1st June 2020 – Week 6 – day 1

Differntial analysis

Input for de analysis – not normalised data (wouldn’t get an accurate outcome)

EdgeR – easier to play with parameters where Deseq2 is more pipelined

Sleuth – at a transcript level

DEXseq – looking at exon usage

Diffbind

Class of object forDEseq – stores raw read counts, differential expression – Ranged summarised experiment – create from range of inputs

Input from Kallisto/salmon (pseudo alignment output)

Main count data – Samples are columns, genes are rows

Bioconductor summarised experiment

Features (gene name or ensemble id/ peak number for attace seeq)

Features data from genes (ie stt and stop etc)

Multiple assays

rowData (gene data information)

for deseq – always need sample information table (tells the algorithm which groups to compare)

Count data –

~ condition+batch #Looking at the effect of the condition on batch

~ X+y

Looking at the interaction effect of DEseq2

Sva/RUVseq looks for other unknown souces of variation (but hard to understand this and what the variable represents)

Main steps of DEseq2 – size factors,

Normalisation accounts for effect of other v highly expressed genes

Calculates a geometric mean – (mean of logs) less skewed from outliers

Remove all genes where any samples have a 0 – focusing here on housekeeping genes for size factors

Size factor calculation –

Step 1. Estimating size factor

2. Dispersion (BCVsquared) variation in counts between replicates – variation between biological replicates

- because there aren’t many replicates per group, the dispersion calculation might not be very accurate so DeSEQ2 looks at other genes with similar levels of expression for their dispersion values

Higher the expression of a gene, lower the dispersion. Black spots are individual gene levels dispersion – how does dispersion change based on mean counts for that gene. Genes that have a lower dispersion tha average are moved towards the trend of dispersion.

After dispersion

Model linear combination of explanatory factors – compare between conditions of interests using wald test

Independent filtering of results – don’t want to test every gene as multiple testing problems

Deseq removes some genes to remove multiple testing burden – not likely to come out as DE for low count genes

Dispersion likely too large to identify statistically significant differences in expression between groups

Dds<- dds[rowSums(counts … #filter

DEseq uses BH multiple correction

logFC shrinkage – separate step after wald test – shrink logFC

some low count genes have v high fold change becase of low expression – squash the low count genes towards axis so when you rank your p values, low count genes don’t dominate

apeglm is default , normal squashes logFC more

DESEq transformed counts – logcounts+pseudocounts

Heteroskedastic – we want the variance to be consistent as expression changes

Transformed counts also normalised to library size

Exercises

Coutns table want features as rows

Metadata want

Pathway analysis – Day 2

Simplify complex gene lists

P values associated with pathways

Testing multiple

Gene set – broad definition

Gene pathway – more defined

Gene sets – papers, transcriptomic/proteomic

Pathway analysis DAVID – not a definitive answer, just for hypothesis generation

Databases are continually changed, others are out of data – want to make sure you choose databases that are up to date

Where experimental data comes from – candidate study?

Some genes are also annotated more than others

Organsisms too

Curation – manual v electronic(paper pulling no manual curation)

Over representation analysis – statistical method – looks at the number of genes in your gene list (ie list of DE genes with log2FC>1 and padj <0.05)

Count number of genes overlapping with gene list. Repeat with background list (ie all genes)

Fishers exact test

Advantages of ORA – don’t need ordered gene lists

Cluster markers – pulled joint list of cluster markers across exeriments (without fold change)

Usually only takes most significant genes

Ignores fold changes – equal importance

Statistical procedures that assume independence

Assumes pathways are independent of each other – GO terms are hierarchical

functional class scoring – GSEA is an example

ie all genes in apoptosis – look at patterns and general trend

Stastical enrichment of gene lists – need a ranked gene list

Gene level statistic (ie fold change)

Pathway level stastic

Significance assessed

GSEA: enrichment score: degree to which gene set is overrepresented at the top or bottom of a ranked list of genes

Correlation of gene with phenotype – explains magnifitude of increment

Enrichment score

Significance of ES – random shuffle which groups phenotype belongs to – null distribution, calculate enrichment score.

Normalised enrichment score and FDR value

Probability density function (value being less than x – PDF v CDF graph)

Genes in gene list randomly raked – more likely to find genes at top?

Max distance between blue n red line – maximum deviation is statistic

Ranked gene list – enrichment score(biggest deviation away from line) ( genes that most contribute to ES are at leading edge)

* This uses all data – small changes in many genes
* Gene sets between labs – more consistent representation of results
* Assumes that most DE genes are most critical – rank rather than expression
* Long lsit of gene sets – overlapping pathways

GSEA can either receive count matrix – fold change differences or can give rnked list

Barcods – all the genes in the matrix – width of barcode is determined by all genes, black line is when each of your genes (from list) are involved in phenotype

XGR

pathway topology

TOPgo/goseq – go terms are hard to do anything with as they’re very vague

Installing new Rversion

Problem is that we need to install all our new packages too

pkgs <- rownames(installed.packages()) #grabs all packages

write.table(x = pkgs, file = "/Users/rhodgson/package\_list.txt", row.names = FALSE, col.names = FALSE, quote = FALSE) #writes to a text file

Close R Studio

<https://cran.r-project.org/bin/macosx/>

click on R-4.0.0.pkg(notarized and signed)

Download. Click through

<https://rud.is/rswitch/guide/>

for if you want to keep more than one version of R

Download RSwitch

Open Rstudio again

Don’t worry about the errors if you haven’t cleared your enviroments

pkgs <- scan("/Users/rhodgson/package\_list.txt", "character")

Need to install BiocManager

install.packages("BiocManager")