OTUs or ASVs?

Profiling Microbial Communities from Amplicon Sequence Data

Morgan Essex, AG Forslund

4 May 2021

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

MDC
MAX DELBRÜCK CENTER
FOR MOLECULAR MEDICINE
IN THE HELMHOLTZ ASSOCIATION

Agenda

- Brief introduction and context
- Discuss Pat Schloss preprint
- Critical discussion

1) Quality of the data:

```
Do you consider appropriate the techniques used (patients selection, sample processing, etc.)?
Were the analysis (statistical methods) and interpretations carefully addressed?
Is the data presented with sufficient detail?
```

2) Support of conclusions:

```
Do the authors provide sufficient evidence to sustain their conclusions?
Does the experimental design counts with appropriate controls?
Could you perceive a systematic bias in the interpretation of the results?
Could you perceive a statistical bias in the analysis of the data?
Is there any bias in the presentation of the results towards a specific conclusion?
```

3) Potential significance:

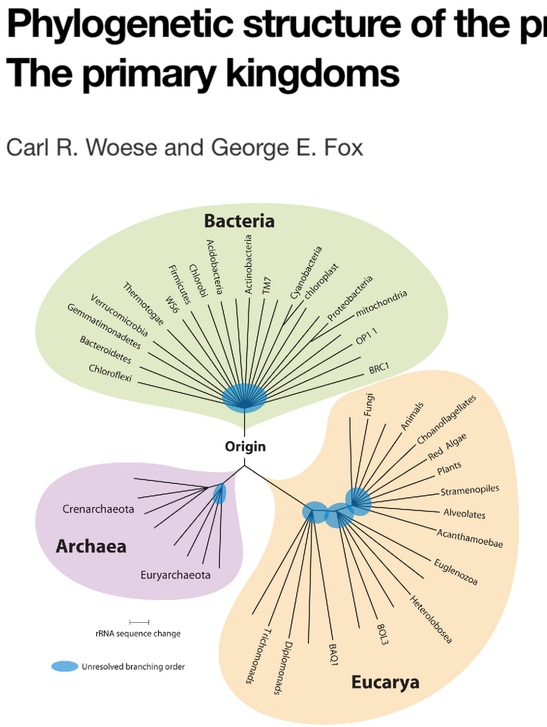
```
Are the results and conclusions relevant to the field?
Do they provide and advance in the understanding of the topic?
```

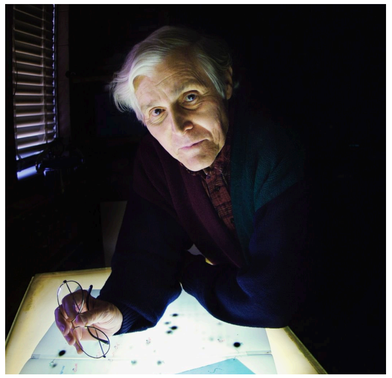
4) Additional comments:

```
Is the manuscript written clearly?
Are the figures and captions clear and detailed enough?
Is the abstract and general overview clear?
What would you suggest to improve this manuscript?
What finding, experimental or analytical method do you consider useful for your project?
```

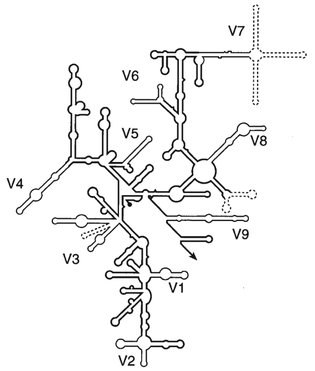
A little bit of history

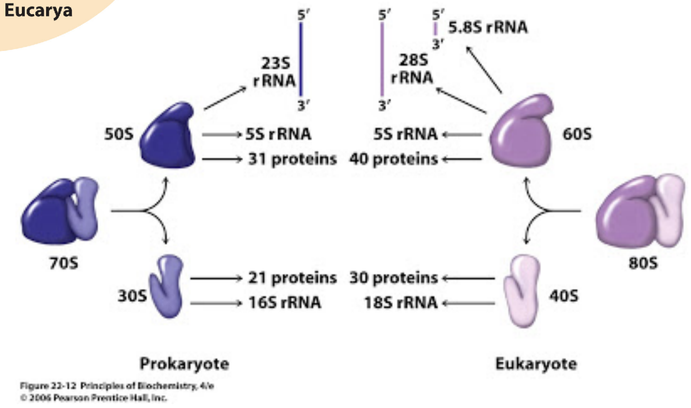
- How to identify and classify living organisms?
  - Taxonomy: similarities and dissimilarities
  - Phylogeny: evolutionary relationships and history
- 16S rRNA: the first marker gene (1977)
  - Found in all known forms of life
  - Copied to each generation with high accuracy
  - Mutation rates make it “evolution’s timekeeper”
- Identification of taxa went from organismal to cellular (prokaryotes vs eukaryotes) to molecular
- Enabled culture-independent profiling of microbial communities and eventually metagenomics





Carl Woese: The Man Who Rewrote the Tree of Life [link]





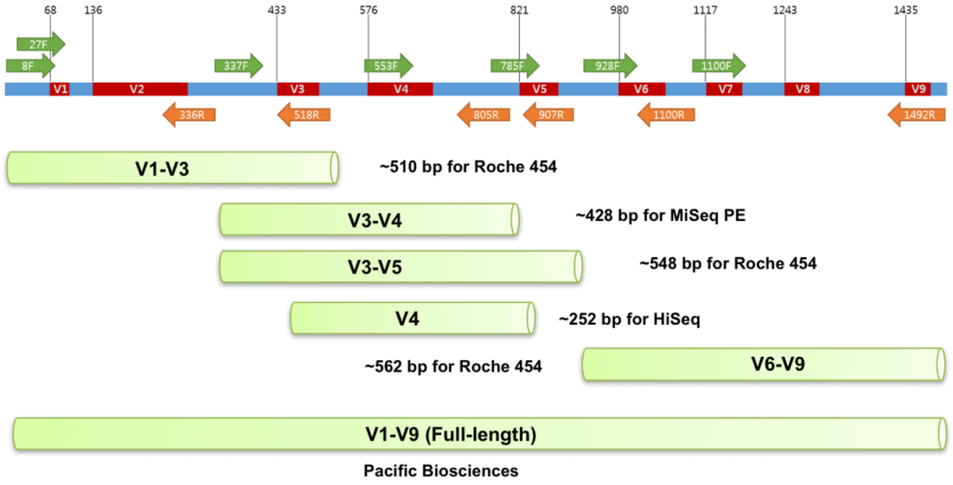
Further complicating matters

- Bacteria and Archaea can have multiple copies of the 16S gene
- Nine hypervariable regions, studies usually sequence one or two
- Species in e.g. some families have 99% similarity across the whole sequence (V1-V9)
  - Enterobacteriaceae, Peptostreptococceae, Clostridiaceae
- V4 may only differ by a couple of nucleotides — is it a different organism or an error?
- Operational taxonomic units (OTUs) are *distance-based clusters* of similar sequence variants intended to mitigate this uncertainty
- Many algorithms and definitions, and expanding

Short report | Open Access | Published: 26 February 2018

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

Stilianos Louca, Michael Doebeli & Laura Wegener Parfrey



Published: 18 August 2013

UPARSE: highly accurate OTU sequences from microbial amplicon reads

Robert C Edgar

Nature Methods **10**, 996–998(2013)

[link] to 2020 benchmark

Software | Open Access | Published: 30 September 2014

LotuS: an efficient and user-friendly OTU processing pipeline

Falk Hildebrand, Raul Tadeo, Anita Yvonne Voigt, Peer Bork & Jeroen Raes

Fundamental classification units

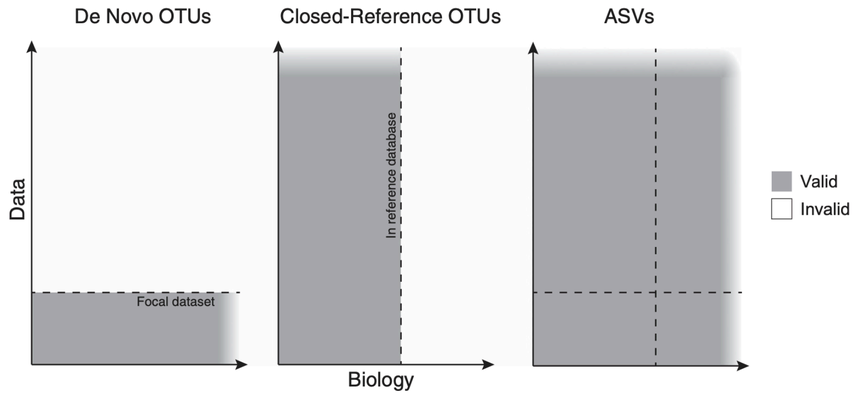
- "OTU-picking" comes after quality control filtering
  - *De novo*: reads are clustered against one another (no reference database needed)
  - Closed-reference: reads not mapped to reference database(s) are thrown out
  - Open-reference: reads not mapped to reference databases are then clustered *de novo*
- Amplicon Sequence Variants (ASVs) aka zero-radius OTUs, represent *de novo* DNA sequences recovered from a high-throughput experiment
  - Discriminate biological and error sequences
  - More reproducible unit than *de novo* OTUs
  - Can offer species-level resolution (can they?)
  - Inference is not performed on reads but on samples (quality matters more?)

DADA2 Amplicon Sequencing. Exactly. Version 1.18

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*



Amplicon sequence variants artificially split bacterial genomes into separate clusters

mothur

Patrick D. Schloss

doi: https://doi.org/10.1101/2021.02.26.433139

This article is a preprint and has not been certified by peer review [what does this mean?].

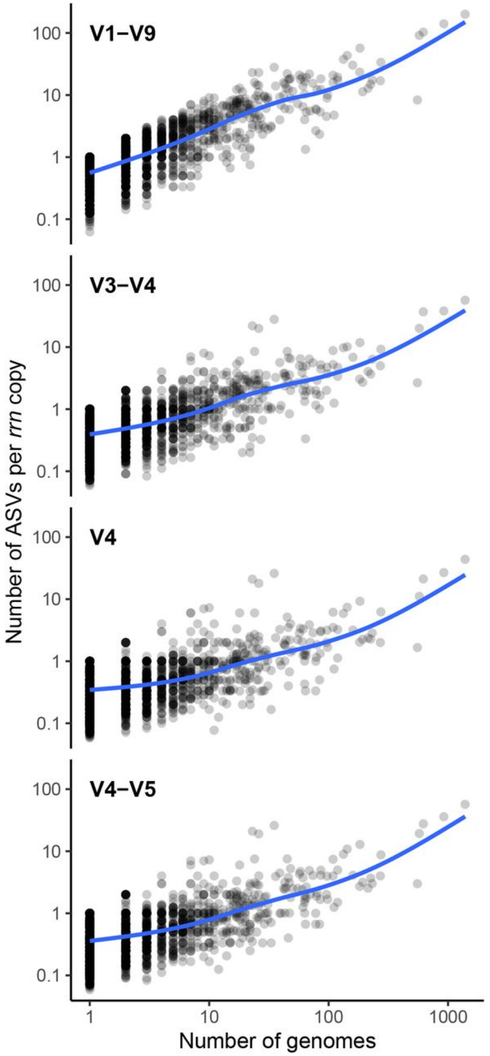
Copy number and intra-genomic variation

Fig. S1

- Bacteria contain multiple, non-identical copies of the 16S gene
- *Staphylococcus aureus* and *S. epidermis*
  - Each have 5 distinct 16S copies = 10 ASVs
  - With 3% distance-based threshold = 1 OTU

- Methods
  - Obtained 20,427 genomes covering 5,972 species represented in the *rrn* copy number database (*rrn*DB) and count variants
  - Median *rrn* copy number per species ranged from 1-19

- As copy number increased, so did number of variants
- Average of 0.58 variants per copy of the full sequence (V1-V9)
- *E. coli* had 1,390 genomes and 1,402 versions of the gene



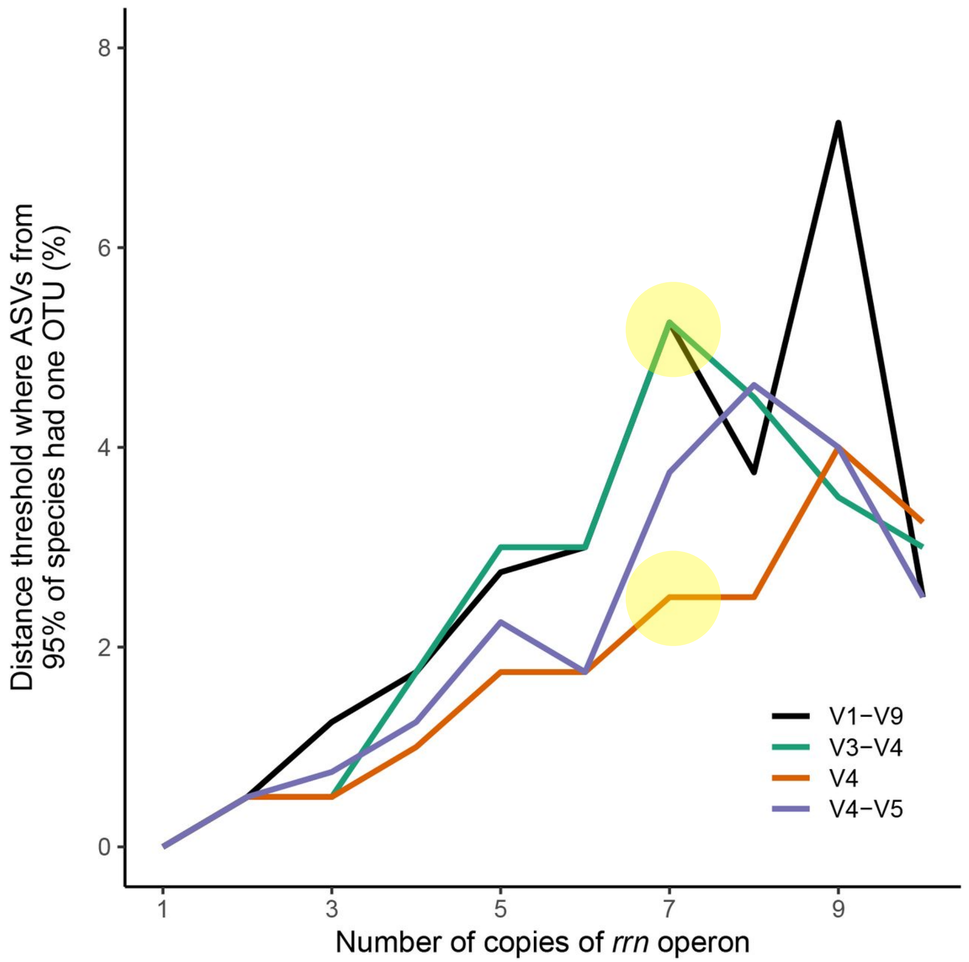
Propensity for genome splitting

Fig. 1

- How does the *rrn* copy number relate to the propensity of a single genome to be split into different clusters?

- Methods
  - Randomly select one genome from each species
  - Calculate distance is required such that 95% of ASVs cluster into a single OTU
  - Repeat 100x —> median shown

- As number of copies increases, a higher distance threshold is needed to avoid splitting genomes
  - *E. coli* (7 copies) required 94.75% sequence identity for full sequence and 97.5% for the V4 region



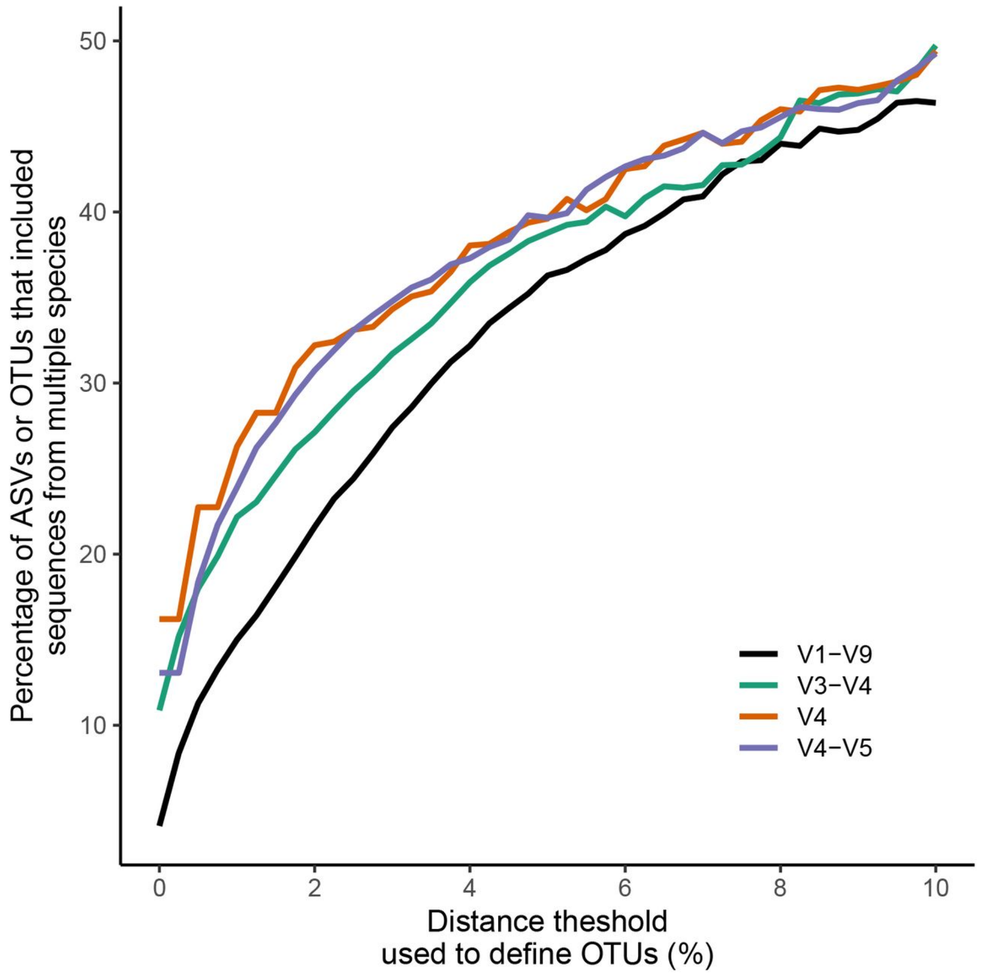
Multiple species in the same OTU

Fig. 2

- Methods
  - Randomly select one genome from each species
  - Measure percentage of ASVs and OTUs that include sequences from multiple species
  - Repeat 100x —> median shown

- No threshold applied:
  - V1-V9: 4.1% of OTUs
  - V4: 16.2%

- Common 3% threshold applied:
  - V1-V9: 27.4% of OTUs
  - V4: 34.3%



Summary

- Defining a meaningful taxonomic unit remains elusive
- Still no biological definition of a species
  - Naming is biased and taxonomic rules inconsistent
- 16S evolves differently in different bacterial lineages
  - Biological definition of OTU limited; at best rooted in theory
- Risk of splitting a genome into multiple clusters is more problematic than clustering together
  - Not plausible that different 16S copies would have different ecologies

- What do you think?

Critical questions

1. **Quality of the data**
    a. Do you consider appropriate the techniques used (patients selection, sample processing, etc.)?
    b. Were the analysis (statistical methods) and interpretations carefully addressed?
    c. Is the data presented with sufficient detail?
2. **Support of conclusions:**
    a. Do the authors provide sufficient evidence to sustain their conclusions?
    b. Does the experimental design counts with appropriate controls?
    c. Could you perceive a systematic bias in the interpretation of the results?
    d. Could you perceive a statistical bias in the analysis of the data?
    e. Is there any bias in the presentation of the results towards a specific conclusion?
3. **Potential significance:**
    a. Are the results and conclusions relevant to the field?
    b. Do they provide and advance in the understanding of the topic?
4. **Additional comments:**
    a. Is the manuscript written clearly?
    b. Are the figures and captions clear and detailed enough?
    c. Is the abstract and general overview clear?
    d. What would you suggest to improve this manuscript?
    e. What finding, experimental or analytical method do you consider useful for your project?

Extra

Operational Genomic Units (OGUs)

OGUs enable effective, phylogeny-aware analysis of even shallow metagenome community structures

Qiyun Zhu, Shi Huang, Antonio Gonzalez, Imran McGrath, Daniel McDonald, Niina Haiminen, George Armstrong, Yoshiki Vázquez-Baeza, Julian Yu, Justin Kuczynski, Gregory D. Sepich-Poore, Austin D. Swafford, Promi Das, Justin P. Shaffer, Franck Lejzerowicz, Pedro Belda-Ferre, Aki S. Havulinna, Guillaume Méric, Teemu Niiranen, Leo Lahti, Veikko Salomaa, Ho-Cheol Kim, Mohit Jain, Michael Inouye, Jack A. Gilbert, Rob Knight

doi: https://doi.org/10.1101/2021.04.04.438427

This article is a preprint and has not been certified by peer review [what does this mean?].

- Metagenomics should develop the same *de novo* phylogenetic inference methods that amplicon analysis has
- OGUs represent individual reference genomes from a database (taxonomy-free)
  - Operational like OTUs, exact like ASVs
- Small conceptual example: taxonomic assignment and phylogenetic placement of O5 not consistent