Single-cell RNA-seq data analysis in Chipster, 4.-5.3.2025

chipster@csc.fi

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 and filtering 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_v5**.

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 v5 -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: Filter cells

Select **both setup_seurat_obj.Robj files** and the tool **Seurat v5 – Filter cells**. Set **Filter out cells which have less than this many genes expressed** = **500** and run the tool ("**Run Tool for Each File**").

-Do you think that the filtering parameters we used are good for this dataset?

3. Combine two samples

Select **both seurat_obj_filter.Robj**ects from the previous step and run the tool **Seurat v5 – Merge & normalise, detect variable genes, regress and PCA** so that you set **Number of variable genes to return = 2000**. From the Run button select the option "**Run tool (1 job)".**

-Seurat developers “neglected to finely tune this parameter for each dataset” and instead gave some default values for different cases. Based on the elbow plot, how many PCs would you continue the analysis with this time?

4. Align the samples, cluster cells and visualize the clusters with UMAP

Select the **seurat_obj_merged.Robj** from the previous step and run the tool **Seurat v5 – Integrate multiple samples** with default parameters. 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_integrated.Robj**. Run **Seurat v5 -Find conserved cluster markers and DE genes in multiple samples** for cluster **3**. Inspect the tables generated by the tool.

-Open **de-list_stim1vsAllOthers.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_integrated.Robj** generated in step 4. Select tool **Seurat v5 - 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 3-6, but this time, use the tool **Seurat v5 -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?

PART III: Joint analysis of six samples – Pseudobulk analysis

To demonstrate the use of the pseudobulk tool, we need more samples. In this session, we have 6 peripheral blood mononuclear cell (PBMC) samples originally from the covid dataset GSE149689. 3 of the samples are normal controls from healthy patients, and 3 are covid samples from patients with COVID-19.
Our collaborators Åsa Björklund and Paulo Czarnewski from NBIS kindly provided these samples to be used in our example sessions. We have already performed QC, filtering, normalization and finding variable genes on these samples for the interest of time. Open the session:
course_single_cell_RNAseq_integrated_analysis_Covid_6samples_Seurat_v5

1. DONE: Setup Seurat object and perform quality control

There are now 6 sample files in hdf5 (.h5) format. Check the names of the files and detect which are covid samples and which are normal.
Select the first sample, **CoV_PBMC_15.h5** and the tool **Seurat v5 -Setup and QC.** Assign the file to **10X or CellBender filtered feature-barcode matrix in hdf5 format**, give
project name = **covid_vs_normal**
sample name = **covid15**
sample group = **COVID**
and **run** the tool.
Repeat this step similarly for the **other samples** (keep the project name the same, but alter the sample and group name as needed, for example for file Normal_PBMC_14.h5 sample name = normal14 and sample group = NORMAL). Pay attention when typing the sample group name.

-How many cells do we have in our dataset?
-What would be optimal parameters here? Can you spot the empties, duplets and broken cells?

2. DONE: Filter cells
Select **all six setup_seurat_obj.Robj files** and the tool **Seurat v5 – Filter cells**. Set
Filter out cells which have more than this many genes expressed = **4100**
Filter out cells which have higher mitochondrial transcript percentage = **20** run the tool ("**Run Tool for Each File**").

-Did the filtering improve the situation? How many cells were removed from each sample? You can try with different threshold also, if you like.

3. DONE: Merge the samples
Select **all six seurat_obj_filter.Robj**ects from the previous step and run the tool **Seurat v5 – Merge & normalise, detect variable genes, regress and PCA** (choose the option "**Run tool (1 job)"**).
-How many cells are left at this point? How many were filtered out?

4. DONE: Align the samples, cluster cells and visualize the clusters with UMAP
Select the **seurat_obj_merged.Robj** from the previous step and run the tool **Seurat v5 – Integrate multiple samples** with default parameters. Open the pdf.

-How many clusters are there in this data?
-Can you spot any UMAP clusters that are only present in one the samples before integration? How about after integration?

5. Differential expression in sample groups
Let's try finding the differentially expressed genes in cluster 3 in few different ways.

First, select **seurat_obj_integrated.Robj** and tool **Seurat -Find DE genes between sample groups** and type:
Name of the sample group to compare with = **COVID**
Name of the sample group to compare to = **NORMAL**.

Then, select again **seurat_obj_integrated.Robj** and tool **Seurat -Find DE genes between chosen sample** and type:
Name of the samples to compare with = **covid15, covid17, covid1**

Name of the samples to compare to = **normal14, normal13, normal5**.
-Compare the two resulting files in Venn diagram. Are the lists identical?

Finally, select again **seurat_obj_integrated.Robj** and tool **Seurat -Find DE genes between sample groups, pseudobulk** and type:
Name of the sample group to compare with = **COVID**
Name of the sample group to compare to = **NORMAL**.
-Compare this gene list to one of the lists above. Are these lists identical?