 doi:10.20944/preprints202010.0160.v1

The Next Million Names for Archaea and Bacteria

Mark J. Pallen[a,b,c]*, Andrea Telatin[a], Aharon Oren[d]

[a]Quadram Institute Bioscience, Norwich Research Park, Norwich, UK.
[b]School of Veterinary Medicine, University of Surrey, Guildford, Surrey, UK.
[c]University of East Anglia, Norwich Research Park, Norwich, UK.
[d]The Institute of Life Sciences, Edmond J. Safra Campus, The Hebrew University of Jerusalem, Jerusalem, Israel.

*Correspondence: mark.pallen@quadram.ac.uk (M.J.Pallen)

Keywords: nomenclature, Candidatus, metagenome-assembled genomes, genome-based taxonomy

 doi:10.20944/preprints202010.0160.v1

Abstract

Latin binomials, popularised in the eighteenth century by the Swedish naturalist Linnaeus, have stood the test of time in providing a stable, clear and memorable system of nomenclature across biology. However, relentless and ever-deeper exploration and analysis of the microbial world has created an urgent unmet need for huge numbers of new names for Archaea and Bacteria. Manual creation of such names remains difficult and slow and typically relies on expert-driven nomenclatural quality control. Keen to ensure the legacy of Linnaeus lives on in the age of microbial genomics and metagenomics, we propose an automated approach, employing combinatorial concatenation of roots from Latin and Greek to create linguistically correct names for genera and species that can be used off the shelf as needed. As proof of principle, we document over a million new names for Bacteria and Archaea. We are confident that our approach provides a road map for how to create new names for decades to come.

The Legacy of Linnaeus

In the eighteenth century, the Swedish naturalist Linnaeus proposed a hierarchical scheme of **taxonomy** (see Glossary) that assigned Latin **binomials** to biological species [1]. Shortly afterwards, the first genus names were applied to bacteria [2]. In the nineteenth century, Darwin's *Origin of Species* provided an evolutionary framework for taxonomy, confidently proclaiming, "Our classifications will come to be, as far as they can be so made, genealogies" [3]. In the twentieth century, Hennig brought clarity to evolutionary taxonomy through the development of *phylogenetic systematics*, now commonly called **cladistics**, which stipulated that biological classifications must represent the phylogenies of organisms and that taxa should represent **monophyletic groups** [4].

Linnaean binomials, drawing on combinations of **Latin** and **Ancient Greek** roots, have stood the test of time in providing a stable, clear and memorable system of nomenclature across biology. Efforts to codify bacterial nomenclature have culminated in the **International Code of Nomenclature of Prokaryotes** (ICNP or "the Code") [5], which sets out the rules for naming species of Archaea and Bacteria. The enduring legacy of Linnaeus in microbiology is evident from the remarkable success of the ICNP in overseeing the valid publication of names for over 3,400 bacterial and archaeal genera and over 20,000 bacterial and archaeal species. These names are superbly well documented in the online List of Prokaryotic names with Standing in Nomenclature (https://www.bacterio.net) [6].

Naming the unnamed millions

However, there are clearly a number of shortcomings to the current approach to nomenclature of Archaea and Bacteria. The most obvious is a deficiency in scale. Comparable efforts for eukaryotes document hundreds of thousands of genera and millions of species (http://www.catalogueoflife.org/annual-checklist/2019/info/about) [7]. Although estimates of the total number of bacterial and archaeal species vary from millions to billions, [8,9], even the most conservative figures amply document an unmet need for many millions of new names for genera and species of Archaea and Bacteria.

Another pressing problem is that most microbiologists follow Shakespeare in possessing, at best, “small Latin, less Greek” [10] and so are poorly equipped for creating well-formed binomials that comply with the rules of Latin grammar and are presented with clear, plausible etymological justifications (Box 1). Despite the publication of several “how-to” guides [11–13], this skills gap has led to propagation of numerous erroneous malformations—a high-profile example is the species epithet *pyloridis,* which even passed validation in the *International Journal of Systematic Bacteriology,* before it had to be corrected according to the rules of Latin grammar to *pylori* [14,15]. What’s more, bacteriologists are bound by the provisions of the Code, which include many detailed difficult rules and recommendations on how names should be formulated [5]. These exacting requirements mean that new names have to undergo time-consuming nomenclatorial quality control by a dwindling pool of experts, who are required to be conversant with classical languages, the Code and contemporary microbiology [16]. These problems are compounded by the custom of creating names on an *ad hoc* as-needed, just-in-time-fashion, which provides a non-stop drip-by-drip flow of work for nomenclatural experts.

Another key challenge stems from the exhilarating success of high-throughput sequencing and bioinformatics, which, twinned with **molecular phylogenetics**, represent a remarkable unifying force across the whole of biology, drawing together all cellular organisms into a single great tree of life (http://tolweb.org/tree/). Within microbiology, such advances have been driven by **culturomics** (high-throughput culture followed by whole-genome sequencing, which has delivered many hundreds of new species) and via metagenomics, which has delivered many thousands of **metagenome-assembled genomes** (MAGs), mostly from uncultured organisms [17,18]. In addition, bioinformatics analyses have enabled the development of a comprehensive **genome-based taxonomy**, GTDB, ranging from species up to domains [19].

While nomenclature has largely kept up with culturomics [20], valid publication of names for bacterial and archaeal species currently requires deposition of

cultured type strains in public repositories. This requirement controversially precludes application of the ICNP rules to uncultured organisms identified and characterised by metagenomics [21,22] (Box 2). One work-around is to apply the designation ***Candidatus*** (abbreviated to *Ca.*) to names for uncultured taxa [23]. Although the resulting names have no standing according to the Code, the *Candidatus* approach provides a clear, memorable and potentially stable nomenclature for uncultured species that mirrors the nomenclature for cultured species. However, so far barely more than 850 species-level *Candidatus* names have been published in the peer-reviewed literature [24].

Similarly, Latin names have yet to be assigned to the vast majority of new species or genera defined by genome-based taxonomies, which include not only those represented solely by MAGs, but also new taxa for which cultured strains are available. Instead, almost all new genera and species identified in these settings have been assigned unstable, confusing and hard to-remember alphanumerical identifiers. For example, in the current release of the GTDB (https://gtdb.ecogenomic.org Release 05-RS95 17th July 2020), there are over five thousand genera and over 186,000 species with only alphanumeric designations, while another 600 genera and 1765 species, split from existing taxa, are identified only by alphabetical suffixes added to existing names.

So, should we conclude that the legacy of Linnaeus is no longer relevant to microbiology in the age of genomes and metagenomes? Should we be happy to refer to a new species as, for example, UBA6965 or sp000063525? We believe the answer is a resounding “no!” However, high-throughput generation of taxa via sequence-based approaches clearly precludes the detailed attention usually applied to the one-by-one construction of Latin binomials. Instead, we propose that the problem can best be solved by automating the creation of well-formed names.

Exploring taxonomic namespace

To meet the need for a stable, clear and memorable nomenclature for the next million bacterial or archaeal species, we propose abandoning the current cottage-industry approach and instead advocate the automated creation of

names *en masse*, in advance of the need to allocate them to biological entities. Here, we are borrowing the concept of **namespace** from computer science—reconfiguring the problem to one of exhaustively exploring taxonomic namespace according to the existing rules to create millions of new names.

How is this even possible? The answer is an approach that reaches back to classical times: joining individual word roots from Greek or Latin together to create compound words with new meanings. For example, the first-century Greek geographer Strabo gave us *Rhinoceros*, combining Ancient Greek roots for “nose” and “horn”, while Linnaeus named the genus *Chrysanthemum* using the Ancient Greek roots for “gold” and “flower”. This principle has been systematised in the Code to create new genus names, with the rule that a connecting vowel -o- is used after Greek roots and -i- after Latin roots. Any new genus name that is created inherits its grammatical properties (including gender and declension) only from the last element in the word formation. As many as four roots have been combined to give us validly published genus names such as *Ectothiorhodospira* from the Greek roots for “outside-sulfur-rose-spiral” or *Allocatelliglobosispora* from the Greek and Latin roots for “another-chainlet-sphere-spore”.

The Great Automatic Nomenclator

We propose to extend this approach so that very large numbers of genus names can be created using an automated combinatorial concatenation of a relatively small set of starting terms. Let’s say we wish to explore a biome-specific generic namespace defined by combinations of three terms (Key Figure 1). If we select ten roots to be deployed in each of the initial, middle and final positions, then it becomes possible to create from just thirty roots ten-times-ten-times-ten = a thousand names with little effort—an approach we have already used to create names for several hundred new genera from the chicken gut microbiome [25].

To automate this approach, we have created a python script named the *Great Automatic Nomenclator* (or *Gan*: https://github.com/telatin/gan) after a short

story by Roald Dahl [26], or *garden* in Hebrew, reflecting its fertile productivity. The script takes as input tables of roots in a specified format and then performs combinatorial concatenation, taking into account ICNP rules governing use or elision of connecting vowels. In addition, because the input roots have already completed linguistic quality control, the new names are grammatically correct and come complete with etymological justifications that can be used in a **protologue**. Currently, Gan version 1.0 still requires detailed curation of the input files, some expertise in bioinformatics and produces rather basic outputs. However, we anticipate that the program will become more user-friendly and productive in subsequent versions.

The power of prefixes

Before exploring the full power of this combinatorial approach, let's take a quick look at an easy win in creating new names that reflect phylogenetic positions. Since the time of Linnaeus, when advances in taxonomy demand that a new taxon be split from an existing taxon, it has been common practice to add a short prefix (or less commonly, a suffix) to the existing name to create a new name, with an etymology that defines the new taxon as "related to but distinct from the" pre-existing taxon. This approach has already seen extensive use for names for Bacteria (https://lpsn.dsmz.de/text/genera-named-after-other-genera), using prefixes such as *neo-* (from the Greek for new) or *allo-* (from the Greek for other).

We have collated a list of thirty eight prefixes that can be used for this purpose (Table 1). We then used GAN to apply these prefixes to validly published names for Bacteria and Archaea, taking care to avoid use of the same prefix if it has already been used in a name. Using this approach, we have been able to create around over 130,000 new genus names and over 700,000 species names, complete with grammatical metadata and etymological justifications, that can be applied to sister taxa related to, but distinct from, already named taxa (Tables S1, File S1, Tables S2.1 and S2.2). Of course, many of these names will never be used, as the namespace is much larger than the number of new sister taxa that are likely to be discovered. However, as proof of immediate utility, this approach could be

applied to all genera marked in GTDB simply with an alphabetical suffix (*Bacillus_A*, *Bacillus_B* etc) to generate well-formed Latin names for over six hundred new genera. It is also worth noting that there are precedents for incorporating more than one prefix into a bacterial name (e.g. *Parapseudoflavitalea or Allopseudarcicella)*, so if we allow two prefixes to be added to all existing names (while avoiding using the same prefix twice), we would be able to generate over 4 million new genus names and 29 million new species names.

Flexible endings

Often in the past, final word elements for bacterial genus names have reflected cellular morphology—as in the ending *coccus* in *Enterococcus*, describing a coccus associated with the gut. However, if we are to create a set of names that can be applied flexibly to any bacterium or archaeon, particularly to uncultured genera, we need to use last word elements that can be used without knowledge of phenotypic characters (e.g. cellular or colonial morphology). We have therefore collated a set of last word elements that can be used in genus names derived from biomes and/or in association with proper nouns (Table 2). Use of such elements brings not just remarkable combinatorial power, but also minimises clashes with botanical and zoological codes.

People and places

Under the auspices of the ICPN, Archaea and Bacteria have often been named after places or people (mythical or real). In 2005, the nomenclature expert Hans Trüper expressed exasperation at excessive use of this approach with place names, which he termed *localimania*. However, the practice continues, with most salient example being use of *Massilia* (Latin name for Marseille) by the IHU Méditerranée Infection—a term that has found it way into over 260 species or genus names [20]. Usefully, these include validly published precedents for combining a proper noun with other roots, e.g. *Methanomassiliicoccus*. The way is thus open for combining names of places associated with identification of new taxa through genomic or metagenomic analyses with additional roots, including our set of last word-elements, e.g.

Brisbanimonas, Brisbanibacterium, for species delineated by the GTDB project in Brisbane (Key Figure 1).

Linnaeus made widespread use of the names drawn from mythology. This practice continues in microbiology. For example, the genus name *Cronobacter* was applied to a pathogen of children after *Cronos*, a Titan who swallowed his children as soon as they were born. Again, this approach provides a precedent for combining a proper noun with other roots in e.g. *Neptunicoccus* or *Poseidonocella* and paves the way for the creation of names for new taxa identified through genomic or metagenomic analyses. For example, combining our flexible end elements with names for over a hundred sea deities drawn from diverse cultures we have been able to create names for over a thousand marine microorganisms (Table S3, File S3).

The ICPN provides rules for a well-established approach for turning surnames into genus names by addition of Latin endings or diminutives—examples include *Escherichia* and *Salmonella.* Recently, this approach has been broadened into combining personal names with other roots, but so far only for a couple of dozen names [27] Application of this approach to the several hundred surnames that have already been used in genus names for bacteria and archaea would allow the creation of many thousands of new names (e.g. *Salmoniimonas*, *Salmoniiplasma*, *Salmoniimicrobium*). However, we note that many of those who created the conceptual and technical framework for microbial taxonomy have yet to be honoured in our discipline—for example, you look in vain for Archaea or Bacteria named after Carl Linnaeus, Charles Darwin or Willi Hennig! We have therefore compiled a gender-balanced list of eighty worthy scientists and have used our program to create 640 new names from this list, including *Darwiniibacterium* and *Hennigiimonas* (Table S5, File S5).

Binomials for Biomes

Another well-established approach for naming new taxa is to describe the habitat or biome in which the organism is found. For example, as we have noted *Enterococcus* describes a coccus found in the gut. However, in the age

of metagenomics and microbiome research, we need new genus names for inhabitants of each microbiome by the dozen or even in the hundreds.

Fortunately, this need can be easily met using our combinatorial approach to link our final word elements to terms that describe an organ or a host—for example, *Intestimonas*, for an microbe associated with the intestine or *Avimonas* for one associated with birds. As common names from classical languages for organs or animals typically provide multiple roots for the same organ/tissue (e.g. *faeci-, merdi, excrementi, stercori-, cacco-* for faeces) or for the same animal (e.g. *galli-*, *pulli-*, *alectryo-*, *cotto-* for chicken), this approach has allowed us to generate thousands of names for new genera from animal microbiomes, drawing on over 200 curated roots specifying organs, tissues or hosts (Table S4). The same approach can be applied to use of classical roots for biomes associated with plants, e.g. *Leguminimicrobium* for a microbe from beans, or with the abiotic environment, e.g. *Oceanimonas* for a microbe from the oceans or *Chthonomicrobium* for a subterranean microbe (Table S6, File S6).

This combinatorial approach proves particularly powerful when, as well as using common names from classical languages, one exploits the genus name of the host as a Neo-Latin term that can combined with other roots, e.g. *Drosophilimonas* or *Arabidopsidimicrobium.* As there are hundreds of thousands of named genera of eukaryotes, this opens up the creation of millions of names for host-associated bacteria genera. Similarly, adopting Neo-Latinised versions of technical terms for a particular biome, e.g. generating roots such as *nasopharyngo-, lotici-*, *bioreactori- or phylloplani-*, brings added precision and enhanced fecundity to the creation of names for the inhabitants of microbiomes.

Stepping up to three roots

The remarkable power of combinatorial concatenation steps up a gear when we move from two roots in a row to three. Here, we propose an approach in which the first root specifies a general context, e.g. a host, a general environment, a person or a place, while the second root specifies a more

specific context, such as an organ or tissue or a specific environment. Using our software on terms for animal hosts and their organs/tissues together with our final word elements, we have generated over a hundred thousand new genus names for inhabitants of animal microbiomes (Table S7, S8, File S7, S8). This approach could also be used for biomes from the abiotic environment, e.g. giving us *Chthonohydromonas* for a microbe from a subterranean water source. However, even more names can be created if existing host genus names or personal nouns are used in the first root position, for example *Triticirhizomicrobium*, *Darwiniintestimonas* or *Brisbaniiterriplasma*.

The species problem

So far, we have concentrated on the creation of genus rather than species names (aside from the use of prefixes). However, a similar principle of combinatorial concatenation of classical roots works here too, even though the grammatical context is slightly different (species epithets are typically genitive nouns or adjectives, even if nouns in the nominative case in apposition are occasionally used). However, unlike a genus name, a species epithet can be used again and again—for example, the epithet *massiliensis* has been used over a hundred times. Thus as typically only a few dozen species names are needed per genus, a pre-formed stock of names can easily be created for each biome by combining just two roots (*merdavium, faecavium, caccavium*, etc for species associated with bird faeces).

Concluding remarks and future perspectives

In 1999, the nomenclature expert Hans Trüper claimed “in view of the million names that will have to be formed in the future… [arbitrary names] are a simple necessity, whether Latin formalists like them or not.” [12]. *Contra* Trüper, here we have shown how combinatorial use of Greek and Latin roots could be used to create millions of well-formed taxonomic names for Bacteria and Archaea. What’s more, we have reduced this principle to practise in the documentation of a million names in the supplementary material.

In so doing, we have outlined a scalable system for filling taxonomic namespace that circumvents onerous and expert-dependent one-by-one creation of names—exploiting computational automation to deliver millions of names that are linguistically correct, meet the requirements of the ICNP and so can be used off the shelf, as needed. We are thus providing added impetus to efforts to create a nomenclature for uncultured organisms and hold a mirror up to the current failure to incorporate uncultured organisms into the Code. We expect that our approach could be broadened to cover the need for well-formed names for across the whole of the Darwin tree of life (https://www.darwintreeoflife.org). We have started a process that raises many questions (see Outstanding Questions), but we predict that one day, naming Bacteria and Archaea might be as easy as using Google Translate. In the meantime, we have provided a template showing how input files for Gan should be formatted (Table S9)

The software we have created for this purpose is freely available. However, it comes with the warning, “Caveat Nomenclator!” in that it will concatenate terms that simply don’t belong together. For example, *Gallidentimonas* might appear to be a well-formed Latin name, but it is nonsensical, as hens don’t have teeth. We must also stress that we have not created or even named any new taxa, merely provided software to generate names that could be used for this purpose—but *only* once they have been published in peer-reviewed journals and have been properly attached to nomenclatural types. For the time being, our names remain naked, as what the jargon calls ***nomina nuda***! The challenge now is for readers in the microbiology community to clothe them with strains, sequences, circumscriptions, positions and ranks.

Box 1: Why it's hard to create well-formed binomials

Latin remains the language of taxonomic nomenclature. This brings the advantages of neutrality and stability, but presents problems for most microbiologists, as Latin is no longer widely taught in schools. Unlike English, Latin is a highly inflected language, where the endings of nouns and adjectives vary according to their role in the sentence and linguistic properties. For example, adjectives change their endings to reflect the gender of noun they are qualifying, e.g. the neuter form *faecale* is used in *Microbacterium faecale*, but the masculine form *faecalis* in *Enterococcus faecalis.* Another problem is many taxonomic names come from Ancient Greek, which has its own alphabet, and so have to be transcribed into the Roman alphabet and then Latinised before use (experts still argue over whether *Acinetobacter* should have been *Akinetobacter*).

These problems are compounded by the fact that bacteriologists are also bound by the Code, which includes sixty-five rules and dozens of recommendations, some requiring subjective judgements, e.g. avoiding names that are "very long or difficult to pronounce". The Code insists that words from languages other than Latin or Greek should be avoided if equivalents exist in Latin or Greek, but allows genus names to be created in an arbitrary manner, so long as they are treated as Latin nouns.

All this makes it difficult for the non-expert to get things right, so that up to half of all newly proposed names for Archaea and Bacteria need to be corrected before use. Common problems include trying to use poorly Latinised English words (e.g. *geesorum* instead of *anserum* for a species associated with geese) or making up nonsensical etymologies [28]. The Code clarifies that names are primarily labels rather than descriptions: "The primary purpose of giving a name to a taxon is to supply a means of referring to it rather than to indicate the characters or the history of the taxon". The Code also explains what is required for names to be validly published, which includes a description of the taxon. Although not specified in the Code, valid publication of names is typically accompanied by a **protologue**, which includes a description of the taxon with an etymology and designation of type material.

Box 2: Culture Wars

Anyone seeking a stable system of nomenclature wouldn't start from where we are now. Rather than one code for all organisms, there are several, with different rules for naming plants, animals, fungi and prokaryotes. Cyanobacteria were for a long time treated as plants and so most still lack validly published names according to the ICNP. Oddly, names for phyla have never been included in the Code. And, for cladists, the term *prokaryote* is deprecated—because prokaryotes are no longer considered a monophyletic group [29]—and so, one could even argue that the ICNP is misnamed and should instead be named the International Code of Nomenclature of Archaea and Bacteria. It is also worth noting that there is no specific term for people who study Archaea and Bacteria—here, we have tended to use "bacteriologist", noting that the term "archaeologist" has been appropriated by another discipline.

Despite claiming "Nothing in this Code may be construed to restrict the freedom of taxonomic thought or action", the ICNP is very particular about what counts as type material in naming a species: only living pure cultures meet the requirement—an ironic contrast to the International Code of Nomenclature for Algae, Fungi and Plants, which requires that type material be dead or at least inert! This requirement for live cultures has attracted controversy. One objection is that almost all microorganisms live in communities and in challenging conditions, so organisms grown in pure culture, at best, provide an incomplete view of the natural world—and, at worst, represent a laboratory artefact [30].

A more pressing objection stems from the fact that most microbial species remain uncultured, even though they are increasingly accessible by metagenomics. Many molecular microbiologists now suggest that uncultured organisms should have their own names and that genome sequences should be acceptable as type material [22]. However, after lengthy debate, a proposal to amend the Code to accommodate this request was rejected in March 2020 [31].

In the meantime, we have to muddle through with *Candidatus* names, which although mentioned in the Code, have no priority. Does this matter? Probably not—as although the *de jure* position is that a *Candidatus* name could be replaced by a fresh name any time in the future, the *de facto* situation is that, for most microbiologists, it will be simply too much effort to create and validly publish new names, when perfectly good names already exist. After all, nearly half of the new names assigned to cultured organisms in journals other than the *International Journal of Systematic and Evolutionary Microbiology* are never followed through to valid publication which requires a request for inclusion in a Validation List in that journal [32]. Reassuringly, online resources as LPSN, NCBI and GTDB already incorporate *Candidatus* names and such *de facto* arrangements do a great job at allowing these names to be used while also preventing confusion over which names have which already been used, whether formally or informally.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2020 doi:10.20944/preprints202010.0160.v1

Table 1a Prefixes used to create new genus or species names

Prefix	Language	Part of speech	Classical term	Definition
alii	Latin	adjective	*alius*	other
allo	Greek	adjective	ἀλλος	other
alteri	Latin	adjective	*alter*	the other
amphi	Greek	preposition	ἀμφί	around, near
cognati	Latin	adjective	*cognatus*	related
crypto	Greek	adjective	κρυπτός	hidden
enantio	Greek	adjective	ἐνἄντΐος	opposite to
epi	Greek	preposition	ἐπί	on top of
extra	Latin	preposition	*extra*	beyond
falsi	Latin	adjective	*falsus*	false
hetero	Greek	adjective	ἕτερος	different
hosper	Greek	adverb	ὥσπερ	like
hyper	Greek	preposition	ὑπέρ	over, beyond
iso	Greek	adjective	*ἴσος*	same as
iuxta	Latin	adjective	*iuxta*	nearby
meso	Greek	adjective	μέσος	middle
meta	Greek	preposition	μετἄ	besides
neo	Greek	adjective	νέος	new
notho	Greek	adjective	νόθος	crossbred
novi	Latin	adjective	*novus*	new
paeni	Latin	adverb	*paene*	almost
para	Greek	preposition	παρά	beside
peri	Greek	preposition	περί	about
praeter	Latin	preposition	*praeter*	beyond
prope	Latin	preposition	*prope*	near
pseudo	Greek	adjective	ψευδής	false
quasi	Latin	conjunction	*quasi*	as if
simili	Latin	adjective	*similis*	similar
tele	Greek	preposition	τῆλε	far away
ultra	Latin	preposition	*ultra*	beyond

Table 1b Final word elements for bacterial or archaeal genera

Application	Element	Language	Part of speech	Gender	Definition
biomes	*adaptatus*	L.	adjectival noun	masc.	adapted to
biomes	*cola*	N.L.	suffix	masc./fem.*	an inhabitant of
biomes	*enecus*	N.L.	noun	masc.	inhabitant
biomes	*habitans*	L.	adjectival noun	masc./fem.*	an inhabitant
biomes	*vicinum*	N.L.	noun	neut.	a neighbour
biomes	*vivens*	N.L.	adjectival noun	masc./fem.*	living
biomes/proper names	*archaeum*	N.L.	noun	neut.	an archaeon#
biomes/proper names	*bacterium*	N.L.	noun	neut.	a bacterium#
biomes/proper names	*microbium*	N.L.	noun	neut.	a microbe
biomes/proper names	*monas*	N.L.	noun	fem.	a monad
biomes/proper names	*morpha*	N.L.	noun	fem.	form, shape
biomes/proper names	*ousia*	N.L.	noun	fem.	essence
biomes/proper names	*plasma*	N.L.	noun	neut.	a form
biomes/proper names	*soma*	N.L.	noun	neut.	a body

* Although either gender can be used, GAN makes it feminine by default
# to be used only for Archaea or Bacteria respectively

Acknowledgements

MJP is supported by the Quadram Institute Bioscience BBSRC-funded Strategic Program: Microbes in the Food Chain (project no. BB/R012504/1) and its constituent project BBS/E/F/000PR10351 (Theme 3, Microbial Communities in the Food Chain). AT gratefully acknowledges the support of the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Programme Gut Microbes and Health BB/R012490/1 and its constituent project BBS/E/F/000PR10353.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2020 doi:10.20944/preprints202010.0160.v1

Glossary

Ancient Greek: (abbreviated as Gr.) a classical language, the language of the many celebrated poets, playwrights and philosophers. Even in ancient times, many Greek words were carried over into Latin and this trend continues in the formation of taxonomic names.

Binomial: a Latinised name for biological species, written in italics and composed of two parts, the first capitalised and identifying the genus, the second identifying the species, e.g. *Escherichia coli.*

***Candidatus*:** a category of name for archaeal or bacterial taxa representing as-yet uncultured organisms. The Code grants no standing to such names, but specifies that they should be prefixed with *Candidatus* (in italics), the genus and species names should be in Roman type and the entire name in quotation marks; for example "*Candidatus* Phytoplasma allocasuarinae".

Cladistics: an approach to biological classification pioneered by the German entomologist Willi Hennig, in which organisms are grouped into monophyletic groups (also called clades) and classification strictly reflects phylogeny.

Culturomics: high-throughput culture of microbes as an approach to discover taxonomic novelty.

Etymology: a justification for the new name, including a description of constituent terms, their origins, grammatical properties and meanings.

Genome-based taxonomy: an approach to molecular phylogenetics in which genome sequences are used to create phylogenies and taxonomies: epitomised by the Genome Taxonomy Database (GTDB).

International Code of Nomenclature of Prokaryotes (the ICNP or 'the Code'): the set of rules and recommendations for naming Bacteria and Archaea maintained by the International Committee on Systematics of Prokaryotes.

Latin: (abbreviated as L.) a classical language, the dominant language of Ancient Rome, which has remained, as New Latin or Neo-Latin (abbreviated as N.L.), the language in which taxonomic names are framed.

List of Prokaryotic names with Standing in Nomenclature: an online database that documents the names of Archaea and Bacteria validly published under the Rules of the ICNP, together with effectively but not validly

published names, *Candidatus* names, and names of cyanobacterial taxa validly published under the Botanical Code.

Metagenome-assembled genome (MAG): a genome sequence that has been reconstructed by assembling and binning metagenomic reads.

Molecular phylogenetics: analysis of inherited differences in informational macromolecular sequences (DNA, RNA or proteins) to identify evolutionary relationships and construct phylogenetic trees.

Monophyletic group: another term for a clade, a taxonomic group that includes a common ancestor and all its descendants.

Nomina nuda (singular: *nomen nudum*), a term for names that look like taxonomic names but have no standing as they have not been published according to the rules of the relevant nomenclatural code.

Namespace: a term borrowed from computing to demarcate the full set of names used to refer to objects, ensuring that all objects have unique names.

Nomenclature: A system for giving names to organisms.

Protologue: a description of a new taxon, which includes the etymology of the name, a description of the taxon and of the designated nomenclatural type (a strain for a new species, but a species for a new genus).

Taxonomy: the branch of biology concerned with the classification, identification and nomenclature of organisms; can also refer to a particular scheme for categorisation.

References

1 Linnaeus (1759) Systema Naturae 10th Edition. Laurentius Salvius. Stockholm.
2 Muller, O.F. (1786) Animalcula infusoria fluviatilia et marina, quae detexit, systematice descripsit et ad vivum delineari curavit. Typis Nicolai Mölleri. Copenhagen and Leipzig.
3 Darwin, C. (1859) On the Origin of Species, 1st Edition. John Murray. London.
4 Hennig, W. (1950) Grundzuge Einer Theorie Der Phylogenetischen Systematik. Deutscher Zentralverlag. Berlin.
5 Parker, C.T. *et al.* (2019) International Code of Nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 69, S1–S111
6 Parte, A.C. *et al.* (2020) List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *Int J Syst Evol Microbiol*
7 Rees, T. *et al.* (2020) All genera of the world: an overview and estimates based on the March 2020 release of the Interim Register of Marine and Nonmarine Genera (IRMNG). *Megataxa* 1, 123–140
8 Dykhuizen, D.E. (1998) Santa Rosalia revisited: why are there so many species of bacteria. *Antonie van Leeuwenhoek* 73, 25–33
9 Amann, R. and Rosselló-Móra, R. (2016) After All, Only Millions. *mBio* 7, e00999–16
10 Jonson, B. (1623) To the memory of my beloved, The Author Mr William Shakespeare: And what he hath left us. In *Mr. William Shakespeares Comedies Histories & Tragedies: Published according to the True Originall Copies* Isaac Iaggard and Edward Blount
11 MacAdoo, T.O. (1993) Nomenclatural Literacy. In *Handbook of new bacterial systematics* (Goodfellow, M. *et al.* eds.), pp. 339–358, Academic Press San Diego, CA
12 Trüper, H.G. (1999) How to name a prokaryote? Etymological considerations, proposals and practical advice in prokaryote nomenclature. *FEMS Microbiology Reviews* 23, 231–249
13 Oren, A. (2015) How to Name New Taxa of Archaea and Bacteria. In *Bergey's Manual of Systematics of Archaea and Bacteria* (Whitman, W.B. *et al.* eds.), pp. 1–24, John Wiley & Sons, Ltd
14 Bacteriology., I.J.O.S. (1985) Validation of the publication of new names and new combinations previously effectively published outside the IJSB. List no. 17. *Int. J. Syst. Bacteriol.* 35, 223–225
15 Marshall, B.J. and Goodwin, C.S. (1987) Revised nomenclature of Campylobacter pyloridis. *International Journal of Systematic and Evolutionary Microbiology* 37, 68–68
16 Oren, A. *et al.* (2015) Wanted: microbiologists with basic knowledge of Latin and Greek to join our 'nomenclature quality control' team. *Int J Syst Evol Microbiol* 65, 3761–3762
17 Lagier, J.C. *et al.* (2018) Culturing the human microbiota and culturomics. *Nat Rev Microbiol* 16, 540–550
18 Parks, D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2, 1533–1542
19 Parks, D.H. *et al.* (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 38, 1079–1086
20 Lagier, J.C. *et al.* (2018) Naming microorganisms: the contribution of the IHU Méditerranée Infection, Marseille, France. *New Microbes New Infect* 26, S89–S95
21 Konstantinidis, K.T. *et al.* (2017) Uncultivated microbes in need of their own taxonomy. *ISME J* 11, 2399–2406
22 Murray, A.E. *et al.* (2020) Roadmap for naming uncultivated Archaea and Bacteria. *Nat Microbiol* 5, 987–994

23 Murray, R.G. and Stackebrandt, E. (1995) Taxonomic note: implementation of the provisional status Candidatus for incompletely described procaryotes. *Int J Syst Bacteriol* 45, 186–187
24 Oren, A. *et al.* (2020) Lists of names of prokaryotic Candidatus taxa. *Int J Syst Evol Microbiol* 70, 3956–4042
25 Gilroy, R. *et al.* (2020) A Genomic Blueprint of the Chicken Gut Microbiome. *Research Square* 10.21203/rs.3.rs-56027/v1,
26 Dahl, R. (1953) The Great Automatic Grammatizator. In *Someone Like You* Alfred A. Knopf
27 Oren, A. *et al.* (2019) Formation of compound generic names based on personal names: a proposal for emendation of Appendix 9 of the International Code of Nomenclature of Prokaryotes. *Int J Syst Evol Microbiol* 69, 594–596
28 Oren, A. (2020) Naming novel prokaryotic taxa discovered in the human gut. *Gut* 69, 969–970
29 Pace, N.R. (2009) It's time to retire the prokaryote. *Microbiology Today* 36, 84
30 Hobman, J.L. *et al.* (2007) Laboratory strains of Escherichia coli: model citizens or deceitful delinquents growing old disgracefully. *Mol Microbiol* 64, 881–885
31 Sutcliffe, I.C. *et al.* (2020) 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* 70, 4416–4417
32 Oren, A. *et al.* (2018) Why are so many effectively published names of prokaryotic taxa never validated. *Int J Syst Evol Microbiol* 68, 2125–2129