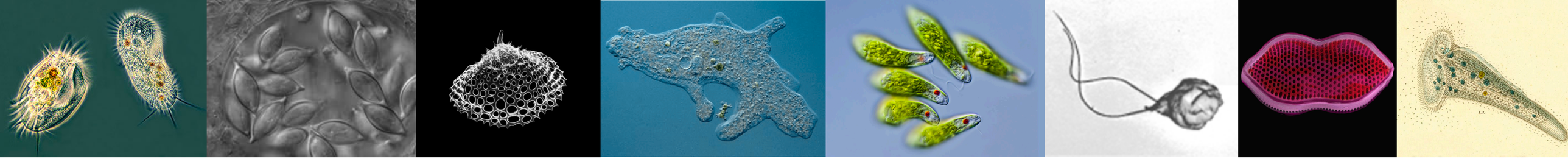


Long-read metabarcoding

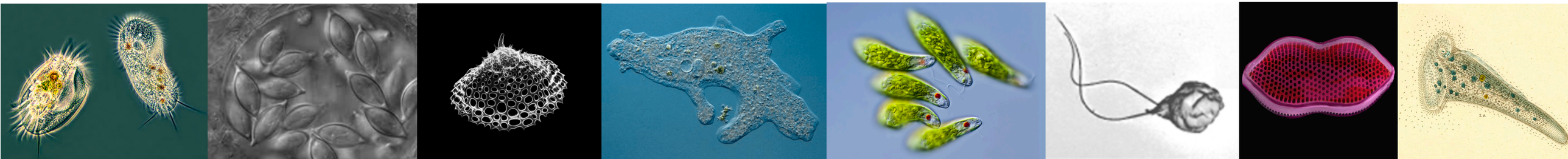
Mahwash Jamy
PhD student, Burki lab

UPPSALA
UNIVERSITET

@JamyMahwash

Your lecturer for the next hour or so

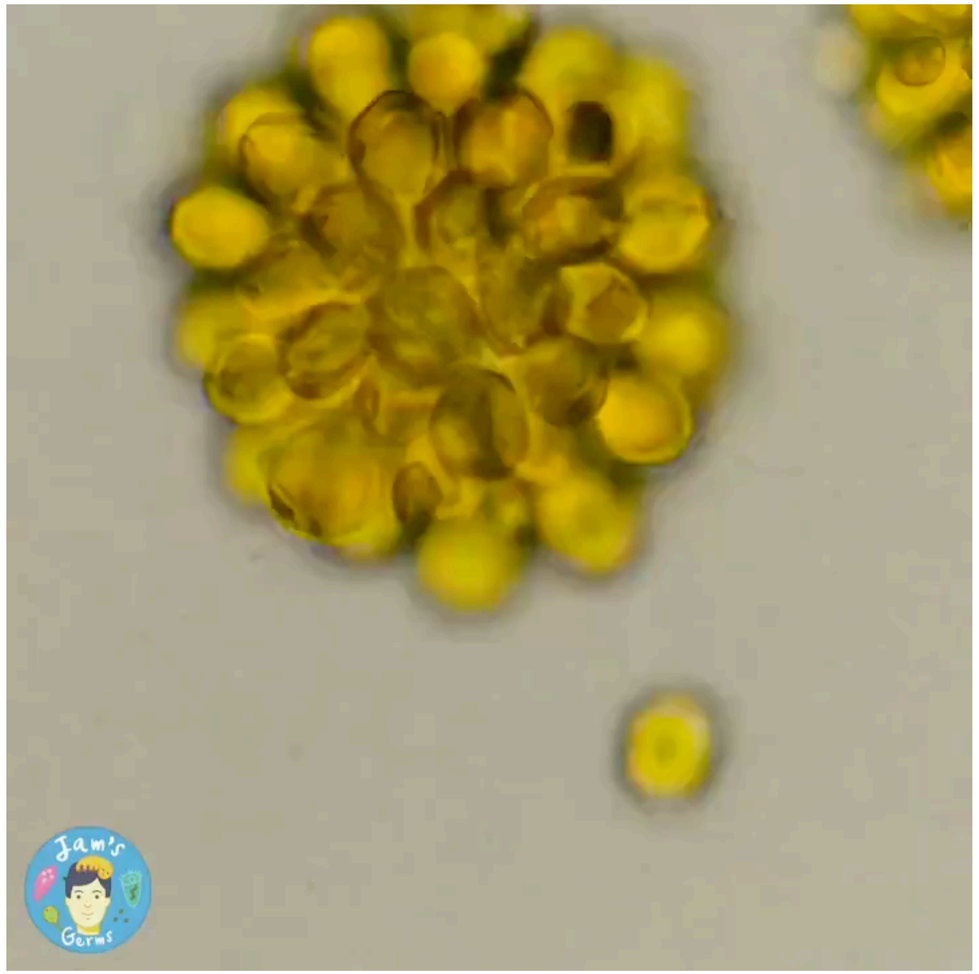
- PhD student with Fabien Burki at Uppsala University
- Thinks protists (microbial eukaryotes) are super cool!
- Assessing protist diversity with environmental sequencing, particularly in a phylogenetic framework



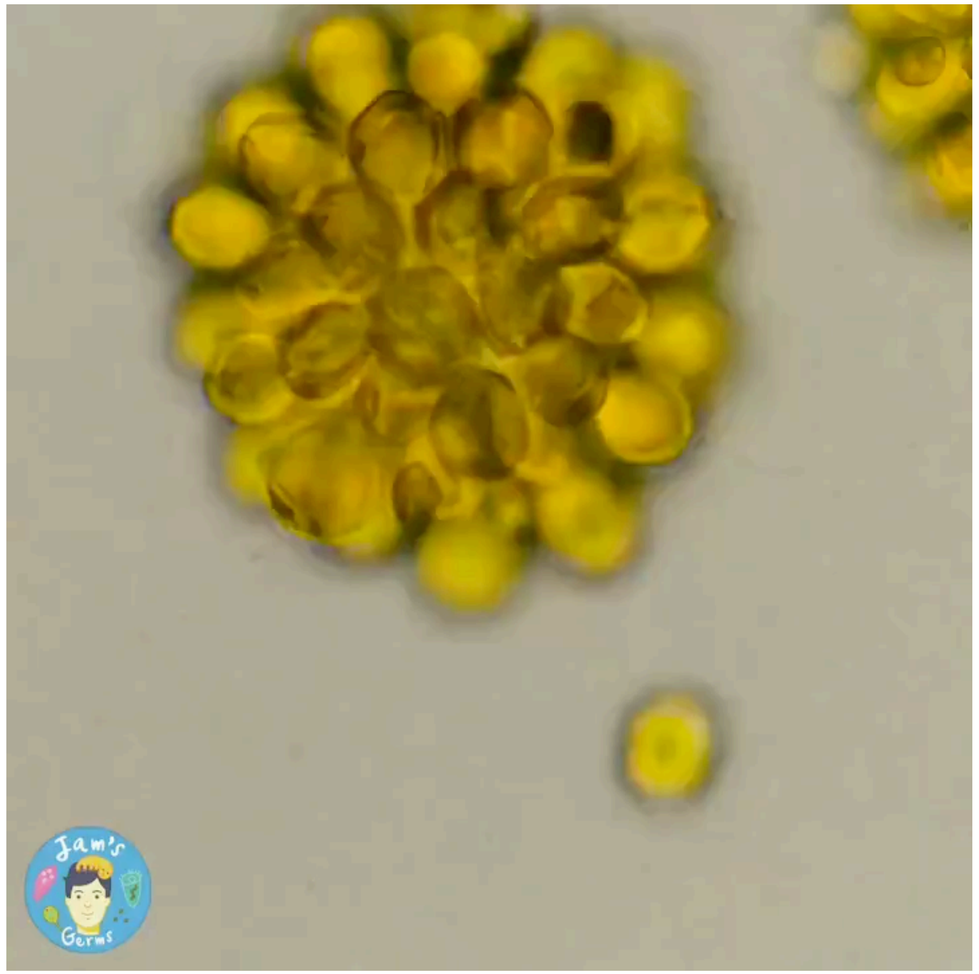
Outline

- What is long-read metabarcoding?
- Why do long-read metabarcoding?
- Which sequencing platform should I use?
- Short break (5 min)
- Cleaning/curating for long-read data
- A case study investigating habitat transitions of protists:
  Using long-read and short-read data together

Any questions?



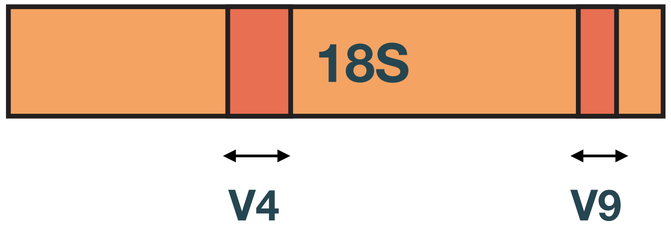
Any questions?



What is long-read metabarcoding?

Metabarcoding with a fragment of the ribosomal SSU gene

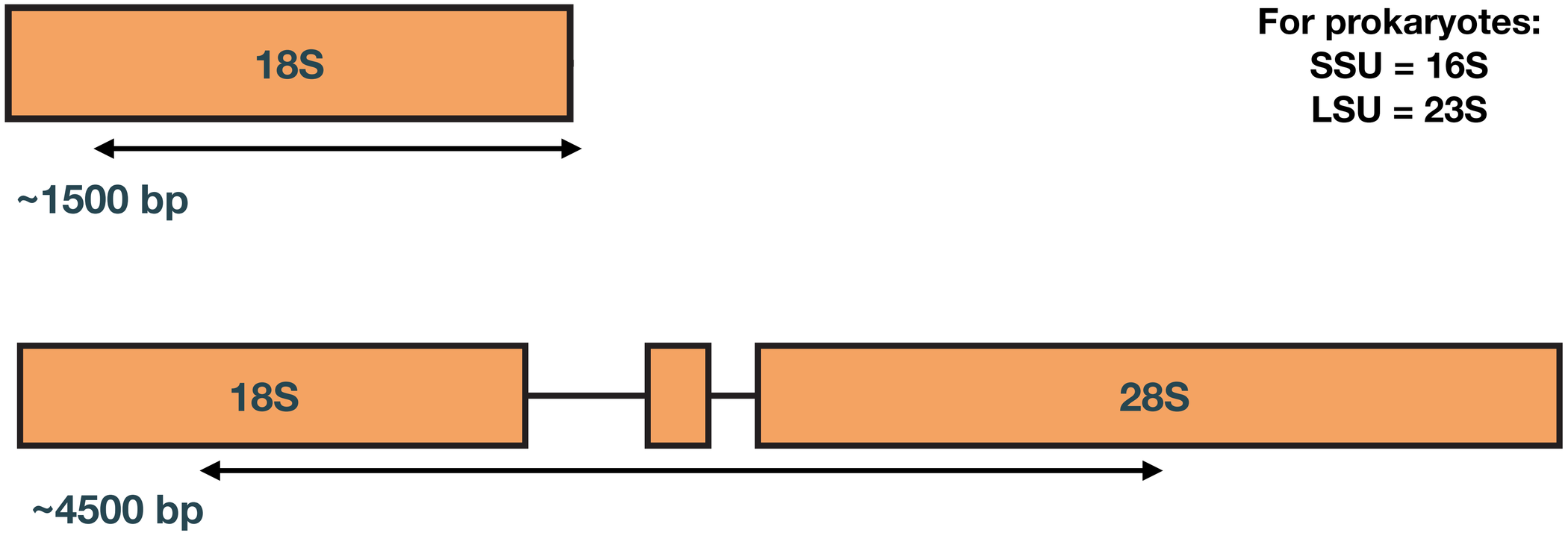
Environmental sample



Sequence > 500 bp fragment with Illumina

Bioinformatic analyses

What is long-read metabarcoding?



For prokaryotes:
SSU = 16S
LSU = 23S

Generally lower-throughput than Illumina sequencing

PACBIO®

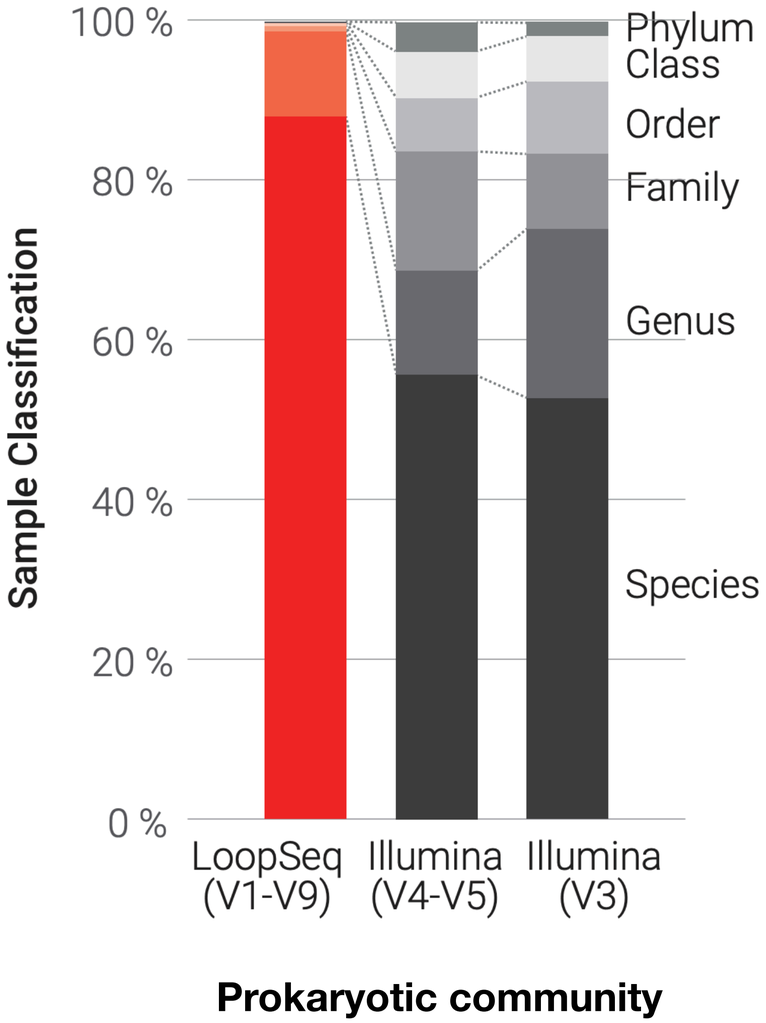
Oxford NANOPORE Technologies

Long-read metabarcoding is lower-throughput than Illumina metabarcoding. So why do it all?

Long reads = increased phylogenetic and taxonomic information

1. Better taxonomic classification

1. Better taxonomic classification



More nucleotide data to tease apart closely related organisms (regardless of taxonomic classification method)

LOOP GENOMICS

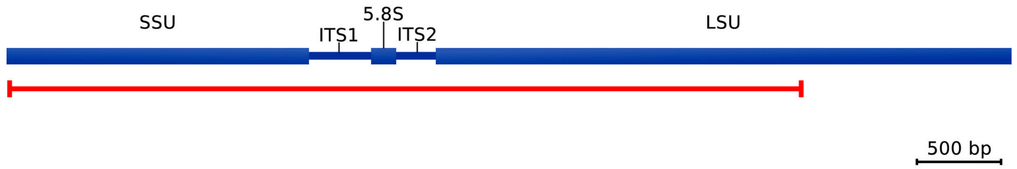
1. Better taxonomic classification

RESOURCE ARTICLE | Full Access

Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments

Felix Heeger, Elizabeth C. Bourne, Christiane Baschien, Andrey Yurkov, Boyke Bunk, Cathrin Spröer, Jörg Overmann, Camila J. Mazzoni, Michael T. Monaghan

First published: 14 August 2018 | https://doi.org/10.1111/1755-0998.12937 | Citations: 32



Heeger et al. 2018, Molecular Ecology Resources

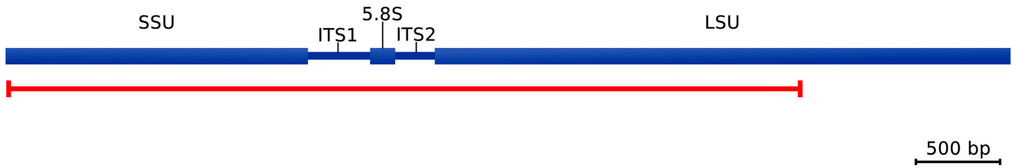
1. Better taxonomic classification

RESOURCE ARTICLE | Full Access

Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments

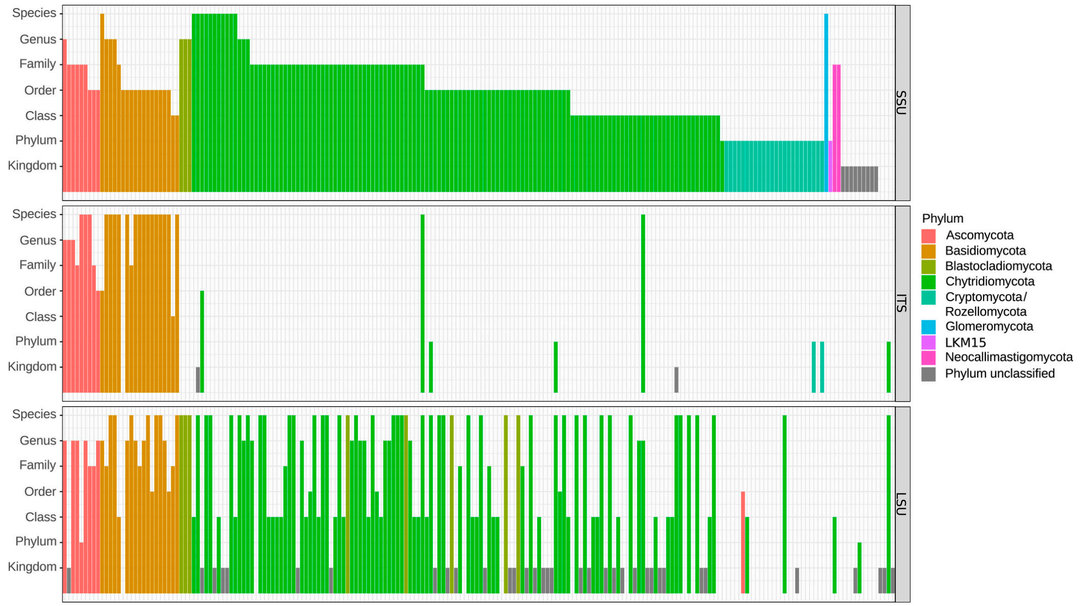
Felix Heeger, Elizabeth C. Bourne, Christiane Baschien, Andrey Yurkov, Boyke Bunk, Cathrin Spröer, Jörg Overmann, Camila J. Mazzoni, Michael T. Monaghan

First published: 14 August 2018 | https://doi.org/10.1111/1755-0998.12937 | Citations: 32



- Able to use different reference databases (UNITE, SILVA, RDP LSU)

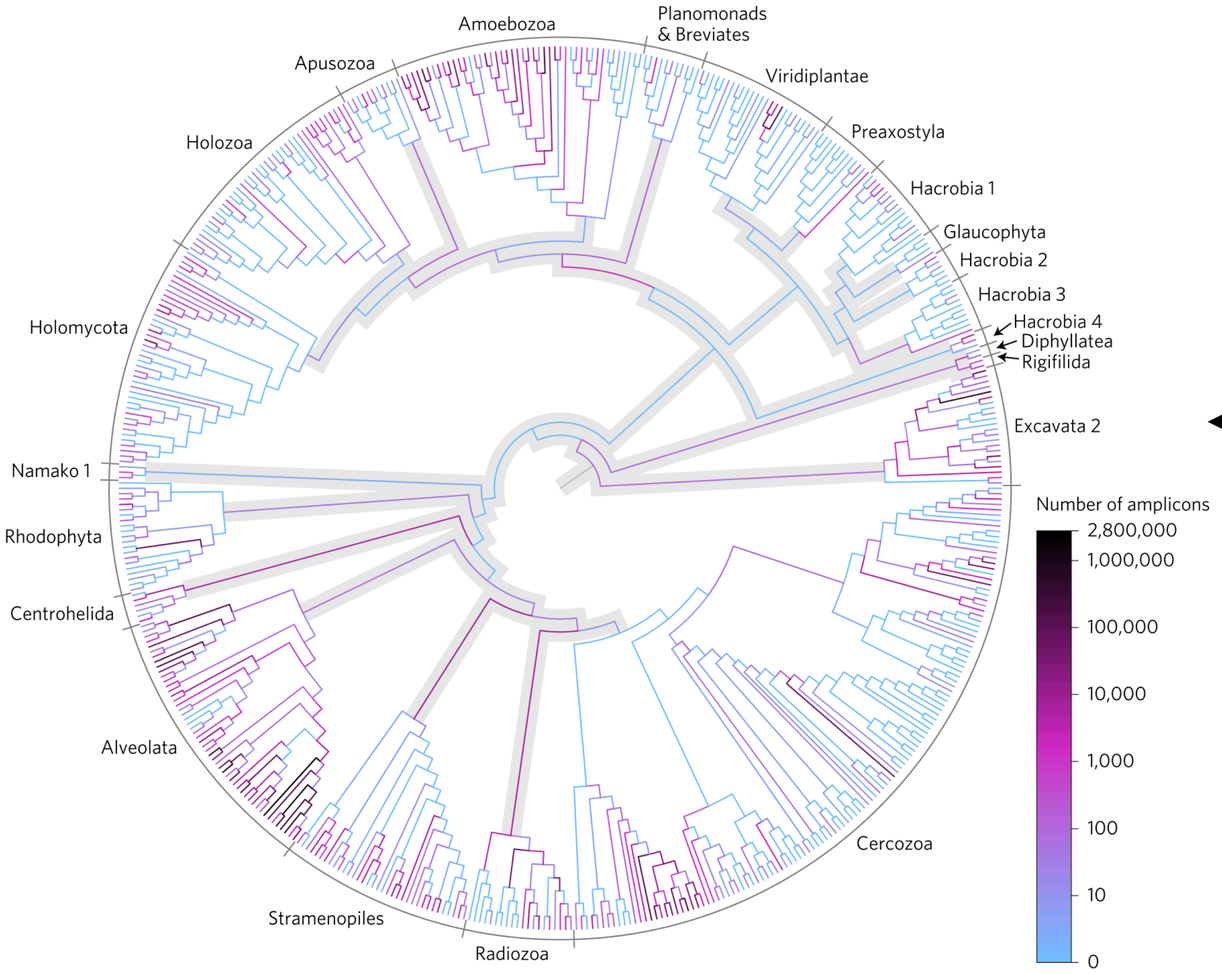
- Able to classify OTUs that are not included in ITS and LSU databases (mostly early diverging fungi)



Heeger et al. 2018, Molecular Ecology Resources

2. Generating reference sequences for phylogenetic placement, and for populating databases

2. Generating reference sequences for phylogenetic placement



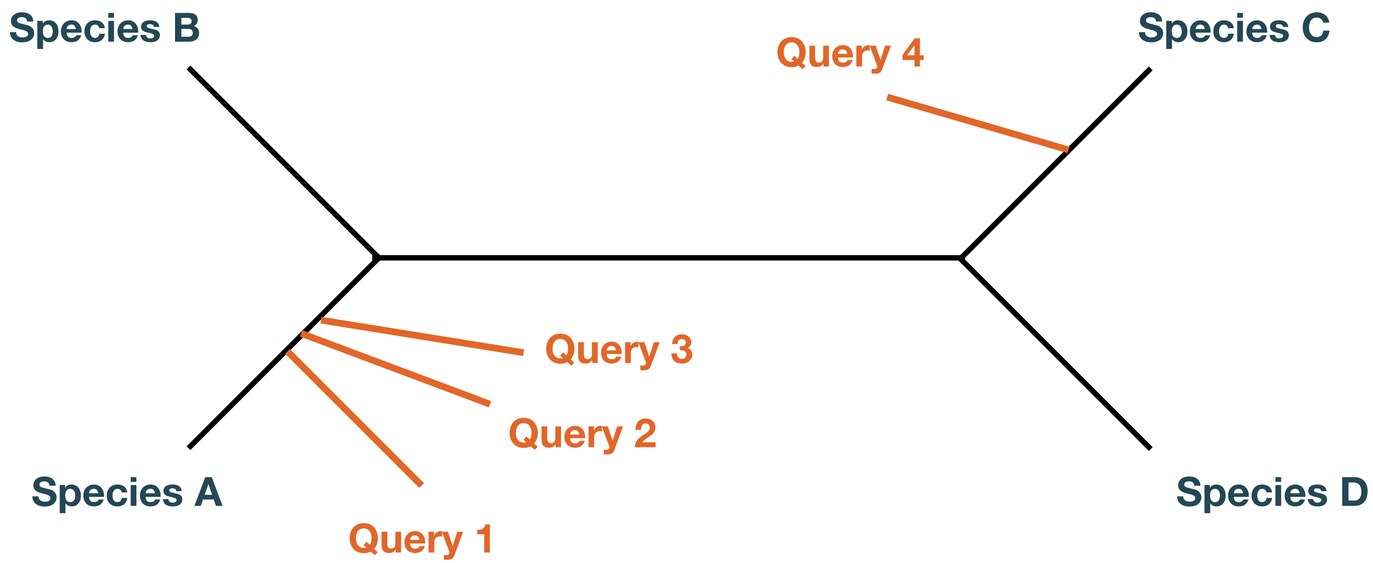
Reference sequences are typically generated by Sanger sequencing which is low-throughput and expensive

Long-read sequencing can generate these reference sequences rapidly and in a cost effective way.

Mahé et al. 2017, Nature Ecology and Evolution

3. Building phylogenies to explore the great environmental diversity

Phylogenetic placement does not solve all problems..



How are queries 1, 2 and 3 related to each other?

3. Building phylogenies to explore the great environmental diversity

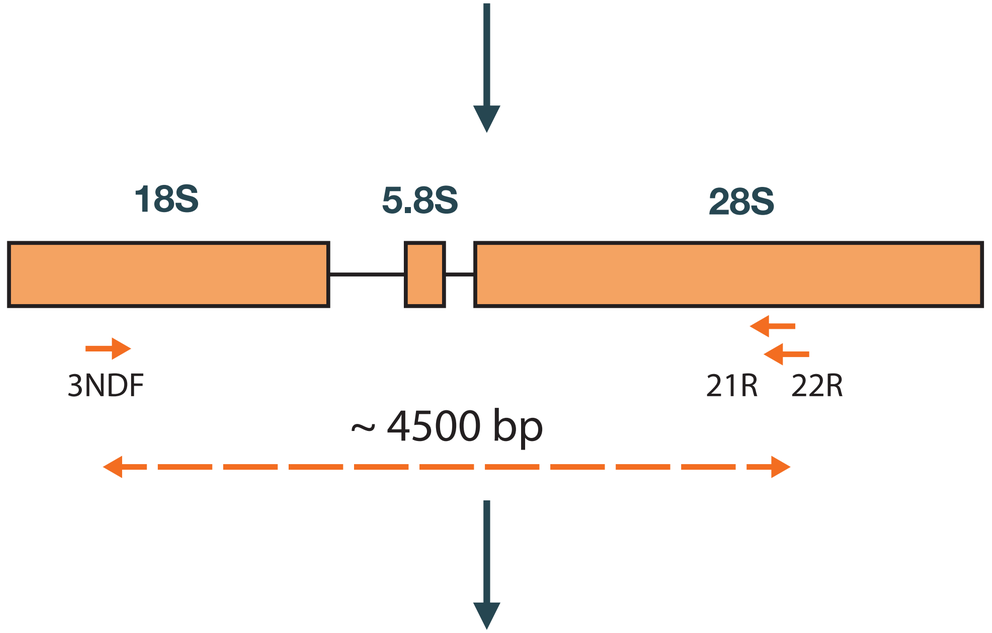
Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity

Mahwash Jamy, Rachel Foster, Pierre Barbera, Lucas Czech, Alexey Kozlov, Alexandros Stamatakis, Gary Bending, Sally Hilton, David Bass ✉, Fabien Burki ✉

First published: 09 November 2019 | https://doi.org/10.1111/1755-0998.13117 | Citations: 10

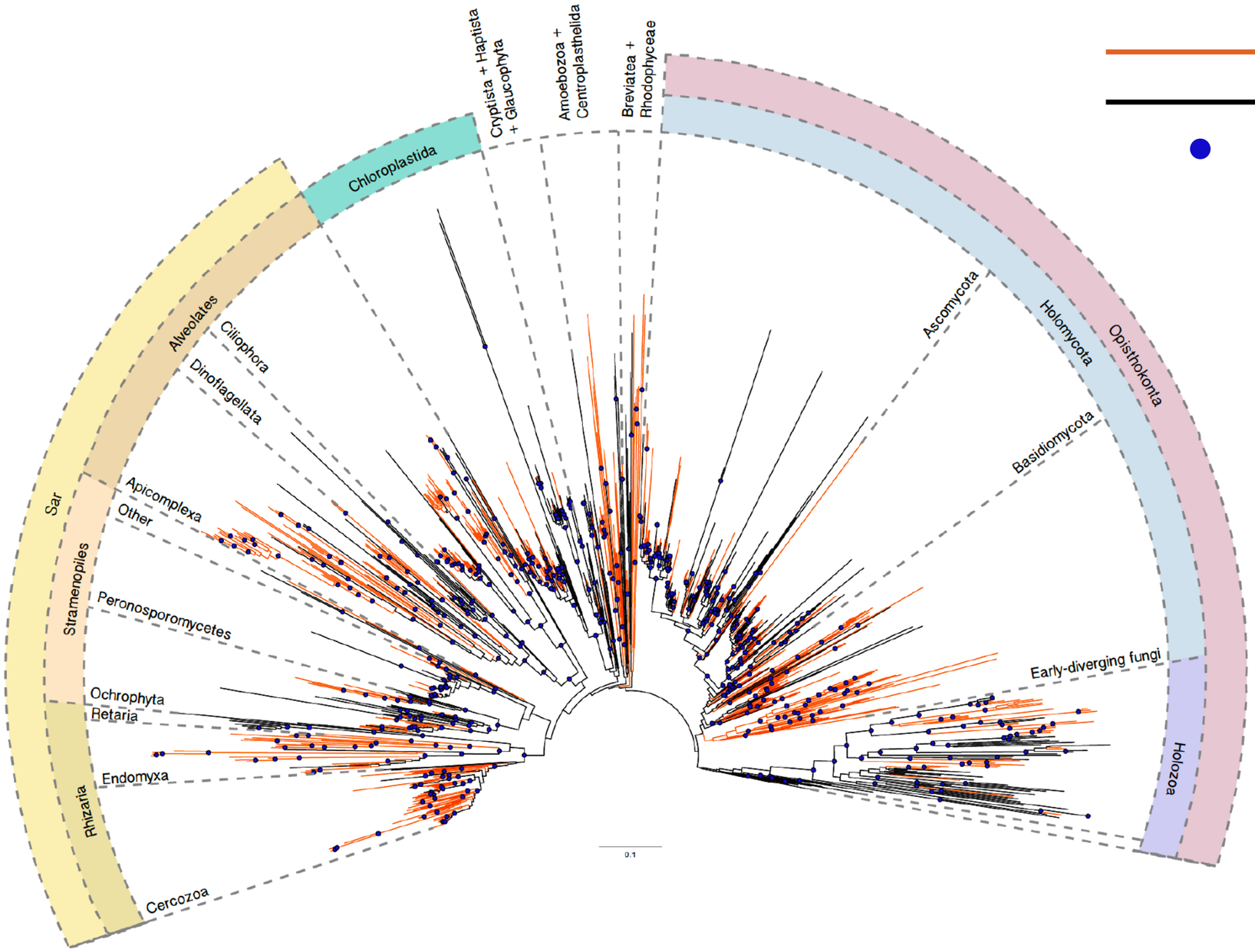
Soil samples (3x)

PACBIO®



650 OTUs

Combined 18S + 28S ML tree



Env. sequences

References

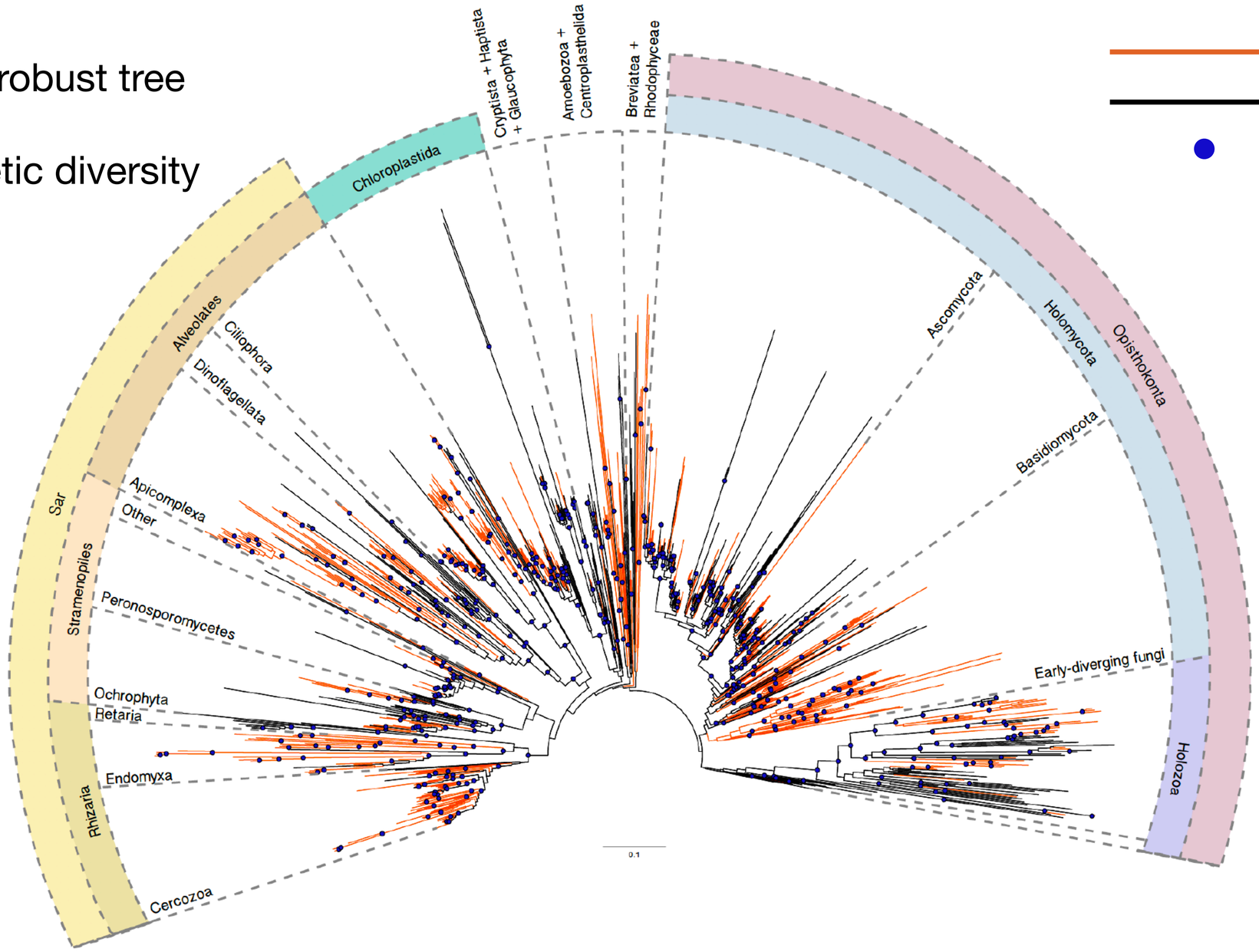
70% bootstrap

1019 taxa
4304 positions

Combined 18S + 28S ML tree

18S-28S dataset = robust tree

Visualise phylogenetic diversity



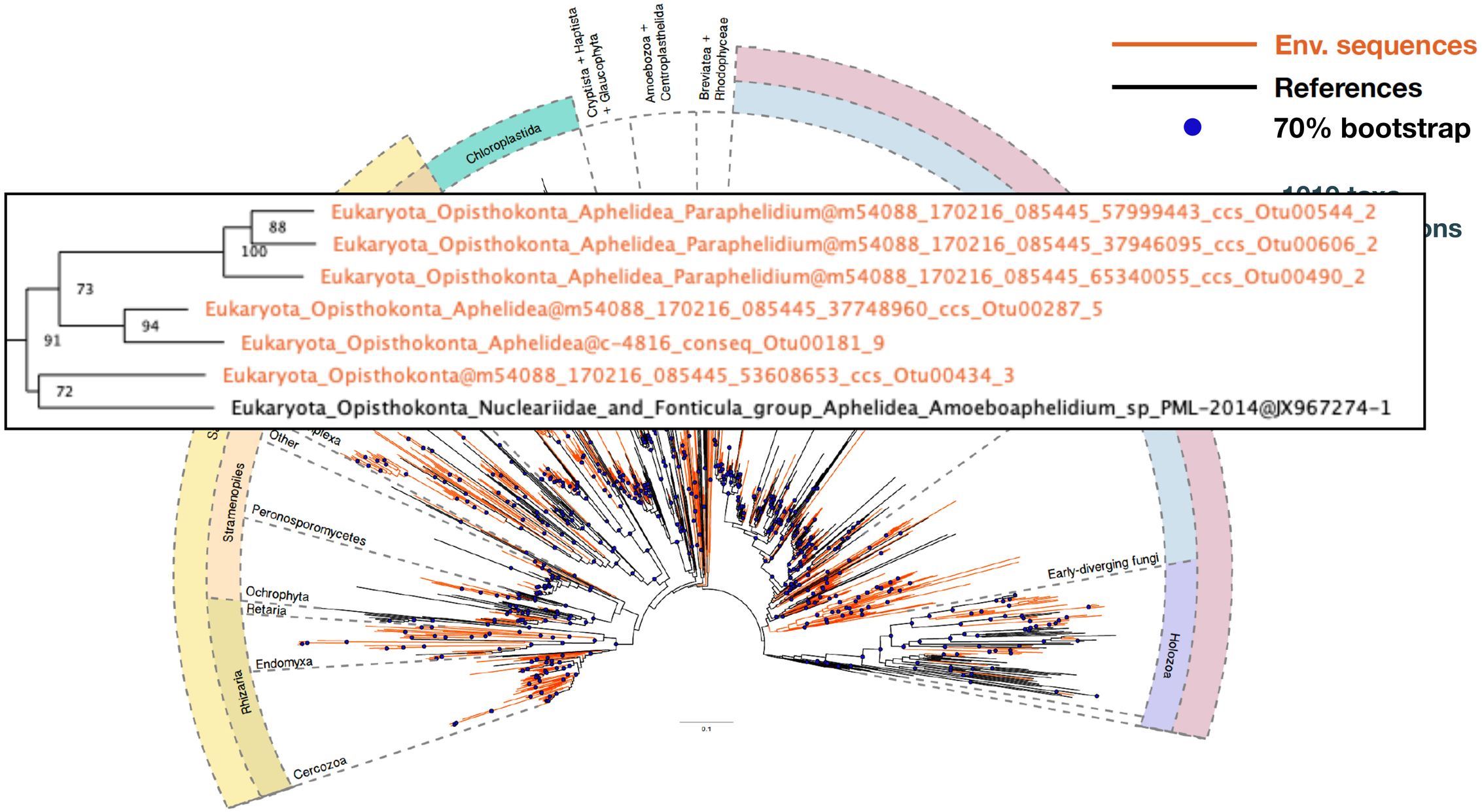
Env. sequences

References

70% bootstrap

1019 taxa
4304 positions

Combined 18S + 28S ML tree



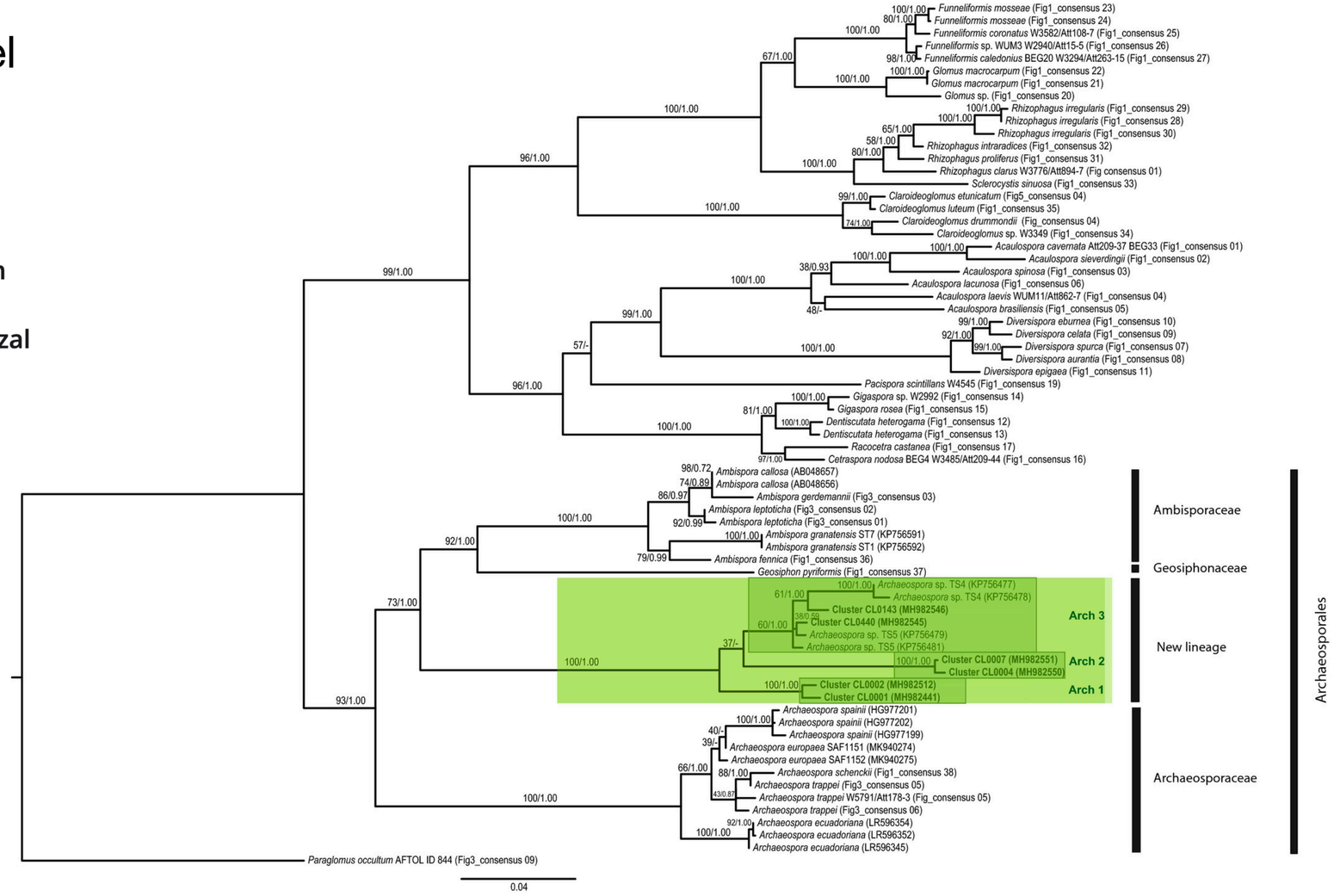
Building phylogenies to explore the great environmental diversity

Particularly useful for discovering novel lineages, and investigating their position in the tree of life.

PacBio sequencing of Glomeromycota rDNA: a novel amplicon covering all widely used ribosomal barcoding regions and its applicability in taxonomy and ecology of arbuscular mycorrhizal fungi

Zuzana Kolaříková ✉, Renata Slavíková, Claudia Krüger, Manuela Krüger, Petr Kohout ✉

First published: 29 March 2021 | https://doi.org/10.1111/nph.17372



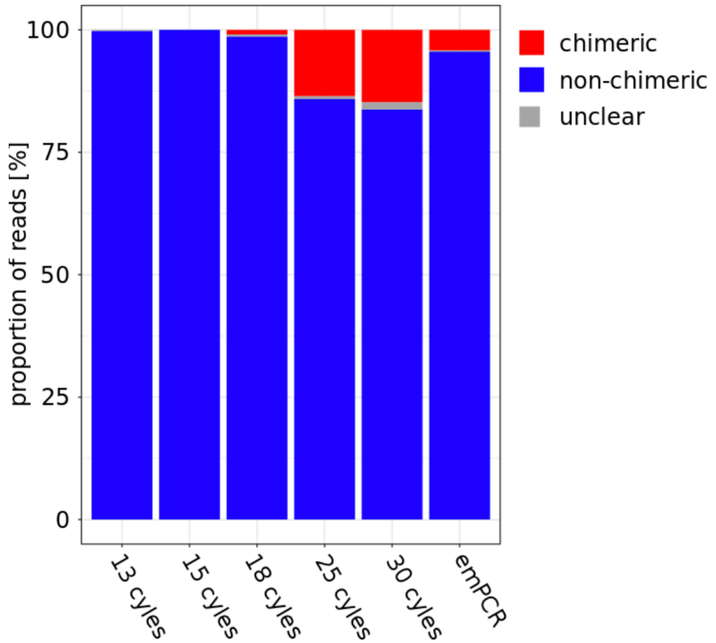
Any questions?

Any questions?

Which sequencing platform should I use?

From DNA to long-reads

- Beware of chimeras!!
- Longer amplicons are more prone to chimeras
- Use fewer PCR cycles if possible



Heeger et al. 2018, Molecular Ecology Resources

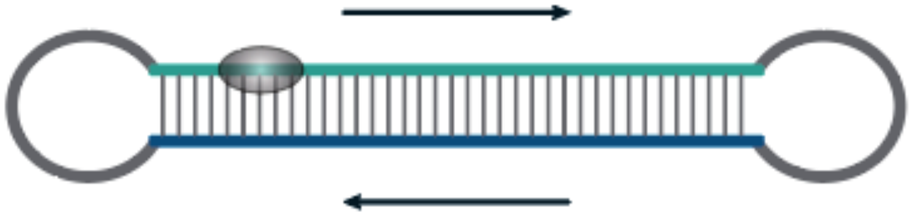
Which sequencing platform should I use?

- PacBio Sequel and Sequel II
- Nanopore
- LoopSeq (synthetic long-reads)

- Error rate
- Portability
- The type of community you want to sequence
- Sequence length/marker

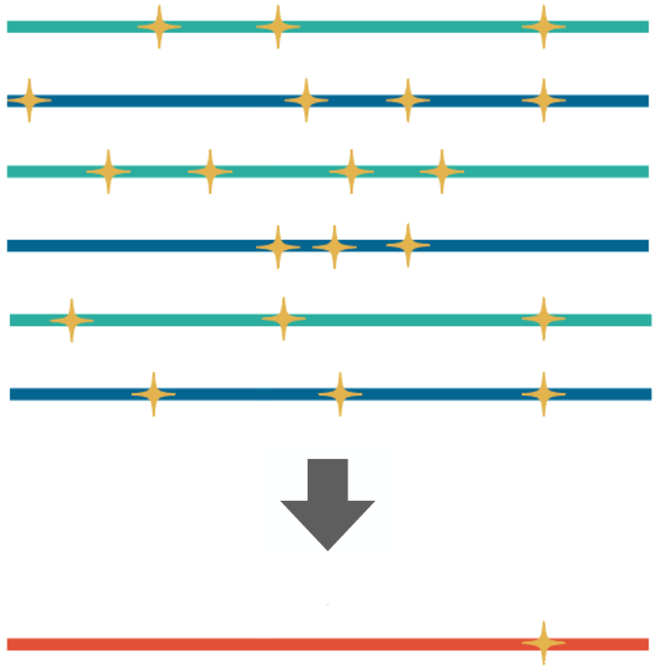
PacBio

SMRTbell template



Hairpin adaptors

Multi-pass sequencing



Subreads = passes
Error rate = 14-15%
BUT randomly distributed

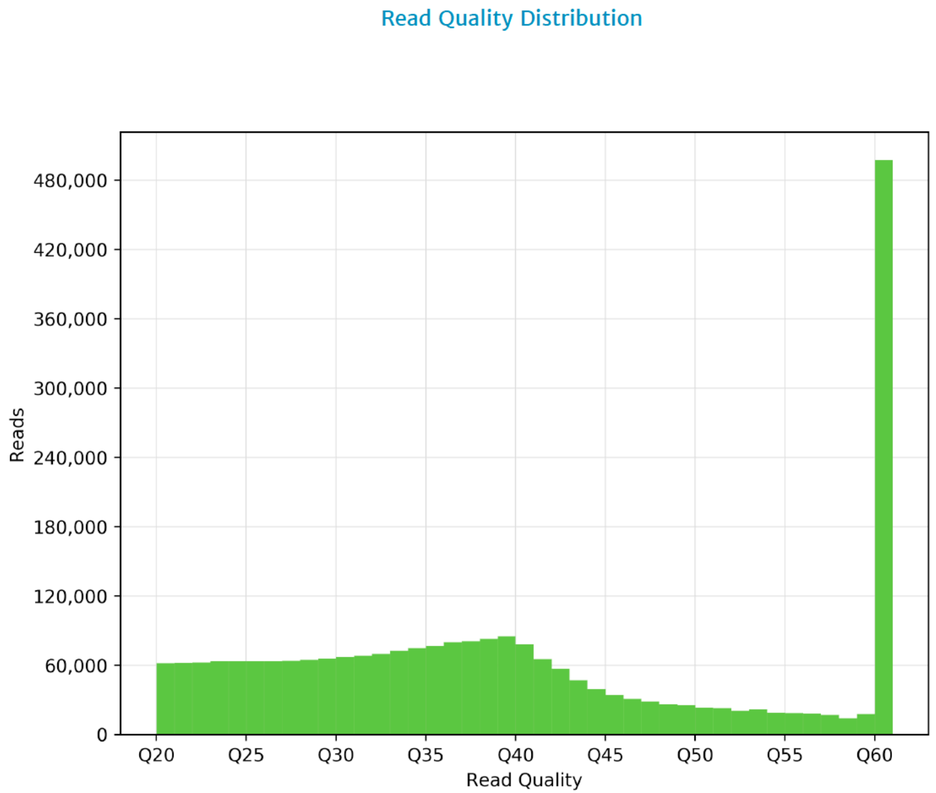
CCS = HiFi

Circular Consensus Sequence (CCS)
Much lower error rate

PacBio

- Use Sequel II if possible (cheaper and more high-throughput than Sequel) ~35,000 CCS per cell in Sequel vs ~3,000,000 CCS per cell for Sequel II

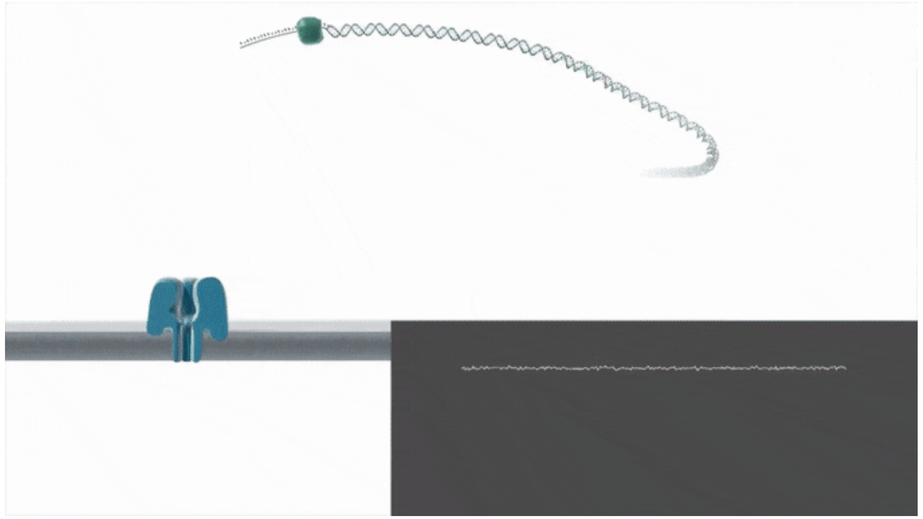
- Works well for complex communities



PacBio

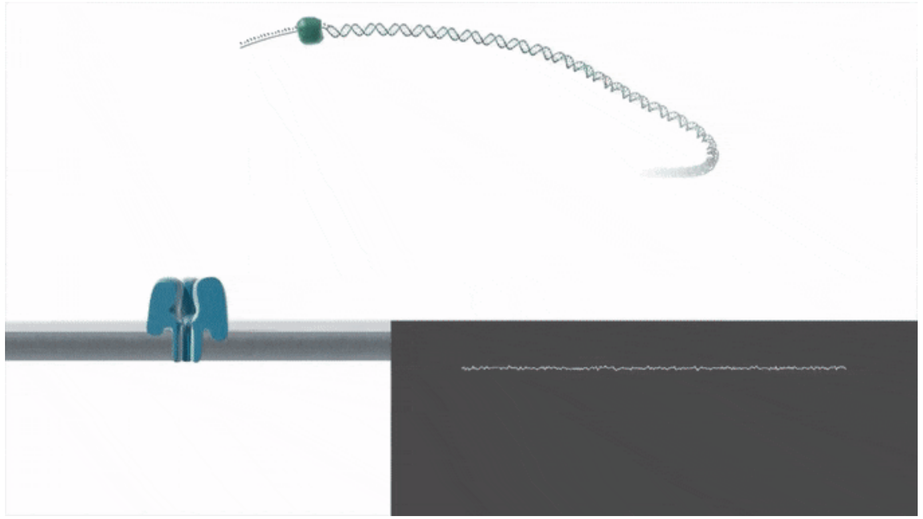
- Error rate. Raw error rate is high. Error rate of CCS is very low.
- Portability. Low
- The type of community you want to sequence. Complex community
- Sequence length/marker. Up to 20 kb

Nanopore



- Avg error rate = 10-15%
- No CCS technology

Nanopore



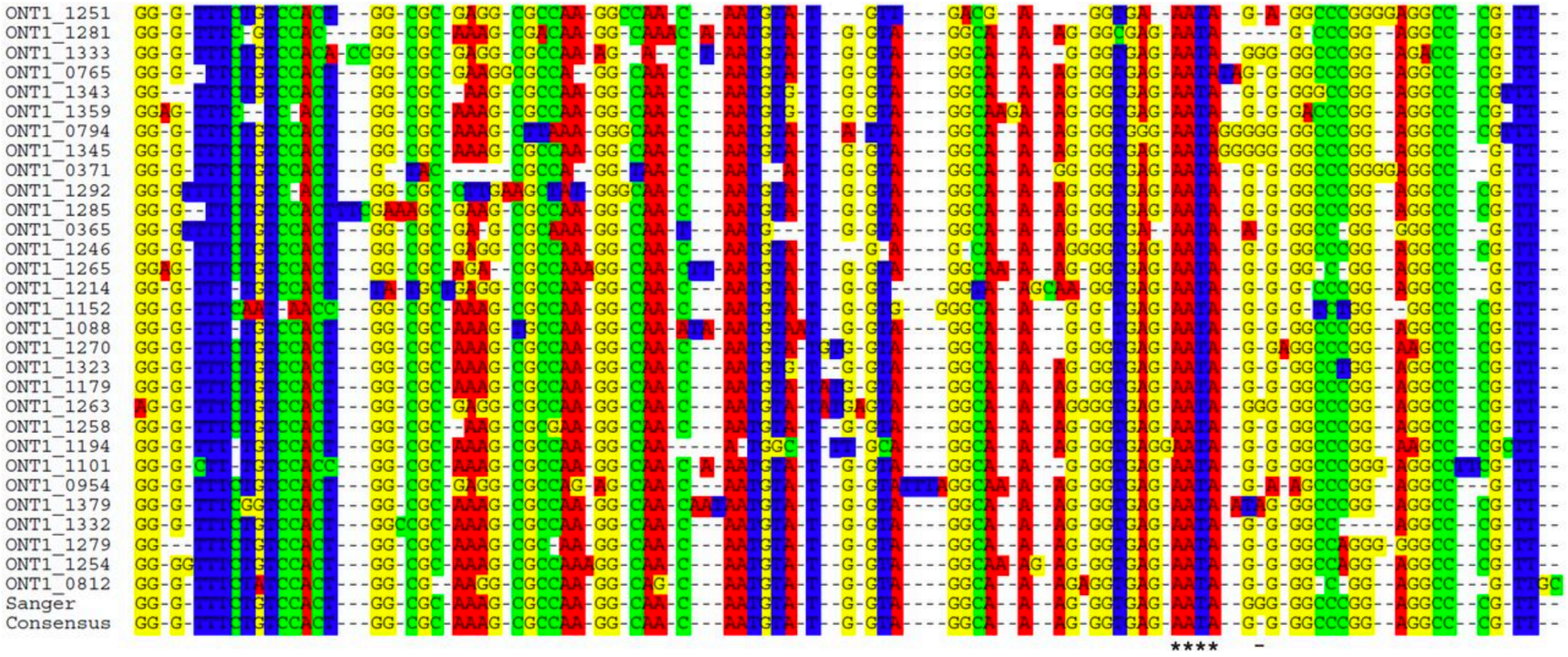
- Avg error rate = 10-15%
- No CCS technology

Nanopore

Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens

Kaire Loit, Kalev Adamson, Mohammad Bahram, Rasmus Puusepp, Sten Anslan, Riinu Kiiker, Rein Drenkhan, Leho Tedersoo

Irina S. Druzhinina, Editor



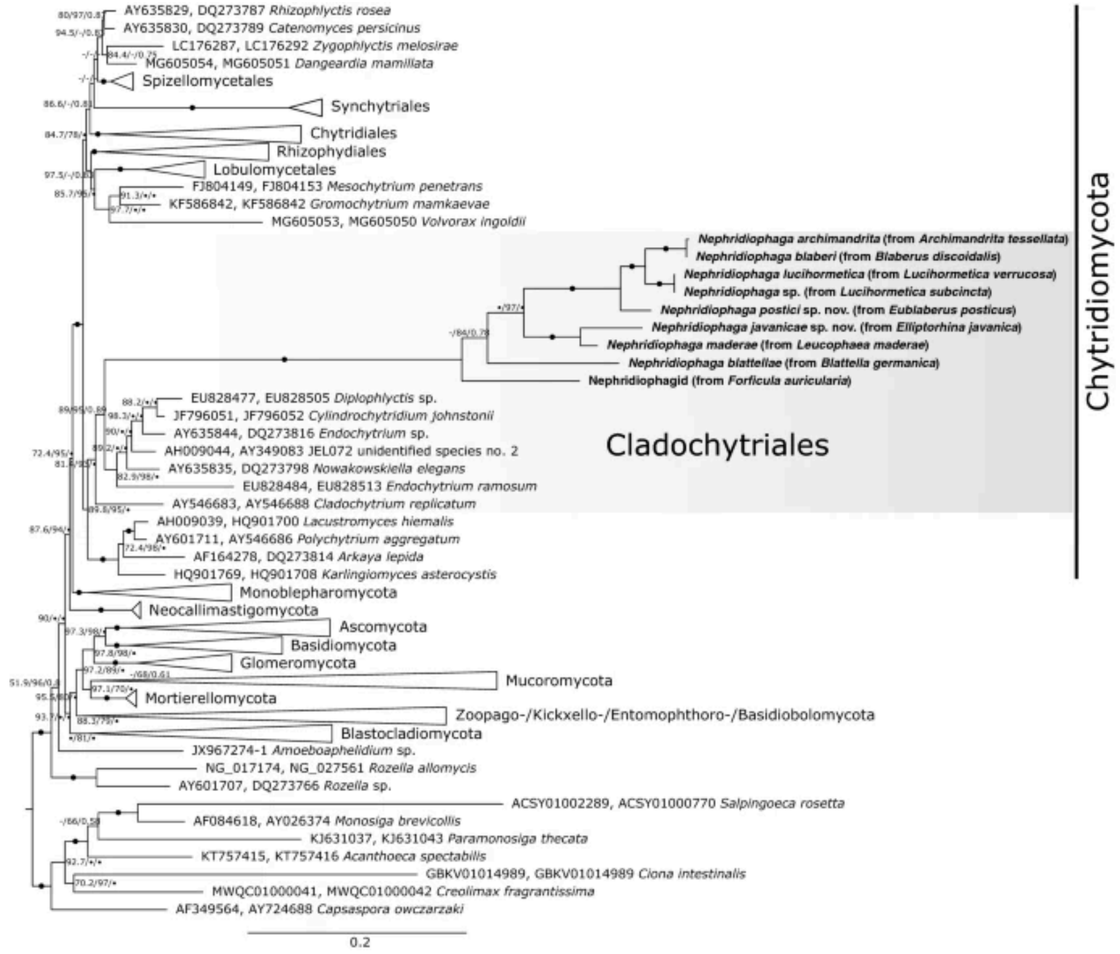
Nanopore

- Error rate. Raw error rate is high. Error rate of processed reads is still too high (compared to Illumina and PacBio).
- Portability. High!! Also very rapid.
- The type of community you want to sequence. Simple community
- Sequence length/marker. Up to 20 kb

Article | Open Access | Published: 11 January 2021

Long rDNA amplicon sequencing of insect-infecting nephridiophagids reveals their affiliation to the Chytridiomycota and a potential to switch between hosts

Jürgen F. H. Strassert, Christian Wurzbacher, Vincent Hervé, Taraha Antany, Andreas Brune & Renate Radek



LoopSeq

LOOP GENOMICS

Synthetic long reads

LoopSeq™ 16S Long Read Kit

$1,600.00

Quantity

1

Add to Cart

SHIPPING INFO −

$100 to any US desitination

PRODUCT INFO +

Quote +

Ultra-accurate Microbial Amplicon Sequencing Directly from Complex Samples with Synthetic Long Reads

Benjamin J Callahan, Dmitry Grinevich, Siddhartha Thakur, Michael A Balamotis, Tuval Ben Yehezkel

doi: https://doi.org/10.1101/2020.07.07.192286

LoopSeq

LOOP GENOMICS

Attach.

Every sample is exposed to millions of unique barcodes, but only one barcode attaches per strand of DNA at the 16S site.

Each molecule is tagged with a unique barcode

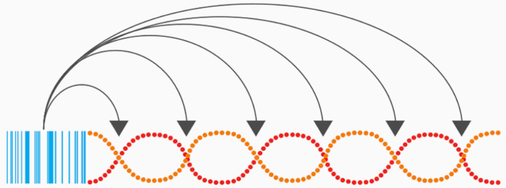
Amplify.

Every molecule, along with its unique barcode, is amplified using PCR.

It’s amplified

Distribute.

For each molecule copy, the barcode is randomly distributed within the molecule.



The barcode is inserted at various points in the SAME molecule

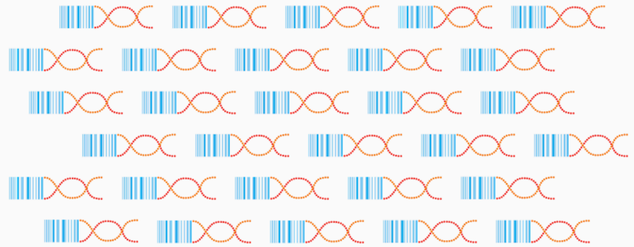
LoopSeq

LOOP GENOMICS

Regular Illumina sequencing

Sequence.

Sequence the segment next to each barcode.



Assemble into long-reads using barcodes

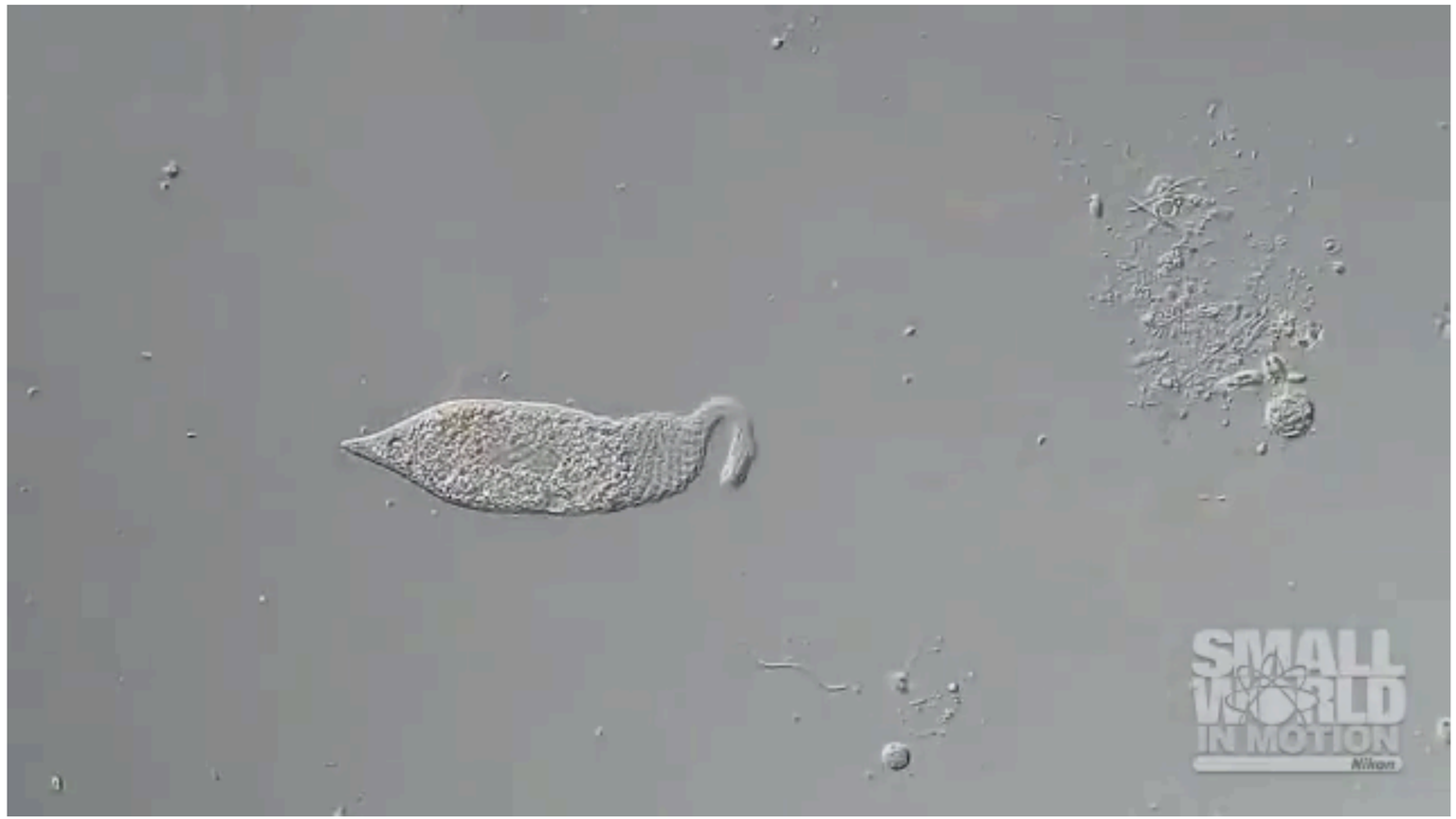
Assemble.

Short reads that share the same barcode are combined algorithmically into a full-length molecule using linked-read de novo assembly.

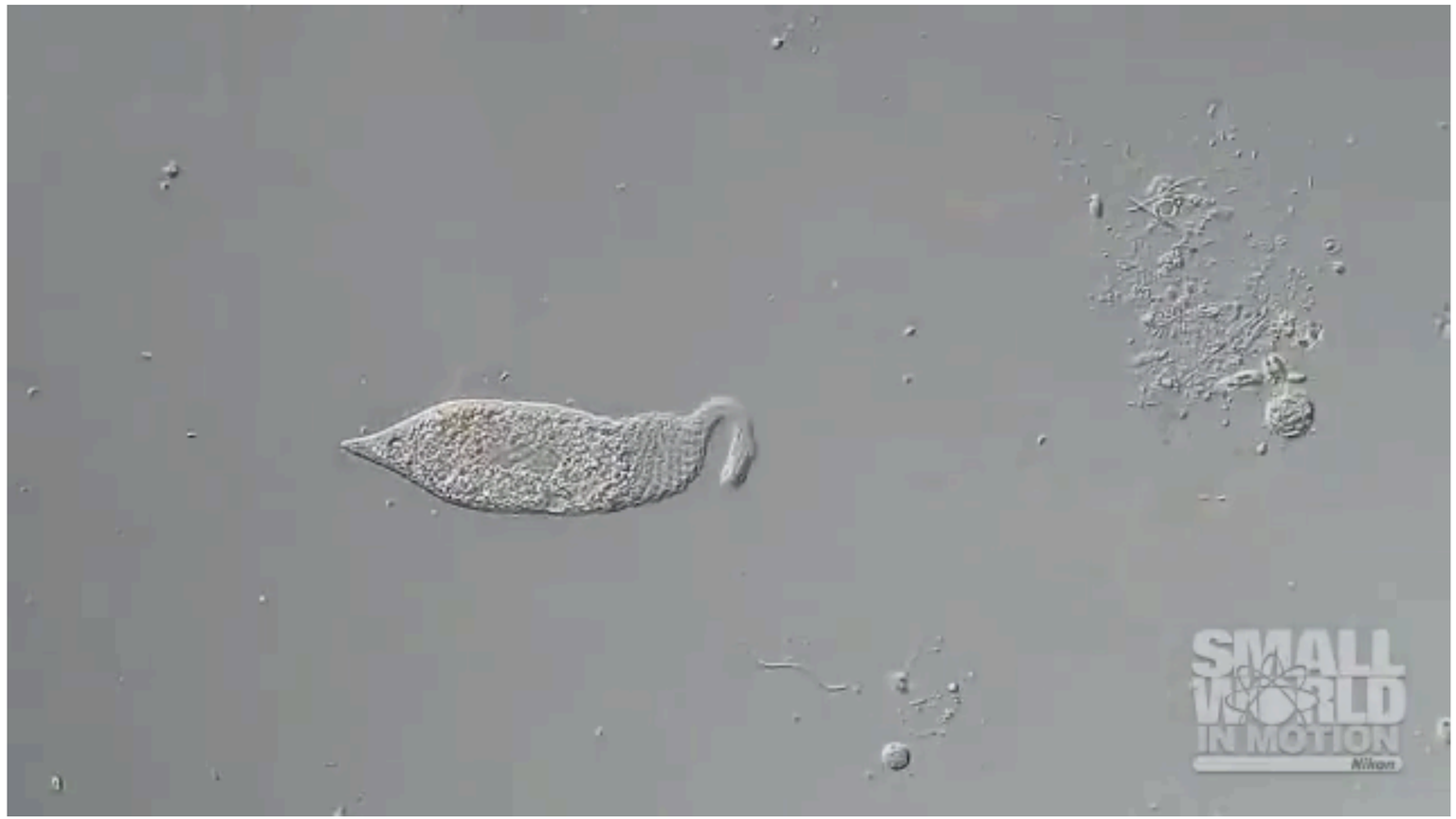
LoopSeq

- Error rate. Very low. ~5 x 10^{-5} per nucleotide
- Portability. Low
- Cost. Kit for 1600 dollars. Plus Illumina sequencing.
- The type of community you want to sequence. Complex community
- Sequence length/marker. 16S (prok). 18S-ITS1-ITS2 (fungal)

Any questions?



Any questions?



Break
5-10 min

Bioinformatic pipeline for long-read data

Bioinformatic analyses of long-read amplicon data

1. Curate reads
2. Taxonomic assignment
3. Phylogenetic analyses of environmental data

Bioinformatic analyses of long-read amplicon data

- Fewer dedicated pipelines for long-read data.

Bioinformatic analyses of long-read amplicon data

- Fewer dedicated pipelines for long-read data.
- DADA2
  - Great if you are using 16S/18S
  - Cannot handle longer reads yet (memory issues)

High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution

Benjamin J Callahan, Joan Wong, Cheryl Heiner, Steve Oh, Casey M Theriot, Ajay S Gulati, Sarah K McGill, Michael K Dougherty

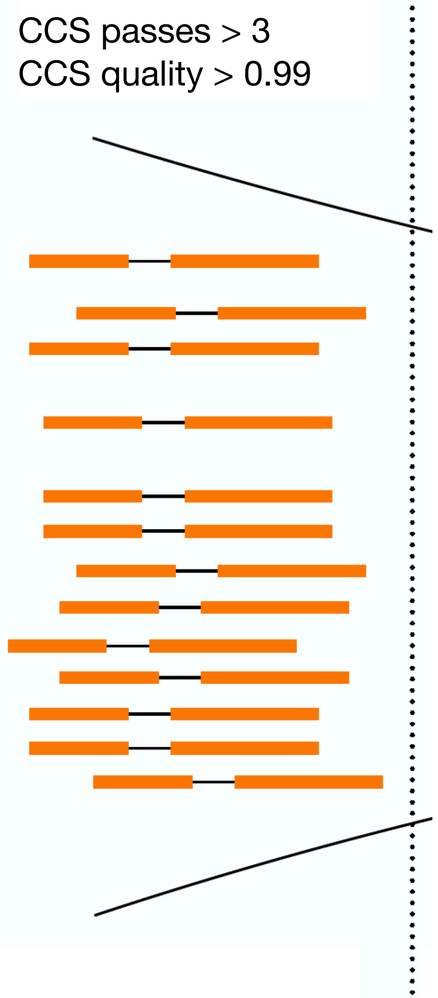
Nucleic Acids Research, Volume 47, Issue 18, 10 October 2019, Page e103, https://doi.org/10.1093/nar/gkz569
Published: 03 July 2019 **Article history**

Bioinformatic analyses of long-read amplicon data

- Fewer dedicated pipelines for long-read data.
- DADA2
  - Great if you are using 16S/18S
  - Cannot handle longer reads yet (memory issues)
- Custom pipelines
  - Depends on taxonomic group, sequencing technology, sequencing depth, length and marker(s) selected

Generate CCS

CCS passes > 3
CCS quality > 0.99

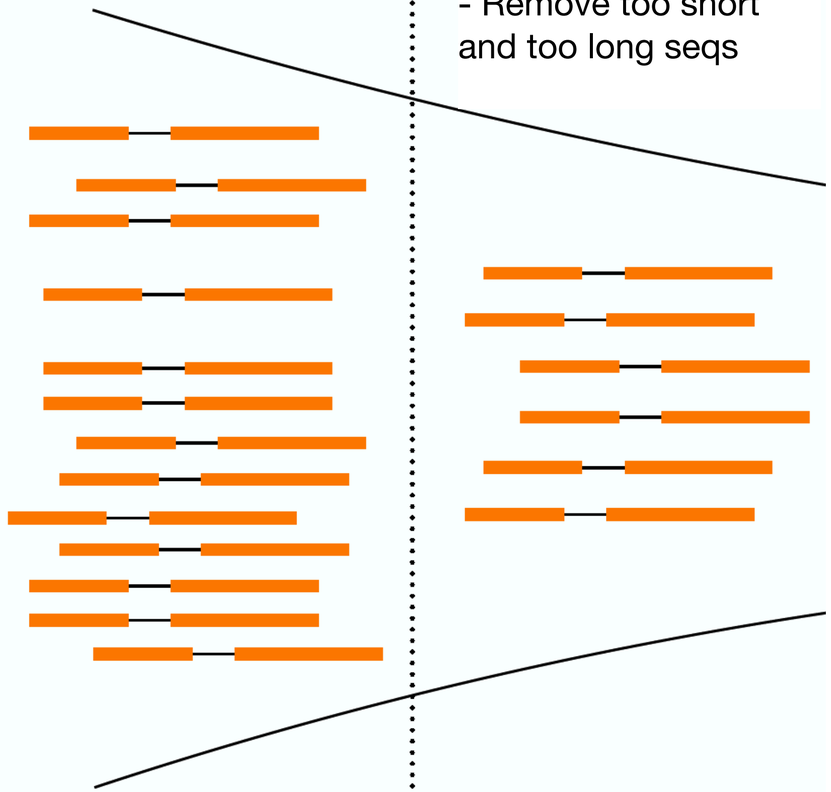


Generate CCS

CCS passes > 3
CCS quality > 0.99

Quality filter

- Both primers present
- No. of expected errors < 4
- Remove too short and too long seqs



Generate CCS

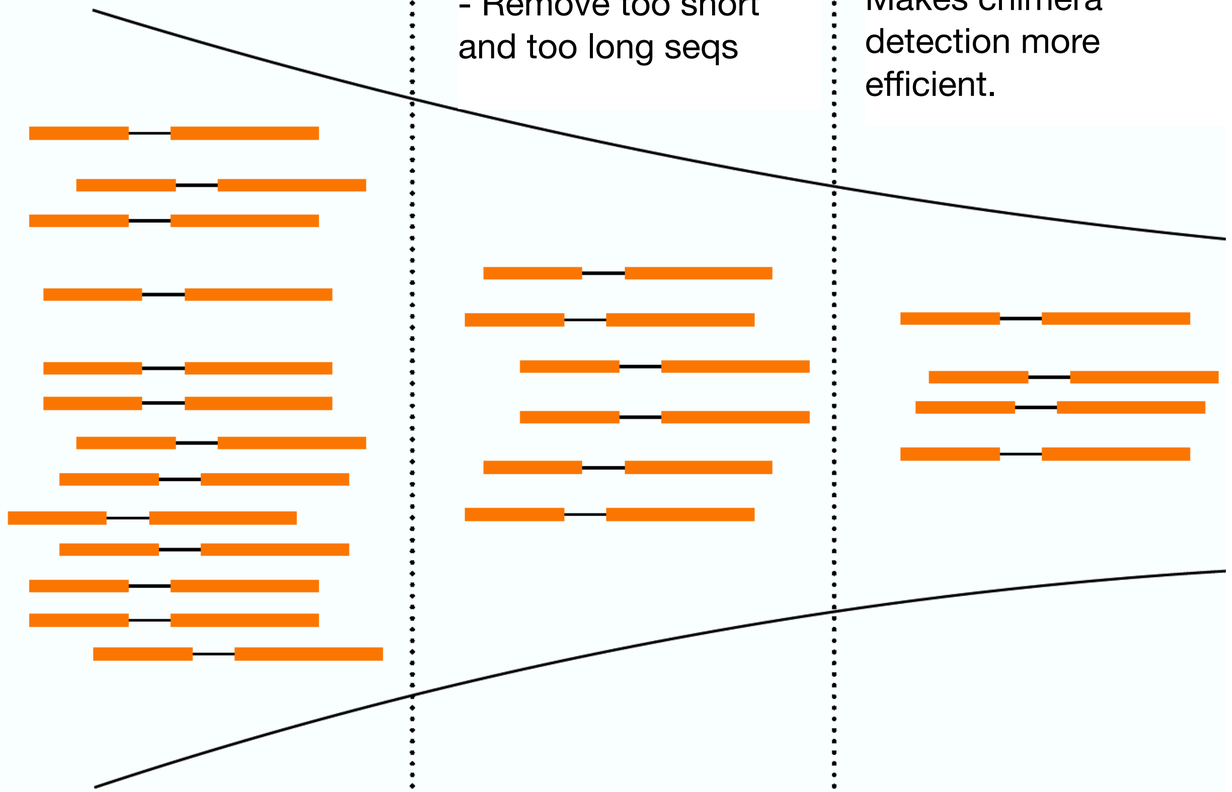
CCS passes > 3
CCS quality > 0.99

Quality filter

- Both primers present
- No. of expected errors < 4
- Remove too short and too long seqs

Precluster + de-novo chimera detection

In order to de-noise. Makes chimera detection more efficient.



Generate CCS

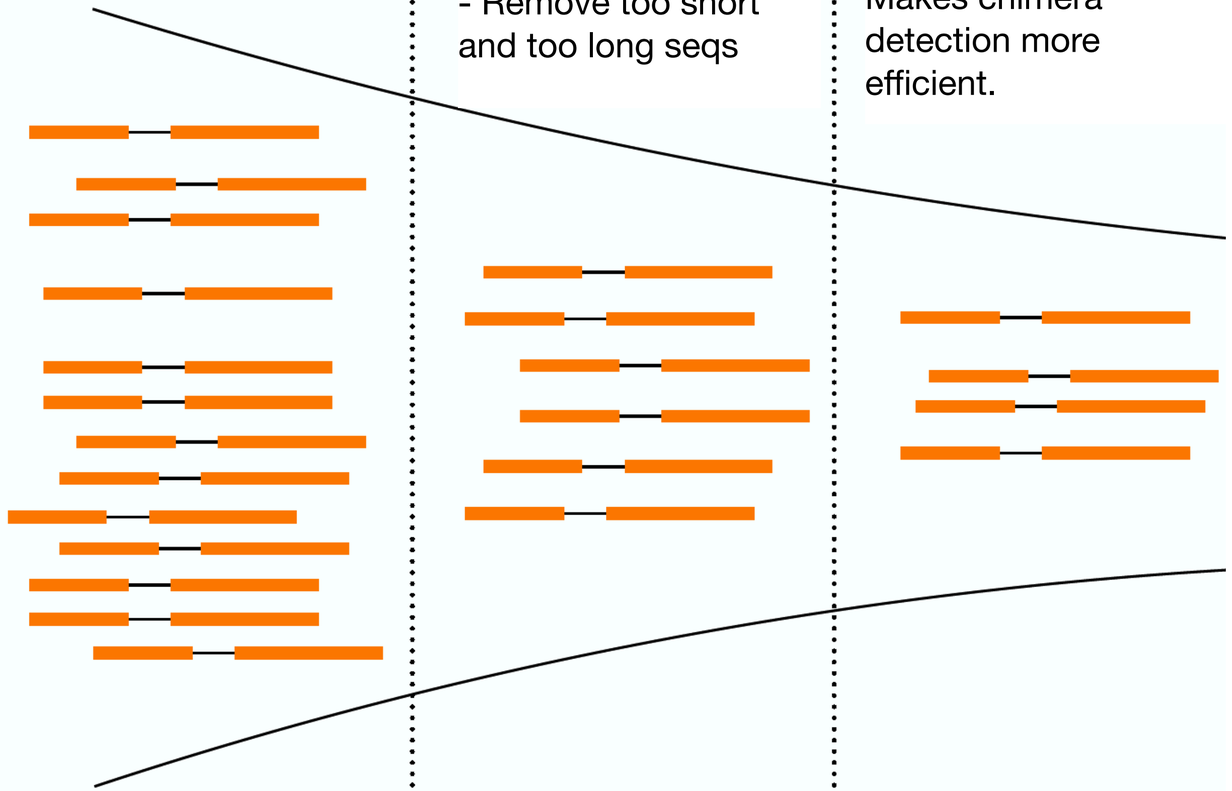
CCS passes > 3
CCS quality > 0.99

Quality filter

- Both primers present
- No. of expected errors < 4
- Remove too short and too long seqs

Precluster + de-novo chimera detection

In order to de-noise. Makes chimera detection more efficient.



Chimera detection

Query

50 bp each			

4x ‘chunks’ **Illumina**

Generate CCS

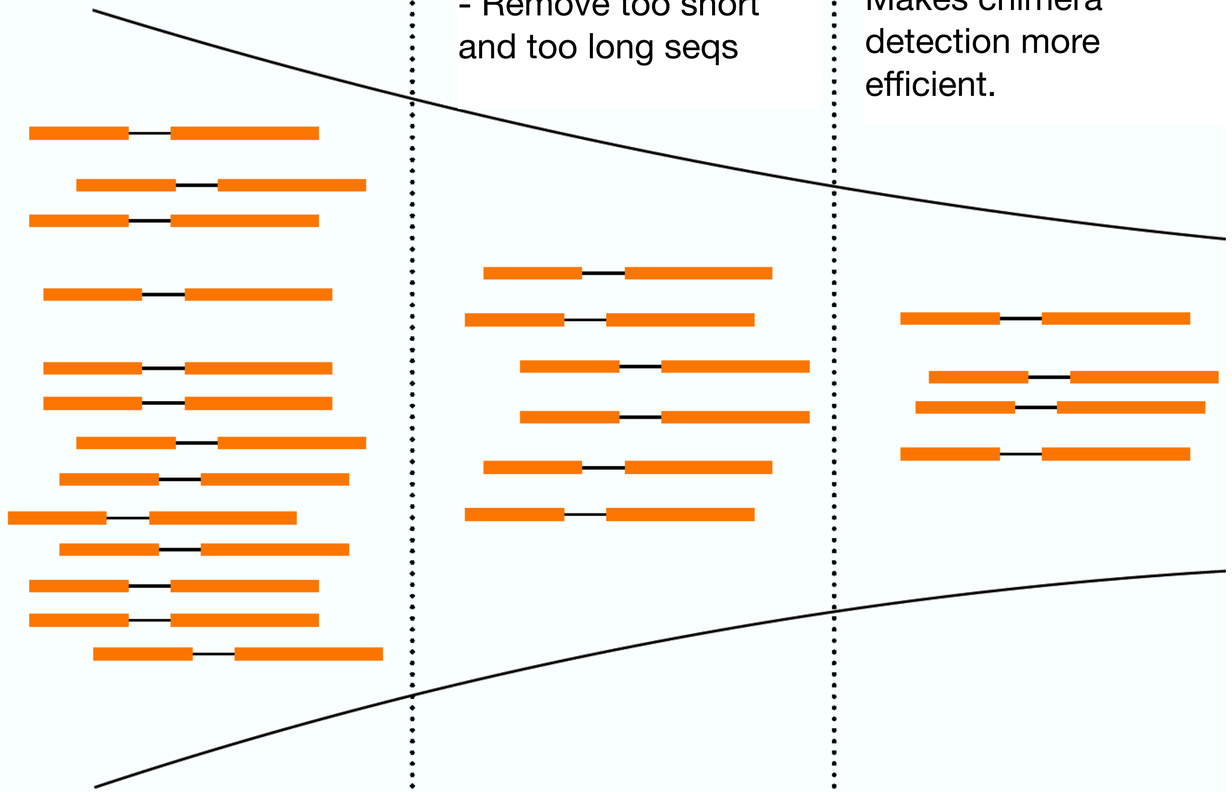
CCS passes > 3
CCS quality > 0.99

Quality filter

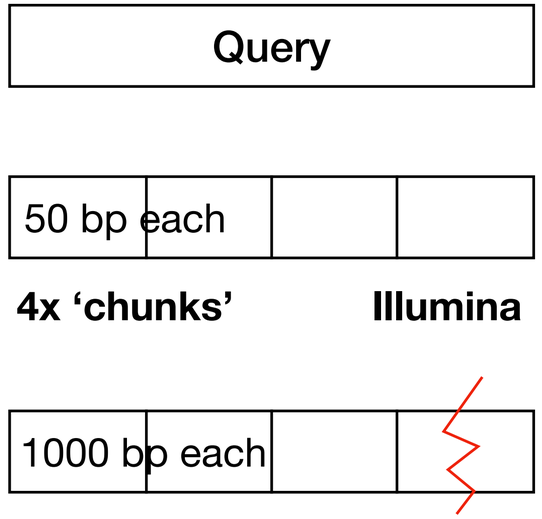
- Both primers present
- No. of expected errors < 4
- Remove too short and too long seqs

Precluster + de-novo chimera detection

In order to de-noise. Makes chimera detection more efficient.



Chimera detection



Generate CCS

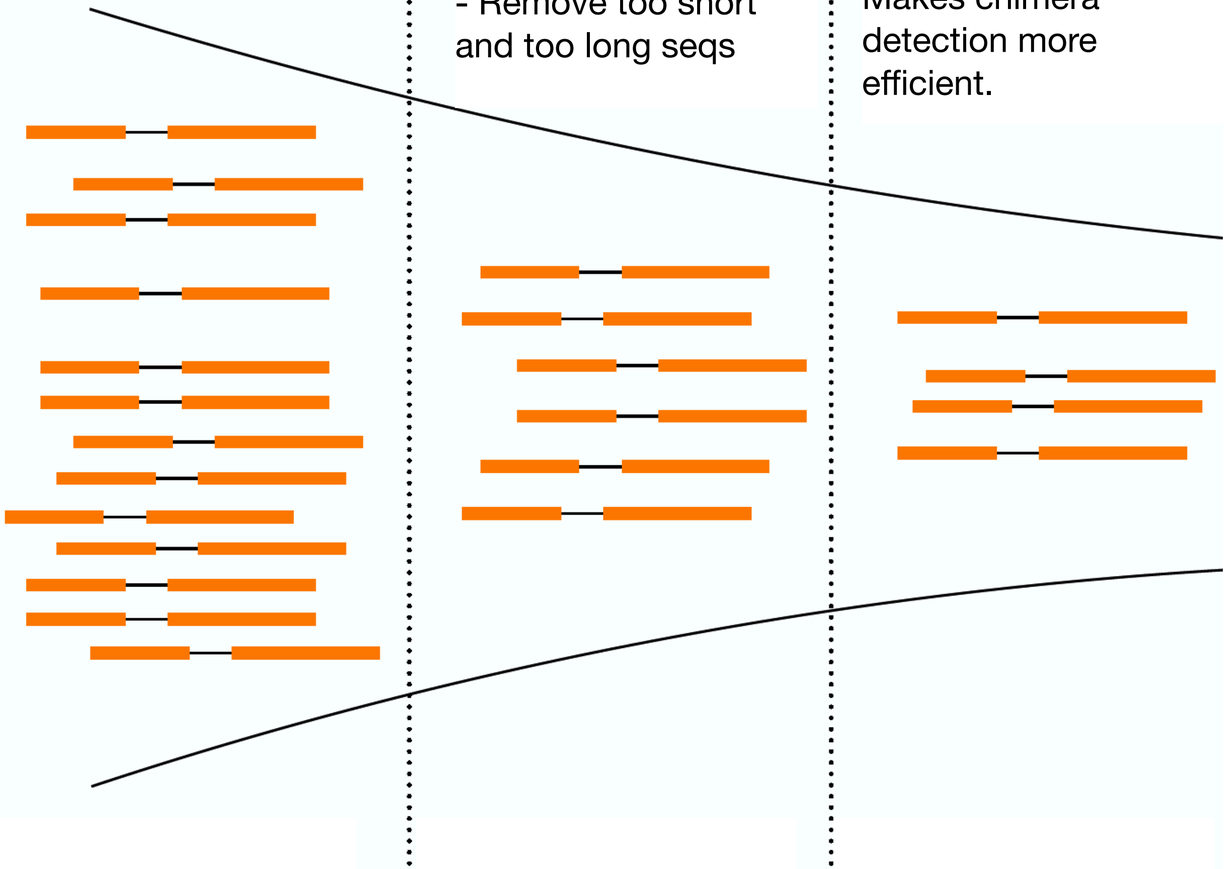
CCS passes > 3
CCS quality > 0.99

Quality filter

- Both primers present
- No. of expected errors < 4
- Remove too short and too long seqs

Precluster + de-novo chimera detection

In order to de-noise. Makes chimera detection more efficient.



Chimera detection

Query

50 bp each

4x ‘chunks’ **Illumina**

1000 bp each

4x ‘chunks’ **PacBio**

Chimera missed

20-30 ‘chunks’ **PacBio**

Chimera detected

Generate CCS

CCS passes > 3
CCS quality > 0.99

Quality filter

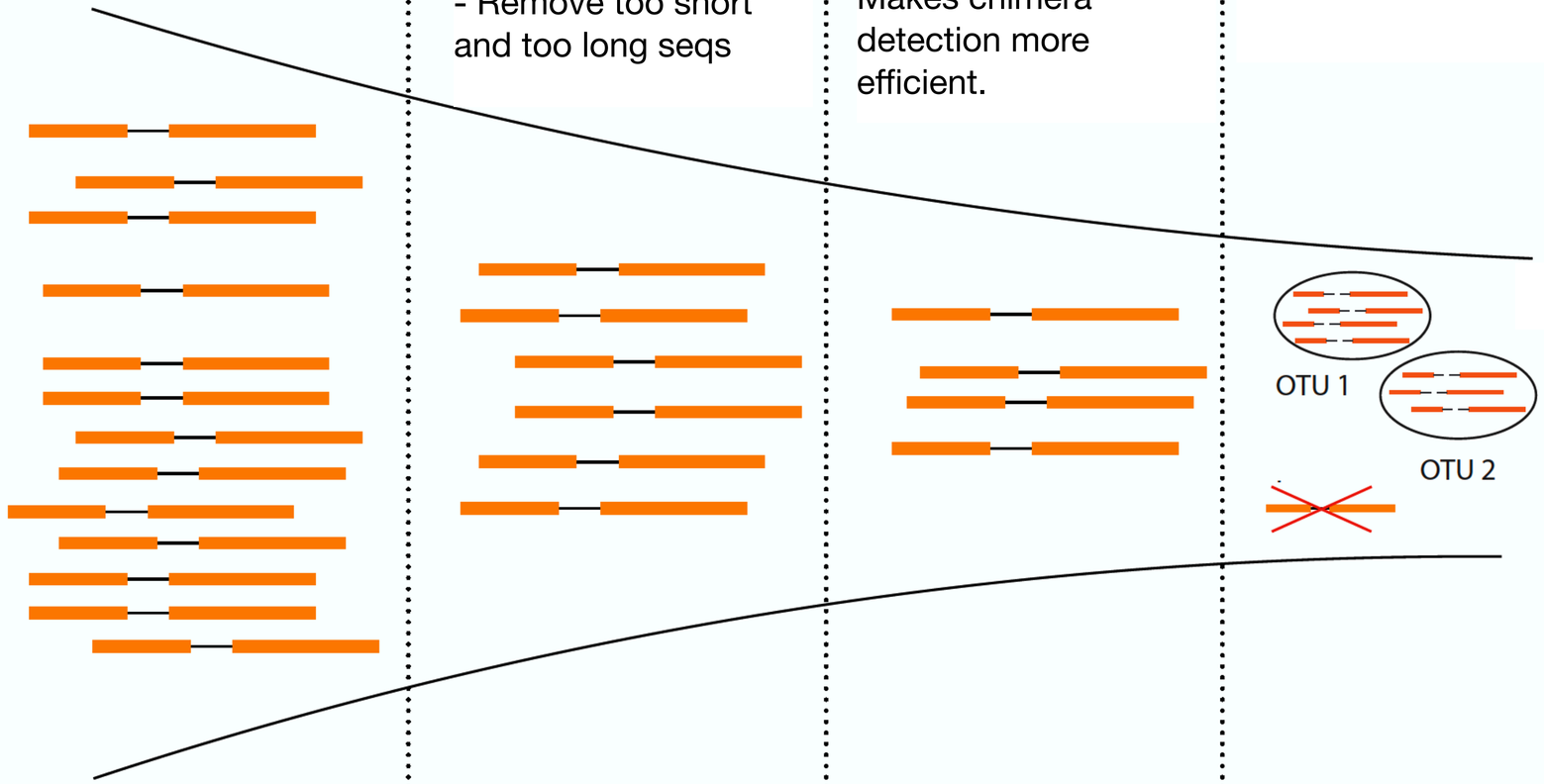
- Both primers present
- No. of expected errors < 4
- Remove too short and too long seqs

Precluster + de-novo chimera detection

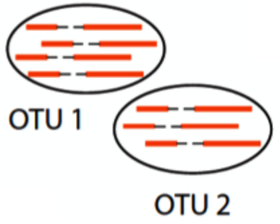
In order to de-noise. Makes chimera detection more efficient.

OTU clustering

97% similarity based on 18S gene.
Discard singletons.



Chimera detection (again)



Clean OTUs!

Any questions?

Any questions?

Investigating habitat transitions of protists: Using long-read and short-read data together

Future directions: combining with Illumina data

- Illumina = more comprehensive, good metadata

Neotropical Forest Soils

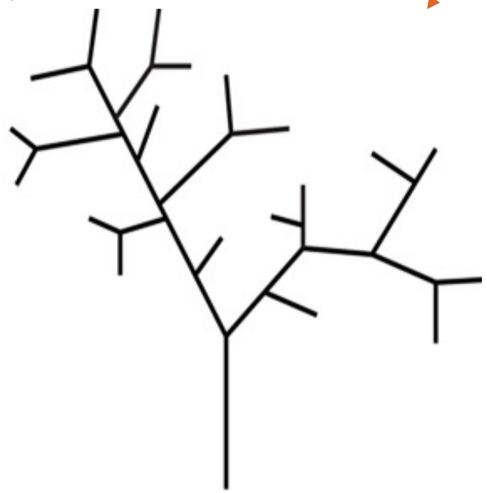
MALASPINA 2010

Global soil surveys

BioMarKs
Biodiversity of Marine euKaryoteS

earth microbiome project

Freshwater surveys

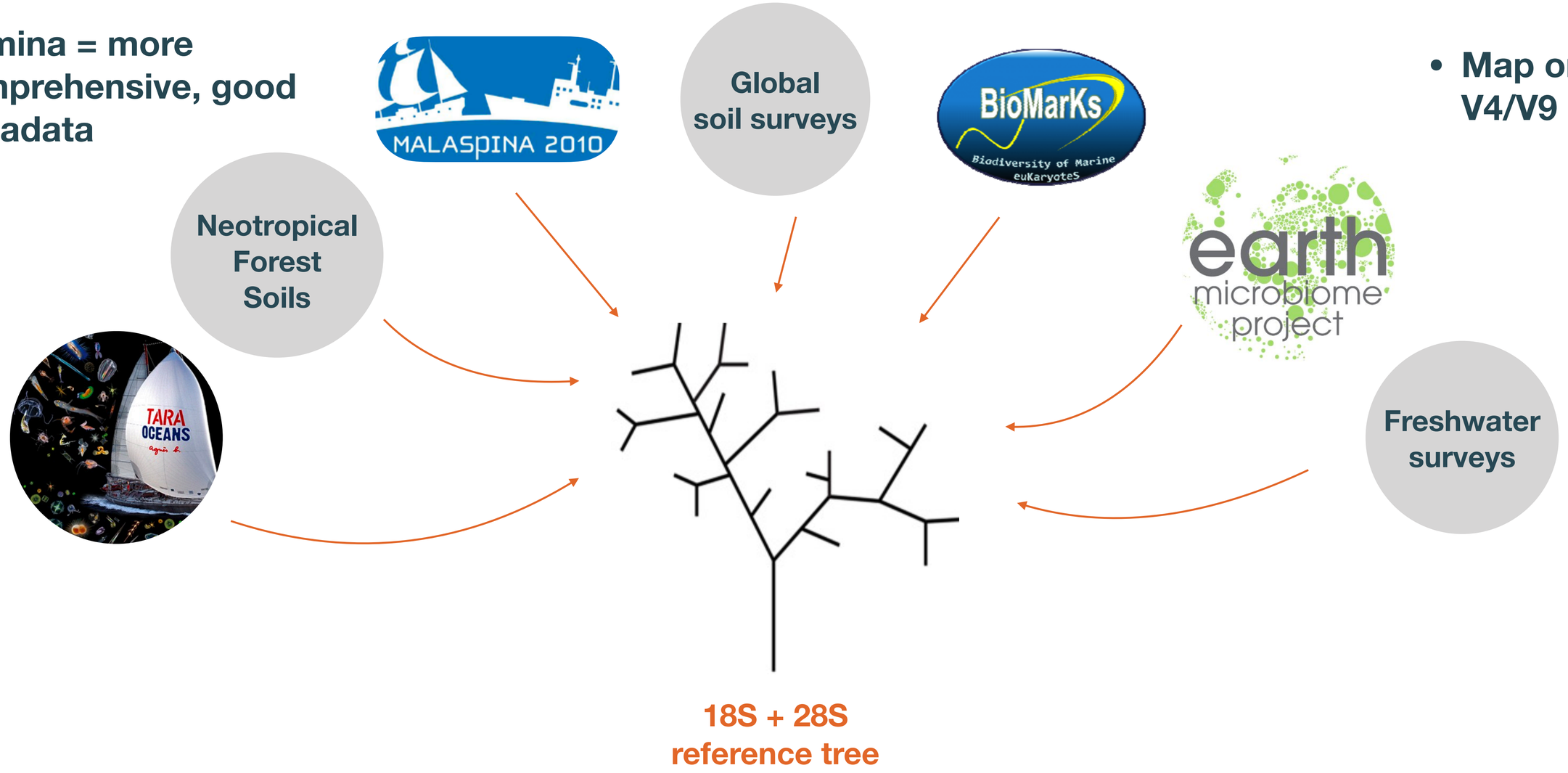


18S + 28S reference tree

- Map on tree V4/V9

Future directions: combining with Illumina data

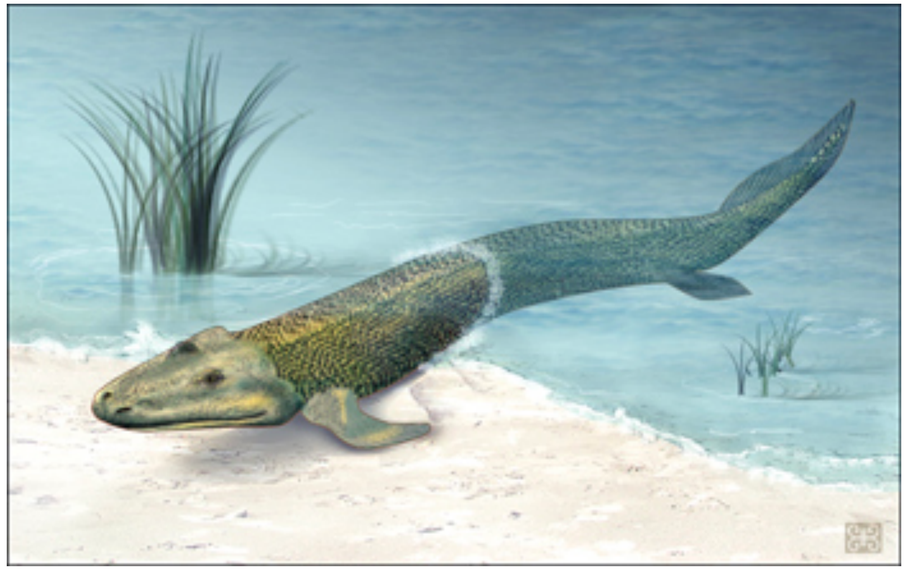
- **Illumina = more comprehensive, good metadata**



- **Map on tree V4/V9**

- Can answer questions of evolutionary nature on a really broad scale

How frequently have protists crossed the marine-non marine habitat boundary?



Credit: *Zina Deretsky, National Science Foundation*

How frequently have protists crossed the salt barrier?

Review

Cell PRESS

Infrequent marine–freshwater transitions in the microbial world

Ramiro Logares[1], Jon Bråte[2], Stefan Bertilsson[1], Jessica L. Clasen[3], Kamran Shalchian-Tabrizi[2] and Karin Rengefors[4]

Extensive dinoflagellate phylogenies indicate infrequent marine–freshwater transitions

Ramiro Logares [a,*], Kamran Shalchian-Tabrizi [b], Andrés Boltovskoy [c], Karin Rengefors [a]

Genetic diversity of goniomonads: an ancient divergence between marine and freshwater species

Sophie von der Heyden , Ema Chao & Thomas Cavalier-Smith

How frequently have protists crossed the salt barrier?

Review

Cell PRESS

Infrequent marine–freshwater transitions in the microbial world

Ramiro Logares[1], Jon Bråte[2], Stefan Bertilsson[1], Jessica L. Clasen[3], Kamran Shalchian-Tabrizi[2] and Karin Rengefors[4]

Extensive dinoflagellate phylogenies indicate infrequent marine–freshwater transitions

Ramiro Logares [a,*], Kamran Shalchian-Tabrizi [b], Andrés Boltovskoy [c], Karin Rengef

They are young, and they are many: dating freshwater lineages in unicellular dinophytes

Anže Žerdoner Čalasan, Juliane Kretschmann, Marc Gottschling

First published: 01 August 2019 | https://doi.org/10.1111/1462-2920.14766 | Citations: 6

Article

Tracing the Origin of Planktonic Protists in an Ancient Lake

Nataliia V. Annenkova [1,*], Caterina R. Giner [2,3] and Ramiro Logares [2,*]

Genetic diversity of goniomonads: an ancient divergence between marine and fresh species

Sophie von der Heyden , Ema Chao & Thomas Cavalier-Sn

Closely related dinoflagellate species in vastly different habitats – an example of a marine–freshwater transition

Nataliia V Annenkova [a], Gert Hansen [b] and Karin Rengefors [c]

[a]Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia; [b]Marine Biological Section, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [c]Aquatic Ecology, Department of Biology, Lund University, Lund, Sweden

How frequently have protists crossed the salt barrier?

Review

Cell PRESS

Infrequent marine–freshwater transitions in the microbial world

Ramiro Logares[1], Jon Bråte[2], Stefan Bertilsson[1], Jessica L. Clasen[3], Kamran Shalchian-Tabrizi[2] and Karin Rengefors[4]

They are young, and they are many: dating freshwater lineages in unicellular dinophytes

Anže Žerdoner Čalasan, Juliane Kretschmann, Marc Gottschling

First published: 01 August 2019 | https://doi.org/10.1111/1462-2920.14766 | Citations: 6

Mostly limited in taxonomic scope

Investigate marine-terrestrial transitions across the eukaryotic tree of life

Are transitions in direction more likely than the other?

Extensive dinoflagellate phylogenies indicate infrequent marine–freshwater transitions

Ramiro Logares [a,*], Kamran Shalchian-Tabrizi [b], Andrés Boltovskoy, Sergio Rossi

Article

Tracing the Origin of Planktonic Protists in an Ancient Lake

Nataliia V. Annenkova [1,*], Catarina P. Ginger [2,3] and Ramiro Logares [2,*]

Genetic diversity of goniomonads: an ancient divergence between marine and fresh species

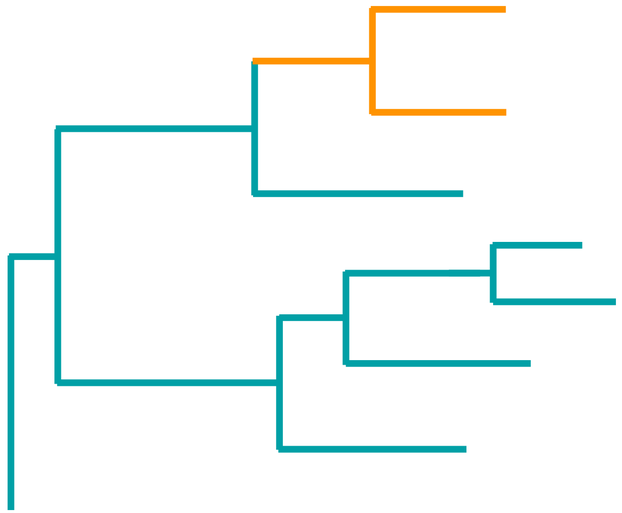
Sophie von der Heyden , Ema Chao & Thomas Cavalier-Sn

Closely related dinoflagellate species in vastly different habitats – an example of a marine–freshwater transition

Nataliia V Annenkova [a], Gert Hansen [b] and Karin Rengefors [c]

[a]Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia; [b]Marine Biological Section, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [c]Aquatic Ecology, Department of Biology, Lund University, Lund, Sweden

Investigating transitions using phylogenies



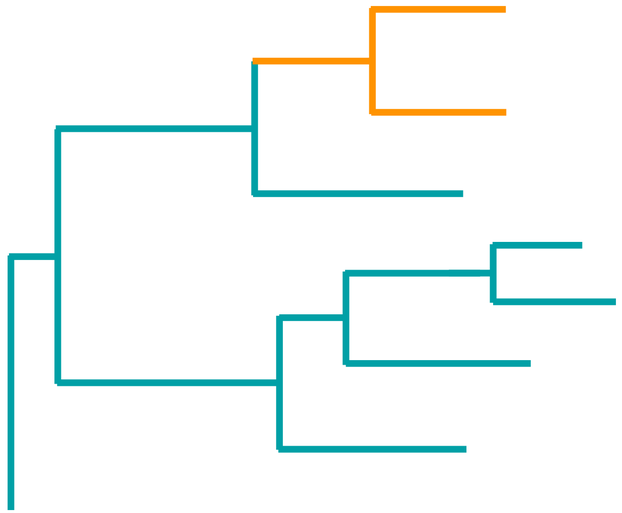
OR

SINGLE TRANSITION

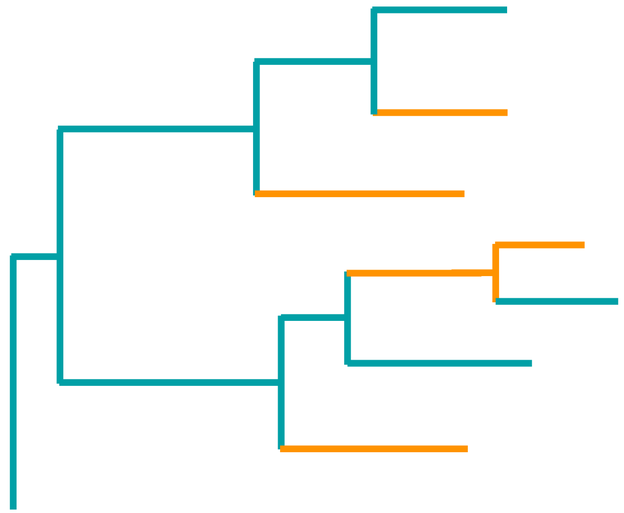
Env 1

Env 2

Investigating transitions using phylogenies



OR



SINGLE TRANSITION

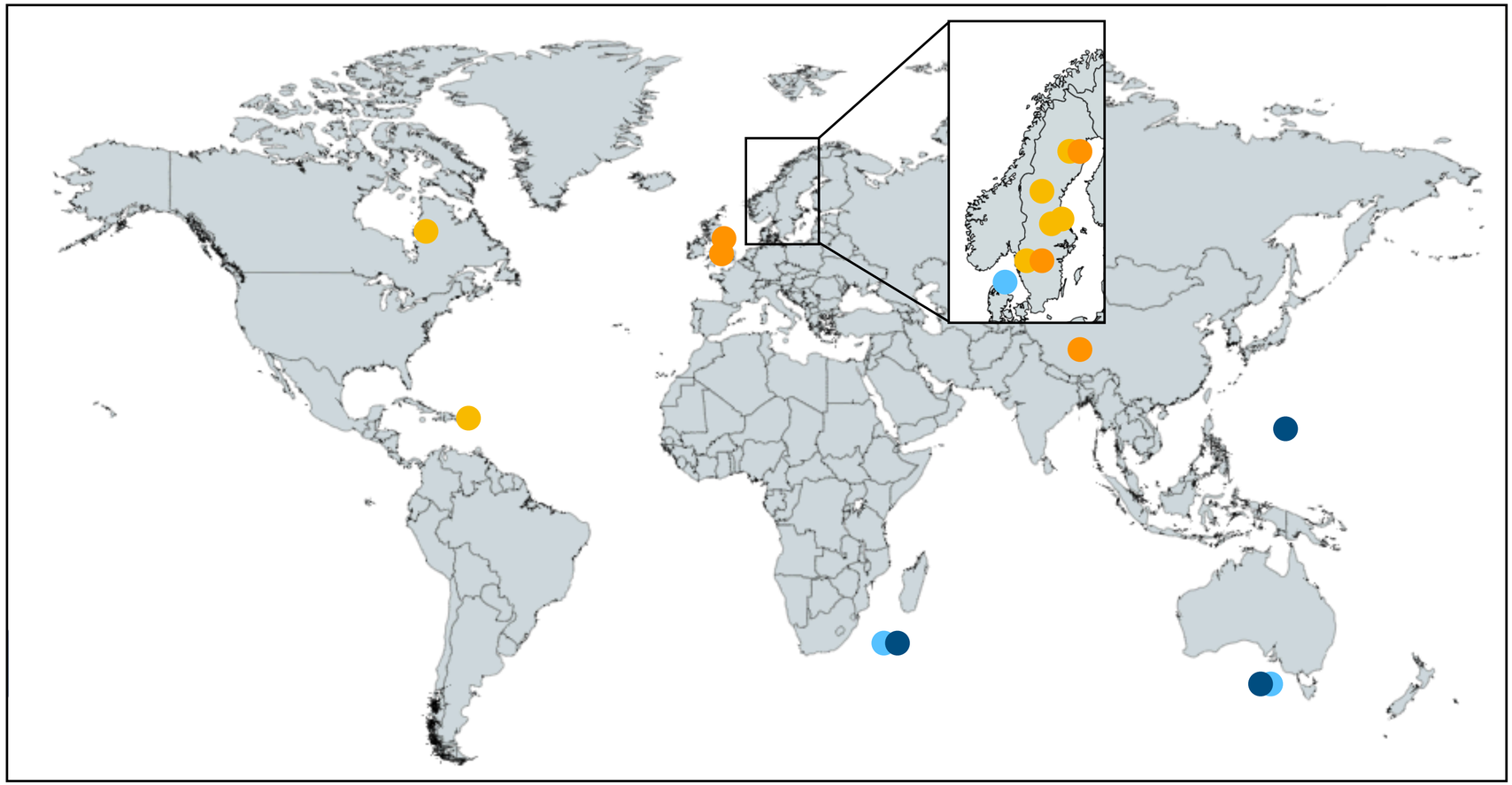
MULTIPLE TRANSITIONS

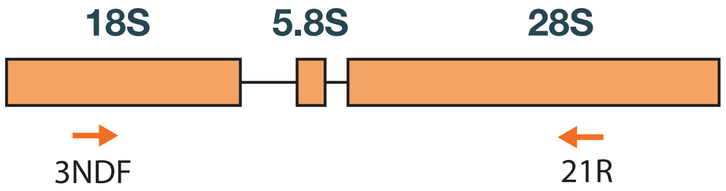
Env 1

Env 2

Study design: Generating a dense 18S-28S eukaryotic dataset

- Soil samples (x7)
- Freshwater samples (x5)
- Marine euphotic (x4)
- Marine aphotic (x5)



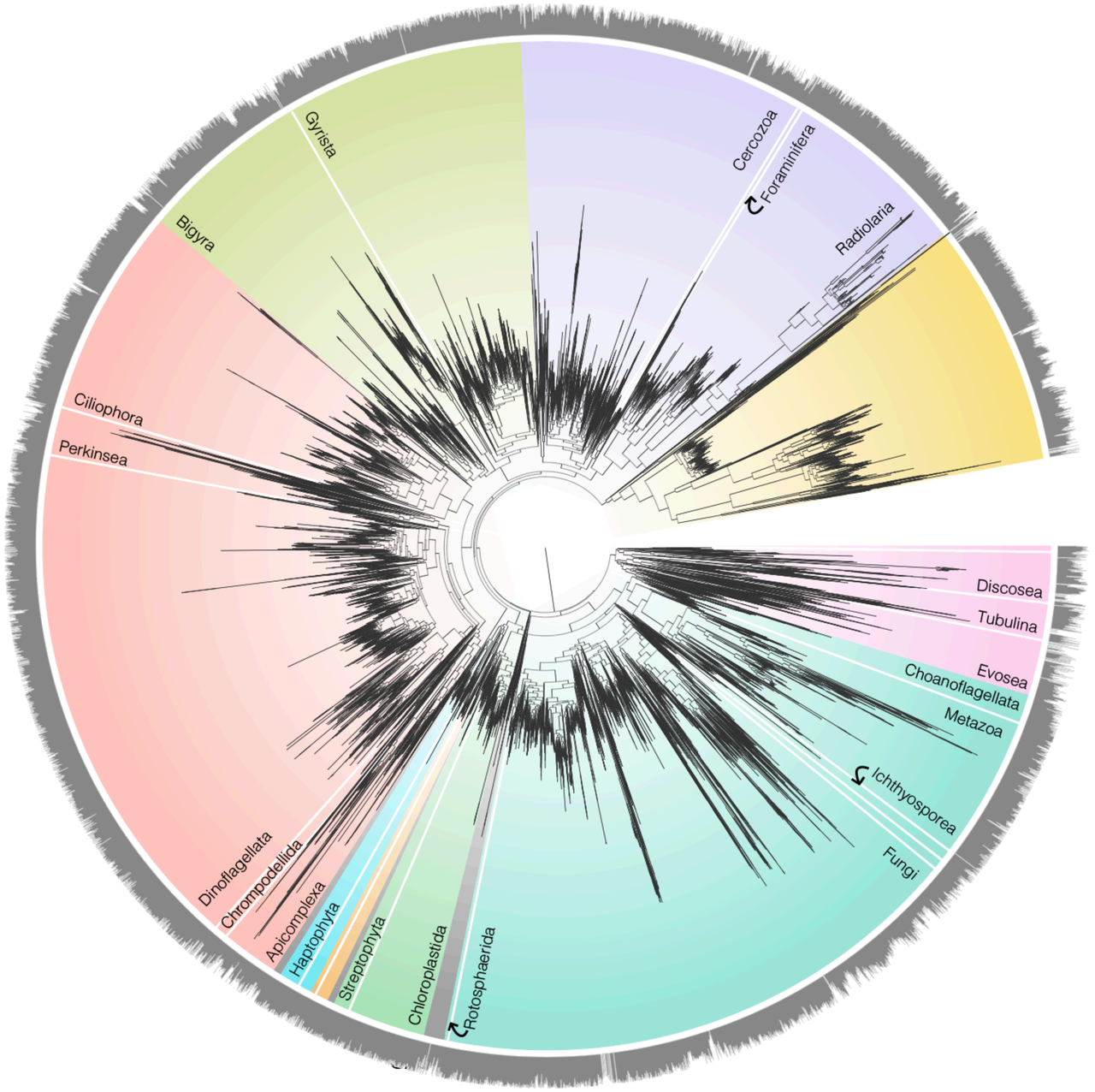


21 samples in total sequenced with PacBio Sequel II

mapchart.net

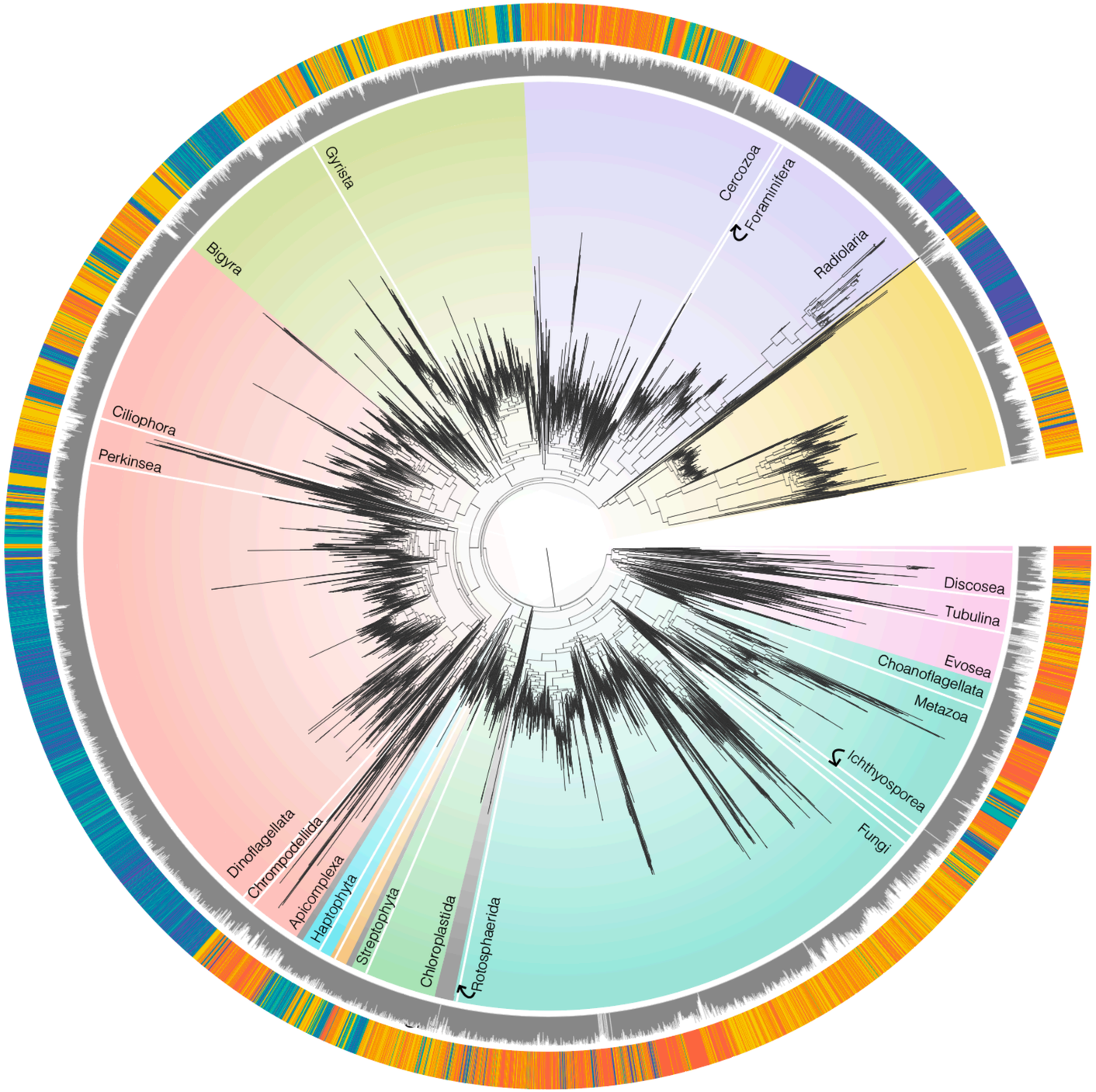
18S+28S tree

16,821 OTUs
7,160 sites



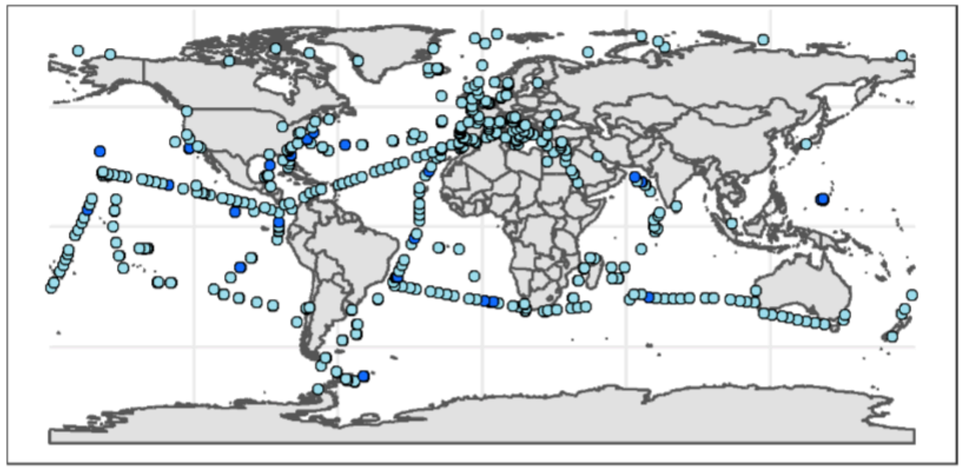
18S+28S tree

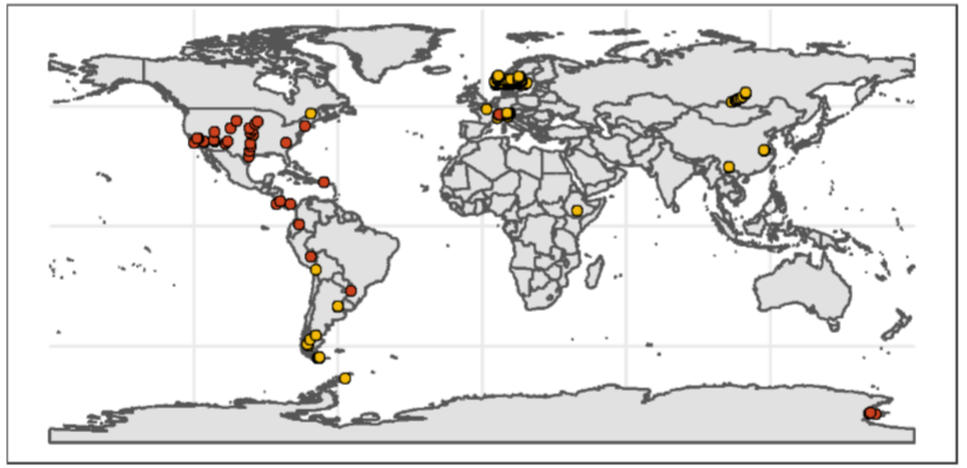
16,821 OTUs
7,160 sites



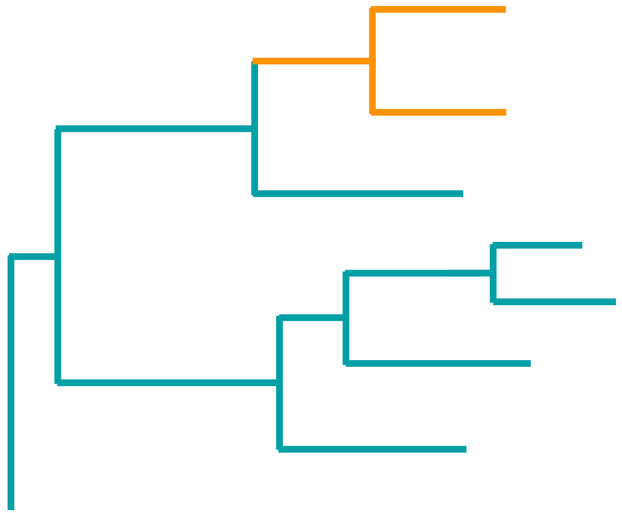
Short-read dataset

- Incorporate available short-read data!
- Clustered into OTUs at 97% similarity





Investigating transitions using phylogenies



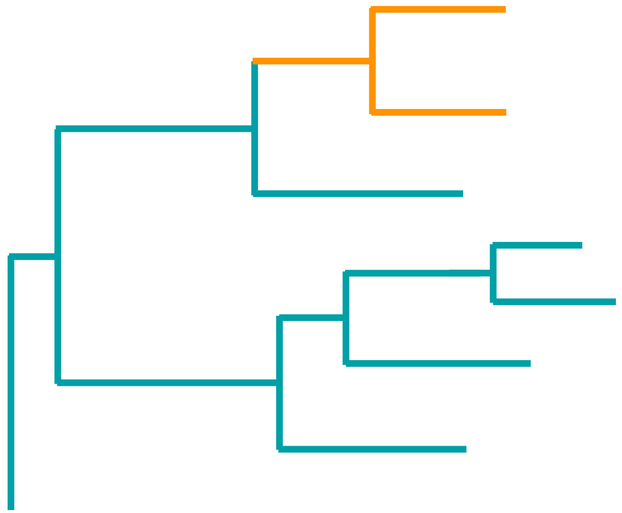
OR

SINGLE TRANSITION

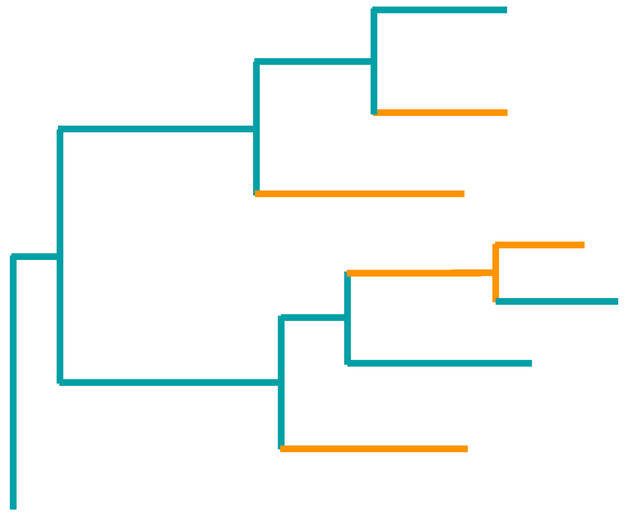
Env 1

Env 2

Investigating transitions using phylogenies



OR



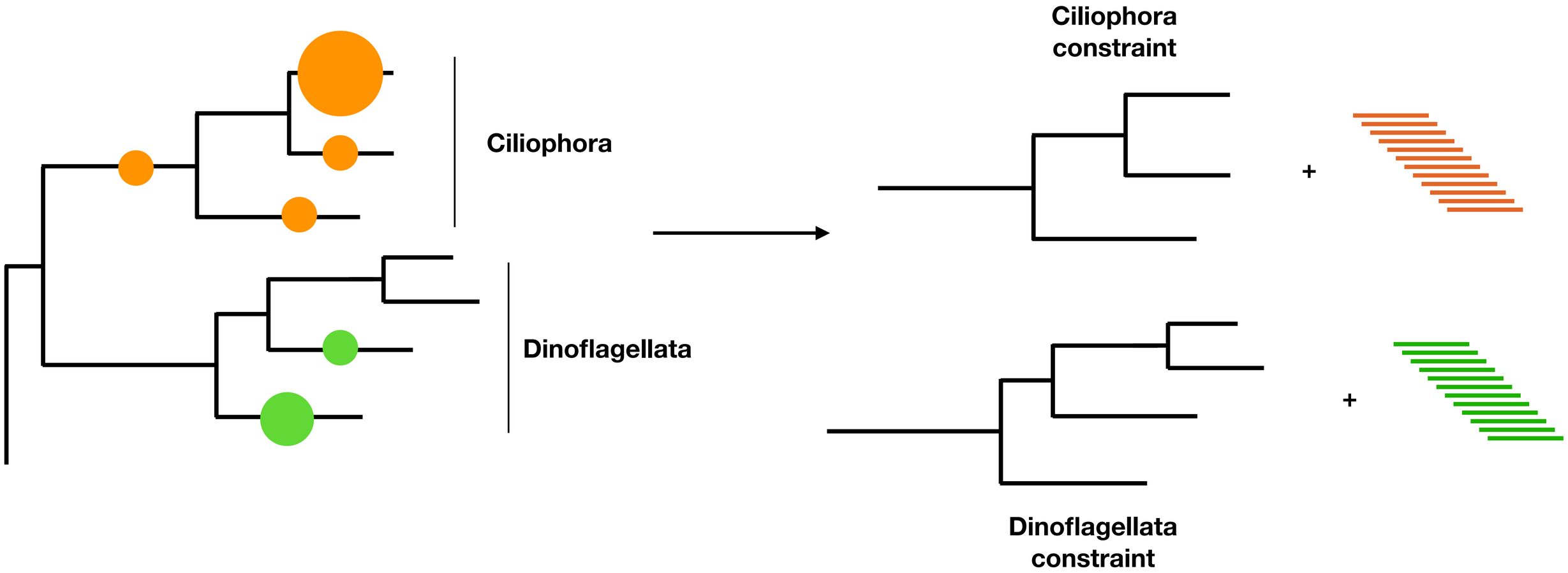
SINGLE TRANSITION

MULTIPLE TRANSITIONS

Env 1

Env 2

Extract clades and infer trees with short reads



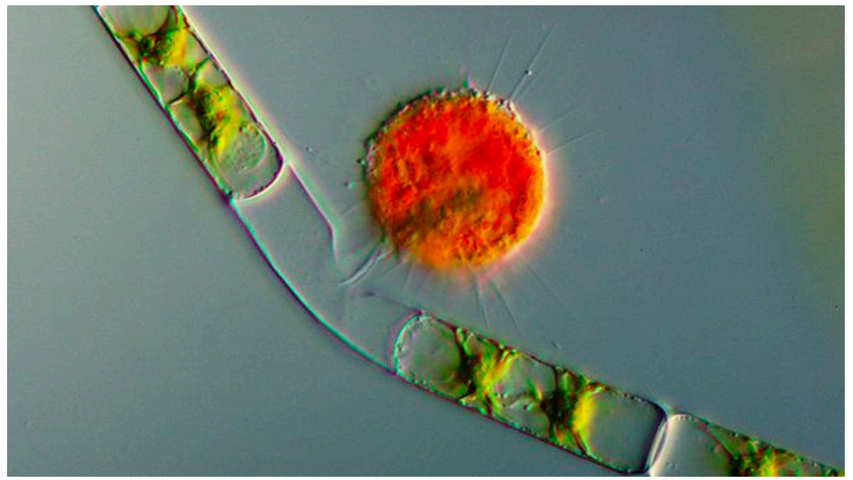
- Extract short-reads for each clade
- Build trees while using PacBio tree as constraint

Short reads data “catches” transitions missed by PacBio data alone

Vampyrellids (Cercozoa)

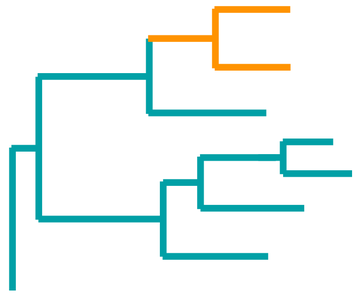
Orange tips = terrestrial, PacBio + short

Blue tips = marine, Illumina only

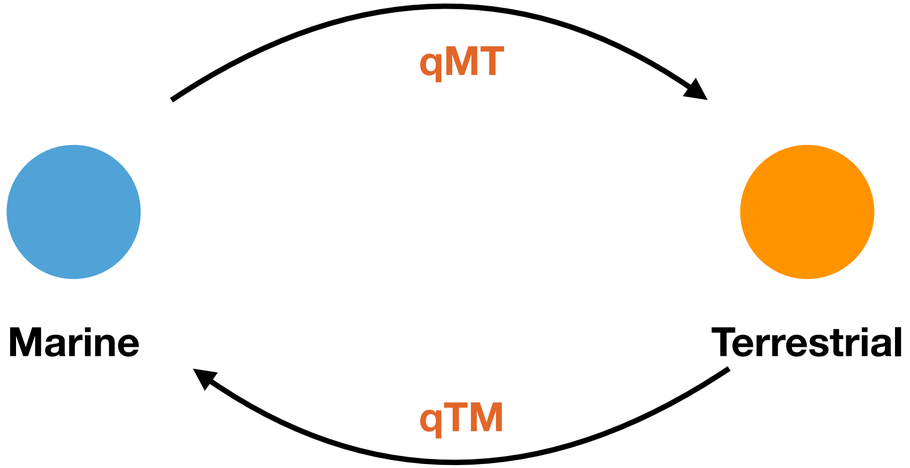


Ancestral state reconstruction

Modelling trait evolution

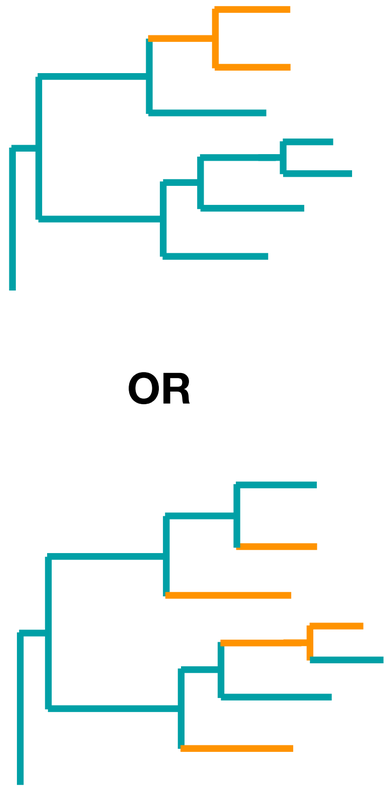


OR

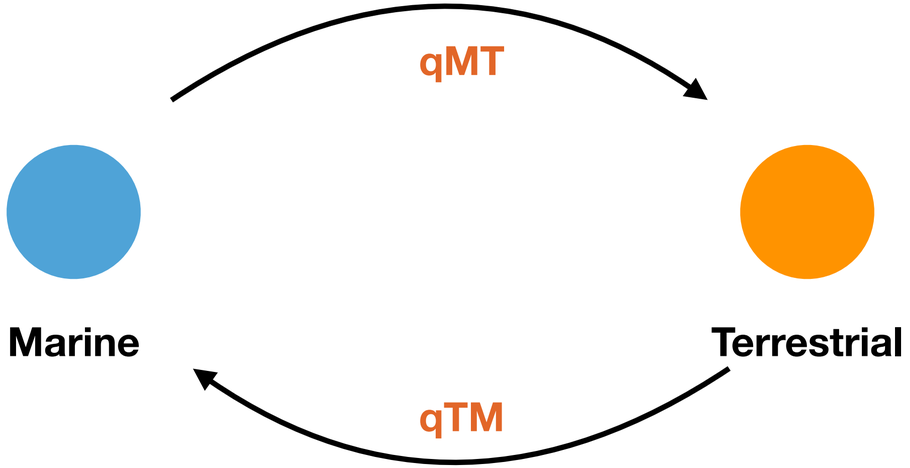


Ancestral state reconstruction

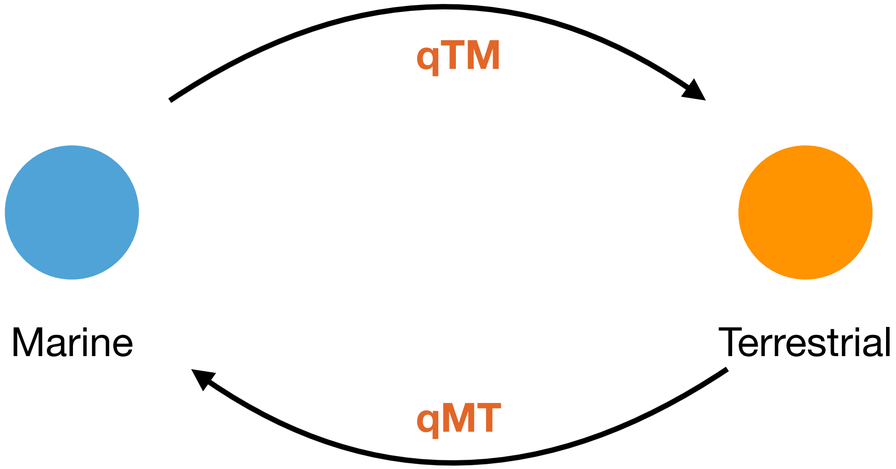
Modelling trait evolution



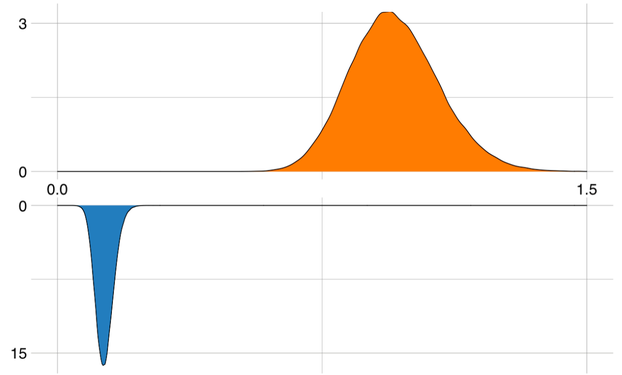
OR



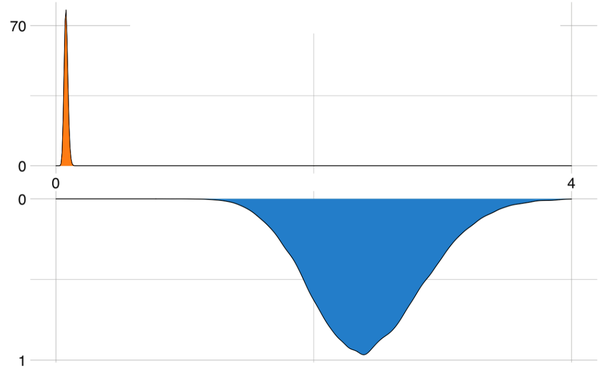
Different eukaryotic clades have different transition patterns and rates



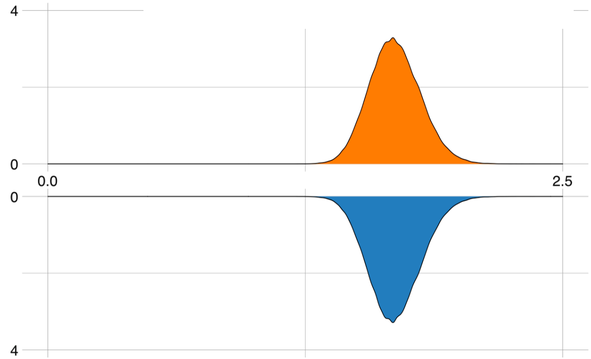
Cercozoa



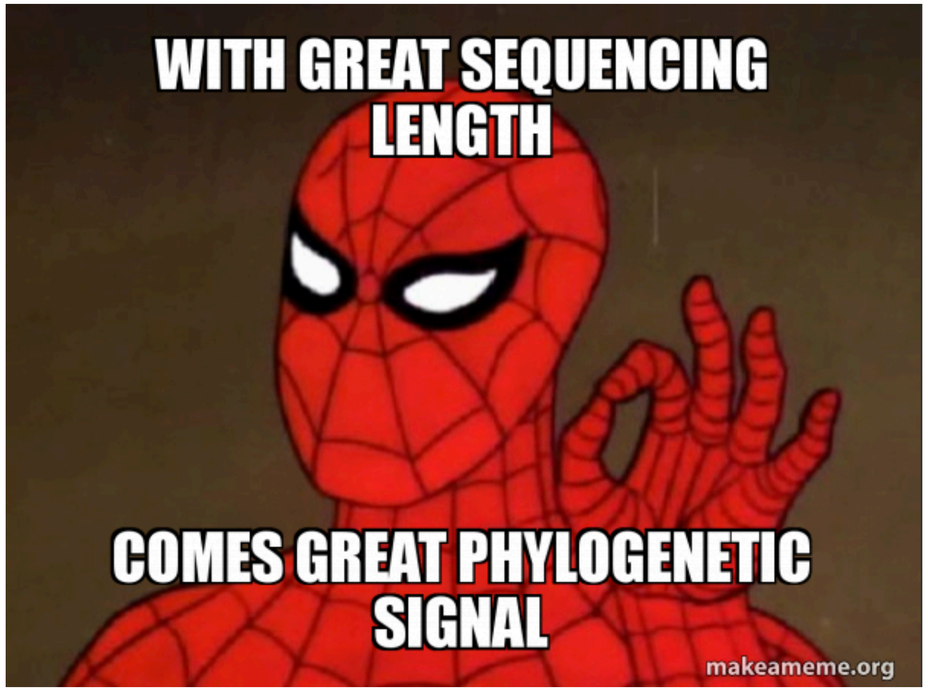
Dinoflagellata



Gyrista



Take-home



Use long-read metabarcoding to enable more phylogenetic analyses.

Use together with short-read data to get the best of both worlds!