Week 5 – Base R

R-help mailing list (questions/ discussions of problems - contribute with solutions)

BiocManager::install(“DESeq2”) #Means from this pacakhe (Bioc) use this function (install) to get package

my\_list[[2]] # to access second item in list (could be an alternative dataframe)

conda env create -n obds-r -f obds\_r.yml #rsynced across yml file containing all the programs and created environment

lapply(vector/list, FUNCTION) #gives you a list back always

sapply #Apply a function to a list but may give you a simplified vector back

Running max on a character vector gives you the last thing alphabetically

1 is max in Boolean vector

Matrix – collection of elements (row names and column names etc)

Matrix(1:10, nrow=2) #fills by columns and gives 2 rows

T(matrix) gets the transpose of the matrix

Iris[1] #extracts with index, or can use $column name – outputs are different though

Rows[1,] #extra rows (before comma)

#Extract using Booleans - $name ==”setosa”, ] #rows

& allows use of multiple condtions

%in% #Says one ofor the other condition

| #means or

View(irs) opens a view data as spreadsheet

Apply #applies a function over rows or collumns (MARGIN specifies rows/columns)

Plotting in R- using base R

Par() sets global paramters

Mfrow

Day2: GGplotting

Each ggplot object needs to have a geom (marks on the plot)

Ggplot(diamonds)#initialise plot but wont generates plot

Ggplot(diamonds, aes(x=carat, y=price)) #generates plot and x/y axis

+ Geom\_point() #makes it into a scatter plot

Colour = cut #colour by cut variable of diamond dataset

Ggplot automatically gives you a legend

Specifying the colour within geom\_point(aes(colour=cut)) will just colour the points

If you specify aes within the ggplot() original function will be passed on to all geoms

Ggplot2 themes: modify anything in plot that is structural (basically not points)

Can use standard themes (ie. Theme\_classic())

Modify theme aspects individually using labs(title, x, y) + theme(axis.title = element\_text(size , face), axis.text

Arrange plots into a grid: cowplot: plot\_grid(plot1,plot2, labels, ncol, nrow)

Also can use gridExtra with grid.arrange()

Split by variable: facet\_grid(cut ~ colour) #cut on y axis, color on x – easy way to look at exercises

Facet\_wrap can only facet by 1 variable (variable name and column num)

Tidyverse:

Transforming data using dplyr, pipe and broom. Organising data with tidy

Tidy data – structure data for analysis

1 varaible per column, each row is single observation, each cell is one value

Piping to combine operators – make a sentence of code

Tibble is a dataframe in R -does less and complains more.

Tidyverse packages: dplyr functions including group\_by(), select(), join()

Tidyr – tidying data into tidy format – gather() #puts multiple columns into key value pairs

1 file: 2 columns per mouse. RFP+ cells and total cells

2 file: mouse id, sex of mouse and day post labelling for measurement

Files are untidy as rows don’t represent an individual observation

Library(tidyverse)

Raw\_data <- read\_csv(“path”)

#Parsed with column specification – col double

Mouse id in column

Want to convert into tidy format to plot the data

Gather function

Test\_with\_na <- raw\_data %>%

Gather(mouse\_typeofcount, count, -crypt\_no)

Typeofcount#key count #value #except for crypt number

Give data, key(new column), value #New column containing values

Na.rm=TRUE #removes any columns where na is value

Separate() Can split the mouse\_typeofcount by

Data\_processed <- data\_processed %>%

Separate(mouse\_typeofcount, c(“Mouse\_ID”, countype”), sep=”\_”)

separate by sep \_

Wwant to use spead function to spread counttype and count

Spread moves the key from count\_type to make two new columns, so give key column to spread (make into two columns) and value column tto become the values

Now for second table of mouse info (mouse id, sex of mouse and date)

Rename columns, then use select to reorder the coumns

Plot histogram of total cells

Wrap by mouse ID to see if total number of crypts varies by mouse or by day of experiment

Facet grid allows split by day and post label

Plot histogram of RFP pos cells, then facet wrap by day post label – peak counts lower in 10, as divided at day 21

Apha =0.2 within geom\_point() #changes transparency of the plot

#Add a column of proportion of RFP cells

Summary stats using group\_by() and

Total cells > 15 #existing columns don’t need to be in quotes

Select specific columns (: can select multiple columns across dataframe)

Slect on value in a column with == “id”

Sleect to remove a column using -col name #selects all columns but col

For the exercise:

Data is RNAseq(mouse CD4/CD8 etc – 3 biological reps, 12 samples)

Two files are obds\_countstable.tsv.gz and obds\_samletable.tsv

Thursday 28th

Built in stats approach – stats are available using fuctions

Each distribution has a route name

Mean – also written as location

Sigma squared is variance

Dnorm is density

Rnorm – test whether your dataset is normally distributed

Can se on slide 6/7 how the function calls change when you change mean of a normal dist

Binomial distribution

R functions for testing all end in .test

If you want to test if data is normally dist, p value is probability that data is Not normally distributed

Each test makes assumptions, weneed to make sure assumptions are met.

Median/mode can also explain distribution

T test assumes normal distribution

Chapiro test tests whether your data is normally distributed

Or if you have log normal distribution then give you a lower p value

Can fool shapiro test with enough points on graph. Qqnorm() gives you a plot for visually expected plot. Plots lowsst value of vector to lowest value of yours – should give diagonal line, log normal distribution will give a curved line,

If assumptions are not met – use a non-parametic test (not as powerful as don’t assume as much)

Can apply non-parametric to normal distribution will just be more conservative

t.test(value ~ group, test\_data) #Test value between group (only works for if you have 2 groups

runif #gives not normal distribution – would use man whitney u test. Test stat is W

Corrections for multiple testing – Bonferroni - taking account of number of tests and times by p values (capped at 1)

Compares P values to expected ones (BH procedure) looks at when p values cross line – non significant. Ranked p values

Inherently