Babraham Bioinformatics

Analysing 10X RNA-Seq data with Seurat

Version 2019-06

(Seurat v3)

Licence

Introduction

In this exercise you are going to see what additional options and levels of control you can have on your data by using R to process the same data we previously looked at in loupe.

We are going to use elements of the Seurat R package to look at this data. We will perform some QC and filtering on the data before selecting interesting genes, doing a dimensionality reduction projection, and finding genes which characterise the different subgroups. Many of the operations will be similar to what you previously did in loupe, but you will see the additional options and flexibility you get in R, but will also see the complexity which comes with this.

Structure of this exercise

You have been provided with a document called “Seurat_workflow.Rmd”. This is an R markdown document which is a mix of R code alongside commentary text. The document is a linear workflow which you are going to run through from beginning to end.

This document is NOT intended to be a recipe for how to analyse single cell data but is instead an illustration of the main parts of a typical workflow, along with a demonstration of how you can directly access the raw data for the experiment if you want to look at other aspects of the data. Each analysis will be somewhat different depending on what you find when you explore it and this document aims to show you what sort of things you can achieve.

You are going to load the document into RStudio and will then execute the different blocks of pre-written code and examine the output. Those who are familiar with R will have the opportunity to change the code and calculate additional result. Everyone should ensure that they understand, at least in general terms, what the code is doing and the meaning and limitations of the results it produces.

If you have any questions about the data, analysis or the R code then please ask.

Loading the document into RStudio

From your start menu you should be able to find the RStudio program (an icon for it will probably also be on your desktop). If you open this you can then go to:

File > Open File

..and then you can select the “seurat_workflow.Rmd” document.

Running the exercise

Within the document you have there are blocks of R code. During the exercise you are going to run these sequentially and look at the results, which will be inserted into the document just underneath the block you run.

Each R code block has a structure which looks like this:

```
```{r}
[Some R code goes here]
```
```

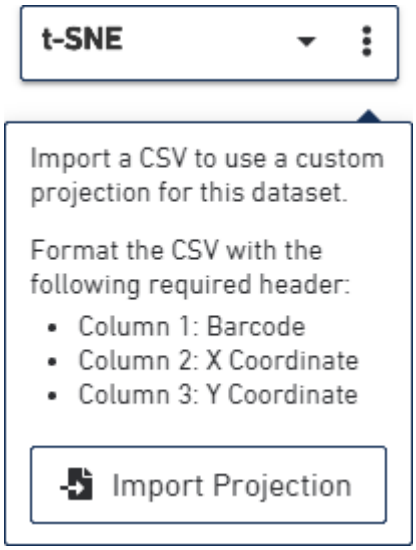
The simplest way to run a block of code is to press the small green “play” arrow at the top right of the block. You can also run it by clicking into the block and pressing Shift + Control + Return.

Most of the blocks will complete within a few seconds. The only one which will take a long time is the block which calculates enriched genes for all of the clusters you define. This will take a few minutes to complete but you should see a progress bar being shown as the calculation is proceeding.

At the end of the exercise there are two parts which step out of the standard Seurat workflow.

1. We use the Sleepwalk package to produce a really useful interative projection which will open in your web browser. In this figure you can mouse over every point in the projection and can see the graph coloured by the calculated distance measures to all other cells. This makes a really nice way to explore the structure of your projection.

2. To link back to the work we did before in loupe at the end of the Seurat exercise we export out our tSNE projection with the additional filtering we did. You can then import this file back into Loupe so you can see the effect this would have on the predictions and clusters you originally looked at. See whether there are clusters which you would no longer consider to be robust. To load the Seurat tSNE data into loupe you click on the 3 dots to the right of where it says “tSNE”, and you will see a menu open with the option to import a projection.



Extensions

If you are comfortable with using R then feel free to try out variations on the main analysis so you can be sure that you understand the code, and can start to get comfortable with configuring your own analysis.

Some suggestions for things you could try:

- The highly expressed genes list shows that a large number of the most highly expressed genes are ribosomal proteins. Go back to the QC section and calculate a value for each cell showing the amount of ribosomal sequences. These are characterised by names which start with "RPL" or "RPS" the pattern to detect these would be "`^RP[LS]`".

- Try changing the parameters of the analysis. The obvious main things you could consider changing would be:

    - The QC cutoffs used to filter your cells
    - The section of variable genes
    - The number of PCs taken forward for distance calculations
    - The resolution of the clusters calculated by FindClusters

- Try replotting the clusters you identified on tSNEs calculated with different perplexity values to see how this changes how distinct they are.

- [Advanced] You can see from a few of the graphs that there is quite a high proportion of the data which "drops out" – ie has no counts at all in many cells. An interesting thing to look at is to calculate for each gene the proportion of cells which "dropped out" for that gene, and compare that to the mean expression level for cells which didn't drop out. You should observe that there is a general negative correlation between these two metrics, but that interestingly there are genes which drop out much more frequently than expected – these are likely to be genes with differential expression in different sub-populations of cells.