Genome Taxonomy Database and SeqCode: Microbial Taxonomy and Nomenclature in the Age of Big Sequence Data

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

‡ The University of Queensland, School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, QLD 4072, Brisbane, Australia

Corresponding author: Maria Chuvochina (m.chuvochina@gmail.com)

Abstract

Microbial taxonomy and nomenclature have been challenged by methodological advances in high-throughput sequencing and high-performance computing. While taxonomy appears to adapt rapidly and has benefited enormously from the availability of whole-genome sequences, nomenclature still struggles to embrace these changes. Here, we present two independent initiatives that have resulted from the transitions of taxonomic practices in microbiology from a phenotypic and single gene-driven framework to a genome-based driven framework.

The first initiative, the Genome Taxonomy Database (GTDB), was developed to address the needs of microbial taxonomists to classify rapidly accumulating genome sequences from both cultured and uncultured microorganisms. Availability of growing numbers of metagenome-assembled genomes (MAGs) and single amplified genomes (SAGs), combined with the genomes from cultured species, created a perfect opportunity for building a consensus classification based on an evolutionary framework. This has been realised in the GTDB, a knowledgebase that provides phylogenetically consistent and rank-normalised taxonomies for bacterial and archaeal genomes. A distinctive feature of GTDB is a complete classification of genomes from species to domain using an automated approach combining average nucleotide identity (ANI) and relative evolutionary divergence (RED), followed by manual curation. GTDB has become an essential taxonomic resource for microbiologists worldwide, attracting ~3,500 users per month. GTDB mainly relies on two public databases, the National Center for Biotechnology Information (NCBI) Assembly database to which GTDB releases are indexed and the List of Prokaryotic names with Standing in Nomenclature (LPSN), as the primary nomenclatural reference. The database operates according to the FAIR (Findable, Accessible, Interoperable, Reusable) data principles and incorporates its own internal (e.g., standards for delineating taxa) as well as external standards. The latter are often directly adopted from the NCBI since it is used as a

primary source of genomes as well as metadata. Examples of such standards include Darwin Core data standards from Biodiversity Information Standards (TDWG), Minimum Information (MI) about any (x) Sequence (MIxS) and MISAG and MIMAG standards (Bowers et al. 2017) from the Genomic Standards Consortium. GTDB is used by many third-party resources and provides direct links to external public resources used for curation and validation of taxonomies. Importantly, GTDB contributes to the further generation of knowledge by enabling users to classify their own genomes within the GTDB taxonomic framework using our open-source GTDB-Tk tool. To our knowledge, GTDB is the only database that provides a comprehensive systematic *de novo* taxonomy for prokaryotes, which serves a multitude of purposes to its global users.

The second initiative, the *Code of Nomenclature of Prokaryotes Described from Sequence Data* or SeqCode, was developed in response to the need for formal naming of uncultured microbial diversity. This need has become even more evident with the establishment of the GTDB taxonomy, which highlighted many issues with nomenclature of uncultured taxa at scale. These include the absence of nomenclatural types, proposed higher taxon names without named children, and the lack of priority for *Candidatus* names (a prefix indicating a provisional status for the names of organisms falling outside the existing Prokaryotic Code). All these issues arise from one core issue: the absence of regulations for naming uncultured taxa because the International Code of Nomenclature of Prokaryotes (ICNP; Oren et al. 2023) only applies to microorganisms able to be obtained in pure culture. To solve this problem and ultimately to be able to express taxonomic affiliations of uncultured taxa in a regulated manner, genome sequences are proposed to serve as nomenclatural types under the SeqCode. This new code has many common aspects with the ICNP and recognises names that are validly published under the ICNP. It operates via an online Registry that allows registration and validation of names following one of two paths:

1. new names are registered and reviewed prior to publication and validated upon the notification about effective publication, or
2. existing names such as names of *Candidatus* taxa are registered and reviewed with a validation certificate granted upon the satisfaction of all checks.

To avoid naming ambiguity and ensure accurate species descriptions, SeqCode requires that genome sequences designated as types satisfy recommendations on minimal standards for DNA sequences, which are largely adopted from the MISAG and MIMAG standards. The SeqCode Registry also embraces FAIR principles, and was developed with interoperable data structures to facilitate the sharing of its names across global biodiversity resources including GTDB. Recently, we illustrated how SeqCode can be applied, along with the ICNP, by proposing new names for GTDB-defined higher taxonomic names under the two codes (Chuvochina et al. 2023). While it is not ideal to operate under two Prokaryotic codes, we believe that this development is a necessary step towards a unified nomenclatural system.

Keywords

genomes, MAGs, taxonomic names, *Candidatus*, prokaryotes

Presenting author

Maria Chuvochina

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Bowers R, Kyrpides N, Stepanauskas R, et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature Biotechnology 35: 725-731. https://doi.org/10.1038/nbt.3893
- Chuvochina M, Mussig AJ, Chaumeil P, Skarshewski A, Rinke C, Parks DH, Hugenholtz P (2023) Proposal of names for 329 higher rank taxa defined in the Genome Taxonomy Database under two prokaryotic codes. FEMS Microbiology Letters https://doi.org/10.1093/femsle/fnad071
- Oren A, Arahal D, Göker M, Moore EB, Rossello-Mora R, Sutcliffe I (2023) International Code of Nomenclature of Prokaryotes. Prokaryotic Code (2022 Revision). International Journal of Systematic and Evolutionary Microbiology 73 (5a). https://doi.org/10.1099/ijsem.0.005585