# Tutorial

## Perform Single Cell RNA Expression and Velocity Analysis

September 5, 2022

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Tutorial

# Perform Single Cell RNA Expression and Velocity Analysis

### Introduction

This tutorial uses part of the data set analyzed in "Key benefits of dexamethasone and antibody treatment in COVID-19 hamster models revealed by single-cell transcriptomics" published by Wyler et al. 2022.

The authors test the effect of two drugs on hamsters infected with COVID-19, both separately and in combination. In this tutorial we use a subset of the single cell data set, replicate the authors' findings summarized in their figure 5, and further expand the analysis. We strongly encourage you to read the article first, as the biology is not covered in detail here.

This tutorial covers the following:

- Downloading and importing the data.
- Performing QC, normalization and creating a UMAP plot.
- Locating the neutrophils using marker genes and predicted cell types.
- Performing RNA velocity analysis on neutrophils and analyzing the identified velocity genes.

### Prerequisites

For this tutorial, you must be working with CLC Genomics Workbench 22.0.2 and CLC Single Cell Analysis Module 22.1.1 or higher. Plugin installation is described in the CLC Genomics Workbench manual.

## Importing expression matrices, cell annotations and clusters

We start by importing the tutorial data using a series of single cell specific importers.

1. Download the data and unzip in a location of your choice. The folder contains the expression matrices (matrix, spliced and unspliced) in a subfolder, cell annotation and cluster files, and a workflow that will be used in this tutorial.
2. Start the CLC Genomics Workbench.
3. Go to the Reference Data Manager, choose **QIAGEN Sets** and download the **Single Cell Mouse (Ensemble** data set. We use the mouse reference data, as mouse and hamsters are closely related, and it contains a pretrained cell type classifier.
4. Import the expression matrix to a folder named **Import** via the toolbar:

   **Import** ( ) | **Import Expression Matrix** ( ) | **Import Expression Matrix in MEX format** ( )

   Configure the wizard as shown in figure 1:

   (a) Set **Gene or transcript track** to 'Mus_musculus_ensembl_v86_Genes' found in the 'CLC_References/mus_musculus/genes/ensembl_v86' folder.

   (b) Set **Cell format** to '{sample}.{barcode}'. This describes how to extract the sample name and barcode from a line in the Barcode file.

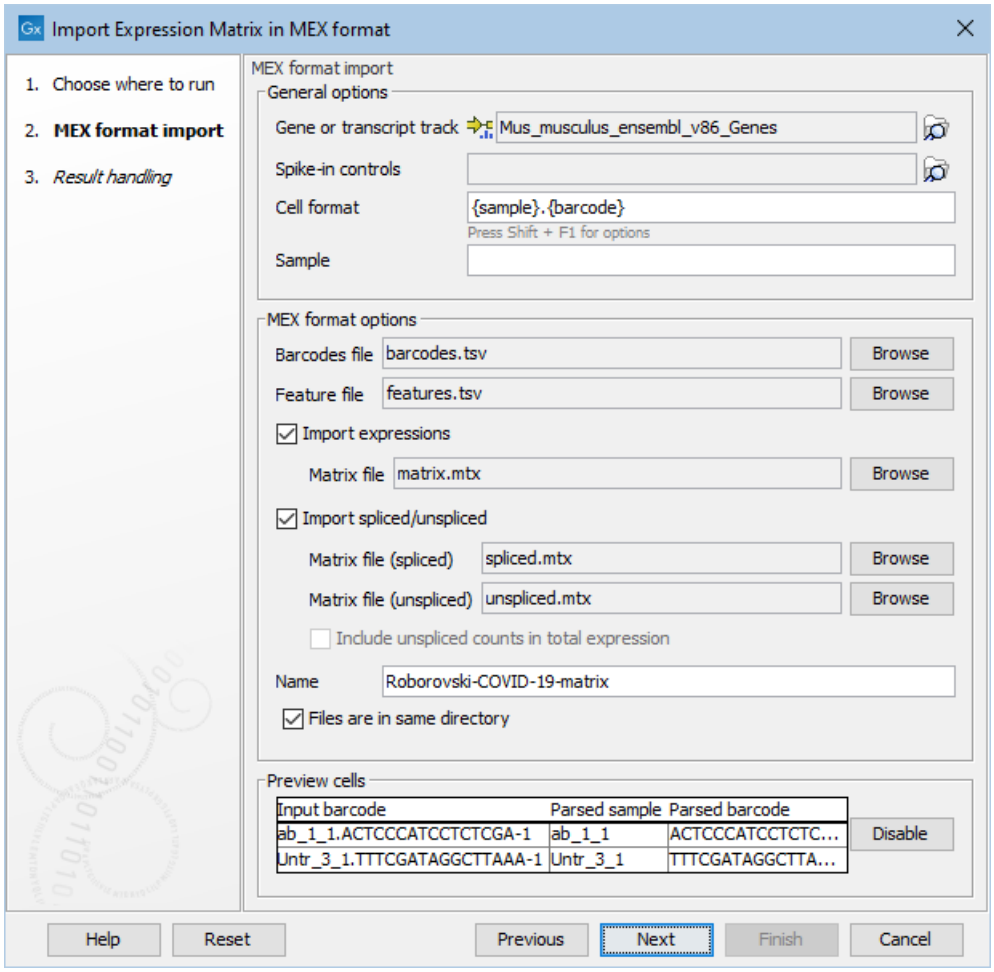


Figure 1: *Import Expression Matrix in MEX format configuration.*

(c) Make sure that **Files are in same directory** is ticked and then set **Barcode file** to Roborovski-COVID-19-matrix/barcodes.tsv file. The first lines will be shown in the **Preview cells** section of the wizard. The remaining options will be filled in automatically.

(d) Tick **Import spliced/unspliced**, which is needed for velocity analysis.

5. Click **Next** and save the results in a new folder called **Import**. You can create the folder using the **New folder** button near the top. Click **Finish**.

6. Now import the cell annotations and clusters provided by the authors using the importers via the toolbar:

   **Import** ( ) | **Import Cell Annotations** ( ) and **Import Cell Clusters** ( )

   See figure 2 and figure 3 for how to configure the importers. Make sure to tick **First column defines sample**, as well as **Map Clusters to QIAGEN Cell Ontology** when importing the clusters. Save the results in the **Import** folder.

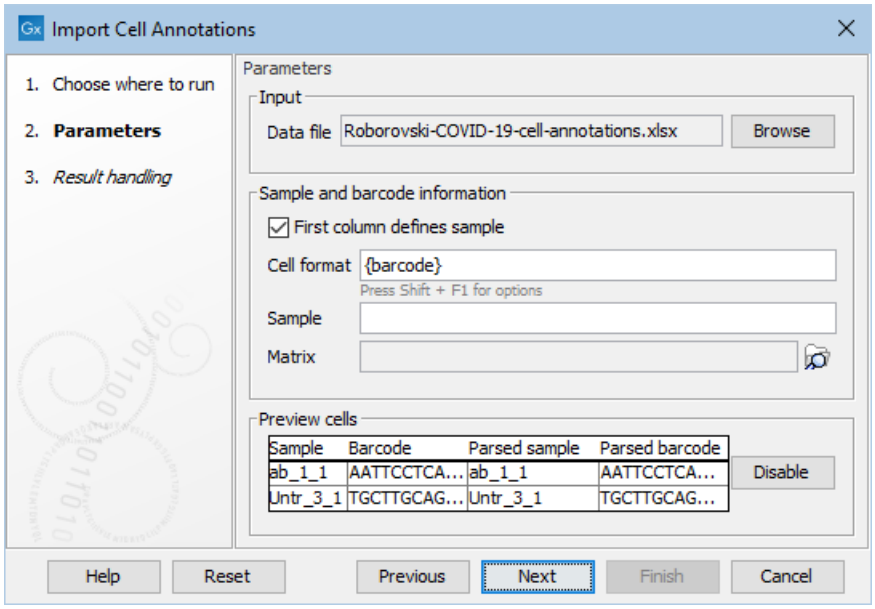


Figure 2: *Import Cell Annotation configuration.*

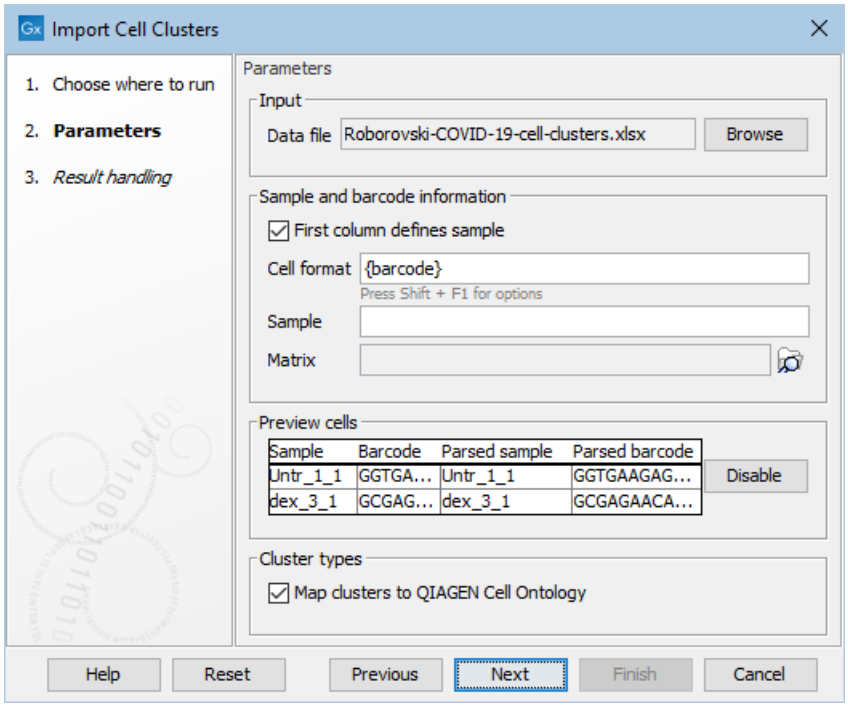


Figure 3: *Import Cell Clusters configuration.*

## Performing QC, normalization and creating a UMAP plot

Before analyzing the RNA expression data, we need to perform QC and normalize the data.

1. Start the **QC for Single Cell** tool from the Single Cell Analysis Toolbox:
   **Gene Expression** ( ) **| Cell Preparation** ( ) **| QC for Single Cell** ( )
2. Select 'Roborovski-COVID-19-matrix' from the **Import** folder.

3. Use **Next** to click through the wizard steps, while leaving the default settings.
4. Save the output in a new folder called **Analysis**.

The tool outputs a filtered matrix and QC metrics, as well as a QC report for each sample found in the input. Open one of the reports and have a look at the different sections. The report indicates that cells have been previously filtered as no cells with few reads or expressed features are present. QC also removes doublets (two cells with the same barcode) and the reports show that, in general, few cells are removed.

We will now normalize the filtered matrix.

1. Start the **Normalize Single Cell Data** tool from the Single Cell Analysis Toolbox:
   **Gene Expression ( ) | Cell Preparation ( ) | Normalize Single Cell Data ( )**
2. Select 'Roborovski-COVID-19-matrix (filtered matrix)' from the **Analysis** folder.
3. Leave default settings, click **Next** and save the output to the **Analysis** folder.

The tool outputs a residual matrix and a report. Open the report and have a look at the different sections. Note that many of the marker genes used by the authors (see `supplementary figure E3A`) for cell type classification are in the list of top variable genes from the report. For example, Camp and S100a8, which are markers for neutrophils, are high in the list.

For further information on how to interpret the results, click **Help** in the lower right part of the window.

We will use the residual matrix for creating a UMAP plot (see figure 4).

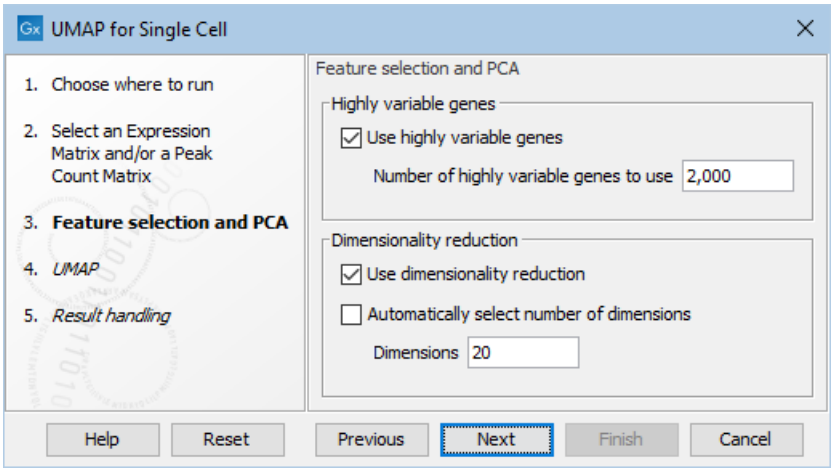


Figure 4: *UMAP for Single Cell configuration.*

1. Start the **UMAP for Single Cell** tool from the Single Cell Analysis Toolbox:
   **Dimensionality Reduction ( ) |UMAP for Single Cell ( )**
2. Select 'Roborovski-COVID-19-matrix (residuals)' from the **Analysis** folder.

3. Tick **Use highly variable genes** and set it to 2,000.
4. Leave the other settings at their defaults.
5. Click **Next** and leave the default settings.
6. Click **Next** and save the output in the **Analysis** folder.

## Locating the neutrophils

Double click 'Roborovski-COVID-19-matrix (UMAP)' when the tool is finished. We will locate the neutrophils in the plot. The authors used Camp and S100a8 as marker genes.

1. In the **Expression data** Side Panel group, write Camp and select the gene in the list below by double clicking it. Add the gene to **Selected features** by clicking the edit icon and choosing **Add to Selected**. Do the same for S100a8. Set the minimum expression of both genes to 50 (figure 5).

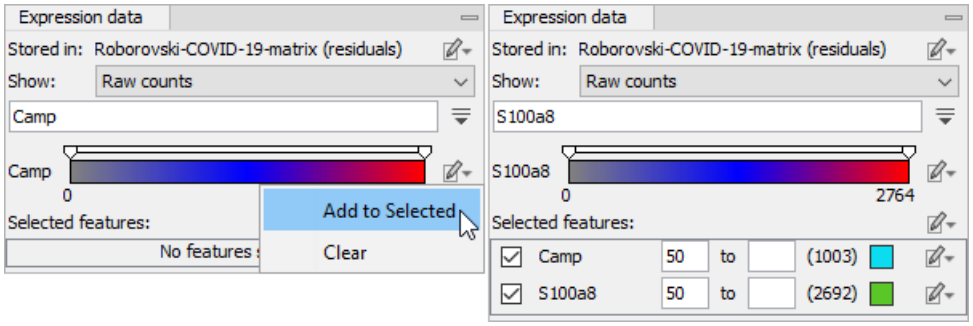


Figure 5: *Selection of marker genes in the 'Expression data' Side Panel group. Left: Adding a feature to 'Selected features'. Right: Minimum expression is set to 50 for both genes.*

2. The cell coloring in the **Coloring and highlighting** Side Panel group should automatically update to **Expression data (selected features)**. Under **Highlight cells**, choose the **Selected in all (intersect)** option. Now only cells with both marker genes are highlighted, see figure 6.
3. From the **Import** folder from the Navigation Area, drag and drop **Roborovski-COVID-19-cell-clusters** and **Roborovski-COVID-19-cell-annotations** into the plot Side Panel **Clusters** and **Cell Annotations** groups, respectively.
4. In the **Clusters** Side Panel group, set **Category** to be 'celltype'. Make sure that **Show colors from** and **Show labels from**, found in **Coloring and highlighting** Side Panel group, are set to **Clusters**. This will color the cells using the authors' annotated cell types, and will add labels to the plots, see figure 7.
5. The cells have also been annotated with the treatment, the hamster replicate, and the viral load. Browse around the clusters and cell annotations and see how this information can add value to the UMAP.
6. Lasso-select the neutrophils cluster by holding down the left mouse button while drawing a circle around the cluster. The cells in the lasso are now colored black to indicate they are selected. Right click on the plot and choose the **Create subset** option (figure 8).

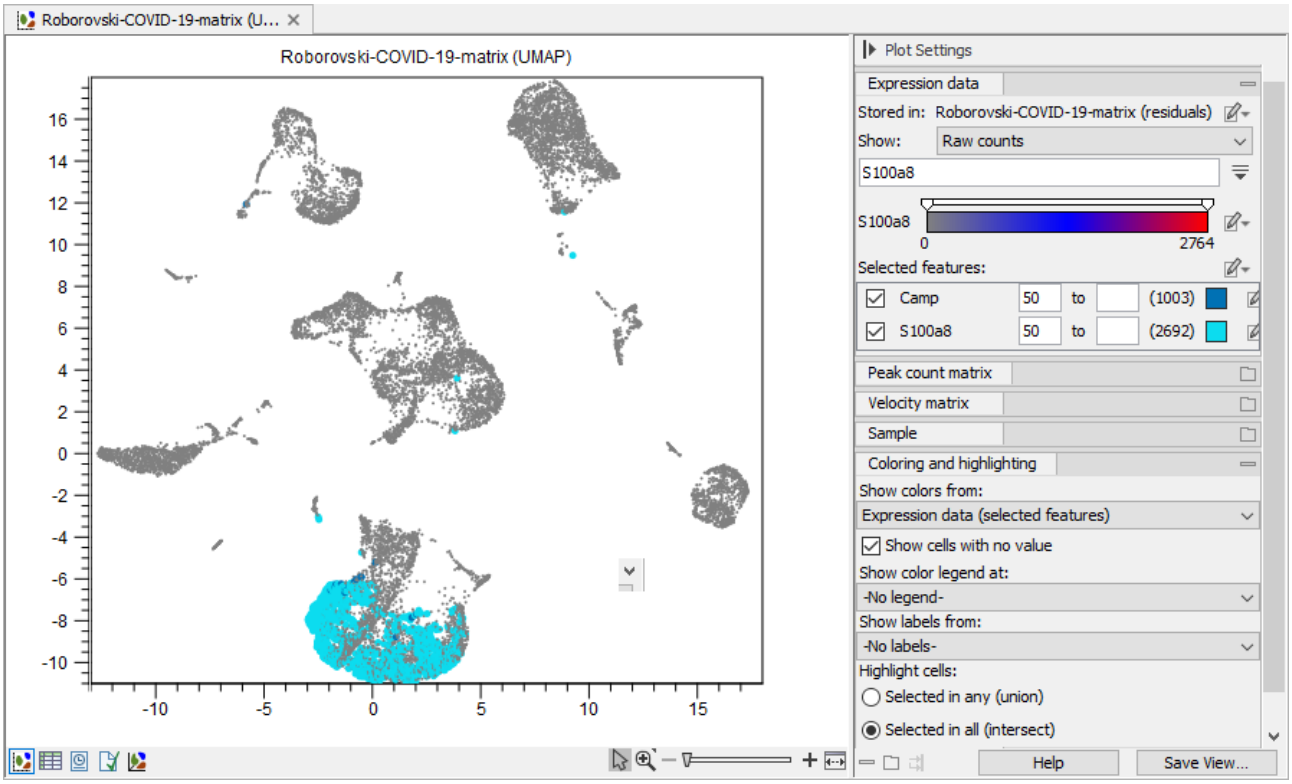


Figure 6: *Cells with expression of at least 50 for both Camp and S100a8.*

Make sure to tick all output options. Save the results in a new folder called **Neutrophils**. We will use these results later.

7. Cell types can also be identified using the **Predict Cell Types** (see figure 9):

   **Gene Expression** ( ) **| Cell Type Classification** ( ) **|Predict Cell Types** ( )

   - Select 'Roborovski-COVID-19-matrix (residuals)' from the **Analysis** folder.
   - Set **Cell type classifier** to 'sc_mouse_cell_type_classifier_v1.2' found in the 'CLC_References/mus_musculus/sc_mouse_cell_type_classifier_v1.2/' folder.
   - As the cells have been collected from the hamsters' caudal lung lobe, set **Tissue type** to 'Lung'. Click **Next**.
   - Optionally, tick **Output the cell annotations**, containing probabilities for the predicted cell types.
   - Save the results in the **Analysis** folder and click **Finish**.

   In the UMAP plot, use the **Clear** option found under the edit menu (small pencil icon in the top right corner) of the **Clusters** Side Panel group. From the **Analysis** folder from the Navigation Area, drag and drop 'Roborovski-COVID-19-matrix (cell types)' into the **Clusters** Side Panel group. Choose the **Cell Type (all)** category and locate the neutrophils. Tick the box and set **Show labels from** to use the clusters. You will find that the prediction points to the same cluster annotated as neutrophils by the authors. Click on the edit menu next to neutrophils and choose **Show in QIAGEN Cell Ontology**. This shows that neutrophils are a

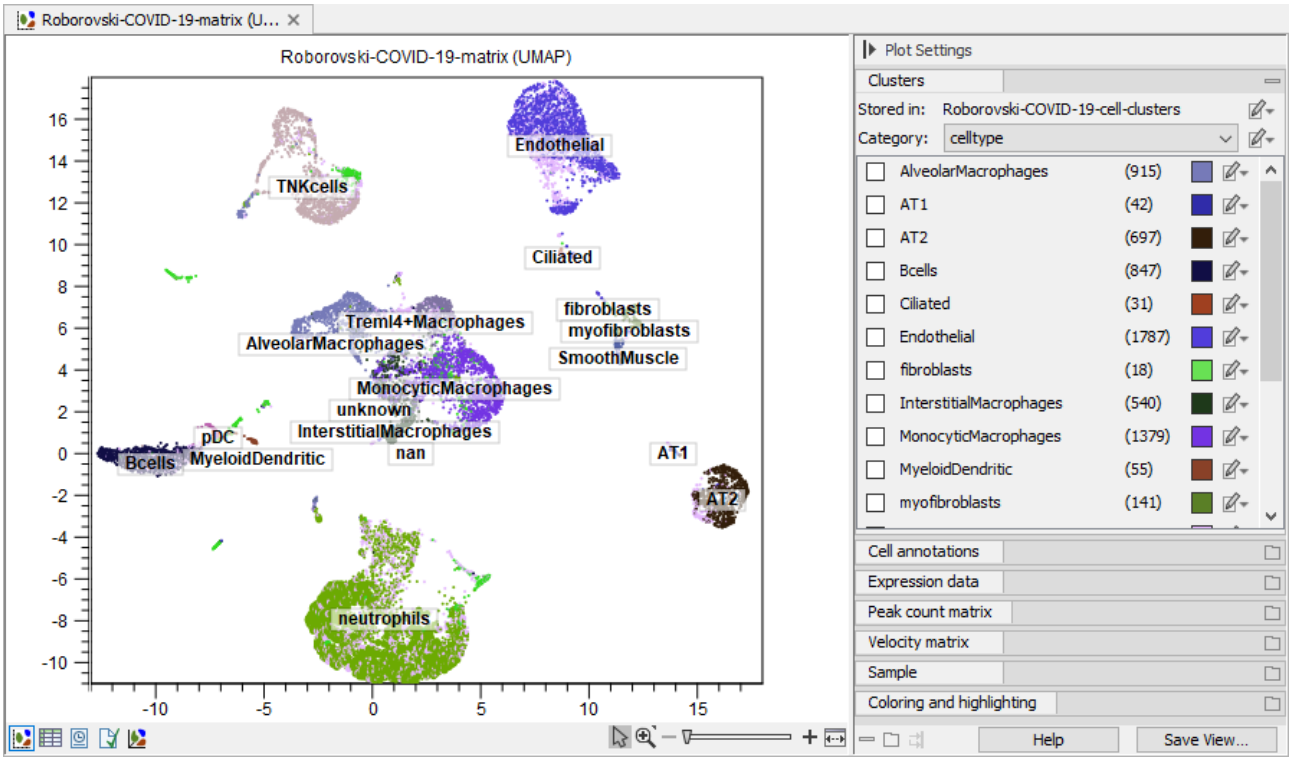


Figure 7: *UMAP with authors' annotated cell types.*

subtype of granulocytes and sister cell type to basophils and eosinophils (figure 10). Tick the boxes from all of these cell types - they are located in the same cluster.

The relation between the predicted and the authors' cell types can be further investigated using `Create Heat Map for Cell Abundance`. Note that for this, the two clusters need to be combined using `Combine Cell Clusters`. Remember that we predicted cell types here using a mouse classifier on a hamster data set.

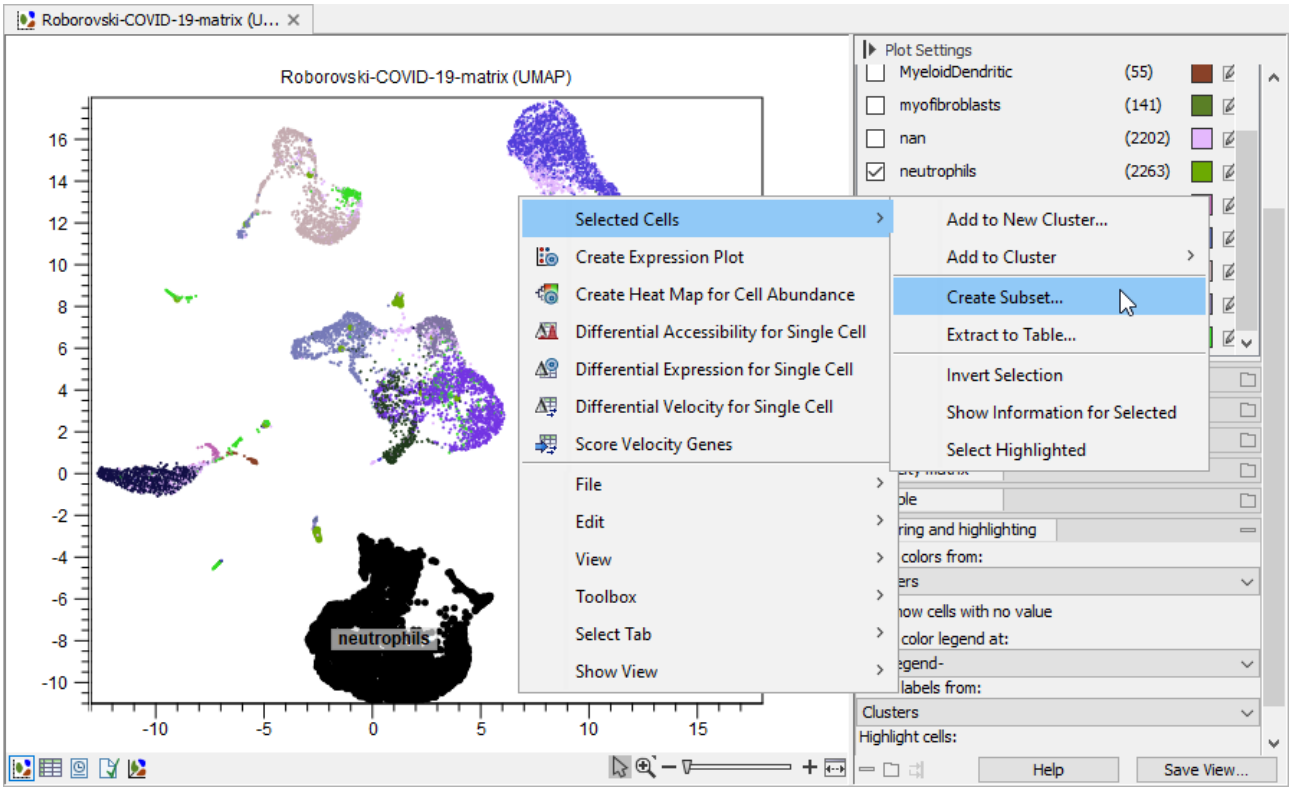


Figure 8: *Create subset of the selected neutrophils.*

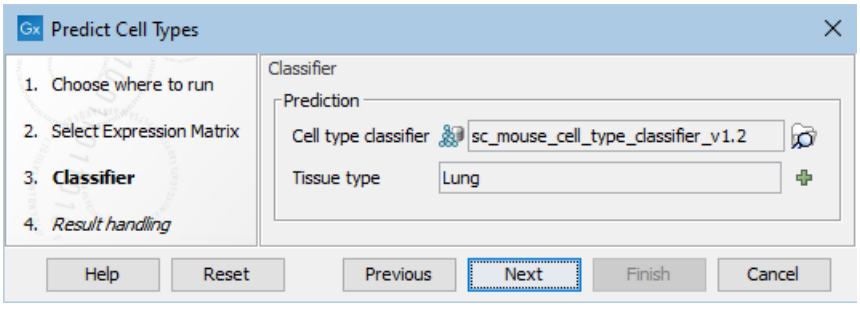


Figure 9: *Predict Cell Types configuration.*

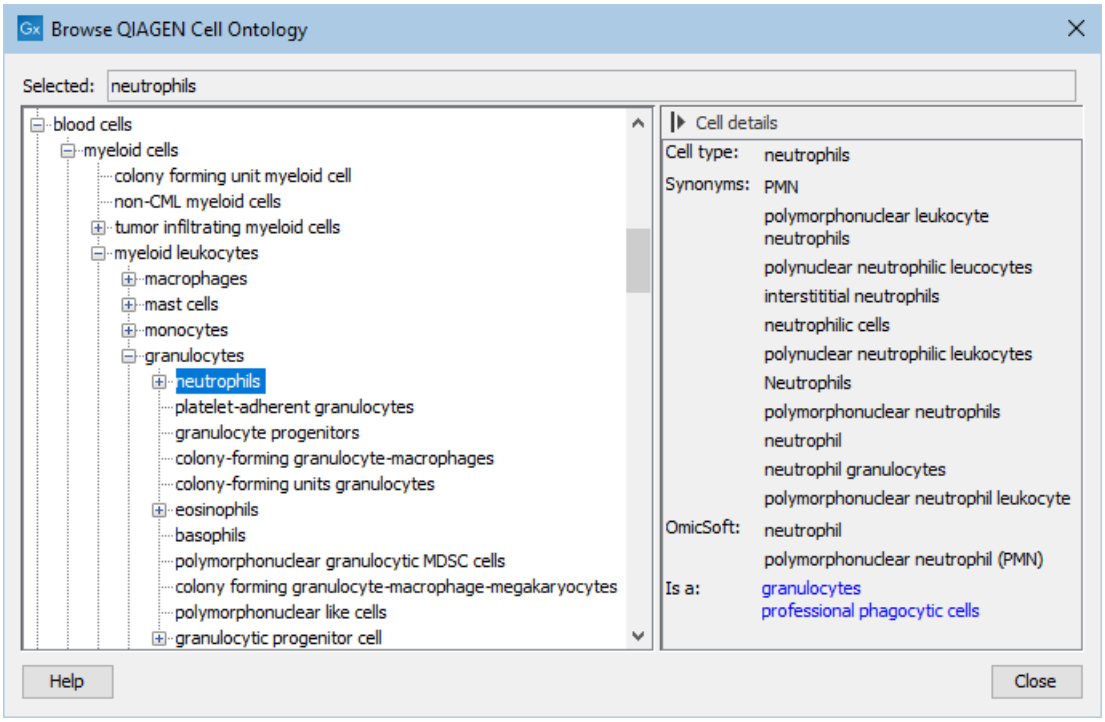


Figure 10: *View of neutrophils in the QIAGEN Cell Ontology.*

## Running the RNA Velocity Analysis workflow

We will now perform RNA velocity analysis for the neutrophils. For this part, we need to import the workflow that was downloaded at the beginning of the tutorial using the **Standard Import** tool via the toolbar:

**Import (  ) | Standard Import (  )**

Add the 'Analyze RNA Velocity in Neutrophils Workflow.clc' file, use the **Automatic import** option and save it in the **Neutrophils** folder.

The workflow contains a set of tools that together will provide a full velocity analysis - double-click it to see the content (figure 11):

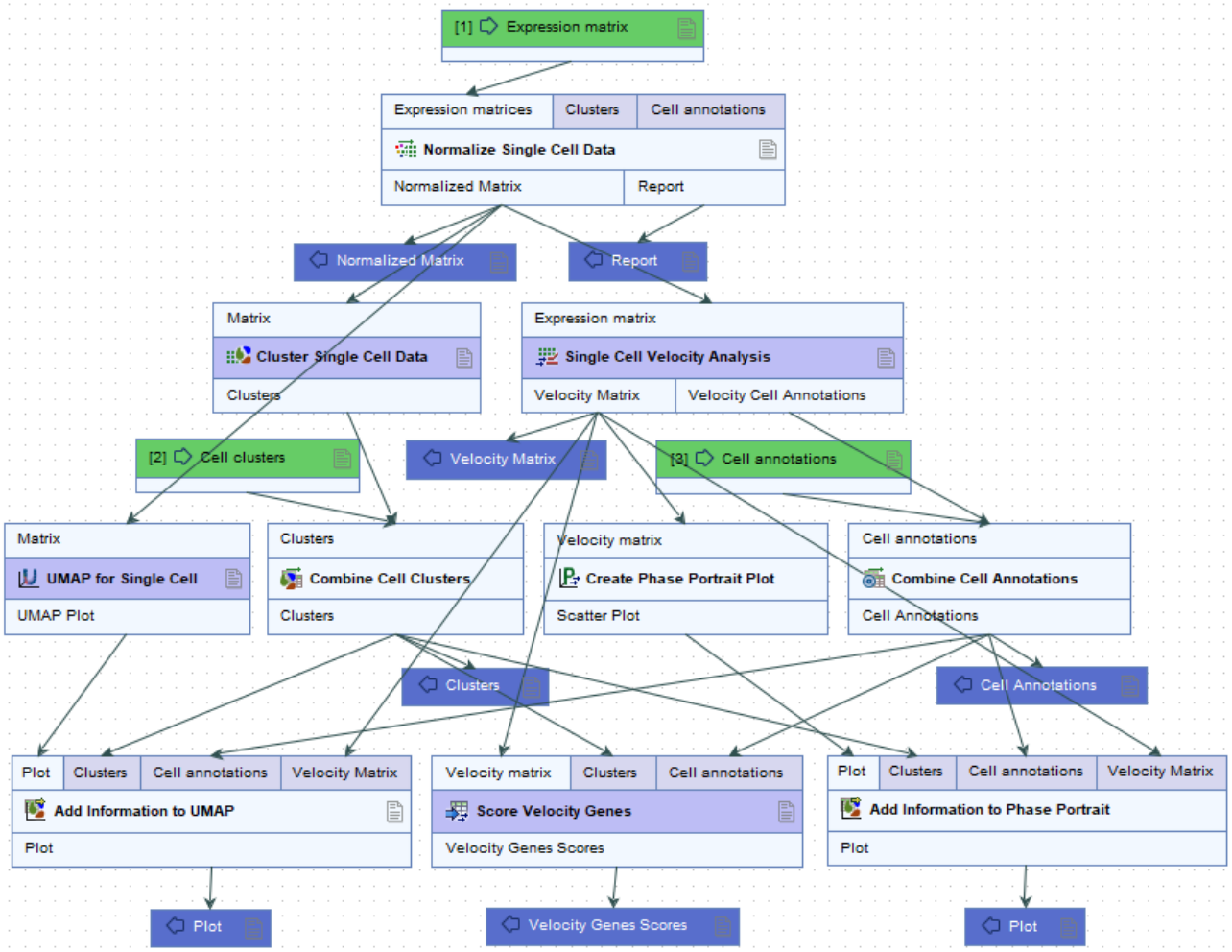


Figure 11: *The 'Analyze RNA Velocity in Neutrophils Workflow'.*

- First, the input expression matrix is normalized (Normalize Single Cell Data).
- The normalized matrix is then used to create a UMAP plot (UMAP for Single Cell), velocity matrix (Single Cell Velocity Analysis) and cell clusters (Cluster Single Cell Data).
- The velocity matrix is further used to create a phase portrait plot (Create Phase Portrait

Plot) and score velocity genes (Score Velocity Genes).

- All cell annotations and clusters are combined (Combine Cell Annotations and Combine Cell Clusters, respectively).
- All relevant information is added to both the UMAP and phase portrait plots (Add Information to Plot).

To execute the workflow:

1. Click the **Run** button to the bottom right.
2. Select 'Roborovski-COVID-19-matrix (residuals, 3495 cell subset)' matrix from the **Neutrophils** folder. Note that the number of cells in the brackets might be different, depending on your lasso selection. Click **Next**.
3. Select 'Roborovski-COVID-19-cell-clusters (3495 cell subset)' and 'Roborovski-COVID-19-cell-annotations (3495 cell subset)' in the next two steps. Click **Next**.
4. In the final wizard step select "Save" and "Create log" before creating a new folder called **Workflow** and clicking **Finish**.

The progress of the workflow is shown in the opened workflow. The **Processes** tab in the lower left corner also shows the progress of the workflow. By clicking the small arrow to the right, you can choose to **Show Log Information**, and see how it gets updated during the execution. You can also click on **Show Results** once the workflow has finished executing. Execution will take a few minutes on a standard laptop. See figure 12 for the expected workflow outputs.

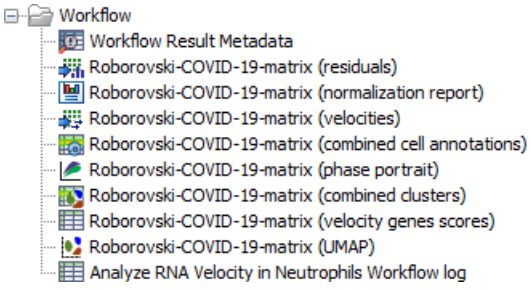


Figure 12: *Outputs created by 'Analyze RNA Velocity in Neutrophils Workflow'.*

## Analyzing the neutrophils

Let us inspect the workflow results.

- Open 'Roborovski-COVID-19-matrix (normalization report)'. The 'Variance of Pearson residual expression' plot indicates that the normalization was successful as the data are tightly clustered around the expected line. Note that the highly variable genes include those identified in the authors' original analysis of the full data (Ccl3, Csf1 and Cxcl10).
- The cell annotations and clusters contain a mixture of original and new annotations/clusters generated by the workflow.

- The UMAP plot summarizes the high-dimensional velocity vector as arrows, providing an easy interpretation of the moving cell system. The plot is already populated with the cell annotations and clusters generated by the workflow, as well as the velocity matrix.

In the following we will reproduce some of the findings and analysis from the original paper summarized in the paper's figure 5.

1. The paper's figure 5.A shows neutrophil sub-clusters identified by the authors. Select the **Leiden resolution = 1.5** category in the **Clusters** Side Panel group and set **Show labels from** to **Clusters** under **Coloring and highlighting** Side Panel group (figure 13). This shows cluster patterns similar to the original figure.

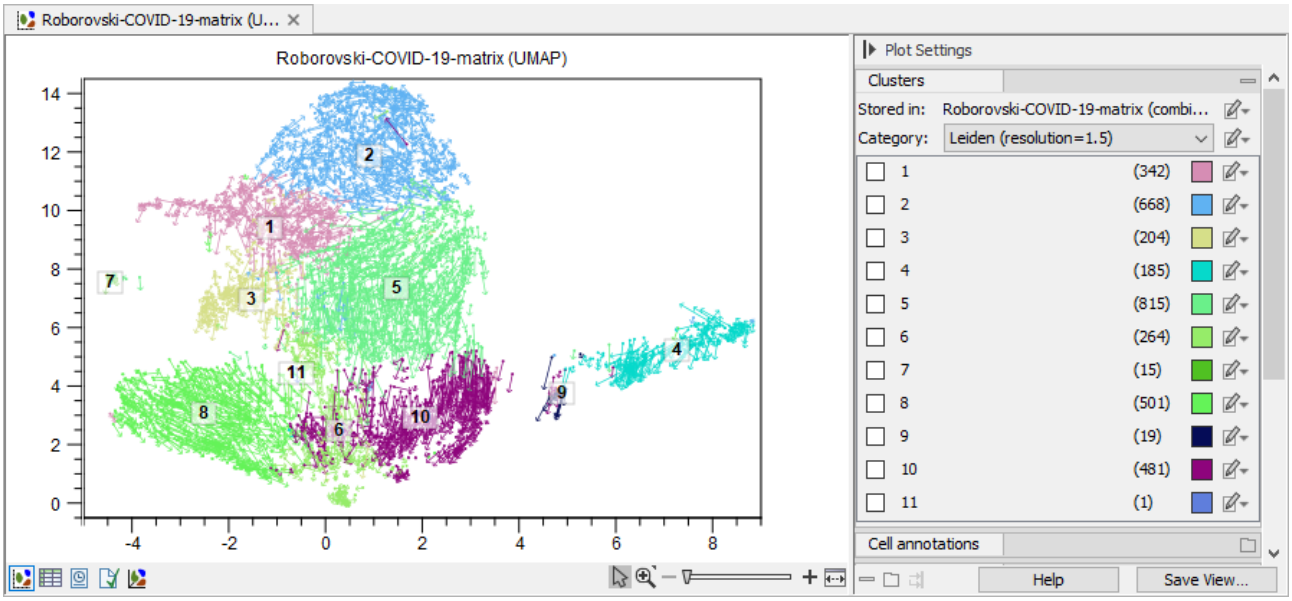


Figure 13: *Output created by the RNA Velocity workflow colored with cluster information and including labels.*

2. Right-click the UMAP plot and select **Create Expression Plot**. The tool is already configured with the cell annotations and clusters present in the plot. Execute the tool using default settings. Open the resulting plots and inspect the genes. Note that Ccl3, Csf1 and Cxcl10 overlap with the genes from the dotplot in the paper's figure 5.D. Use the side panel options to customize the plot, see Create Expression Plot for more details.

3. To obtain a plot that is closer to the paper's figure 5.D, we only select the relevant genes, see figure 14. Note that our cluster 6 has similar expression to cluster 6 in the paper's figure 5.D (it is just a coincidence that the clusters have the same names in both places). Go back to the UMAP plot and check where cluster 6 is located. To visualize it more easily, you can change its associated color by double clicking on the colored rectangle to the right, and choosing another color. Even though cluster 6 is a bit more scattered in the UMAP plot, it has roughly the same location as cluster 6 in figure 5.D. Additionally, the velocity arrows of the surrounding cells point towards cluster 6, as they do in the paper.

4. Let us investigate the relation between the viral load and the velocity dynamics. In the **Cell annotations** Side Panel group of the UMAP plot, choose the **log 10 SCoV2_sum**

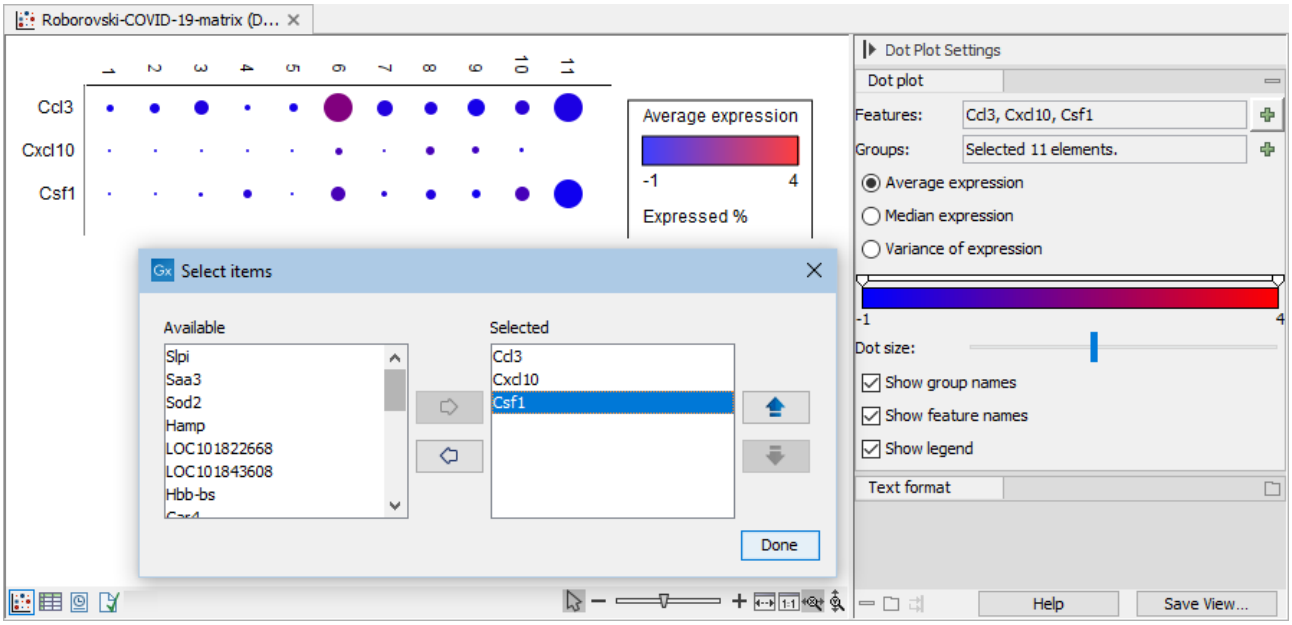


Figure 14: *Dot plot summarizing expression of Ccl3, Csf1 and Cxcl10.*

category. Change **Show colors from** under **Coloring and highlighting** Side Panel group to **Cell annotations**, and set **Show color legend at** to **Upper right**. The color palette can be changed to be similar to `the paper's figure 5.B` by clicking the gradient. The arrows above the gradient can also be moved to make it easier to spot the cells with higher values, see figure 15.

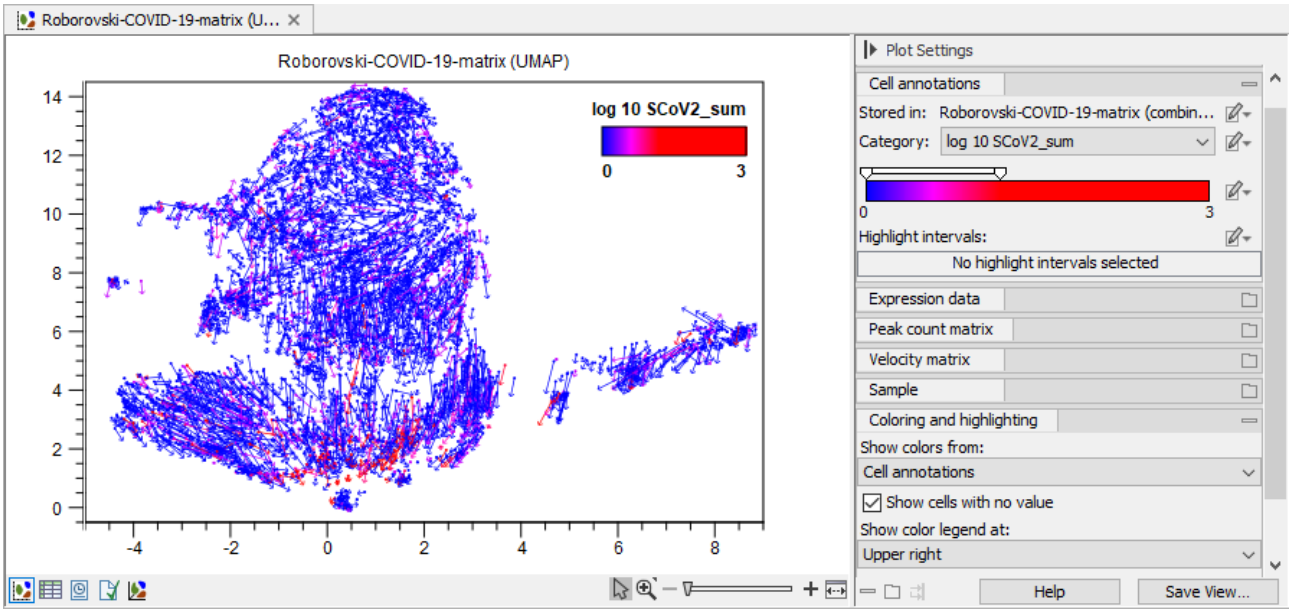


Figure 15: *Viral load is higher in the cells that are in the lower part of the plot, in the vicinity of cluster 6.*

The velocity arrows are also pointing towards the lower end of the plot, which is consistent with the results from the original article. Note that the size of the dots and arrows can be adjusted in the **Sizes** Side Panel group. This can be helpful for better visualization of the direction and magnitude of the velocity arrows.

5. The paper's figure 5.E shows the log 2 fold changes of the cell counts in the different clusters, for all three treatments compared to untreated. The authors note that in cluster 6, cells treated with dexamethasone are less abundant, particularly so for the combination treatment. Let us investigate this for the combination treatment (figure 16):

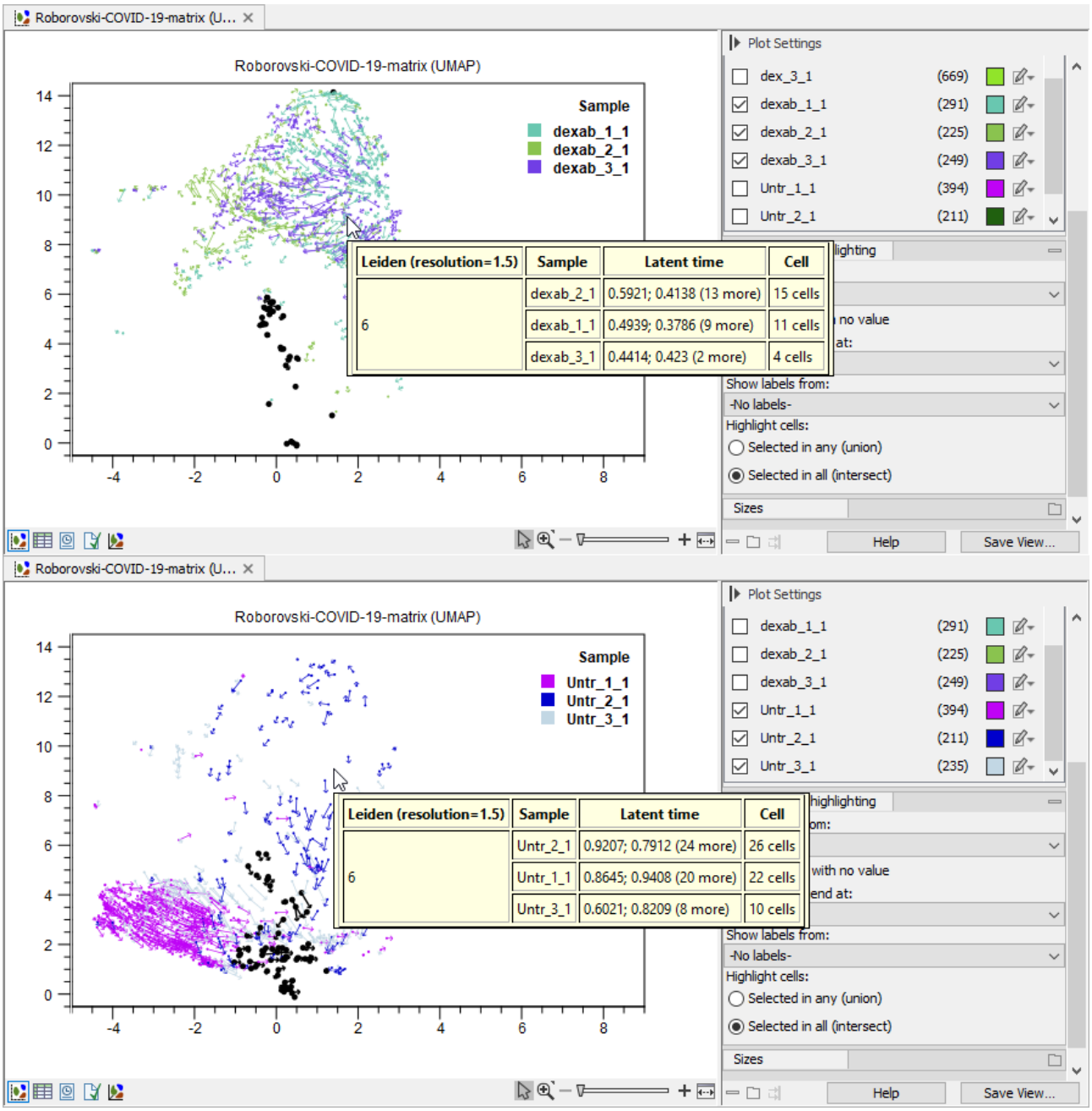


Figure 16: *Comparison of combination treatment and no treatment in cluster 6.*

- In the **Clusters** Side Panel group, tick cluster 6.
- In the **Sample** Side Panel group, tick the desired samples, corresponding to the combination treatment (samples with names beginning "dexab") or no treatment (samples with names beginning "Untr").
- In the **Coloring and highlighting** Side Panel group, set **Show colors from** to **Sample**, untick **Show cells with no value** (this will hide the cells from the remaining samples), set **Show color legend at** to **Upper right** and **Highlight cells** to **Selected in all (intersect)**. The cells that are from the selected clusters and cluster 6 will be shown as larger dots.
- To more easily spot the highlighted cells, right click on the plot, and choose **Selected cells** / **Select highlighted**. The cells are now black.
- Right click on the plot again, and choose **Selected cells** / **Show information for selected**. This will display a summary for the cells that are in the selected samples and cluster 6 (see figure 16).

We can see that the cells with the combination treatment are less abundant than the untreated cells in cluster 6 (see figure 16. To compare two UMAP plots at once, a second copy was opened by double-clicking the UMAP result in the Navigation Area.). We can calculate the frequencies of the cells and we obtain that 6.9% (58 out of 840) of the untreated cells are in cluster 6, while only 3.9% (30 out of 765) of the cells with the combination treatment are in cluster 6.

## Additional analysis

The workflow also outputs the list of velocity genes with scores related to how big of an impact they have in the overall velocity system, and a phase portrait. Let us investigate them.

1. Open the 'Roborovski-COVID-19-matrix (velocity genes scores)' from the **Workflow** folder. The table contains the 181 identified velocity genes: genes for which there is sufficient evidence that the expression is not in a steady-state and the gene is actively being up- or down-regulated. The table reports the velocity likelihood calculated when using all cells ('Roborovski-COVID-19-matrix (velocities)' column) or only the cells present in different clusters. The higher the likelihood, the more evidence for the gene to be a velocity gene. We can identify the three genes we have previously investigated (Ccl3, Csf1 and Cxcl10) using the **Filter** functionality (figure 17).

   Csf1 is the only gene of these three that was detected as a velocity gene. We can visualize its expression in the UMAP plot (figure 18). By unticking **Show cells with no value**, only the cells with some expression of Csf1 are shown.

   We can see that mainly the cells at the bottom of the figure express Csf1, where cluster 6 and the cells with high viral load are located. 'Roborovski-COVID-19-matrix (velocity genes scores)' reports the likelihoods calculating using the treatment. We can also use the identified clusters:

   (a) Right click on the UMAP plot and choose **Score Velocity Genes**. The tool is already configured with the cell annotations and clusters present in the plot.

   (b) Run the tool with **Score velocity genes for** "Leiden (resolution=1.5)", and otherwise using the default settings.

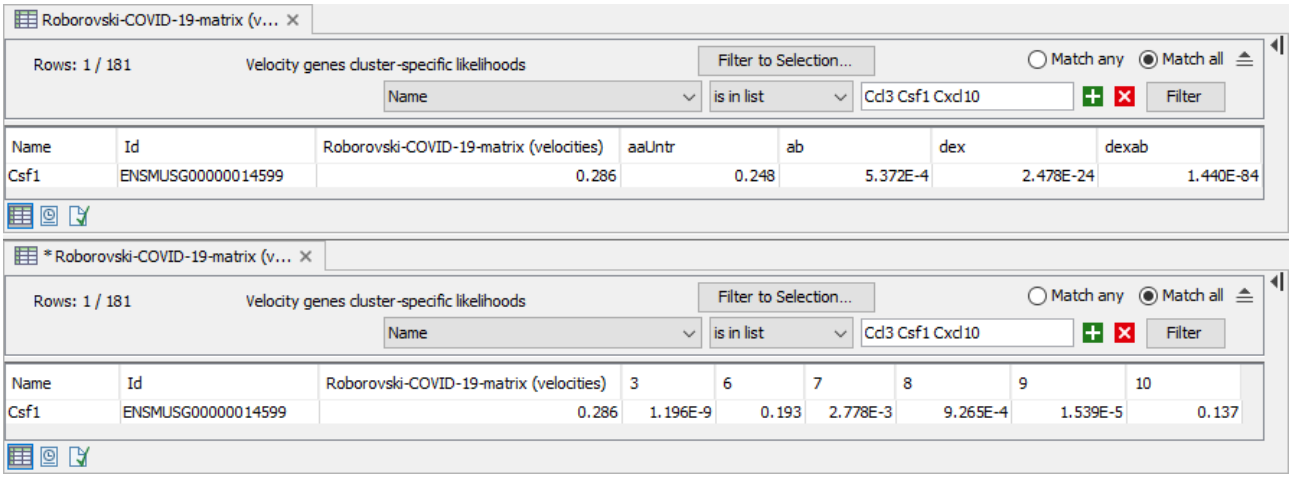


Figure 17: *Velocity likelihoods. Top: using the treatments. Bottom: using the clusters. Clusters with likelihood below* $10^{-70}$ *are not shown.*

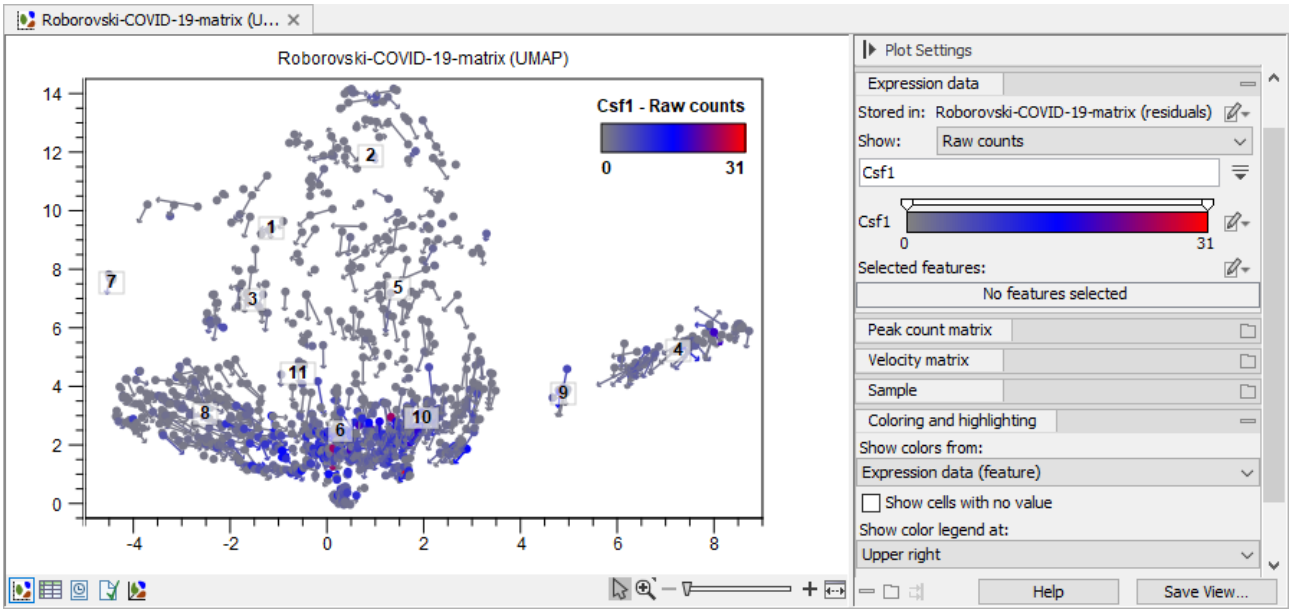


Figure 18: *Csf1 expression.*

(c) Filter the resulting table to only show Csf1.

The two tables (figure 17) show a strong velocity signal was present for the Csf1 gene when looking at the subset of cells that were not treated with dexamethasone/clusters 6, 7, 8 and 10.

2. Let us now open 'Roborovski-COVID-19-matrix (phase portrait)' from the **Workflow** folder. The plot shows the imputed spliced and unspliced counts for all genes for which this was calculated (as determined by the options of Single Cell Velocity Analysis). For the velocity genes, it also shows the estimated steady-state ratio and inferred dynamics. Cells above/below the steady-state are up-/down-regulated, while the dynamics curve shows the

predicted state of the system.

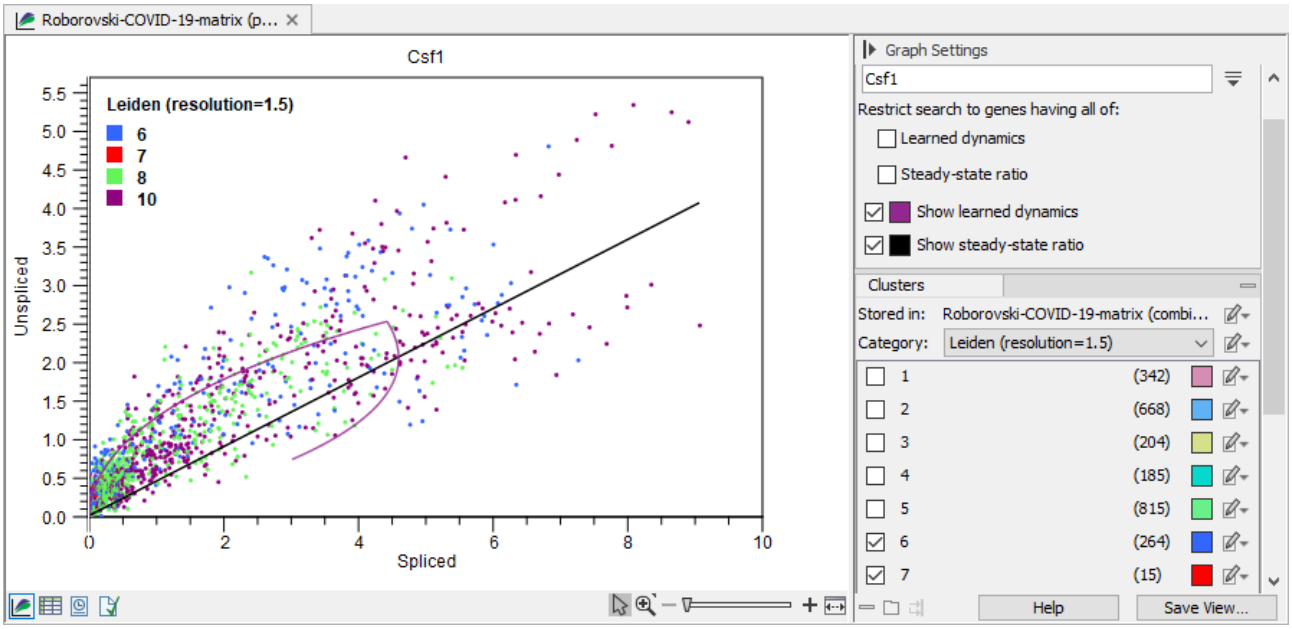


Figure 19: *Phase portrait for Csf1 showing cells in clusters 6, 7, 8 and 10.*

We can open the phase portrait for Csf1 by typing the gene name in the text box in the top right and double clicking on the gene in the list below. Set the category to **Leiden (resolution=1.5)** in the **Clusters** Side Panel group. Change **Show colors from** to **Clusters**, untick **Show cells with no value** and set **Show color legend at** to **Upper left** in the **Coloring and highlighting** Side Panel group. We can investigate which clusters are contributing to the clear up-regulation of this gene by ticking and unticking the different clusters. We find that the same clusters that have a high likelihood for this gene show clear signs of up-regulation (figure 19).

3. 'Roborovski-COVID-19-matrix (velocity genes scores)' indicates that Hmgb2 (High-mobility group protein B2) is the top velocity gene, with the highest likelihood (note that the table can be sorted using a certain column by clicking on the column name). The phase portrait indicates that the gene is being down-regulated (figure 20). This gene is involved in the end-joining process of e.g. V(D)J recombination, indicating involvement in immune response.

The data can be analyzed beyond what is shown in this tutorial. For example, differentially expressed genes between the clusters can be identified using Differential Expression for Single Cell. The UMAP plot can be further investigated by coloring the cells using different sources of information. For example, the cell annotations contain a predicted 'Latent time', which approximates the real time the cells experience as they differentiate.

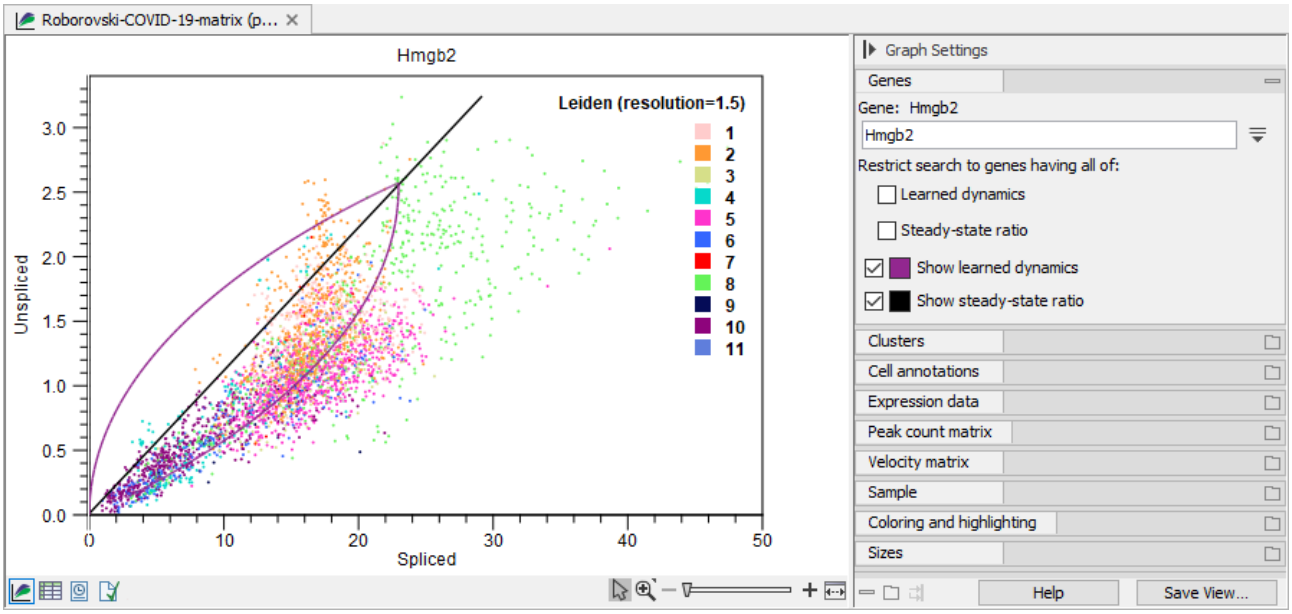


Figure 20: *Phase portrait for Hmbg2.*