Comparing Community Measures in Lake Microbial Ecology: Metagenomes and Metatranscriptomes and Amplicons, oh my!

Julia Kristina Nuy (julia.nuy@uni-due.de)
Universitat Duisburg-Essen Fakultat fur Biologie https://orcid.org/0000-0003-0270-0070
Till Bornemann
Universitat Duisburg-Essen
Daniela Beisser
Universitat Duisburg-Essen Fakultat fur Biologie
Alexander J. Probst
Universitat Duisburg-Essen
Jens Boenigk
Universitat Duisburg-Essen Fakultat fur Biologie

Short report

Keywords: Metagenomes, Metatranscriptomes, 16SrRNA amplicons, freshwater microbiome

DOI: https://doi.org/10.21203/rs.3.rs-29111/v1

Abstract

Lake ecosystems are hotspots on Earth for biogeochemical cycling yet linking their microbiome to physicochemical parameters remains a challenge. Here, we assess the quality of 16S rRNA gene-based metatranscriptomics, assembled metagenomics and 16S rRNA gene amplicon sequencing for 21 lake ecosystems across Europe. We identified method-dependent, massive differences between community composition and proportional activity for key taxa like Alphaproteobacteria suggesting different ecological conclusions for the same lake ecosystems. In redundancy analysis (RDA), environmental parameters explained the greatest amount of the variance in metatranscriptomes suggesting that the active community is heavily influenced by environmental parameters. While amplicon data recruited the least amount of environmental variables in RDA (pH and temperature), four additional parameters explained the sequenced metagenomes. These results suggest that metagenomes and metatranscriptomes are currently the best methods for linking lake microbiomes to physicochemical parameters and can be used as proxies for designing future ecological surveys.

Background

Most studies on microbial communities are reliant on the analysis of next generation sequencing data, which can be generated from metagenomic DNA or RNA. While the widely distributed application of amplicon sequencing is mainly used to target specific hypervariable regions of 16S rRNA genes that are PCR-amplified, metagenomes provide the entire genomic blueprints. For investigating the active community, metatranscriptomics has proven to be useful for analyzing ribosomal and messenger RNA (Mills et al. 2012) (rRNA and mRNA, respectively). Due to the massive generation of genomes from metagenomes, databases have drastically expanded enabling taxonomic profiling using assembled metagenomes and metatranscriptomes, yet a thorough comparison of these two technologies to amplicon sequencing has not been performed. Of particular interest are lake ecosystems that are reflected by high microbial activity, seasonal changes and high biomass production.

In order to address this gap of knowledge, we leveraged samples from 21 different lakes across Europe and performed 16S rRNA gene sequencing, metagenomics and metatranscriptomics (rRNA) encompassing 3.31 Gbps, 170.91 Gbps and 34.11 Gbps of sequencing depth, respectively. In contrast to previous studies (Tessler et al. 2017; Tedersoo et al. 2015; Guo et al. 2016) we assembled the metagenomic data into scaffolds (379 Mbps +/- 248) and determined the community composition via annotated ribosomal protein subunit 3 (rpS3) genes, a taxonomic marker that has been well established previously (Sharon et al. 2015).

Methods

Sampling

Sampling of 21 European freshwater lakes was conducted in August 2012. The samples were taken by daylight from the shore of each lake or pond collecting epilimnial water up to 0.5 m depth (table S1). Samples for genomic DNA extraction and RNA extraction were filtered onto 0.2 µm nucleopore filters. To obtain similar biomass per sampling site, water was filtered until the filters were blocked by biomass. Biomass filters for genomic DNA were immediately preserved below − 80 °C in a cryoshipper (Chart/MVE, Ball Ground, USA) to avoid DNA degradation. Samples for RNA extraction were stabilized in RNA-stabilization solution (LifeGuard Soil Preservation Solution, MoBio Laboratories Inc., Carlsbad, CA) to prevent RNA degradation, and subsequently stored below − 80 °C in a cryoshipper (Chart/MVE, Ball Ground, USA) to ensure continuous cooling until analysis.

DNA extraction

Genomic DNA

Genomic DNA was extracted from the biomass filters using the my-Budget DNA Mini Kit (Bio-Budget Technologies GmbH, Krefeld, Germany) following the protocol of the manufacturer with minor adaptations and under sterile conditions. We changed the protocol as follows: The filters were homogenzied in 800 µl Lysis Buffer TLS within lysing Matrix E tubes (MP Biomedicals, Santa Ana, California, USA) using a FastPrep (MP Biomedicals, Santa Ana, California, USA) for three times for 45 seconds at 6 m/s. The samples were subsequently incubated for 15 min at 55° C. The quality of the DNA was checked using a NanoDrop™ ND-2000 UV-Vis spectrophotometer (Thermo Fisher Scientifics, Waltham, Massachusetts, USA).

RNA

The extraction of RNA from the biomass filters was carried out under sterile conditions. The filters were removed from the RNA stabilization solution and, after grinding in liquid nitrogen, incubated in Trizol Solution for 5 min at room temperature (Life Technologies, Paisley, Scotland – protocol modified). The samples were afterwards incubated in Chloroform at room temperature for 15 min. To separate the RNA from the residual cell contents, the samples were centrifuged and the top phase (containing RNA) was transferred to a new tube and isopropanol was added. The RNA was allowed to precipitate for 1 h at -20 °C. The samples were centrifuged, the supernatant was discarded and the pellet was washed three times in 75% ethanol. The RNA pellet was dissolved in DEPC-treated (diethylpyrocarbonate) water and stored below − 80 °C. To check the quality and quantity of RNA, a spectrophotometer (NanoDrop ND-2000 Spectrophotometer; Thermo Fisher Scientific, and the program NanoDrop 2000 / 2000c Operating Software, version 1.6 (September 2014), Thermo Fisher Scientific, Wilmington, DE, U.S.A.) were used. To investigate DNA contamination and RNA degradation, RNA was checked on a 1% agarose gel by us and assessed prior to cDNA library preparation by the sequencing company.

DNA preparation Sequencing and Filtering

PCR for amplicon based analysis

(Published in Nuy et al 2020) PCR amplifications targeted the V2-V3 region of the 16S rRNA gene using the primers 104F (5'- GGC GVA CGG GTG MGT AA-3') and 515R (5'- TTA CCG CGG CKG CTG GCA C-3') (Lange et al. 2015). The selected forward primer contains two wobble positions to catch a broader spectrum of taxonomic groups. Each sample was amplified twice using primers with different sample identifiers following the AmpliconDuo protocol (A and B variant) (Lange et al. 2015). For the PCR reaction 1 µl of DNA template in 25 µl PCR reaction with 0.4 units of Phusion DNA polymerase (Thermo Fisher Scientifics, Waltham, Massachusetts, USA), 0.25 µM primers, 0.4 mM dNTPs and 1 x Phusion buffer (Thermo Fisher Scientifics, Waltham, Massachusetts, USA) was used. The PCR protocol consisted of 35 cycles, including a denaturation step at 98° C for 30 s, annealing step at 72° C for 45 s, and an elongation step at 72° C for 30 s. Finally, the PCR was completed by a final extension step at 72° C for 10 min. Samples were pooled in equimolar ratios and commercially sequenced. The data are available at NCBI under Bioproject ID PRJNA559862.

Amplicon Sequencing Illumina

Sequencing was conducted using paired-end (2 × 300 bp) HiSeq 2500 Illumina sequencing in “rapid-run” mode (Fasteris, Geneva, Switzerland). Adapter removal, quality trimming and demultiplexing of indexed sequences was performed by the sequencing company (Fasteris, Geneva, Switzerland). Thereupon, base quality of raw sequence reads was checked using the FastQC software (v0.11.8,Andrews 2018). The raw sequences were quality filtered to remove reads with an average Phred quality score below 25 using PRINSEQ-lite (v0.20.4;Schmieder und Edwards 2011). Additionally, all reads with at least one base with a Phred quality score below 15 were removed. The paired-end reads were assembled and quality filtered with the tool PANDASeq (v2.10;Masella et al. 2012). The remaining reads were dereplicated and chimeras were removed using UCHIME with default parameters (usearch v7.0.1090;Edgar et al. 2011). Finally, a split-sample filtering protocol for Illumina amplicon sequencing (AmpliconDuo) was used as described in Lange et al. (2015) for the removal of sequence artefacts and sequences were discarded that were not found in both sample branches (A and B variant). Filtered reads were clustered in OTUs via the software SWARM using default settings (v2.2.2; (Mahé et al. 2014). Taxonomic assignment was performed using the database SILVA (SILVA SSURef release 132). *Betaproteobacteriales* were reassigned to the class *Betaproteobacteria*.

Metagenomic Shotgun-Sequencing

Metagenomic samples were sequenced at BGI on an Illumina Hiseq XTen sequencer producing 150 bp paired-end samples. At BGI, the genomic DNA was quality tested and qualified samples were used for sequencing library preparation. The DNA was fragmented by shearing to the desired size (350 bp) by Covaris S/E210 or Bioruptor. The resulting overhangs were transformed into blunt ends by using T4 DNA polymerase, Klenow Fragment and T4 Polynucleotide Kinase. Sequencing adapters were ligated onto both ends of the DNA fragments. Target fragments were purified via gel-electrophoresis. Index tags were introduced into the adapter sequence to allow pooling. Finally, the libraries were quality tested and sequenced. Clean and demultiplexed samples were provided. Sequences are available at NCBI as Bioproject PRJNA578039 (shallow) and Bioproject PRJNA602302 (deep)

Raw sequence data were trimmed based on quality scores using the read trimmer, bbduk, high quality sequences were filtered using Sickle (https://github.com/najoshi/sickle; default parameters). The filtered and trimmed reads were assembled to scaffolds using metaSPADES (v3.12, Nurk et al. 2017) using default settings. Sequencing coverage was determined for each assembled scaffold by mapping reads from the sample to the assembly using Bowtie2(Langmead und Salzberg 2012)(–sensitive). Gene prediction was performed using prodigal Hyatt et al. 2010.Gene annotation was conducted with diamond blast(Buchfink et al. 2015) against the UniRef100 database(Suzek et al. 2015) (evalue cutoff 10^{-5}). To evaluate community composition, the assemblies were filtered for ribosomal protein S3 genes (rpS3). RpS3 genes were clustered at 99% sequence similarity using usearch (v11, Edgar 2010). The clustering resulted in centroid sequences that were taxonomically compared with usearch to a database consisting of rpS3 sequences compiled from previous studies(Anantharaman et al. 2016; Hug et al. 2016; Probst et al. 2018). Reads were mapped on the rpS3 genes to calculate coverage per gene per sample. All read mappings were performed using Bowtie 2Langmead und Salzberg 2012 with parameter –sensitive. All samples were normalized by the number of reads of each library.

Metatranscriptomic Shotgun sequencing - RNA

Preparation of cDNA libraries as well as sequencing was performed by the sequencing service Eurofins (Eurofins MWG GmbH, Ebersberg, Germany). Two amplified short-insert (150–400 bp) cDNA libraries (poly-A enriched mRNA and total RNA) were prepared per sample, individually indexed for Illumina HiSeq-2000 sequencing with the 2 × 100 bp paired-end module (v3 chemistry), and finally demultiplexed. All samples (but S031BU, mean Q: 28.24) obtained a mean base pair quality above 35. Raw sequence data were published in the NCBI sequence read archive under the BioProject ID PRJNA345457(Grossmann et al. 2016). Ribosomal RNA (rRNA) was used for further analyses.

We used the quality control tool FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to analyze the quality of the raw reads. Removal of adapter sequences and trimming of bad quality bases with a quality score of below 20 was carried out using cutadapt software v1.3(Martin 2011). Sequences with a read length of below 30 base pairs were simultaneously discarded. The amount of rRNA was determined by mapping the reads to the SILVA 16S/18S SSU rRNA gene database (Yilmaz et al. 2014; Smith und Milne 1981; Quast et al. 2013). The index was built from the downloaded SILVA database release 119.

To determine the taxonomic composition from rRNA reads, the reads were grouped according to their phylum and for Proteobacteria to their class, following the list of prokaryotic names and nomenclature Parte 2018. The taxonomically grouped reads were then counted.

Assessment of abiotic factors

The assessment of abiotic factors can be obtained from Grossmann et al. (2016) except for area and time. Time of sampling was added to abiotic factors (Table S3). Surface area calculation of each water body was conducted on the basis of satellite images from Google Mymaps (Map data 2017 GeoBasis-

DE/BKG (2009), Google). Surface area was estimated using Image J (v1.8.172). Coordinates of the sampling sites were calculated with a GPS device.

Statistical tests

Prior to statistical tests each dataset was normalized to percent relative abundances.

Barplots and PCA were calculated on the basis of percent relative abundances using the vegan package (Oksanen et al. 2015) and the ggplot2 package in R (R Core Team 2019). Additionally, we analyzed the relationship of read abundances per phylum in pairwise Pearson correlations of all datasets. P-values were adjusted with the Benjamin Hochberg correction.

To investigate dissimilarities in communities and between community data and environmental data, we calculated Bray Curtis dissimilarities of community matrices and Euclidean distances of the environmental matrix. To test the relationship of environmental and community data, we used the Mantel test evaluating the entire environmental matrix and the BioEnv correlating the best Model of significant environmental samples with Spearman rank correlation. The explanatory power of the environmental dataset on the community datasets was revealed with redundancy analyses RDA on the basis of Euclidian distances. Models for significant environmental factors were calculated with ordistep (Oksanen et al. 2015).

The primer bias per phylum was inferred by calculating the weighted Primer score using Primerprospector (Starke und Morais 2019). Gene copy numbers were obtained from the rrnDB (Stoddard et al. 2015). The mean of all entries per phylum was used to correct for the gene copy number.

Results And Discussion

Our comparisons revealed 16 phyla shared among all technologies, and another 13 phyla only found with one or two of the technologies that all belonged to the rare biosphere, meaning taxa with low read abundances (Fig. 1A, B, C, Table S1). An earlier study showed that less than 50% of phlya were detected with metagenomes in comparison to amplicons (Tessler et al. 2017), while other studies demonstrated the exact opposite (Poretsky et al. 2014; Guo et al. 2016; Tessler et al. 2017). However, comparisons of short reads to public databases is problematic; e.g., 150 bps-reads only cover approximately 15% of the average prokaryotic gene (Xu et al. 2006) leading to inherent biases in taxonomic calling. Interestingly, the percent relative abundance of the individual phyla varied greatly between the technologies, suggesting biases either arising from PCR-amplification or from measuring only active community members in metatranscriptomes (Fig. 1A, B, C). Ordination analyses coupled to Mantel-tests based on Bray-Curtis dissimilarities and Spearman rank correlations of taxon composition clearly showed a greater similarity between metatranscriptomes and metagenomes (r_S=0.4122, p = 0.001) and metatranscriptomes and amplicons (r_S=0.1928, p = 0.018), than between metagenomes and amplicons (r_S=0.04281, p = 0.309) in taxon composition (Fig. 1D). Main differences between metagenomes and amplicon sequencing data were attributable to different abundances of phyla that recruited low to average numbers of reads as reported previously (Poretsky et al. 2014). Proportional overrepresentation or underrepresentation of phyla in amplicons in comparison to metagenomes were analyzed using the pairwise t-test with paired samples. The analysis revealed a proportional greater detection of Bacteroidetes, Alphaproteobacteria and Cyanobacteria and a lower detection of Betaproteobacteria, Planctomycetes and Actinobacteria in amplicons (Supplemental Material 1). While relative abundances of assembled metagenomes have been shown to significantly correlate with quantitative digital droplet PCR measurements (Probst et al. 2018), amplicon data suffer from primer bias (Probst et al. 2015; Starke und Morais 2019), amplification biases (Acinas et al. 2005), chimeras (Haas et al. 2011) and variable numbers of 16S rRNA gene copies per genome (Farrelly et al. 1995). In direct abundance correlations only low-abundant Gemmatimonadetes and fairly high abundant Betaproteobacteria showed concurrences between the two technologies, i.e., amplicon and metagenomes (Figures S1, S2). To overcome potential biases in amplicon data *in silico*, we corrected the amplicon dataset for 16S rRNA gene copy number and the specific primer bias by using the weighted primer score. As evidenced in Figure S3 and Table S2, a correction of relative abundances by gene copy number and weighted primer score (Figures S4, S5) did not result in an observable (significant) shift of community composition. Thus, we conclude that either amplification biases or suitability of the chosen hypervariable region for taxonomic calling (Yang et al. 2016) are the major components distorting the community profile derived from amplicon data.

The active community reflected by metatranscriptomes was significantly similar to communities assessed with metagenomes (Fig. 1), although the metatranscriptomic data was only based on transcribed 16S rRNA genes. Relating the relative abundance of a given taxon detected with metatranscriptomes to DNA-based methods (amplicon data or metagenome data) can result in meaningful ecological statements regarding its activity at the time of sampling. We determined great differences between the proportion of the active population in the sampled ecosystems when using metagenomic or amplicon data. For the twelve most abundant phyla only Actinobacteria, Cyanobacteria and Gemmatimonadetes had similar proportions for metatranscriptomes with metagenomes, and metatranscriptomes with amplicons, while the percent of the population that was active for nine other phyla (including Verrucomicrobia) was inconclusive (Figures S1, S2). For instance, metatranscriptomics coupled to metagenomics of Alphaproteobacteria, which are important for nitrogen cycling in lakes (Newton et al. 2011), suggests that the majority of the population was inactive. In contrast, using amplicon data as the baseline, a high percentage of the population was transcribing 16S rRNAs. These results suggest that the selection of the DNA profiling method heavily influences the inferences of the active members of a microbiome.

To receive a broader picture of the impact of the sequencing strategies on ecological conclusions, we compared the three datasets performing diverse methods from community ecology, i.e. RDA, Mantel tests and BioEnv (Fig. 2, Suppl Mat. 2). In line with Crump et al. (2007) and Souffreau et al. (2015), we assume that physico-chemical factors have a great effect on community composition. Therefore, we expect that a high proportion of variance is explained by physico-chemical factors. In general, the Mantel tests and BioEnv analysis (Fig. 2, Supplemental Material 2) revealed a strong link of the physico-chemical matrix with the metagenomic and metatranscriptomic dataset, while the amplicon dataset showed a negative correlation in the Mantel test and positive correlation in the BioEnv analysis. The RDA conducted with amplicon community data as response matrix and environmental data as explanatory variables revealed data that only 15% of variance could be explained (Figs. 2, S6). For metagenomic and metatranscriptomic community 43% and 51% of the total variance were explained by physicochemical data, respectively. The significant factors for the metatranscriptomic dataset were total phosphorus (TP), dissolved phosphorus (DP), elevation and potassium (K), while the metagenomic dataset was additionally to these factors significantly correlated with temperature and pH. Significant factors in the amplicon dataset were temperature and pH only.

Conclusion

Our study involving amplicon data, metagenomes, and metatranscriptomes from 21 different lakes across Europe were coupled to physicochemical measurements in order to understand which method is best used to assess microbiome composition, proportional activity and ecology. Our data suggests that the marker gene abundance based on assembled metagenomes is best explained by six out of twelve environmental variables, which exceed numbers for amplicon and transcriptome data. The latter showed the highest amount of explained variance in RDA suggesting that a combination of metagenomics and metatranscriptomics is best for assessing community composition and activity in aquatic freshwater ecosystems.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The datasets generated and analysed during the current study are available in the NCBI repository under the BioProject ID PRJNA345457 for metatranscriptomic data, as Bioproject PRJNA578039 and PRJNA602302 for metagenomic and Bioproject ID PRJNA559862 for amplicon data.

Competing interests

The authors declare that they have no competing interests.

Funding

We thank the German Research Foundation (DFG) for the funding of the projects BO 3245/19 and BO3245/17.

Authors' contributions

JB contributed to the sampling of the lakes. **JKN** designed the study and performed all analysis performed in this study. **JKN** and **DB** processed the amplicon and the metatranscriptomic data. **TB** and **AJP** processed the metagenomic data. **JKN** and **AJP** wrote the manuscript. All authors approved final versions of the manuscript.

Acknowledgments

We thank the team for publication services from the University of Duisburg - Essen for providing funds for the open access publication fee.

References

1. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol. 2005;71(12):8966–9. DOI:10.1128/AEM.71.12.8966-8969.2005.
2. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature communications. 2016;7:13219. DOI:10.1038/ncomms13219.
3. Andrews S. (2018): FastQC: a quality control tool for high throughput sequence data. Online verfügbar unter http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
4. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015;12(1):59–60. DOI:10.1038/nmeth.3176.
5. Crump RC, Adams HE, Hobbie JE, Kling GW. Biogeography of bacterioplankton in lakes and streams of an Arctic tundra catchment. Ecology. 2007;88(6):1365–78. DOI:10.1890/06-0387.
6. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1. DOI:10.1093/bioinformatics/btq461.
7. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27(16):2194–200. DOI:10.1093/bioinformatics/btr381.
8. Farrelly V, Rainey FA, Stackebrandt E. Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. Appl Environ Microbiol. 1995;61(7):2798–801.
9. Grossmann L, Beisser D, Bock C, Chatzinotas A, Jensen M, Preisfeld A, et al. Trade-off between taxon diversity and functional diversity in European lake ecosystems. Molecular ecology. 2016;25(23):5876–88. DOI:10.1111/mec.13878.

10. Guo J, Cole JR, Zhang Q, Brown C, Titus; Tiedje JM. Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. Appl Environ Microbiol. 2016;82(1):157–66. DOI:10.1128/AEM.02772-15.
11. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome research. 2011;21(3):494–504. DOI:10.1101/gr.112730.110.
12. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. (2016): A new view of the tree of life. Nature microbiology 1, 16048. DOI:10.1038/nmicrobiol.2016.48.
13. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 2010;11:119. DOI:10.1186/1471-2105-11-119.
14. Lange A, Jost S, Heider D, Bock C, Budeus B, Schilling E, et al. AmpliconDuo: A Split-Sample Filtering Protocol for High-Throughput Amplicon Sequencing of Microbial Communities. PloS one. 2015;10(11):e0141590. DOI:10.1371/journal.pone.0141590.
15. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357–9. DOI:10.1038/nmeth.1923.
16. Mahé F, Rognes T, Quince C, Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ. 2014;2:e593. DOI:10.7717/peerj.593.
17. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011;17(1):10. DOI:10.14806/ej.17.1.200.
18. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinform. 2012;13:31. DOI:10.1186/1471-2105-13-31.
19. Mills HJ, Reese BK, Shepard AK, Riedinger N, Dowd SE, Morono Y, Inagaki F. (2012): Characterization of Metabolically Active Bacterial Populations in Subseafloor Nankai Trough Sediments above, within, and below the Sulfate-Methane Transition Zone. In: *Frontiers in microbiology* 3, S. 113. DOI: 10.3389/fmicb.2012.00113.
20. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A guide to the natural history of freshwater lake bacteria. Microbiology molecular biology reviews: MMBR. 2011;75(1):14–49. DOI:10.1128/MMBR.00028-10.
21. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017;27(5):824–34. DOI:10.1101/gr.213959.116.
22. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’Hara RB, et al. (2015): Vegan: community ecology package. Ordination methods, diversity analysis and other functions for community and vegetation ecologists. In: *R-package version 2.3-1.* https://CRAN.R-project.org/package=vegan.
23. Parte AC. LPSN - List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. Int J Syst Evol MicroBiol. 2018;68(6):1825–9. DOI:10.1099/ijsem.0.002786.
24. Poretsky R, Rodriguez -R, Luis M, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PloS

one. 2014;9(4):e93827. DOI:10.1371/journal.pone.0093827.

25. Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CMK, Emerson JB, et al. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nature microbiology. 2018;3(3):328–36. DOI:10.1038/s41564-017-0098-y.
26. Probst AJ, Weinmaier T, DeSantis TZ, Santo Domingo JW, Ashbolt N. New perspectives on microbial community distortion after whole-genome amplification. PloS one. 2015;10(5):e0124158. DOI:10.1371/journal.pone.0124158.
27. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2013;41 (Database issue):D590-6. DOI:10.1093/nar/gks1219.
28. R Core Team. (2019): R: A language and environment for statistical computing. Hg. v. Foundation for Statistical Computing. Vienna, Austria. Online verfügbar unter https://www.R-project.org/.
29. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4. DOI:10.1093/bioinformatics/btr026.
30. Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, et al. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome research. 2015;25(4):534–43. DOI:10.1101/gr.183012.114.
31. Smith JDavid, Milne, Peter J. Spectrophotometric determination of silicate in natural waters by formation of α-molybdosilicic acid and reduction with a tin(IV)-ascorbic acid-oxalic acid mixture. Anal Chim Acta. 1981;123:263–70. DOI:10.1016/S0003-2670(01)83179-2.
32. Souffreau C, van der Gucht K, van Gremberghe I, Kosten S, Lacerot G, Lobão L Meirelles et al. Environmental rather than spatial factors structure bacterioplankton communities in shallow lakes along a 6000 km latitudinal gradient in South America. Environ Microbiol. 2015;17(7):2336–51. DOI:10.1111/1462-2920.12692.
33. Starke R, Morais D. (2019): Gene copy normalization of the 16S rRNA gene cannot outweigh the methodological biases of sequencing.
34. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. (2015): rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development (Database issue).
35. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926–32. DOI:10.1093/bioinformatics/btu739.
36. Tedersoo L, Anslan S, Bahram M, Põlme S, Riit T, Liiv I, et al. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MC. 2015;10:1–43. DOI:10.3897/mycokeys.10.4852.
37. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho, Luiz Felipe M, et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing.

Scientific reports. 2017;7(1):6589. DOI:10.1038/s41598-017-06665-3.

38. Xu L, Chen H, Hu X, Zhang R, Zhang Z, Luo ZW. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Molecular biology evolution. 2006;23(6):1107–8. DOI:10.1093/molbev/msk019.
39. Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinform. 2016;17:135. DOI:10.1186/s12859-016-0992-y.
40. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic acids research. 2014;42(Database issue):D643-8. DOI:10.1093/nar/gkt1209.

Figures

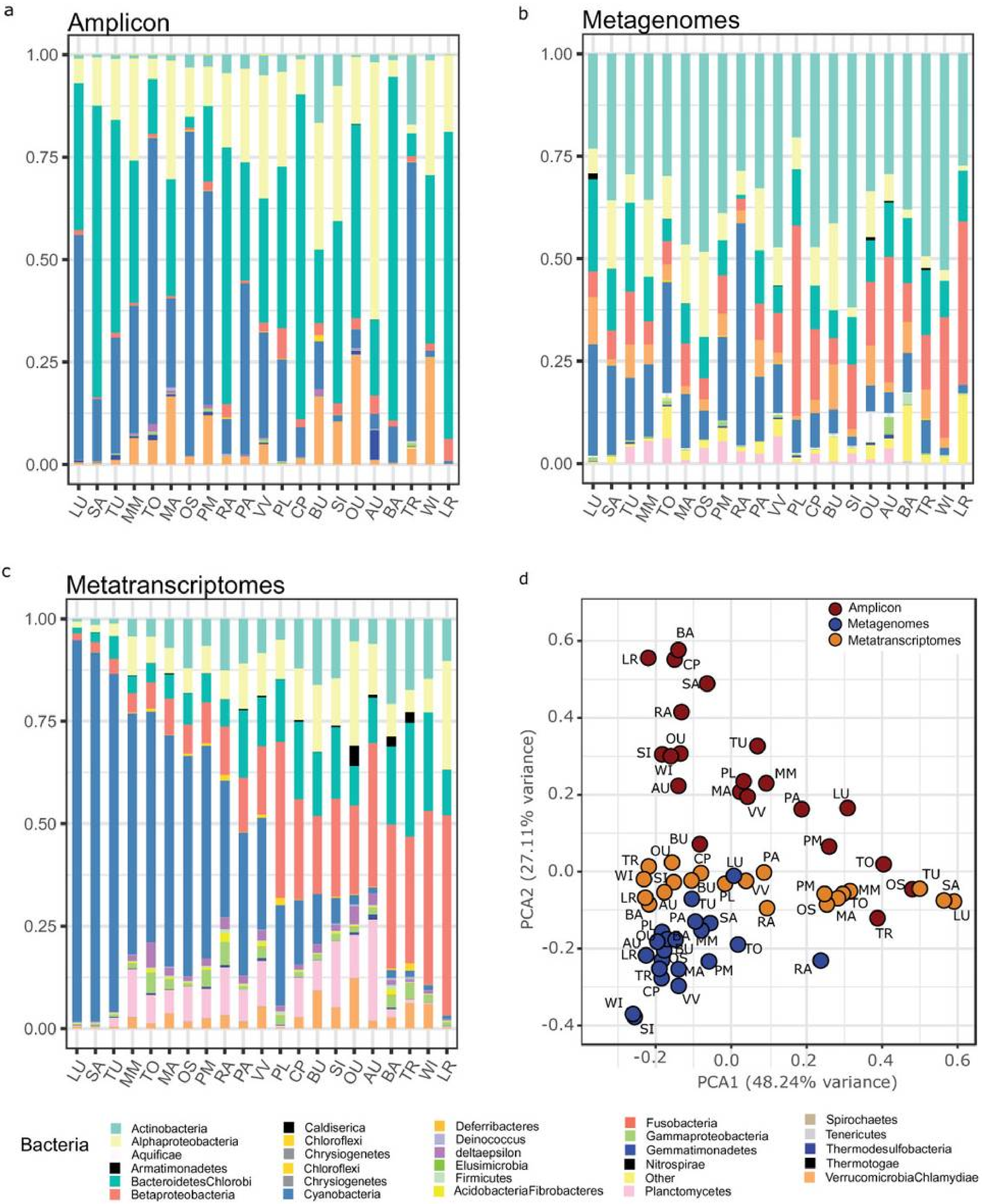
a
Amplicon
b
Metagenomes
c
Metatranscriptomes
1.00
0.75
0.50
0.25
0.00
LU SA TU MM TO MA OS PM RA PA VV PL CP BU SI OU AU BA TR WI LR
d
Amplicon
Metagenomes
Metatranscriptomes
PCA2 (27.11% variance)
PCA1 (48.24% variance)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
0.4
0.6
Bacteria
Actinobacteria
Alphaproteobacteria
Aquificae
Armatimonadetes
BacteroidetesChlorobi
Betaproteobacteria
Caldiserica
Chloroflexi
Chrysiogenetes
Chloroflexi
Chrysiogenetes
Cyanobacteria
Deferribacteres
Deinococcus
deltaepsilon
Elusimicrobia
Firmicutes
AcidobacteriaFibrobacteres
Fusobacteria
Gammaproteobacteria
Gemmatimonadetes
Nitrospirae
Other
Planctomycetes
Spirochaetes
Tenericutes
Thermodesulfobacteria
Thermotogae
VerrucomicrobiaChlamydiae

Figure 1

Dissimilarity of bacterial communities per site. Percent relative abundances of bacterial phyla obtained from a) 16SrRNA amplicon sequences, b) rpS3 gene from assembled metagenomes and c) ribosomal RNA of metatranscriptomes are shown ordered by the percent relative abundance of Cyanobacteria in c). The occurring phyla are similar among different sequencing techniques and genetic material, but vary in composition (shared phyla are shown in Table S1). d) The dissimilarities of bacterial communities per site are visualized in a principal component analysis. Axis 1 and 2 explain 48% and 27%, respectively. Bacterial communities obtained from amplicons, metagenomes and metatranscriptomes are separated along the vertical axis. As also supported by Mantel tests with Spearman rank correlation of Bray-Curtis Similarities, the metagenomic and metatranscriptomic datasets reveal a significant significant relationship. Metagenomic and amplicons are most distantly positioned and thus significantly dissimilar (rs=0.04281, p=0.309), while the communities obtained from metatranscriptomes and amplicon are significantly correlated (rs=0.1928, p=0.018).

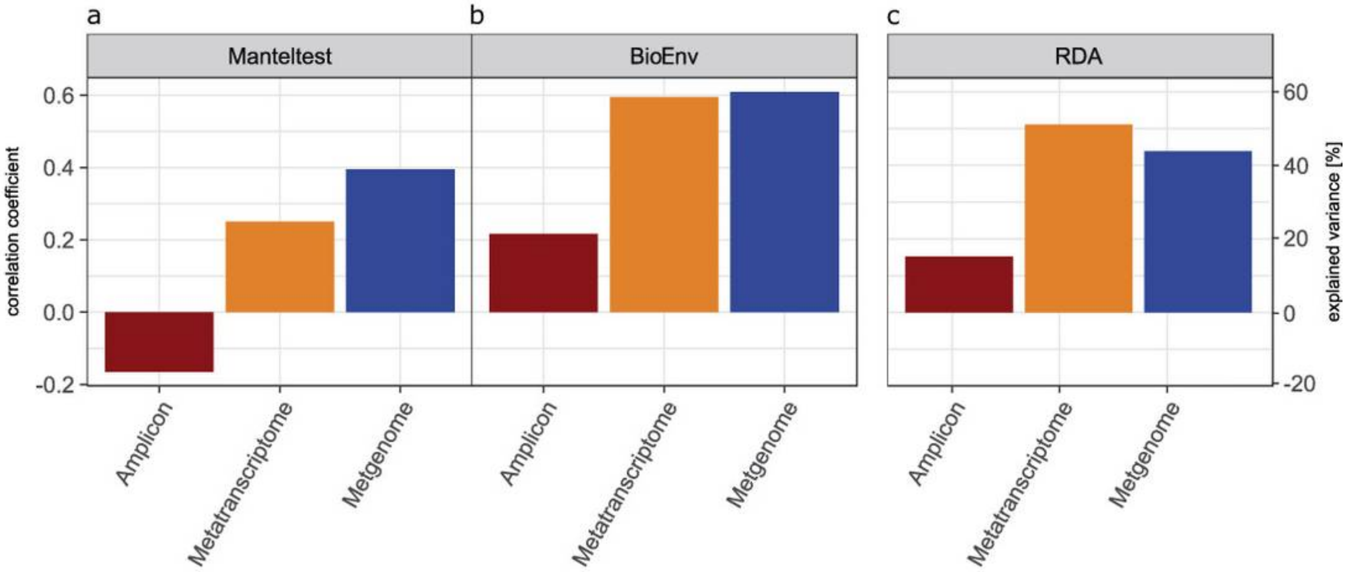

Figure 2

Relation of environmental factors to community data. Bray-Curtis dissimilarities of community data, i.e., amplicon, metatranscriptomes and metagenomes and euclidean distances of environmental data were calculated for a) Mantel tests based on Pearsons correlation b) BioEnv (Suppl.Mat.1) and c) RDA (Figure S6). Overall, amplicons showed the lowest correlations and only ~15% of variation could be explained by environmental metadata. Metagenomes had the highest correlation coefficients with environmental data as exemplified in a) and b). Variance in metatranscriptomic data could be best explained by environmental factors c). Altogether, shotgun sequencing methods revealed regardless of the genetic material stronger links to environmental variables. In line with that, a higher number of significant environmental variables for explaining the community composition was found for shotgun data as compared to amplicons. Further, the variables contributing significantly to explaining the metatranscriptomic and metagenomic community are well known community shaping factors (Newton et al. 2011). A similar result was found by Tessler et al. (2017) and related this result to the lower number of phlya involved in the ordination process potentially leading to clearer divisions by site. As we found the strongest link of metagenomic and metatranscriptomic data to environmental metadata for sites which had the highest number of phlya, we doubt that Tessler's et al. (2017) explanation applies for our study.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS3.xlsx
- FigureS3.pdf
- FigureS5.pdf

- FigureS2.pdf
- FigureS1.pdf
- TableS1.xlsx
- Data2.txt
- FigureS4.pdf
- TableS2.xlsx
- Data1.xlsx
- FigureS6.pdf