NamesforLife Release 20190614

During June 2019, there were a total of 83 changes as compared to **NamesforLife Release 20190521**, including 59 novel taxa, 2 replacement names, 2 rank elevations, no rank reductions, 3 new combinations, no corrections, 1 neotype/proxy type, and 9 transfers of taxa and 7 changes in the preferred names appearing in the NamesforLife condensed taxonomy.

NamesforLife maintains and distributes two views of the monthly state of prokaryotic taxonomy, based on recent updates of the published record and supporting data. The taxonomies are purely 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*. These taxonomies represent a consensus view of experts who apply a contemporary method 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	49	39	39	38	35	38
Classes	175	97	98	90	74	91
Orders	403	244	244	222	159	224
Families	819	569	566	495	343	510
Genera	3,726	3,058	3,014	2,198	963	2,289
Species/Subsp.	20,486	16,323	16,124	8,277	3,754	9,417

[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 to correctly interpreting the names appearing in older literature. It also includes a growing subset of published *Candidatus* taxa.

[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 sets 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% 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 20190614** contains 211,861 records including 10,396 assemblies that were verified as sourced from 8,277 type strains with validly published names. This number was reduced to 7,894 type strains when excluded assemblies were removed from consideration.

Five hundred and thirty-five type strain genome records were re-annotated to reflect recent changes in nomenclature. An additional 162,848 assemblies could be associated with a total of 4,141 taxa with validly published names at varying levels of taxonomic resolution. The taxonomic coverage of the combined data set remained at 9,417 validly named species/subspecies. There were 38,619 records in which the nomenclature was re-annotated. Of the remaining sequence records, 6,039 were identified as *Candidatus* taxa and could be placed into 489 discrete "groups" at varying levels of taxonomic resolution. This decrease represents a sharp drop from what has been observed in prior months and reflects our effort to refine the groupings of candidate taxa based on depositor assigned names. It is likely that this number will continue to fluctuate over the next few months as we work towards

achieving a more uniform naming of candidate taxa. The remaining 32,541 sequences were associated with names that have no standing in the nomenclature of prokaryotes. As in previous months we have also observed that some assemblies present in a prior month have been suppressed or excluded. In **NamesforLife Release 20190614**, 49 assemblies present in **NamesforLife Release 20190501** do not appear this release.

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

Reported exclusion category[a]	N4L type	N4L non-type	N4L *Candidatus*	N4L invalid
assembly from type material	9988	41	81	170
assembly from synonym type material	68	15	26	1
assembly from proxytype material	6	6	12	5
assembly designated as neotype	8	0	0	0
assembly from pathotype material	0	4	8	0
untrustworthy as type	76	1	2	0
derived from environmental source	1	19	68	461
derived from metagenome	0	406	5947	1876
derived from single cell	0	8	240	680
derived from surveillance project	0	1836	1836	0
metagenome	0	406	5947	1876
chimeric	0	0	0	0
contaminated	25	7	17	21
mixed culture	3	0	0	0
unverified source organism	1	0	0	1
hybrid	0	0	0	0
misassembled	3	1	1	5
validation errors	1	7	7	1
genome length too large	37	62	104	7
genome length too small	10	18	67	3
partial	0	1	36	9
high contig L50	63	62	100	58
low contig N50	99	213	772	265
abnormal gene to sequence ratio	6	17	31	8
low gene count	0	0	3	0
low quality sequence	27	21	43	30
many frameshifted proteins	126	167	215	47
missing ribosomal protein genes	2	2	7	0
missing rRNA genes	17	2	13	2
missing tRNA genes	12	8	39	2

[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.