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

Eren et al. 2015

In a nutshell…

- Categorizing DNA sequences into ecologically meaningful units
- Marker gene analyses of microbial diversity

Metagenomic analysis: General pipeline

AG-AGCTGCTCA
AG-AGCTGCTCA
AG-AGCTGCTCA
AGGAGGTGCA-A
AGGAGGTGCA-A
AAA--CTGCTCA
AGGAGCT--TCA

Approaches

1. Alignment against genomic reference databases
2. De novo clustering by sequence similarity of operational taxonomic units (**OTUs**)

Similarity matrix

	U1	U2	U3
U1	1	0.1	0.7
U2	0.1	1	0.3
U3	0.7	0.3	1

Alignment against databases VS OTUs clustering?

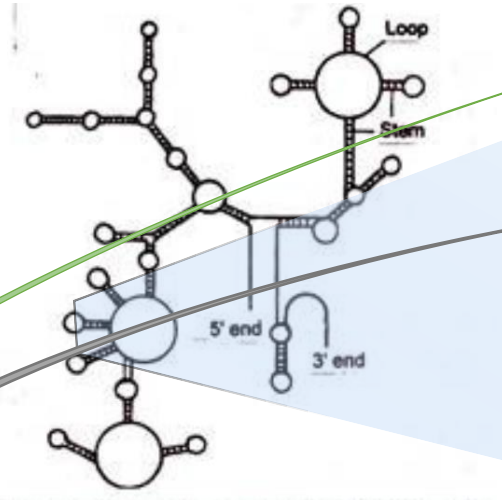
- Reference databases are very sparse compared to astonishing number of uncultured microbes
- In contrast, OTUs clustering is database independent

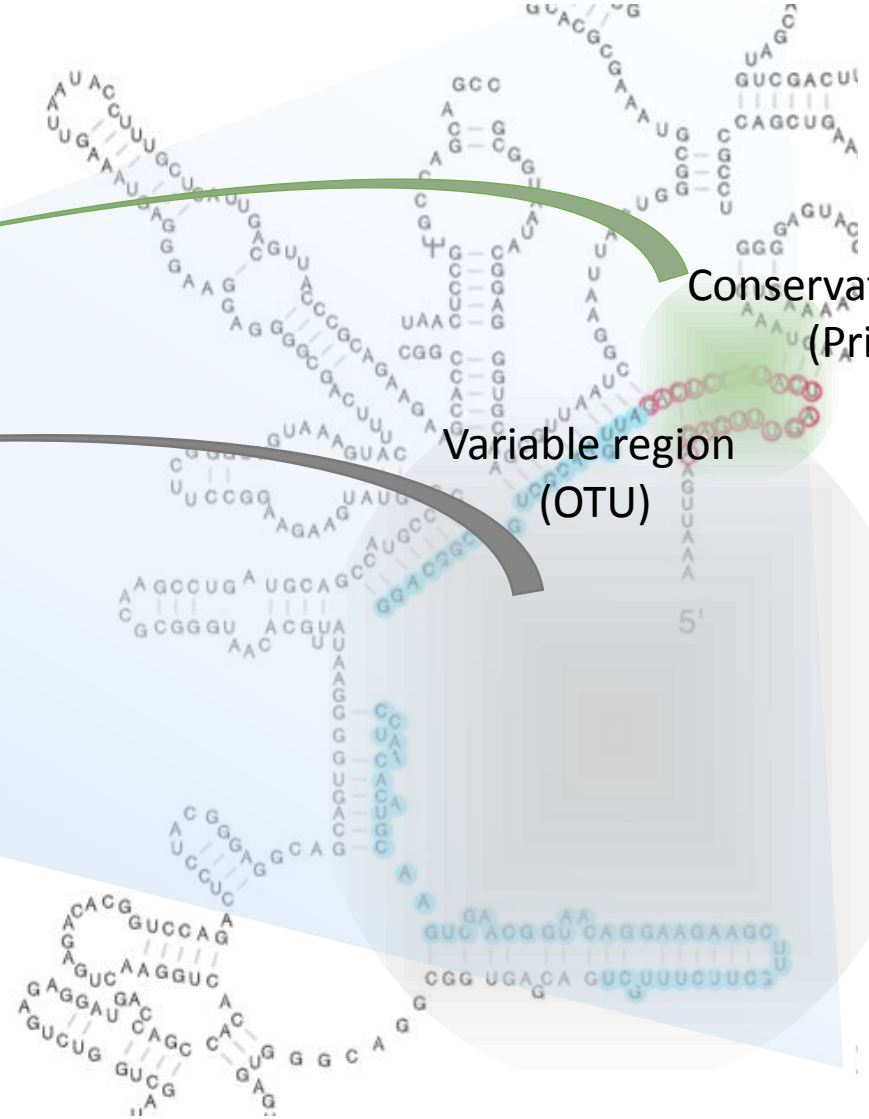
Targeted Sequencing

- Relies on the assumption that some specific regions can be used to distinguish between different organisms
- **16S rRNA** (ribosomal RNA) is present in most Prokaryotes.
- Typically, in contain conservative region and variable region
- We can utilize the conservative regions as ”anchors” to detect (and amplify) variable regions that come right after them
- These variable regions are used to distinguish different organisms
- **OTUs!**

Targeted Sequencing - Illustration

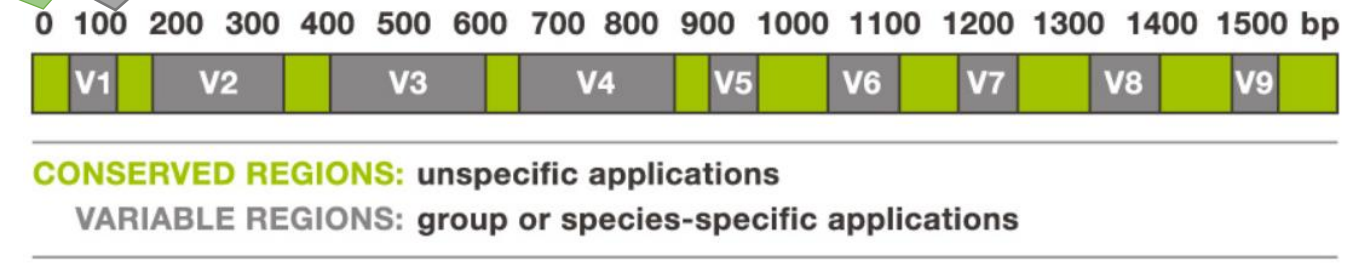
16S rRNA
(simplified)





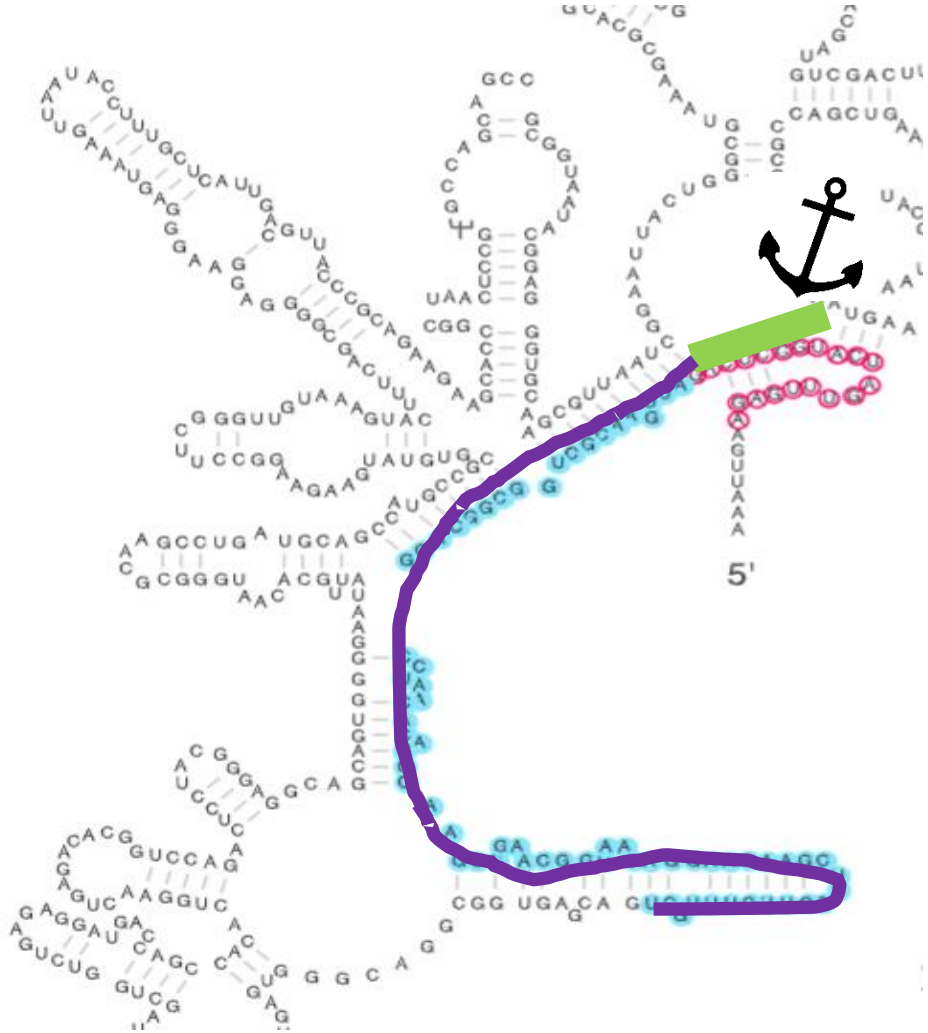
Conservative Region
(Primer)

Variable region
(OTU)

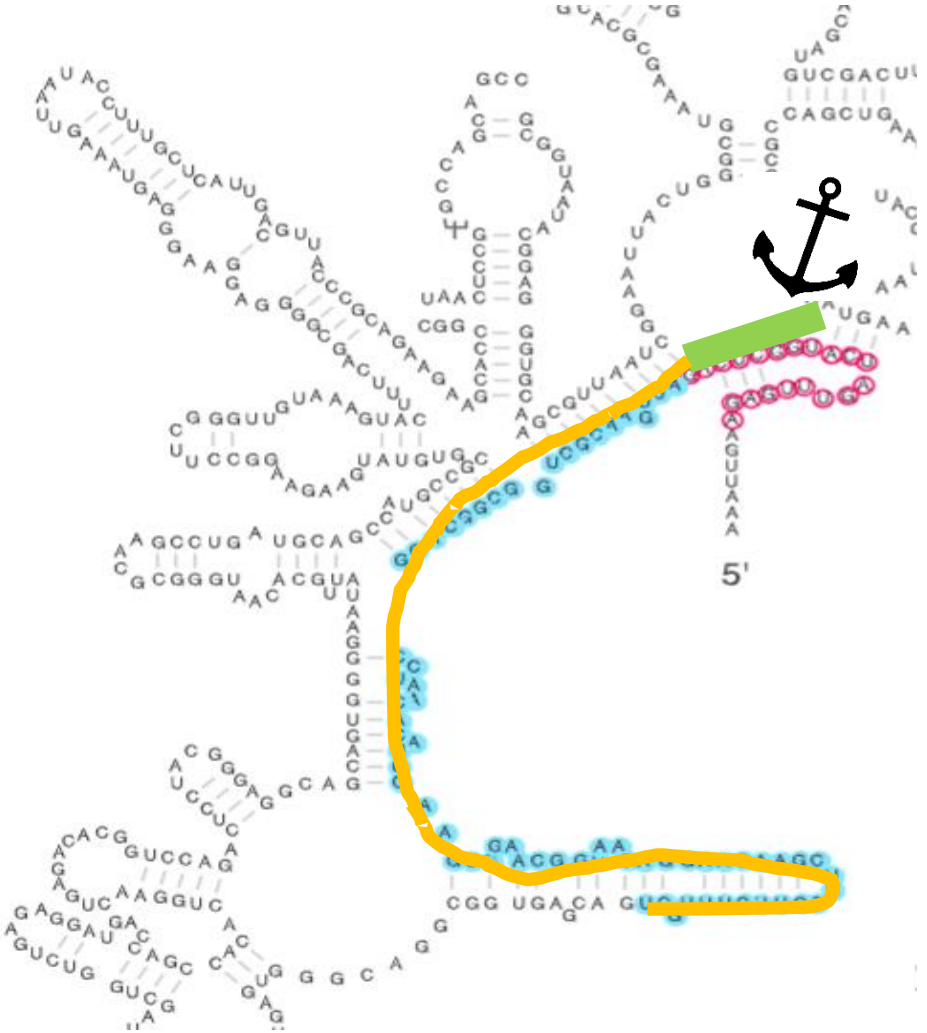


Targeted Sequencing - Illustration

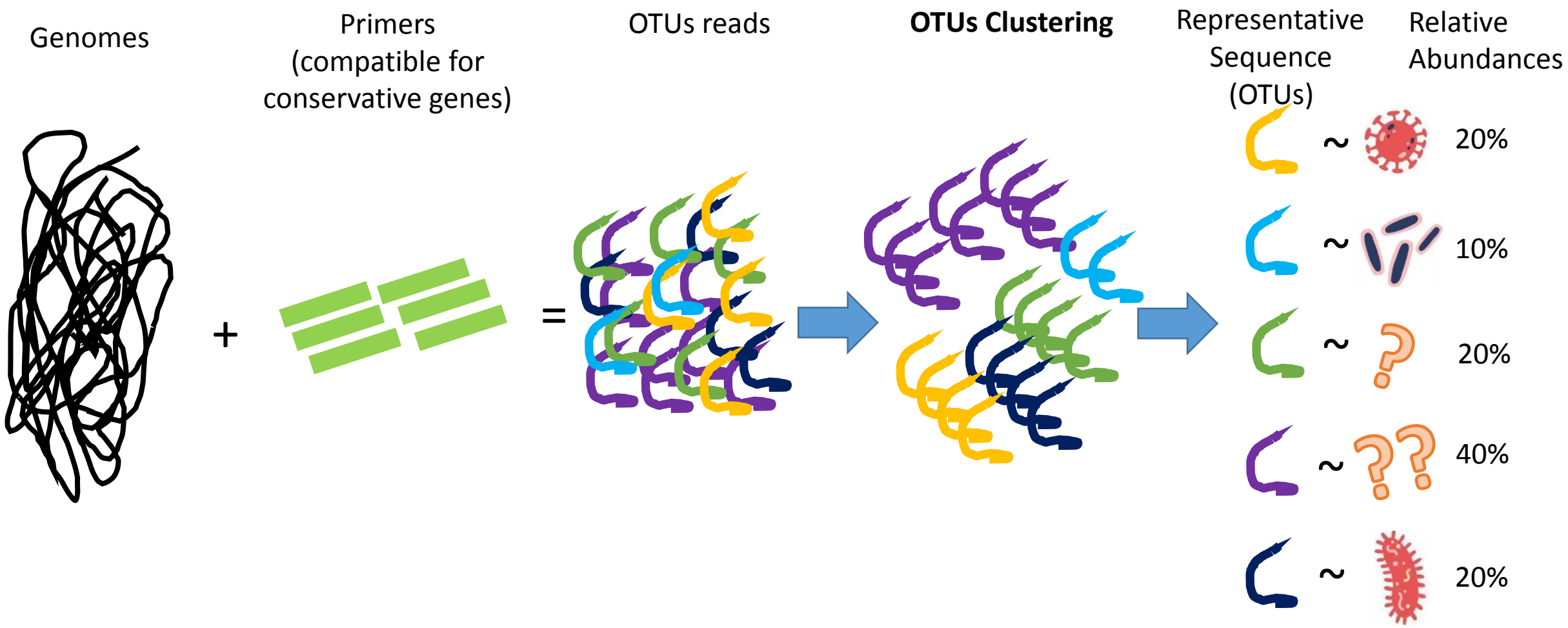
Bacteria type A



Bacteria type B



Targeted Sequencing - Illustration



OTU clustering – how?

1) Compare all the sequences to each other

2) Group similar sequences
   - 3% commonly considered as technical/sequencing error

3) Output a representative sequence from each group

- Some of the more recent methods use more heuristic/greedy approaches for efficiency

OTUs clustering - issues

- 97% similarity might be too relaxed, yielding clusters (rep. OTUs) that aggregate distinct organisms.

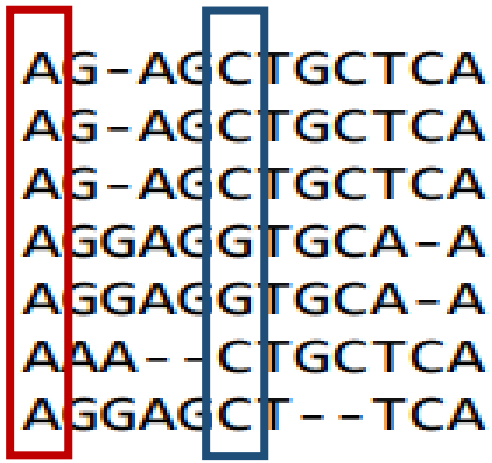
- Pairwise similarity could be computationally heavy task

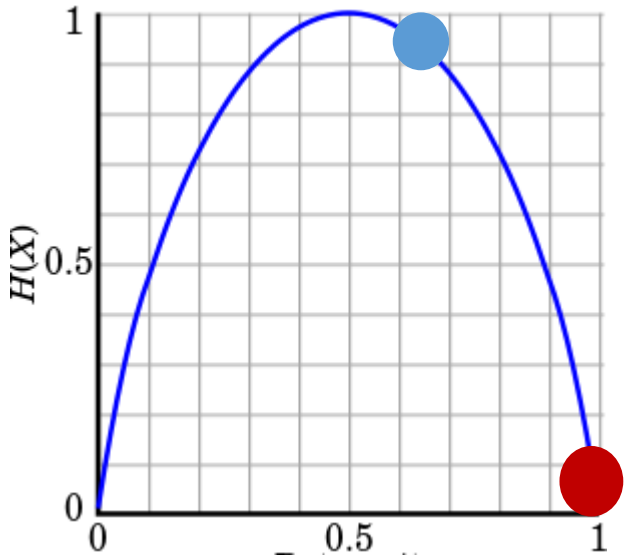
Oligotyping (Eren et al. 2013)

- Some nucleotide are more ”meaningful” for OTU clustering task.
- Oligotyping uses Shannon entropy to identify biologically meaningful signals (e.g SNP) from noise (e.g seq. errors)
- No pairwise sequence similarity required

Shannon Entropy (information theory)

- **Entropy** is level of uncertainty
- Shannon Entropy: $S = -\sum_i P_i \log P_i$





- The higher the entropy in a position– the “valuable” that position is

Oligotyping (con’d)

- Pros:
  - facilitates the identification of closely related but distinct organisms that may differ by as little as one nucleotide out of hundreds (overcome the 3% barrier)
  - Computationally efficient compared to naïve pairwise similarity

Oligotyping (con’d)

- Cons:
  - Computationally efficient over closely related taxa, but not over distantly related organisms.

    Therefore, a preprocessing step is required in order to identify closely related taxa and then apply Oligotyping on each taxon separately

Minimum Entropy Decomposition (MED) Eren et al. 2015

- An improvement of Oligotyping

- Can be applied directly over entire dataset instead of group of closely related taxa - no preprocessing step is required!

How MED works?

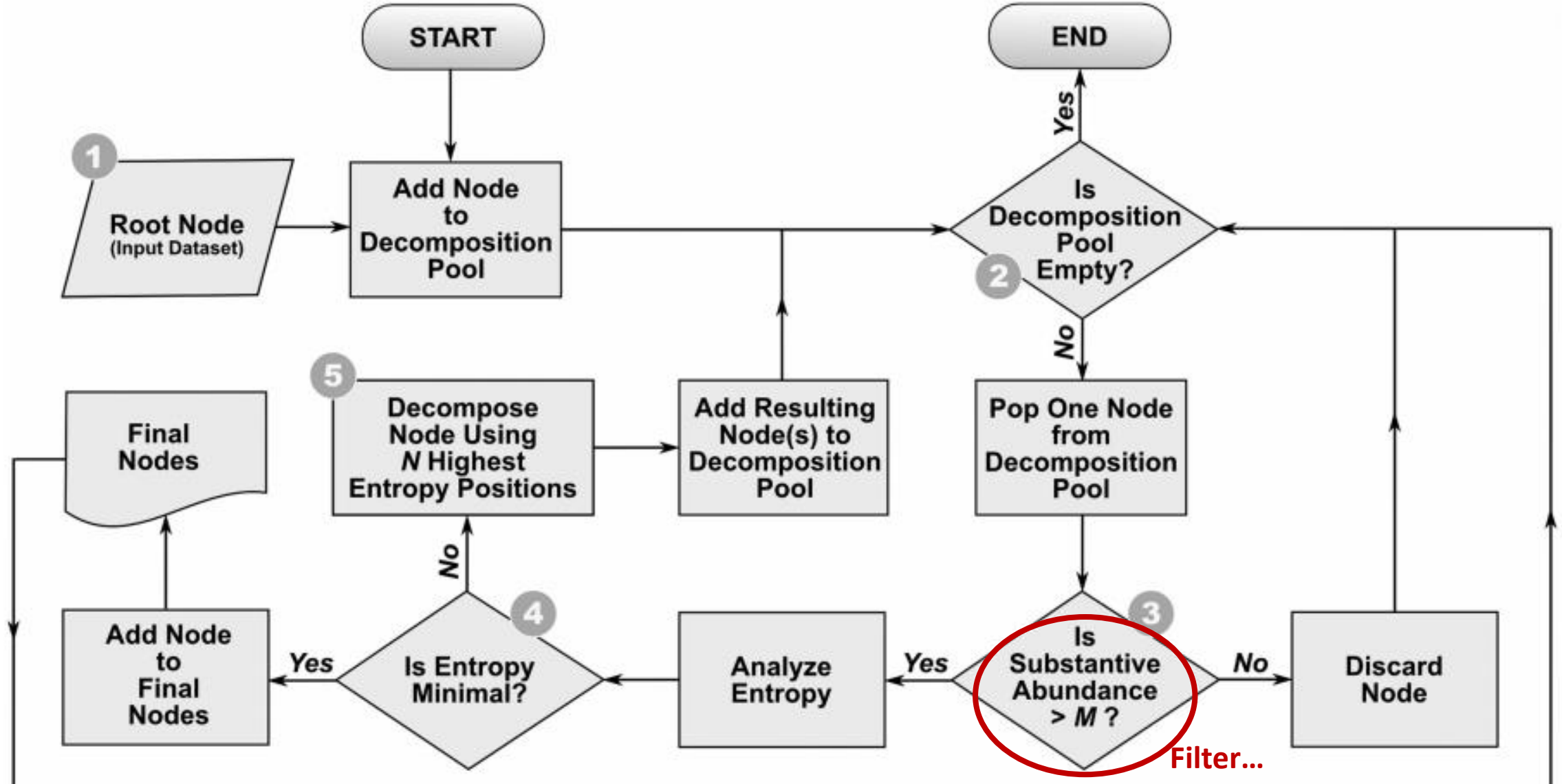


Illustration in context

Nodes decomposition goes from higher and leftmost position

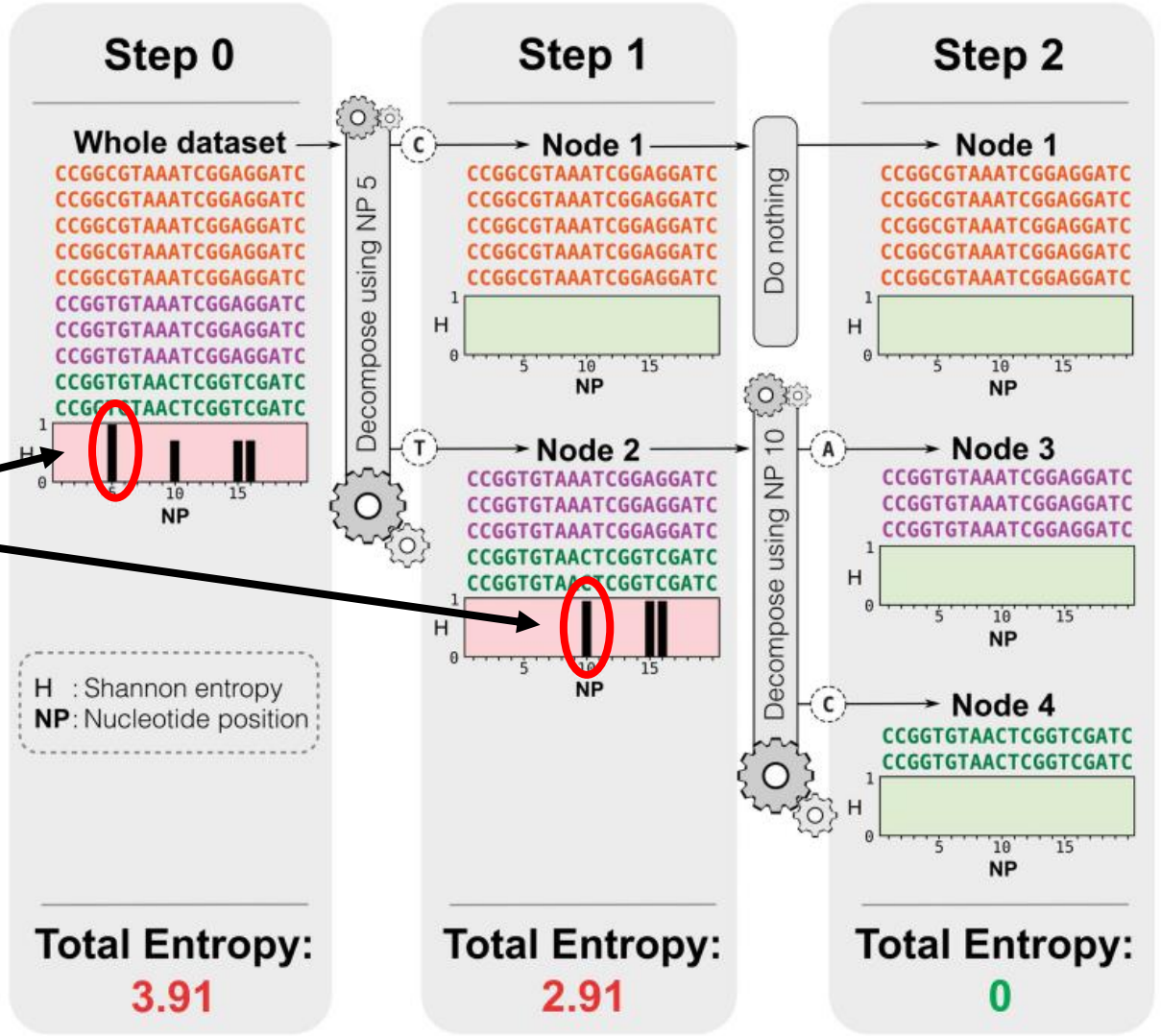
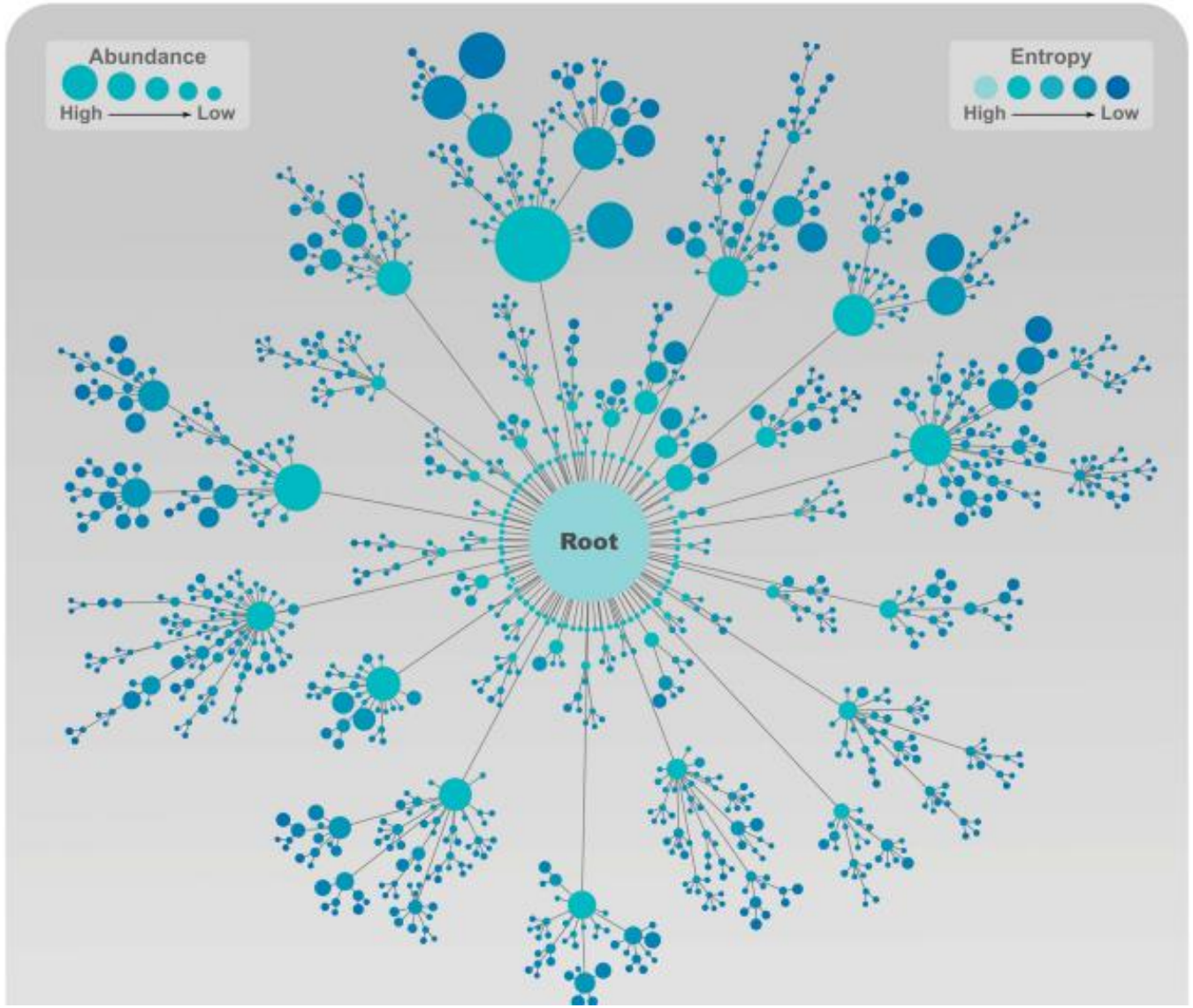


Illustration of results



Comparison

Set-up

- Datasets:
  - Microbiome from deep-sea sponge cryptic species
    - 6 samples of H. dedritifera,
    - 13 samples of H. cf. dedritifera
    - 5 samples of local water samples
  - Oral microbiome of The Human Microbiome Project Consortium
    - 6 million samples

Evaluation criteria

- Sponges microbiome:
  - Distinguish between different microbiome communities
  - ➢ Adversaries: OTUs clustering, GAST (old-school taxonomy assignment)

- Oral cavity microbiome:
  - Explained cluster variances calculated for each method
  - Ratio of between-group variance to within-group variance
  - Variation associated with different sites in oral cavity
  - Recovery of previously identified taxa against a reference database
  - Adversaries: OTUs clustering, GAST, Oligotyping (for variances), BLAST (for recovery task)

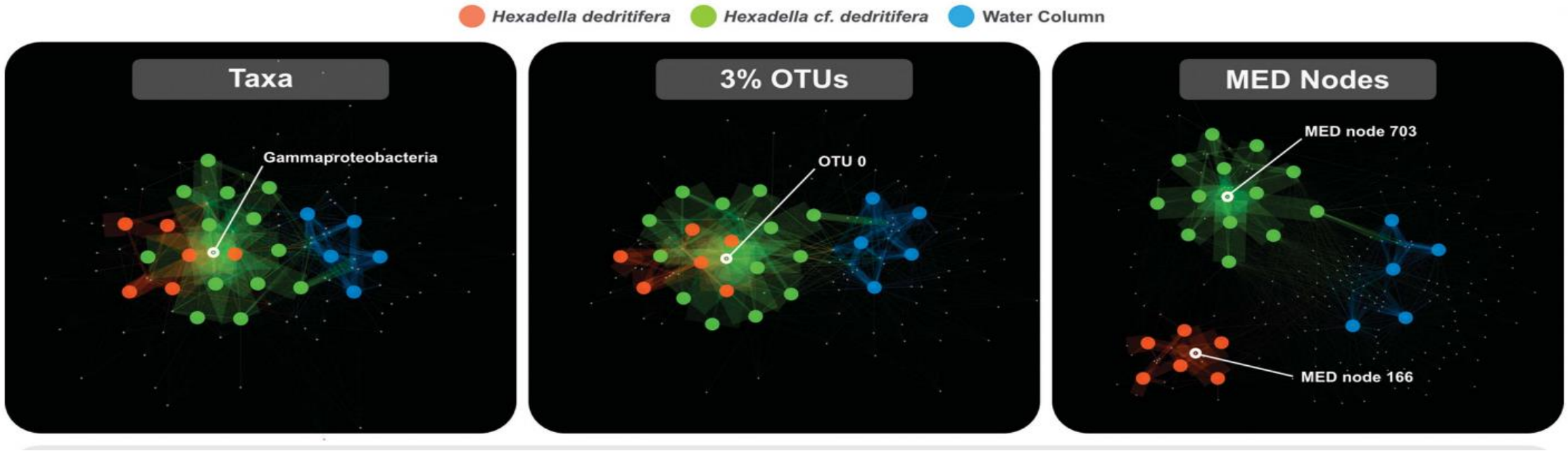
Number of clusters

- Oral Microbiome:
  - GAST: 122 OTUs: 329, MED 858, Oligotyping: 481

- Sponges Microbiome:
  - GAST: 80, OTUs: 91, MED: 187

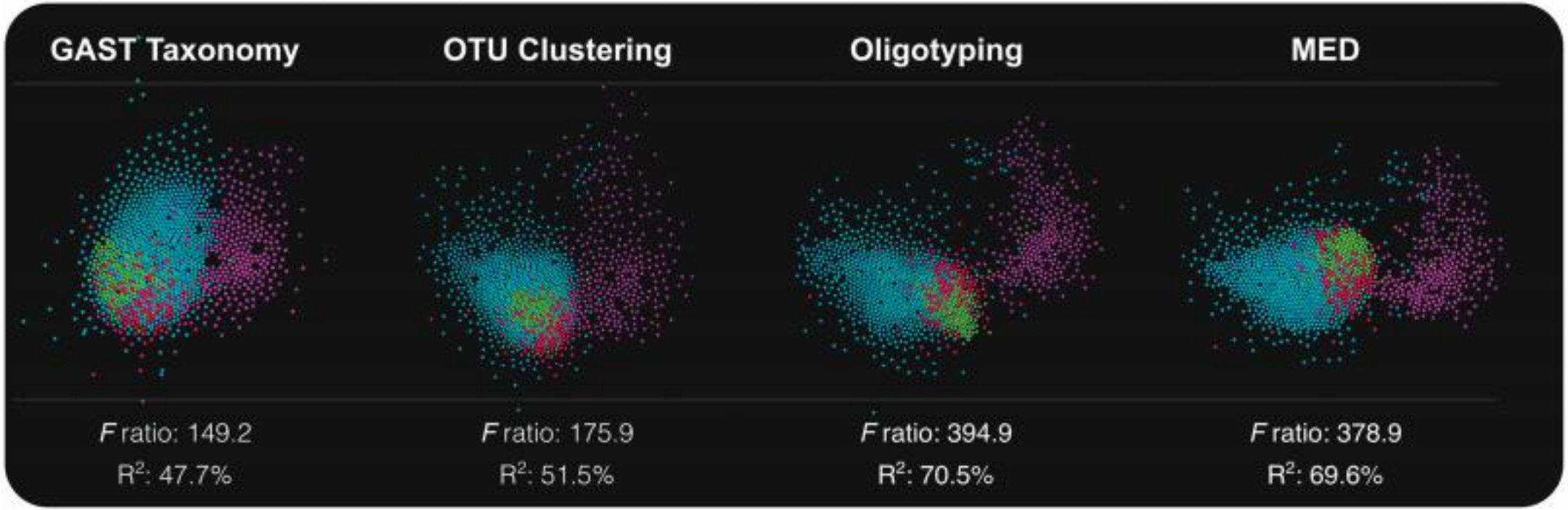
Distinguish between different microbiome communities

- 3 nt (99.2%) differences between MED 703 to MED 166 communities

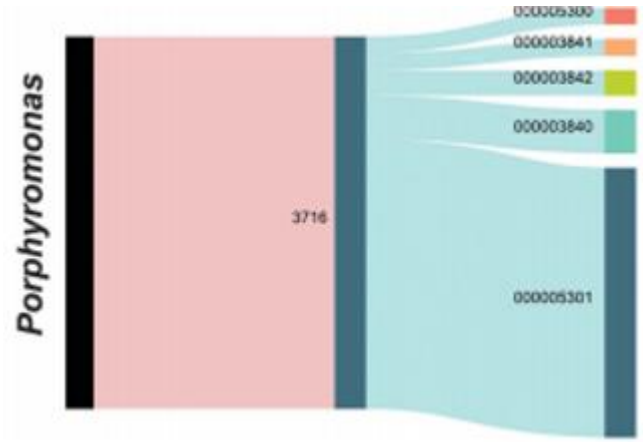


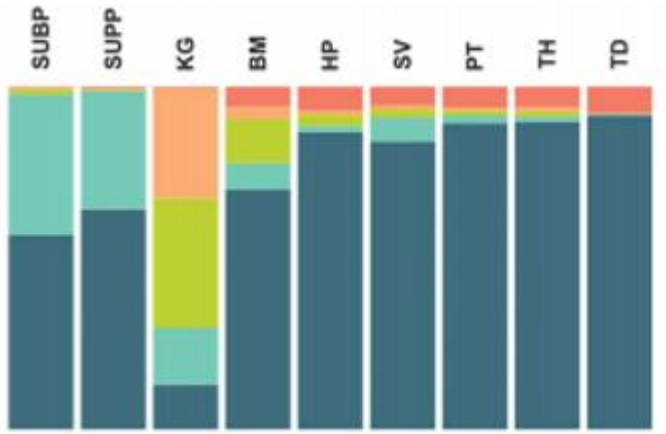
Ratio of between-group variance to within-group variance (F Ratio)

Explained cluster variances calculated for each method (R^2)



Variation associated with different sites in oral cavity





Recovery of previously identified taxa against a reference database

- Number of taxa to be found: 248
- Identified taxa: BLAST: 67 (27%), OTUs clustering: 112 (45%), MED: 235 (95%)
- Limiting # of MED clusters to be the same as OTUs' (n=329) results less impressive, but still better result (138)

Final words…

- MED (and Oligotyping) is a cornerstone in the evolution of the old-school OTUs clustering approach

- Much more clusters - might prone to over-sensitivity.
  - “False positive” evaluation (Artificial generated data?)
  - Omit 20% of samples and measure how much cluster changes

10X 4  ☺