 doi:10.20944/preprints202203.0069.v1

Could a Focus on the "Why" of Taxonomy Help Taxonomy Better Respond to the Needs of Science and Society?

Leighton Pritchard[1], C. Titus Brown[2], Bailey Harrington[1], Lenwood S. Heath[3], N. Tessa Pierce-Ward[2], Boris A. Vinatzer[4]

1 Strathclyde Institute for Pharmacy and Biomedical Sciences (SIPBS), University of Strathclyde, 161 Cathedral Street, Glasgow, G4 0RE, UK
2 Department of Population Health and Reproduction, UC Davis, Davis, CA 95616, USA
3 Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA
4 School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061, USA

Corresponding author: Boris A. Vinatzer
Email: vinatzer@vt.edu

Abstract

Genomics has put prokaryotic rank-based taxonomy on a solid phylogenetic foundation. However, most taxonomic ranks were set long before the advent of DNA sequencing and genomics. In this concept paper, we thus ask the simple yet profound question: Should prokaryotic classification schemes besides the current phylum-to-species ranks be explored, developed, and incorporated into scientific discourse? Could such alternative schemes provide better solutions to the basic need of science and society for which taxonomy was developed, namely, precise and meaningful identification? A neutral genome-similarity based framework is then described that could allow alternative classification schemes to be explored, compared, and translated into each other without having to choose only one as the gold standard. Classification schemes could thus continue to evolve and be selected according to their benefits and based on how well they fulfill the need for prokaryotic identification.

Keywords: prokaryotic taxonomy, classification, identification, genomics

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 March 2022 doi:10.20944/preprints202203.0069.v1

The *why* of taxonomy

In an insightful article in 2021 the International Society of Microbial Ecology Journal, Hugenholtz and colleagues provided a comprehensive review of the history of prokaryotic taxonomy and highlighted current and future challenges, particularly in regard to the unculturable majority [1]. The article contributed rich context for the ongoing debate over taxonomic nomenclature and built a strong argument for genome-based taxonomy. However, we believe that improving taxonomy using genomics should not stop us from more fundamentally rethinking both its structure and applications and answering the question of *why* we practice taxonomy in the first place.

The *why* of taxonomy becomes clear when we consider all three elements of taxonomy: classification, nomenclature, and *identification*. Importantly, only when taxonomy leads to identification of an unknown as a member of a named group, a taxon, with distinct and relevant characteristics can it answer to scientific and societal needs. In this context, classification becomes a prerequisite for meaningful identification by creating clear and distinct boundaries of practical relevance between groups of microbes. The need for nomenclature also follows from identification as clear naming is critical for effective communication, and possibly societal action, following identification.

Here are some examples of what we intend by meaningful identification of microbes. In basic science, identification may simply consist in finding the position of an organism in a phylogenetic tree to reveal its evolutionary relationship to other organisms. Beyond single organisms, reliable identification of community members may delineate community structure to understand system-level responses central to the environmental roles of microbes. From a societal perspective, meaningful identification of an unknown as a member of a taxon may predict a threat, such as the potential to cause disease, and may trigger regulatory action under national or international laws, such as import/export restrictions or implementation of quarantine or isolation. On the other hand, identification as a member of a group with beneficial characteristics, may lead to intellectual property protection. Therefore, from the perspective of *identification*, it becomes clear that taxonomy answers important needs in both basic science and society.

Questioning the current practice of taxonomy

Also, when we focus on identification, aspects of current taxonomy that some taxonomists may consider to be unchangeable facts and needs of taxonomy can be seen to serve only the *how* of the current praxis of taxonomy, but not the *why* of taxonomy itself. In our view, several elements of classical taxonomy can be reassessed on these grounds, such as: the reliance on historic and subjective taxonomic ranks; making a distinction between “taxonomy” and “strain typing”;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 March 2022 doi:10.20944/preprints202203.0069.v1

considering species to be the smallest recognizable taxonomic unit; requiring name-bearing type strains to describe species; the attachment to Latin binomials; and the insistence on a scheme of stable and unique hierarchical names to describe a collection of items that could be partitioned in many other ways. In the current era of databases that can provide persistent stable IDs to individuals, we believe it is time for microbiology to adopt the advantages that technology brings to data organization, while preserving the best features of classical taxonomy.

Assumptions relating to genome-based taxonomy can also be questioned if we look at the *why* of taxonomy. For example, an assumption that phylogenetic clades based on the alignment of core genes or proteins should be the only basis for the circumscription of named groups is essentially an update of binomial nomenclature, privileging vertical transfer of genomic information. However, in many societally-important circumstances, the phenotype of interest may be governed by horizontally-transferred genes [2]. In these circumstances, a classification that prioritizes the phenotype and considers lateral transfer of genomic information might be more useful. Such a classification might be better facilitated by a system of individual genome accessions with flexible labelling (like tags in a Google Mail inbox) than a hierarchical naming scheme. Hugenholtz and colleagues argue that hierarchical taxonomic ranks and binomial species are a necessity because of biologists' reluctance to take up new systems, such as the rank-free PhyloCode [3]. However, reluctance to change is not an argument against change and, in any case, might very well apply more to taxonomists than to biologists.

Nonetheless, there are challenges inherent to taxonomy that cannot be avoided. Circumscription of existing groups requires revision as new knowledge becomes available, and such reclassification necessarily requires translation from one classification system to another. Also, there is more than one reasonable hierarchical classification system. For this reason, and because we do not have perfect knowledge and understanding of the hierarchical process of evolution, it is inevitable that no single human-created model will capture all useful categorizations of organisms — regardless of the claim that "biologists now agree that taxonomy should be based on evolutionary relationships as the most natural way of arranging organisms".

Where to go from here?

If we accept that our current taxonomic system is conditioned by history and just one of many reasonable choices, perhaps we could dare to try some unconventional solutions to answer the *why* of taxonomy within the landscape of genomic data. For example, what if we were to eliminate the distinction between taxonomy and within-species strain typing, and acknowledge that there is a fluid transition between species and strain? Would it help us understand biology better if we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 March 2022 doi:10.20944/preprints202203.0069.v1

were to give species complexes, within-species clades, and other monophyletic groups the same importance as current taxonomic ranks, as proposed in the PhyloCode [3]? What if we could use a neutral framework of genome identifiers to explore and compare new ways to infer evolutionary relationships between organisms, beyond core gene phylogenies? What value could be gained from approaches that combine genome similarity with similarity in gene content reflecting adaptation to different ecological niches? Is it possible to construct complementary phylogenetic or even non-phylogenetic classification systems, similar to library cataloging systems? Specialized schemes could be based on the content of functional classes of genes, such as pathogenicity/virulence genes that are important in a biosecurity context, where the risk is governed by gene content more than evolutionary relationship.

To implement, compare, and perhaps unify these alternative classification schemes, we would need a “Rosetta Stone" to translate between them. We are thus building a “genomic coordinate system” as part of the *genomeRxiv* platform using the Life Identification Number (LIN) approach [4, 5] analogous to a map grid reference, which hierarchically subdivides and labels the entire prokaryotic genome space into uniquely-labelled volumes (or voxels) of sequence-similar genomes that, at their finest resolution, contain a single genome uniquely identified with its “coordinate” (its LIN). We believe that this is a practical, quantitative, automatable, stable, and robust solution to the problem of translating among classification schemes, for example, between validly published prokaryotic named species and genome-based species clusters [6]. More in general, classifications of prokaryote genomes made by one scheme, such as descent from a common ancestor, or “bona fide species definitions,” can then be expressed as combinations of uniquely-labelled voxels and compared to similar combinations obtained by alternative classification schemes, such as presence or absence of specific genes.

This coordinate scheme is hierarchical, but purely descriptive. It neither requires nor imposes an evolutionary model. It is neutral on the questions of nomenclature and classification. It requires no consensus on a single scheme but would instead enable meaningful translation among alternative schemes. Our proposal is also by nature able to accommodate the many, as yet unknown and unclassified, prokaryotes that currently pose a significant problem for nomenclature and classification. So long as a genome sequence is available, a coordinate in genome space can be assigned and used as an identifier even before the genome is classified as a member of an already described, or still to be named, taxonomic group. If this genome is of an emerging pathogen, the identifier can be used for clear communication about an ongoing disease outbreak from the moment the genome has been sequenced without having to wait for a validly published name.

In conclusion, we propose to treat genome sequence data neutrally to build a genotypic framework. We do not propose to privilege a specific set of genes, or a specific evolutionary model or reconstruction method as an immutable truth, or "gold standard", against which all other schemes would be measured. Instead, we propose a whole-genome framework on which alternative choices of phylogenetic and phenotypic classification schemes can be compared. Like a coordinate system on a map, this framework provides an address for each genome and links to any classification information within any taxonomic system. We expect this framework to provide a landscape on which classification systems that best respond to the needs of science and society can continue to evolve based on the latest biological, biotechnological, and ecological discoveries.

Acknowledgements

Development of the *genomeRxiv* platform is supported by the US National Science Foundation (DBI-2018522) and the UK Biotechnology and Biological Sciences Research Council (BB/V010417/1).

Competing interests: *Life Identification Number* and *LIN* are registered trademarks of This Genomic Life Inc. Lenwood S Heath and Boris A Vinatzer report in accordance with Virginia Tech policies and procedures and their ethical obligation as researchers, that they have a financial interest in This Genomic Life Inc that may be affected by the publication of this manuscript. They have disclosed those interests fully to Virginia Tech, and they have in place an approved plan for managing any potential conflicts arising from this relationship.

References

1. Hugenholtz P, Chuvochina M, Oren A, Parks DH, Soo RM. Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J* 2021; **15**: 1879–1892.
2. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. *Nat Rev Genet* 2015; **16**: 472–482.
3. Cantino PD, de Queiroz** K. International Code of Phylogenetic Nomenclature (PhyloCode) : Version 6*, 1st Edition. 2020. CRC Press.
4. Vinatzer BA, Tian L, Heath LS. A proposal for a portal to make earth's microbial diversity easily accessible and searchable. *Antonie Van Leeuwenhoek* 2017; **110**: 1271–1279.
5. Tian L, Huang C, Mazloom R, Heath LS, Vinatzer BA. LINbase: a web server for genome-

 doi:10.20944/preprints202203.0069.v1

based identification of prokaryotes as members of crowdsourced taxa. *Nucleic Acids Res* 2020.

6. Sanford RA, Lloyd KG, Konstantinidis KT, Löffler FE. Microbial Taxonomy Run Amok. *Trends Microbiol* 2021; **29**: 394–404.