Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-133-RC2, 2021

Comment on essd-2021-133

Anonymous Referee #2

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-RC2, 2021

The Arctic is undergoing major changes and data collection there is always challenging. Thus, Arctic datasets are of great value and interest. The manuscript of E. Egge and coauthors includes, as the title states, an 18S V4 rDNA metabarcoding dataset of protist diversity in the Atlantic inflow to the Arctic Ocean, through the year and down to 1000 m depth. This dataset is important and relevant. Yet, this dataset has major problems: there is inconsistency in the sampling points and the filtration procedures, as those change every month. There are also some other minor concerns in this version of the manuscript. Please see comments below.

-Concerns about the project: Samples were taken in 2014. Why is it important to publish the data, now in 2021, and in this journal? Is the dataset complete, or do you have more data from more recent samplings? May this dataset overlap with other datasets retrieved from other polar expeditions in this area?

-Concerns about the sampling stations: The distribution of the sampled stations seems chaotic . Why is not more consistent? Please explain the criteria for selecting the stations. Please provide information on the ice cover of each month (e.g. ice cover maps). What was the distance between the station and the ice cover? Please include this information in the dataset.

-Concerns about the size fractions: Why does the size of the fractions vary among months? This must be justified. Maybe the authors aimed to increase the resolution of the plankton fractions in May, August and November. The only consistent and comparable fraction is the 0.45-3 µm fraction. This is a strong point of the dataset and should be highlighted in the article. A proposal to increase the robustness of the results would be to join the 3-10 µm, 10-50 µm and 50-200 µm samples into one (3-200 µm). In total, there would be only two size fractions (0.45-3 µm and 3-180/200 µm), but will be possible to compare the larger size fractions.

-Concerns about the sequencing: There is little consistency in the sequencing process. Why were the samples sent to two different sequencing centers? Why two different sequencing machines (HiSeq vs MiSeq)? Why were some samples replicated? Why was the number of PCR cycles changed in some samples? Etc.

-Concerns about the dataset: When the project is searched at the ENA browser (https://www.ebi.ac.uk/ena/browser/view/PRJEB40133) the result is: "No records were found for PRJEB40133." This means the project does not exist, and thus the dataset is not available. Dataset is 155 samples (140+15), and 199 sequencing events. At the ASV document, there are only 198 columns, so one sequencing event is missing. The environmental table includes many variables, but accessory information is lacking on what each variable means, and its units. The latitude and longitude is wrong at the first rows.

Specific comments:

-How to refer to samples: In this manuscript it happens that individual DNA samples (where each sample corresponds to a particular station, depth and fraction) are named as "sample_sizefract". This nomenclature is misleading because it appears to refer to a size fraction in general (and thus encompasses multiple samples). I propose to refer to all DNA samples simply as "sample" or as "DNA sample". Explain this in the text if necessary. To avoid further confusions, I recommend name any other type of sample (replica, niskin bottle, chlorophyll ...) with its particular specification.

-L.40: Please explain better about the challenges. Include if necessary, arguments about adverse weather.

-L.71-77: The information in this paragraph is confusing. If light is important, the dataset should include: sampling time, daylight hours, and number of hours of sunshine during the sampling day.

-Section 2.2.3: Please say that the results of Inorganic nutrients and Chlorophyll a are in Figure 2. Where is the methodology on chlorophyll measurements, cell counts, nutrients and other environmental variables?

-Section 3.1.1: Did you use a rosette? How many bottles per rosette? What time were the casts released?

-L.105-108: There is repeated information in this section.

-L.106, 107, 111: Why is the buffer temperature important?

-L.109: was the nylon mesh previously sterilized?

-L.124: “Subsequently 4 µL RNase was added”. Why?

-L. 129-130: Please briefly comment on the advantage of this primer over others that are also commonly used.

-L.143-151: This section needs many clarifications, e.g:

-L.143-145: Why were the samples sent to two different sequencing centers? Why two different sequencing machines (HiSeq vs MiSeq)?

-L.146-147: What kind of issues did you have with Illumina Miseq in 2015, and why are they relevant? Why was it only done with two runs? Include "center" next to “GATC”, since as it is it leads to confusion.

-L.148: What were these “few samples”? and why were replicas made?

-L.149-150: And why increasing the cycles was a solution? Why not pooling samples?

-L.151: Table 1 should include: type of platform used (Hiseq / MiSeq), site (Oslo / Germany), PCR cycles (25-30)…

-L.161-162: Are there versions of the PR2 databases? which one have you used?

-L.168: better explain this “merged” and why.

-L.169: Why doesn't the 3-180 µm fraction have 40,000 reads, as the others? Why 3-180

µm fraction in Figure 3 has variable number of reads?

-L.170: please indicate if this subsampling was made with a specific function in R

-L.176-178: For consistency, previous sections should explain where the sequencing data is (ENA, link…).

-L.181: I guess "44 samples from niskin bottles" is more accurate (instead of "44 Niskin samples")

-L.182: at this point the reader cannot understand what this code means: "May_P4_net_10_50 failed"

-L.182: “These samples are in the following referred to as ’sample_sizefract”.Where?

-L.181, 185, 187 and others: change "sample_sizefract" to "sample" or "DNA sample". See general comments.

-L.185: please explain what is ENA.

-L.187: why they were merged? Please explain this.

-L.191: the removal of singletons was not mentioned in section 4.2

-L.192 and 193: I imagine that when you say “size fraction” you mean “sample”.

-L.194: I recommend to separate the numbers with ";" instead of "," .

-L.195-196: Please clarify this point.

-L.200: how was the slope calculated?

-L.202: why is this correlation important?

-L.204-205: both figures, Figure 4 and Figure A1, show: metabarcoding reads, ASV, division and class levels. Please rephrase.

-L.206-207: Is the name of the fractions (pico-, nano- and micro-) important? Then explain them at the beginning of the manuscript, and include them in datasets and graphs.

-L.209: Please explain why in Table A2 some groups do not have any >0 values.

-L.210: Please explain which groups are heterotrophic or parasitic.

-L.211: At this point data is "relative abundances", not "read abundances".

-L.210-255: please explain where the values of “richness” comes from.

-L.219: please explain which groups are phototrophs

-L.227: please explain which Division / class corresponds to Diatoms

-L.257: This dataset is descriptive and does not include patterns or dynamics.

-L.258: This dataset is not about food webs.

-L.256-260: An interesting argument to add is that this dataset is a baseline for future studies aiming to determine temporal changes.

-L.262: https://www.ebi.ac.uk/ena/browser/view/PRJEB40133 says: No records were found for PRJEB40133.

-L.263: I recommend to remove this part "corresponding to the size-fractionated plankton samples"

-All figures and tables: captions need improvement. Should provide more information and guidance. Here some comments:

-Figure 2: some profiles are missing, e.g.: there are at least 6 stations in March, and at least 4 go deep. Something is wrong with the y-axis: Is it logarithmic?

-Figure 3: change “sample_sizefract” by “sample” or “DNA sample”. Figure caption needs improvement. The reader here is lost. Avoid using references like "asvtab3_merged_subsamp_readnum.txt". Better if you make other types of references, for example: "see details at Table X ..."

-Figure 4: remove codes and put an understandable legend (e.g. "1m" instead of 0001). If in the bars, the order of protist groups is from left to right, the same order should be in the color code of the legend, but it is the other way around. The two fractions 3-180 µm and 3-10 µm are very different, and should appear separately (different columns).

-Figure A1: this figure needs to be linked with the main text.

-Table A1: include the N (number of samples) included in each size fraction.

-Table A2: Please explain why in Table A2 some groups do not have any >0 values.