SeqCode, a nomenclatural code for prokaryotes described from sequence data

Prepared as a Brief Communication for Nature Microbiology

Brian P. Hedlund[1,] Maria Chuvochina[2], Philip Hugenholtz[2], Konstantinos T. Konstantinidis[3], Alison E. Murray[4], Marike Palmer[1], Donovan H. Parks[2], Alexander J. Probst[5], Anna-Louise Reysenbach[6], Luis M. Rodriguez-R[7], Ramon Rossello-Mora[8], Iain C. Sutcliffe[9], Stephanus N. Venter[10] and William B. Whitman[11*]

[1] School of Life Sciences, University of Nevada, Las Vegas, NV, USA

[2] The University of Queensland, School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, Brisbane, Australia

[3] School of Civil and Environmental Engineering, Georgia Tech, Atlanta, GA, USA

[4] Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV, USA

[5] Department of Chemistry, Environmental Microbiology and Biotechnology (EMB), Group for Aquatic Microbial Ecology and Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Essen, Germany.

[6] Biology Department, Portland State University, Portland, OR, USA

[7] Department of Microbiology and Digital Science Center (DiSC), University of Innsbruck, Innsbruck, Austria

[8] Marine Microbiology Group, Department of Animal and Microbial Diversity, Mediterranean Institute of Advanced Studies (CSIC-UIB), Esporles, Illes Balears, Spain

[9] Faculty of Health & Life Sciences, Northumbria University, Newcastle upon Tyne, UK

[10] Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa

[11] Department of Microbiology, University of Georgia, Athens, GA, USA

For correspondence: William B. Whitman, whitman@uga.edu

Abstract

Most prokaryotes are not available as pure cultures and therefore ineligible for naming under the International Code of Nomenclature of Prokaryotes. Here we summarize the development of the SeqCode, a code of nomenclature under which genome sequences can serve as nomenclatural types. The SeqCode operates through self-registration (https://seqco.de/), provides a reproducible and objective framework for all prokaryotes, regardless of cultivability, and facilitates communication across microbiological disciplines.

Manuscript body

It is widely recognized that the requirement of the International Code of Nomenclature of Prokaryotes (ICNP) for deposition of axenic and viable cultures as nomenclatural types has hindered the development of a nomenclature for uncultured and fastidious cultured prokaryotes (Archaea and Bacteria) and thus effective communication of microbial diversity (Konstantinidis et al., 2017; Murray et al., 2020). For example, as-yet-uncultivated taxa account for ~85% of the phylogenetic diversity of prokaryotes (Nayfach et al., 2021), and named prokaryotes account for only <0.2% of total species (Sutcliffe et al., 2021). By excluding the uncultured majority, a substantial portion of the tree of life is relegated to poorly ordered, ambiguous, and often synonymous names or alphanumeric codes, the latter of which have limited mnemonic value (Miller 1956).

To address this problem, Murray et al. (2020) proposed two paths, which were endorsed by 121 authors and signatories from 22 countries and six continents (Murray et al., 2020). 'Plan A' was based on proposals by Whitman (2015) that DNA sequences could serve as nomenclatural types and be incorporated into the existing ICNP infrastructure. However, the International Committee on Systematics of Prokaryotes (ICSP) rejected Whitman's proposal (Sutcliffe et al., 2020), thus triggering "Plan B", which called for a new code of nomenclature (Murray et al., 2020). To further engage the community in the implementation of "Plan B", we organized a series of online workshops (https://www.isme-microbes.org/reports-sponsored-events) that garnered 848 registrants from a broad range of microbiology disciplines and 42 countries. Ninety percent of participants reported that they would use a new code that accepts DNA sequences as types (https://www.isme-microbes.org/sites/default/files/reports/Path_forward_Naming_Uncultivated.pdf). Given strong participation and near-unanimous support, we acted on a variety of community recommendations (Table S1) to complete the SeqCode (formally The International Code of Nomenclature of Prokaryotes Described from Sequence Data; see Additional Information) and made progress on systems to implement it.

The SeqCode uses genome sequence data as common currency for typification of both cultivated and uncultivated microorganisms and follows the tenets of the ICNP by observing similar rules of priority. In essence, these rules state that the earliest validly published name for a taxon in a

particular position is the correct name, observing historical precedent and stabilizing nomenclature. The SeqCode also recognizes the priority of ICNP names provided they do not violate the priority of SeqCode names, thus minimizing divergence between the systems. Taxonomic names will be captured in the SeqCode Registry, a simple self-registration portal through which names and nomenclatural types (e.g., genome sequences for species) are registered, validated, and linked to metadata. In the best-case scenario, data will be entered and reviewed prior to publication, allowing automated checks and curators to guide users through the naming process. Following peer review and publication of the manuscript describing the taxa, the manuscript Digital Object Identifier (DOI) is entered into the Registry, completing the valid publication of the name/s (Figure 1, Path 1). However, the SeqCode also enables registration of previously published names, such as *Candidatus* names that conform to its rules. In that case, the *Candidatus* designation could be dropped, and the names given priority under the SeqCode (Figure 1, Path 2; see Additional Information). While the SeqCode itself is necessarily comprehensive, we have also developed resources to guide the community, including a glossary and examples (see Supplementary Information). Table 1 summarizes recommended minimal standards for sequences and reporting requirements. We endorse high quality standards for use of the SeqCode but expect standards to evolve to keep pace with community feedback and methodological improvements.

Potential users may ask: (i) What is the difference between *Candidatus* status and valid publication under the SeqCode? In reply, *Candidatus* is a provisional status lacking priority and standing and is relegated to a non-legislative appendix of the ICNP. *Candidatus* status was developed for organisms for which "more than a mere nucleic acid sequence is available". Since its inception, visualization of the taxon in a natural sample has been recommended (Murray and Stackebrandt 1995; Parker et al., 2019), but this is rarely implemented. It has been argued that *Candidatus* names should be granted priority under the ICNP (Whitman et al., 2019); however, this proposal was also rejected (Sutcliffe et al., 2020). As a result, many *Candidatus* names may prove to be ephemeral. (ii) What are the consequences for taxonomic names that are published in primary literature but not validly published under the SeqCode? Although the community is free to publish taxonomic names that do not comply with codes of nomenclature, we argue that codes of nomenclature and taxonomic frameworks serve the greater community by promoting objectivity, best practices, communication, and data interoperability. However, the unique restrictions of the ICNP regarding viable and accessible type strains have alienated many microbiologists and engendered a sense of normalcy in publishing names outside of the regulation of the ICNP. The SeqCode addresses this problem by providing an efficient and user-friendly resource that serves the common interests of the wider research community. The SeqCode embraces Findability, Accessibility, Interoperability, and Reusability (FAIR) principles, and the Registry was developed with interoperable data structures to promote sharing SeqCode names across global biodiversity inventories within microbiology and the broader biology research communities (e.g., NCBI (Schoch et al., 2020), GTDB (Parks et al., 2018),

MiGA (Rodriguez-R et al., 2016), LPSN (Parte et al., 2020), Catalogue of Life (Roskov et al., 2019), Global Biodiversity Information Facility (GBIF 2020)).

In closing, we emphasize a few important points. First, the SeqCode is not intended to discourage cultivation. Cultivation of mixed or pure cultures enables testing properties predicted from genomes under controlled conditions. Furthermore, investigators are strongly encouraged to deposit strains to culture collections to improve strain stability and availability, enable assessment of reproducibility of phenotypic traits, provide resources for biochemistry and biotechnology, and promote international cooperation. Second, like all other codes of nomenclature, the SeqCode does not provide rules or recommendations on the delineation of taxa. Existing and improving approaches and data structures are available for that purpose (e.g., Parks et al., 2020; Rodriguez-R et al., 2018), and proposals for novel taxa must be settled through peer review. Finally, this is the first version of the SeqCode and we hope that it will evolve as the community engages in further developing the system. Because of our desire to serve the broad microbiology research community, we will engage the community to gather feedback and develop bylaws for SeqCode administration. This code is driven by bottom-up desires to improve communication across the microbial sciences. Thus, we view this 'SeqCode v1.0' as a necessary first step toward a unified system of nomenclature to communicate the full diversity of prokaryotes, and we will cooperate with the community toward building this vision.

Acknowledgements

Large portions of the text of the SeqCode were derived from the ICNP, and the editors gratefully acknowledge the many authors who contributed to that code. Funding was provided by the US National Science Foundation (DEB 1841658, DEB 1557042, and EAR 1516680), the US National Institute of General Medical Sciences (GM103440) from the National Institutes of Health, the Spanish Ministry of Science, Innovation and Universities (PGC2018-096956-B-C41), the Australian Research Council (FL150100038), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, SFB 1439/1 2021 – 426547801) also supported with European Regional Development Funds (FEDER), and the International Society for Microbial Ecology (ISME). We also thank all participants in the SeqCode workshops, especially guest speakers who graciously shared their expertise: Jongsik Chun, Nicole Dubilier, Emiley Eloe-Fadrosh, Chris Lane, Juncai Ma, Edward Moore, Aharon Oren, Jörg Overmann, Susanne Renner, Vincent Robert, Conrad Schoch, Scott Tinghe, Linhuan Wu, and Arvind Varsani.

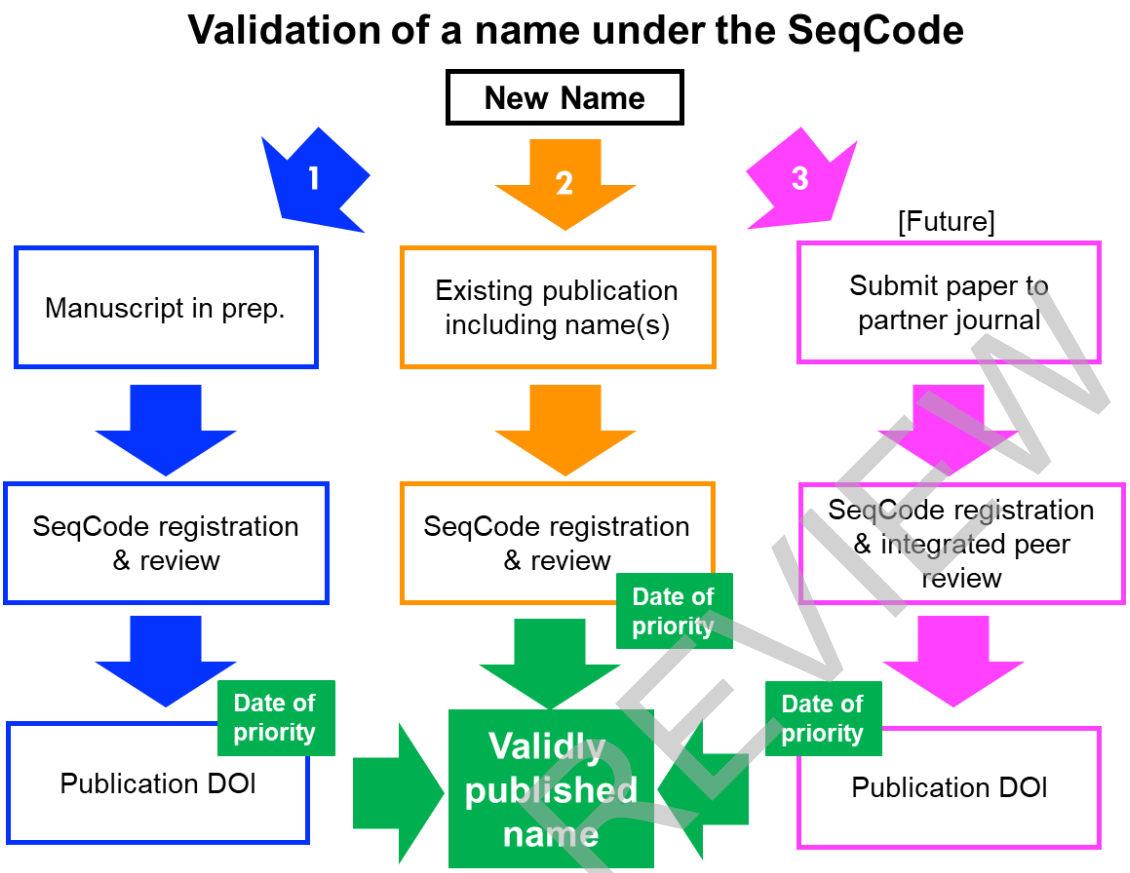

Figure 1. Validation process of a name under the SeqCode. Currently, two mechanisms exist, with a third possible in the future. The recommended mechanism (left arrow, Path 1) involves draft registration of the name and metadata into the SeqCode Registry prior to publication. Automated data quality and name synonymy checks in conjunction with curator review will lead to provisional acceptance of proposals that comply with SeqCode rules. Completion of the registration process requires the DOI of the effective publication. Once the proposal is accepted and the DOI entered, the registration is complete, marking the time and date of priority. The second (middle arrow, Path 2) is for names that are already published, such as *Candidatus* names. It requires draft registration of the name and metadata into the SeqCode Registry. SeqCode curators review compliance with the SeqCode rules before accepting the proposal. Acceptance of the proposal completes registration and marks the time and date of priority. At that point, the *Candidatus* designation can be removed. The third mechanism could be developed in partnership with one or more journals in the future (right arrow, Path 3) and would involve simultaneous peer review and SeqCode Registry curator review as an integrated path to the validation of proposed names. Issue of the DOI of the accepted paper marks the time and date of priority.

Table 1. Data quality and reporting requirements and recommendations for an isolate genome, MAG, or SAG to serve as the nomenclatural type for a species named under the SeqCode. Requirements will be checked as part of the validation process on the SeqCode Registry. Recommendations are suggested best practices to guide authors and peer reviewers to ensure high quality data supporting species to be named. See Supplementary Information for examples.

Information	Requirements	Recommendations
Included in publication proposing new species names under SeqCode[a]		
Name	Required for all names	1. Etymologies for all proposed names are recommended. 2. Names with mnemonic cues are recommended.
Interpretation of biological properties	None	Indicate inferred or demonstrated physiological traits and ecological information, such as habitat in the manuscript body and/or protologue.
Designated genome	None	1. Indicate access to genomic assembly (e.g., INSDC accession). 2. Indicate access to raw data (e.g., SRA accession). 3. Demonstrate compliance with GSC standards for isolate genomes (Field et al., 2008) and high-quality SAGs and MAGs (Bowers et al., 2017). 4. Include as much metadata as possible in the publication (see Field et al., 2008).
Evidence of the species, taxonomic rank, and position	None	1. Demonstrate the uniqueness of the species with respect to existing named species and justify the taxonomic rank and position (e.g., Jain et al., 2018, Karthikeyan et al., 2019; Parks et al., 2020; Rodriguez-R et al., 2018). 2. For MAGs and SAGs, compare multiple high-quality genomes representing the species in more than one sample (e.g., Supplemental Information).[b]
Data quality[c] and availability necessary for completion of SeqCode Registry		
Type genome assembly quality	1. >90% complete and <5% contaminated; 16S and 23S rRNA genes >75% complete (modified from Bowers et al., 2017). 2. Isolate genome read coverage ≥50x (Field et al., 2008).	1. >80% of tRNAs present (modified from Bowers et al., 2017). 2. High genome integrity (contig # <100; N50 >25 kb; max. contig >10 kb). 3. MAG/SAG read coverage ≥10x.
INSDC data availability	1. Assembly available in INSDC database. 2. Raw data available in INSDC databases (e.g., Sequence Read Archive)[d].	1. Data submission using MIxS Checklists in INSDC databases (https://gensc.org/mixs/). 2. Include as much metadata as possible in INSDC.
SeqCode Registry	Type genome assembly and raw data INSDC	Provide as much contextual data as possible to facilitate downstream genome comparisons with

	accession numbers, taxon name, etymology, rank.	respect to provenance.

a. There are purposefully few requirements for the effective publication to accommodate existing and future publications that don't adhere to all recommendations. Critical data will be captured on the SeqCode Registry (Figure 1).
b. Comparison of multiple high-quality genomic assemblies from multiple samples can support the non-chimeric nature of MAGs and provide confidence of the assembly for both MAGs and SAGs.
c. Data quality can be assessed by automated pipelines or other approaches. Exceptions for lower data quality should be justified by authors in the effective publication.
d. Not required for names effectively published before January 1, 2023, to allow for existing published names (e.g., existing *Candidatus* names) and names currently undergoing peer review to be validated under the SeqCode.

References

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725-31.

GBIF: The Global Biodiversity Information Facility (2020) What is GBIF? Available from https://www.gbif.org/what-is-gbif

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008 May; 26:541–547.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:1–8.

Konstantinidis KT, Rosselló-Móra R, Amann R. Uncultivated microbes in need of their own taxonomy. ISME J. 2017;11:2399–406.

Miller G. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 1956;63(2):81–97.

Murray RGE, Stackebrandt E. Taxonomic note: implementation of the provisional status candidatus for incompletely described procaryotes. Int. J. Syst. Bacteriol. 1995;45:186–7.

Murray AE, Freudenstein J, Gribaldo S, Hatzenpichler R, Hugenholtz P, Kämpfer P, et al. Roadmap for naming uncultivated archaea and bacteria. Nat Microbiol. 2020;5:987-94.

Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol 2021;39:499-509.

Parker CT, Tindall BJ, Garrity, GM. International code of nomenclature of prokaryotes. Int J Syst Evol Microbiol; 2019; 69:S1-S111.

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38;1079–86.

Parte AC, Sardà Carbasse J, Meier-Kolthoff JP, Reimer LC, Göker M. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. Int J Syst Evol Microbiol; 2020;70: 5607-12.

Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, Konstantinidis KT. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res. 2018;46:W282–8.

Roskov Y, Ower G, Orrell T, Nicolson D, Bailly N, Kirk PM, et al. Species 2000 & ITIS Catalogue of Life, 25th March 2019. Digital resource at www.catalogueoflife.org/col. 2021;Species 2000: Naturalis, Leiden, the Netherlands. ISSN 2405-8858.

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database. 2020 baaa062.

Sutcliffe IC, Rosselló-Mora R, Trujillo M. Addressing the sublime scale of the microbial world: reconciling an appreciation of microbial diversity with the need to describe species. New Microbes New Infect. 2021;43:100931.

Sutcliffe IC, Dijkshoorn L, Whitman WB. Minutes of the International Committee on Systematics of Prokaryotes online discussion on the proposed use of gene sequences as type for naming of prokaryotes, and outcome of vote. Int. J. Syst. Evol. Microbiol. 2020;70:4416-7.

Whitman WB. Genome sequences as type material for taxonomic descriptions. System. Appl. Microbiol. 2015;38:217-222.

Whitman WB, Sutcliffe I, Rosselló-Mora R. Proposal for changes in the International Code of Nomenclature of Prokaryotes: granting priority to Candidatus names. Int J Syst Evol Microbiol. 2019;69(7):2174-2175.