

Generating a README.txt file for the Dryad submission of data from study “DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life”

Description of README file

```
line0 <- paste("This README file was generated on", Sys.Date(), "by Luna L. Sanchez-Reyes, <https://orcid.org/0000-0001-7668-2528>")
```

General Information on data set

```
line1 <- "
```

GENERAL INFORMATION

1. Title of Dataset

```
Data from: DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life.
```

2. Author Information

```
Corresponding Researcher
```

```
name: Luna L. Sanchez-Reyes
```

```
institution: University of California, Merced, USA
```

```
email: sanchez.reyes.luna@gmail.com
```

```
Co-researcher 1
```

```
name: Emily Jane McTavish
```

```
institution: University of California, Merced, USA
```

```
Co-researcher 2
```

```
name: Brian C. O'Meara
```

```
institution: University of Tennessee, Knoxville, USA
```

3. Date of data collection: 2022-01-28

4. Geographic location of data collection: Online

5. Funding sources that supported the collection of the data: National Science Foundation, USA

6. Recommended citation for this dataset :

Sanchez-Reyes LL, McTavish EJ, & Mearns BC. (2019). DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life. BioRxiv, 782094.

Data from: DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life. Dryad, Dataset

"

Original file names, rename

```
file.copy(from = "../tables/table-fringillidae-small-example.csv",
          to = "../dryad/Sanchez-Reyes_etal_2022_table_1.csv",
          overwrite = TRUE)
file.copy(from = "../tables/table-fringillidae-small-example-summary.csv",
          to = "../dryad/Sanchez-Reyes_etal_2022_table_2.csv",
          overwrite = TRUE)
file.copy(from = "../tables/table-fringillidae-all-congruified.csv",
          to = "../dryad/Sanchez-Reyes_etal_2022_supplementary_table_S1.csv",
          overwrite = TRUE)
file.copy(from = "../tables/table-fringillidae-all-congruified.pdf",
          to = "../dryad/Sanchez-Reyes_etal_2022_supplementary_table_S1.pdf",
          overwrite = TRUE)
file.copy(from = "../tables/table-fringillidae-all-summary.csv",
          to = "../dryad/Sanchez-Reyes_etal_2022_supplementary_table_S2.csv",
          overwrite = TRUE)
file.copy(from = ".../tables/table-fringillidae-all-summary.pdf",
          to = "../dryad/Sanchez-Reyes_etal_2022_supplementary_table_S2.pdf",
          overwrite = TRUE)
file.copy(from = "../figures/figure-workflow/median_chronogram.tre",
          to = "../dryad/Sanchez-Reyes_etal_2022_figure_1_chronogram_mock__
            example.tre",
          overwrite = TRUE)
file.copy(from = "../figures/figure-small-example/median_chronogram.tre",
          to = "../dryad/Sanchez-Reyes_etal_2022_figure_3_chronogram_small__
            example.tre",
          overwrite = TRUE)
file.copy(from = "../figures/figure-fringillidae/median_chronogram.tre",
          to = "../dryad/Sanchez-Reyes_etal_2022_figure_5_chronogram_finches__
            example.tre",
          overwrite = TRUE)
file.copy(from = "../figures/fringillidae-topologies/fringillidae_monophyletic__
            _topology.tre",
          to = "../dryad/Sanchez-Reyes_etal_2022_figure_4A_topology_finches__
            mrca.tre",
          overwrite = TRUE)
file.copy(from = "../figures/fringillidae-topologies/fringillidae_paraphyletic__
            _topology.tre",
          to = "../dryad/Sanchez-Reyes_etal_2022_figure_4B_topology_finches__
            ncbi.tre",
```

```

        overwrite = TRUE)

files_in_path <- list.files(path = "../figures/figure-fringillidae/source_
    chronograms")

# get source chronograms newick file names
newick_files <- files_in_path[grep("[.] tre", files_in_path)]

# order files numerically
newick_files <- stringr::str_sort(newick_files, numeric = TRUE)

index <- 1
for (file_name in newick_files) {
    index <- index + 1
    file.copy(from = paste0("../figures/figure-fringillidae/source_chronograms/"
        , file_name),
        to = paste0("../dryad/Sanchez-Reyes_etal_2022_supplementary_figure_S
            ", index, "_source.tre"),
        overwrite = TRUE)
    write(paste0("https://raw.githubusercontent.com/LunaSare/datelifMS1/main/
        dryad/Sanchez-Reyes_etal_2022_supplementary_figure_S", index, "_source.
        tre"), file = "../dryad/urls.txt", append=TRUE)
}

# get cross validated chronograms newick file names
files_in_path <- list.files(path = "../figures/fringillidae-cross-validation")

newick_files <- files_in_path[grep("[.] tre", files_in_path)]

# order files numerically
newick_files <- stringr::str_sort(newick_files, numeric = TRUE)

index <- 1
for (file_name in newick_files) {
    index <- index + 1
    file.copy(from = paste0("../figures/fringillidae-cross-validation/", file_
        name),
        to = paste0("../dryad/Sanchez-Reyes_etal_2022_supplementary_figure_S
            ", index, "_cross_validated.tre"),
        overwrite = TRUE)
    write(paste0("https://raw.githubusercontent.com/LunaSare/datelifMS1/main/
        dryad/Sanchez-Reyes_etal_2022_supplementary_figure_S", index, "_cross_
        validated.tre"), file = "../dryad/urls.txt", append=TRUE)
}

```

Description of data files

line2.1 <- "DATA_&_FILE_OVERVIEW"

1. Description of dataset

These data were generated to investigate and showcase the performance of the datelife R package (<<https://github.com/phylostatic/datelife>>).

We showcased the application of the package with one mock example and two different biological examples.

The first biological example uses `datelife` on a small sample of bird species. The second one uses `datelife` on bird species belonging to the family `Fringillidae` of 'true finches', following the NCBI taxonomy. The mock example was created by replacing species names from the small example with letters, and reducing the number of source chronograms.

We evaluated the performance of the package `datelife` with two analysis: a benchmarking analysis to measure computing time of functions, and a cross validation analysis to test the accuracy and precision of the functions.

2. File List:

File 1 Name: `Sanchez-Reyes_etal_2022_table_1.csv`

File 1 Description: CSV (comma separated values) file with node age results from `datelife` 's small example, presented in Table 1 of the manuscript. It contains node names, taxon names defining the nodes, the corresponding node ages (in Million years), and references of source chronograms that were congruified to the chosen tree topology to extract node ages.

File 2 Name: `Sanchez-Reyes_etal_2022_table_2.csv`

File 2 Description: CSV (comma separated values) file summarizing age data (in Million years) from table 1, per node. These data are presented in Table 2 of the manuscript.

File 3 Name: `Sanchez-Reyes_etal_2022_supplementary_table_S1.csv`

File 3 Description: CSV (comma separated values) file with node age results from `datelife` 's `Fringillidae` example. It contains node names, taxon names defining the nodes, the corresponding node ages (in Million years), and references of source chronograms that were congruified to the chosen tree topology (shown in Figure 4B of the manuscript) to extract node ages.

File 4 Name: `Sanchez-Reyes_etal_2022_supplementary_table_S1.pdf`

File 4 Description: PDF version of file `Sanchez-Reyes_etal_2022_supplementary_table_S1.csv`

File 5 Name: `Sanchez-Reyes_etal_2022_supplementary_table_S2.csv`

File 5 Description: CSV (comma separated values) file summarizing age data (in Million years) from table S1, per node.

File 6 Name: `Sanchez-Reyes_etal_2022_supplementary_table_S2.pdf`

File 6 Description: PDF file version of file `Sanchez-Reyes_etal_2022_supplementary_table_S2.csv`.

File 7 Name: `Sanchez-Reyes_etal_2022_figure_1_chronogram_mock_example.tre`

File 7 Description: Newick version of chronogram from the mock example shown in Figure 1 of the manuscript (<https://www.biorxiv.org/content/10.1101/782094v2>).

File 8 Name: `Sanchez-Reyes_etal_2022_figure_3_chronogram_small_example.tre`

File 8 Description: Newick version of chronogram from the small example shown in Figure 3 of the manuscript (<https://www.biorxiv.org/content/10.1101/>

782094v2).

File_9_Name: Sanchez-Reyes_etal_2022_figure_4A_topology_finches_mrca.tre
File_9_Description: Newick version of topology extracted from the Open Tree of Life synthetic tree, shown in Figure 4A of the manuscript (<https://www.biorxiv.org/content/10.1101/782094v2>).

File_10_Name: Sanchez-Reyes_etal_2022_figure_4B_topology_finches_ncbi.tre
File_10_Description: Newick version of topology extracted from the Open Tree of Life synthetic tree, shown in Figure 4B of the manuscript (<https://www.biorxiv.org/content/10.1101/782094v2>).

File_11_Name: Sanchez-Reyes_etal_2022_figure_5_chronogram_finches_example.tre
File_11_Description: Newick version of chronogram from the finch example shown in Figure 5 of the manuscript (<https://www.biorxiv.org/content/10.1101/782094v2>).

File_12_Name: Sanchez-Reyes_etal_2022_supplementary_figure_S1.pdf
File_12_Description: PDF file showing results of cross validation analysis of datelife 's chronogram generating functions.

```
line2.2 <- c()
# short study references:
names_phylo_all <- c("Barker et al. 2013", # 1
  "Barker et al. 2015 chronogram 1", # 2
  "Barker et al. 2015 chronogram 2", # 3
  "Burns et al. 2014", # 4
  "Claramunt et al. 2015", # 5
  "Gibb et al. 2015", # 6
  "Hedges et al. 2015 chronogram 1", # 7
  "Hedges et al. 2015 chronogram 2", # 8
  "Hooper et al. 2017", # 9
  "Jetz et al. 2012 chronogram 1", # 10
  "Jetz et al. 2012 chronogram 2", # 11
  "Kimball et al. 2019 chronogram 1", # 12
  "Kimball et al. 2019 chronogram 2", # 13
  "Oliveros et al. 2019", # 14
  "Price et al. 2014 chronogram 1", # 15
  "Price et al. 2014 chronogram 2", # 16
  "Roquet et al. 2014 chronogram 1", # 17
  "Roquet et al. 2014 chronogram 2", # 18
  "Uyeda et al. 2017")

file_index <- 12
fig_index <- 1
for (name in names_phylo_all) {
  file_index <- file_index + 1
  fig_index <- fig_index + 1
  x1 <- paste0("File_", file_index, "_Name: Sanchez-Reyes_etal_2022_
    supplementary_figure_S", fig_index, ".jpg")
  x2 <- paste0("File_", file_index, "_Description: JPG file of results of
    cross validation analysis of datelife chronogram generating functions,
    using a tree topology from", name, ". Comparison of original chronogram
    (black) and the chronogram obtained using datelife (gray).\n")
}
```

```

file_index <- file_index + 1
x3 <- paste0("File_", file_index, "_Name:Sanchez-Reyes_etal_2022_
supplementary_figure_S", fig_index, "_cross_validated.tre")
x4 <- paste0("File_", file_index, "_Description:Newick_file_of_chronogram_
cross_validated_using_dates_obtained_with_datelife ,shown_in_gray_on_
supplementary_figure_Sanchez-Reyes_etal_2022_supplementary_figure_S",
fig_index, ".jpg\n")
file_index <- file_index + 1
x5 <- paste0("File_", file_index, "_Name:Sanchez-Reyes_etal_2022_
supplementary_figure_S", fig_index, "_source.tre")
x6 <- paste0("File_", file_index, "_Description:Newick_file_of_source_
chronogram_from", name, "_used_for_the_cross_validation_analysis_shown_on_
supplementary_figure_Sanchez-Reyes_etal_2022_supplementary_figure_S",
fig_index, ".jpg\n")

line2.2 <- c(line2.2, x1, x2, x3, x4, x5, x6)
}

```

Methods that generated the data files

```

line3 <- "
METHODOLOGICAL_INFORMATION

```

All age data and chronograms were obtained using the R package datelife v0.6.5
<https://CRAN.R-project.org/package=datelife> and the chronogram database
 OpenTreeChronograms v2022.1.28
 "

Individual descriptions of each data file

```

var_description_1 <- c("Node names in tree topology, as defined by taxon A and
taxon B. ",
"Scientific name of the first element of a taxon pair
that defines a node from a tree topology.",
"Scientific name of the second element of a taxon pair
that defines a node from a tree topology.",
"Ages of nodes defined by taxon A and taxon B, in
Million years.",
"Reference of the study that published the chronogram
from which the node ages were extracted.")

var_description_2 <- c("Names of nodes from the tree topology used for the
dating analysis.",
"Minimum node age, in Million years.",
"First quartile of node age distributions, in Million
years.",
"Mean node age, in Million years.",
"Median node age, in Million years.",
"Third quartile of node age distributions, in Million
years.",
"Maximum node age, in Million years.",
"Variance of node age, in Million years.",

```

```

"Standard deviation of node age, in Million years.")

csv_file_names <- c("Sanchez-Reyes_etal_2022_table_1.csv",
  "Sanchez-Reyes_etal_2022_table_2.csv",
  "Sanchez-Reyes_etal_2022_supplementary_table_S1.csv",
  "Sanchez-Reyes_etal_2022_supplementary_table_S2.csv")
tre_file_names <- c("Sanchez-Reyes_etal_2022_figure_1_chronogram_mock_example.
  tre")

text <- c()
var_descriptions <- list(var_description_1, var_description_2, var_description
  _1, var_description_2)
names(var_descriptions) <- csv_file_names

for (file_name in csv_file_names) {
  # read the csv file in:
  csv_data <- utils::read.csv(file = paste0("../dryad/", file_name), header =
    TRUE, row.names = NULL)
  # remove the column with row numbers:
  csv_data <- csv_data[, -1]
  # create the description text
  var_list <- gsub("[.] ", "\n", colnames(csv_data))
  var_list <- paste0("\t", var_list, ":", var_descriptions[[file_name]])
  text <- c(text,
    c(paste("\nDATA-SPECIFIC INFORMATION FOR:", file_name, "\n"),
      paste("1. Number of variables/columns:", ncol(csv_data), "\n"),
      paste("2. Number of cases/rows:", nrow(csv_data), "\n"),
      "3. Variable List:", var_list,
      "\n4. Missing data codes: None",
      "\n5. Abbreviations used: NA = not applicable\n"))
}

line4 <- text

```

Writing everything to TXT file

```

writeLines(text = c(line0,
  line1,
  line2.1,
  line2.2,
  line3,
  line4), con = "../dryad/README_Sanchez-Reyes_etal_2022_
  data.txt")

```