Gene regulatory network inference

Fabricio Almeida-Silva

2022-08-09

Contents

```
library(here)
library(BioNERO)
library(igraph)
library(ggstatsplot)
library(ggpubr)
set.seed(123)
dup_palette <- c("#1984c5", "#ffb400")
```

1 Overview

Here, we will describe the code to infer gene regulatory networks (GRN) for each species from the RNA-seq data obtained from EBI's Expression Atlas.

2 GRN inference

Here, we will define a function to process the input for GRN inference with *BioNERO* and infer a GRN for each species manually (not in a loop) to avoid errors in one data set leading to loss of all previously inferred GRNs.

```
# Load expression data
load(here("data", "expression_data.rda"))

# Load TFs
load(here("data", "tfs.rda"))

# Define function to infer GRNs
get_GRN <- function(species) {

    # Get expression matrix for species x
    exp <- expression_data[[species]]
    exp <- BioNERO::exp_preprocess(exp, Zk_filtering = FALSE)
    exp <- as.data.frame(exp)

    # Get vector of TFs for species x
    vtfs <- tfs[[species]]$Gene
    vtfs <- vtfs[vtfs %in% rownames(exp)]  # keep only TFs in expression matrix
    vtfs <- unique(vtfs)

    if (length(vtfs) > 1) {
        message(species, ": ", length(vtfs), " TFs")
    } else {
        stop("Found less than 2 TFs.")
    }

    # Infer GRN
    grn <- BioNERO::grn_infer(exp, method = "genie3", regulators = vtfs,
        nTrees = 1000)
    return(grn)
```

```
}

# Infer GRNs
grn_osa <- get_GRN("Osativa")
grn_zma <- get_GRN("Zmays")
grn_gma <- get_GRN("Gmax")
grn_sly <- get_GRN("Slycopersicum")
grn_ptr <- get_GRN("Ptrichocarpa")
grn_ath <- get_GRN("Athaliana")
grn_vvi <- get_GRN("Vvinifera")
```

We now have our GRNs as fully connected graphs. Let's filter them by picking only the top *N* edges (based on edge rankings). To decide *N* for each species, we will break each GRN into 20 increasingly larger graphs and assess the scale-free topology fit for each graph. Then, we will pick the maximum value of *N* that makes the graph fit a scale-free topology ($R^2 = 0.75$).

```
# Filter network based on scale-free topology fit Osativa
png(filename = here("products", "plots", "grn_filtering_osativa.png"),
    width = 8, height = 6, units = "in", res = 300)
grn_filtered_osa <- grn_filter(grn_osa, nsplit = 20)
dev.off()

grn_filtered_osa <- grn_osa[1:1799884, ]


## Zmays
png(filename = here("products", "plots", "grn_filtering_zmays.png"),
    width = 8, height = 6, units = "in", res = 300)
grn_filtered_zma <- grn_filter(grn_zma, nsplit = 20)
dev.off()

grn_filtered_zma <- grn_zma[1:2045899, ]

## Vvinifera
png(filename = here("products", "plots", "grn_filtering_vvinifera.png"),
    width = 8, height = 6, units = "in", res = 300)
grn_filtered_vvi <- grn_filter(grn_vvi, nsplit = 20)
dev.off()

grn_filtered_vvi <- grn_vvi[1:1214095, ]

## Gmax
png(filename = here("products", "plots", "grn_filtering_gmax.png"),
    width = 8, height = 6, units = "in", res = 300)
grn_filtered_gma <- grn_filter(grn_gma, nsplit = 20)
dev.off()

grn_filtered_gma <- grn_gma[1:2268359, ]

## Slycopersicum
```

```
png(filename = here("products", "plots", "grn_filtering_slycopersicum.png"),
    width = 8, height = 6, units = "in", res = 300)
grn_filtered_sly <- grn_filter(grn_sly, nsplit = 20)
dev.off()

grn_filtered_sly <- grn_sly[1:1344344, ]


## Ptrichocarpa
png(filename = here("products", "plots", "grn_filtering_ptrichocarpa.png"),
    width = 8, height = 6, units = "in", res = 300)
grn_filtered_ptr <- grn_filter(grn_ptr, nsplit = 20)
dev.off()

grn_filtered_ptr <- grn_ptr[1:1275872, ]


## Athaliana
png(filename = here("products", "plots", "grn_filtering_athaliana.png"),
    width = 8, height = 6, units = "in", res = 300)
grn_filtered_ath <- grn_filter(grn_ath, nsplit = 20)
dev.off()

grn_filtered_ath <- grn_ath[1:2503623, ]

# Save filtered GRNs to an .rda file
grns <- list(Osativa = grn_filtered_osa, Zmays = grn_filtered_zma,
    Vvinifera = grn_filtered_vvi, Gmax = grn_filtered_gma, Slycopersicum = grn_filtered_sly,
    Ptrichocarpa = grn_filtered_ptr, Athaliana = grn_filtered_ath)
grns <- lapply(grns, function(x) return(x[, 1:2]))  # remove 'Weight' column

save(grns, file = here("products", "result_files", "grns.rda"),
    compress = "xz")
```

To summarize, the numbers of top N edges for each species were:

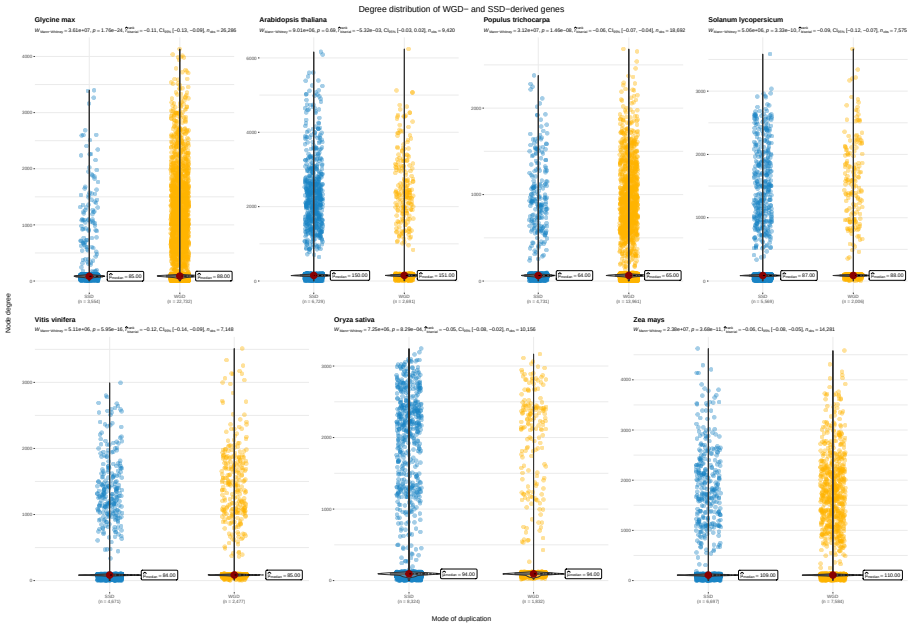
Species	Edge number (Millions)
Osativa	1.799884
Zmays	2.045899
Vvinifera	1.214095
Gmax	2.268359
Slycopersicum	1.344344
Ptrichocarpa	1.275872
Athaliana	2.503623

3 Exploring degree distributions

Finally, let's compare the degree distributions of WGD- and SSD-derived gene pairs.

```
load(here("data", "duplicated_genes.rda"))
load(here("products", "result_files", "grns.rda"))

# Create data frame of degree distribution for each species
# by duplication mode
degree_distros <- Reduce(rbind, lapply(seq_along(grns), function(x) {
    species <- names(grns)[x]
    net <- grns[[x]]
    dups <- duplicated_genes[[species]]

    # Get degree distribution as a data frame
    degree <- graph_from_data_frame(net[, 1:2]) %>%
        degree()
    degree_df <- data.frame(gene = names(degree), degree = degree) %>%
        dplyr::inner_join(., dups, by = "gene") %>%
        dplyr::select(gene, degree, type, peak) %>%
        dplyr::mutate(species = species)

    return(degree_df)
}))

# Visualize degree distributions as a violin plot
plot_violin <- function(data) {
    p <- ggbetweenstats(data = data, x = type, y = degree, type = "nonparametric",
        pairwise.display = "all", p.adjust.method = "BH") + ggplot2::scale_color_manual(values = dup_palette)
        labs(x = "", y = "") + theme(plot.subtitle = element_text(size = 8.5))
    return(p)
}
sdegree_distros <- split(degree_distros, degree_distros$species)

plot_degree_gma <- plot_violin(sdegree_distros$Gmax) + labs(title = "Glycine max")

plot_degree_ath <- plot_violin(sdegree_distros$Athaliana) + labs(title = "Arabidopsis thaliana")

plot_degree_ptr <- plot_violin(sdegree_distros$Ptrichocarpa) +
    labs(title = "Populus trichocarpa")

plot_degree_sly <- plot_violin(sdegree_distros$Slycopersicum) +
    labs(title = "Solanum lycopersicum")

plot_degree_vvi <- plot_violin(sdegree_distros$Vvinifera) + labs(title = "Vitis vinifera")

plot_degree_osa <- plot_violin(sdegree_distros$Osativa) + labs(title = "Oryza sativa")

plot_degree_zma <- plot_violin(sdegree_distros$Zmays) + labs(title = "Zea mays")

# Combine plots
p_deg_upper <- ggarrange(plot_degree_gma, plot_degree_ath, plot_degree_ptr,
    plot_degree_sly, nrow = 1)
p_deg_lower <- ggarrange(plot_degree_vvi, plot_degree_osa, plot_degree_zma,
    nrow = 1)
```

```
p_deg_final <- ggarrange(p_deg_upper, p_deg_lower, nrow = 2)
p_deg_final <- annotate_figure(p_deg_final, top = text_grob("Degree distribution of WGD- and SSD-derived gene
    size = 15), bottom = text_grob("Mode of duplication"), left = text_grob("Node degree",
    rot = 90))
p_deg_final
```



As in the PPI network, although the P-values for some comparisons were significant, the effect size is negligible, indicating that degree does not influence motif frequencies. The small P-value is likely an artifact generated by the large sample size of the comparisons.

```
# Saving figure and degree distros object Degree distros
save(degree_distros, file = here("products", "result_files",
    "degree_distros_grn.rda"), compress = "xz")

## Plot
ggsave(p_deg_final, file = here("products", "plots", "grn_degree_distros.png"),
    width = 22, height = 15, dpi = 300)
```

Session info

This document was created under the following conditions:

```
sessioninfo::session_info()
## - Session info -----------------------------------------------------------------
##  setting  value
##  version  R version 4.2.1 (2022-06-23)
##  os       Ubuntu 20.04.4 LTS
##  system   x86_64, linux-gnu
##  ui       X11
##  language (EN)
```

```
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       Europe/Brussels
##  date     2022-08-09
##  pandoc   2.18 @ /usr/lib/rstudio/bin/quarto/bin/tools/ (via rmarkdown)
##
## - Packages ---------------------------------------------------------------
##  package              * version    date (UTC) lib source
##  abind                  1.4-5      2016-07-21 [1] CRAN (R 4.2.0)
##  annotate               1.74.0     2022-04-26 [1] Bioconductor
##  AnnotationDbi          1.58.0     2022-04-26 [1] Bioconductor
##  assertthat             0.2.1      2019-03-21 [1] CRAN (R 4.2.0)
##  backports              1.4.1      2021-12-13 [1] CRAN (R 4.2.0)
##  base64enc              0.1-3      2015-07-28 [1] CRAN (R 4.2.0)
##  bayestestR             0.12.1     2022-05-02 [1] CRAN (R 4.2.0)
##  Biobase                2.56.0     2022-04-26 [1] Bioconductor
##  BiocGenerics           0.42.0     2022-04-26 [1] Bioconductor
##  BiocManager            1.30.18    2022-05-18 [1] CRAN (R 4.2.0)
##  BiocParallel           1.30.3     2022-06-05 [1] Bioconductor
##  BiocStyle            * 2.25.0     2022-06-15 [1] Github (Bioconductor/BiocStyle@7150c28)
##  BioNERO              * 1.4.0      2022-04-26 [1] Bioconductor
##  Biostrings             2.64.0     2022-04-26 [1] Bioconductor
##  bit                    4.0.4      2020-08-04 [1] CRAN (R 4.2.0)
##  bit64                  4.0.5      2020-08-30 [1] CRAN (R 4.2.0)
##  bitops                 1.0-7      2021-04-24 [1] CRAN (R 4.2.0)
##  blob                   1.2.3      2022-04-10 [1] CRAN (R 4.2.0)
##  bookdown               0.27       2022-06-14 [1] CRAN (R 4.2.0)
##  boot                   1.3-28     2021-05-03 [1] CRAN (R 4.2.0)
##  broom                  0.8.0      2022-04-13 [1] CRAN (R 4.2.0)
##  cachem                 1.0.6      2021-08-19 [1] CRAN (R 4.2.0)
##  car                    3.1-0      2022-06-15 [1] CRAN (R 4.2.0)
##  carData                3.0-5      2022-01-06 [1] CRAN (R 4.2.0)
##  checkmate              2.1.0      2022-04-21 [1] CRAN (R 4.2.0)
##  circlize               0.4.15     2022-05-10 [1] CRAN (R 4.2.0)
##  cli                    3.3.0      2022-04-25 [1] CRAN (R 4.2.0)
##  clue                   0.3-61     2022-05-30 [1] CRAN (R 4.2.0)
##  cluster                2.1.3      2022-03-28 [1] CRAN (R 4.2.0)
##  coda                   0.19-4     2020-09-30 [1] CRAN (R 4.2.0)
##  codetools              0.2-18     2020-11-04 [1] CRAN (R 4.2.0)
##  colorspace             2.0-3      2022-02-21 [1] CRAN (R 4.2.0)
##  ComplexHeatmap         2.12.0     2022-04-26 [1] Bioconductor
##  correlation            0.8.1      2022-05-20 [1] CRAN (R 4.2.0)
##  cowplot                1.1.1      2020-12-30 [1] CRAN (R 4.2.0)
##  crayon                 1.5.1      2022-03-26 [1] CRAN (R 4.2.0)
##  data.table             1.14.2     2021-09-27 [1] CRAN (R 4.2.0)
##  datawizard             0.4.1      2022-05-16 [1] CRAN (R 4.2.0)
##  DBI                    1.1.3      2022-06-18 [1] CRAN (R 4.2.0)
##  DelayedArray           0.22.0     2022-04-26 [1] Bioconductor
##  DESeq2                 1.36.0     2022-04-26 [1] Bioconductor
##  digest                 0.6.29     2021-12-01 [1] CRAN (R 4.2.0)
##  doParallel             1.0.17     2022-02-07 [1] CRAN (R 4.2.0)
```

```
##  dplyr                 1.0.9     2022-04-28 [1] CRAN (R 4.2.0)
##  dynamicTreeCut        1.63-1    2016-03-11 [1] CRAN (R 4.2.0)
##  edgeR                 3.38.1    2022-05-15 [1] Bioconductor
##  effectsize            0.7.0     2022-05-26 [1] CRAN (R 4.2.0)
##  ellipsis              0.3.2     2021-04-29 [1] CRAN (R 4.2.0)
##  evaluate              0.15      2022-02-18 [1] CRAN (R 4.2.0)
##  fansi                 1.0.3     2022-03-24 [1] CRAN (R 4.2.0)
##  farver                2.1.0     2021-02-28 [1] CRAN (R 4.2.0)
##  fastcluster           1.2.3     2021-05-24 [1] CRAN (R 4.2.0)
##  fastmap               1.1.0     2021-01-25 [1] CRAN (R 4.2.0)
##  foreach               1.5.2     2022-02-02 [1] CRAN (R 4.2.0)
##  foreign               0.8-82    2022-01-13 [1] CRAN (R 4.2.0)
##  formatR               1.12      2022-03-31 [1] CRAN (R 4.2.0)
##  Formula               1.2-4     2020-10-16 [1] CRAN (R 4.2.0)
##  genefilter            1.78.0    2022-04-26 [1] Bioconductor
##  geneplotter           1.74.0    2022-04-26 [1] Bioconductor
##  generics              0.1.2     2022-01-31 [1] CRAN (R 4.2.0)
##  GENIE3                1.18.0    2022-04-26 [1] Bioconductor
##  GenomeInfoDb          1.32.2    2022-05-15 [1] Bioconductor
##  GenomeInfoDbData      1.2.8     2022-05-06 [1] Bioconductor
##  GenomicRanges         1.48.0    2022-04-26 [1] Bioconductor
##  GetoptLong            1.0.5     2020-12-15 [1] CRAN (R 4.2.0)
##  ggnetwork             0.5.10    2021-07-06 [1] CRAN (R 4.2.0)
##  ggnewscale            0.4.7     2022-03-25 [1] CRAN (R 4.2.0)
##  ggplot2             * 3.3.6     2022-05-03 [1] CRAN (R 4.2.0)
##  ggpubr              * 0.4.0     2020-06-27 [1] CRAN (R 4.2.0)
##  ggrepel               0.9.1     2021-01-15 [1] CRAN (R 4.2.0)
##  ggsignif              0.6.3     2021-09-09 [1] CRAN (R 4.2.0)
##  ggstatsplot         * 0.9.3     2022-05-27 [1] CRAN (R 4.2.0)
##  GlobalOptions         0.1.2     2020-06-10 [1] CRAN (R 4.2.0)
##  glue                  1.6.2     2022-02-24 [1] CRAN (R 4.2.0)
##  GO.db                 3.15.0    2022-05-06 [1] Bioconductor
##  gridExtra             2.3       2017-09-09 [1] CRAN (R 4.2.0)
##  gtable                0.3.0     2019-03-25 [1] CRAN (R 4.2.0)
##  here                * 1.0.1     2020-12-13 [1] CRAN (R 4.2.0)
##  Hmisc                 4.7-0     2022-04-19 [1] CRAN (R 4.2.0)
##  htmlTable             2.4.0     2022-01-04 [1] CRAN (R 4.2.0)
##  htmltools             0.5.2     2021-08-25 [1] CRAN (R 4.2.0)
##  htmlwidgets           1.5.4     2021-09-08 [1] CRAN (R 4.2.0)
##  httr                  1.4.3     2022-05-04 [1] CRAN (R 4.2.0)
##  igraph              * 1.3.2     2022-06-13 [1] CRAN (R 4.2.0)
##  impute                1.70.0    2022-04-26 [1] Bioconductor
##  insight               0.17.1    2022-05-13 [1] CRAN (R 4.2.0)
##  intergraph            2.0-2     2016-12-05 [1] CRAN (R 4.2.0)
##  IRanges               2.30.0    2022-04-26 [1] Bioconductor
##  iterators             1.0.14    2022-02-05 [1] CRAN (R 4.2.0)
##  jpeg                  0.1-9     2021-07-24 [1] CRAN (R 4.2.0)
##  KEGGREST              1.36.2    2022-06-09 [1] Bioconductor
##  knitr                 1.39      2022-04-26 [1] CRAN (R 4.2.0)
##  labeling              0.4.2     2020-10-20 [1] CRAN (R 4.2.0)
##  lattice               0.20-45   2021-09-22 [1] CRAN (R 4.2.0)
```

```
##  latticeExtra          0.6-29     2019-12-19 [1] CRAN (R 4.2.0)
##  lifecycle             1.0.1      2021-09-24 [1] CRAN (R 4.2.0)
##  limma                 3.52.2     2022-06-19 [1] Bioconductor
##  locfit                1.5-9.5    2022-03-03 [1] CRAN (R 4.2.0)
##  magrittr              2.0.3      2022-03-30 [1] CRAN (R 4.2.0)
##  Matrix                1.4-1      2022-03-23 [1] CRAN (R 4.2.0)
##  MatrixGenerics        1.8.1      2022-06-26 [1] Bioconductor
##  matrixStats           0.62.0     2022-04-19 [1] CRAN (R 4.2.0)
##  memoise               2.0.1      2021-11-26 [1] CRAN (R 4.2.0)
##  mgcv                  1.8-40     2022-03-29 [1] CRAN (R 4.2.0)
##  minet                 3.54.0     2022-04-26 [1] Bioconductor
##  munsell               0.5.0      2018-06-12 [1] CRAN (R 4.2.0)
##  NetRep                1.2.4      2020-10-07 [1] CRAN (R 4.2.0)
##  network               1.17.2     2022-05-21 [1] CRAN (R 4.2.0)
##  networkD3             0.4        2017-03-18 [1] CRAN (R 4.2.0)
##  nlme                  3.1-158    2022-06-15 [1] CRAN (R 4.2.0)
##  nnet                  7.3-17     2022-01-13 [1] CRAN (R 4.2.0)
##  paletteer             1.4.0      2021-07-20 [1] CRAN (R 4.2.0)
##  parameters            0.18.1     2022-05-29 [1] CRAN (R 4.2.0)
##  patchwork             1.1.1      2020-12-17 [1] CRAN (R 4.2.0)
##  performance           0.9.1      2022-06-20 [1] CRAN (R 4.2.0)
##  pillar                1.7.0      2022-02-01 [1] CRAN (R 4.2.0)
##  pkgconfig             2.0.3      2019-09-22 [1] CRAN (R 4.2.0)
##  plyr                  1.8.7      2022-03-24 [1] CRAN (R 4.2.0)
##  png                   0.1-7      2013-12-03 [1] CRAN (R 4.2.0)
##  preprocessCore        1.58.0     2022-04-26 [1] Bioconductor
##  purrr                 0.3.4      2020-04-17 [1] CRAN (R 4.2.0)
##  R6                    2.5.1      2021-08-19 [1] CRAN (R 4.2.0)
##  RColorBrewer          1.1-3      2022-04-03 [1] CRAN (R 4.2.0)
##  Rcpp                  1.0.8.3    2022-03-17 [1] CRAN (R 4.2.0)
##  RCurl                 1.98-1.7   2022-06-09 [1] CRAN (R 4.2.0)
##  rematch2              2.1.2      2020-05-01 [1] CRAN (R 4.2.0)
##  reshape2              1.4.4      2020-04-09 [1] CRAN (R 4.2.0)
##  RhpcBLASctl           0.21-247.1 2021-11-05 [1] CRAN (R 4.2.0)
##  rjson                 0.2.21     2022-01-09 [1] CRAN (R 4.2.0)
##  rlang                 1.0.3      2022-06-27 [1] CRAN (R 4.2.1)
##  rmarkdown             2.14       2022-04-25 [1] CRAN (R 4.2.0)
##  rpart                 4.1.16     2022-01-24 [1] CRAN (R 4.2.0)
##  rprojroot             2.0.3      2022-04-02 [1] CRAN (R 4.2.0)
##  RSQLite               2.2.14     2022-05-07 [1] CRAN (R 4.2.0)
##  rstatix               0.7.0      2021-02-13 [1] CRAN (R 4.2.0)
##  rstudioapi            0.13       2020-11-12 [1] CRAN (R 4.2.0)
##  S4Vectors             0.34.0     2022-04-26 [1] Bioconductor
##  scales                1.2.0      2022-04-13 [1] CRAN (R 4.2.0)
##  sessioninfo           1.2.2      2021-12-06 [1] CRAN (R 4.2.0)
##  shape                 1.4.6      2021-05-19 [1] CRAN (R 4.2.0)
##  statmod               1.4.36     2021-05-10 [1] CRAN (R 4.2.0)
##  statnet.common        4.6.0      2022-05-02 [1] CRAN (R 4.2.0)
##  statsExpressions      1.3.2      2022-05-20 [1] CRAN (R 4.2.0)
##  stringi               1.7.6      2021-11-29 [1] CRAN (R 4.2.0)
##  stringr               1.4.0      2019-02-10 [1] CRAN (R 4.2.0)
```

```
##  SummarizedExperiment   1.26.1     2022-04-29 [1] Bioconductor
##  survival               3.3-1      2022-03-03 [1] CRAN (R 4.2.0)
##  sva                    3.44.0     2022-04-26 [1] Bioconductor
##  tibble                 3.1.7      2022-05-03 [1] CRAN (R 4.2.0)
##  tidyr                  1.2.0      2022-02-01 [1] CRAN (R 4.2.0)
##  tidyselect             1.1.2      2022-02-21 [1] CRAN (R 4.2.0)
##  utf8                   1.2.2      2021-07-24 [1] CRAN (R 4.2.0)
##  vctrs                  0.4.1      2022-04-13 [1] CRAN (R 4.2.0)
##  WGCNA                  1.71       2022-04-22 [1] CRAN (R 4.2.0)
##  withr                  2.5.0      2022-03-03 [1] CRAN (R 4.2.0)
##  xfun                   0.31       2022-05-10 [1] CRAN (R 4.2.0)
##  XML                    3.99-0.10  2022-06-09 [1] CRAN (R 4.2.0)
##  xtable                 1.8-4      2019-04-21 [1] CRAN (R 4.2.0)
##  XVector                0.36.0     2022-04-26 [1] Bioconductor
##  yaml                   2.3.5      2022-02-21 [1] CRAN (R 4.2.0)
##  zeallot                0.1.0      2018-01-28 [1] CRAN (R 4.2.0)
##  zlibbioc               1.42.0     2022-04-26 [1] Bioconductor
##
##  [1] /home/faalm/R/x86_64-pc-linux-gnu-library/4.2
##  [2] /usr/local/lib/R/site-library
##  [3] /usr/lib/R/site-library
##  [4] /usr/lib/R/library
##
## ------------------------------------------------------------------------------
```