The species-level, open-reference 16S rRNA NGS reads taxonomy assignment pipeline

Version 20170120

Tsute Chen, Forsyth Institute Copyright 2018

Contact: tchen@forsyth.org

Reads were BLASTN-searched against a combined set of 16S rRNA reference sequences that consist of the HOMD (version 15.1 http://www.homd.org/index.php?name=seqDownload&file&type=R), HOMD 16S rRNA RefSeq Extended Version 1.1 (EXT), GreenGene Gold (GG) (http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/gold_strains_gg16S_aligned.fasta.gz) , and the NCBI 16S rRNA reference sequence set (ftp://ftp.ncbi.nlm.nih.gov/blast/db/16SMicrobial.tar.gz). The original number of reference sequences are 998 (HOMD), 495 (EXT), 3,940 (GG), and 19,670 (NCBI) respectively. These sequences were screened and combined to remove short sequences, chimera, duplicated and sub-sequences, as well as sequences with poor taxonomy annotation (e.g., without species information). The final combined set contains a total of 21,815 sequences representing a total of 14,649 oral and non-oral microbial species.

The NCBI BLASTN version 2.2.28+ (16) was used with the default parameters. Reads with ≥ 98% sequence identity to the matched reference and ≥ 98% alignment length (i.e., ≥ 98% of the read length that was aligned to the reference and was used to calculate the sequence percent identity) were classified based on the taxonomy of the reference sequence with highest sequence identity. If a read matched with reference sequences representing multiple species equally (i.e., equal percent identity and alignment length) it was subject to chimera checking. Non-chimeric reads with multi-species best hits were considered valid and were assigned as a different species with multiple species names. Unassigned reads (i.e., ≤ 98% identity or ≤ 98% alignment length) were pooled together and subject to the *de novo* chimera checking and sequence quality screening using the USEARCH program version v8.1.1861. The *de novo* chimera checking was done using 98% as the sequence identity cutoff. Non-chimeric unassigned reads that are ≥ 200 bases were then subject to species-level *de novo* operational taxonomy unit (OTU) calling \with 98% as the sequence identity cutoff using USEARCH. Representative reads from each of the OTUs/species were BLASTN-searched against the same reference sequence set again to determine the closest species for these potential novel species. All assigned reads were subject to several down-stream bioinformatics analyses, including alpha and beta diversity assessments, provided in the QIIME (Quantitative Insights Into Microbial Ecology) software package version 1.9.1. Samples with < 500 read counts were excluded in the QIIME analysis. The phylogenetic tree required for constructing the UniFrac-based matrices used in some of the beta diversity analyses, was built dynamically from reference sequences with matched reads (no novel species identified in the *de novo* OTU calling stage were included in the tree due to the lack of full length sequences). The reference sequences were aligned with the software MAFFT version 7.149b

prior to tree construction using the QIIME treeing script. Down-stream analyses were done for a range of minimal read count per OTU/species (MC): 1, 2, 5, 10 50, and 100 separately.

Designations used in the taxonomy:

1. **Taxonomy levels:**

k__: domain/kingdom
p__: phylum
c__: class
o__: order
f__: family
g__: genus
s__: species

Example:
k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia;s__faecis

2. **Unique level identified – known species:**

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Roseburia;s__hominis

The above example shows some reads match to a single species (all levels are unique)

3. **Non-unique level identified – known species:**

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Roseburia;s__multispecies_spp123_3

The above example "s__multispecies_spp123_3" indicates certain reads equally match to **3** species of the genus Roseburia; the "spp123" is a temporally assigned species ID.

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__multigenus;s__multispecies_spp234_5

The above example indicates certain reads match equally to 5 different species, which belong to multiple genera.; the "spp234" is a temporally assigned species ID.

4. Unique level identified – unknown species, potential novel species:

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Roseburia;s__ hominis_nov_97%

The above example indicates that some reads have no match to any of the reference sequences with sequence identity ≥ 98% and percent coverage (alignment length) ≥ 98% as well. However this groups of reads (actually the representative read from a *de novo* OTU) has 96% percent identity to *Roseburia hominis*, thus this is a potential novel species, closest to *Roseburia hominis*. (But they are not the same species).

5. Multiple level identified – unknown species, potential novel species:

k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Roseburia;s__ multispecies_sppn123_3_nov_96%

The above example indicates that some reads have no match to any of the reference sequences with sequence identity ≥ 98% and percent coverage (alignment length) ≥ 98% as well. However this groups of reads (actually the representative read from a *de novo* OTU) has 96% percent identity equally to 3 species in *Roseburia.* Thus this is no single closest species, instead this group of reads match equally to multiple species at 96%. Since they have passed chimera check so they represent a novel species. “sppn123” is a temporary ID for this potential novel species.