GTDB-Tk 2: memory friendly classification with the Genome Taxonomy Database

Pierre-Alain Chaumeil, Aaron J. Mussig, Maria Chuvochina, Christian Rinke, Philip Hugenholtz, Donovan H. Parks

The University of Queensland, St Lucia, QLD, Australia

School of Chemistry and Molecular Biosciences,Australian Centre for Ecogenomics

What is GTDB-Tk?

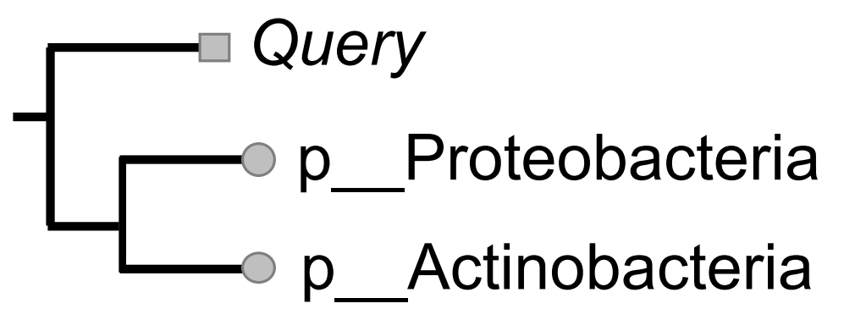
The Genome Taxonomy Database Toolkit (GTDB-Tk):

- Provides automated and objective taxonomic classification of bacterial and archaeal genomes.
- Places genomes into domain-specific, concatenated protein reference trees.
- Used to assign taxonomic classifications to tens of thousands of bacterial and archaeal metagenome-assembled genomes (MAGs) recovered from environmental and human-associated samples (*Chaumeil et al., 2019; Almeida et al., 2021; Nayfach et al., 2021; Chen et al., 2021*).

Why do we need GTDB-Tk v2?

- GTDB-Tk places genomes into GTDB reference trees using the maximum-likelihood placement tool pplacer (Matsen et al. 2010).
- When using the GTDB R07-RS207 bacterial reference tree comprised of 62,291 genomes, pplacer requires **~320 GB** of RAM.
- GTDB-Tk v2 **reduces memory requirements** by dividing the GTDB bacterial reference tree into class-level subtrees.

How does GTDB-Tk classify my genomes?

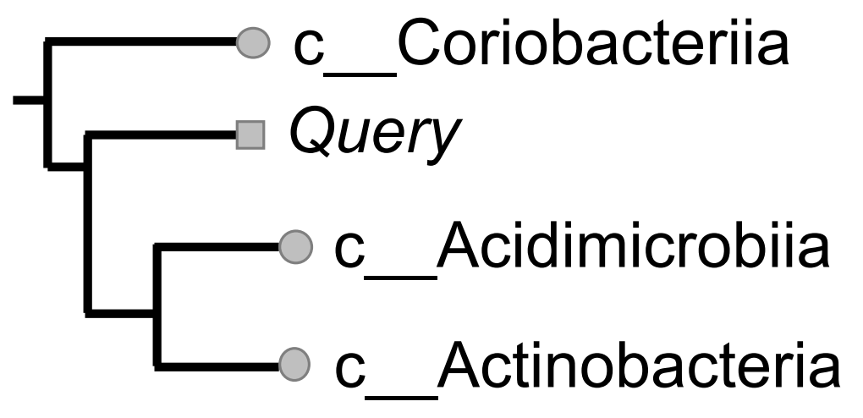


Query

p__Proteobacteria

p__Actinobacteria

Query genome represents a new phylum.



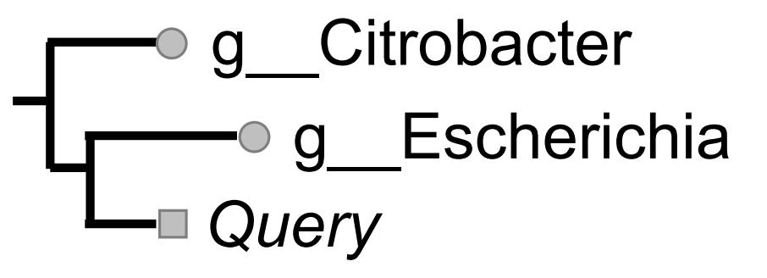
c__Coriobacteriia

Query

c__Acidimicrobiia

c__Actinobacteria

Query genome represents a novel class within the phylum *Actinobacteria*.

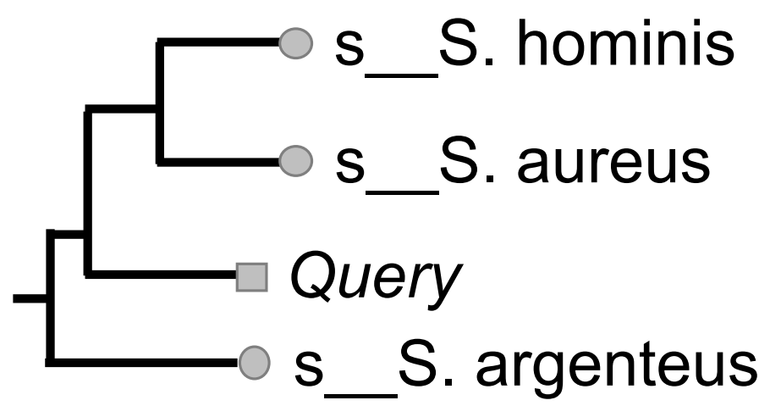


g__Citrobacter

g__Escherichia

Query

Query genome will be classified as either a novel basal *Escherichia* species or a novel genus in the family *Enterobacteriaceae* depending on its RED value.



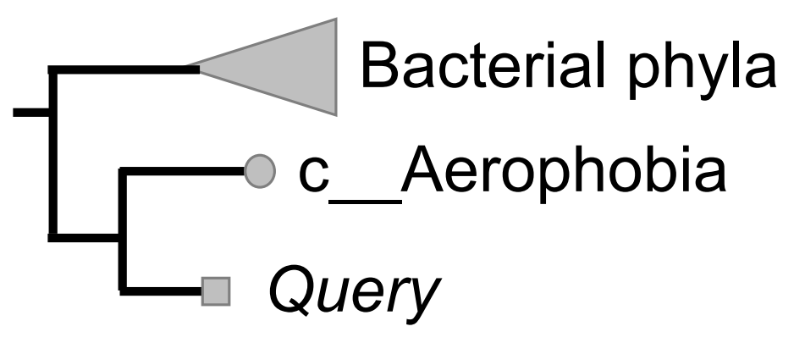
s__S. hominis

s__S. aureus

Query

s__S. argenteus

The query genome is assigned to the closest *Staphylococcus* species if the ANI is above the species ANI circumscription radius or is otherwise classified as a novel species.

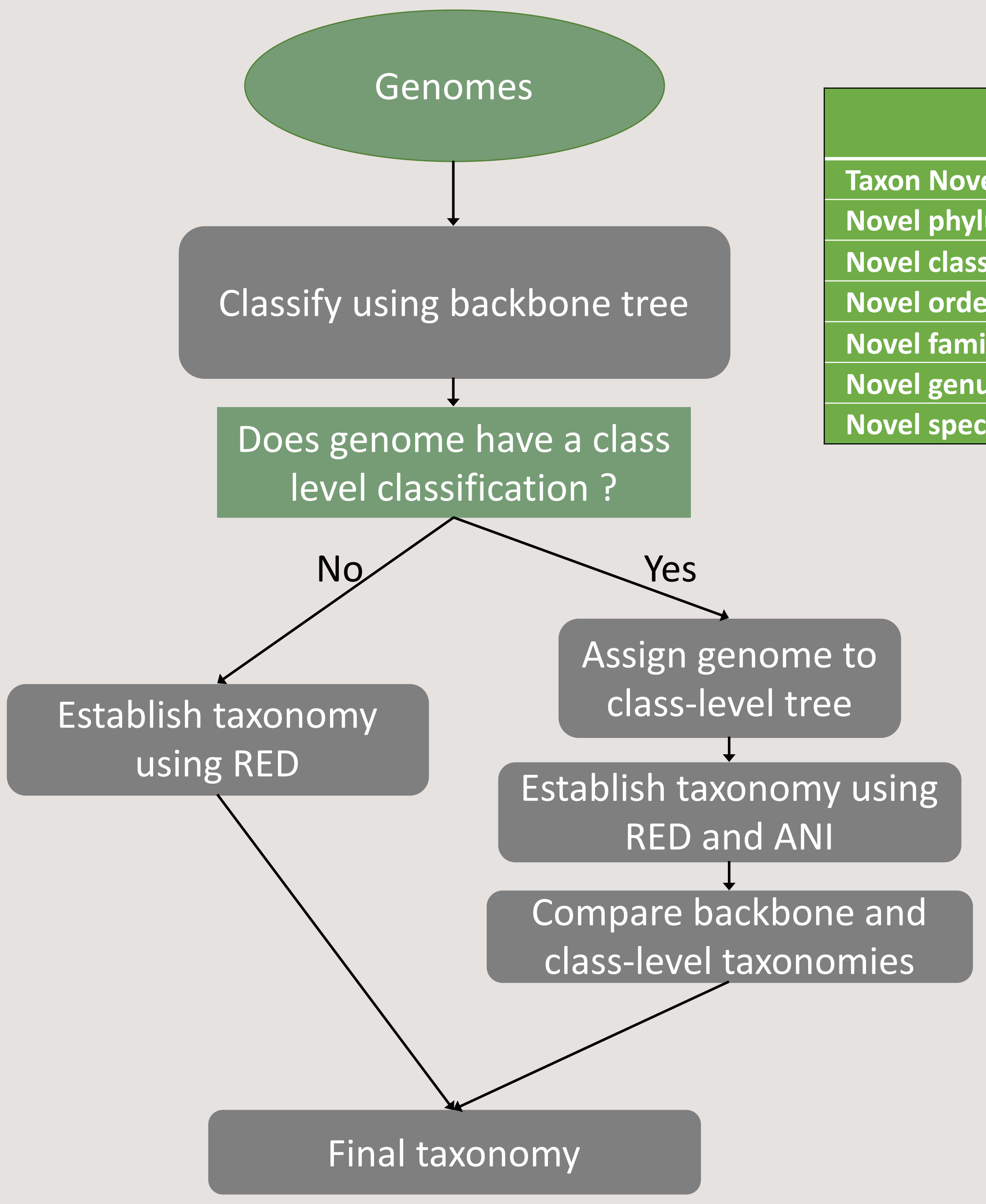


Bacterial phyla

c__Aerophobia

Query

Aerophobia is the only class within the *Aerophobota* phylum and as such, the query genome may be classified as the most basal order in *Aerophobia*, a novel class within the *Aerophobota*, or a novel phylum depending on its RED value.



Genomes

Classify using backbone tree

Does genome have a class level classification ?

No

Yes

Establish taxonomy using RED

Assign genome to class-level tree

Establish taxonomy using RED and ANI

Compare backbone and class-level taxonomies

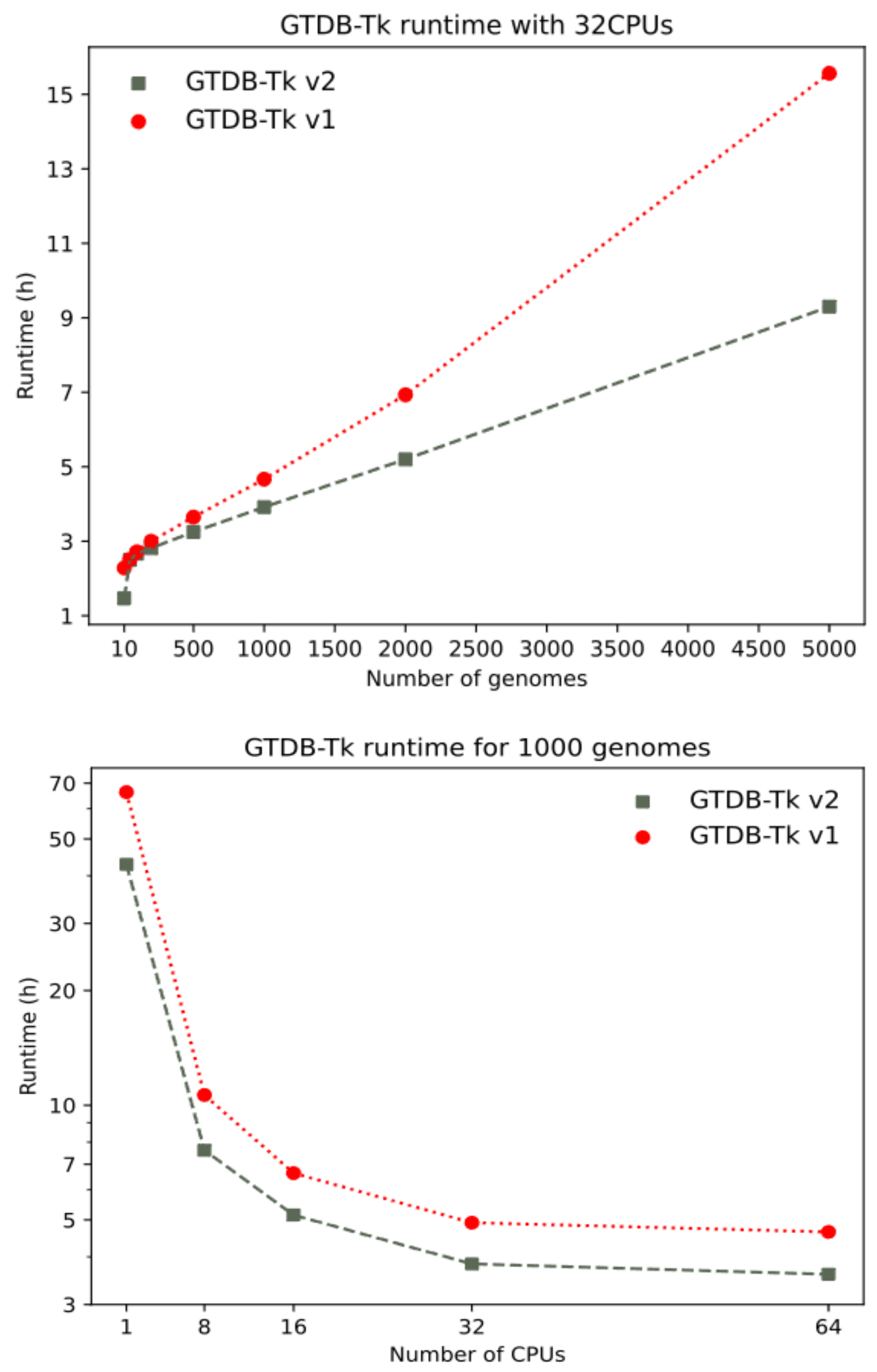
Final taxonomy

		GTDB-Tk v2 classifications relative to GTDB-Tk v1 classifications			
Taxon Novelty	No. genomes	Congruent	Conflict	Underclassified	Overclassified
Novel phylum	3	2	0	0	1
Novel class	42	35	2	2	2
Novel order	144	143	0	0	1
Novel family	543	540	0	1	2
Novel genus	3,222	3,219	0	1	0
Novel species	12,756	12,576	0	0	0

Performance

Memory: 55 GB vs 320GB with GTDB-Tk v1

Speed: Up to 40% faster than GTDB-Tk v1



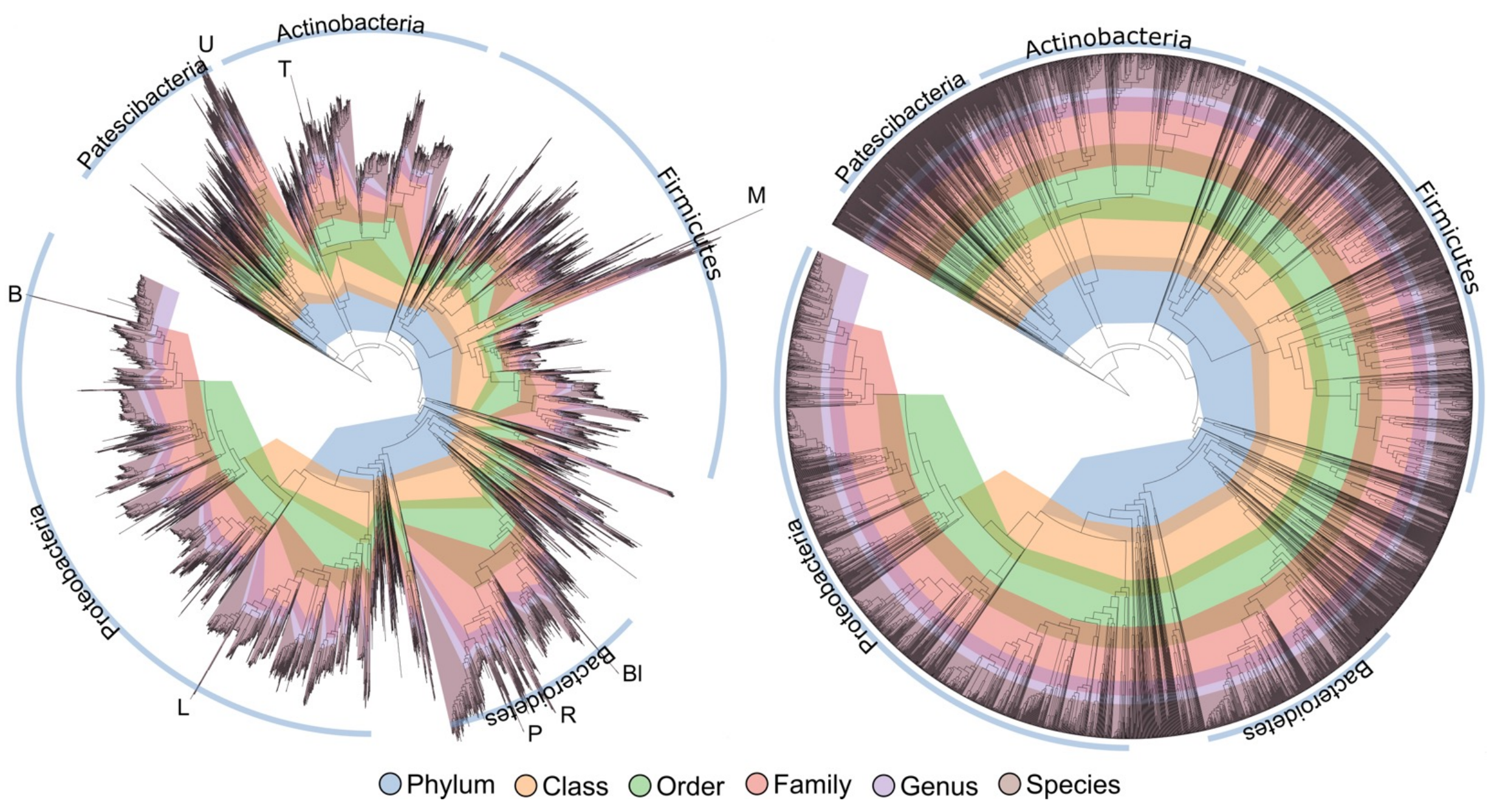
Accuracy

- Using **16,710** bacterial genomes from the GEMs dataset (*Nayfach et al., 2021*), **only 12** genomes (0.07%) did not have identical classifications between GTDB-Tk v1 and GTDB-Tk v2 (**See Table**).
- Using **23,548** genomes introduced in GTDB R07-RS207 classified using GTDB-Tk R06-RS202, **only 13** genomes (0.06%) did not have identical classifications between GTDB-Tk v1 and GTDB-Tk v2.

Relative Evolutionary Divergence (RED)

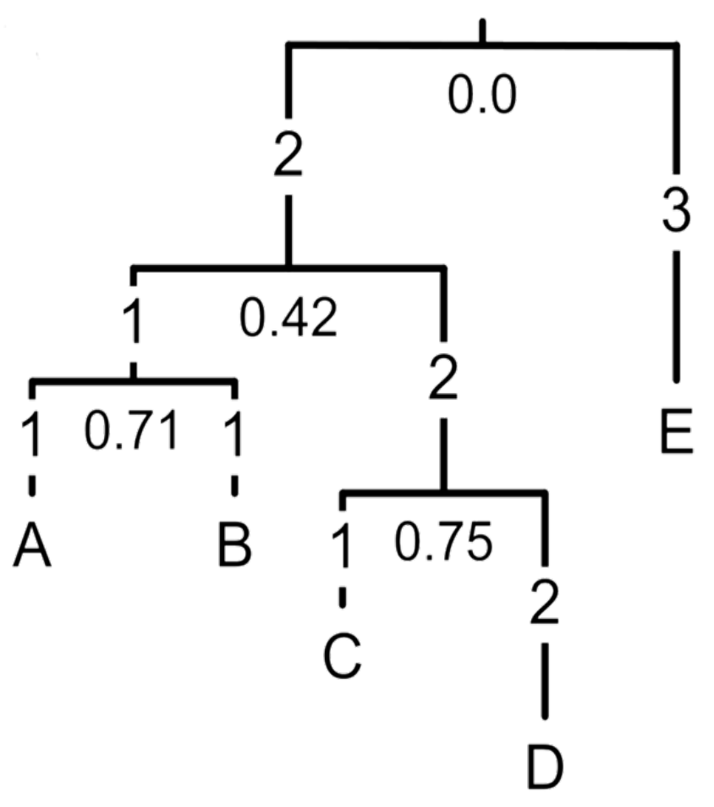
Goal: Ancestors of equal rank should have co-existed.

RED approximates this goal by normalizing the tree to account for varying rates of evolution and then using concentric bands in this normalized tree to define the desired placement of taxa at different ranks.



RED for a node, n, is p + (d/u) × (1 – p), where

- p is the RED of its parent
- d is the branch length to its parent
- u is the average branch length from the parent node to all extant taxa descendant from node to calculate.

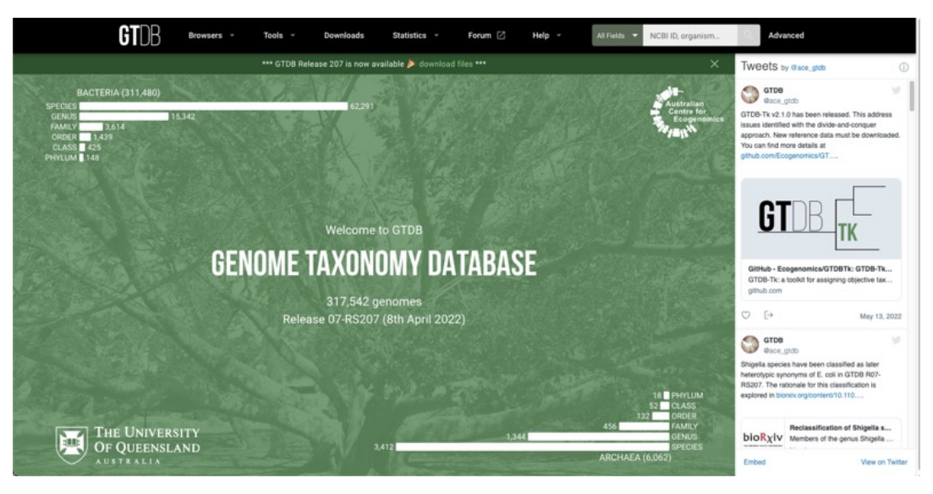


Example: the parent node of leaves C and D has a RED value of **0.75** = (0.42 + (2/3.5) × (1 – 0.42))

GTDB Team and Resources

GTDB website

gtdb.ecogenomic.org

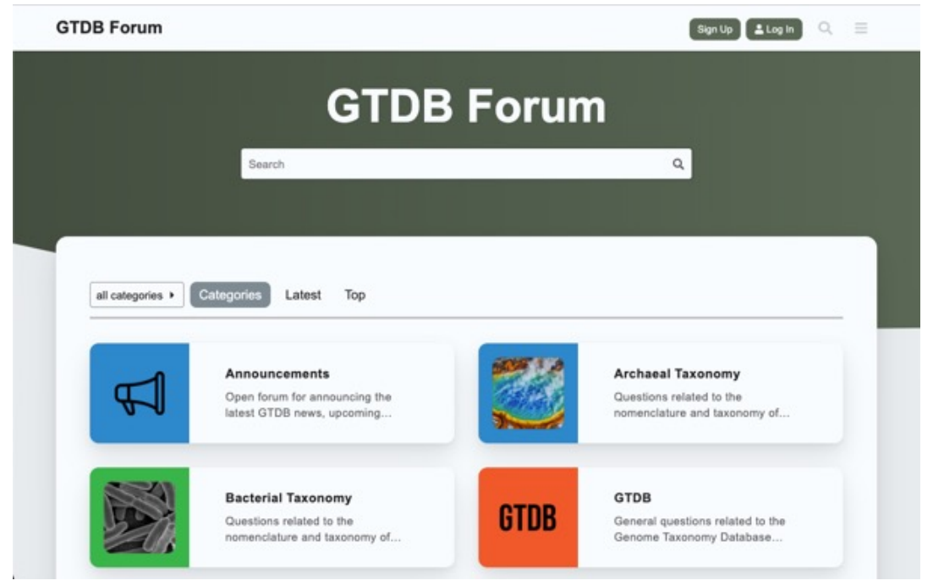


Documentation for GTDB-Tk

ecogenomics.github.io/GTDBTk

Open forum for announcing the latest GTDB news and discussing GTDB data

forum.gtdb.ecogenomic.org

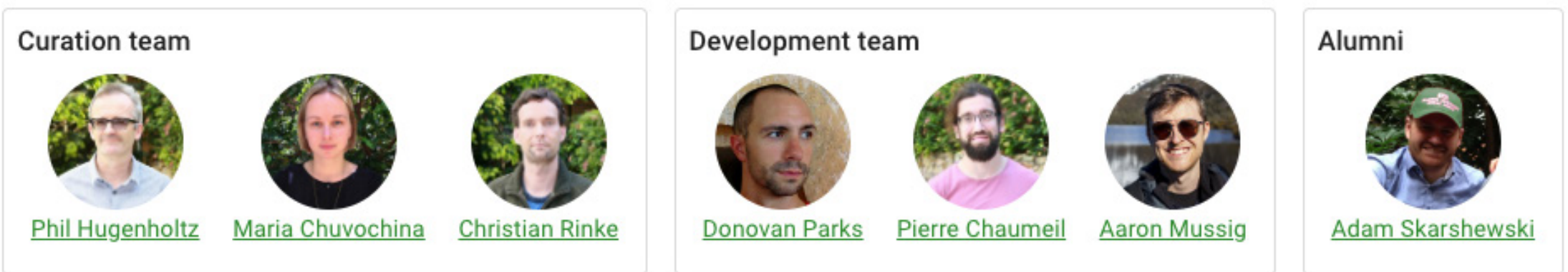


Announcements
Open forum for announcing the latest GTDB news, upcoming...

Archaeal Taxonomy
Questions related to the nomenclature and taxonomy of...

Bacterial Taxonomy
Questions related to the nomenclature and taxonomy of...

GTDB
General questions related to the Genome Taxonomy Database...



Curation team: Phil Hugenholtz, Maria Chuvochina, Christian Rinke

Development team: Donovan Parks, Pierre Chaumeil, Aaron Mussig

Alumni: Adam Skarshewski

p.chaumeil@uq.edu.au

@ace_gtdb

gtdb.ecogenomic.org

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

References

- Chaumeil, P.-A et al.(2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics https://doi.org/10.1093/bioinformatics/btz848.
- Almeida, A.et al. (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39, 105–114. https://doi.org/10.1038/s41587-020-0603-3.
- Nayfach, S. et al. (2021) A genomic catalog of Earth's microbiomes. Nat Biotechnol 39, 499–509. https://doi.org/10.1038/s41587-020-0718-6.
- Matsen, F.A., Kodner, R.B. & Armbrust, E. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11, 538 (2010). https://doi.org/10.1186/1471-2105-11-538.
- Parks, D.H. et al. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res., 50 (D1), D785-D784.

Download this poster

GTDB–Tk v2 preprint

Australian Centre for Ecogenomics