Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-133-RC1, 2021

Comment on essd-2021-133

Anonymous Referee #1

Referee comment on "An 18S V4 rRNA metabarcoding dataset of protist diversity in the Atlantic inflow to the Arctic Ocean, through the year and down to 1000□m depth" by Elianne Egge et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-133-RC1, 2021

General comments

In this paper, the authors constructed an matabarcoding dataset of marine protists communities in the Northern Svalbard region of the Arctic Ocean. These data comprise samples collected at some stations from the surface to the 1000 m depth every two to three months. The total number of amplicon sequence libraries is huge and the data would be enough to address a variety of ecological questions. Environmental metadata was also prepared for each of the sampling event. I agree that the dataset can be valuable as the study area is key to understand the connectivity between Arctic and Atlantic Oceans. However, I have some serious concerns regarding the sampling and sequencing strategies used in this study.

Major issues

In my opinion, sampling strategy is inappropriate to evaluate seasonal change or size-dependence of the microbial assemblages. For example, as the sampling locations are distributed among season, it would be difficult to decipher if the observed variation of microbe is owing to the season or just to the location. Similarly, size fractionation was not consistent across seasons. Samples were taken from 3-180 μm fraction in Jan to Mar, while 3-10, 10-50, and 50-200 μm fractions were applied for May to Nov. Sequencing platform was also inconsistent across samples (ie., MiSeq and HiSeq) and the different criteria were used for the downstream sequencing processing. Thus, the users will have difficulty in interpreting their results as they have to consider the possible effects of the different location, size fraction, and sequence platform. These methodological discontinuities would collectively diminish the overall quality of the dataset, and consequently the strength of the conclusion of the analysis.

The accuracy of the eukaryotic community profiling strongly depends on the choice of

primer pair. I checked the ability of the primer set used in this study by in silico PCR analysis using a primer test tool (Silva TestPrime: https://www.arb-silva.de/search/testprime/), and found that the primer may amplify only 2% of known haptophyte sequences and 0.5% of Rhizaria sequences in the Silva SSU database. As these lineages are important components of marine microbial eukaryotes, the sequencing libraries were potentially biased due to the mismatches of some specific species. Although it is impossible to amplify all the rRNA genes in the environments, some primers have shown to be highly universal for both eukaryotes and prokaryotes (Parada et al., 2015, 10.1111/1462-2920.13023; McNichol et al., 2021, 10.1128/mSystems.00565-21). These primer may be a standard method for monitoring overall communities of prokaryotes and eukaryotes across space and time. Although there is no golden standard in a choice of the primer pair so far, it's worth pointing out.

Specific comments

L8-9: Please specify size range of each of picoplankton, nanoplankton, etc.

L59-65 and the related metadata: The sampling conditions (location, depth and filter size) were bit complicated and sometimes inconsistent. For example, readers cannot recognize which size fractions and depths were applied for each season and site from this information (metadata is not suitable to see this kind of information). I recommend the authors to make a table summarizing the sampling site, season, depth and size fraction.

L88-89: A bit difficult to find out what the authors mean (i.e., "opposite patterns" of what?)

L129: The authors should justify the use of this primer set. As described earlier, this primer may not be a universal for some of the eukaryotic lineages, such as Haptophyta and Rhizaria.

L146-147: Please specify sequence length (e.g., 150 bp, PE) of MiSeq and HiSeq sequencing.

L203: It is difficult to know if there is a seasonality of eukaryotic communities from the Figure 4. I would recommend the authors to add an ordination plot such as NMDS to overview the variation of the community composition across season, size, and depths.

Environmental data file: Some parameters lack the unit (e.g., counting data of bacteria

and virus).