The Journal of Open Source Software

genomesizeR: An R package for genome size prediction

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

1 Bioeconomy Science Institute, Rotorua, New Zealand

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Published: unpublished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 Archaea, Bacteria, or Eukaryote organisms identified at any taxonomic level.

This R package 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.

`genomesizeR` retrieves the taxonomic classification of input queries, estimates the genome size of each query, and provides 95% confidence intervals for each estimate. Some plotting functions are also provided to visualise the results.

Statement of need

The size of genomes and their evolution can provide important insights into evolutionary and ecological processes influencing both species and the environments 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). Among many others, these findings demonstrate the opportunities associated with including genome size as a key trait in studies on communities to provide insights into ecological and evolutionary processes.

However, characterizing genome size remains challenging. The exponentially growing genome databases are an inexpensive resource unlocking a myriad of research opportunities, but 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 can be statistically inferred by using data from related taxa where this information is available, 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 any organism.

Given the growing availability 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 NCBI.

Methods

NCBI database filtering and processing

The reference database is built by querying all genome metadata information from the curated NCBI RefSeq database (O'Leary et al., 2016). This raw database is then filtered and prepared to include more pre-computed information to be used by the package.

Bayesian method

The reference database of genome sizes was split by superkingdom (Bacteria, Archaea, Eukaryotes). A distributional Bayesian linear hierarchical model using the `brm` function from the `brms` package (Bürkner, 2021) was fitted to each superkingdom dataset. The general model structure is outlined below and corresponds exactly to the most complex model, implemented for the Bacteria superkingdom. This general model was simplified by dropping the class group effect in the standard deviation model for the Eukaryote superkingdom, and dropping both the class and phylum group effect in the standard deviation model for the Archaea superkingdom. The Archaea model is therefore not addressed using a distributional model, as the response variance has no predictor. The model is as follows:

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

where G_i is the genome size of species i in units of 10 Mbp. The model uses taxonomic levels as predictors, and is described in more detail in the package vignettes.

The estimation process uses Stan's Hamiltonian Monte Carlo algorithm with the No-U-Turn Sampler (NUTS, Hoffman & Gelman, 2014; Stan Development Team, 2025).

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.

Frequentist method

A frequentist linear mixed-effects model (LMM) 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 effects of genus and family for genus $g[i]$ and family $f[i]$ and e_i is the residual error of observation i.

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.

Method validation and comparison

The strengths and limitations of each method are outlined in Table 1. The weighted mean method is less reliable but 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 1: Comparison of method behaviour and applicability. Legend: "+" = generally suitable / performs well in this case; "++" = particularly suitable / performs very well in this case (relative to the other methods).

	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

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).

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. https://doi.org/10.1016/j.copbio.2013.09.012

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(47), 1593–1623. http://jmlr.org/papers/v15/hoffman14a.html

Moran, N. A. (2002). Microbial minimalism: Genome reduction in bacterial pathogens. *Cell*, *108*(5), 583–586. https://doi.org/10.1016/S0092-8674(02)00665-7

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. https://doi.org/10.1093/nar/gkv1189

Stan Development Team. (2025). *Stan reference manual, version 2.37.0*. https://mc-stan.org

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. https://doi.org/10.1073/pnas.0705902104