Introduction to R

Practice exercises

Kim Dill-McFarland, kadm@uw.edu

version October 28, 2021

Contents

Setup

Open the Intro R Rproject and start a new working script. Install any new packages, load packages, set a seed, and load data as we did in the workshop.

```
#RNAseq expression and metadata
load("data/RSTR_data_clean_subset.RData")
```

Exercises

Day 1: Base R

Project setup

1. What are the benefits of storing data in `RData` versus tables (`csv`, `tsv`, etc)?
2. Imagine a hypothetical project with the following data and results. How would you choose to setup your Rproject directory and sub-directories? This is something that may evolve over time, but it is helpful to start with a defined structure to make it easier for you and others to find things.
   - `.RData` file containing all cleaned data for the project
   - 2 `.csv` of raw RNAseq counts and sample metadata (what was cleaned to make the `.RData`)
   - 4 `.csv` with linear model results
   - 25 `.png` plots of gene expression, individual genes
   - 1 `.png` plot of gene expression, faceted with many genes
   - 2 `.R` scripts, 1 for linear modeling and 1 for making plots
   - 1 `.Rmd` report summarizing and interpreting the results

Data types

1. What is the difference between a `character` and `facter`?
2. What data type does R classify the date `2021.06`? What about `2021/06`? If it is not classified as a "date", how could this impact downstream analyses? Try to predict the outcomes before checking in R.
   - Challenge: Checkout the package `lubridate` for functions to effectively work with dates in R.

3. You have an `S3` list object named `myData`and it contains 2 data frames named `A` and `B`. Within `B` there is a column named `variable1`. How do you access this variable?

Subsetting and filtering

Using `dat`:

1. What is the mean library size `lib.size`?
2. Try running `summary(dat$targets)`. What kinds of data does it provide? Why are the results different for different variables?
3. How many libraries have a library size `lib.size` greater than 5 million and a normalization factor `norm.factors` less than 1?
4. Challenge: Using the function `grepl`, how many libraries are from a donor with an `RSID` that starts with "RS1025"?

R session

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.1.1  magrittr_2.0.1  fastmap_1.1.0   tools_4.1.1
##  [5] htmltools_0.5.2 yaml_2.2.1      stringi_1.7.5   rmarkdown_2.11
##  [9] knitr_1.36      stringr_1.4.0   xfun_0.27       digest_0.6.28
## [13] rlang_0.4.12    evaluate_0.14
```