Tree root microbiome programme

SCION
FORESTS ▪ PRODUCTS ▪ INNOVATION

# genomesizeR: an R package for predicting the size of any species' genome using taxonomic information

Celine Mercier, Joane Elleouet, Steve A Wakelin

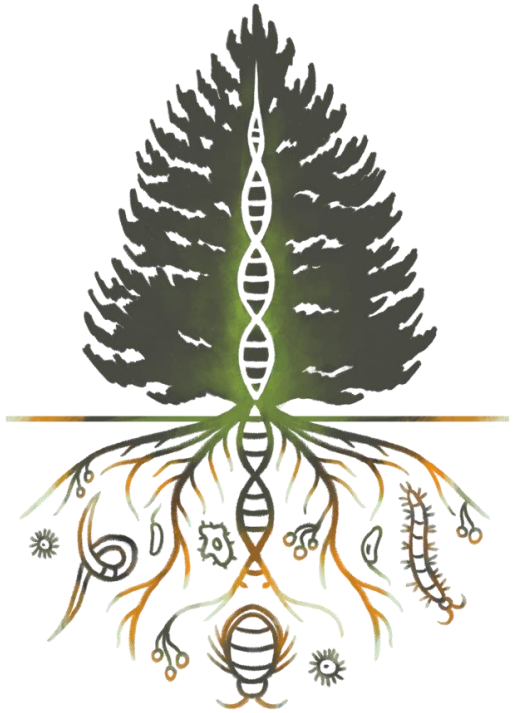
eDNA Conference, February 2025 in Wellington

# Background

How are genome sizes distributed in an eDNA sample?

↓

**Can we estimate the genome size of any (partially) identified organism ?**



Tree root microbiome programme

Tree root microbiome programme | SCION

# Background - general model

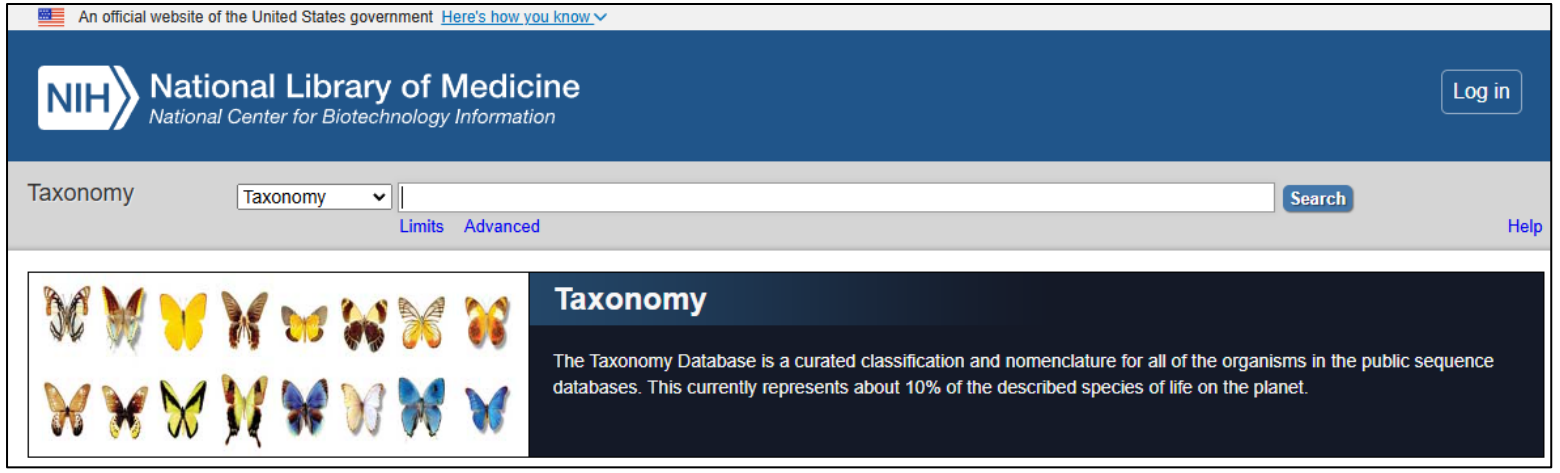
- Model: ***genome size ~ taxonomy***

- Model assumptions:
  - Genome size and phylogeny are correlated
  - Phylogeny and taxonomy are correlated
  - Reference database complete and accurate enough for correct predictions

Tree root microbiome programme | SCION

# Reference data

- NCBI RefSeq: curated collection of genome assemblies
  - 53,016 bacterial spp. references
  - 1,922 eukaryotic spp. references (incl. 596 fungal)
  - 1,146 archean spp. references
- NCBI Taxonomy: taxonomic tree approximating the evolutionary relationships among all the organisms included in GenBank/RefSeq




Tree root microbiome programme | SCION

# Three estimation methods

- Weighted means
- Frequentist reduced hierarchical model (linear mixed model, LMM)
- Bayesian hierarchical model

# Weighted means method

- Weighted means of the nearest taxonomic neighbours (ignores information above order rank)
- Weights based on taxonomic distance
- Quite straight-forward, but underestimates confidence intervals
- Best approach for well-known taxa, or if your queries are lists of several taxa (e.g. blast output)

# Frequentist reduced hierarchical model (LMM)

- Frequentist linear mixed-effects model
- Uses nested genus and family information (known genome sizes)
- Mostly reliable confidence intervals

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

where $\alpha_0$ 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 $i$.

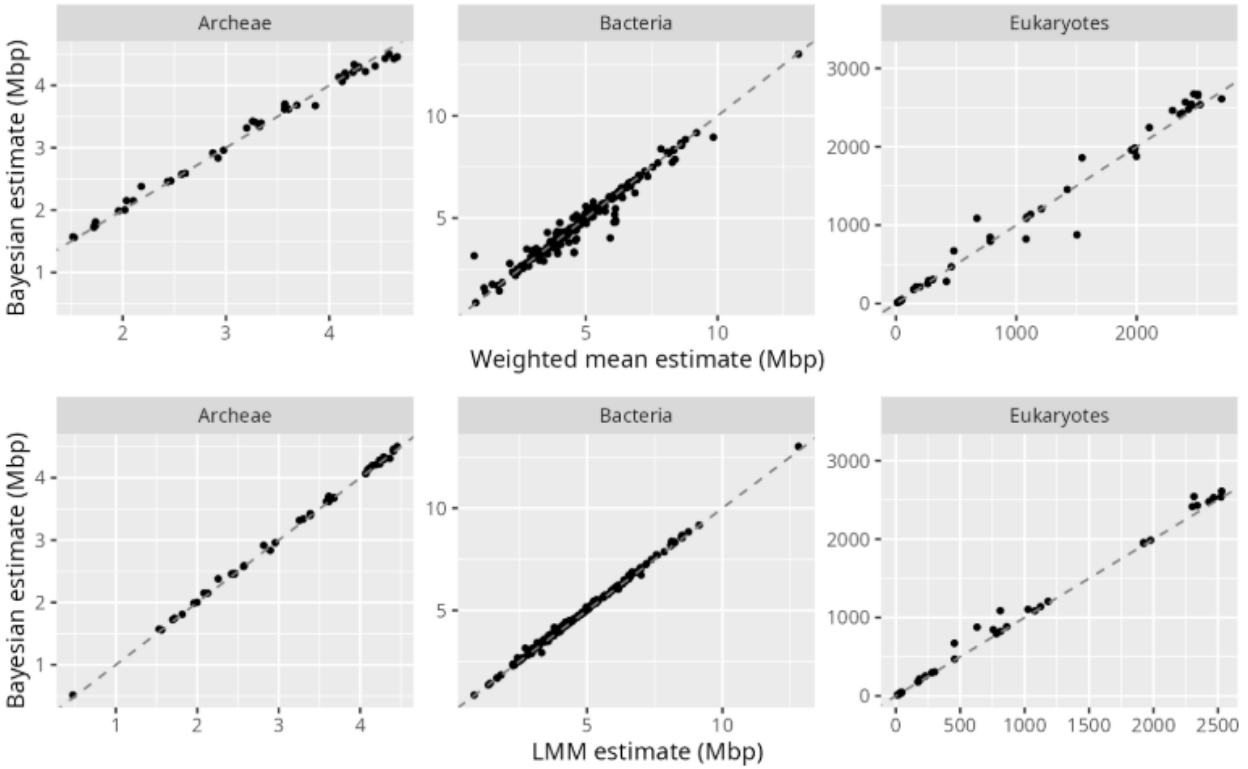
Tree root microbiome programme | SCION

# Bayesian hierarchical model

- Distributional Bayesian linear hierarchical model
  - General model structure: $log\left(G_i\right) \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$
  - Uses brms package, Stan's Hamiltonian Monte Carlo algorithm with the U-turn sampler
- One model per superkingdom
- Uses all known genome sizes at all main taxonomic levels
- Reliable confidence intervals

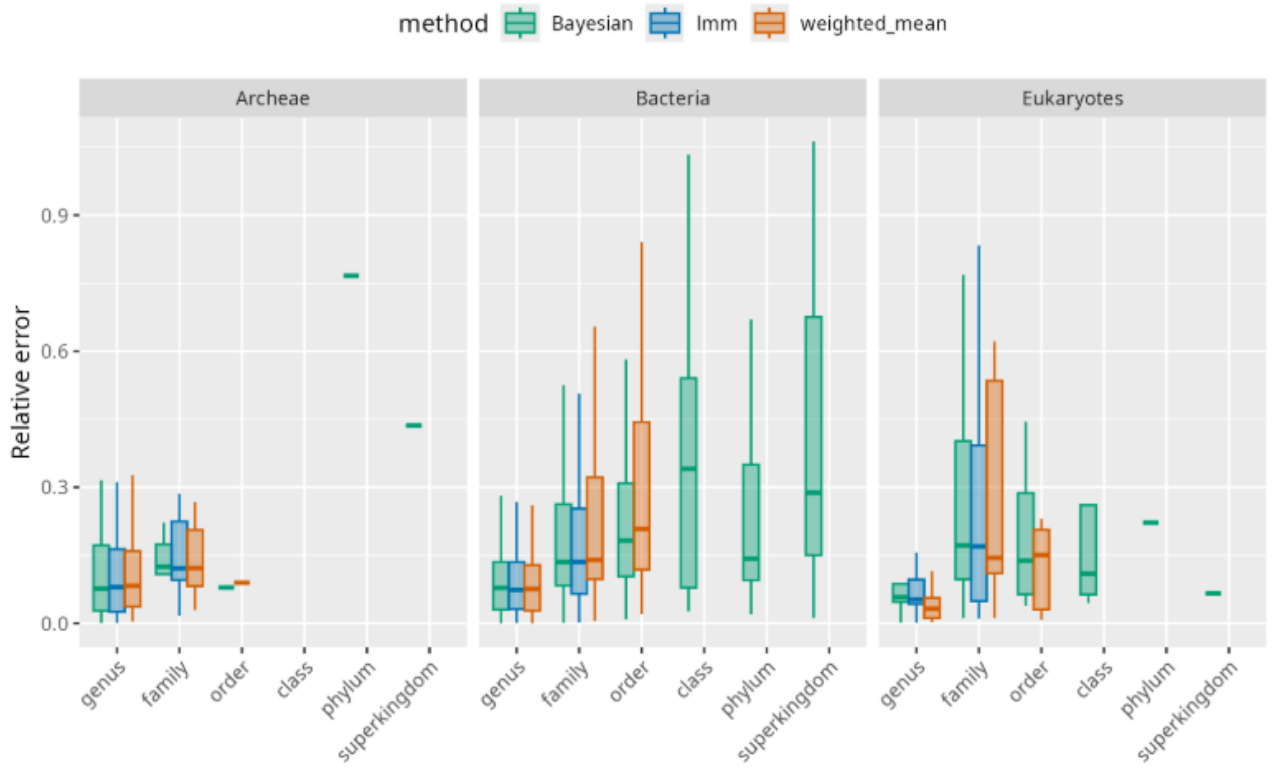
Tree root microbiome programme | SCION

# Estimation comparison between the different methods



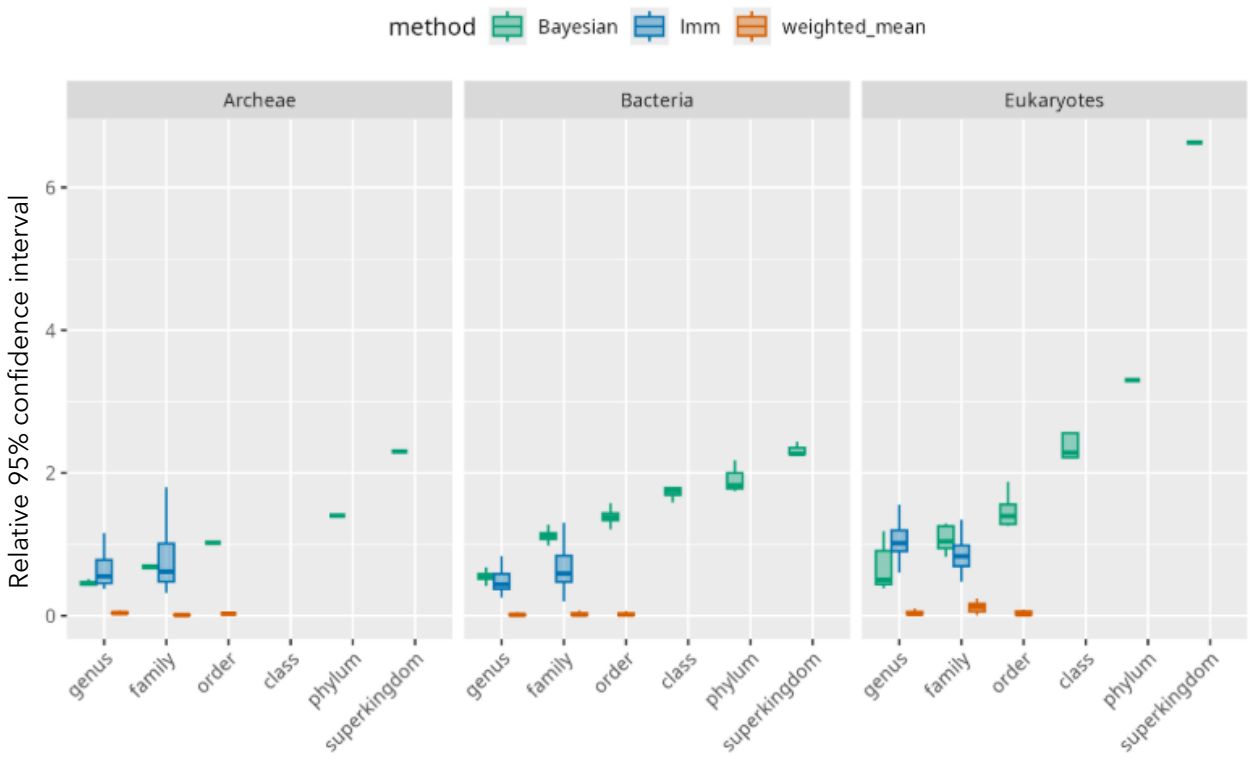

Tree root microbiome programme | SCION

# Relative error with the different estimation methods on validation set




Tree root microbiome programme | SCION

# Confidence intervals with the different estimation methods




| Method | 95%CI coverage for match above family | 95%CI coverage for match at or below family |
|---|---|---|
| Bayesian | 96/99 (97%) | 326/340 (96%) |
| lmm | 0/0 (NaN%) | 283/340 (83%) |
| weighted_mean | 1/60 (2%) | 27/337 (8%) |

Tree root microbiome programme | SCION

# Method comparison summary

| | 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 |

Tree root microbiome programme | SCION

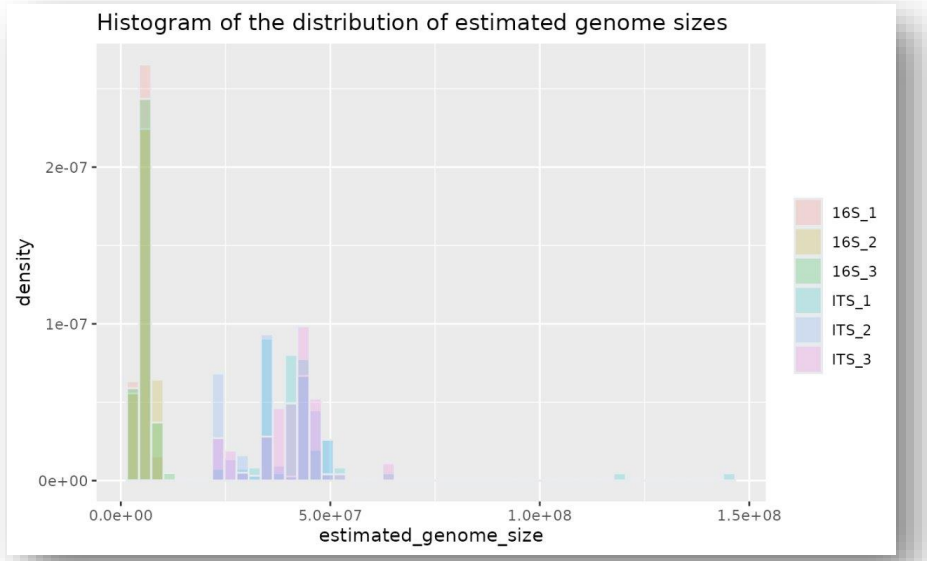
# genomesizeR package

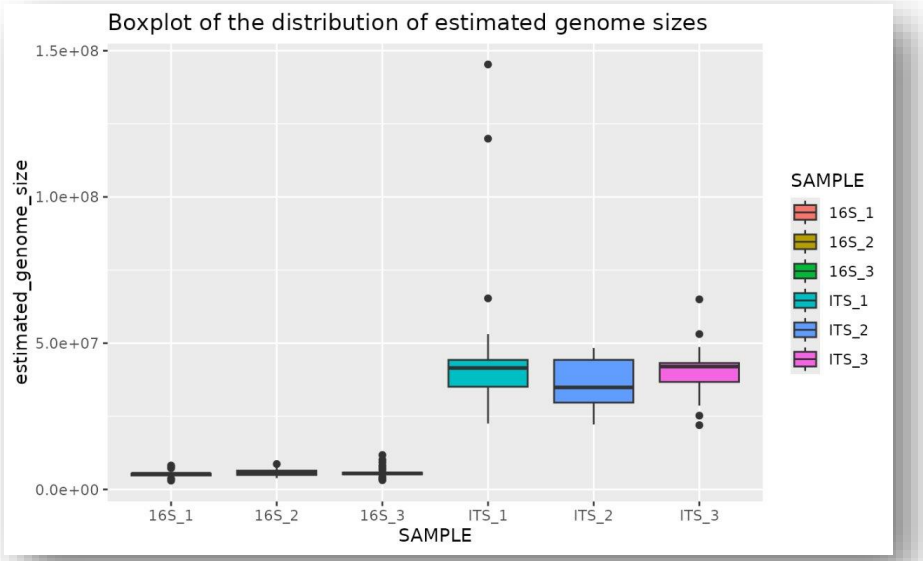
- R package, available on GitHub
- Input format: taxonomy table .csv (phyloseq/mothur format), or any csv-like file or data frame with a column containing either NCBI taxids or taxon names
- Reference and taxonomy databases provided (hosted on zenodo)
- Output format: data frame
- Prediction and visualisation functions

Tree root microbiome programme | SCION

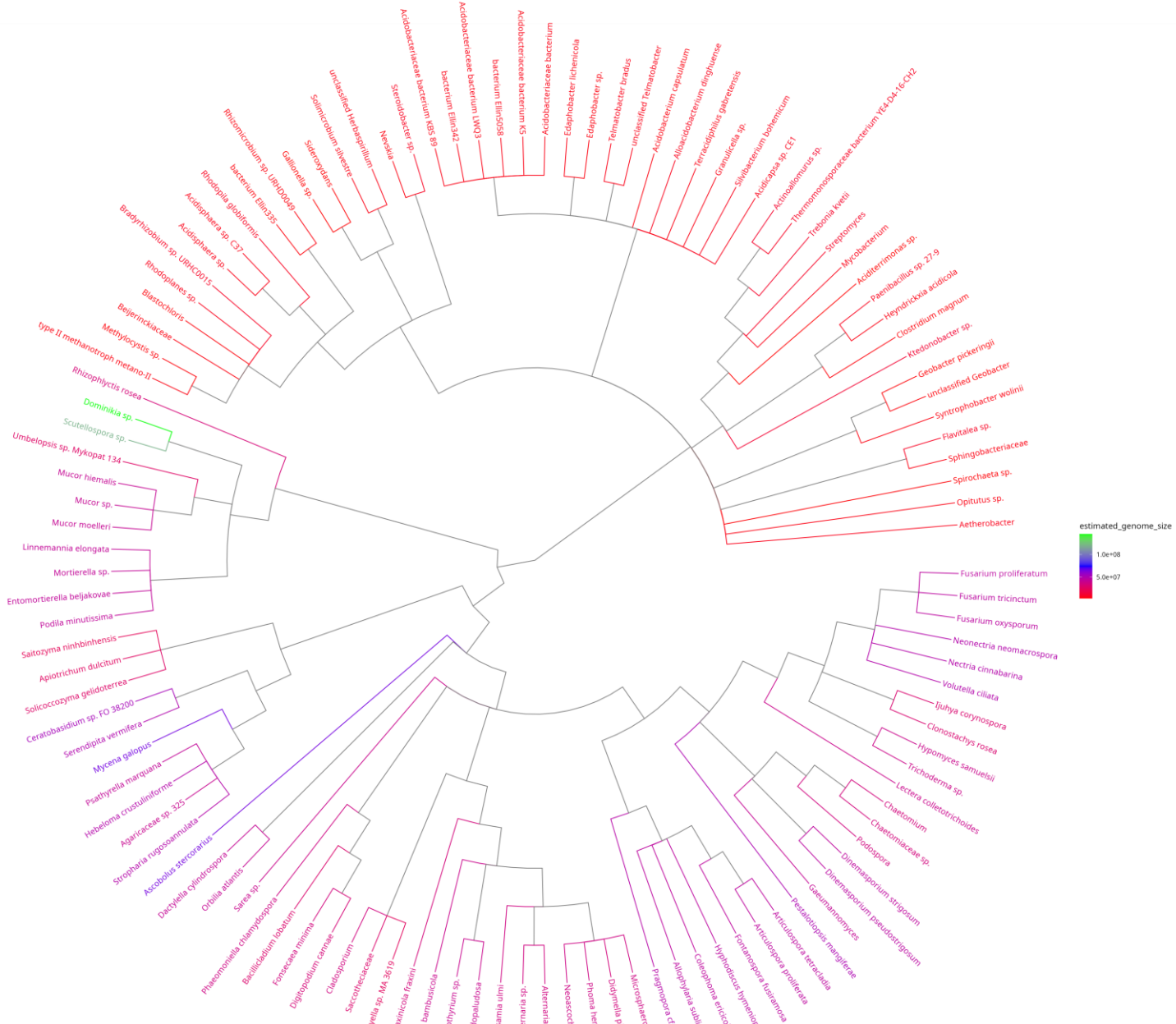
# genomesizeR package: visualisation functions

- Visualisation functions to plot histograms, boxplots and taxonomic trees from the results




Tree root microbiome programme | SCION




Tree root microbiome programme | SCION

# Thank you

https://github.com/ScionResearch/genomesizeR

genomesizeR: An R package for genome size prediction. 2024. Celine Mercier, Joane S. Elleouet, Loretta Garrett, Steve A Wakelin.
*bioRxiv 2024.09.08.611926*; doi: https://doi.org/10.1101/2024.09.08.611926

Acknowledgements: Steve A Wakelin, Loretta Garrett, Sean Husher
Funding: MBIE Endeavour Fund, New Zealand Forest Growers Levy Trust

Celine.Mercier@scionresearch.com

Tree root microbiome programme | SCION

www.scionresearch.com


Prosperity from trees *Mai i te ngahere oranga*

# NCBI data

$$genome\ size \sim taxonomy?$$

- diverse reconstruction levels and methods
- Data for 55,240 species after filtering (2.7%)
  - Bacteria: 10.2%
  - Archaea: 8.5%
  - Eukaryota: 0.1%

- Superkingdom
- Kingdom
- Subkingdom
- Superphylum
- Phylum
- Subphylum
- Infraphylum
- Superclass
- Class
- Subclass
- Infraclass
- Cohort
- Subcohort
- Superorder
- Order
- Suborder
- Infraorder
- Parvorder
- Superfamily
- Family
- Subfamily
- Tribe
- Subtribe
- Genus
- Subgenus
- Section
- Subsection
- Series
- Subseries
- species group
- **species**
- subgroup
- Species
- forma specialis
- Subspecies
- Varietas
- Subvariety
- Forma
- Serogroup
- Serotype
- strain
- isolate

Tree root microbiome programme | SCION

# Hierarchical model

## Predictors

- Completely nested categories
- Common to most organisms
- Data at species resolution

- **Superkingdom**
- Kingdom
- Subkingdom
- Superphylum
- **Phylum**
- Subphylum
- Infraphylum
- Superclass
- **Class**
- Subclass
- Infraclass
- Cohort
- Subcohort
- Superorder
- **Order**
- Suborder
- Infraorder
- Parvorder
- Superfamily
- **Family**
- Subfamily
- Tribe
- Subtribe
- **Genus**
- Subgenus
- Section
- Subsection
- Series
- Subseries
- species group
- **species**
- subgroup
- Species
- forma specialis
- Subspecies
- Varietas
- Subvariety
- Forma
- Serogroup
- Serotype
- strain
- isolate

## Response

- log(size * 10Mbp)

Tree root microbiome programme | SCION

# Hierarchical models

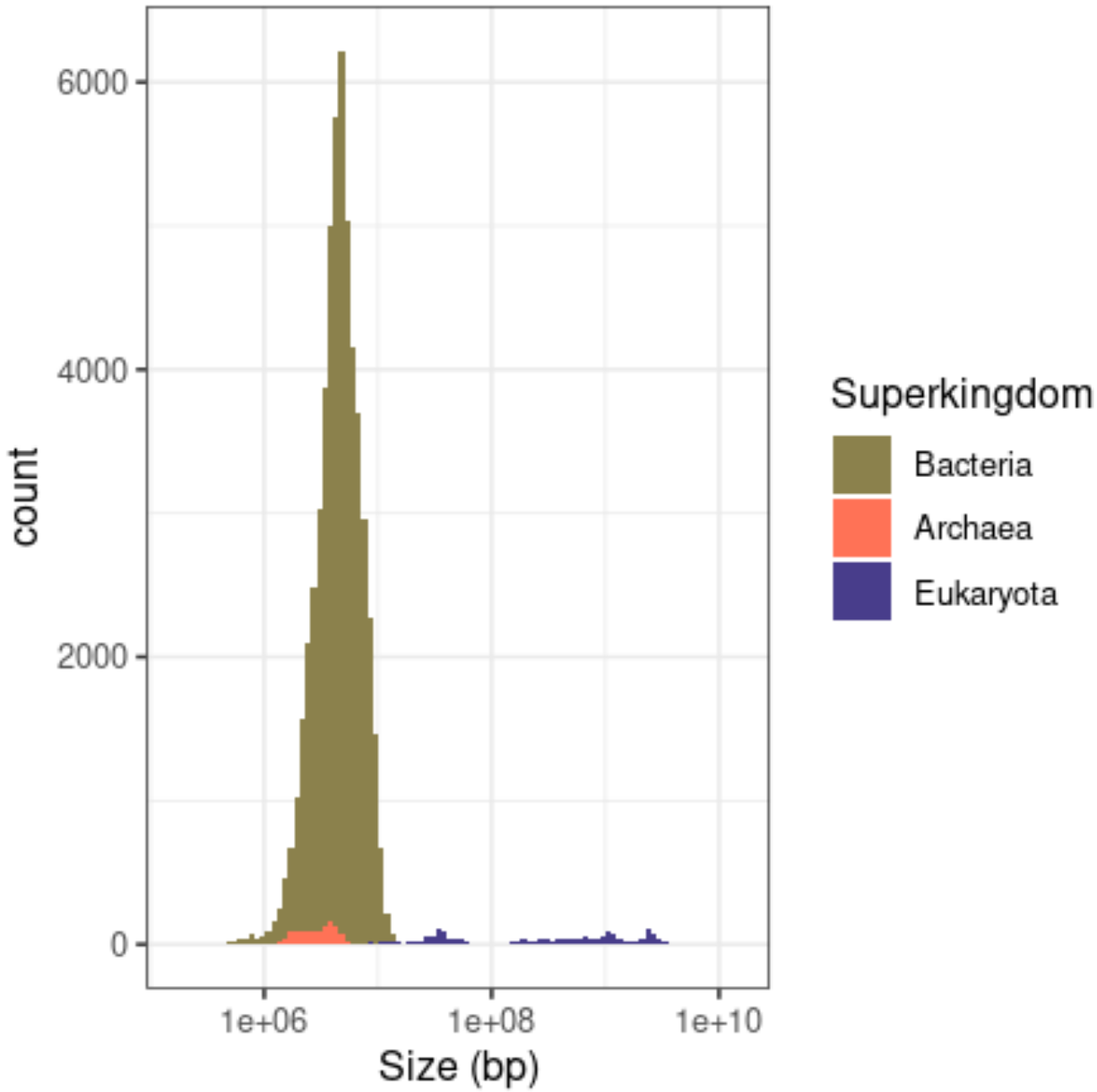
## Predictors

- Phylum
- Class
- Order
- Family
- Genus

## Response

- log(size * 10Mbp)

**1 model per superkingdom**



# Hierarchical models

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

$$\mu_i = \alpha_0 + \alpha_{genus_{g[i]}} + \alpha_{family_{f[i]}} + \alpha_{order_{o[i]}} + \alpha_{class_{c[i]}} + \alpha_{phylum_{p[i]}}$$

$$\alpha_{genus_{g[i]}} \sim \mathcal{N}(0, \sigma^2_{genus})$$

$$\alpha_{family_{f[i]}} \sim \mathcal{N}(0, \sigma^2_{family})$$

$$\alpha_{order_{o[i]}} \sim \mathcal{N}(0, \sigma^2_{order})$$

$$\alpha_{class_{c[i]}} \sim \mathcal{N}(0, \sigma^2_{class})$$

$$\alpha_{phylum_{p[i]}} \sim \mathcal{N}(0, \sigma^2_{phylum})$$

Tree root microbiome programme | SCION

# Hierarchical models


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

$$\mu_i = \alpha_0 + \alpha_{genus_{g[i]}} + \alpha_{family_{f[i]}} + \alpha_{order_{o[i]}} + \alpha_{class_{c[i]}} + \alpha_{phylum_{p[i]}}$$

$$\alpha_{genus_{g[i]}} \sim \mathcal{N}(0, \sigma^2_{genus})$$

$$\alpha_{family_{f[i]}} \sim \mathcal{N}(0, \sigma^2_{family})$$

$$\alpha_{order_{o[i]}} \sim \mathcal{N}(0, \sigma^2_{order})$$

$$\alpha_{class_{c[i]}} \sim \mathcal{N}(0, \sigma^2_{class})$$

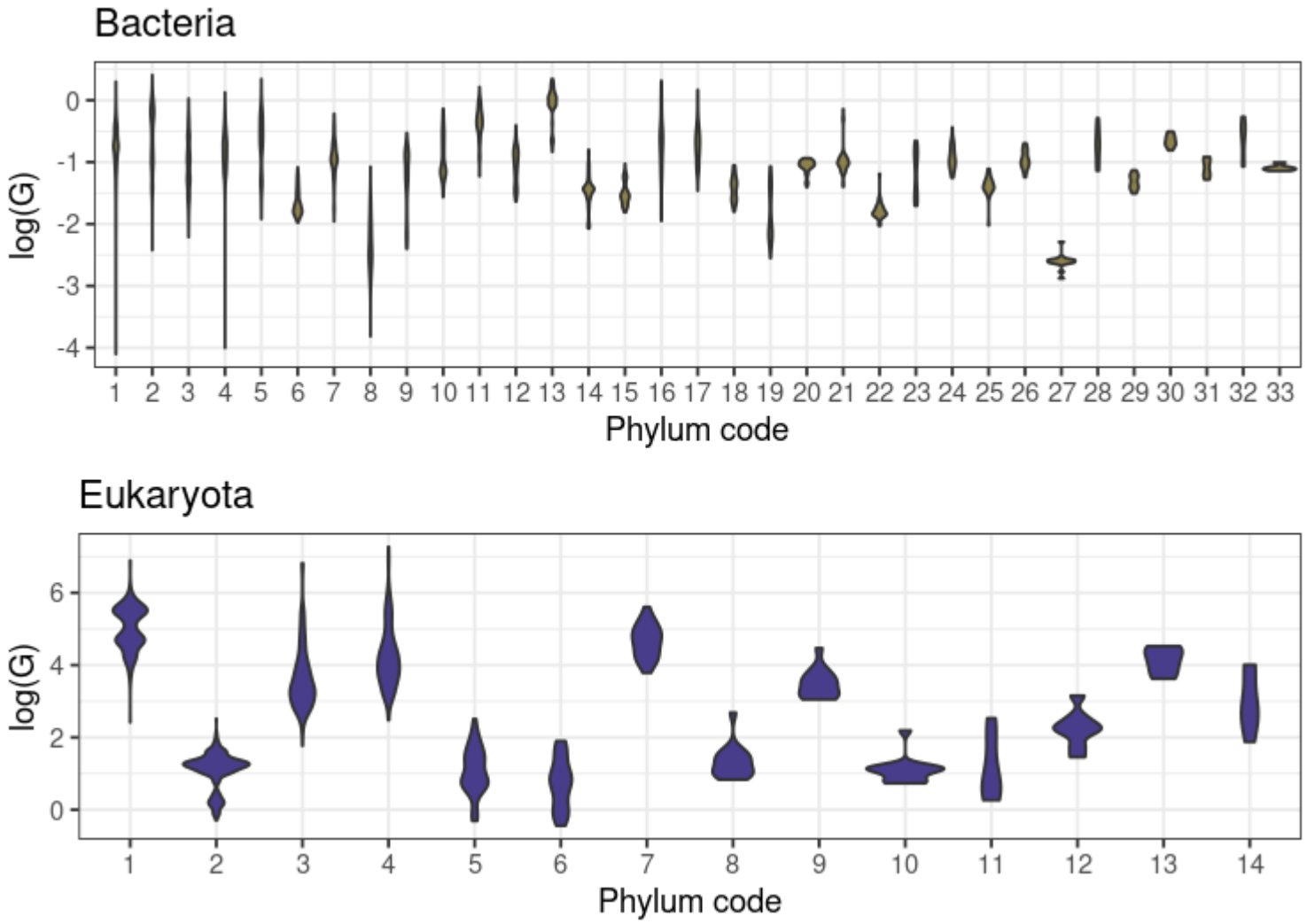
$$\alpha_{phylum_{p[i]}} \sim \mathcal{N}(0, \sigma^2_{phylum})$$

$$\alpha_0 \sim \mathcal{N}(0, 5)$$

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

Tree root microbiome programme | SCION

# Observed variation by phylum




Tree root microbiome programme | SCION

# Hierarchical distributional models

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

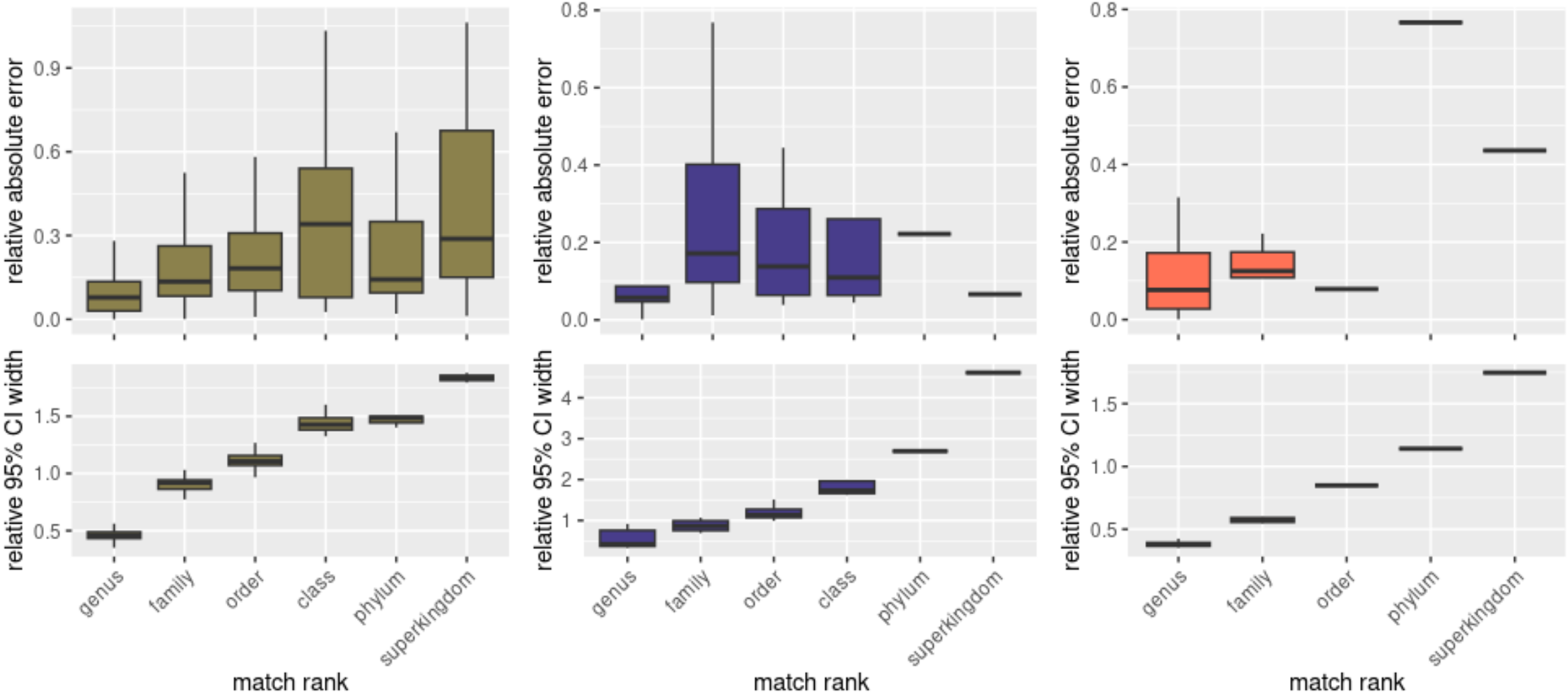
$$log(\sigma_i) = \lambda_0 + \lambda_{class_{c[i]}} + \lambda_{phylum_{p[i]}}$$
$$\lambda_{class_{c[i]}} \sim \mathcal{N}(0, s^2_{class})$$
$$\lambda_{phylum_{p[i]}} \sim \mathcal{N}(0, s^2_{phylum})$$

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

Tree root microbiome programme | SCION

| | Bacteria | Eukaryota | Archaea |
|---|---|---|---|
| Best model (LOOIC) | Distributional $\sigma^2 = f(\lambda_{phylum}, \lambda_{class})$ | Distributional $\sigma^2 = f(\lambda_{phylum})$ | Non-distributional |
| 95%CI cov. | 315/325 (97%) | 63/65 (97%) | 44/49 (90%) |
| 90%CI cov. | 306/325 (94%) | 62/65 (95%) | 43/49 (89%) |

**Validation set results**



# genomesizeR: An R package for genome size prediction

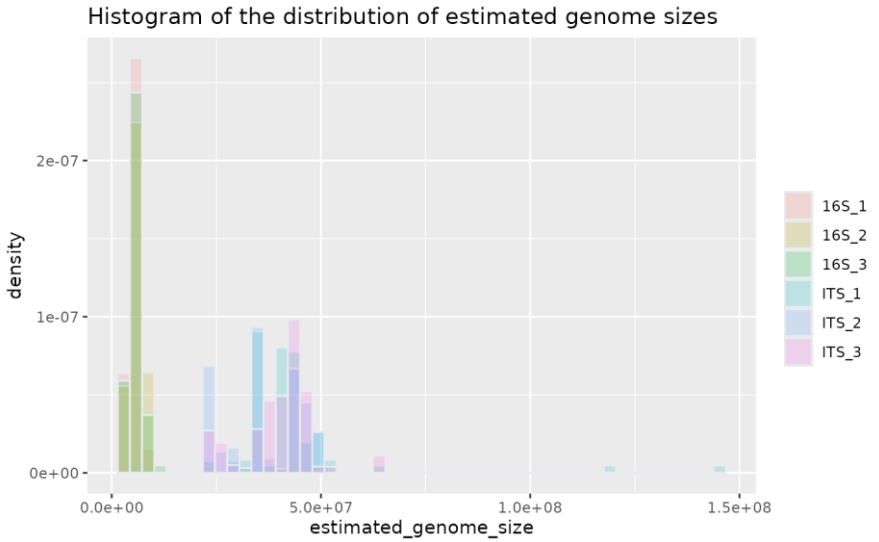
## Input:

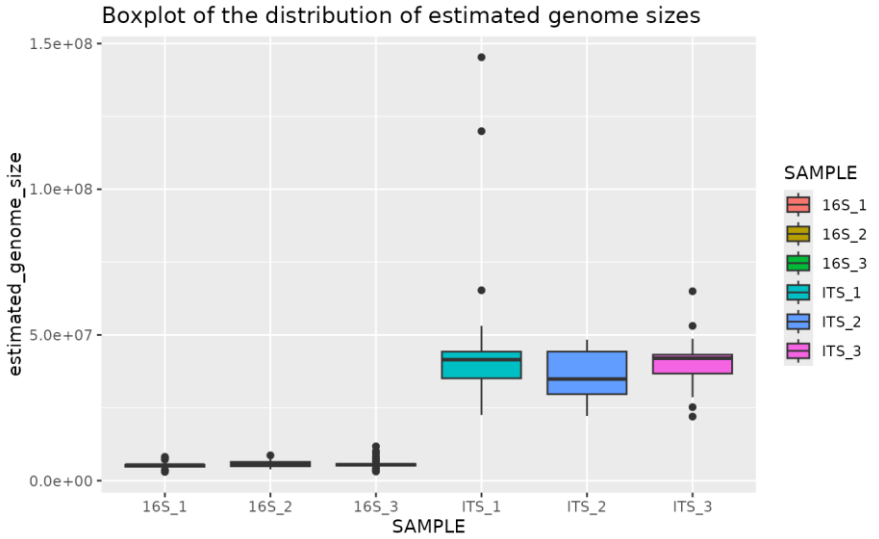
- phyloseq/mothur/… ‘taxonomy table’
- csv with NCBI taxids or taxon names

| COUNT | SAMPLE | SCIENTIFIC_NAME | TAXID |
|---|---|---|---|
| 75 | 16S_2 | Mycobacterium | 1763 |
| 39 | 16S_2 | unclassified Bradyrhizobium | 2631580 |
| 22 | 16S_2 | Acidicapsa sp. CE1 | 1078845 |
| 18 | 16S_2 | Bacillaceae | 186817 |
| 15 | 16S_2 | Silvibacterium bohemicum | 1577686 |

## Output:

Same + output columns; Visualisation functions



# phylogeny as a predictor in statistical models

| | Commonly | This work |
|---|---|---|
| **Objective** | Causal inference | Prediction |
| **Estimand of interest** | Model parameters $\beta$ | Predicted response $\hat{y}$ |
| **subject** | Restricted group of species | All living organisms |
| **Phylogeny** | confound | Main predictor |
| **Data missingness** | low | high |
| **Data** | phylogenetic distance (continuous) | **Taxonomy ?** (discrete) |
| **Modelling method** | Regression – hierarchical modelling with 1 random effect with complex variance-covariance structure | Regression – hierarchical modelling with nested random effects. |

Tree root microbiome programme | SCION

# Validation and prediction

## Validation

Using training and validation sets

## Prediction

- If species in the model database → using the observed species-level mean
- If species not in the model database
- If entry not defined at the species level

Prediction at match levels assuming "new" lower-level groups

Tree root microbiome programme | SCION