Assessing long-read sequencing with Nanopore R9, R10 and PacBio CCS to obtain high-quality metagenome assembled genomes from complex microbial communities

Mantas Sereika[a*], Rasmus Hansen Kirkegaard[a*], Søren Michael Karst[a], Thomas Yssing Michaelsen[a], Emil Aare Sørensen[a] and Mads Albertsen[a**]

[a]Center for microbial communities, Aalborg University, Denmark

*These authors contributed equally to the paper
**Corresponding author ma@bio.aau.dk

Abstract

Short-read DNA sequencing has led to a massive growth of genome databases but mainly with highly fragmented metagenome assembled genomes from environmental systems. The fragmentation is a result of closely related species, strains, and genome repeats that cannot be resolved with short reads. To confidently explore the functional potential of a microbial community, high-quality reference genomes are needed. In this study, we evaluated the use of different combinations of short (Illumina) and long-read technologies (Nanopore R9.4, R10.3, and PacBio CCS) for recovering high-quality metagenome assembled genomes (HQ MAGs) from a complex microbial community (anaerobic digester). Depending on the sequencing approach, 33 to 86 HQ MAGs (encompassing up to 34 % of the assembly and 49 % of the reads) were recovered using long reads, with Nanopore R9 featuring the lowest sequencing costs per HQ MAG recovered. PacBio CCS was also found to be an effective platform for genome-centric metagenomics (74 HQ MAGs) and produced HQ MAGs with the lowest fragmentation (median of 9 contigs) as a stand-alone technology. Using PacBio CCS MAGs as reference, we show that, although a high number of high-quality MAGs can be generated using Nanopore R9, systematic indel errors are still present, which

26 can lead to truncated gene calling. However, polishing the Nanopore MAGs with short-read Illumina
27 data, enabled recovery of MAGs with similar quality as MAGs from PacBio CCS.

28 Importance

29 Multiple approaches exists for performing genome-centric metagenomics to recover high-quality
30 genomes from microbial communities. The plethora of possible sequencing strategies can be
31 cumbersome for designing metagenomics projects, especially under limited budget conditions.
32 Here, we performed single sequencing runs of the same complex microbial community sample using
33 Illumina MiSeq, Oxford Nanopore MinION and PacBio Sequel II to assess the performance of each
34 platform. We show that, in general, long read sequencing significantly outperforms the short read
35 Illumina platform for recovering microbial genomes. In addition, we observed that the hybrid
36 Nanopore-Illumina approach recovers genomes of comparable quality to PacBio CCS, while
37 maintaining lower sequencing costs per genome recovered. For this reason, we find the
38 Nanopore R.9.4.1 and supplemental Illumina read polishing to be, at the moment of writing, the
39 most cost-effective sequencing strategy for *de novo* acquisition of high-quality genomes.

40

Introduction

Bacteria live in almost every environment on Earth and a recent analysis of the global microbial diversity estimated a total of 10^{12} species (1). To obtain representative genomes culturing has been used to isolate specific microbes, but the throughput is highly limited. Recently, genome recovery from metagenomes with short, high-quality reads has drastically improved genome recovery for uncultivated species (2–4), and has spurred multiple tools for automated binning of draft genomes (5–8). This has led to studies reporting thousands or even hundreds of thousands of genomes (9–13). However, the recovered metagenome assembled genomes (MAGs) are often very fragmented, incomplete, and contaminated with genome fragments from other microbes. Hence, even though the pace of genome sequencing is increasing, only a tiny fraction of the estimated microbial diversity have representative genomes (14) with 47,894 species having representative genomes in the Genome Taxonomy Database (GTDB) version 202 (15).

Recovering genomes from metagenomes using short reads is challenging due to the presence of genetic repeat elements. From an assembly point-of-view, repeat elements are classified as identical DNA segments that are not able to be spanned by the length of the sequenced reads. For example, the 16S ribosomal RNA gene is highly conserved between species, can be present in multiple copies within species, and is much longer than the typical short reads. This either breaks up the assembly graph or produce chimeric segments (16, 17). In complex microbial communities a large source of repeat elements originates from closely related microbial strains that share large parts of their genomic content. The repeat elements have a coverage that is different from the rest of the genome, and are thus often not picked up by automated binning tools (11). In addition,

63 genomes can also be contaminated with fragments from other organisms that just happen to
64 correlate with the binned genome (11, 14, 18). Tools have been developed to estimate the
65 completeness and the level of contamination and even "decontaminate" the bins (16, 18). However,
66 they rely heavily on good reference databases and assumptions about universal essential single copy
67 genes that are known not to be truly universal (19, 20). Maybe the only way to be certain about the
68 completeness of a genome is to produce closed MAGs as in the case of circular chromosomes (21).
69 This is especially important as repeat regions may contain biological information that is essential to
70 understand a microbial community (22), as have been shown for some antimicrobial resistance
71 genes (23).

72

73 Long-range information is, to a large degree, capable of solving the genome fragmentation caused
74 by repeat elements (24). Different approaches have been developed to generate such data, e.g.
75 mate-pair sequencing, synthetic long-read sequencing, and long-read single molecule sequencing,
76 such as offered by Pacific Biosciences sequencing (PacBio) and Oxford Nanopore Technologies (25–
77 32). All of these techniques have been used to recover genomes from metagenomes in
78 environments of varying complexity and recent technological breakthroughs deliver vastly
79 increased data yields, which makes even higher complexity samples tractable (12, 33–38). However,
80 long read sequencing technologies feature drawbacks of their own, as the PacBio platform is
81 relatively costly (39, 40), while Nanopore sequencing exhibits systematic errors in homopolymer
82 regions that can lead to insertions and deletions in the assemblies (41, 42).

83

Here, we assess the recent DNA sequencing technologies in the context of genome-centric, high-throughput metagenomics under realistic scenarios for a typical small research project. We compare metagenome assembly, binning, rRNA and protein gene recovery between single flow cell sequencing runs of Illumina MiSeq and Oxford Nanopore MinION R.9.4.1 and R.10.3 (referred to as R9 and R10 hereafter) and PacBio Sequel II (circular consensus sequencing).

Results and discussion

DNA of a microbial community from an anaerobic digester was sequenced via Illumina MiSeq (2x300 bp), PacBio CCS and the Oxford Nanopore MinION platform, using R9 and R10 chemistry flow cells (**Fig 1a**). The reads from different sequencing platforms were then assembled using Megahit for short reads and metaFlye for long reads. Despite being the same sample of extracted DNA, direct comparisons are difficult as the additionally size selection of the PacBio CCS dataset both increased the read length (**Fig 1b**) and altered the relative abundances of the species in the sample (**Fig S1a**). The Illumina data also features variation in relative abundances (**Fig S1b**), presumably due to GC bias (43). Only the R9 and R10 relative abundance data correlated well (**Fig S1c**), even though they had a 2.7-fold difference in sequencing yield (**Table 1**). Hence, the comparisons are influenced by differences in sequencing depth, read length, library preparation technique (44) and changes to relative abundances of the community members, but still represent a realistic scenario for a typical research project.

106 **Table 1. General statistics for sequenced reads and assembled contigs.**

Data type	Feature	Illumina	Nanopore R9	Nanopore R10	PacBio CCS
Reads	Total count	47,091,904	10,266,261	3,646,771	992,914
	Total yield (Mb)	13,285	35,237	13,009	15,309
	N50 (kb)	0.3	5.9	6.4	15.4
	Modal read accuracy (%)*	99.98	96.34	95.81	99.97
Contigs**	Total count	145,876	24,680	13,132	8,989
	Circular, > 0.5 Mb count	0	7	5	9
	Total size (Mb)	409	754	395	606
	N50 (kb)	3.5	79.9	71.7	172.5
Reads mapped to contigs (%)		88.1	93.5	93.2	95.2

107 *Derived from read Phred scores.
108 **Assembly metrics are presented after removing < 1 kb contigs.

109

110 To assist automated contig binning, we performed Illumina sequencing of 9 additional samples from
111 the same anaerobic digester spread over 9 years (**Supp Table 1**) and used the coverage profiles as
112 input for binning. Furthermore, to evaluate the impact of micro-diversity on MAG quality, we
113 calculated the polymorphic site rates for each MAG as a simple proxy for presence of micro-
114 diversity (10).

115

116 After performing automated contig binning it is evident that micro-diversity has a large impact on
117 MAG fragmentation, but that long-read sequencing results in much less fragmentation (**Fig 1c-d**) of
118 bins at higher amounts of micro-diversity (**Fig 2a**). Despite large differences in read length for
119 Nanopore and PacBio CCS data (N50 read length 6 kbp vs. 15 kbp), only small differences in bin
120 fragmentation were observed, as compared to the Illumina-based results.

121

122 After performing bin quality estimates, the most HQ MAGs were recovered by using Nanopore
123 sequencing on a R9 flow cell (n=86), supplemented with Illumina read polishing, while also featuring
124 the second lowest laboratory costs per recovered HQ MAG (**Table 2**). The second highest number
125 of HQ MAGs (n=74) was achieved via PacBio CCS, which also exhibited the least HQ MAG

126 fragmentation in terms median contigs per bin (n=9), although the overall sequencing costs per
127 HQ MAG recovered were greater, compared to other long read approaches.

128

129 **Table 2.** Results for automated contig binning. Genome quality was defined based on the MiMAG
130 (14) standard using CheckM (18): HQ MAGs are defined by completeness of >90 %, <5 %
131 contamination, at least 18 distinct tRNA genes and the 5S, 16S, 23S rRNA genes occurring at least
132 once. MQ MAGs by completeness >50 % and contamination <10 %, while low quality MAGs are
133 defined by completeness <50 % and contamination <10 %. MAGs with contamination estimates
134 higher than 10 % were classified as contaminated. *Fraction of the assembled contigs (> 1kb) that
135 were placed into bins. **Sequencing costs refers to the expenses encountered at the time of
136 conducting the experiments and may differ for other research groups, depending on agreements
137 with sequencing service providers.

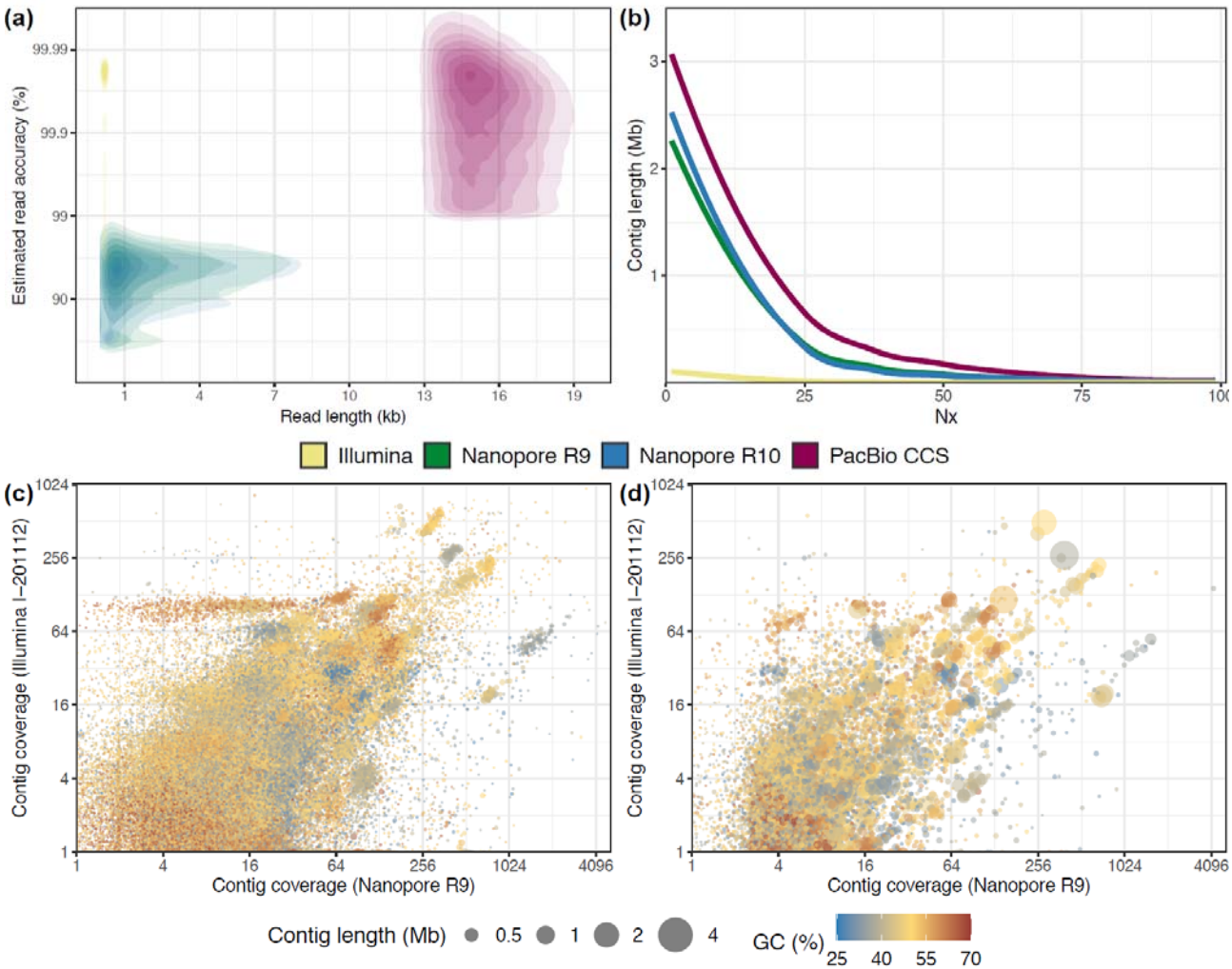
Feature	Illumina MiSeq	Nanopore R9	Nanopore R9 + Illumina	Nanopore R10	Nanopore R10 + Illumina	PacBio CCS
HQ MAGs	8	64	86	33	45	74
MQ MAGs	83	114	95	64	63	72
LQ MAGs	3	28	26	18	12	22
Contaminated MAGs	10	6	13	6	6	14
Reads binned (%)	76	86	85	86	86	83
Contigs binned (%)*	55	68	71	71	74	73
Contigs per HQ MAG (median)	184	15	16	15	18	9
Contigs binned in HQ MAGs (%)	4	22	30	22	29	34
Reads binned in HQ MAGs (%)	16	46	49	39	41	48
Sequencing costs ($)**	1,200	811	2,011	811	2,011	4,420
Cost per HQ MAG ($)	150	13	23	25	45	60

138

139 After further examining MAG quality estimates between the clustered bins (see Materials and
140 methods for clustering details) of different sequencing platforms, MAGs from PacBio CCS were
141 found to feature greater mean bin completeness estimates (**Fig 2b**) than the Illumina-only
142 approach (93.3 ± 6.2 vs. 89.3 ± 8.2, respectively, paired t-test p=0.0001). Furthermore, polishing of
143 Nanopore assemblies with Illumina reads was observed to slightly improve the mean bin
144 completeness for both R9 (from 91.9 ± 5.4 to 93.5 ± 4.7 %, p=0.0009) and R10 (from 91.9 ± 6.0 to
145 93.1 ± 5.9 %, p=0.021) chemistries. The observed increase in bin completeness from Illumina read

146 polishing is assumed to be caused by improved recovery of protein sequences (bin completeness
147 estimation is based on the presence of marker proteins (18)), due to correction of Nanopore
148 sequencing systematic errors, such as homopolymer truncation that can artificially induce
149 frameshifts, and thus introduce early stop codons in genes (45, 46).

150



151
152 **Figure 1. Overview of general sequencing results from short and long read platforms. a)** Read
153 length and estimated accuracy (inferred from Phred scores, log-scaled) distributions of the read
154 datasets. **b)** Contig Nx plot for the assembled metagenomes (after 1 kb length filtering). Differential
155 coverage plots are presented for **c)** Illumina and **d)** Nanopore R9 metagenome assemblies.

156

Although Illumina read polishing increases the overall completeness estimates of Nanopore-assembled bins, the improvements are case-specific, due to differences in coverage and homopolymer content and the degree of micro-diversity between MAGs. To investigate this further, PacBio CCS bins were assumed as "reference quality" and compared to bins from Nanopore sequencing with and without short read polishing (see Materials and Methods for selection criteria). Applying Illumina read polishing was found to reduce the rate of indels (by at least 20 %) in 18 of 18 Nanopore R9 bins and 14 of 15 R10 bins, independent of coverage (**Fig 2c**). Interestingly, the estimated reduction in mismatches (at least 20 %) was less pronounced, as it was observed in 6 of 18 Nanopore R9 bins and 8 of 15 for R10, presumably due to Nanopore R10 bins featuring overall lower long read coverage (**Fig 2d**). Hence, even though it is possible to acquire HQ MAGs from Nanopore-only assemblies, as demonstrated in this and other studies (47, 48), supplemental Illumina read polishing still aids in recovering more HQ MAGs (**Table 2**) from complex microbial communities as well as producing presumably less erroneous genome sequences.

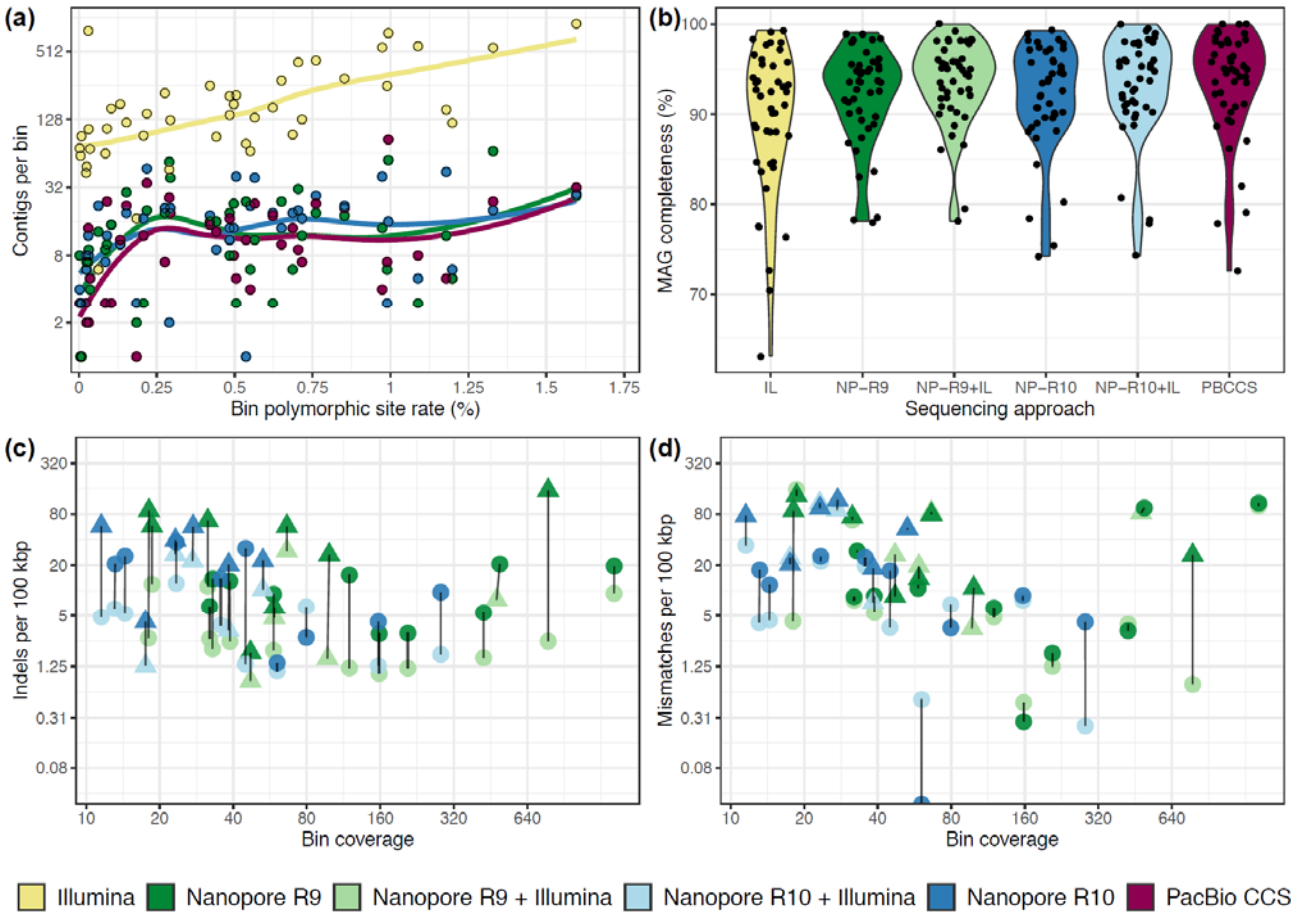


Figure 2. Comparison of bins from different sequencing approaches. a) MAG fragmentation (log-scaled) at different bin SNP rates in PacBio CCS MAGs**. b)** Genome bin completeness estimates for different sequencing platforms. IL — Illumina, NP — Nanopore, PBCCS — PacBio CCS. Bin **c)** indel and **d)** mismatch rates (log-scaled) for MAGs from Nanopore sequencing with and without Illumina read polishing, compared to MAGs from PacBio CCS. The presented bin coverage on the x axis (log-scaled) is for the corresponding Nanopore chemistry type. HQ MAGs are represented by circle, while triangles denote MQ MAGs. For all figures, only the bins that were clustered together between all the different sequencing platforms (see Materials and methods) are presented.

Reduction of indel rates in genomes is expected to yield less truncated protein sequences (49). To investigate this we used the IDEEL test (12) that estimates protein length by comparing to known sequences. In the MAGs using Nanopore R9 or R10 data alone, respectively 15 % and 14 % of protein sequences were estimated to exhibit a greater than 20 % truncation, while proteins from the hybrid Nanopore and Illumina approach were found to be of comparable length to that of PacBio CCS (3-4 %, **Table 3**). Furthermore, Illumina-only bins were observed to feature slightly increased estimated

186 protein truncation levels than PacBio CCS and the hybrid approach, although the following is
187 expected to be influenced by the generally shorter contig lengths, leading to a higher count of
188 fragmented protein sequences (50).

189

190 **Table 3. Protein sequence distribution (%) at different estimated truncation rates.**

Query-to-Subject length ratio intervals	Illumina	Nanopore R9	Nanopore R9 + Illumina	Nanopore R10	Nanopore R10 + Illumina	PacBio CCS
1.0-0.8	90.89	85.01	96.69	85.56	96.37	96.38
0.8-0.6	3.03	4.51	1.30	4.51	1.40	1.41
0.6-0.4	2.48	4.40	0.90	4.35	1.00	0.97
0.4-0.2	2.29	4.21	0.77	3.91	0.83	0.87
0.2-0.0	1.31	1.87	0.33	1.66	0.40	0.37

191

192 A known source of errors in Nanopore-based assemblies is systematic homopolymer miscalling.
193 Similarly to the IDEEL test, using MAGs from PacBio CCS as a reference, Nanopore-only MAGs were
194 found to feature increased homopolymer errors at longer lengths, especially for cytosine and
195 guanine homopolymers (**Fig S3**), which coincides with read-level error rates of Nanopore
196 sequencing (51). As expected, sequences from Nanopore R10 featured lower rates of homopolymer
197 miscalling than R9 and the hybrid assemblies featured reduced error profiles, similar to that of
198 Illumina-only assemblies.

199

200 To reiterate, although the benefits of the hybrid Nanopore and Illumina approach have been
201 reported in numerous projects (52–59), our study utilized PacBio CCS as a means of establishing
202 reference sequences for examining the effects of short read polishing on Nanopore bins that were
203 recovered from a complex microbial community. Given that short read length can be a cause for
204 mapping issues, often leading to errors and biases in genome recovery (60–62), we observed
205 evidence that Illumina read polishing of Nanopore bins still significantly reduces the estimated rate

206 of indels, especially homopolymer errors, in a high complexity metagenome. Furthermore, although
207 multiple bioinformatics workflows for frame-shift correction of Nanopore assemblies without short
208 reads have been developed (48, 63–65), the indel corrections are not *de novo*, as the workflows are
209 based on comparisons to reference databases, and hence erroneous sequences of novel genes
210 might not get corrected or proteins with biologically occurring frame-shift mutations might be
211 falsely converted to a full-length state, making the method suboptimal for characterising novel
212 microbial species. Nonetheless, using databases to correct Nanopore sequencing errors could still
213 be a useful and cost-effective approach, when *de novo* recovery of novel genomic sequences is not
214 relevant for the research project.
215
216 Lastly, the presence of repetitive DNA sequences, including genes in multiple copies, can lead to
217 contig breaks in the assembly (66). Ribosomal RNA genes were chosen as an example of repeat gene
218 as it is also the most widely used marker gene for amplicon sequencing, phylogenetic analysis and
219 FISH (67–69). Hence, sufficient recovery and binning of rRNA genes is an important aspect of high
220 throughput genome-centric metagenomics in order to connect MAGs to additional information.
221 Comparing 44 MAGs captured in all assemblies (see Materials and Methods), 13 out of 44 Illumina
222 bins featured all rRNA genes, while it was 40 for Nanopore R9, 41 for Nanopore R10 and 41 for
223 PacBio CCS (**Fig S4 a-c**). In the Illumina assembly, the ratio of 16S to 23S rRNA gene counts was
224 0.76 (**Fig S4 d-e**) indicating either collapsed or fragmented genes as a result of micro-diversity,
225 whereas for long-read technologies a comparable ratio of ~1 was observed (**Supp Table 2**).
226
227

Conclusion

Long read DNA sequencing improves assembly contiguity over short read assemblies and, as a result, produces significantly greater numbers of high-quality genomes from complex microbial communities. Recovery of repeated rRNA genes via long read sequencing is also vastly improved over the short read approach. Nevertheless, short read sequencing is still useful, as an economical way of acquiring time series data for binning MAGs and it can be combined with Nanopore sequencing to recover more high-quality genome bins and full protein sequences from metagenomes. The PacBio CCS platform features overall superior read metrics in terms of accuracy and length, although the relatively higher sequencing costs can act as a bottleneck for large scale genome-centric metagenomics projects. Despite the multiple differences between Nanopore and PacBio CCS platforms, recovering genomes from complex metagenomes via Nanopore R9 sequencing with supplemental short read Illumina polishing was, at the moment of writing, found to be the optimal strategy, balancing laboratory costs, adequate sequencing depth and minimal sequencing errors in the assembled genomes.

Materials and methods

Sampling

Sludge biomass was sampled from the anaerobic digester at Fredericia wastewater treatment plant (Latitude 55.552219, Longitude 9.722003) at multiple time points and stored as frozen 2 mL aliquots at -20°C.

DNA extraction

DNA was extracted from the anaerobic digester sludge using DNeasy PowerSoil Kit (QIAGEN, Germany) following the manufacturer's protocol. The extracted DNA was then size selected using the SRE XS (Circulomics, USA), according to the manufacturer's instructions.

DNA QC

DNA concentrations were determined using Qubit dsDNA HS kit and measured with a Qubit 3.0 fluorimeter (Thermo Fisher, USA). DNA size distribution was determined using an Agilent 2200 Tapestation system with genomic screentapes (Agilent Technologies, USA). DNA purity was determined using a NanoDrop One Spectrophotometer (Thermo Fisher, USA).

Nanopore DNA sequencing

Library preparation was carried out using the SQK-LSK109 kit (Oxford Nanopore Technologies, UK) following the manufacturer's protocol. The DNA libraries were sequenced using the R.9.4.1 and the R.10.3 MinION flowcells (Oxford Nanopore Technologies, UK) on a MinION Mk1B (Oxford Nanopore Technologies, UK) device. After sequencing, the MinION flowcells were washed using the Flow Cell Wash Kit (EXP-WSH002, Oxford Nanopore Technologies, UK) and the same library was loaded again to generate additional sequencing data.

Illumina DNA sequencing

The Illumina libraries were prepared using the Nextera DNA library preparation kit (Illumina, USA) following the manufacturer's protocol and sequenced using the Illumina MiSeq platform.

PacBio CCS

Size-selected DNA sample was sent out to the DNA Sequencing Center at Brigham Young University. The DNA sample was fragmented with Megaruptor (Diagenode, Belgium) to 15 kb and size-selected using the Blue Pippin (Sage Science, USA). The sample was then prepared using SMRTbell Express Template Preparation Kit 1.0 (PacBio, USA) according to manufacturers' instructions. Sequencing was performed on the Sequel II system (PacBio, USA) using the Sequel II Sequencing Kit 1.0 (PacBio, USA) with the Sequel II SMRT Cell 8M (PacBio, USA) for a 30 hour data collection time.

Data analysis

Read processing

Illumina reads were trimmed for adapters using Cutadapt (v. 1.16 (70)). The generated raw Nanopore data was basecalled in super-accurate mode with using Guppy (v. 5.0.7, https://community.nanoporetech.com/downloads) with dna_r9.4.1_450bps_sup.cfg model for R9 and dna_r10.3_450bps_sup.cfg model for R10 chemistry. Adapters for nanopore reads were removed using Porechop (v. 0.2.3 (71) and read with Phred quality scores below 7 were filtered out using Filtlong (v. 0.1.1, https://github.com/rrwick/Filtlong). The CCS tool (v. 6.0.0, https://ccs.how/) was utilized with the sub-read data from PacBio CCS to produce HiFi reads. Read statistics were acquired via NanoPlot (v. 1.24.0 (72))

Metagenome assembly and binning

Long reads were assembled using Flye (v. 2.9-b1768 (73)) with the "--meta" setting enabled and the "--nano-hq" option for assembling Nanopore reads, whereas "--pacbio-hifi" and "--min-overlap 7500 --read-error 0.01" options were used for assembling PacBio CCS reads, as it resulted in more HQ MAGs than using the default settings. Polishing tools for Nanopore-based assemblies: Minimap2 (v. 2.17 (74)), Racon (used thrice, v. 1.3.3 (75)), and Medaka (used twice, v. 1.4.1, https://github.com/nanoporetech/medaka). Nanopore assemblies were additionally polished with Illumina reads using Racon. The trimmed Illumina reads were also assembled using Megahit (v. 1.1.4 (76))

Automated binning was carried out using MetaBAT2 (v. 2.12.1 (77)), with "-s 500000" settings, MaxBin2 (v. 2.2.7) and Vamb (v. 3.0.2) with "-o C --minfasta 500000" settings. Contig coverage profiles from different sequencer data as well as 9 additional time series Illumina datasets of the same anaerobic digester were used for generating the bins. The binning output of different tools was then integrated and refined using DAS Tool (v. 1.1.2 (78)). CoverM (v. 0.6.1, https://github.com/wwood/CoverM) was applied to calculate the bin coverage ("-m mean" settings) and relative abundance ("-m relative_abundance") values.

Bin processing

The completeness and contamination of the genome bins was estimated using CheckM (v. 1.1.2 (18)). The bins were classified using GDTB-Tk (v. 1.5.0 (79), R202 database). Bin protein sequences were predicted using Prodigal (v. 2.6.3, https://github.com/hyattpd/Prodigal) with "-p meta" setting, while rRNA genes were predicted using Barrnap (v. 0.9, https://github.com/tseemann/barrnap) and tRNAscan-SE (v. 2.0.5 (80)) was used for tRNA

predictions. Bin quality was determined following the Genomic Standards Consortium guidelines, wherein a MAG of high quality featured genome completeness of more than 90 %, less than 5 % contamination, at least 18 distinct tRNA genes and the 5S, 16S, 23S rRNA genes occurring at least once (14). MAGS with completeness above 50 % and contamination below 10 % were classified as medium quality, while low quality MAGs featured completeness below 50 % and contamination below 10 %. MAGs with contamination estimates higher than 10 % were classified as contaminated.

Illumina reads were mapped to the metagenome assembly using Bowtie2 (v. 2.4.2 (81)) with the "--very-sensitive-local" setting. The mapping was converted to BAM and sorted using SAMtools (v. 1.9 (82)). Single nucleotide polymorphism rate was then calculated using CMseq (v. 1.0.3 (10)) from the mapping using poly.py script with "--mincov 10 --minqual 30" settings.

Bins were clustered using dRep (v. 2.6.2 (83)) with "-comp 50 -con 10 -sa 0.95" settings. Only the bins that featured higher coverage than 10 in their respective sequencing platform and a higher Illumina read coverage than 5 for bins from the hybrid approach were included in downstream analysis. For IDEEL test (12), the predicted protein sequences from clustered bins were searched against the UniProt TrEMBL (84) database (release 2021_01) using Diamond (v. 2.0.6 (85)). Query matches, which were not present in all datasets, were omitted to reduce noise.

QUAST (v. 4.6.3 (86)) was applied on the clustered bins with less than 0.5 % SNP rate to acquire mismatch and indels metrics. Cases with Quast parameters "Genome Fraction" of less than 75 % and "Unaligned length" of more than 250 kb were omitted to reduce noise. For homopolymer analysis, the clustered bins were mapped to each other using "asm5" mode of Minimap2 and

Counterr (v. 0.1, https://github.com/dayzerodx/counterr) was used on the mapping files to get homopolymer calling errors. For QUAST and Counterr, PacBio CCS bins were used as reference sequences.

Data availability

The sequencing data and bins are available at the ENA with bio project ID: PRJEB48021. The code, datasets used to generate the figures and additional material are available at https://github.com/Serka-M/Digester-MultiSequencing.

Acknowledgements

We would like to acknowledge the plant operators at Fredericia wastewater treatment plant for supplying the sample material.

Funding information

The study was funded by research grants from VILLUM FONDEN (15510) and the Poul Due Jensen Foundation (Microflora Danica).

Author contributions

MS performed DNA extraction and size selection. RHK did Illumina and Nanopore sequencing of the sample. RHK and MS performed the bioinformatics. MS, RHK and MA wrote the manuscript. SMK, TYM and EAS contributed to experiment design, result interpretation and writing of the manuscript. All authors reviewed the manuscript.

Conflict of interest

EAS, SMK, MA, RHK are also employed at DNASense ApS. TYM is employed at Lyras A/S. The remaining authors declare no conflict of interest.

References

1. Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences of the United States of America 113:5970–5.

2. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev V V, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43.

3. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome research 23:111–20.

4. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature biotechnology 31:533–8.

5. Wu Y, Tang Y, Tringe SG, Simmons BA, Singer SW. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:26.

6. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165.

7. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nature methods 11:1144–6.

379 8. Miller IIJ, Rees ER, Ross J, Miller IIJ, Baxa J, Lopera J, Kerby RL, Rey FE, Kwan JC. 2019.
380 Autometa: automated extraction of microbial genomes from individual shotgun
381 metagenomes. Nucleic acids research 47:e57.

382 9. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated
383 genomes of the global human gut microbiome. Nature 568:505–510.

384 10. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A,
385 Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C,
386 Segata N. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over
387 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell
388 176:649-662.e20.

389 11. Parks DH, Rinke C, Chuvochina M, Chaumeil P, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson
390 GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially
391 expands the tree of life. Nature microbiology 2:1533–1542.

392 12. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. 2019. Compendium of
393 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme
394 discovery. Nature biotechnology 37:953–961.

395 13. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH,
396 Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen M. 2021. Connecting structure
397 to function with the recovery of over 1000 high-quality metagenome-assembled genomes
398 from activated sludge using long-read sequencing. Nat Commun 12:2009.

399 14. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-smith M, Doud D, Reddy TBK, Schulz F,
400 Jarett J, Rivers AR, Eloe-fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED,
401 Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA,
402 Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema
403 TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon
404 KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A,
405 Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017.
406 Minimum information about a single amplified genome (MISAG) and a metagenome-
407 assembled genome (MIMAG) of bacteria and archaea. Nature biotechnology 35:725–731.

408 15. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete
409 domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 38:1079–1086.

410 16. Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, Dangl JL, Ivanova N,
411 Woyke T, Kyrpides N, Pati A. 2016. ProDeGe: a computational protocol for fully automated
412 decontamination of genomes. The ISME journal 10:269–72.

413 17. Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes
414 from long-read sequencing and assembly. Current Opinion in Microbiology 23:110–120.

415 18. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the
416 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
417 Genome research 25:1043–55.

418 19. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC,
419 Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of
420 domain Bacteria. Nature 523:208–11.

421 20. Sekiguchi Y, Ohashi A, Parks DH, Yamauchi T, Tyson GW, Hugenholtz P. 2015. First genomic
422 insights into members of a candidate bacterial phylum responsible for wastewater bulking.
423 PeerJ 3:e740.

424 21. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete
425 genomes from metagenomes. Genome Res 30:315–333.

426 22. Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing:
427 computational challenges and solutions. Nat Rev Genet 13:36–46.

428 23. Berbers B, Saltykova A, Garcia-Graells C, Philipp P, Arella F, Marchal K, Winand R, Vanneste
429 K, Roosens NHC, De Keersmaecker SCJ. 2020. Combining short and long read sequencing to
430 characterize antimicrobial resistance genes on plasmids applied to an unauthorized
431 genetically modified Bacillus. Scientific Reports 10:4310.

432 24. Warwick-dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB,
433 Temperton B. 2019. Long-read viral metagenomics captures abundant and microdiverse
434 viral populations and their niche-defining genomic islands. PeerJ 1–28.

435 25. Ng P, Tan JJS, Ooi HS, Lee YL, Chiu KP, Fullwood MJ, Srinivasan KG, Perbost C, Du L, Sung W-
436 K, Wei C-L, Ruan Y. 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for

437 the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic acids research
438 34:e84.

439 26. Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto GA, Foquet
440 M, Turner SW. 2008. Selective aluminum passivation for targeted immobilization of single
441 DNA polymerase molecules in zero-mode waveguide nanostructures. Proceedings of the
442 National Academy of Sciences of the United States of America 105:1176–81.

443 27. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT,
444 Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson
445 H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M.
446 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads.
447 Nature biotechnology 36:338–345.

448 28. Kasianowicz JJ, Brandin E, Branton D, Deamer DW. 1996. Characterization of individual
449 polynucleotide molecules using a membrane channel. Proceedings of the National Academy
450 of Sciences of the United States of America 93:13770–3.

451 29. McCoy RC, Taylor RW, Blauwkamp T a., Kelley JL, Kertesz M, Pushkarev D, Petrov D a.,
452 Fiston-Lavier A-S. 2014. Illumina TruSeq synthetic long-reads empower de novo assembly
453 and resolve complex, highly-repetitive transposable elements. PloS one 9:e106689.

454 30. Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, Sidow A, Dekas AE, Batzoglou S,
455 Bhatt AS. 2018. High-quality genome sequences of uncultured microbes by assembly of read
456 clouds. Nature biotechnology 36:1067–1080.

457 31. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. 2016. Synthetic long-read
458 sequencing reveals intraspecies diversity in the human microbiome. Nature biotechnology
459 34:64–9.

460 32. Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, Amirebrahimi M,
461 Thomas BC, Burstein D, Tringe SG, Williams KH, Banfield JF. 2015. Accurate, multi-kb reads
462 resolve complex populations and detect rare microorganisms. Genome research 25:534–43.

463 33. White RA, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, Fansler SJ,
464 Glaesemann KR, Glass K, Jansson JK. 2016. Moleculo Long-Read Sequencing Facilitates
465 Assembly and Genomic Binning from Complex Soil Metagenomes. mSystems 1:1–15.

466 34. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, Pope PB.
467 2016. Improved metagenome assemblies and taxonomic binning using long-read circular
468 consensus sequence data. Scientific reports 6:25373.

469 35. Daims H, Lebedeva E V, Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M,
470 Vierheilig J, Bulaev A, Kirkegaard RH, von Bergen M, Rattei T, Bendinger B, Nielsen PH,
471 Wagner M. 2015. Complete nitrification by Nitrospira bacteria. Nature 528:504–9.

472 36. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh
473 JY, Tong C, Ng OT, Barkham T, Young B, Marimuthu K, Chng KR, Sikic M, Nagarajan N. 2019.
474 Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants
475 and mobile elements in human microbiomes. Nature biotechnology 37:937–944.

476 37. Slaby BM, Hackl T, Horn H, Bayer K, Hentschel U. 2017. Metagenomic binning of a marine
477 sponge microbiome reveals unity in defense but metabolic specialization. The ISME journal
478 11:2465–2478.

479 38. Arumugam K, Bessarab I, Haryono MAS, Liu X, Zuniga–Montanez RE, Roy S, Qiu G, Drautz–
480 Moses DI, Law YY, Wuertz S, Lauro FM, Huson DH, Williams RBH. 2021. Recovery of
481 complete genomes and non-chromosomal replicons from activated sludge enrichment
482 microbial communities with long read metagenome sequencing. npj Biofilms Microbiomes
483 7:23.

484 39. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M.
485 2021. High-accuracy long-read amplicon sequences using unique molecular identifiers with
486 Nanopore or PacBio sequencing. Nat Methods 18:165–169.

487 40. Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. Genomics, Proteomics &
488 Bioinformatics 13:278–289.

489 41. Koren S, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2019. Reply to “Errors in long-read
490 assemblies can critically affect protein prediction”. Nature biotechnology 37:127–128.

491 42. Nicholls SM, Quick JC, Tang S, Loman NJ. 2019. Ultra-deep, long-read nanopore sequencing
492 of mock microbial community standards. GigaScience 8:1–9.

493 43. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, Rasmussen M, Zervas A,
494 Hansen LH. 2020. GC bias affects genomic and metagenomic reconstructions,
495 underrepresenting GC-poor organisms. Gigascience 9:giaa008.

496 44. Wick RR, Judd LM, Wyres KL, Holt KEY 2021. 2021. Recovery of small plasmid sequences via
497 Oxford Nanopore sequencing. Microbial Genomics 7:000631.

498 45. O'Donnell CR, Wang H, Dunbar WB. 2013. Error analysis of idealized nanopore sequencing.
499 Electrophoresis 34:2137–2144.

500 46. Scheunert A, Dorfner M, Lingl T, Oberprieler C. 2020. Can we use it? On the utility of de
501 novo and reference-based assembly of Nanopore data for plant plastome sequencing. PLoS
502 One 15.

503 47. Cusco A, Perez D, Viñes J, Francino O. 2020. Long-Read Metagenomics to Retrieve High-
504 Quality Metagenome-Assembled Genomes from Canine Feces. Research Square
505 https://doi.org/10.21203/rs.3.rs-60068/v1.

506 48. Huang Y-T, Liu P-Y, Shih P-W. 2021. Homopolish: a method for the removal of systematic
507 errors in nanopore sequencing by homologous polishing. Genome Biol 22:95.

508 49. Hu J, Ng PC. 2012. Predicting the effects of frameshifting indels. Genome Biol 13:R9.

509 50. Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. 2017.
510 De novo PacBio long-read and phased avian genome assemblies correct and add to
511 reference genes generated with intermediate and short reads. Gigascience 6:1–16.

512 51. Delahaye C, Nicolas J. 2021. Sequencing DNA with nanopores: Troubles and biases. PLoS
513 One 16:e0257521.

514 52. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome
515 assemblies from short and long sequencing reads. PLoS computational biology
516 13:e1005595.

517 53. Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. 2020. Inclusion of Oxford
518 Nanopore long reads improves all microbial and viral metagenome-assembled genomes
519 from a complex aquifer system. Environ Microbiol 22:4000–4013.

520 54. Liu L, Wang Y, Che Y, Chen Y, Xia Y, Luo R, Cheng SH, Zheng C, Zhang T. 2020. High-quality
521 bacterial genomes of a partial-nitritation/anammox system by an iterative hybrid assembly
522 method. Microbiome 8:155.

523 55. Karlsson E, Lärkeryd A, Sjödin A, Forsman M, Stenberg P. 2015. Scaffolding of a bacterial
524 genome using MinION nanopore sequencing. Scientific Reports 5:11996.

525 56. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, AbuOun M,
526 Stubberfield E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum
527 MF, Walker AS, Stoesser N, On Behalf Of The Rehab Consortium null. 2019. Comparison of
528 long-read sequencing technologies in the hybrid assembly of complex bacterial genomes.
529 Microb Genom 5.

530 57. Gan HM, Tan MH, Austin CM, Sherman CDH, Wong YT, Strugnell J, Gervis M, McPherson L,
531 Miller AD. 2019. Best Foot Forward: Nanopore Long Reads, Hybrid Meta-Assembly, and
532 Haplotig Purging Optimizes the First Genome Assembly for the Southern Hemisphere
533 Blacklip Abalone (Haliotis rubra). Front Genet 10.

534 58. Shin SC, Kim H, Lee JH, Kim H-W, Park J, Choi B-S, Lee S-C, Kim JH, Lee H, Kim S. 2019.
535 Nanopore sequencing reads improve assembly and gene annotation of the Parochlus
536 steinenii genome. Sci Rep 9:5095.

537 59. Haghshenas E, Asghari H, Stoye J, Chauve C, Hach F. 2020. HASLR: Fast Hybrid Assembly of
538 Long Reads. iScience 23.

539 60. Maguire F, Jia B, Gray KL, Lau WYV, Beiko RG, Brinkman FSL. 2020. Metagenome-assembled
540 genome binning methods with short reads disproportionately fail for plasmids and genomic
541 Islands. Microbial Genomics, 6:e000436.

542 61. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013.
543 Characterizing and measuring bias in sequence data. Genome Biology 14:R51.

544 62. Modlin SJ, Robinhold C, Morrissey C, Mitchell SN, Ramirez-Busby SM, Shmaya T, Valafar F.
545 2021. Exact mapping of Illumina blind spots in the Mycobacterium tuberculosis genome
546 reveals platform-wide and workflow-specific biases. Microbial Genomics 7.

547 63. Arumugam K, Bağcı C, Bessarab I, Beier S, Buchfink B, Górska A, Qiu G, Huson DH, Williams
548 RBH. 2019. Annotated bacterial chromosomes from frame-shift-corrected long-read
549 metagenomic data. Microbiome 7:61.

550 64. Huson DH, Albrecht B, Bağcı C, Bessarab I, Górska A, Jolic D, Williams RBH. 2018. MEGAN-LR:
551 new algorithms allow accurate binning and easy interactive exploration of metagenomic
552 long reads and contigs. Biology Direct 13:6.

553 65. Hackl T, Trigodet F, Eren AM, Biller SJ, Eppley JM, Luo E, Burger A, DeLong EF, Fischer MG.
554 2021. proovframe: frameshift-correction for long-read (meta)genomics. bioRxiv
555 2021.08.23.457338.

556 66. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg
557 M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. 2019. Tandem repeats
558 lead to sequence assembly errors and impose multi-level challenges for genome and protein
559 databases. Nucleic Acids Research 47:10994.

560 67. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. 2018. Retrieval
561 of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without
562 primer bias. Nat Biotechnol 36:190–195.

563 68. Amann RI, Krumholz L, Stahl DA. 1990. Fluorescent-oligonucleotide probing of whole cells
564 for determinative, phylogenetic, and environmental studies in microbiology. J Bacteriol
565 172:762–770.

566 69. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W,
567 Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic
568 frameworks. Nucleic Acids Res 42:D643–D648.

569 70. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
570 reads. EMBnet.journal 17:10.

571 71. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with
572 multiplex MinION sequencing. Microbial genomics 3:e000132.

573 72. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing
574 and processing long-read sequencing data. Bioinformatics 34:2666–2669.

575 73. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using
576 repeat graphs. Nature biotechnology 37:540–546.

577 74. Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long
578 sequences. Bioinformatics (Oxford, England) 32:2103–10.

579 75. Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly
580 from long uncorrected reads. Genome research 27:737–746.

581 76. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node
582 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
583 Bioinformatics (Oxford, England) 31:1674–6.

584 77. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive
585 binning algorithm for robust and efficient genome reconstruction from metagenome
586 assemblies. PeerJ 7:e7359.

587 78. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery
588 of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat
589 Microbiol 3:836–843.

590 79. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P.
591 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises
592 the tree of life. Nature biotechnology 36:996–1004.

593 80. Chan PP, Lowe TM. 2019. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences.
594 Methods Mol Biol 1962:1–14.

595 81. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods
596 9:357–9.

597 82. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
598 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
599 format and SAMtools. Bioinformatics (Oxford, England) 25:2078–9.

600 83. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic
601 comparisons that enables improved genome recovery from metagenomes through de-
602 replication. The ISME journal 11:2864–2868.

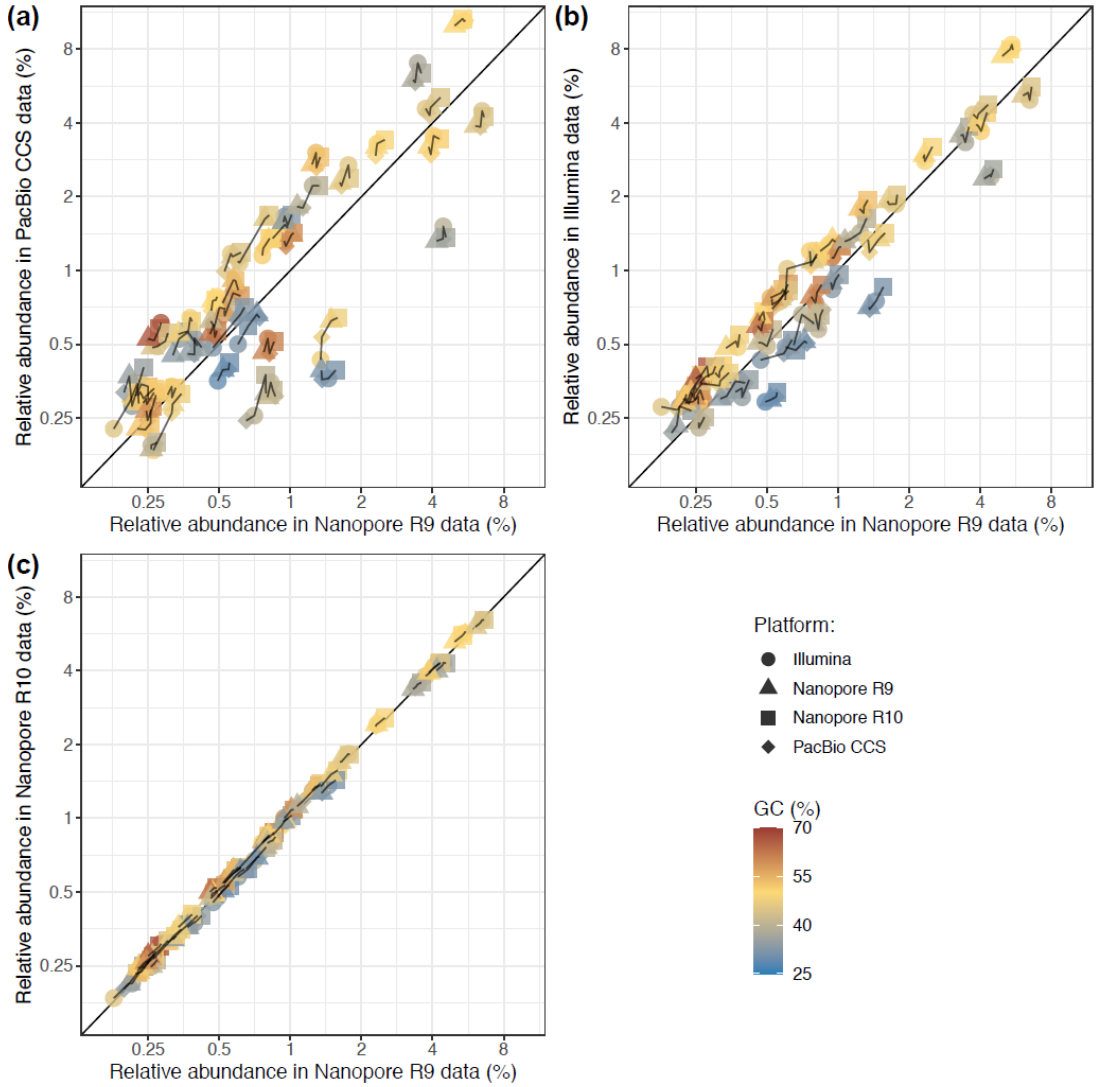
603 84. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169.

604 85. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND.
605 Nat Methods 12:59–60.

606 86. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for
607 genome assemblies. Bioinformatics 29:1072–1075.

608
609

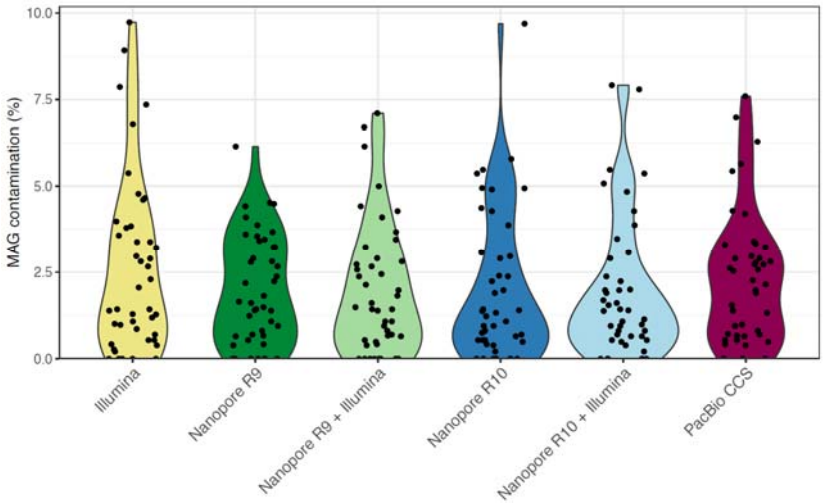
SUPPLEMENT



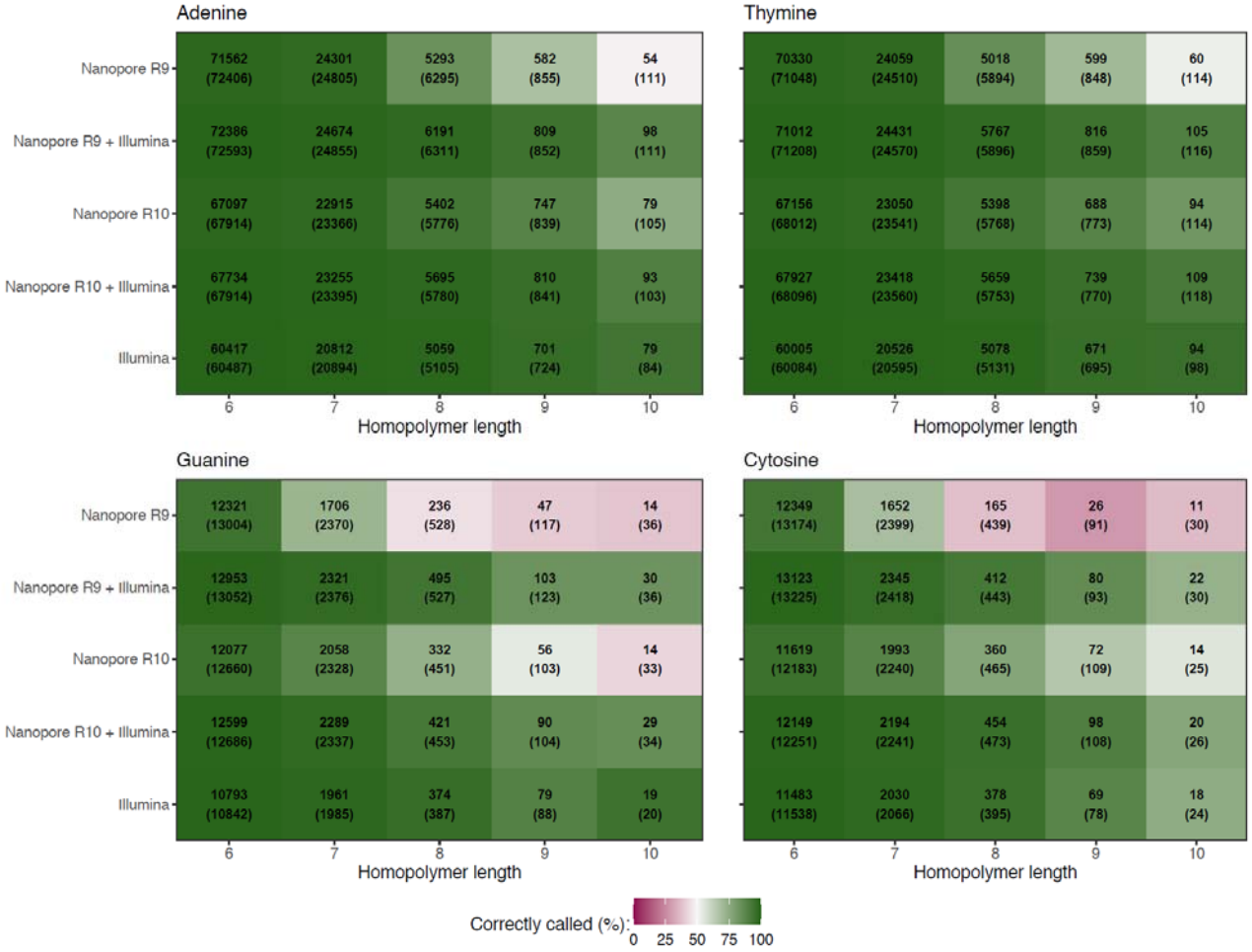
Supplemental Figure 1. Comparison of bin relative abundances between different sequencing platforms. Relative abundance values (log-scaled) are presented between the Nanopore R9 data and **a)** PacBio CCS, **b)** Illumina, **c)** Nanopore R10. Only the bins that were clustered together between different platforms are presented in the plots and are interlinked.

Supplemental Table 1. Overview of read datasets used in the study

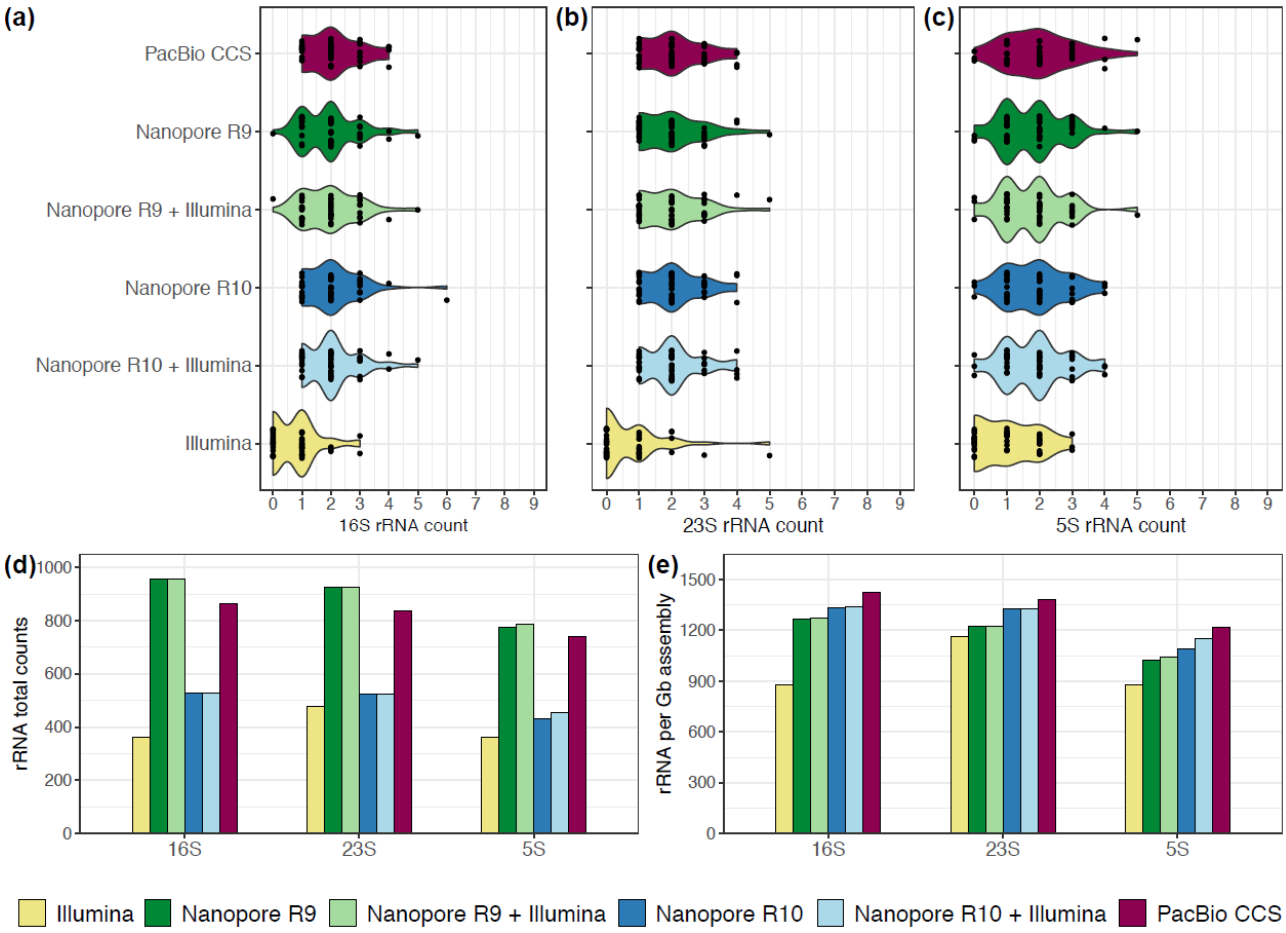
Read dataset	Instrument	Yield (Gb)	Read N50 (kb)	Read count	ENA sample ID
IL-201104	Illumina HiSeq	6.2	0.15	42,727,130	ERS7673063
IL-201112	Illumina HiSeq	11.4	0.15	79,619,634	ERS7673064
IL-201301	Illumina HiSeq	7.5	0.25	31,702,618	ERS7673065
IL-201308	Illumina HiSeq	6.7	0.25	28,067,586	ERS7673066
IL-201502	Illumina HiSeq	5.3	0.25	22,351,578	ERS7673067
IL-201702	Illumina HiSeq	15.9	0.25	66,225,442	ERS7673068
IL-201705	Illumina HiSeq	4.9	0.25	20,492,240	ERS7673069
IL-201707	Illumina HiSeq	5.5	0.25	23,663,146	ERS7673070
IL-201804	Illumina MiSeq	3.2	0.3	11,981,252	ERS7673071
IL-202001	Illumina MiSeq	13.3	0.3	47,091,904	ERS7673072
PB-202001	PacBio Sequel II	15.3	15.4	992,914	ERS7673073
R9-202001	MinION Mk1B	35.2	5.9	10,266,261	ERS7673074
R10-202001	MinION Mk1B	13.0	6.4	3,646,771	ERS7673075



Supplemental Figure 2. Bin contamination estimates. Only the bins that were clustered together between different platforms are presented in the plot.



Supplemental Figure 3. Homopolymer calling estimates in metagenomes from different sequencing platforms. Values in the heatmap show observed homopolymer counts estimated to be called correctly at given sequence length. The total count of homopolymers (called correctly and incorrectly) are in brackets. Only the contigs for bins that were clustered together between different platforms were used to generate values for the plot.



Supplemental Figure 4. Recovery of rRNA genes from complex microbial communities via different sequencing techniques. Distribution of rRNA gene counts, found in MAGs from different sequencing methods, are presented for **a)** 16S, **b)** 23S and **c)** 5S rRNA genes. Only the MAGs, which clustered together between all the different sequencing approaches, were included in the plots. **d)** Counts for total rRNA sequences that were recovered from the assembled metagenomes. **e)** Recovered rRNA gene counts, normalized to assembly size.

Supplemental Table 2. Statistics for rRNA sequence recovery from metagenome assemblies.

Feature	Illumina	Nanopore R9	Nanopore R10	PacBio CCS
Sequencing yield (Gb)	13.3	35.2	13.0	15.3
Assembly size (Gb)	0.41	0.75	0.40	0.61
16S rRNA count	360	958	528	862
16S RNA per Gb yield	27.1	27.2	40.6	56.3
16S RNA per Gb assembly	878	1,277	1,320	1,413
23S rRNA count	476	924	524	838
23S RNA per Gb yield	35.8	26.2	40.3	54.7
23S RNA per Gb assembly	1,160	1,232	1,310	1,373
5S rRNA count	360	774	432	704
5S RNA per Gb yield	27.1	22.0	33.2	46.0
5S RNA per Gb assembly	878	1,032	1,080	1,154