Naming the unnamed: Over 45,000 *Candidatus* names for unnamed Archaea and Bacteria in the Genome Taxonomy Database

MarkJ.Pallen[1,2,3][orcid.org/0000-0003-1807-3657]*,
Nabil-Fareed Alikhan[2] [orcid.org/0000-0002-1243-0767]

[1] Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, Norfolk, United Kingdom
[2] Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, United Kingdom
[3] School of Veterinary Medicine, University of Surrey, Guildford, Surrey, United Kingdom

***Corresponding author**
m.pallen@uea.ac.uk

Keywords
Bacterial nomenclature; archaeal nomenclature; genome taxonomy; shotgun metagenomics; Candidatus names

Abstract

Thousands of new bacterial and archaeal species and higher-level taxa are discovered each year through the analysis of genomes and metagenomes. The Genome Taxonomy Database (GTDB) provides hierarchical sequence-based descriptions and classifications for new and as-yet-unnamed taxa. However, bacterial nomenclature, as currently configured, cannot keep up with the need for new well-formed names. Instead, microbiologists have been forced to use hard-to-remember alphanumeric placeholder labels. Here, we exploit an approach to the generation of well-formed arbitrary Latinate names at a scale sufficient to name tens of thousands of unnamed taxa within GTDB. These newly created names represent an important resource for the microbiology community, facilitating communication between bioinformaticians, microbiologists and taxonomists, while populating the emerging landscape of microbial taxonomic and functional discovery with accessible and memorable linguistic labels.

Data summary

Input files for this study were obtained from the following sources

- Whitaker's Latin stems: http://archives.nd.edu/whitaker/wordsall.zip
- English Wiktionary headwords: https://dumps.wikimedia.org/enwiktionary/20210920/enwiktionary-20210920-pages-articles-multistream-index.txt.bz2
- genus names compiled by Global Biodiversity Information Facility: https://hosted-datasets.gbif.org/datasets/backbone/backbone-current-simple.txt.gz
- GDTB metadata and taxonomy files: https://data.gtdb.ecogenomic.org/releases/release202/202.0/
- Zenodo files: https://zenodo.org/record/5652886, includes input files ar_genus_endings.txt, bac_genus_endings.txt and species_endings.txt and all output files specified in the manuscript and the shell script *GTDB_renamer.sh*.

Python scripts and the *GTDB_renamer.sh* shell script used in this analysis are available on GitHub

- https://github.com/quadram-institute-bioscience/namingGTDB

Supplementary files:

- **Table S1** Creation and use of well-formed arbitrary, meaningless Latinate names in naming Archaea and Bacteria (Excel file with multiple tabs, carrying data from output files in the Zenodo archive)
- **Table S2** New names for Archaea and Bacteria
- **File S1** Archaeal Protologues (Word file).
- **File S2** Bacterial Protologues (Word file).
- **File S3** Renamed GTDB taxonomy file for use with the GTDB Toolkit (.tsv file)

Introduction

We microbiologists live in an age of genomic plenty [1]. Genome and metagenome analyses have fuelled exponential growth in the identification of new taxa of Archaea and Bacteria [2]. In addition, ready availability of genome sequences has primed the development of comprehensive sequence-based taxonomies, such as the Genome Taxonomy Database (GTDB) [3, 4]. However, such exhilarating success in discovering and classifying new microorganisms has created an urgent new challenge: how are we going to name all these newfound microbial taxa?

Linnaean binomials, typically 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 the tree of life [5, 6]. However, in the absence of mechanisms for the speedy creation of well-formed Latin names at scale, the GTDB team has adopted a system of alphanumeric placeholders for newly delineated taxa, which are hard to remember and are easily confused. The current system includes alphanumerical species epithets (e.g. sp011333035) derived from numerical identifiers of representative genomes and alphanumeric designations for genera and higher-level taxa (e.g. CG2-30-70-394), selected arbitrarily from identifiers in public sequence databases. These unnamed placeholder taxa now clearly outnumber those with names: in the latest release of the GTDB taxonomy [7] around a quarter of species (12,398 or 25.9%) have been assigned Latin names, while roughly three quarters of the species (35,496 or 74.1%) are identified only by alphanumeric placeholders labels.

Nomenclature of Archaea and Bacteria is governed by *The International Code for Nomenclature of Prokaryotes* (the ICNP*)* [8]. Most names for Archaea and Bacteria have been applied to taxa that can be maintained in stable culture. However, since the 1980s, there has been a growing recognition that uncultured taxa can nonetheless be identified and classified via analysis of macromolecular sequences [9–11]. To accommodate such uncultured taxa, Murray and Schleifer proposed a new category of taxonomic names, which they termed *Candidatus* [12]. Recommendations on the use of the category *Candidatus* are now included in the ICNP, alongside clarification that such names are provisional and enjoy no standing in nomenclature. In a recent review, one of us (Pallen) concluded that the category *Candidatus* continues to serve a useful function in providing well-formed names for uncultured taxa [13].

Most names for taxa of Archaea and Bacteria are descriptive or named after people or places. However, the use of arbitrary names has a long history in taxonomy and is clearly sanctioned within the ICNP [8, 14]. With these considerations in mind, here we address the challenge of "naming the unnamed" by creating nearly 60,000 arbitrary well-formed Latin names for Archaea and bacteria and then using a subset of them to assign *Candidatus* names to all unnamed taxa in GTDB.

Methods

Creation of arbitrary Latinate names

The workflow used to create and assign arbitrary Latinate names is shown in Figure 1. The workflow can be run using the Linux shell script *GTDB_renamer.sh*. To ensure selection of initial word-components consistent with the phonotactics of Latin (the rules governing how sounds are combined to create syllables), a publicly available set of Latin stems compiled by the Latinist William Whitaker was downloaded as STEMLIST.GEN from http://archives.nd.edu/whitaker/wordsall.zip in September 2021.

The Python script *stemlist_clean.py* was used to extract the relevant column from the file, to convert all entries to lower case and to remove duplicates. The resulting output file *Whitakers_stems.txt* was used as input for the Python script *openings_creator.py*. This script created a de-replicated set of five-letter Latin opening word-components that ended in consonants. The script excluded strings if they

- contained tandem vowels or difficult-to-pronounce consonant clusters
- contained word components that might carry unwanted connotations (e.g. vulgarities)
- began or ended with the letters j, k, w, y, z, absent from the original Latin alphabet.

To ensure that the opening word components used in name formation were free of meaning, the list was purged of entries identical to ~6 million headwords in the English Wiktionary, which include words from English, Latin and many other languages. In addition, the script excluded opening word-components identical to genus names already in use in taxonomy, by searching against a set of unique genus names compiled by the Global Biodiversity Information Facility. Wiktionary headwords were obtained from an index file downloaded in September 2021 from https://dumps.wikimedia.org/enwiktionary/20210920/enwiktionary-20210920-pages-articles-multistream-index.txt.bz2. Headwords were extracted from this file using the Python script *enwiki_clean.py* to give a file *enwiki_terms.txt* containing sorted de-replicated words in lowercase alphabetical characters. The entries in the file *GBIF_clean.txt* were extracted using a Python script *GBIF-clean.py* from the file downloaded via https://hosted-datasets.gbif.org/datasets/backbone/backbone-current-simple.txt.gz. The files *enwiki_terms.txt* and *GBIF_clean.txt* were concatenated into a file excluded_terms.txt, which was used by the *openings_creator.py* script.

To select final word components for names that comply with phonotactic and grammatical norms of Latin, a set of distinctive feminine Latin suffixes was selected from a list compiled by Wikipedia (https://en.wiktionary.org/wiki/Category:Latin_feminine_suffixes), alongside short nouns suitable for use as non-specific descriptors of microbes (Table 1). For ease of use, these final word components have been restricted to the feminine gender. To facilitate the use of genus names in creating names of higher-level taxa by addition of suffixes to the stem, five short final word-components belonging to the first declension were selected for use in bacterial genus names and incorporated into the file *bac_genus_endings.txt*. The single final word component *archa* (derived via Latinisation of the Greek noun, ἀρχή, meaning "beginning, origin") was reserved for the creation of names for archaeal genera and deployed in the file *ar_genus_endings.txt*. The remaining final word-components, which belong to the first and third declensions, were incorporated into the file *species_endings.txt*.

The Python script *name-creator.py* was run three times to combine all opening word-components with all final word-components in the *bac_genus_endings.txt*, *ar_genus_endings.txt* and *species_endings.txt files* to create the output files *bac_genus_names.txt*, *ar_genus_endings.txt* and *species_names.txt*. As with the

Assignment of arbitrary names to unnamed taxa

The Python script, *name_table_maker.py* was used to extract all alphanumeric taxon names from the latest versions of the GTDB taxonomy files for Archaea and Bacteria (*ar122_taxonomy_r202.tsv* and *bac120_taxonomy_r202.tsv*) and to sort them by rank in the taxonomic hierarchy (i.e. listing phylum designations before class designations etc). The script used the *archaeal_genus_names.txt*, *bacterial_genus_names.txt* and *species_names.txt* files as input and after randomizing the order of the genus names and species names and then sorting genus names by simplicity (deploying those with no double consonants first), the script assigned Latin names to the unnamed taxa Archaea and Bacteria identified in the GTDB taxonomy files.

The Python script *taxon_renamer.py* was then used to replace all alphanumeric designations with Latinate names in the GTDB taxonomy and metadata files for Archaea and Bacteria to create taxonomy files suitable for use with the GTDB toolkit [15], together with metadata files providing relevant details on the newly named taxa. New names were marked by the addition of an exclamation mark so that they could be easily distinguished from existing names in the taxonomy and metadata files. However, unmarked versions of the files were created suitable for use by the GTDB toolkit in assigning genomes to taxa.

The Python scripts *archaeal_protologue_maker.py* and *bacterial_protologue_maker.py* were used to output relevant descriptive information from the renamed GDTB metadata files in Rich Text Format (RTF) to create protologues suitable for publication of *Candidatus* names for use in the scientific literature and in public databases.

Results

Creation of arbitrary Latinate names

After extracting all unique initial five-letter-strings from the Latin stems compiled by William Whitaker, excluding unwanted clusters of letters and ensuring that all the strings did not form meaningful words in English, Latin and other major languages, we were left with 1,474 suitable opening word-components (Table S1). When these were combined with our curated set of final word-components used in Latin—while taking care to exclude terms with pre-existing meanings or genus names already used in taxonomy—we created 1,473 arbitrary, meaningless names for archaeal genera, 9,970 arbitrary, meaningless names for bacterial genera and 48,530 species epithets suitable for naming species from either domain (Table S1).

Assignment of arbitrary names to unnamed taxa

After retrieving ~42,000 alphanumeric placeholder names from the GTDB taxonomy files, we built name-replacement tables showing the placeholder names alongside replacement well-formed Latin names (Table S1). Chuvochina *et al* have emphasized the importance of specifying type material for uncultured taxa [16]. GTDB makes an arbitrary choice as to which placeholder labels should be propagated from the level of genus up to higher taxa, which, in turn, specifies which genus should be considered to act as type genus for those higher taxa. We respected the choices made by GTDB for type genera, replacing alphanumeric designations with arbitrary names wherever they occur in the taxonomy hierarchy, using the new name for each type genus to build names by addition of suffixes specified in the ICNP and elsewhere [8, 17].

Following the GTDB's lead in making an arbitrary choice of type material, we made an arbitrary choice as to which species should be designated the type species for a previously unnamed genus. GTDB also adopts the approach of giving each unnamed species a unique alphanumeric designation. Although, under the rules of the ICNP, there is no need for each species epithet to be unique, for simplicity and consistency, we have followed GTDB's practice and assigned a unique species epithet to each of the unnamed species in the current database.

The name-replacement tables were used to replace placeholders with the new names in the GTDB taxonomy and metadata files. In so doing, we assigned 650 new archaeal genus names and 1,806 new archaeal species names along with 8,607 new bacterial genus names and 31,072 new bacterial species names. Unused genus names and species epithets were retained for use with later releases of the database. When the new genus names were used to build new names for higher-level taxa, we created new names for 42 bacterial and 4 archaeal phyla (Table 2); 14 archaeal classes, 64 archaeal orders and 251 archaeal families; plus 2050 bacterial families, 749 bacterial orders and 188 bacterial classes. For 320 taxa, GTDB uses placeholders for higher-level taxa that have not been applied to a genus. In these cases, we have replaced the placeholders with new Latinate names in the taxonomy files, but have left these out of the protologues, as we anticipate that these anomalies will be addressed in subsequent releases of the GTDB.

Although the renamed GTDB metadata files contain a rich set of descriptive data for the newly named taxa, mindful that some authorities prefer to see descriptions of new taxa in a text-based format, we have created two RTF files—a file format that can be opened by most word processors—containing protologues for all the new *Candidatus* taxa (File S1, File S2). However, as these traditional protologues take up over 8,000 pages, they cannot be presented in the main body of this manuscript. The *GTDB_renamer.sh* script also created tables of new names from these protologue files, with the new names ordered by taxonomic rank (phyla shown in Table 2, all taxa shown in Table S2). We confirmed that the renamed GTDB taxonomy file (File S3) worked as expected by running the GTDB toolkit over a set of metagenome-associated genomes from the horse gut (data not shown). This required replacing the original taxonomy file used by the GTDB Toolkit with renamed GTDB taxonomy file, while retaining the original file name expected by the program, *gtdb_taxonomy.tsv.*

Discussion

Conventional approaches to the creation of new names for Archaea and Bacteria generally rely on descriptive names or names associated with people or places. Over the last ten years, such approaches have delivered around a thousand new validly published species names per year and tens-to-hundreds of *Candidatus* names annually [13, 18]. However, given a backlog of > 32,000 well classified but unnamed species in GTDB, using conventional approaches at the current rate of progress would take at least thirty years to name all these unnamed taxa—by which time, we would be faced with the problem of naming hundreds of thousands more newly discovered species! Recently, Pallen *et al* [19] described an approach enabling automated creation of descriptive names *en masse*. However, deployment of functionally descriptive names at the scale needed here would require reconstruction of the phenotypic properties of tens of thousands of species, which is a non-trivial and error-prone task. Similarly, assigning names based on habitat requires exhaustive searches of genome metadata to ensure names are accurate and precise. In addition, there is a trade-off here between the semantic specificity of a descriptive name and its length and usability, as is evident from recently proposed names such as *Hoministercoradaptatus ammoniilyticus*, *Anthropogastromicrobium aceti* and *Porcipelethomonas ammoniilytica* [20].

Here, given the failure of conventional approaches, we have taken what might seem like a radical step in creating and applying arbitrary well-formed Latinate names to unnamed uncultured taxa of Archaea and Bacteria. However, formation of names in an arbitrary fashion has a long tradition in taxonomy. Linnaeus created arbitrary names via anagrams, e.g. the genus name *Mahernia* as an imperfect anagram of *Hermannia* [6]. In the 1830s, the eminent English botanist John Lindley wrote: “So impossible is it to construct generic names that will express the peculiarities of the species they represent, that I agree with those who think a good, well-sounding, unmeaning name as good as any that can be contrived” [21]. Soon after, Scottish naturalist George Johnston created the arbitrary genus name *Carinella* for a marine annelid [22].

In 1869, in his groundbreaking *Lois de Nomenclature Botanique* [23], the Swiss botanist Alphonse De Candolle wrote: “Generic names are drawn from certain characters, from certain appearances… and even from combinations of letters that are quite arbitrary. All that is required of a name is that it shall lead neither to confusion nor to error…” Early 20th-century biologist William Kearfott created over a hundred arbitrary rhyming species epithets—e.g. *bana, cana, dana; bobana, cocana, dodana; boxcana, coxcana, doxcana*—many of which still belong to validly published names in use today [24].

In 1952, palaeobiologist Raymond Casey invented the Greek-sounding name *Gythemon* for an extinct clam, cited in the Zoological Code as a name built from ‘an arbitrary combination of letters’ [25]. A few years later, the botanist Gordon Rowley professed that “names are more often than not mere handles, and the most that we can ask of a handle is that it be neat and easy to grasp”. The current version of the ICNP also makes clear "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" and names "may… be composed in an arbitrary manner" [8].

In fact, there are many precedents for the arbitrary formation of names in bacterial nomenclature, including names derived from organizational acronyms, such as *Cedecea* (from CDC), names derived in an arbitrary fashion from personal names, such *Simkania* (after Simona Kahane) or names created from arbitrary contractions of functional descriptions, such as *Methermicoccus* [14].

Principle 3 of the ICNP [8] states "scientific names of all taxa are Latin or latinized words treated as Latin regardless of their origin". In English, Latinate terms form a lexical stratum with a distinctive phonotactic "Latinity", associated with scientific or scholarly writing [26, 27]. Here, to comply with the requirement of the ICNP, while also respecting the look and feel of the language, we have used words or syllables derived from Latin as components of new names. The result is a set of names that recall the familiarity and gravitas of Latin, even though they are devoid of etymology or meaning.

According to the strictest interpretation, *Candidatus* names can be assigned only to uncultured taxa and, here, we have made the assumption that all placeholder names in GTDB are associated with uncultured taxa. However, as Pallen [13] has argued elsewhere, even if a small number of the renamed taxa do already have cultured representatives, we can safely fall back upon the broader definition of *Candidatus* as a category ‘used for describing prokaryotic entities ... for which characteristics required for description according to the Code are lacking’.

The fact that *Candidatus* names have no standing in nomenclature is often seen as a deficiency of the current system of bacterial nomenclature [28]. However, the provisional status of the names proposed here can be seen as helpful, in that if, in the future, those working on a given taxon wish to propose a new name, they are perfectly entitled to do so within the current system—albeit with the proviso that the opening principle of the ICNP is to “aim for stability in names”. Aside from such changes, the vast majority of the *Candidatus* names proposed here are likely to remain highly stable—given that only 0.26% of genomes are on average assigned to a different species cluster from one release of GTDB to the next [7]—and so can now be safely adopted by databases and used in the scientific literature.

The scale of our efforts here, together with the fact that we are naming taxa that have already been delineated and classified by others, begs the question "Who has the right to create and assign taxonomic names?" Although the ICNP says nothing on this issue, traditionally the task of naming new cultured species has fallen to those who isolate and discover the species and deposit type material in culture collections. As this often requires a lot of work, the act of naming can be seen as a reward for the "sweat of one's brow". However, it remains unclear whether similar principles can be applied to the naming of uncultured species defined only by sequence analysis, when who can say who should be rewarded with a stake in the process: those who collected samples, those who sequenced them, those who binned reads into metagenome-assembled genomes or those who performed the sophisticated phylogenetic analyses delineating and classifying uncultured taxa? In any case, so far, none of these parties has shown interest in—or developed competing methodologies for—creating new names at scale.

It is worth stressing that the age of microbial discovery is far from over. Each new GTDB release is going to bring a fresh deluge of new taxa needing new names. Fortunately, additional names remain available from the set created here, which can be used in the near future. Further ahead, new names can be created following the principles established here with only minor changes to procedures or input files (e.g. including Greek stems, longer initial word-components or additional final word-components).

Drawing on the encouraging precedent with new names for SARS variants of concern [29], we hope that the names proposed here are rapidly adopted by the scientific community and used widely. After all, the alternative remains continuing use of confusing, hard-to-remember alphanumeric designations into the indefinite future. Instead, we propose a system of bacterial nomenclature fit for the age of genomics and ambitious enough to cope with the exciting discoveries yet to come.

Author statements

Authors and contributors

Conceptualization, MJP; Data curation, Methodology: MJP and NFA; Writing: MJP.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

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) and by the Medical Research Council CLIMB-BIG-DATA grant MR/T030062/1. NFA is supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1).

Acknowledgements

We thank Andrea Telatin for help in creating the first version of the *taxon_renamer.py* Python script. We thank Phil Hugenholtz for providing feedback on an early draft of the manuscript. We thank Rachel Gilroy for confirming that the renamed GTDB taxonomy files worked as expected with the GTDB Toolkit.

Figures and tables

Table 1 Final word-components used to create arbitrary names

The feminine Latin suffixes were selected from a list compiled by Wikipedia (https://en.wiktionary.org/wiki/Category:Latin_feminine_suffixes) and were used alongside short nouns suitable for use as non-specific descriptors of microbes. The resulting names, were all defined as feminine Neo-Latin nouns, which when used as species epithets, are deployed as nouns used in apposition.

Component (nom., gen.)	Declension	Application	Derivation
archa, archae	1	Archaeal genera	Latinised derivative of Gr. fem. n. ἀρχή, beginning, origin
ana, anae	1	Bacterial genera	Latin suffix to a noun stem to form an adjective
aria, ariae	1	Bacterial genera	Latin suffix used to form abstract nouns from other nouns
ella, ellae	1	Bacterial genera	Latin suffix used to form a diminutive of a noun
ia, iae	1	Bacterial genera	Latin suffix used to form an abstract noun
osa, osae	1	Bacterial genera	Latin suffix used to form adjectives from nouns meaning "full of"
ula, ulae	1	Bacterial genera	Latin suffix used to form a diminutive of a noun
astra, astrae	1	Species	Latin suffix of nouns, expressing resemblance
atica, aticae	1	Species	Latin suffix used to form adjectives indicating a relation to the root noun
entia, entiae	1	Species	Latin suffix used to form an abstract noun
etta, ettae	1	Species	Latin suffix used to form a diminutive of a noun
ibra, ibrae	1	Species	Latin suffix of nouns denoting instrument, vessel, place or person
ibula, ibulae	1	Species	Latin suffix of nouns denoting instrument, vessel, place or person
icella, icellae	1	Species	Connecting vowel -i-; L. fem. n. *cella*, a cell
icula, iculae	1	Species	Latin suffix used to form a diminutive of a noun
ifica, ificae	1	Species	Latin suffix forming adjectives that denote bringing or making.
iforma, iformae	1	Species	Connecting vowel -i-; L. fem. n. *forma*, a form
igena, igenae	1	Species	Latin suffix meaning "born from, sprung form"
ilega, ilegae	1	Species	Latin suffix forming adjectives related to the concept of collecting
ilenta, ilentae	1	Species	Latin suffix forming adjectives meaning "abounding in, full of"
imonia, imoniae	1	Species	Latin suffix used to form abstract nouns from adjectives
isca, iscae	1	Species	Late Latin suffix used to form adjectives
issa, issae	1	Species	Late Latin suffix used to form feminine forms of masculine nouns.
itia, itiae	1	Species	Latin suffix to form an abstract noun describing the condition of being something.
itica, iticae	1	Species	Latin suffix added to nouns to form adjectives
itoga, itogae	1	Species	Connecting vowel -i-; L. fem. n. *toga*, a covering, garment, used non-specifically in many bacterial names
itura, iturae	1	Species	Latin suffix used to form a noun relating to an action.
ivita, ivitae	1	Species	Connecting vowel -i-; L. fem. n. *vita*, life
ousia, ousiae	1	Species	Gr. fem. n. *ousia*, essence
ago, aginis	3	Species	Latin suffix , forms nouns describing objects, plants, and animals.
atio, ationis	3	Species	Latin suffix added to a verb to form an abstract noun

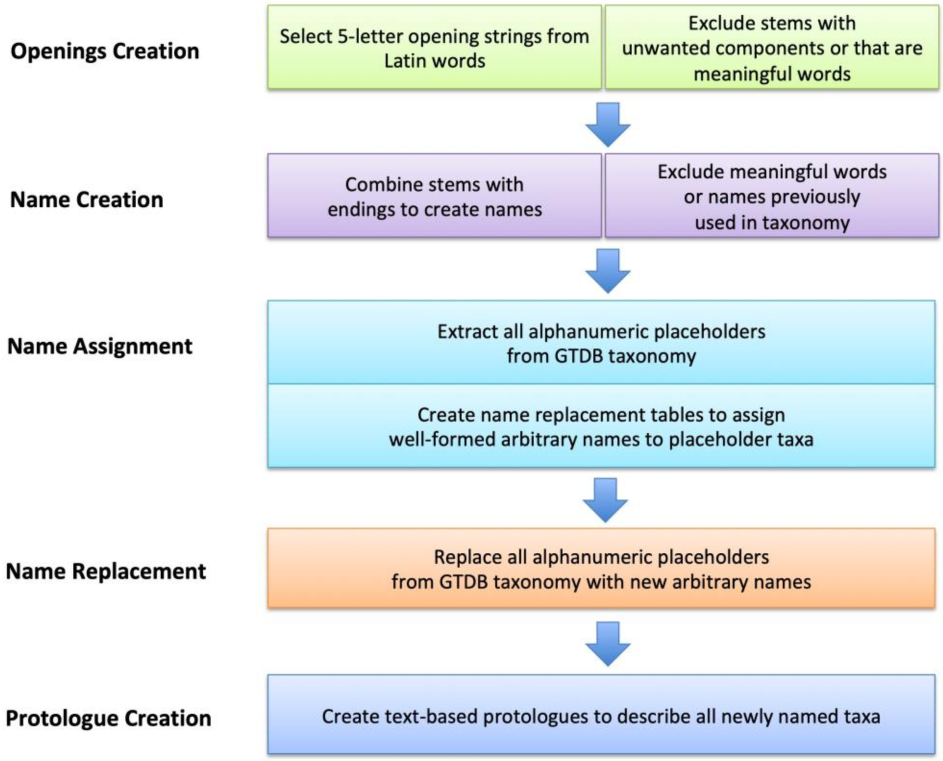
ax, acis	3	Species	Latin suffix used to form adjectives expressing a tendency or inclination
edo, edinis	3	Species	Latin suffix used to form abstract nouns
imonas, imonadis	3	Species	Connecting vowel -i-; L. fem. n. *monas*, a unit, monad
itas, itatis	3	Species	Latin suffix used to form abstract nouns indicating a state of being.
itio, itionis	3	Species	Latin suffix used to form a noun relating to an action.
itrix, itricis	3	Species	Latin suffix used to form a feminine agent noun
itudo, itudinis	3	Species	Latin suffix forming an abstract noun indicating a state
ugo, uginis	3	Species	Latin suffix forms nouns denoting coating of material

Table 2 Newly Named Phyla

New name	GTDB Placeholder	Domain
Candidatus Bifararchota *phyl. nov.*	EX4484-52	Archaea
Candidatus Funivarchota *phyl. nov.*	SpSt-1190	Archaea
Candidatus Genufarchota *phyl. nov.*	QMZS01	Archaea
Candidatus Nutamarchota *phyl. nov.*	PWEA01	Archaea
Candidatus Bicoriota *phyl. nov.*	B130-G9	Bacteria
Candidatus Bifisiota *phyl. nov.*	GCA-001730085	Bacteria
Candidatus Casigariota *phyl. nov.*	UBA9089	Bacteria
Candidatus Ceracanota *phyl. nov.*	T1Sed10-126	Bacteria
Candidatus Cibaranota *phyl. nov.*	DUMJ01	Bacteria
Candidatus Ciminosota *phyl. nov.*	ARS69	Bacteria
Candidatus Cinipanota *phyl. nov.*	SLNR01	Bacteria
Candidatus Cocibosota *phyl. nov.*	CG2-30-53-67	Bacteria
Candidatus Colepanota *phyl. nov.*	CG2-30-70-394	Bacteria
Candidatus Conipanota *phyl. nov.*	JABMQX01	Bacteria
Candidatus Culilariota *phyl. nov.*	UBA8481	Bacteria
Candidatus Cululellota *phyl. nov.*	JACIXR01	Bacteria
Candidatus Cululosota *phyl. nov.*	FEN-1099	Bacteria
Candidatus Decunulota *phyl. nov.*	QNDG01	Bacteria
Candidatus Defacidota *phyl. nov.*	CLD3	Bacteria
Candidatus Defarosota *phyl. nov.*	DTU030	Bacteria
Candidatus Deforanota *phyl. nov.*	CAIJMQ01	Bacteria
Candidatus Dehinanota *phyl. nov.*	OLB16	Bacteria
Candidatus Desonellota *phyl. nov.*	AABM5-125-24	Bacteria
Candidatus Domigosota *phyl. nov.*	UBA10199	Bacteria
Candidatus Feditulota *phyl. nov.*	SM23-31	Bacteria
Candidatus Foricariota *phyl. nov.*	RUG730	Bacteria
Candidatus Funamanota *phyl. nov.*	UBA2233	Bacteria
Candidatus Lanifulota *phyl. nov.*	4572-55	Bacteria
Candidatus Laxitanota *phyl. nov.*	CG03	Bacteria
Candidatus Lenunidota *phyl. nov.*	CSSED10-310	Bacteria
Candidatus Levifellota *phyl. nov.*	SpSt-1050	Bacteria
Candidatus Minagiota *phyl. nov.*	CSP1-3	Bacteria
Candidatus Musulosota *phyl. nov.*	J088	Bacteria
Candidatus Nicotellota *phyl. nov.*	SpSt-318	Bacteria
Candidatus Relapellota *phyl. nov.*	BMS3Abin14	Bacteria
Candidatus Resuliota *phyl. nov.*	RBG-13-66-14	Bacteria
Candidatus Rinocanota *phyl. nov.*	SZUA-182	Bacteria
Candidatus Rubefidota *phyl. nov.*	JABDJQ01	Bacteria
Candidatus Satacellota *phyl. nov.*	DQWO01	Bacteria
Candidatus Semimosota *phyl. nov.*	DSWW01	Bacteria
Candidatus Semunulota *phyl. nov.*	UBA6262	Bacteria
Candidatus Subadanota *phyl. nov.*	T1SED10-198M	Bacteria
Candidatus Televariota *phyl. nov.*	SpSt-956	Bacteria
Candidatus Venunellota *phyl. nov.*	UBA3054	Bacteria
Candidatus Volubiota *phyl. nov.*	RBG-13-61-14	Bacteria
Candidatus Bicoriota *phyl. nov.*	B130-G9	Bacteria

Figure 1

Schematic workflow for creation and assignment of arbitrary Latinate names. Opening 5-letter strings from Latin words were selected and curated to exclude unwanted components and meaningful words. These were combined with Latin suffixes or words to create names, which were then used to name placeholder taxa in GTDB files and create text-based protologues.


Openings Creation
Select 5-letter opening strings from Latin words
Exclude stems with unwanted components or that are meaningful words
Name Creation
Combine stems with endings to create names
Exclude meaningful words or names previously used in taxonomy
Name Assignment
Extract all alphanumeric placeholders from GTDB taxonomy
Create name replacement tables to assign well-formed arbitrary names to placeholder taxa
Name Replacement
Replace all alphanumeric placeholders from GTDB taxonomy with new arbitrary names
Protologue Creation
Create text-based protologues to describe all newly named taxa

References

1. **Loman NJ, Pallen MJ.** Twenty years of bacterial genome sequencing. *Nat Rev Microbiol.* 2015;13:787-794. DOI: 10.1038/nrmicro3565
2. **Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ *et al.*** Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol.* 2017;2:1533-1542. DOI: 10.1038/s41564-017-0012-7
3. **Rosselló-Móra R.** Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. *Environ Microbiol.* 2012;14:318-334. DOI: 10.1111/j.1462-2920.2011.02599.x
4. **Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ *et al.*** A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol.* 2020;38:1079-1086. DOI: 10.1038/s41587-020-0501-8
5. **Austin D.** The Nuance and Wit of Carolus Linnaeus. *The Palmetto.* 1993;13:8.
6. **Linnaeus C**. *Systema Naturae 12th Edition.* Stockholm: Laurentius Salvius; 1759
7. **Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA *et al.*** GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 2021gkab776. DOI: 10.1093/nar/gkab776
8. **Parker CT, Tindall BJ, Garrity GM.** International Code of Nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology.* 2019;69:S1-S111. DOI: DOI 10.1099/ijsem.0.000778
9. **Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA.** Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol.* 1986;40:337-365. DOI: 10.1146/annurev.mi.40.100186.002005
10. **Woese CR.** Bacterial evolution. *Microbiol Rev.* 1987;51:221-271.
11. **Relman DA.** The identification of uncultured microbial pathogens. *J Infect Dis.* 1993;168:1-8. DOI: 10.1093/infdis/168.1.1
12. **Murray RG, Schleifer KH.** Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes. *Int J Syst Bacteriol.* 1994;44:174-176. DOI: 10.1099/00207713-44-1-174
13. **Pallen MJ.** The status *Candidatus* for uncultured taxa of Bacteria and Archaea: a SWOT analysis. *Int J Syst Evol Microbiol.* 2021;71:005000.
14. **Pallen MJ.** Bacterial nomenclature in the era of genomics. *New Microbes and New Infections.* 2021;44:100942.
15. **Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH.** GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics.* 2019DOI: 10.1093/bioinformatics/btz848
16. **Chuvochina M, Rinke C, Parks DH, Rappé MS, Tyson GW *et al.*** The importance of designating type material for uncultured taxa. *Syst Appl Microbiol.* 2019;42:15-21. DOI: 10.1016/j.syapm.2018.07.003
17. **Whitman WB, Oren A, Chuvochina M, da Costa MS, Garrity GM *et al.*** Proposal of the suffix -ota to denote phyla. Addendum to 'Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes'. *Int J Syst Evol Microbiol.* 2018;68:967-969. DOI: 10.1099/ijsem.0.002593
18. **LPSN.** List of Prokaryotic names with Standing in Nomenclature. 2021. https://www.bacterio.net
19. **Pallen MJ, Telatin A, Oren A.** The Next Million Names for Archaea and Bacteria. *Trends Microbiol.* 2021;29:289-298. DOI: 10.1016/j.tim.2020.10.009
20. **Hitch TCA, Riedel T, Oren A, Overmann J, Lawley TD *et al.*** Automated analysis of genomic sequences facilitates high-throughput and comprehensive description of bacteria. *ISME Communications.* 2021;1DOI: 10.1038/s43705-021-00017-z
21. **Lindley J**. *An Introduction to Botany.* Longman, Orme, Brown, Green, and Longmans; 1839
22. **Johnston G.** Illustrations in British Zoology. *Magazine of natural history and journal of zoology,*

botany, mineralogy, geology and meteorology. 1833;6:233-235.

23. **de Candolle A**. *Lois de la Nomenclature Botanique.* Masson; 1867
24. **Kearfott WD.** New North American Tortricidae. *Transactions of the American Entomological Society.* 1907;33:1-98.
25. **Rawson PF, Rushton AWA, Simpson MI.** Raymond Charles Casey 10 October 1917 – 26 April 2016. *Biogr. Mems Fell. R. Soc.* 2020;68:71-86. DOI: 10.1098/rsbm.2019.0050
26. **Chomsky N, Halle M**. *The Sound Pattern of English.* MIT Press (MA); 470: 1968
27. **Hayes B**. Comparative phonotactics. 2016. p. 265-285.
28. **Murray AE, Freudenstein J, Gribaldo S, Hatzenpichler R, Hugenholtz P *et al.*** Roadmap for naming uncultivated Archaea and Bacteria. *Nat Microbiol.* 2020;5:987-994. DOI: 10.1038/s41564-020-0733-x
29. **Konings F, Perkins MD, Kuhn JH, Pallen MJ, Alm EJ *et al.*** SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nat Microbiol.* 2021;6:821-823. DOI: 10.1038/s41564-021-00932-w 10.1038/s41564-020-0770-5