Marine microeukaryote metatranscriptomics





Natalie Cohen

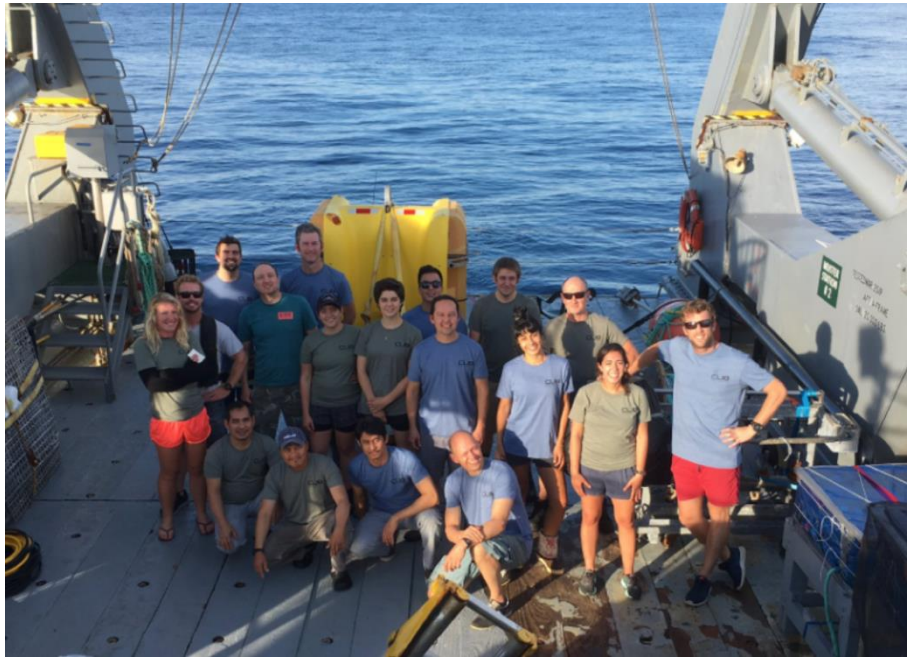
Croucher Summer Course 2025



Natalie Cohen
BS Penn State – Biology
Ph.D UNC Chapel Hill – Marine Science
Postdoc training – Woods Hole Oceanographic
Assistant Professor @ UGA – Skidaway Institute

Hobbies:
Eating
Aerial gymnastics
Running

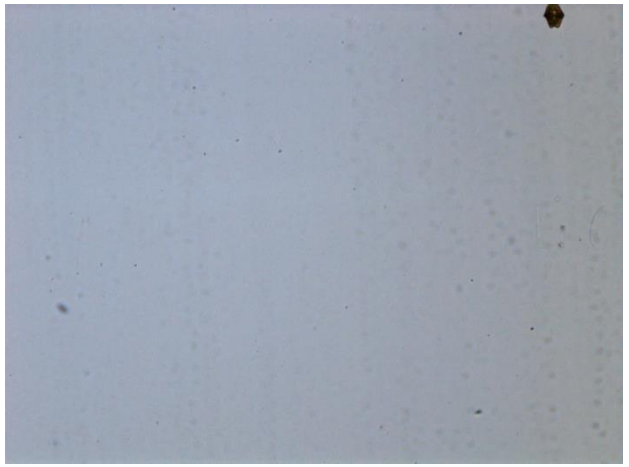






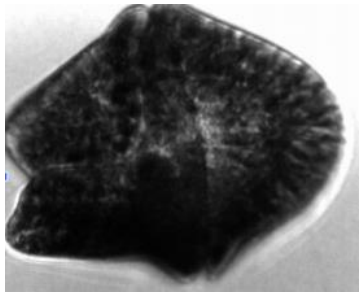
Monitoring harmful algal blooms

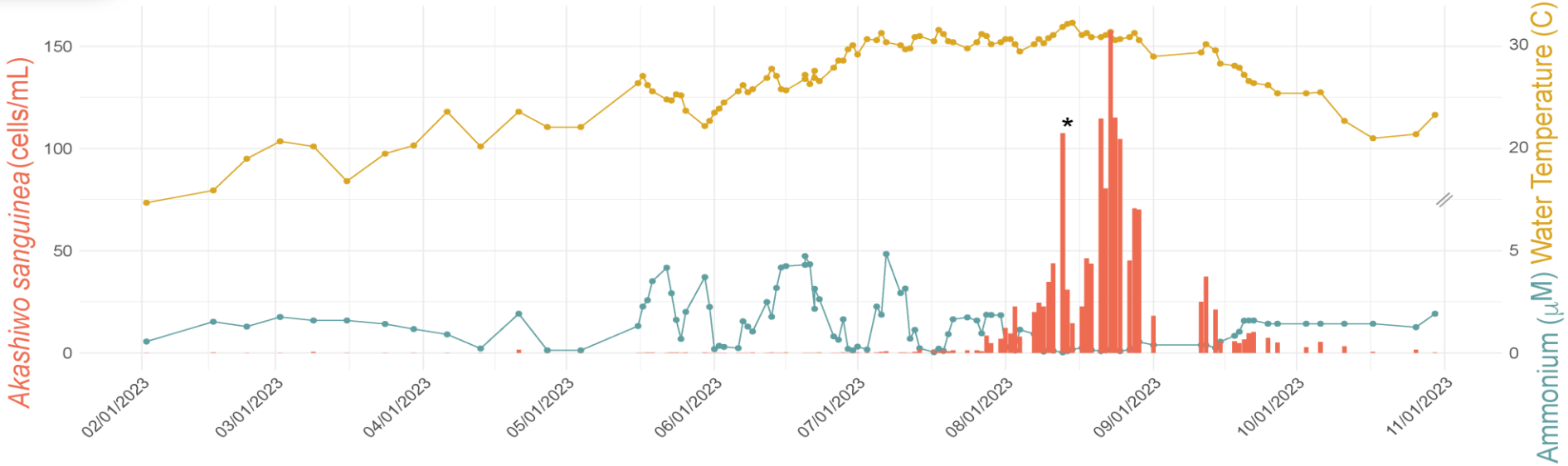




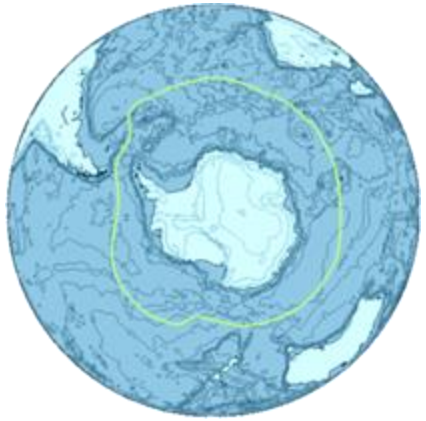
Akashiwo sanguinea



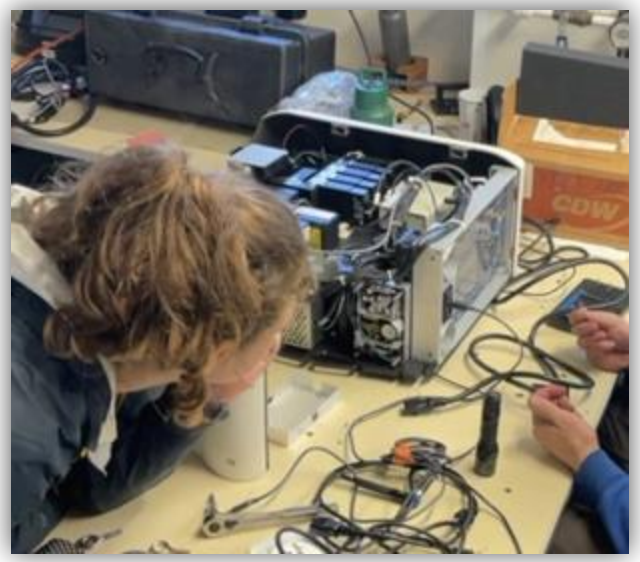


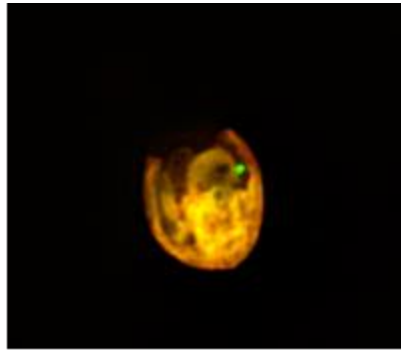


Examining mixotrophy under nutrient limitation



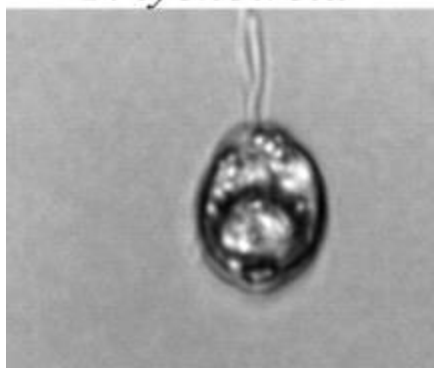
Southern Ocean is low in trace metals (Fe, Mn)

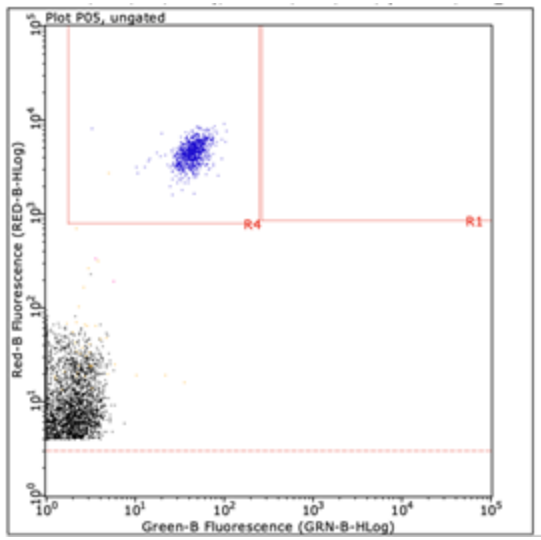


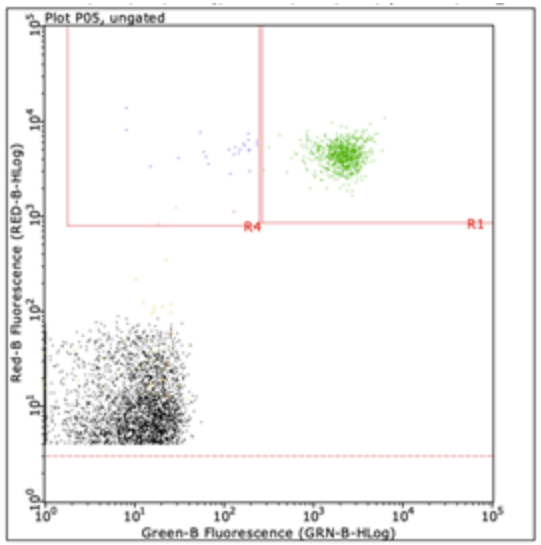




P. tychotreta

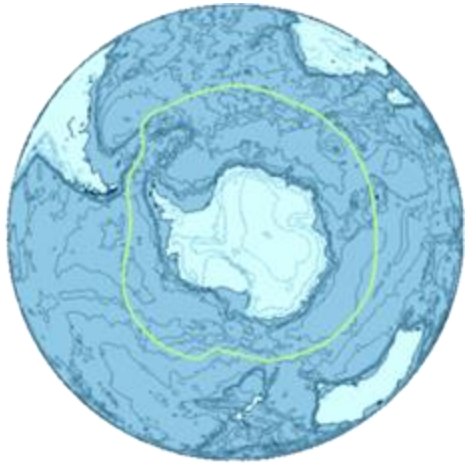






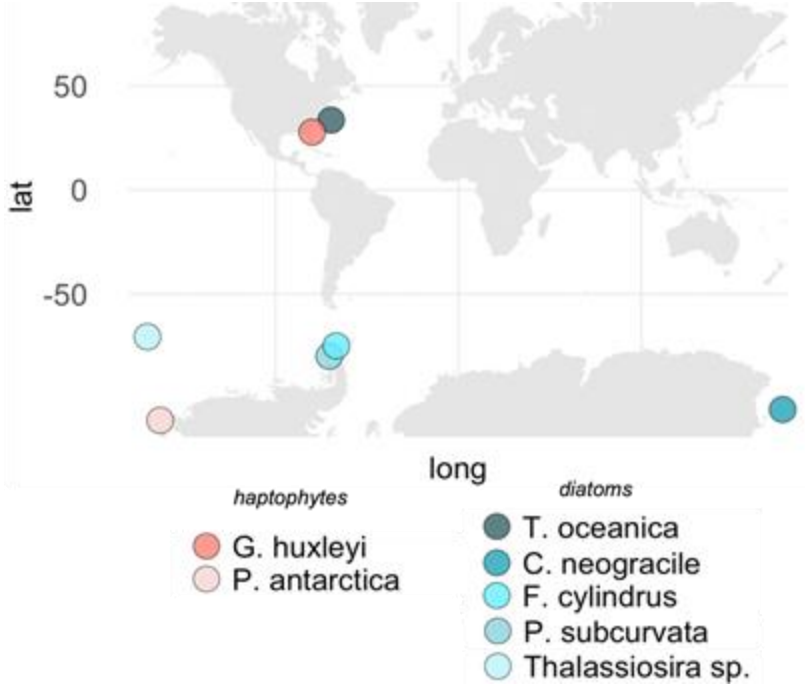


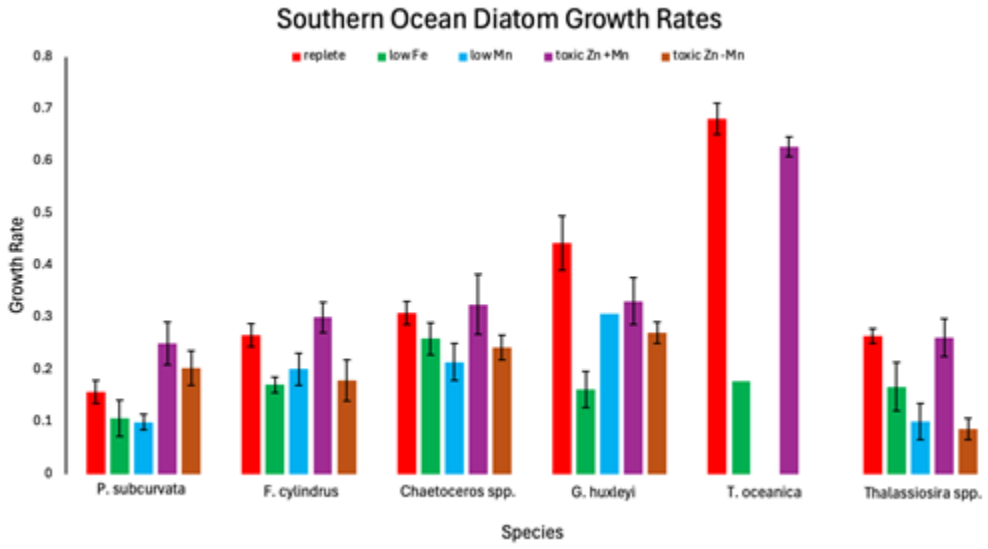
Metal physiology in the Southern Ocean



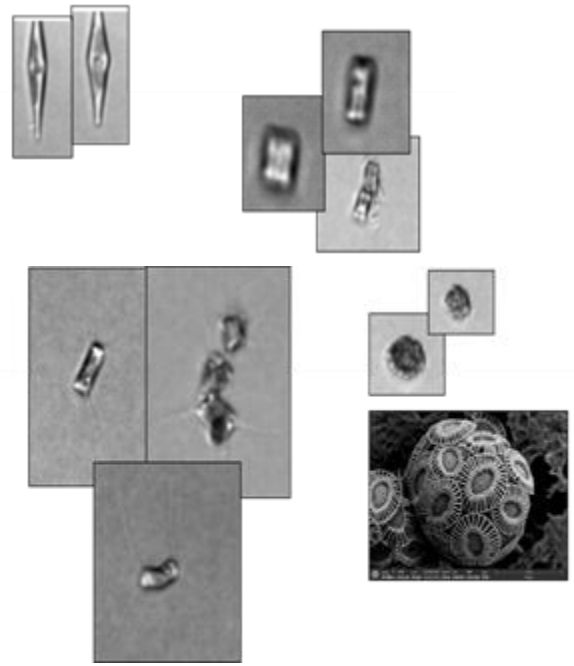
Southern Ocean is low in trace metals (Fe, Mn)

..and sometimes high in Zn!



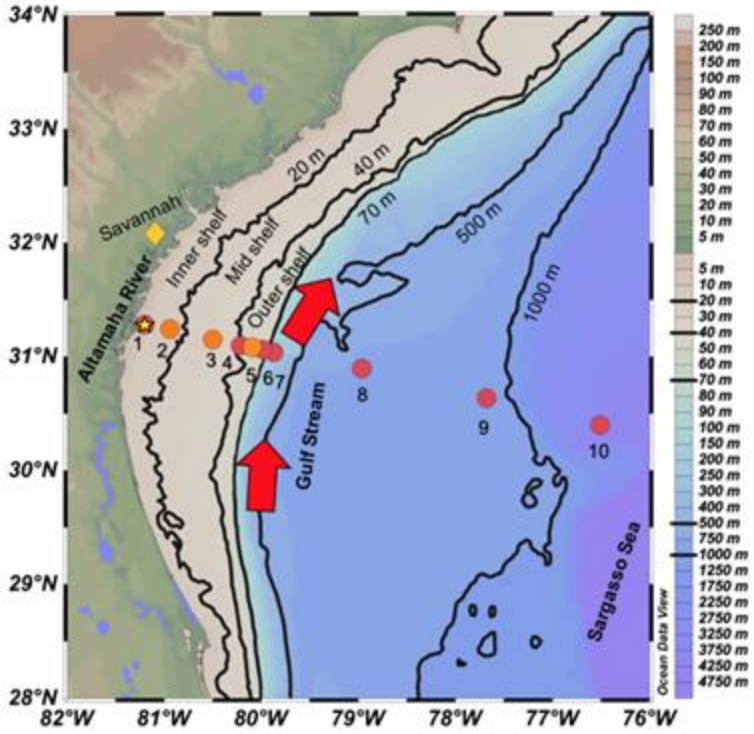


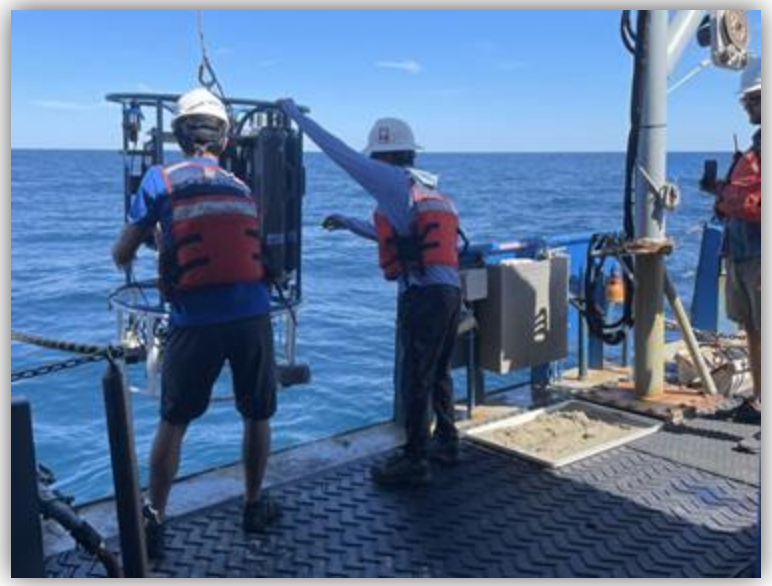




Plankton trophic interactions & metal ecology in continental shelf ecosystems

Adam Greer, Dan Ohnemus, Marc Frischer



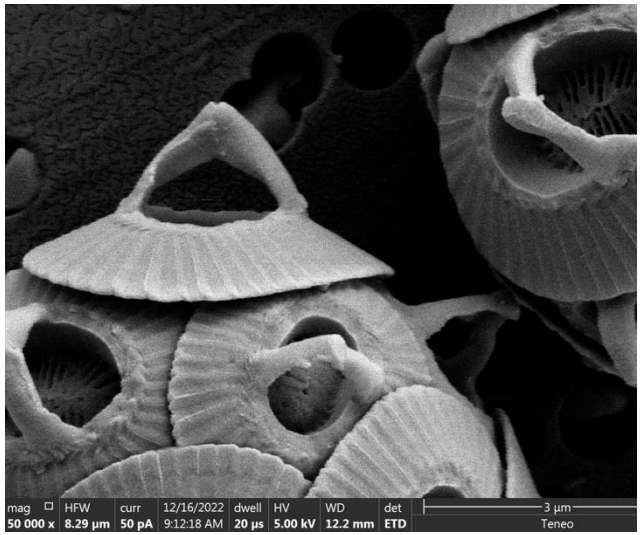


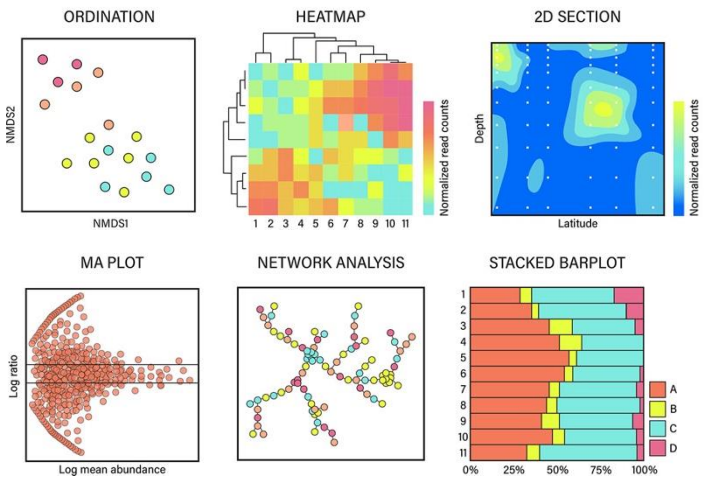


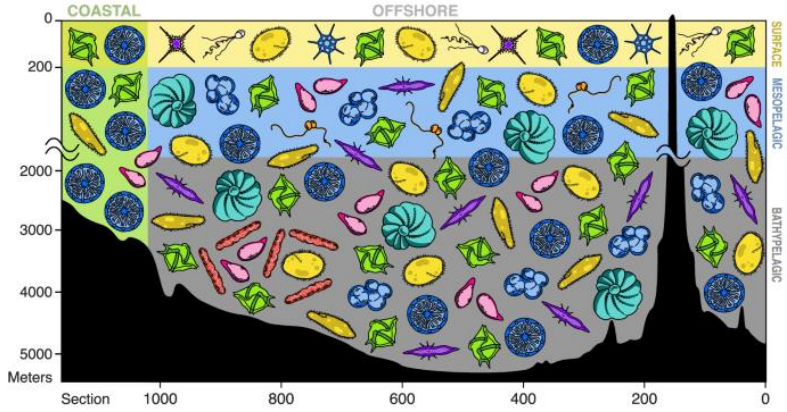




Outline







Inferring physiology from omics

Primary production, nutrient biomarkers, links to ocean biogeochemistry

Best practices in metatranscriptomics

From sample collection to data visualization

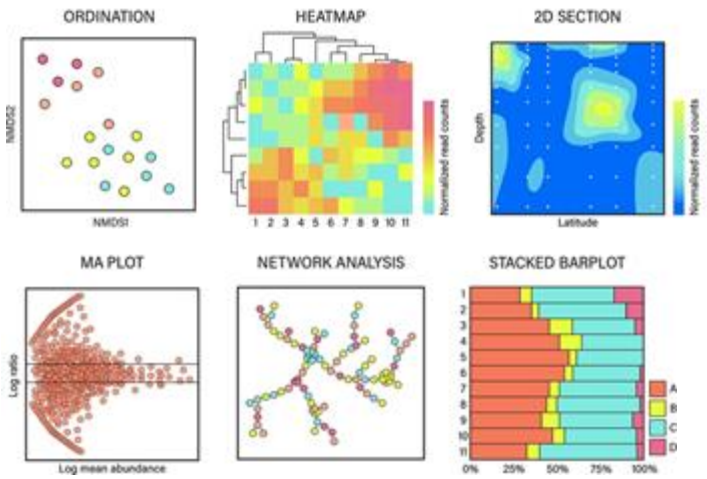
Microeukaryote metabolism across the ocean

Biogeography & ocean gradients

Outline

Inferring physiology from omics

Primary production, nutrient biomarkers, links to ocean biogeochemistry

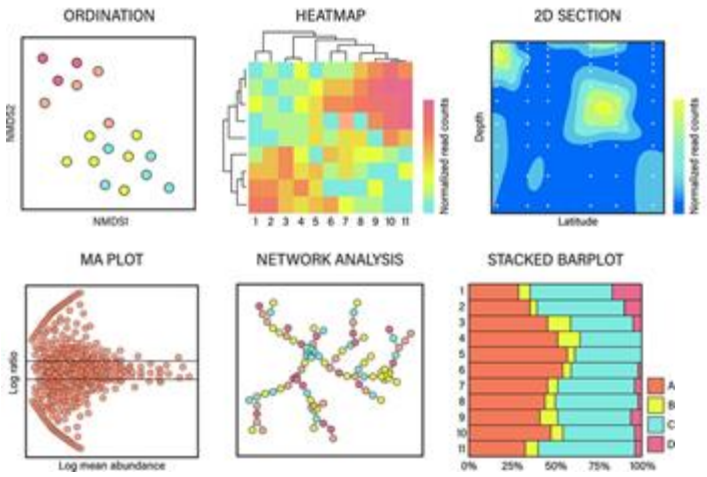


Best practices in metatranscriptomics

From sample collection to data visualization

Microeukaryote metabolism across the ocean

Biogeography & ocean gradients

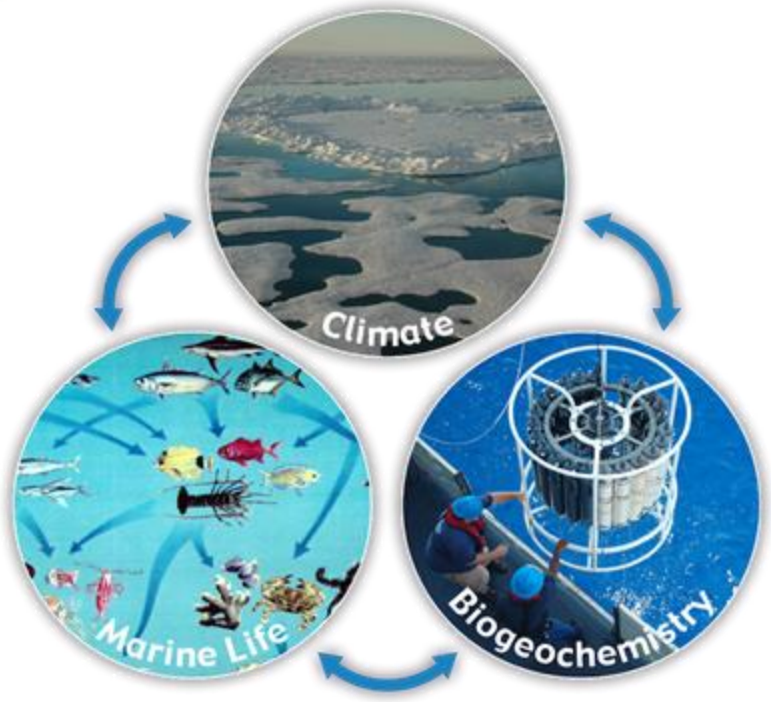


Best practices in metatranscriptomics

From sample collection to data visualization

- Intercomparison efforts
- Lab/field collection
- RNA preparation
- Bioinformatic pipeline
- Databases
- Normalization & differential expression
- Data visualization approaches

Ocean Carbon & Biogeochemistry (OCB) working group: **meta-eukomics**



Biological pump

Marine ecosystems

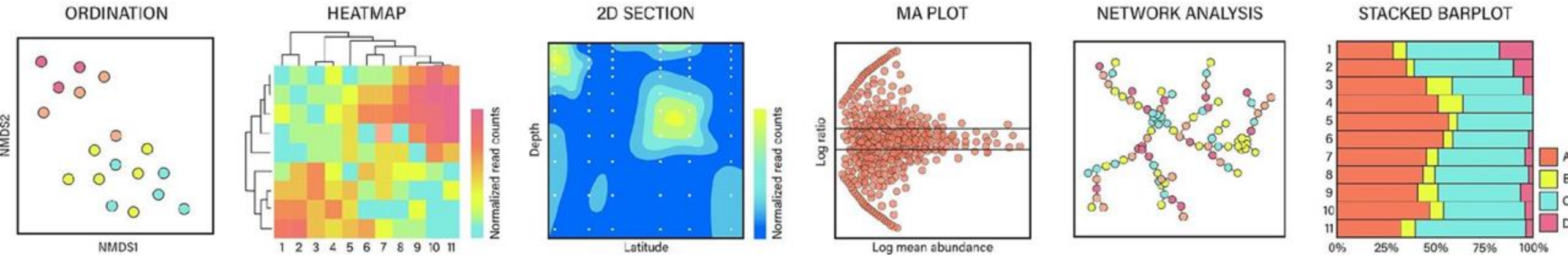
Ocean observing

Ocean carbon uptake

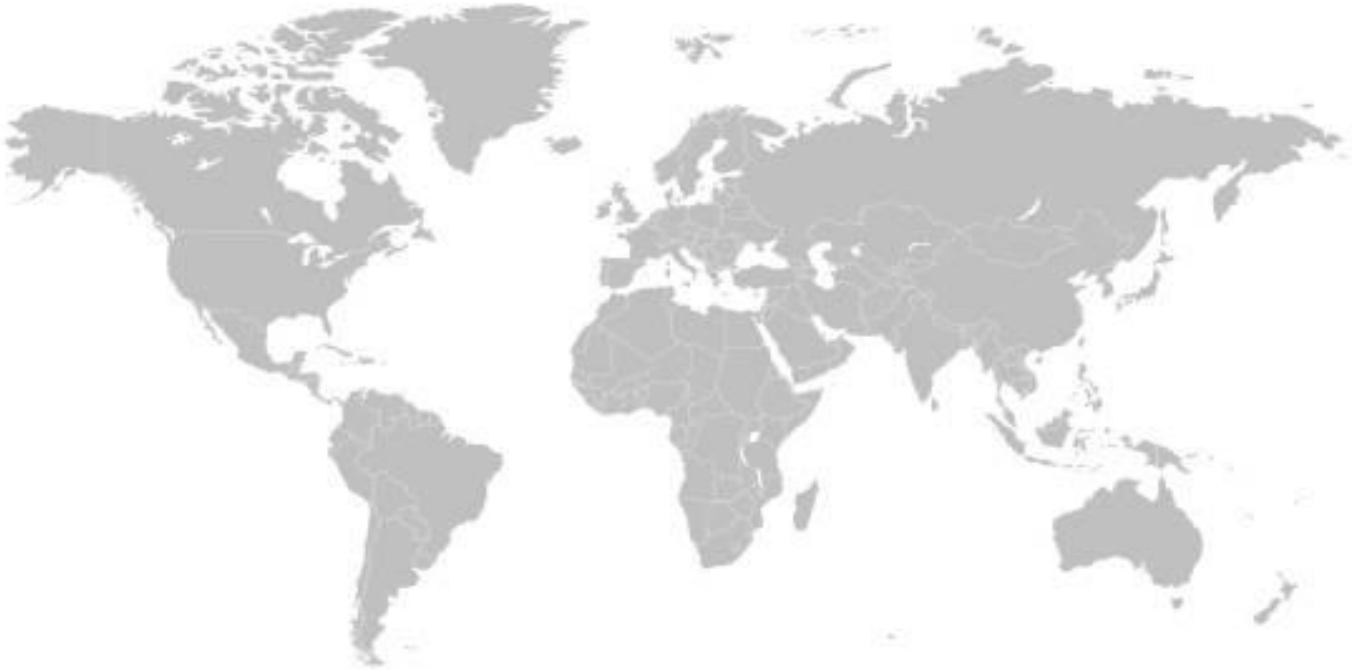
Changing ocean chemistry

Estuarine & coastal carbon fluxes

Website: ocb.org/metat-intercomparison/



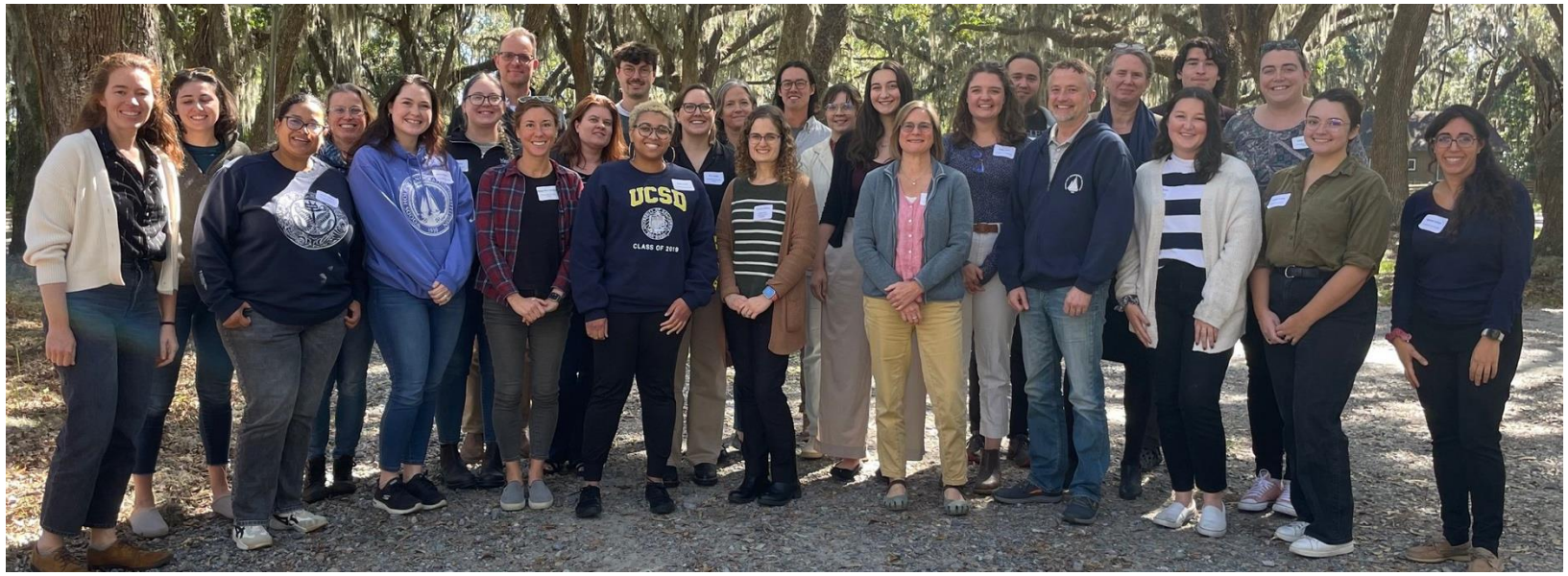
How does sample collection, RNA extraction, sequencing, and bioinformatics influence biological interpretations?



MetaT Intercomparison Working Group

Goal: Intercomparison of eukaryotic metatranscriptomic **RNA extraction procedures** and **bioinformatic analyses** by lab

OUTCOME: White paper summary of results and best practices



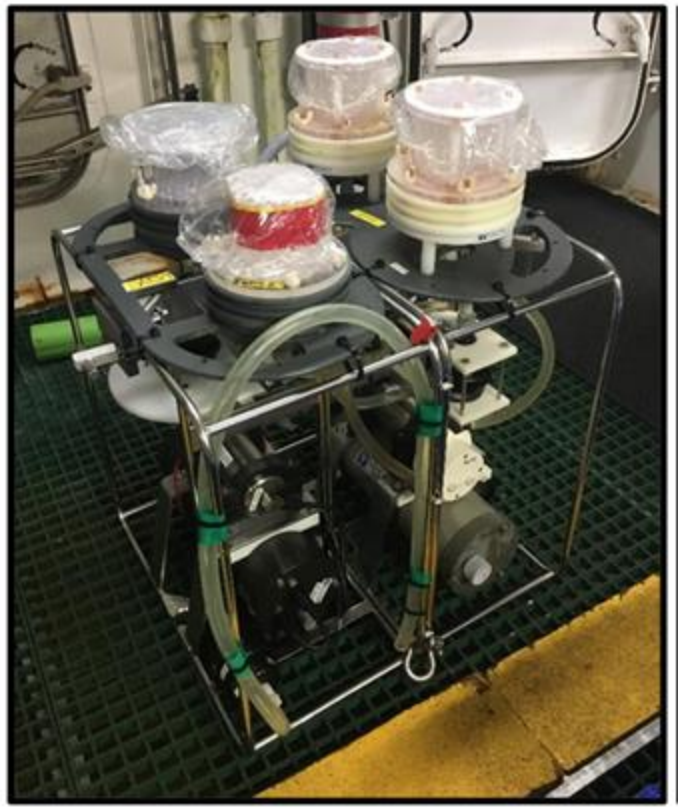
Community-wide goals:

- Learn how existing and developing omic tools can be used to further our understanding of microbial diversity and function
- Understand the utility and limitations of the methods
- Foster exchange between intercomparison efforts

Previous intercomp efforts

OCB Ocean nucleic acids 'omics intercalibration and standardization workshop report (Biogeoscapes):
Euks/proks/viruses, metaG, metaT, metaB

https://www.us-ocb.org/ocean-nucleic-acids-omics-workshop/



Mak Saito: OCB 2020 OMIS

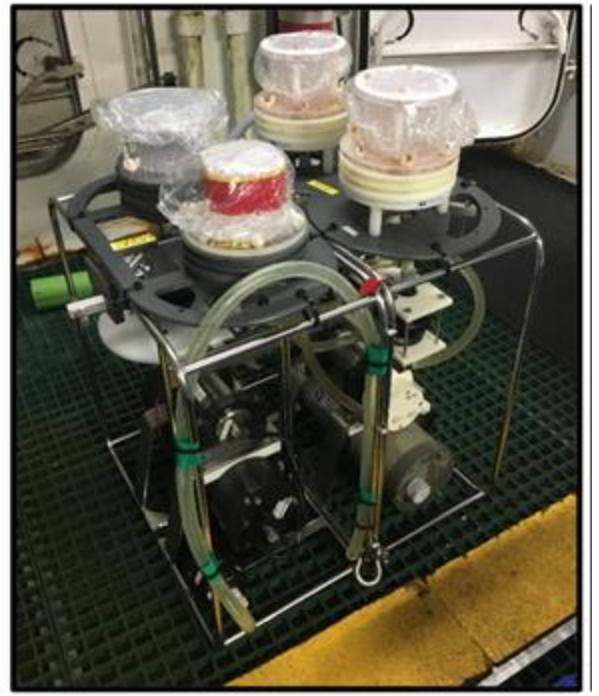
OCB Ocean metaproteomic intercomparison study:
Wet lab protein extractions and mass spectrometry

https://www.us-ocb.org/intercomparison-and-intercalibration-metaproteomics/

Saito et al. 2024, *Biogeosciences*

Intercomparison Plan

- Surface samples collected in May 2023 by Mak Saito on Costa Rica Dome Upwelling cruise

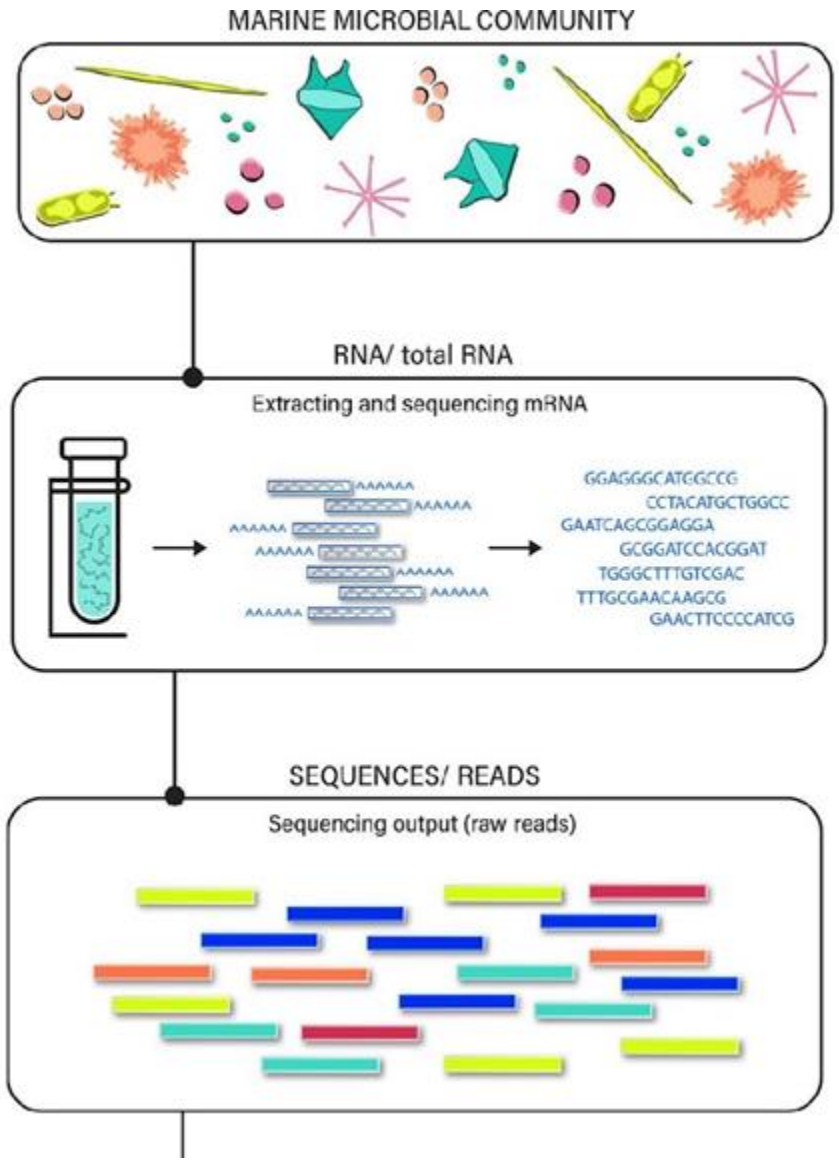


- 4 McLane pumps with 0.2 – 51 um filter fraction (Supor)
  Sliced into sections & sent to participating labs

- Illumina sequencing using an experimental design matrix
  - RNA extraction procedure
  - Poly-A selection vs. rRNA depletion
  - Sequencing facility

- Informatics intercomparison will be performed to determine extent of variability due to bioinformatic choices alone (assembly, annotation, etc)

Physical sample processing



Step 1. Sample collection:
Filtration mechanism, duration, filter type, preservation

Step 2. RNA extraction:
Sample handling, extraction kit, physical disruptions, DNA removal

Step 3. Sequence prep:
Library preparation, rRNA removal

Step 4. Sequencing:
Sequence facility, platform and throughput

Sample collection (lab & field)

- **Filtration mechanism: vacuum pump, peristaltic pump, in situ pumps & filtration duration**
- Filter material and porosity
- Preservation method: flash freezing, RNAlater, time in storage



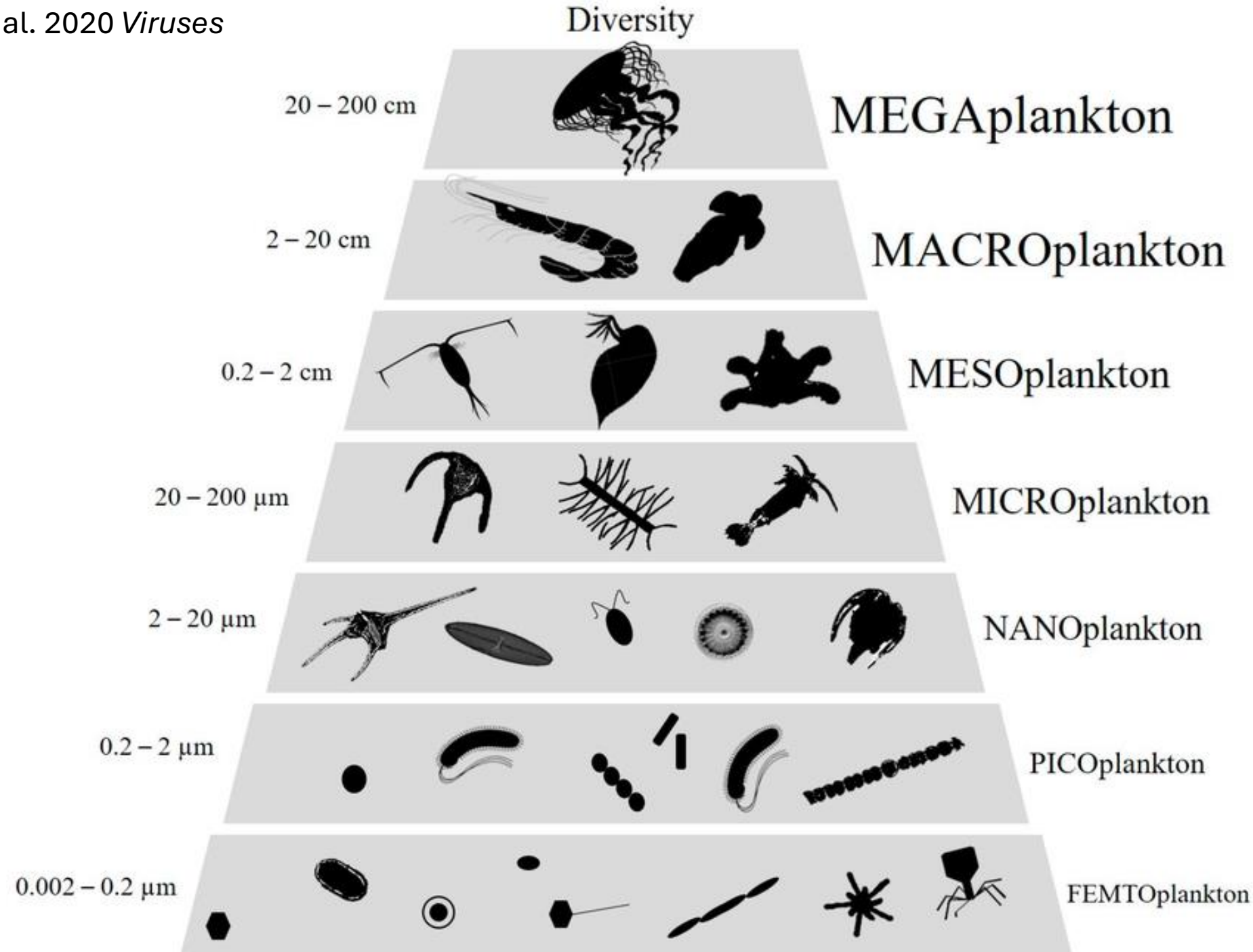




Sample collection (lab & field)

- Filtration mechanism: vacuum pump, peristaltic pump, in situ pumps & filtration duration
- **Filter material and porosity**
- Preservation method: flash freezing, RNAlater, time in storage

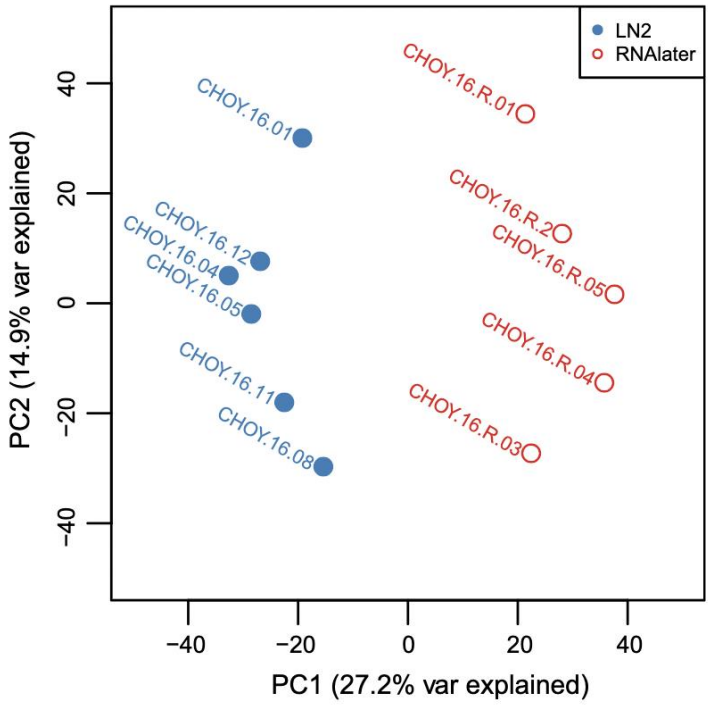
Colombet et al. 2020 *Viruses*



Sample collection (lab & field)

- Filtration mechanism: vacuum pump, peristaltic pump, in situ pumps & filtration duration
- Filter material and porosity
- **Preservation method: flash freezing, RNAlater, time in storage**

Passow et al. 2018 *Molecular Ecology Resources*

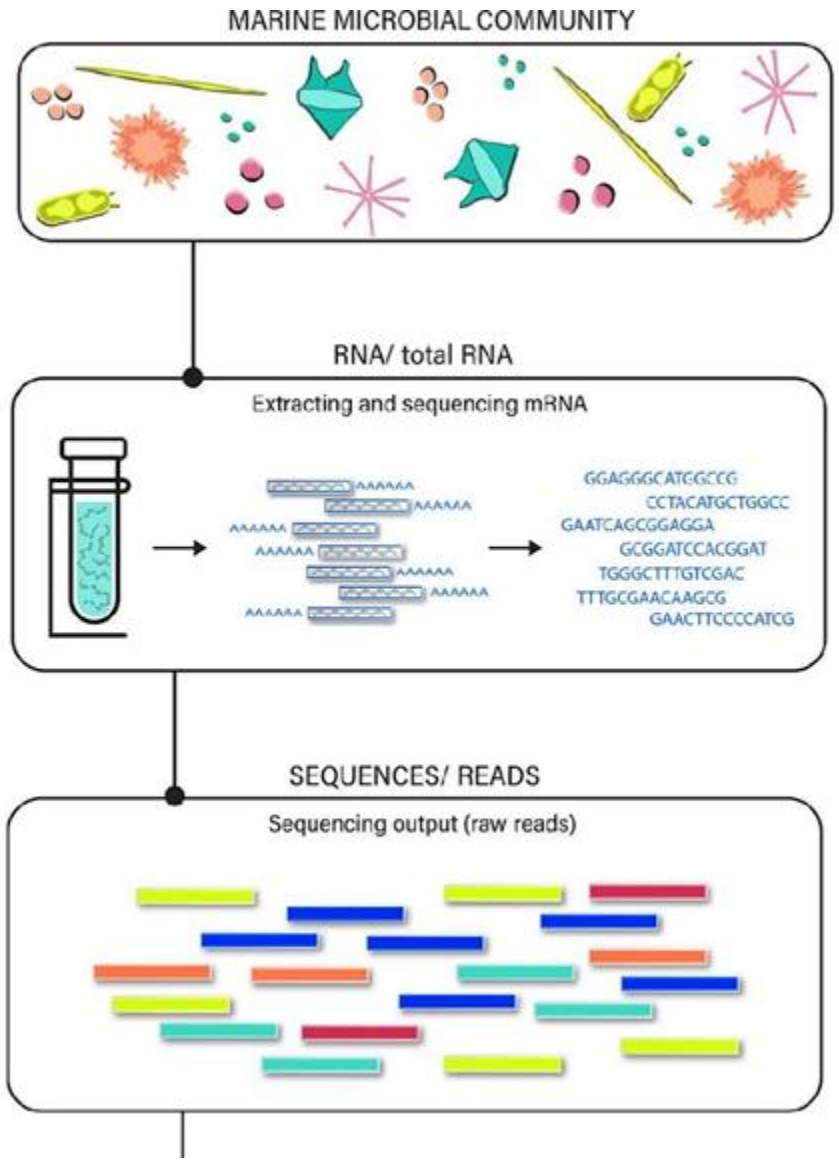


Liquid nitrogen: genes with elevated GC content showed higher expression

RNAlater: higher expression of genes involved in RNA processing

RNAlater may cause a physiological response that can bias results

Physical sample processing



Step 1. Sample collection:
Filtration mechanism, duration, filter type, preservation

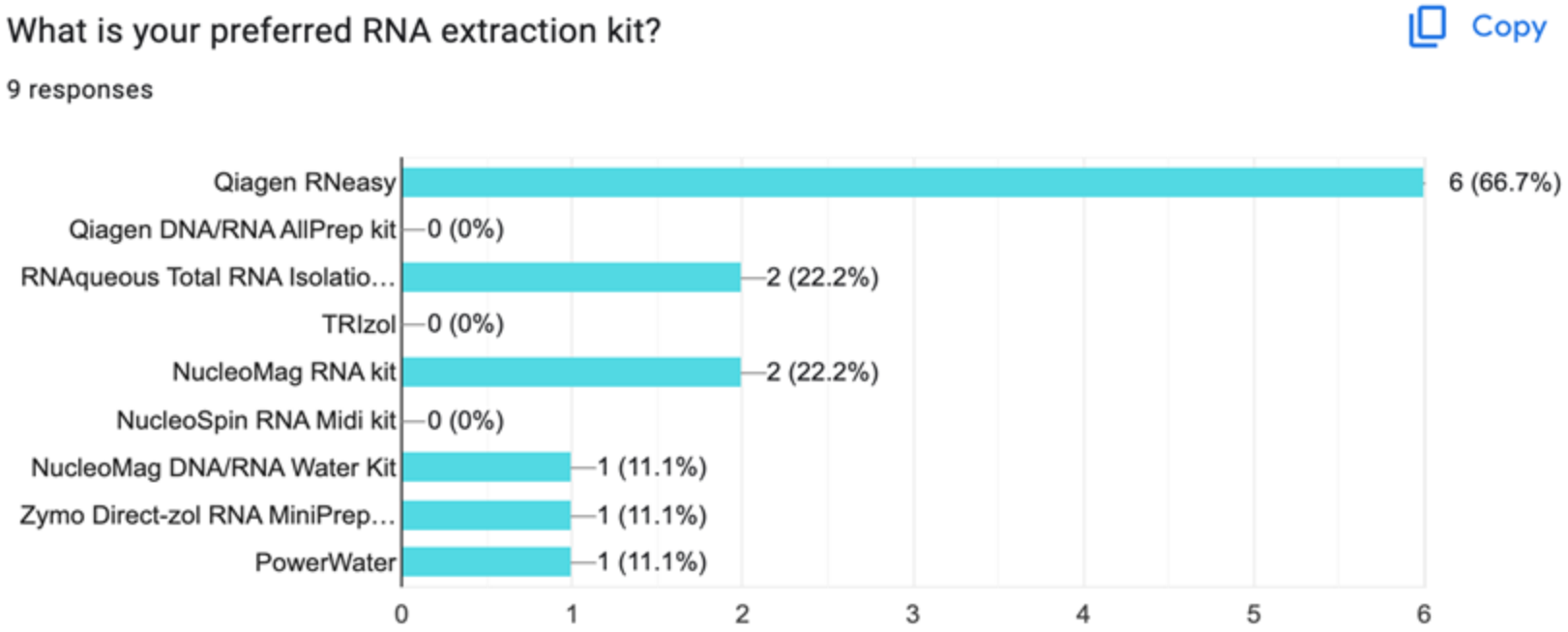
Step 2. RNA extraction:
Sample handling, extraction kit, physical disruptions, DNA removal

Step 3. Sequence prep:
Library preparation, rRNA removal

Step 4. Sequencing:
Sequence facility, platform and throughput

RNA extraction

- **Variety of kits commercially available**
- Common modifications: beads, physical disruption of cells
- DNA removal
- Quantitative standards



RNA extraction

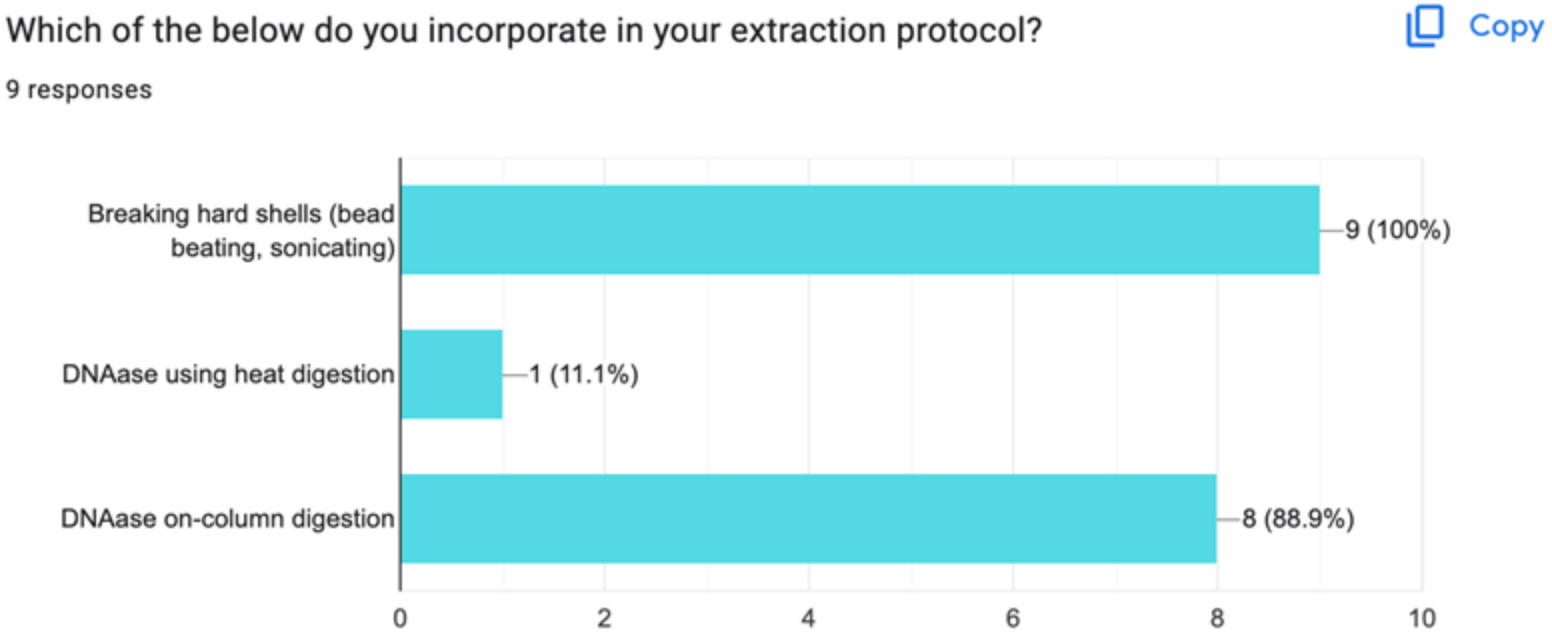
- Variety of kits commercially available
- **Common modifications: beads, physical disruption of cells**
- DNA removal
- Quantitative standards





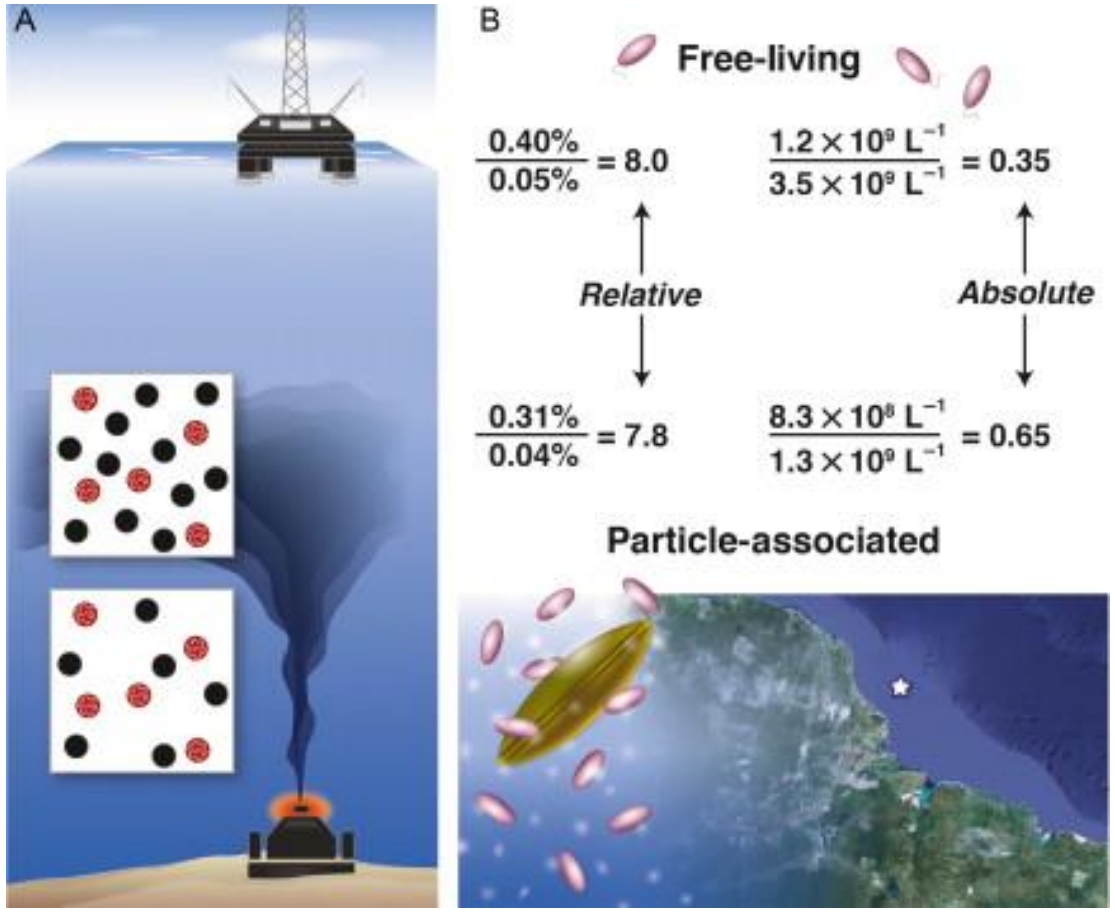
RNA extraction

- Variety of kits commercially available
- Common modifications: beads, physical disruption of cells
- **DNA removal: DNAase**
- Quantitative standards



RNA extraction

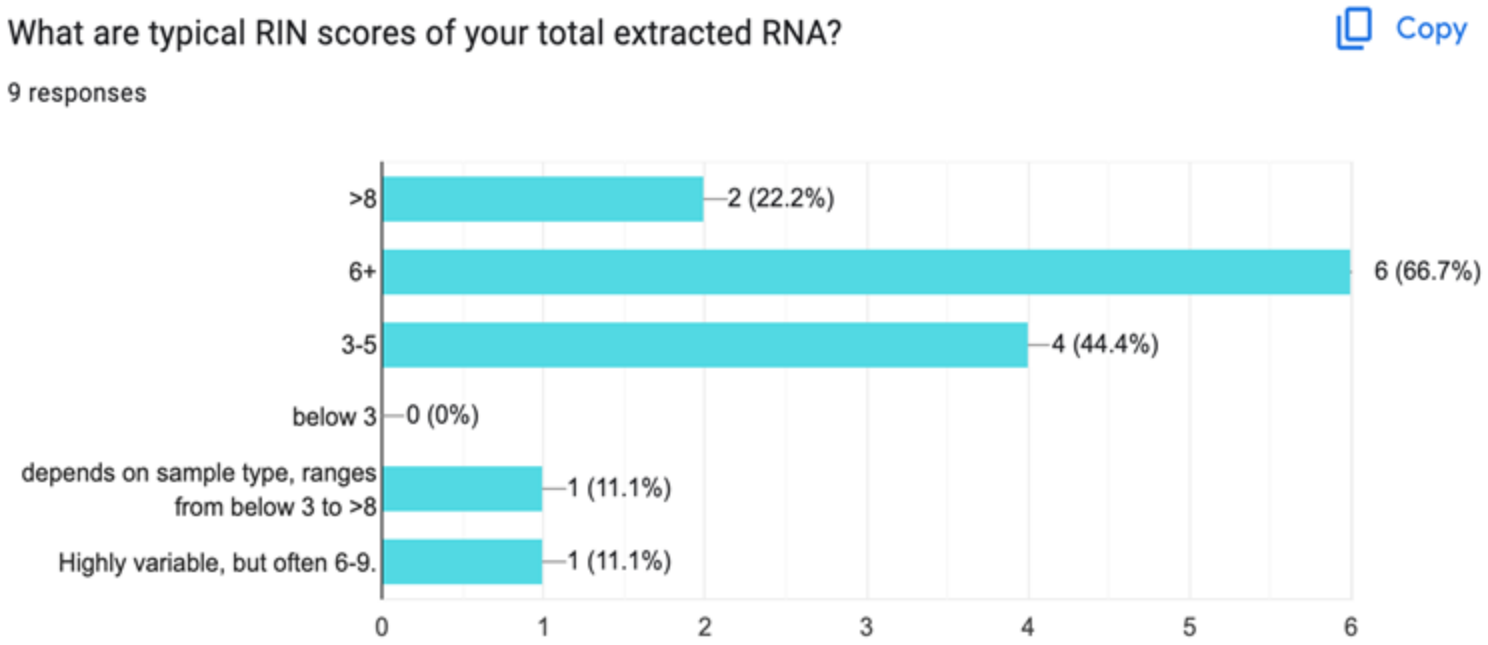
- Variety of kits commercially available
- Common modifications: beads, physical disruption of cells
- DNA removal
- **Quantitative standards**



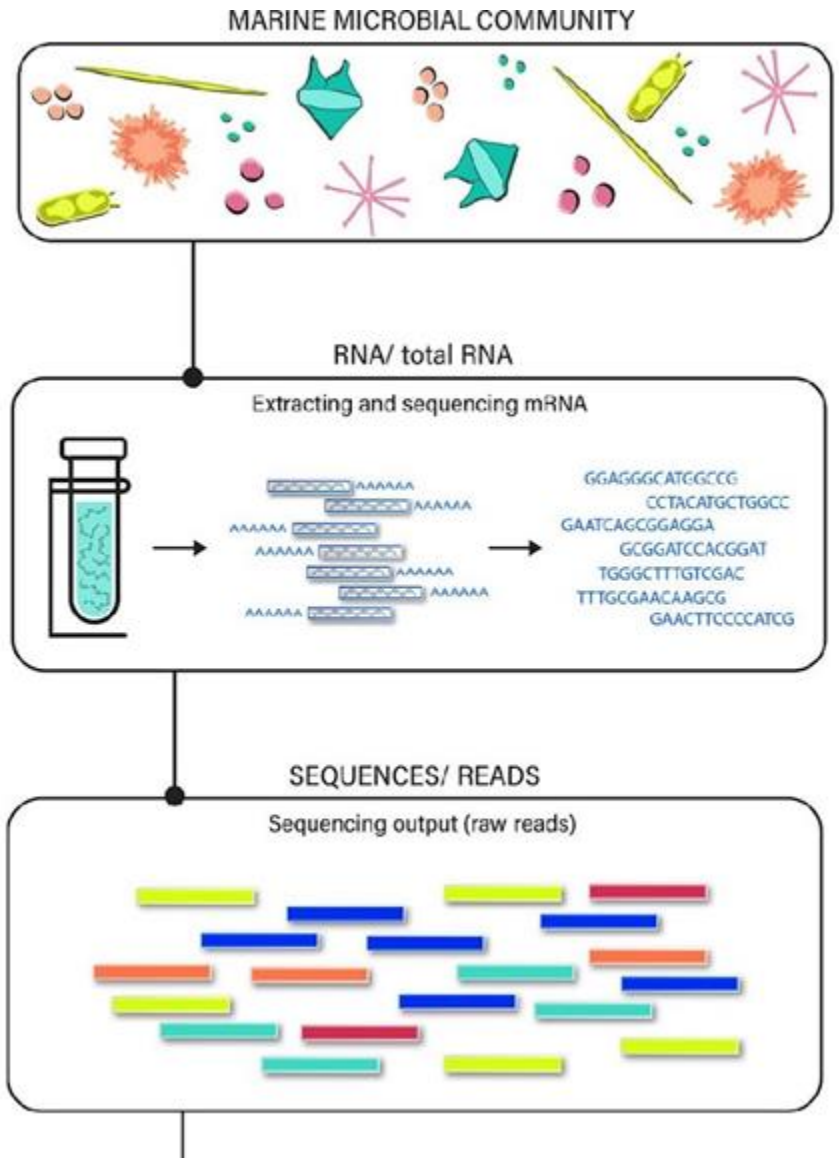
Satinsky et al. 2013 *Methods in Enzymology*

RNA extraction

- Variety of kits commercially available
- Common modifications: beads, physical disruption of cells
- DNA removal
- Quantitative standards
- **Assessing RNA quality**:
  - Nanodrop
  - Qubit
  - Bioanalyzer



Physical sample processing



Step 1. Sample collection:
Filtration mechanism, duration, filter type, preservation

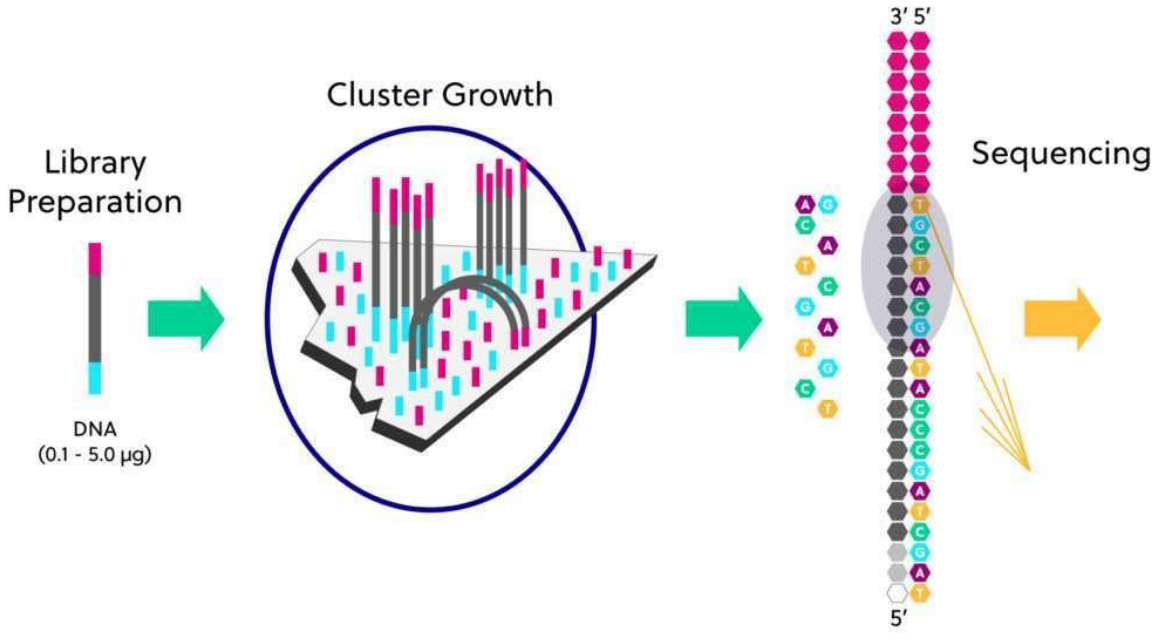
Step 2. RNA extraction:
Sample handling, extraction kit, physical disruptions, DNA removal

Step 3. Sequence prep:
Library preparation, rRNA removal

Step 4. Sequencing:
Sequence facility, platform and throughput

Sequencing options

- **Platforms**
  - **Short read sequencing (Illumina, AVITI)**
  - **Long read sequencing (PacBio, Oxford Nanopore)**
- Low input options
- mRNA enrichment



Sequencing options

- **Platforms**
  - **Short read sequencing (Illumina, AVITI)**
  - **Long read sequencing (PacBio, Oxford Nanopore)**
- Low input options
- mRNA enrichment

De Ronne et al. 2025 *Genome*

Table 1. Comparative analysis of triple-digest genotyping-by-sequencing performance and concordance between the NovaSeq and AVITI sequencer for 40 *Cannabis* samples.

		NovaSeq			AVITI		
Step	Statistics	Overall	Average	SD*	Overall	Average	SD
Sequencing and mapping	Average Q-score	–	35.8	0.1	–	43.4	0.6
	Mean mapping quality	–	35.3	1.2	–	35.8	1.2
	Proportion of read with insertion (%)	–	6.2	1.1	–	7.0	1.3
	Proportion of read with deletion (%)	–	4.3	0.47	–	4.3	0.45
	Error rate (%)	–	0.013	0.0007	–	0.012	0.0007
	Homopolymer indels (%)	–	45.0	3.3	–	42.6	3.8
	Mean coverage (X)	–	0.17	0.09	–	0.27	0.15

Sequencing options

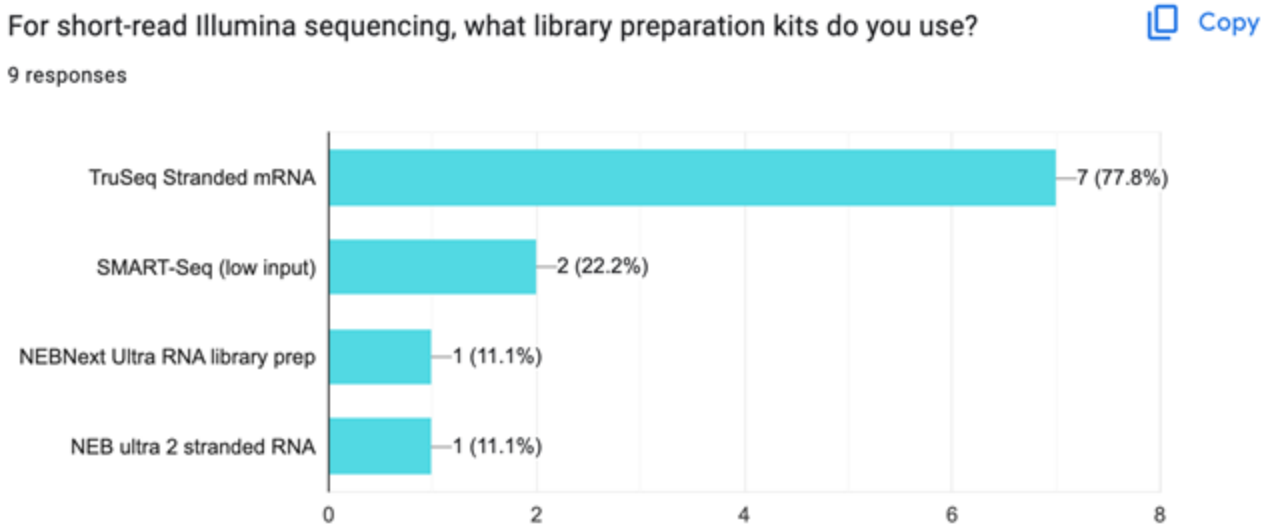
- **Platforms**

Sperfeld et al. 2022 *Applied & Environmental Microbiology*

	LR transcr.	Hybrid transcr.	SR transcr.
1) Basic statistics			
Counts	14,665	42,686	41,913
Bases (Mb)	24,291,537	60,039,032	54,236,986
N50 length (bp)	1,883	1,859	1,689
Mean length (bp)	1,656	1,406	1,294
GC content	67.4%	68.9%	69.0%
2) Protein-coding transcripts (>100 aa; ANGEL predicted)			
Total	13,972	37,086	35,927
Relative	95.3%	86.9%	85.7%
3) BUSCO completeness (n = 255 eukaryotic reference genes)			
Detected: complete + fragmented	44.7% (114)	83.2% (212)	83.2% (212)
Complete: single-copy	29.4% (75)	55.3% (141)	60.8% (155)
Complete: duplicated-copy	7.1% (18)	16.9% (43)	11.4% (29)
Fragmented	8.2% (21)	11.0% (28)	11.0% (28)

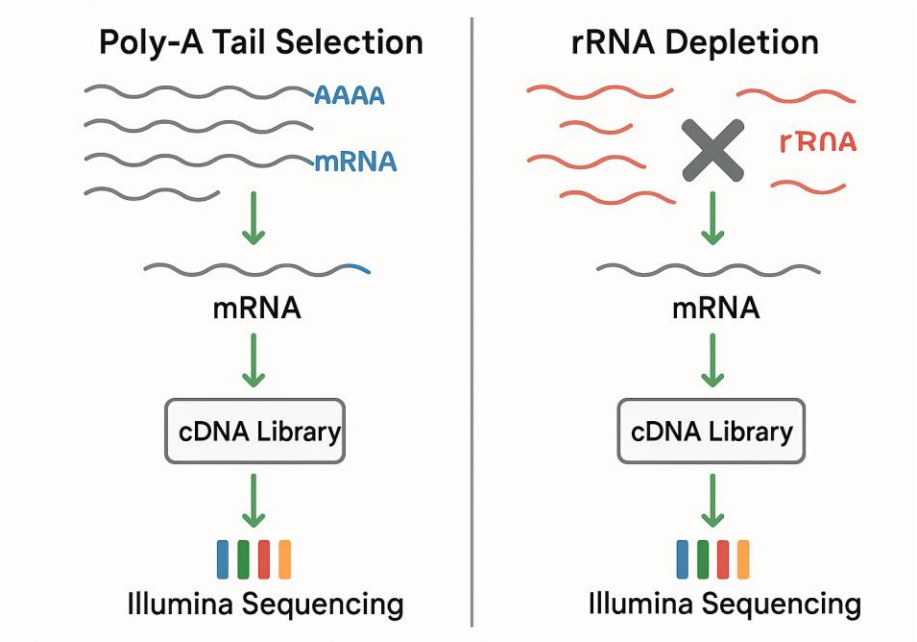
Sequencing options

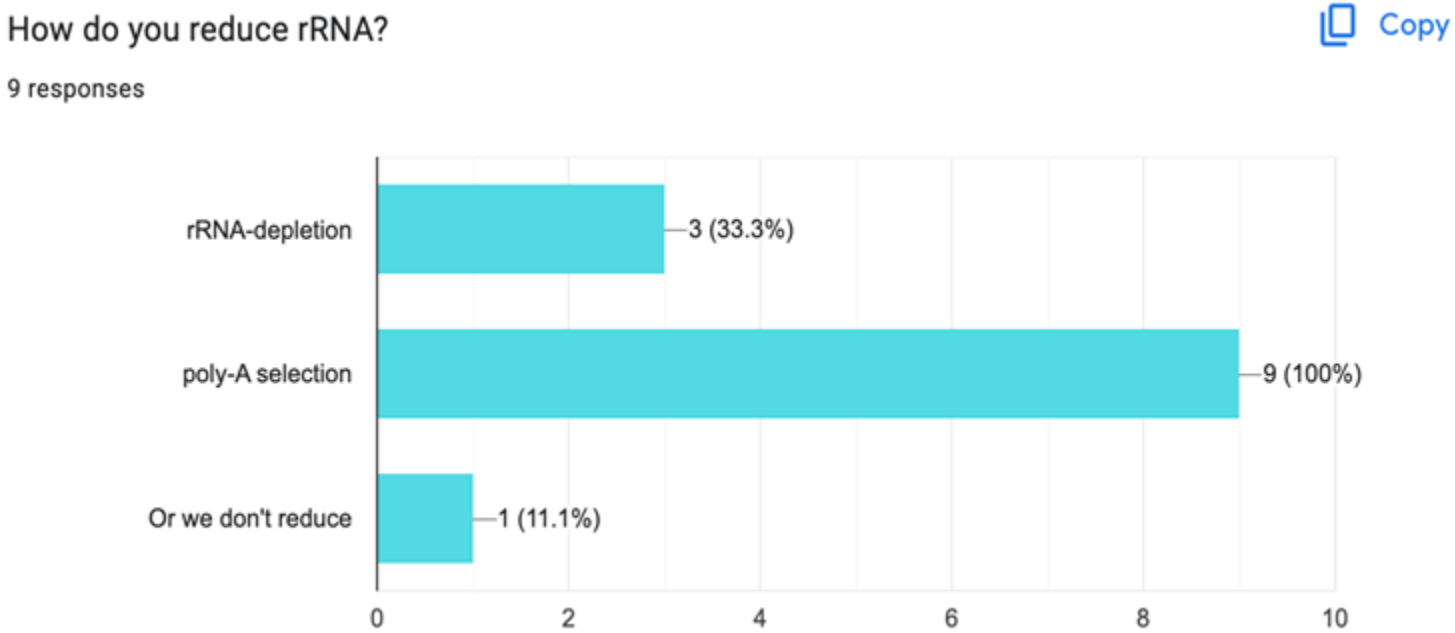
- Platforms
  - Short read sequencing (Illumina, AVITI)
  - Long read sequencing (PacBio, Oxford Nanopore)
- **Low input options: SMARTer-seq v4 (expensive). Accepts pg levels of total RNA vs. ~100 ng needed for standard kits**
- mRNA enrichment



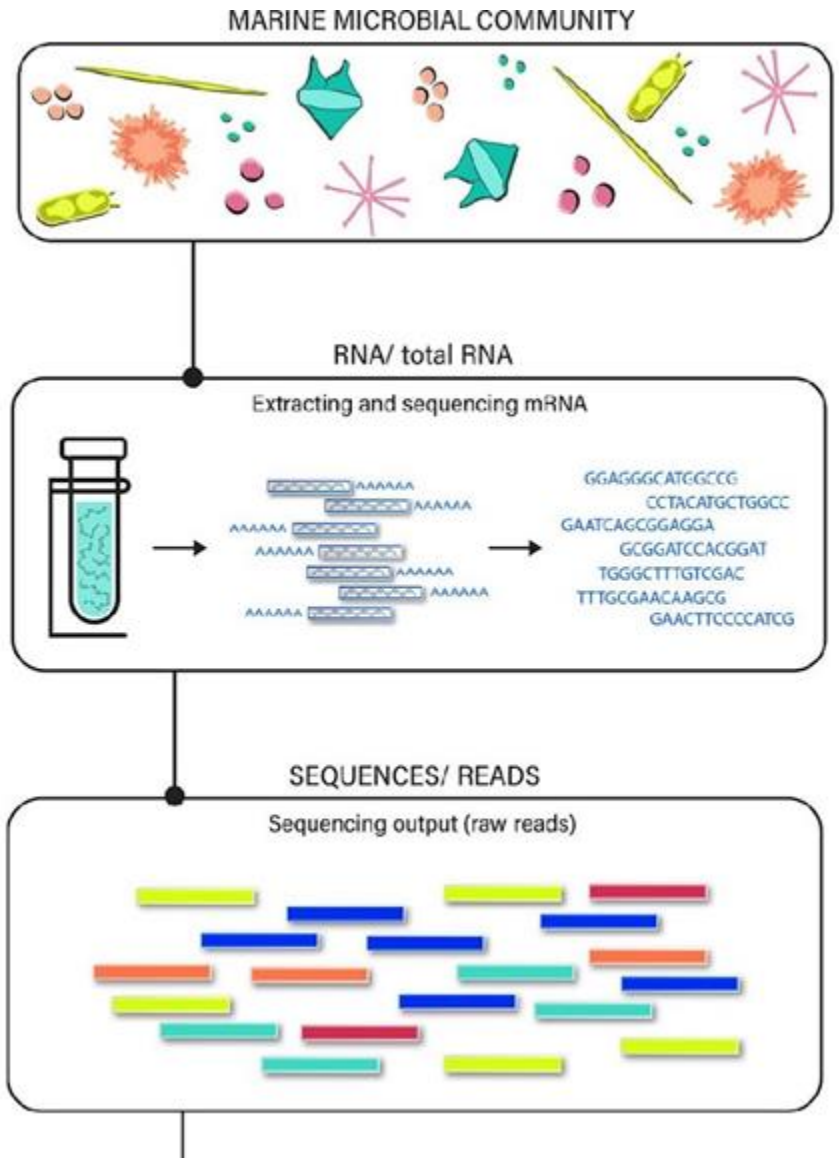
Sequencing options

- Platforms
  - Short read sequencing (Illumina, AVITI)
  - Long read sequencing (PacBio, Oxford Nanopore)
- Low input options: SMARTer-seq v4 (expensive)
- **mRNA enrichment: Poly-A tail selection or rRNA depletion**





Physical sample processing



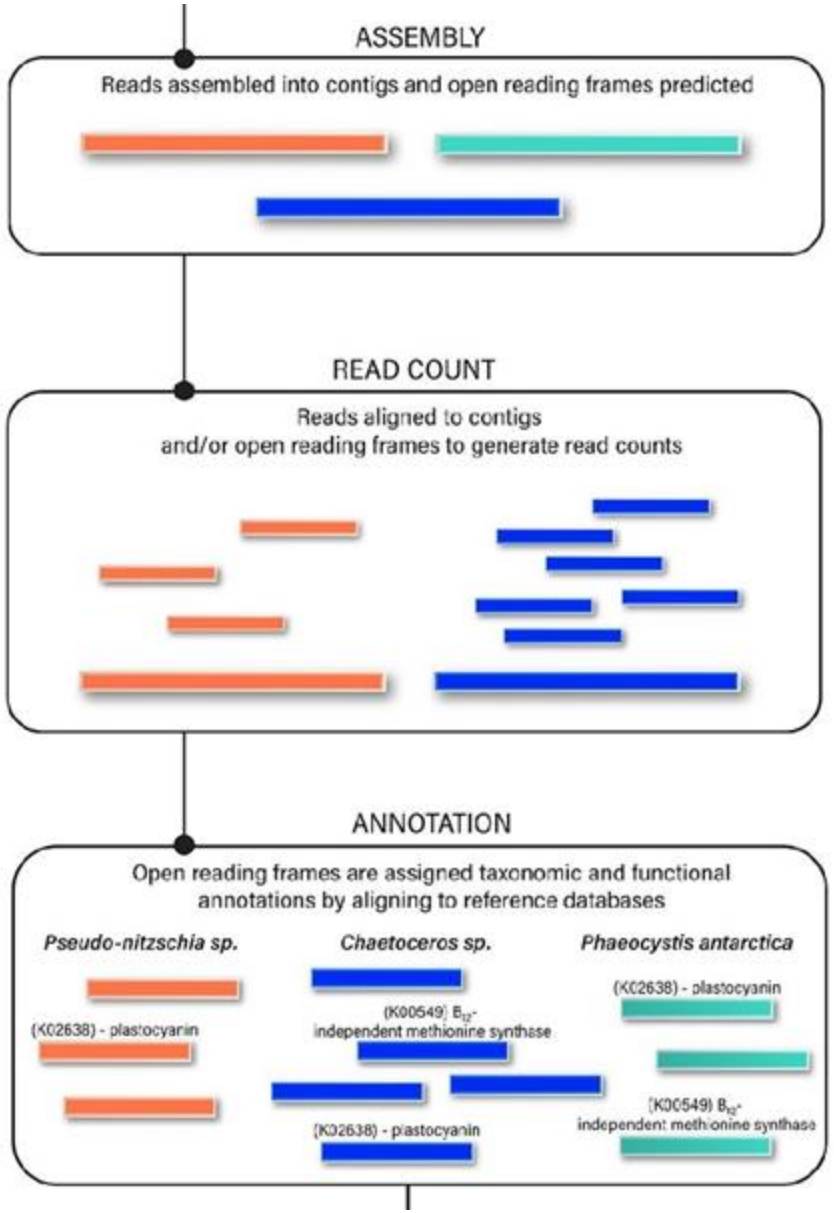
Step 1. Sample collection:
Filtration mechanism, duration, filter type, preservation

Step 2. RNA extraction:
Sample handling, extraction kit, physical disruptions, DNA removal

Step 3. Sequence prep:
Library preparation, rRNA removal

Step 4. Sequencing:
Sequence platform, length and throughput

Bioinformatics



Step 5. Assembly:
contaminant removal, trimming, assembly vs. direct mapping, assembler choice

Step 6: Annotation:
Ortholog clustering, taxonomic database, functional database, alignment method

Step 7+: Differential expression:
normalization method, biostatistical tests

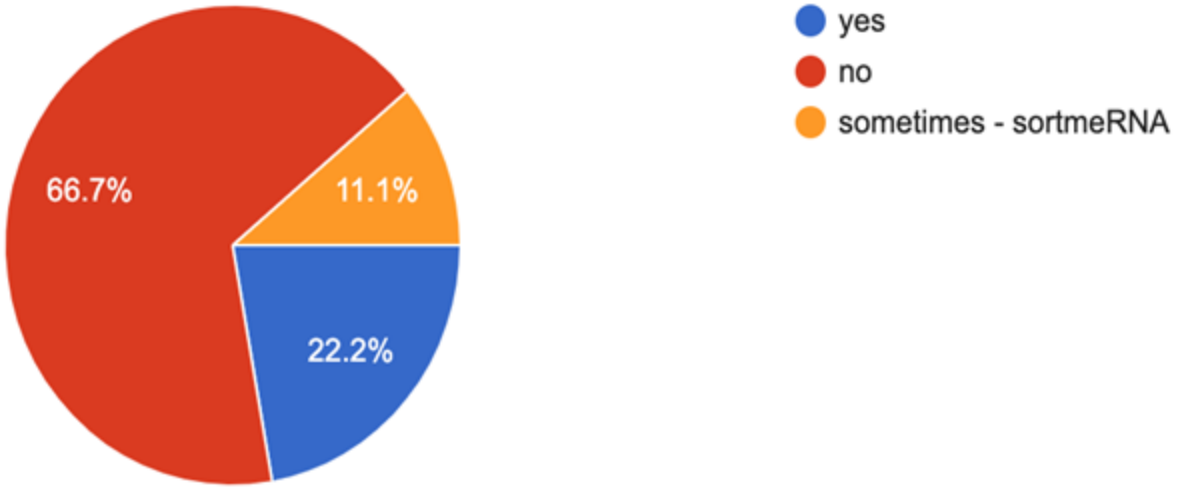
Assembly

- **Contaminant removal**
- Trimming
- Assembly or direct mapping

Do you remove rRNA from sequences (e.g., using RiboDetector)?

Copy

9 responses



Assembly

- **Contaminant removal**
- Trimming
- Assembly or direct mapping

```
sourmash gather -k 31 \
    ecoli-reads.sig ecolidb

loaded query: ecoli_ref-5m.khmer.fq.gz... (k=31, DNA)
loaded 1 databases.

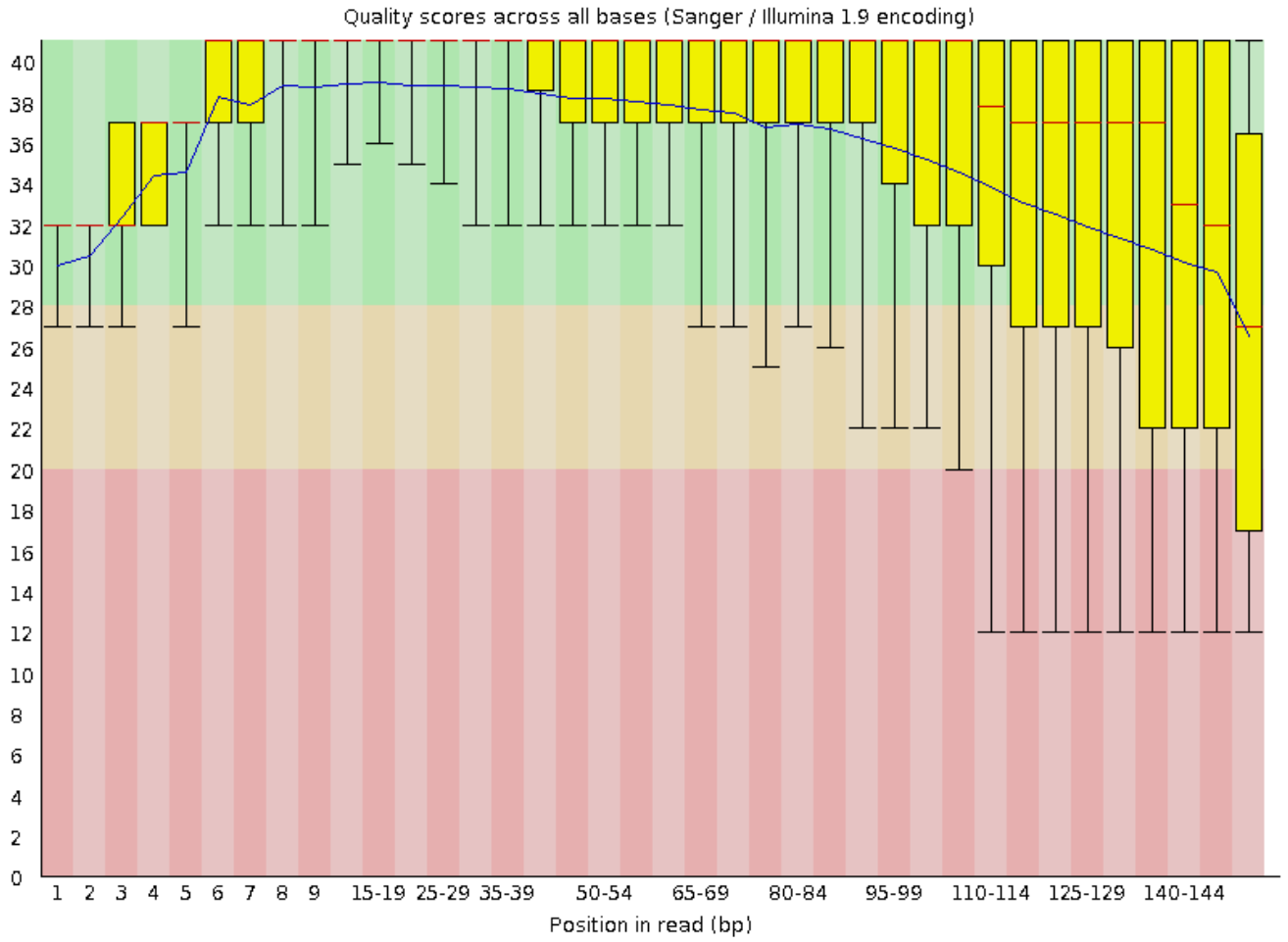
overlap     p_query p_match
---------   ------- -------
4.1 Mbp       65.4%   83.5%    NZ_JMGW01000001.1 Escherichia coli 1-...
2.4 Mbp        2.7%    3.3%    NZ_GG749254.1 Escherichia coli FVEC14...
3.4 Mbp        1.4%    1.7%    NZ_MOGK01000001.1 Escherichia coli st...
3.1 Mbp        0.6%    0.7%    NZ_LVOV01000001.1 Escherichia coli st...
3.1 Mbp        0.3%    0.4%    NZ_MIWP01000001.1 Escherichia coli st...
3.0 Mbp        0.3%    0.4%    NZ_APWY01000001.1 Escherichia coli 17...
3.5 Mbp        0.2%    0.2%    NZ_JNLZ01000001.1 Escherichia coli 3-...
4.0 Mbp        0.2%    0.2%    NZ_GG774190.1 Escherichia coli MS 196...
2.3 Mbp        0.1%    0.2%    NZ_KB732756.1 Escherichia coli KTE66 ...
2.0 Mbp        0.1%    0.1%    NZ_BBUW01000001.1 Escherichia coli O1...
2.3 Mbp        0.1%    0.1%    NZ_MOZX01000101.1 Escherichia coli st...
2.3 Mbp        0.1%    0.1%    NZ_JSMW01000001.1 Escherichia coli st...
4.0 Mbp        0.1%    0.1%    NZ_JMGU01000001.1 Escherichia coli 2-...
2.0 Mbp        0.0%    0.0%    NZ_MOZC01000010.1 Escherichia coli st...
2.1 Mbp        0.0%    0.0%    NZ_MKJG01000001.1 Escherichia coli st...
1.8 Mbp        0.0%    0.0%    NZ_AEKA01000453.1 Escherichia sp. TW1...
2.6 Mbp        0.0%    0.0%    NZ_LEAD01000071.1 Escherichia coli st...
3.5 Mbp        0.0%    0.0%    NZ_MIWF01000001.1 Escherichia coli st...

found 18 matches total;
the recovered matches hit 71.5% of the query
```

Pierce et al. 2019 *F1000 Research*

Assembly

- Contaminant removal
- **Trimming**
- Assembly or direct mapping

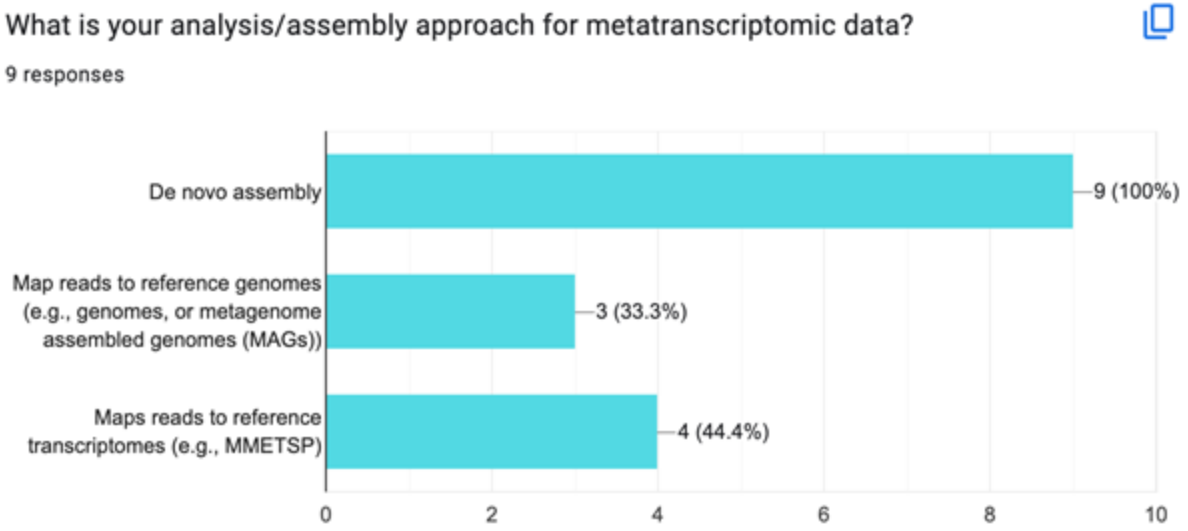


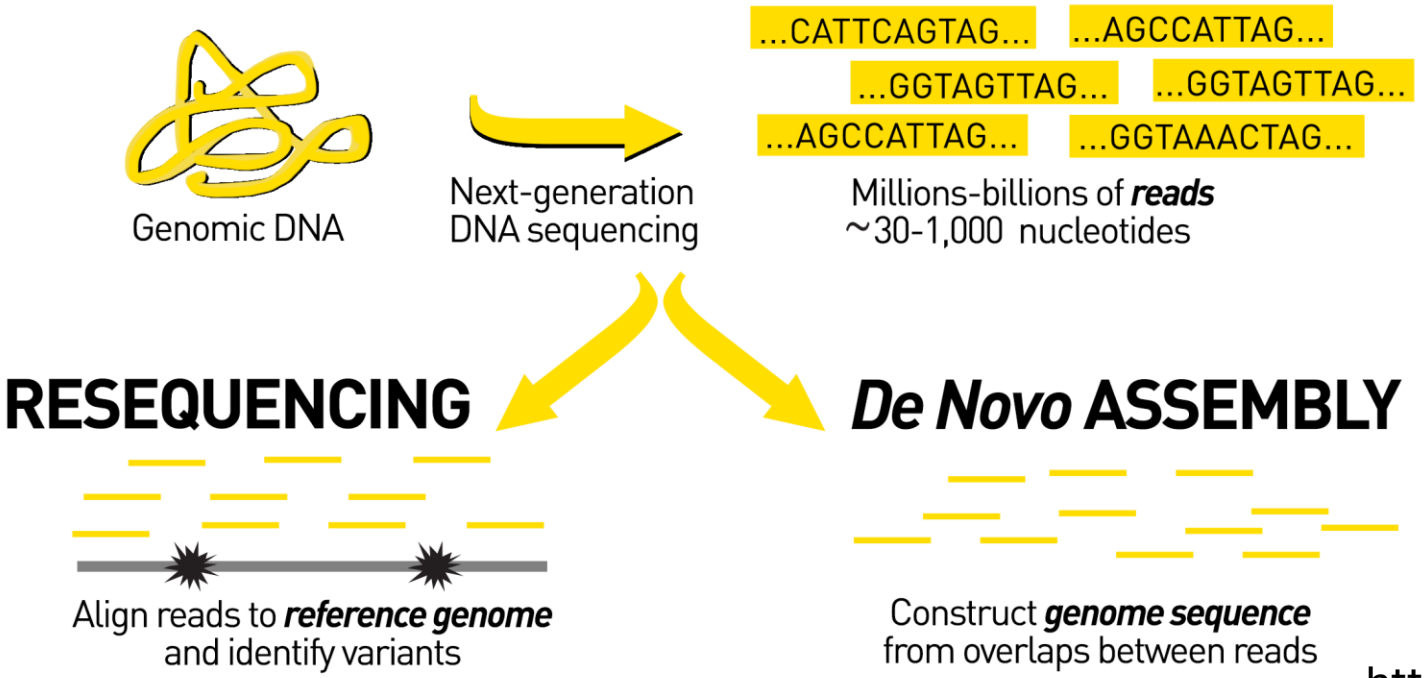
Aggressive cutting could alter gene expression results

(Williams et al. 2016 *BMC Bioinformatics)*

Assembly

- Contaminant removal
- Trimming
- **Assembly or direct mapping**
  - Assembler choice

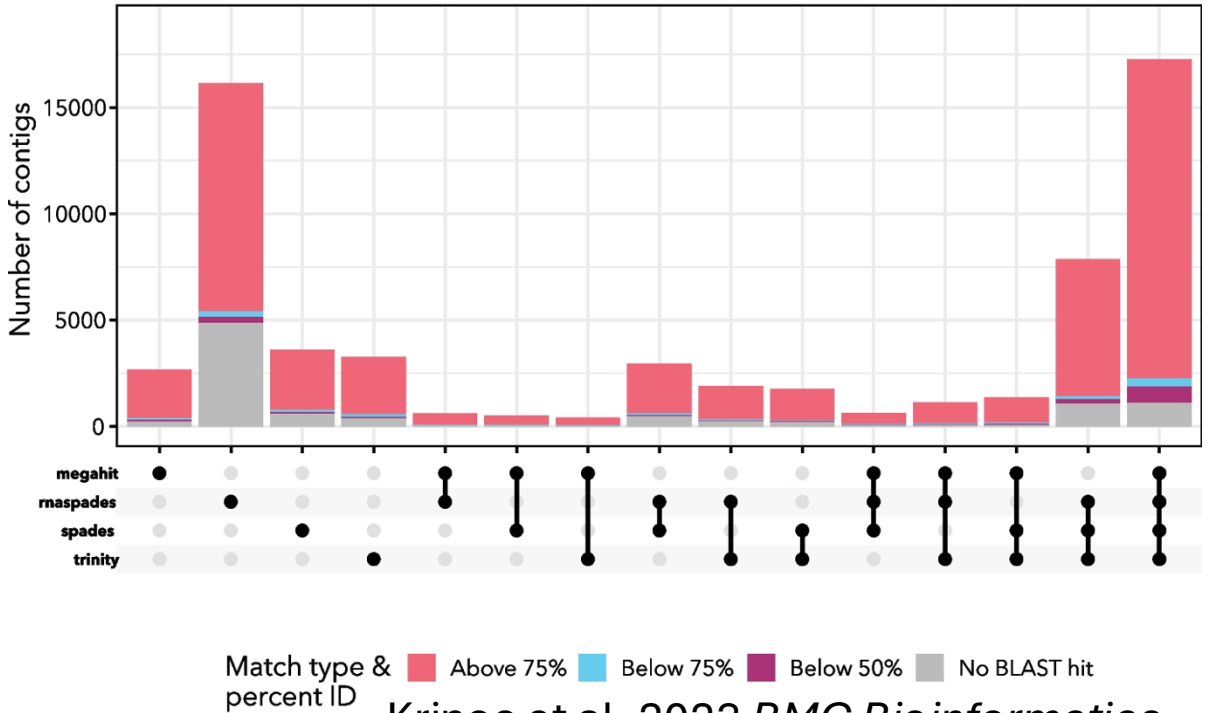




https://www.hudsonalpha.org

Assembly

- Contaminant removal
- Trimming
- Assembly or direct mapping
  - **Assembler choice**

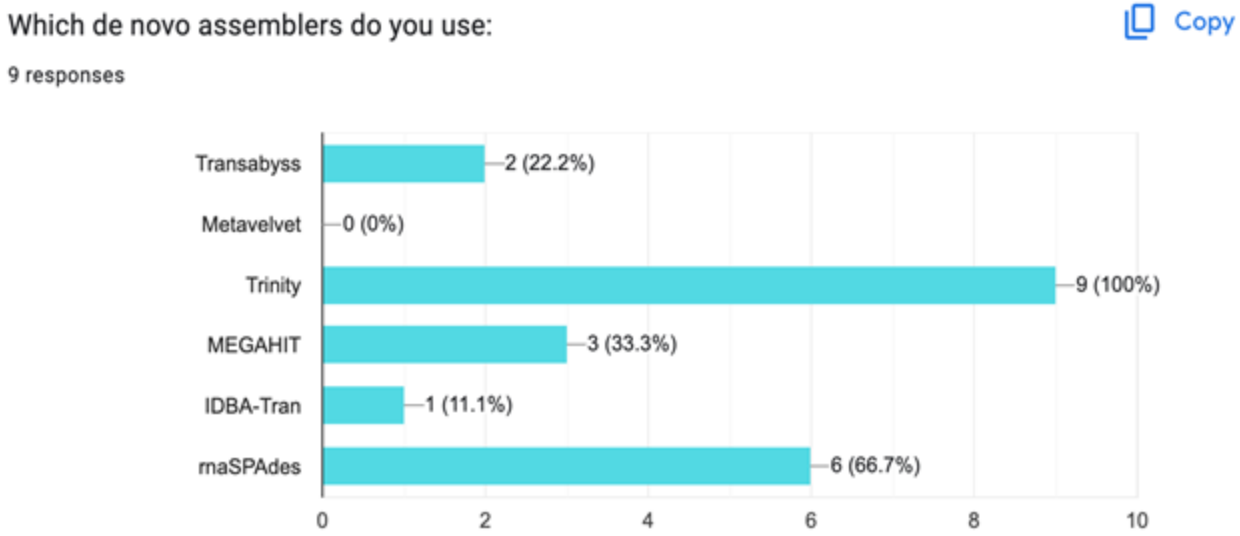


Krinos et al. 2023 *BMC Bioinformatics*

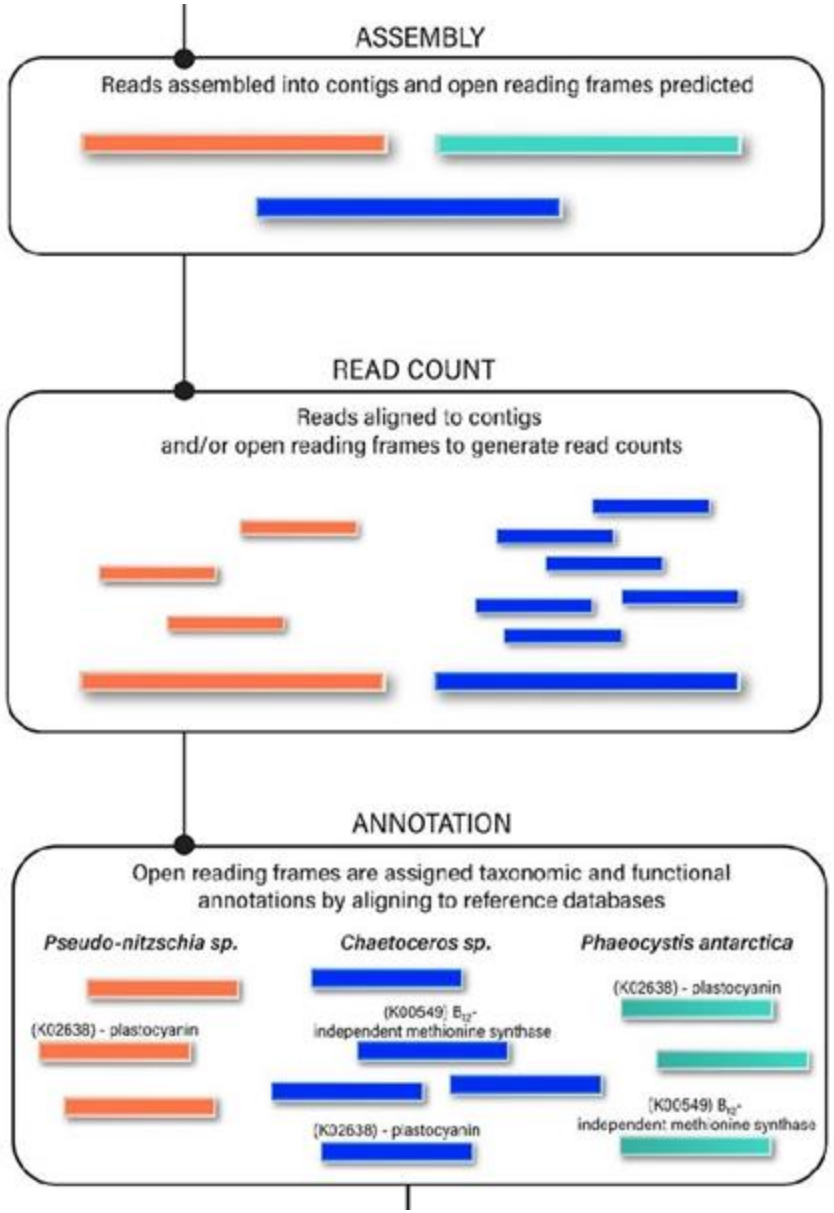
rnaSpades produces the most novel contigs, megahit the least

Megahit performs the fastest, Trinity is slow

Multi-assembly approach improves predictions



Bioinformatics



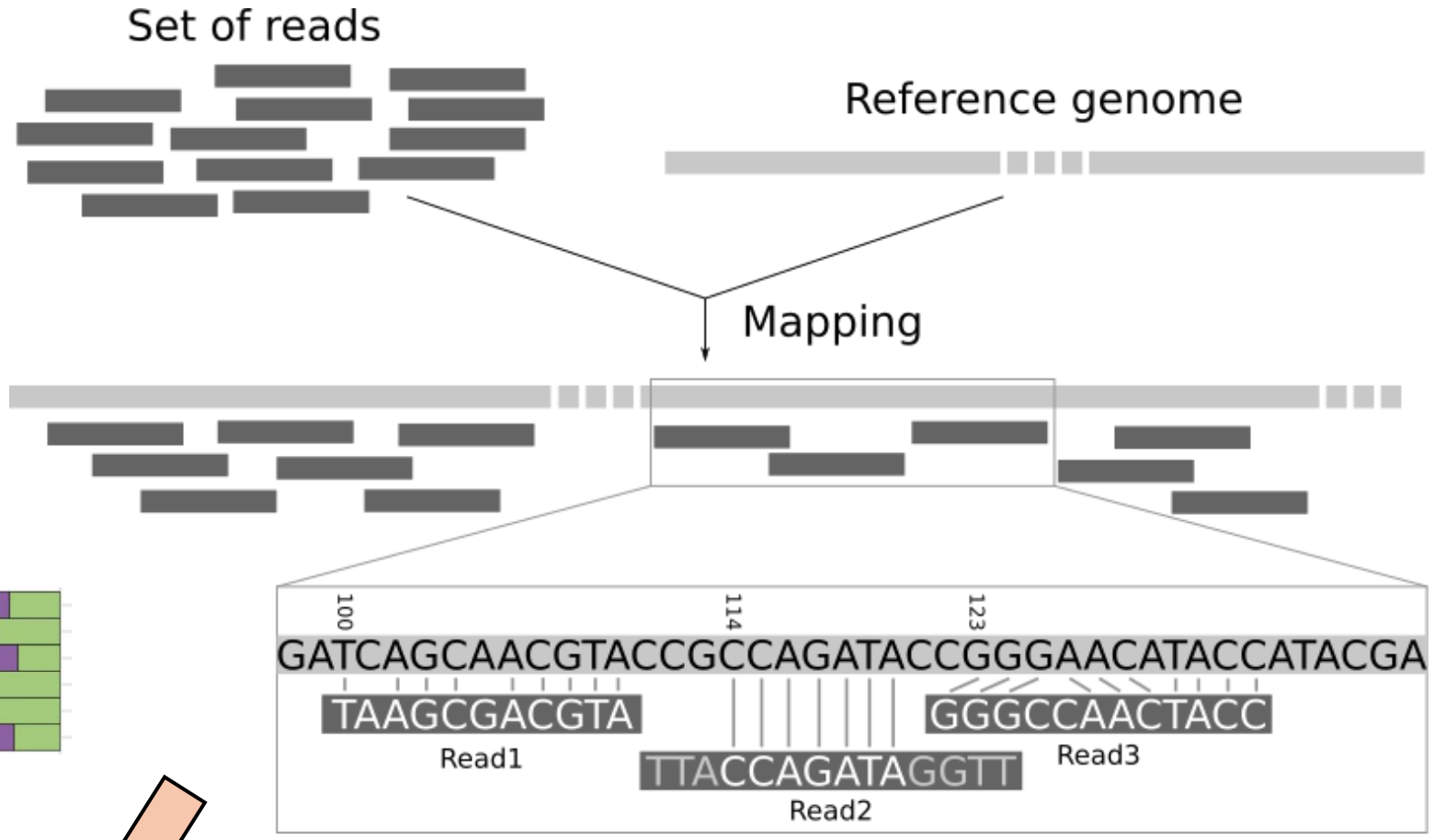
Step 5. Assembly:
contaminant removal, trimming, assembly vs. direct mapping, assembler choice

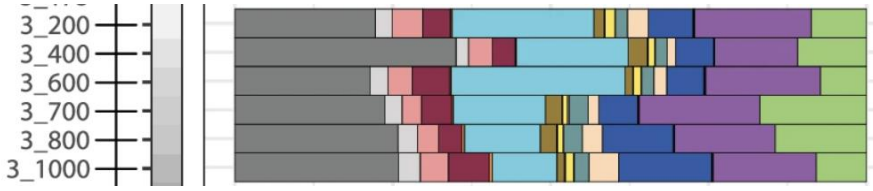
Step 6: Annotation:
Ortholog clustering, taxonomic database, functional database, alignment method

Step 7+: Differential expression:
normalization method, biostatistical tests

Annotation

- **Mapping reads**
- Taxonomy
- Function

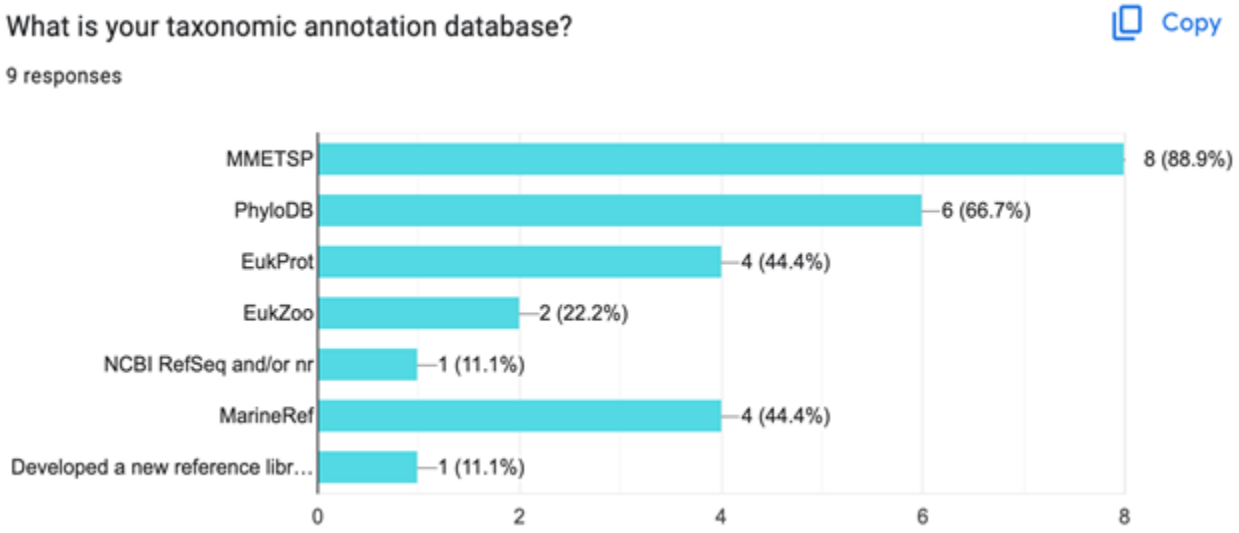


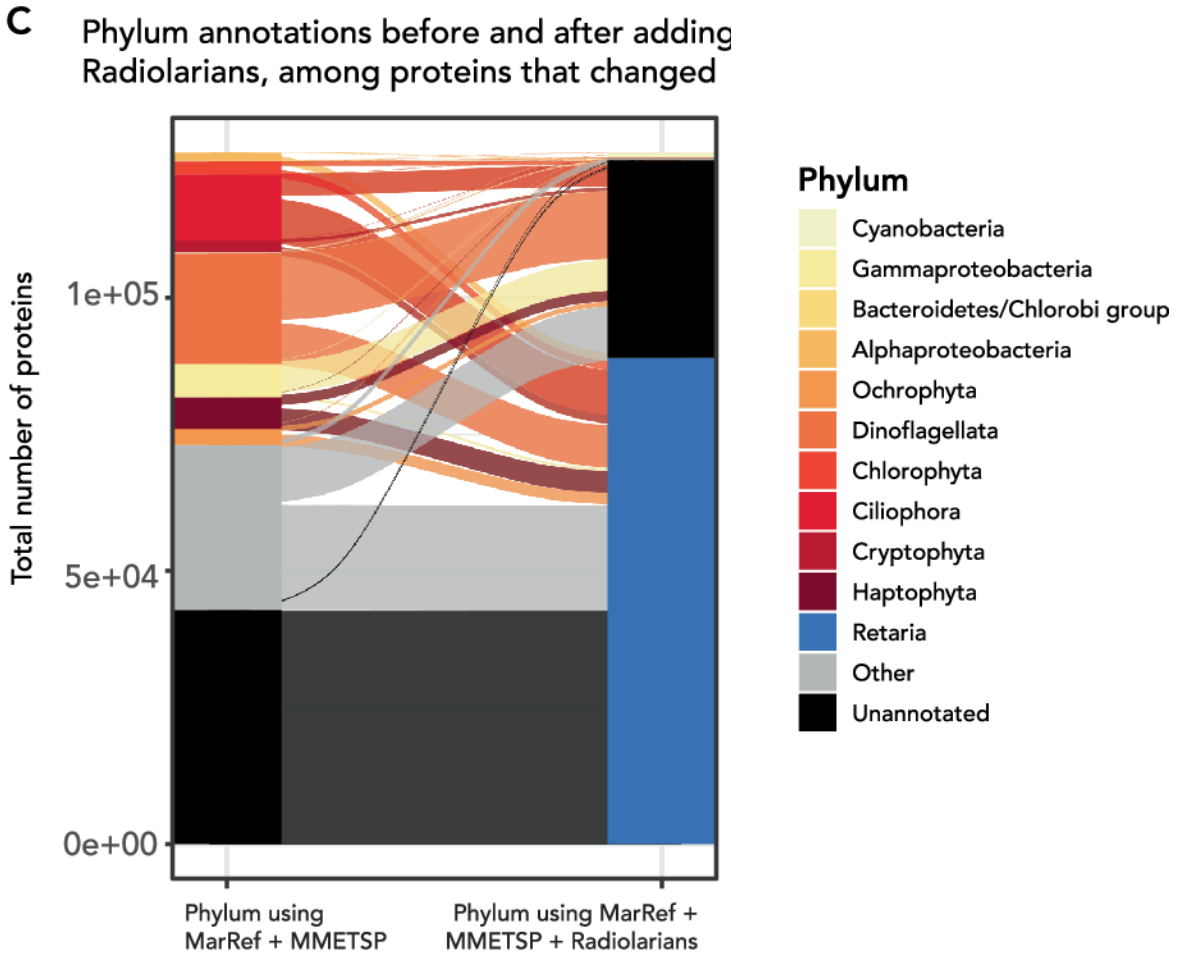


https://training.galaxyproject.org

Annotation

- Mapping reads
- **Taxonomy**
  - Database choice
  - Alignment method
- Function

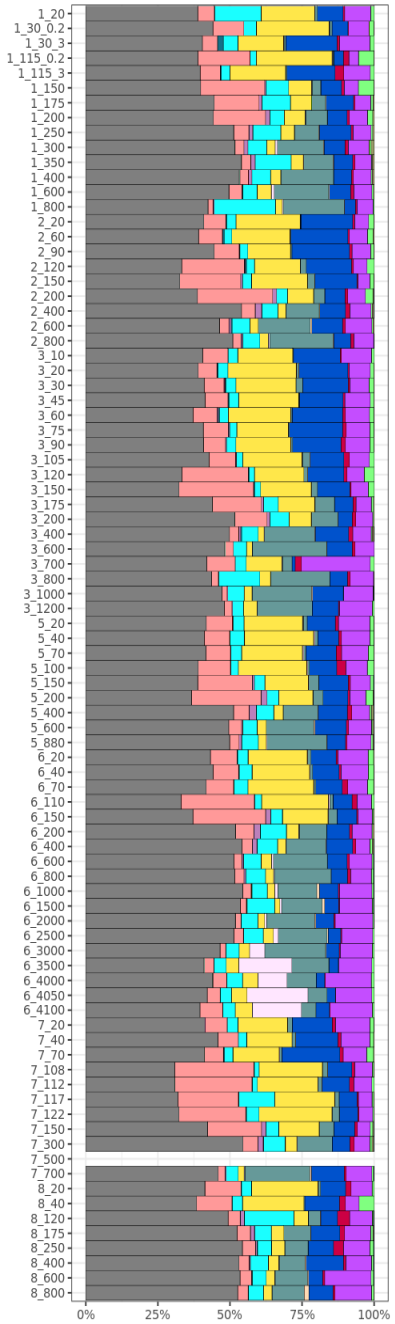




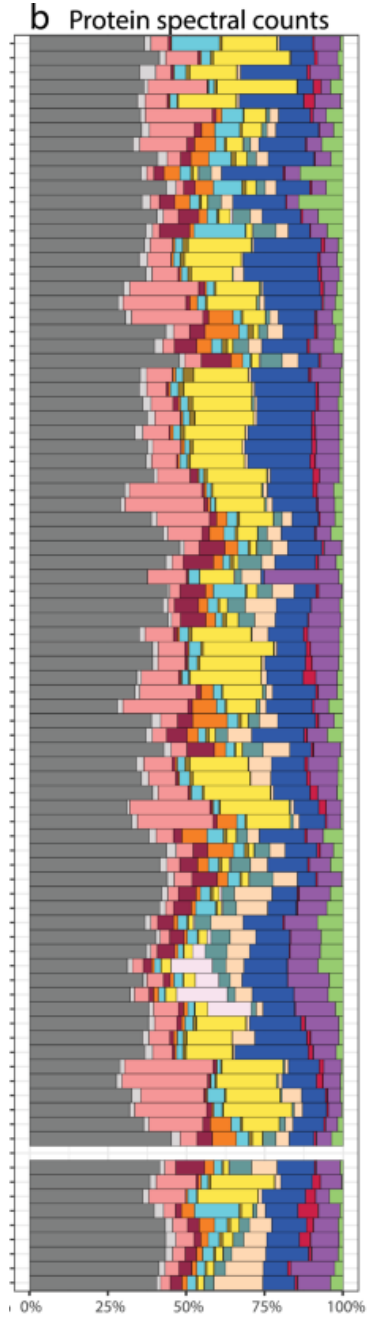
Krinos et al. 2024 *Nature Communications*

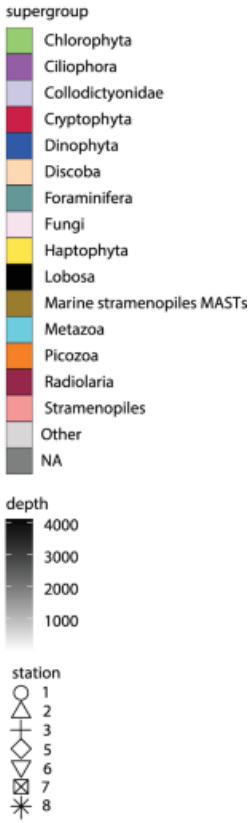
Also consider how database composition affects lineage calls

Before Eukprot



After Eukprot

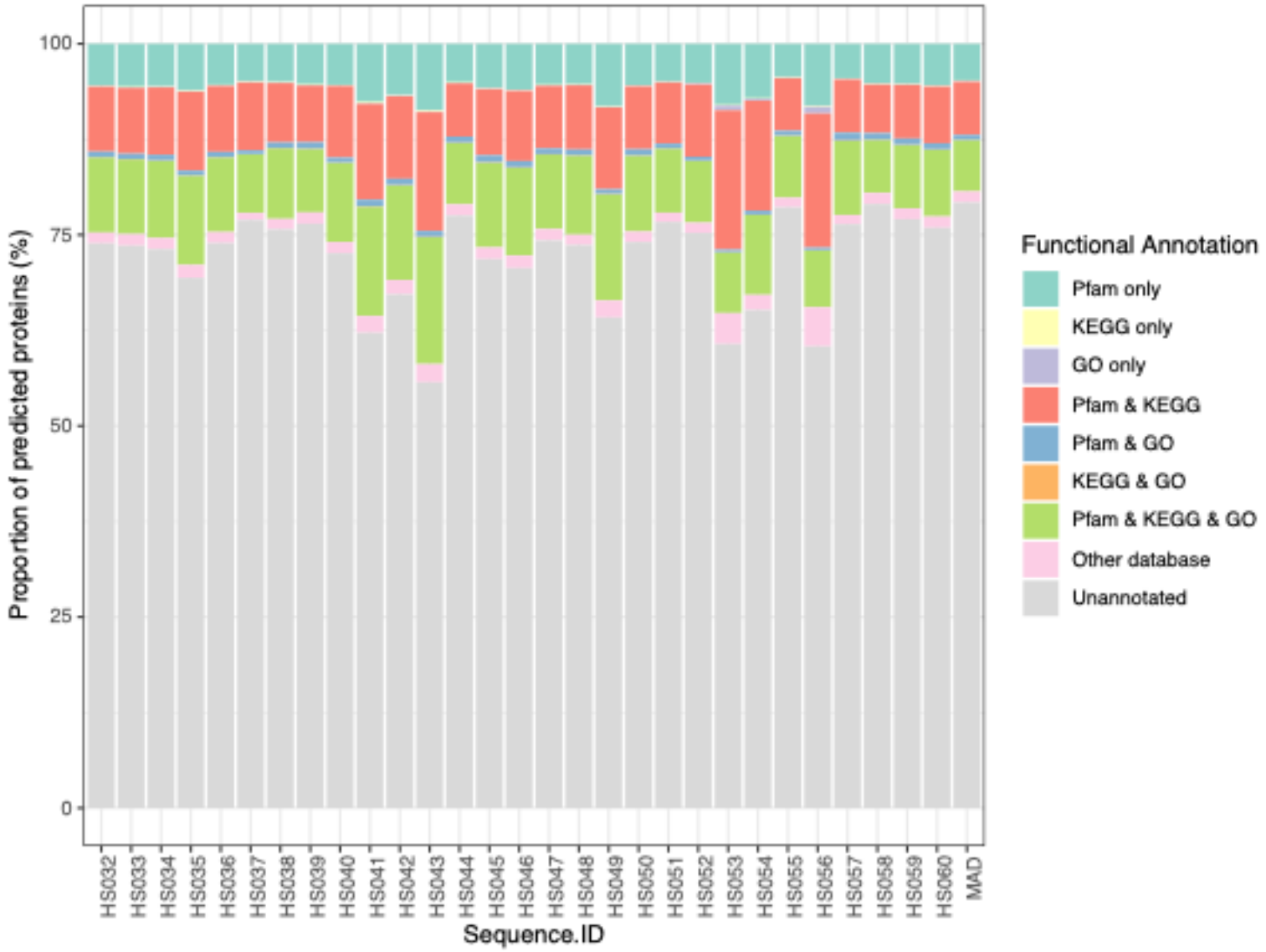




Cohen et al. 2024 *Nature Communications*

Annotation

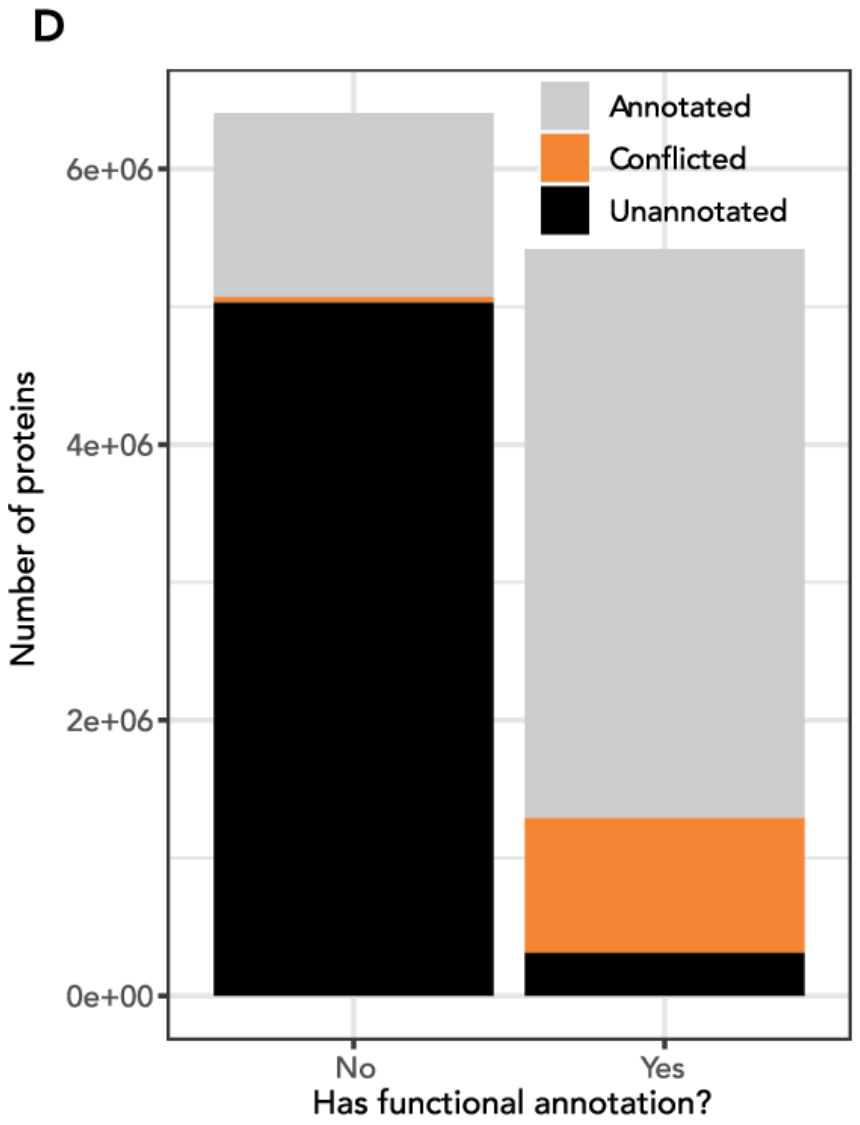
- Mapping reads
- Taxonomy
- **Function**
  - Most content is unannotated
  - KEGG
    - Organized in higher module levels
    - Relies on sequence homology
  - Pfam
    - Identifies conserved protein domains



Katie Roche, OCB working group

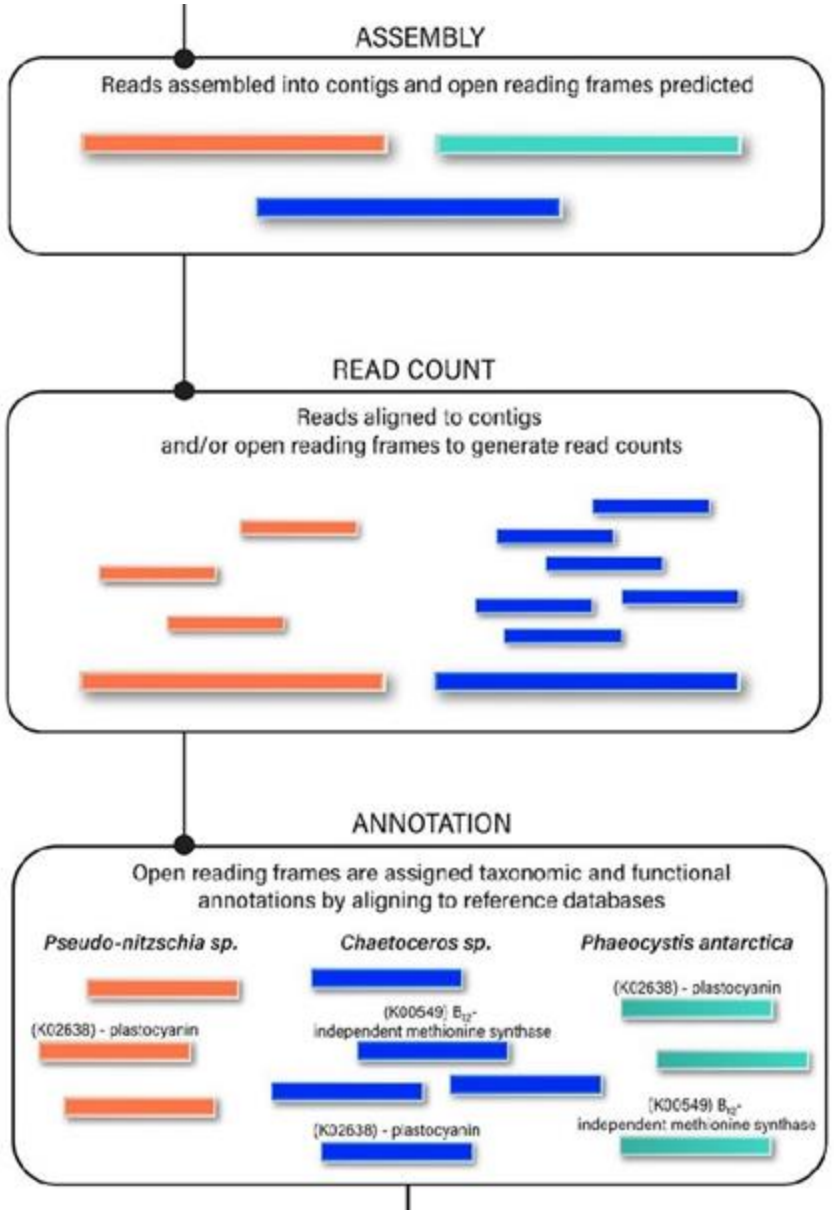
Annotation

- Mapping reads
- Taxonomy
- **Function**
  - Most content is unannotated
  - KEGG
    - Organized in higher module levels
    - Relies on sequence homology
  - Pfam
    - Identifies conserved protein domains



Krinos et al. 2024 *Nature Communications*

Bioinformatics



Step 5. Assembly:
contaminant removal, trimming, assembly vs. direct mapping, assembler choice

Step 6: Annotation:
Ortholog clustering, taxonomic database, functional database, alignment method

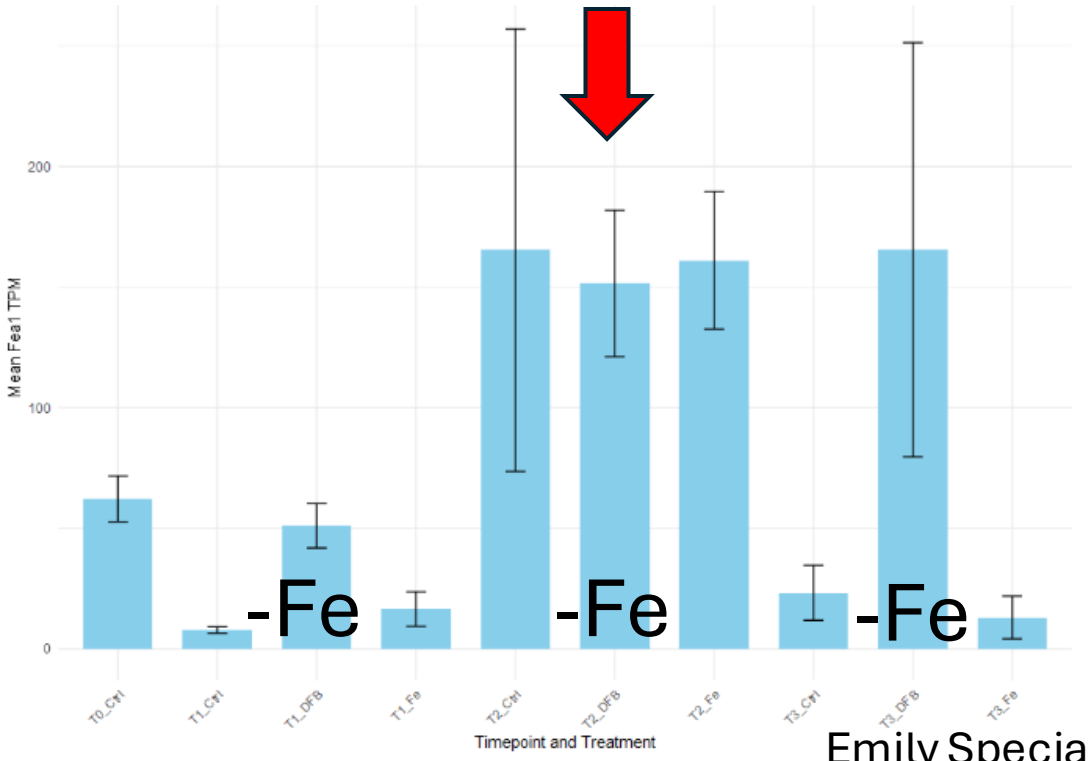
Step 7+: Differential expression
normalization method, biostatistical tests

Differential expression

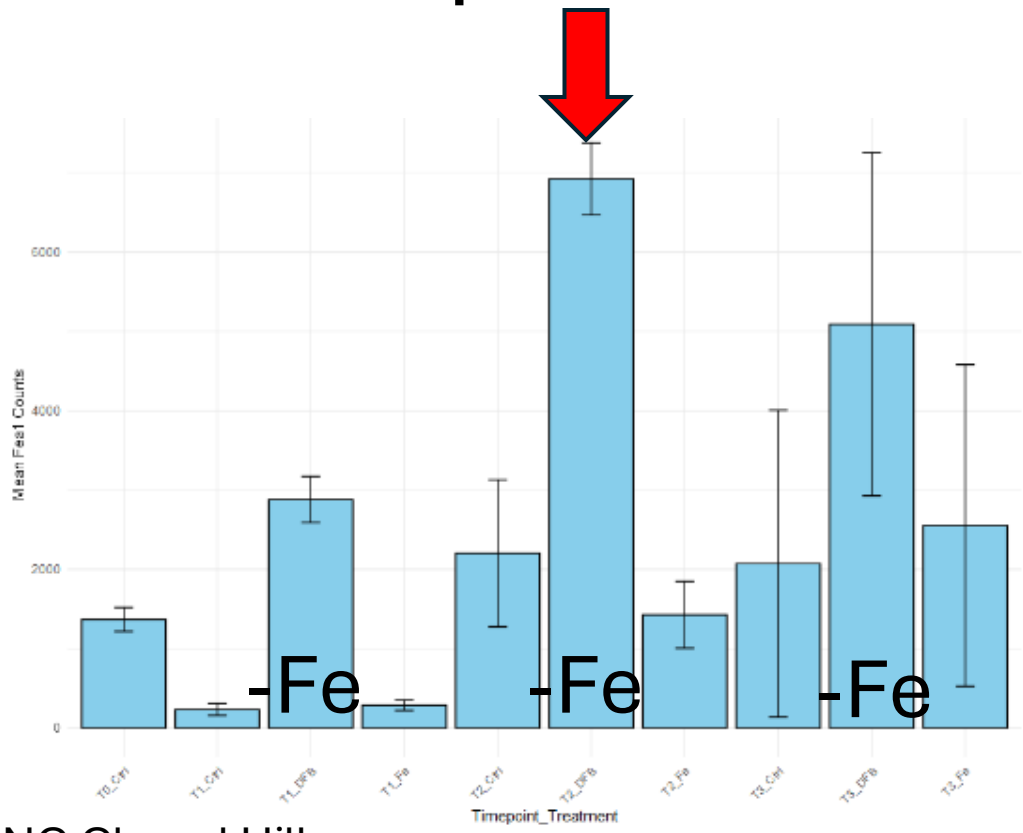
- **Normalization**
- Biostatistical tests
- Visualization

TMM (edgeR) or mean of ratios (DEseq2): adjusts for RNA composition with normalization factors
TPM/RPKM: Relative to total reads
Copies/L: absolute quantification

TPM



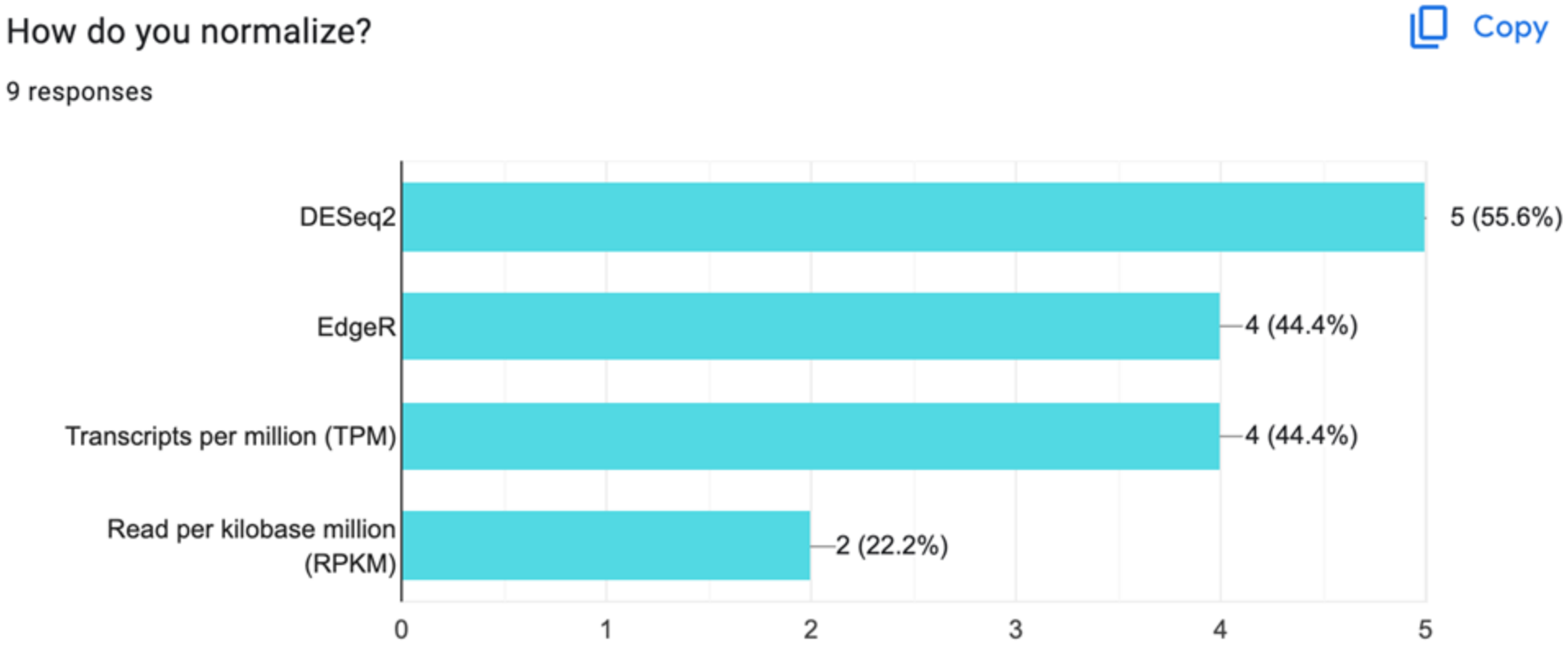
DESeq2 normalization



Differential expression

- **Normalization**
- Biostatistical tests
- Visualization

TMM (edgeR) or mean of ratios (DEseq2): adjusts for RNA composition with normalization factors
TPM/RPKM: Relative to total reads
Copies/L: absolute quantification



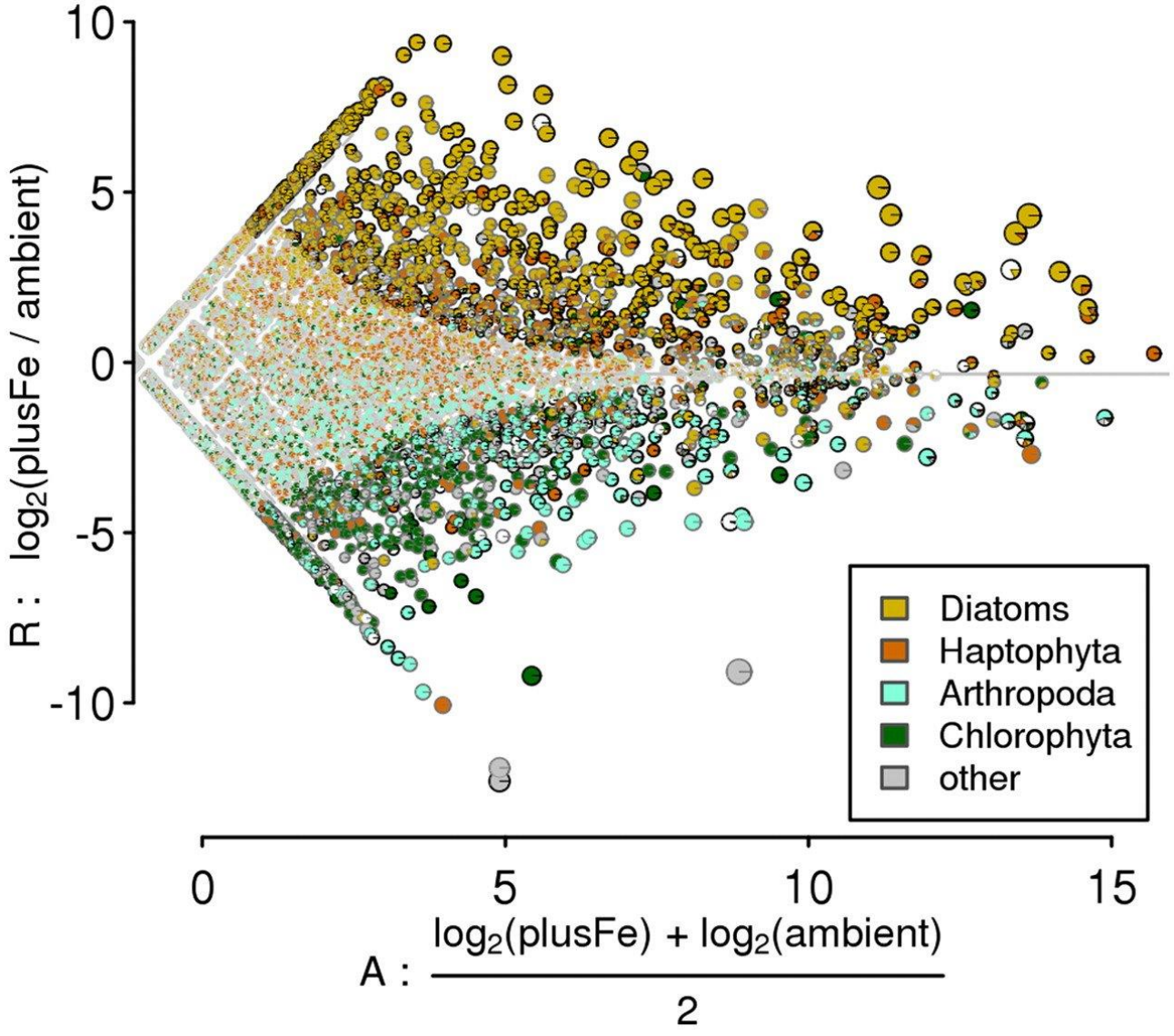
Differential expression

- Normalization
- **Biostatistical tests**
- Visualization

EdgeR & DEseq2 estimate differential expression

Assumes that most genes are **not** differentially expressed

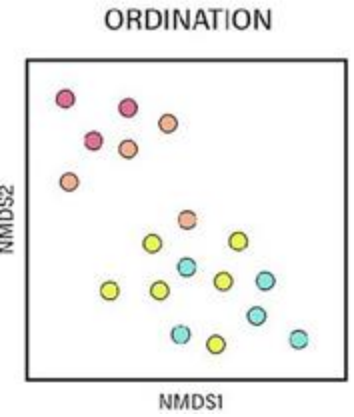
Applies Exact test (edgeR), Wald test or likelihood ratio test (Deseq2)

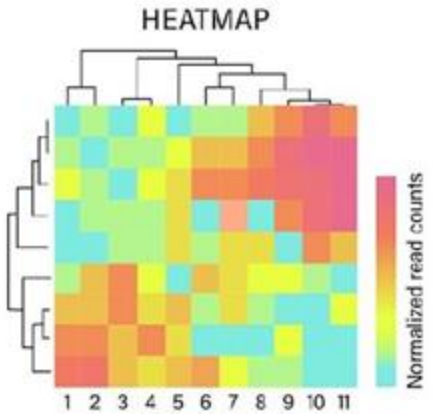


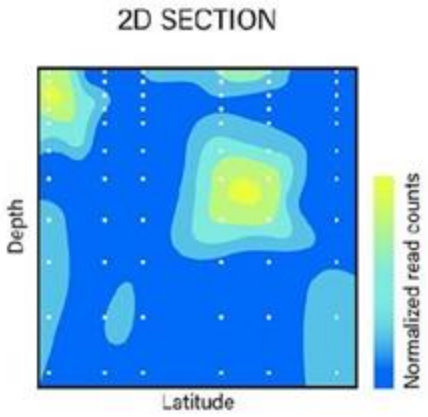
Marchetti et al. 2012 *PNAS*

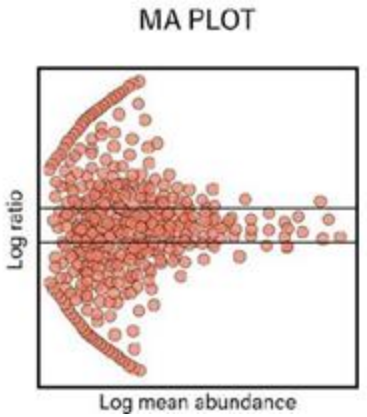
Differential expression

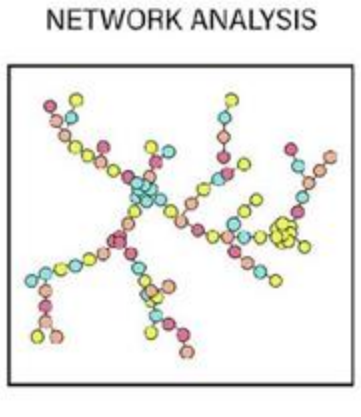
- Normalization
- Biostatistical tests
- **Visualization**

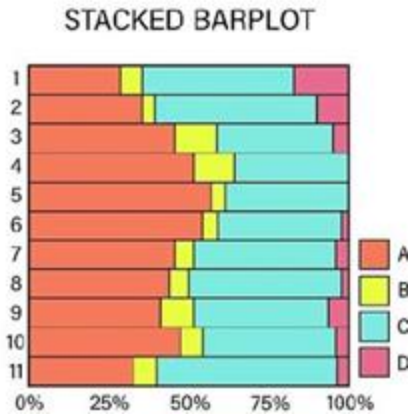










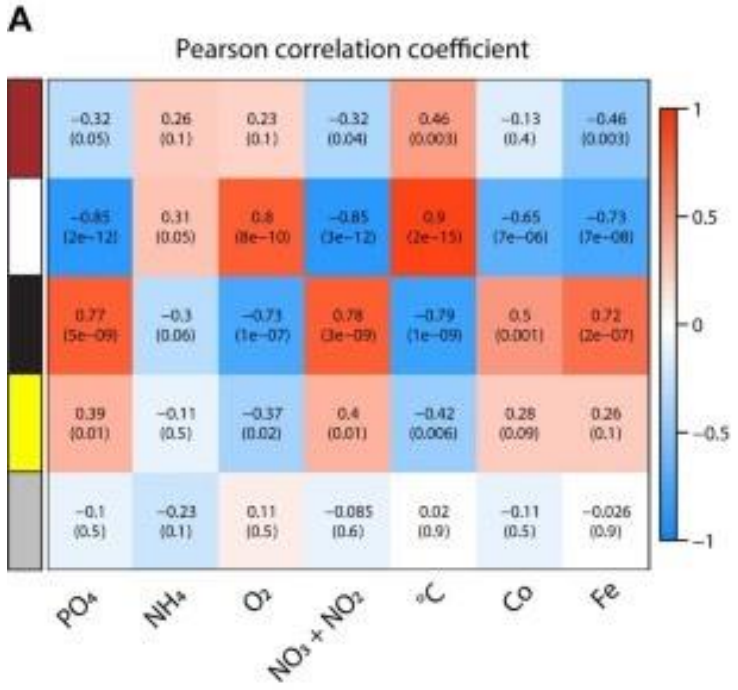


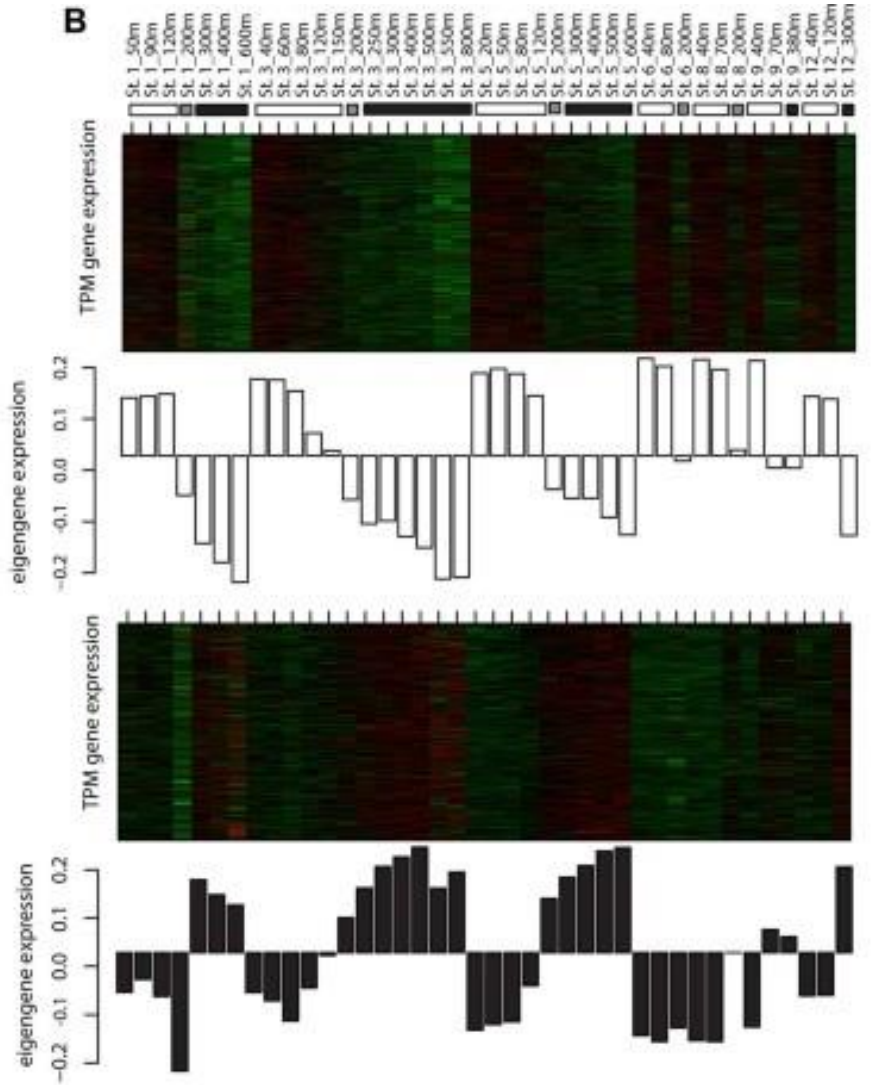
Differential expression

- Normalization
- Biostatistical tests
- **Visualization**

Identified differences in surface and mesopelagic dinoflagellate gene expression

Network analysis





Cohen et al. 2021 *Nature Microbiology*

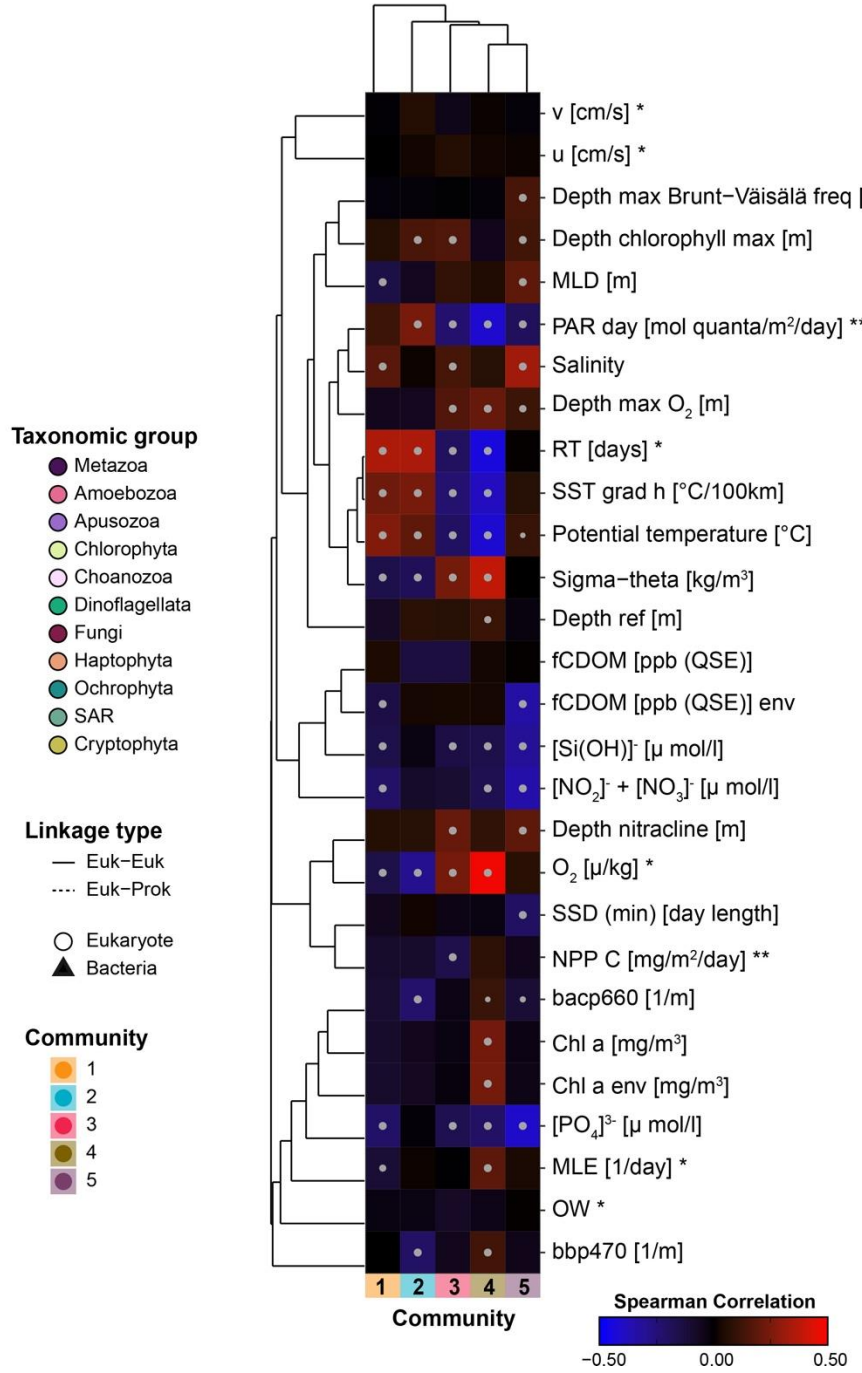
Differential expression

- Normalization
- Biostatistical tests
- **Visualization**

Network analysis

Identified communities that co-occur across the surface metagenomes (TARA)





Alexander et al. 2023 *mBio*

Differential expression

- Normalization
- Biostatistical tests
- **Visualization**

See PNAS site:
https://www.pnas.org/doi/full/10.1073/pnas.2200014119

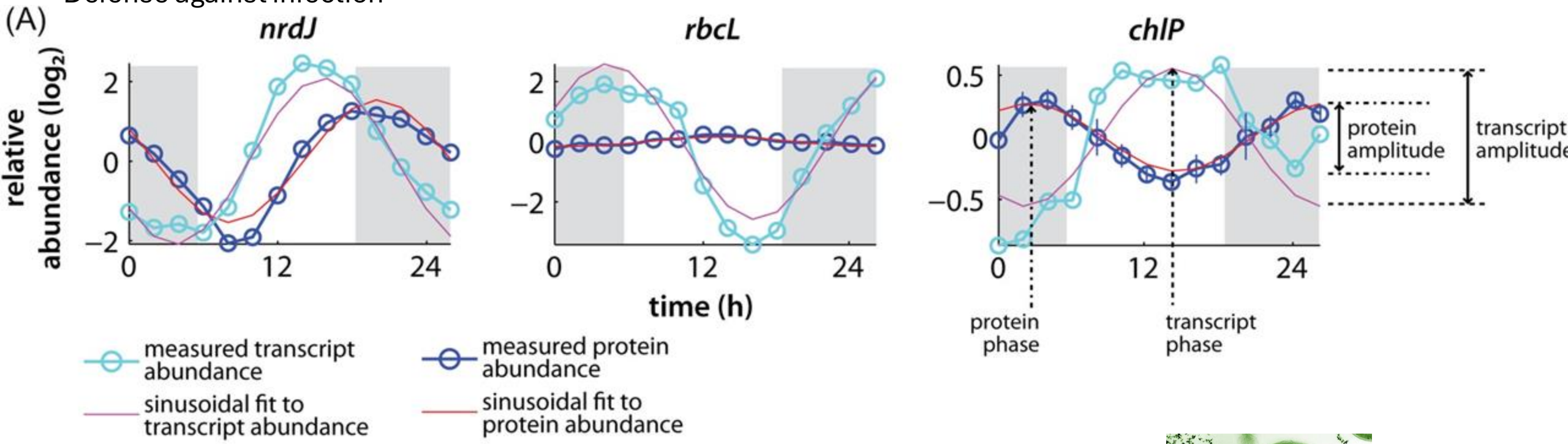
Saunders et al. 2022 *PNAS*

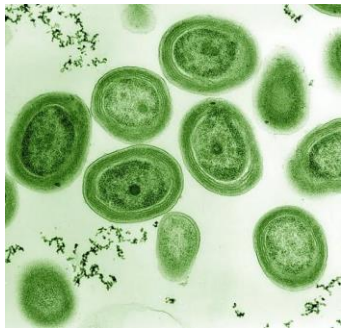
Multi-omic integration: transcripts vs. proteins

Ribonucleotide reductase – nucleic acid repair
Defense against infection

Rubisco

Chlorophyll





Boyd & Van Mooy 2025 *Limnology & Oceanography,* Waldbauer et al. (2012)

Looking forward

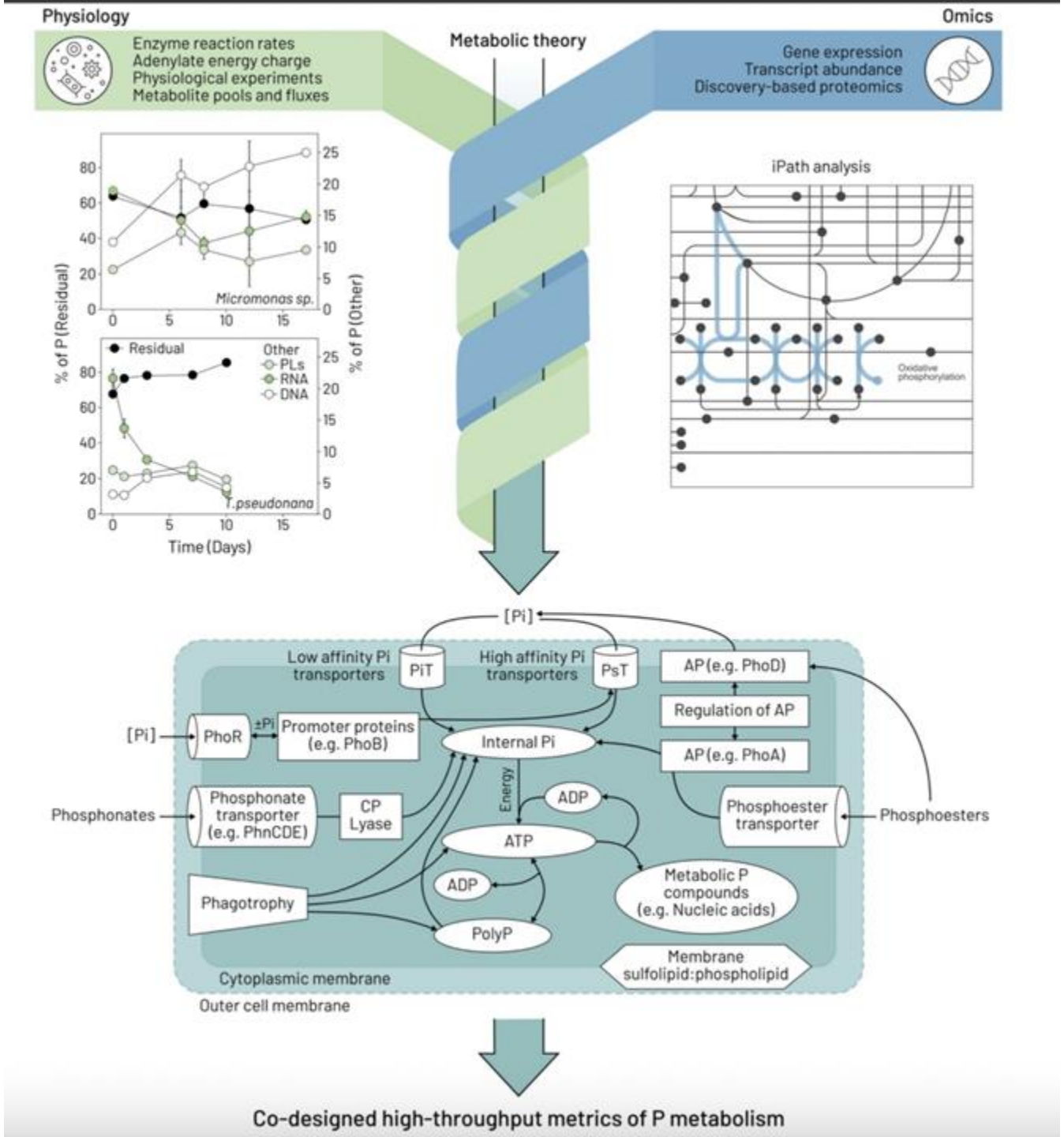
Physiology data is central to understanding biogeochemistry

Experiments should be co-designed with physiology metrics and omics integrated

*

Nucleic acid based omics is limited by:

- Poor reference databases
- Large number of unannotated genes
- Disconnects with measured biomass/rates
- Compositional data
- Not often reported in absolute quantities (“relative abundance”)



Strzepek et al. 2022 *Journal of Plankton Research*

What are the most important considerations?

- Identify priority question, and design sampling scheme around this
  - Group targeting? One cruise or many?
- Sequence length matters
- Choose taxonomic database wisely
- Explore multiple visualization formats