Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system

Will A. Overholt, Martin Hölzer, Patricia Geesink, Celia Diezel, Manja Marz, and Kirsten Küsel

1 Institute of Biodiversity, Aquatic Geomicrobiology, Friedrich Schiller University, Jena, Germany.
2 RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University, Jena, Germany.
3 European Virus Bioinformatics Center, Friedrich Schiller University, Jena, Germany.
4 FLI Leibniz Institute for Age Research, Jena, Germany.
5 German Center for Integrative Biodiversity Research Halle-Jena-Leipzig, Leipzig, Germany.

Summary

Assembling microbial and viral genomes from metagenomes is a powerful and appealing method to understand structure–function relationships in complex environments. To compare the recovery of genomes from microorganisms and their viruses from groundwater, we generated shotgun metagenomes with Illumina sequencing accompanied by long reads derived from the Oxford Nanopore Technologies (ONT) sequencing platform. Assembly and metagenome-assembled genome (MAG) metrics for both microbes and viruses were determined from an Illumina-only assembly, ONT-only assembly, and a hybrid assembly approach. The hybrid approach recovered 2× more mid to high-quality MAGs compared to the Illumina-only assembly, ONT-only assembly, and a hybrid assembly approach. The hybrid approach recovered 2× more mid to high-quality MAGs compared to the Illumina-only approach and 4× more than the ONT-only approach. A similar number of viral genomes were reconstructed using the hybrid and ONT methods, and both recovered nearly fourfold more viral genomes than the Illumina-only approach. While yielding fewer MAGs, the ONT-only approach generated MAGs with a high probability of containing rRNA genes, 3× higher than either of the other methods. Of the shared MAGs recovered from each method, the ONT-only approach generated the longest and least fragmented MAGs, while the hybrid approach yielded the most complete. This work provides quantitative data to inform a cost–benefit analysis of the decision to supplement shotgun metagenomic projects with long reads towards the goal of recovering genomes from environmentally abundant groups.

Introduction

Shotgun metagenomics is a powerful method that conceptually allows all the genomes from all the organisms and their associated viruses within a sample to be determined given sufficient sequencing depth (Tyson et al., 2004; Venter et al., 2004; Handelsman et al., 2007). In practice, metagenomic data typically represent hundreds to thousands of microorganisms and viruses at different coverage levels depending on the community structure within the sample (richness, evenness, and genome size variation). These data enable the determination of the community composition (who is there) and total community function (what are they capable of doing). In addition to this wealth of information, one of the most beneficial outcomes of shotgun metagenomic projects is the ability to assemble high quality, complete or nearly complete, genomes from organisms not yet amenable to cultivation (Tyson et al., 2004; Luo et al., 2012). And indeed, such metagenome-assembled-genomes (MAGs) have provided information leading to the cultivation of organisms of interest (Gutleben et al., 2018; Cross et al., 2019; Imachi et al., 2020), along with the discoveries of new metabolic processes (Daims et al., 2015), novel insights into the ecology and evolution of globally abundant groups (Delmont et al., 2018), and uncovering a wide diversity of novel phylum-level lineages that have restructured the current understanding of the tree of life (Brown et al., 2015; Hug et al., 2016).

MAGs not only represent bacteria and archaea but also include viruses [termed uncultivated viral genomes (UViGs)], which are an integral part of most microbial communities (Dutilh et al., 2014; Brum et al., 2015; Paez-Espino et al., 2016; Roux et al., 2019). Viruses are the...
most abundant biological entities in many ecosystems and can exert proportionately large effects on ecosystem functions (Fuhrman, 1999; Breitbart et al., 2002). UVIGs have led to the discovery of megaphages (with genomes >540 kb) from human and animal gut microbiomes (Devoto et al., 2019), provided first insights into their global distribution (Al-Shayeb et al., 2019), and confirmed that environmental cyanophages contribute to global marine photosynthesis rates (Fridman et al., 2017). There have been recent advances in both identifying and reporting UVIGs; however, the identification of viruses from metagenomes and their distinction from prophages continues to be challenging (Nooij et al., 2018; Roux et al., 2019; Warwick-Dugdale et al., 2019).

One exciting avenue in reconstructing de novo MAGs has been assembling and binning ONT long reads directly, and a hybrid approach whereby long read sequences act as scaffolds to short-read sequences to help improve continuity and bridge repeat regions within a genome (Chen et al., 2019; Kolmogorov et al., 2019). The concept of hybrid assemblies has been present since the advent of second-generation sequencing technologies (Roche 454, Illumina, SOLID) (Goldberg et al., 2006), but the breakthroughs in third-generation sequencing technologies, particularly PacBio and ONT, have improved the practicality of such approaches by providing access to much longer reads along with opening the door for long-read only assembly approaches (Scholz et al., 2014; Frank et al., 2016; Bertrand et al., 2019; Kolmogorov et al., 2019). Furthermore, these single-molecule long read sequences may also enable the recovery of closely related strains within metagenomes, either with a hybrid approach or with long-read only approaches (Stewart et al., 2018; Somerville et al., 2019).

The goal of this study was to compare a hybrid assembly approach, incorporating ONT long reads, to an Illumina-only short-read approach and an ONT-only long-read approach with respect to the recovery of high-quality MAGs from a groundwater ecosystem, along with the underlying assemblies. We leveraged the well-characterized monitoring transect of the Hainich Critical Zone Exploratory (CZE) that encompasses 15 monitoring wells spread across a hillslope covered by mixed beech forest, pasture, and cropland (Küsel et al., 2016). Groundwater from this site contains a wide diversity of microbial life. The microorganism component is dominated by Patescibacteria, uncultured organisms that are often missed by routine amplicon datasets (Brown et al., 2015; Herrmann et al., 2019; Wegner et al., 2019; Geesink et al., 2020), and the viral component has only recently started to be explored (Kallies et al., 2019). A previous, gene-centric-based metagenomics project has identified dominant metabolic pathways within the aquifer (Wegner et al., 2019). However, it has been challenging to directly link these key metabolic pathways to the specific microorganisms mediating them, which is critical for our understanding of the ecology of the site.

Materials and methods

Sample collection, DNA extraction

Groundwater was collected in the Hainich Critical Zone Exploratory (NW Thuringia, Germany), from shallow groundwater resources in Upper Muschelkalk bedrock that has been extensively described (Küsel et al., 2016; Lehmann and Totsche, 2020). In brief, the Hainich CZE contains a multistory, fractured aquifer system within the hillslope, composed of alternating layers of limestone and mudstone (Kohlhepp et al., 2017). The Upper-Muschelkalk aquifer system is characterized by the limestone-dominated main aquifer (Trockitenkalk formation, moTK; formerly referred to as the HTL) that is predominantly oxic and the mudstone-dominated hanging strata (including Meissner formation, moM; formerly the HTU) that is anoxic (Kohlhepp et al., 2017). Groundwater (115 l) was collected from well H52 (moM) on 11 December 2018 and was sequentially filtered through 0.2 μM and 0.1 μM PTFE filters (142 mm, Omnispore Membrane, Merck Millipore, Germany). Filters were immediately frozen on dry ice and transported to a −80°C freezer.

The DNA extraction was performed as previously described, using a phenol-chloroform-based method without mechanical lysis to minimize fragmentation (Taubert et al., 2018). Following extraction, the Zymo DNA Clean & Concentrator kit was used to purify and concentrate the DNA for both Illumina and Oxford Nanopore Technology (ONT) sequencing. DNA concentrations of 8.89 ng μl⁻¹ (0.1 μM filter fraction) and 37.7 ng μl⁻¹ (0.2 μM filter fraction) were measured using a Qubit 4 Fluorometer (Invitrogen).

Illumina metagenome preparation and initial processing

Illumina libraries from both filter fractions were generated using the NEBNext Ultra II FS DNA library preparation kit following the recommended protocol. Size selection was performed using the AMPure XP beads (Beckman Coulter). The average insert sizes were 509 bp (0.2 μM fraction) and 392 bp (0.1 μM fraction) as determined with an Agilent Bioanalyzer using a DNA7500 chip. The sequencing was performed in-house on an Illumina MiSeq with 2 × 300 bp v3 chemistry. The 0.2 μM filter fraction DNA sample generated 17 660 385 paired-end sequences (10.6 Gbp), while the smaller 0.1 μM fraction sample generated 15 546 350 (9.33 Gbp) raw sequences. These library sizes are smaller than typical metagenomic datasets, but Nonpareil estimates average coverages of 0.88 and 0.7 for the 0.1 μM and 0.2 μM fractions.
respectively, suggesting that the community diversity is well covered (Rodriguez-R and Konstantinidis, 2013; Rodriguez-R et al., 2018).

Adapter sequences from the Illumina reads were removed using BBduk using kmer searching (k = 23, hdist = 1) and reads were trimmed with a Phred score of 20, allowing to trim both sides of the read. Reads shorter than 50 bp were discarded (Bushnell, 2014). The average estimated insert sizes of the sequences were 234 bp and 214 for the 0.2 μM fraction and the 0.1 μM fraction samples, respectively. Raw reads were deposited at the ENA under accession PRJEB35315.

Oxford Nanopore metagenome preparation

We performed Nanopore sequencing of the same DNA from the 0.2 μM fraction on a single MinION flow cell (FLO-MIN106 with an R9.4.1 pore) using the 1D genomic DNA by ligation kit (SQK-LSK109, ONT) following manufacturers’ instructions with minor adaptations. In short, the initial g-TUBE shearing step was omitted and potential nicks in DNA and DNA ends were repaired in a combined step using NEBNext FFPE DNA Repair Mix and NEBNext Ultra II End repair/da-tailing Module (New England Biolabs, USA) and doubling the incubation time. A subsequent AMPure bead (Agencourt AMPure XP, Beckman Coulter) purification was followed by the ligation of sequencing adapters onto prepared ends. A second clean-up step with AMPure beads was performed and sequencing buffer and loading beads were added to the library. An initial quality check of the flow cell (ID: FAK43462) showed 1761 active pores at the start of sequencing. We loaded the DNA with a concentration of 98 ng μL⁻¹ (measured by Qubit 3 Fluorometer; Thermo Fisher Scientific) and a total amount of ~1.4 μg. The sequencing run stalled after 18 h and was restarted for another 24 h using the MinKNOW software.

Basecalling was performed using the Guppy software (v2.3.1) with the high-accuracy model r9.4.1_450bps_large_flipflop. Called reads were classified as either pass or fail depending on their mean quality score. A total of 2 380 279 reads were basecalled and of these 2 081 879 (87.5%) were passed as satisfying the quality metric. The passed reads contain a total of 11.58 Gb of DNA sequence with a mean read length of 5560 nt. This passed-fraction amounts to 91.8% of the total DNA nucleotide bases sequenced. We deposited the raw signal files (FAST5) and basecalled reads (FASTQ) at the ENA under accession PRJEB35315.

Metagenome assembly and binning

We used metaSPAdes v3.13.0 (Nurk et al., 2017) to assemble the hybrid and Illumina-only data to be able to directly compare both approaches. The sequences from the 0.2 μm sample were individually assembled with the default parameters specified by the ‘-meta’ flag with only the inclusion of the long reads with the ‘-nanopore’ flag being different. The assembly statistics for the two methods were calculated using MetaQUAST with the default settings (Mikheenko et al., 2016). A third assembly was generated from the ONT-only reads using metaFlye v.2.5 (Kolmogorov et al., 2019) with the following flags: --meta --plasmids -g700m. Subsequently, minimap2 v.2.17 (Li, 2018) was used to align the reads to the draft assembly with the -x map-only parameter. Finally, the assembly was polished with medaka v.0.9.2 (https://github.com/nanoporetech/medaka) using the model r941_min_high. We did not short-read-polish the metaFlye assembly additionally because we wanted to compare a purely ONT-based assembly.

We binned the scaffolds/contigs that were longer than 1000 nucleotides (nt) from each of the three assemblies using MaxBin2 (Wu et al., 2016) and MetaBAT2 (Kang et al., 2015, 2019) included in the MetaWRAP binning module with the ‘--universal’ flag (Uritsky et al., 2018). Differential coverage information was included using only Illumina QA/QC reads from both filter fractions. Additionally, scaffolds >3000 nt from each assembly were binned with BinSanity using the ‘wf’ workflow and the log normalized coverage file produced with the ‘BinSanity-profile’ command (Graham et al., 2017). The MetaWRAP ‘Bin_refinement’ module was used to dereplicate bins produced from the three different binning methods, using the MetaWRAP scoring algorithm which favours low redundancy values while also selecting for higher percent completion. We found that the MetaWRAP ‘Reassemble_bins’ module reduced the quality of our bins and thus proceeded with the two sets of refined bins that were at least 50% complete with less than 10% redundancy.

Each collection was imported into Anvi’o and MAG statistics were exported using the anvi-summarize command (Murat Eren et al., 2015). Our completeness and redundancy/contamination criteria were initially assessed using estimations from CheckM (Parks et al., 2015), while the estimations exported from Anvi’o were used to calculate the values represented in Fig. 1, visualized using the R package ggplot2 v.3.1.0 (Wickham, 2009; Wickham et al., 2019a). The Anvi’o completeness estimations were lower than the CheckM values while the redundancy/contamination estimations were higher, and therefore 74 hybrid, 39 Illumina-only, and 18 ONT-only MAGs were included, out of the original 82, 44, and 35 that met the CheckM specifications. Each MAG was screened for rRNA genes using Barmap, with each hit required to be at least 20% of the full gene length (Seemann, 2015). The 16S rRNA genes were taxonomically classified using
the online portal for IDTAXA (Murali et al., 2018). All statistics were compiled from the automatically refined MAGs and not from manually curated MAGs as we considered this the most direct comparison. While we consider it essential that MAGs be manually curated before publishing (Bowers et al., 2017; Shaiber and Eren, 2019), using the automated results for this specific comparison minimizes added bias.

To identify MAGs that were recovered from both assembly methods, FastANI was run on all pairwise comparisons of the Illumina-only and ONT-only MAGs to the hybrid MAGs that were initially assessed using CheckM (82 hybrid, 44 Illumina-only, 35 ONT-only) (Jain et al., 2018). All MAGs that were recovered from all three assemblies had an average nucleotide identity (ANI) of at least 96.8% to the corresponding hybrid MAG (with most being >99%), and in all cases, secondary hits were <82% ANI (Supporting Information Table S1). These shared MAGs were investigated in more detail as they enabled a direct comparison between the methods. While most of the data was normal or nearly-normal, we used Friedman tests (non-parametric one-way repeated measures ANOVA) followed by post hoc Wilcoxon T-tests with Benjamini–Hochberg (false discovery rate) corrections to test the differences in completeness, redundancy/contamination, genome length, number of scaffolds, and N50. The tests were conducted in R with the packages ‘rstatix’, ‘tidyr’, and ‘dplyr’ (R Core Team, 2014; Wickham and Henry, 2019; Wickham et al., 2019b; Kassambara, 2020). The proportion of MAGs containing each of the three rRNA genes was tested using a chi-square test of homogeneity followed by a post hoc
pairwise proportion test with FDR correction using the same packages.

The scaffolds from the hybrid and Illumina-only MAGs that were recovered from all three methods were aligned to the corresponding ONT-only MAGs using mummer wrapped within the mummer2circos.py (https://github.com/metagenlab/mummer2circos) script (Kurtz et al., 2004). All nanopore reads >1 kbp were aligned to the same ONT-MAGs using minimap2 with the ‘-axe map-ont’ flag (Li, 2018). Log2 scaled coverage profiles were generated from QAQC Illumina reads for each filter fraction using pileup.sh across 1 kb sized sections from the BBTools suite (Bushnell, 2014), after aligning the reads to the ONT-bin scaffolds using bmap.sh and sorting the alignments with SAMtools (Li et al., 2009). MAGs were taxonomically classified using GTDB-TK v0.3.2, following the classify workflow (Hyatt et al., 2010; Matsen et al., 2010; Price et al., 2010; Eddy, 2017; Jain et al., 2018; Parks et al., 2018; Chaumeil et al., 2020).

To disentangle the improvements in MAG recovery between the hybrid and Illumina-only methods due to longer ONT reads instead of simply having a higher sequencing depth, we subsampled the ONT reads into four different sets and then re-ran the SPAdes assemblies and binning steps. As described above, the full hybrid assembly approach was based on 17 660 385 Illumina paired-end sequences (10.6 Gbp), and 2 081 879 ONT sequences (11.58 Gbp). The four additional ONT sets were (i) all nanopore reads >10,000 nt which lead to 349 321 sequences (6.8 Gbp), (ii) ONT reads >20 000 nt (114 853 sequences; 3.6 Gbp), (iii) ONT reads >50 000 nt (7848 sequences; 0.48 Gbp), and (iv) randomly subsampled to 25% of the initial number of sequences (595 070 sequences, 3.15 Gbp). Each of the resulting assemblies was binned and analysed following the same procedures as above (Supporting Information Table S2).

Viral comparison

We used VirSorter v1.0.5 to search for putative virus and prophage sequences in the assemblies (Roux et al., 2015). Individually for each assembly approach and for all contigs with a length > 2000 nt that have been identified as potential viruses (categories 1, 2, and 3) or prophages (categories 4, 5, and 6) by VirSorter, we ran the VIRify pipeline (https://github.com/EBI-Metagenomics/emg-viral-pipeline) to assign taxonomic classifications to contigs if possible. The taxonomic assignment of VIRify is based on a collection of virus-specific protein sequences encapsulated in profile hidden Markov models (HMMs). For each HMM, taxonomic information about the related virus order, subfamily, family, or genus is used to assign a taxonomy to each contig based on a voting system. In short, the pipeline predicts ORFs using prodigal v2.6.3 (Hyatt et al., 2010) and then screens the resulting protein sequences with hmmscan v3.2.1 against the HMMs. The hmmscan output is then parsed and filtered to provide a taxonomic assignment for each contig based on the reported hits against the HMM database. Finally, the taxonomic assignments were visualized using Sankey plots and editing in Inkscape (https://inkscape.org).

Results and discussion

Metagenomes were generated using groundwater collected from a shallow aquifer system using both Illumina short-read and Oxford Nanopore Technology (ONT) sequencing in an effort to quantify the improvements and utility of long read sequences for metagenomic assembly and binning steps. One of the most striking findings is that the inclusion of long reads for the hybrid method more than doubled (74 versus 39) the number of bacterial and archaeal MAGs that were at least 50% complete and less than 10% redundant (assessed with Anvi’o), compared with using assemblies generated from only Illumina short-read sequences when using the same assembly and binning algorithms. The ONT-only approach that we applied yielded the fewest MAGs that met our cut-off criteria (Fig. 1 A). However, these ONT-only MAGs were much less fragmented than either the illumina-only or hybrid approaches (Fig. 1 B). Compared to the Illumina-only approach, the hybrid MAGs were less fragmented, with more than 50% of the hybrid MAGs represented by less than 100 scaffolds as compared to only 5% of the Illumina-only MAGs (Fig. 1 B). The ONT MAGs exhibited much higher N50 values with all MAGs having an N50 > 100 kbp compared to only one hybrid MAG and none of the Illumina-only MAGs (Fig. 1 C). In comparison to the illumina-only method, the hybrid approach recovered additional MAGs with lower coverage (minimum 1.9× versus 4.2×), while compared to the ONT method, the improvements of the hybrid MAGs can be attributed to a better recovery of populations that were close to our quality cutoffs, particularly when estimated by Anvi’o (Fig. 1 E).

In addition to the bacterial and archaeal MAGs, we identified 278 putative viral sequences (258 phages/viruses, 20 prophages) in the hybrid assembly, 73 (71 viruses, 2 prophages) in the Illumina-only assembly, and 332 (166 viruses, 166 prophages) in the ONT-assembly (Fig. 1 F). The viral sequences include both bacteriophages (phages), putative eukaryotic viruses (Nimaviridae, Supporting Information Fig. S1), and prophage sequences. Interestingly, the number of complete viral contigs of the VirSorter category 1 (confident virus assignment) was the lowest in the ONT-assembly (2), followed by 7 in the Illumina assembly, and 22 in the
hybrid assembly. Conversely, the number of putative prophage sequences was by far the highest in the ONT-assembly (166 compared to 20 in the hybrid and only 2 in the Illumina-only), although the majority of these calls were not confident (Fig. 1 F). The integration of long read data into the hybrid-assembly process helped to recover more potential viral sequences from the sample, as well as an increase in viral sequences that were confidently taxonomically identified, while the ONT-only assembled data alone recovered many more putative (but low confidence) prophage sequences. We assume that the difference in virus versus prophage sequences between the methods to be due to unequal differential coverage profiles caused by underlying prevalence levels within the dominant groundwater populations. Stated differently, the much longer ONT-only scaffolds were more likely to represent single organism MAGs while the hybrid assemblies that rely upon an initial short-read based assembly may reflect a consensus and population-level genome which would not include the variable regions. Alternatively, these differences may be due to underlying differences in the sequences recovered between the two methods.

When searching for the same MAGs recovered from each method, we relaxed our inclusion criteria to use the CheckM completeness/redundancy estimates rather than the more conservative Anvi’o ones, improving the statistical power of the dataset. For the bacteria and archaea, there were 21 MAGs that were recovered by all three methods (Fig. 2). Every MAG reconstructed from the Illumina-only assembly was also recovered from the hybrid assembly, likely since both utilize the same Illumina data to generate the initial assembly with SPAdes, while the ONT-long reads in the hybrid assembly are used to resolve the underlying Illumina-based De Bruijn graph. There were 29 unique MAGs to the hybrid approach and five unique to the ONT-only approach. We performed an in-depth comparison of the 21 bacteria and archaea MAGs that were initially recovered with all three assemblies (Fig. 2, Table 1; Supporting Information Table S1). The three methods generated significant differences in the estimated completeness of the shared MAGs (Friedman Stat = 21, \( P < 0.0001 \)), with follow-up post hoc Wilcoxon t-tests suggesting that the hybrid MAGs were the most complete while the ONT MAGs were the least complete (Table 1). Surprisingly, the ONT MAGs had the largest genomes, on average 179 kbp longer than the hybrid ones, and 457 kbp longer than the Illumina-only ones. This discrepancy between shorter MAGs associated with higher completeness is likely due to the higher indel (insertion/deletion) error rates associated with non-Illumina polished ONT-only data, that strongly impact ORF prediction programs, and lead to mistranslated or partial gene calls that do not match the amino acid-based HMM profiles used to find single marker copy genes (Watson and Warr, 2019). To support this hypothesis, we compared the predicted ORF length distributions for the 21 shared MAGs that were recovered from each method and found significant differences between all three sets (Kruskal–Wallis test, statistic = 10 302, \( P = 0 \)) with the ONT-only MAGs exhibiting slightly longer median predicted ORF sizes than the hybrid MAGs (\( P = 3.8e-7; \) median = 686, 659 respectively) and the ONT-only the smallest (\( P = 0; \) median = 410) (Supporting Information Fig. S2). Accordingly, the effect size (Wilcoxon test) between the hybrid and Illumina-only bins was much smaller (\( r = 0.019 \)), than what was observed between the ONT-only bins and the hybrid (\( r = 0.26 \)) or the Illumina-only (\( r = 0.27 \)) (Supporting Information Fig. S2). Furthermore, the ONT-only MAGs had 63 453 predicted ORFs while the same MAGs from the hybrid method consisted of 40 589 ORFs and the Illumina-only MAGs had 34 755. This increase in potentially truncated gene calls is likely the reason we were unable to assign any virus-specific marker HMMs with the ViRity pipeline to contigs derived from the ONT-only assembly. There were no significant differences identified in the redundancy estimates, further suggesting that the discrepancy in MAG size cannot be explained by a higher prevalence of chimeric MAGs in the hybrid approach. As suggested by Fig. 1, these ONT-only MAGs were significantly less fragmented and with significantly larger N50s (Table 1).

The ONT-only MAGs had a significantly higher proportion of 16S and 23S rRNA genes than either of the versions recovered with the hybrid or Illumina-only method (Table 2). Furthermore, there was no significant difference found in the proportion of the rRNA genes between the hybrid and Illumina-only recovered MAGs. That the hybrid assembly method was not nearly as successful as the ONT-only method in recovering these challenging-to-assembly genomic regions suggests that metaSPAdes is not able to fully leverage all information contained within the long reads, likely because the SPAdes algorithm

![Fig 2. Comparison of the number of MAGs recovered using each method and those shared between the methods. The quality cutoffs for inclusion were >50% complete and <10% redundant as assessed using CheckM. [Color figure can be viewed at wileyonlinelibrary.com]](image-url)
starts from a short-read-based graph structure even if long-read information is provided. Due to the difference in the total number of MAGs recovered by each method, we also investigated their performances when including all the MAGs that passed the CheckM quality cutoffs, rather than just the shared MAGs. Even though the hybrid method recovered more than $2 \times$ more MAGs than the ONT method, approximately 30 MAGs had rRNA genes with each method (Supporting Information Table S3). Comparing the hybrid to Illumina-only MAGs, we recovered 160% more 5S genes along with 250% more 16S genes and 23S genes using the hybrid approach (Supporting Information Table S3). However, there was no significant difference in the length of these genes (Supporting Information Table S3). Furthermore, while there were no significant differences in the length of the 16S genes found within hybrid and ONT-only MAGs, the 23S genes were significantly longer in the ONT-only MAGs. Finally, we checked if the taxonomies of the 16S genes were congruent with the GTDB_TK, full MAG-based, taxonomy (Supporting Information Table S4). Of the 95 MAGs from the three methods, only two exhibited a confident 16S assignment different from the MAG taxonomy (one from the ONT-only method, and one from the Illumina-only approach).

Due to the importance of rRNA genes, particularly the 16S gene, in microbial ecology, we also screened the three assemblies for these genes and their lengths (Supporting Information Table S5). The ONT-only assembly had the most 16S, 23S, and 5S genes; with the 16S and 23S genes both found to be significantly longer than in either of the other assemblies (Supporting Information Table S5). The hybrid assembly 23S genes were significantly longer than those found in the Illumina-only assembly, while there were no significant differences identified between the lengths of 16S genes.

We also briefly compared the underlying assemblies based on the number of contigs, the N50, average contig size, and the total assembly size (Supporting Information Table S6). When considering all contigs produced, the hybrid assembly was 0.7 Gb followed by the Illumina assembly at approximately 0.6 Gbp, and the ONT-only at 0.4 Gbp. Similar to the MAG statistics, the Illumina-only assembly was highly fragmented, with both an average
ONT hybrid assembly improvements in groundwater

contig size and N50 of 600 bp. Conversely, the ONT-only assembly had an average size of 40 kbp and an N50 of 90 kbp. When removing all contigs <1 kbp, the ONT-only assembly was approximately the same size as the hybrid assembly, both of which were 2× larger than the Illumina-only assembly (Supporting Information Table S6). The largest contig recovered from the ONT-only assembly was 2.4 Gbp, which was 10× longer than the longest hybrid contig and >25× the largest Illumina-only contig.

The Illumina dataset used here is smaller than typical Illumina metagenomes, particularly with the advent of the Illumina NovaSeq, and it is unusual to recover equal sequencing depths between Illumina and ONT-based reads. To explicitly test whether the improvements in MAG-retrieval using the hybrid approach over the Illumina-only method was due to the increase in sequencing depth provided by the ONT reads, we subsampled using four different strategies. These ranged from only using ONT reads >50 000 nt (7848 sequences; 0.48 Gbp) to all reads >10 000 nt (349 321; 6.8 Gbp), along with a random subsampling to simulate a less successful sequencing run (595 070 sequences; 3.15 Gbp) (Supporting Information Table S2). The randomly subsampled ONT set yielded 62 MAGs in the hybrid approach that were >50% complete and <10% redundant (as assessed by CheckM), 18 more MAGs than the MiSeq assembly alone. Including only the ~8000 reads with more than 50 000 nt resulted in 58 MAGs passing our standards. We interpret these results to show that the longer nanopore reads do help recover MAGs, and a large improvement can be expected even if the ONT run was not so successful in terms of yield, while also acknowledging that the greater sequencing depth alone supplied by the ONT reads partially contributes to these results.

The hybrid assembly recovered the highest diversity of MAGs, representing 17 bacterial and archaeal phyla, while 14 were represented from the Illumina-only assembly, and 12 from the ONT-only assembly. The missing phyla from the Illumina-only assembly were Bacteroidota (2 MAGs), Micrarchaeota (1), and Verrucomicrobiota (1). The ONT-only method was also missing MAGs from the Verrucomicrobiota, along with Elusimicrobiota, Gemmatimonadota, GCW2-55-46, and MBNT15 (Supporting Information Table S7). The (super) phylum with the greatest improvement in the number of assembled MAGs was the Patescibacteria (formerly the Candidate Phyla Radiation group), with over double (38 versus 17 and 13) the number of MAGs recovered from the hybrid approach compared to the Illumina-only or ONT-only that met our criteria (Supporting Information Table S7). In comparing the hybrid approach to the Illumina-only approach we show that the ONT long read sequences bridge missing gaps in the Illumina-only MAGs (Fig. 3), thereby improving the contiguity and increasing the genome length across fewer scaffolds. It is noteworthy that these benefits are not restricted to only the most abundant organisms and even relatively few long reads mapping to populations can improve chances to recover them (Fig. 3 B).

A similar comparison was performed using the most abundant and least abundant Patescibacteria MAGs that were recovered using all three approaches. Supporting Information Figure S3 highlights regions that the hybrid approach failed to incorporate the longer read information. Again we focused on the Patescibacteria, since this group had the largest differences in the number of recovered MAGs, contains almost no cultivated representatives, and they are often highly abundant in groundwater systems (Brown et al., 2015; Pedron et al., 2019). In addition, previous research from the Hainich CZE has demonstrated that Patescibacteria can represent up to 79% of the groundwater community (Herrmann et al., 2019).

A visualization between the hybrid and Illumina-only MAGs was also performed for three other dominant bacterial phyla (Nitrospirota, Actinobacteriota, Proteobacteria) that had previously been shown to be important in the functioning of the groundwater (Wegner et al., 2019; Yan et al., 2019) (Supporting Information Fig. S4). However, for all 17 phyla we detected, the hybrid approach always yielded the most MAG representatives relative to the other two approaches (Supporting Information Table S7). There were almost double the number of Nitrospirota associated MAGs recovered with the hybrid approach compared to the other two methods, and these were amongst the largest of the MAGs recovered (Supporting Information Table S7; Fig. S4). Within the Actinobacteriota and Proteobacteria, a similar pattern was observed. The MAG metrics within each phylum fell within the confidence intervals of the full sample set with the exception of the Actinobacteriota that showed greater than expected improvements with the hybrid approach compared to the Illumina-only method (Supporting Information Fig. S4).

The majority of viral sequences identified in all three assemblies were not classified using ViRify (Supporting Information Fig. S1). However, a similar pattern to the classified microbial diversity was observed in the classified viral diversity, with the hybrid approach representing the highest diversity of named viruses (with hits for Podoviridae, Myoviridae, Microviridae, Nimaviridae, and Poxviridae), followed by the Illumina-only assembly (Podoviridae, Microviridae, Poxviridae). We were unable to classify any of the ONT-only putative viral sequences, even though this method yielded the most viral sequences overall (Fig. 1; Supporting Information Fig. S1). Again, we suspect this is due to difficulties in
identifying high-quality ORFs that are then used to compare to taxonomic databases and this is reflected in the lower confidence in identified viral sequences within the ONT-only assembly (Fig. 1 F). Our results further underline the importance of polishing an ONT metagenomic assembly with short reads, if available, to further reduce the error rate (mainly indels) and thus improve subsequent sequence-based analysis steps (Watson and Warr, 2019).

Our results from these diverse and understudied groundwater ecosystems extend findings recently published that utilized mock communities and spiked-in complex human gut microbiome communities which compared a hybrid approach to a short-read only approach (Bertrand et al., 2019). In addition, we demonstrated the recovery of a wider diversity of microorganisms and viruses using a hybrid approach which lead to the recovery of three additional phyla and six classes. While the ONT-only assembly generated the fewest number of MAGs, the importance of a low degree of fragmentation, high N50 values, and a preponderance of rRNA genes should not be overlooked. With respect to the hybrid approach, the improvements towards contiguity and completeness of recovered MAGs are likely generalizable as...
they are reflected within the results presented by Bertrand et al. (2019). While not explicitly tested in our study, the hybrid approach has been previously shown to result in fewer misassemblies using a mock community where the genome contents are known a priori (Bertrand et al., 2019).

One of the limitations in interpreting our results was the inclusion of two different assembly methods between the hybrid/Illumina-only results and the ONT-only ones. The ability to use metaSPAdes enabled us to constrain the improvements in the hybrid-based assembly to the inclusion of the ONT-long reads, and mapping back those same reads showed how those improvements were achieved. However, some of the differences observed between the hybrid and the ONT-only MAGs may be due to software differences between metaSPAdes and metaFlye. The decision to use the coverage provided by the Illumina samples to aid in the binning of the ONT-only assembly may also influence our results. A few factors went into this decision, but in our opinion, the largest biases would come from having two samples provide coverage information for the binning algorithms for the hybrid and Illumina-only assemblies, while we only had a single ONT sample. To mitigate this effect, we opted to use the same set of reads to generate the coverage profiles for all contigs from each of the three assemblies. A secondary issue was that the binning tools utilized do not natively support long-read coverage information, meaning a different suite of read mappers would be used for only the ONT-based method. Only using the ONT-based coverage information from the single sample was tested by generating the coverage profiles for metabat2 and maxbin2 with minimap2, which resulted in the recovery of 8 of the 35 MAGs (above our QA/QC threshold) that were produced when we used the coverage information from the two Illumina samples. We suspect this was due to a loss of information rather than any consequence of the ONT-only reads but cannot rule it out at this time.

In addition, the higher error rate associated with indels impacts the ONT-only MAGs to a greater extent in this study due to our >50% completeness and <10% redundancy thresholds that are based on single copy marker genes. If these genes are not predicted correctly within the contigs, they will be missed by the HMM models used to assess completeness and contamination metrics (Watson and Warr, 2019) as well as methods to taxonomically classify MAGs. Missed gene calls and a higher proportion of truncated ORFs may also have severe consequences for functionally annotating and metabolically profiling ONT-only based MAGs, which we only cursorily touch upon here. Along with the shallower sequencing depth in the ONT-only approach, these types of sequence errors may contribute to the recovery of fewer MAGs in the ONT-only method. Furthermore, the higher error rates associated with the ONT-reads may also partially explain why the hybrid method did not perform better in recovering MAGs with a similar number of contigs and proportion of rRNA genes as compared to the ONT-only method, i.e. ONT read errors may have exceeded the sequence similarities required by metaSPAdes to bridge two Illumina-based contigs, reducing its efficiency. Alternatively, particularly in regions of low Illumina sequence coverage these errors may have been included in the hybrid assembly.

Unlike more traditional, gene-centric-based analyses that provide insights into the metabolic repertoire of an ecosystem, a genome-centric approach enables research questions directed towards population niche differentiation, determination of microbial groups that are bio-indicators for a specific metabolism, and potential microbial networks and microbial interactions between syntrophs and/or auxotrophs along with the phages that control population sizes and alter or enhance biogeochemical cycling rates (Anantharaman et al., 2016; Howard-Varona et al., 2017). Such an approach further offers insights into the evolutionary history of archaeal, bacterial, or viral groups and the ecological consequences if such groups were to be lost or invade the system. In particular, the identification of novel viruses from metagenomic data and the way they interact with other microbes extends our understanding of complex environmental systems (Roux et al., 2016). As a final consideration, the information contained within high-quality MAGs may offer a road map to cultivation, which in turn allows hypothesis testing and verification of in silico predictions (Cross et al., 2019).

Conclusions

To improve the recovery of metagenome-assembled-genomes we find that the addition of ONT long read sequences doubled the number of bacterial and archaeal MAGs, represented more phylogenetic diversity, and improved all measured quality metrics as compared with an Illumina short-read approach only. In addition, nearly four-fold more putative viral sequences were identified including 10x more putative prophages. Furthermore, ONT-only assembly and binning greatly complement a hybrid approach, providing less depth but improving genome continuity and most measured genome metrics for the shared MAGs. Ultimately, a multifaceted approach utilizing a dereplicated set of recovered MAGs from multiple assembling and binning methods is likely to improve our ability to reconstruct genomes from the environment and may enable the recovery of closely related strains within a sample or project. ONT-only projects will likely increase as this technology becomes more available and is also effective at reconstructing dominant populations.
on their own. Polishing the ONT-reads using Illumina information either prior to or after ONT-only assembly can improve the recovery of high-quality MAGs that meet common quality threshold criteria by reducing indel errors and improving ORF predictions.

Considerations on supplementing Illumina paired-end metagenomic projects with ONT reads include the DNA extraction method used, the total amount of DNA available, and the cleanliness and fragmentation of the extract. The additional amount of hands-on time needed to prepare a sequencing run with the ONT MinION is comparably low. The current library preparation time using the revised genomic DNA by ligation kit takes about 3 h (including elongated incubation times). Shorter protocols such as the rapid sequencing kit (e.g. SQK-RAD004, ONT store) are also available to the ONT-community. A sequencing run lasts between 24 and 48 h, or until no active pores are available anymore and the data can be immediately analysed depending on the available hardware. There are cost concerns with supplementing an already expensive metagenomic sequencing project with ONT long-read sequences, considering a complete run on one MinION flow cell, including library preparation, currently costs 750 € per sample. However, the improvements documented here provide better genome context for both microbial and viral comparative genomic projects, more single marker copy genes for detailed phylogenomic studies, more complete metabolic reconstructions, and an end-product that is more useful to the greater scientific community. The results from this study are likely conservative and we expect there to be further improvements as ONT sequence quality increases with more accurate base calling algorithms and as more assembly and binning algorithms are developed to take advantage of all the information provided by long reads.

Acknowledgements

The authors thank Falko Gutmann, Perla Abigail Figueroa-Gonzalez, Till Bornemann and Heiko Minkmar for water sampling and water filtration. The authors would also like to thank Alexander Probst for his helpful comments and discussion, and Stefan Riedel for operating the Illumina MiSeq. This study was conducted within the Collaborative Research Centre AquaDiva (CRC 1076 AquaDiva) of the Friedrich Schiller University Jena and was funded by the Deutsche Forschungsgemeinschaft. Martin Hölzer additionally appreciates the support of the Joachim Herz Foundation by the add-on fellowship for interdisciplinary life science. Patricia Geesink was further supported by a Strategy and Innovation Grant from the Free State of Thuringia (41-5507-2016), the Leibniz Research Cluster InfectoOptics (SAS-2015-HKI-LWC), and the Excellence Cluster Balance of the Microverse (EXC 2051/1) funded by Deutsche Forschungsgemeinschaft.

References

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., et al. (2020) Clades of huge phages from across Earth’s ecosystems. Nature, 578: 425–431. https://doi.org/10.1038/s41586-020-2007-4
Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7: 13219.
Bertrand, D., Shaw, J., Kalathiyanpam, M., Ng, A.H.O., Kumar, M.S., Li, C., et al. (2019) Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol 37: 937–944.
Bowers, R.M., Kyprides, N.C., Stepansauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35: 725–731.
Breitbart, M., Salamon, P., Andersen, B., Mahaffy, J.M., Segall, A.M., Mead, D., et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99: 14250–14255.
Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., et al. (2015) Unusual biology across a group comprising more than 15% of domain bacteria. Nature 523: 208–211.
Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., et al. (2015) Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science 348: 1261498.
Bushnell, B. (2014) BBTools software package. URL http://sourceforge.net/projects/bbmap
Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D. H. (2020) GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics 36: 1925–1927.
Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. Genome Research 30: 315–333. http://dx.doi.org/10.1101/gr.258640.119.
Cross, K.L., Campbell, J.H., Balachandran, M., Campbell, A. G., Cooper, S.J., Griffen, A., et al. (2019) Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol 37: 1314–1321.
Daims, H., Lebedeva, E.V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., et al. (2015) Complete nitrification by Nitrospira bacteria. Nature 528: 504–509.
Delmont, T.O., Quince, C., Shaiber, A., Esen, Ö.C., Lee, S. T., Rappé, M.S., et al. (2018) Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol 3: 804–813.
Devoto, A.E., Santini, J.M., Olm, M.R., Anantharaman, K., Munk, P., Tung, J., et al. (2019) Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol 4: 693–700.
Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., et al. (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5: 4498.
Eddy, S. (2017) HMMER3: a new generation of sequence homology search software.

Frank, J.A., Pan, Y., Tooming-Kluenderud, A., Eijsink, V.G.H., McHardy, A.C., Nederbragt, A.J., and Pope, P.B. (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep 6: 25373.

Fridman, S., Flores-Urbe, J., Larom, S., Alalouf, O., Liran, O., Yacoby, I., et al. (2017) A myxovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. Nat Microbiol 2: 1350–1357.

Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399: 541–548.

Geesink, P., Wegner, C.-E., Probst, A.J., Hermann, M., Dam, H.T., Kaster, A.-K., and Küsel, K. (2020) Genome-inferred spatio-temporal resolution of an uncultivated Roizmanbacterium reveals its ecological preferences in groundwater. Environ Microbiol 22: 726–737.

Goldberg, S.M.D., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., et al. (2006) A sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103: 11240–11245.

Graham, E.D., Heidelberg, J.F., and Tully, B.J. (2017) BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. PeerJ 5: e3035.

Gutleben, J., Chiab De Mares, M., van Elsas, J.D., Smit, H., Overmann, J., and Sipkema, D. (2018) The multi-omics promise in context: from sequence to microbial isolate. Crit Rev Microbiol 44: 212–229.

Handelsman, J., Tiedje, J., Alvarez-Cohen, L., Ashburner, M., Cann, I.K.O., Delong, E.F., et al. (2007) The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet, Washington, DC: The National Academies Press

Herrmann, M., Wegner, C.-E., Taubert, M., Geesink, P., Lehmkanf, K., Yan, L., et al. (2019) Predominance of Cand. Patescibacteria in groundwater is caused by their oligotrophic conditions. Front Microbiol 10: 1407.

Howard-Varona, C., Hargreaves, K.R., Abedon, S.T., and Sullivan, M.B. (2017) Lysogeny in nature: mechanisms, impact and ecology of temperate phages. Front Microbiol 8: 1511–1520.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., et al. (2016) A new view of the tree of life. Nat Microbiol 1: 16048.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.

Imachi, H., Nobu, M.K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., et al. (2020) Isolation of an archaeon at the prokaryote-eukaryote interface. Nature 577: 519–525.

Jain, C., Rodríguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9: 5114.

Kallies, R., Hölzer, M., Brizola Toscan, R., Nunes da Rocha, U., Anders, J., Marz, M., and Chatzinotas, A. (2019). Evaluation of Sequencing Library Preparation Protocols for Viral Metagenomic Analysis from Pristine Aquifer Groundwaters. Viruses 11: 484–502. http://dx.doi.org/10.3390/v11060484.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ Preprints 7: e7359.

Kassambara, A. (2020) rstatix: Pipe-Friendly Framework for Basic Statistical Tests.

Kohlhepp, B., Lehmann, R., Seeber, P., Küsel, K., Trumbore, S.E., and Totsche, K.U. (2017) Aquifer configuration and geostuctural links control the groundwater quality in thin-bedded carbonate–silicilastic alternations of the Hainich CZE, Central Germany. Hydrol Earth Syst Sci 21: 6091–6116.

Kolmogorov, M., Rayko, M., Yuan, J., Polevиков, Е., and Pevzner, P. (2019) metaFlye: scalable long-read metagenome assembly using repeat graphs. bioRxiv: 637637. https://doi.org/10.1101/637637.

Kurtz, S., Phlippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.

Küsel, K., Totsche, K.U., Trumbore, S.E., Lehmann, R., Steinhäuser, C., and Herrmann, M. (2016) How deep can surface signals be traced in the critical zone? Merging biodiversity with biogeochemistry research in a central German Muschelkalk landscape. Front Earth Sci China 4: 32.

Lehmann, R., and Totsche, K.U. (2020) Multi-directional flow dynamics shape groundwater quality in sloping bedrock strata. J Hydrol 580: 124291.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Luo, C., Tsementzi, D., Kyrpides, N.C., and Konstantinidis, K.T. (2012) Individual genome assembly from complex community short-read metagenomic datasets. ISME J 6: 896–901.

Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010) Pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11: 538.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016) MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32: 1088–1090.

Murri, A., Bhargava, A., and Wright, E.S. (2018) IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. Microbiome 6: 140.

Murat Eren, A., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015)
(categories 1, 2, and 3) in the hybrid and MiSeq assembly approach. The VIRify pipeline (https://github.com/EBI-Metagenomics/emg-viral-pipeline) was used to assign a taxonomy for each contig based on a set of virus-specific protein-profile HMMs. The image shows the viral ranks that were identified in both assemblies and their overlap.

**Figure S2.** Predicted lengths of the ORFs from the 21 shared MAGs recovered from each of the assembly approaches. The outliers are not shown. Median values are indicated in the center of each boxplot. The number of ORFs predicted (n) for each are depicted below the x-axis, and the results from post hoc Wilcoxon (Mann–Whitney) tests are shown above the boxplots (false discovery rate corrected p values, the effect size).

**Figure S3.** Genome circos plots for the most (A) and least (B) covered Patescibacteria MAGs retrieved by all three assembly methods. The outer ring in blue represents the ONT assembly derived scaffolds, followed by the aligned hybrid scaffolds, and then the Illumina-only scaffolds. The nanopore reads were mapped with minimap2 and coloured based on length. The coverage values are log2 scaled and calculated for each 1 kb segment of the ONT-derived scaffolds with pileup.sh from BBTools of the Illumina reads. The values below each plot represent the mean coverage from the 0.2 μM fraction Illumina MiSeq reads and the 0.1 μM fraction reads, respectively. The hybrid genome size is indicated in the middle of each plot.

**Table S1.** Comparison of the MAGs (bins) that were recovered from all three methods.

**Table S2.** The effect of subsampling the ONT reads on the number of MAGs that were >50% complete and <10% redundant (as assessed by checkM).

**Table S3.** The number and length of the rRNA genes recovered from the MAGs with >50% completion and <10% redundancy, as assessed by checkM. Differences in the length

**Table S4.** Congruency between the bin taxonomy and the 16S gene taxonomies (if present). The 16S taxonomy was determined using the DECIPHER online portal using both

**Table S5.** The number and length of the rRNA genes recovered from each assembly. Differences in gene lengths were tested for using Kruskal-Wallis tests followed by

**Table S6.** Comparison of assembly statistics between the three methods. Statistics calculated for (top) all contigs, (middle) contigs larger than 1 kb, (bottom) contigs larger

**Table S7.** Number of MAGs recovered classified into bacterial phyla by GTDB_TK.

© 2020 The Authors. *Environmental Microbiology* published by Society for Applied Microbiology and John Wiley & Sons Ltd., *Environmental Microbiology*, 22, 4000–4013