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

Pierre-Alain Chaumeil, Aaron J. Mussig, Philip Hugenholtz, Donovan H. Parks

Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences
The University of Queensland, Australia

What is GTDB-Tk

The Genome Taxonomy Database Toolkit (GTDB-Tk) is a computationally efficient toolkit that provides automated and objective taxonomic classification of bacterial and archaeal genomes by placing them into domain-specific, concatenated protein reference trees.

It has been used to assign taxonomic classifications to tens of thousands of bacterial and archaeal metagenome-assemble 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).

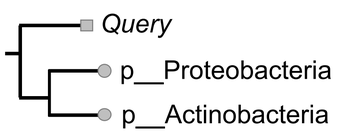
Why do we need GTDB-Tk v2?

GTDB-Tk is placing genomes into the GTDB reference trees using the maximum-likelihood (ML) placement tool pplacer (Matsen et al. 2010).

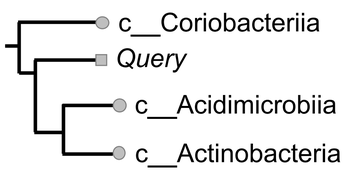
when using the GTDB R07-RS207 bacterial reference tree comprised of 62,291 genomes, pplacer requires ~320 GB of RAM .

Here we show that of GTDB-Tk v2 addresses the memory requirements by dividing the GTDB bacterial reference tree into class-level subtrees

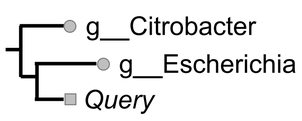
How does GTDB-Tk classify my genomes?



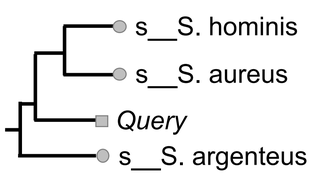
Query genome represents a new phylum



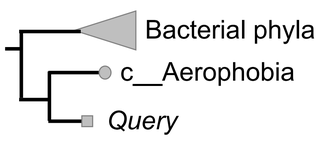
Query genome represents a novel class within the phylum *Actinobacteria.*



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.

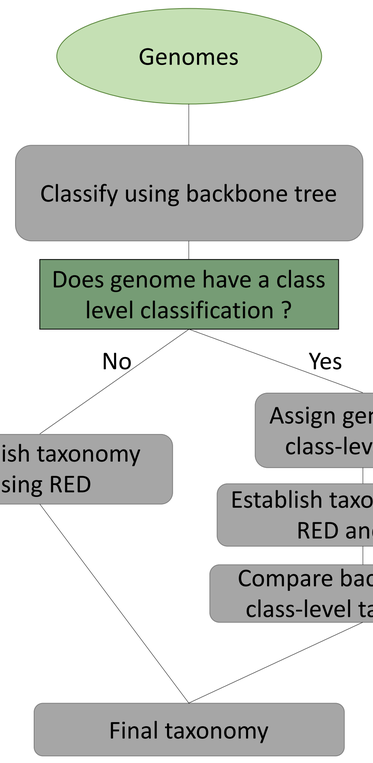


The query genome is assigned to the closest *Staphylococcus* species if the ANI is above the species ANI circumscription radius or is otherwise classified as a novel species.



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.

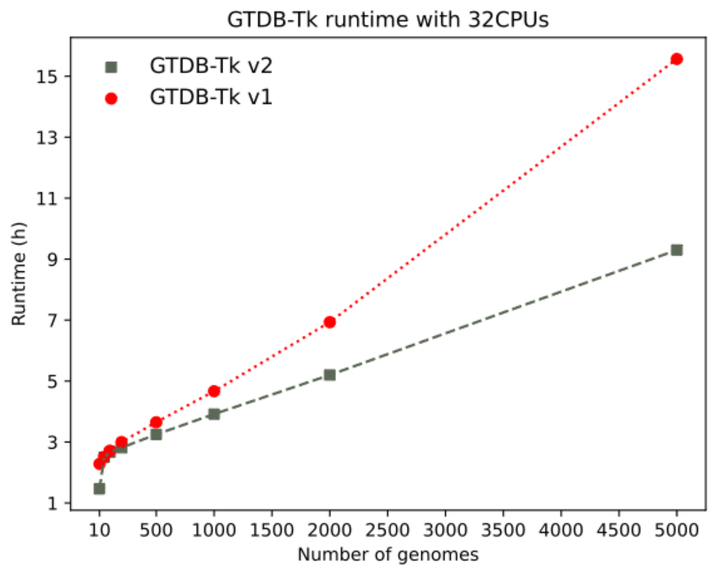
p.chaumeil@uq.edu.au
@ace_gtdb
gtdb.ecogenomic.org

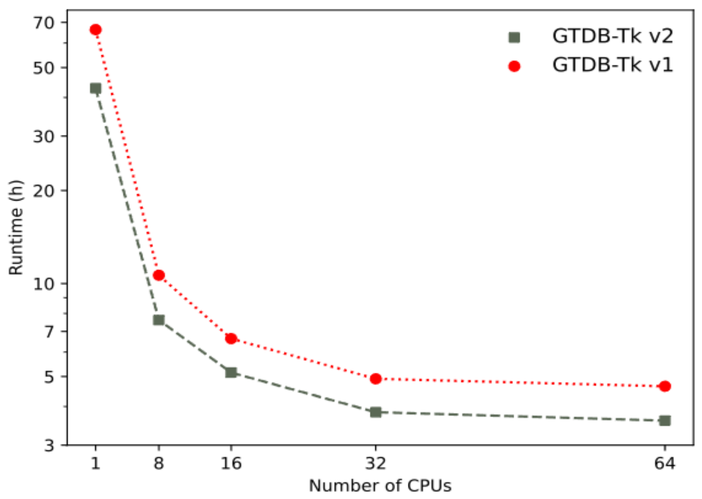


Performance

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

Speed: Faster than GTDB-Tk v1 (up to 40% faster)





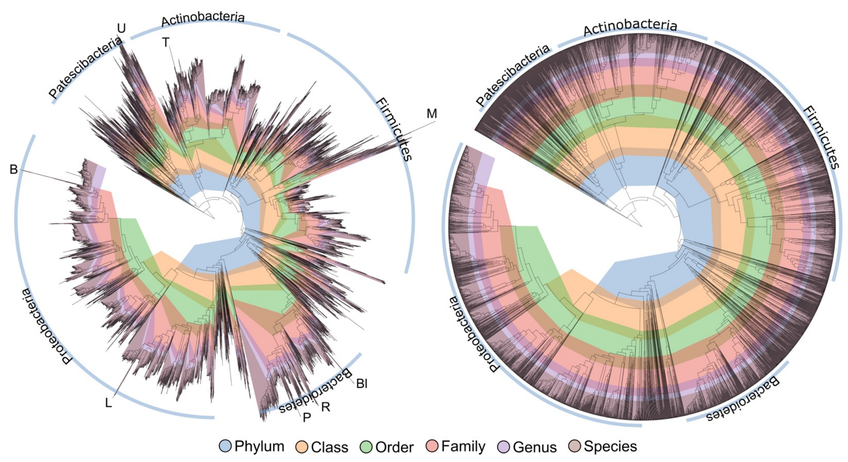
Accuracy

- Using **16,710** bacterial genomes from the GEMs dataset (Nayfach et al., 2021) , **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,**13** genomes (0.06%) did not have identical classifications between GTDB-Tk v1 and GTDB-Tk v2

		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

Relative Evolutionary Divergence (RED)



Relative Evolutionary Divergence (RED)

Formula : p + (d/u) × (1 – p),

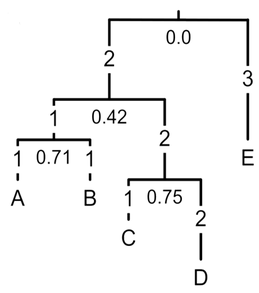
p is the RED of its parent

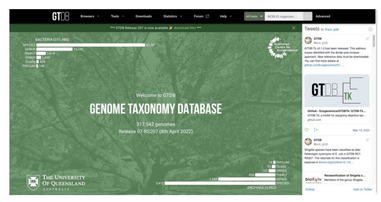
d is the branch length to its parent

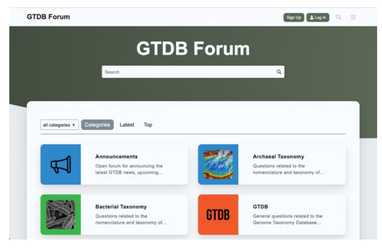
u is the average branch length from the parent node to all extant taxa descendant from node to calculate.

Example, the parent node of leaves C and D has a RED value

of 0.75 (0.42 + (2/3.5) × (1 – 0.42)),







Curation team: Phil Hugenholtz, Maria Chuvochina, Christian Rinke

Development team: Donovan Parks, Pierre Chaumeil, Aaron Mussig

Alumni: Adam Skarshewski

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.

Australian Centre for Ecogenomics

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA