Single cell RNA-seq data analysis in Chipster 24.9.2019

chipster@csc.fi

PART I: Finding clusters of cells and marker genes for them

In this tutorial we detect subgroups of peripheral blood mononuclear cells (PBMCs), and we also want to find marker genes for the different cell types. The 10X Genomics data set used in the exercises is available at https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html. The tar package containing the three 10X Genomics output files has been already imported in Chipster for you.

Open Chipster: Go to **https://chipster.csc.fi/,** click on **Launch Chipster**, and log in.

1. Open example session
Click **Open example session** and select the session **course_single_cell_RNAseq_Seurat**.

2. Setup Seurat object & perform quality control
Select the **files.tar.gz.** Select tool **Single cell RNA-seq / Seurat v3 -Setup and QC.** Check the parameters, and **name your project** to PBMC. **Run** the tool.
Open the **QCplots.pdf** in **external web browser**. Look at all the pages.

- What would be the optimal limits for the number of genes (nFeature_RNA) and mitochondrial transcript percentage (percent.mt)?
- How many cells are there?

3. Filter cells, normalize expression values, regress out unwanted variation and detect variable genes
Select **seurat_obj.Robj** created in the previous step, and the tool **Seurat v3 - Filter, normalize, regress and detect variable genes.** Are the default cell filtering parameters good for this dataset, based on the QC plots? While the tool is running, click the **More help** button and learn about the four steps this tool performs.
- What are those steps?

Once the tool is done, open the **Dispersion_plot.pdf** in **external browser** and check also the second page.
- How many variable genes are there?
- How many cells were filtered out?

Bonus exercise: Repeat the run but keep cells which express at least 500 genes. How many cells are left now?

4. Principal component analysis
Select **seurat_obj.Robj** from the previous step and run the tool **Seurat v3 - PCA.**
Open **PCAplots.pdf** in **external browser**. Look at the PC heatmaps and the elbow plot.
-How many principal components should we use for clustering? Would 10 be ok?

5. Clustering and detection of cluster marker genes
Select **seurat_obj.Robj** from the previous step and the tool **Seurat v3 – Clustering and detection of cluster marker genes**. In the parameters, set **Number of principal components to use** = **10**.

-Open **clusterPlot.pdf** in external browser. Does the coloring (clustering results) match the grouping shape found by tSNE and UMAP? How many clusters are there?
- Open **markers.tsv** as a **spreadsheet**. What gene is the most specific marker for the first cluster?

Repeat the run but change the perplexity parameter to 5. How does the tSNE plot change? Remove the result files of this extra run.

Repeat the run and set **Which test to use for finding marker genes = MAST**.
-How many marker genes were found?

Compare the marker genes found by the default Wilcoxon test and MAST: Select both **tsv** files and the interactive visualization **Venn diagram**.

-How many genes were found by both methods? Make a new file containing only those genes: Click on the intersect area, go to **Selected tab** on the right, and click **Create dataset from selected.**

6. Retrieve marker genes for cluster 3

Select **markers.tsv** and run **Utilities / Filter table by column value** using the following parameters:

Column to filter by = cluster
Does the first column have a title = no
Cutoff = 3
Filtering criteria = equal-to

-How many marker genes were found for cluster 3? Keep only those genes whose adjusted p-value is < 0.05 (you can highlight the higher ones, right-click and select “Create dataset from unselected”).

7. Visualize markers

Choose **seurat_obj.Robj** generated in step 5. Select tool **Seurat v3 -Visualize cluster marker genes**. Type a marker **gene name** to the parameter field (try for example with MS4A1, LYZ and PF4). Open a **biomarker_plot.pdf** in external browser.

-Is your gene a good marker for that cluster?

8. Quality control: Color UMAP based on mitochondrial transcript percentage and sequencing depth

Choose **seurat_obj.Robj** generated in step 5. Select tool **"Seurat v3 -Visualise features in UMAP plot"** and set **Feature = percent.mt**. Repeat the run so that you set **Feature = nCount_RNA.**

-Are the clusters evenly colored?

PART II: Joint analysis of two samples: Finding common cell types and performing comparative analysis

In this tutorial we compare two samples of PBMCs: control cells and interferon beta stimulated cells. We want to find cluster marker genes that are conserved between the samples, and genes which change expression in response to interferon. We also want to know if this differential expression is specific to a particular cell type.
The data is available at https://satijalab.org/seurat/v3.0/immune_alignment.html. For the interest of time the first two steps were made for you (they are the same what you practiced with one sample above). Open example session **course_single_cell_RNAseq_Seurat_integrated**.

1. DONE: Import gene expression matrices for both samples to Chipster, setup Seurat object, and perform quality control

Select the **immune_control_expression_matrix.txt.gz** and the tool **Seurat v3 -Setup and QC**. Assign the file to **DGE matrix**. Give **project name** = **PBMC_CTRL** and **sample name** = **CTRL**. Require that a gene is expressed in at least **5** cells. Repeat this step similarly for the **immune_stimulated_expression_matrix.txt.gz**, put set **project name** = **PBMC_STIM** and **sample name** = **STIM**.

-How many cells we now have in our dataset?

2. DONE: Filtering, normalization, regression and detection of variable genes

Select **both seurat_obj.Robj files** and the tool **Seurat v3 - Filtering, normalization, regression and detection of variable genes**. Set **Filter out cells which have less than this many genes expressed** = **500**. Click **Run for each**.

-Compare the most variable genes in each dataset. Are there similarities? Differences?

3. Combine two samples

Select **both seurat_obj.Robj**ects from the previous step and run the tool **Seurat v3 –Combine two samples.**

4. Align the samples, cluster cells and visualize the clusters with UMAP
Select the combined **seurat_obj.Robj** from the previous step and run the tool **Seurat v3 –Integrated analysis of two samples**. Open the pdf in external browser.

- How many clusters are there in this data? Do the clusters (colors) separate in the UMAP plot? How many stimulated cells are in the smallest cluster?

5. Find conserved cluster markers and genes which are differentially expressed
Select the **seurat_obj.Robj** from the previous step. Run **Seurat v3 -Find conserved cluster markers and DE genes in two samples** for cluster **3**. Inspect the tables generated by the tool.

Select **de-list.tsv** and the tool **Utilities / Filter table by column value**, and filter the table by column **p_val_adj**, so that you get only genes whose adjusted p-value is **smaller-than 0.05**. Remember to set **Does the first column have a title** = **no**.

-How many genes in this cluster changed expression in response to the interferon stimulation?

Select **conserved_markers.tsv**, and run the tool **Utilities / Filter table by column value** twice to get the list of **p_val_adj <0.05** for columns **CTRL_p_val_adj** and **STIM_p_val_adj.** Select the resulting tables, and draw a **Venn diagram**. Select the intersect area. On the right, go to **Selected tab** and click **Create dataset from selected.**
-How many conserved biomarkers were recognized for cluster 3?

6. Visualize markers and differentially expressed genes
Choose **seurat_obj.Robj** generated in step 4. Select tool **Seurat v3 - Visualize genes with cell type specific responses in two samples**. Type gene names to the parameter field, try for example:
CD3D, GNLY, IFI16, ISG15, CD14, CXCL10
Open **split_dot_plot.pdf** in external browser.

-Is GNLY a conserved cluster marker? If so, for which cluster?
-Which of the genes respond to the treatment regardless of the cell-type?
-In which clusters is the expression of CXCL10 elevated due to the treatment?