ECOSYSTEM ECOLOGY

Bioinformatics Course

Session 2

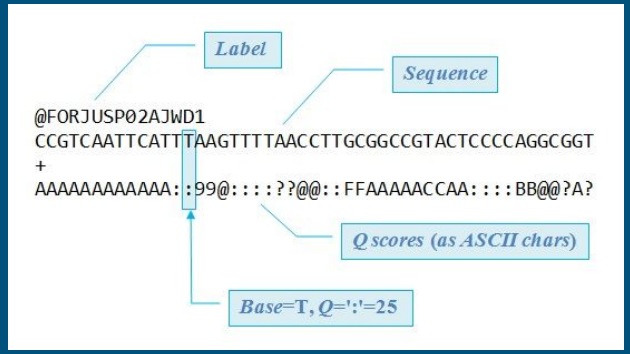
Theory: **Clustering** vs **Denoising**, reference **databases**

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

Recap from Session 1

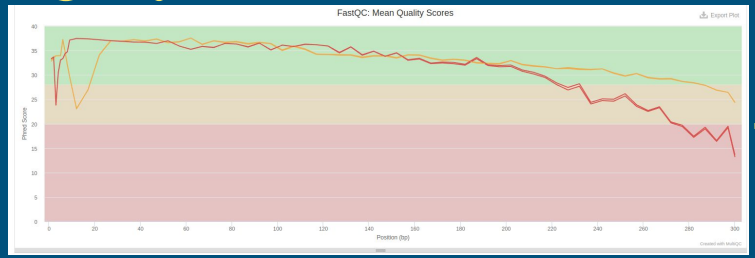
The **metabarcoding** technology

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!



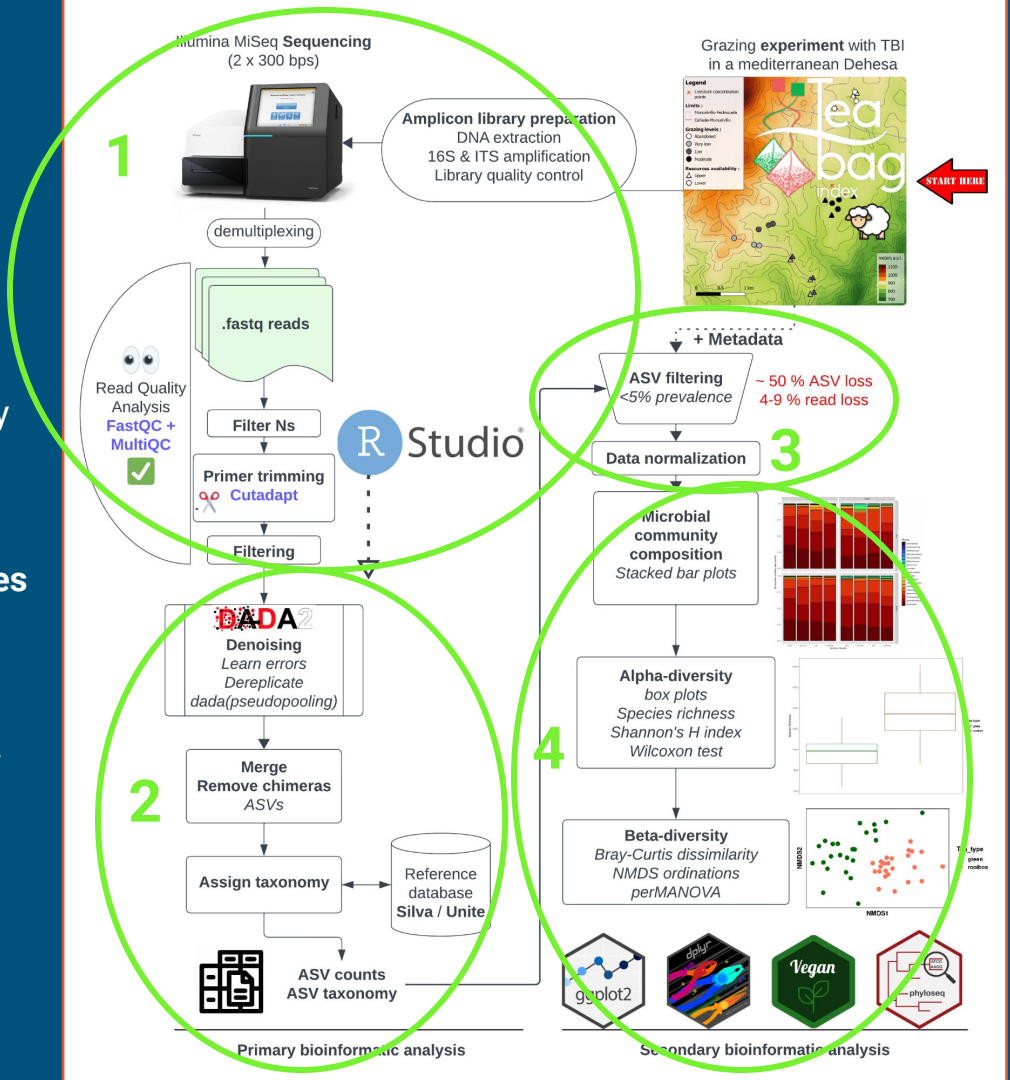
Course **plan**

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

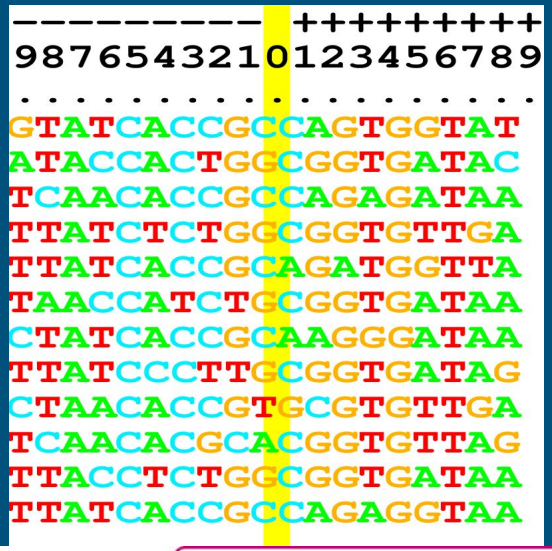
Session 2
Theory: **Denoising**, **ASVs** vs **OTUs**, reference **databases**
Practice: **DADA2**, **assign taxonomy**

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



In this Session 2 . . .

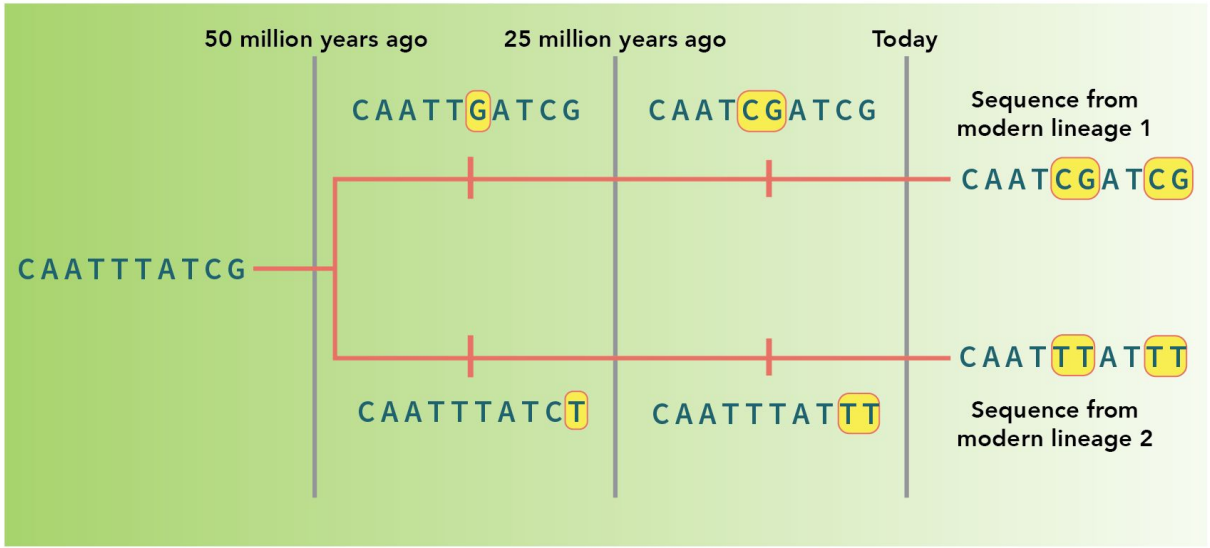


	Group	Taxonomy	X2014_winter_FL	X2014_winter_PA	X2015_
1	ASV33112	Bacteria(100);Proteobacteria(100);Alphaproteobacteri...	1793	152	
2	ASV122970	Bacteria(100);Cyanobacteria(100);Cyanobacteria(100);...	492	112	
3	ASV148428	Bacteria(100);Proteobacteria(100);Gammaproteobacte...	0	252	
4	ASV212114	Bacteria(100);Cyanobacteria(100);Cyanobacteria(100);...	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	hhhhhhhhhhghhghhhhhfhhhhhfffffe'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?

Evolutionary genetic divergence

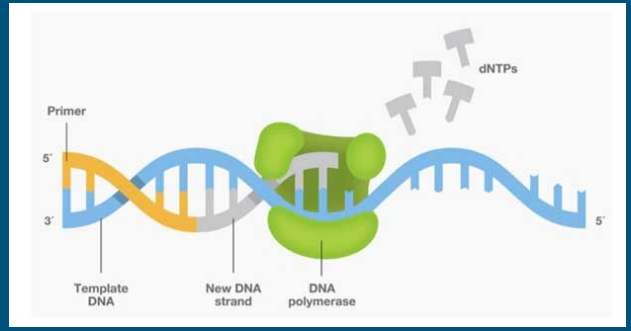


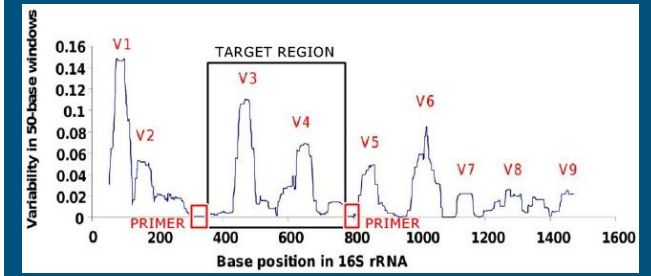
The shared ancestor of the two modern lineages lived 50 million years ago.

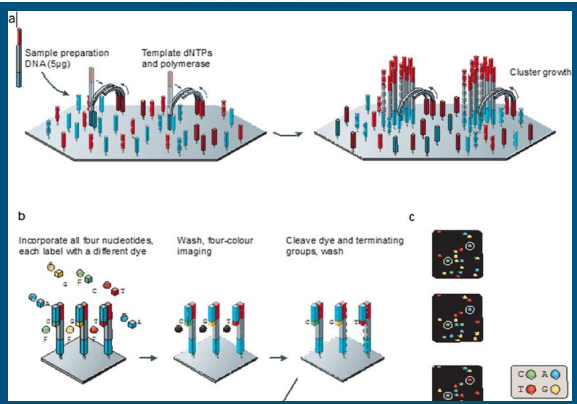
After 25 million years, the two descendent lineages have diverged. Each has had a single base mutation, so that the lineages now differ by two bases.

After 50 million years, the two descendent lineages have diverged further by another base mutation.

The lineages now differ by four bases.







PROBLEM

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

Clustering reads into OTUs

Quantitative Insights Into Microbial Ecology

qiime

distance

= nº of nucleotides differences between 2 aligned sequences

Assumption:

Biologically **“real” sequences** will always be **more abundant** that their derived artifactual PCR-produced sequences. By clustering similar sequences together, into a “centroid” sequence, **the influence of “fake”, artifactual sequences becomes negligible**.

Identity threshold: >97%

Question: if clustered sequences are **100 bps** long, and the identity threshold is **97%, how many nucleotides will differ between them**?

Clustering reads into OTUs

vsearch

De novo clustering

swarm

Heavy computation

Closed-reference clustering

Aligns reads to the available sequences in the database
Discards sequences unaligned to reference database.

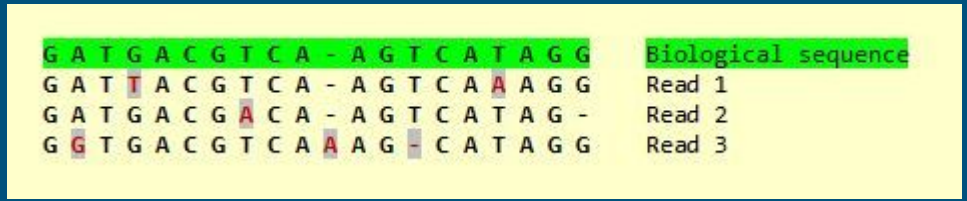
Reference BIAS

Open-reference clustering

Reference-aligned sequences get clustered into “known” OTUs **+** *de novo* clustering on unaligned sequences

Most OTU clustering methods generate a **consensus** (or centroid) **sequence…**

Are these “real” biological sequences?



Denoising into ASVs

BioInfo4EEG

DADA2

Brief Communication | Published: 23 May 2016

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes

Nature Methods **13**, 581–583 (2016) | Cite this article

109k Accesses | **117** Altmetric | Metrics

Deblur

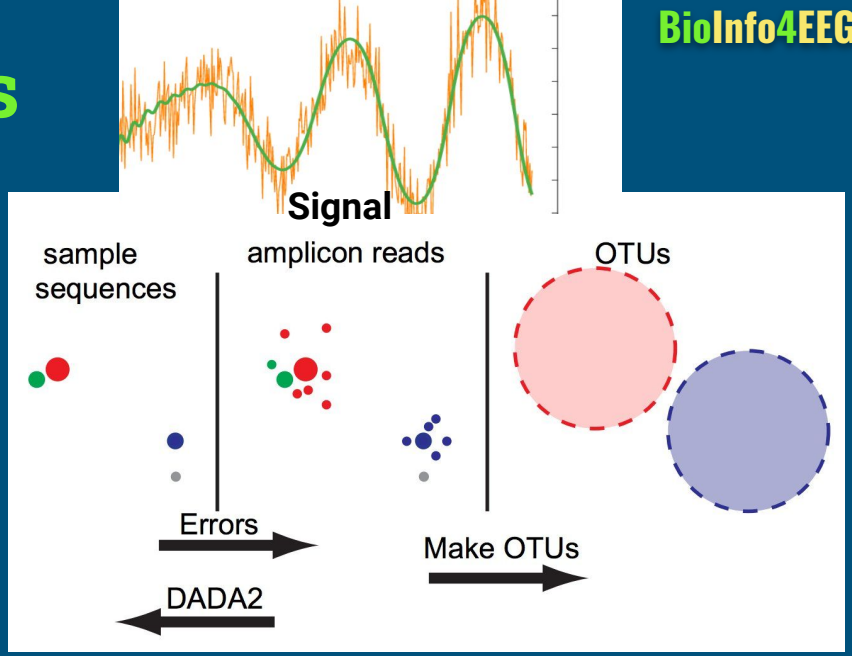
Editor's Pick | Observation | 7 March 2017

Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns

Authors: Amnon Amir, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, Luke R. Thompson, Embriette R. Hyde, Antonio Gonzalez, Rob Knight | AUTHORS INFO & AFFILIATIONS

https://doi.org/10.1128/msystems.00191-16 · Check for updates

1,158 / 46,674 | CITE | PDF/EPUB



ACT**G**GAGTCCAG**G**TACC **Seq 1** - 3 503 counts

G>C

ACT**C**GAGTCCAG**G**TACC **Seq 2** - 1 500 counts

G>T

ACTGGAGTCCAG**T**TACC **Seq 3** - 3 counts

Denoising methods estimate the **probability** of an input sequence being either a “real” biological sequence or a “fake” artifactual one by considering:

1. **Sequence counts (n° of reads)**
2. **Per nucleotide quality (.fastq phred score)**

QUESTION: What would you choose?

Clustering into **OTUs**

vs

Denoising into **ASVs**

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:

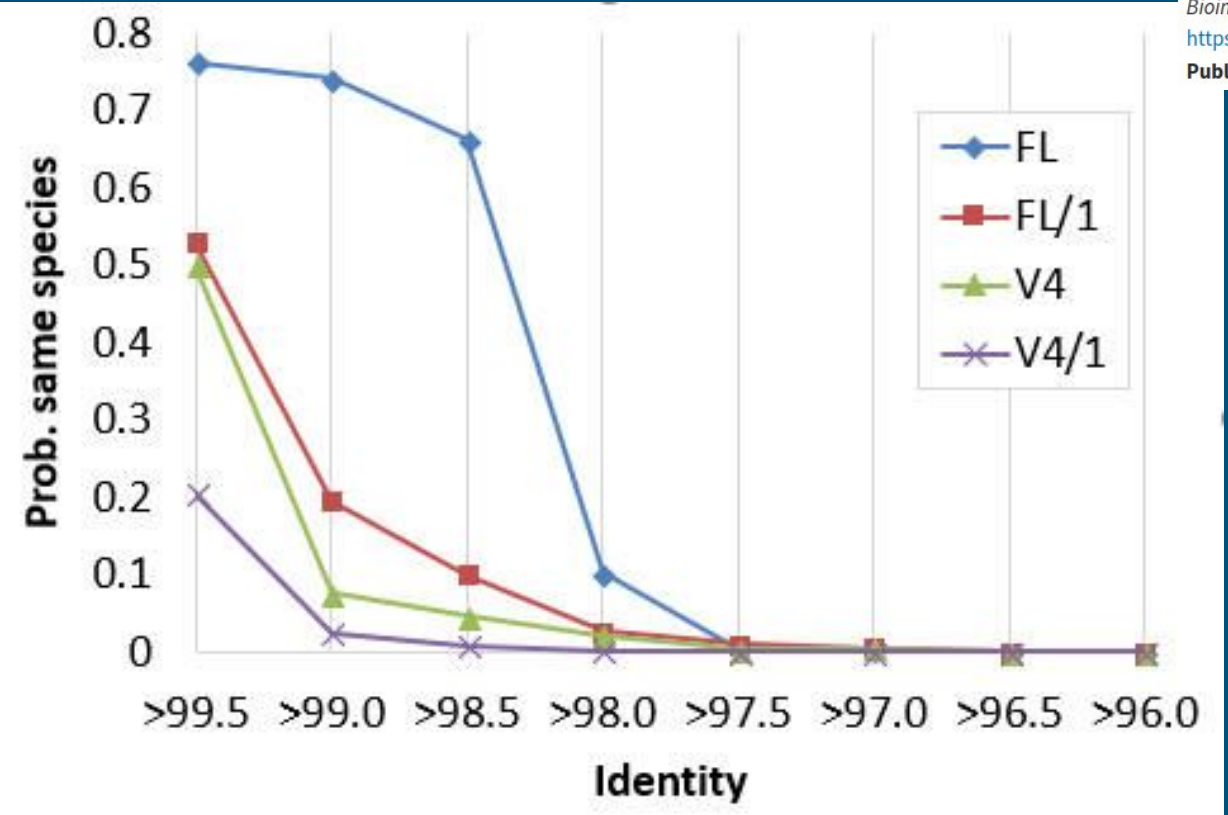
- Elimination of noise, errors
- Are **real biological sequences**
- **Comparability** between studies and datasets
- **Higher taxonomic resolution** (different ASVs from even 1 nucleotide transition)

Disadvantages:

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

Species concept & Operational Taxonomic Units

BioInfo4EEG



JOURNAL ARTICLE

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

Robert C Edgar

Bioinformatics, Volume 34, Issue 14, July 2018, Pages 2371–2375,
https://doi.org/10.1093/bioinformatics/bty113
Published: 28 February 2018 **Article history**

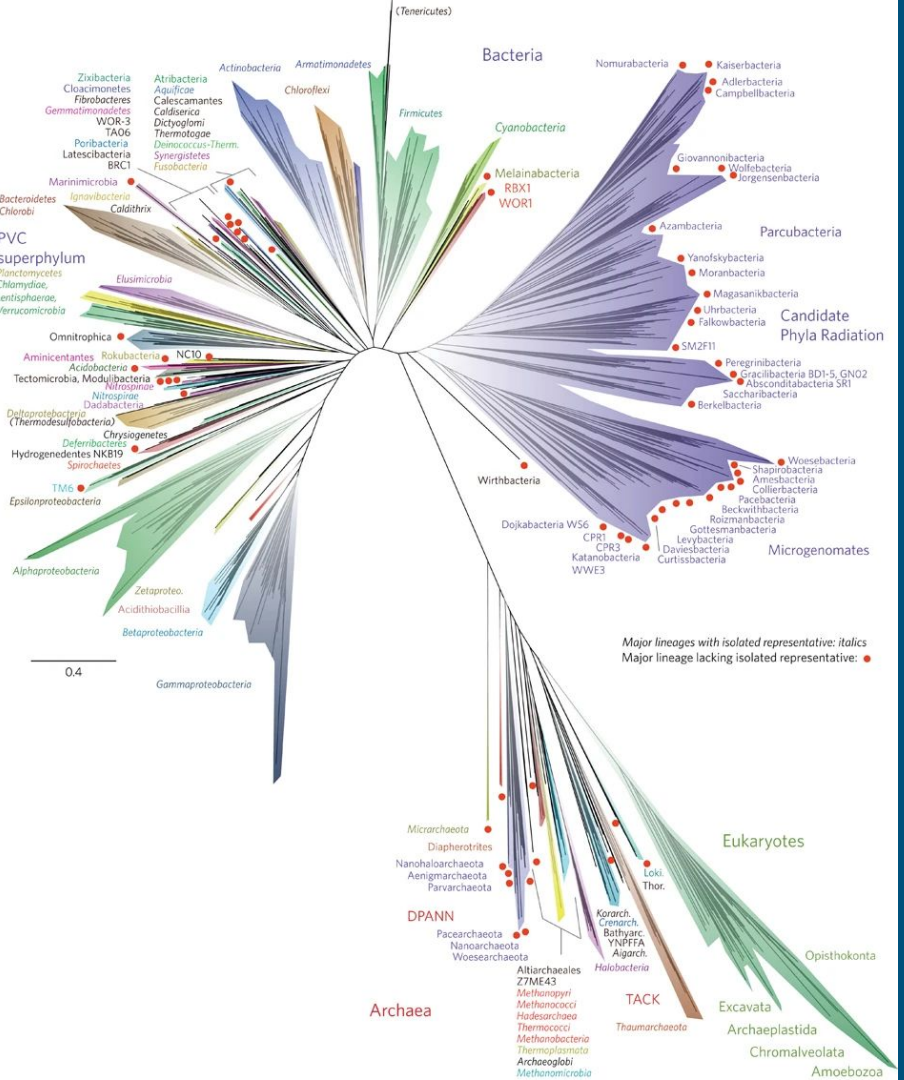
What is a species?

Are species real?

What is an animal species?

Morphology?
Phylogeny?
Ecology?

What is a microbial species?
fungal
bacterial



What is a microbial species?
fungal
bacterial

"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)

Wrapping your head microbial taxonomy is not easy…

Letter | Open access | Published: 11 April 2016

A new view of the tree of life

Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas & Jillian F. Banfield

Nature Microbiology **1**, Article number: 16048 (2016) | Cite this article

The OTU vs ASV debate

16S - Prokaryota

BioInfo4EEG

Perspective | Open access | Published: 21 July 2017

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

Benjamin J Callahan ✉, Paul J McMurdie & Susan P Holmes

The OTU vs ASV debate

16S - Prokaryota

BioInfo4EEG

Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold

Marlène Chiarello, Mark McCauley, Sébastien Villéger, Colin R. Jackson

Published: February 24, 2022 • https://doi.org/10.1371/journal.pone.0264443

RESEARCH ARTICLE

ASV vs OTUs clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies

Andrea Fasolo, Saptarathi Deb, Piergiorgio Stevanato, Giuseppe Concheri, Andrea Squartini*

Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units

Lisa Joos, Stien Beirinckx, Annelies Haegeman, Jane Debode, Bart Vandecasteele, Steve Baeyen, Sofie Goormachtig, Lieven Clement & Caroline De Tender

BMC Genomics **21**, Article number: 733 (2020) | Cite this article

The OTU vs ASV debate

ITS - Fungi
SSU - AM Fungi

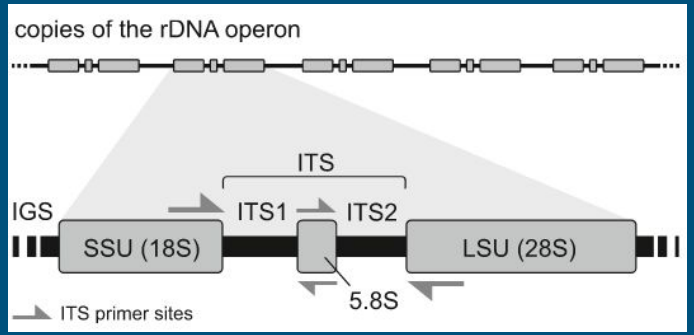
Best practices in metabarcoding of fungi: From experimental design to results

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)

The OTU vs ASV debate

ENVIRONMENTAL MICROBIOLOGY

Research article | Open Access

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

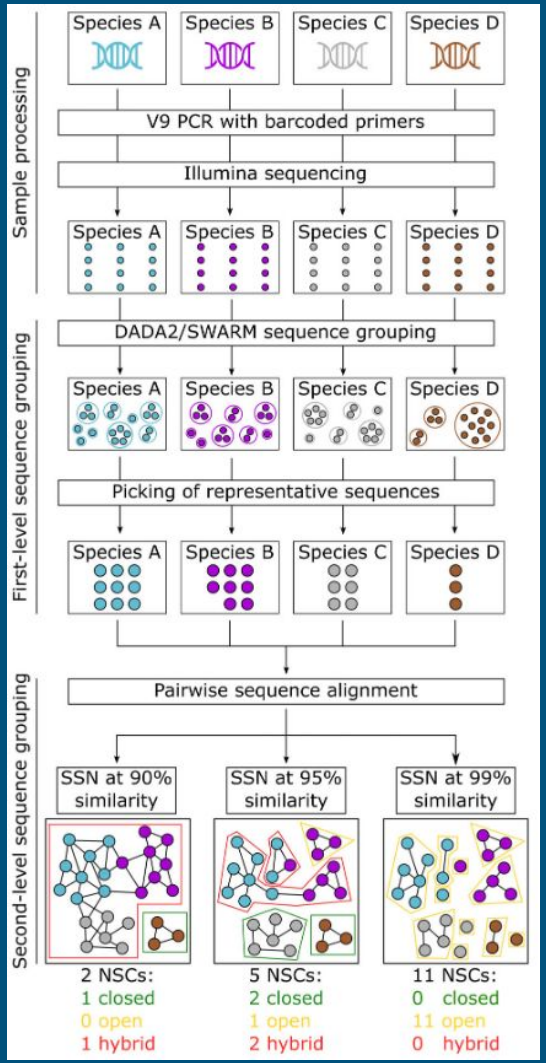
Dominik Forster, Guillaume Lentendu, Sabine Filker, Elyssa Dubois, Thomas A. Wilding, Thorsten Stoeck

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.

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, ...

The OTU vs ASV debate

COI - Animals

BioInfo4EEG

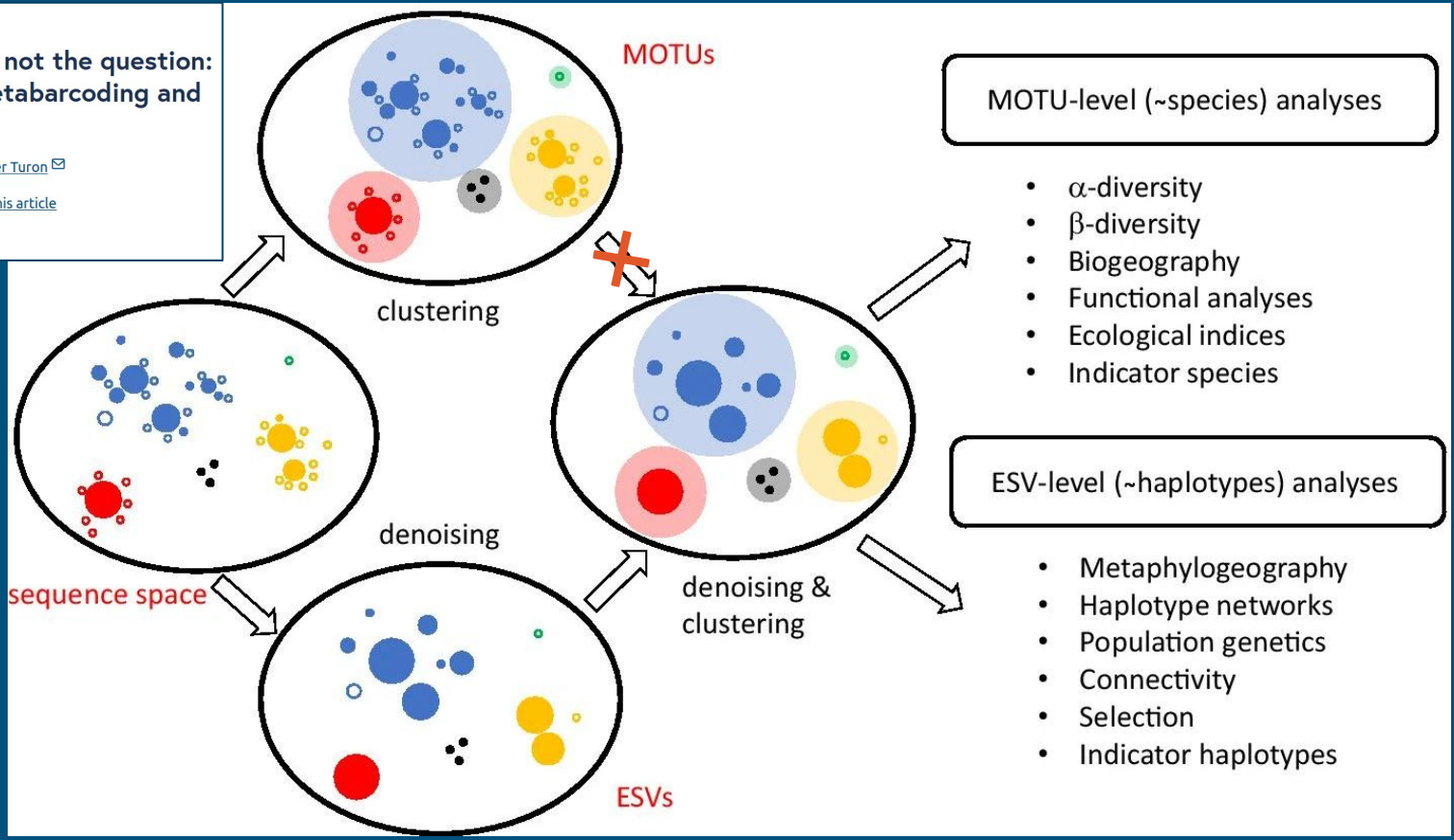
DnoisE tool

Research article | Open access | Published: 05 April 2021

To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography

Adrià Antich, Creu Palacin, Owen S. Wangensteen & Xavier Turon

BMC Bioinformatics **22**, Article number: 177 (2021) | Cite this article



Databases

BOLD SYSTEMS

IDENTIFICATION V4 TAXONOMY WORKBENCH LOGIN

24,868,408
Specimen Records

19,565,646
Specimens with Barcodes

356,304
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:

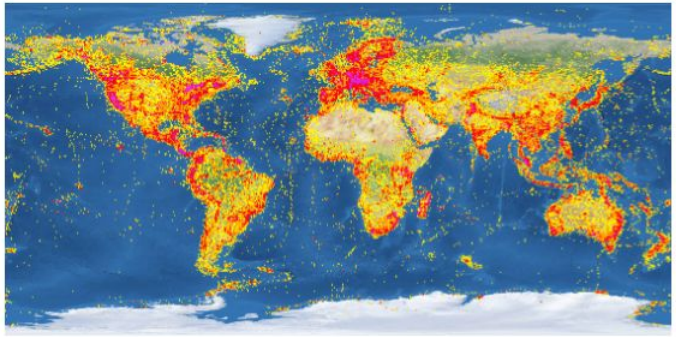
- Bryophyta [24474]
- Chlorophyta [23242]
- Lycopodiophyta [1]
- Magnoliophyta [10687]
- Pinophyta [14]
- Pteridophyta [1307]

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

silva
high quality ribosomal RNA databases

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

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

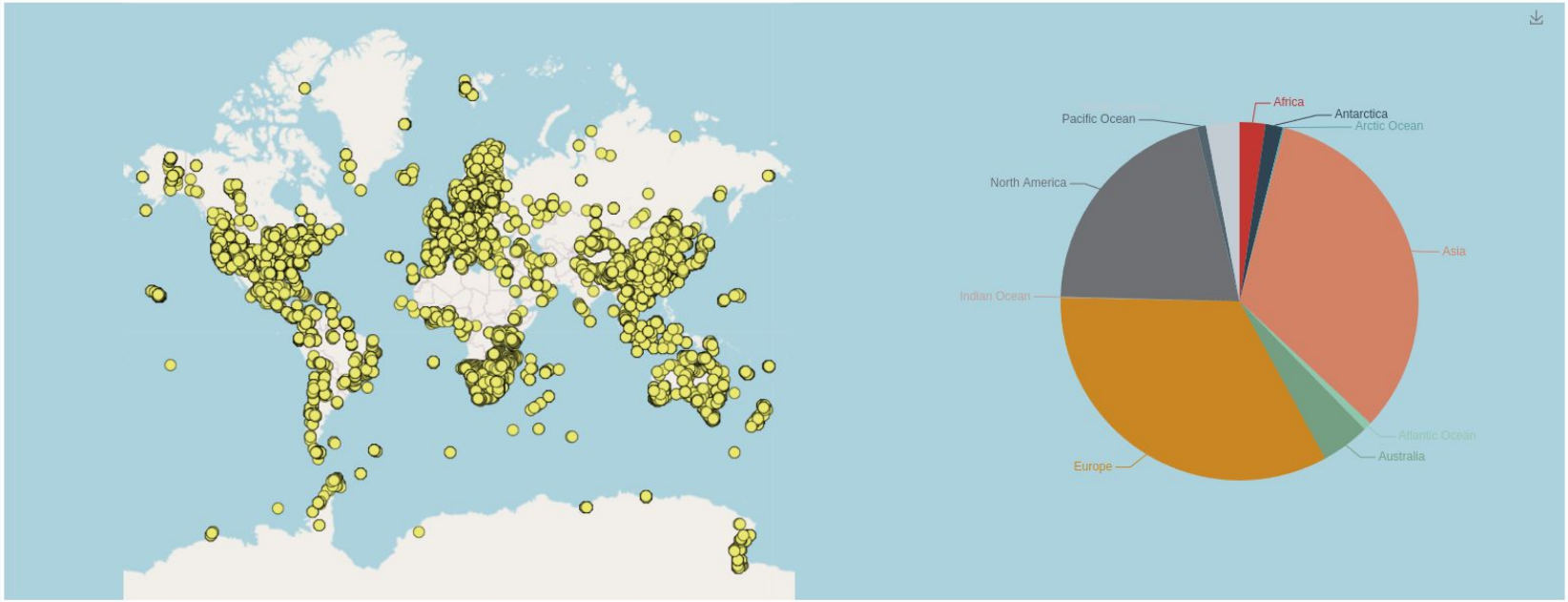
Version no	Release date	Taxon group	No of RefS	No of RepS	Release status	Link	Notes
10.0	2024-04-04	Fungi	18 895	74 190	Current	**https://doi.org/10.15156/BIO/2959332**	**When using this resource, please cite it as follows:** Abarenkov, Kessy; Zirk, Allan; Piirmann, Timo; Pöhönen, Raivo; Ivanov, Filipp; 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.

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

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)

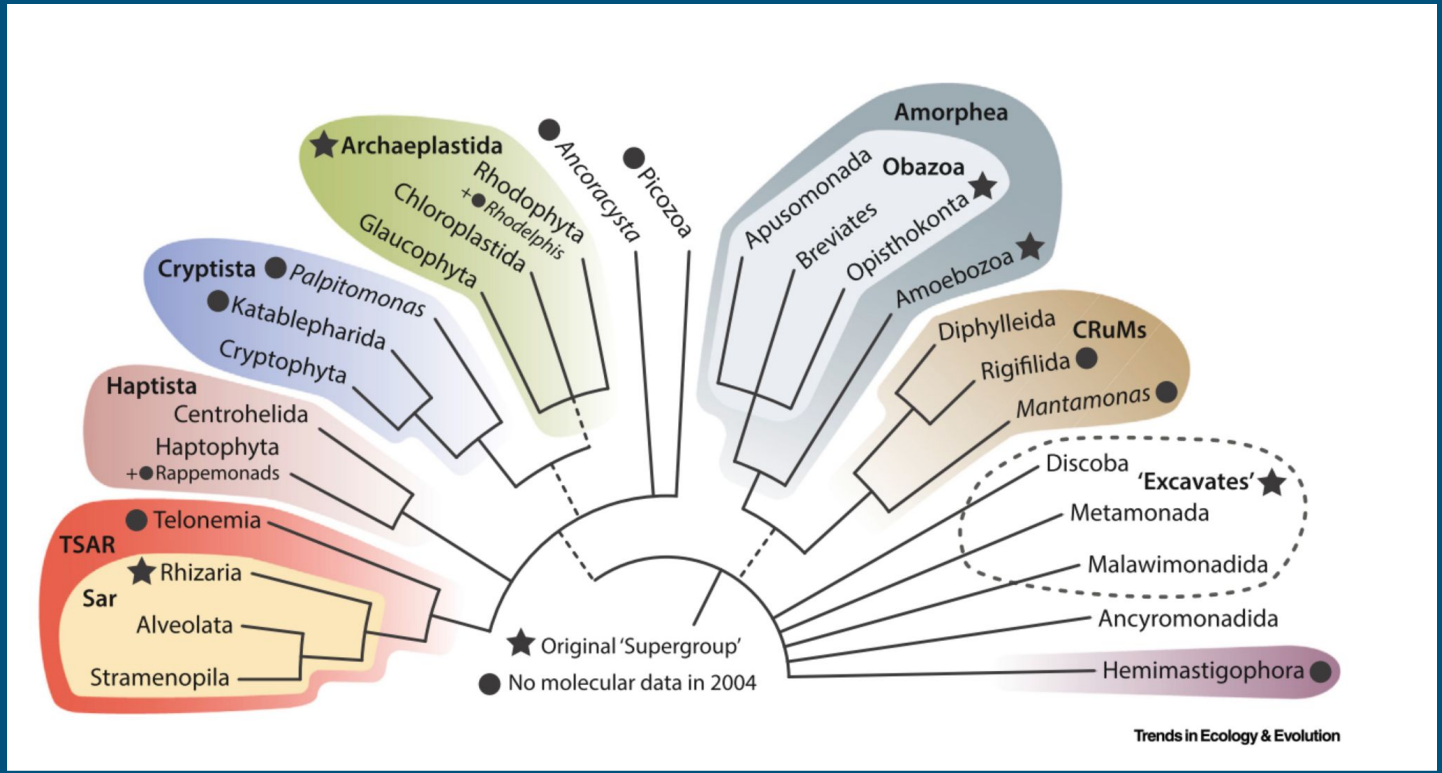


To access published data on arbuscular mycorrhizal fungal communities, click here!

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

Databases

18S - Protists



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 SAfiA-75-219 ORF1 gene, complete cds	Torque teno virus	5236	5540	95%	0.0	93.50%	3707	MN765939.1
Torque teno virus isolate SAfiA-765-2 ORF1 gene, complete cds	Torque teno virus	5227	5472	95%	0.0	93.36%	3694	MN768131.1
Torque teno virus isolate SAfiA-562-2 ORF1 gene, complete cds	Torque teno virus	5123	5476	93%	0.0	93.52%	3628	MN765915.1
Torque teno virus isolate SAfiA-304-7 ORF1 gene, complete cds	Torque teno virus	5090	5219	87%	0.0	94.55%	3370	MN765876.1
Torque teno virus isolate SAfiA-640-1 ORF1 gene, complete cds	Torque teno virus	5068	5202	90%	0.0	93.37%	3506	MN765928.1
Torque teno virus isolate SAfiA-98-7 ORF1 gene, complete cds	Torque teno virus	4800	4877	91%	0.0	91.44%	3549	MN765950.1
Torque teno virus strain ydyzj-6462, complete genome	Torque teno virus	4737	4988	96%	0.0	90.82%	3732	MT783405.1
Torque teno virus isolate SAfiA-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 SAfiA-319-3 ORF1 gene, complete cds	Torque teno virus	4536	4639	80%	0.0	93.65%	3093	MN765881.1
Torque teno virus isolate SAfiA-5A-0 ORF1 gene, complete cds	Torque teno virus	4525	4874	85%	0.0	93.02%	3290	MN767702.1
Torque teno virus isolate SAfiA-401-1 ORF1 gene, complete cds	Torque teno virus	4503	4779	84%	0.0	92.88%	3287	MN765893.1
Torque teno virus isolate SAfiA-630A-11 ORF1 gene, complete cds	Torque teno virus	4473	4639	81%	0.0	93.10%	3147	MN765927.1
Torque teno virus isolate SAfiA-811-176 ORF1 gene, complete cds	Torque teno virus	4464	4539	80%	0.0	93.05%	3099	MN768143.1
Torque teno virus isolate SAfiA-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!