A Comparison of NamesforLife 16S rDNA data vs. Silva v.132 and Greengenes 13.5.99

For the last 25 years, 16S rDNA sequence similarity has been the principle means of classifying and identifying *Bacteria* and *Archaea*. Whereas "full length" sequences were considered mandatory just a few years ago, the advent of next generation sequencing led to a proliferation of methods using short amplicons of the variable regions in the 16S gene for such applications. Short amplicons have proven adequate for many purposes, including community analyses and comparison of the taxonomic composition of metagenomes and microbiomes. Regardless of the sequencing methodology or the purpose of an analysis, the amplicons will ultimately be compared to external reference data to establish taxonomic affiliation. A number of algorithms have emerged for this purpose along with recommended thresholds for defining the ranks of taxa. However, little attention has been given to either the taxonomic coverage or quality of the sequence annotations of those reference sets. This raises questions about the precision and accuracy of identification of unknown samples based on reference sources that may be incorrectly labeled or incomplete. What might the impact be on the results of analyses when economic consequences of a decision are high or public health or safety at stake? Here, we provide some answers to these questions using the NamesforLife 16S rDNA (N4L 16S) data set.

The NamesforLife 16S data set

The N4L 16S data set is a professionally curated collection of 16S gene sequences from the type species of prokaryotes. The data set is re-annotated monthly to integrate changes in taxonomy and nomenclature of *Bacteria* and *Archaea* with validly published names[1][1]. Each sequence is assigned to an appropriate *NamesforLife Exemplar* [2] which uniquely identifies the strain while maintaining its historical and current nomenclature, type status and taxonomic placement. An *Exemplar Information Object* aggregates descriptive data and literature references about the corresponding strain and establishes links to related resources in public databases and culture collections.

Sequence preprocessing

Before addition to the high quality N4L 16S (*HQ16S*) data set, sequences are screened for length, ambiguous bases and alignment quality. The alignment function in mothur [2] is used to generate a quality score and a best match to a reference alignment. In cases where more than one sequence is available for a given type strain, the highest scoring one is chosen to produce a non-redundant data set of the validly named species for which a viable type strain is available from one or more public culture collections. As of May 2019, the N4L 16S data set provided coverage of 16,027 of 16,277 type strains of *Bacteria* and *Archaea*. Those not represented in the data are species or subspecies having validly published names but lacking a viable type strain and requirements for viable deposits (Rules 27 and 30, ICNP).

As a preliminary step in the comparison of the N4L 16S data set to the non-redundant Silva and Greengenes reference data sets distributed by mothur, a self-comparison was made to identify sequences which showed a best-match to a closely related species or subspecies rather than to self. As these mismatches are to type strains of different species/subspecies, they are considered to be unpublished heterotypic synonyms (Rule 24a Note 3, ICPN) that warrant further investigation.

A second preliminary step was removal of mitochondrial, chloroplast, cyanobacterial and eukaryotic 18S rDNA sequences from the Silva and Greengenes data sets. This was followed by a review of the names at each rank in the corresponding sequence metadata to establish which were validly published names or invalid names. Sequence identifiers were also compared to those in NamesforLife *Exemplar Information*

[1] A validly published name is one that appears in the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) and conforms to the Rules of the *International Code of Nomenclature of Prokaryotes* (ICPN) [1].

Objects to determine which sequences could be positively identified as sourced from a type strain, regardless of the name appearing in the Silva or Greengenes metadata. Sequences identified as *Candidatus* taxa (Appendix 11, ICPN) were also flagged. Finally, Silva and Greengenes sequence metadata were expanded to include this information in support of more fine-grained analyses.

Sequence identification

Preprocessed Silva and Greengenes unaligned sequences were compared to the aligned NamesforLife reference data to identify the best match between each query and template sequence. The alignment report was expanded to incorporate nomenclatural data retrieved from NamesforLife via the Exemplar DOI; thereby permitting a direct comparison of the names from the best matching type strain sequence to those of the query sequence. Although the names of higher taxa are not independent of the names of the lower taxa (Principle 8, Rule 23a, ICPN), they are often treated as such; therefore, comparisons were made at the ranks of domain – genus (Silva) and domain to species (Greengenes). Agreement was scored using an additive heuristic, weighted by rank (class = 0.1, order = 1, family = 10, genus = 100, species = 200). This provides a way to objectively assess levels of support for names and putative identity of a sequence in Silva or Greengenes for the purpose of reannotation.

Results of this comparison for type strain sequences in Silva v.132 are presented in Table 1. The type strains are a special case, because the correct names[2] at any point in time are known, even if the type strains have been involved in one or more taxonomic rearrangements resulting in nomenclatural changes. Of the 109,776 non-redundant sequences with validly published names in Silva v.132, 7,522 could be positively identified as originating from type strains of species/subspecies with validly published names. As the *correct name* is known, a correctly annotated sequence record should bear that name at each rank in a taxonomy for it to be deemed legitimate (Rule 23, Note 5). What was found is that only 2,402/7,522 type-strain sequences bore the correct name at the four ranks evaluated. If scoring was restricted to genus – order, the number increases to 3,945. These results reveal that the maximum accuracy of correctly identifying type strains using Silva v.132 as the reference taxonomy in its current state would be in the range of 32–52%. Search scores (similarity in *k*-mers, *k*=8) and alignment scores (Needleman-Wunch method) are often used to evaluate a match between query and template sequences and the maximum scores found are quite high (>99.5). This is true even when the genus, family, order or class names are incorrect or missing. This suggests that a high score by either of these criteria allows one to correctly match a sequence to the appropriate taxon, but the probability of assigning the wrong name(s) to a sequence of known identity is quite high for Silva v.132 in its current state. The reasons are two-fold; the majority of annotations are incorrect and taxonomic coverage is incomplete. This suggests that a single metric to establish identity of an unknown may result in classification errors if the reference sequences are incorrectly annotated and/or provide incomplete coverage of the taxonomic space of interest.

Similar results were observed for non-type sequences with validly published names (Table 2). While the exact identity of individual sequences could not be ascertained, a similar pattern emerged. Only 36% of the non-type sequences bore the correct annotation at all four levels. The proportion increased to 45% if only the genus and family names were considered. The range of search scores varied more widely, but neither the search scores nor the sequence similarity scores showed a strong correlation to the likely accuracy of the assigned name appearing in the N4L 16S dataset.

This pattern was also found in the comparisons of Greengenes to the N4L 16S data set (Tables 3–4). The taxonomic resolution of Greengenes is finer than Silva, resulting in an additional annotation score, but

[2] The correct name is defined under the ICPN and maintained in the NamesforLife database.

there were fewer verified type strain sequences (3,473). A further complicating factor is the age of the last update (2013) and low taxonomic coverage.

Sequence reannotation and taxonomic coverage

The results of preprocessing show that only 31.9% of the type strain sequences in Silva and 6.3% of the type strain sequences in Greengenes are correctly annotated at the levels of taxonomic resolution claimed. Even if judged at the only the highest level of taxonomic resolution, accuracy in annotations of the type strain sequences included in Silva and Greengenes is 85% and 10.8%, respectively. However, these levels drop when taxonomic coverage of all *Bacteria* and *Archaea* with validly published names are considered (Silva 61.6% at the genus level and Greengenes 3.2% species level, 23.3% at the genus level). This raises questions about the reliability of identifications of known or unknown taxa when using these resources and the probability of incorrectly asserting taxonomic novelty.

Reannotation of the sequence metadata in Silva and Greengenes can address these problems by expanding it to include the correct current nomenclature and taxonomic information. For type strain sequences, this can proceed directly using curated data from NamesforLife *Exemplar Information Objects*. A similar approach can be used for non-type sequences that are annotated with validly published names at the genus (Silva) or species (Greengenes) level. When unambiguous sequence information is not provided, a combination of the annotation, sequence similarity and search scores can guide the process. However, setting arbitrary cut-off levels imposes a risk, as was seen in the examination of type strains. *Whereas sequence similarity may be used to discriminate among closely and distantly related sequences, it is not guaranteed to return the correct name.* This is especially true when sequence metadata are not curated and regularly updated. A remedy for this situation is to accept the identity of the template sequence with the best matching score in the N4L 16S data set as the most probable identity. While this may or may not yield the correct name initially, this improves following repeated rounds of reannotation and expansion of a data resource. This approach can also be used to assess taxonomic coverage.

Taxonomic coverage of the N4L 16S data set is shown in Table 5. The *Complete Taxonomy* includes all published synonyms, homonyms, and documented illegitimate, rejected, orthographically or grammatically incorrect and invalid names. This taxonomy is used to establish nomenclatural accuracy, and to interpret names in older literature (prior to January 1, 1980). It also includes a subset of published *Candidatus* taxa, some of which are now in culture. The *Condensed Taxonomy* is a view of the current state of prokaryotic taxonomy and nomenclature. It leverages features of the NamesforLife Information Architecture. Each species/subspecies is uniquely represented as a single point in the taxonomic hierarchy, based on its most recent validly published name or revision in its circumscription or properties. The methods used to create and maintain the *HQ16S* are discussed above.

Tables 6 and 7 show the effect of reannotation of the Silva and Greengenes reference data sets in relation to the NamesforLife taxonomy. Sequence metadata were examined at all ranks to establish which taxa are covered (presence/absence). Type strains and non-type strains were examined separately and combined. For Silva, coverage has been extended to the species level, but coverage of taxonomic type strains is only 44%. Taxonomic coverage increases to 78.3% when sequences from type and non-type strains are included. Tables 8 and 9 show the comparable results for Greengenes. Reannotation increases the number of type strain sequences from 3.25% coverage to 20.9% and combined type and non-type coverage from 12.9% coverage to 73.4% coverage.

Conclusions

Recently, considerable attention has been focused on the perceived quality of the taxonomies used to interpret and annotate genome, metagenome and microbiome data. Invariably, the focus of such studies is either on algorithms that are claimed to improve taxonomic accuracy and precision or on refined cut-offs for establishing taxon boundaries and assertions of identity. It is curious that these studies invariably fail to consider that taxonomies are neither static nor complete. Taxonomies and taxonomic names are in a constant state of flux and undergo frequent expansion and revision to reflect current opinion in the field. This information must to be continuously fed back into the reference taxonomies so that inferences reflect current taxonomic knowledge and link to the relevant data and literature persistently. When this is not done readers and end-users cannot make informed judgements based on current knowledge.

Since the beginning of this year, 602 validly published names of new taxa appeared in the taxonomic literature, along with 84 emendations of existing taxa. In addition to the published emendations, there have been many hundreds of implicit emendations arising from these published taxonomic proposals, which are not documented. Implicit emendations arise when the circumscription of higher taxa expand or contract in response to published proposals of subordinate taxa but lack a corresponding published description. Failure to account for these changes in reference data can result in a variety of classification errors including incorrect assertions of novelty, misidentifications, proliferation of taxonomic synonyms and incorrect inferences about the properties of the claimed taxa. The trend over the last 25 years has been reliance on measures of sequence similarity to identify *Bacteria* and *Archaea.* However, it is a name that is invariably used to access our knowledge about the phenotype, ecology, beneficial or hazardous properties of a strain or its membership in a taxon, pangenome, metagenome, or microbiome. As seen here, names can be ambiguous or incorrect, even if the scoring criteria used are within accepted bounds.

These are well-known problems that are solved by the NamesforLife Information Architecture [3], which provides a way to maintain reference data that accurately reflects past, present and future states of the taxonomy and nomenclature of prokaryotes in accordance with the Rules of the ICPN. NamesforLife DOIs are tightly integrated into the taxonomic literature and updates occur in synchrony with validation of new names and announcement of taxonomic rearrangements [4]. NamesforLife annotation and web services also provide direct access to up-to-date taxonomic information that can be used to augment digital resources to provide publishers, data providers, readers and end users of data products with the precise information they need, when they need it and in a way that is readily and persistently accessible [5,6].

References

1. Parker CT, Tindall BJ, Garrity GM. International Code of Nomenclature of Prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 2015. doi:10.1099/ijsem.0.000778.
2. Garrity G, Lyons C. Systems and methods for resolving ambiguity between names and entities. 2011. https://patents.google.com/patent/US7925444B2. Accessed 24 Jun 2019.
3. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol*. 2009;75:7537–41.
4. Garrity GM, Lyons C. Future-Proofing Biological Nomenclature. *OMICS: A Journal of Integrative Biology*. 2003;7:31–3. doi:10.1089/153623103322006562.
5. Parker CT, Lyons CM, Roston GP, Garrity GM. Systems and methods for automatically identifying and linking names in digital resources. 2019. https://patents.google.com/patent/US10204168B2. Accessed 24 Jun 2019.
6. Parker CT, Lyons CM, Roston GP, Garrity GM. Systems and methods for automatically identifying and linking names in digital resources. 2017. https://patents.google.com/patent/US9672293B2. Accessed 24 Jun 2019.

For information about NamesforLife services and data products visit https://www.namesforlife.com or email info@namesforlife.com. This white paper is accessible online as WP-N4L-20190702.

Table 1. Comparison of 16S rRNA sequence annotations between Silva v.132 (type strains) and NamesforLife Release 20190521.

Annotation score[1]	Class[2]	Order	Family	Genus	Species	Type strains (n)[3]	Search Score (range)[4]	Sequence similarity (range)[5]
111	-	x	x	x	-	1,543	75.60 - 99.58	96.19 - 99.58
110.1	x	-	x	x	-	413	86.45 - 99.44	97.64 - 99.44
110	-	-	x	x	-	761	90.57 - 99.44	98.47 - 99.44
101.1	x	x	-	x	-	615	78.75 - 99.72	95.42 - 99.72
101	-	x	-	x	-	148	96.22 - 99.37	99.72 - 99.37
100.1	x	-	-	x	-	510	77.47 - 99.42	92.88 - 99.42
100	-	-	-	x	-	303	91.36 - 99.51	98.69 - 99.51
11.1	x	x	x	-	-	200	84.70 - 99.38	95.88 - 99.38
11	-	x	x	-	-	172	88.89 - 99.51	98.06 - 99.51
10.1	x	-	x	-	-	40	96.19 - 99.23	99.57 - 99.23
10	-	-	x	-	-	141	94.94 - 99.24	99.17 - 99.24
1.1	x	x	-	-	-	84	68.00 - 99.35	93.14 - 99.35
1	-	x	-	-	-	12	64.85 - 98.95	91.08 - 98.95
0.1	x	-	-	-	-	142	81.77 - 99.17	96.48 - 99.17
0	-	-	-	-	-	36	95.49 - 99.23	99.72 - 99.23

[1] Annotation score – an additive weighted score of annotation similarities between two annotated taxonomies; species names identical, 300; genus, 100; family, 10; order, 1; class, 0.1.

[2] Maximum taxonomic resolution in Silva v.132 is at the genus level.

[3] Type strain – number of type strains at each annotation category that was positively identified via NamesforLife Exemplar Objects with links to the published taxonomic proposals.

[4] Search score range – Search Score computed in mothur (ver 1.42.1) using *k*-mer, *k* = 8; score is for percent matching *k*-mers between query and template sequences, range based on scores appearing in each annotation category.

[5] Alignment similarity between query sequence (Silva v.132) and aligned template sequence (NamesforLife non-redundant type strain 16S reference set); Needleman-Wunch alignment method.

Table 2. Comparison of 16S rRNA sequence annotations between Silva v.132 (non-type strains) and NamesforLife Release 20190521.

Annotation score	Class	Order	Family	Genus	Species	Non-type strains (n)	Search Score (range)	Sequence similarity (range)
111	-	x	x	x	-	9,348	28.94 - 99.37	74.25 - 99.37
110.1	x	-	x	x	-	10,479	33.77 - 99.56	80.13 - 99.56
110	-	-	x	x	-	5,901	68.87 - 98.73	94.72 - 98.73
101.1	x	x	-	x	-	7,844	28.10 - 99.51	80.88 - 99.51
101	-	x	-	x	-	2,698	81.35 - 99.10	98.93 - 99.10
100.1	x	-	-	x	-	7,178	25.40 - 99.16	81.43 - 99.16
100	-	-	-	x	-	8,016	41.99 - 99.58	80.67 - 99.58
11.1	x	x	x	-	-	2,566	86.75 - 98.26	98.69 - 98.26
11	-	x	x	-	-	1,557	83.75 - 94.60	97.99 - 94.60
10.1	x	-	x	-	-	1,849	78.04 - 96.03	96.63 - 96.03
10	-	-	x	-	-	1,173	57.66 - 95.31	91.48 - 95.31
1.1	x	x	-	-	-	1,637	35.32 - 99.21	79.40 - 99.21
1	-	x	-	-	-	180	69.97 - 86.99	95.47 - 86.99
0.1	x	-	-	-	-	2,521	28.06 - 99.04	74.96 - 99.04
0	-	-	-	-	-	2,498	28.65 - 99.29	75.45 - 99.29

Table 3. Comparison of 16S rRNA sequence annotations between Greengenes 13.5.99 (type strains) and NamesforLife Release 20190521.

Annotation score	Class	Order	Family	Genus	Species	Type strains (n)	Search Score (range)	Sequence similarity (range)
311.1	x	x	x	x	x	247	77.66 - 99.20	97.23 - 100.0
311	-	x	x	x	x	20	82.31 - 89.55	97.43 - 97.67
310.1	x	-	x	x	x	33	84.50 - 90.82	97.66 - 97.96
310	-	-	x	x	x	57	88.45 - 93.32	96.92 - 97.52
301.1	x	x	-	x	x	25	75.05 - 82.18	95.43 - 97.14
301	-	x	-	x	x	1	75.05 - 94.96	95.43 - 100.0
300.1	x	-	-	x	x	41	92.40 - 92.47	100.0 - 100.0
300	-	-	-	x	x	15	87.44 - 91.22	97.29 - 100.0
211	-	x	x	x	-	2	89.69 - 95.59	99.09 - 100.0
111.1	x	x	x	x	-	906	60.17 - 76.81	83.98 - 92.13
111	-	x	x	x	-	236	86.74 - 86.97	96.98 - 97.72
110.1	x	-	x	x	-	100	73.79 - 83.75	94.92 - 96.10
110	-	-	x	x	-	256	68.60 - 69.46	91.54 - 91.55
101.1	x	x	-	x	-	148	51.89 - 77.66	88.75 - 94.13
101	-	x	-	x	-	15	93.06 - 95.18	99.85 - 100.0
100.1	x	-	-	x	-	88	82.82 - 83.57	96.83 - 97.28
100	-	-	-	x	-	45	86.37 - 88.32	92.77 - 98.36
11.1	x	x	x	-	-	566	62.80 - 76.64	89.79 - 93.61
11	-	x	x	-	-	112	88.26 - 90.09	97.81 - 98.67
10.1	x	-	x	-	-	63	80.63 - 85.52	95.89 - 97.69
10	-	-	x	-	-	115	86.00 - 91.27	97.88 - 98.04
1.1	x	x	-	-	-	140	83.57 - 87.44	93.59 - 94.6
1	-	x	-	-	-	10	93.49 - 94.39	100.0 - 100.0
0.1	x	-	-	-	-	168	87.38 - 88.36	94.17 - 94.33
0	-	-	-	-	-	64	88.36 - 99.02	94.17 - 100.0

Table 4. Comparison of 16S rRNA sequence annotations between preprocessed Greengenes 13.5.99 (non-type strains) and NamesforLife Release 20190521.

Annotation score	Class	Order	Family	Genus	Species	Non-type strains (n)	Search Score (range)	Sequence similarity (range)
311.1	x	x	x	x	x	5,124	51.89 - 99.51	83.98 - 100.0
311	-	x	x	x	x	621	88.35 - 89.33	98.34 - 98.34
310.1	x	-	x	x	x	1,915	81.18 - 81.64	96.94 - 96.94
310	-	-	x	x	x	397	76.64 - 76.87	96.21 - 96.21
301.1	x	x	-	x	x	339	67.33 - 68.94	94.85 - 94.86
301	-	x	-	x	x	6	74.89 - 75.47	94.81 - 94.81
300.1	x	-	-	x	x	298	92.79 - 93.40	99.51 - 99.51
300	-	-	-	x	x	103	90.00 - 90.82	98.73 - 98.73
211.1	x	x	x	x	-	5	89.97 - 91.71	98.70 - 98.70
211	-	x	x	x	-	5	87.41 - 88.31	98.68 - 98.68
111.1	x	x	x	x	-	31,177	71.29 - 72.06	93.13 - 93.13
111	-	x	x	x	-	4,713	93.70 - 94.15	98.85 - 98.85
110.1	x	-	x	x	-	6,090	80.32 - 80.44	96.86 - 96.86
110	-	-	x	x	-	3,972	92.82 - 92.90	99.37 - 99.37
101.1	x	x	-	x	-	4,413	90.25 - 90.43	98.66 - 98.66
101	-	x	-	x	-	58	91.50 - 91.77	98.61 - 98.61
100.1	x	-	-	x	-	2,397	76.38 - 76.39	94.95 - 94.95
100	-	-	-	x	-	1,662	88.88 - 89.22	98.83 - 98.83
11.1	x	x	x	-	-	4,923	85.99 - 86.32	98.19 - 98.19
11	-	x	x	-	-	1,244	80.38 - 82.79	96.57 - 96.57
10.1	x	-	x	-	-	2,992	81.08 - 81.30	96.36 - 96.36
10	-	-	x	-	-	1,092	83.65 - 83.97	97.15 - 97.15
1.1	x	x	-	-	-	2,893	92.30 - 92.41	98.56 - 98.56
1	-	x	-	-	-	45	86.62 - 90.54	98.48 - 98.49
0.1	x	-	-	-	-	2,372	67.52 - 68.23	92.90 - 92.92
0	-	-	-	-	-	2,585	67.52 - 68.23	92.90 - 92.92

Table 5. Taxonomic coverage NamesforLife 16S Reference Set Release 20190521.

N4L	Complete taxonomy	Condensed Taxonomy	HQ16S
Phylum	49	39	39
Class	195	97	98
Order	404	244	244
Family	817	567	564
Genus	3,712	3,047	3,003
Species/Subspecies	20,437	16,277	16,072

Table 6. Taxonomic coverage pre-processed Silva v.132 16S Reference Set.

Silva v132 input	type	non-type	combined
Phylum	34	38	38
Class	82	92	93
Order	203	225	230
Family	445	500	520
Genus	1,851	2,321	2,473
Species/Subspecies	0	0	0

Table 7. Taxonomic coverage re-annotated Silva v.132 16S Reference Set.

Silva v.132 – re-annotated	type	non-type	combined
Phylum	35	38	38
Class	85	93	96
Order	208	228	238
Family	465	522	546
Genus	2,030	2,586	2,796
Species/Subspecies	7,199	10,987	12,751

Table 8. Taxonomic coverage pre-processed Greengenes 13.6.99 16S Reference Set.

Greengenes input	type	non-type	combined
Phylum	28	36	36
Class	55	74	74
Order	121	161	163
Family	263	365	373
Genus	699	1,381	1,465
Species/Subspecies	529	1,813	2,095

Table 9. Taxonomic coverage re-annotated Greengenes 13.6.99 16S Reference Set.

Greengenes re-annotated	type	non-type	combined
Phylum	31	36	37
Class	66	77	82
Order	158	175	199
Family	362	414	468
Genus	1,330	2,075	2,437
Species/Subspecies	3,405	10,247	11,942