Copy_of_Deseq2_R

January 1, 2022

[ ]:
```
?system
version
```

```
Warning message in seq_len(head.end.idx):
"first element used of 'length.out' argument"
ERROR while rich displaying an object: Error in seq_len(head.end.idx): argument
must be coercible to non-negative integer

Traceback:
1. FUN(X[[i]], …)
2. tryCatch(withCallingHandlers({
 .     if (!mime %in% names(repr::mime2repr))
 .         stop("No repr_* for mimetype ", mime, " in repr::mime2repr")
 .     rpr <- repr::mime2repr[[mime]](obj)
 .     if (is.null(rpr))
 .         return(NULL)
 .     prepare_content(is.raw(rpr), rpr)
 . }, error = error_handler), error = outer_handler)
3. tryCatchList(expr, classes, parentenv, handlers)
4. tryCatchOne(expr, names, parentenv, handlers[[1L]])
5. doTryCatch(return(expr), name, parentenv, handler)
6. withCallingHandlers({
 .     if (!mime %in% names(repr::mime2repr))
 .         stop("No repr_* for mimetype ", mime, " in repr::mime2repr")
 .     rpr <- repr::mime2repr[[mime]](obj)
 .     if (is.null(rpr))
 .         return(NULL)
 .     prepare_content(is.raw(rpr), rpr)
 . }, error = error_handler)
7. repr::mime2repr[[mime]](obj)
8. repr_html.help_files_with_topic(obj)
9. repr_help_files_with_topic_generic(obj, Rd2HTML)

               _
platform       x86_64-pc-linux-gnu
arch           x86_64
os             linux-gnu
system         x86_64, linux-gnu
```

```
status
major          4
minor          1.1
year           2021
month          08
day            10
svn rev        80725
language       R
version.string R version 4.1.1 (2021-08-10)
nickname       Kick Things
```

Analyze gene count data using Deseq2

[ ]:
```
install.packages("rgl", repos = "http://cran.rstudio.com/")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘lazyeval’, ‘htmlwidgets’, ‘crosstalk’
```

[ ]:
```
# Install specific version of XML for compatibility with following packages␣
 ↪(CosRank and Deseq2)
packageurl <- "https://cran.r-project.org/src/contrib/Archive/XML/XML_3.98-1.20.
 ↪tar.gz"
install.packages(packageurl, repos=NULL, type="source")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

[ ]:
```
install.packages("ConsRank", repos = "http://cran.rstudio.com/")
library("ConsRank")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘rlist’, ‘proxy’, ‘gtools’


Loading required package: rgl

Warning message in rgl.init(initValue, onlyNULL):
"RGL: unable to open X11 display"
Warning message:
```

```
"'rgl.init' failed, running with 'rgl.useNULL = TRUE'."

Attaching package: ‘ConsRank’


The following object is masked from ‘package:base’:

    labels
```

[ ]:
```
system("add-apt-repository -y ppa:marutter/rrutter")
system("add-apt-repository -y ppa:marutter/c2d4u")
system("apt-get update")
system("apt install -y r-cran-rstan")
```

[ ]:
```
install.packages("ggplot2")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

[ ]:
```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

[ ]:
```
BiocManager::install(version="3.13", ask=FALSE)
```

```
'getOption("repos")' replaces Bioconductor standard repositories, see
'?repositories' for details

replacement repositories:
    CRAN: https://cran.rstudio.com


Bioconductor version 3.13 (BiocManager 1.30.16), R 4.1.1 (2021-08-10)

Installing package(s) 'BiocVersion'

Installing package(s) 'pbdZMQ', 'XML', 'backports', 'cli', 'crayon',
  'generics', 'openssl', 'sessioninfo', 'tzdb', 'usethis', 'uuid', 'nlme'
```

[ ]:
```
BiocManager::install("DESeq2", version="3.13")
```

```
'getOption("repos")' replaces Bioconductor standard repositories, see
'?repositories' for details

replacement repositories:
    CRAN: https://cran.rstudio.com


Bioconductor version 3.13 (BiocManager 1.30.16), R 4.1.1 (2021-08-10)

Installing package(s) 'DESeq2'

also installing the dependencies ‘bitops’, ‘formatR’, ‘plogr’, ‘png’,
‘Biostrings’, ‘RCurl’, ‘GenomeInfoDbData’, ‘zlibbioc’, ‘matrixStats’,
‘lambda.r’, ‘futile.options’, ‘RSQLite’, ‘KEGGREST’, ‘xtable’, ‘GenomeInfoDb’,
‘XVector’, ‘MatrixGenerics’, ‘DelayedArray’, ‘futile.logger’, ‘snow’, ‘BH’,
‘AnnotationDbi’, ‘annotate’, ‘S4Vectors’, ‘IRanges’, ‘GenomicRanges’,
‘SummarizedExperiment’, ‘BiocGenerics’, ‘Biobase’, ‘BiocParallel’, ‘genefilter’,
‘locfit’, ‘geneplotter’, ‘RcppArmadillo’


Old packages: 'backports', 'cli', 'crayon', 'generics', 'openssl',
  'sessioninfo', 'tzdb', 'usethis', 'uuid', 'nlme'
```

[ ]:
```r
# Download the Data
```

[ ]:
```r
countsName <- "https://raw.githubusercontent.com/Mangul-Lab-USC/
 ↪Biomedical_workshop_assignment_part2_datasets/master/Alberttl%40usc.edu/
 ↪gene_matrix.csv"
download.file(countsName, destfile = "gene_matrix.csv", method = "auto")

countData <- read.csv('gene_matrix.csv', header = TRUE, sep = ",")
head(countData)

metaDataName <- "https://raw.githubusercontent.com/Mangul-Lab-USC/
 ↪Biomedical_workshop_assignment_part2_datasets/master/Alberttl%40usc.edu/
 ↪meta_data.csv"
download.file(metaDataName, destfile = "meta_data.csv")
metaData <- read.csv('meta_data.csv', header = TRUE, sep = ",")
metaData
```

A data.frame: 6 × 9

	ensgene <chr>	SAMPLE1 <dbl>	SAMPLE2 <dbl>	SAMPLE3 <dbl>	SAMPLE4 <dbl>	SAMPLE5 <dbl>	S
1	ENSG00000000003	723	486	904	445	1170	1
2	ENSG00000000005	0	0	0	0	0	0
3	ENSG00000000419	467	523	616	371	582	7
4	ENSG00000000457	347	258	364	237	318	4
5	ENSG00000000460	96	81	73	66	118	9
6	ENSG00000000938	0	0	1	0	2	0

A data.frame: 8 × 2

id <chr>	status <chr>
SAMPLE1	control
SAMPLE2	treated
SAMPLE3	control
SAMPLE4	treated
SAMPLE5	control
SAMPLE6	treated
SAMPLE7	control
SAMPLE8	treated

[ ]:

[ ]:
```
countData <- read.csv('gene_matrix.csv', header = TRUE, sep = ",",row.
↪names=NULL)
head(countData)
```

A data.frame: 6 × 9

	ensgene <chr>	SAMPLE1 <dbl>	SAMPLE2 <dbl>	SAMPLE3 <dbl>	SAMPLE4 <dbl>	SAMPLE5 <dbl>	S
1	ENSG00000000003	723	486	904	445	1170	1
2	ENSG00000000005	0	0	0	0	0	0
3	ENSG00000000419	467	523	616	371	582	7
4	ENSG00000000457	347	258	364	237	318	4
5	ENSG00000000460	96	81	73	66	118	9
6	ENSG00000000938	0	0	1	0	2	0

[ ]:
```
metaData <- read.csv('meta_data.csv', header = TRUE, sep = ",",row.names=NULL)
metaData
```

A data.frame: 8 × 2

id <chr>	status <chr>
SAMPLE1	control
SAMPLE2	treated
SAMPLE3	control
SAMPLE4	treated
SAMPLE5	control
SAMPLE6	treated
SAMPLE7	control
SAMPLE8	treated

Differential analysis Calculation with Salmon data using DESEQ2

Construct DESEQDataSet Object

[ ]:
```
library(ggplot2)
library( "DESeq2" )
```

[ ]:
```
dds <- DESeqDataSetFromMatrix(countData=countData,
                              colData=metaData,
                              design=~status, tidy = TRUE)
```

```
converting counts to integer mode

Warning message in DESeqDataSet(se, design = design, ignoreRank):
"some variables in design formula are characters, converting to factors"
```

[ ]:
```
#Design specifies how the counts from each gene depend on our variables in the␣
 ↪metadata
#For this dataset the factor we care about is our treatment status (status)
#tidy=TRUE argument, which tells DESeq2 to output the results table with␣
 ↪rownames as a first #column called 'row.

#let's see what this object looks like
dds
```

```
class: DESeqDataSet
dim: 38694 8
metadata(1): version
assays(1): counts
rownames(38694): ENSG00000000003 ENSG00000000005 … ENSG00000283120
  ENSG00000283123
rowData names(0):
colnames(8): SAMPLE1 SAMPLE2 … SAMPLE7 SAMPLE8
colData names(2): id status
```

Run DESEQ function

[ ]:
```
dds <- DESeq(dds)
#estimateSizeFactors
#This calculates the relative library depth of each sample

#estimateDispersions
#estimates the dispersion of counts for each gene

#nbinomWaldTest
#calculates the significance of coefficients in a Negative Binomial GLM using␣
 ↪the size and dispersion outputs
```

```
estimating size factors
```

```
estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing
```

Results Table

[ ]:
```
res <- results(dds, tidy=TRUE)
head(res)
```

A data.frame: 6 × 7

	row <chr>	baseMean <dbl>	log2FoldChange <dbl>	lfcSE <dbl>	stat <dbl>	pvalue <dbl>
1	ENSG00000000003	747.2088950	-0.35104497	0.1682782	-2.0860993	0.036969
2	ENSG00000000005	0.0000000	NA	NA	NA	NA
3	ENSG00000000419	520.1285176	0.20575958	0.1008784	2.0396802	0.041382
4	ENSG00000000457	322.6646922	0.02418624	0.1450333	0.1667634	0.867556
5	ENSG00000000460	87.6829843	-0.14751290	0.2568407	-0.5743361	0.565740
6	ENSG00000000938	0.3192055	-1.73263987	3.4936009	-0.4959467	0.619932

Summary of differential gene expression

[ ]:
```
summary(res)
```

```
    row              baseMean        log2FoldChange      lfcSE
Length:38694       Min.   :     0.0   Min.   :-6.031   Min.   :0.057
Class :character   1st Qu.:     0.0   1st Qu.:-0.425   1st Qu.:0.174
Mode  :character   Median :     1.1   Median :-0.009   Median :0.445
                   Mean   :   587.3   Mean   :-0.007   Mean   :1.136
                   3rd Qu.:   202.2   3rd Qu.: 0.306   3rd Qu.:1.847
                   Max.   :329278.9   Max.   :16.411   Max.   :3.534
                                      NA's   :13435    NA's   :13435
     stat              pvalue           padj
Min.   :-15.939   Min.   :0.000   Min.   :0.000
1st Qu.: -0.643   1st Qu.:0.167   1st Qu.:0.202
Median : -0.028   Median :0.533   Median :0.605
Mean   :  0.052   Mean   :0.495   Mean   :0.539
3rd Qu.:  0.593   3rd Qu.:0.800   3rd Qu.:0.864
Max.   : 35.672   Max.   :1.000   Max.   :1.000
NA's   :13435     NA's   :13577   NA's   :23548
```

Sort summary list by p-value

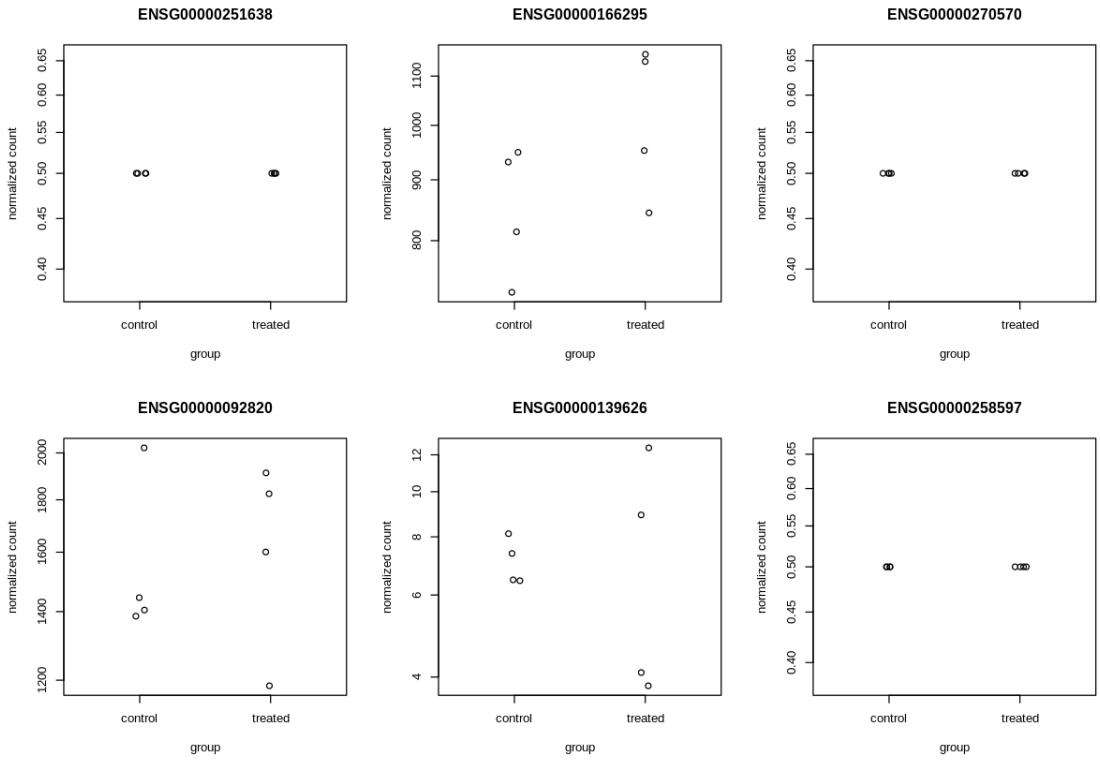
[ ]:
```
res <- res[order(res$pvalue),]
head(res)
```

A data.frame: 6 × 7

	row <chr>	baseMean <dbl>	log2FoldChange <dbl>	lfcSE <dbl>	stat <dbl>	pvalue <dbl>
3430	ENSG00000107611	110061.5	15.93021	0.4465760	35.67190	1.078397
6893	ENSG00000134463	111507.7	16.17015	0.4659220	34.70571	6.461656
35429	ENSG00000274540	110383.5	16.14559	0.4666590	34.59827	2.682747
4585	ENSG00000116580	111171.6	16.16846	0.4689464	34.47827	1.698301
24028	ENSG00000230667	110897.6	16.14944	0.4697168	34.38122	4.812873
2939	ENSG00000104626	111359.2	16.41089	0.5001353	32.81289	3.856213

Plot Counts

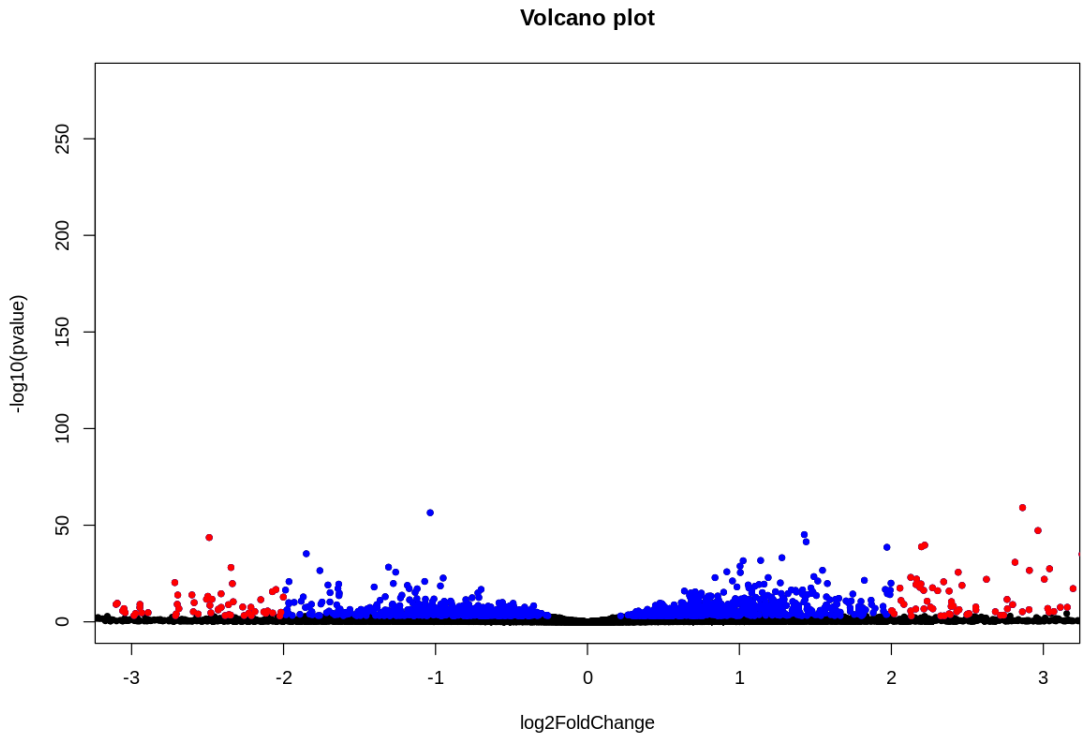
[ ]:
```
#we can use plotCounts fxn to compare the normalized counts
#between treated and control groups for our top 6 genes
par(mfrow=c(2,3))

plotCounts(dds, gene="ENSG00000251638", intgroup="status")
plotCounts(dds, gene="ENSG00000166295", intgroup="status")
plotCounts(dds, gene="ENSG00000270570", intgroup="status")
plotCounts(dds, gene="ENSG00000092820", intgroup="status")
plotCounts(dds, gene="ENSG00000139626", intgroup="status")
plotCounts(dds, gene="ENSG00000258597", intgroup="status")
#Next steps in exploring these data...BLAST to database to find associated gene
 ↪function
```



Volcano Plot

[ ]:
```
library(repr)
options(repr.plot.width=10, repr.plot.height=7)
```

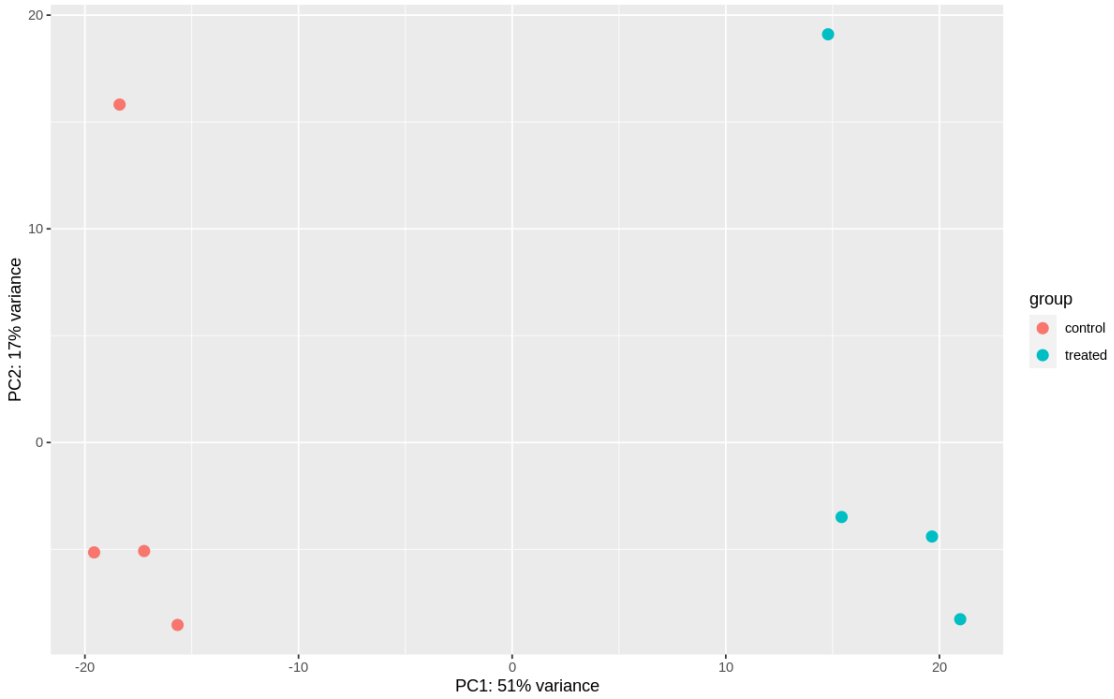
[ ]:
```
#reset par
par(mfrow=c(1,1))
# Make a basic volcano plot
with(res, plot(log2FoldChange, -log10(pvalue), pch=20, main="Volcano plot",␣
 ↪xlim=c(-3,3)))

# Add colored points: blue if padj<0.01, red if log2FC>1 and padj<0.05)
with(subset(res, padj<.01 ), points(log2FoldChange, -log10(pvalue), pch=20,␣
 ↪col="blue"))
with(subset(res, padj<.01 & abs(log2FoldChange)>2), points(log2FoldChange,␣
 ↪-log10(pvalue), pch=20, col="red"))
```



[ ]:
```
#First we need to transform the raw count data
#vst function will perform variance stabilizing transformation
```

```
vsdata <- vst(dds, blind=FALSE)
```

[ ]:
```
plotPCA(vsdata, intgroup="status")
#using the DESEQ2 plotPCA fxn we can look at how our samples group by treatment
```



[ ]:
```

```