BMC Genomics

RESEARCH

Open Access

Dancing the Nanopore limbo – Nanopore metagenomics from small DNA quantities for bacterial genome reconstruction

Sophie A. Simon[1*], Katharina Schmidt[1], Lea Griesdorn[1], André R. Soares[1,2], Till L. V. Bornemann[1,2] and Alexander J. Probst[1,2*]

Abstract

Background While genome-resolved metagenomics has revolutionized our understanding of microbial and genetic diversity in environmental samples, assemblies of short-reads often result in incomplete and/or highly fragmented metagenome-assembled genomes (MAGs), hampering in-depth genomics. Although Nanopore sequencing has increasingly been used in microbial metagenomics as long reads greatly improve the assembly quality of MAGs, the recommended DNA quantity usually exceeds the recoverable amount of DNA of environmental samples. Here, we evaluated lower-than-recommended DNA quantities for Nanopore library preparation by determining sequencing quality, community composition, assembly quality and recovery of MAGs.

Results We generated 27 Nanopore metagenomes using the commercially available ZYMO mock community and varied the amount of input DNA from 1000 ng (the recommended minimum) down to 1 ng in eight steps. The quality of the generated reads remained stable across all input levels. The read mapping accuracy, which reflects how well the reads match a known reference genome, was consistently high across all libraries. The relative abundance of the species in the metagenomes was stable down to input levels of 50 ng. High-quality MAGs (> 95% completeness, ≤ 5% contamination) could be recovered from metagenomes down to 35 ng of input material. When combined with publicly available Illumina reads for the mock community, Nanopore reads from input quantities as low as 1 ng improved the quality of hybrid assemblies.

Conclusion Our results show that the recommended DNA amount for Nanopore library preparation can be substantially reduced without any adverse effects to genome recovery and still bolster hybrid assemblies when combined with short-read data. We posit that the results presented herein will enable studies to improve genome recovery from low-biomass environments, enhancing microbiome understanding.

Keywords Nanopore sequencing, Low DNA input, Low-biomass, Long reads, Hybrid assembly, Prokaryotes, Genomics

*Correspondence:
Sophie A. Simon
sophie.simon@uni-due.de
Alexander J. Probst
alexander.probst@uni-due.de

[1]Environmental Metagenomics, Faculty of Chemistry, Research Center One Health Ruhr of the University Alliance Ruhr, University of Duisburg-Essen, Essen, Germany
[2]Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Essen, Germany

Introduction

Metagenomics has expanded our knowledge of the diversity of Bacteria, Archaea, and Viruses across all environments on Earth [cf., [1–3]]. Recovering MAGs has become a key practice to assess the taxonomic and functional diversity of microorganisms in various environments while circumventing the need to cultivate the respective microbes for genomic studies [4]. Genome-resolved metagenomics provides detailed insights into crucial microbially-driven roles across various ecological processes, such as nutrient cycling [5] or their interactions with other organisms [6]. The completeness and contamination of a reconstructed MAG are directly correlated with the accuracy of downstream genomic predictions, which can elucidate the metabolic potential [7, 8], evolutionary relationships [9], horizontal gene transfer [10], and other genomic traits [cf., [11–13]]. Using short-read sequencing platforms, the first circularized, closed MAGs (cMAGs) were recovered about ten years ago [14, 15]. In September 2019, the number of publicly available cMAGs was 59 [16] compared to the thousands deposited MAGs in public databases [cf., [17, 18]]. Using a hybrid approach of short Illumina reads and long Nanopore reads, Singleton et al. have recently been able to recover almost the same amount of circularized, complete MAGs (57) within one study and from the same ecosystem [19].

Oxford Nanopore Technologies (ONT) has emerged as powerful platform for metagenomics as long reads produced by this sequencing technology span large areas of genomes, covering otherwise problematic genomic regions. Examples for such regions are highly repetitive and/or conserved elements including multiple copies of the same transposable element in one genome [20], for which short read-based *de novo* assemblies are more likely to fail. Long-read metagenomics has successfully been used to for cMAG recovery from diverse samples, including activated sludge [21, 22], water bodies [23], sediment [24] and feces [25, 26]. However, all aforementioned ecosystems were of high biomass, providing sufficient DNA amounts for Nanopore sequencing.

ONT recommends using 1000 ng high-molecular weight (HMW) DNA as input for preparing sequencing libraries using the ligation sequencing kits SQK-LSK109 and SQK-LSK110 and loading between 5 and 50 fmol DNA library on a R9.4.1 Flow Cell. Since these DNA quantities and molarities are often difficult to extract from environmental samples, we investigated how Nanopore sequencing quality, stability of coverage distribution, assembly, and binning quality varies with reduced DNA input quantities. To this end, we sequenced a microbial standard of high molecular weight DNA reducing the amount of input DNA for library preparation from 1000 down to 1 ng in eight steps. Evaluating the results at multiple levels of a genome-resolved metagenomics pipeline (from reads to assemblies and MAGs) we demonstrate that the required amount for successful assembly and genome reconstruction can be significantly reduced. Based on these results, we recommend including Nanopore long reads in every assembly-based metagenomic study including those from low-biomass environments.

Results

To determine the lower limit of DNA input for Nanopore sequencing, we generated 27 long read metagenomes (nine input levels, three replicates each), whereas the ZymoBIOMICS HMW DNA standard served as input material. This mock community was composed of seven bacterial strains and one yeast strain (https://files.zymoresearch.com/protocols/_d6322_zymobiomics_hmw_dna_standard.pdf [02.02.23]); the range of input varied from 1000 ng (recommended) down to 1 ng. Each input quantity was analyzed in triplicates and is herein reported with mass and replication number, e.g., 350_3 for 350 ng input material and replicate #3. The amount of DNA library loaded onto the Flow Cell was reduced from the recommended amount of 4.7 to 9.88 fmol prepared from 1000 ng down to 0.136 fmol prepared from 10 ng DNA input; five sequencing libraries were below the detection limit of the Qubit DNA HS assay but still successfully sequenced (Table S1). We aimed for 1 Gbp of raw sequencing depth per metagenome; down to 50 ng input material, 1 Gb was achieved for all but three metagenomes (350_2, 200_3, 100_3). For the nine metagenomes with less DNA input, only one sequencing run with 1 ng input material (1_1) met the target output. The target sequencing depth of 1 Gb was achieved for metagenomes with 1000−350 ng DNA input in less than 6 h, both on washed and new Flow Cells. For sequencing runs with a lower DNA input that met the target sequencing depth, a run time of up to ~22.5 h was necessary. Figure S1 and Table S2 give an overview about the sequencing depth in dependency to sequencing run time and prior Flow Cell usage. We assessed the results of the different input levels based on read quality, stability of the coverage distribution across the individual microbial genomes, the assembly quality, and the quality of *de novo* reconstructed bacterial genomes. In addition, we leveraged publicly available Illumina data of the mock community to determine if long-reads from low input material can bolster hybrid assemblies of microbial communities [27].

Sequencing quality remains stable irrespective of DNA input quantity

The Q-Score of the generated reads varied around 12.61 (±1.11) and remained stable across the entire dilution series (Fig. 1A). No difference in read quality was evident

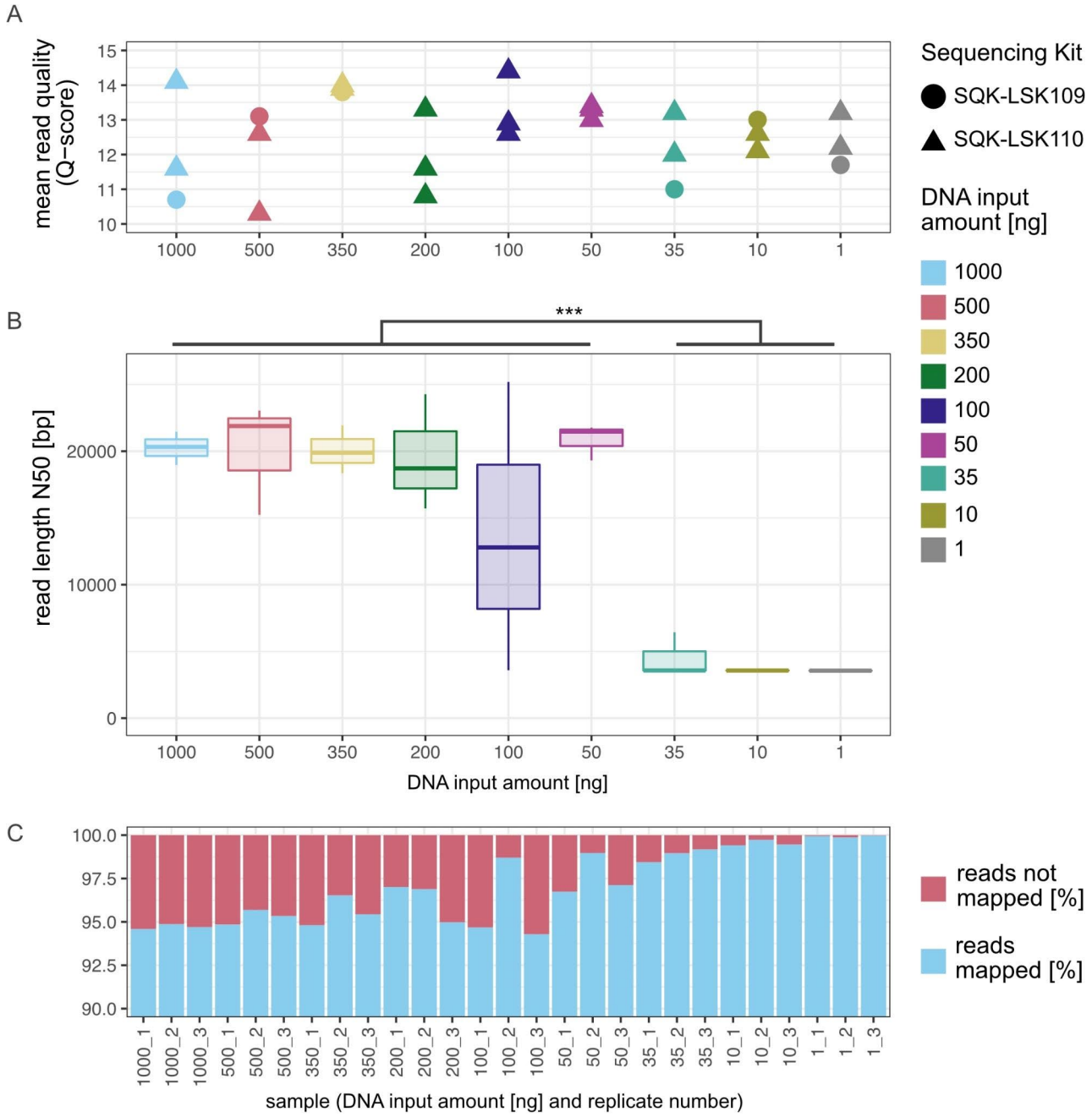

Fig. 1 Sequencing summary. (**A**) Mean read quality (Q-Score) of all Nanopore metagenomes differentiated by the used sequencing kit. (**B**) Boxplot showing how the read length, represented by the N50, changes over the course of the input reduction. The read length N50 decreases significantly when sequencing less than 50 ng. To test for significance the values for the N50 were grouped into ≥ 50 ng and ≤ 35 ng. Both groups are not normally distributed (Shapiro Wilk test; *p*-value ≥ 50 ng: < 0.001; *p*-value ≤ 35 ng: < 0.01), with a subsequent Wilcoxon Rank test the significance was verified (*p*-value < 0.001). (**C**) Proportions of sequencing reads that map to reference genomes changes. The mapping rate varied between 94.29 and 99.98%

between the two sequencing kits used herein (SQK-LSK109/110), rendering all metagenomes comparable with each other. The length of nanopore reads, represented using read length N50, had a mean of 19,106 bp (±4991 bp) in the range of 1,000 ng down to 50 ng (Fig. 1B). For lower input levels, we observed a significant decrease (U-test, *p*-value < 0.001) of the read length N50 down to an average of 3883 bp (±957 bp) (Fig. 1B) compared to the 1000 ng – 50 ng range. The read mapping accuracy by read-to-reference alignments varied between 94.68% and 99.98% for all samples, demonstrating fairly consistent sequencing quality irrespective of input DNA quantity with the exception of an improved mapping rate of 1 ng samples compared to 1000 ng (Kruskal-Wallis

p-value 0.0614 across all samples, post-hoc Dunn's test *p*-value 0.0298 for this comparison only; Fig. 1C).

Community composition remains stable down to 50 ng DNA input

To compare the coverage or percent relative abundance of the eight species in the metagenomes with their theoretical abundance in the community standard, we mapped all reads to the reference genomes using minimap2 [28]. We were able to detect all eight species included in the community standard in each of the 27 metagenomes (Fig. 2A). Except for one outlier (100_3), all samples down to an input of 50 ng showed a significant correlation (p-value < 0.05, Benjamini-Hochberg adjusted) between the determined percent relative abundances of the species and the theoretical abundance of the individual species (Pearson R, p-values and q-value for each sample are given in Table S3). When further reducing the DNA input, the relative abundance of *Salmonella enterica* and *Escherichia coli* increased tremendously to more than 95% in combination. A principal coordinate analysis, PCoA, of the relative abundance distribution between samples and theoretical composition shows clustering of replicates for inputs ranging from 50 to 1000 ng, with the distance between replicates increasing inversely with input amount (Fig. 2B). Metagenomes generated from 1 ng −35 ng input material were strongly separated along the first axis of the PCoA (one 100 ng library represents an outlier, see above), which agrees with the strong distortion of the community below 100 ng input material observed in the bar plot (Fig. 2A). Surprisingly, in the range of 1 to 35 ng of input DNA, distances between replicates decrease significantly, suggesting high reproducibility in this input range. These results indicate that a quantitative statement about relative percent microbial abundances is possible with a DNA input reduced down to 50 ng.

A read of length is a joy forever: Nanopore reads derived from down to 1 ng DNA input improve hybrid assemblies

For Nanopore reads assembled using metaFlye, the average N50 of the assembly was 3.89 Mb (±1.46 Mb) down to an input level of 50 ng. Fragmentation of the assemblies was the greatest for libraries generated from low DNA quantity (1 ng −35 ng), while high-input libraries showed the greatest N50. However, Nanopore reads of replicates 2 and 3 corresponding to 1 ng input could hardly be assembled, e.g., the assembly of replicate 2 of 1 ng only had a total size of 31,902 bp distributed over 4 contigs (Table S4). To determine if low input material could still improve hybrid assemblies, we assembled our Nanopore sequencing data with publicly available Illumina data [27] of the mock community for all 27 metagenomes using hybridSPAdes [29]. The N50 of the hybrid assemblies was 1.06 Mb (±0.259 Mb) for input levels down to 50 ng. At lower amounts, the N50 dropped drastically compared to assemblies of higher input material to an average of 0.375 Mb (±0.246 Mb). The lowest N50 of a hybrid assembly was 136 kb, achieved with Nanopore reads obtained from 1 ng DNA, but still higher than that of a short-read only assembly (120 kb; Fig. 3A).

While the GC content of the hybrid and Illumina-only assembly was very constant (on average 47.41% (±0.04%), the more variable GC content of Nanopore assemblies was 46.87% (±3.11%). Down to 50 ng input, the GC content of the Nanopore assemblies is at 48.38% (±1.45%), and then decreased to 43.84% (±3.40%). Nanopore reads showed less variation in GC content than the assemblies (44.05±1.01%; Figure S1).

Analyses of contig/scaffold lengths support the conclusions of N50 analyses. Using only Nanopore reads down to 50 ng input material, contig lengths that corresponded to the respective expected genome sizes of mock community species assembled. For example, in twelve of the 27 assemblies, the greatest contig was 6.79 Mb in length, which is the genome size of *Pseudomonas aeruginosa*. With hybrid assemblies, scaffolds of around 6.6 Mb could be detected in 17 of 27 assemblies. This approach also enabled the reconstruction of long scaffolds at 35 ng input material, which was not possible for short or long reads alone (3.9 and 6.6 Mbps respectively). Indeed, the longest scaffold of the short-read only assembly has a length of 1.49 Mb, which is substantially shorter than the smallest genome in the mock community (*Staphylococcus aureus*, 2.730 Mbps). The longest scaffolds resulting from hybrid assemblies fed with Nanopore reads from 1 ng and 10 ng input DNA ranged between 1.50 Mb and 2.09 Mb (highest for 1 ng 1.59 Mbps), which is still greater than short-read assembly alone (Fig. 3B).

Recovery of near-complete MAGs down to 35 ng input material

After assembly, the Nanopore-only metagenomes were manually binned using uBin [30]. uBin uses 51 universal bacterial single-copy genes as markers [31]. As depicted in Fig. 4A, recovery of near complete genomes (≥95% completeness, ≤ 5% contamination) was successful down to 50 ng of input material. Remarkably, we could also bin 4 high quality MAGs with more than 95% completeness from replicate 35_3, i.e., with 35 ng input material only. Analysis of Average Nucleotide Identity (ANI) and length consensus—a metric comparing how well the sequences of a MAG and reference sequence align to each other—confirmed the high quality of the recovered MAGs (Fig. 4B). These comparisons showed that even down to 10 ng input, bins with a length consensus > 75% and an ANI of >99.7% could be obtained for Nanopore-only metagenomes (Fig. 4B).

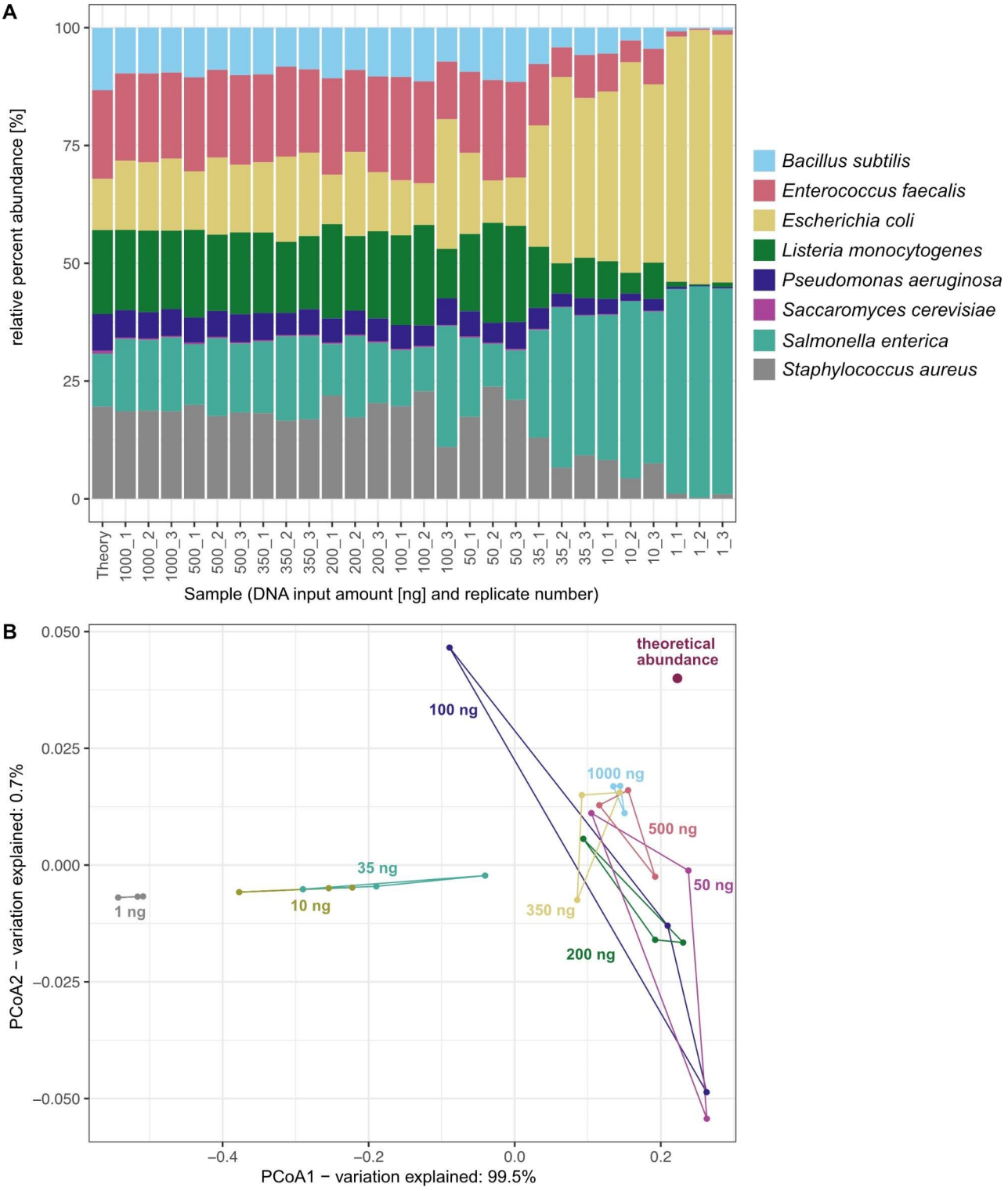

Fig. 2 Abundance analysis. (**A**) Relative abundance and taxonomic assignment based on mapping Nanopore sequences to each of the eight reference genomes. (**B**) PCoA based on Bray-Curtis dissimilarities of percent relative abundance of species calculated from read coverage in comparison to the theoretical composition of the sequenced microbial standard (ZYMO)

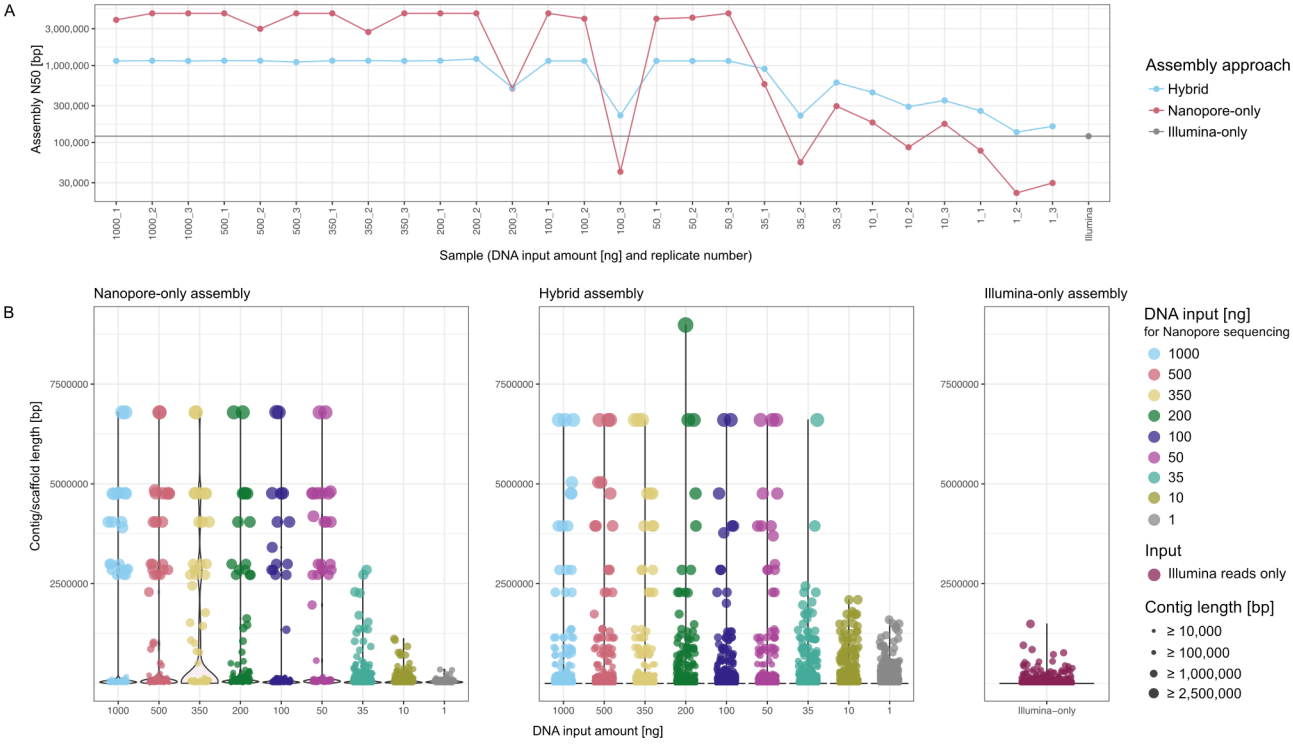

Fig. 3 Comparison of undertaken assembly approaches. (**A**) Representation of the N50 across Nanopore-only, hybrid and Illumina-only assemblies. (**B**) Overview showing the contig/scaffold to length distribution faceted by assembly approach

Discussion

In this study, we report how lower than recommended DNA input affects Nanopore long-read metagenomics by comparing different levels of metagenomic analysis from read QC to retrieval of MAGs. Many previously published Nanopore-based metagenomics studies lack information about the quantity of DNA used for library preparation [19, 24–26, 32]. To the best of our knowledge, other published metagenomic studies quantified the amount of input DNA and met or exceeded the recommended amount of 1000 ng DNA for input material. These studies covered various high-biomass environments including human oral cavity [33, 34], enrichments of artificially contaminated irrigation water [35], groundwater aquifer [36], sand filters of a drinking water treatment plant [37], and activated sludge [21] with up to 4.5 µg of DNA for library preparation. Noteworthy, Chen et al. spiked environmental DNA with DNA of a known bacterium as the authors described their DNA amount as insufficient for Nanopore sequencing [23]. In similar fashion, the so-called CarrierSeq protocol includes the addition of 1000 ng of Lambda phage DNA to the target DNA, which could be metagenomic DNA for sequencing; in such an approach Mojarro et al. used 0.2 ng *Bacillus subtilis* DNA as the target, which resulted in only 777 Nanopore-reads (out of 718,432) that mapped to the target genome of *B. subtilis* [38]. In a study of low-biomass Mars analog soil, CarrierSeq was used for astrobiological investigations also resulting in limited amount of sequence information [39]. With the exception of the aforementioned study, all studies listed above utilized the recommended quantity of DNA from microbiomes that are of high biomass or easily accessibility for repeated sampling. However, there are numerous examples of low-biomass ecosystems, for which DNA extractions do not yield the recommended DNA input for ONT library preparation and excessive sampling is difficult.

While the most prominent ecosystems on Earth (soil, ocean) usually harbor high biomass, smaller ecosystems with low biomass are far more diverse and numerous and yet important for ecosystem services and human health. The air microbiome, for instance, is an ecosystem that constantly and globally surrounds humans, influences the Earth's surface but remains little explored regarding the genetic diversity resolved at species level [40]. The gaseous state, and the low biomass, atmospheric turbulences, temperature differences and day and night shifts complicate sampling [40, 41]. Consequently, we suggest that low-input Nanopore sequencing combined with short reads in a hybrid assembly as presented herein could substantially enhance the study of the air microbiome. In similar fashion, research on glacier and desert microbiomes could be facilitated via hybrid assemblies.

Glacial ice for example, contains minute microbial cell concentrations when compared to many other environments, making low-biomass metagenomics of their

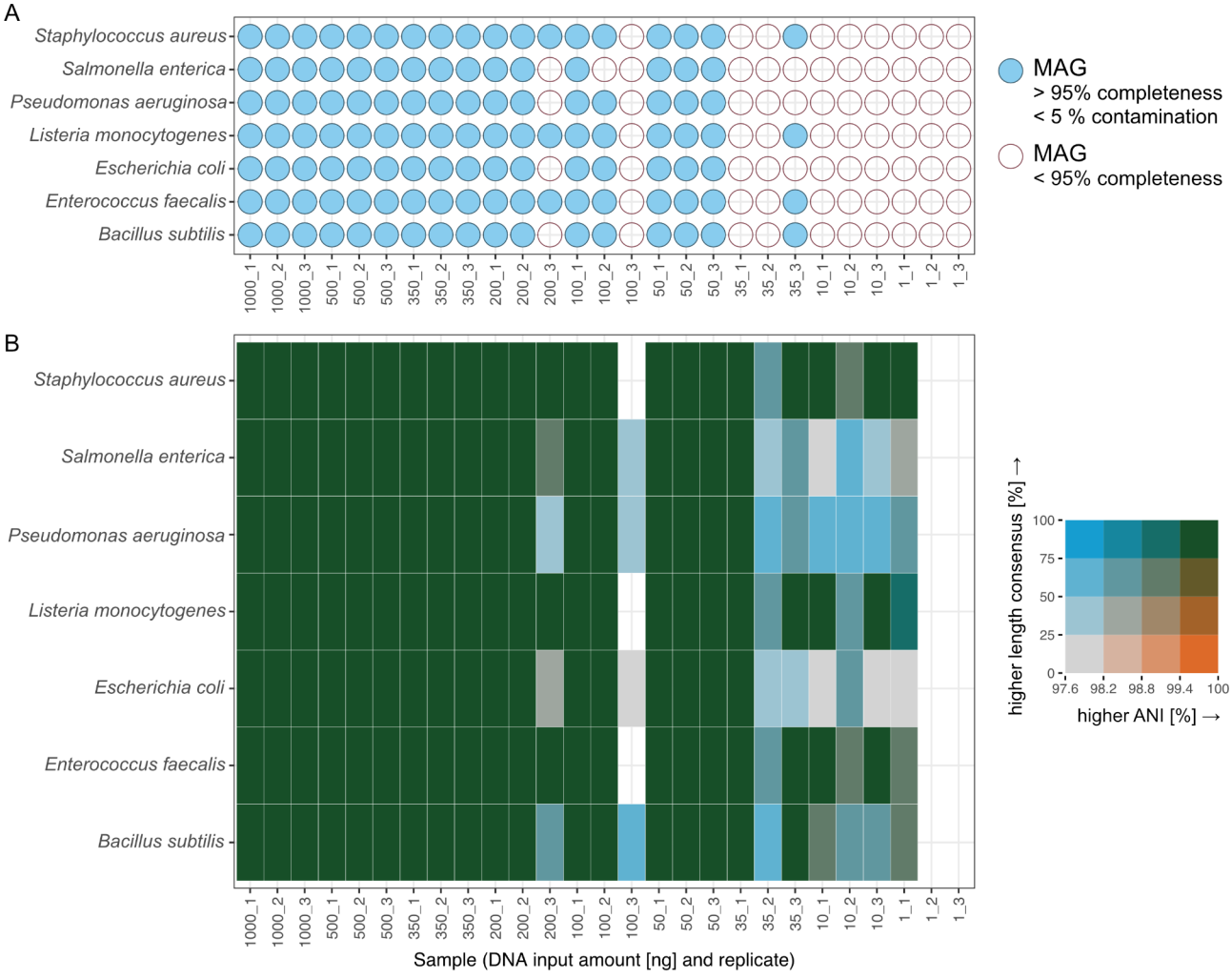

Fig. 4 Quality of metagenome assembled genomes. (**A**) Bubble plot showing which genomes were near-full length recovered based on single copy genes. (**B**) Bivariate heatmap showing the relationship between length consensus and ANI per bin. The increments of the legend were distributed equally on both axes

unique microbial communities necessary [42, 43]. This is also the case for desert soils, where little DNA can be recovered, and resulting genomes from short-read assemblies are generally highly fragmented [44]. Contamination control of spacecraft assembled in cleanrooms is of uttermost importance to protect the integrity of life-detection missions destined to other celestial bodies [45]; however, cleanroom environments are of low-biomass, and application of metagenomics to cleanroom samples usually requires whole genome amplification prior to sequencing [46] distorting community compositions [47] and thus hampering the recovery of genomes from metagenomes. In addition, research on environments for which the matrix strongly hinders nucleic acid extraction could also benefit from approaches developed for low-biomass environments. For instance, chemical processes (e.g., induced by pH) can lead to DNA hydrolysis, denaturation or depurination [48], or nucleic acids are captured by adsorption effects [49]. Furthermore, there are ecosystems where quantity as well as access is restricted (e.g., the surface of the human eye [50]); at the same time, high spatial resolution of high-biomass environments like soil could also be achieved via lowering input material.

Contamination remains a problem in metagenomic studies of low-biomass environments, which can be introduced during sampling or processing the samples and severely skews the result of respective metagenomic studies [51, 52]. Eisenhofer et al. have compiled a detailed checklist of good measures that help to conduct low-biomass studies with as minimal risk of contamination possible [53]. Another level of contamination control can be introduced by computational identification of microbial contaminants [54]. Nevertheless, the numerous examples of ecosystems on Earth stated before could substantially benefit from long-read low-input metagenomics

to bolster our understanding of microbial diversity and evolution.

Although our study has successfully demonstrated that low DNA input Nanopore long-read metagenomics is possible, we acknowledge limitations in terms of extrapolating our findings to complex environmental samples. DNA isolated from environmental samples is most likely fragmented and from lower purity than the mock community DNA used in this study, complicating sequencing, and *de novo* assemblies due to shorter and/or a lower sequencing depth. Additionally, a heterogeneous distribution of microbial cells and variations in the environmental sample matrix can negatively affect DNA extractions and reduce the accuracy of our insights into the microbial diversity of the sampled ecosystem.

The used microbial DNA standards consists of a very limited number of known strains in defined and fairly even distributed proportions (here between 7.8% and 19.6%), offering a simplified representation of a microbial community. By contrast, environmental samples are most likely more complex, including a vast array of unknown microorganisms with varying abundances. Irrespective of the ecosystem studied, different strains of microbial species occur in microbiomes increasing metagenome complexity. A high strain-heterogeneity might hamper *de novo* assemblies [55], which is not addressed in this mock-community based study. Nevertheless, using a microbial DNA standard with known composition and abundances was required to controlled and reproducible reference workflows.

Based on our results, we recommend considering the addition of ONT reads to assembly-based metagenomic studies at each DNA concentration investigated, i.e., from 1000 ng down to 1 ng, and perform hybrid assemblies. Recovery of circularized, closed genomes and the improvement in scaffold length clearly shows that hybrid approaches of short and long reads should be the current standard. To close genomes of, e.g., slow growing prokaryotic isolates, co-cultures or low complex enrichment cultures using Nanopore sequencing 35–50 ng DNA are sufficient as starting material.

Outlook

We here provide a thorough workflow for low biomass long-reads metagenomics, yet another layer of complexity might be introduced with new ONT chemistries that have become available while the study was executed or will become available in the future. For instance, with the new Q20+chemistry, the recommended input amount for a library preparation is still 1000 ng of HMW DNA or >100 ng of fragmented DNA. Recently, the recommended amount of final DNA to be loaded onto the Flow Cells has remained constant or even decreased compared to the previously used combination of LSK109/110 and R9 Flow Cells, which is promising for the future application of Nanopore sequencing for exploring low-biomass ecosystems. Previously, 5–50 fmol DNA library were recommended, but at the moment of writing this paper (February 2023), Oxford Nanopore Technologies recommends only 5–10 or 10–20 fmol for the new chemistries. Another interesting development that was recently announced by ONT for handling small DNA samples is Library Recovery. The DNA library is sequenced on one Flow Cell and as soon as its sequencing performance decreases or ends, the library is aspirated out of the original Flow Cell with a pipette tip and transferred to a new, freshly primed Flow Cell to continue sequencing (https://community.nanoporetech.com/docs/prepare/library_prep_protocols/library-recovery-from-flow-cells/v/lir_9178_v1_revb_11jan2023 [18.01.23]). Exploring these new techniques in combination with low DNA input for library preparation will be a challenging but also promising task with the goal to further explore biodiversity and genetic content of low-biomass environments.

Materials and methods

Microbial community standard

The ZymoBIOMICS HMW DNA Standard #D6322 (Zymo Research, USA) was used to determine the input limits for Nanopore gDNA sequencing without the need for whole-genome amplification. The DNA standard is composed of eight microbial species: seven bacteria and one yeast (Table S5). Used reference genome sequences are provided by Zymo and available at https://s3.amazonaws.com/zymo-files/BioPool/D6322.refseq.zip [07.12.2022].

Library preparation and QC

The input DNA amount of the standard was reduced stepwise from the recommended 1000 ng to 1 ng and included the following quantities of DNA: 1000 ng – 500 ng – 350 ng – 200 ng – 100 ng – 50 ng – 35 ng – 10 ng – 1 ng. DNA amounts in the dilution series were verified via Qubit. Each input quantity was analyzed in triplicates reported with mass and replicated number, e.g., 350_3 for 350 ng input material and replicate #3.

Library preparation was performed using the SQK-LSK109 and SQK-LSK110 sequencing kits (Oxford Nanopore Technologies, UK) with minor deviations from the manufacturer's instructions: Both clean-up steps with AmPure XP beads (Beckman Coulter, USA) were extended by 5 min (10 min in total). To enrich for long fragments, AmPure XP beads were washed using the long fragment buffer (LFB) in the respective step. Additionally, elution of the DNA library from the AmPure XP beads was performed at 37 °C as recommended for HMW DNA. DNA concentration and quality of prepared libraries were determined using Qubit 3.0 fluorometer

(Thermo Fisher Scientific, USA) with the dsDNA HS array and Agilent TapeStation genomic DNA screen tapes. The maximum possible amount (12 µL) of DNA library was always loaded onto the Flow Cells. Prepared libraries were stored at -80 °C if not subsequently sequenced.

Nanopore sequencing

Sequencing was performed using a MinION™ Mk1B (ONT) equipped with FLO-MIN106D Flow Cells. Sequencing runs were supervised by MinKNOW v21.02.1. Sequencing runs were stopped manually after achieving 1 Gb per sample or if even after >12 h the sequencing output had reduced so much that we did not expect to reach 1 Gb.

Basecalling and read QC

Generated Nanopore raw reads were basecalled using Guppy v 6.1.3 in its super-accurate mode enabled using the dna_r9.4.1_450bps_sup.cfg model (https://community.nanoporetech.com/downloads [21.12.22]). Guppy basecalling in super-accurate mode automatically excluded reads with a Q-Score lower than 10. Statistics about the sequencing run and the sequencing reads were acquired using Nanoplot v 1.39.0 using the sequencing summary file generated by Guppy as input [56]. Basecalled Nanopore reads were filtered with Filtlong v 0.2.0 (https://github.com/rrwick/Filtlong [11.08.22]) to remove reads shorter than 2000 bps using –min_length 2000.

Estimation of sequencing coverage

In order to estimate the coverage and the relative abundance of the individual species, respectively, Nanopore reads were mapped to the provided reference genome sequences using minimap2 v. 2.24-r1122 using the option --ax map-ont, --secondary=no and --sam-hit-only, whereas to quantify unmapped reads the --sam-hit-only flag was omitted [28]. Samtools v. 1.10 was used to manipulate resulting SAM files [57].

Long read metagenomic assembly and binning

Long reads were assembled using Flye v. 2.9-b1768 with --meta and --nano-raw options enabled [58]. Open reading frames were predicted using prodigal v.2.6.3 [59] in its meta mode and annotated using DIAMOND v.0.9.9 [60] by searching against UniRef100 (e-value cutoff: 0.00001) [61]. Nanopore reads were mapped against the contigs to determine contig coverage using minimap2 as described above alongside uBin helper scripts to determine length and GC content per contig for manual binning (https://github.com/ProbstLab/uBin-helperscripts [18.01.23]). Manual binning and estimation of completeness and contamination of the genome bins were carried out using uBin [30]. Plasmids were not included in the bins.

Assessment of MAG quality

The average nucleotide identity (ANI) between reference genomes and bins was calculated using fastani v.1.33 [62]. Agreement of reference genome and reconstructed MAG was determined by aligning them to each other using compare-sets (https://github.com/CK7/compare-sets [18.01.23]). For Assessing ANI and length consensus only the chromosomal genome was considered, i.e., plasmids were excluded from the reference genomes for this step.

Illumina assembly and hybrid assembly

Illumina reads were taken from Sereika et al. deposited at ENA with the BioProject ID RJEB48692 [27]. Quality control of Illumina reads was performed using BBduk (Bushnell B. – sourceforge.net/projects/bbmap/ [21.12.22]) and Sickle [63]. Reads were assembled using metaSPAdes 3.15.4 [64]. Hybrid assemblies, i.e. combinations of Illumina reads and Nanopore reads, were performed using hybridSPAdes3.15.4 with the --nanopore option [29]. Statistics like N50 and GC-content of all assemblies were assessed using SeqKit v.2.3.0 [65].

Data visualization

Statistical evaluation and data visualization was done in R [66] using the packages tidyverse [67], ggplot2 [68], ggalt [69], ggnewscale [70], rcartocolor [71], ape [72] and biscale [73].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09853-w.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

We thank Sabrina Eisfeld and Ines Pothmann for laboratory maintenance and Ken Dreger for server administration and maintenance. Maximiliane Ackers is acknowledged for administrative support.

Author contribution

Sophie A. Simon: Conceptualization, Methodology, Formal Analysis, Investigation, Project administration, Visualization, Writing – Original Draft. Katharina Schmidt: Investigation, Formal Analysis. Lea Griesdorn: Investigation. André Rodrigues Soares: Visualization, Writing – Review & Editing. Till L. V. Bornemann: Provided software, Writing – Review & Editing. Alexander J. Probst: Conceptualization, Methodology, Supervision, Project administration, Writing – Original Draft, Funding acquisition.

Funding

This study was funded by the German Federal Ministry of Education and Research within the project "MultiKulti" (BMBF funding code: 161L0285E). Open Access funding enabled and organized by Projekt DEAL.

Data Availability
Sequencing data has been deposited to SRA and are available under the BioProject PRJNA944173.

Declarations

Ethics approvals and consent to participate
Not applicable.

Consent for publication
Not applicable.

Conflict of interest
The authors declare that there are no conflicts of interest.

Received: 3 March 2023 / Accepted: 28 November 2023
Published online: 01 December 2023

References

1. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:16048.
2. Spang A, Caceres EF, Ettema TJG. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science. 2017;357:eaaf3883.
3. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, et al. The global Virome Project. Science. 2018;359:872–4.
4. Handelsman J, Metagenomics. Application of Genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68:669–85.
5. Murakami T, Takeuchi N, Mori H, Hirose Y, Edwards A, Irvine-Fynn T, et al. Metagenomics reveals global-scale contrasts in nitrogen cycling and cyanobacterial light-harvesting mechanisms in glacier cryoconite. Microbiome. 2022;10:50.
6. Schwank K, Bornemann TLV, Dombrowski N, Spang A, Banfield JF, Probst AJ. An archaeal symbiont-host association from the deep terrestrial subsurface. ISME J. 2019;13:2135–9.
7. Dombrowski N, Teske AP, Baker BJ. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. Nat Commun. 2018;9:4999.
8. Farag IF, Biddle JF, Zhao R, Martino AJ, House CH, León-Zayas RI. Metabolic potentials of archaeal lineages resolved from metagenomes of deep Costa Rica sediments. ISME J. 2020;14:1345–58.
9. Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication rates in microbial communities. Nat Biotechnol. 2016;34:1256–63.
10. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. PLOS Comput Biol. 2015;11:e1004095.
11. Danko D, Bezdan D, Afshin EE, Ahsanuddin S, Bhattacharya C, Butler DJ, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell. 2021;184:3376–3393e17.
12. Long AM, Hou S, Ignacio-Espinoza JC, Fuhrman JA. Benchmarking microbial growth rate predictions from metagenomes. ISME J. 2021;15:183–95.
13. Tamarit D, Caceres EF, Krupovic M, Nijland R, Eme L, Robinson NP, et al. A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal viruses. Nat Microbiol. 2022;7:948–52.
14. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. Untangling genomes from metagenomes: revealing an uncultured class of Marine Euryarchaeota. Science. 2012;335:587–90.
15. Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D, et al. Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. Nat Commun. 2013;4:2120.
16. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. Genome Res. 2020;30:315–33.
17. 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. Nat Commun. 2016;7:13219.
18. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2021;39:499–509.
19. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nat Commun. 2021;12:2009.
20. Probst AJ, Banfield JF. Homologous recombination and transposon propagation shape the Population structure of an organism from the deep subsurface with minimal metabolism. Genome Biol Evol. 2018;10:1115–9.
21. Arumugam K, Bessarab I, Haryono MAS, Liu X, Zuniga–Montanez RE, Roy S, et al. Recovery of complete genomes and non-chromosomal replicons from activated sludge enrichment microbial communities with long read metagenome sequencing. Npj Biofilms Microbiomes. 2021;7:23.
22. Liu L, Yang Y, Deng Y, Zhang T. Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. Microbiome. 2022;10:209.
23. Chen Y-H, Chiang P-W, Rogozin DY, Degermendzhy AG, Chiu H-H, Tang S-L. Salvaging high-quality genomes of microbial species from a meromictic lake using a hybrid sequencing approach. Commun Biol. 2021;4:996.
24. Sereika M, Petriglieri F, Jensen TBN, Sannikov A, Hoppe M, Nielsen PH, et al. Closed genomes uncover a saltwater species of Candidatus Electronema and shed new light on the boundary between marine and freshwater cable bacteria. ISME J. 2023. https://doi.org/10.1038/s41396-023-01372-6.
25. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol. 2020;38:701–7.
26. Cuscó A, Pérez D, Viñes J, Fàbregas N, Francino O. Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces. BMC Genomics. 2021;22:330.
27. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19:823–6.
28. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
29. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 2016;32:1009–15.
30. Bornemann TLV, Esser SP, Stach TL, Burg T, Probst AJ. uBin – a manual refining tool for genomes from metagenomes. Environ Microbiol. 2023;:1462–2920.16351.
31. Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, et al. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. Environ Microbiol. 2017;19:459–74.
32. Galata V, Busi SB, Kunath BJ, de Nies L, Calusinska M, Halder R, et al. Functional meta-omics provide critical insights into long- and short-read assemblies. Brief Bioinform. 2021;22:bbab330.
33. Trigodet F, Lolans K, Fogarty E, Shaiber A, Morrison HG, Barreiro L, et al. High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. Mol Ecol Resour. 2022;22:1786–802.
34. Yahara K, Suzuki M, Hirabayashi A, Suda W, Hattori M, Suzuki Y, et al. Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. Nat Commun. 2021;12:27.
35. Maguire M, Kase JA, Roberson D, Muruvanda T, Brown EW, Allard M, et al. Precision long-read metagenomics sequencing for food safety by detection and assembly of Shiga toxin-producing Escherichia coli in irrigation water. PLoS ONE. 2021;16:e0245172.
36. Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. Environ Microbiol. 2020;22:4000–13.
37. Poghosyan L, Koch H, Frank J, van Kessel MAHJ, Cremers G, van Alen T, et al. Metagenomic profiling of ammonia- and methane-oxidizing microorganisms in two sequential rapid sand filters. Water Res. 2020;185:116288.
38. Mojarro A, Hachey J, Ruvkun G, Zuber MT, Carr CE. CarrierSeq: a sequence analysis workflow for low-input nanopore sequencing. BMC Bioinformatics. 2018;19:108.
39. Mojarro A, Hachey J, Bailey R, Brown M, Doebler R, Ruvkun G, et al. Nucleic acid extraction and sequencing from low-Biomass Synthetic Mars Analog soils for *in situ* life detection. Astrobiology. 2019;19:1139–52.
40. Drautz-Moses DI, Luhung I, Gusareva ES, Kee C, Gaultier NE, Premkrishnan BNV, et al. Vertical stratification of the air microbiome in the lower Troposphere. Proc Natl Acad Sci. 2022;119:e2117293119.

41. Luhung I, Uchida A, Lim SBY, Gaultier NE, Kee C, Lau KJX, et al. Experimental parameters defining ultra-low biomass bioaerosol analysis. Npj Biofilms Microbiomes. 2021;7:37.
42. Zhong Z-P, Tian F, Roux S, Gazitúa MC, Solonenko NE, Li Y-F, et al. Glacier ice archives nearly 15,000-year-old microbes and phages. Microbiome. 2021;9:160.
43. Liu Y, Ji M, Yu T, Zaugg J, Anesio AM, Zhang Z, et al. A genome and gene catalog of glacier microbiomes. Nat Biotechnol. 2022;40:1341–8.
44. Hwang Y, Schulze-Makuch D, Arens FL, Saenz JS, Adam PS, Sager C, et al. Leave no stone unturned: individually adapted xerotolerant Thaumarchaeota sheltered below the boulders of the Atacama Desert hyperarid core. Microbiome. 2021;9:234.
45. Rummel JD. Planetary protection policy overview and application to future missions. Adv Space Res. 1989;9:181–4.
46. Weinmaier T, Probst AJ, La Duc MT, Ciobanu D, Cheng J-F, Ivanova N, et al. A viability-linked metagenomic analysis of cleanroom environments: eukarya, prokaryotes, and viruses. Microbiome. 2015;3:62.
47. 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:e0124158.
48. Gates KS. An overview of chemical processes that damage Cellular DNA: spontaneous hydrolysis, Alkylation, and reactions with radicals. Chem Res Toxicol. 2009;22:1747–60.
49. Jiang S, Zhuang J, Wang C, Li J, Yang W. Highly efficient adsorption of DNA on Fe3+–iminodiacetic acid modified silica particles. Colloids Surf Physicochem Eng Asp. 2012;409:143–8.
50. Ozkan J, Nielsen S, Diez-Vives C, Coroneo M, Thomas T, Willcox M. Temporal Stability and Composition of the ocular surface Microbiome. Sci Rep. 2017;7:9880.
51. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
52. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8:24.
53. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass Microbiome studies: issues and recommendations. Trends Microbiol. 2019;27:105–17.
54. Liu Y, Elworth RAL, Jochum MD, Aagaard KM, Treangen TJ. De novo identification of microbial contaminants in low microbial biomass microbiomes with Squeegee. Nat Commun. 2022;13:6799.
55. Hug LA, Thomas BC, Sharon I, Brown CT, Sharma R, Hettich RL, et al. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environ Microbiol. 2016;18:159–73.
56. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018;34:2666–9.
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009;25:2078–9.
58. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020;17:1103–10.
59. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
60. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
61. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23:1282–8.
62. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
63. Joshi N, Fass J, Sickle. A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. 2011.
64. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
65. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and Ultrafast Toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11:e0163962.
66. R Core Team. R: A Language and Environment for Statistical Computing. 2021.
67. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4:1686.
68. Wickham H. ggplot2. Cham: Springer International Publishing; 2016.
69. Rudis B, Bolker B, Schulz Jggalt. Extra Coordinate Systems, "Geoms", Statistical Transformations, Scales and Fonts for "ggplot2.&#8221.
70. Campitelli E. ggnewscale: Multiple Fill and Colour Scales in "ggplot2." 2022.
71. Nowosad J. "CARTOColors" Palettes. 2018.
72. Paradis E, Claude J, Strimmer K. APE: analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20:289–90.
73. Prener C, Grossenbacher T, Zehr A, biscale. Tools and Palettes for Bivariate Thematic Mapping. 2022.

Publisher's Note