

TF Classification

Arend Vancraeynest

2022-03-20

```
library(here)
```

Overview

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

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-z]{3,4}\\.[a-zA-Z]+)_((cv|ssp)\\. [a-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("^.*gene=", "", all_seqs, ignore.case = TRUE)
    unique_seqs <- gsub("\\s.*$", "", unique_seqs)

    # This also reduces all scaffolds annotated as 'scaffoldXXX_XX' to 'scaffoldXXX' (by Haudry et al.)
    # which results in an extremely high loss of data far below the expected number
    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$Coding_genes <- sapply(combined_list, "[", 2)
df$Annotated_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 headers
names(df) <- gsub("_", " ", names(df))

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

```

##	Species name	Genome assembly version	Coding genes
## 1	Aeth. arabicum	GCA 000411095.1	8649
## 2	Aeth. arabicum	Aarabicum v3.1	24932
## 3	Ara. lyrata cv. MN47	Alyrata 384 v2.1	31073
## 4	Ara. thaliana	Athaliana 447 Araport11	27654
## 5	Boe. retrofracta	GCA 015832515.1	27048
## 6	Boe. stricta	Bstricta 278 v1.2	27416
## 7	Bra. carinata cv. zd-1	GCA 016771965.1	97149
## 8	Bra. juncea cv. AU213	GCA 020002505.1	99904
## 9	Bra. juncea cv. T84-66	GCA 020002515.1	100829
## 10	Bra. napus cv. ZS11	GCA 000686985.2	101942
## 11	Bra. nigra cv. NI100	CGI Ni100 ONT Assembly v2	59851
## 12	Bra. nigra cv. Sangam	GCA 016432835.1	47953
## 13	Bra. oleracea cv. HDEM	GCA_900416815.2	61279
## 14	Bra. oleracea cv. JZS	GWHAAS000000000 v2.0	59064
## 15	Bra. rapa cv. Chiifu	GCA 000309985.3	46250
## 16	Bra. rapa cv. FPsc	Brapa FPsc 277 v1.3	40492
## 17	Cam. sativa cv. DH55	GCA 000633955.1	89418
## 18	Cap. grandiflora	Cgrandiflora 266 v1.1	24805
## 19	Cap. rubella cv. MonteGargano	Crubella 474 v1.1	27682
## 20	Car. hirsuta	Chirsuta MPIPZ v1.0	29458
## 21	Eut. salsugineum	Esalsugineum 173 v1.0	26351
## 22	Lea. alabamica	GCA 000411055.1	7561
## 23	Raph. raphanistrum ssp. landra	GWHANWL000000000 v1.0	43703
## 24	Raph. raphanistrum ssp. raphanistrum	GWHANWMO000000000 v1.0	42319
## 25	Raph. sativus cv. XYB36-2	Rapsa_Xiang v1.0	43239
## 26	Sch. parvula	Sparvula 574 v2.2	26847
## 27	Sys. irio	GCA 000411075.1	8309
##	Annotated proteins	References	
## 1	37839	Haudry et al, 2013	
## 2	24932	Nguyen et al, 2019	

```
## 3 33132 Rawat et al, 2015
## 4 27654 Cheng et al, 2017
## 5 28268 Kliver et al, 2018
## 6 29812 Lee et al, 2017
## 7 97149 Song et al, 2021
## 8 99904 Yang et al, 2021
## 9 100829 Yang et al, 2021
## 10 101942 Sun et al, 2017
## 11 59851 Perumal et al, 2020
## 12 47953 Paritosh et al, 2020
## 13 61279 Belser et al, 2018
## 14 59064 Cai et al, 2020
## 15 46250 Zhang et al, 2018
## 16 40492 Nordberg et al, 2014
## 17 94495 Kagale et al, 2014
## 18 26561 Slotte et al, 2013
## 19 27682 Slotte et al, 2013
## 20 37997 Gan et al, 2016
## 21 26351 Yang et al, 2013
## 22 38676 Haudry et al, 2013
## 23 46064 Zhang et al, 2021
## 24 44930 Zhang et al, 2021
## 25 43239 Zhang et al, 2015
## 26 26847 Oh et al, 2014
## 27 49956 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