Supplementary material

for "Metagenome-Assembled Genome abundances do not specifically correspond to 18S V9 metabarcode abundances in *Tara* Oceans dataset"

The output data files and Figures are available online by the link: https://github.com/ZavadikD/MAG-barcode_correspondence . The code and datasets produced are available on GitHub.

Table S1.

The data used for the analysis. The content of each dataset and the exact information which has been used is indicated, along with the hyperlinks to the datasets and, where relevant - publications linked to the datasets.

Content description	Data used	Link to the data storage	Reference	File/variable name
Data on SMAGs	Sheet 4, columns 1-2 SMAG name and its taxonomic assignation; Sheet 7 - SMAG abundances across stations	https://www.genoscope.cns.fr/tara/localdata/data/SMAGs-v1/Supplemental_Tables.zip	https://doi.org/10.1101/2020.10.15.341214	Table_S04_distributions_nr_SMAGs.xlsx
Data on V9 barcodes - organised at OTU level	Representative V9 barcode abundances across stations, taxonomic assignations, and other metadata	https://zenodo.org/record/6794351#.YxIe4NJBxhE - newer version https://doi.org/10.5281/zenodo.3768510 - older version	https://doi.org/10.1126/science.1261605	globaldataset.otus.v20171106.withfunctions.pub.tsv
Sampling event information for SMAGs	Sample.ID and Sample.mat columns	https://doi.pangaea.de/10.1594/PANGAEA.840718?format=html#download	https://doi.org/10.1594/PANGAEA.853810 https://doi.org/10.1594/PANGAEA.859953	TARA_CONTEXT_95_SEQ_STATIONS.txt.gz
Sampling event information for V9 metabarcoding dataset	Sample_seq_id and Sample.mat columns	https://doi.pangaea.de/10.1594/PANGAEA.843017	https://doi.org/10.1126/science.1261605	TARA_CONTEXT_95_SEQ_STATIONS_V9.tsv
A "vocabulary" table with taxonomic assignments of barcodes and SMAGs matched with each other	(Columns 1 and 2)	https://github.com/ZavadikD/MAG-barcode_correspondence/blob/main/vocabulary_taxa.csv	this article	vocabulary_taxa.csv
A "vocabulary" table with metagenome identifiers matched with station/sampling event identifiers	(Columns 1 and 5)	https://github.com/ZavadikD/MAG-barcode_correspondence/blob/main/TAGs-18S-V4_NAME-PROJ-MAT-READ-CAB_nico.list	this article	TAGs-18S-V4_NAME-PROJ-MAT-READ-CAB_nico.list

Table S2.

The datasets produced during the analysis. Data for each taxonomic group is written into a separate file with the file prefix being the name of the taxonomic group.

Columns	Rows	Measure of	file name suffix
station IDs	SMAGs from the subset	SMAG assembly abundance	dataSMAGs_df.tsv
station IDs	V9 barcodes from the subset	V9 barcode abundance	data_V9_relabund.tsv
"md5sum" - ID of representative V9 barcode,"SMAG" - SMAG id,"rho" - correlation/proportionality coefficient estimate,"pid" - %identity of the representative barcode with the closest match to reference V9 sequence,"lineage" - taxonomic group of assigned to the representative barcode,"sequence" - the sequence of V9 barcode,"refs" - closest reference V9 sequence,"abundance" - total abundance of the V9 barcodes of the OTU	SMAG-barcode pairs	all pairwise barcode-SMAGs correlation/proportionality + miscellaneous information	all_scores_relabund.tsv
		top-3 barcodes with the highest correlation/proportionality value for each SMAG + miscellaneous information	best_3_scores_relabund.tsv

Appendix S1.

Note on the "propr" function optimal settings.
"Propr" function allows computing a false discovery rate. This has been used to evaluate the performance of proportionality coefficient calculation after different 0 replacement and log-transformation procedures. In Fig.S1 there is a comparison of FDR of four different setups tested - using either clr or iqlr transformation, and 0s being replaced either by the smallest value or with Box-Cox transformation, "alpha" argument of the "propr" function being set to 0.5, as recommended in "propr" package Vignettes. As seen from Fig.S1., Box-Cox transformation results in many taxa showing FDR >= 0.05, even with the proportionality cutoff being 0.25. This is especially clear if Box-Cox transformation is combined with iqlr transformation. Clr and Iqlr transformations perform almost equally if 0s are replaced with the smallest value (Fig.S1.). Finally, the preference was given to iqlr transformation for it being sub-compositionally coherent.

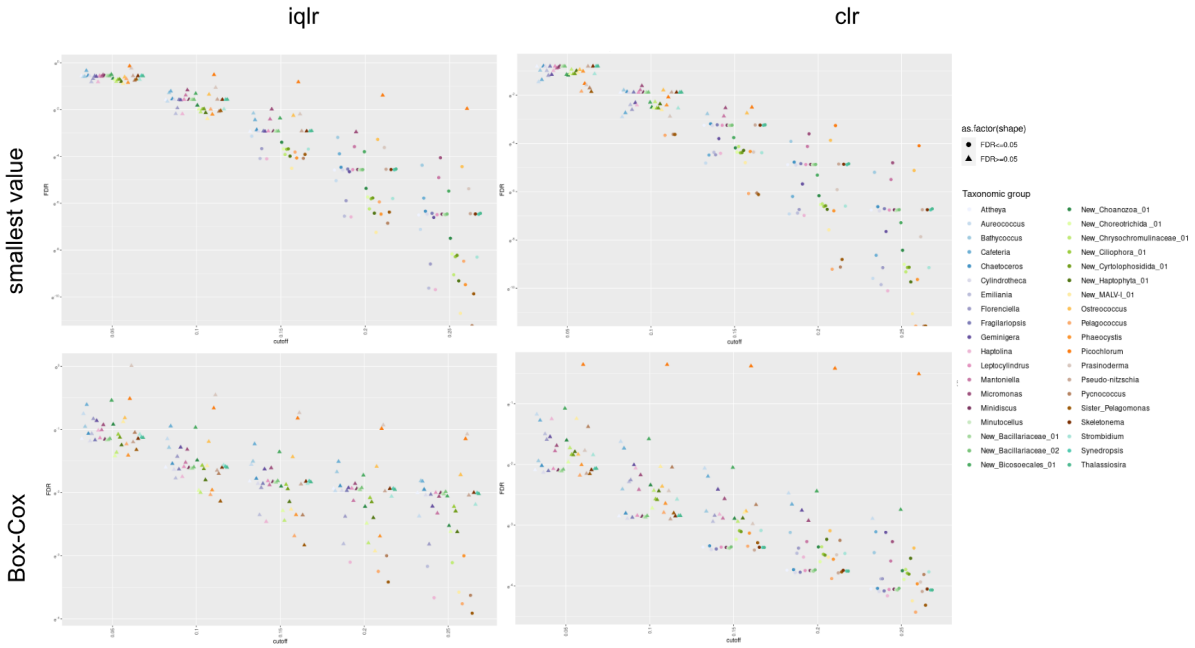

Fig. S1. FDR values for different data preprocessing setups in "propr" function arguments. FDR values are represented on the y-axis; proportionality value cutoffs for which FDR is estimated are plotted on the x-axis. Colours depict the taxonomic groups to which MAG is assigned. The shape of the points indicates if FDR is higher or lower than the threshold value of 0.05. Transformation procedure in indicated in column names of the grid, and 0 replacement procedure - as row names.