82H: Stem Cell RNA-Seq Analysis Log

Emily Fradley

09/11/2020

Contents

1 Set Up, Import and Tidy

We will be using R package `scran` (Lun, McCarthy, and Marioni 2016) to analyze differences in expression data between LT-HSCs, HSPCs and Prog cells.

First we need to import the data, tidy the format and check it's structure to determine whether it is appropriate for the analysis we are going to do and to be able to interpret the results later. The data is in the form of 3 different data sets, one for each cell type.

1.1 Data Stucture

First we will check to see what the dimensions of the imported data looks like. Visually looking at the data shows us that each **gene is a row** and **each cell is a column**. We can find out the exact number of each using the code below:

```r
#Use ncol and nrow to extract the number of cells and genes respectively
#This can be double checked against the numbers reported by the str function

#LTHSCs number of cells and genes
ncol(lthsc)
nrow(lthsc)

#HSPCs number of cells and genes
ncol(hspc)
nrow(hspc)

#Prog number of cells and genes
ncol(prog)
nrow(prog)
```

So all the data sets have **423 genes** or rows and the **number of cells** (i.e. columns) are respectively:
- **LTHSCs: 155**

- HSPCs: 701
- Prog Cells: 798

Next we can look to see if any of the genes or cells from the different data sets overlap using the `intersect` function.

By comparing the rows and columns of each data set, we can see that the number of overlapping rows is 423 for all 3 data sets. This is the same as the total number of rows of all the data sets so we have confirmed that all the data is showing **expression data for the same genes**.

For the overlap of the columns (i.e. the individual cells) we find that we have overlap between the data sets. This means **all the cells in each set are different from each other** and **different from the cells in the other data sets**.

Now we know that all the data sets contain the same 423 genes but all have expression data for unique cells. This is what we expected.

1.1.1 Note

The original collection of the expression data contained over 2000 genes, our data has been cut down to include just over 400 genes. This has likely excluded any genes that showed no expression across the board of the different HSCs as any genes that weren't expressed would have been pointless to include in the analysis as we are looking to differentiate the cells based on differing expression patterns. There could be other reasons for the removal of this data though ***

1.2 Normalised or Log Data

To find genes of interest we are going to use the `findMarkers` function but to do that the expression data needs to be in **normalised log format**. The data has already been normalised and corrected for biases so all we need to do is check whether the data is in linear or log format.
We can do this by checking the the maximum value of the expression data. We would expect linear data to return a very large number for this (in the thousands). In contrast we would expect log data to return a relatively small maximum value.

The highest expression figures for the different data sets are as follows:
HSPCs: 17.637044
LTHSC: 15.0144247
Prog Cells: 16.0601145

This shows us that our results are most certainly log transformed and we can continue with the analysis of the data without having to transform it further.

1.3 Summary Statistics

Finally we want to summarize the data we have for each gene and cell in a way that is easy to absorb (i.e. a plot). To do this we first have to generate some summary data including the mean and the standard deviation for each cell.

After that we want to use the summary statistics to produce plots that will summarize the data sets for us and give us an idea of the overall expression patterns we have.

```r
names(prog_sum)
```

```
## [1] "Mean" "SD"
```

References

Lun, Aaron T. L., Davis J. McCarthy, and John C. Marioni. 2016. "A Step-by-Step Workflow for Low-Level Analysis of Single-Cell Rna-Seq Data with Bioconductor." *F1000Res.* 5: 2122. https://doi.org/10.12688/f1000research.9501.2.