

# Dryad Data Package for study “DateLife: leveraging databases and analytical tools to reveal the dated Tree of Life”

This README file was generated on 2022-07-13 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](mailto: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:

Sánchez-Reyes LL, McTavish EJ & O’Meara 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

## 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 containing results from the small example. Names and ages (in Million years) of nodes from source chronograms that have been congruified to nodes of the chosen tree topology.

File 2 Name: Sanchez-Reyes\_etal\_2022\_table\_2.csv

File 2 Description: CSV (comma separated values) file containing a summary of node age (in Million years) data from table S1.

File 3 Name: Sanchez-Reyes\_etal\_2022\_figure\_1\_chronogram\_mock\_example.tre

File 3 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 4 Name: Sanchez-Reyes\_etal\_2022\_figure\_3\_chronogram\_small\_example.tre

File 4 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 6 Name: Sanchez-Reyes\_etal\_2022\_figure\_4A\_topology\_finches\_mrca.tre

File 6: 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 7 Name: Sanchez-Reyes\_etal\_2022\_figure\_4B\_topology\_finches\_ncbi.tre

File 7 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 5 Name: Sanchez-Reyes\_etal\_2022\_figure\_5\_chronogram\_finches\_example.tre

File 5 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 XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S1.pdf

File XXX Description: PDF file showing results of cross validation analysis of datelife's chronogram generating functions.

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S2.jpg

File XXX Description: JPG file of 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.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Barker et al. 2015, chronogram 1. 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.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Barker et al. 2015, chronogram 2. 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.jpg  
File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Burns et al. 2015. 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.jpg  
File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Claramunt et al. 2015. 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.jpg  
File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Gibb et al. 2015. 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.jpg  
File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Hedges et al. 2015, chronogram 1. 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.jpg  
File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Hedges et al. 2015, chronogram 2. 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.jpg  
File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Hooper et al. 2017. 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.jpg  
File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Jetz et al. 2012, chronogram 1. 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.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Jetz et al. 2012, chronogram 2. 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.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Kimball et al. 2019, chronogram 1. 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.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Kimball et al. 2019, chronogram 2. 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.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Oliveros et al. 2019. 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.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Price et al. 2014, chronogram 1. 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\_S17.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Price et al. 2014, chronogram 2. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S17.tre

File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S17 (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S18.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating

functions, using a tree topology from Roquet et al. 2014, chronogram 1. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S18.tre

File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S18 (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S19.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Roquet et al. 2014, chronogram 2. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S19.tre

File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S19 (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S20.jpg

File XXX Description: JPG file of results of cross validation analysis of datelife's chronogram generating functions, using a tree topology from Uyeda et al. 2017. Comparison of original chronogram (black) and the chronogram obtained using datelife (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_figure\_S20.tre

File XXX Description: Newick file of chronogram obtained with datelife, shown in supplementary Figure S20 (gray).

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_table\_S1.csv

File XXX Description: CSV (comma separated values) file

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_table\_S1.pdf

File XXX Description: PDF file of

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_table\_S2.csv

File XXX Description: CSV (comma separated values) file

File XXX Name: Sanchez-Reyes\_etal\_2022\_supplementary\_table\_S2.pdf

File XXX Description: PDF file of r

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
```

## 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 QuantStudio™ 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).

```
files <- c()
text <- c()
for (file in 1:10) {

  text <- paste(text,
                paste("DATA-SPECIFIC INFORMATION FOR:", file, "\\newline"))
}
text
#> [1] " DATA-SPECIFIC INFORMATION FOR: 1 \\newline DATA-SPECIFIC INFORMATION FOR: 2 \\newline DATA-SPE
```

DATA-SPECIFIC INFORMATION FOR: 1  
 DATA-SPECIFIC INFORMATION FOR: 2  
 DATA-SPECIFIC INFORMATION FOR: 3  
 DATA-SPECIFIC INFORMATION FOR: 4  
 DATA-SPECIFIC INFORMATION FOR: 5  
 DATA-SPECIFIC INFORMATION FOR: 6  
 DATA-SPECIFIC INFORMATION FOR: 7  
 DATA-SPECIFIC INFORMATION FOR: 8  
 DATA-SPECIFIC INFORMATION FOR: 9  
 DATA-SPECIFIC INFORMATION FOR: 10  
 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 Fusobacterium 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 (mucinous), 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 *Fusobacterium 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 eval: 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).