NamesforLife Release 20201224

During 2020, there were a total of 7,388 changes in prokaryotic taxonomy and nomenclature as compared to **Release 20191204** including; 1,559 novel taxa, 24 replacement names, 15 rank elevations, 33 rank reductions, 750 new combinations, 3 correction, 2 neotypes, 4,179 transfers of taxa and 822 changes in the preferred names appearing in the NamesforLife condensed taxonomy. This represents a sharp rise over what has been observed during the past 5 years, in which the publication of new taxa ranged from 1,042 – 1,395 per year. In addition, there were 863 emendation of existing names. It is noteworthy that these data apply only to those taxa with validly published names and do not include 1,274 names appearing in the Lists of Candidatus Taxa, Nos. 1 and 2.

NamesforLife maintains and distributes two views of the state of prokaryotic taxonomy, based on monthly updates of the published record and supporting data. The taxonomies are hierarchical and based on the validly published names (those which conform to the principles and rules set forth in the *International Code of Nomenclature of Prokaryotes*) appearing in the *International Journal of Systematic and Evolutionary Microbiology*. NamesforLife taxonomies represent a consensus view of experts who apply contemporary methods of classification including phylogenetic analysis of the small ribosomal subunit (16S rRNA gene), pairwise comparison of genome sequences and phenotypic properties.

Table 1. A summary of the current state of taxonomy of prokaryotes with validly published names.

	Complete Taxonomy [a]	Condensed Taxonomy [b]	HQ16S [c]	Genomes (type) [d]	Genomes (non-type) [e]	Genomes (combined)
Phyla	57	44	44	44	41	44
Classes	207	119	120	116	109	117
Orders	459	275	275	258	226	261
Families	953	665	661	604	508	624
Genera	3,838	3,339	3,293	2,699	1,862	2,841
Species/Subsp.	22,900	17,773	17,518	10,953	4,888	12,128

[a] The *Complete Taxonomy* includes all published synonyms, homonyms and names that may be considered illegitimate, rejected, orthographically or grammatically incorrect or not validly published for a documented reason. It is used to establish nomenclatural accuracy and determining the correct current state of a name and correctly interpreting the names appearing in older literature. It also includes a growing subset of published *Candidatus* taxa. Only those that are considered preferred names are presented here.

[b] The *Condensed Taxonomy* is a view of the current state of prokaryotic taxonomy and nomenclature that leverages features of the NamesforLife Information Architecture. Each species/subspecies is uniquely represented in a single point in the hierarchy, based on its most recent validly published name or revision in its circumscription or properties. Mapping to earlier states and all associated data and literature is addressed using NamesforLife DOIs.

[c] The *HQ16S* data set consists of curated, high-quality 16S rRNA gene sequences used in the published descriptions of type strains of species/subspecies of bacteria and archaea with validly published names. Linking to verified deposits of viable type material in over 125 culture collections as well as earlier synonyms, the *HQ16S* dataset allows for accurate identification and naming of 98.56% of bacteria and archaea with validly published names.

[d-e] NamesforLife genome sequence data is a continuously updated version of publicly available prokaryotic genome assemblies. **Release 20201224** contains 295,784 records including 14,747 assemblies that were verified as sourced from 10,953 type strains with validly published names; 6,302 type strains were represented by two or more genome assemblies. This number was reduced to 14,145 type strains when excluded assemblies were removed from consideration. This represents 61.5% coverage of the species and subspecies with validly published names, which is up from 51.2% coverage at the outset of 2020.

In this release, 5,988 type strain genome records were reannotated to reflect changes in nomenclature and to correct type strains that were erroneously assigned to heterotypic synonyms. An additional 238,746 assemblies could be associated with 4,888 taxa with validly published names at the species/subspecies level and represents a change from our previous inclusion of higher taxa with validly published names in this category. The taxonomic

coverage of the combined data has increased to 10,953 validly named taxa. Of the remaining sequence records 12,544 were identified as *Candidatus* taxa that could be placed into 870 discrete "groups" at varying levels of taxonomic resolution. Of those, 460 have names that appear on the ***Lists of names of prokaryotic* Candidatus *taxa*** in the ***International Journal of Systematic and Evolutionary Microbiology,*** representing 27.7% coverage of the Candidatus taxa named to date**.** The remaining 29,747 sequences were associated with 2,741 names that have no standing in the nomenclature of prokaryotes or represent higher taxa (genus – phylum) to which the genome has been identified.

Table 2. Summary of reported NCBI exclusions for N4L re-annotated genomes Release 20201224.

Reported exclusion category [a]	N4L type	N4L non-type	N4L *Candidatus*	N4L invalid
Total assemblies	14,747	238,746	12,544	29,747
assembly from type material	13,777	756	9	808
assembly from synonym type material	202	42	0	0
assembly from proxytype material	0	0	0	0
assembly designated as neotype	9	1	0	0
assembly from pathotype material	0	40	0	0
assembly designated as reftype	0	22	1	2
untrustworthy as type	374	18	9	27
abnormal gene to sequence ratio	5	155	11	1
chimeric	0	1	0	0
contaminated	65	1514	7	27
derived from environmental source	2	14,846	2,847	10,461
derived from metagenome	1	29,298	8,004	12,415
derived from single cell	0	882	330	351
derived from surveillance project	0	0	0	0
fragmented assembly	136	5,710	1,065	2,035
genome length too large	21	693	12	9
genome length too small	13	689	43	11
genus undefined	30	23,681	8,149	11,385
hybrid	0	1	0	0
low gene count	0	0	3	0
low quality sequence	12	197	15	7
many frameshifted proteins	211	3,608	5	20
metagenome	1	29,298	8,004	12,415
misassembled	3	38	0	1
missing ribosomal protein genes	1	94	6	1
missing rRNA genes	20	82	10	6
missing strain identifier	6	326	13	12
missing tRNA genes	5	176	22	5
mixed culture	0	3	0	0
partial	3	104	37	31

[a] Exclusions and relation to type material reported by NCBI and mapped to N4L categories of re-annotated genome assemblies. Genomes assemblies may have zero, one or more than class of exclusion reported.

For more information about NamesforLife services and data products visit https://namesforlife.com.