The Journal of Open Source Software

genomesizeR: An R package for genome size prediction

Celine Mercier [1*] **and Joane Elleouet** [1*]

1 Scion, New Zealand Forest Research Institute, New Zealand * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

The genome size of organisms present in an environment can provide many insights into evolutionary and ecological processes at play in that environment. The genomic revolution has enabled a rapid expansion of our knowledge of genomes in many living organisms, and most of that knowledge is classified and readily available in the databases of the National Center for Biotechnology Information (NCBI). The `genomesizeR` tool leverages the wealth of taxonomic and genomic information present in NCBI databases to infer the genome size of Archeae, Bacteria, or Eukaryote organisms identified at any taxonomic level. This R package uses statistical modelling on data from the most up-to-date NCBI databases and provides three statistical methods for genome size prediction of a given taxon, or group of taxa. A straightforward 'weighted mean' method identifies the closest taxa with available genome size information in the taxonomic tree, and averages their genome sizes using weights based on taxonomic distance. A frequentist random effect model uses nested genus and family information to output genome size estimates. Finally a third option provides predictions from a distributional Bayesian multilevel model which uses taxonomic information from genus all the way to superkingdom, therefore providing estimates and uncertainty bounds even for under-represented taxa.

All three methods use:

- A list of queries; a query being a taxon or a list of several taxa. The package was designed to make it easy to use with data coming from environmental DNA experiments, but works with any table of taxa.
- A reference database containing all the known genome sizes, built from the NCBI databases, with associated taxa, provided in an archive to download.
- A taxonomic tree structure as built by the NCBI, provided in the same archive.

`genomesizeR` retrieves the taxonomic classification of input queries, estimates the genome size of each query, and provides 95% confidence intervals for each estimate.

Statement of need

The size of microbial genomes and its evolution can provide important insights into evolutionary and ecological processes influencing both microbial species and the environments in which they inhabit. The shedding of unnecessary genetic elements and their associated biosynthetic pathways, for example, is a common phenomenon observed in organisms with a high degree of host symbiosis (Brader et al., 2014; Moran, 2002; Vandenkoornhuyse et al., 2007). Genome size reduction has also been observed in organisms experiencing arid environments (Liu et al., 2023), or a narrow range of substrates or metabolic options (Tyson et al., 2004). Among many others, these findings demonstrate the opportunities associated with including genome size as a key trait in microbial communities to provide insights spanning niche size, co-evolution, adaption, and metabolic flexibility of the microbiomes present, but also stability, and ecophysiological and functional complexity of abiotic and biotic environments.

However, characterizing genome size for all organisms in a microbiome remains challenging. Methods in the past have included the use of DNA staining through to cell enumeration, flow cytometry, culturing, gel electrophoresis, and DNA renaturation kinetics. All have merits and limitations (Raes et al., 2007). The widespread availability and use of microbiome related tag-amplicon DNA sequencing has tremendously increased our scientific knowledge of microbial genomics. There is also an opportunity to explore the rapidly expanding archives of short read DNA libraries (i.e. extant 16S and ITS amplicon sequences).

Alternatively, when well documented and archived in user-friendly and publicly available databases, the exponentially growing genomic knowledge of micro-organisms is an inexpensive resource unlocking a myriad of research opportunities in all fields of environmental sciences, from human and agricultural microbiomes through to aquatic, soil, and atmospheric ecosystems. Combining available genome size information to data and metadata on community structure from existing projects can add further scientific value to investment already made in these projects, at no added cost.

However, genome size estimates for many taxa found in environmental samples are missing from public databases, or fully unknown. The evolutionary rule that phylogenetically related organisms share genetic similarities can be exploited and genome size for taxa with unknown genome size can be statistically inferred from related taxa with known genome size, using taxonomy as a proxy for phylogeny. Another challenge is the precision of identification: some taxa can only be identified at high taxonomic levels. Statistical methods can also be used to infer their genome size range from databases. To our knowledge, there is no convenient and fast way to obtain genome size estimates with uncertainty bounds for all organisms identified or partially identified in an environmental sample.

Using the increased prevalence of whole-genome information for all organisms, we have therefore developed `genomesizeR`, allowing the inference of genome size of many queries at once, based on taxonomic information and available genome data from the National Center for Biotechnology Information (NCBI).

Methods

NCBI database filtering and processing

The reference database used is built by querying all genome metadata information from the curated NCBI RefSeq database (O'Leary et al., 2016). Filters are applied to only keep full genomes, and discard data that the NCBI has tagged as anomalous, or abnormally large or small.

This raw database is then prepared to include more pre-computed information to be used by the package. Genome sizes are aggregated to the species level by iteratively averaging all entries below, hence the package can only provide estimates at the level of species and above. Average genome sizes and their associated standard error values are also pre-computed, to be used by the weighted mean method.

Bayesian method

The NCBI database of species with known genome sizes was split by superkingdom (Bacteria, Archeae, Eukaryotes). A distributional Bayesian linear hierarchical model using the `brm` function from the `brms` package was fitted to each superkingdom dataset. The general model structure is outlined below.

$$log(G_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

where G_i is the genome size of species i in the units of 10 Mbp. The model uses predictors for both mean and standard deviation. The mean is modelled as follows:

$$\mu_i = \alpha_0 + \alpha_{genus_{g[i]}} + \alpha_{family_{f[i]}} + \alpha_{order_{o[i]}} + \alpha_{class_{c[i]}} + \alpha_{phylum_{p[i]}}$$
$$\alpha_{genus_{g[i]}} \sim \mathcal{N}(0, \sigma^2_{genus})$$
$$\alpha_{family_{f[i]}} \sim \mathcal{N}(0, \sigma^2_{family})$$
$$\alpha_{order_{o[i]}} \sim \mathcal{N}(0, \sigma^2_{order})$$
$$\alpha_{class_{c[i]}} \sim \mathcal{N}(0, \sigma^2_{class})$$
$$\alpha_{phylum_{p[i]}} \sim \mathcal{N}(0, \sigma^2_{phylum})$$

The following prior distributions are used:

$$\alpha_0 \sim \mathcal{N}(0, 5)$$
$$(\sigma_{genus}, \sigma_{family}, \sigma_{order}, \sigma_{class}, \sigma_{phylum}, s_{class}, s_{phylum}) \sim \mathcal{N}^+(0, 1)$$

Differences in genome size variability was observed among taxa, therefore the model also adds predictors to the standard deviation of the response. The standard deviation is modelled as follows:

$$log(\sigma_i) = \lambda_0 + \lambda_{class_{c[i]}} + \lambda_{phylum_{p[i]}}$$
$$\lambda_{class_{c[i]}} \sim \mathcal{N}(0, s^2_{class})$$
$$\lambda_{phylum_{p[i]}} \sim \mathcal{N}(0, s^2_{phylum})$$

with priors

$$\lambda_0 \sim \mathcal{N}(0, 1)$$
$$(s_{class}, s_{phylum}) \sim \mathcal{N}^+(0, 1)$$

$\mathcal{N}^+$ is the normal distribution truncated to positive values. $g[i]$,$f[i]$,$o[i]$,$c[i]$ and $p[i]$ are respectively the index for the genus, family, order, class, and phylum of entry i in the species-level database. Note that taxonomic groups are naturally nested within each other and the indices are designed to be unique to the particular taxonomic group it represents.

The estimation process uses Stan's Hamiltonian Monte Carlo algorithm with the U-turn sampler.

Posterior predictions are obtained using the `predict` function from the `brms` package, and 95% credible intervals are obtained using 2.5% and 97.5% quantiles from the posterior distribution.

Queries corresponding to identified species with an available genome size estimate in the NCBI database get allocated the genome size value of the database (averaged at the species level) and 95% confidence intervals are calculated as XXX

Frequentist method

A frequentist linear mixed-effects model using the `lmer` function from the `lme4` package (Bates et al., 2015) was fitted to the NCBI database of species with known genome sizes. The model is as follows:

$$log(G_i) = \alpha_0 + \alpha_{genus_{g[i]}} + \alpha_{family_{f[i]}} + e_i$$

where α_0 is the overall mean, $\alpha_{genus_{g[i]}}$ and $\alpha_{family_{f[i]}}$ are random effect of genus and family for genus $g[i]$ and family $f[i]$ and e_i is the residual error of observation i.

The estimation process using the restricted maximum likelihood method (REML). A prediction interval is computed using the `predictInterval` function from the `merTools` package (Knowles & Frederick, 2024). As higher nested levels (order, class) are not taken into account in the model, predictions are not produced for queries above the family level.

Weighted mean method

The weighted mean method computes the genome size of a query by averaging the known genome sizes of surrounding taxa in the taxonomic tree, with a weighted system where further neighbours have less weight in the computed mean. The identification of related taxa is limited to levels below and including order.

In the calculation of the prediction for a given query, the weight for a related taxon in the database is calculated as

$$w_i = \frac{1}{d_i + 1}$$

where d_i is the distance of the related taxon to the query in number of nodes in the taxonomy.

The confidence interval is calculated as:

$$CI = 1.96 \times \sqrt{\hat{\mu}} \tag{1}$$

where $\hat{\mu}$ is the computed weighted mean.

For queries relating to well-characterised species where many genetic studies have been performed, such as model organisms, this might lead to more precise predictions than the two other methods. This method can also perform better than the others if your queries consist of lists of taxa (for example, an output of *blastn* where several matches can be obtained for each query). Otherwise, we suggest using one of the other methods, as the confidence intervals calculated are less reliable for the weighted mean method.

Implementation

The main steps of all methods are multithreaded on POSIX systems using the packages parallel (R Core Team, 2024) and doParallel (Corporation & Weston, 2022).

The packages ncbitax (Machne, 2022) and CHNOSZ (Dick, 2019) are used to read the taxonomy data, dplyr (Wickham et al., 2023) and biomformat (McMurdie & Paulson, 2024) are used for some of the formatting, and pbapply (Solymos & Zawadzki, 2023) is used to display the progress bar.

The R package accepts as input formats the common 'taxonomy table' format used by popular packages such as phyloseq (McMurdie & Holmes, 2013) and mothur (Schloss et al., 2009), and any file or data frame with a column containing either NCBI taxids or taxon names. The output format is a data frame with the same columns as the input, with some added columns providing information about the estimation and the quality of the estimation. The user can also choose a simple output format only containing the estimation information.

Several plotting functions using the ggplot2 (Wickham, 2016) and ggtree (Yu, 2020) packages are also provided to visualise the results.

Method comparison

The applicability of each method varies. The Bayesian method outputs results for any taxon that is recognised in the NCBI taxonomy. The frequentist random effects model method only outputs results for queries that have a match at the species, genus, or family level. The weighted mean method only performs an estimation for queries that have at least two matches at the species, genus, family, or order level. Below is a comparison of estimates for an example set of bacteria and fungi queries where the highest level of match with the database is the family level. Note that there are fewer successful estimations with the weighted mean method than with the two model-based methods.

Figures below show that estimates and the width of confidence intervals differ between methods (figures Figure 1, Figure 2 and Figure 3).

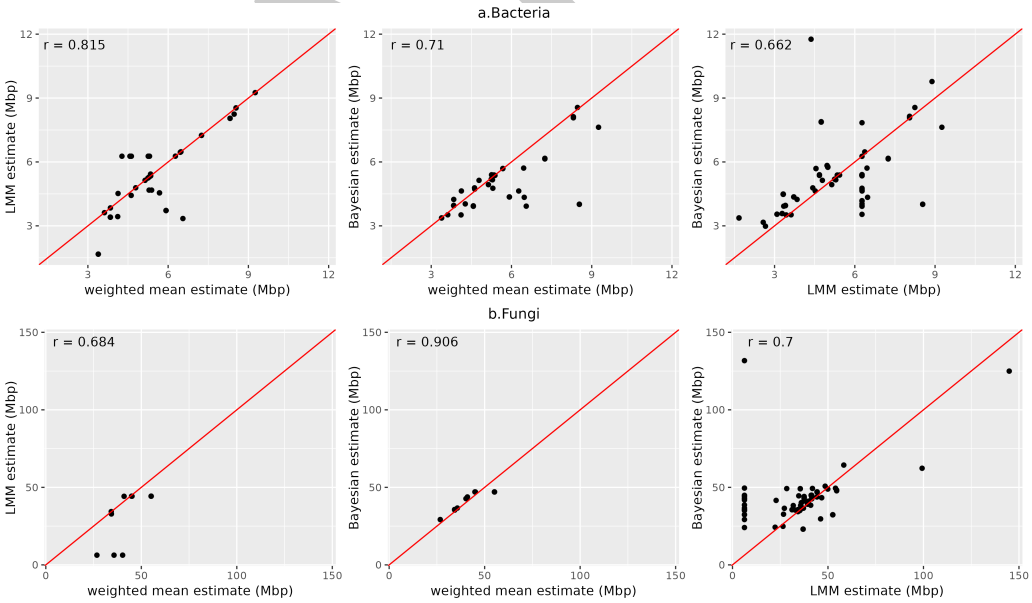

Figure 1: Pairwise comparison of estimates from different methods for a. bacteria and b. fungi. Pearson's correlation coefficient is displayed at the top left.

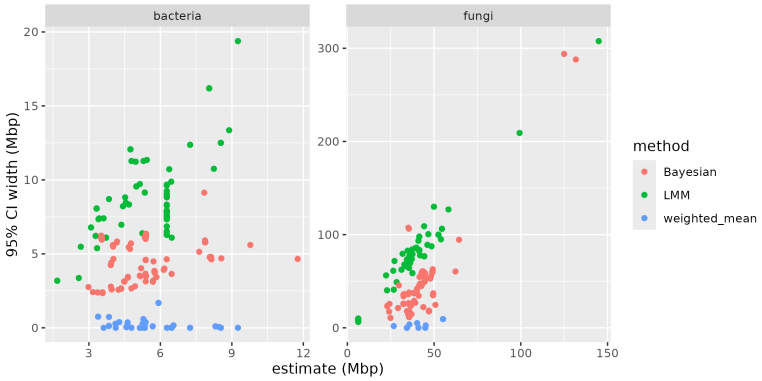

Figure 2: Pairwise comparison of 95% confidence intervals from different methods for a. bacteria and b. fungi.

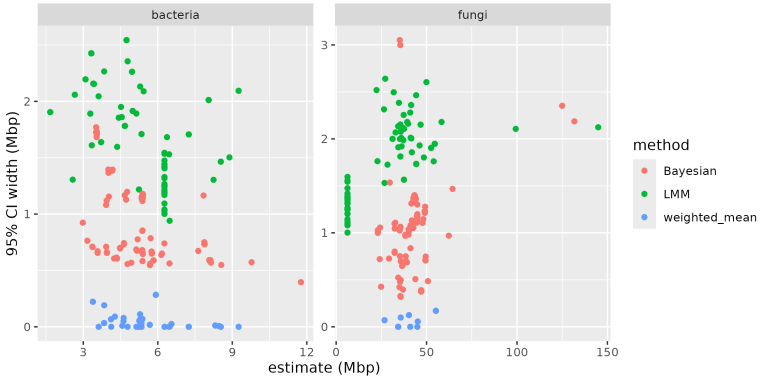

Figure 3: Pairwise comparison of relative 95% confidence intervals (scaled by estimated size) from different methods for a.bacteria and b.fungi.

Example

This example data is a subset of the dataset from Labouyrie et al. (2023).

First, the genome sizes are predicted from the taxa:

```
results = estimate_genome_size(example_input_file, refdata_archive_path,
sep='\t', match_column='TAXID', output_format='input')
```

```
################################################################################
# Genome size estimation summary:
#
#  32.22222 % estimations achieving required precision
#
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  2260662   4850009  15733757  23739845  39882934 161191581

# Estimation status:
Confidence interval to estimated size ratio > ci_threshold  | OK
                                                        122  | 58
```

Then, the results can be visualized using the plotting functions provided. Figure 4 shows a histogram and a boxplot of the estimated genome sizes for each sample. Figure 5 shows a tree

showing the taxonomic relationships as well as the estimated genome sizes.

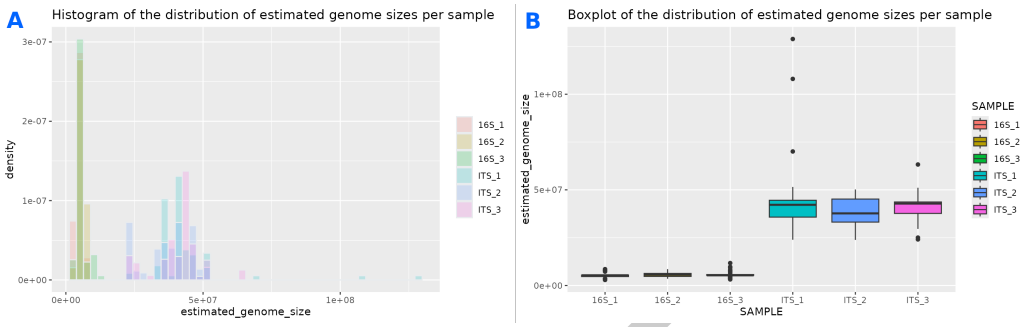

Figure 4: Histogram (A) and boxplot (B) of estimated genome sizes for each sample

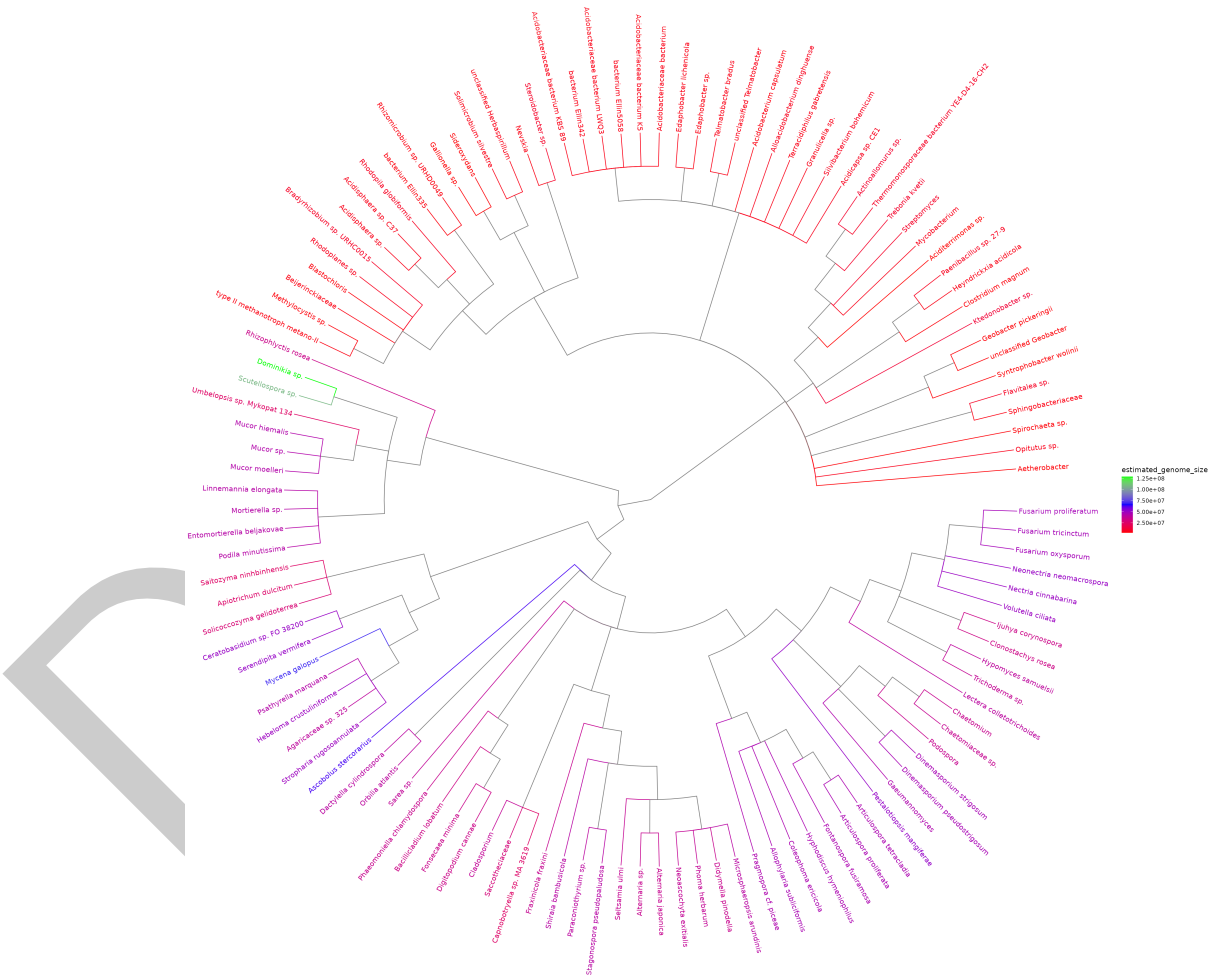

Figure 5: Tree representing taxonomic relationships and estimated genome sizes between queries

Availability

Project name: genomesizeR Project home page: https://github.com/ScionResearch/genomesizeR Operating system(s): Platform independent Programming language: R License: GNU General Public License

Acknowledgements

We acknowledge contributions from Sean Husheer. The authors declare that they have no conflict of interest. Funding for this research came from the Tree-Root-Microbiome programme, which is funded by MBIE's Endeavour Fund and in part by the New Zealand Forest Growers Levy Trust (C04X2002). We make no warranties regarding the accuracy or integrity of the Data. We accept no liability for any direct, indirect, special, consequential or other losses or damages of whatsoever kind arising out of access to, or the use of the Data. We are in no way to be held responsible for the use that you put the Data to. You rely on the Data entirely at your own risk.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brader, G., Compant, S., Mitter, B., Trognitz, F., & Sessitsch, A. (2014). Metabolic potential of endophytic bacteria. *Current Opinion in Biotechnology*, *27*, 30–37.

Corporation, M., & Weston, S. (2022). *doParallel: Foreach parallel adaptor for the 'parallel' package*. https://CRAN.R-project.org/package=doParallel

Dick, J. M. (2019). CHNOSZ: Thermodynamic calculations and diagrams for geochemistry. *Front. Earth Sci.*, *7*.

Knowles, J. E., & Frederick, C. (2024). *merTools: Tools for analyzing mixed effect regression models*. https://github.com/jknowles/mertools

Labouyrie, M., Ballabio, C., Romero, F., Panagos, P., Jones, A., Schmid, M. W., Mikryukov, V., Dulya, O., Tedersoo, L., Bahram, M., Lugato, E., Heijden, M. G. A. van der, & Orgiazzi, A. (2023). Patterns in soil microbial diversity across europe. *Nat. Commun.*, *14*(1), 3311.

Liu, H., Zhang, H., Powell, J., Delgado-Baquerizo, M., Wang, J., & Singh, B. (2023). Warmer and drier ecosystems select for smaller bacterial genomes in global soils. *iMeta*, *2*(1), e70.

Machne, R. (2022). *Ncbitax: Simple base r parser for NCBI taxonomy*. https://github.com/raim/ncbitax

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4), e61217. http://dx.plos.org/10.1371/journal.pone.0061217

McMurdie, P. J., & Paulson, J. N. (2024). *Biomformat: An interface package for the BIOM file format*. https://doi.org/10.18129/B9.bioc.biomformat

Moran, N. A. (2002). Microbial minimalism: Genome reduction in bacterial pathogens. *Cell*, *108*(5), 583–586.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K.

D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, *44*(D1), D733–45.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raes, J., Korbel, J. O., Lercher, M. J., Von Mering, C., & Bork, P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biology*, *8*, 1–11.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Solymos, P., & Zawadzki, Z. (2023). *Pbapply: Adding progress bar to '*apply' functions*. https://CRAN.R-project.org/package=pbapply

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43.

Vandenkoornhuyse, P., Mahé, S., Ineson, P., Staddon, P., Ostle, N., Cliquet, J.-B., Francez, A.-J., Fitter, A. H., & Young, J. P. W. (2007). Active root-inhabiting microbes identified by rapid incorporation of plant-derived carbon into RNA. *Proceedings of the National Academy of Sciences*, *104*(43), 16970–16975.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org

Yu, G. (2020). Using ggtree to visualize data on Tree-Like structures. *Curr. Protoc. Bioinformatics*, *69*(1). https://doi.org/10.1002/cpbi.96