Babraham
Bioinformatics

Final Exercise:
Using the R Tidyverse Packages

Version 2019-04

Licence

Part 1: Data Preparation

In this exercise, you will combine the knowledge of the different parts of the tidyverse to do a more complete processing and visualisation of a more complex dataset.

You have been provided with three files;

1. `methylation.txt` has promoter and gene body methylation values for all genes from oocytes
2. `expression.txt` has gene expression values (log2 FPKM) for oocytes
3. `final_annotation.txt` has chromosomal location information for the genes

Combine all of the data into a single tibble

To be able to do an integrative analysis of this data we need to make up a single tibble which has all of the data for methylation and expression in it, as well as the gene annotation information.

We only want to work with genes for which we have measures for both methylation and expression, so we can remove any genes which only have measures for some data types.

In the final tibble we construct we should have columns of:

- Probe
- Expression
- Gene_body_meth
- Promoter_meth
- Chromosome
- Start
- End
- Strand

Clean up the data

To get to a cleaner set we are going to apply some filters

1. Remove any genes whose total length is less than 10kb
2. Remove any genes whose promoter methylation is -1 (ie unmeasured)
3. Remove any genes whose name starts with “Gm” (use `substr` in a `filter` statement along with `!=`)

Do some summarisations

Find the average expression, gene body and promoter methylation per chromosome and annotate this with the number of genes present.

Make up a tibble with columns;

1. Chromosome
2. Number of genes on the `+` strand

3. Number of genes on the – strand
4. Difference between the + and – strand

Add a column which says whether each gene is highly expressed. High expressed counts as an expression level >0.

Calculate the median Gene Body meth and Promoter meth for genes which are and aren’t high expressed.

Section 2: Plotting and visualisation

Distributions

- Draw out the distributions of values for the methylation and expression datasets. In each case make the graph look nice and add a suitable title.
  - On the graph of expression levels draw a strong vertical line at 0 to show where we made the distinction between high and low expressed genes.

Summaries

- From the original methylation data (not the cleaned one you made before). Plot out a bar graph showing the mean gene body methylation level per chromosome.

- Plot out a stripchart, split by chromosome of the promoter methylation levels in the 500 highest expressed genes.
  - Repeat the above plot but using the 500 lowest expressed genes

Comparisons

- Plot out a scatterplot of the relationship between the Promoter methylation and the Gene body methylation.
  - Colour the plot above by expression level.
  - If you're really feeling bold, change the colour scale for the expression to be a diverging scale starting blue, going through white and ending up red.

- Plot out the relationship between expression and gene body methylation as a scatterplot. To summarise this you can try adding a geom_density_2d representation over the top of the scatterplot.

- Draw side-by-side violin plots of the distributions of gene body methylation for the high and low expressed sets you defined before.

- Draw a barplot showing the mean gene body methylation +/- stdev for the high and low expressed groups. You'll need to do a new summary to calculate these values and then plot them.