Intro to Amplicon Sequence Variants Analysis

Ashley Shade

EDAMAME 29 June 2018

The "OTU"
operational taxonomic unit

- Species = basic unit of classification
- Defined somewhat arbitrarily
- Typical = 97% sequence identity
  - Originally, identity based on *full length* 16S rRNA gene
  - roughly equivalent to genus level
  - Does not well-distinguish "taxa" for all bacteria (*e.g., Streptomyces*)
  - Used in part to minimize species inflation by including amplification and sequencing errors
- Different methods of defining OTUs will result in different numbers of taxa! Different numbers of taxa mean different perspectives of diversity!

Approaches to Picking OTUs

- Reference based : percent identity to defined taxa populating a reference database
  - Pros: you know the taxa are "real"!
  - Cons: Weird environments don't have many representatives in databases, only as good as your database, could end up throwing out a lot of real stuff, use of "representative sequences" of the cluster

- De novo : percent identity to other sequences in the dataset; taxonomic assignment to the OTUs happens afterwards
  - Pros: Good for weird environments with low representation in databases
  - Cons: Computationally expensive, "greedy" algorithms can artificially inflate diversity

- Open reference : cluster against a reference db first, and anything that doesn't hit gets clustered de novo
  - Best of both worlds? Now can optimized so that new de novo OTUs are added to the original database and used subsequently in "reference" clustering
  - See Rideout et al. 2014 PeerJ

What are amplicon sequence variants?

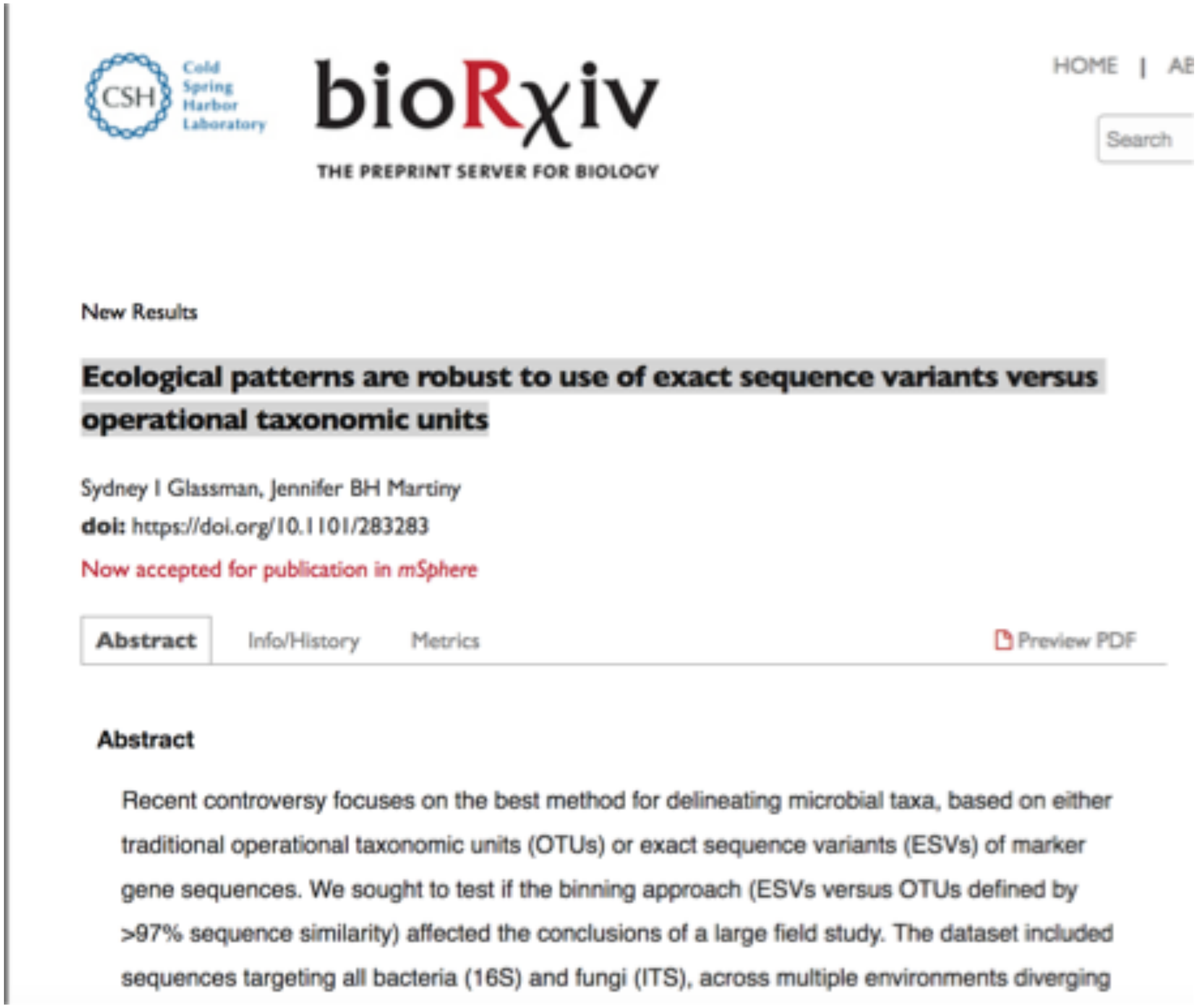
- Essentially, OTUs clustered at 100% sequence identity, after *rigorous and conservative* quality control of the data

- Direct consequences:
  - OTU inflation ?
    - Multiple 16S rRNA copies from the same organism may get split into different OTUs
    - If the quality control is too relaxed and keeps trashy data
  - OTU deflation?
    - If the quality control algorithm is too conservative and throws out real data
    - If there were a LOT of errors in the data and they were ”corrected”

Other names for sequence variants

- Exact sequence variants (ESV) - Glassman and Martiny 2018
- Amplicon sequence variants (ASV) – Callahan and Holmes / Phyloseq folks
- Sequence variants (SV) –- QIIME folks
- ZOTUs (zero-radius OTUs) – usearch/unoise, Edgar et al.
- Oligotypes – Eren et al.

Robust to ecological patterns


Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | AB
Search

New Results

Ecological patterns are robust to use of exact sequence variants versus operational taxonomic units

Sydney I Glassman, Jennifer BH Martiny

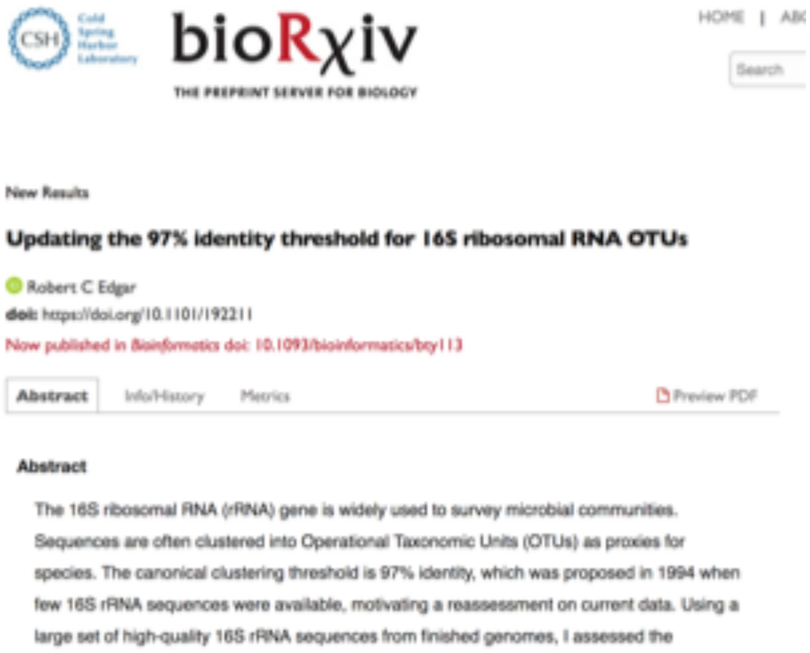
doi: https://doi.org/10.1101/283283

Now accepted for publication in *mSphere*

Abstract
Info/History
Metrics
Preview PDF

Abstract

Recent controversy focuses on the best method for delineating microbial taxa, based on either traditional operational taxonomic units (OTUs) or exact sequence variants (ESVs) of marker gene sequences. We sought to test if the binning approach (ESVs versus OTUs defined by >97% sequence similarity) affected the conclusions of a large field study. The dataset included sequences targeting all bacteria (16S) and fungi (ITS), across multiple environments diverging


Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABO
Search

New Results

Updating the 97% identity threshold for 16S ribosomal RNA OTUs

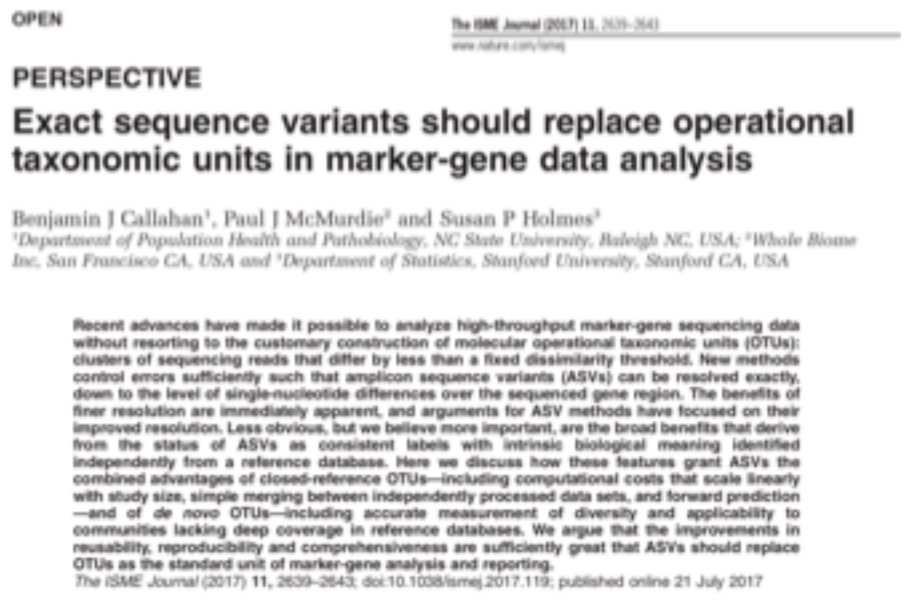
Robert C Edgar

doi: https://doi.org/10.1101/192211

Now published in *Bioinformatics* doi: 10.1093/bioinformatics/bty113

Abstract

The 16S ribosomal RNA (rRNA) gene is widely used to survey microbial communities. Sequences are often clustered into Operational Taxonomic Units (OTUs) as proxies for species. The canonical clustering threshold is 97% identity, which was proposed in 1994 when few 16S rRNA sequences were available, motivating a reassessment on current data. Using a large set of high-quality 16S rRNA sequences from finished genomes, I assessed the



OPEN

The ISME Journal (2017) 11, 2639–2643
www.nature.com/ismej

PERSPECTIVE

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan[1], Paul J McMurdie[2] and Susan P Holmes[3]
[1]*Department of Population Health and Pathobiology, NC State University, Raleigh NC, USA;* [2]*Whole Biome Inc, San Francisco CA, USA and* [3]*Department of Statistics, Stanford University, Stanford CA, USA*

Recent advances have made it possible to analyze high-throughput marker-gene sequencing data without resorting to the customary construction of molecular operational taxonomic units (OTUs): clusters of sequencing reads that differ by less than a fixed dissimilarity threshold. New methods control errors sufficiently such that amplicon sequence variants (ASVs) can be resolved exactly, down to the level of single-nucleotide differences over the sequenced gene region. The benefits of finer resolution are immediately apparent, and arguments for ASV methods have focused on their improved resolution. Less obvious, but we believe more important, are the broad benefits that derive from the status of ASVs as consistent labels with intrinsic biological meaning identified independently from a reference database. Here we discuss how these features grant ASVs the combined advantages of closed-reference OTUs—including computational costs that scale linearly with study size, simple merging between independently processed data sets, and forward prediction—and of *de novo* OTUs—including accurate measurement of diversity and applicability to communities lacking deep coverage in reference databases. We argue that the improvements in reusability, reproducibility and comprehensiveness are sufficiently great that ASVs should replace OTUs as the standard unit of marker-gene analysis and reporting.
The ISME Journal (2017) **11**, 2639–2643; doi:10.1038/ismej.2017.119; published online 21 July 2017

Work suggesting that “tighter” clusters should be used instead of 97%

Advantages of using sequence variants

- Can be re-used and compared directly across studies: “consistent labels”
- For some taxa, represent the most taxonomically resolved autonomous unit
- Some quality control and “clustering” methods afford a high level of confidence
  - Note that the most rigorous of these methods are not really clustering in the traditional sense
- “de novo”, so not reliant on limited databases, but precisely linked to any reference at 100%

Disadvantages of ASV

- Quality control **IS SO IMPORTANT** – errors can propagate within and across datasets quickly if not careful about QC
- For some taxa, other less resolved clustering (e.g., 97%) is biologically appropriate for ecologically meaningful taxa
- For some taxa, (e.g., Streptomyces), ASV are no better at resolving than 97%
- The most rigorous methods of ASV analysis are more computationally expensive
- It is cheap and easy to re-pick OTUs/ASVs for each new study/meta-analysis, and arguably needs to be done anyway to ensure integrity and rigor
- Sequencing effort matters, and must be consistent across studies

Denoising: essential for ASV
Errors in amplicon HTS

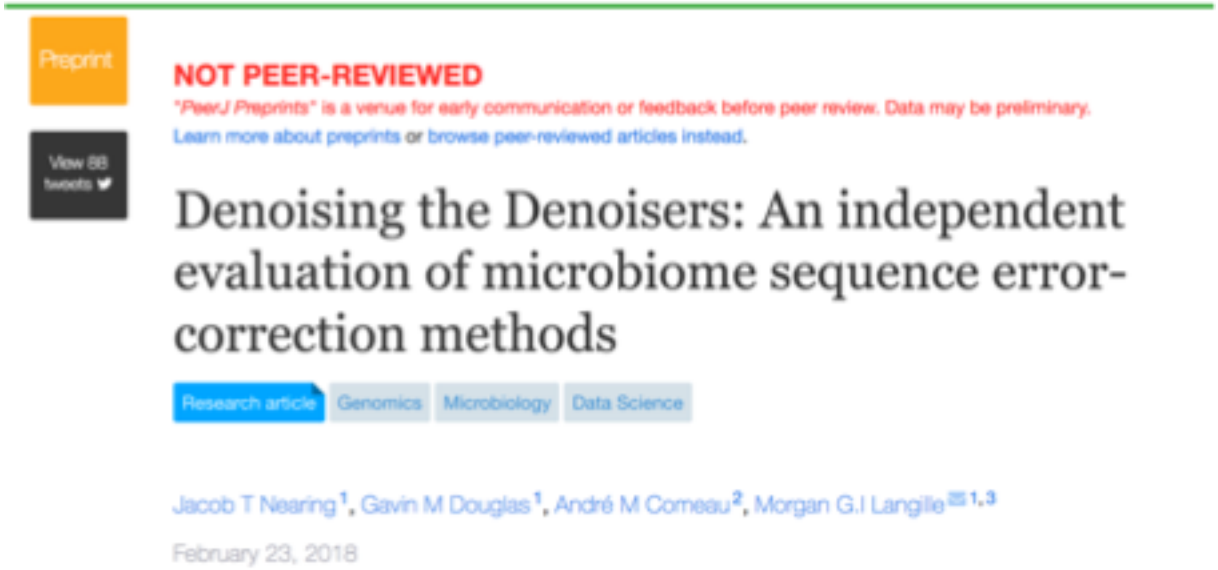
- Amplification errors
  - Substitutions – incorrect base pairings (*point error*)
  - Gaps – polymerase slips (*point error*)
  - Chimeras – combined biological templates by polymerase
- Sequencing errors
  - Insertions and deletions ("indels", *point errors*)
  - Spurious base calls (*point error*)
- Contaminants
  - In reagents or kits (or from researcher!)
  - "Cross-talk" between Illumina lanes

Edgar 2018 bioRXiv

Tools for determining sequence variants

- **Oligotyping/ MEN** (Eren et al. 2013) – includes quality filtering but works best when pre-QC steps are taken
- **QIIME2**: DEBLUR for QC using information theory method, and then q2-vsearch for 100% clustering
- **unoise3**: “zOTUs” – includes quality filtering, pools all samples to error correct using heuristic model and determine variants
- **DADA2**: “amplicon sequence variants” - includes quality filtering; uses single samples to build an error correction model

Comparing the denoising methods



NOT PEER-REVIEWED

"PeerJ Preprints" is a venue for early communication or feedback before peer review. Data may be preliminary.

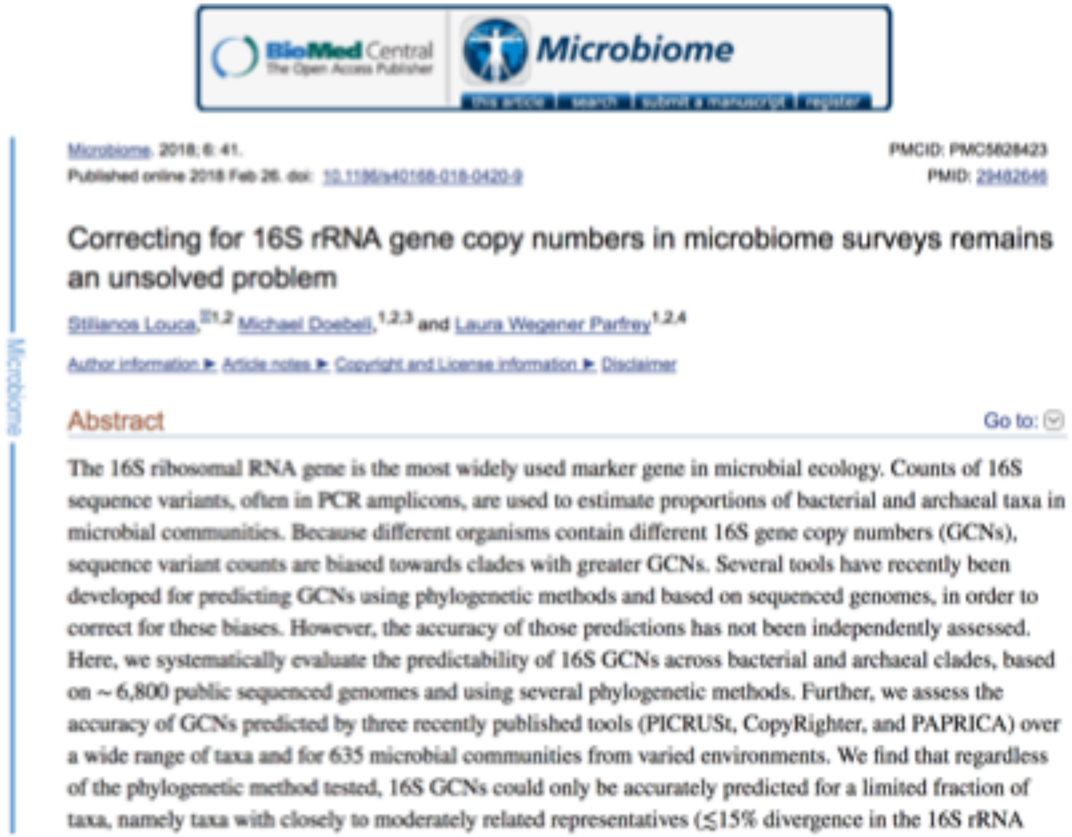
Learn more about preprints or browse peer-reviewed articles instead.

Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction methods

Jacob T Nearing[1], Gavin M Douglas[1], André M Comeau[2], Morgan G.I Langille[1,3]

February 23, 2018

Correcting for ribosomal copy number?



Microbiome. 2018; 6: 41.
Published online 2018 Feb 26. doi: 10.1186/s40168-018-0420-9

PMCID: PMC5828423
PMID: 29482646

Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem

Stilianos Louca,[1,2] Michael Doebeli,[1,2,3] and Laura Wegener Parfrey[1,2,4]

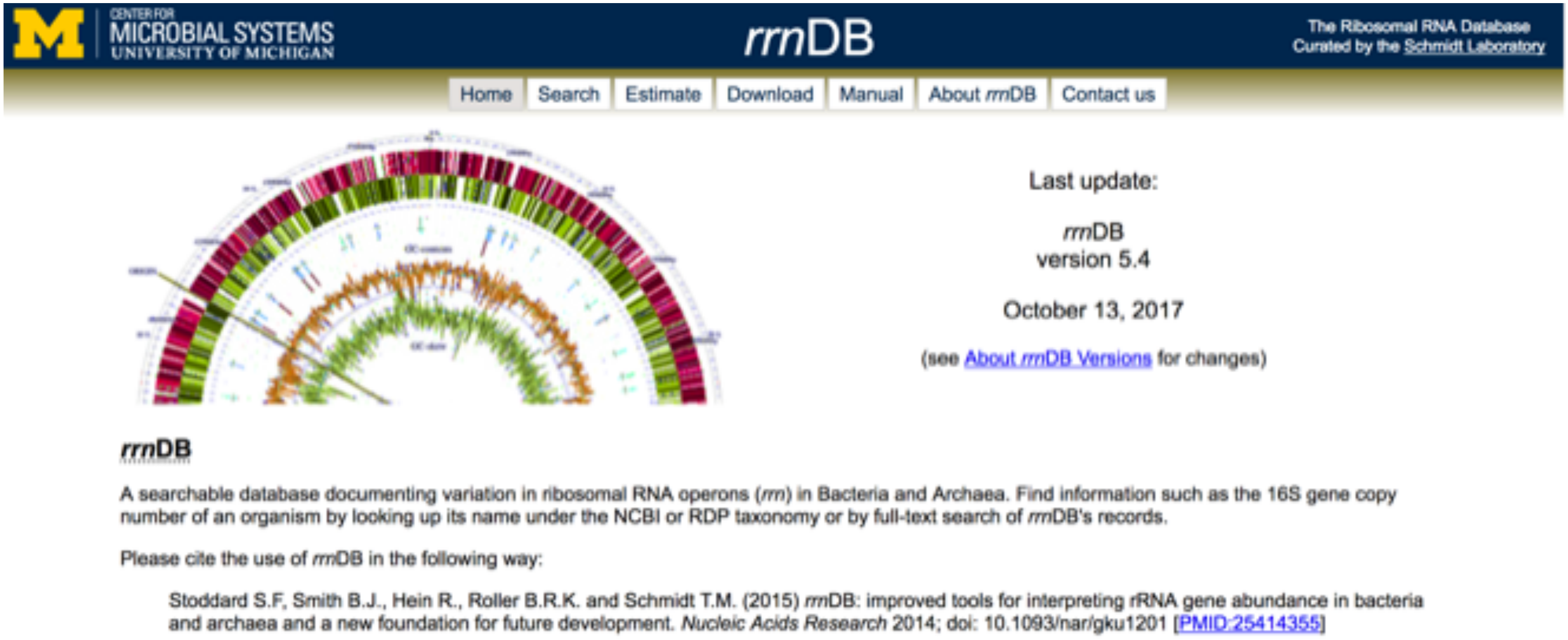
Author information ► Article notes ► Copyright and License information ► Disclaimer

Abstract

Go to: ⌄

The 16S ribosomal RNA gene is the most widely used marker gene in microbial ecology. Counts of 16S sequence variants, often in PCR amplicons, are used to estimate proportions of bacterial and archaeal taxa in microbial communities. Because different organisms contain different 16S gene copy numbers (GCNs), sequence variant counts are biased towards clades with greater GCNs. Several tools have recently been developed for predicting GCNs using phylogenetic methods and based on sequenced genomes, in order to correct for these biases. However, the accuracy of those predictions has not been independently assessed. Here, we systematically evaluate the predictability of 16S GCNs across bacterial and archaeal clades, based on ~6,800 public sequenced genomes and using several phylogenetic methods. Further, we assess the accuracy of GCNs predicted by three recently published tools (PICRUSt, CopyRighter, and PAPRICA) over a wide range of taxa and for 635 microbial communities from varied environments. We find that regardless of the phylogenetic method tested, 16S GCNs could only be accurately predicted for a limited fraction of taxa, namely taxa with closely to moderately related representatives (≲15% divergence in the 16S rRNA

Tools are being curated!



Last update:

*rrn*DB
version 5.4

October 13, 2017

(see About *rrn*DB Versions for changes)

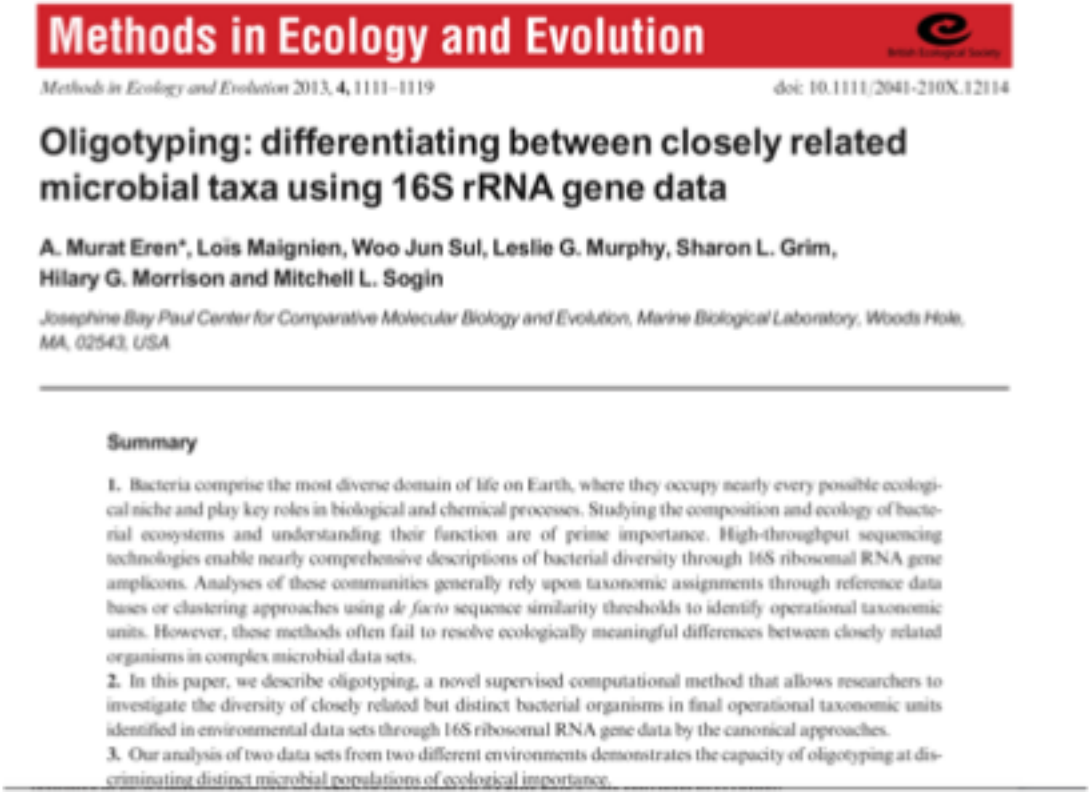
_rrn_DB

A searchable database documenting variation in ribosomal RNA operons (*rrn*) in Bacteria and Archaea. Find information such as the 16S gene copy number of an organism by looking up its name under the NCBI or RDP taxonomy or by full-text search of *rrn*DB's records.

Please cite the use of *rrn*DB in the following way:

> Stoddard S.F, Smith B.J., Hein R., Roller B.R.K. and Schmidt T.M. (2015) *rrn*DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* 2014; doi: 10.1093/nar/gku1201 [PMID:25414355]

Oligotypes: the classic approach to ASV



Methods in Ecology and Evolution

Methods in Ecology and Evolution 2013, **4**, 1111–1119 doi: 10.1111/2041-210X.12114

Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data

A. Murat Eren*, Lois Maignien, Woo Jun Sul, Leslie G. Murphy, Sharon L. Grim, Hilary G. Morrison and Mitchell L. Sogin

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, 02543, USA

Summary

1. Bacteria comprise the most diverse domain of life on Earth, where they occupy nearly every possible ecological niche and play key roles in biological and chemical processes. Studying the composition and ecology of bacterial ecosystems and understanding their function are of prime importance. High-throughput sequencing technologies enable nearly comprehensive descriptions of bacterial diversity through 16S ribosomal RNA gene amplicons. Analyses of these communities generally rely upon taxonomic assignments through reference data bases or clustering approaches using *de facto* sequence similarity thresholds to identify operational taxonomic units. However, these methods often fail to resolve ecologically meaningful differences between closely related organisms in complex microbial data sets.

2. In this paper, we describe oligotyping, a novel supervised computational method that allows researchers to investigate the diversity of closely related but distinct bacterial organisms in final operational taxonomic units identified in environmental data sets through 16S ribosomal RNA gene data by the canonical approaches.

3. Our analysis of two data sets from two different environments demonstrates the capacity of oligotyping at discriminating distinct microbial populations of ecological importance.

Oligotyping

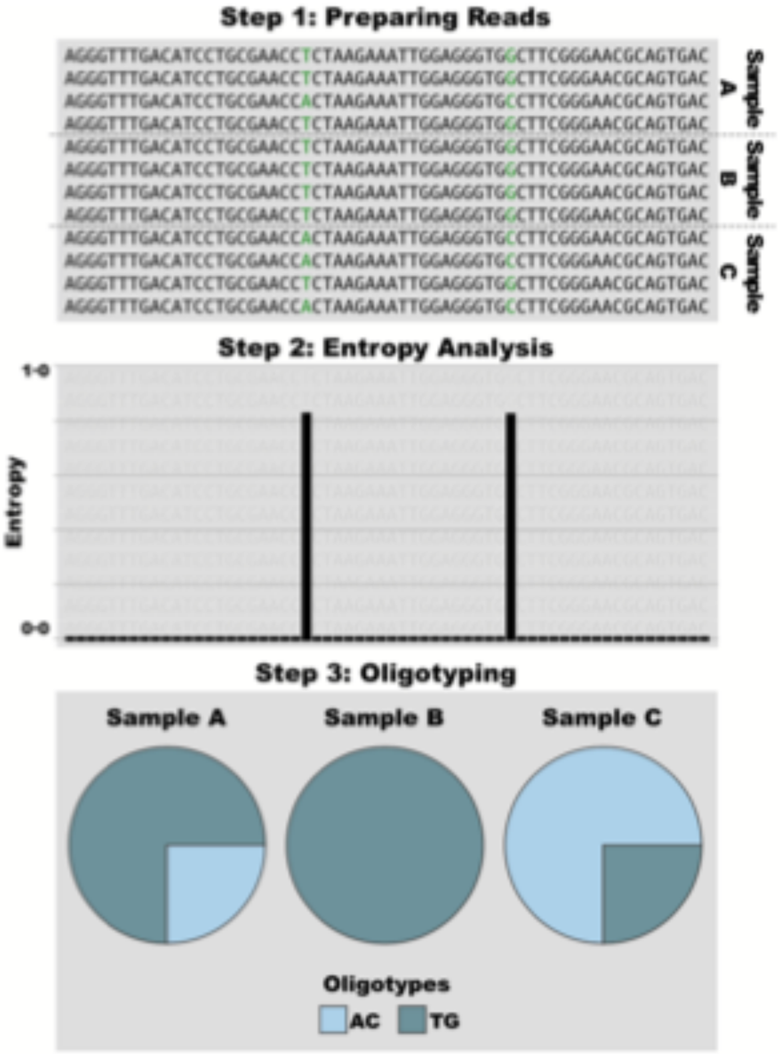
- Capitalizes on most “informative” variable sites along a nucleotide sequence
  - “systematic” identification of nucleotide positions that are informative
- Premise
  - Distinguish sequencing errors that result in insertions or deletions (indels) from real mutations that result in sequence variation from a shared evolutionary history
  - Key component: reads must be the same length (good for Illumina) and, if possible **high quality**
  - Uses Shannon Entropy to determine informative sequence variants

Shannon Entropy

- Quantify the uncertainty at each nucleotide position
- Think about: the amount of information that can be gleaned from a nucleotide position.
  - Information has an element of "surprise"
- Probability distribution: one event has a probability of 1, and all else of 0.
- Maximum entropy = log(n), where n is number of events
  - Events are like surprises – shifts in nucleotide content at the same nucleotide position, across similar sequences
  - For a nucleotide position that contains all 4 bases, the entropy is log(4)
  - For a position that contains only 2 bases, the entropy is log(2)
  - For a position that contains 1 base, the entropy is log(1) == 0
- Positions with maximum entropy are flagged as informative for oligotyping

Oligotyping workflow

These seqs are all clustered as the same OTU


Step 1: Preparing Reads
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCACTAAGAAATTGGAGGGTGCCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCACTAAGAAATTGGAGGGTGCCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCACTAAGAAATTGGAGGGTGCCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCTCTAAGAAATTGGAGGGTGGCTTCGGGAACGCAGTGAC
AGGGTTTGACATCCTGCGAACCACTAAGAAATTGGAGGGTGCCTTCGGGAACGCAGTGAC
Sample A
Sample B
Sample C
Step 2: Entropy Analysis
1·0
Entropy
0·0
Step 3: Oligotyping
Sample A
Sample B
Sample C
Oligotypes
AC
TG

Oligotyping is a supervised, iterative process

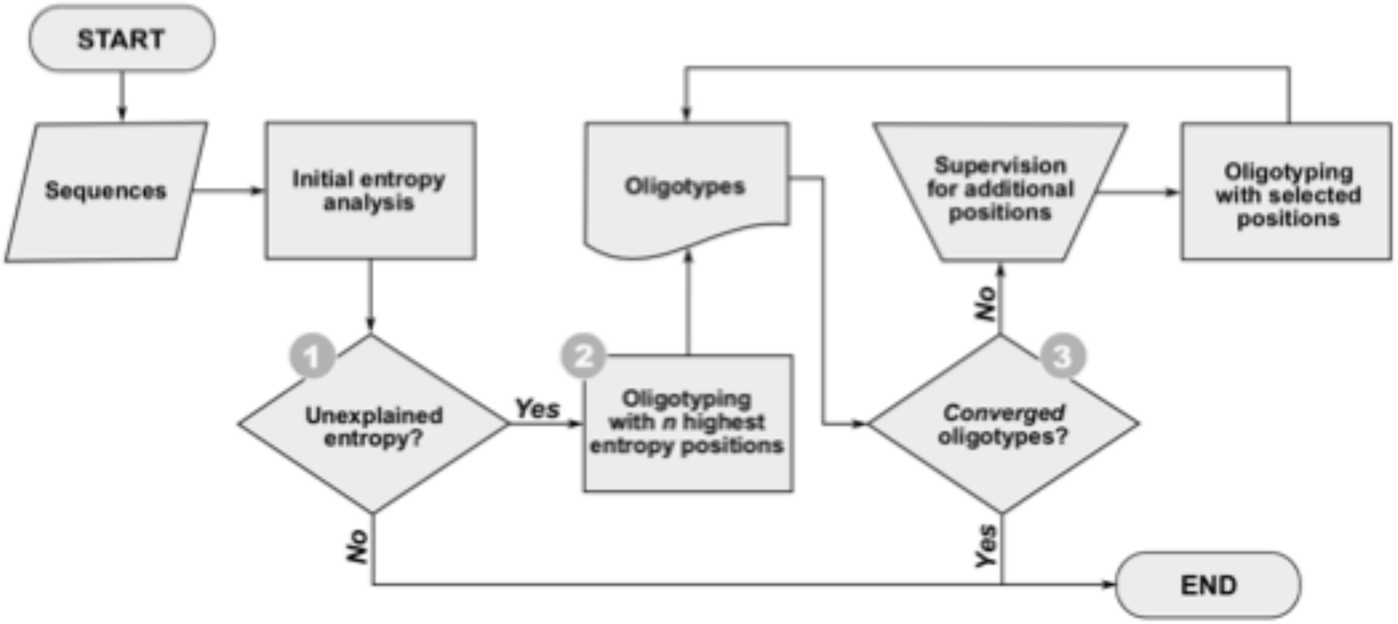
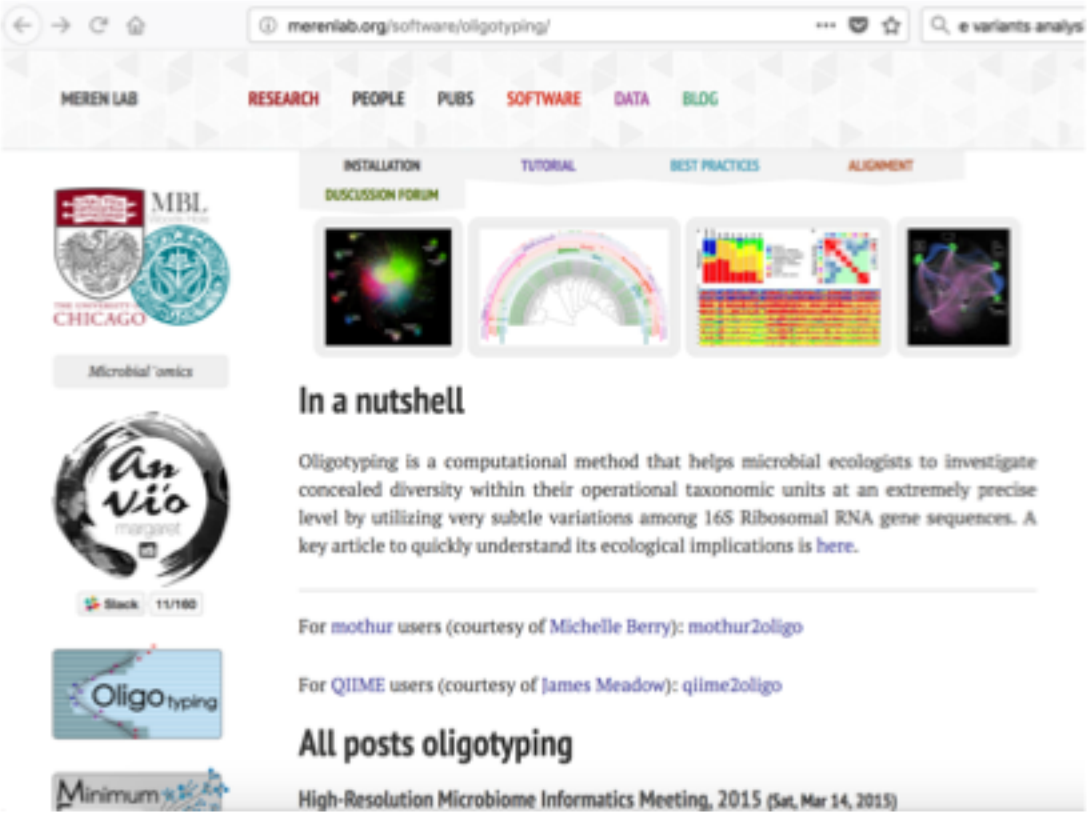

START
Sequences
Initial entropy analysis
1
Unexplained entropy?
Yes
No
2
Oligotyping with *n* highest entropy positions
Oligotypes
Converged oligotypes?
3
No
Yes
Supervision for additional positions
Oligotyping with selected positions
END

Fig. 1. Flowchart of an oligotyping analysis

http://merenlab.org/software/oligotyping/



In a nutshell

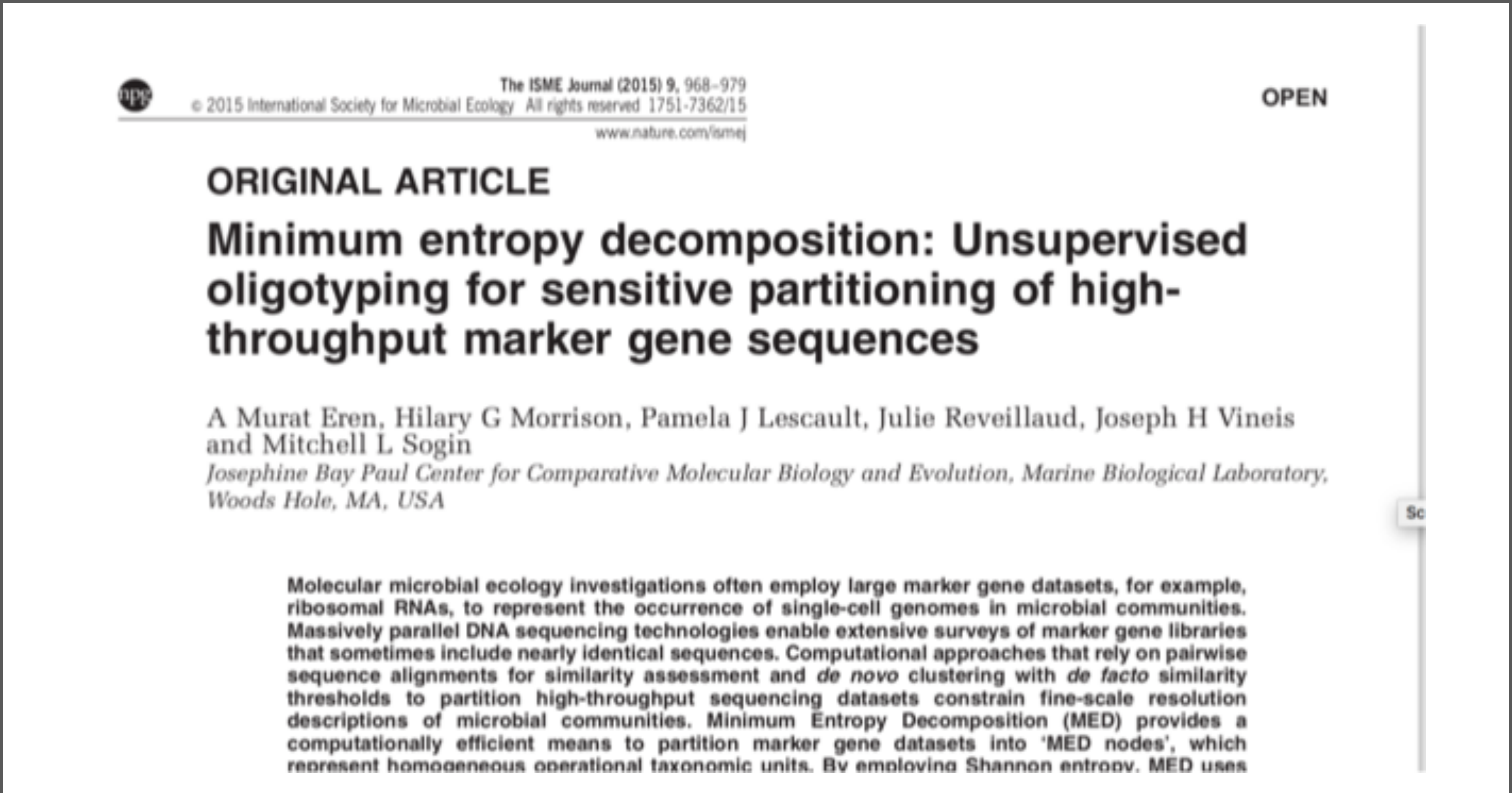
Oligotyping is a computational method that helps microbial ecologists to investigate concealed diversity within their operational taxonomic units at an extremely precise level by utilizing very subtle variations among 16S Ribosomal RNA gene sequences. A key article to quickly understand its ecological implications is here.

For mothur users (courtesy of Michelle Berry): mothur2oligo

For QIIME users (courtesy of James Meadow): qiime2oligo

All posts oligotyping

High-Resolution Microbiome Informatics Meeting, 2015 (Sat, Mar 14, 2015)



The ISME Journal (2015) 9, 968–979
© 2015 International Society for Microbial Ecology All rights reserved 1751-7362/15
www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences

A Murat Eren, Hilary G Morrison, Pamela J Lescault, Julie Reveillaud, Joseph H Vineis and Mitchell L Sogin
Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA

Molecular microbial ecology investigations often employ large marker gene datasets, for example, ribosomal RNAs, to represent the occurrence of single-cell genomes in microbial communities. Massively parallel DNA sequencing technologies enable extensive surveys of marker gene libraries that sometimes include nearly identical sequences. Computational approaches that rely on pairwise sequence alignments for similarity assessment and *de novo* clustering with *de facto* similarity thresholds to partition high-throughput sequencing datasets constrain fine-scale resolution descriptions of microbial communities. Minimum Entropy Decomposition (MED) provides a computationally efficient means to partition marker gene datasets into 'MED nodes', which represent homogeneous operational taxonomic units. By employing Shannon entropy, MED uses