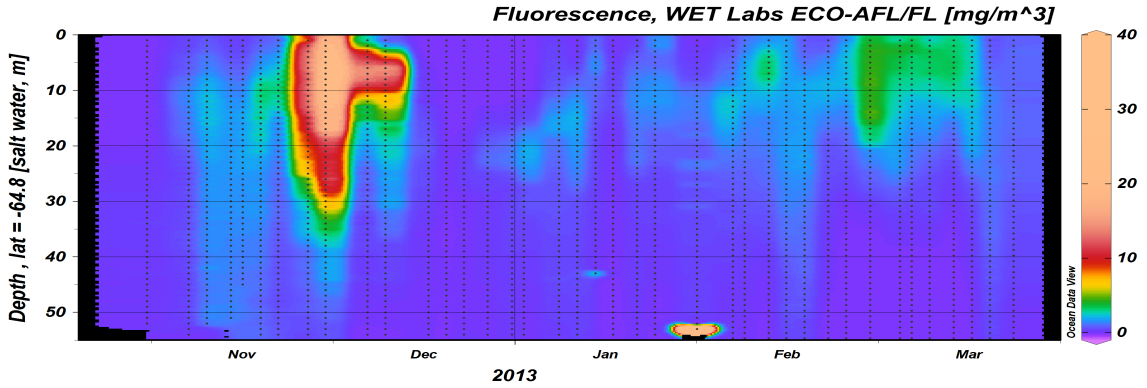
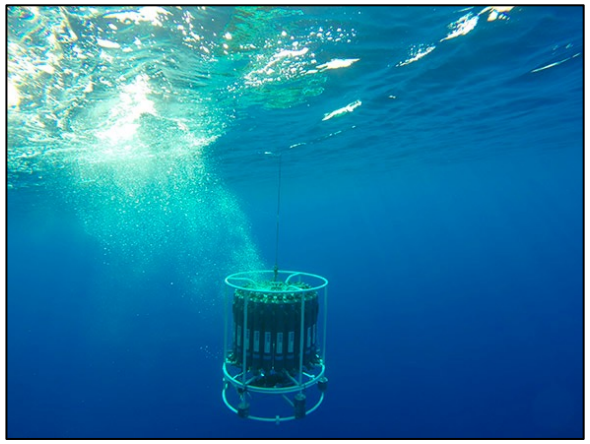
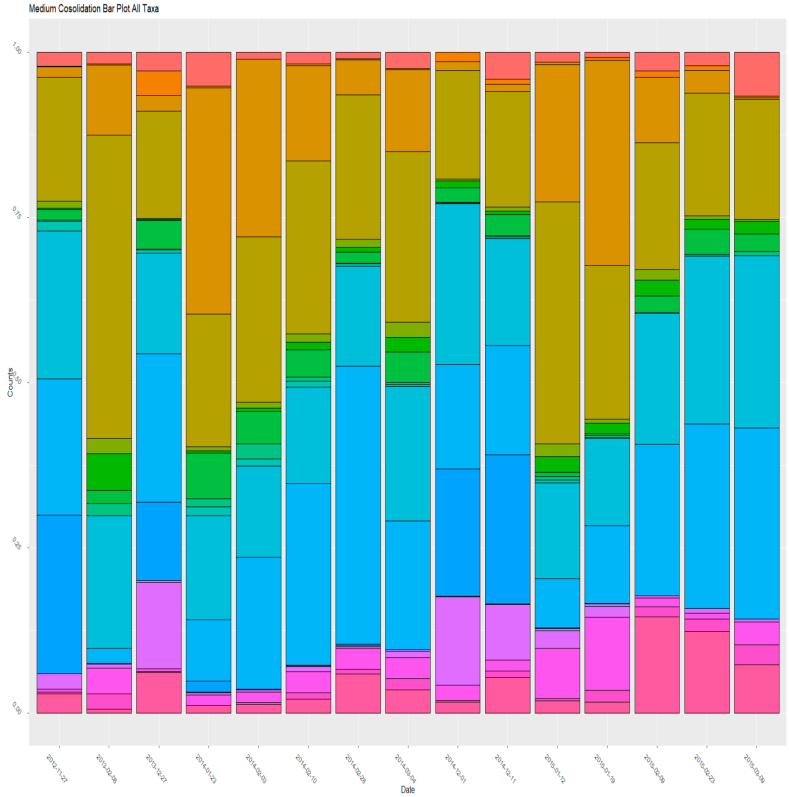
Opening Thought Question – 2/14

Why was the discovery of UCYN-A so important for understanding productivity (i.e. adding biomass) in the ocean?

Metagenomics

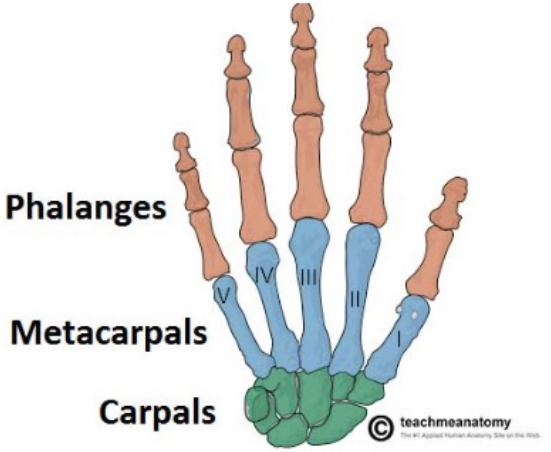






“Meta” means “Beyond”

- Metaphysics: Aristotle’s notes on mystical subjects, placed after his notes on physics
- Metacarpal: The bones beyond the carpals
- Metadata: Data the describes the central data
- Metamorphosis: Beyond the original body plan

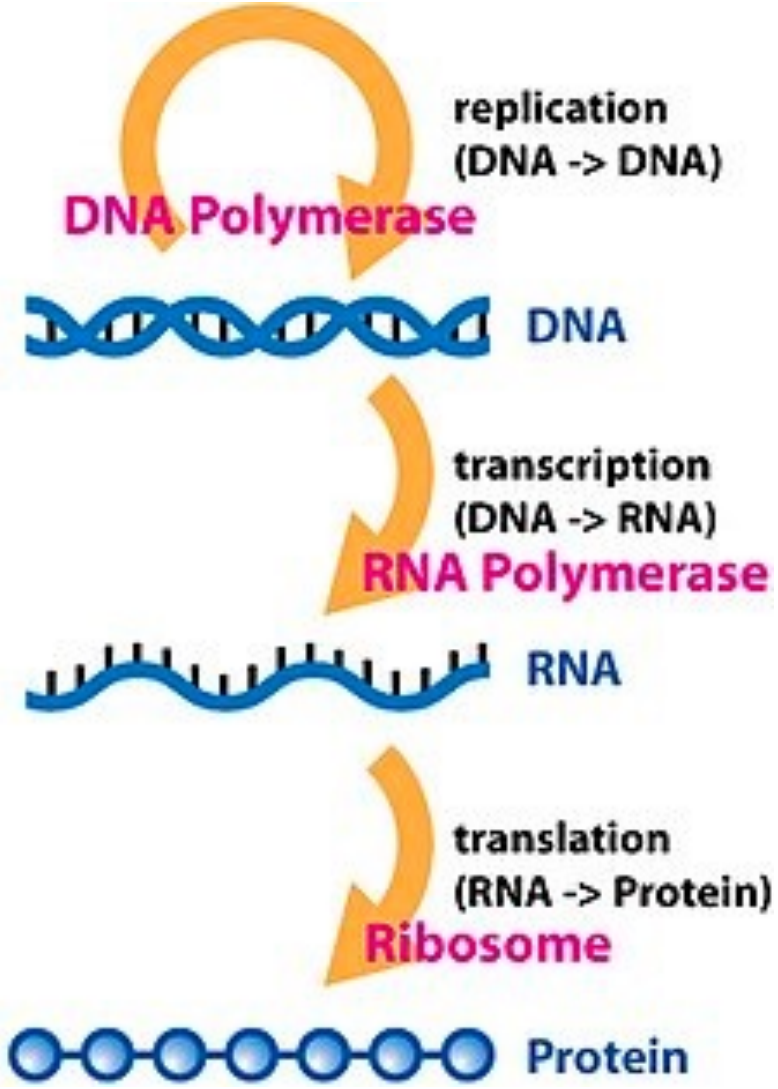


Meta*omics: beyond *omics of an individual organism or species

- Genomics → Metagenomics:
  - Genomics of a (microbial) community
- Transcriptomics → Metatranscriptomics
  - What genes are transcribed (expressed)?
  - Sometimes requires time series, especially for cyanobacteria
- Proteomics → Metaproteomics
  - What proteins are produced?

Meta*omics: beyond *omics of an individual organism or species

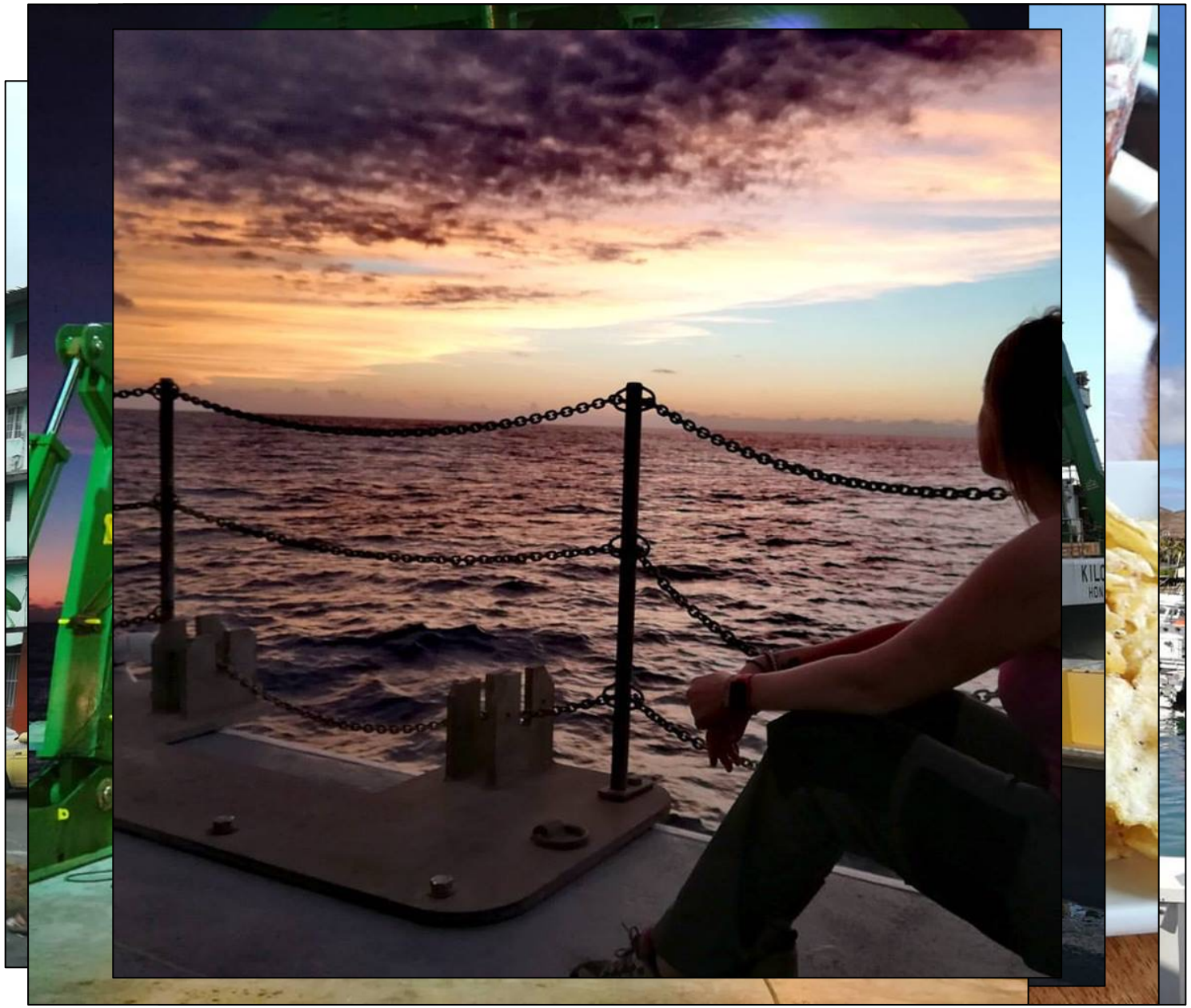
- Genomics → Metagenomics:
  - Genomics of a (microbial) community
- Transcriptomics → Metatranscriptomics
  - What genes are transcribed (expressed)?
  - Sometimes requires time series, especially for cyanobacteria
- Proteomics → Metaproteomics
  - What proteins are produced?



The steps of a metagenomic pipeline

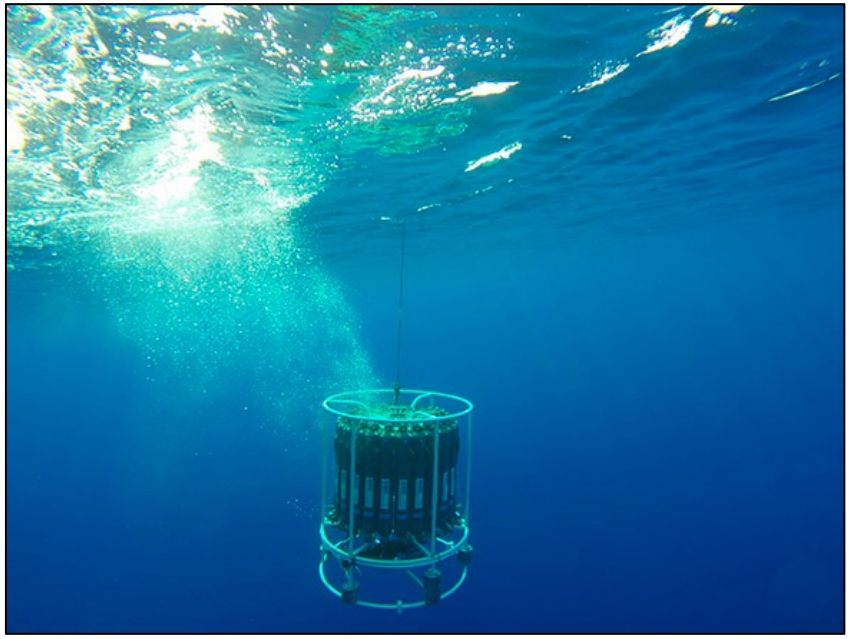
1) Collect
2) Sequence
3) Merge
4) Quality trim
5) Identify
6) Analyze

Collecting: My colleague's recent Facebook photos



Collecting: CTD

- C = Conductivity
  - Water doesn't conduct electricity, salt water does
  - Proxy for salinity
- T = Temperature
- D = Depth
- Bottles are open at top & bottom, programmed to close at desired depth

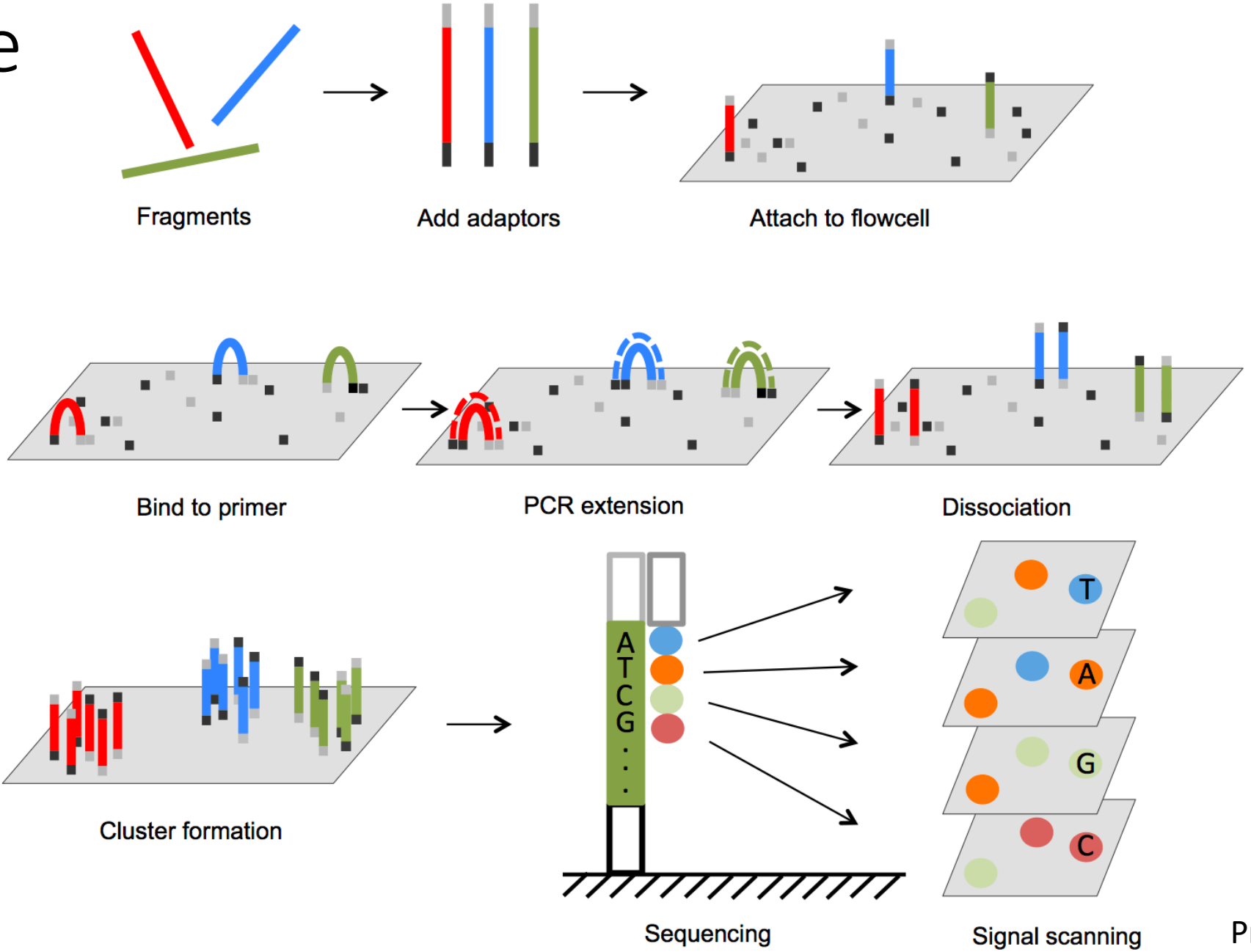


Collect: Filter water for DNA & RNA

The steps of a metagenomic pipeline

1) Collect 
2) Sequence 
3) Merge
4) Quality trim
5) Identify
6) Analyze

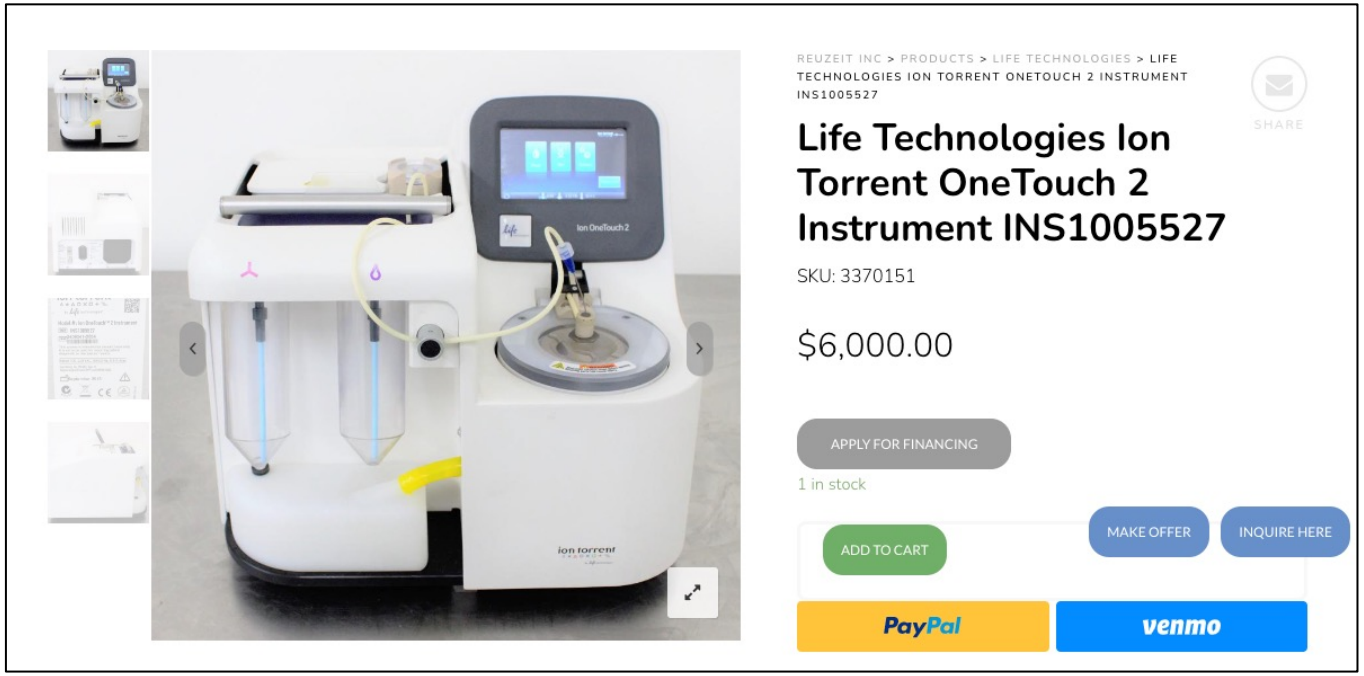
Sequence



Praxilabs.com

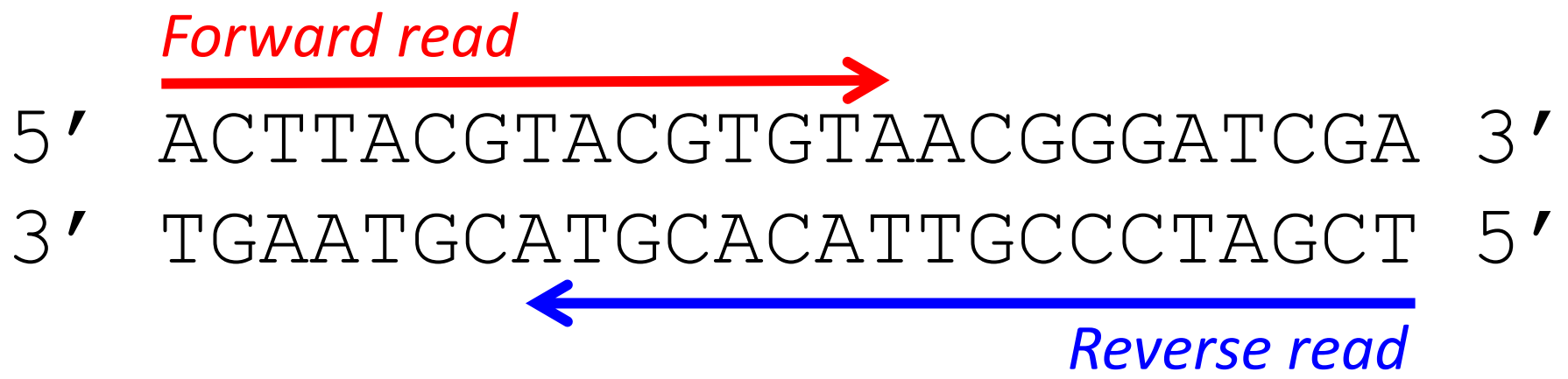
Sequence


GeneLab
Open Science for Life in Space
NASA GeneLab
OPEN SCIENCE FOR LIFE IN SPACE
NovaSeq 6000


REUZEIT INC > PRODUCTS > LIFE TECHNOLOGIES > LIFE TECHNOLOGIES ION TORRENT ONETOUCH 2 INSTRUMENT INS1005527
SHARE
Life Technologies Ion Torrent OneTouch 2 Instrument INS1005527
SKU: 3370151
$6,000.00
APPLY FOR FINANCING
1 in stock
ADD TO CART
MAKE OFFER
INQUIRE HERE
PayPal
venmo

Sequence: paired-end reads

- Next-Gen sequencing can only reliably read a few hundred bases from 5’ of a fragment
- Best quality (lowest probability of error) is near 5’ end of fragment, gets progressively worse in 3’ direction
- Therefore quality near 3’ can be pretty bad
- So sequence from both ends



Sequence: paired-end fastq files

@Molecule 1 defline
Molecule 1 forward seq
+
Molecule 1 quality
@Molecule 2 defline
Molecule 2 forward seq
+
Molecule 2 quality
@Molecule 3 defline
Molecule 3 forward seq
+
Molecule 3 quality
. . .

@Molecule 1 defline
Molecule 1 reverse seq
+z
Molecule 1 quality
@Molecule 2 defline
Molecule 2 reverse seq
+
Molecule 2 quality
@Molecule 3 defline
Molecule 3 reverse seq
+
Molecule 3 quality
. . .

Sequence: the fastq format

- Each record has 4 fields:
  - Defline: starts with @, uniquely identifies the record, looks random
  - Sequence: combination of A/C/G/T, multiple lines ok
  - +
  - Quality: probability of error, encoded as visible character

Sequence: the fastq format

```
@HWI-M01367R:73:000000000-A5WWT:1:1101:17207:1618#0/1
GGACTACTCGGGTATCTAATCCTGTTTGCTCCCCATGCTTTCGTTCCTTAGTGT
CAGTATTGTCCCAGATAGTCGCTTTCGCTCCTGGTATTCCTTCTAATATCTTTG
GATTTTATTCCTACACTAGAAATTCTACTATCCTCTTCCAAACTCAAGATAATC
AGTATTGAAACAACTTTCAAGGTTAAGCCCTGAGATTTCCTTTTCAACTTAATT
ATCCACCTACGAACCCTTTACGCCCAGTTATTTA
+
A3ABBFBFCAABGEGDGFFGGBGDB5FGHFB4FEAEGFHFGHGCGFHGGHGBDB
AF5F5FHHHD5DEFEHFHHGGCEHGDEDE11FGHH5FH4BEGHHBFGHHHGGHF
CCGGHHHHB@4BEGHHHHHGHHHHGHHGGFHHHHHHFHGHGDGFHDFGHHHFGF
HGHHHGGFFFFFACGHHHBGH1<FFDFFFEHACG1FHHHFGGHHFGGHHHHHHH
HHE0GHHHGBCDHCCGHHGHFEGGGGGGGGGGGG
```

defline

nucleotide sequence

\+ sign

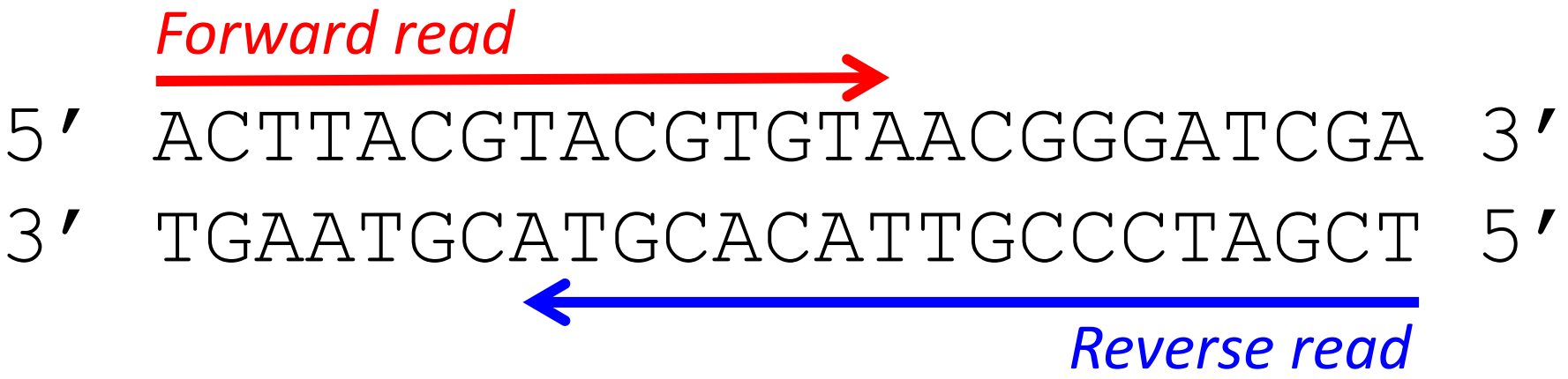
quality

The steps of a metagenomic pipeline

1) Collect
2) Sequence 
3) Merge 
4) Quality trim
5) Identify
6) Analyze

Merge

(from before)

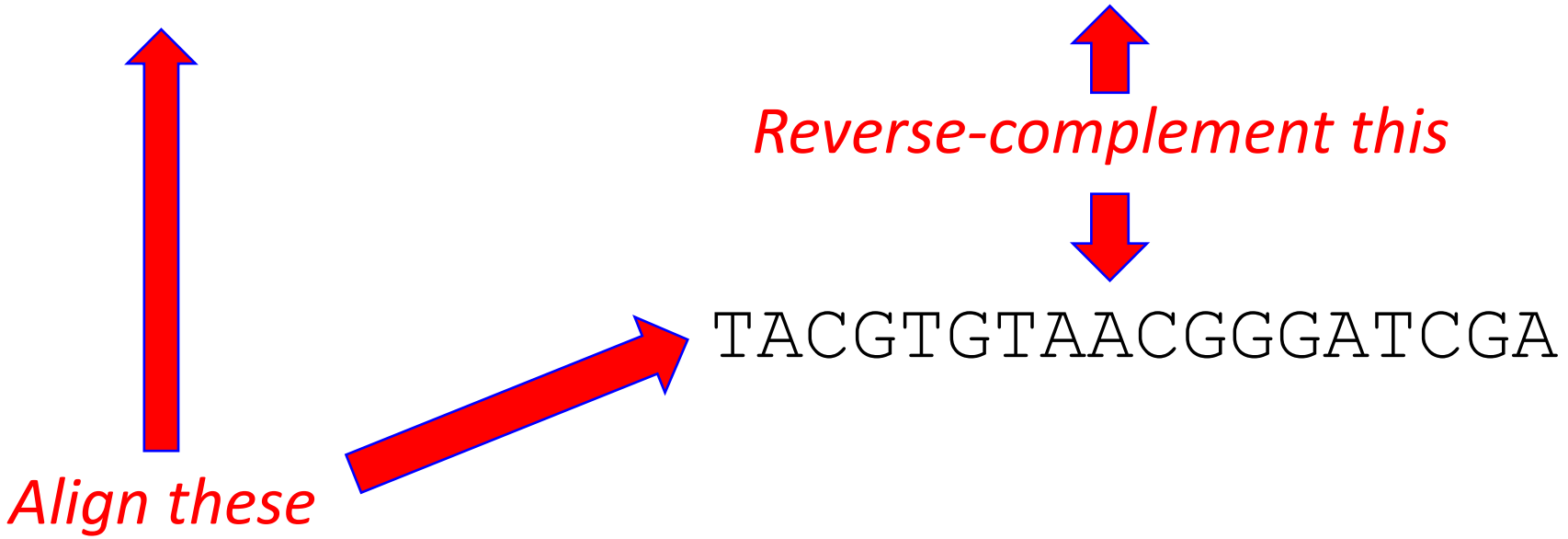


Given : ACTTACGTACGTGT and TCGATCCCGTTACACGTA

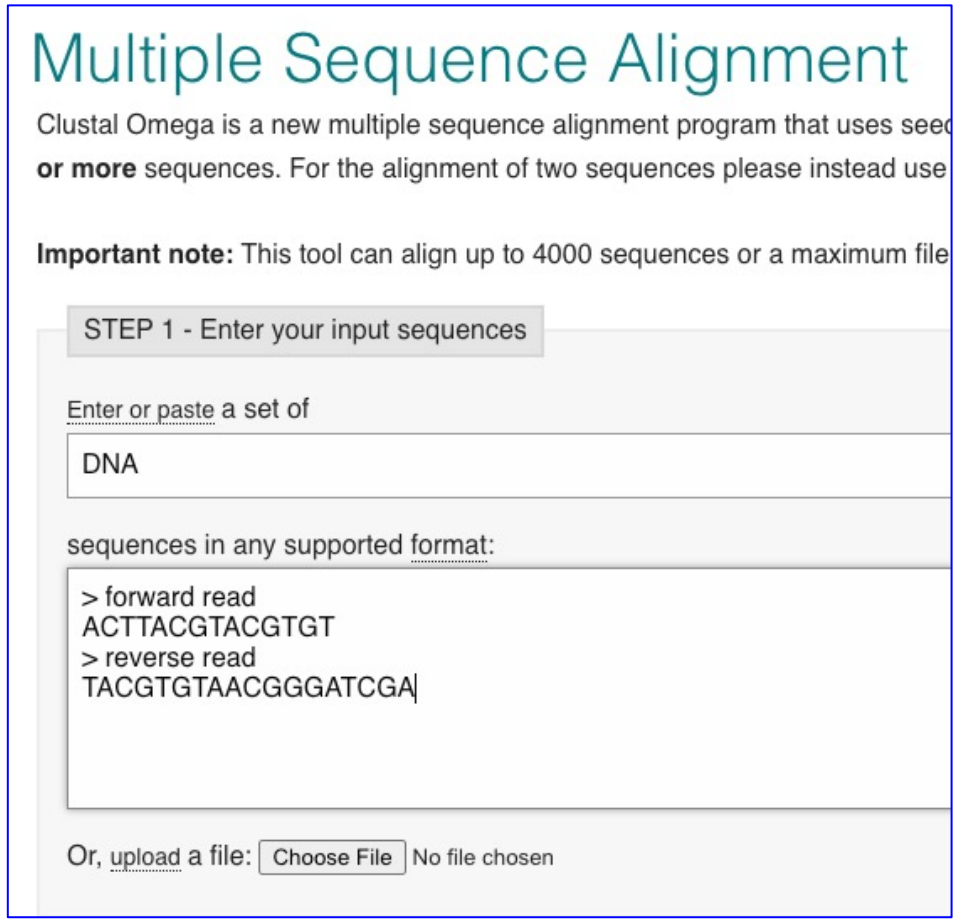
➔ Reconstruct the original molecule

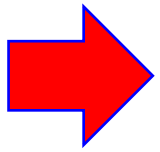
Merge

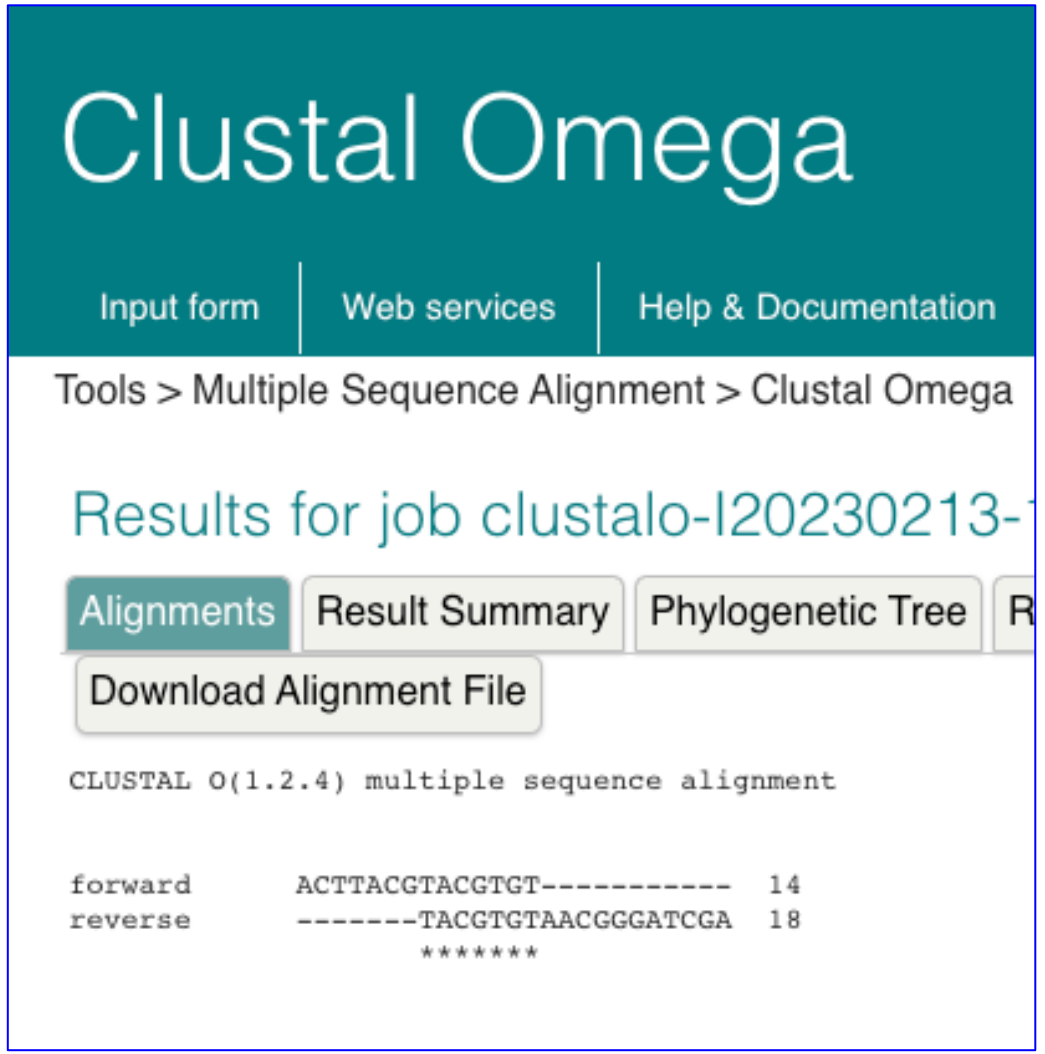
Given : `ACTTACGTACGTGT` **and** `TCGATCCCGTTACACGTA`



Merge: specialty complicated software, but it's like Clustal:

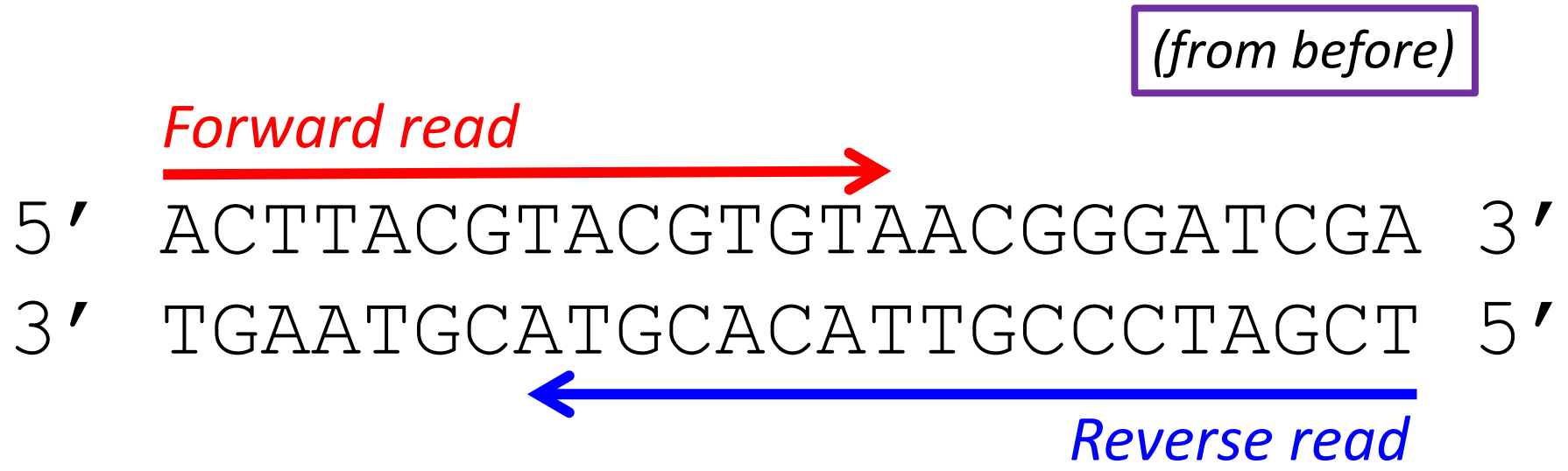






Merge

```
forward ACTTACGTACGTGT---------- 14
reverse -------TACGTGTAACGGGATCGA 18
                *******
```



Merge

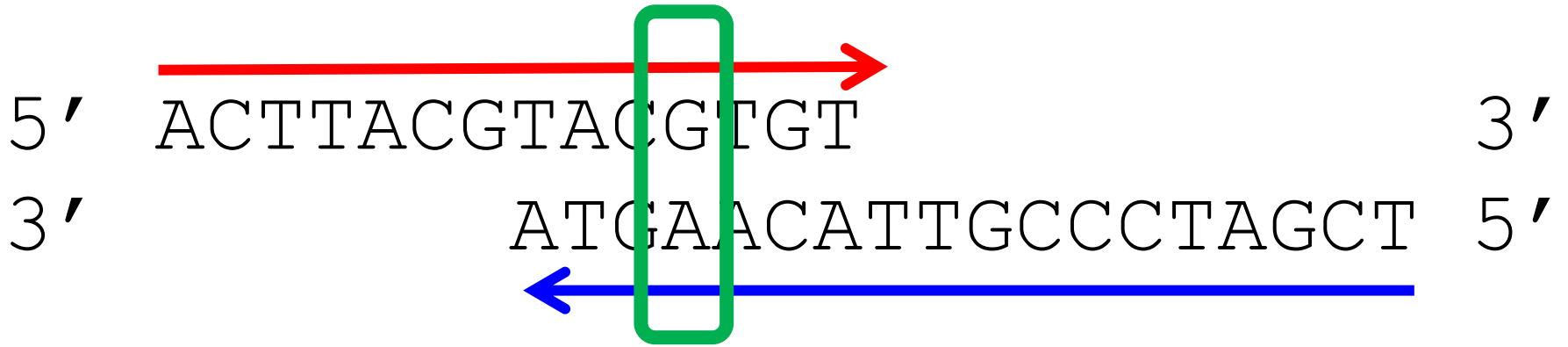
- 2 original paired reads have become 1 merged read
- Merge software can recompute prob(error) at overlap sites

The steps of a metagenomic pipeline

1) Collect
2) Sequence
3) Merge ✓
4) Quality trim ←
5) Identify
6) Analyze

Quality trim

- Reads that are too short ➔ discard
- Reads with poor quality ➔ discard or delete low-quality ends
- Delete adaptors
- Convert to fasta (you've seen fasta format) … each record has 2 fields:
  - Defline starts with > (it was @ in fastq)
  - In addition to A/C/G/T, sequence can contain N, meaning unknown



The steps of a metagenomic pipeline

1) Collect
2) Sequence
3) Merge
4) Quality trim 
5) Identify ⬅
6) Analyze ⬅

Identify

- You guessed it: BLAST (or similar, e.g. Diamond)
- Database can be GenBank or a smaller but more specialized/reliable database
- Can’t use NCBI compute resources
- Possible project: install BLAST software, maybe also GenBank database or create your own database

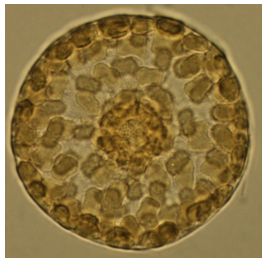
Analyze

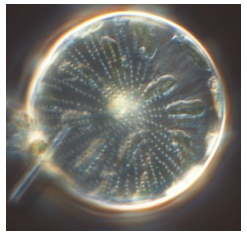
- The fun part
- Might need every technique you ever heard of
- Might need to learn new ones
- Might need to invent new ones

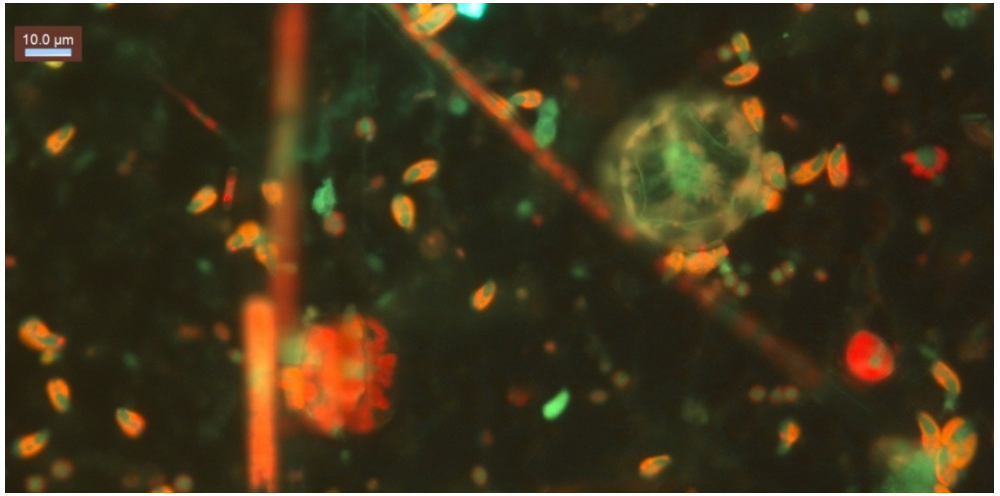
Enough theory.
Time for a case study

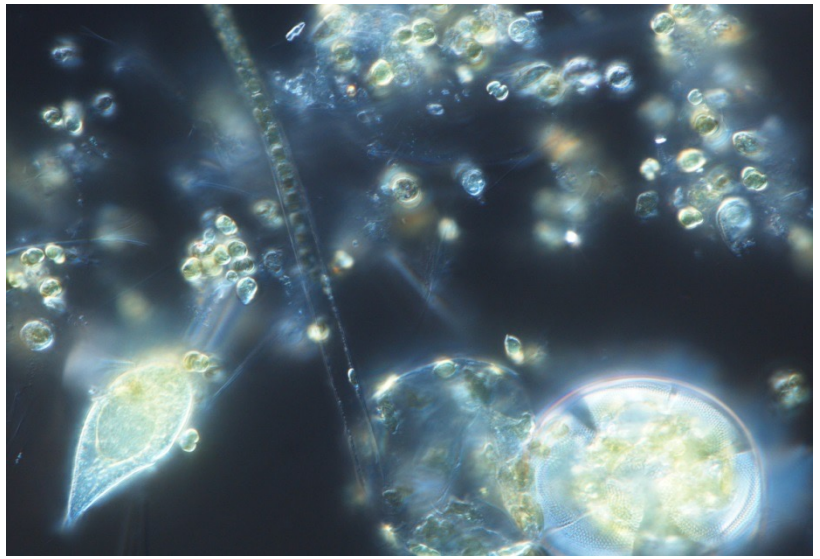
Investigating marine microbial populations in coastal Antarctic via molecular genetic data

Shellie Bench, PhD

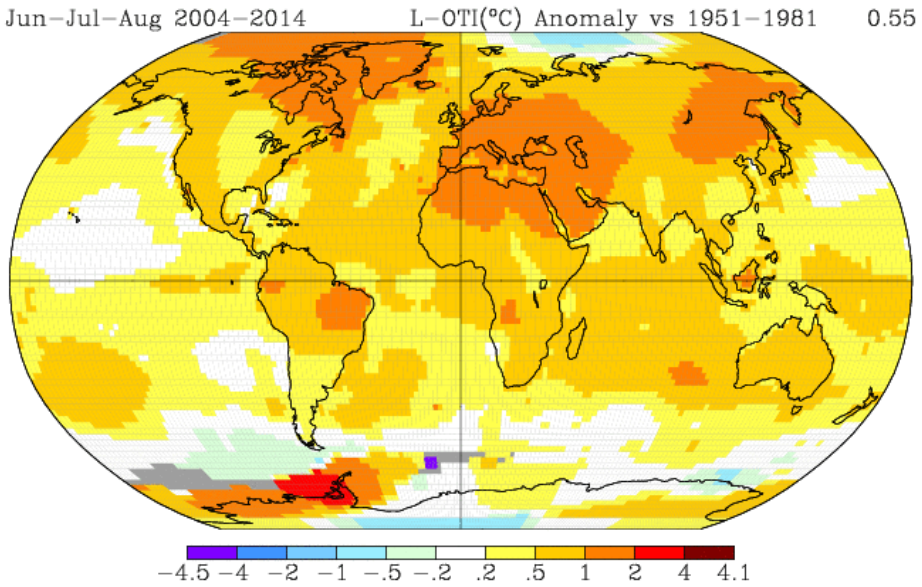


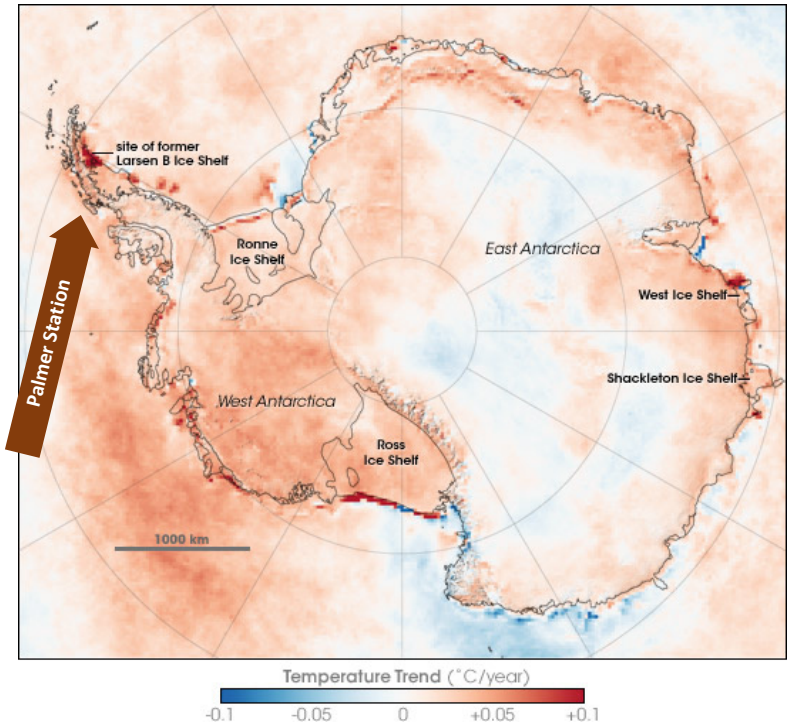






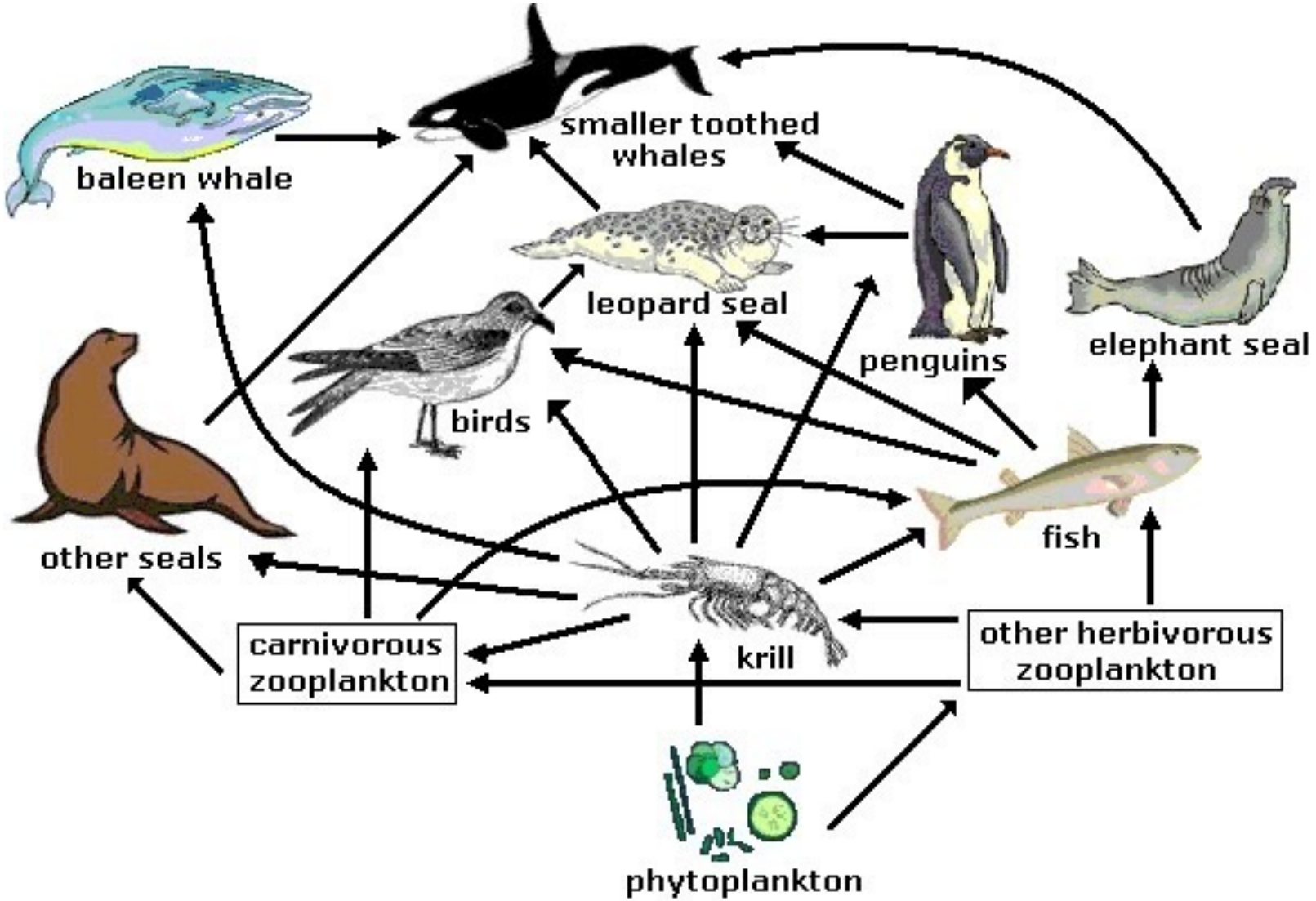
West Antarctic Peninsula: Motivation





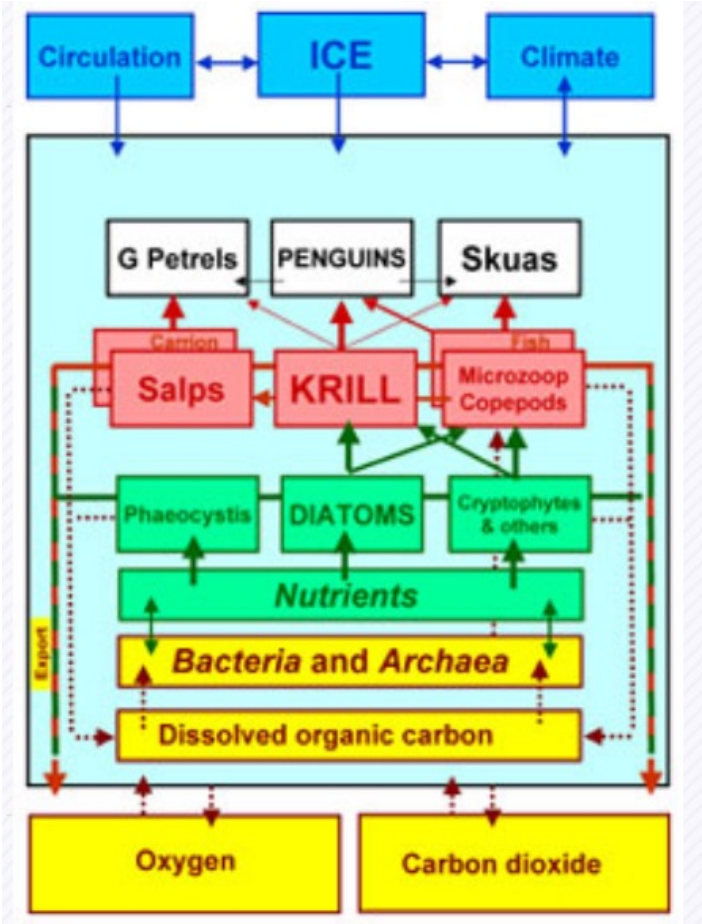
- Environment is warming rapidly
  - Delicate ecosystem
  - Early indicator for other regions

Phytoplankton changes impact entire ecosystem

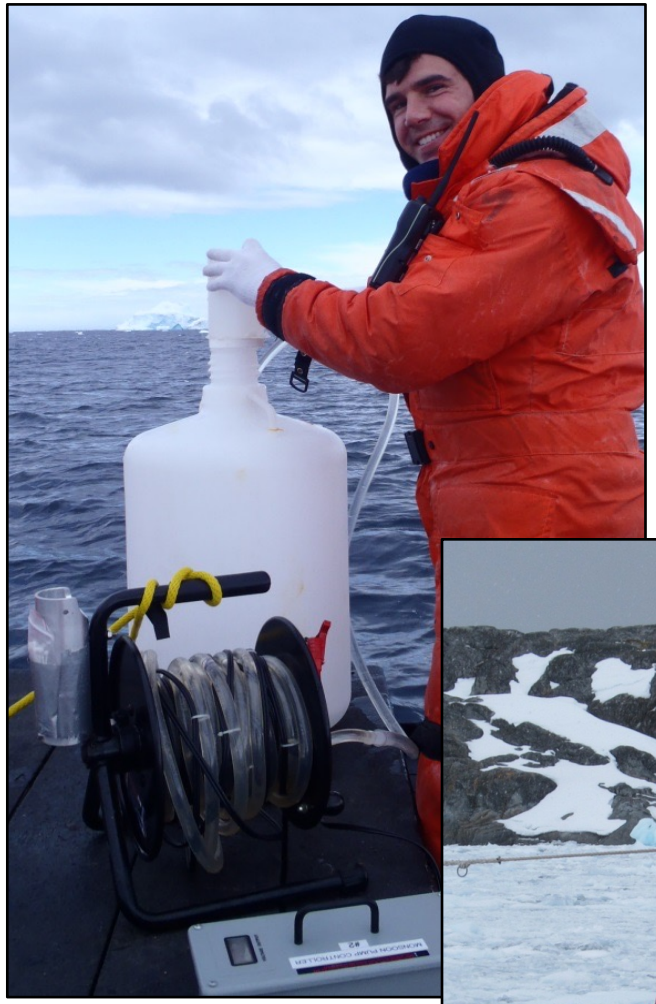


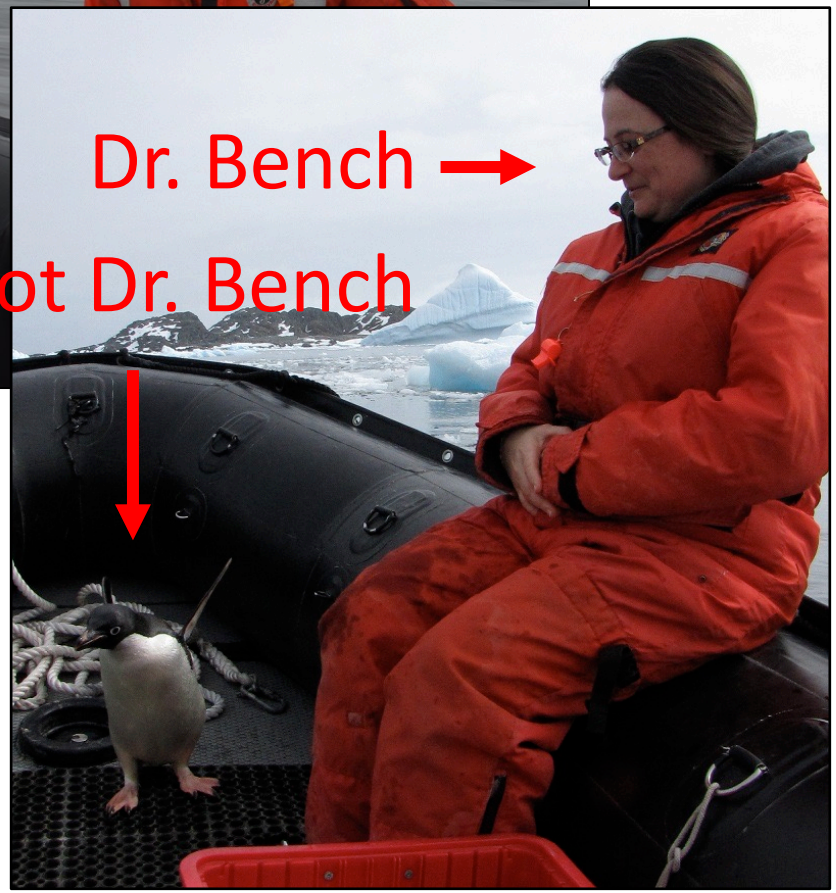
Palmer Station LTER provides context

- 20+ years of ecosystem data capturing the entire food chain
  - From chemistry and microbes to penguins and whales
- Phytoplankton measurements focused on bulk community properties
  - Flow cytometry counts, chlorophyll, and primary production
  - Rough taxonomy indicators (pigments)
  - No molecular data from phytoplankton
- My project goal was to provide molecular data over a 3 year time series
  - To identify correlations with environmental variables and ultimately predict ecosystem impacts




YAMAHA




Dr. Bench
Not Dr. Bench

What: 16S and 18S genes

Where: Palmer Station

Depth = 10m

When: 3 Austral summers:

2013-14

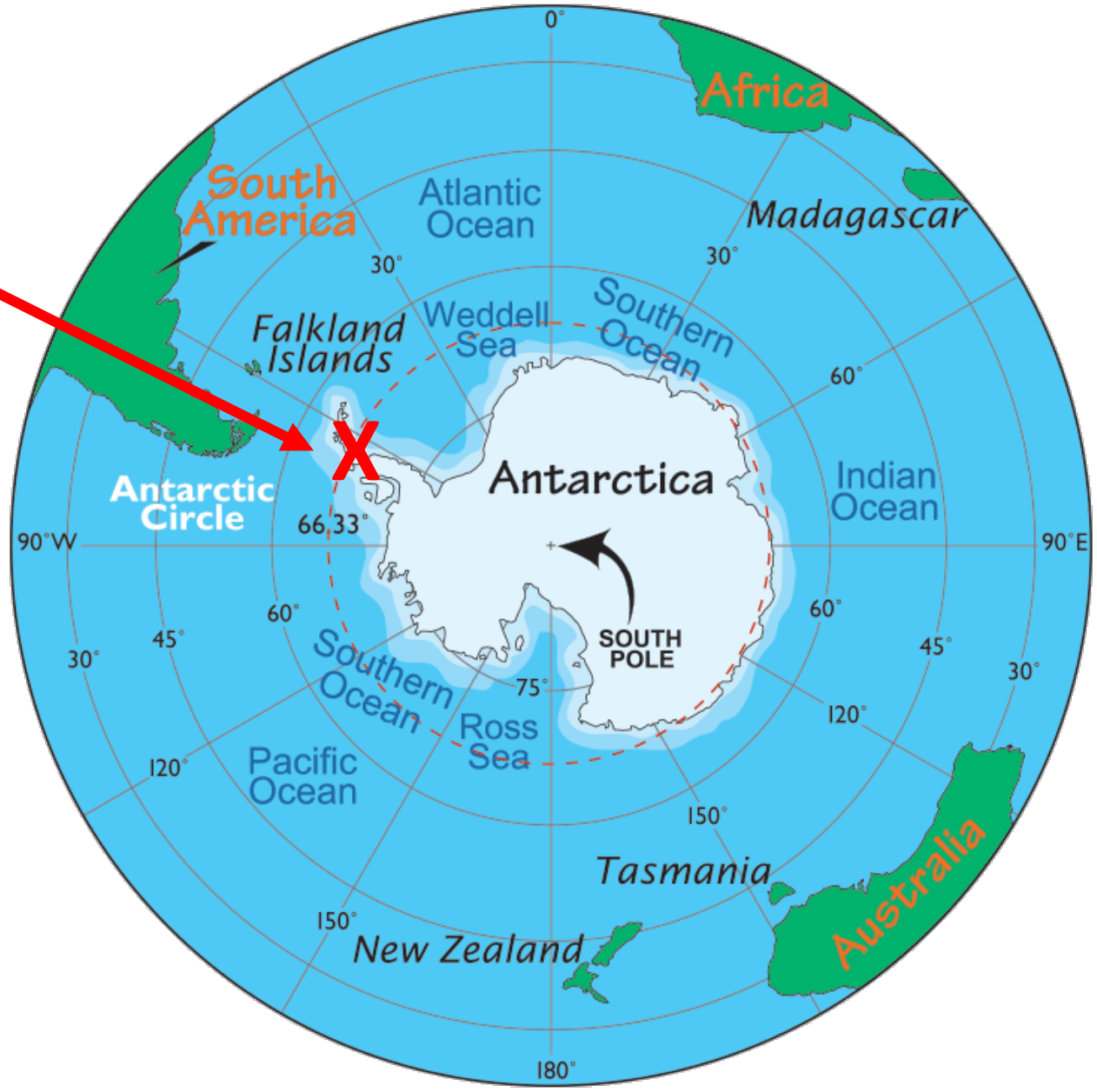
2014-15

2015-16

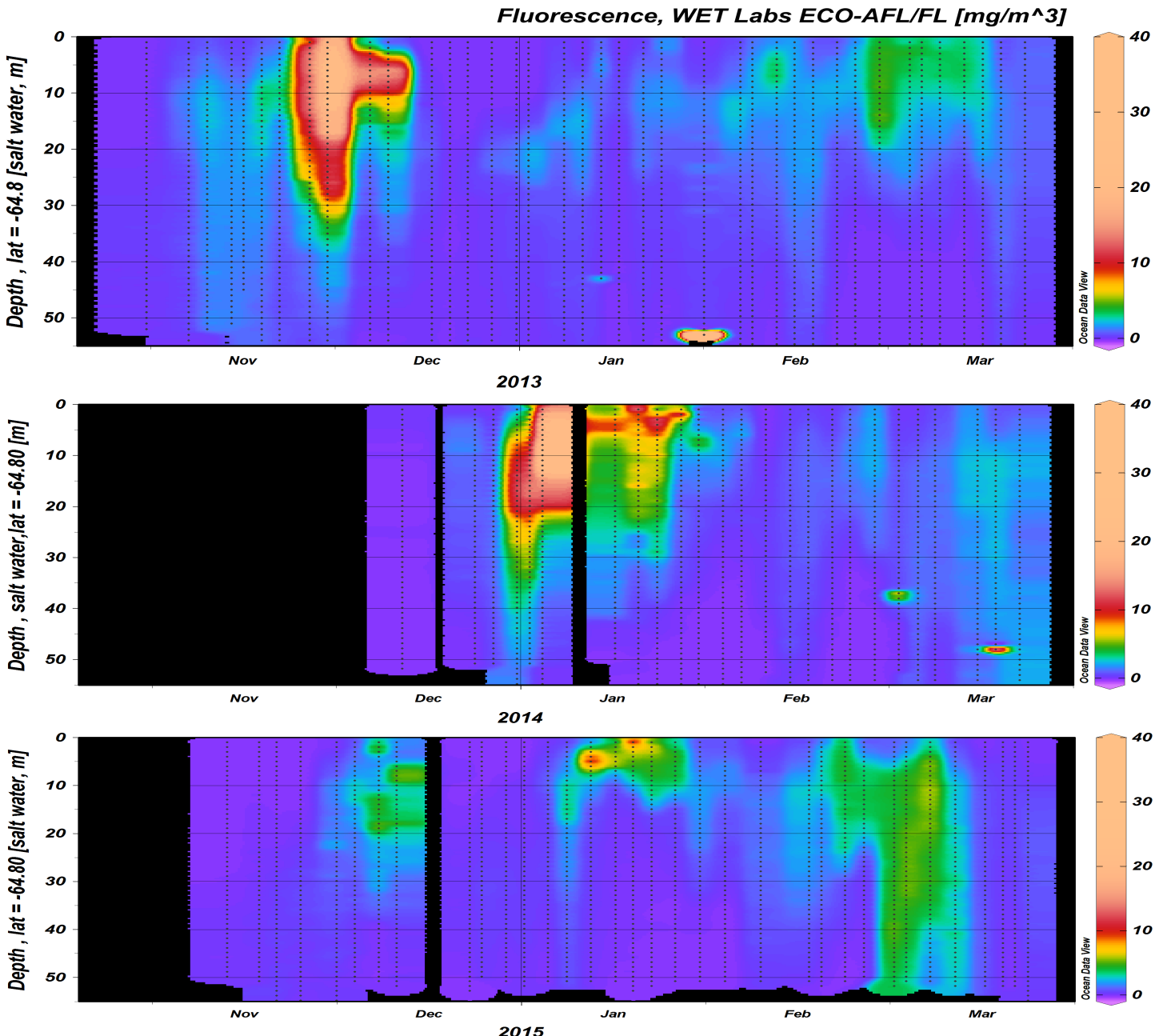
Who: Bench

Later: Heller, Siman-Tov

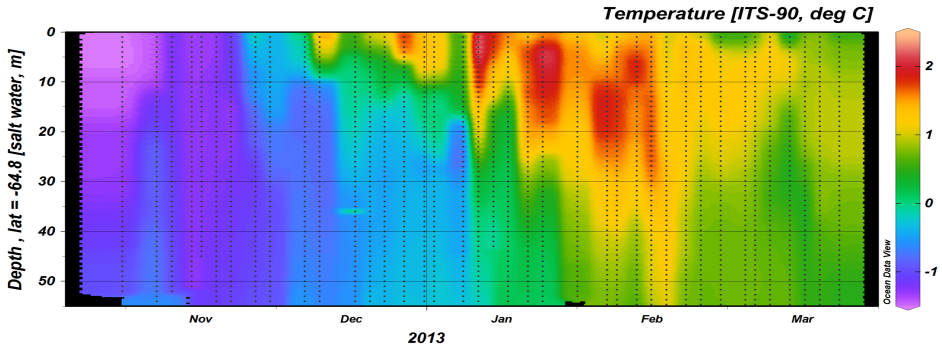
Salter

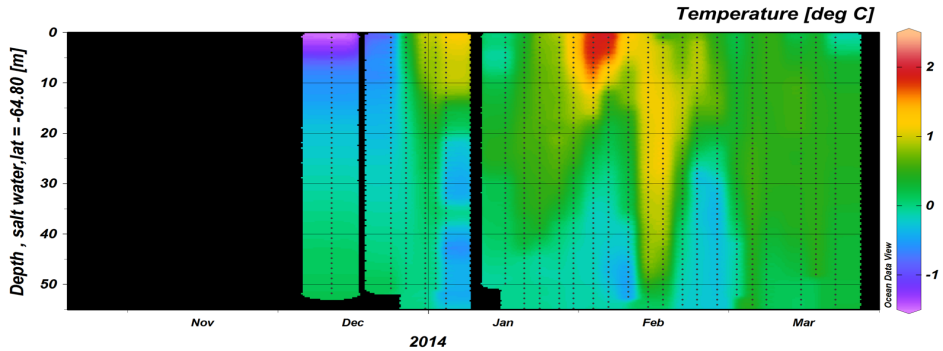


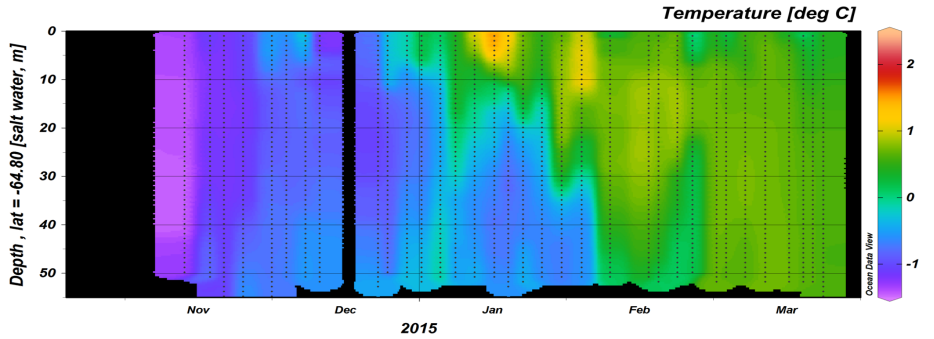
Fluorescence: proxy for biomass



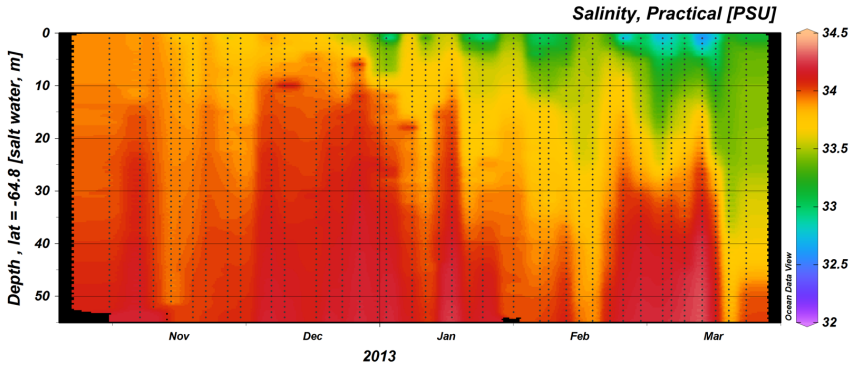
Temperature

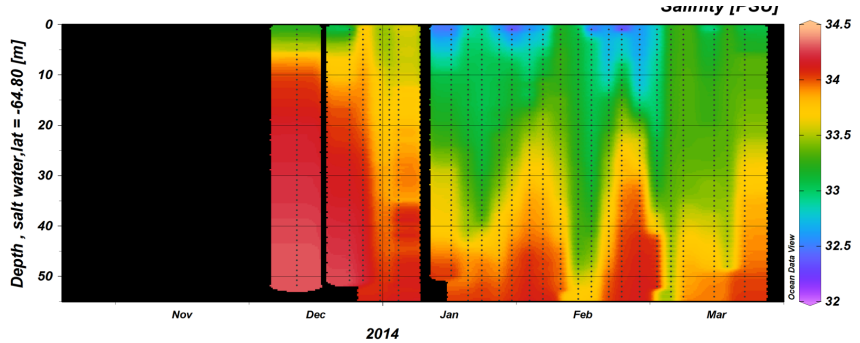


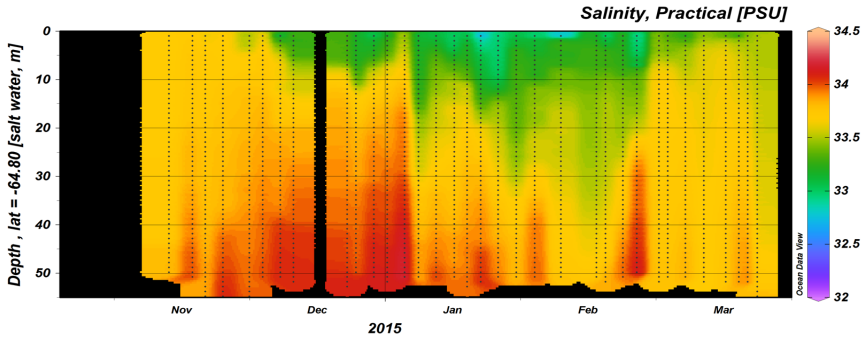




Salinity







Why does colder mean less salty?

- "Colder" *by Antarctic standards*
- Water freezes to join growing sea ice and terrestrial land ice
- Not cold enough to freeze salt water
- Brine rejection: ocean H2O molecules attach to growing ice, leaving salt behind
- Water becomes saltier, denser →sinks
- Never touch a brinicle!

This video: https://www.youtube.com/watch?v=lAupJzH31tc

Higher resolution video at: https://www.youtube.com/watch?v=BtQhb8sWJNw

Opening Thought Question – 2/16

Why is studying phytoplankton blooms important?

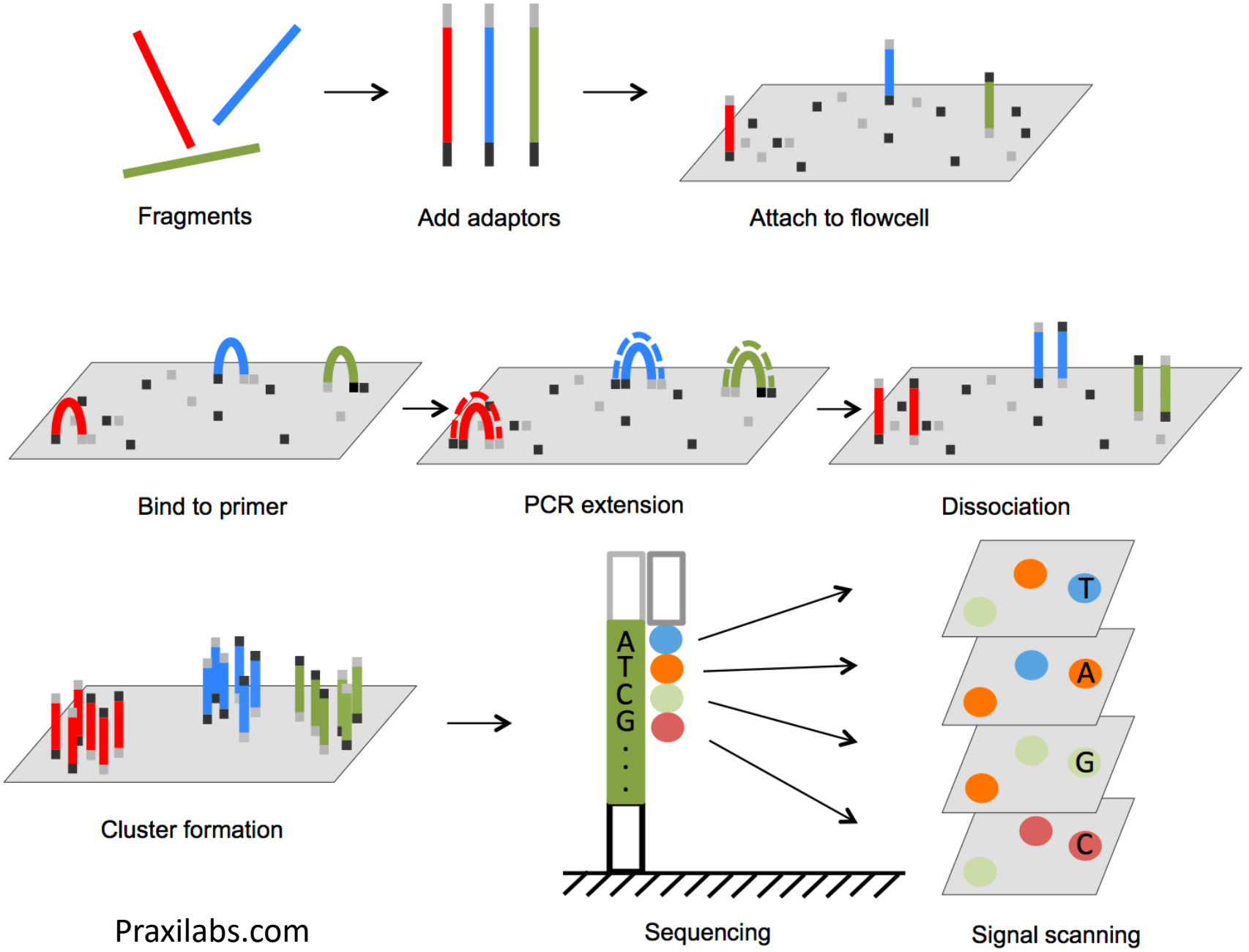
The data

- 15 data sets (“libraries”)
  - 2012-13: 2 libraries
  - 2013-14: 6 libraries
  - 2014-15: 7 libraries
- 18S rRNA genes from microbial eukaryotes
- Pending analysis: 16S rRNA genes from prokaryotes
- Merging & trimming: nothing in particular to report

Identification what's the deal with all these carp?

- BLASTed all the merged/trimmed fastas against GenBank … took ~2 months on the CoS HPC.
- Shockingly many hits to carp.
- USGS: "The common carp (*Cyprinus carpio*) is a native of Asia. It is now found on every continent except Antarctica (Jester 1974) and in all 48 contiguous States (Sigler 1958). The northern limit to carp distribution appears to be the 18° C isotherm (Keleher 1956)."

Adaptors … I'm *so* embarrassed!



Praxilabs.com

Adaptors … I'm *so* embarrassed!

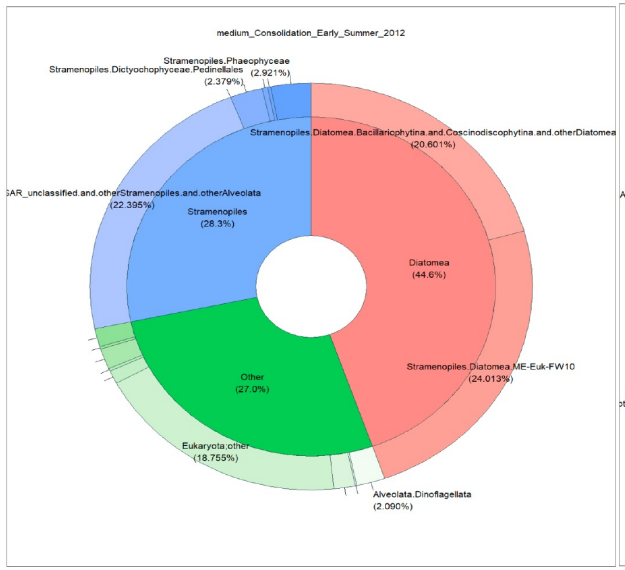
- I forgot to delete them from Antarctic libraries.
- Whoever submitted *Cyprinus carpio* to GenBank used the same adaptor sequences.
  - And also forgot to delete them.
- So *our* bogus adaptor sequences in our queries, hit *their* bogus adaptor sequences in GenBank.
- And that's time I'll never get back.

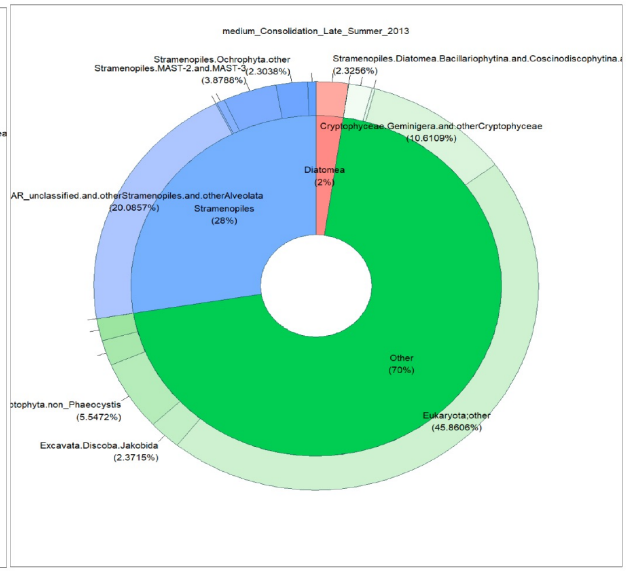
How to make sense of millions of BLAST hits

- Can’t think about species … too many species
- Ditto genus
- Family: better
- Order: even better, but sometimes family differences matter
- Class: better than order, but often order/family differences matter
- Phylum: much too broad
- ➔ Determine groups of interest, ignore taxonomic rank
  - Order *Jakobida*
  - Subphylum *Ochrophytina*
  - *Holozoa* (unranked), a kind of *Ophisthokonta* (broad group in domain *Eukaryota*)
  - *Ophisthokonta* that aren’t *Holozoa*

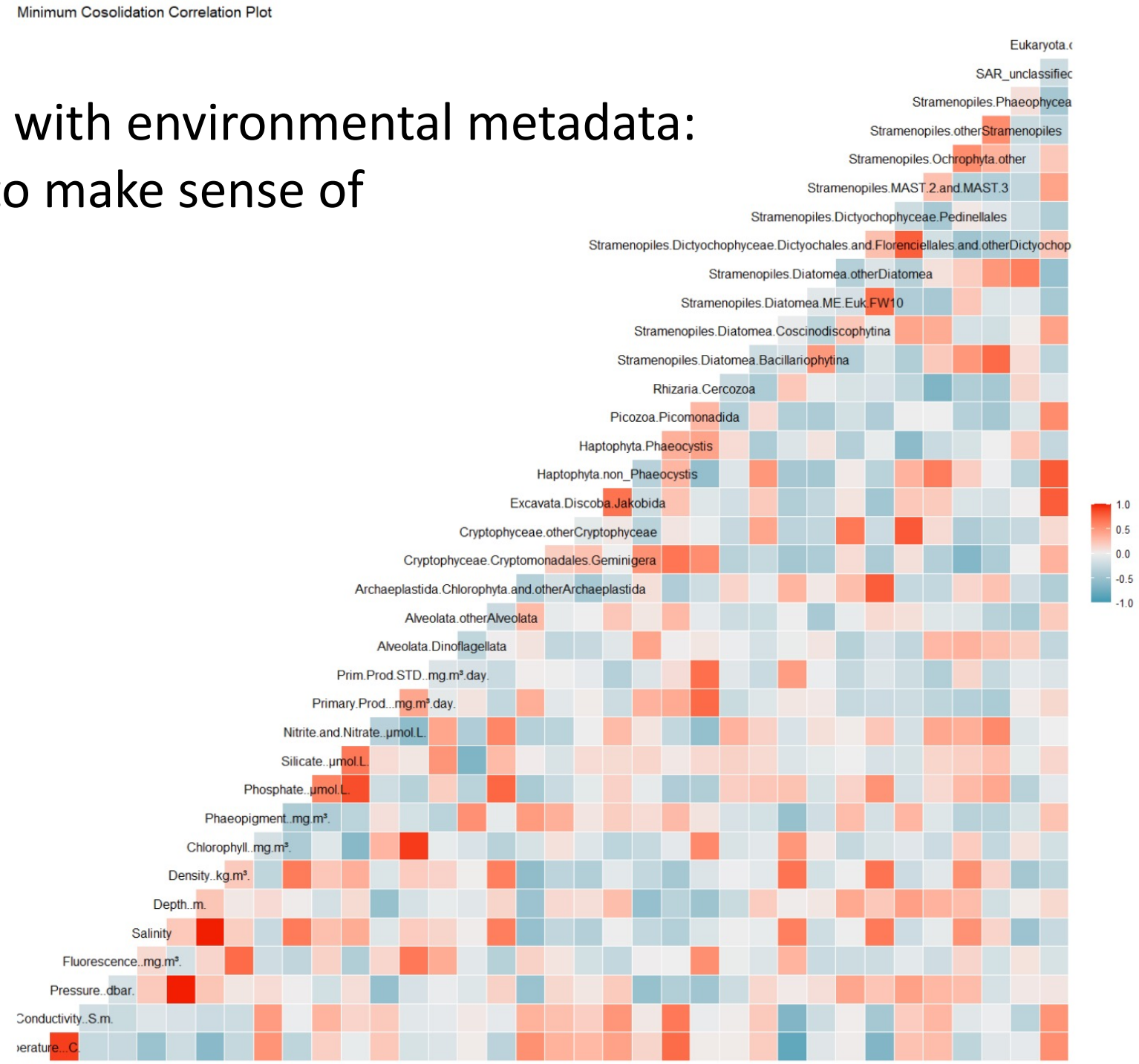
Eventually

- 28 categories
  - Some are clades (monophyletic)
  - Some are polyphyletic
- This can only be done by an experienced ecologist
- Without the insight captured in the list of categories, analysis would be impossible



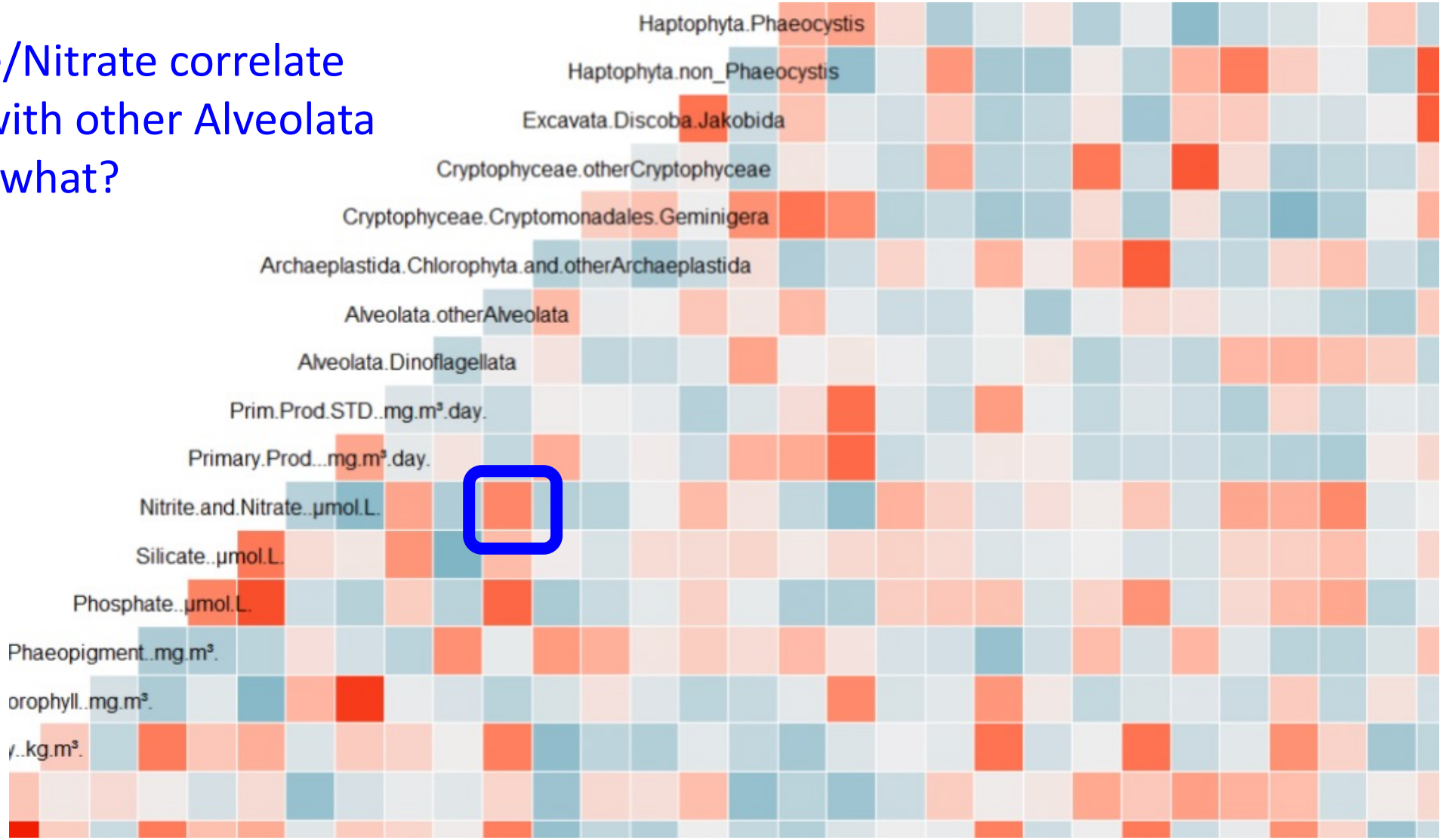


Correlating with environmental metadata: Too much to make sense of



E.g. just 1 of 650 numbers:

Nitrate/Nitrate correlate sorta with other Alveolata but so what?

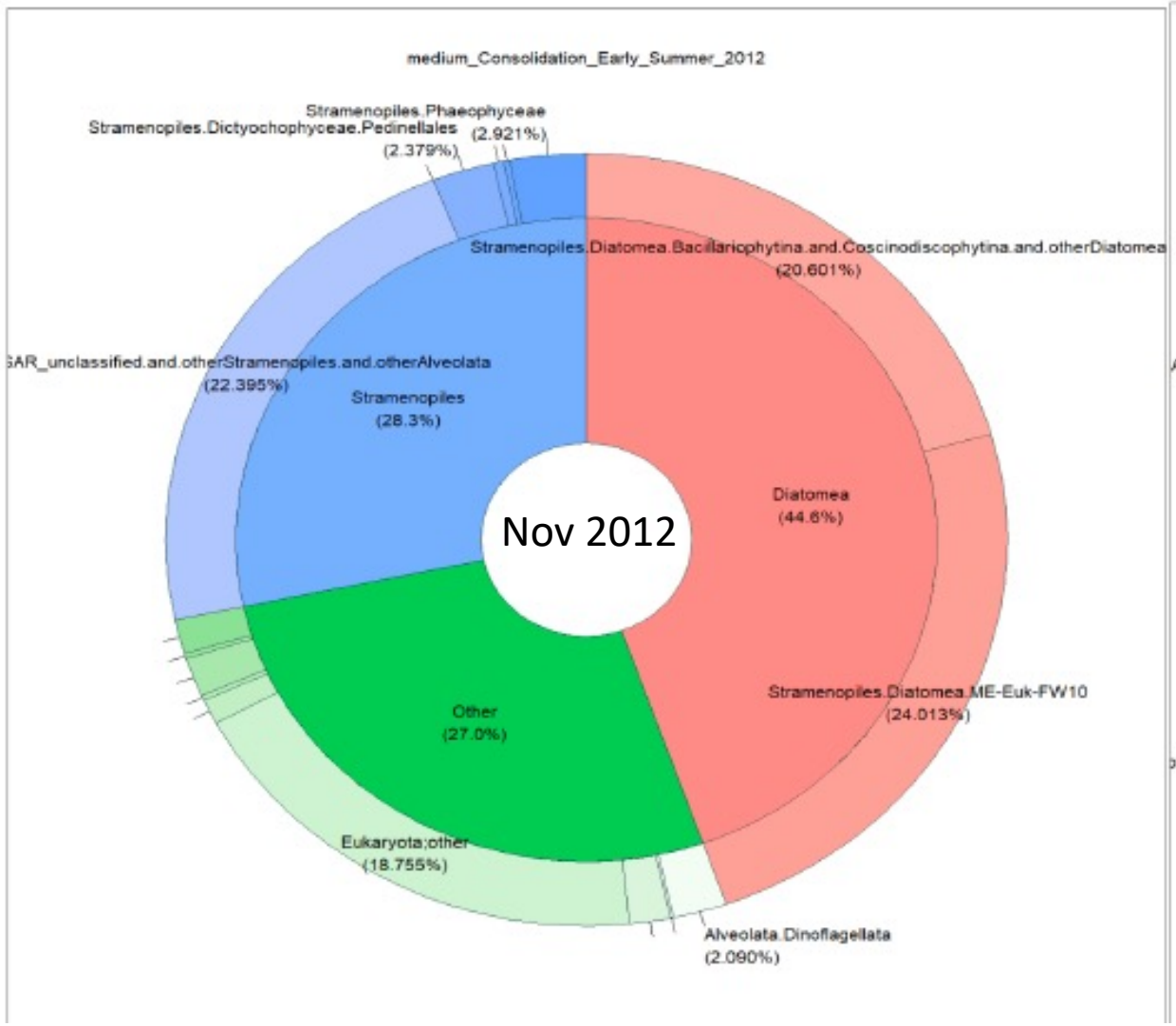


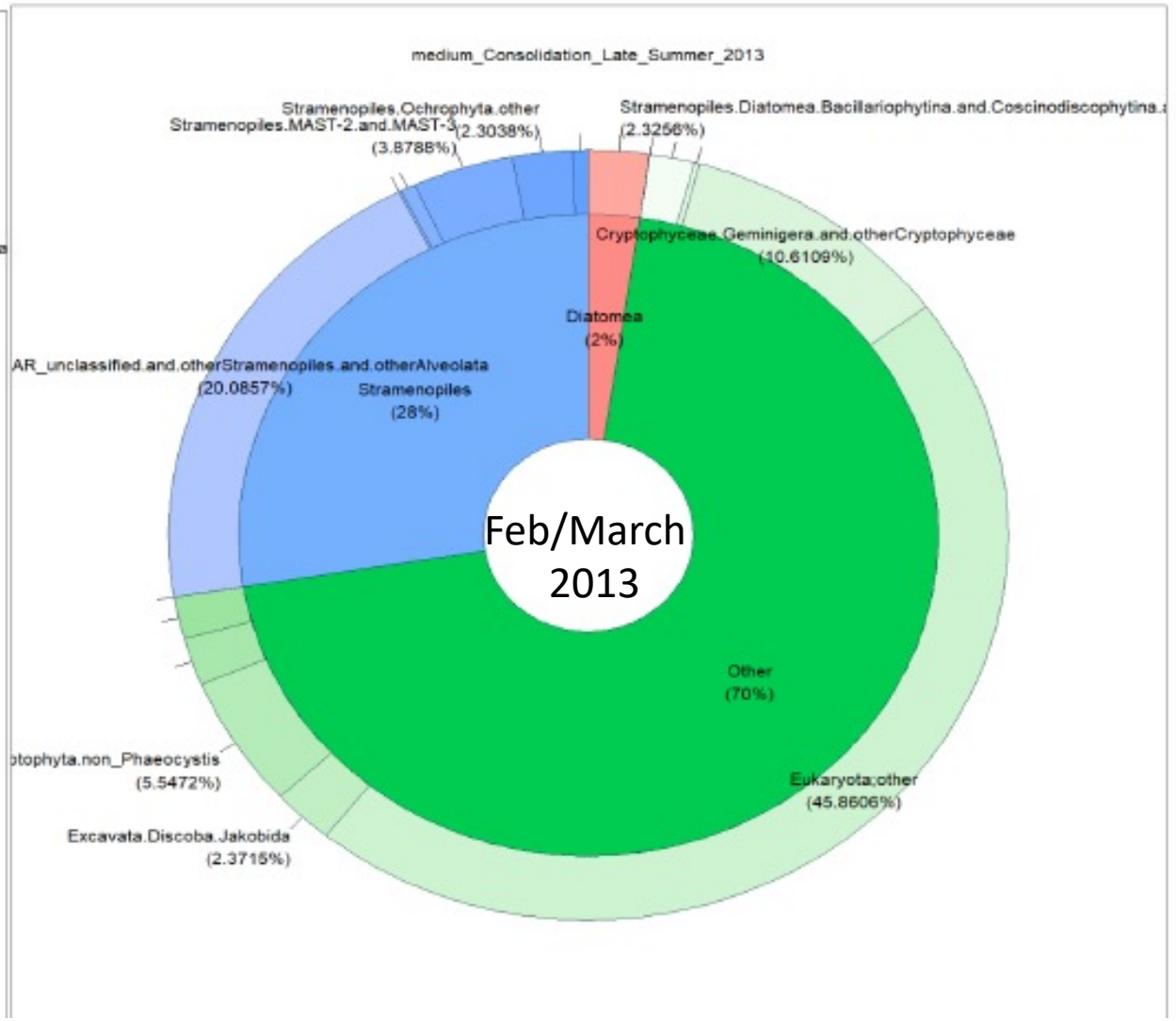
Too many categories!

- Idan’s insight:
  - Merge some categories
  - See if patterns become visible
- 3 “consolidations”:
  - Original (no consolidation): 28 categories
  - Minimum consolidation: 21 categories
  - Medium consolidation: 16 categories
  - Maximum consolidation: 12 categories

Now we can visualize groups over time

= diatoms

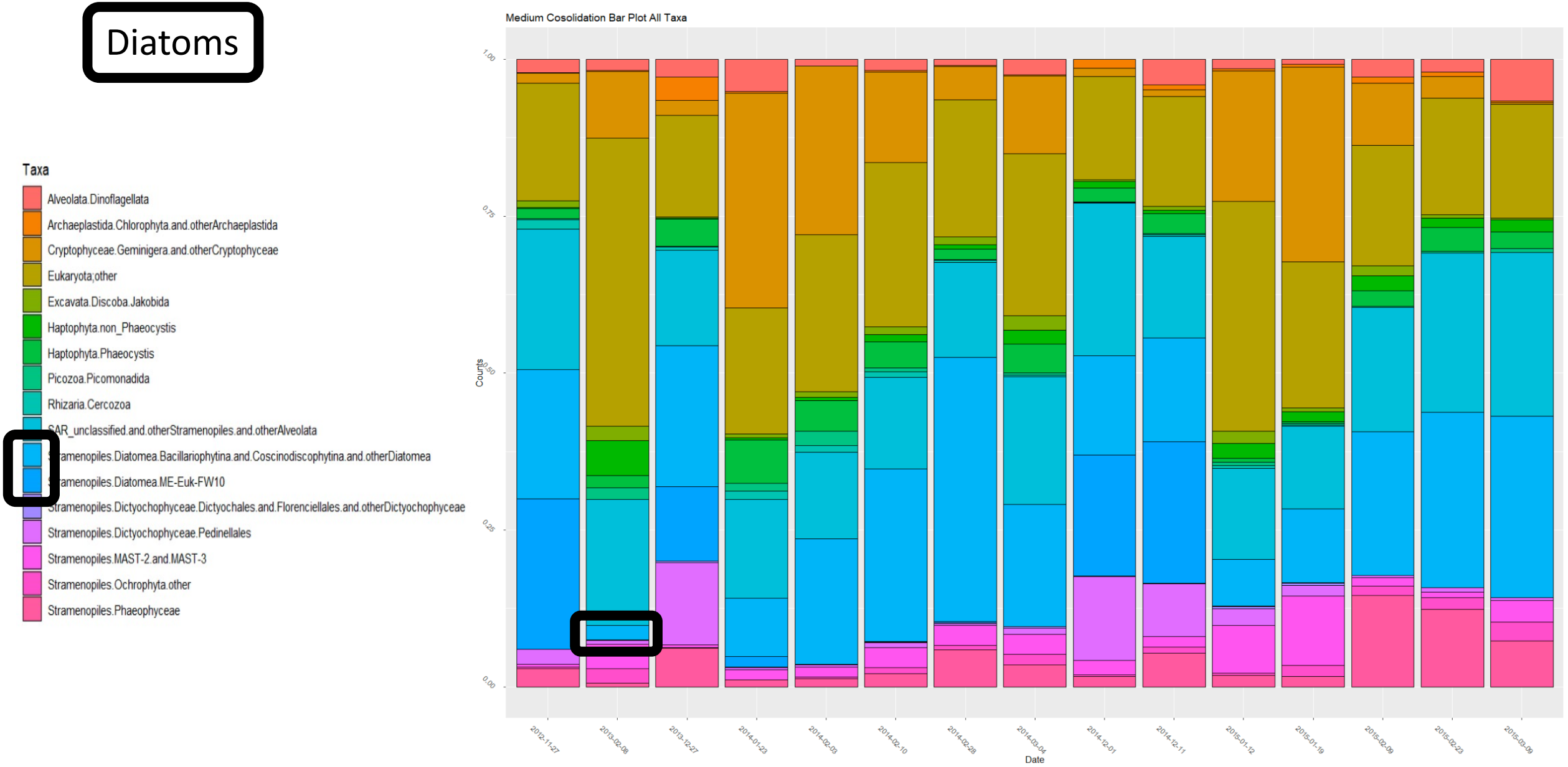




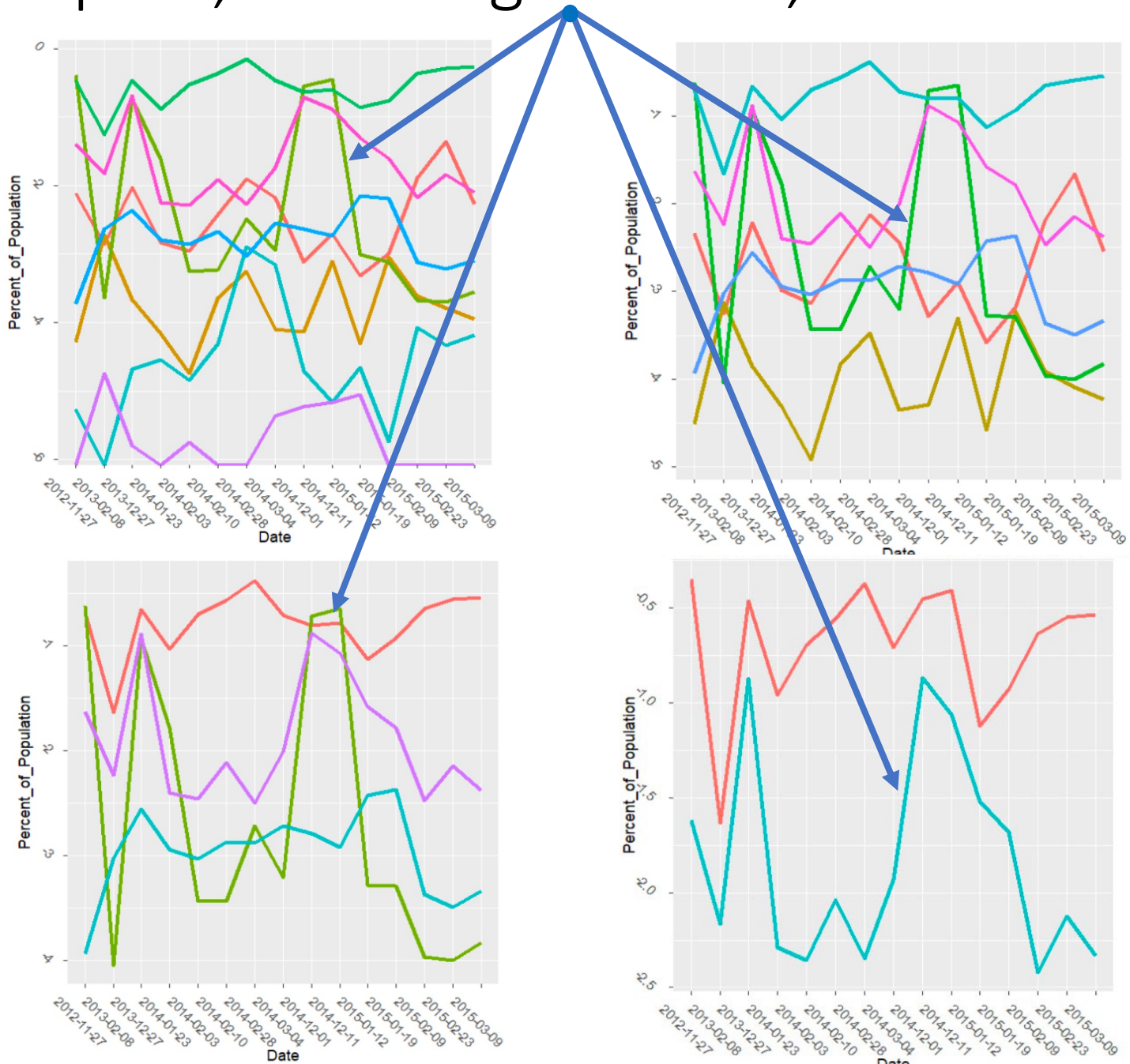
Siman-Tov, Idan MSBI report 2022

Medium consolidation, all taxa

Diatoms

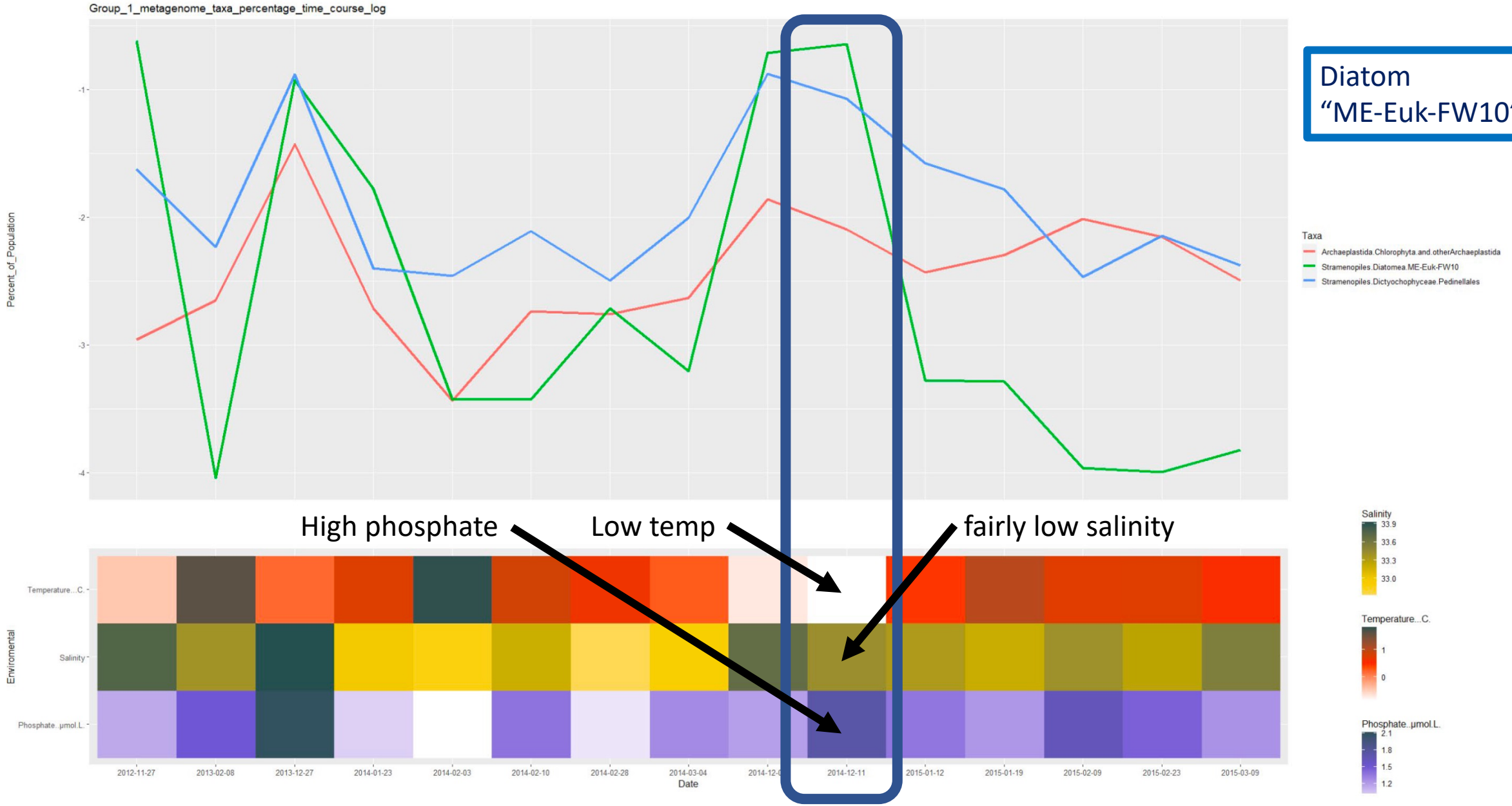


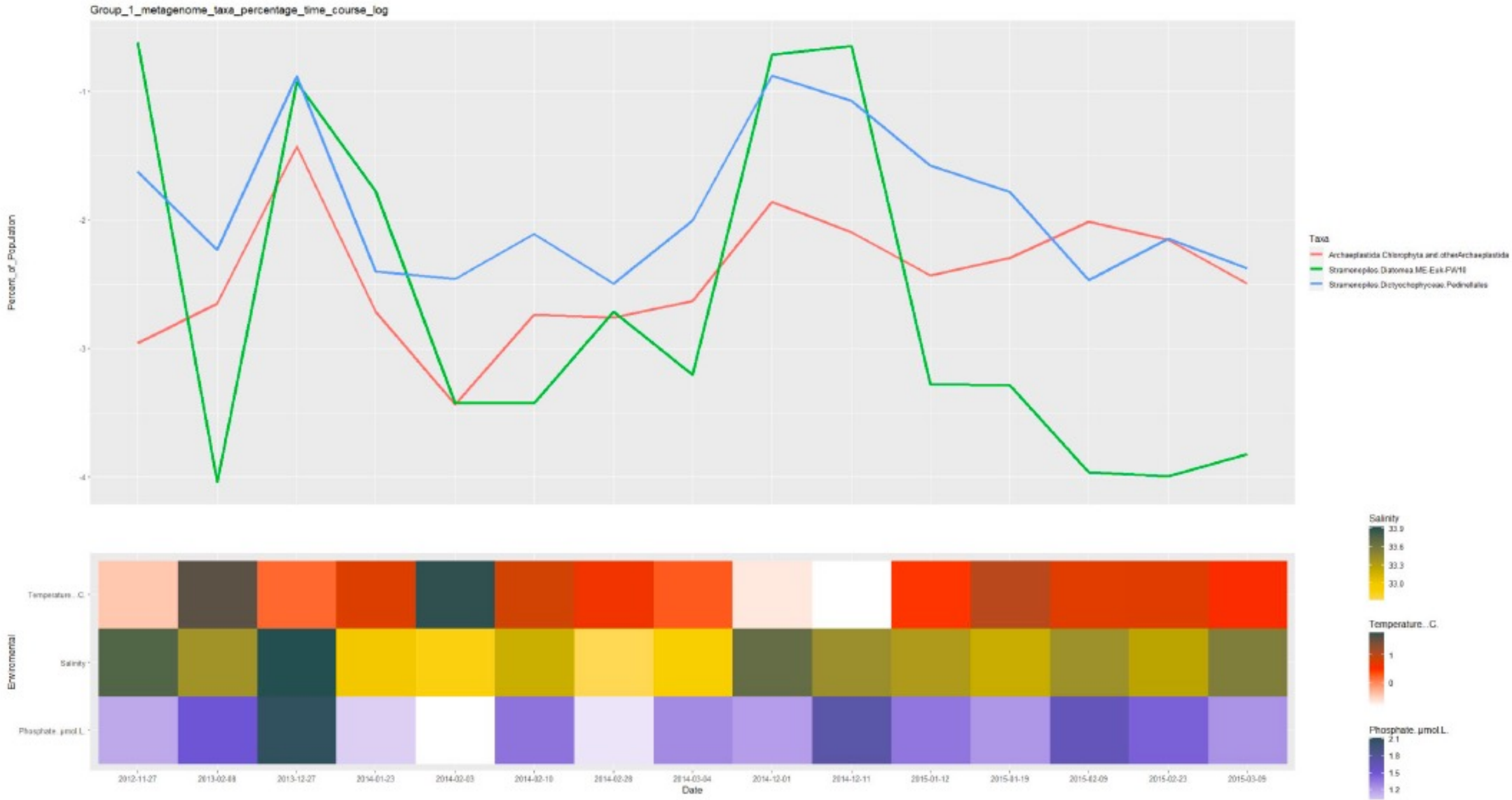
Stramenopiles, including diatoms, at 4 consolidations



It's diatom "ME-Euk-FW10"

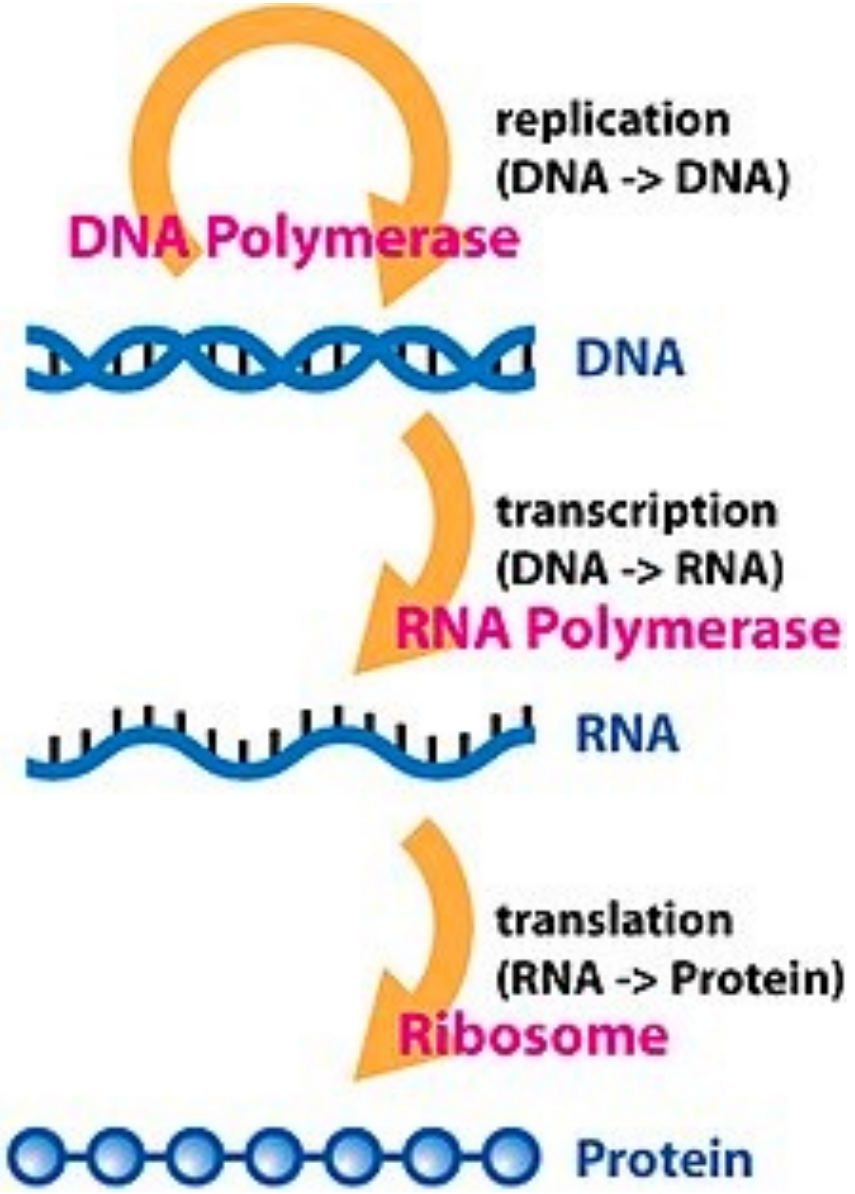
Really nice combination of line chart and heat map




Group_1_metagenome_taxa_percentage_time_course_log
Percent_of_Population
-1
-2
-3
-4
Taxa
Archaeplastida.Chlorophyta and othesArchaeplastida
Stramenopiles.Diatomea.ME.Esk.PW18
Stramenopiles.Dictyochophyceae.Pedinellales
Enviromental
Temperature...C.
Salinity
Phosphate..µmol.L.
2012-11-27
2013-02-08
2013-12-27
2014-01-23
2014-02-03
2014-02-10
2014-02-28
2014-03-04
2014-12-01
2014-12-11
2015-01-12
2015-01-19
2015-02-09
2015-02-23
2015-03-05
Date
Salinity
33.9
33.6
33.3
33.0
Temperature...C.
1
0
Phosphate..µmol.L.
2.1
1.8
1.5
1.2

Metatrascriptomics

- Metagenomics tells you the genetic potential of a community
- Metatranscriptomics tells you the genetic activity of a community

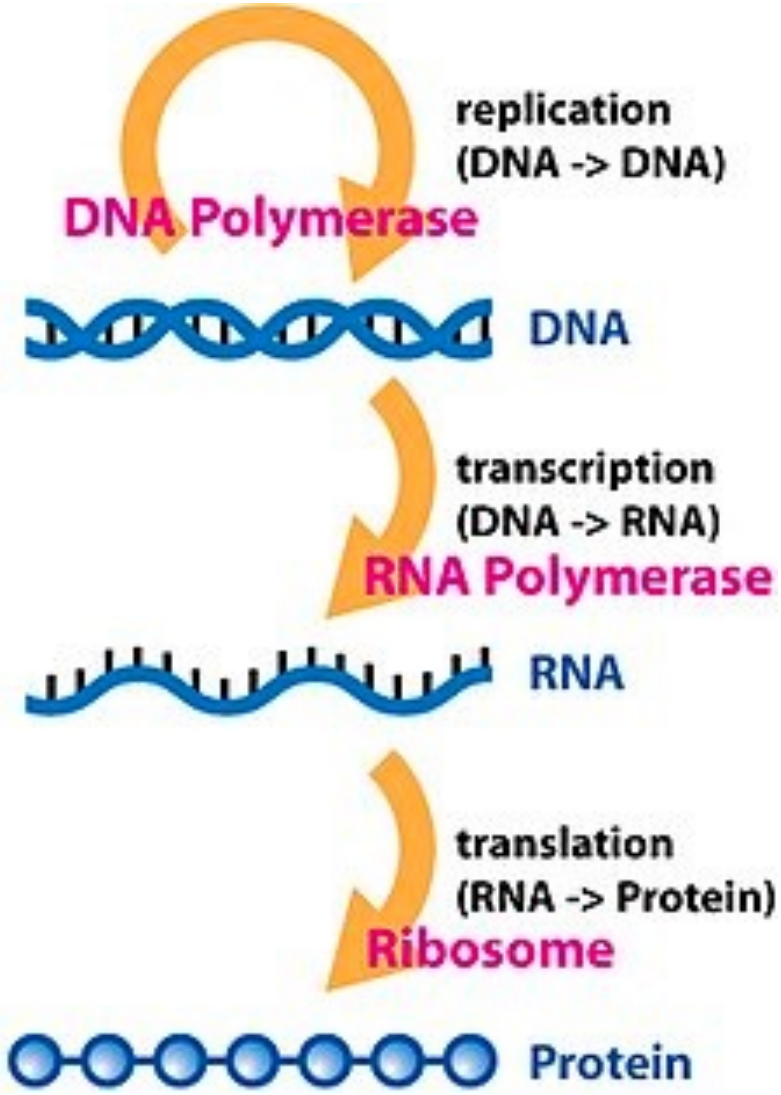


Why a gene might be present but not expressed

- It’s not really a gene, it’s a pseudogene
  - Small mutation, e.g. frame shift
  - Looks like a gene to annotation software
  - No longer under selective pressure to be useful ➔ subject to random mutations
- It wasn’t expressed *at the moment you sequenced it*

Why a gene might be present but not expressed

- Slide title is too binary.
- Genes are more subtle than “expressed” or “not expressed”
- Better question: *how productive* is a gene?
- So sequence mRNA rather than DNA, and count molecules
- Bad news / Good news
  - The bad news: you can’t sequence mRNA
  - The good news: you can sequence cDNA



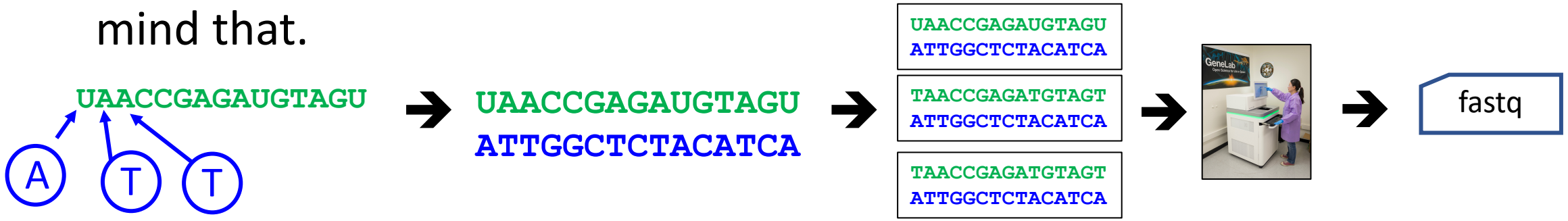
Complementary DNA ("cDNA")

- Sequencers need double-stranded molecules.

ATTGGCTCTACATCA
TAACCGAGATGTAGT
} DNA

- mRNA is a single-stranded copy of a DNA strand (but U instead of T).

- So hybridize nucleotides to mRNA ➔ double-stranded cDNA, amplify, and sequence. 1 molecule has Uracil on 1 strand. Sequencers don't mind that.



One thing we can learn from a marine metatranscriptomic experiment

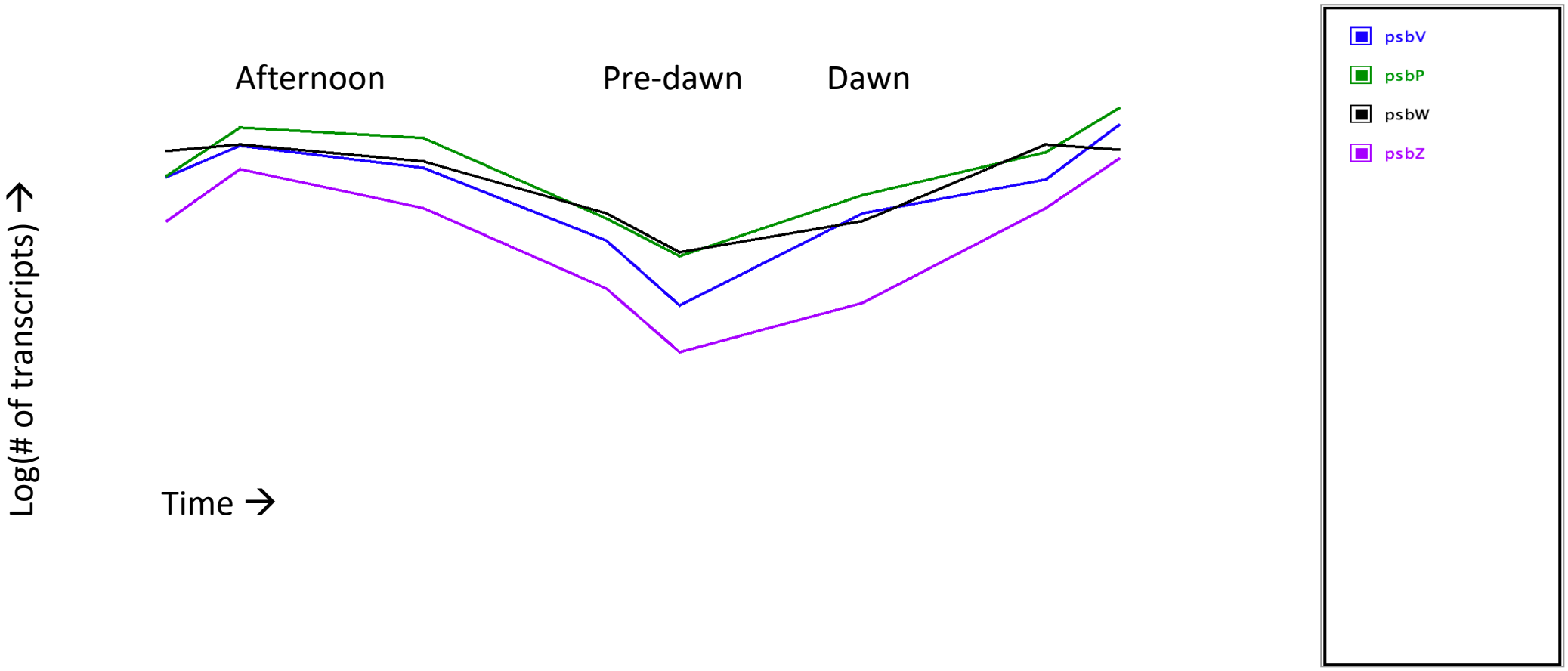
- Diel expression of photosystem genes
- “Diel” = fluctuating on a (roughly) repeating 24-hour cycle
- Why?
- Photosystem proteins are
  - Expensive
  - Short-lived (half-life is sometimes 12 hours)
- Just-in-time manufacturing: why Japanese cars have outsold American cars for the past 50 years

$>>$

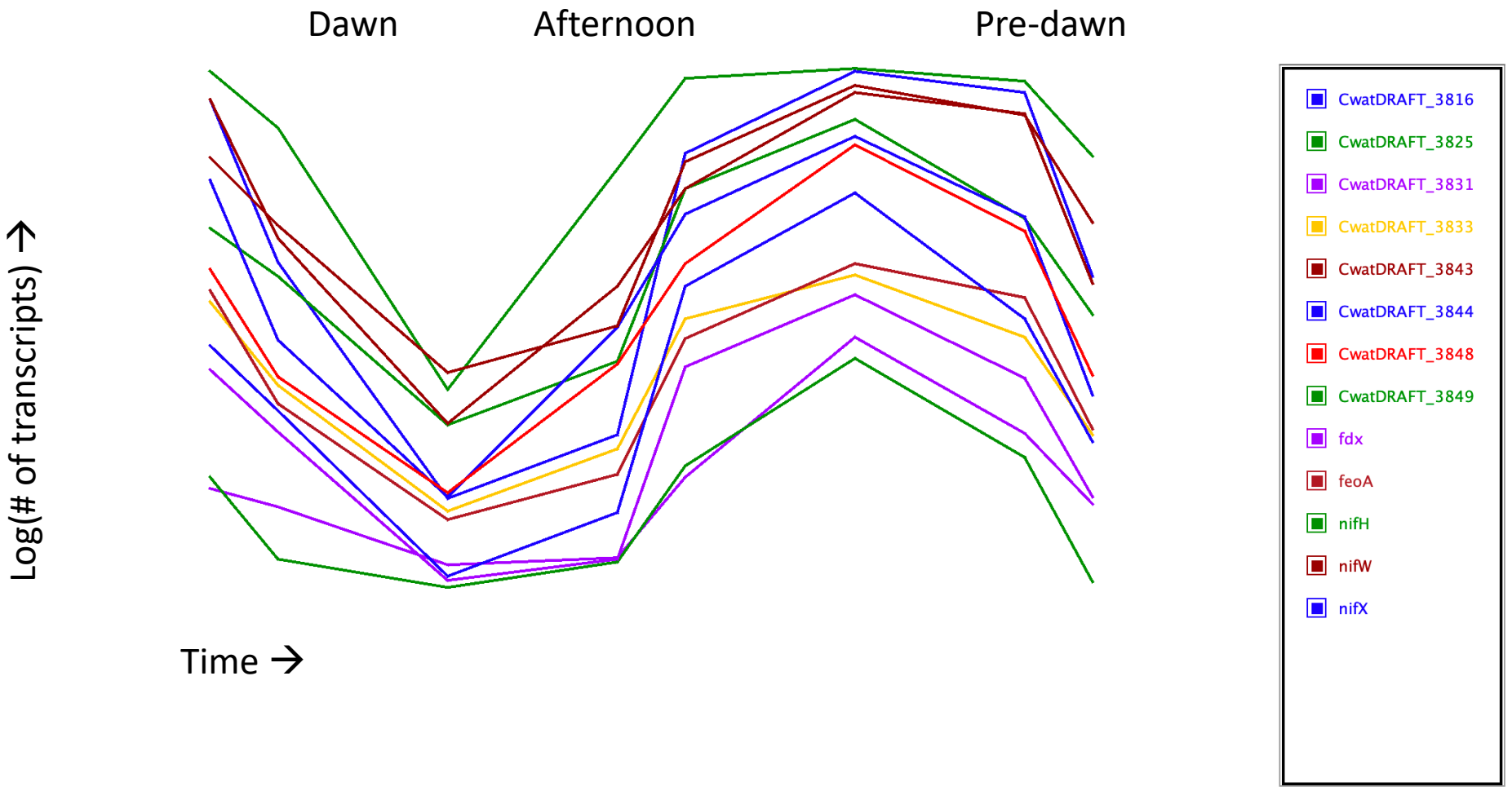
3 possible strategies for expressing photosystem genes

- WORST: Express at constant level all the time ("constitutive" expression)
  - Maximum waste: protein made at sunset will mostly be gone by sunrise
- BETTER: Sunrise triggers expression, sunset turns it off
  - Lost opportunity at, and shortly after, dawn: first no proteins, then only a little
  - Wasted manufacture at, and shortly before, sunset: no sunlight for those proteins to harvest
- BEST: Just-in-time manufacturing
  - Start production a few hours before sunrise
  - End a few hours before sunset
  - Need an alarm clock

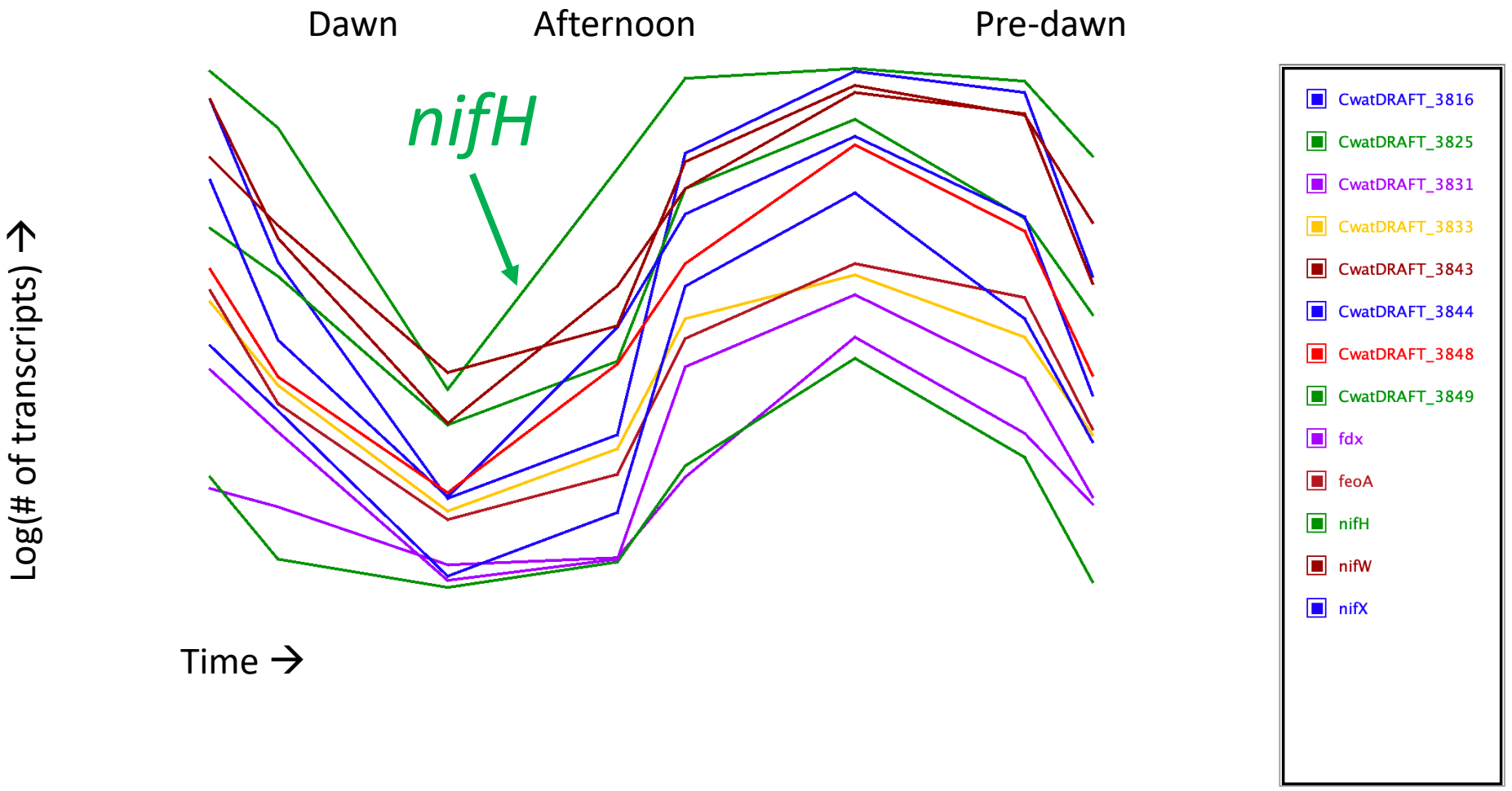
4 photosystem genes in marine bacterium *Crocosphaera watsonii*



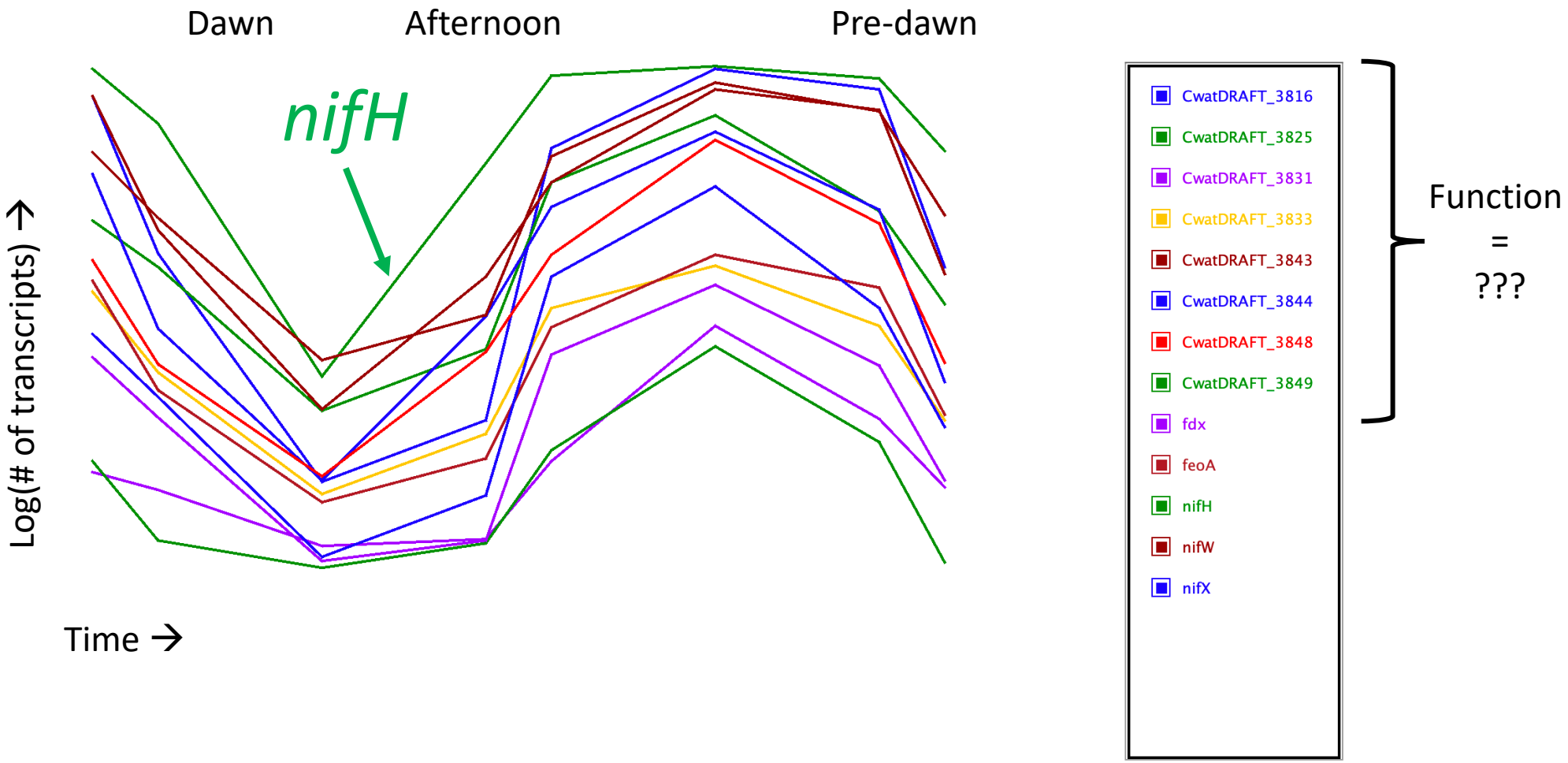
Do any other marine prokaryote genes show diel expression?



Do any other marine prokaryote genes show diel expression?

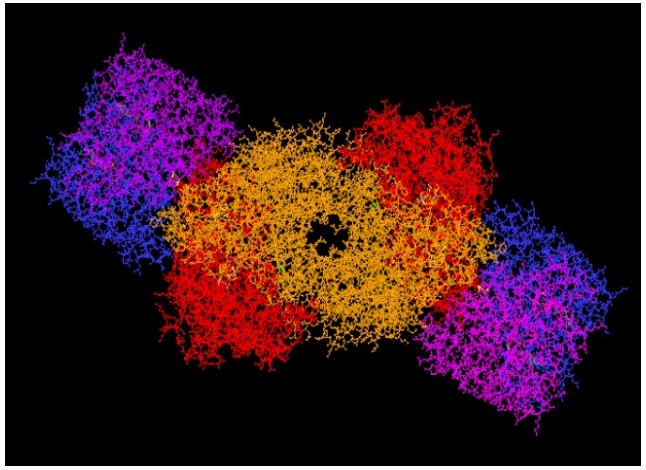


Do any other marine prokaryote genes show diel expression?



Why do nitrogen fixation genes need to fluctuate like this?

- Hint: it's not because they need sunlight.
- They need ATP, which is produced from sunlight by photosynthesis.



$$N_2 + 8e^- + 16ATP + 8H^+ \rightarrow 2NH_3 + 16ADP + 16P_i + H_2$$

- But ATP sticks around, doesn't need sunlight.

The bad news about nitrogenase

Long ago, when the air was clean, an organism invented a new way to use an abundant energy source.

The bad news about nitrogenase

Long ago, when the air was clean, an organism invented a new way to use an abundant energy source.

And that changed everything. Even the color of the sky.

The bad news about nitrogenase

Long ago, when the air was clean, an organism invented a new way to use an abundant energy source.

And that changed everything. Even the color of the sky.

I know what you're thinking, but this time you're wrong…

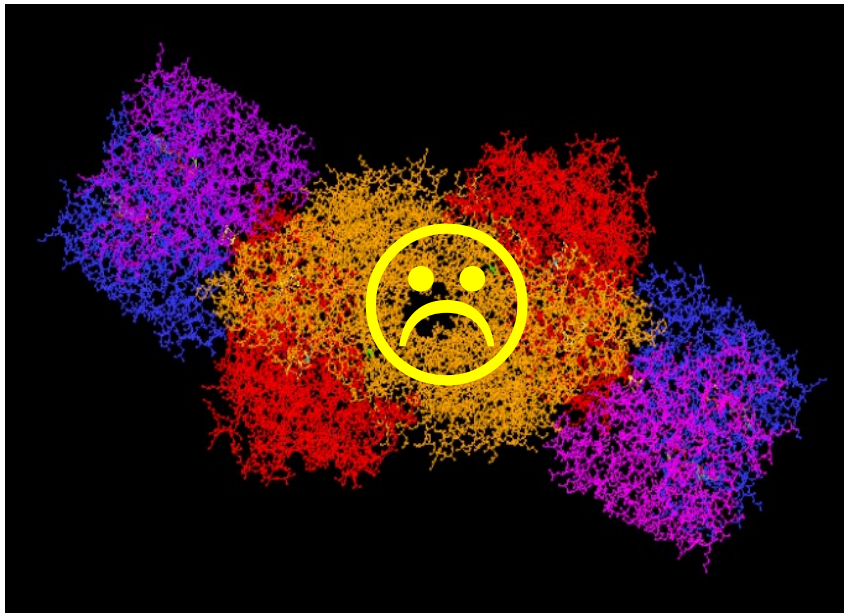
Before:

After:

O2 is highly reactive

Iron + O2 ➔ rust

Dry forest + O2 ➔ wildfire



Nitrogenase + O2
➔
broken nitrogenase

Before:

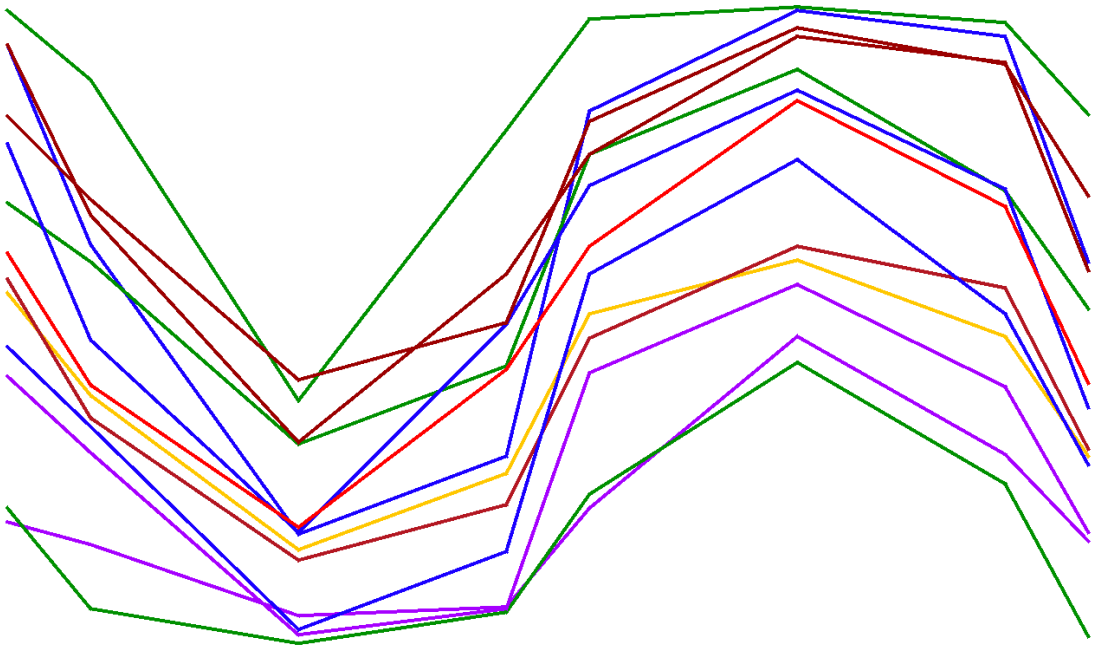
- Nitrogen fixation is energetically expensive
- Owning fixed nitrogen confers membership in the low levels of the food web

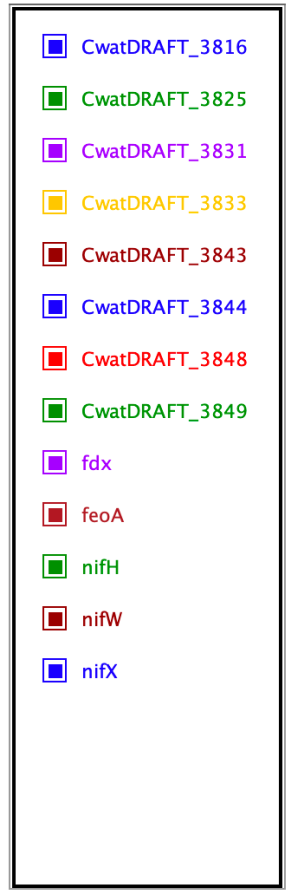
After:

- 🡸 All that, and you have to hide from oxygen: sequestration in time or space

Sequestration in time

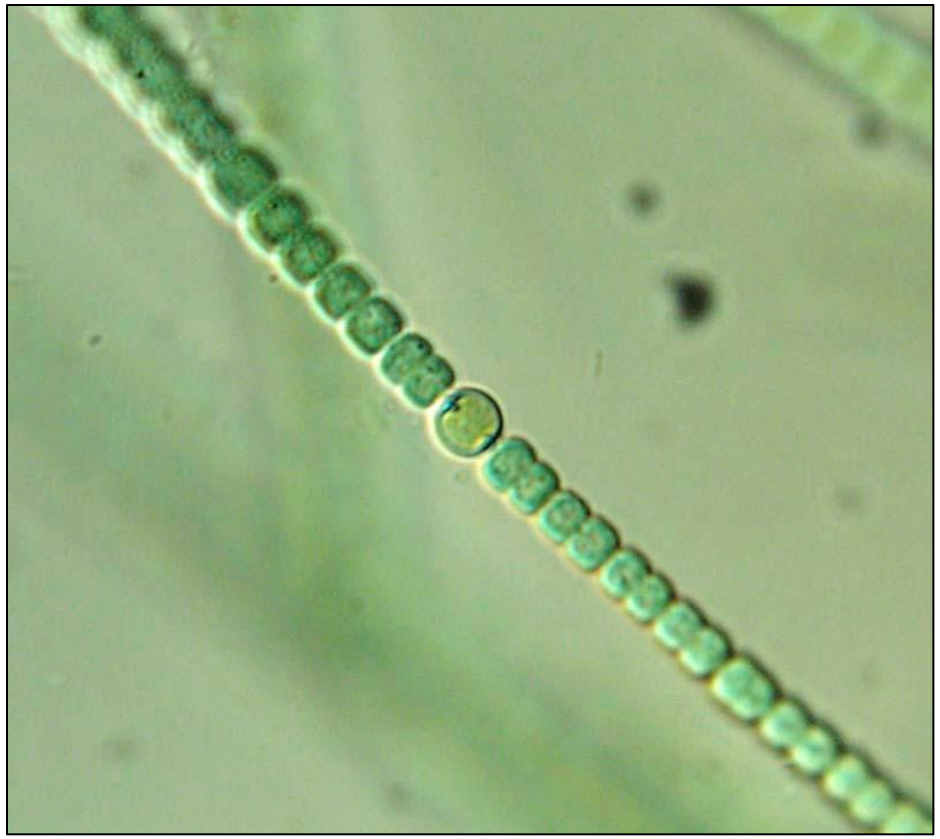
- Only fix nitrogen when you're not photosynthesizing





Sequestration in space: Heterocysts

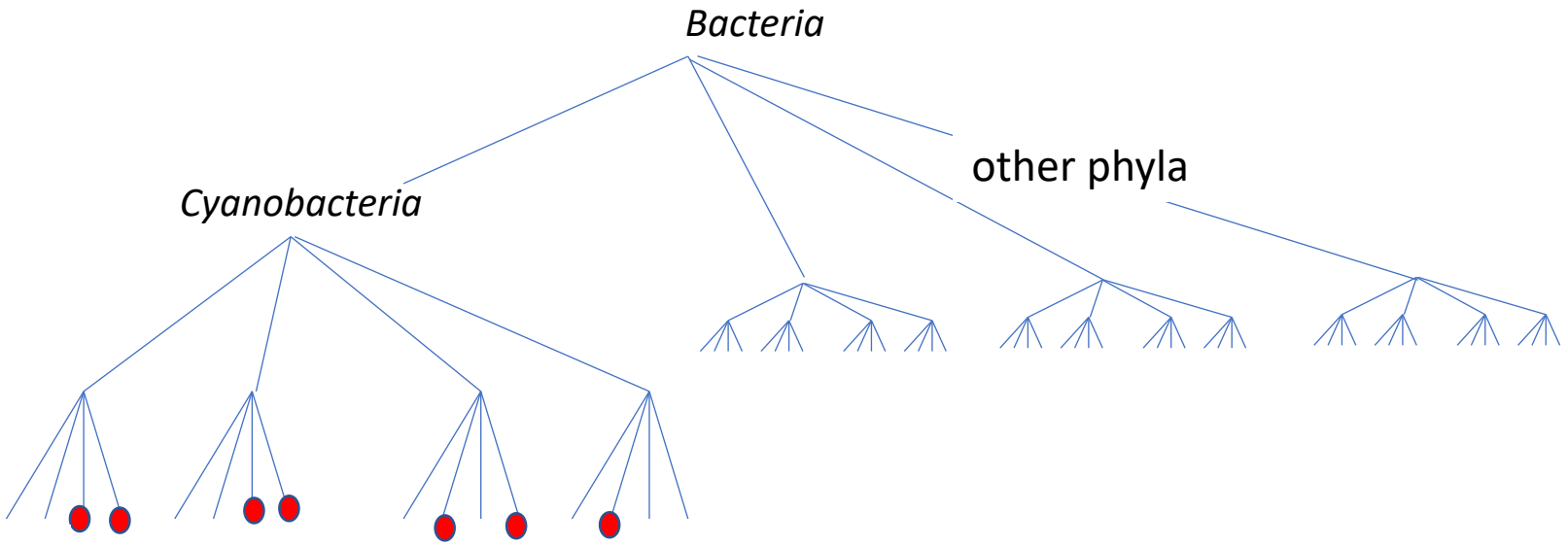
- Strategy used by filamentous cyanobacteria
- Most cells only photosynthesize
- A few cells (“heterocysts”) specialize in nitrogen fixation and don’t photosynthesize



Genus *Anabaena*

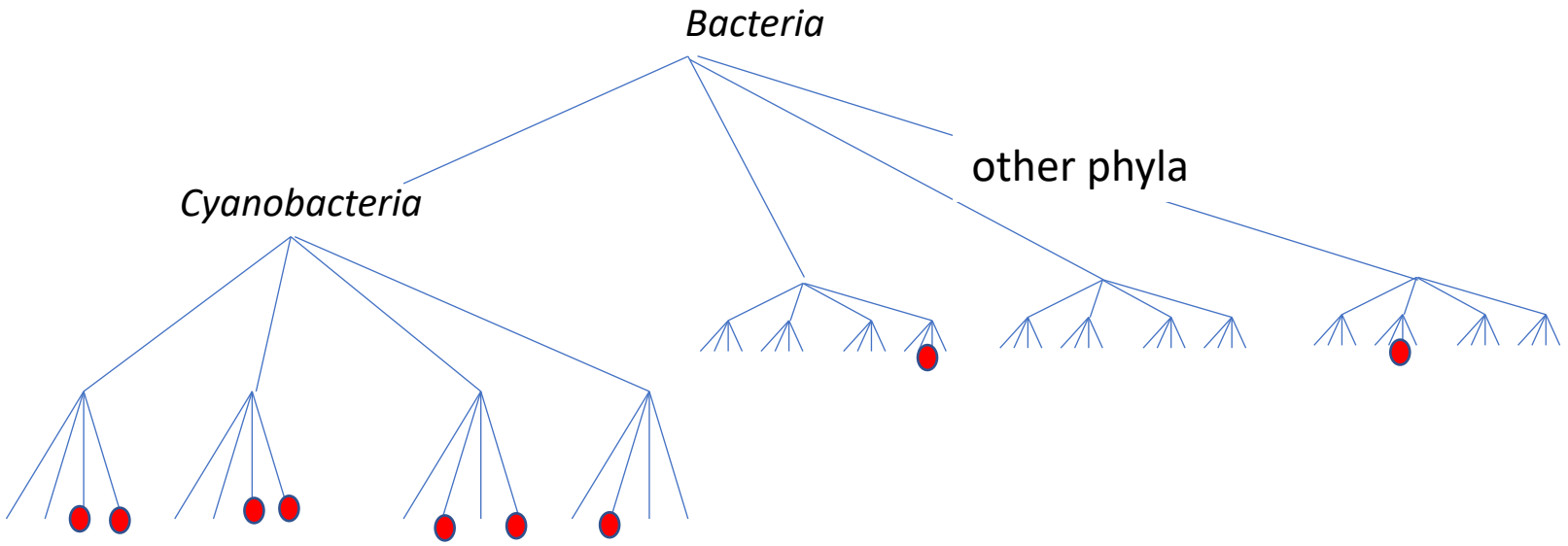
Today: nitrogen fixation is sparsely distributed in the prokaryote tree of life

- Until ~8 years ago: Phylum *Cyanobacteria* were believed to be the only diazotrophs (nitrogen fixers) due to need for photosynthesis)



Today: nitrogen fixation is sparsely distributed in the prokaryote tree of life

- Until ~8 years ago: Phylum *Cyanobacteria* were believed to be the only diazotrophs (nitrogen fixers) due to need for photosynthesis)
- Today: Other phyla may also (sparsely) contain diazotrophs



2016:

Trends in Microbiology

CellPress

Review

Marine Non-Cyanobacterial Diazotrophs: Moving beyond Molecular Detection

Deniz Bombar,[1] Ryan W. Paerl,[1] and Lasse Riemann[1,*]

916 Trends in Microbiology, November 2016, Vol. 24, No. 11 http://dx.doi.org/10.1016/j.tim.2016.07.002

"The belief is that cyanobacteria are the only relevant N2-fixing (diazotrophic) organisms. It has, however, now become evident that non-cyanobacterial diazotrophs, bacteria and archaea with ecologies fundamentally distinct from those of cyanobacteria, are widespread and occasionally fix N2 at significant rates."

2021

The ISME Journal (2021) 15:124–128
https://doi.org/10.1038/s41396-020-00765-1

ISME

ARTICLE

Check for updates

Intriguing size distribution of the uncultured and globally widespread marine non-cyanobacterial diazotroph Gamma-A

Francisco M. Cornejo-Castillo[1] · Jonathan P. Zehr[1]

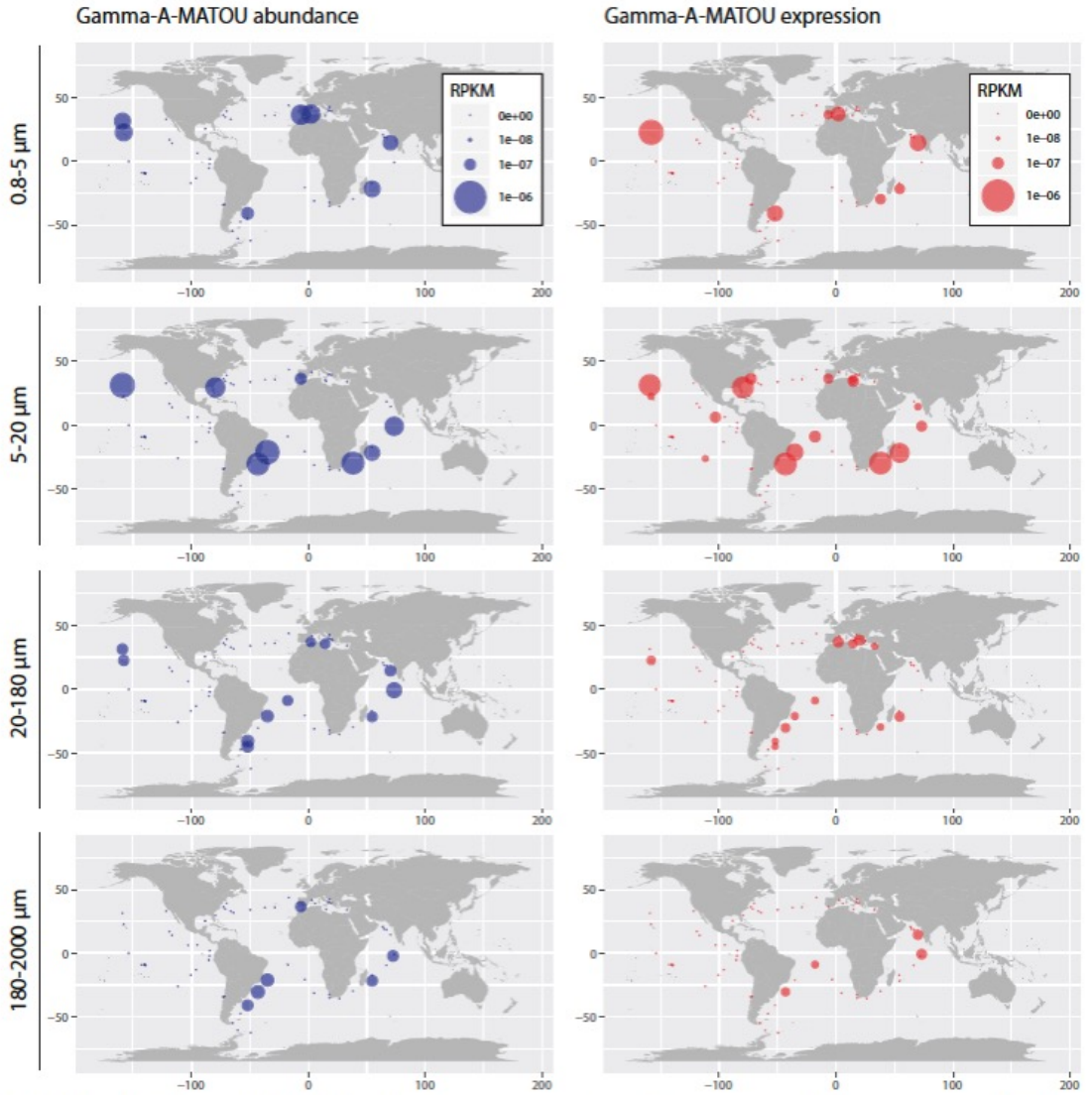
Received: 10 February 2020 / Revised: 18 August 2020 / Accepted: 27 August 2020 / Published online: 11 September 2020

Abstract

Non-cyanobacterial diazotrophs (NCDs) have recently emerged as potentially important contributors to marine nitrogen fixation. One of the most widely distributed NCDs is Gamma-A, yet information about its autecology is still scarce and

Gamma-A

Gamma-A Metagenomic Reads



Gamma-A Transcripts

Fig. 1 Distribution of Gamma-A-MATOU in the surface ocean. Abundance (metagenome-based; left panel) and expression (metatranscriptome-based; right panel) of Gamma-A-MATOU across size fractions are shown. The area of the bubble is proportional to the abundance of metagenomic reads (blue) or transcripts (red) of Gamma-A-MATOU for each sample. Abundances of metagenomic and metatranscriptomic reads are expressed as RPKM (Reads Per Kilobase covered per Million of mapped reads).

2021

FEMS MICROBIOLOGY REVIEWS

Article Contents

JOURNAL ARTICLE CORRECTED PROOF

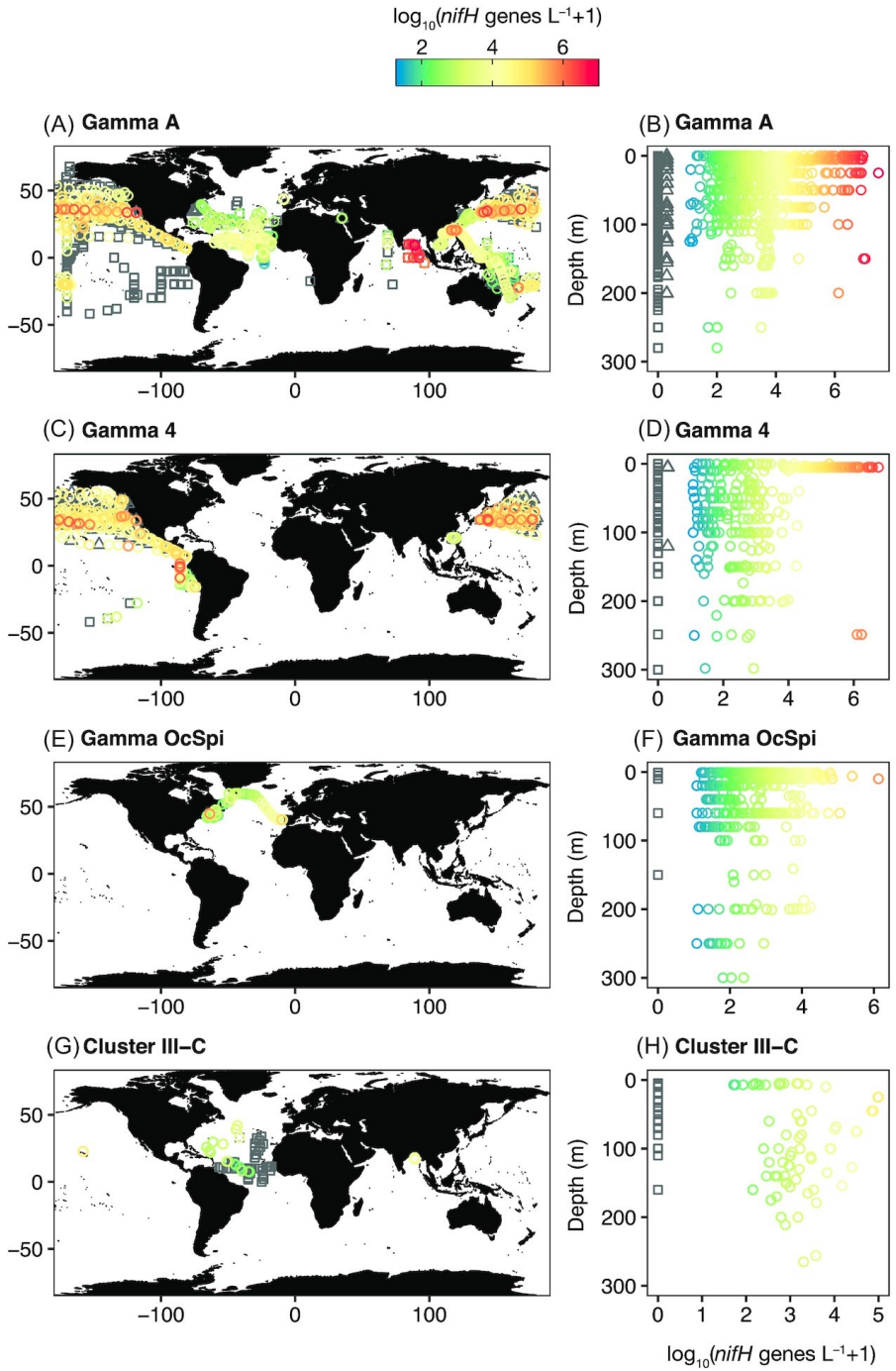
Non-cyanobacterial diazotrophs: global diversity, distribution, ecophysiology, and activity in marine waters

Kendra A Turk-Kubo ✉, Mary R Gradoville, Shunyan Cheung, Francisco M Cornejo-Castillo, Katie J Harding, Michael Morando, Matthew Mills, Jonathan P Zehr Author Notes

FEMS Microbiology Reviews, fuac046, https://doi.org/10.1093/femsre/fuac046

Published: 23 November 2022 **Article history** ▾

Gamma-A, Gamma 4, Gamma OcSpi, Cluster III-C



→

The hunt for nitrogen fixing prokaryotes continues.