New method promises species-level assignment from algorithmic classifiers

Isabel F. Escapaa
Yanmei Huang
Tsute Chen
Maoxuan Lin
Alexis Kokaras
Floyd E. Dewhirst
Katherine P. Lemon

Video Byte

Keywords: Microbiome, training set, naïve Bayesian RDP Classifier, species-level taxonomy, 16S rRNA gene, V1-V3, habitat-specific database, eHOMD, nasal, aerodigestive tract, amplicon sequence variant, ASV, k-mer, microbiota, Human Oral Microbiome Database, eHOMD, supraspecies, metagenomic whole genome sequencing, metagenomics, resolution

Posted Date: October 28th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-99200/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Abstract

Classifying organisms is easier, cheaper, and more accurate than ever thanks to state-of-the-art gene sequencing techniques and powerful machine learning algorithms. Now, researchers report a method that boosts classification resolution beyond the genus level, a well-known barrier faced by 16S rRNA gene sequencing. The principles behind the method are simple: make classification algorithms smarter by training them on richer genetic reference databases. That starts with collecting high-quality genetic datasets for a given habitat, and then identifying reference sequences within these datasets. These sequences help capture the natural genetic variability within each group of organisms in the selected habitat. The team tested their method by creating a training set for a familiar microbial habitat: the aerodigestive tract. Feeding that data into the naïve Bayesian RDP Classifier yielded species-level classification, breaking the genus-level barrier that has traditionally limited 16S rRNA gene microbiome studies. With further refinement, this new approach could help classify and assign sequences to known taxonomical structures.