README for Dryad Data Package from study "DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life"

This README file was generated on 2022-07-02 by Luna L. Sánchez Reyes, https://orcid.org/0000-0001-7668-2528

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:

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/phylotastic/datelife). We showcased the application of the package with one mock example and two different biological examples. The mock example was GENERATED WITH 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. We investigate 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:

   File 2 Name: Sanchez-Reyes_etal_2022_table_2.csv File 2 Description:

   File 3 Name: Sanchez-Reyes_etal_2022_figure_1_chronogram_mock_example.tre File 3 Description:

   File 4 Name: Sanchez-Reyes_etal_2022_figure_3_chronogram_small_example.tre File 4 Description:

   File 6 Name: Sanchez-Reyes_etal_2022_figure_4A_topology_finches_mrca.tre File 6: Description:

   File 7 Name: Sanchez-Reyes_etal_2022_figure_4B_topology_finches_ncbi.tre File 7 Description:

   File 5 Name: Sanchez-Reyes_etal_2022_figure_5_chronogram_finches_example.tre File 5 Description:

   File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S1.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions.

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S2.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Barker et al. 2013. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S2.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S2 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S3.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S3.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S3 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S4.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S4.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S4 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S5.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S5.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S5 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S6.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S6.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S6 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S7.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S7.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S7 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S8.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S8.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S8 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S9.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S9.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S9 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S10.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S10.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S10 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S11.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S11.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S11 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S12.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S12.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S12 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S13.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S13.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S13 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S14.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S14.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S14 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S15.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S15.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S15 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S16.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S16.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S16 (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.pdf File XXX Description: Results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from XXX. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_figure_S.tre File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S (gray).

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_table_S1.csv File XXX Description:

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_table_S1.pdf File XXX Description:

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_table_S2.csv File XXX Description:

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_table_S2.pdf File XXX Description:

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_ File XXX Description:

File XXX Name: Sanchez-Reyes_etal_2022_supplementary_ File XXX Description:

3. Original names and locations of files at https://github.com/LunaSare/datelifeMS1:

```
cp tables/table-fringillidae-all.pdf dryad/Supplementary_Table_S1.pdf
cp tables/table-fringillidae-all-summary.pdf dryad/Supplementary_Table_S2.pdf
cp figures/figure-cross-validation/fig-cross-validation-xy-plots-diffs.pdf dryad/Supplementary_Figure_S
cp figures/figure-cross-validation/cross_validation_1.jpg dryad/Supplementary_Figure_S2.jpg
cp figures/figure-cross-validation/cross_validation_2.jpg dryad/Supplementary_Figure_S3.jpg
cp figures/figure-cross-validation/cross_validation_3.jpg dryad/Supplementary_Figure_S4.jpg
cp figures/figure-cross-validation/cross_validation_4.jpg dryad/Supplementary_Figure_S5.jpg
cp figures/figure-cross-validation/cross_validation_5.jpg dryad/Supplementary_Figure_S6.jpg
cp figures/figure-cross-validation/cross_validation_6.jpg dryad/Supplementary_Figure_S7.jpg
cp figures/figure-cross-validation/cross_validation_7.jpg dryad/Supplementary_Figure_S8.jpg
cp figures/figure-cross-validation/cross_validation_8.jpg dryad/Supplementary_Figure_S9.jpg
cp figures/figure-cross-validation/cross_validation_9.jpg dryad/Supplementary_Figure_S10.jpg
cp figures/figure-cross-validation/cross_validation_10.jpg dryad/Supplementary_Figure_S11.jpg
cp figures/figure-cross-validation/cross_validation_11.jpg dryad/Supplementary_Figure_S12.jpg
cp figures/figure-cross-validation/cross_validation_12.jpg dryad/Supplementary_Figure_S13.jpg
cp figures/figure-cross-validation/cross_validation_13.jpg dryad/Supplementary_Figure_S14.jpg
cp figures/figure-cross-validation/cross_validation_14.jpg dryad/Supplementary_Figure_S15.jpg
cp figures/figure-cross-validation/cross_validation_15.jpg dryad/Supplementary_Figure_S16.jpg
cp figures/figure-cross-validation/cross_validation_16.jpg dryad/Supplementary_Figure_S17.jpg
cp figures/figure-cross-validation/cross_validation_17.jpg dryad/Supplementary_Figure_S18.jpg
cp figures/figure-cross-validation/cross_validation_18.jpg dryad/Supplementary_Figure_S19.jpg
cp figures/figure-cross-validation/cross_validation_19.jpg dryad/Supplementary_Figure_S20.jpg
```

METHODOLOGICAL INFORMATION

Bacterial detection was performed on the ViiA7 Real-time PCR System (Thermo Fisher Scientific) using PrimeTime® qPCR primers, probes and mastermix (IDT) according to the manufacturer's instructions. Reactions were performed using 1X PrimeTime® Gene Expression Master Mix, 1X PrimeTime®qPCR Assay and up to 10ng of DNA. Cycling conditions were 95°C for 3 minutes, 60 cycles of 95°C for 5 seconds and 60°C for 30 seconds. Amplification results were reviewed using QuantStudioTM Real-Time PCR Software version 1.1 (Thermo Fisher Scientific). Amplification of beta-actin and prostaglandin transporter (PGT) was used to determine relative abundance.

Diversity profiling was performed by AGRF (Australian Genome Research Facility, Melbourne Australia). Samples were amplified with universal primers to the V1-V3 region of the bacterial 16S gene (forward AGAGTTTGATCMTGGCTCAG; reverse GWATTACCGCGGCKGCTG). Amplicons were indexed using the Nextera XT Index Kit (Illumina, San Diego, CA, USA) followed by Paired End sequencing on a MiSeq next generation sequencer (Illumina). Paired-end reads were assembled by aligning the forward and reverse reads using PEAR1 (version 0.9.5). Primers were identified and trimmed. Trimmed sequences were processed using Quantitative Insights into Microbial Ecology (QIIME 1.8) USEARCH (version 8.0.1623) and UPARSE software. Sequences were quality filtered and sorted by abundance after removal of full-length duplicate sequences. Singletons or unique reads were discarded. Sequences were clustered and then chimera filtered using "rdp_gold" database as reference. Reads were mapped back to Operational Taxonomic Units with a minimum identity of 97% and taxonomy was assigned using the QIIME 1 default classifier, pre-trained against Greengenes database5 (Version 13_8, Aug 2013).

DATA-SPECIFIC INFORMATION FOR: Rye_2021_a_Cohort_IDs.xlsx

1. Number of variables: 4
2. Number of cases/rows: 73
3. Variable List: Screening_cohort: Patient included in screening cohort; yes/no Screening_cohort_ID: Patient ID number for screening cohort Site_inv_cohort: Patient included in site investigation cohort; yes/no Site_inv_cohort_ID: Patient ID number for site investigation cohort
4. Missing data codes: None
5. Abbreviations used: N/A; not applicable
6. Other relevant information: 20 of 21 patients in the screening cohort whose tumour tested positive for F. nucleatum, B. fragilis or both species were included in the site investigation cohort, along with 31 additional patients. Tumour material for one screening cohort patient whose tumour was positive for both species (patient 24) was not available for further DNA extraction.

DATA-SPECIFIC INFORMATION FOR: Rye_2021_b_Screening_qPCR_data.xlsx

1. Number of variables: 9
2. Number of cases/rows: 42
3. Variable List: Screening_cohort_ID: Screening cohort patient ID number B-actin_mean_Ct: Mean Ct for beta-actin; no. cycles PGT_mean_Ct: Mean Ct for prostaglandin transporter; no. cycles F_nucleatum_mean_Ct: Mean Ct for Fusobaterium nucleatum; no. cycles B_fragilis_gyrase_mean_Ct: Mean Ct for Bacteroides fragilis gyrase; no. cycles B_fragilis_bft_mean_Ct: Mean Ct for Bacteroides fragilis toxin; no. cycles B_breve_mean_Ct: Mean Ct for Bifidobacterium breve; no. cycles C_showae_mean_Ct: Mean Ct for Campylobacter showae; no. cycles L_buccalis_mean_Ct: Mean Ct for Leptotrichia buccalis ; no. cycles

4. Missing data codes: No Amp; No amplification
5. Abbreviations used: Amp; amplification, BFT; Bacteroides fragilis toxin, PGT; prostaglandin transporter
6. Other relevant information: Reactions were performed in duplicate. Results were reported where one or both samples amplified. Values represent the mean Ct where both samples amplified or the individual Ct for single amplifications.

DATA-SPECIFIC INFORMATION FOR: Rye_2021_c_Site_inv_species_by_site.xlsx

1. Number of variables: 9 + comments field
2. Number of cases/rows: 436
3. Variable List: Site_inv_cohort_ID: Site investigation cohort patient ID number Tumour_ID: Tumour ID number for patients with synchronous tumours (n = 5); 1/2 Site: Area of tissue targeted; Normal (proximal), Normal (distal), Normal (adjacent), Tum luminal surface, Tumour, Tumour (mutinous), Invading margin, Inflammation, Stroma, Lymph node Site_code: Numerical code assigned to site; 1-11 Site_code_2: Second site if same area used to represent more than one site; 1-11 B_actin_mean_Ct: Mean Ct for beta-actin; no. cycles PGT_mean_Ct: Mean Ct for prostaglandin transporter; no. cycles F_nucleatum_mean_Ct: Mean Ct for Fusobaterium nucleatum; no. cycles B_fragilis_mean_Ct: Mean Ct for Bacteroides fragilis gyrase; no. cycles Comments: Used to flag values excluded from analysis
4. Missing data codes: No Amp; No amplification Single Amp; Single amplification N/A: not applicable
5. Abbreviations used: Amp; amplification, PGT; prostaglandin transporter, Tum; tumour
6. Other relevant information: Target sites included: normal tissue from the proximal and distal surgical margins of the resection specimen, normal tissue adjacent to the tumour region, the tumour luminal surface, central tumour, mucinous tumour (where applicable), invading margin, sites of inflammation, stroma, lymph nodes containing tumour deposits and metastatic sites where available. Metastatic site data have been excluded from this dataset to minimise any risk of deductive disclosure of patient identities. B-actin and PGT reactions were performed in duplicate and all samples amplified. F. nucleatum and B. fragilis reactions were performed in triplicate and results reported where two or more samples amplified. Mean cT values were excluded from analysis if the SD of the replicate Ct values > 5 (highlighted red).

DATA-SPECIFIC INFORMATION FOR: Rye_2021_d_Site_inv_16S_qPCR_data.xlsx

1. Number of variables: 4
2. Number of cases/rows: 51
3. Variable List: Site_inv cohort_ ID: Site investigation cohort patient ID number PGT_mean_Ct: Mean Ct for prostaglandin transporter; no. cycles 16S_TFS_mean_Ct: Mean Ct for 16S rRNA using primer set 1 16S_IDT_mean_Ct: Mean Ct for 16S rRNA using primer set 2
4. Missing data codes: None
5. Abbreviations used: IDT; Integrated DNA Technologies, PGT; prostaglandin transporter, TFS; Thermo Fisher Scientific
6. Other relevant information: Primer set 1: Obtained from Integrated DNA Technologies; Sequences as published in Nadkarni et al. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. Microbiology. 2002;148(Pt 1):257-66. doi: 10.1099/00221287-148-1-257. PubMed PMID: 11782518. Primer set 2: Obtained from Thermo Fisher Scientific; Assay ID Ba04230899_s1. Performed on DNA samples extracted from the tumour luminal surface.

DATA-SPECIFIC INFORMATION FOR: Rye_2021_e_Seq_data_absolute_abundance.xlsx

1. Number of variables: 7

2. Number of cases/rows: 308
3. Variable List: OTU_ID: OTU ID number Pt_32: Absolute abundance of each OTU for patient 32; no. sequence reads Pt_26: Absolute abundance of each OTU for patient 26; no. sequence reads Pt_35: Absolute abundance of each OTU for patient 35; no. sequence reads Pt_28: Absolute abundance of each OTU for patient 28; no. sequence reads Pt_29: Absolute abundance of each OTU for patient 29; no. sequence reads Consensus_Lineage: Consensus lineage
4. Missing data codes: None
5. Abbreviations used: OTU; Operational Taxonomic Unit, Pt; patient
6. Other relevant information: Performed on DNA samples extracted from the tumour luminal surface. Taxonomy assigned using the QIIME 1 default classifier, pre-trained against Greengenes database5 (Version 13_8, Aug 2013).

DATA-SPECIFIC INFORMATION FOR: Rye_2021_f_Seq_data_otu_table.xlsx

1. Number of variables: 6
2. Number of cases/rows: Sheet 1 'Phylum_level': 13 Sheet 2 'Class_level': 21 Sheet 3 'Order_level': 28 Sheet 4 'Family_level': 49 Sheet 5 'Genus_level': 85 Sheet 6 'Species_level': 101
3. Variable List: Taxon: Assigned taxon Pt_32: Relative abundance of each taxon for patient 32; proportion of total sequence reads Pt_26: Relative abundance of each taxon for patient 26; proportion of total sequence reads Pt_35: Relative abundance of each taxon for patient 35; proportion of total sequence reads Pt_28: Relative abundance of each taxon for patient 28; proportion of total sequence reads Pt_29: Relative abundance of each taxon for patient 29; proportion of total sequence reads
4. Missing data codes: None
5. Abbreviations used: Pt; patient
6. Other relevant information: Performed on DNA samples extracted from the tumour luminal surface. Taxonomy assigned using the QIIME 1 default classifier, pre-trained against Greengenes database5 (Version 13_8, Aug 2013).

DATA-SPECIFIC INFORMATION FOR: Rye_2021_g_Seq_data_mg_blast.xlsx

1. Number of variables: 16
2. Number of cases/rows: 366
3. Variable List: subject: OTU ID gi: Nucleotide BLAST accession number evalue: Expectation Value bit-score: Bit score score: Raw alignment score alignment-length: Alignment length; number of bases %identity: Percentage of bases identical to OTU group reference sequence identical: Number of bases identical to OTU group reference sequence positives: Number of bases with a positive match to OTU group reference sequence %positives: Percentage of bases with a positive match to OTU group reference sequence scientific-names: Scientific name of OTU classification group common-names: Common name of OTU classification group subject-blast-names: Grouping name subject-super-kingdoms: Kingdom names subject-title: Descriptive title for OTU group OTU-seq: OTU reference Sequence
4. Missing data codes: None
5. Abbreviations used: OTU; Operational Taxonomic Unit
6. Other relevant information: Taxonomy assigned using the QIIME 1 default classifier, pre-trained against Greengenes database5 (Version 13_8, Aug 2013).