BioInfo4EEG

ECOSYSTEM ECOLOGY

Bioinformatics Course

Session 2

Theory: **Clustering** vs **Denoising**, reference **databases**

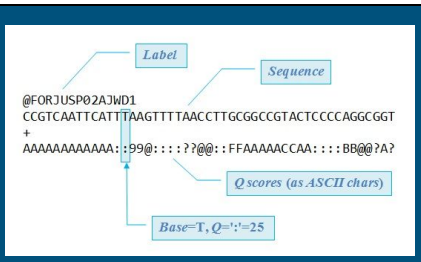
Workshop: **DADA2**, assign **taxonomy**

BioInfo4EEG

Recap from Session 1

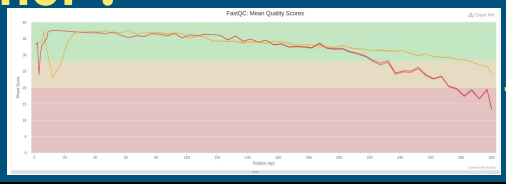
The **metabarcoding** technology

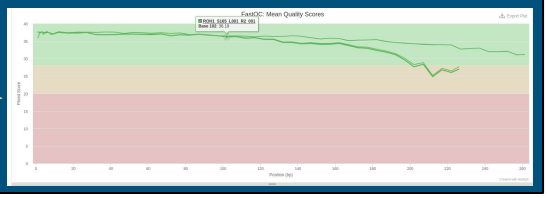
THEORY



- Objective: **Identification** of **taxa** in a **community**
- Workflow: Sample > DNA extraction > Sequencing (Illumina, PCR) > Bioinformatics (.fastq reads)
- Advantages: Identify unculturable species otherwise impossible to detect. Identify through DNA sequence (cryptic taxa). Generates a lot of information > "Complete" community picture.
- Limitations: Microbial Dark Matter. Unknown taxa (classical taxonomist are still needed!). Incomplete reference databases. Barcode BIAS. Lots of information (we have to filter out the noise).

WORKSHOP!





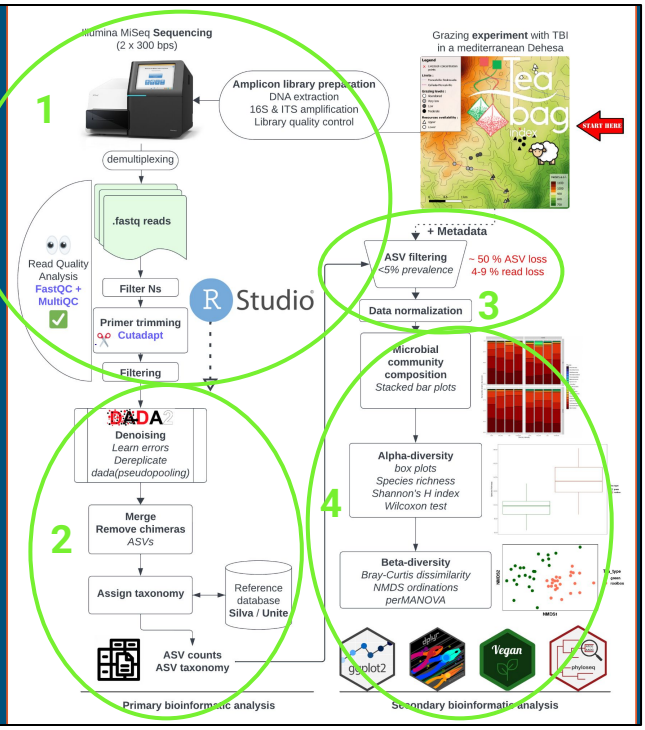
Course **plan**

Session 1
Theory: Understanding the **metabarcoding** technology
Practice: Initial processing of **.fastq** reads

Session 2
Theory: **Denoising**, **ASVs** vs **OTUs**, reference **databases**
Practice: **DADA2**, **assign taxonomy**

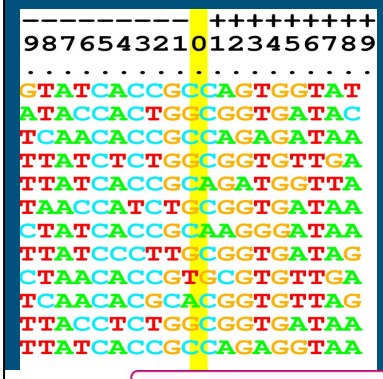
Session 3
Theory: Waste not, want not. **Normalization**. **Filtering**.
Practice: ASV **filtering** & **normalization**

Session 4
Theory: **Downstream** analysis, microbial ecology
Practice: R downstream options and ramifications



BioInfo4EEG

In this Session 2 . . .



	Group	Taxonomy	X2014_winter_FL	X2014_winter_PA	X2015_
1	ASV33112	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...	1793	152	
2	ASV122970	Bacteria(100);Cyanobacteria(100);Cyanobacteria(100);...	492	112	
3	ASV148428	Bacteria(100);Proteobacteria(100);Gammaproteobacte...	0	252	
4	ASV212114	Bacteria(100);Cyanobacteria(100);Cyanobacteria(100);...	574	184	
5	ASV9620	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...	464	70	
6	ASV147186	Bacteria(100);Proteobacteria(100);Betaproteobacteria(...	0	40	
7	ASV89359	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...	142	8	
8	ASV1061	Bacteria(100);Proteobacteria(100);Gammaproteobacte...	0	4	
9	ASV328581	Bacteria(100);Bacteroidetes(100);Bacteroidia(100);Bac...	2	540	
10	ASV86104	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...	72	6	
11	ASV57649	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...	69	24	
12	ASV172568	Bacteria(100);Proteobacteria(100);Gammaproteobacte...	130	7	
13	ASV237646	Bacteria(100);Bacteroidetes(100);Flavobacteriia(100);F...	0	0	
14	ASV67428	Bacteria(100);Planctomycetes(100);OM190(100);OM1...	0	16	

Identifier	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence	TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign	+
Quality scores	hhhhhhhhhhghghhhhhfhhhhhfffffe'ee[`X]b[d[ed'[Y[~Y
Identifier	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence	GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign	+
Quality scores	hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

Question: How do we tell species/taxa apart from each other? What makes each row in the community matrix exist?

Objective: from a series of DNA sequences, obtain **"real"** biological sequences, with an associated **taxonomic classification** and per sample **abundance** (read counts).

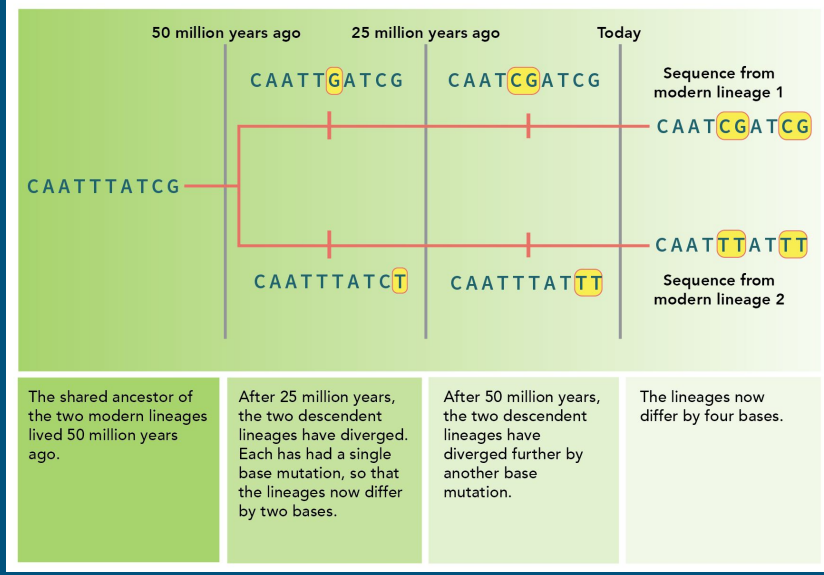
We will learn the core concepts of **clustering** and **denoising**, as well as **taxonomic assignment** using **reference databases**.

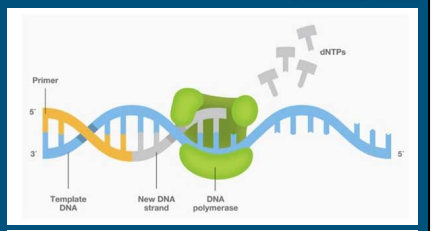
From last session, we understand how DNA sequences are obtained and what they represent (molecular barcodes). We also saw that many PCR cycles were needed, where the polymerase produces errors, mutations.

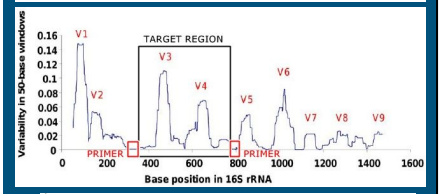
Question: what makes a species different from another? How can we tell them apart?

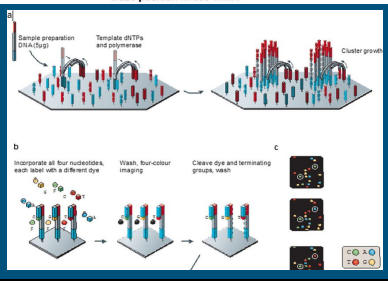
BioInfo4EEG

Evolutionary genetic divergence









The **genomes** of living organisms display **differences** in their **DNA sequences** (as well as in their proteins). Some of these differences, under pressure of natural selection, will be passed on to the next generation, and so on, producing diverse **lineages**, and groups of similar individuals (or organisms) that are classified into **taxa** by taxonomists. Throughout evolution, DNA sequences **diverge**. This is the basis of **phylogeny**.

When comparing DNA sequences between organisms, we can determine their **percentage of identity**, or, inversely, their **variability**. This variability in the tree of life is what we **take advantage of in metabarcoding technology** to classify DNA sequences into one taxon or another (as well as in phylogeny and many other fields)

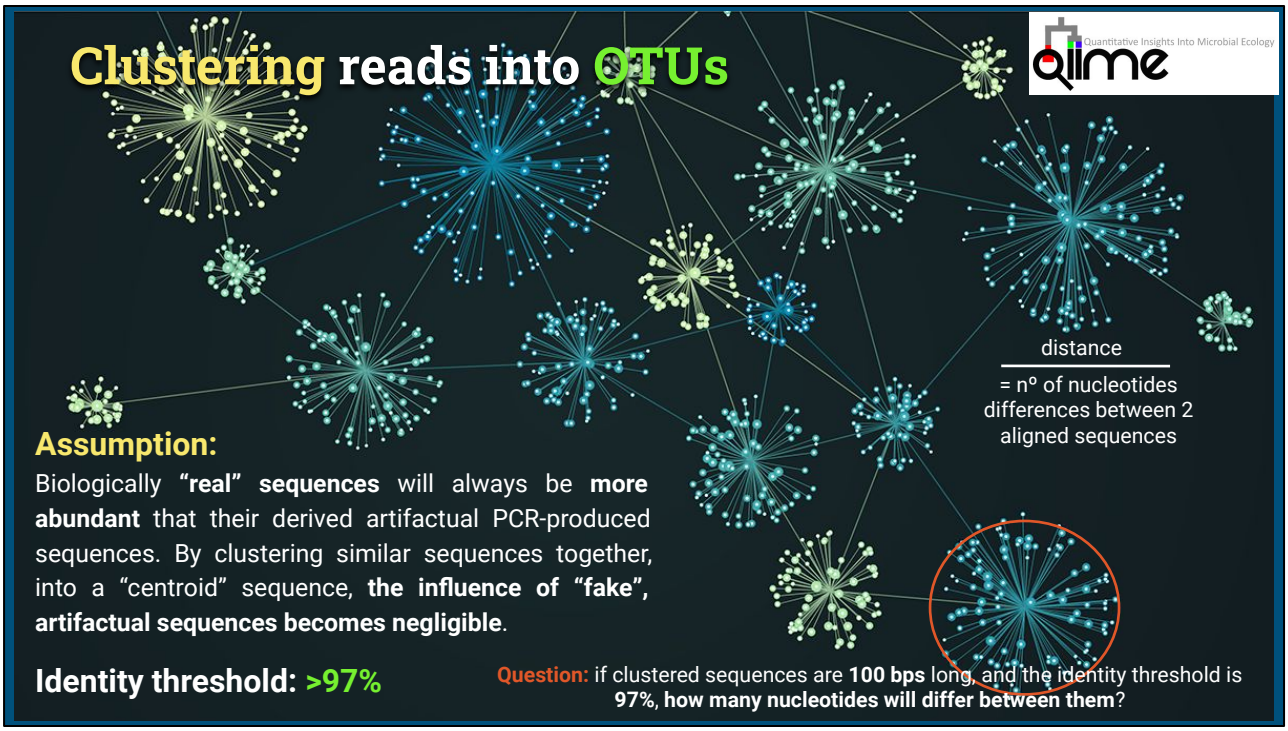
Thanks life for the **errors** of the **polymerases enzymes** (by far **not the only cause** of genetic variability!). Thanks to error induced variability we have biodiversity!

However, **this blessing is also a curse**, since **we utilize the same imperfect molecular machinery to obtain our DNA-seq data**! The same biouniversal phenomenon that causes genetic drift, molecular phylogeny, species, biodiversity, evolution, etc is **also adding NOISE to our data**. Remember there are extensive PCR cycles reactions required by the Illumina sequencing technology. So..

¿How do you tell a real biological variation apart from an artificial PCR product?

PROBLEM

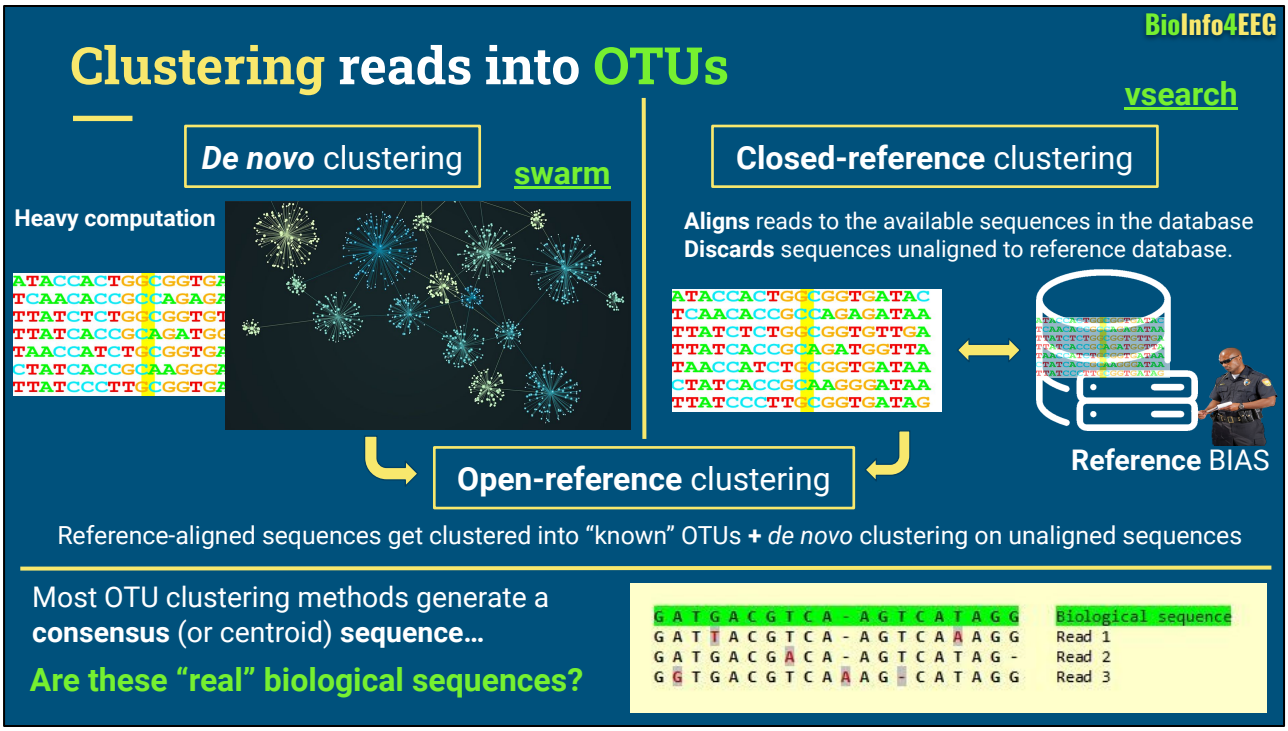
How do you tell a **"real biological variant"** apart from an **artifactual PCR product**?



Clustering similar sequences into **OTUs** (Operational Taxonomic Units) a.k.a **mOTUs** (Molecular OTUs): Incorporating the errors…

Historically implemented in QIIME1, this was the **first method** widely used to assess the problem of artifactual, PCR-produced, sequences. OTUs are clustered based on the **idea** that **the "real" biological sequence will be way more abundant than its "fake" artifactual sequence**. Assuming this, we can **cluster** together similar sequences, **under a threshold of sequence identity (usually 97% identical sequences)**, into an OTU, because **the effect of "fake" sequences will be negligible**. Typically, by clustering the sequences together, a **OTU consensus sequence** is built.

Historically, OTUs were built by clustering sequences at 95% similarity. However, with time and re-evaluation, t**he consensus threshold for clustering OTUs is nowadays >97%, even 99%** (for V4 16S!).

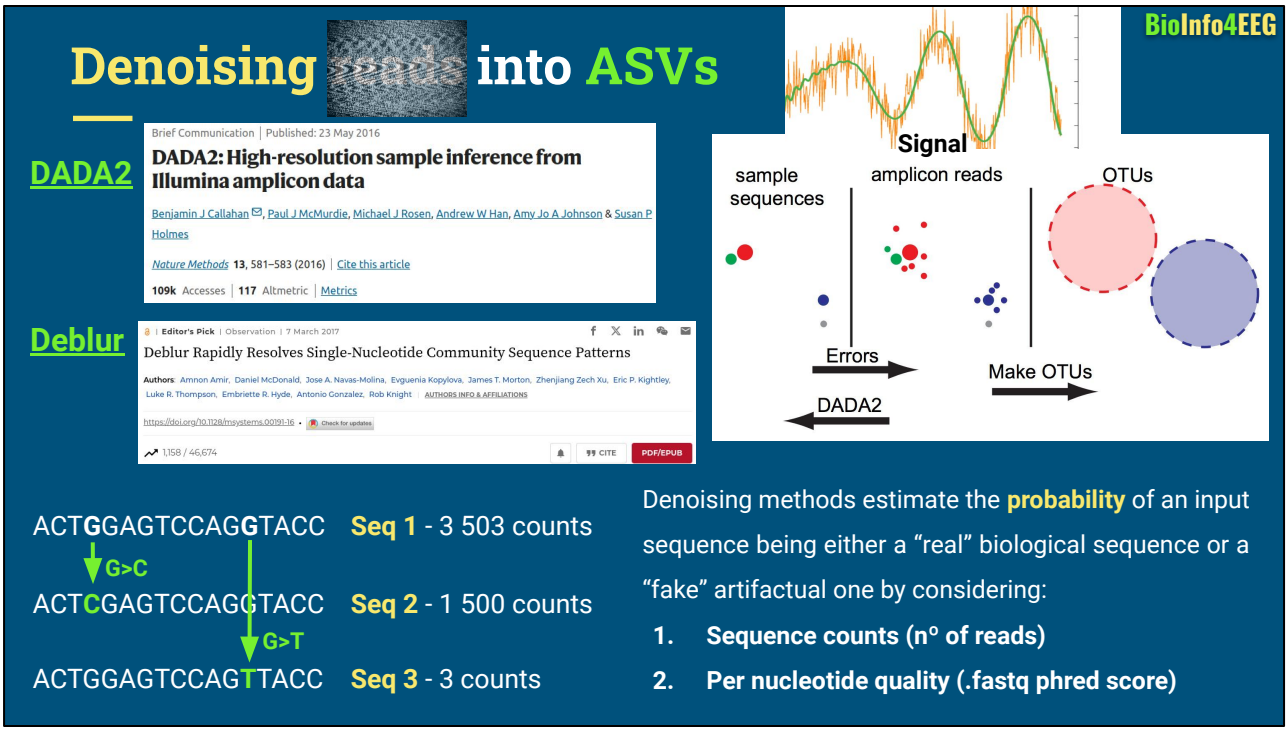


De novo clustering of sequences is the most computationally demanding method for generating OTUs, so other alternatives appeared:

Closed-reference clustering aligns the sequences to reference sequences found in databases, and clusters the sequences to known, pre-existing OTUs. This method discards sequences that don't align to the reference database (bellow the OTU % of similarity threshold ~97%) and has the danger of falling into **reference bias**. This is even more probable for less sampled environmental DNA, as compared to human gut microbiome, for example.

Open-reference clustering combines both techniques. It aligns sequences to the reference database and clusters them into OTUs based on the reference sequences. Sequences that poorly align to the reference database are not discarded but instead clustered *de novo*.

Many OTU clustering methods generate what is know as a **consensus sequence**, a sequence that can be thought of as the “average” sequence made up by all the reads that are incorporated into the cluster. **Can we consider this consensus sequence a “real” biological sequence?**



Around 2016, new tools (like ***DADA2*** or ***Deblur***) emerged trying to tackle the problem of **artifactual reads** in sequencing data. These tools are known as **denoisers** because they aim at eliminating "fake" reads, removing that "noise" from the dataset and thus **obtaining a signal that more accurately represents the "true" biological information originally found in the sample**.

The **main idea** behind DADA2 algorithm is the following: technology-produced "fake" reads happen independently between and within "real" biological reads because of **random polymerase errors** (the polymerase does not choose to make a mistake at a certain nucleotide base). However, **repeated observations of a read happening** (read counts, or abundance) suggest that the

read in question is probably a “real” sequence. Equally, it is highly improbable that the nucleotide transition happened repeatedly, multiple times, from the original “real” sequence.

= = = = = = Example from the slide = = = = = =

- ***Seq 1*** is the most abundant read in our dataset, having been read 3 503 times (counts). **We consider it a “real” biological sequence**.
- One could think that ***Seq 2***, differing only from Seq 1 by a G>C transition in position 3, is a product of a polymerase error in position 3. However, **how probable is that the G>C transition in position 3 has happened 1 500 times (Seq 2 counts) because of PCR? It is more likely that Seq 2 is a “real” biological variant than a “fake” PCR product**.
- What about ***Seq 3***? It also differs in just one nucleotide with respect to Seq 1 (G>T transition in position 13). But, **because of its low read count (read only 3 times), it is more probable that Seq 3 is indeed a PCR error**.

= =

Furthermore, DADA2 also incorporates the **information** from the **.fastq quality score** in each of its transitions to its models and algorithms in order to **resolve input sequence data to exact sequence variants**. Erroneous, artifactual sequences get “corrected” and are considered a count of the sequence they

originated from.

For more info:

- DADA2 nature's methods paper
- A short podcast with DADA2 developper, Benjamin Callahan, can be found here. He explains the thought process behind DADA2 error modelling and denoising algorithm.

QUESTION: What would you choose?

Clustering into **OTUs**
vs
Denoising into **ASVs**

BioInfo4EEG

OTUs *vs* ASVs

OTUs

Benefits:

- Reference-based clustering in well-known environments (i.e. human gut microbiota)
- Fast computation (with reference)
- Approximation to a **concept** of **species**, a solid **basal taxonomic unit**

Disadvantages:

- Consensus sequences: **are they "real"?**
- Not comparable between studies (with de novo clustering), **not reproducible**
- Incorporation of polymerase **errors**

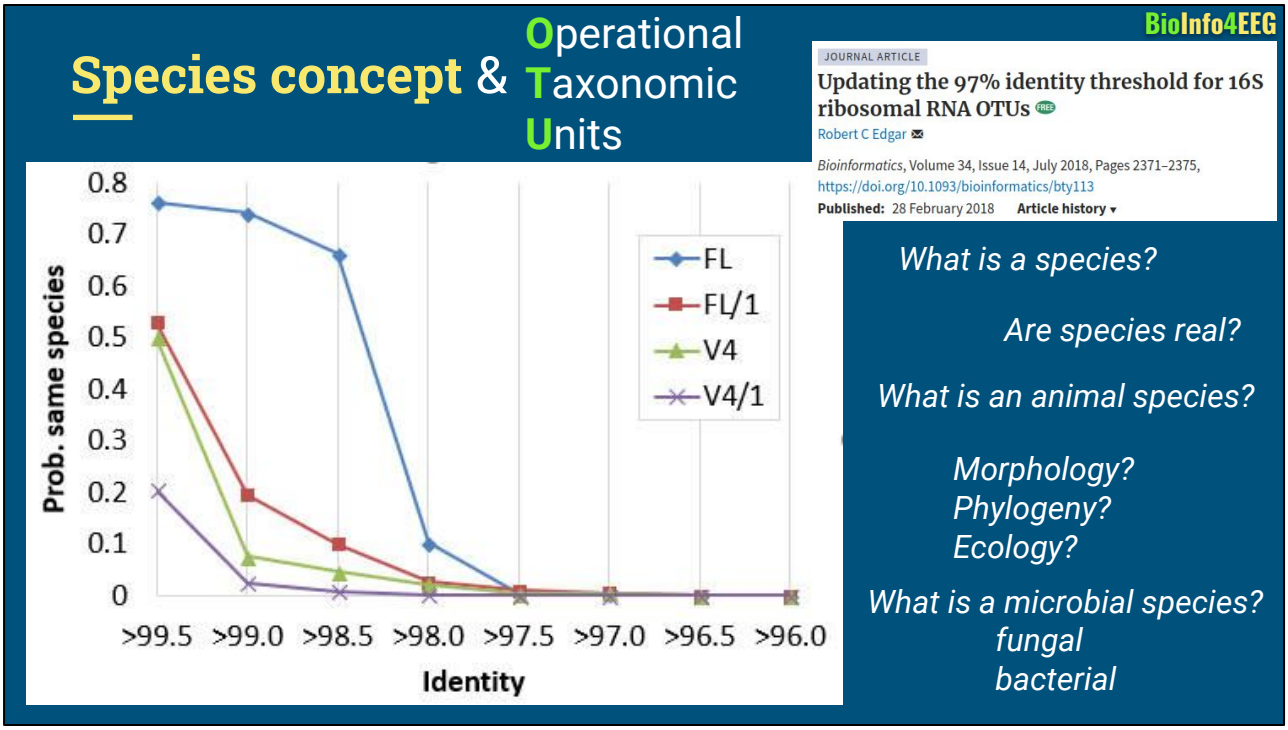
ASVs

Benefits:

- Elimination of noise, errors
- Are **real biological sequences**
- **Comparability** between studies and datasets
- **Higher taxonomic resolution** (different ASVs from even 1 nucleotide transition)

Disadvantages:

- More computationally demanding
- Higher taxonomic resolution: **is it too much resolution**? Does it set a basal taxonomic unit for studying microbial communities?



Why are OTUs clustered at a **certain** identity threshold?

The threshold of sequence similarity at which sequences are clustered together into an OTU tries to imitate the **species concept**. However, **the definition of what is a specie is not set in stone**: it greatly **varies** across phylogenetic, ecological and molecular **interpretations**, as well as across all the **different life kingdoms and domains**. For example, it is estimated that if the criterion of species delineation for bacteria (~97% similarity for the entire 16S DNA region) was applied to animals, the whole order of primates would be considered a single species (James T. Staley 1997).

Furthermore, as seen in the graph, **percentages of identity** between and within species will vary **depending on the barcode used**. The graph demonstrates that a **97% identity threshold** will "correctly" **cluster prokaryotic species together** (and apart from each other), but only for the **full length 16S region (~ 1 600 bps)!** Instead, for the **shorter V4 region** of the 16S barcode (**~ 254 bps**), a **threshold closer to ~ 98-99 %** seems more adequate for clustering species into OTUs. That is because shorter genomic regions acting as barcodes have less opportunity of being different between species.

This only scratches the surface of a "pandora box" that is the species concept in **biology** and its derived fields & across the tree of life:
Are species real? What is the concept of a species? What is an animal species? What is a microbial species?

Personally, I like to think of living organisms as a whole as a continuum of variation in their DNA that makes all living organisms slightly different. However, species concepts are useful in the study of evolution and ecosystems. The time scale at which species appear and disappear is difficult to grasp. Molecular evolution is always happening, but, relative to our reality, it manifests itself really slowly.

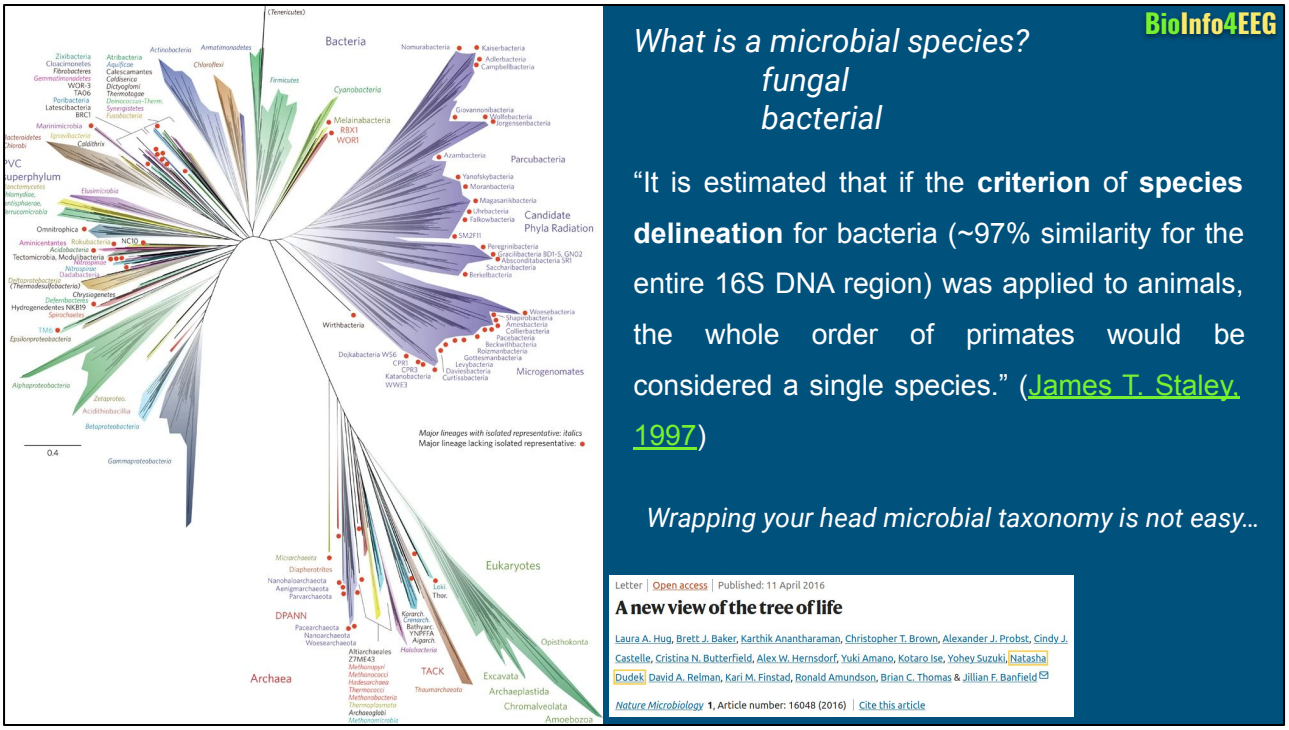
Fall into the species concept rabbit hole:

- Species Wikipedia page
- Species Concepts and Species Delimitation
- What are Bacterial Species?
- Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology
- Biodiversity: are microbial species threatened?: Commentary

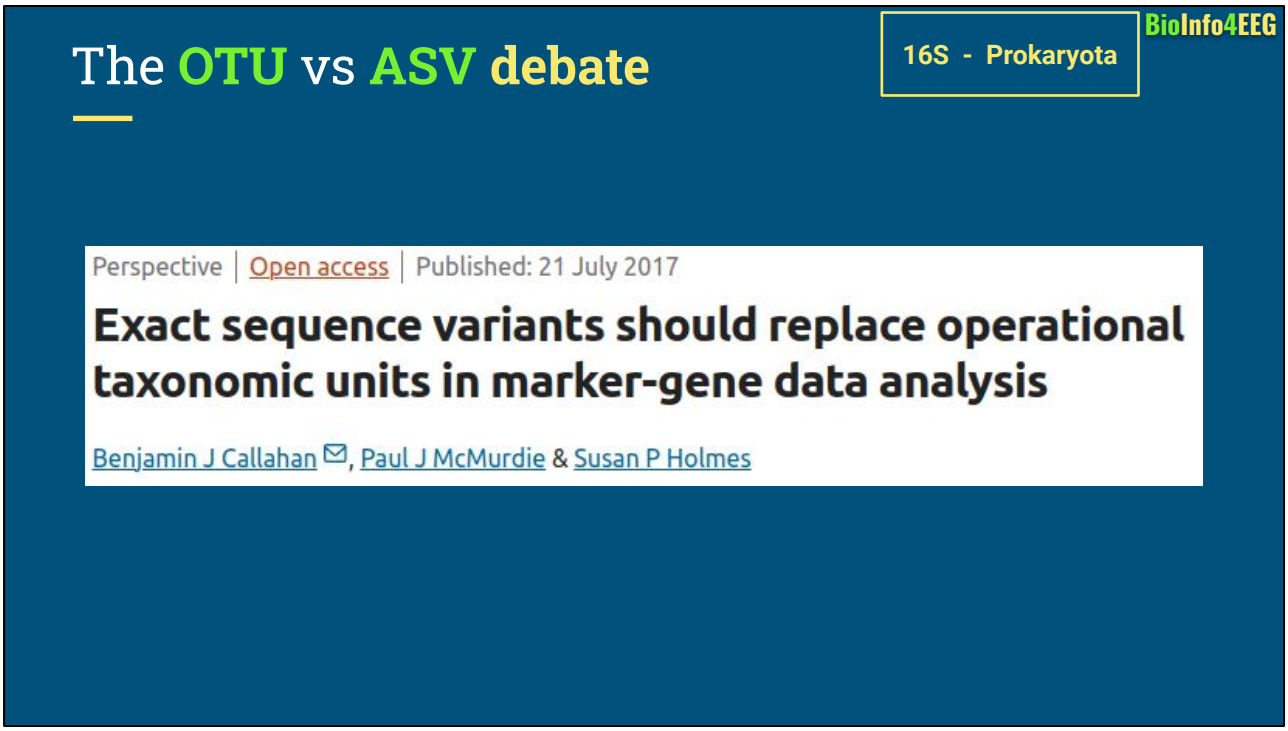
In the end, a big part of the analysis will rely on clustering taxa together under the same order, family or genus. Nonetheless, it is important to **set a basal unit in taxonomic analysis**. For example, when a botanist goes to the field and starts counting species in his plot, he has (most of the time) a clear idea of the basal taxonomic unit he is using: clearly defined and delineated species. Those species most probably **do show molecular differences within the same species**, discrepancies in DNA between barcodes of individuals (haplotypes). However, a taxonomic unit is needed for the study of ecology.

As a reminder, there is a consensus that unless sequencing longer regions of DNA (whole 16S, entire ITS and SSU, etc), a confident taxonomic assignment to the species level based on small

hypervariable regions (i.e. V3-V4 region, either ITS1 or ITS2) is impossible to achieve.



The threshold of sequence similarity at which sequences are clustered together into an OTU tries to imitate the **species concept**. However, **the definition of what is a specie is not set in stone**: it greatly **varies** across phylogenetic, ecological and molecular **interpretations**, as well as across all the **different life kingdoms and domains**. For example, it is estimated that if the criterion of species delineation for bacteria (~97% similarity for the entire 16S DNA region) was applied to animals, the whole order of primates would be considered a single species (James T. Staley 1997).



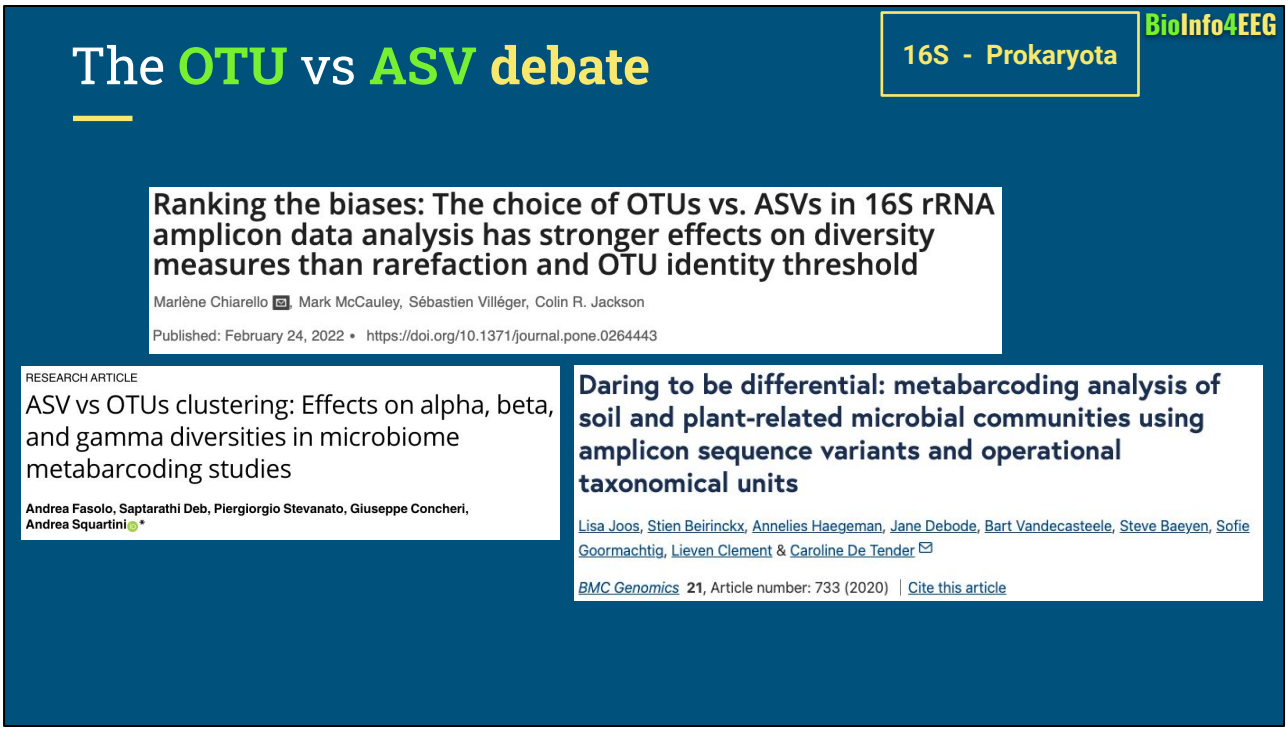
Closely after the development of DADA2 and other similar denoisers, an opinion formed that **ASVs**, or **ESV**s (Exact Sequence Variants) **should replace OTUs**, because, compared to OTUs, ASVs offer:

1) **Higher taxonomic resolution**, resolving differences in even 1 nucleotide between sequences, enabling the differentiation at the species and even strain level (with longer barcodes!)
2) **Reproducibility**.
3) **"Real" biological meaning**: they effectively are sequences present in our study sample.

From the paper on the slide: "Appealing terminology such as 'resolution of exact sequence variants' **does not eliminate the**

limitations inherent to representing a complex biological organism by a short genetic barcode.”

Is it worth debating OTUs vs ASVs? Does it really matter?



Is it worth debating OTUs vs ASVs? Does it really matter?

Yes, it does.

From this paper:

In 16S metabarcoding analysis, the choice between clustering into OTUs or denoising into exact sequences has an effect on biodiversity metrics, a key result in almost all environmental studies. This effect of choosing one method or another is more pronounced in highly diverse, environmental samples (in this case river sediment) than in more specialized, niche environments (a mussel's gut microbiome).

The OTU vs ASV debate

BioInfo4EEG

ITS - Fungi
SSU - AM Fungi

Best practices in metabarcoding of fungi: From experimental design to results

Leho Tedersoo[1,2] | Mohammad Bahram[1,3] | Lucie Zinger[4,5] | R. Henrik Nilsson[6] | Peter G. Kennedy[7] | Teng Yang[8] | Sten Anslan[9] | Vladimir Mikryukov[1,9]

exceeds a user-settable parameter (BAND_SIZE). The default value of this parameter was chosen to minimally impact the alignment of sequences with few indels, such as ribosomal RNA genes. Both heuristics can be disabled by the user, and the default values should be re-examined if the algorithm is applied to genetic regions with significantly different characteristics, such as the indel-rich ITS region.

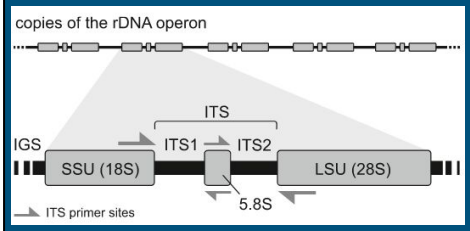
The DADA2 paper acknowledges that **parameters** in its algorithm can and should be **modified** for **non-16S barcode** data.

Many experts **discourage the use of exact sequence variants** (ESVs) as the baseline taxonomic unit (~species) in fungal metabarcoding studies. **Why?**

ITS alchemy: On the use of ITS as a DNA marker in fungal ecology

Håvard Kauserud

Sections for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, Norway



Internally Transcribed Spacer (ITS):

- Variable length of the rDNA operon ⇒ conflict with DADA2
- Intraspecific variability
- Intragenomic variability
- Intra-individual variability (heterokaryotic fungi)

For metabarcoding analysis of ITS regions in the identification of **fungal** communities, there are many experts **opposing** the use of **denoisers'** (like DADA2, Deblur or UNOISE) **output - exact sequence variants - as equivalent to fungal species**. This reluctance towards a "blind" use of denoisers is understandable since those tools were created and developed with the more popular 16S barcode in mind (a lot of research has been done on the human gut microbiome and its relation with health). Fungi experts are against the idea that clustering should be avoided.

ITS means *Internally Transcribed Spacer*, and defines the genomic regions between ribosomal subunits (SSU, LSU) that gets transcribed to RNA. This ITS region is within the

The ITS regions in fungi are characterized by:

- Variable length of the ITS region
- Repeated copies

From Tederso best practices:

By reanalysing a data set from Furneaux et al. (2021), we show that the DADA2 ITS pipeline and UNOISE ESV approaches reduce phylogenetic richness by disproportionately eliminating rare members of the unicellular fungal groups, *Glomeromycota* and nonfungal eukaryotes (Figure Box 2). In terms of community composition, the results are similar between ESV and OTU-based approaches (Glassman & Martiny, 2018; Porter & Hajibabaei, 2020), because these are driven by abundant taxa. **We conclude that ESV approaches overestimate richness of common fungal species (due to haplotype variation) but underestimate richness of rare species (by removing rare variants; see also Joos et al. 2020). ESV approaches can nevertheless be useful for studying allele or haplotype distribution of various common species based on eDNA (Zizka et al., 2020).**

References:

- Best practices in the metabarcoding of fungi (Tedersoo)
- ITS alchemy (Kauserud)
- Qiime2 forum thread: **ASV vs OTU for fungal ITS**

The OTU vs ASV debate

ENVIRONMENTAL MICROBIOLOGY

Research article | Open Access

First published: 30 July 2019 | https://doi.org/10.1111/1462-2920.14764 | Citations: 32

The copyright line for this article was changed on 13 October 2020 after original online publication.

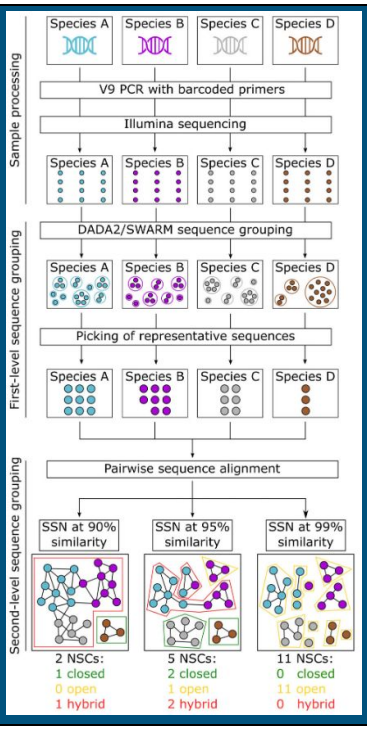
SECTIONS PDF TOOLS SHARE

Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants

Dominik Forster, Guillaume Lentendu, Sabine Filker, Elyssa Dubois, Thomas A. Wilding, Thorsten Stoeck

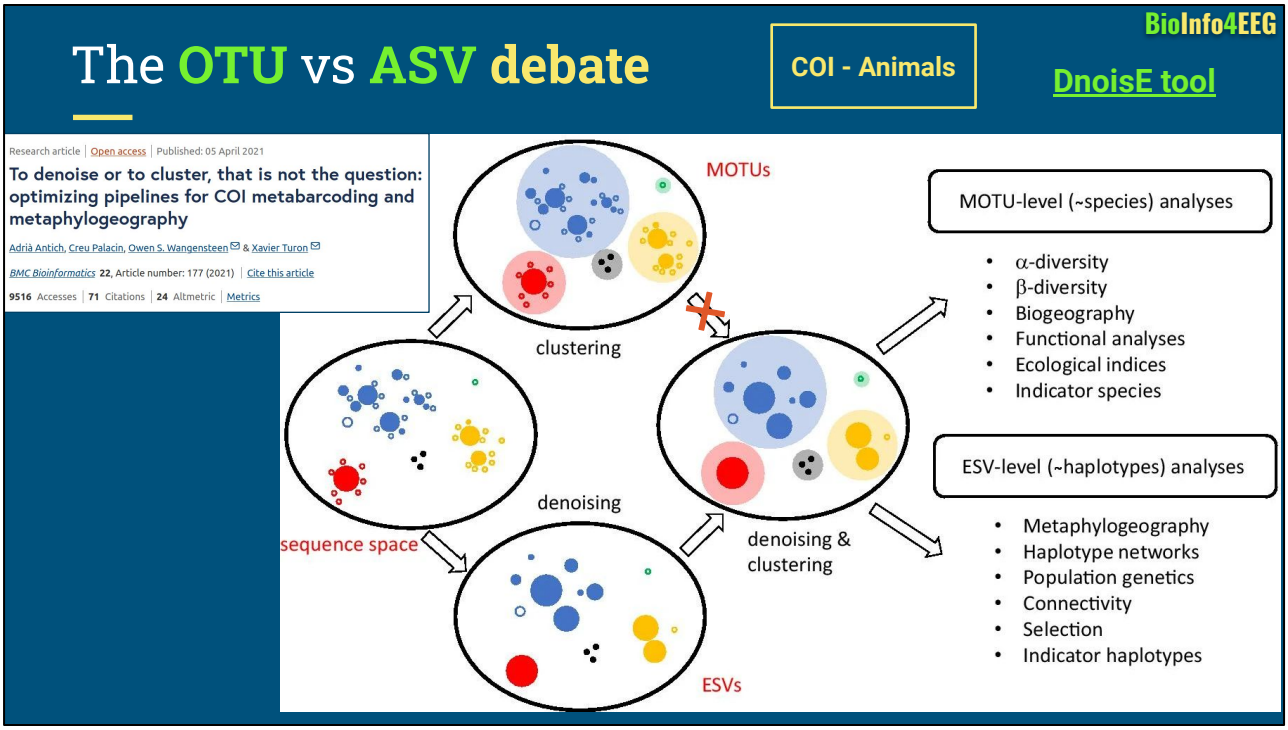
Summary

Effective and precise grouping of highly similar sequences remains a major bottleneck in the evaluation of high-throughput sequencing datasets. Amplicon sequence variants (ASVs) offer a promising alternative that may supersede the widely used operational taxonomic units (OTUs) in environmental sequencing studies. We compared the performance of a recently developed pipeline based on the algorithm DADA2 for obtaining ASVs against a pipeline based on the algorithm SWARM for obtaining OTUs. Illumina-sequencing of 29 individual ciliate species resulted in up to 11 ASVs per species, while SWARM produced up to 19 OTUs per species. To improve the congruency between species diversity and molecular diversity, we applied sequence similarity networks (SSNs) for second-level sequence grouping into network sequence clusters (NSCs). At 100% sequence similarity in SWARM-SSNs, NSC numbers decreased from 7.9-fold overestimation without abundance filter, to 4.5-fold overestimation when an abundance filter was applied. For the DADA2-SSN approach, NSC numbers decreased from 3.5-fold to 3-fold overestimation. Rand index cluster analyses predicted best binning results between 97% and 94% sequence similarity for both DADA2-SSNs and SWARM-SSNs. Depending on the ecological questions addressed in an environmental sequencing study with protists we recommend ASVs as replacement for OTUs, best in combination with SSNs.



BioInfo4EEG

Protists 18S, …



I love the title of Antich et al. paper: **To denoise or to cluster, that is not the question:...**

This view goes against choosing either clustering or denoising, but rather

Although metabarcoding of cytochrome C oxidase gene (COI) for identification of animals strives away from the ecosystem ecology group's focus, the approach described in Adria Antich's paper can be in some ways carried over to metabarcoding analysis of other barcodes, especially ITS.

Databases

BOLD SYSTEMS

IDENTIFICATION V4 TAXONOMY WORKBENCH LOGIN

24,868,408	19,565,646	356,304
Specimen Records	Specimens with Barcodes	Species with Barcodes

Animals:

- Acanthocephala [3703]
- Annelida [151628]
- Arthropoda [21366479]
- Brachiopoda [551]
- Bryozoa [7900]
- Chaetognatha [2161]
- Chordata [1054916]
- Cnidaria [49981]
- Ctenophora [1309]
- Cycliophora [546]
- Echinodermata [70367]
- Entoprocta [121]
- Gastrotricha [1796]
- Gnathostomulida [50]
- Hemichordata [437]
- Kinorhyncha [836]
- Mollusca [325443]
- Nematoda [108185]
- Nematomorpha [541]
- Nemertea [9436]
- Onychophora [2123]
- Phoronida [253]
- Placozoa [37]
- Platyhelminthes [56598]
- Porifera [15032]
- Priapulida [347]
- Rhombozoa [48]
- Rotifera [18172]
- Tardigrada [6266]
- Xenacoelomorpha [990]

Plants:

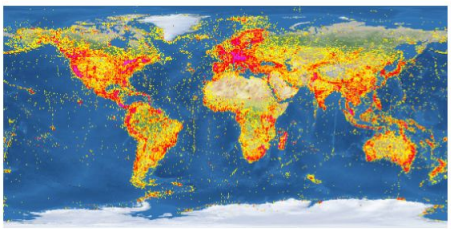
- Bryophyta [24474]
- Chlorophyta [23242]
- Lycopodiophyta [1]
- Magnoliophyta [10687]
- Pinophyta [14]
- Pteridophyta [1307]

Fungi:

- Ascomycota [239829]
- Basidiomycota [140585]
- Chytridiomycota [42958]
- Glomeromycota [3572]
- Myxomycota [235]
- Zygomycota [1225]

Protists:

- Chlorarachniophyta [67]
- Ciliophora [1271]
- Heterokontophyta [9808]
- Pyrrophycophyta [2339]
- Rhodophyta [64845]



Databases

16S - Prokarya

BioInfo4EEG

Cabezas *et al. Environmental Microbiome* (2024) 19:88
https://doi.org/10.1186/s40793-024-00634-w

Environmental Microbiome

RESEARCH Open Access

MIMt: a curated 16S rRNA reference database with less redundancy and higher accuracy at species-level identification

M. Pilar Cabezas[1,2], Nuno A. Fonseca[3,4] and Antonio Muñoz-Mérida[3,4*]

Databases are not perfect. Some taxa are unresolved (unclassified) at certain taxonomic levels and there are repeated entries.

Databases

ITS - Fungi

unite community

BioInfo4EEG

General FASTA release (**download**)

This release consists of a single FASTA file: the RepS/RefS of all SHs, adopting the dynamically use of clustering thresholds whenever available. The format of the FASTA header is:

```
>Glomeraceae|AM076560|SH146432.05FU|refs|k__Fungi;p__Glomeromycota;c__Glomeromycetes;o__Glomerales;f__Glomeraceae;g__;s__uncultured_Glomus
```

This is the file we recommend for local BLAST searches against the SHs.

List of all general FASTA releases

Version no	Release date	Taxon group	No of RefS	No of RepS	Release status	Link	Notes
10.0	2024-04-04	Fungi	18 895	74 190	Current	**https://doi.org/10.15156/BIO/2959332**	**When using this resource, please cite it as follows:** Abarenkov, Kessy; Zirk, Allan; Piirmann, Timo; Pöhönen, Raivo; Ivanov, Filipp; Nilsson, R. Henrik; Kõljalg, Urmas (2024): UNITE general FASTA release for Fungi. Version 04.04.2024. UNITE Community. **https://doi.org/10.15156/BIO/2959332** Includes singletons set as RefS (in dynamic files).
10.0	2024-04-04	Fungi	18 895	140 300	Current	**https://doi.org/10.15156/BIO/2959333**	**When using this resource, please cite it as follows:** Abarenkov, Kessy; Zirk, Allan; Piirmann, Timo; Pöhönen, Raivo; Ivanov, Filipp; Nilsson, R. Henrik; Kõljalg, Urmas (2024): UNITE general FASTA release for Fungi 2. Version 04.04.2024. UNITE Community. **https://doi.org/10.15156/BIO/2959333** Includes global and 3% distance singletons.
10.0	2024-04-04	All eukaryotes	19 302	122 914	Current	**https://doi.org/10.15156/BIO/2959334**	**When using this resource, please cite it as follows:** Abarenkov, Kessy; Zirk, Allan; Piirmann, Timo; Pöhönen, Raivo; Ivanov, Filipp; Nilsson, R. Henrik; Kõljalg, Urmas (2024): UNITE general FASTA release for eukaryotes. Version 04.04.2024. UNITE Community. **https://doi.org/10.15156/BIO/2959334** Includes singletons set as RefS (in dynamic files).
10.0	2024-04-04	All eukaryotes	19 302	232 937	Current	**https://doi.org/10.15156/BIO/2959335**	**When using this resource, please cite it as follows:** Abarenkov, Kessy; Zirk, Allan; Piirmann, Timo; Pöhönen, Raivo; Ivanov, Filipp; Nilsson, R. Henrik; Kõljalg, Urmas (2024): UNITE general FASTA release for eukaryotes 2. Version 04.04.2024. UNITE Community. **https://doi.org/10.15156/BIO/2959335** Includes global and 3% distance singletons.

BioInfo4EEG

GlobalFungi Database

NOTE: best viewed using Firefox, Chrome or Edge on PC

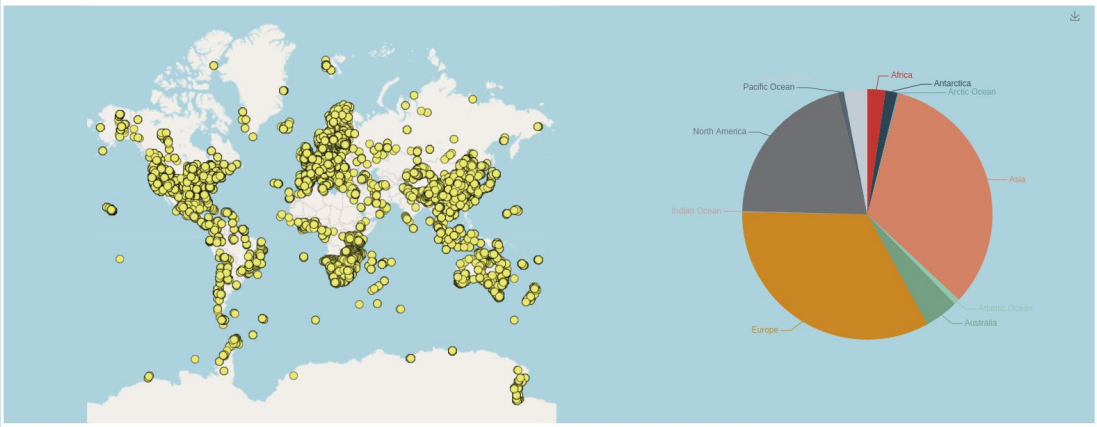
- Home
- Taxon search
- Sequence search
- Geosearch
- Studies
- Results
- How to cite
- About GlobalFungi
- Join mailing list
- Help
- Submit your study
- Leave a message
- Collaborators

Site design + programming
Tomas Vetrovsky
Daniel Morais
(c) 2020

Welcome to GlobalFungi!

GlobalFungi dataset release 5.0 (16.11.2023). Taxonomy based on UNITE version 10.0 (4.4.2024).
Actual number of samples in the database: 84972; actual number of studies included: 846.
Number of ITS sequence variants: 593 399 355; number of ITS1 sequences 1 233 820 630; number of ITS2 sequences 3 474 636 588.

GlobalFungi Twitter page | YouTube tutorials: | How to use GlobalFungi Database (tutorial) | How to Submit your Study (tutorial)

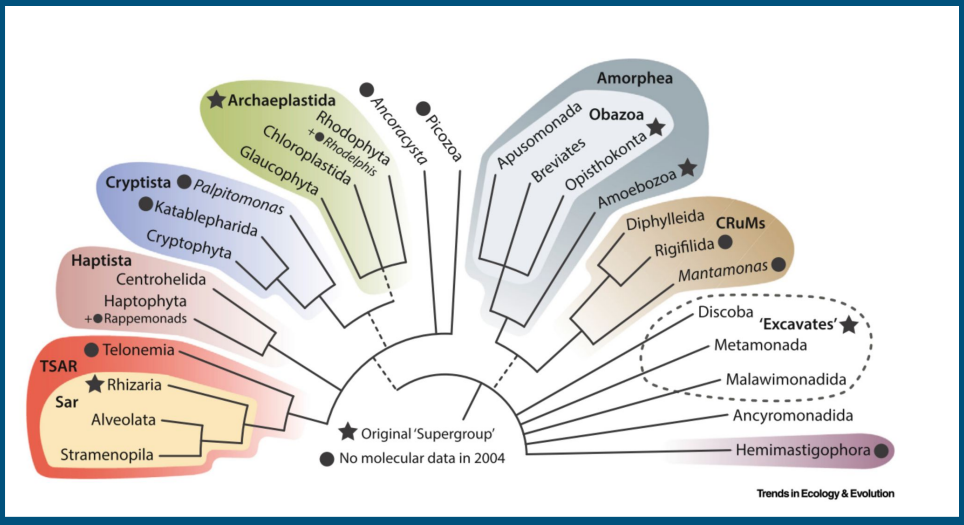


To access published data on arbuscular mycorrhizal fungal communities, click here!

BioInfo4EEG

Databases

18S - Protists



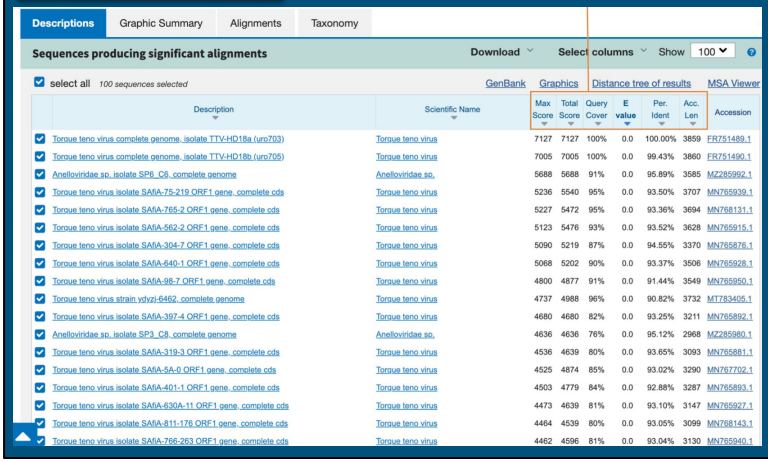
BioInfo4EEG

Aligners & Classifiers

Research Article | 15 August 2007

Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy

Authors: Qiong Wang, George M. Garrity, James M. Tiedje, James R. Cole | AUTHORS INFO & AFFILIATIONS



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Select columns — Show 100

select all 100 sequences selected — GenBank | Graphics | Distance tree of results | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Torque teno virus complete genome, isolate TTV-HD18a (uro703)	Torque teno virus	7127	7127	100%	0.0	100.00%	3859	FR751489.1
Torque teno virus complete genome, isolate TTV-HD18b (uro705)	Torque teno virus	7005	7005	100%	0.0	99.43%	3860	FR751490.1
Anelloviridae sp. isolate SP6_C6, complete genome	Anelloviridae sp.	5688	5688	91%	0.0	95.89%	3585	MZ285992.1
Torque teno virus isolate SAfA-75-219 ORF1 gene, complete cds	Torque teno virus	5236	5540	95%	0.0	93.50%	3707	MN765939.1
Torque teno virus isolate SAfA-765-2 ORF1 gene, complete cds	Torque teno virus	5227	5472	95%	0.0	93.36%	3694	MN768131.1
Torque teno virus isolate SAfA-562-2 ORF1 gene, complete cds	Torque teno virus	5123	5476	93%	0.0	93.52%	3628	MN765915.1
Torque teno virus isolate SAfA-304-7 ORF1 gene, complete cds	Torque teno virus	5090	5219	87%	0.0	94.55%	3370	MN765876.1
Torque teno virus isolate SAfA-640-1 ORF1 gene, complete cds	Torque teno virus	5068	5202	90%	0.0	93.37%	3506	MN765928.1
Torque teno virus isolate SAfA-98-7 ORF1 gene, complete cds	Torque teno virus	4800	4877	91%	0.0	91.44%	3549	MN765950.1
Torque teno virus strain vdyzj-6462, complete genome	Torque teno virus	4737	4988	96%	0.0	90.82%	3732	MT783405.1
Torque teno virus isolate SAfA-397-4 ORF1 gene, complete cds	Torque teno virus	4680	4680	82%	0.0	93.25%	3211	MN765892.1
Anelloviridae sp. isolate SP3_C8, complete genome	Anelloviridae sp.	4636	4636	76%	0.0	95.12%	2968	MZ285980.1
Torque teno virus isolate SAfA-319-3 ORF1 gene, complete cds	Torque teno virus	4536	4639	80%	0.0	93.65%	3093	MN765881.1
Torque teno virus isolate SAfA-5A-0 ORF1 gene, complete cds	Torque teno virus	4525	4874	85%	0.0	93.02%	3290	MN767702.1
Torque teno virus isolate SAfA-401-1 ORF1 gene, complete cds	Torque teno virus	4503	4779	84%	0.0	92.88%	3287	MN765893.1
Torque teno virus isolate SAfA-630A-11 ORF1 gene, complete cds	Torque teno virus	4473	4639	81%	0.0	93.10%	3147	MN765927.1
Torque teno virus isolate SAfA-811-176 ORF1 gene, complete cds	Torque teno virus	4464	4539	80%	0.0	93.05%	3099	MN768143.1
Torque teno virus isolate SAfA-766-263 ORF1 gene, complete cds	Torque teno virus	4462	4596	81%	0.0	93.04%	3130	MN765940.1

Naïve Classifiers have to be trained!