Leveraging large datasets to discover protistan diversity across scales

Massachusetts Institute of Technology

WOODS HOLE OCEANOGRAPHIC INSTITUTION
Our Ocean. Our Planet. Our Future.®

Arianna I. Krinos

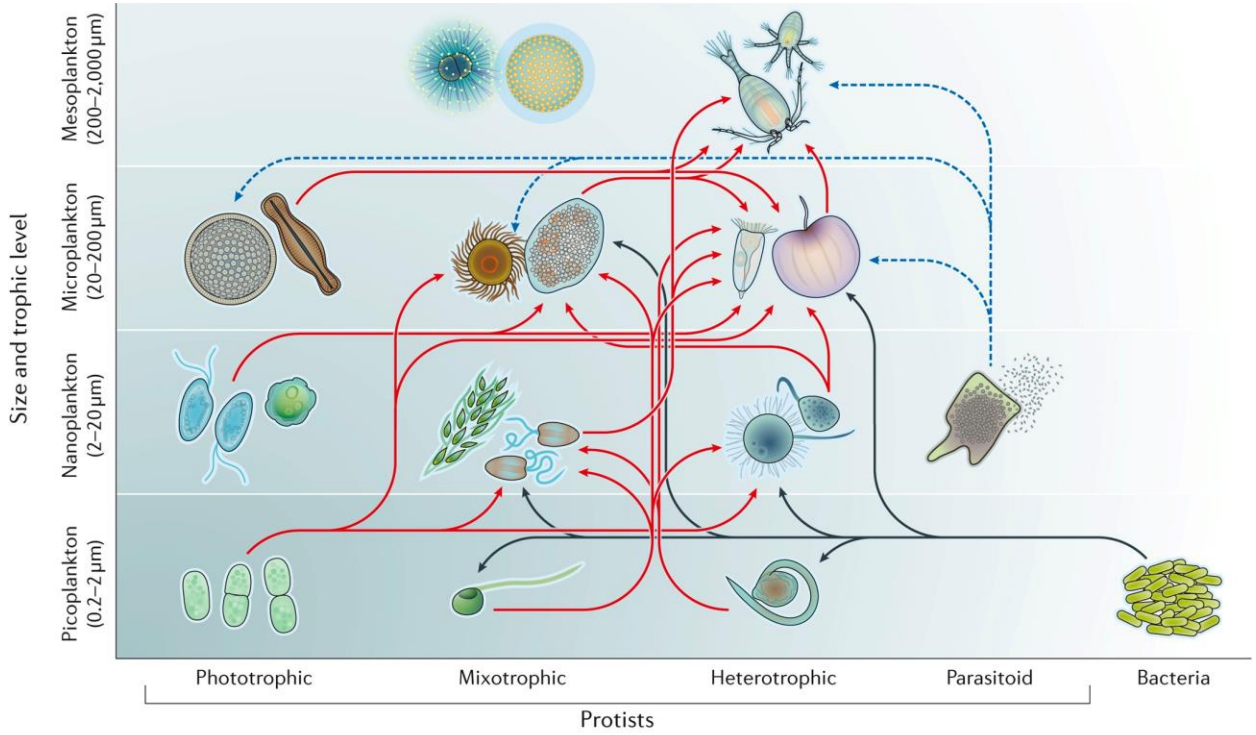
DOE CSGF

cbiomes
Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems

Margaret Mars Brisbin, Natalie Cohen, Sarah Hu, Weixuan Li, Sara Shapiro, Stephanie Dutkiewicz, Frederik Schulz, Michael Follows, and Harriet Alexander

Protists are essential players in global biogeochemical cycles

- Single-celled microbial eukaryotes are an essential link in the food web between numerically-abundant bacteria & higher trophic levels
- Phytoplankton are photosynthesizers that contribute to primary production - they fill the role that plants fill on land and on coasts in the open ocean

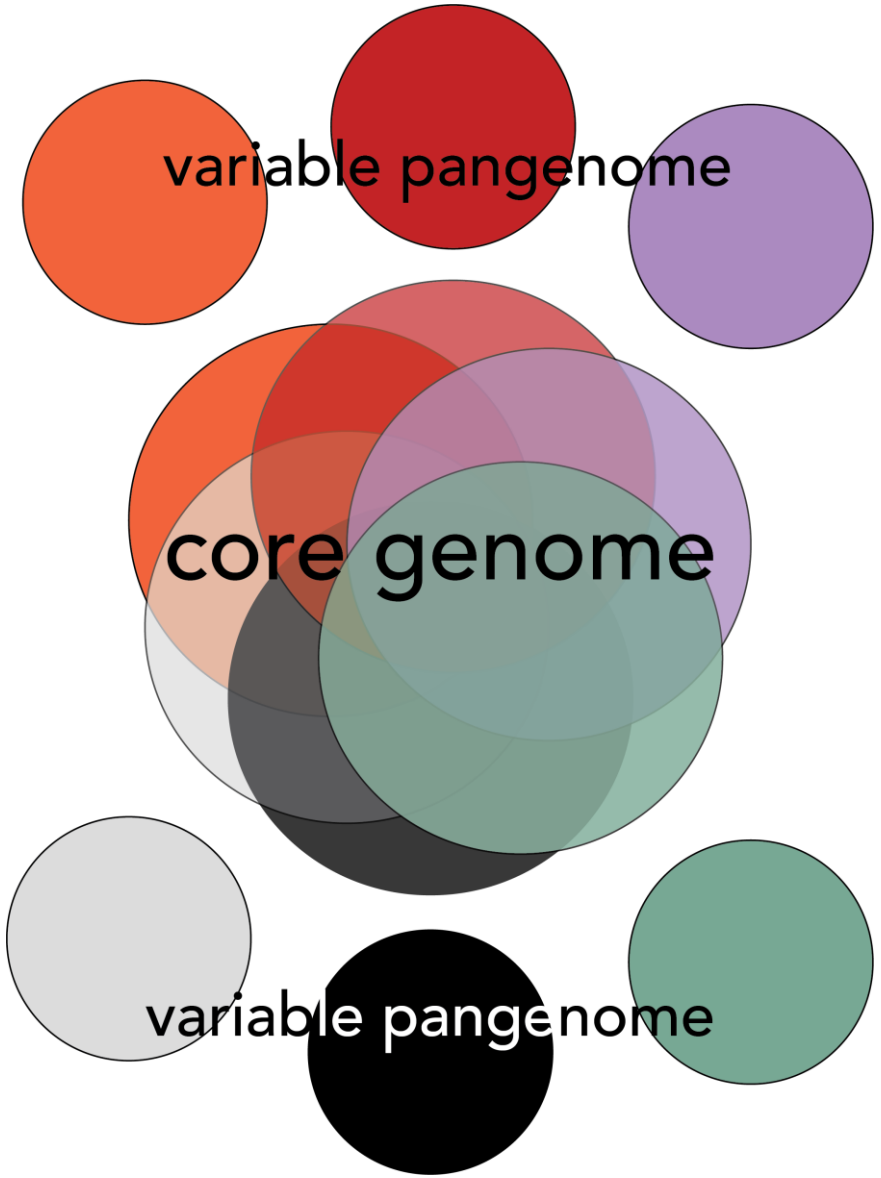


Caron et al. 2017

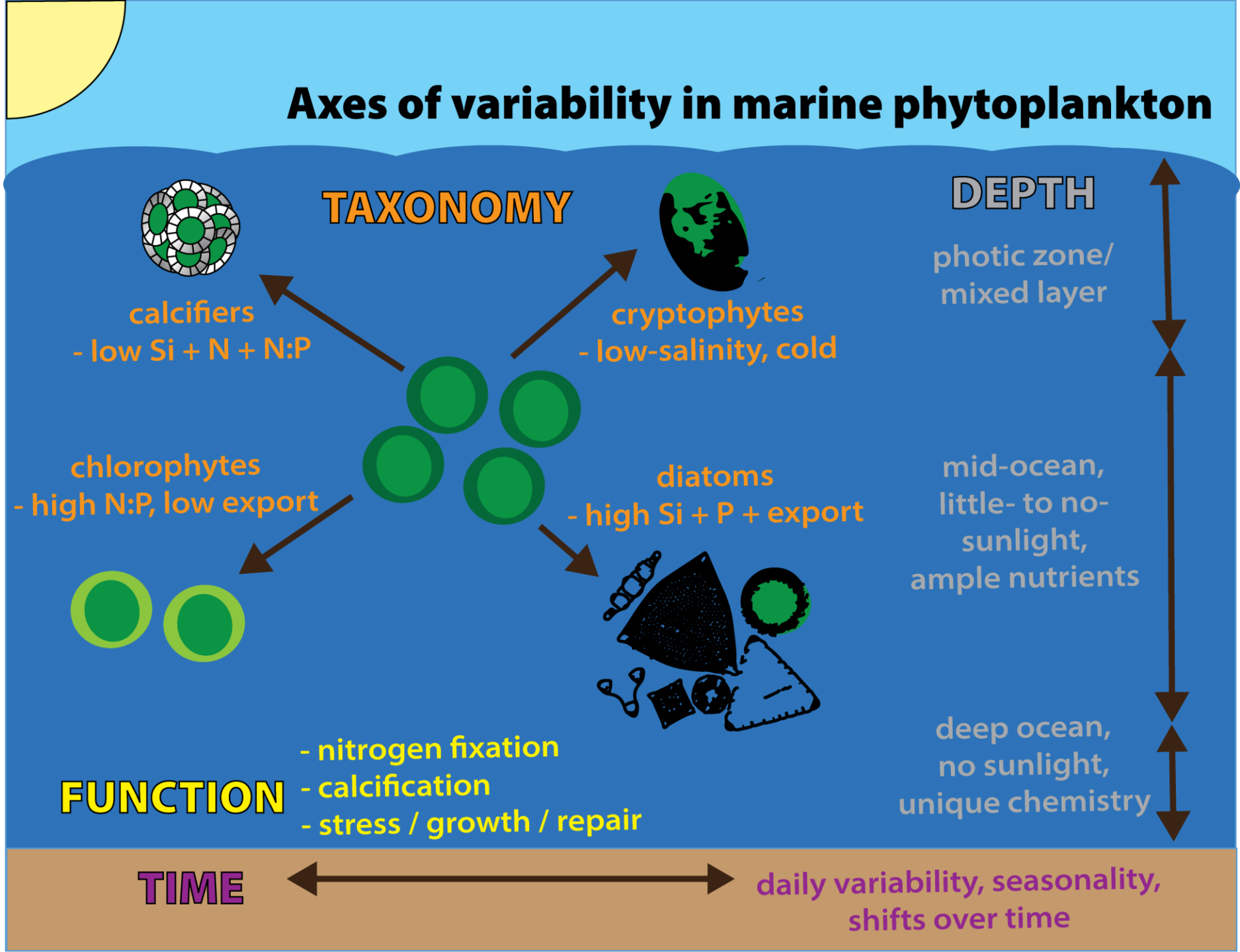
Nature Reviews | Microbiology

But their distribution, ecology, and genetics is <u>complicated</u>

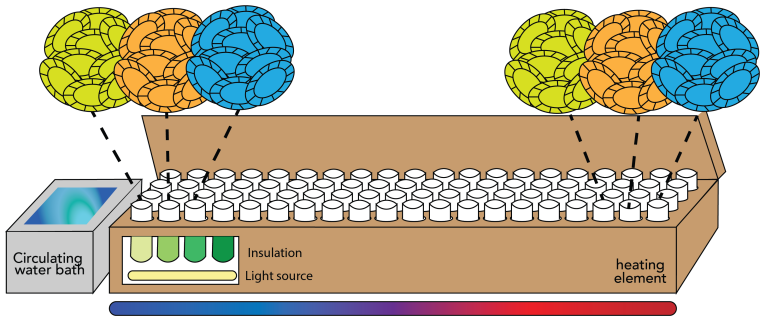
- ”Biological species” is complicated by the fact that a growing number of taxa have a variable “pangenome”
- Having many shared genes means that you’re all part of the same group; having a set of genes you don’t share defines you as a **strain**



In situ, phytoplankton genetic diversity is one part of many axes of variability



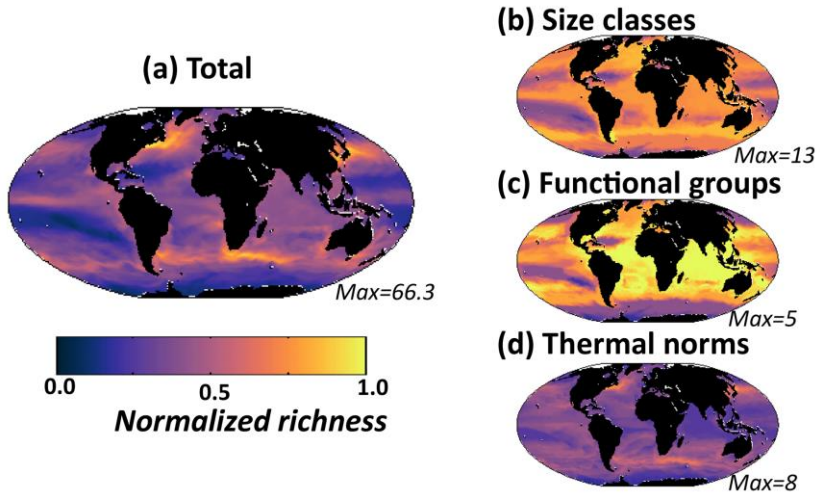
To study protists, we'll need **cross-scale tools.**



Lab physiology experiments

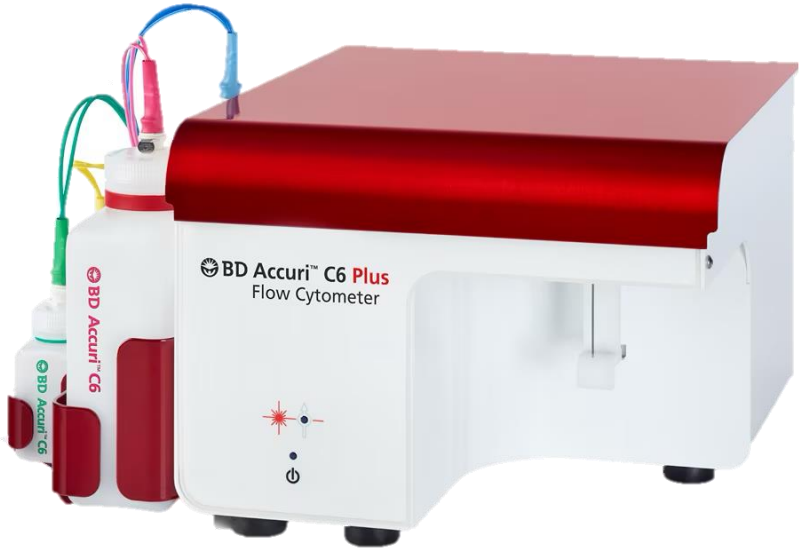
Genetic sequencing approaches and tool development

Field observations

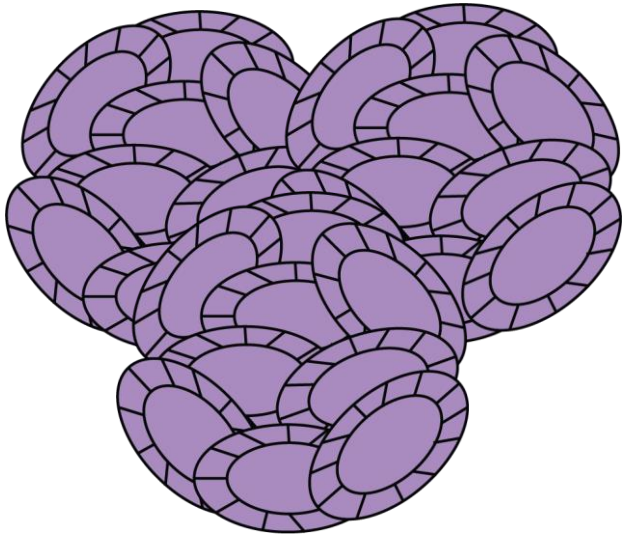


Ecosystem modeling (Dutkiewicz et al. 2020)

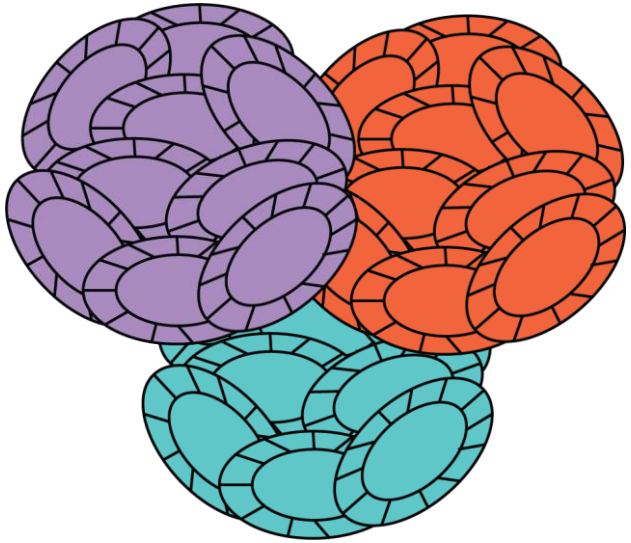
And a lot of data



Daily measurements of cell geometry and culture density via flow cytometry

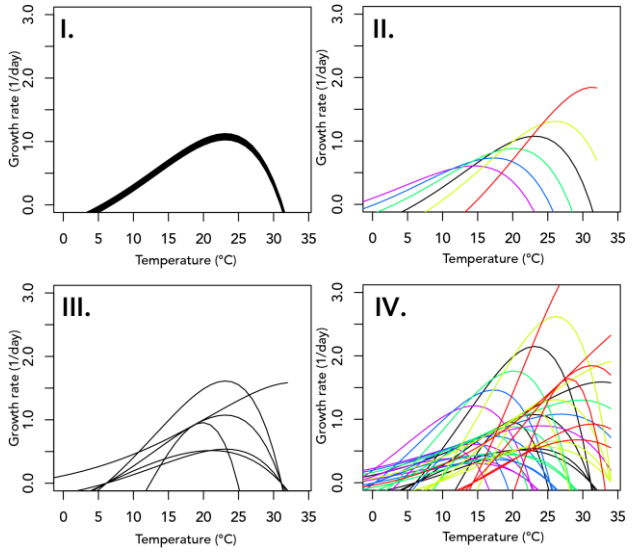


Population-level



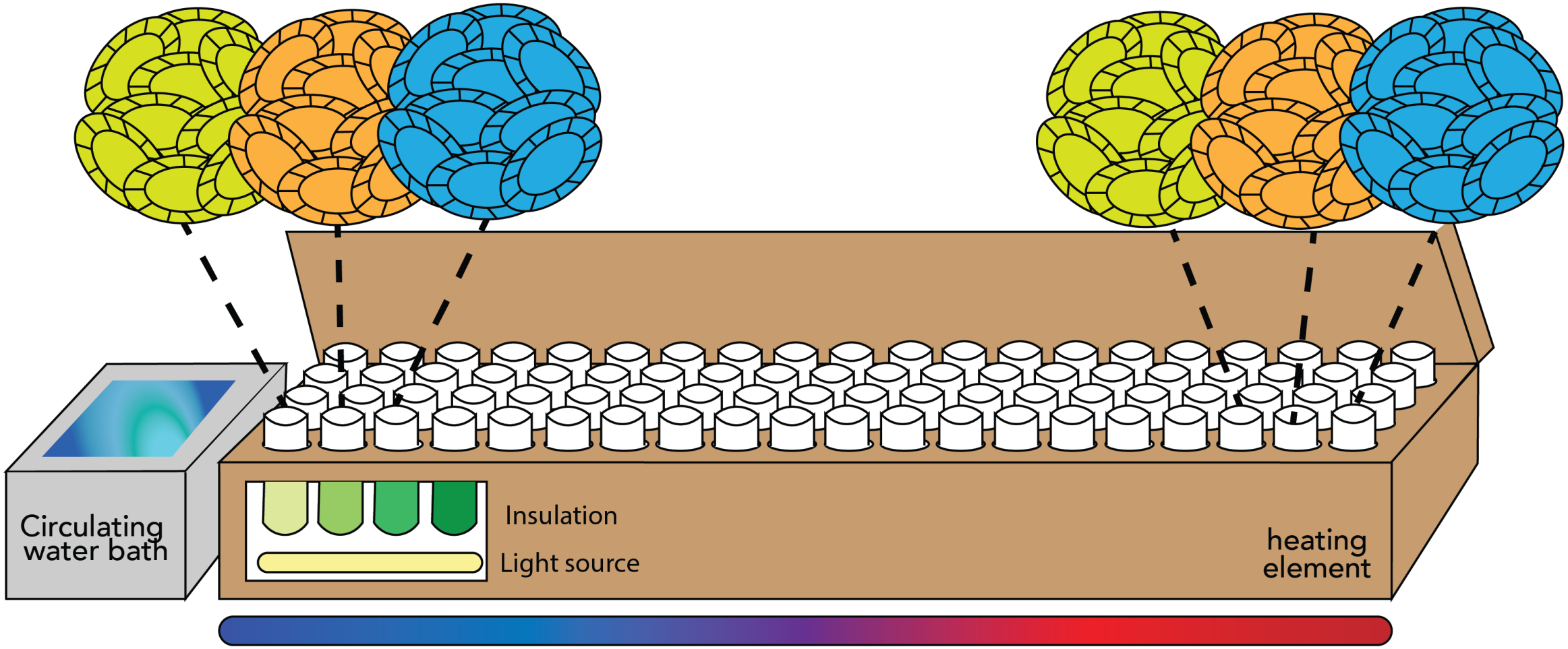
Community-level

Large sequencing datasets

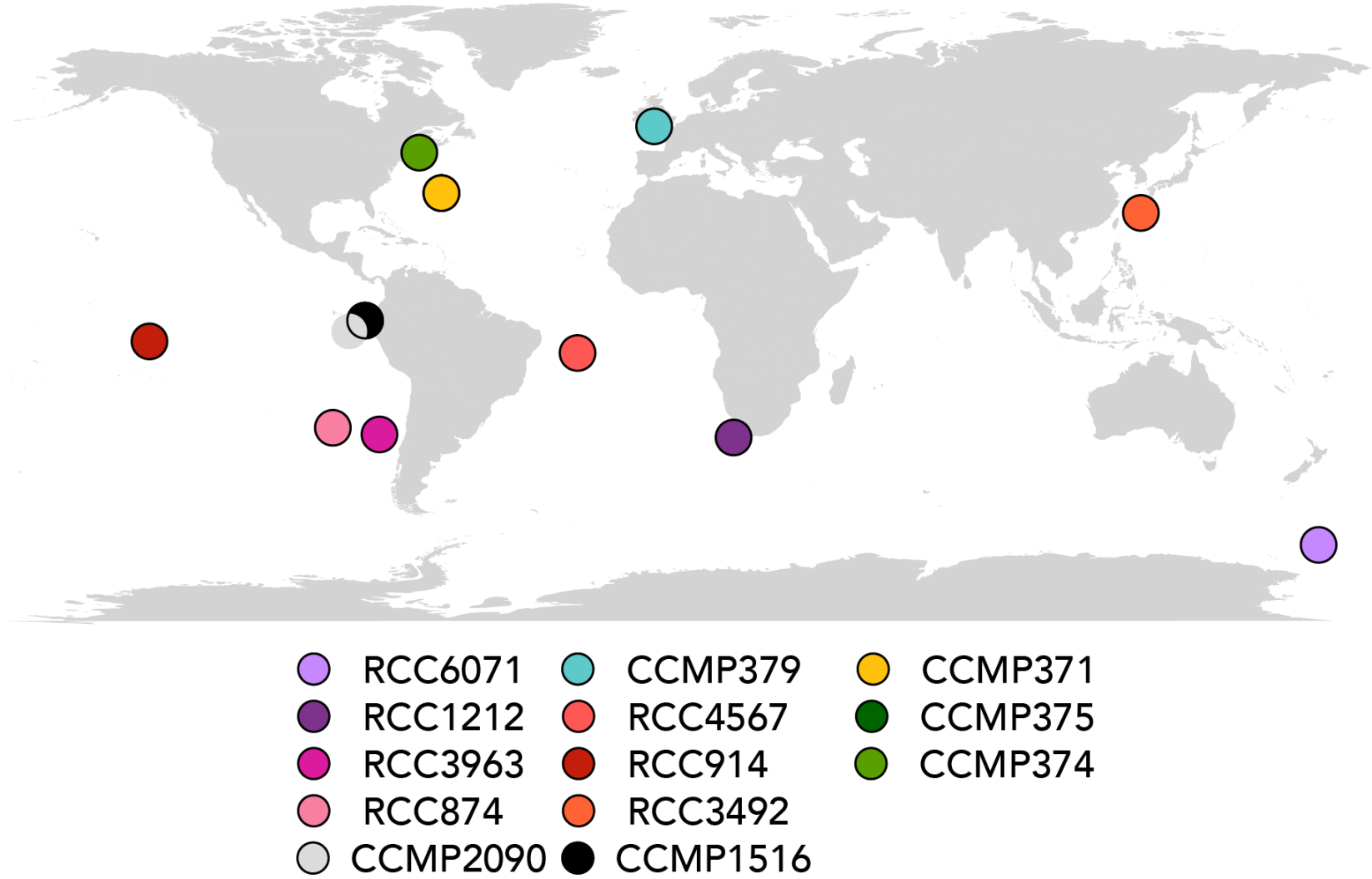


Ecosystem model outputs

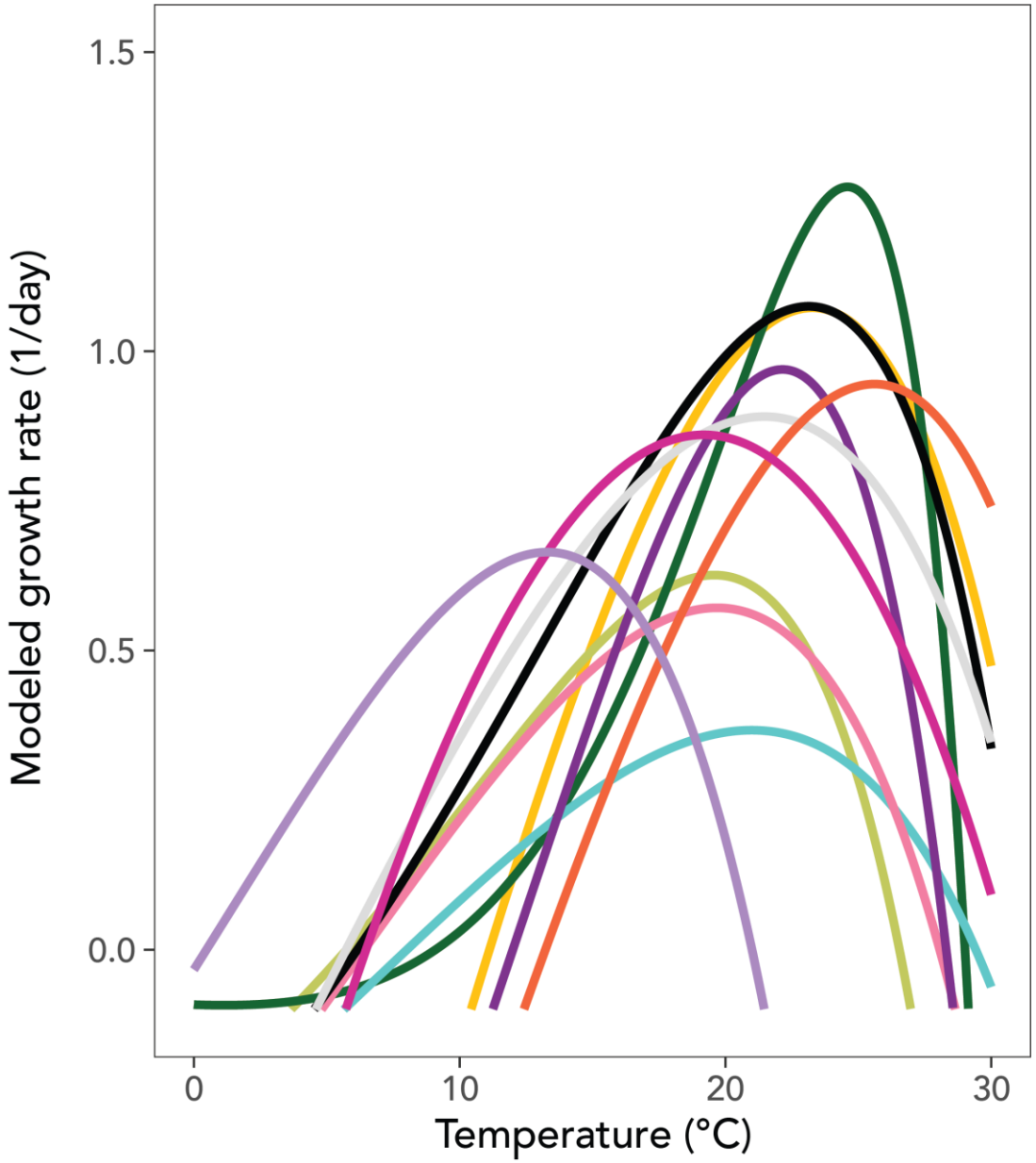
How do strains of *Emiliania huxleyi* acclimate to local environmental temperature?



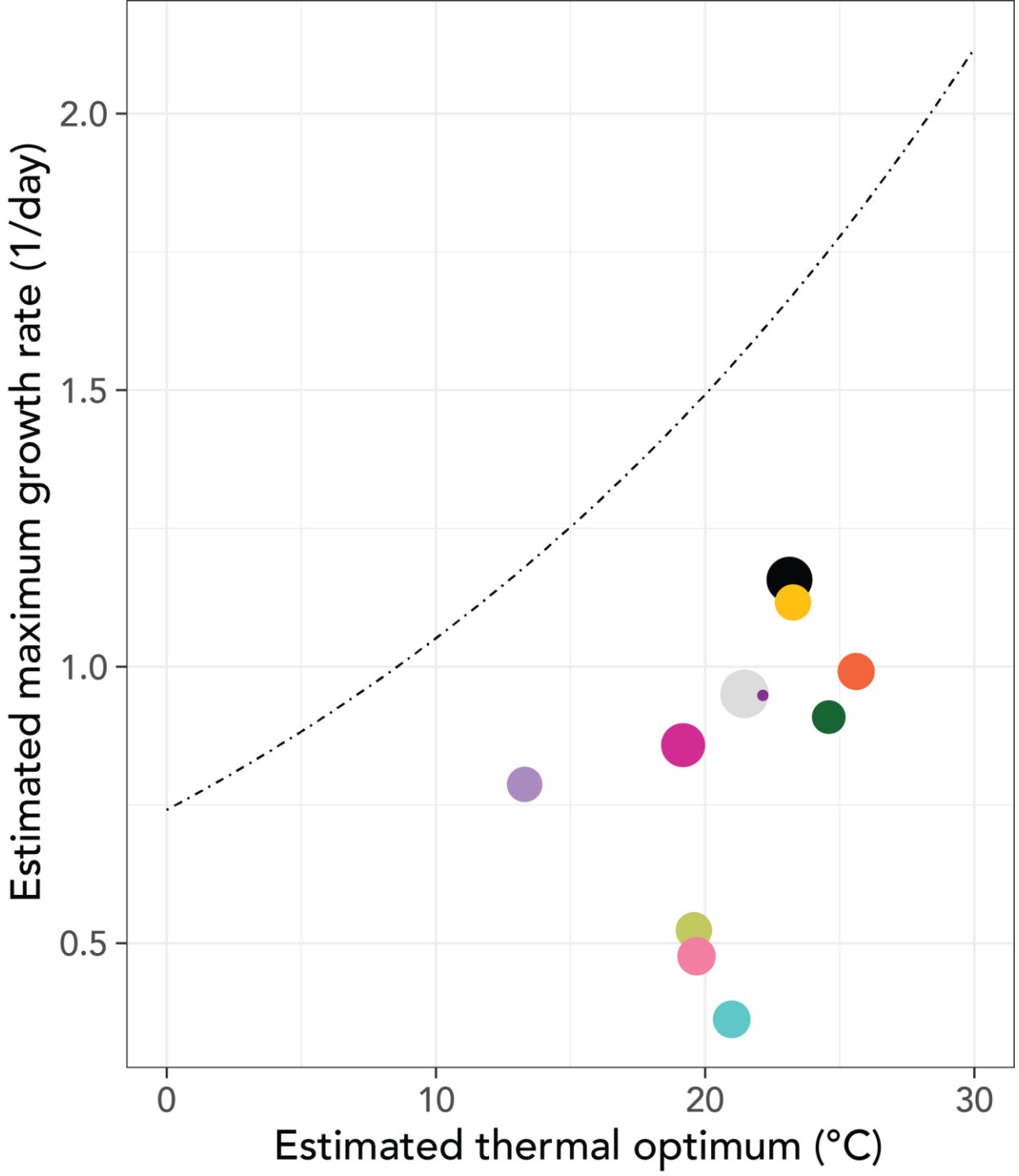
Strains of *Emiliania huxleyi* isolated from across the global ocean



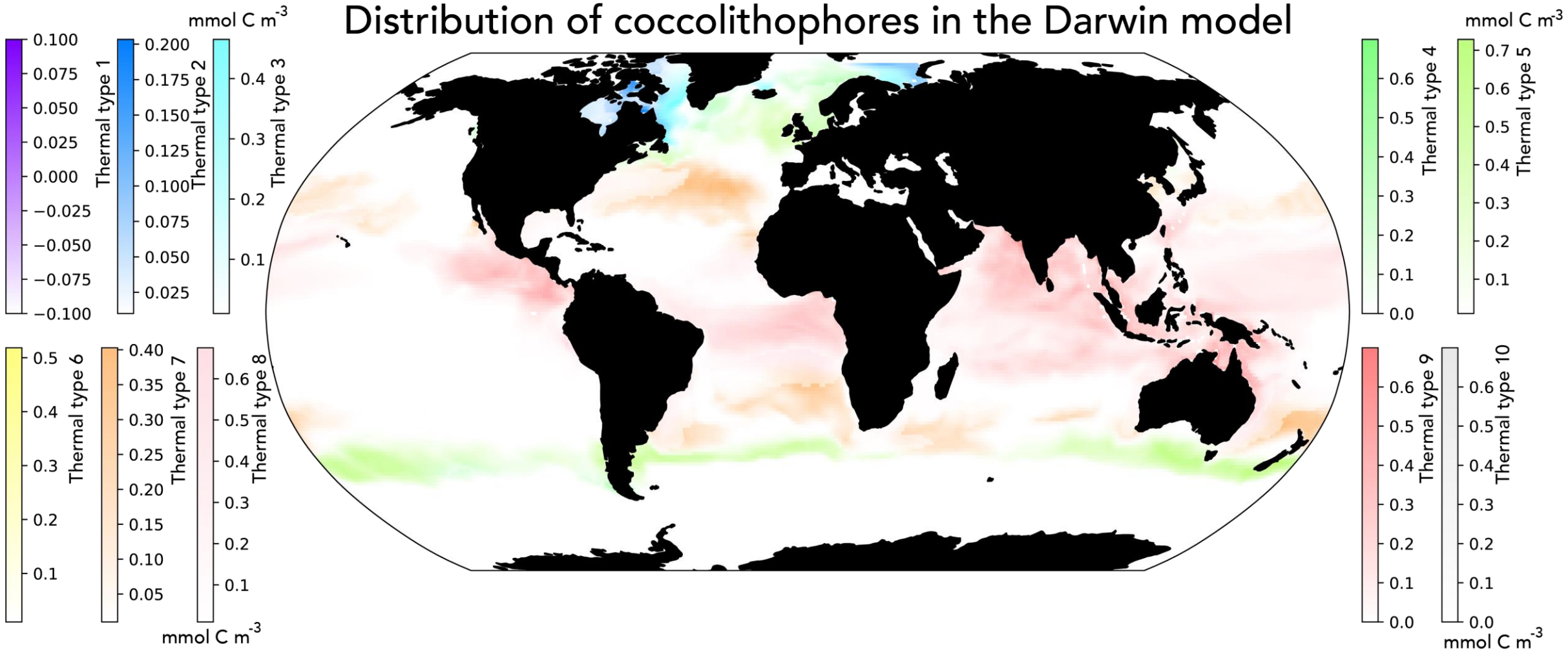
Thermal response curves vary in shape and according to original isolation latitude



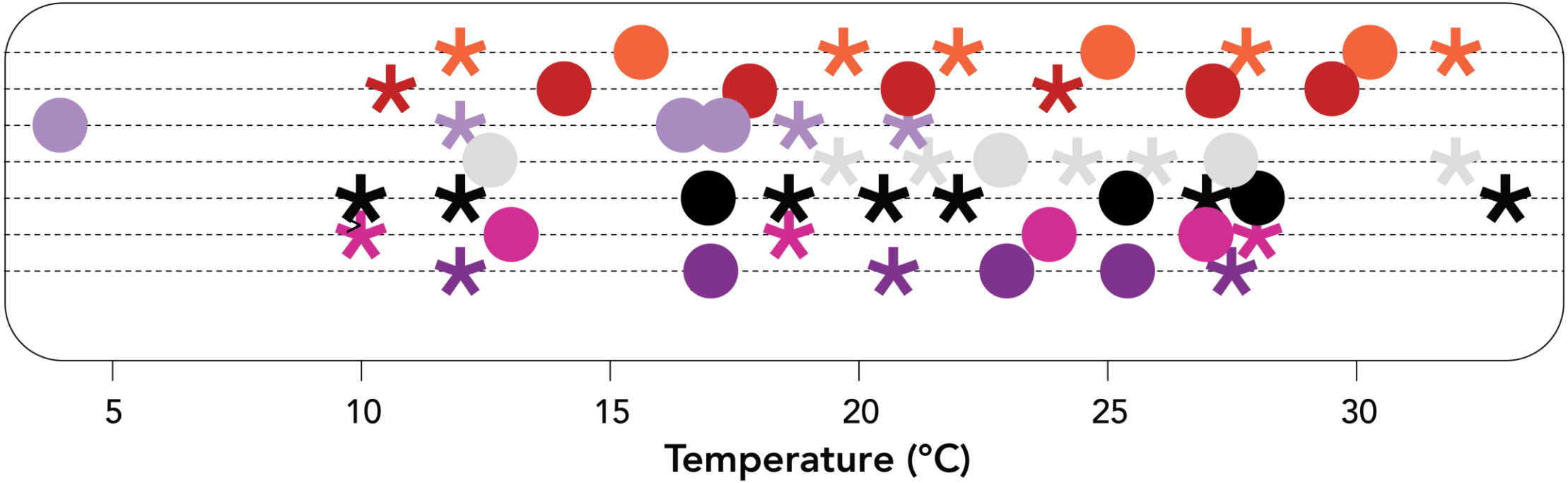
And differ from the growth rate – temperature scaling we expect from theory



We can encode these insights into ecosystem models that can predict coccolithophore distribution

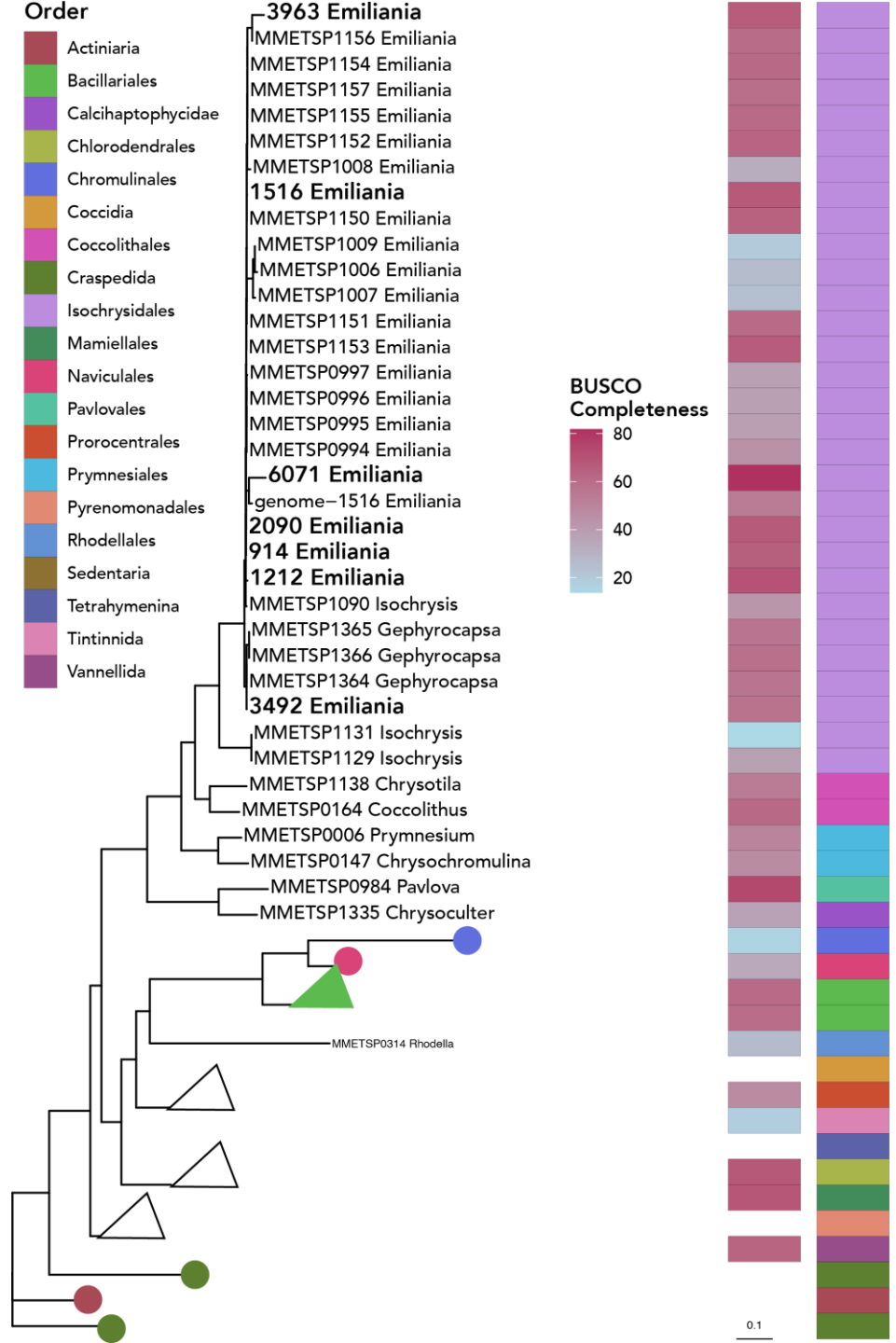


Sequencing transcriptomes can provide insight into the mechanism of thermal response

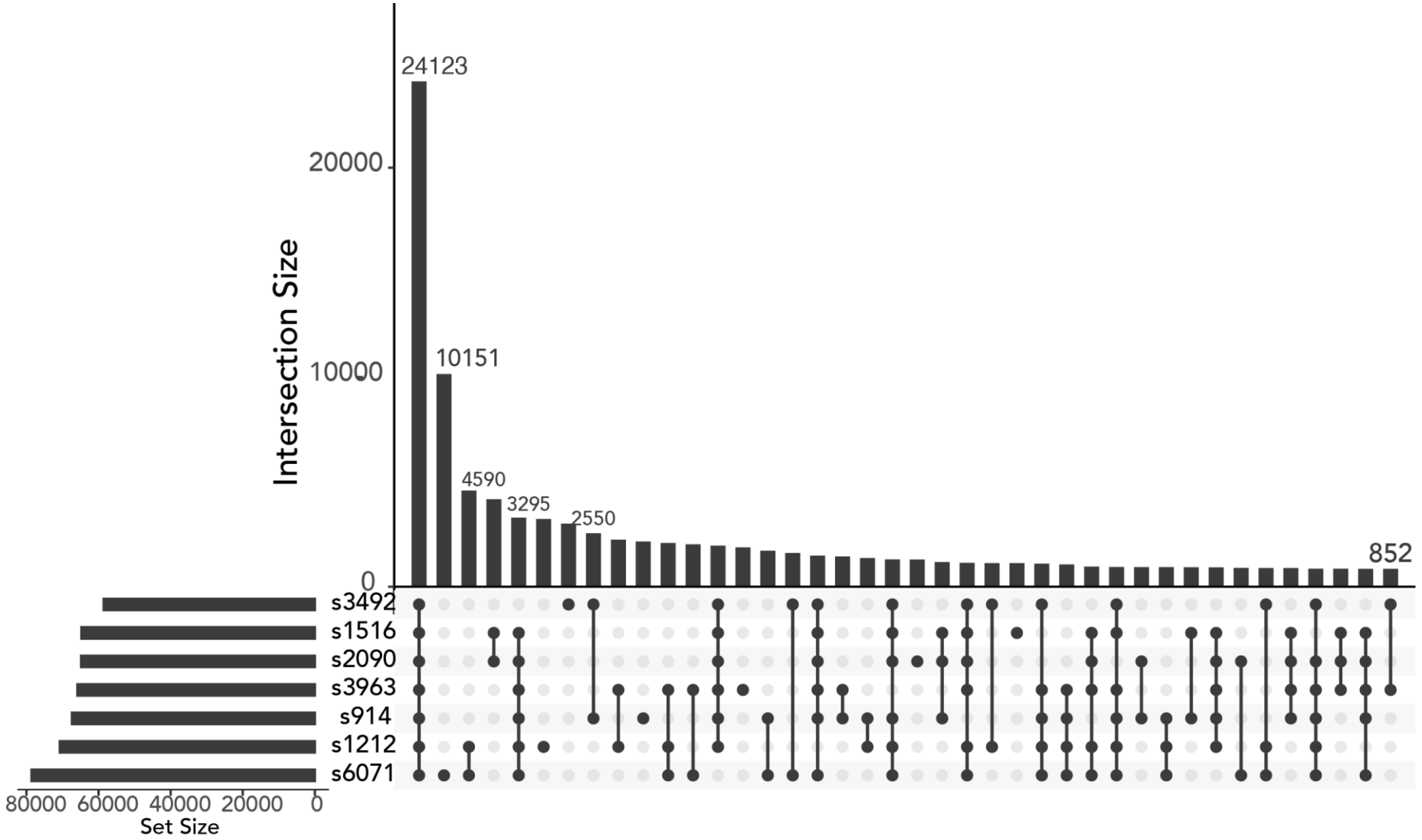


Total number of raw read sequences: >1,300,000,000
Number of predicted transcripts: 981,810; 1,161,365

Strains of *Emiliania huxleyi* with different expressed traits have similar core genes



But vastly different pools of overlapping gene content



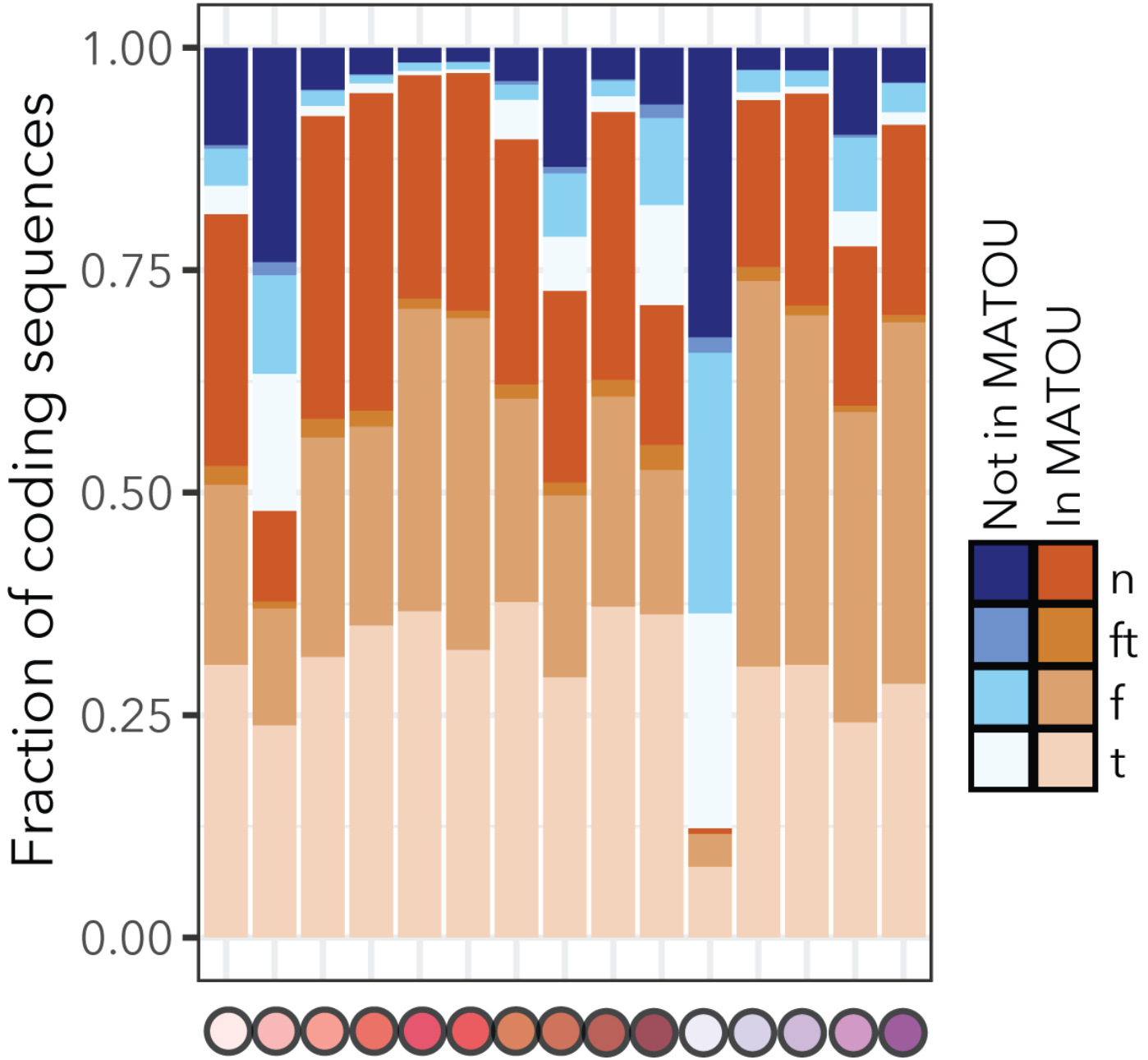
Tools are seldom built for eukaryotes

- To process metatranscriptomes, I built an open-source pipeline called eukrhythmic
- For taxonomic annotation of assembled sequences from metagenome-assembled genomes and metatranscriptomes, I built a Python package called EUKulele

Krinos et al. 2023

Krinos et al. 2021

Reassembling global ocean samples using the eukrhythmic pipeline expands gene recovery in environmental datasets

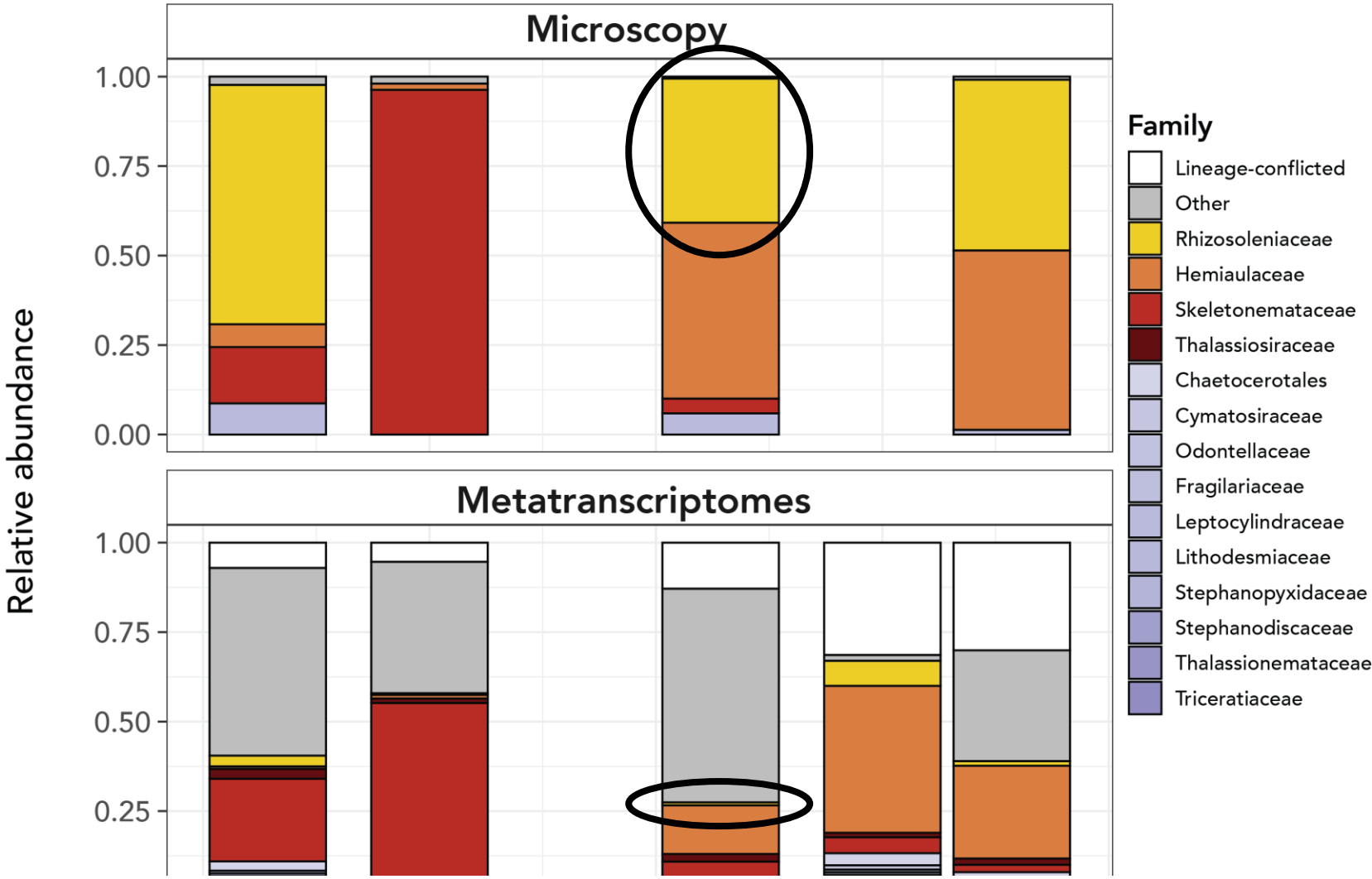


Krinos et al. 2023

Just these two samples from the Tara Oceans dataset require up-to-date computing resources

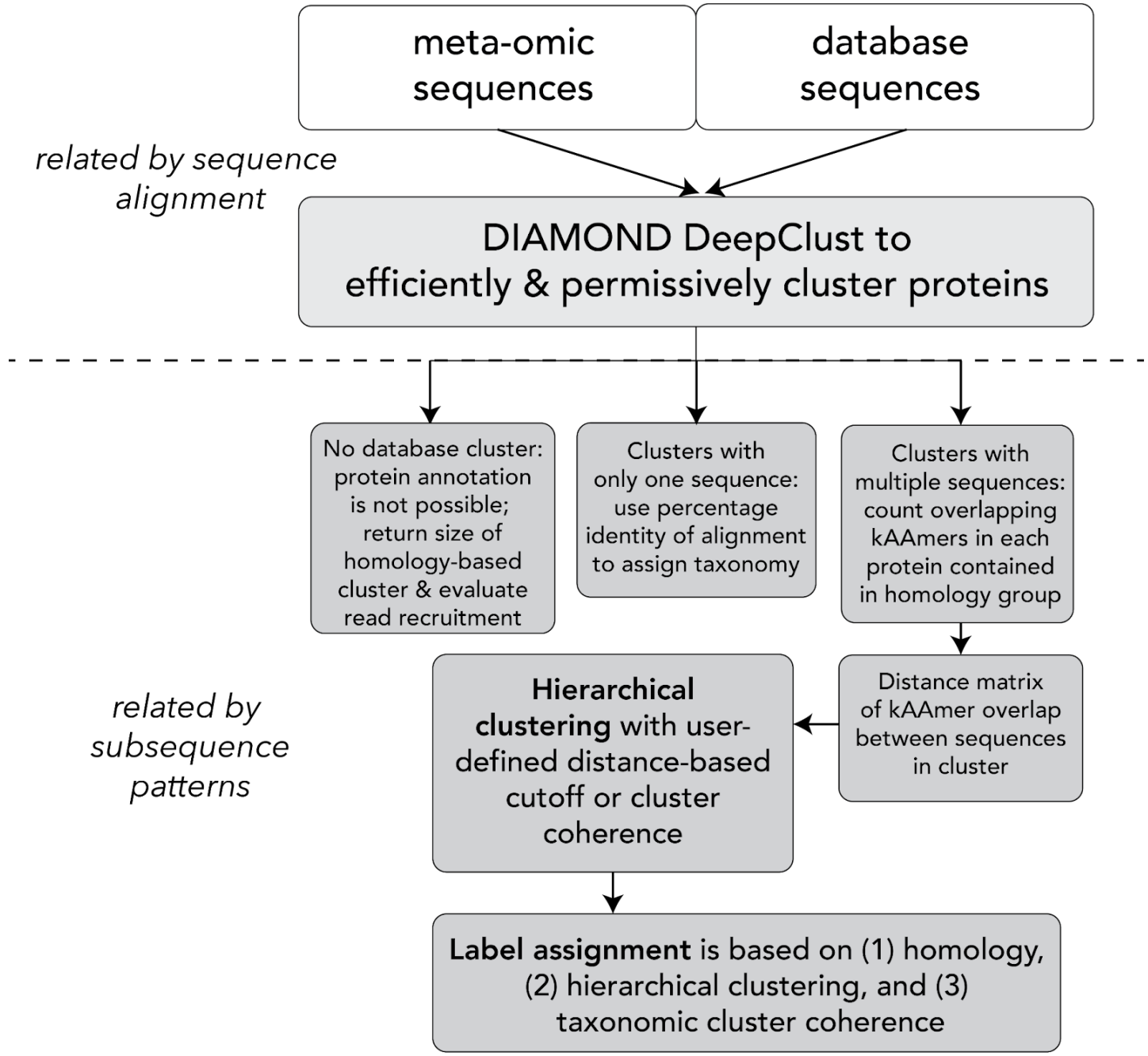
- Total size of raw sequencing reads: 94Gb
- CPU hours required for assembly using just one of the four assembly algorithms and one of the samples: 20 hours over 16 cores using 500GB of RAM
- Total number of predicted genes from one resulting assembly: **308,532**
- Total number of predicted genes from all assemblies on previous plot: **15,241,002**

Comparing traditional multi-omic annotations to microscopic counts



How can we develop new algorithms to process taxonomic annotations of protists in a high-throughput way?

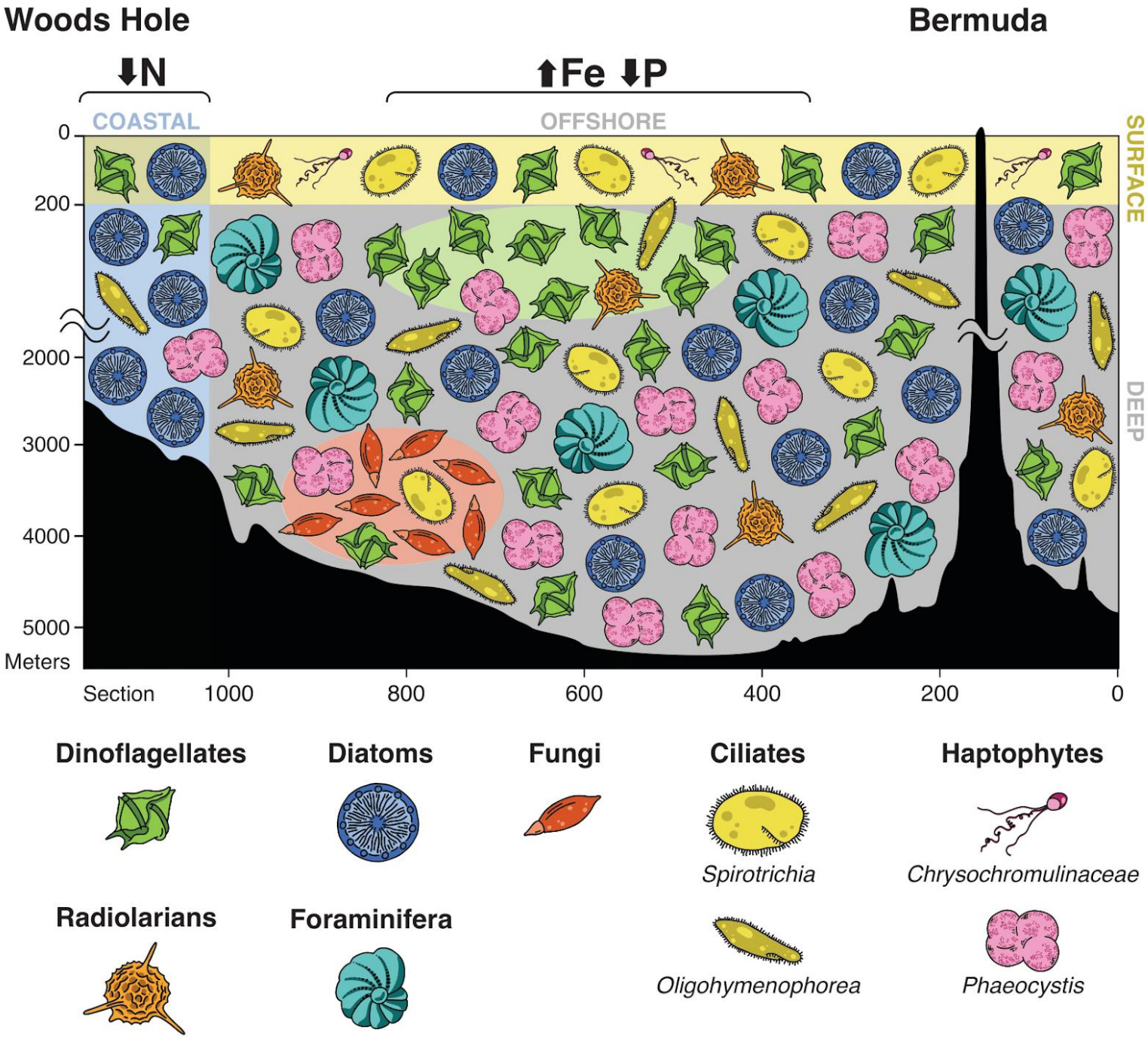
tax-aliquots: identifying taxonomic identity of eukaryotic communities via clustering



High-performance computing enables this refinement

- Just in our database, we have **15,514,482** sequences that contain **194,481** unique 4-base amino acid kAAmers and **625,624,476** unique 7-base amino acid kAAmers
- Making comparisons between the training data alone requires **2.41e14** computations

High-performance computing enables important <u>ecosystem insights</u> about where protists are found and what ecosystem services they're providing



Thank you!

- **Dr. Harriet Alexander**, WHOI
- **Dr. Mick Follows**, MIT
- Dr. Natalie Cohen, UGA
- Dr. Maggi Mars Brisbin, WHOI
- Dr. Sarah Hu, TAMU
- Dr. Frederik Schulz, JGI
- Dr. Stephanie Dutkiewicz, MIT
- Dr. Mak Saito, WHOI
- Dr. Tatiana Rynearson, URI
- Dr. Sonya Dyhrman, Columbia
- Sheean Haley, Columbia
- Sara Shapiro, WHOI
- Weixuan Li, MIT
- Quinn Perian, MIT

Contact: @ariannakrinos

akrinos@mit.edu