nature portfolio

https://doi.org/10.1038/s41586-024-08313-z

Supplementary information

Diversity and biogeography of the bacterial microbiome in glacier-fed streams

In the format provided by the authors and unedited

Supplementary Information

for

Diversity and biogeography of the bacterial microbiome in glacier-fed streams

Leïla Ezzat[1,2*], Hannes Peter[1], Massimo Bourquin[1], Susheel Bhanu Busi[3,4], Grégoire Michoud[1], Stilianos Fodelianakis[1], Tyler J. Kohler[1,5], Thomas Lamy[2], Aileen Geers[1], Paraskevi Pramateftaki[1], Florian Baier[1], Ramona Marasco[6], Daniele Daffonchio[6], Nicola Deluigi[1], Paul Wilmes[4], Michail Styllas[1,7], Martina Schön[1], Matteo Tolosano[1], Vincent De Staercke[1], and Tom J. Battin[1*]

***Corresponding Author:** tom.battin@epfl.ch; leila.ezzat@gmail.com

(1) River Ecosystems Laboratory, Alpine and Polar Environmental Research Center, Ecole Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland

(2) MARBEC, University of Montpellier, CNRS, Ifremer, IRD, Montpellier, France

(3) UK Centre for Ecology and Hydrology (UKCEH), Wallingford, United Kingdom

(4) Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

(5) Department of Ecology, Faculty of Science, Charles University, Prague, Czechia

(6) Biological and Environmental Science and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

(7) Institut de Physique du Globe de Paris, IPGP, Paris, France

Table of Contents

Supplementary Note 1. Detailed results from variance partitioning for the GFS microbiome.

Overall, the model explained 54.9% of the total variability of the GFS microbiome composition, which was spatially structured across scales ranging from the linear gradient (explained variance: 15.9%, $F_{3,134}=9.8$, $p=0.001$), among (20.8%, $F_{9,131}=5.97$, $p=0.001$) and within (16.4%, $F_{60,122}=1.7$, $p=0.001$) mountain ranges. Interestingly, spatial processes played a significant role in explaining the composition of the GFS microbiome. The explained variance is primarily driven by pure spatial processes both among and within mountain ranges (11.4 % and 12.2%, respectively) and by spatially structured environment (among: 9.4%; within: 4.2%). Pure environmental processes explained only 1.8% of the total variance.

Supplementary Method 1. Detailed methods regarding the additional denoising approaches used to analyse the dataset.

UNOISE3

Raw reads were trimmed and primers were removed using cutadapt (v2.10)[53]. Then they were merged with vsearch (v 2.7.0)[100]with a maximum number of mismatch in the alignments of 10 and a minimum size of 400 bp. Low-quality sequences were removed with the option fastq_filter of vsearch accounting for a maximum expected error of less than 1. Sequences were dereplicated with the derep_fulllength command, clustered with cluster_unoise and chimeras were removed using default parameters, all using vsearch. The reads were mapped back to the centroids with an identity threshold of 99%. Further analyses (e.g., taxonomy assignment) were processed in QIIME2[54] described in the *Method* section.

DADA2

A total of 883 amplicon sequence libraries were produced, comprising 868 benthic sediment samples, six blanks and nine mock communities. Paired-end sequencing generated a total of 158,774,383 reads across the 883 libraries, with an average read number of 163,696.9 ± 48,997.5 for sediment samples. Raw sequences were first trimmed of primers using the plugin Trimmomatic[101]. Briefly, reads were truncated in 4 pb sliding windows when the mean quality dropped below a Phred score of 15. The three leading and trailing nucleotides were removed, and reads shorter than 200 pb were discarded. Amplicon sequences were processed using Quantitative Insights into Microbial Ecology 2 (QIIME2, 2020.8) workflow. The demux plugin was employed to visualize interactive quality plots and assess read quality. Using DADA2 default parameters, primers were trimmed, and reads were subsequently denoised and joined to produce amplicon sequence variants (ASVs). At the beginning of each forward and reverse read, 17 and 21 nucleotides were removed, respectively. Forward and reverse reads were truncated at 280 and 220 bases, respectively. These steps resulted in a total of 104,487,984 reads across 883 samples and 259,022 ASVs. Taxonomy was then assigned against the SILVA reference database (v138)[55] using the algorithm classify-sklearn. The ASV and taxonomy tables and metadata were transferred to the R (v4.1) environment for subsequent statistical analyses. Eukaryote, Archaea, Mitochondria, Chloroplast and blank-related sequences were discarded, resulting in 98,097,921 reads across 868 sediment samples and a total of 255,303 ASVs. We then applied a prevalence threshold to our dataset so that only ASVs present in at least two out of the six sediment patches (i.e., ecological replicates) were kept across our samples, independently of their presence or absence in other GFSs. GFSs exhibiting less than 3 ecological replicates were discarded. Read counts were subsequently averaged across replicates for a given GFS and multiplied by six prior to rarefaction. This led to a total of 71,387 ASVs across 152 samples.

Supplementary Method 2. Details on the comparative analysis of the accuracy of denoising approaches tested within this study.

To assess the accuracy of the denoising approaches, we analyzed different types of blanks and mock communities. We included six blanks in the sequencing run to evaluate both the background noise inherent to the sequencing process and the potential contamination that could arise from the reagents (i.e., used in DNA extraction or library preparation). Consequently, the 43 ASVs present in the blanks were filtered from the final ASV table. We also sequenced three types of mock communities to assess the inherent noise generated by the different denoising methods. Two mock communities were purchased from Zymo, either in the form of DNA (cat 6305) or as cells (cat 6300) and we created a lab internal mock community composed of 7 strains. The latter were prepared from pure cultures obtained from a Swiss GFS and were cultured at 10°C in R2A liquid media. For taxonomic identification of these mock-strains, DNA was extracted, the 16S rRNA gene amplified and Sanger-sequenced for identification (using the 27F and 1492R primer pair for amplification and Sanger sequencing in both directions). To obtain mock communities, strains were grown in R2A medium and optical density (OD600nm) was monitored. After reaching a plateau, the cultures were diluted to the same OD600 and mixed to obtain the mock community. DNA was then extracted from this mock and the cells mock community from Zymo using the same protocol as for the GFS. Our results showed that Deblur[34] accurately identified the known microorganisms and effectively assessed the expected composition of the mock communities, with very few other ASVs present (8-10)(*Extended Data Figure 11*). In addition, Deblur demonstrated very few contaminations compared to the other denoising approaches examined within this study (See *Supplementary Table 12*). Indeed, all ASVs belonged to the mock strains except one associated to the *Paenibacillus* genus and one unknown Gallionellaceae, although both contaminants had only very few reads of less than 1% compared to the lowest reads for an expected community member. Further all genera had only one AVS with the exception of *Salmonella* that had 2 in some cases (*Supplementary Table 12*). To further evaluate the identification of key players in the results, we analyzed the mock community derived from isolates obtained from Swiss GFS (*Extended Data Fig.11d*). Our results indicate that all isolates were detected, and Deblur was confirmed to be the best approach generating fewer ASVs (17 and 15 instead of a theoretical 7). All ASVs belonged to the expected genera using Deblur but for 4 and 5 strains, we detected two ASVs instead of one. Furthermore, all three methods demonstrated similar levels of diversity based on the Shannon index (*Extended Data Figure 8*), while UNOISE3[35] significantly overestimated per-sample diversity when comparing the number of observed ASVs. After evaluating the three methods, we opted to proceed with Deblur, as it best resolved the taxonomy of the mock communities and has been demonstrated to be one of the most robust denoising approaches[102] - a good compromise between UNOISE3 and DADA2.

Supplementary Method 3. Processing of cryospheric samples.

Raw sequence data of other cryospheric ecosystems, retrieved from the meta-analysis of Bourquin et al.,[21], were re-processed using Deblur to allow comparability with our dataset. The filtered GFS and cryobiome datasets were subsequently merged and rarefied to 10,185 reads per sample (due to the lower number of reads in other cryospheric samples) and included 268 samples and 62,188 ASVs. We then computed ratios of alpha diversity (ASV richness) and Shannon H between GFS samples and samples from other cryospheric ecosystems (*Supplementary Table 4*).

Bibliography

100. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

101. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

102. Nearing, J. T., Douglas, G. M., Comeau, A. M. & Langille, M. G. I. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**, e5364 (2018).