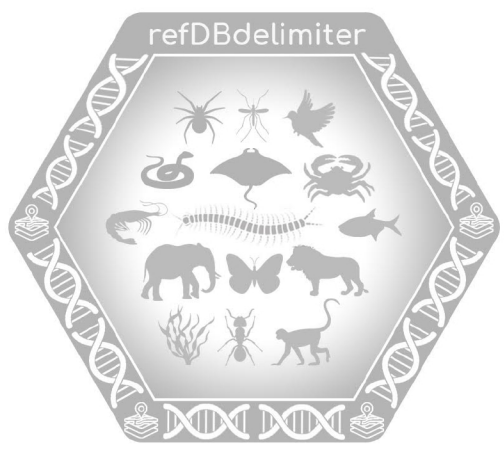

RefDBdelimiter: *An R Package for Refining Molecular Reference Databases with Occurrence Data and Performing Molecular Identification for Metabarcoding Analysis*

Fabricio dos Anjos Santa Rosa

OVERVIEW

This study presents an integrated workflow designed for streamlined acquisition and filtering phylogenetic data based on occurrence, and taxonomic assignment of Operational Taxonomic Units (OTUs) generated by metabarcoding approach. The workflow comprises ten sequential steps aimed at ensuring data accuracy and relevance. Initially, occurrence data is acquired from the Global Biodiversity Information Facility (GBIF) database, followed by rigorous filtering to exclude erroneous entries and standardize information. Subsequently, taxa-specific occurrence data are selected based on predefined geographic regions. Concurrently, sequences data are retrieved from the National Center for Biotechnology Information (NCBI) database and undergo stringent filtering procedures, including the exclusion of unverified sequences and standardization of sequence headers. Taxa-specific sequences are then selected to align with the occurrence data. This integrated approach enables the establishment of a local curated database, facilitating efficient data management and analysis. Furthermore, the workflow incorporates the assignment of taxonomic identifications to OTUs generated through metabarcoding workflows, leveraging the curated database for accurate taxonomic inference. The presented workflow offers a comprehensive framework for biodiversity studies, enhancing data reliability and enabling more robust analyses of species distributions and genetic diversity.

• Install and/or Load packages:

```
pack <- c('dplyr', 'tibble','devtools','plyr','readr','taxize','seqinr','ggplot2',
          'ape','DECIPHER','phangorn','phytools','tidyr','stringr','taxizedb',
          'rgbif','rmarkdown','sass')
vars <- pack[!(pack %in% installed.packages()[, "Package"])]
if (length(vars != 0)) {
  install.packages(vars, dependencies = TRUE)
}

sapply(pack, require, character.only = TRUE)
```

• Define the needed directories:

```
Directory <- paste0("C:/Users/Administrador/Downloads")
setwd("C:/Users/Administrador/Downloads")
```

• Install “wsl” command (This command creates a Linux subsystem in any windows machine):

```
# Check if WSL is installed
wsl_path <- Sys.which("wsl")

# If WSL is not installed, run the command to install it
if (nzchar(wsl_path)) {
  print("WSL is installed.")
} else {
  print("WSL is not installed. Installing...")
  system("wsl --install")
}
```

```
## [1] "WSL is installed."
```

• Get gbif database based on continent and taxa (i.e: Plants from South America):

```
setwd("C:/Users/Administrador/Downloads")

source("C:/Users/Administrador/Downloads/functions.R")

generate_gbif_taxa_dataset_gibi(continent = "SOUTH_AMERICA", scientificName = "Plantae", taxa_n = 300, gbif_taxa_
dataset = "C:/Users/Administrador/Downloads/gbif_taxa_dataset.txt")
```

```
## [1] "GBIF data acquisition completed successfully."
```

• Format headers of cleaned ncbi database:

```
setwd("C:/Users/Administrador/Downloads/")

source("C:/Users/Administrador/Downloads/functions.R")

format_ncbi_database_gibi(raw_database = "its_database.fasta", database_cleaned = "database_cleaned_formated.fast
a", min_sequence_length = 100, pattern = "UNVERIFIED")
```

```
## [1] "NCBI data cleaning is completed successfully."
```

• Subset NCBI databse based on GBIF:

```
setwd("C:/Users/Administrador/Downloads/")

subset_ncbi_based_on_gbif_gibi(gbif_database = "gbif_taxa_dataset.txt", cleaned_ncbi_database = "database_cleaned
_formated.fasta", ncbi_database_based_on_gbif = "its2_database.fasta")
```

```
## [1] "NCBI subset is completed successfully."
```

• Database production:

```
setwd("C:/Users/Administrador/Downloads/")

create_blast_db_gibi(database = "its2_database.fasta", parse_seqids = T, title = "ITS2_database", database_type =
"nucl")
```

```
## [1] "Database creation completed successfully."
```

• Taxonomic assignment of Operational Taxonomic Units (OTUs) based on a curated database:

```
source("C:/Users/Administrador/Downloads/functions.R")

setwd("C:/Users/Administrador/Downloads/")
Database_File <- paste0("/mnt/c/Users/Administrador/Downloads/its2_database.fasta")
Directory <- "C:/Users/Administrador/Downloads/"

blast_gibi(Directory, Database_File, max_target_seqs = 50, perc_identity = 90, qcov_hsp_perc = 95, num_threads =
6, Specie_Threshold = 98, Genus_Threshold = 95, Family_Threshold = 90)
```

```
## Database already exists, returning old file
```

```
## [1] "End of Run"
```

• See the Taxonomic assignment results:

```
taxonomic_assignment <- read.table("taxonomic_assignment.txt", header = TRUE, sep = "")

head(taxonomic_assignment[,1:10])
```

```
##   qseqid                seqid       Phylum         Class        Order
## 1   Otu1   Epidendrum fulgens Streptophyta Magnoliopsida  Asparagales
## 2   Otu2 Taliparitì tiliaceum Streptophyta Magnoliopsida     Malvales
## 3   Otu3      Justicia carnea Streptophyta Magnoliopsida     Lamiales
## 4   Otu4  Passiflora caerulea Streptophyta Magnoliopsida Malpighiales
## 5   Otu5   Cymbalaria muralis Streptophyta Magnoliopsida     Lamiales
## 6   Otu6      Justicia carnea Streptophyta Magnoliopsida     Lamiales
##           Family      Genus              Species pident Sample1
## 1    Orchidaceae Epidendrum   Epidendrum fulgens    100       0
## 2      Malvaceae Taliparitì Taliparitì tiliaceum    100       5
## 3    Acanthaceae   Justicia      Justicia carnea    100       0
## 4 Passifloraceae Passiflora  Passiflora caerulea    100      66
## 5 Plantaginaceae Cymbalaria   Cymbalaria muralis    100       0
## 6    Acanthaceae   Justicia      Justicia carnea    100      12
```

Detailed scripts about the entire metabarcoding workflow can be found on my GitHub profile