Species –Level Open Reference Taxonomy Assignment for NGS 16S rRNA Reads
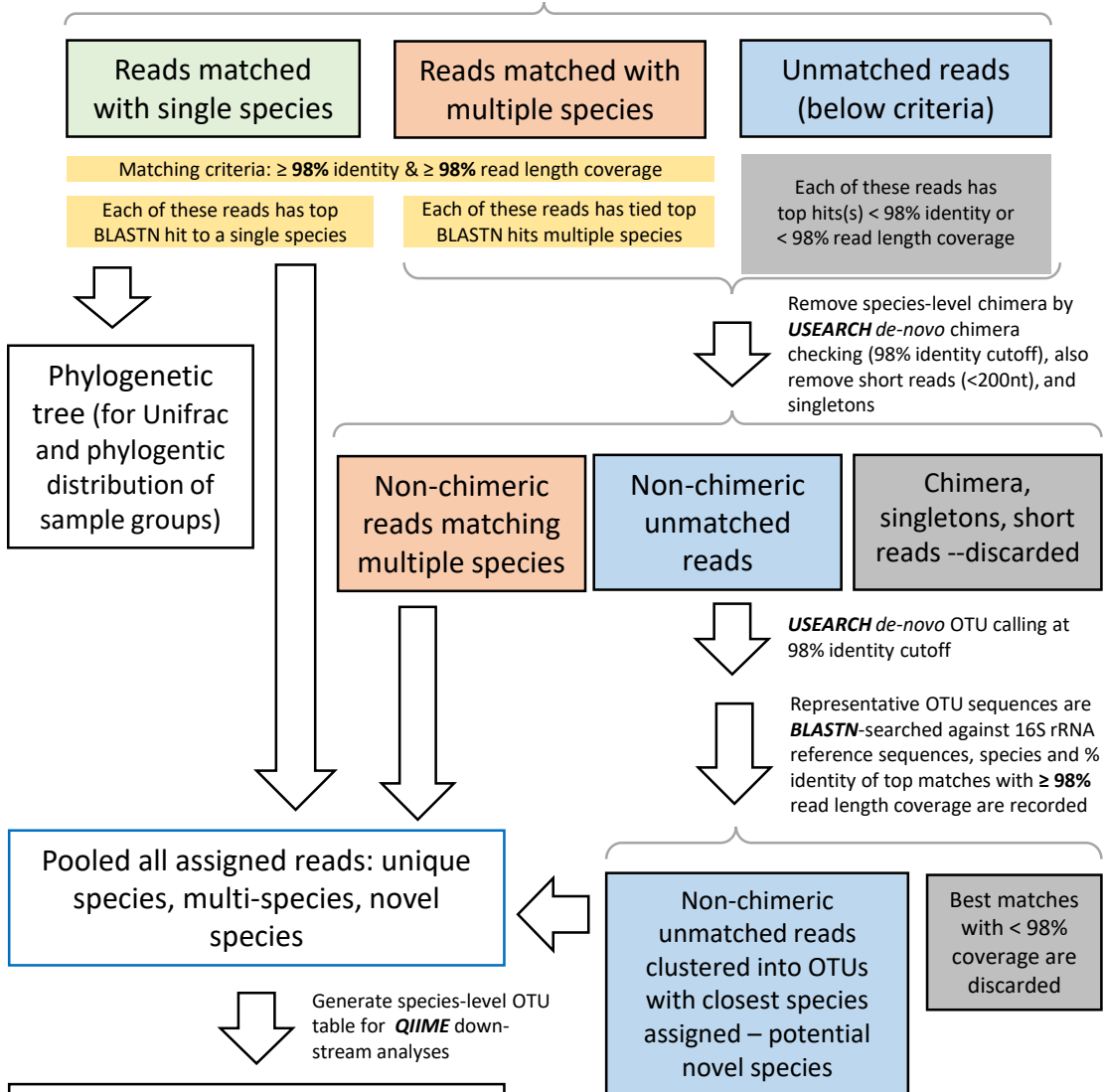
Tsute Chen & Nezar Al-hebshi
Version 2018-06-04

Raw sequence reads → DADA2 amplicon sequence variant (AVS) analysis **†** → DADA2 AVS

† DADA2 includes quality filtering, trimming, merging (overlapped paired reads) or concatenation (non-overlap reads), and chimera removal at variant level.

BLASTN search against
16S rRNA reference sequences *

Reads matched with single species

Reads matched with multiple species

Unmatched reads (below criteria)

Matching criteria: ≥ **98%** identity & ≥ **98%** read length coverage

Each of these reads has top BLASTN hit to a single species

Each of these reads has tied top BLASTN hits multiple species

Each of these reads has top hits(s) < 98% identity or < 98% read length coverage

Phylogenetic tree (for Unifrac and phylogentic distribution of sample groups)

Remove species-level chimera by *USEARCH* de-novo chimera checking (98% identity cutoff), also remove short reads (<200nt), and singletons

Non-chimeric reads matching multiple species

Non-chimeric unmatched reads

Chimera, singletons, short reads --discarded

USEARCH de-novo OTU calling at 98% identity cutoff

Representative OTU sequences are *BLASTN*-searched against 16S rRNA reference sequences, species and % identity of top matches with ≥ **98%** read length coverage are recorded

Pooled all assigned reads: unique species, multi-species, novel species

Non-chimeric unmatched reads clustered into OTUs with closest species assigned – potential novel species

Best matches with < 98% coverage are discarded

Generate species-level OTU table for *QIIME* down-stream analyses

OTU tables, taxonomy plots, alpha diversity analysis, and optionally - 2D & 3D beta diversity analysis.

QIIME analyses are done with different species abundance levels, from 1 (All), 10 (MC10), and 100 (MC100) minimal read count (MC) per species

The 16S rRNA reference sequences include:
- **HMT RefSeq V15.1**: **998** sequences
- **HOMD RefSeq Extended V1.11**: **151** sequences
- **GreenGeneGold V1**: **2,623** sequences;
- **NCBI 16S rRNA Reference**: **18,044** sequences

Total unique sequences: **21,816**
Total species: **14,651**
Note: In this version sequences < 1000nt have been removed.