genomesizeR: An R package for genome size prediction

14 June 2024

Summary

The purpose of this R package is to implement tools that allow the inference of genome size based on taxonomic information and available genome data from the National Center for Biotechnology Information (NCBI).

This R package offers three different methods for genome size prediction: a Bayesian linear hierarchical model, a frequentist linear mixed-effects model, and a weighted mean method.

The methods use:

- A list of queries; a query being a taxon or a list of several taxa.
- A reference database containing all the known genome sizes, built from the NCBI databases, with associated taxa.
- A taxonomic tree structure as built by the NCBI

`genomesizeR` estimates the genome size of each query, with a confidence interval on the estimation.

Statement of need

Ask Steve?

Methods

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.

$$log(G_i) \sim \mathcal{N}(\mu_i, \sigma_i^2) \tag{1}$$

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

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

with priors

$$\alpha_0 \sim \mathcal{N}(0, 5)(\sigma_{genus}, \sigma_{family}, \sigma_{order}, \sigma_{class}, \sigma_{phylum}, s_{class}, s_{phylum}) \sim \mathcal{N}^+(0, 1) \tag{3}$$

The standard deviation is modelled as follows:

$$log(\sigma_i) = \lambda_0 + \lambda_{class_{c[i]}} + \lambda_{phylum_{p[i]}} \lambda_{class_{c[i]}} \sim \mathcal{N}(0, s^2_{class}) \lambda_{phylum_{p[i]}} \sim \mathcal{N}(0, s^2_{phylum}) \tag{4}$$

with priors

$$\lambda_0 \sim \mathcal{N}(0, 1)(s_{class}, s_{phylum}) \sim \mathcal{N}^+(0, 1) \tag{5}$$

$\mathcal{N}^+$ 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 genome, family, order, class, and phylum of observation i. 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 credible intervals are obtained using quantiles from the posterior distribution.

Note that the model was fitted at the species level. The base dataset for the model used a version of the NCBI database where all entries below the species level were iteratively averaged, starting from the smallest level, until one value was obtained per species.

Frequentist method

A frequentist linear mixed-effects model using the lmer function from the lme4 package was fitted to the NCBI database of species with known genome sizes. The model is as follows:

$$log(G_i) = \alpha_0 + \alpha_{genus_{g[i]}} + \alpha_{family_{f[i]}} + e_i \tag{6}$$

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). Models with additional higher nested levels (order, class) were tested but led to convergence failure of the REML algorithm. A prediction interval is computed using the predictInterval function from the merTools package.

As for the Bayesian method, the model was fitted at the species level. The base dataset for the model used a version of the NCBI database where all entries below the species level were iteratively averaged, starting from the smallest level, until one value was obtained per species.

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 confidence interval is calculated as:

$$CI = 1.96 \times \sqrt{\hat{\mu}} \tag{7}$$

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

Note that this method allows for estimation below the species level. For queries relating to well-characterised species where many genetic studies have been performed, this might lead to more precise predictions than the two other methods. Otherwise, we suggest using one of the other methods.

Implementation

The main steps of all methods are multithreaded on POSIX systems (not Windows) using the ### package(s) (ref(s)).

The R package accepts the common 'taxonomy table' format used by popular packages such as phyloseq (ref) and mothur (ref), and any file or dataframe with a colum containing either NCBI taxids or taxon names as input formats. 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 package (ref) are also provided to visualise the results.

Comparison of methods

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 means 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 means 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).

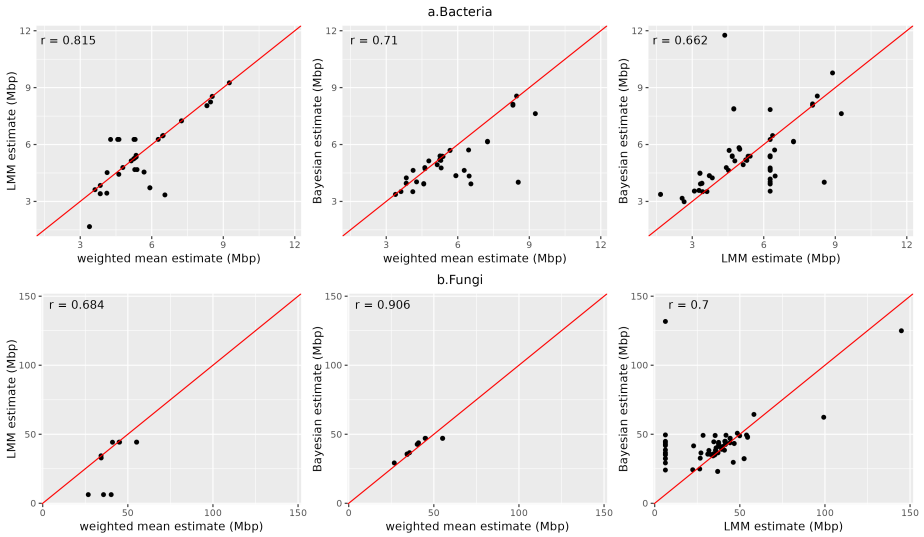

Figure 1: Pairwise comparison of estimates from different methods for a. bacteria and b. fungi. Pearson's correlation coefficient is displayed at the top left.

Example

This example data is a subset of the dataset from (ref).

First, the genome sizes are predicted from the taxa:

```
results = estimate_genome_size(example_input_file, format='csv', sep='\t', match_column='TA

################################################################################
# Genome size estimation summary:
#
```

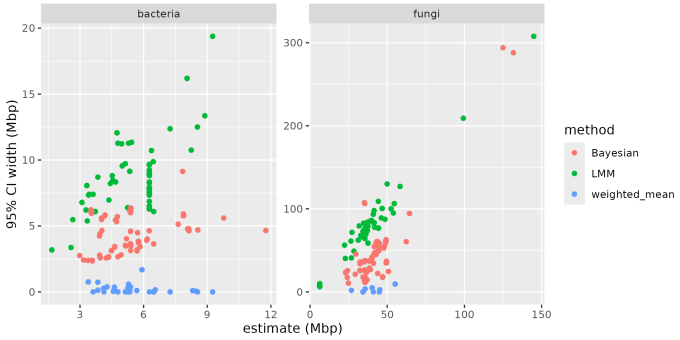

Figure 2: Pairwise comparison of 95% confidence intervals from different methods for a. bacteria and b. fungi.

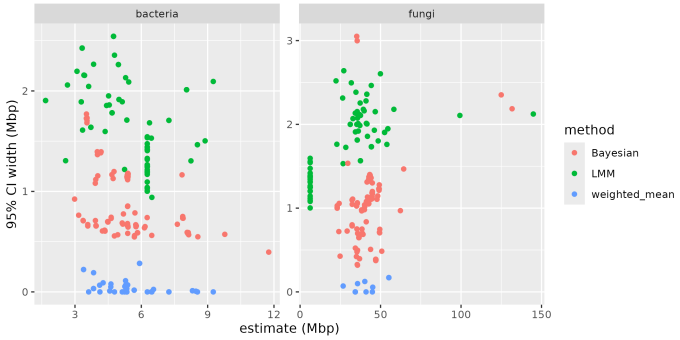

Figure 3: Pairwise comparison of relative 95% confidence intervals (scaled by estimated size) from different methods for a.bacteria and b.fungi.

```
#  8.888889 % estimations achieving required precision
#
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  2893119   5344026  17768591  24126621  42194240 128877284

# Estimation status:
Confidence interval to estimated size ratio > ci_threshold
                                                        164
```

Then, the results can be visualized using the plotting functions provided:

```
plotted_df = plot_genome_size_histogram(results)
```

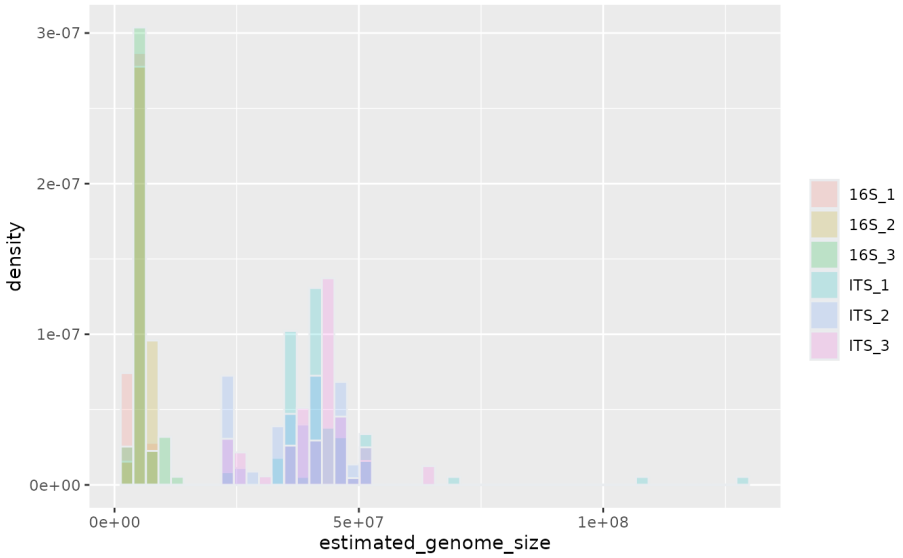

Figure 4: Histogram of estimated genome sizes for each sample

Citations

Citations to entries in paper.bib should be in rMarkdown format.

If you want to cite a software repository URL (e.g. something on GitHub without a preferred citation) then you can do it with the example BibTeX entry below for @fidgit.

For a quick reference, the following citation commands can be used: - `@author:2001` -> "Author et al. (2001)" - `[@author:2001]` -> "(Author et al., 2001)" - `[@author1:2001; @author2:2001]` -> "(Author1 et al., 2001; Author2 et al., 2002)"

Figures

Figures can be included like this: Caption for example figure. and referenced from text using section .

Figure sizes can be customized by adding an optional second parameter: Caption for example figure.

Acknowledgements

We acknowledge contributions from Sean Husher.

References