Compare DE genes found by edgeR and DESeq2 in treatment versus control comparison using the same reference genome assembly

Ann Loraine

2023-09-27

Prelude - Define variables

```
output_fname_prefix = "results/CvT-compare-"
output_fname_suffix = ".txt"
assemblies = c("SL4","SL5")
methods = c("edgeR","DESeq2")
```

Introduction

This Markdown compares edgeR to DESeq2 differential expression results across assemblies SL4 and SL5.

SL4 corresponds to IGB Quickload S_lycopersicum_Sep_2019 and SL5 corresponds to IGB Quickload S_lycopersicum_Jun_2022.

Questions:

- How similar or different were the differential expression results found by edgeR and DESeq2 differential expression analyses?

To answer the question, output a data file suitable for an analyst to open, explore, and analyze using Excel.

Summarize the file structure, such as column headings and their meanings, in the Discussion section below.

Analysis and Results

Load custom functions for analyzing and visualizing differential expression:

```
source("Common.R")
```

Load required edgeR and DESeq2 generated data:

```
dfs=list()
for (method in methods) {
  for (assembly in assemblies) {
    thing = getCvS(method=method,assembly=assembly,
                   standardize=T)
    thing$comp = paste(thing$group1,thing$group2,
                      sep="-")
    key = paste(method,assembly,sep=".")
    row.names(thing)=paste(thing$comp,thing$gene,sep=".")
    dfs[[key]]=thing
  }
}
```

In the previous code chunk, we loaded 4 differential expression results sets into memory.

Define false discovery rate cutoff variable "Q" for deciding differential expression.

```
Q = 0.05
```

The above value of 0.05 represents our target fraction of false discoveries within a results data set.

Report the number of test results obtained per comparison:

```
for (key in names(dfs)) {
  thing = dfs[[key]]
  print("COMPARISON")
  print(key)
  print(table(thing$comp))
}
```

```
## [1] "COMPARISON"
## [1] "edgeR.SL4"
## 
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##            9369            9438            9386            9447            8461
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.15-V.34.15 V.28.30-V.34.30
##            8508            8552            8500            8393            8496
## V.28.45-V.34.45 V.28.75-V.34.75
##            8569            8594
## [1] "COMPARISON"
## [1] "edgeR.SL5"
## 
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##            9168            9246            9197            9238            8271
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.15-V.34.15 V.28.30-V.34.30
##            8317            8377            8301            8223            8287
## V.28.45-V.34.45 V.28.75-V.34.75
##            8388            8420
## [1] "COMPARISON"
## [1] "DESeq2.SL4"
## 
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##           17381           12673           11800           12875           16399
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.15-V.34.15 V.28.30-V.34.30
##           16283           11713           14104           16108           16256
## V.28.45-V.34.45 V.28.75-V.34.75
##           13172           11663
## [1] "COMPARISON"
## [1] "DESeq2.SL5"
## 
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##           11656           13308           13994           13481           15968
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.15-V.34.15 V.28.30-V.34.30
##           11025           15374           12013           15729           15841
## V.28.45-V.34.45 V.28.75-V.34.75
##           14667           10249
```

The preceding code chunk shows the number of tests performed by each method, and each genome assembly.

Now, show how many tests had adjusted p value of 0.05 or smaller.

DE genes found by method, assembly:

```
for (key in names(dfs)) {
  thing = dfs[[key]]
  print("COMPARISON")
  print(key)
  v = thing$padj<=Q
  print(table(thing$comp[v]))
}
```

```
## [1] "COMPARISON"
## [1] "edgeR.SL4"
##
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##              24              55              85             136               1
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.15-V.34.15 V.28.30-V.34.30
##              23              27              51               1               9
## V.28.45-V.34.45 V.28.75-V.34.75
##              34              50
## [1] "COMPARISON"
## [1] "edgeR.SL5"
##
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##              29              57              71             129               1
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.15-V.34.15 V.28.30-V.34.30
##              24              28              55               2               6
## V.28.45-V.34.45 V.28.75-V.34.75
##              34              54
## [1] "COMPARISON"
## [1] "DESeq2.SL4"
##
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##              19              49              80             137               1
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.30-V.34.30 V.28.45-V.34.45
##              18              25              48               5              27
## V.28.75-V.34.75
##              43
## [1] "COMPARISON"
## [1] "DESeq2.SL5"
##
## A.28.15-A.34.15 A.28.30-A.34.30 A.28.45-A.34.45 A.28.75-A.34.75 F.28.15-F.34.15
##              23              49              63             134               1
## F.28.30-F.34.30 F.28.45-F.34.45 F.28.75-F.34.75 V.28.30-V.34.30 V.28.45-V.34.45
##              14              22              47               5              26
## V.28.75-V.34.75
##              41
```

As shown in the above output, the edgeR method called more genes as differentially expressed than the

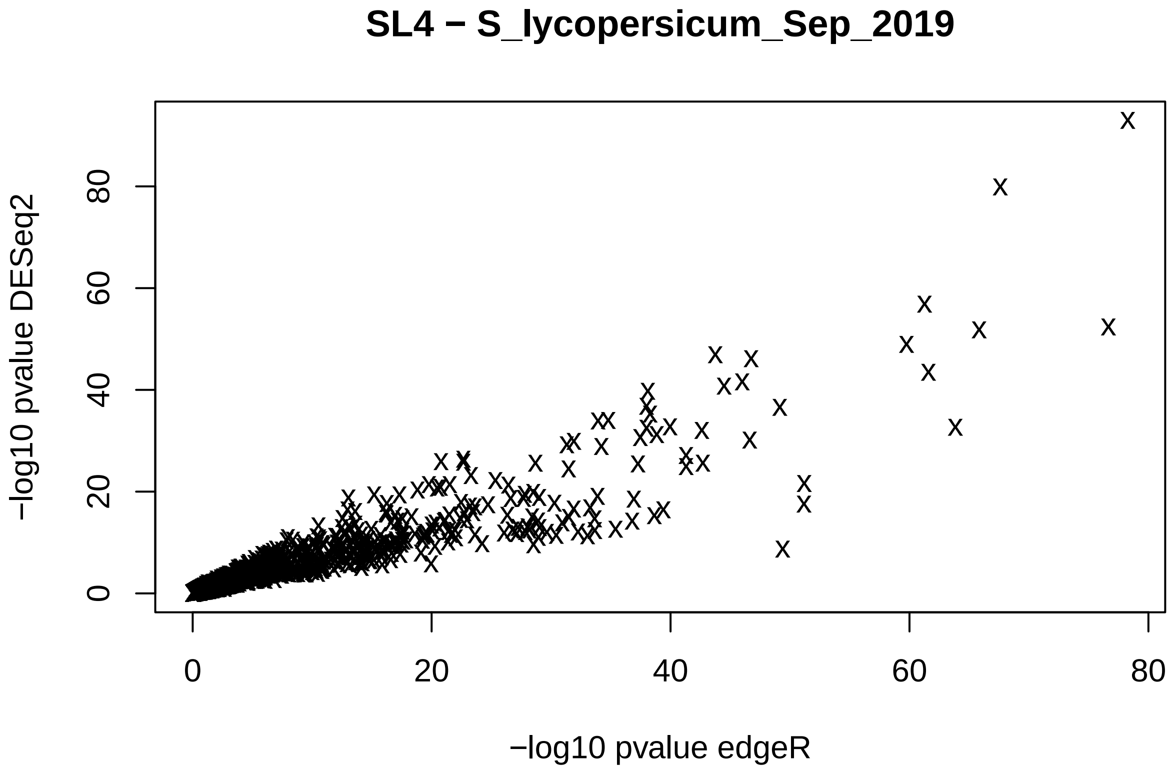
DESeq2 method did for most of the treatment duration time points.

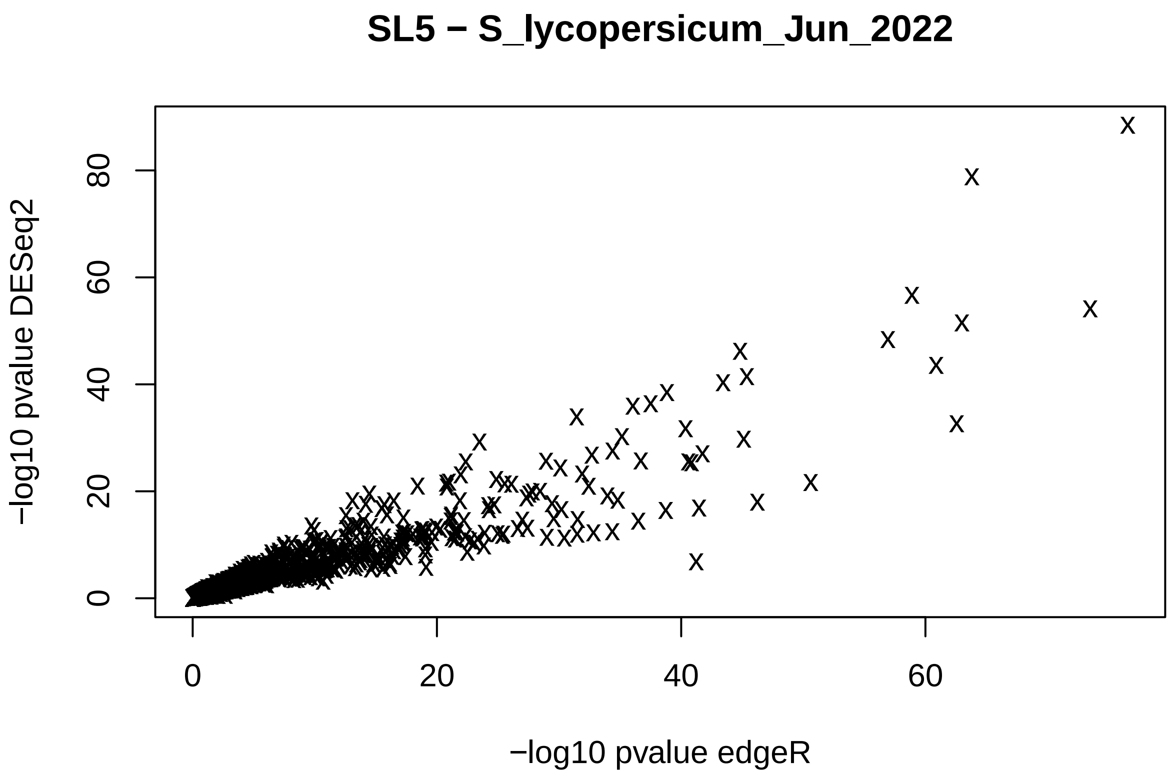
Both methods computed log2 fold-changes and nominal p values for the comparisons. How similar were these values among the genes found to be differentially expressed?

For genes present in edgeR and DESeq2 results sets, show the correlation (if any) between log2 fold-change and nominal p values produced by the two methods.

Compare nominal p values:

```r
assembly_synonyms = c("S_lycopersicum_Sep_2019",
                      "S_lycopersicum_Jun_2022")
names(assembly_synonyms)=assemblies
for (assembly in assemblies) {
  key1 = paste(methods[1],assembly,sep=".")
  key2 = paste(methods[2],assembly,sep=".")
  thing1 = dfs[[key1]]
  thing2 = dfs[[key2]]
  v = intersect(row.names(thing1),row.names(thing2))
  plot(-log10(thing1[v,"pvalue"]),
       -log10(thing2[v,"pvalue"]),
       xlab=paste("-log10 pvalue",methods[1]),
       ylab=paste("-log10 pvalue",methods[2]),
       main=paste(assembly,assembly_synonyms[assembly],sep=" - "),
       pch="x")
}
```

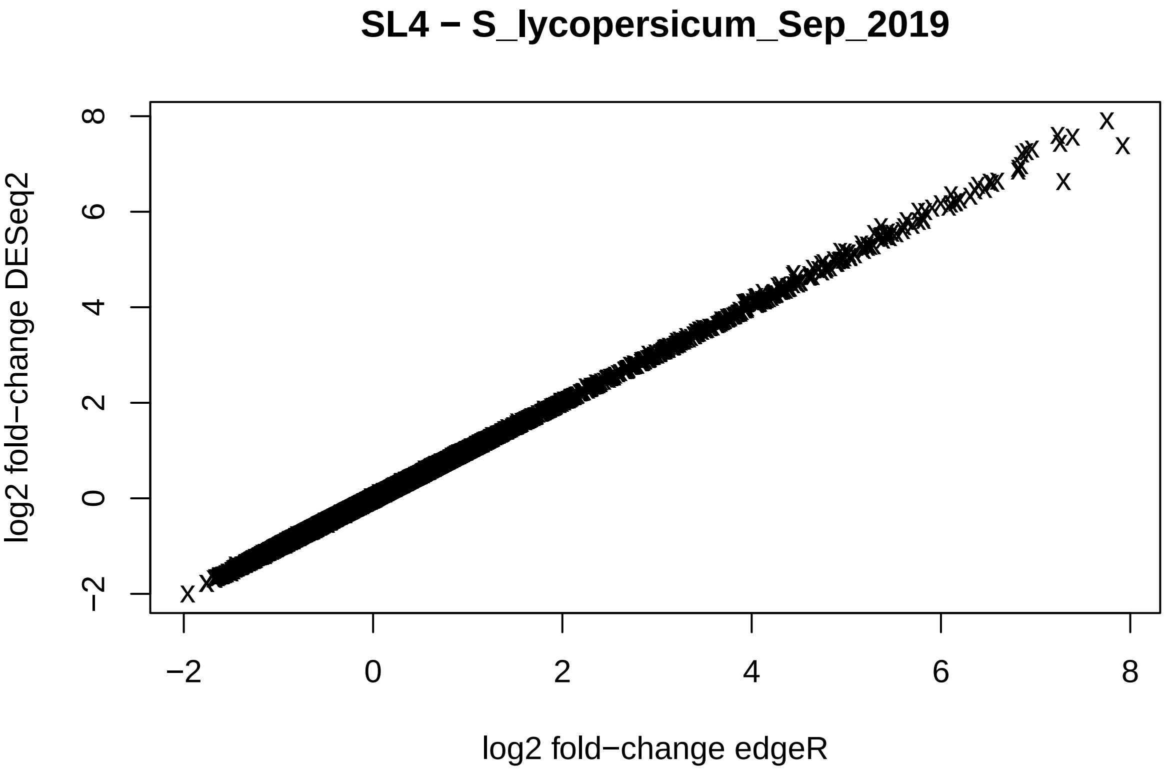


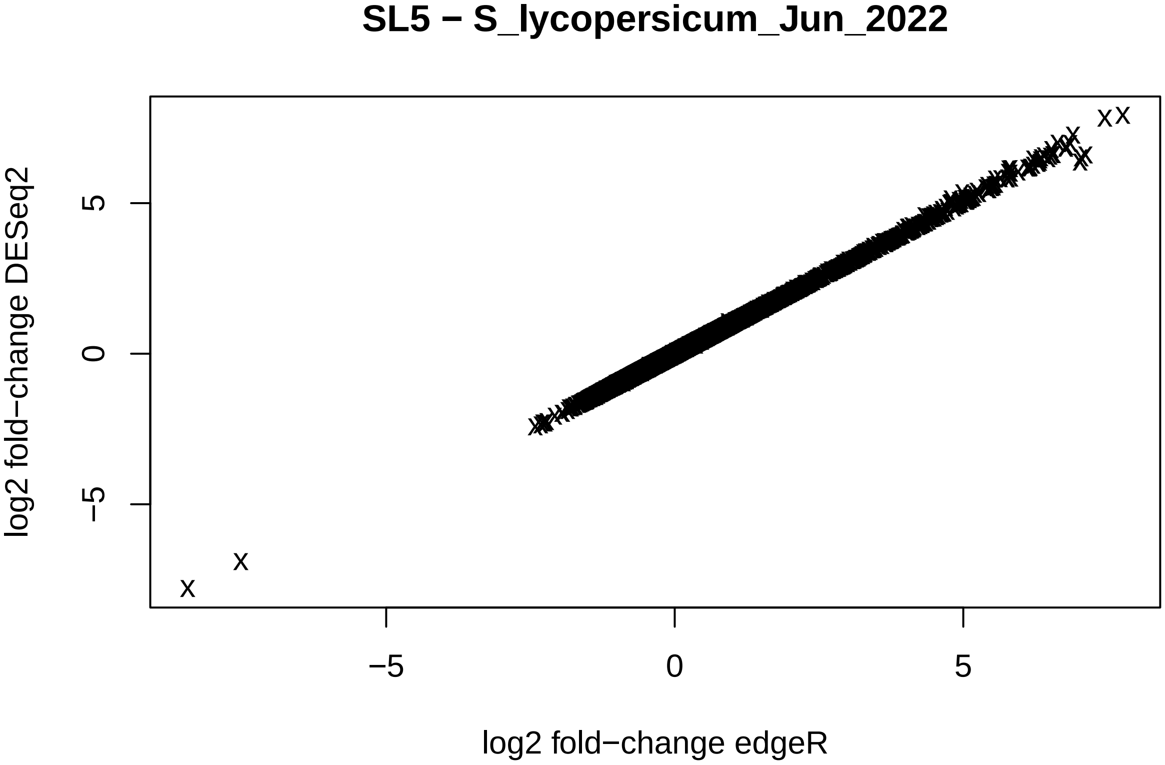


The above plot shows a pretty good correspondence between p values calculated by the two methods.

Examine correspondence of fold-changes:

```
for (assembly in assemblies) {
  key1 = paste(methods[1],assembly,sep=".")
  key2 = paste(methods[2],assembly,sep=".")
  thing1 = dfs[[key1]]
  thing2 = dfs[[key2]]
  v = intersect(row.names(thing1),row.names(thing2))
  plot(thing1[v,"logFC"],
       thing2[v,"logFC"],
       xlab=paste("log2 fold-change",methods[1]),
       ylab=paste("log2 fold-change",methods[2]),
       main=paste(assembly,assembly_synonyms[assembly],sep=" - "),
       pch="x")
}
```


SL4 − S_lycopersicum_Sep_2019


SL5 − S_lycopersicum_Jun_2022

The above plot shows very tight correspondence between log2 fold-changes between the two methods.

Having a combined data frame / spreadsheet of the DE results obtained from both methods might help analysts examine individual genes. Let's create a data frame for each assembly that combines the results and includes all the DE genes from each method from the same genome assembly.

```
for (assembly in assemblies) {
  key1 = paste(methods[1],assembly,sep=".")
  key2 = paste(methods[2],assembly,sep=".")
  thing1 = dfs[[key1]]
  thing1 = thing1[thing1$padj<=Q,]
  thing2 = dfs[[key2]]
  thing2 = thing2[thing2$padj<=Q,]
  thing3 = merge(thing1,thing2,
                by="row.names",
                all.x=T,
                all.y=T,
                suffixes=c(paste0(".",methods[1]),
                           paste0(".",methods[2])))
  if (assembly=="SL4") {
    to_write = thing3[,c(2,3,4,5,6,7,14,15,16,19)]
    names(to_write)[c(1:3,10)]=c("gene","group1","group2","description")
    all.sl4 = to_write
  }
  if (assembly=="SL5") {
    to_write = thing3[,c(2,3,4,5,6,7,15,16,17,20,21)]
    names(to_write)[c(1:3,10,11)]=c("gene","group1","group2","description","SL4")
    all.sl5 = to_write
  }
  output_fname=paste0(output_fname_prefix,assembly,output_fname_suffix)
  write.table(to_write,file=output_fname,sep="\t",quote=F,
            row.names = F)
}
```

The preceding code chunk identified all comparisons that produced adjusted p value less than or equal to 0.05, found by one or both methods, for the two genome assemblies. Data for the two genome assemblies were considered separately.

SL4 assembly results:

- edgeR produced 496 comparisons with Q <= 0.05.
- DeSeq2 produced 452 comparisons with Q <= 0.05.
- There were 410 comparisons with Q <= 0.05 found by each method.

SL5 assembly results:

- edgeR produced 490 comparisons with Q <= 0.05.
- DESeq2 produced 425 comparisons with Q <= 0.05.
- There were 387 comparisons with Q <= 0.05 found by each method.

Discussion

Both methods identified different but overlapping sets of comparisons with adjusted pvalues indicating differential expression.

The fold-changes calculated by the two methods were very close, whereas the nominal p values were more different but still very similar.

Two data files with comparisons made from both methods were written out, one for each assembly. Each data file name contained prefix results/CvT-compare- and a different suffix indicating the genome assembly.

Columns included:

- gene - the gene measured
- group1 - the group tested, the control
- group 2 - the treatment group testing, the treatment
- padj, pvalue, logFC - adjusted p value, nominal pvalue, log2 fold-change for each method, as indicated in the column name (e.g., padj.edgeR or padj.DESeq2)
- description - gene description from SL4, if available; NA if not
- SL4 - only in the SL5 spreadsheet; the SL4 gene name if available

---

Conclusions

- edgeR and DESeq2 produced similar results
- Both methods produced the same trend with longer treatment times producing more differences in expression between treatment and control

---

Session Info

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16)
## Platform: x86_64-apple-darwin20 (64-bit)
## Running under: macOS Big Sur 11.7.9
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] EnhancedVolcano_1.18.0      ggrepel_0.9.3
##  [3] ggplot2_3.4.3               DESeq2_1.40.2
##  [5] SummarizedExperiment_1.30.2 Biobase_2.60.0
##  [7] MatrixGenerics_1.12.3       matrixStats_1.0.0
##  [9] GenomicRanges_1.52.0        GenomeInfoDb_1.36.3
## [11] IRanges_2.34.1              S4Vectors_0.38.1
## [13] BiocGenerics_0.46.0         edgeR_3.42.4
## [15] limma_3.56.2                readxl_1.4.3
## [17] readr_2.1.4                 stringr_1.5.0
##
## loaded via a namespace (and not attached):
```

```
##  [1] gtable_0.3.4            xfun_0.40               lattice_0.21-8
##  [4] tzdb_0.4.0              vctrs_0.6.3             tools_4.3.1
##  [7] bitops_1.0-7            generics_0.1.3          parallel_4.3.1
## [10] tibble_3.2.1            fansi_1.0.4             pkgconfig_2.0.3
## [13] Matrix_1.6-1            lifecycle_1.0.3         GenomeInfoDbData_1.2.10
## [16] compiler_4.3.1          munsell_0.5.0           codetools_0.2-19
## [19] htmltools_0.5.6         RCurl_1.98-1.12         yaml_2.3.7
## [22] pillar_1.9.0            crayon_1.5.2            BiocParallel_1.34.2
## [25] DelayedArray_0.26.7     abind_1.4-5             tidyselect_1.2.0
## [28] locfit_1.5-9.8          digest_0.6.33           stringi_1.7.12
## [31] dplyr_1.1.3             fastmap_1.1.1           grid_4.3.1
## [34] colorspace_2.1-0        cli_3.6.1               magrittr_2.0.3
## [37] S4Arrays_1.0.6          utf8_1.2.3              withr_2.5.0
## [40] scales_1.2.1            rmarkdown_2.24          XVector_0.40.0
## [43] cellranger_1.1.0        hms_1.1.3               evaluate_0.21
## [46] knitr_1.44              rlang_1.1.1             Rcpp_1.0.11
## [49] glue_1.6.2              rstudioapi_0.15.0       R6_2.5.1
## [52] zlibbioc_1.46.0
```