Intercomparison of metatranscriptomic methods for characterizing microbial eukaryote contributions to the biological carbon pump

Organizing Committee: Harriet Alexander (Woods Hole Oceanographic Institution), Natalie Cohen (University of Georgia Skidaway Institute of Oceanography), Sarah Hu (Texas A&M), Adrian Marchetti (University of North Carolina at Chapel Hill)

1. Scientific Summary and Rationale

High-throughput sequencing has become a standard measurement in the field of biological oceanography. Metagenomics, metatranscriptomics, metabarcoding, and eDNA sampling are increasingly being performed as core measurements on individual cruises as well as being incorporated into large oceanographic surveys (*e.g.*, Tara Oceans, Bio-GO-SHIP, BioGeoTraces[1–3]). However, there is currently no single gold standard practice for the collection, preservation, or processing of these samples—nor is there a sense of how variable these measures might be among labs or research cruise endeavors. Intercomparison and intercalibration of 'omic studies is a necessary next step for our field—following in the steps of international programs such as GEOTRACES.

Metatranscriptomics (metaT) is an approach that allows expressed genes, or transcripts, to be interrogated in diverse, natural microbial assemblages. As metaT focuses on the expressed fraction of gene content in mixed assemblages, it has proved a particularly useful tool to survey and study microbial eukaryotes, given that the large genome sizes of eukaryotic organisms can make metagenomic approaches intractable. In the marine ecosystem, this tool has been applied over the past decade (*e.g.*,[4–6], with the methodology and research scope continuing to evolve alongside technical advances in sequencing technology and bioinformatic capabilities[7]. MetaT has revealed insights into environmental drivers of phytoplankton biogeography[8], algal bloom formation[9,10], and patterns in diel metabolism[11]. The approach is interdisciplinary in nature and involves oceanographic sample collection, lab-based RNA extractions, knowledge of gene sequencing platforms, and computational and bioinformatic manipulation. Variations in metaT methodology exist across labs in virtually every step of the process, including how samples are collected, RNA extraction protocol, sequencing preparations, and bioinformatic choices (assembly, annotation database and alignment methods, read mapping)**.** These unknowns prevent direct comparisons of results across studies performed in different locations, times, and by different research groups (or even within research groups). Ultimately, our ability to reliably determine dominant community members present and their contributions to ocean carbon cycling is hindered by our inability to assess technical variance in our sampling methods.

Work within other sequencing-based 'omic domains has demonstrated that choices made impact community composition and recovery within datasets. In assessing extraction methodology for eDNA community composition metrics, Anderson et al. [12] observed that extraction choices such as bead treatment resulted in the absence and reduction of certain eukaryotic taxa (diatoms and chlorophytes). Similarly, an investigation in filtering choices for metagenomic sequencing showed that volume filtered and filter type did not significantly impact downstream community composition metrics, but that size fractionation did [13]. However, to our knowledge, there has been no marine-focused intercomparison effort to examine the upstream metaT sampling, extraction, and sequencing impacts on data interpretation. Work to-date on analytical differences in metaT has highlighted the variability present in metaT bioinformatic analytical choices [14]. The oceanographic community has recently recognized the urgent need for 'omic inter-standardization before an international program can be launched, in which plankton metabolism would be compared across space and time to gain biogeographical and geochemical insights [15].

2. Scientific Justification and Relevance to OCB

Microeukaryotes are core components of the biological carbon pump through their roles in primary production, carbon export, and trophic transfer in the marine food web [16], and metaT has become a standard approach used to understand their assemblages and metabolic underpinnings

[5,17]. However, critical gaps exist in our understanding of how methodological practices influence downstream biological interpretations, including estimates of community composition and metabolic function. **Our goal is to determine the extent of variability in existing metatranscriptomic pipelines through a deliberate community-led intercomparison to build international confidence in methodological choices.** The outcomes of this working group will be to develop a best practices "how-to" guide, which would increase accessibility to non-experts, and to foster connections within the marine microbial 'omic community, allowing for growth in our understanding of the strengths and limitations of this method.

This effort complements another recent OCB working group, "*Intercomparison of Ocean Metaproteomic Analyses*", which led the first comparison of mass spectrometry-based marine metaproteomics. Our working group in particular would benefit sequence-based 'omic users working with other (non-eukaryotic) organisms, as we collectively consider the factors that most contribute to variability in sequencing results. Many of these ideas were discussed at length at the OCB "*Ocean nucleic acids 'omics intercalibration and standardization"* workshop in January 2020, in which the community acknowledged the urgent need for 'omic standardization in order to interpret 'omic data among distinct lab groups, and before an international field survey and process study program could be orchestrated [15]. This is now a reality, with the recent National Science Foundation AccelNet award "*Development of an International Network for the Study of Ocean Metabolism and Nutrient Cycles on a Changing Planet*", of which PI Alexander is an organizing committee member. A targeted metaT working group is therefore timely, and intercomparison results are needed to move forward with a broad scale international ocean survey program.

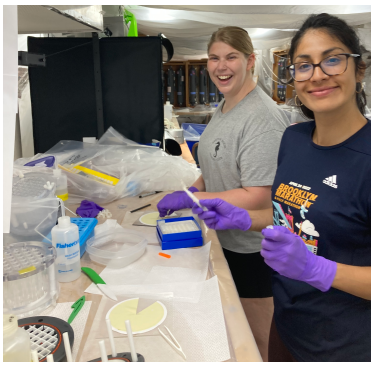

Fig. 1. McLane pump filters being sliced into smaller fractions onboard a research cruise led by Mak Saito in May-June 2023. Filters were used to concentrate biomass from Costa Rica Upwelling surface waters, and fractions were sent to working group participants in July 2023.

3. Proposed OCB Activity

I. Intercomparison data generation

Following a BioGeoSCAPES intercomparison webinar in January 2023, where M. Saito (WHOI) offered the opportunity to collect large-volume samples as part of a funded cruise, co-PIs Alexander and Cohen developed a US-based eukaryotic metaT intercomparison trial, which was performed without specified funds. A group of 12 individual labs were identified based on their expertise in eukaryotic metaT and invited to join a community-led microeukaryote metaT intercomparison. Additional members were identified with an advertisement circulated on Twitter/X to recruit researchers outside of this network. The group was US-focused to keep the number of participants manageable, but there is interest in broadening the effort to include international colleagues, and we view this OCB call as a means to include non-US participants. If funding for international partners becomes a limitation, we will encourage these individuals to attend events remotely.

Four McLane pumps were used to concentrate biomass (0.2 - 51 µm; 142 mm) at ~35m from the Costa Rica Upwelling Zone, targeting a deep chlorophyll maximum where microeukaryotes are expected to be abundant (Fig. 1). These filters serve as reference source material for the metaT intercomparison effort, and were sectioned in slices and immediately stored at -80C. Filter slices were sent to participating labs, with 20 remaining in archive and remain available, should others want to join the working group.

Participating labs agreed to fully document their RNA extraction process, with enough detail to enable reproducibility. RNA will either be prepared using poly-adenylated tail selection or ribosomal RNA depletion, and sequenced at either Columbia University or an alternative sequencing center (Fig. 2). Our design matrix therefore will ideally be used to isolate sources of variation, including extraction method, library preparation type, and sequencing center, and will additionally provide

baseline information regarding the extent of variability across labs. The primary question we seek to address is: **Do users recover a similar composition of microeukaryotic taxa and metabolic function, regardless of the RNA preparation method?** Once sequence data is available, a bioinformatic comparison can take place by analyzing all sequences using a consistent bioinformatic method (assembly, database choice, alignment method), and by allowing participants to use their own preferred method. Secondarily, we also hope to address bioinformatic considerations, such as normalization practices, assembly methods, and database reliance, to explore how these practices influence our results.

II. Planned activities

With OCB support we propose to bring together representatives of the original labs that participated in the effort, in addition to inviting new, international partners who will be chosen using a publicly advertised application process. Funds from OCB will be used to host **two in-person meetings** to interpret and process generated metaT data. The first in-person meeting will be a "Hackathon," in which members will work with raw sequence data produced and work through their preferred bioinformatic pipeline to generate the comparisons. The second in-person meeting will be a "Synthesis Session" used to interpret results, synthesize best-practice recommendations for metaT sample processing and bioinformatic analyses, outline a research article describing the results from the intercalibration exercises, and outline a white paper describing best practices, and finally form a group to contribute to a National Science Foundation proposal effort.

The first Hackathon meeting will occur over 2 days and bring interested parties together physically and provide dedicated time and attention to process and intercompare the results of bioinformatic processing by individual lab groups, with the goal of comparing community taxonomy and function across lab-generated samples. The in-person nature will naturally facilitate exchange between groups and accelerate the bioinformatic analyses of the intercomparison data. Organizers Alexander and Cohen will oversee the bioinformatic processing of *all* Columbia-generated sequences using a standard bioinformatic pipeline. This will allow for isolated effects of informatics to be inferred, whereas the former individual-user method will provide an estimate of the full range of variability produced during the wet lab and computational procedure. Harriet, Natalie, and Sarah

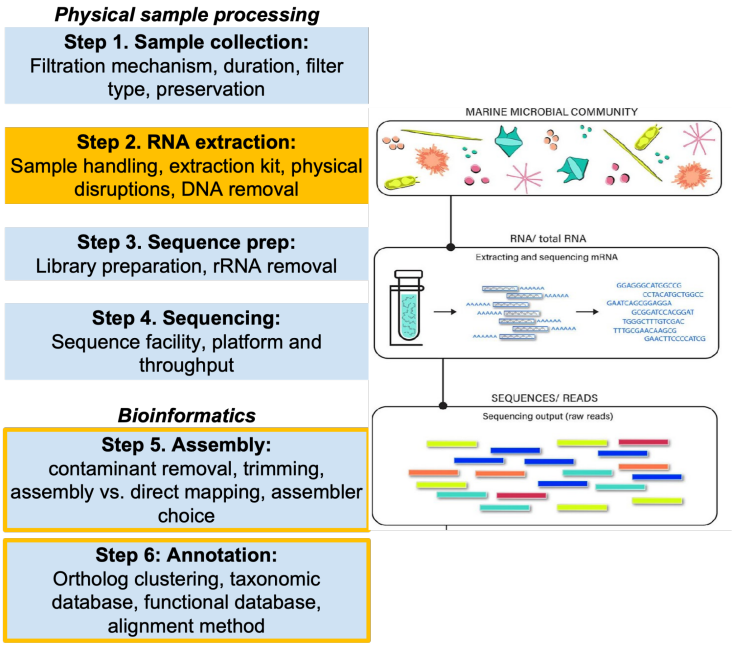

Fig. 2. Overview of the metatranscriptomic collection and analysis process. Our intercalibration exercise focuses on the steps highlighted in orange (extraction and bioinformatic analyses). Briefly, each participant was assigned at least two filter slices that will be used to test various comparisons. The intercomparison will begin at Step 2: RNA extraction, given that samples have already been collected using *in-situ* McLane pumps in Summer 2023 (see Fig. 1). We will also test the effects of poly-adenylated tail selection vs. rRNA depletion, with Andrew Allen (Scripps/JCVI) performing riboPOOL RNA depletions and the Columbia Genome Center performing the Ribozero protocol (Step 3: Sequence prep). Other participants will compare different sequencing facilities by sending one sample to the Columbia Genome Center (funded by H. Alexander), and the other to a lab-preferred sequencing center (e.g., University of Washington, Genewiz; self-funded) (Step 4: Sequencing). After sequences are obtained, all samples will be pooled and used with a consistent pipeline (Krinos et al. 2022) to determine the effects of Steps 1-4. In addition, users will be asked to process a designated sample using their own computational procedure to isolate sources of variation in the bioinformatic steps (Steps 5 & 6).

have experience organizing and leading in-person coding workshops as official instructors with The Carpentries organization, which aims to teach foundation coding practices to learners of all backgrounds. Open science educational principles, taught and practiced by The Carpentries, will be followed to ensure participants of the workshop have the resources needed to process their sequences.

We will broadly advertise an application to select 10 early career researchers (late graduate students and postdoc scholars) to participate in both meetings. Selection will be based on graduate program of study and applicability to the intercomparison, interest level in bioinformatics, and career stage, emphasizing the inclusion of students with backgrounds traditionally underrepresented in STEM. We will advertise on platforms intentionally to seek out diverse EC participants, including those who follow Black in Marine Science (BIMS) (@BlackinMarSci) and Latinas in Earth and Planetary Sciences (@GeoLatinas) on Twitter/X. Attendance at the Hackathon will serve as advanced computational biology training for EC researchers, while participation in the second meeting will lead to the inclusion of their ideas and intercomparison results in a journal publication.

In addition to invite-only in-person meetings, we plan to host public, bimonthly virtual webinars to keep our working group stimulated throughout the year and to engage with the broader microbial 'omic scientific community. We will invite domestic and international speakers inside and outside of our working group. In particular, we plan to invite individuals who have led or participated in other forms of intercomparison exercises in 'omics or ocean science more broadly, or who have domain specialty in a topic area tangentially relevant to the working group (*e.g.*, taxonomic annotation in metagenomics) to share 20-minute oral presentations outlining their findings. These talks will be followed by breakout group discussions. Information gained during these activities will be used to direct discussions during in-person meetings. We envision this being a space where marine microbial ecologists can learn from past efforts and brainstorm what a future coordinated international field 'omic survey would entail.

4. Planned Outcomes & Benefits

- **Publication of intercomparison results:** We will publish a research article covering the results of our first metaT intercomparison efforts in an open access journal. This publication will assess the impact of extraction protocols, library prep techniques, and bioinformatic analyses on the taxonomic and functional composition of metaT datasets.
- **Best-practices guide**: Given the interdisciplinary nature of metaT, users require a particular set of skills to work with the data, which can be a barrier to broad use. In recent years, efforts by individual community members to make code and bioinformatic pipelines available have greatly aided in making the tool more accessible. However, we lack a roadmap for earlier steps in the process, and we lack an understanding of how distinct bioinformatic approaches influence interpretations, although some progress has been made [18]. A direct product of the intercomparison will be an extended update to a recent marine microeukaryote metaT review [7] with verified recommendations beginning at the sample handling steps, and outlining suggested materials, protocols, and important considerations, similar to the widely regarded Geotraces "cookbook" (https://www.geotraces.org/methods-cookbook/).
- **Preliminary data for an NSF proposal**: We will use the results from this initial metaT intercomparison effort to define the scope and metaT parameters to be tested for a larger scale nucleic acid omics intercalibration and standardization effort [15].

5. Budget

Potential locations for these meetings include the University of Georgia Skidaway Institute in Savannah GA, University of North Carolina at Chapel Hill in Chapel Hill NC, and Texas A&M in College Station TX. These campuses have hosted national meetings in the past, and are within an hour drive to airports, and are equipped with the infrastructure needed to facilitate small working group events. We have selected these locations to optimize for presence on diverse,

undergraduate campuses but recognize these campuses are all located at institutes in the US South. These states unfortunately have unsupportive and, in some cases, aggressive policies towards women's healthcare and the LGBTQ community, and a history of aggression towards people of color. With this in mind, we will take considerable steps to ensure the spaces occupied during the meeting are inclusive, supportive, and free of bigotry. Communication guidelines will be discussed with participants and campus staff. Gender-neutral restrooms will be available on site. In case these steps are not sufficient and participants remain uncomfortable visiting potential meeting locations, we will send out anonymous surveys to determine the regions that are off limits to participants.

Item	**Cost**
Meeting Travel	$1,000 per person x 14 participants + 10 early career x 2 in-person events
Meeting Spaces (Wifi access, overhead projection, and AV support)	$4,000 for meeting space costs x 2 in-person events
Catering	$4,000 x 2 in-person events
Manuscript submission	$4,000
TOTAL	**$68,000**

6. References

1. Ustick, L. J. *et al.* Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* **372**, 287–291 (2021).
2. Biller, S. J. *et al.* Marine microbial metagenomes sampled across space and time. *Scientific Data* **5**, 180176 (2018).
3. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
4. Marchetti, A. *et al.* Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences* (2012) doi:10.1073/pnas.1118408109.
5. Alexander, H., Jenkins, B. D., Rynearson, T. A. & Dyhrman, S. T. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proceedings of the National Academy of Sciences* **112** , E2182–E2190 (2015).
6. Kolody, B. C. *et al.* Diel transcriptional response of a California Current plankton microbiome to light, low iron, and enduring viral infection. *ISME J.* (2019) doi:10.1038/s41396-019-0472-2.
7. Cohen, N. R., Alexander, H., Krinos, A. I., Hu, S. K. & Lampe, R. H. Marine Microeukaryote Metatranscriptomics: Sample Processing and Bioinformatic Workflow Recommendations for Ecological Applications. *Frontiers in Marine Science* **9**, (2022).
8. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**, 373 (2018).
9. Gong, W. *et al.* Molecular insights into a dinoflagellate bloom. *ISME J.* **11**, (2017).
10. Ji, N. *et al.* Metatranscriptome analysis reveals environmental and diel regulation of a Heterosigma akashiwo (raphidophyceae) bloom. *Environ. Microbiol.* (2018) doi:10.1111/1462-2920.14045.
11. Coesel, S. N. *et al.* Diel transcriptional oscillations of light-sensitive regulatory elements in open-ocean eukaryotic plankton communities. *Proceedings of the National Academy of Sciences* **118**, (2021).
12. Anderson, S. R. & Thompson, L. R. Optimizing an enclosed bead beating extraction method for microbial and fish environmental DNA. *Environ. DNA* **4**, 291–303 (2022).
13. Pascoal, F. *et al.* Inter-comparison of marine microbiome sampling protocols. *ISME Commun* **3**, 84 (2023).
14. Krinos, A. I. *et al.* Missing microbial eukaryotes and misleading meta'omic conclusions. *bioRxiv* (2023) doi:10.1101/2023.07.30.551153.
15. Berube, P. *et al. Roadmap Towards Communitywide Intercalibration and Standardization of Ocean Nucleic Acids 'Omics Measurements*. 50 http://dx.doi.org/10.1575/1912/28054 (2022) doi:10.1575/1912/28054.
16. Worden, A. Z. *et al.* Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* Preprint at https://doi.org/10.1126/science.1257594 (2015).
17. Cohen, N. R. *et al.* Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. *Nature Microbiology* (2021) doi:10.1038/s41564-020-00814-7.
18. Krinos, A. I., Cohen, N. R., Follows, M. J. & Alexander, H. Reverse engineering environmental metatranscriptomes clarifies best practices for eukaryotic assembly. *BMC Bioinformatics* **24**, 74 (2023).