**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 Latin 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 on April 22 2022

- Whitaker's Latin stems: http://archives.nd.edu/whitaker/old/wordsall.zip
- English Wiktionary headwords: https://dumps.wikimedia.org/enwiktionary/latest/enwiktionary-latest-pages-articles-multi-stream-index.txt.bz2
- Genus names compiled by Global Biodiversity Information Facility: https://hosted-datasets.gbif.org/datasets/backbone/current/simple.txt.gz
- GTDB metadata and taxonomy files: https://data.gtdb.ecogenomic.org/releases/release207/207.0/
- Zenodo files: https://zenodo.org/record/6477137, includes output files specified in the manuscript and the shell script GTDB_renamer.sh

Python scripts and the shell script used in this analysis are available on GitHub:

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

**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, the ready availability of genome sequences has primed the development of comprehensive sequence-based taxonomies, such as the Genome Taxonomy Database (GTDB) [3–5]. However, such exhilarating success in discovering and classifying new micro-organisms 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 [6, 7]. 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 alphanumeric designations for genera and higher-level taxa (e.g. CG2-30-70-394), selected arbitrarily from identifiers in public sequence databases, together with unique alphanumerical species epithets (e.g. sp011333035) derived from numerical identifiers of representative genomes. These unnamed placeholder taxa now decisively outnumber those with Latin names. For example, in the latest GTDB release (r207), over 75% of species (~50000 out of ~65000) are identified only by alphanumeric placeholders while 81 phyla remain without well-formed names [8].

Nomenclature of Archaea and Bacteria is governed by the International Code for Nomenclature of Prokaryotes (ICNP) [9]. 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 [10–12]. To accommodate such uncultured taxa, Murray and Schleifer proposed a new category of taxonomic names, which they termed Candidatus [13]. 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 (M.J.P.) concluded that the category Candidatus continues to serve a useful function in providing well-formed names for uncultured taxa [14].

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 [9, 15]. With these considerations in mind, here we address the challenge of ‘naming the unnamed’ by creating over 65000 Candidatus names for unnamed Archaea and Bacteria in the GTDB.

**METHODS**

**Correction of anomalies**

Despite the importance of designating type material for uncultured taxa [16], the current release of GTDB contains some anomalies. In particular, some placeholders for high-level taxa have not been derived from those for lower-level taxa (e.g. the placeholder FCPU426 is used to identify a phylum, even though there is no class, order, family or genus with that designation). In addition, some placeholder names have been applied to higher-level taxa even though a taxon contains genera with that designation. To address the challenge of ‘naming the unnamed’ by creating over 65000 Candidatus names for unnamed Archaea and Bacteria in the GTDB.

We first used a script anomaly_table_maker.py to detect and tabulate such anomalies. However, as automated correction of such anomalies proved unworkable, we relied on detailed manual inspection of anomalous names and phylogenies to compile a list of proposed substitutions (Table S1, available with the online version of this article). These were then applied to the GTDB taxonomy files via the anomalies_clean.py script to generate corrected taxonomies. These corrected taxonomies were used by the script build_protologues_from_named_genera.py to generate protologues for new names built from existing genus names (File S1).

**Creation of arbitrary Latinate names**

The workflow used to create and assign arbitrary Latinate names is shown in Fig. 1. The workflow can be run using the Bash shell script GTDB_renamer.sh once the required scripts and programmes are in place.

To create a set of >4 million restricted terms that should not be used in name creation for the workflow, we downloaded and parsed:

- a publicly available set of Latin stems compiled by the Latinist William Whitaker (downloaded as STEMLIST.GEN from http://archives.nd.edu/whitaker/old/wordsall.zip);
- headwords in the English Wiktionary, which includes millions of words from English and other languages (downloaded from https://dumps.wikimedia.org/enwiktionary/latest/enwiktionary-latest-pages-articles-multistream-index.txt.bz2);
- millions of names already in use in taxonomy compiled by the Global Biodiversity Information Facility, downloaded from https://hosted-datasets.gbif.org/datasets/backbone/current/simple.txt.gz.

The Python script input_terms_clean.py was used to extract relevant information from these input files, to convert all entries to lower case and to remove duplicates, thereby creating the output file excluded_terms.txt, which was used by the subsequent name creation scripts.

Each name was built from a combination of an opening word-component and a final word-component. Our initial efforts at creating arbitrary names relied on extracting initial five-letter strings from existing Latin words to create opening word-components [17]. However, this led to a set of genus names that were less distinctive than those currently in use as validly published names (File S2). We therefore took a number of steps to improve the distinctiveness of genus names.
Rather than relying on strings found in Latin words, our Python script, *genus_stem_creator.py* built opening word-components from all potential strings of consonants/consonant clusters (C) and vowels (V) found in the original Latin alphabet (i.e. excluding j, k, w, z), conforming to the format VCVC and CVCVC.

To avoid creating names that might sound confusingly similar, we exploited a variation of the phonetic-matching algorithm Soundex in a two-step process. First, we built all possible strings avoiding consonants representing sounds with similar sites of articulation (i.e. included 'b' but not 'p', 'c' but not 'g', 'd', but not 't', 'm' but not 'n', 'l' but not 'r'). Then, in each resulting string, we randomly replaced letters 'b', 'f', 'c', 'd', 'm' and 'l' with similar sounding alternatives or left them unchanged.

The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other [18]. To enhance the distinctiveness of the genus stems, our initial set was run through the script *name_curator.py*, which produced a set of curated stems separated by a Levenshtein distance of at least 2.

To create final word components for names that comply with phonotactic and grammatical norms of Latin, Latin suffixes were 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 were restricted to the feminine gender and first declension. The single final word component *archa* (derived via Latinization of the Greek noun, ἀρχή, meaning ‘beginning, origin’) was reserved for the creation of names for archaeal genera, while 26 final word-components were selected for use in bacterial genus names (all of which are separated by a Levenshtein distance of at least 2).

Next, we used the scripts *bac_genus_name_creator.py* and *ar_genus_name_creator.py* to create the files *archaeal_genus_names.txt*, *bacterial_genus_names.txt*. The scripts generated all combinations of opening word-components with final word-components, which were then searched against the excluded terms to ensure that they remained free of meaning and had not been previously
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:Latin_feminine_suffixes] and have been used alongside short nouns suitable for use as non-specific descriptors of microbes. The resulting names are all defined as Neo-Latin feminine first declension nouns in the nominative singular. When used as species epithets, these names are deployed as nouns in apposition, thereby avoiding any need for agreement in gender between genus name and species epithet.

| Component (nom., gen.) | Application | Derivation |
|------------------------|-------------|------------|
| archa, archae           | Archaean genus names | Latinized derivative of Gr. fem. n. ἀρχή, beginning, origin |
| ana, anae               | Bacterial genus names; species names | Latin suffix to a noun stem to form an adjective |
| ari, ariae              | Bacterial genus names; species names | Latin suffix used to form abstract nouns from other nouns |
| ella, ellae             | Bacterial genus names; species names | Latin suffix used to form a diminutive of a noun |
| ia,iae                 | Bacterial genus names; species names | Latin suffix used to form an abstract noun |
| osa, osae              | Bacterial genus names; species names | Latin suffix used to form adjectives from nouns meaning 'full of' |
| uda, uae               | Bacterial genus names | Latin suffix used to form a diminutive of a noun |
| astra, astrae           | Bacterial genus names | Latin suffix of nouns, expressing resemblance |
| atica, aticae           | Bacterial genus names | Latin suffix used to form adjectives indicating a relation to the root noun |
| entia, entiae           | Bacterial genus names | Latin suffix used to form an abstract noun |
| etta, ettae             | Bacterial genus names | Latin suffix used to form a diminutive of a noun |
| ibra, ibrae             | Bacterial genus names | Latin suffix of nouns denoting instrument, vessel, place or person |
| ibula, ibulae           | Bacterial genus names | Latin suffix of nouns denoting instrument, vessel, place or person |
| icella, icellae         | Bacterial genus names | Connecting vowel -i-; L. fem. n. cella, a cell |
| icula, iculae           | Bacterial genus names | Latin suffix used to form a diminutive of a noun |
| ifica, ificae           | Bacterial genus names | Latin suffix forming adjectives that denote bringing or making. |
| iformae, iforane        | Bacterial genus names | Connecting vowel -i-; L. fem. n. forma, a form |
| igena, iginae           | Bacterial genus names | Latin suffix meaning 'born from, sprung from' |
| ilega, ilegae            | Bacterial genus names | Latin suffix forming adjectives related to the concept of collecting |
| itenta, itentae         | Bacterial genus names | Latin suffix forming adjectives meaning 'abounding in, full of' |
| isca, iscae             | Bacterial genus names | Late Latin suffix used to form adjectives |
| issa, issae             | Bacterial genus names | Late Latin suffix used to form feminine forms of masculine nouns. |
| itia, itiae             | Bacterial genus names | Latin suffix to form an abstract noun describing the condition of being something. |
| itoga, itogae           | Bacterial genus names | Connecting vowel -i-; L. fem. n. toga, a covering, garment, used non-specifically in many bacterial names |
| itura, iturae            | Bacterial genus names | Latin suffix used to form a noun relating to an action. |
| ivita, ivitae            | Bacterial genus names | Connecting vowel -i-; L. fem. n. vita, life |
| ousia, ousiae            | Bacterial genus names | Gr. fem. n. ousia, essence |

used in taxonomy. The resulting names were randomly shuffled before sorting by length and ending to ensure that they gave rise to short but distinctive names when applied to higher-level taxa.

As there is far less need for distinctiveness among species epithets, we adopted a more relaxed approach to generating opening word-components, using the species_name_creator.py script to create on all possible CVC, VCVC and CVCVC combinations
and then combine them with each of five endings: -ia, -ana, -osa, -aria, -ella. This script also ensured that the resulting species names were not found in the excluded terms or among the newly created genus names.

**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* 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 then paired up Latin names with the alphanumeric taxon names 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 [19], 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 GTDB metadata files to create protologues rendering the names suitable for use in the scientific literature and in public databases.

**RESULTS**

**Correction of anomalies**

We identified 3608 anomalous phylogenies within GTDB release R207, which led us to propose 221 well-formed Latin names for taxa previously identified only by GTDB placeholders (Table S1; File S1). This includes the two bacterial phyla, previously designated WOR-3 and SZUA-79, which, in line with the recent rule change [20], we have named *Candidatus* Hydrothermoarchaeum after the genus *Candidatus* Hydrothermus and *Candidatus* Acidulodesulfobacteriota after the genus *Candidatus* Acidulodesulfobacterium [16, 21, 22]. In addition, we found and replaced 224 placeholders not built from genus designations.

**Creation of arbitrary Latin names**

Paying careful attention to phonetic and orthographic distinctiveness, we created 1670 genus stems from strings of consonants and vowels from the Latin alphabet. Combining these with a distinctive set of final word-components drawn from Latin—while taking care to exclude terms with pre-existing meanings or genus names already used in taxonomy—we created 1670 arbitrary, meaningless names for archaeal genera and nearly 30000 arbitrary, meaningless names for bacterial genera (Table S1). Using similar but more relaxed criteria, we created over 2 million species epithets suitable for naming species from either domain.

**Assignment of arbitrary names to unnamed taxa**

After retrieving nearly 65000 alphanumeric placeholder names from the GTDB taxonomy files, we built name-replacement tables showing the placeholder names alongside replacement well-formed Latin names (Table S2). We respected the arbitrary choices made by GTDB for type genera, replacing alphanumeric designations wherever they occur in the taxonomy hierarchy through addition of suffixes to genus names, as specified in the ICNP and elsewhere [9, 23]. Following the GTDB's lead in making arbitrary choices of type material for high-level taxa, we also made an arbitrary choice as to which species should be designated the type species for a previously unnamed genus.

GTDB 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 well-formed Latinate *Candidatus* names to: three archaeal and 76 bacterial phyla; 14 archaeal and 267 bacterial classes; 68 archaeal and 1043 bacterial orders; 360 archaeal and 2745 bacterial families; 1104 archaeal and 11158 bacterial genera; and 2813 archaeal and 45178 bacterial species (phylla shown in Table 2, all taxa shown in Table S2). Unused genus names and species epithets were retained for use with later releases of the database.

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 protologues for all the new *Candidatus* taxa (File S1, File S2). However, as these protologues take up over 14000 pages, they cannot be presented in the main body of this manuscript.

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
Table 2. Newly named phyla

| Domain    | New name                      | GTDB placeholder |
|-----------|-------------------------------|------------------|
| Archaea   | Candidatus Axalarchota        | EX4484-52        |
| Archaea   | Candidatus Iduparchota        | Splt-1190        |
| Archaea   | Candidatus Oferarchota        | B1Sed10-29       |
| Bacteria  | Candidatus Acidulodesulfo bacteriota | SZUA-79    |
| Bacteria  | Candidatus Afabiota           | CG2-30-53-67     |
| Bacteria  | Candidatus Asafioota          | UBP6             |
| Bacteria  | Candidatus Axaliota           | KSB1             |
| Bacteria  | Candidatus Bimaifiota         | UBP7             |
| Bacteria  | Candidatus Binidiota          | TA06             |
| Bacteria  | Candidatus Busufiota          | TJSed10-126      |
| Bacteria  | Candidatus Cemuriota          | 4572-55          |
| Bacteria  | Candidatus Cobisiota          | SLNR01           |
| Bacteria  | Candidatus Conosiota          | RBG-13-61-14     |
| Bacteria  | Candidatus Canadiota          | DRYD01           |
| Bacteria  | Candidatus Cutipioota         | UBA2233          |
| Bacteria  | Candidatus Cuxuifiota         | B130-G9          |
| Bacteria  | Candidatus Dalofiota          | JACPSX01         |
| Bacteria  | Candidatus Debeiota           | AABMS5-125-24    |
| Bacteria  | Candidatus Dobariota          | UBA6262          |
| Bacteria  | Candidatus Dufaniota          | JAHJDO01         |
| Bacteria  | Candidatus Dufopiota          | VGIIX01          |
| Bacteria  | Candidatus Dumaciota          | RUG730           |
| Bacteria  | Candidatus Dunuliota          | UBP13            |
| Bacteria  | Candidatus Dupeciota          | SZUA-182         |
| Bacteria  | Candidatus Efretiota          | FCPU426          |
| Bacteria  | Candidatus Enediota           | CSP1-3           |
| Bacteria  | Candidatus Esaciota           | UBA10199         |
| Bacteria  | Candidatus Falabiota          | NPL-UPA2         |
| Bacteria  | Candidatus Femapiota          | JABDJOQ01        |
| Bacteria  | Candidatus Figanistiota       | JACPQY01         |
| Bacteria  | Candidatus Fitomiota          | UBA8481          |
| Bacteria  | Candidatus Fixabiota          | JADJOY01         |
| Bacteria  | Candidatus Geseifiota         | JAFGBW01         |
| Bacteria  | Candidatus Getofiota          | CSSED10-310      |
| Bacteria  | Candidatus Galusiota          | TISED10-198M     |
| Bacteria  | Candidatus Hacexiota          | JACIXR01         |
| Bacteria  | Candidatus Hicupiota          | UBA3054          |
| Bacteria  | Candidatus Honifiota          | JAGOBFX01        |
| Bacteria  | Candidatus Hudomiota          | BMS3Abi14        |
| Domain     | New name              | GTDB placeholder |
|------------|-----------------------|------------------|
| Bacteria   | Candidatus Hufasiota  | JAFGOL01         |
| Bacteria   | Candidatus Hurebiota  | JAFGN01          |
| Bacteria   | Candidatus Hutubiota  | SpSt-318         |
| Bacteria   | Candidatus Hydrothermota | WOR-3       |
| Bacteria   | Candidatus Ibiota     | UBA8248          |
| Bacteria   | Candidatus Idrufiota  | UBP14            |
| Bacteria   | Candidatus Idupiota   | CG03             |
| Bacteria   | Candidatus Ifutiota   | JAAHXH01         |
| Bacteria   | Candidatus Lagoxiota  | UBP15            |
| Bacteria   | Candidatus Lanaxiota  | BS750m-G34       |
| Bacteria   | Candidatus Locusiota  | UBP18            |
| Bacteria   | Candidatus Macifiota  | DUMJ01           |
| Bacteria   | Candidatus Moxediota  | PUNC01           |
| Bacteria   | Candidatus Naraxiota  | JACRZ01          |
| Bacteria   | Candidatus Niriota    | QNDG01           |
| Bacteria   | Candidatus Niteriota  | 4484–113         |
| Bacteria   | Candidatus Ocupiota   | CAJMMQ01         |
| Bacteria   | Candidatus Ogusiota   | RBG-13-66-14    |
| Bacteria   | Candidatus Omexiota   | UBA9089          |
| Bacteria   | Candidatus Omubiota   | CG2-30-70-394    |
| Bacteria   | Candidatus Ostexiota  | UBA1439          |
| Bacteria   | Candidatus Podoexiota | FEN-1099         |
| Bacteria   | Candidatus Pudofiota  | JACPUC01         |
| Bacteria   | Candidatus Puresiota  | HKB111           |
| Bacteria   | Candidatus Robemioti  | CLD3             |
| Bacteria   | Candidatus Rosetiota  | TA06_A           |
| Bacteria   | Candidatus Rudexiota  | JACQOV01         |
| Bacteria   | Candidatus Saxciota   | JACPWU01         |
| Bacteria   | Candidatus Sifexiota  | J088             |
| Bacteria   | Candidatus Soduniota  | UBA6266          |
| Bacteria   | Candidatus Sogexiota  | UBP17            |
| Bacteria   | Candidatus Sonubiota  | SM23-31          |
| Bacteria   | Candidatus Sucudiota  | BS750m-G25       |
| Bacteria   | Candidatus Tefagiota  | JAGOEH01         |
| Bacteria   | Candidatus Tiberiota  | GCA-001730085    |
| Bacteria   | Candidatus Tufoliota  | ARS69            |
| Bacteria   | Candidatus Tusxiota   | UBP4             |
| Bacteria   | Candidatus Tuxefiota  | JABMQX01         |

Table 2. Continued
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 10 years, such approaches have delivered around a thousand new validly published species names per year and tens-to-hundreds of *Candidatus* names annually [14, 24]. However, given a backlog of >50,000 well-classified but unnamed species in GTDB, using conventional approaches at the current rate of progress would take at least half a century to name all these unnamed taxa—by which time, we would be faced with the problem of naming thousands or even millions more newly discovered species!

Recently, Pallen *et al.* [25] 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 distinctive 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* [26].

Here, given the failure of conventional approaches, we have taken what might seem like a radical step in creating and applying arbitrary well-formed Latin 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* [7]. 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’ [27]. Soon after, Scottish naturalist George Johnston created the arbitrary genus name *Carinella* for a marine annelid [28].

In 1869, in his groundbreaking *Lois de Nomenclature Botanique* [29]—precursor of all subsequent codes of nomenclature—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’.

In the early twentieth-century, the American entomologist William Kearfott created over a hundred arbitrary rhyming species epithets, including *bana*, *cana*, *dana*; *bobana*, *cocana*, *dodana*; *boxcana*, *coxcana*, *doxcana*—many of which are still in use today [30]. 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’ [31]. A few years later, 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 first International Bacteriological Code of Nomenclature made clear in the 1940s that bacterial names ‘may be composed in an arbitrary manner’—a stipulation carried forward into subsequent versions of the code, including the ICNP [9, 32]. Drawing on this point, the bacterial nomenclature expert Hans Trüper quoted the code to make clear: ‘genus names or specific epithets “may be taken from any source and may even be composed in an arbitrary manner”… These’rubber’ paragraphs open up a box of unlimited possibilities for people whose Latin is at the end. But in view of the million names that will have to be formed in the future, they are a simple necessity—whether Latin formalists like them or not.’

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* [15].

| Domain | New name         | GTDB placeholder |
|--------|------------------|------------------|
| Bacteria | *Candidatus Udisiota* | JdFR-76        |
| Bacteria | *Candidatus Urusiota* | OLB16          |
| Bacteria | *Candidatus Usiniota* | DTU030         |
| Bacteria | *Candidatus Usuriota* | SAR324         |

Table 2. Continued
However, there are limits to the practice of arbitrary name creation, in that Principle 3 of the ICNP [9] states 'scientific names of all taxa are Latin or latinized words treated as Latin regardless of their origin.' In English, Latinate terms often form a lexical stratum with a distinctive phonotactic 'Latinity', associated with scientific or scholarly writing [33, 34]. To comply with this requirement of the ICNP, while also respecting the look and feel of the language, we have combined strings of letters from the Latin alphabet with declinable feminine suffixes. 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 placeholder names in GTDB are generally associated with uncultured taxa. However, as Pallen [14] has argued elsewhere, even if some renamed taxa do 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 [35]. 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 [5]—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 newly cultured species has fallen to those who isolate and discover the species and deposit type material in culture collections. As this often requires a substantial effort, 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.

One limitation of this work is that, although we have replaced all alphanumeric placeholders, some GTDB taxa nonetheless remain without well-formed names, because they have been split from existing named taxa and given provisional designations based on a well-formed name plus an alphabetical suffix (e.g. Clostridium_A, Clostridium_B). However, as most of these split taxa have cultured representatives, such a venture would require careful designation of type strains, species and genera, placing it beyond the scope of the present study. Replacement of names for split taxa thus remains a task for the future.

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. allowing longer or more complex initial word-components or additional final word-components).

Drawing on the encouraging precedent with new names for SARS variants of concern [36], 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.

**Note added in proof**

Since this manuscript was submitted, Williams et al (https://doi.org/10.1111/1462-2920.16026) have proposed the following Candidatus phylum names:

- *Ca.* Hinthialibacterota for OLB16, which we have designated *Ca.* Urusiota.
- *Ca.* Electryoneota for AABM5-125-24, which we have designated *Ca.* Debefiota
- *Ca.* Lernaellota for FEN-1099, which we have designated *Ca.* Podoxiota

As there is no convention for establishing priority for Candidatus names, we will leave it to the scientific community to decide which names for these phyla should be used in the future.
Acknowledgements

We thank Andrea Telatin for help in creating the first version of the taxon_renamer.py Python script. We thank Philip Hugenholtz for providing feedback on an early draft of the manuscript. We thank Aharon Oren for commenting on the protologues for new names for anomalous placeholder taxa built from existing genus names. We thank Rachel Gilroy for confirming that the renamed GTDB taxonomy files worked as expected with the GTDB Toolkit. The computational results presented in FileS2 have been achieved using the LEO HPC infrastructure of the University of Innsbruck.

Author contributions

Conceptualization: M.J.P. Data curation and methodology: M.J.P., L.M.R.-R. and N.F.A. Writing: M.J.P.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol 2015;13:787–794.
2. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcock BJ, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2017;2:1533–1542.
3. Rosselló-Móra R. Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ Microbiol 2012;14:318–334.
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.
5. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, 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 2022;50:D785–D794.
6. Austin D. The nuance and wit of Carolin linnaeus. The Palmetto 1993;13:8.
7. Linnaeus C. Systema Naturae. 12th edn. Stockholm: Laurentius Salvius; 1759.
8. GTDB. Release 207 Statistics. 2022.
9. Parker CT, Tindall BJ, Garrity GM. International code of nomenclature of prokaryotes. Int J Syst Evol Microbiol 2019;69:51–511.
10. 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.
11. Woese CR. Bacterial evolution. Microbiol Rev 1987;51:221–271.
12. Relman DA. The identification of uncultured microbial pathogens. J Infect Dis 1993;168:1–8.
13. Murray RG, Schleifer KH. Taxonomic notes: a proposal for recording the properties of putative taxa of prokaryotes. Int J Syst Bacteriol 1994;44:174–176.
14. Pallen MJ. The status Candidatus for uncultured taxa of Bacteria and Archaea: SWOT analysis. Int J Syst Evol Microbiol 2021;71:005000.
15. Pallen MJ. Bacterial nomenclature in the era of genomics. New Microbes New Infect 2021;44:100942.
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.
17. Pallen M, Alikhan N-F. Naming the unnamed: over 45,000 candidatus names for unnamed Archaea and Bacteria in the genome taxonomy database. BIOLOGY. 2021;2021110557.
18. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 1966;10:707–710.
19. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics 2019;btz848.
20. Oren A, Arahal DR, Rosselló-Móra R, Sutcliffe IC, Moore ERB. Emendation of rules 5b, 8, 15 and 22 of the international code of nomenclature of Prokaryotes to include the rank of phylum. Int J Syst Evol Microbiol 2021;71:004861.
21. Tan S, Liu J, Fang Y, Hedlund BP, Lian Z-H, et al. Insights into ecological role of a new deltaproteobacterial order Candidatus Acidulodesulfobacterales by metagenomics and metatranscriptomics. ISME J 2019;13:2044–2057.
22. Oren A, Garrity GM. Candidatus List No. 2. Lists of names of prokaryotic Candidatus taxa. Int J Syst Evol Microbiol 2021;71:004671.
23. Whitman WB, Oren A, Chuvochina M, da Costa MS, Garrity GM, et al. Proposal of the suffix -ota to denote phyla. addendum to "proposals to include the rank of phylum in the International Code of Nomenclature of Prokaryotes." Int J Syst Evol Microbiol 2018;68:967–969.
24. LPSN. List of Prokaryotic names with Standing in Nomenclature. https://www.bacterio.net [accessed 22 April 2022].
25. Pallen MJ, Telatin A, Oren A. The next million names for Archaea and Bacteria. Trends Microbiol 2021;29:289–289.
26. 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 COMMUN 2021;1.
27. Lindley J. An Introduction to Botany. London: Longman, Orme, Brown, Green, and Longmans; 1839.
28. Johnston G. Illustrations in british zoology. Magazine of natural history and journal of zoology, botany, mineralogy, geology and meteorology 1833;6:235–235.
29. de Candolle A. Lois de la Nomenclature Botanique. Masson, 1867.
30. Kearfott WD, New North American Tortricidae. Trans Amer Entomol Soc 1907;33:1–98.
31. Rawson PF, Rushton AWA, Simpson ML, Raymond Charles Casey. 10 October 1917–26 April 2016. Biogr Mem Fell R Soc 2020;68:71–86.
32. Buchanan RE, St John-Brooks R, Breed RS. International Bacteriological Code of Nomenclature. J Bacteriol 1948;55:287–306.
33. Chomsky N. Halle M. The Sound Pattern of English. MIT Press (MA), 1968, p. 470.
34. de Candolle A. Lois de la Nomenclature Botanique. Paris. V. Masson et fils. 1867.
35. Murray AE, Freudenstein J, Gribaldo S, Hugenholtz P. Roadmap for naming uncultivated Archaea and Bacteria. Nat Microbiol 2020;5:987–996.
36. 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.