TF Classification

Arend Vancraeynest

2022-03-19

```
library(here)
```

Overview

Here, you must describe the goal of the analyses documented in this documented.

Proteome metadata

Ready the data for R manipulation

```
library(Biostrings)
proteomes <- list.files(here("data", "Proteomes"), full.names = TRUE)

seqs <- lapply(proteomes, readAAStringSet)

# A file containing the genome assembly versions and reference papers has been made manually
extra_data <- read.table(here("data", "Genome_references.txt"), header = FALSE, sep = "\t")
```

Generate the dataframe

```
# Initiate the dataframe based on the species names contained within the file names
# (genus abbreviated in 3 or 4 letters, species epithet in full, divided by a '.')
df <- data.frame(Species_name=gsub(".*/([a-zA-Z]{3,4}\\.[a-zA-Z]+)_(cv\\.[a-zA-Z0-9-]+)?.*", "\\1 \\2",
df$Species_name <- gsub("\\.", ". ", df$Species_name)

# Add the Genome Assembly version
df$Genome_assembly_version <- extra_data$V1

# Count the number of proteins and genes for each species and append to an initiated dataframe
check_isoforms <- function(seqlist = NULL) {
  combined_list <- lapply(seq_along(seqlist), function(x){
    all_seqs <- names(seqlist[[x]])
    proteins <- length(all_seqs)

    # Substring the unique gene names from the different types of annotation
    unique_seqs <- gsub(" pacid.*$", "", all_seqs)
    unique_seqs <- gsub("^.*gene=", "", unique_seqs)
    unique_seqs <- gsub(" (start.*|Protein)$", "", unique_seqs)
    unique_seqs <- gsub(r"(\t.*)", "", unique_seqs)

    # In case you want to keep all scaffolds
    #unique_seqs <- gsub("(?<=([0-9]{5}))(_|\\.)t?[0-9]+(.p)?$", "", unique_seqs, perl = T)
```

```
    # Otherwise
    unique_seqs <- gsub("(_|\\.)t?[0-9]+(.p)?$", "", unique_seqs)

    # Only retain unique gene handles
    unique_seqs <- unique(unique_seqs)
    genes <- length(unique_seqs)

    return(c(proteins, genes))
  }
  )
  df$Genes <<- sapply(combined_list, "[[", 2)
  df$Proteins <<- sapply(combined_list, "[[", 1)
}

check_isoforms(seqs)

# Add the reference papers
df$References <- extra_data$V2
```

```
# Tidy up the dataframe to make it more reader-friendly by replacing underscores by spaces in the heade
names(df) <- gsub("_", " ", names(df))


library(stringr)
print(df, right = F)
```

```
##    Species name                 Genome assembly version   Genes  Proteins
## 1  Aeth. arabicum               GCA 000411095.1            8649  37839
## 2  Ara. lyrata cv. MN47         Alyrata 384 v2.1          31073  33132
## 3  Ara. thaliana                Athaliana 447 Araport11   27654  27654
## 4  Boe. retrofracta             GCA 015832515.1           27048  28268
## 5  Boe. stricta                 Bstricta 278 v1.2         27416  29812
## 6  Bra. juncea cv. AU213        GCA 020002505.1           99904  99904
## 7  Bra. juncea cv. T84-66       GCA 020002515.1          100829 100829
## 8  Bra. napus cv. ZS11          GCA 000686985.2          101942 101942
## 9  Bra. nigra cv. NI100         CGI Ni100 ONT Assembly v2  59851  59851
## 10 Bra. nigra cv. Sangam        GCA 016432835.1           47953  47953
## 11 Bra. oleracea cv. JZS        GWHAASO00000000           59064  59064
## 12 Bra. rapa cv. Chiifu         GCA 000309985.3           46250  46250
## 13 Bra. rapa cv. FPsc           Brapa FPsc 277 v1.3       40492  40492
## 14 Cam. sativa cv. DH55         GCA 000633955.1           89418  94495
## 15 Cap. grandiflora             Cgrandiflora 266 v1.1     24805  26561
## 16 Cap. rubella cv. MonteGargano Crubella 474 v1.1        27682  27682
## 17 Eut. salsugineum             Esalsugineum 173 v1.0     26351  26351
## 18 Lea. alabamica               GCA 000411055.1            7561  38676
## 19 Raph. sativus cv. XYB36-2    Rapsa_Xiang v1.0          43239  43239
## 20 Sch. parvula                 Sparvula 574 v2.2         26847  26847
## 21 Sys. irio                    GCA 000411075.1            8309  49956
##    References
## 1  Haudry et al, 2013
## 2  Rawat et al, 2015
## 3  Cheng et al, 2017
## 4  Kliver et al, 2018
## 5  Lee et al, 2017
```

```
## 6  Yang et al, 2021
## 7  Yang et al, 2021
## 8  Sun et al, 2017
## 9  Perumal et al, 2020
## 10 Paritosh et al, 2020
## 11 Cai et al, 2020
## 12 Zhang et al, 2018
## 13 Nordberg et al, 2014
## 14 Kagale et al, 2014
## 15 Slotte et al, 2013
## 16 Slotte et al, 2013
## 17 Yang et al, 2013
## 18 Haudry et al, 2013
## 19 Zhang et al, 2015
## 20 Oh et al, 2014
## 21 Haudry et al, 2013
```

Session info

This document was created under the following conditions:

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Dutch_Belgium.1252  LC_CTYPE=Dutch_Belgium.1252
## [3] LC_MONETARY=Dutch_Belgium.1252 LC_NUMERIC=C
## [5] LC_TIME=Dutch_Belgium.1252
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] stringr_1.4.0       Biostrings_2.62.0   GenomeInfoDb_1.30.1
## [4] XVector_0.34.0      IRanges_2.28.0      S4Vectors_0.32.3
## [7] BiocGenerics_0.40.0 here_1.0.1
##
## loaded via a namespace (and not attached):
##  [1] crayon_1.5.0           digest_0.6.29          rprojroot_2.0.2
##  [4] bitops_1.0-7           magrittr_2.0.2         evaluate_0.15
##  [7] zlibbioc_1.40.0        rlang_1.0.2            stringi_1.7.6
## [10] cli_3.2.0              rstudioapi_0.13        rmarkdown_2.13
## [13] tools_4.1.2            RCurl_1.98-1.6         xfun_0.30
## [16] yaml_2.3.5             fastmap_1.1.0          compiler_4.1.2
## [19] htmltools_0.5.2        knitr_1.37             GenomeInfoDbData_1.2.7
```