Wk2_codingexercise_answers

1/26/2022

CODING EXERCISES

- Today we'll follow the LULU tutorial, but use our own data
  - "wK2_refseqs.fasta" instead of "centroids_test.txt"
  - "wk2_seqtab.csv" instead of "otutable_test.txt"

Use local BLAST to make the database

- Get "wk2_refseqs.fasta" file from GitHub using the download.file command and save it to your working directory
- Make sure you have local BLAST installed (https://www.ncbi.nlm.nih.gov/books/NBK1762/)
- Paste the following code into a terminal window to make your BLAST database

```
C:\...>makeblastdb -in wK2_refseqs.fasta -parse_seqids -dbtype nucl
```

- Paste the following code into a terminal window to make your BLAST the seqs against themselves

```
C:\...>blastn -db wK2_refseqs.fasta -outfmt "6 qseqid sseqid pident" -out match_list.txt
-qcov_hsp_perc 80 -perc_identity 84 -query wK2_refseqs.fasta
```

Load the files you need into R

```
ASV <- read.csv(file="wk2_seqtab.csv",
                stringsAsFactors=FALSE,
                header=TRUE,
                row.names=1)

matchlist <- data.table::fread(file="match_list.txt",
                   stringsAsFactors = FALSE,
                   header=FALSE)

#check that there are 3 cols & first two ASV cols are characters to match ASVtable designations
str(matchlist)
```

```
## Classes 'data.table' and 'data.frame':   190 obs. of  3 variables:
##  $ V1: chr  "ASV1" "ASV1" "ASV1" "ASV1" ...
##  $ V2: chr  "ASV1" "ASV6" "ASV2" "ASV4" ...
##  $ V3: num  100 96.9 96.5 90.2 100 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Use LULU to curate the results

```
library("lulu")

#run lulu to curate the ASV table
curated_result <- lulu(ASV, matchlist)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## [1] "progress: 3%"
## [1] "progress: 7%"
## [1] "progress: 10%"
## [1] "progress: 13%"
## [1] "progress: 17%"
## [1] "progress: 20%"
## [1] "progress: 23%"
## [1] "progress: 27%"
## [1] "progress: 30%"
## [1] "progress: 33%"
## [1] "progress: 37%"
## [1] "progress: 40%"
## [1] "progress: 43%"
## [1] "progress: 47%"
## [1] "progress: 50%"
## [1] "progress: 53%"
## [1] "progress: 57%"
## [1] "progress: 60%"
## [1] "progress: 63%"
## [1] "progress: 67%"
## [1] "progress: 70%"
## [1] "progress: 73%"
## [1] "progress: 77%"
## [1] "progress: 80%"
## [1] "progress: 83%"
## [1] "progress: 87%"
## [1] "progress: 90%"
## [1] "progress: 93%"
## [1] "progress: 97%"
## [1] "progress: 100%"
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
#save lulu results
ASV_lulu <- curated_result$curated_table
write.csv(ASV_lulu, "MetabAbund_ASV_lulu.csv")

ASV_lulu_discards <- curated_result$discarded_otus
write.csv(ASV_lulu_discards, "MetabAbund_ASV_lulu_discards.csv")

ASV_lulu_retained <- curated_result$curated_otus
write.csv(ASV_lulu_retained, "MetabAbund_ASV_lulu_retained.csv")

curated_result$curated_count
```

```
## [1] 29
```

```
#retained 29/30 ASVs

curated_result$discarded_count
```

```
## [1] 1
```

```
#discarded 1/30 ASVs
```

Session info

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
```

```
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_1.0.7 lulu_0.1.0
##
## loaded via a namespace (and not attached):
##  [1] knitr_1.36        magrittr_2.0.1    tidyselect_1.1.1  R6_2.5.1
##  [5] rlang_0.4.11      fastmap_1.1.0     fansi_0.5.0       stringr_1.4.0
##  [9] tools_4.1.1       data.table_1.14.2 xfun_0.26         utf8_1.2.2
## [13] DBI_1.1.1         htmltools_0.5.2   ellipsis_0.3.2    assertthat_0.2.1
## [17] yaml_2.2.1        digest_0.6.28     tibble_3.1.5      lifecycle_1.0.1
## [21] crayon_1.4.1      purrr_0.3.4       vctrs_0.3.8       glue_1.4.2
## [25] evaluate_0.14     rmarkdown_2.11    stringi_1.7.5     compiler_4.1.1
## [29] pillar_1.6.4      generics_0.1.1    pkgconfig_2.0.3
```