NamesforLife Release 20200703

During the first six months of 2020, there were a total of 5,035 changes in prokaryotic taxonomy and nomenclature as compared to **Release 20191204** including; 1,014 novel taxa, 23 replacement names, 7 rank elevations, 26 rank reductions, 462 new combinations, 1 correction, one neotype, 3,100 transfers of taxa and 518 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. It is noteworthy that these data apply only to those taxa with validly published names and do not include the published names of > 1,500 Candidatus taxa at any rank.

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	52	38	40	40	34	40
Classes	185	95	100	97	70	97
Orders	442	256	261	243	159	243
Families	903	619	621	563	345	572
Genera	3,719	3,219	3,185	2,515	1,136	2,591
Species/Subsp.	22,116	17,290	17,103	10,040	4,211	10,939

[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 98.73% 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 20200703** contains 255,886 records including 13,320 assemblies that were verified as sourced from 10,040 type strains with validly published names; 2,284 type strains were represented by two or more genome assemblies. This number was reduced to 9,856 type strains when excluded assemblies were removed from consideration. This represents 57.0% coverage of the species and subspecies with validly published names, which is down slightly from Release 20200530.

In this release, 89 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 160,847 assemblies could be associated with 4,294 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. Those assemblies are now

included with those having names that are not-validly published. The taxonomic coverage of the combined data has increased to 10,897 validly named taxa. There were 11,823 records in which the nomenclature was re-annotated. Of the remaining sequence records 10,743 were identified as *Candidatus* taxa that could be placed into 717 discrete "groups" at varying levels of taxonomic resolution. Of those, 500 have names that appear on the ***Lists of names of prokaryotic* Candidatus *taxa*** in the ***International Journal of Systematic and Evolutionary Microbiology***, representing 69.8% coverage of the Candidatus taxa named to date**.** The remaining 70,975 sequences were associated with 3,723 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 20200703.

Reported exclusion category [a]	N4L type	N4L non-type	N4L *Candidatus*	N4L invalid
assembly from type material	12,422	71	8	784
assembly from synonym type material	201	27	0	0
assembly from proxy type material	0	0	0	0
assembly designated as neotype	9	1	0	0
assembly from pathotype material	0	39	0	1
assembly designated as ref type	0	24	1	2
untrustworthy as type	340	11	4	28
derived from environmental source	2	96	2,767	11,874
derived from metagenome	0	2,237	6,897	34,346
derived from single cell	0	29	316	1,209
derived from surveillance project	0	0	0	0
chimeric	0	1	0	0
contaminated	61	1,339	6	195
mixed culture	0	2	0	1
unverified source organism	0	4	0	4
hybrid	0	1	0	0
misassembled	3	28	0	7
validation errors	0	0	0	0
genome length too large	24	639	8	14
genome length too small	13	633	46	15
partial	3	76	35	57
high contig L50	0	0	0	0
low contig N50	0	0	0	0
abnormal gene to sequence ratio	3	147	11	9
low gene count	0	0	3	0
low quality sequence	12	148	15	56
many frameshifted proteins	131	1,474	5	83
missing ribosomal protein genes	1	80	7	1
missing rRNA genes	14	81	11	8
missing tRNA genes	4	179	23	7
fragmented assembly	136	1,627	828	4,925

[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.