genomesizeR: An R package for genome size prediction

Celine Mercier [1], **Joane Elleouet** [1], **Sean Husheer** [1,2], **Loretta Garrett** [1], **and Steve A Wakelin** [1]

1 Scion Research, PO Box 3020, Rotorua 3010, New Zealand **2** Havelock North Research Centre, The New Zealand Institute for Plant and Food Research Limited, Havelock North, New Zealand

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. Several plotting functions are also provided to visualise the results.

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

42 trait in microbial communities to provide insights spanning niche size, co-evolution, adaption,
43 and metabolic flexibility of the microbiomes present, but also stability, and ecophysiological
44 and functional complexity of abiotic and biotic environments.

45 However, characterizing genome size for all organisms in a microbiome remains challenging.
46 The exponentially growing genome databases are an inexpensive resource unlocking a myriad
47 of research opportunities in all fields of environmental sciences, but genome size estimates for
48 many taxa found in environmental samples are missing from public databases, or fully unknown.
49 The evolutionary rule that phylogenetically related organisms share genetic similarities can be
50 exploited, and genome size for taxa with unknown genome size can be statistically inferred
51 from related taxa with known genome size, using taxonomy as a proxy for phylogeny. Another
52 challenge is the precision of identification: some taxa can only be identified at high taxonomic
53 levels. Statistical methods can also be used to infer their genome size range from databases.
54 To our knowledge, there is no convenient and fast way to obtain genome size estimates with
55 uncertainty bounds for all organisms identified or partially identified in an environmental sample.

56 Using the increased prevalence of whole-genome information for all organisms, we have therefore
57 developed genomesizeR, allowing the inference of genome size of many queries at once, based
58 on taxonomic information and available genome data from the NCBI.

Methods

NCBI database filtering and processing

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

65 This raw database is then prepared to include more pre-computed information to be used by
66 the package. Genome sizes are aggregated to the species level by iteratively averaging all
67 entries at lower ranks, and estimated standard errors for each species with multiple genome size
68 entries are stored. The package can therefore only provide estimates at the level of species and
69 above. Average genome sizes and their associated standard error values are also pre-computed,
70 to be used by the weighted mean method.

Bayesian method

72 The NCBI database of species with known genome sizes was split by superkingdom (Bacteria,
73 Archeae, Eukaryotes). A distributional Bayesian linear hierarchical model using the brm
74 function from the brms package (Bürkner, 2017) was fitted to each superkingdom dataset. The
75 general model structure is outlined below and corresponds exactly to the most complex model,
76 implemented for the Bacteria superkingdom. This general model was simplified by dropping
77 the class group effect in the standard deviation model for the Eukaryote superkingdom, and
78 dropping both the class and phylum group effect in the standard deviation model for the
79 Archeae superkingdom. The latter is therefore not addressed using a distributional model, as
80 the response variance has no predictor.

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

81 where G_i is the genome size of species i in the units of 10 Mbp. The model uses predictors
82 for both mean and standard deviation. The model is described in more detail in the package
83 vignettes.

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

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

88 Queries corresponding to identified species with an available genome size estimate in the NCBI
89 database get allocated the genome size value of the database (averaged at the species level)
90 and 95% confidence intervals are calculated based on the standard error of the mean of all
91 genome sizes available for that species in the processed NCBI database.

Frequentist method

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

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

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

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

102 Queries corresponding to identified species with an available genome size estimate in the NCBI
103 database get allocated the genome size value of the database (averaged at the species level)
104 and 95% confidence intervals are calculated based on the standard error of the mean of all
105 genome sizes available for that species in the processed NCBI database.

Weighted mean method

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

111 The pseudocode describing the algorithm for the estimate and confidence interval computation
112 is available in the package vignettes.

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

Implementation

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

122 The packages ncbitax (Machne, 2022) and CHNOSZ (Dick, 2019) are used to read the
123 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.

Method validation and comparison

Each method was quality-controlled by splitting the species-level database into a training set and a validation set. Sampling the validation set involved a level of stratification to obtain a sample representing all match ranks (rank of the closest taxon represented in the training set). Estimation error and confidence bounds were assessed for each method on the same validation set.

Not all queries are estimable by all methods. As stated above, the identification of taxa related to a query in the weighted mean method is limited to ranks below and including order, and only performs an estimation for queries that have at least two such matches. Valid estimations from the frequentist LMM method are limited to match ranks below and including family. Only the Bayesian method can output results for any taxon that is recognised in the NCBI taxonomy.

Figure 1 shows summaries of estimation error relative to the size of the genome, and 95% confidence interval widths relative to the size of the genome, for all successful estimations from the validation set, per method, match rank, and superkingdom. The magnitude and variability of estimation error increases as match rank increases, for all methods. Relative errors seem similar across methods for low match ranks (genus and family), with a slightly better performance of the weighted mean method for match ranks at the genus levels in the Eukaryote superkingdom. Confidence interval widths also increases as match rank increases for the model-based methods, while the weighted mean method has narrow confidence intervals across match ranks. Although the Bayesian method seems to have narrower confidence bands than the frequentist method for match ranks at the genus level (especially for Eukaryotes), it has the widest intervals for queries at a higher match rank.

To assess the adequacy of estimated confidence intervals in each method, the number of true genome sizes contained between the estimated confidence bounds across the validation set was added up across two groups of low match ranks (genus and family) and high match ranks (above family). We would expect an adequate model to have around 95% of true values falling within 95% confidence intervals. The Bayesian method performs as expected across the two groups (Table 1). The frequentist LMM method only has a coverage of 83% over the estimable group, suggesting that 95% confidence intervals might not always be fully reliable. The weighted mean method does not perform well in confidence interval estimation.
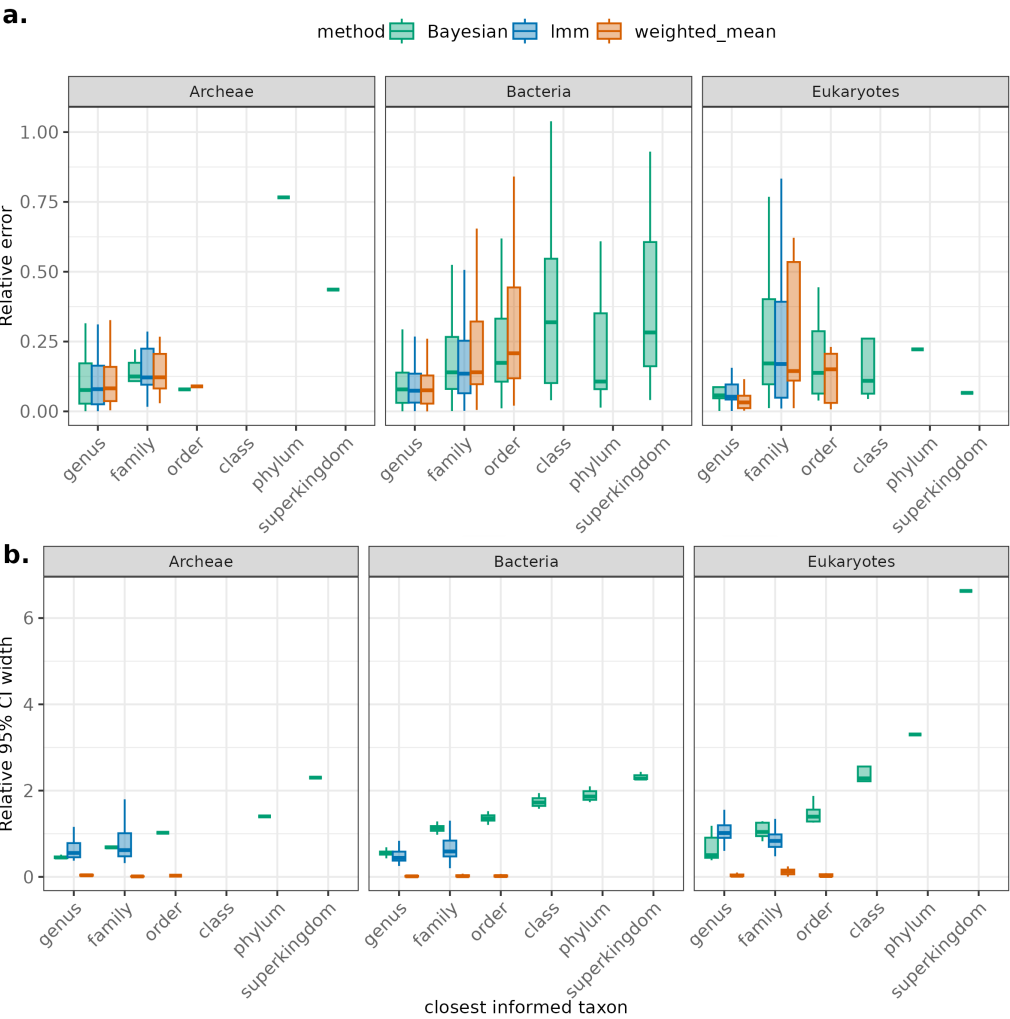
a.

b.

Figure 1: Boxplot of (a) estimation error (b) 95% confidence interval widths, for all successful estimations from the validation set. In both cases values are relative to the size of the genome.

Table 1: Observed coverage of 95% confidence intervals for each method and two different rank groups, obtained from the validation process.

Method	Match rank genus or family	Match rank above family
Bayesian	96/99 (97%)	326/340 (96%)
Frequentist LMM	-	283/340 (83%)
Weighted mean	1/60 (2%)	27/337 (8%)

The strengths and limitations of each method are outlined in Table 2. We emphasise that the weighted mean method is only suitable for taxa that are well-characterised at low taxonomic levels, and when uncertainty bounds are of minor interest. The major advantage of the method is that it can be used on queries with several potential taxonomic matches. The Bayesian method is the most reliable method especially for quantifying uncertainty around estimated means, and obtaining estimates for taxa that are not well represented at low ranks in the NCBI database.

Table 2: Comparison of method behaviour and applicability

	CI estimation	Model information	Behaviour with well-studied organisms	Query is a list of several taxa	Minimum number of references needed for estimation
Bayesian	very reliable	any rank	+	+	1
LMM	mostly reliable	up to family level	+	+	1
Weighted mean	unreliable	up to order level	++	++	2

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

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.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*, 1–28.

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

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/

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