82H: Stem Cell RNA-Seq Analysis Log

Emily Fradley

09/11/2020

Contents

1 Set Up, Tidy, and Summerise

We will be using R package `scran` (Lun, McCarthy, and Marioni 2016) to analyse 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 Structure

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 Linear 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.

1.4 Visual Summary

NEED TO COME BACK TO THE FIRST PLOTS
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. We will use 2 plots:

A good way to view the all of the expression data at once is to use a heat map. It's easier to digest visually but still gives us a good idea of expression patterns and early differences in the cell types.

The second graph is a parallel coordinate graph that we will use to plot the expression datas against each other.

2 Results

2.1 findMarkers

To use `findMarkers` we first need to combine the different cell type datasets so we can compare them. Then we apply the function to return statistical analysis results that tell use if there is difference between each genes expression in the different cell types. The results will return for each gene the **p-value**, **q (FDR) value**, and the **logFC (fold change)**.
Now that we have these results saved we need to filter them to find our genes of interest.

2.2 Filtering Results

2.2.1 Statistical Significance

To determine which genes returned **statistically significant** results we need to filter them. This would usually done by p-value but since we have such a large data set we need to adjust the p-values to account for false positive results. The `findMarkers` function has already calculated the FDR (False Discovery Rate) or q value for us so now all we have to do is filter the results so we **only keep the genes with a FDR<0.01**. Simply this means we accept that the filtered results will include genes that are, at most, 1% false positives.

For context we filtered the results with FDR of 1%, 5% and 10% to see the difference in the number of statistically significant genes we are left with. The results can be seen in the table below:

\begin{table}

\caption{(#tab:FDR table)The number of genes that have a statistically significant difference in expression between the cell types HSPC & LTHSC, LTHSC & Progenitor, and LTHSC & Progenitors respectively when the maximum FDR value is set to 1%, 5%, and 10%.}

	1%	5%	10%
HSPC/LTHSC	89	125	151
HSPC/Prog	250	282	307
LTHSC/Prog	209	243	264

\end{table}

2.2.2 Relative Differences

Next we need to take the results that we know are statistically significant and filter them by the actual difference in expression between cell populations. If the difference is too small it wouldn't matter if it is statistically significant or not as it will be unmeasurable in a practical setting and won't help us differ the cell types. The aim is to find genes that will act as indicators/markers for the cell type by their expression.

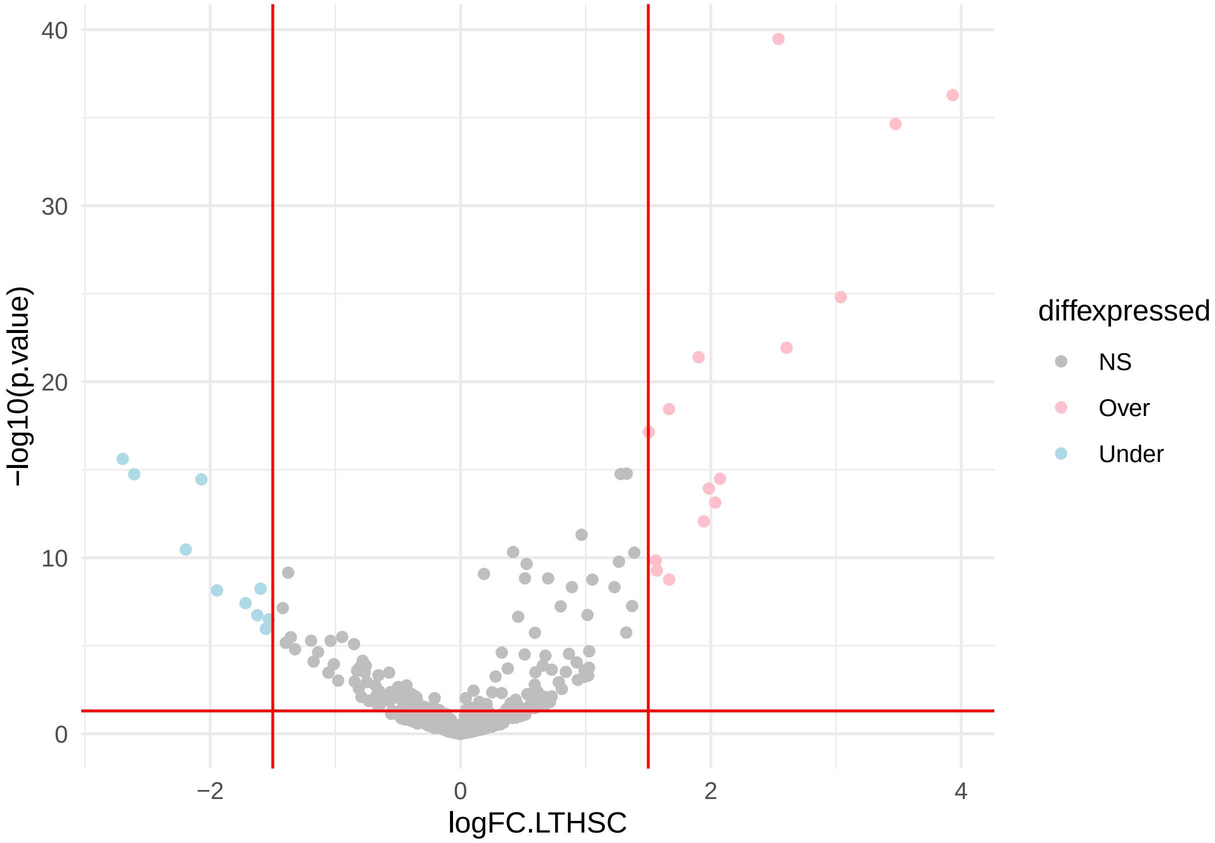
To asses the difference in the expression we use the `logFC` part of our `findMarkers` results. This will tell us the difference between our log2 expression values for the genes. That is a difference of 2 would represent a 2 fold (2x) increase in expression.

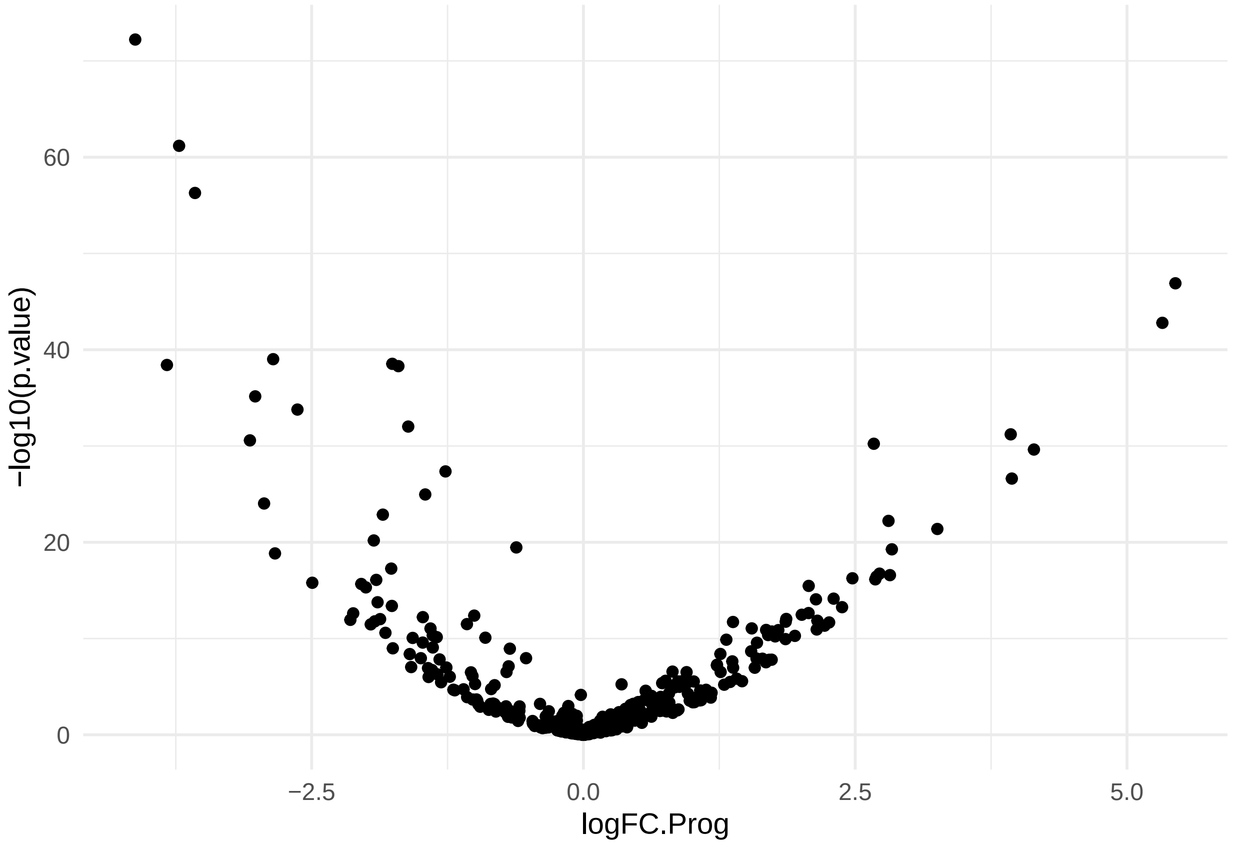
2.2.3 Absolute Expression

To help us get a better understanding of the actual expression levels to ensure we again have high enough expression to be measurable in practice we need to look at the genes **absolute expression**. We can do this by using the mean expressions of each gene that we already calculated earlier.

2.3 Visual Results

To view both the statistical significance (p-value) and the the biological significance (log2fold change) at the same time we can use a volcano plot.





We can also visualise results using an MA plot.

2.4 Differentially Expressed (DE) Genes

For this project we will be investigating **LT-HSC** which means we want to identify genes that are different between the LTHSCs and the other 2 cell types.

2.5 Biological Investigation

Now we can use our list of differentially expressed (DE) genes and look at their biological function and orthologs to see if we can find any correlation between the statistically interesting genes and the biologically interesting ones. First we want to run a GO enrichment analysis. Since the genes of our original dataset were already filtered to only include secretome genes then we cannot assume that all the information given by the g:profiler software is inherent to our cell type. We must separate out that information from the information on enriched genes.

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.