Single-cell RNA-seq data analysis in Chipster, 29.-30.5.2023

chipster@csc.fi

PART I: One sample analysis - 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/articles/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 training session

Click **Sessions**, go to **Training sessions** and select **course_single_cell_RNAseq_Seurat**.
Rename the session course_single_cell_RNAseq_Seurat_**your_first_name**

2. Setup Seurat object & perform quality control

Select the **files.tar.gz** and the tool **Single-cell RNA-seq / Seurat v4 -Setup and QC.** Check the parameters, and set **Project name for plotting** = **PBMC**. **Run** the tool.
Select the **QCplots.pdf** and click **Open in new tab**. 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, scale data, regress out unwanted variation, and detect highly variable genes

Select **setup_seurat_obj.Robj** and the tool **Seurat v4 – Filter cells, 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 **Info** to open the tool manual page and learn about the steps this tool performs.
- What are those steps?

Once the tool is done, select the file **Dispersion_plot.pdf** and click **Open in new tab.** Check also the second page.
- How many cells were filtered out?
- What are the ten most highly variable genes?

4. Principal component analysis

Select **seurat_obj_preprocess.Robj** from the previous step and run the tool **Seurat v4-PCA** so that you set **Number of PCs to compute = 20**.
Select **PCAplots.pdf** and click **Open in new tab**. Look at the PC heatmaps and the elbow plot.
-How many principal components should we use for clustering? Would 10 be ok?

5. Clustering

Select **seurat_obj_PCA.Robj** from the previous step and run the tool **Seurat v4 -Clustering** using the following parameters:

Number of principal components to use = **10**
Resolution for granularity = 0.5

-Open **clusterPlot.pdf**. Does the coloring (= clustering) match the grouping found by tSNE and UMAP? How many clusters are there?

6. Detection of cluster marker genes

Select **seurat_obj_clustering.Robj** from the previous step and the tool **Seurat v4 -Find differentially expressed genes between the clusters.** In the parameters, set the parameters as indicated below and run the tool.
Find all markers = FALSE
Cluster of interest = 3
Limit testing to genes which are expressed in at least this fraction of cells = 0.25

Check which markers show higher than 4-fold difference in expression between cluster 3 and all other cells. Select **markers.tsv** and run the tool **Filter table by column value** from the **Utilities category** using the following parameters:

Column to filter by = avg_log2FC
Does the first column lack a title = yes
Cutoff = 2 (why do we put 2 here if we want a 4-fold difference?)
Filtering criteria = larger-than

-How many genes do you get?

7. Visualize markers

Choose **seurat_obj_clustering.Robj** generated in step 5. Select tool **Seurat v4 -Visualize genes**. Type a marker **gene name(s)** in the parameter field. Try for example with MS4A1, LYZ and PF4. You can enter several gene names at the same time, separated by comma (,). Set the parameters

Add labels on top of clusters in plot = yes
Plotting order of cells based on expression = yes
Give a list of average expression and percentage of cell expressing in each cluster = yes

-Are the genes you selected good markers and for which clusters (check both the plots and the tables)?

8. Annotate clusters

Choose **seurat_obj_clustering.Robj** generated in step 5. Run tool **SingleR cluster annotation.**

-Open **SingleR_annotations_plots.pdf** and see how the clusters are annotated. What are the cells in cluster 3?

9. Rename clusters

Select **seurat_obj_clustering.Robj** generated in step 5 and **cluster_names.tsv** which contains cluster annotations based on marker genes from the Seurat vignette. Check that the files are correctly assigned. Run the tool **Seurat v4 -Rename clusters** and open the result file **clusterPlotRenamed.pdf**. How well do these annotations match with what you got from SingleR in the previous exercise?

10. Color named clusters based on mitochondrial transcript percentage

Choose **seurat_obj_renamed.Robj** generated in step 9. Select tool **Seurat v4 -Visualise features in UMAP plot** and set

Feature = percent.mt

Add labels on top of clusters in plot = yes

-Are the clusters evenly colored? Is this what you would expect?

11. Subset based on gene expression

Choose again **seurat_obj_clustering.Robj** generated in step 5. Run tool **Seurat v4- Subset Seurat objects based on gene expression** for gene MS4A1. Then select **Extract information from Seurat object.**

-How many genes are left in the subset?

12. Share a session with a colleague (in this case with Eija and Maria)

Make sure that no file is selected. Go to the **Session info** panel, click the **three dots** next to the session name, and select **Share**. In the new window that opens

-click **Add rule**.

-In the **UserID** field, enter **jaas/support_session_owner**

-set **Rights** = **Read-only** (you don't want us to mess up your session)

-Click **Save**.

-Click **Close**.

Check what is your own UserID: Click on your **username** (top right corner) and select **Account**.

13. Bonus exercise: Repeat the analysis with SCTransform

Repeat steps 3-5, but this time, use the tool **Seurat v4 -SCTransform: Filter cells, normalize, regress and detect variable genes** for normalization. Make the following changes:

In step 4, set **Number of PCs to compute = 50**.

In step 5, set

Normalisation method used previously = SCTransform

Number of PCs to use = 30

Resolution = 0.8

-How many clusters are found now? Which cluster seems to correspond to the cluster 3 obtained with the global scaling normalization previously?

Repeat step 6 but look for marker genes for cluster **2**.

-How many marker genes do you get? Are they the same genes as what you got for cluster 3 earlier? (Tip: use the Venn diagram to compare the markers to those found when the global scaling normalization was used)?

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 cells stimulated with interferon beta. 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/articles/integration_introduction.html. We have already performed QC, filtering, normalization and finding variable genes on these samples for the interest of time (you practiced these steps in the previous exercise sheet with one sample). Open the example session **course_single_cell_RNAseq_integrated_Seurat_v4.1.1**.

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 v4 -Setup and QC**. Assign the file to **DGE table in tsv format**. 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 do we have in our dataset?
-Do you notice anything strange with this dataset?

2. DONE: Filtering, normalization, regression and detection of variable genes
Select **both setup_seurat_obj.Robj files** and the tool **Seurat v4 – Filter cells, normalize, regress and detect variable genes**. Set **Filter out cells which have less than this many genes expressed** = **500** and run the tool ("Run Tool for Each File").

-Compare the most variable genes in each dataset. Are there similarities? Differences?
-Do you think that the filtering parameters we used are good for this dataset?

3. DONE: Combine two samples
Select **both seurat_obj_preprocess.Robj**ects from the previous step and run the tool **Seurat v4 –Combine multiple samples** -this time only once, so choose the option "**Run tool (1 sample)".**

4. Align the samples, cluster cells and visualize the clusters with UMAP
Select the **seurat_obj_combined.Robj** from the previous step and the tool **Seurat v4 – Integrated analysis of multiple samples with parameters**. Set **Number of PCs to use = 30** and **resolution for granularity = 0.5**. Open the pdf.

- 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 **seurat_obj_combined_integrated.Robj**. Run **Seurat v4 -Find conserved cluster markers and DE genes in multiple samples** for cluster **3**. Inspect the tables generated by the tool.

-Open **de-list.tsv** How many genes in this cluster changed expression in response to the interferon stimulation?
-Open **conserved_markers.tsv**. How many conserved biomarkers were recognized for cluster 3?

6. Visualize markers and differentially expressed genes

Choose **seurat_obj_combined_integrated.Robj** generated in step 4. Select tool **Seurat v4 - Visualize genes with cell type specific responses in multiple samples**. Type gene names in the parameter field, try for example: CD3D, GNLY, IFI16, ISG15, CD14, CXCL10. Use comma (,) as a separator.
Open **split_dot_plot.pdf** in new tab.

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

7. Send a support request to the Chipster team

In the top panel, click **Contact**. Click the **Contact support** button. In the small window that opens,

-Click **Attach a copy of your last session** XXX
-Enter your **email address**
-Write a small **message** (you can tell a joke for example)

8. At home exercise: Repeat analysis with SCTransform

Repeat steps 2-6, but this time, use the tool **Seurat v4 -SCTransform: Filter cells, normalize, regress and detect variable genes**. Note, that in all the tools after that, you need to select: **Normalisation method used previously = SCTransform**. After filtering the markers, you can again use Venn diagram to compare those to the ones you got with other methods.

-Compare **integrated_plot.pdf** Do they differ?