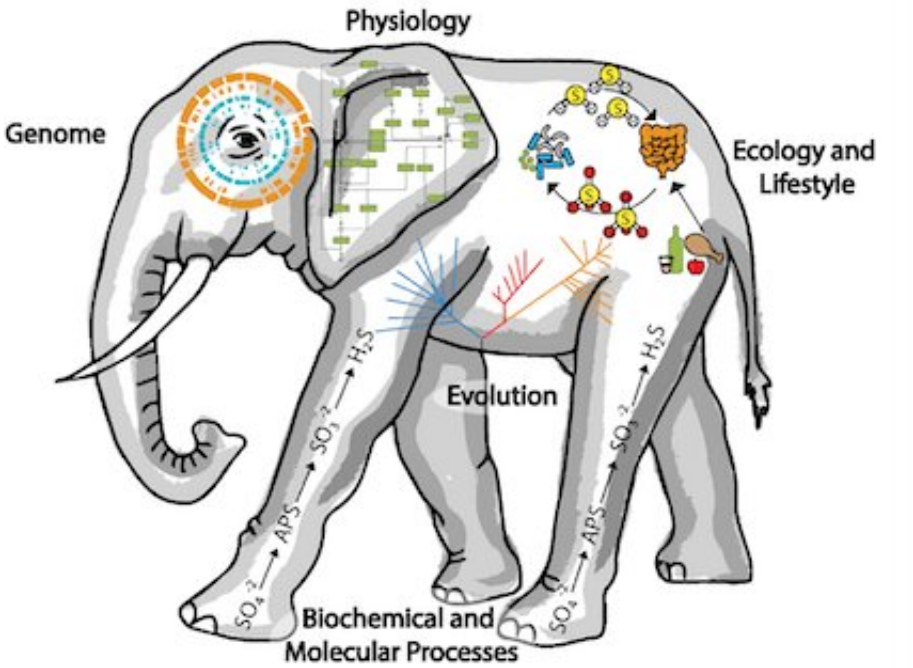
# SeqCode provides a path to name uncultivated prokaryotes

July 24 2023




The name elephant. Investigators from different disciplines examine different portions of a microorganism and develop their own opinions about its nature. Precise scientific names enable the unification of information from disparate disciplines and laboratories to create a single, multifaceted vision. Credit: William Whitman and Brian Hedlund

Most prokaryotes have never been isolated in pure culture and cannot be

named under the International Code of Nomenclature of Prokaryotes (ICNP). However, many studies have described uncultured bacteria and archaea though metagenome-assembled genome sequences (MAGs) or single-cell amplified genomes (SAGs), often in combination with physiological or ecological data.

These studies have greatly expanded our view of prokaryotic life, its metabolic capabilities and its roles in the environment, but communication about these uncultured bacteria and archaea is difficult without formal taxonomic names.

The Code of Nomenclature of Prokaryotes Described from Sequence Data (also called the SeqCode) provides a way to create scientifically precise taxonomic names for these uncultured prokaryotes by using genome sequences instead of cultures as nomenclatural types.

Nomenclatural types, or types, are elements to which a taxonomic name is attached and are also used to anchor taxonomic names of plants, animals and fungi. These elements can be compared to those of unidentified samples to decide whether or not the samples should have the same name. SeqCode names are compatible with the ICNP and allow creation of a unified taxonomy for cultured and uncultured prokaryotes. The SeqCode also sets standards for metagenome-assembled genomes (MAGs) and single-cell amplified genomes (SAGs) to be used in systematics, which ensures that nomenclatural types are of sufficient quality to unambiguously identify the taxon, create a meaningful taxonomy and support high-quality bioinformatic analyses.

## Why base names on genome sequence?

Taxonomic names are important for science communication, and SeqCode provides the means to name uncultured prokaryotes. Without a generally agreed upon naming system, history has shown that

nomenclature becomes imprecise, redundant and confusing. Precise naming is also critical for creation of stable and effective bioinformatic databases and for large-scale analyses common in bioinformatics. Without a single, permanent name for taxa, databases would require frequent curation, which would be increasingly difficult as the information content grows. Thus, a naming system is also important for the application of modern informatics tools.

Genome sequences are particularly well suited as nomenclatural types for prokaryotes. Previously, deposition of type strains in culture collections was necessary for naming new species of prokaryotes. Strains had to be cultivated from environmental samples and usually isolated as pure cultures, and then compared by a large number of mostly phenotypic tests in a process called polyphasic taxonomy to determine if they represented novel species.

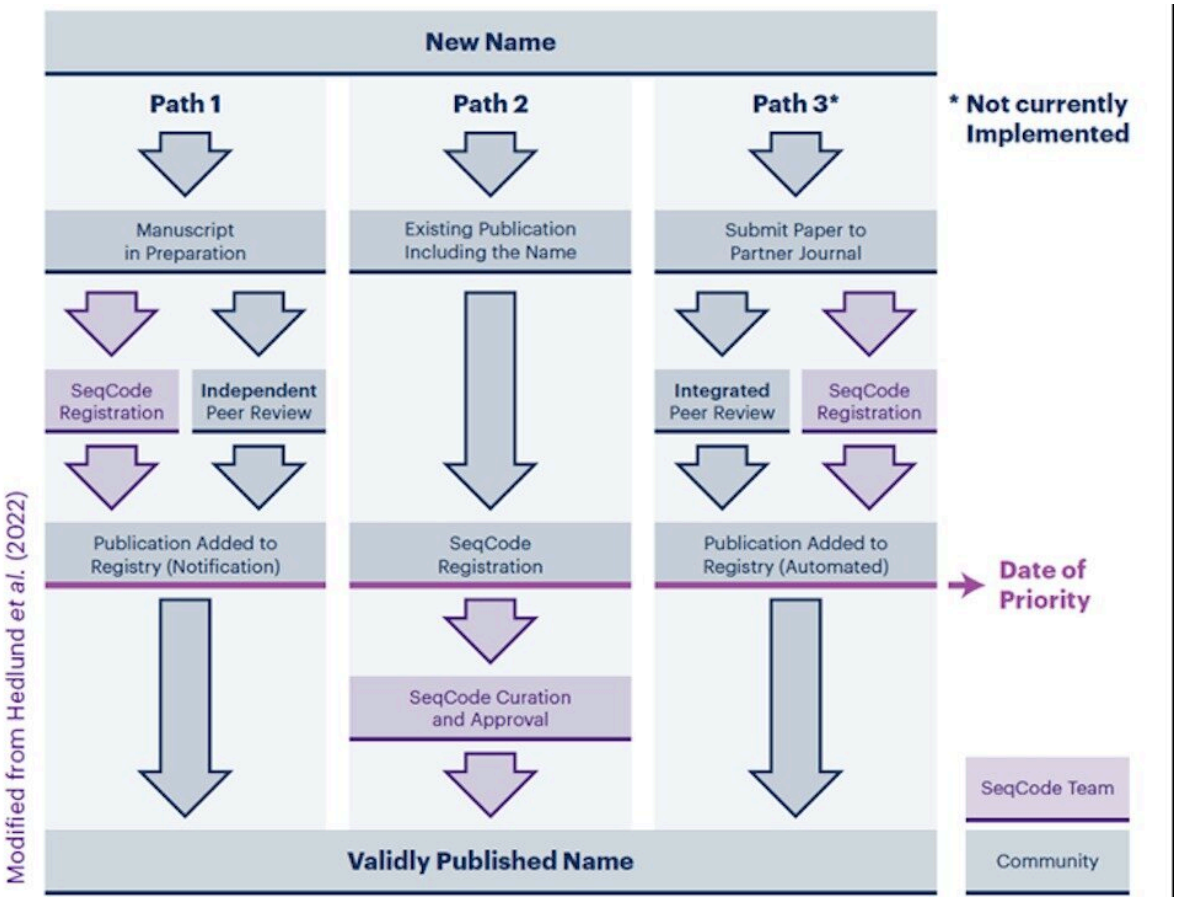
When molecular methods were first introduced, it was discovered that these earlier methods were inaccurate, often grouping unrelated organisms together and failing to recognize closely related species. First, 16S rRNA sequences were used to reconstruct the phylogenetic relationships among type strains and guide their taxonomy. In the last decade, the genomes of most type strains have been sequenced, which allowed a more precise delineation of the phylogenetic relationships between species and greatly enhanced the basic knowledge of their properties.

Because the genome sequence is the blueprint for the properties of the entire organism, it unites the enormous literature on the physiology, molecular biology and ecology of prokaryotes with the properties of specific organisms. In addition, it informs us of key lifestyle features, such as bioenergetics, physiology, differentiation, motility and phage infections. Linking names to genome sequences allows important inferences from one scientific discipline to inform the other disciplines.

In addition, there are also practical reasons to use genome sequences as types instead of strains. Sequences can be stored on computers and easily shared. Strains must be stored in culture collections, remain viable for the future and are comparatively difficult to share.

## Uncultured prokaryotes are common

Metagenome sequencing has revealed an enormous diversity of prokaryotes, most of which have never been cultured. Direct sampling of DNA from many biomes has identified MAGs representing 135 phyla. In contrast, only about 40 phyla are represented by cultures deposited in culture collections and named under the ICNP.



Validation process of names under the SeqCode. Source: Modified from Hedlund et al./Nature Microbiology, 2022

Phyla are the most distantly related lineages commonly used in prokaryotic systematics. A familiar phylum includes the Firmicutes (now also called the Bacillota), which includes genera such as Bacillus, Clostridium and Staphylococcus. Phyla are ancient groups that diverged billions of years ago. As such, members of different phyla vary greatly in their growth properties, cell structures and metabolisms. Since representatives of less than 1/3 of the phyla have been isolated as pure cultures, there is an enormous gap in our understanding of the breadth of prokaryote diversity.

Similarly, although estimates of the total number of prokaryote species vary widely, it is unlikely that more than 0.2 % have ever been cultured and named under the ICNP. At the current rate of description, it will take over 1,000 years to describe all the prokaryotic species believed to exist. Even in the human gut, which is probably the most thoroughly studied microbiome, many of the strains identified by metagenomic sequencing cannot be classified in known species or genera. In one study , 696 strains of prokaryotes were identified, but only 460 and 321 could be assigned to genera and species described under the ICNP, respectively. Thus, even in a well-studied microbiome closely associated with the human body, a large fraction of the prokaryotes remains unnamed.

## How to register names under the SeqCode

The SeqCode establishes the SeqCode Registry to create and validate names of prokaryotes based on genome sequences. Genome sequences are derived from genome assemblies, which are computed from the raw

sequence data to represent the best estimate of the actual genome sequence. For metagenomic studies comprising large-scale sequencing of environmental DNA, the genome assemblies must be >90% complete and 10-fold to serve as a type. In addition, the assembly of the type sequence and raw data must be available in an International Nucleotide Sequence Database Collaboration (INSDC) database, such as GenBank and the Short Read Archive. The raw sequences are required so that, as new assembly methods are developed, the sequence can be improved.

In a typical investigation, a MAG, SAG or other genome sequence is compared to the databases to determine if close relatives have been previously described. A common criterion that is used is the new sequence must possess an average nucleotide identity or ANI of