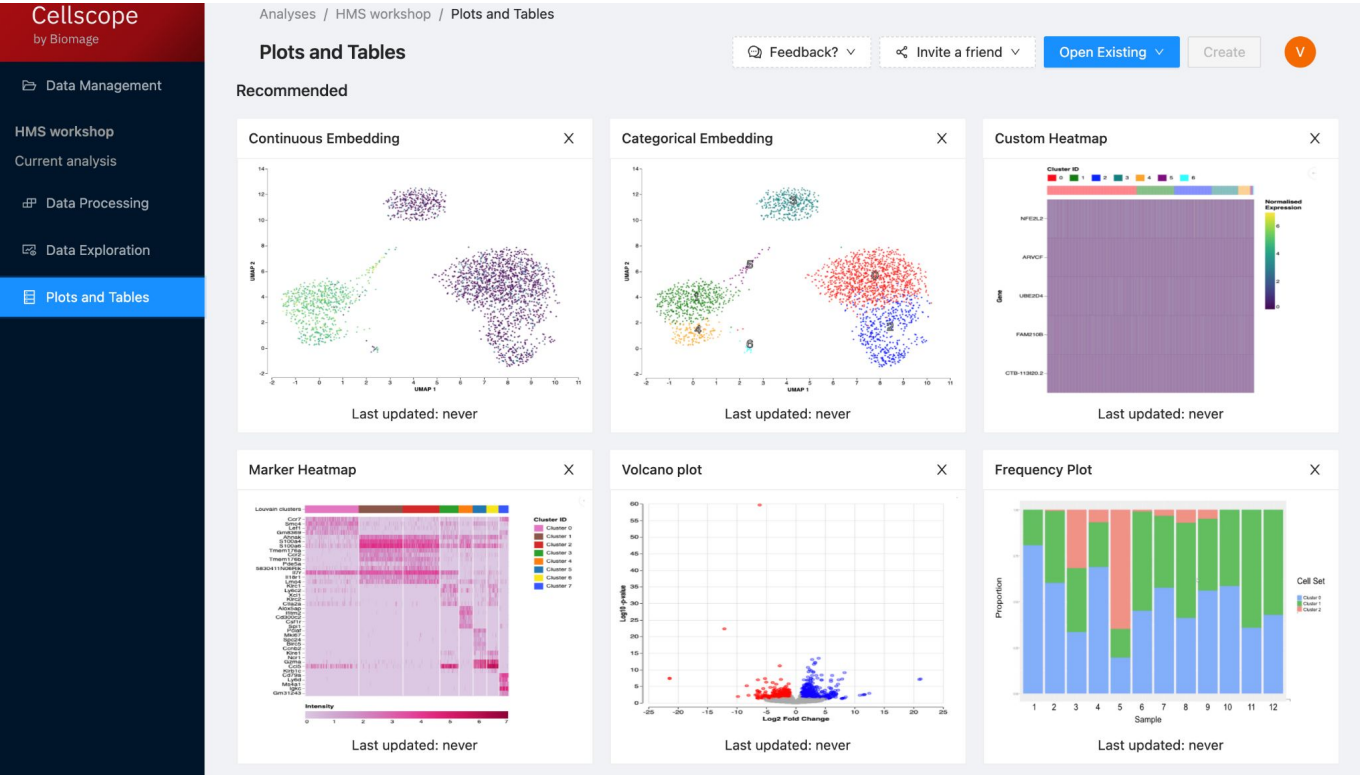
Click twice to visualize expression in the UMAP

Click to add genes to the heatmap

Answer 2 - Annotate the unknown clusters



Plots & Tables module



- range of plots pre-loaded with your data
- full customization of plot styling
- export differentially expressed gene lists
- export publication-quality figures

-> Create and customise figures for publication

Exercise 3 - Visualise differences between cell types [5mins]

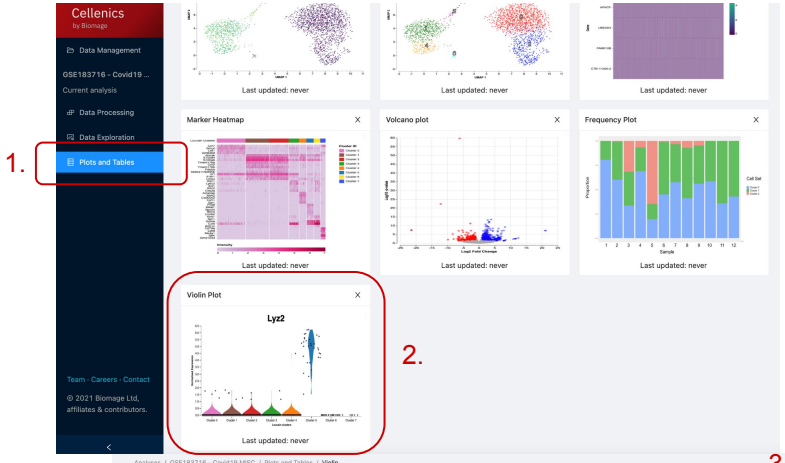
Now that we have annotated one or more clusters, let's create some plots to sustain our findings.

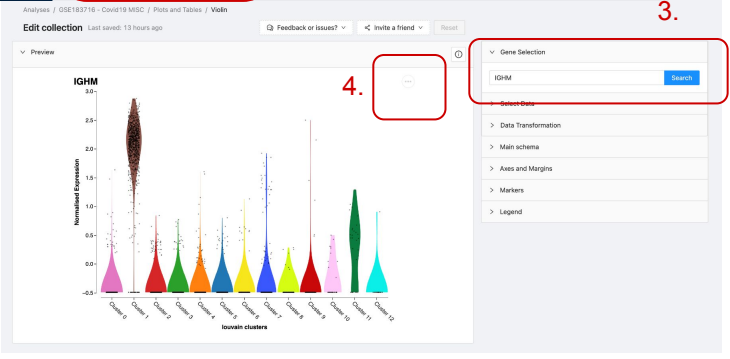
Create and download:

a) a Violin Plot, and

b) a Continuous Embedding plot

...to justify the annotation.





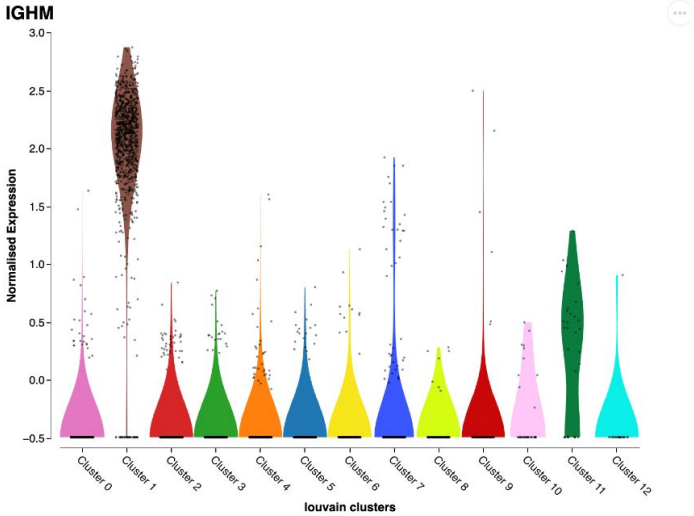
INSTRUCTIONS:

1. Go to the Plots and Tables module
2. Select the Violin plot (or continuous embedding)
3. In the "Gene selection" menu, input one of the genes that you found in the Differential Expression results (e.g. IGHM)
4. Download the plot using the "..." menu.

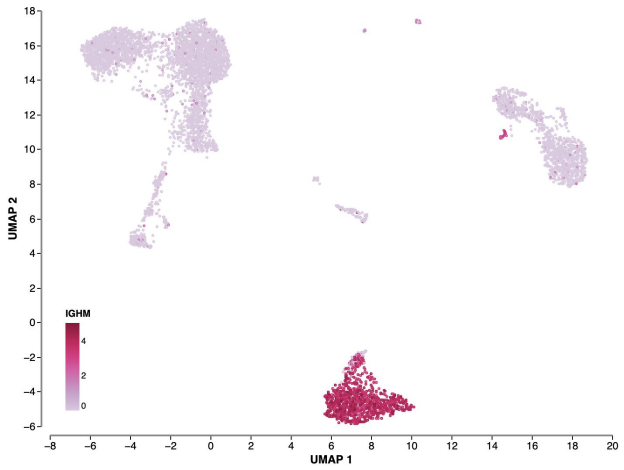
When you have completed it, write "Done 3" in chat.

Answer 3 - Visualise differences between the cell types

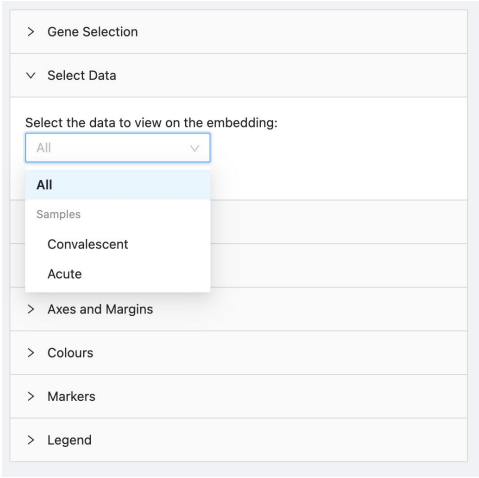
A **violin plot** is a great idea to showcase the difference between clusters



A **continuous embedding plot** can also be used to visualise gene expression



FYI: In both plots, you can view all cells or specific samples using the “select data” menu.

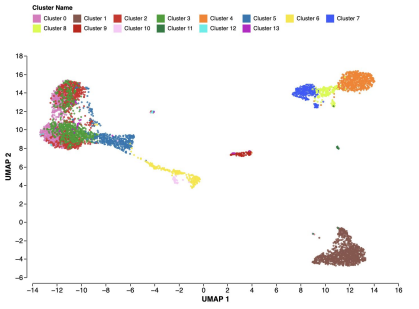


Exercise 4 - Investigate the unknown population [20mins]

a) Investigate the population of cells that is excluded in filter 4.

What are they? Create a plot to illustrate your conclusion.

The old embedding:



INSTRUCTIONS:

1. Go to Data Processing module
2. Navigate to step 4 (genes vs UMIs filter)
3. Disable this filter - you'll need to re-run the pipeline
4. Look in Data Exploration for a new cluster, use DE tool to identify them
5. Go to Plots & Tables - create a plot to illustrate your findings

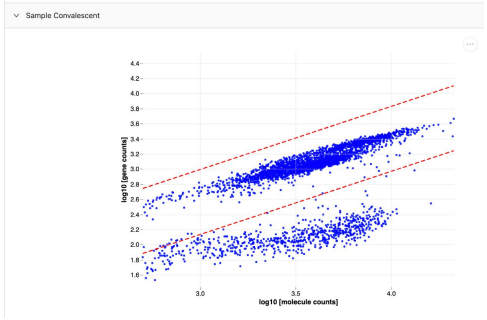
When you have completed it, write "Done 4" in chat.

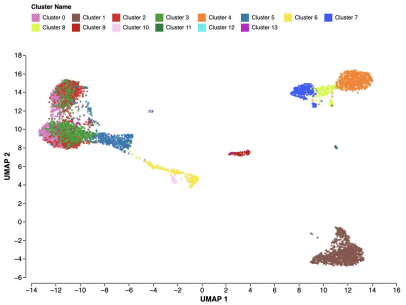
Answer 4 - Investigate the unknown population

a) Investigate the population of cells that is excluded in filter 4.

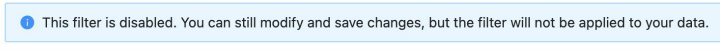
What are they? Create a plot to illustrate your conclusion.

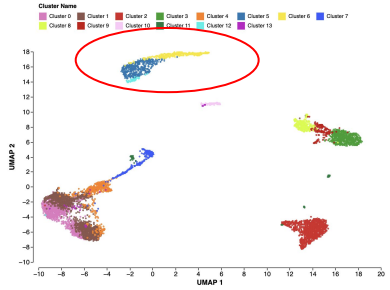
Exclude the population:





Include the population:

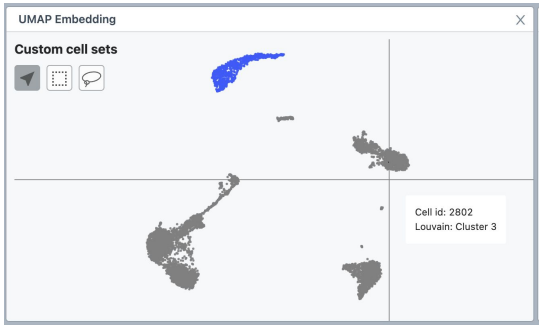


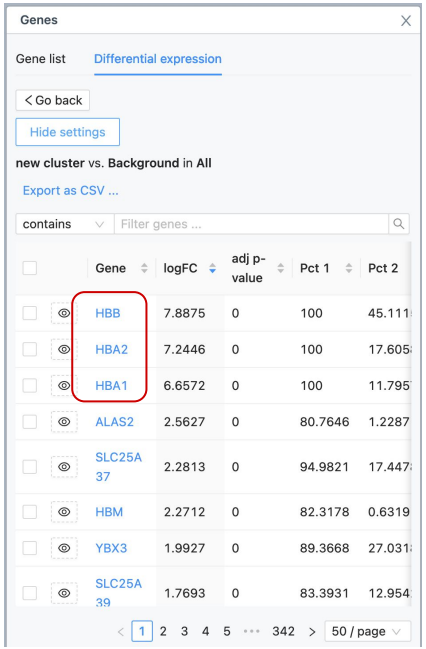


Answer 4 - Investigate the unknown population

a) Investigate the population of cells that is excluded in filter 4.

What are they? Create a plot to illustrate your conclusion.





HBB = Hemoglobin subunit beta
HBA2 = Hemoglobin subunit alpha 2
HBA1 = Hemoglobin subunit alpha 1

Red Blood Cells

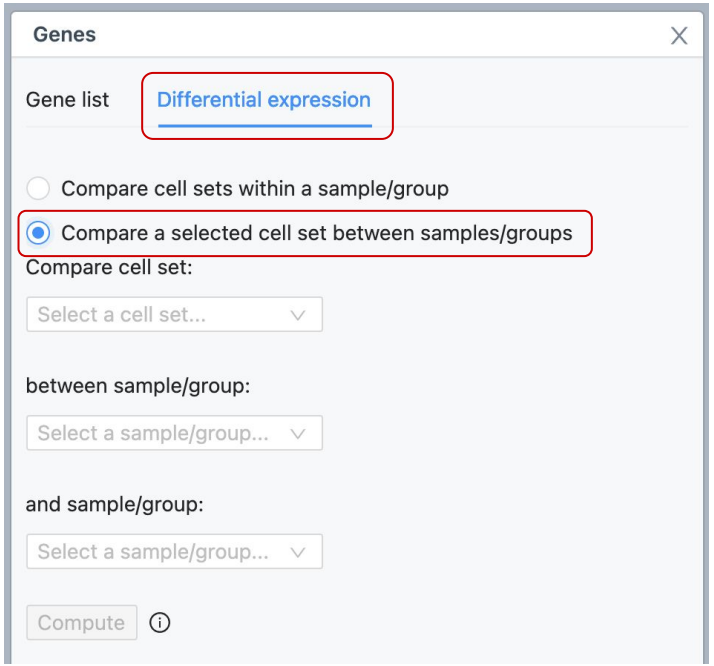
Exercise 5 - Find differences between the samples [10mins]

Determine if the two samples are from the same patient.

INSTRUCTIONS:

1. Perform Differential Expression, comparing ‘All’ cells between the two samples.
2. Sort the results by ascending / descending logFC.
3. Explore the top 5 genes.
4. Generate a plot (e.g. volcano plot) to illustrate your findings.

Write Done 5 in the chat



Answer 5 - Find differences between the samples

Determine if the two samples are from the same patient.

The Convalescent sample expresses various Y-linked genes (e.g RPS4Y1).

These Y-linked genes are only expressed in the Convalescent sample and not in the acute sample.

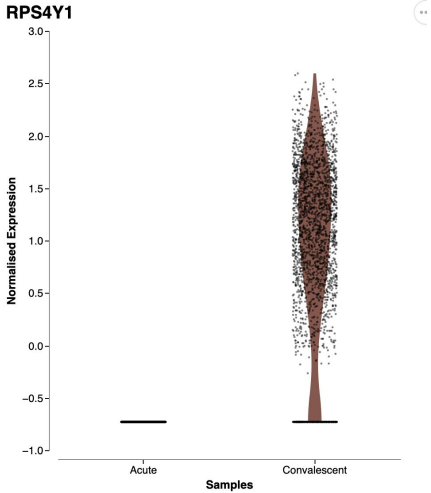
These findings lead us to believe that the two samples actually come from two different patients: the Acute sample is from a female and the Convalescent sample is from a male.

DE results, ordered by descending logFC.

Acute vs. **Convalescent** in **All**

Gene	logFC	adj p-value	Pct 1	Pct 2	AUC
RPS4Y1	-1.7745	0	0	84.9763	0.0751

Violin plot showing RPS4Y1 expression in the acute and convalescent samples.

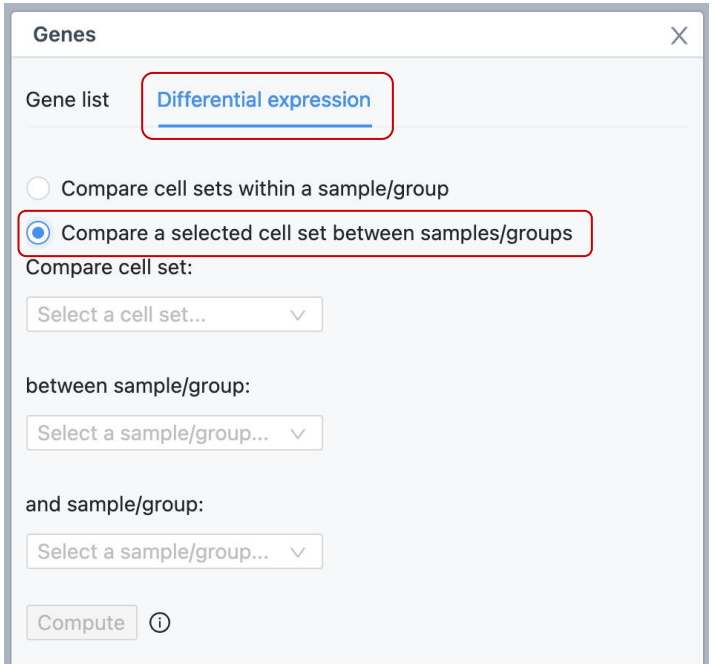


Homework - Find differences between the samples

Find differences in the Killer cell cluster (yellow cluster from exercise 2) between the acute and convalescent samples.

INSTRUCTIONS:

1. Perform Differential Expression, comparing the Killer cell cluster between the two samples.
2. Sort the results by ascending / descending logFC.
3. Explore the top 5 genes.
4. Generate a plot (e.g. volcano or violin plot) to illustrate your findings.

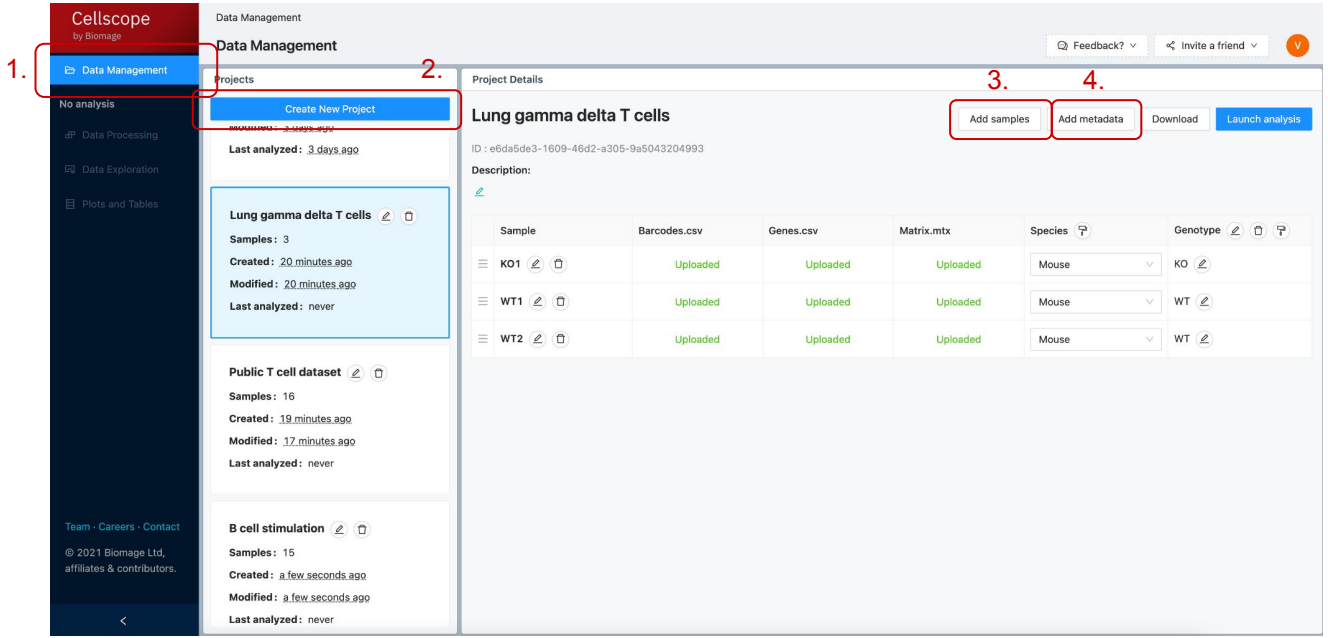


Data upload

To upload your own dataset

INSTRUCTIONS:

1. Go to Data Management module
2. Create and name a new Project
3. Click “Add samples” to upload the files.
4. Click “Add metadata” to allocate metadata to each sample



Supported file types

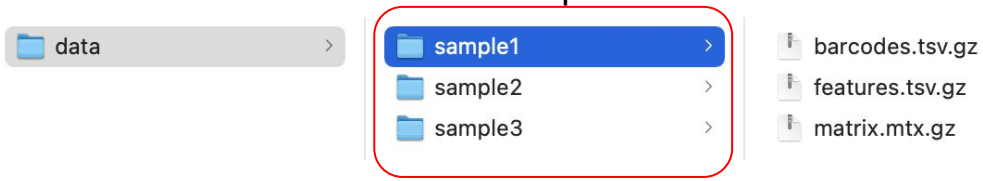
Datasets generated using 10X Chromium technology

Use the raw (unfiltered) count matrices that are output from Cell Ranger - These are usually stored in a folder named ‘raw_feature_bc_matrix’.

For each sample you need 3 files:

1. `features.tsv` or `features.tsv.gz` or `genes.tsv` or `genes.tsv.gz`
2. `barcodes.tsv` or `barcodes.tsv.gz`
3. `matrix.mtx` or `matrix.mtx.gz`

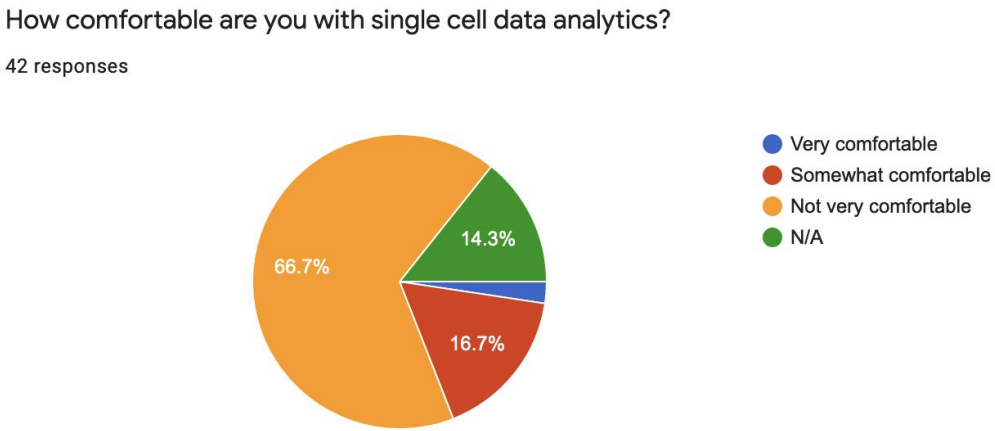
Include each of these 3 files in a folder that is named with the sample ID:



Upload the folders (named with the sample IDs) to Cellenics

Need help? Email: vicky@biomage.net

Summing up



What we hope to have achieved in this workshop:

- That you understand how Cellenics can support your single cell analysis
- That you feel comfortable exploring a dataset in Cellenics
- That you know how to get help with your analysis in Cellenics

OUR GOAL

We will help you to get your analysis done!

ACTIONS:

1. Upload your own data or any dataset that you are interested in exploring.
2. Book a free 45-minute meeting with Vicky (biologist) and Oliver (bioinformatician) to discuss your analysis needs:

https://calendly.com/vicky-morrison/cellenics-onboarding

vicky@biomage.net and oliver@biomage.net

Thank you!

References

-Centers for Disease Control and Prevention. Multisystem Inflammatory Syndrome (MIS) https://www.cdc.gov/mis/index.html. Page last reviewed February 24, 2021.

-Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10

-Giacalone, M., Scheier, E., & Shavit, I. (2021). Multisystem inflammatory syndrome in children (MIS-C): a mini-review. International Journal of Emergency Medicine, 14(1), 1-4.

-Henderson, L. A., & Yeung, R. S. (2021). MIS-C: early lessons from immune profiling. Nature Reviews Rheumatology, 17(2), 75-76.

-Kleiveland, C. R. (2015). Peripheral blood mononuclear cells. The impact of food bioactives on health, 161-167.

-National Cancer Institute - https://www.cancer.gov/

-Peripheral Blood | Handbook. https://www.miltenyibiotec.com/US-en/resources/macs-handbook/human-cells-and-organs/human-cell-sources/blood-human.html

-The innate and adaptive immune landscape of SARS-CoV-2-associated multisystem inflammatory syndrome in children (MIS-C) from acute disease to recovery. (GEO accession GSE183716)

-Weakley, S. M., Wang, H., Yao, Q., & Chen, C. (2011). Expression and function of a large non-coding RNA gene XIST in human cancer. World journal of surgery, 35(8), 1751-1756.

Homework answer - Find differences between the samples

b) Find differences in the Killer cell cluster (yellow cluster from exercise 2) between the acute and convalescent samples.

There are some really interesting immunological genes!

Granzyme is involved in the killing of virally-infected cells

Gene	logFC	adj p-value	Pct 1	Pct 2
GNLY	-2.074	5.8184e-12	46.0123	94.285
CCL4	-1.5134	4.9096e-21	24.5399	75.714
GZMB	-1.4853	5.8184e-12	43.5583	85.714
RPS4Y1	-1.4721	7.3392e-35	0	73.571
SPON2	-1.233	8.9661e-13	28.2209	72.142
PRF1	-1.2131	3.3308e-15	61.3497	91.428

A negative regulator (inhibitor) of the NFkB pathway. NFkB is a transcription factor that drives production of inflammatory cytokines.

Gene	logFC	adj p-value	Pct 1	Pct 2
LTB	1.2767	4.3857e-9	65.6442	35.714
CD52	0.9932	4.6798e-8	87.1166	62.142
CD3D	0.8539	0.0000098105	74.8466	40.714
NFKBIA	0.8537	2.0777e-11	59.5092	17.142
IL7R	0.8443	0	55.2147	27.857
RPS4X	0.8184	2.2481e-21	100	97.142