

1 **Oxford Nanopore R10.4 long-read sequencing enables near-perfect**
2 **bacterial genomes from pure cultures and metagenomes without**
3 **short-read or reference polishing**

4 Mantas Sereika^{a*}, Rasmus Hansen Kirkegaard^{a,b*}, Søren Michael Karst^a, Thomas Yssing
5 Michaelsen^a, Emil Aarre Sørensen^a, Rasmus Dam Wollenberg^c and Mads Albertsen^{a**}

6 ^aCenter for microbial communities, Aalborg University, Denmark

7 ^bJoint Microbiome Facility, University of Vienna, Austria

8 ^cDNASe ApS, Denmark

9 *These authors contributed equally to the paper

10 **Corresponding author ma@bio.aau.dk

11

12 **ABSTRACT**

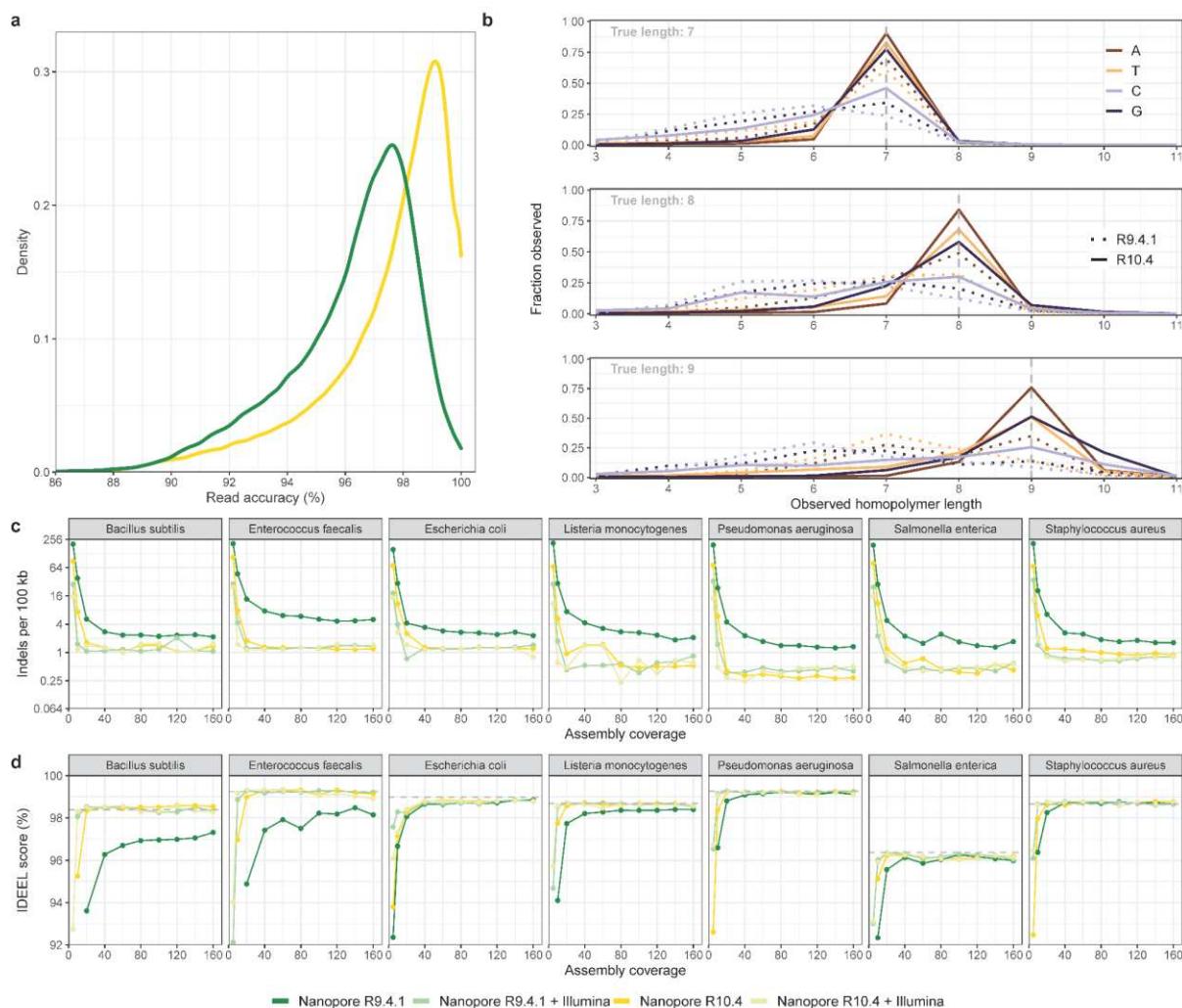
13 **Long-read Oxford Nanopore sequencing has democratized microbial genome**
14 **sequencing and enables the recovery of highly contiguous microbial genomes from**
15 **isolates or metagenomes. However, to obtain near-perfect genomes it has been**
16 **necessary to include short-read polishing to correct insertions and deletions derived**
17 **from homopolymer regions. Here, we show that Oxford Nanopore R10.4 can be used to**
18 **generate near-perfect microbial genomes from isolates or metagenomes without short-**
19 **read or reference polishing.**

20 **MAIN TEXT**

21 Bacteria live in almost every environment on Earth and the global microbial diversity is
22 estimated to entail more than 10^{12} species¹. To obtain representative genomes, sequencing
23 of pure cultures or genome recovery directly from metagenomes are often employed²⁻⁴. High-
24 throughput short-read sequencing has for many years been the method of choice^{5,6} but fails
25 to resolve repeat regions larger than the insert size of the library⁷. This is especially
26 problematic in metagenome samples where related species or strains often contain long
27 sequences of near-identical DNA. More recently, long-read sequencing has emerged as the
28 method of choice for both pure culture genomes^{8,9} and metagenomes¹⁰⁻¹². PacBio HiFi reads

29 combine low error rates with relatively long reads and generate near-perfect microbial
30 genomes from pure cultures or metagenomes^{13–15}. Despite very high-quality raw data, the
31 relatively high cost pr. base remains an economic hindrance for many research projects. A
32 widely used alternative is Oxford Nanopore sequencing which offers low-cost long-read data.
33 However, numerous studies have shown that despite vast improvements in raw error rates,
34 assembly consensus sequences still suffer from insertion and deletions in homopolymers that
35 often cause frameshift errors during gene calling^{16–18}. A commonly adopted solution has been
36 to include short-read data for post-assembly error correction^{12,19}, although it increases the cost
37 and complexity overhead. Another solution has been to apply reference-based polishing to
38 correct frameshift errors^{20–22}, but while it provides a practical solution, which allows gene
39 calling, it does not provide true near-perfect genomes.

40 We first evaluated the ability for Oxford Nanopore R9.4.1 and R10.4 data to obtain near-
41 perfect microbial genomes through sequencing of the ZymoBIOMICS HMW DNA Standard
42 #D6322 (Zymo mock) consisting of 7 bacterial species and 1 fungus. A single PromethION
43 R10.4 flowcell generated 52.3 gbp of data with a modal read accuracy of 99 % (**Figure 1A**,
44 **Table S1**). In contrast to R9.4.1 data, we do not see any significant improvement in assembly
45 quality for R10.4 by the addition of Illumina polishing (**Figure 1C, Figure S1**). This indicates
46 that near-perfect microbial reference genomes can be obtained from R10.4 data alone at a
47 coverage of approximately 40x. The improvement in assembly accuracy from R9.4.1 to R10.4
48 is largely due to an improved ability to call homopolymers, as R10.4 is able to correctly call
49 the length of the majority of homopolymers up to a length of 10 (**Figure 1B, Figure S2-3**). In
50 general, a homopolymer length of more than 10 is very rare in bacteria, with an estimate of
51 less than 10 per species on average¹⁸.



52

53 **Figure 1:** Sequencing and assembly statistics for the Zymo mock. **A)** Observed raw read
54 accuracies measured through read-mapping. **B)** Observed homopolymer length of raw reads
55 compared to the reference genomes (see **Figure S2-3** for a complete overview). **C)** Observed
56 indels of de novo assemblies per 100 kbp at different coverage levels, with and without Illumina
57 polishing. Note that the reference genomes available for the Zymo mock are not identical to
58 the sequenced strains (**Table S3**). **D)** IDEEL²³ score calculated as the proportion of predicted
59 proteins which are $\geq 95\%$ the length of their best-matching known protein in a database¹⁶. The
60 dotted line represents the IDEEL score for the reference genome.

61 To assess the performance of state-of-the-art sequencing technologies in recovering near-
62 perfect microbial genomes from metagenomes we sequenced activated sludge from an
63 anaerobic digester using single runs of Illumina MiSeq 2x300 bp, PacBio HiFi, and Oxford
64 Nanopore R9.4.1 and R10.4. Despite being the same sample, direct comparisons are difficult
65 as the additional size selection of the PacBio CCS dataset both increased the read length
66 (**Figure S4**) and altered the relative abundances of the species in the sample (**Figure S5**).
67 Furthermore, Nanopore R9.4.1 produced more than twice the amount of data compared to the

68 other datasets, while the Illumina data featured variations in relative abundances presumably
69 due to GC bias (**Figure S5**). To assist automated contig binning, we performed Illumina
70 sequencing of 9 additional samples from the same anaerobic digester spread over 9 years
71 (**Table S2**) and used the coverage profiles as input for binning using multiple different
72 approaches. Furthermore, to evaluate the impact of micro-diversity on MAG quality, we
73 calculated the polymorphic site rates for each MAG as a simple proxy for the presence of
74 micro-diversity⁶.

75 After performing automated contig binning it is evident that micro-diversity has a
76 large impact on MAG fragmentation, but that long-read sequencing data results in much less
77 fragmentation of bins at higher amounts of micro-diversity (**Figure S6**). Despite large
78 differences in read length for Nanopore and PacBio CCS data (N50 read length 6 kbp vs. 15
79 kbp), only small differences in bin fragmentation were observed, as compared to the
80 Illumina-based results (**Table 1, Figures S6**).

81

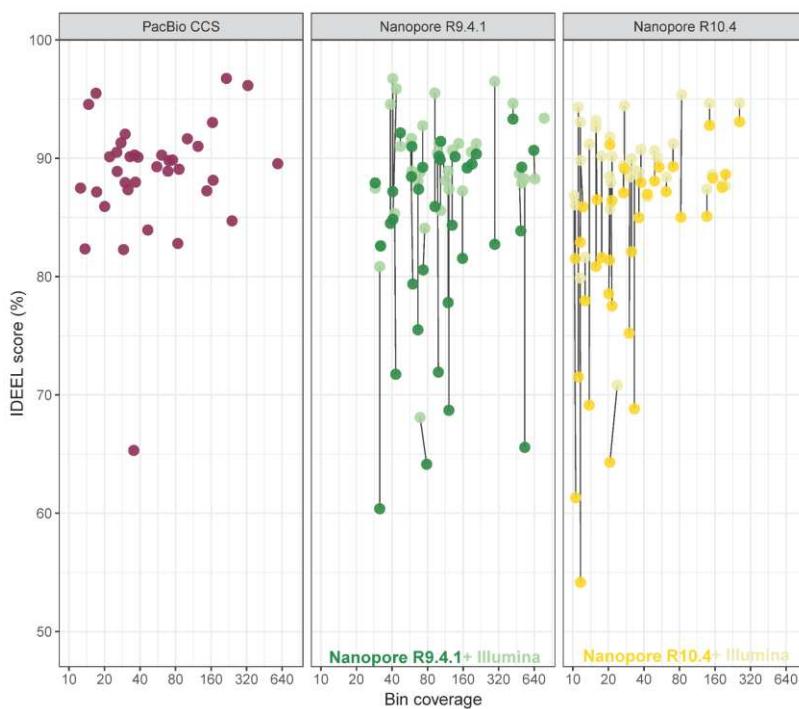
82 **Table 1:** Sequencing and assembly statistics for the anaerobic digester sample using
83 different technologies and approaches. *Costs refer to the expenses encountered at the time
84 of conducting the experiments and may differ for other research groups.

| | Illumina MiSeq | R9.4.1 / +Illumina | R10.4 / +Illumina | PacBio HiFi |
|---|-------------------|-----------------------|----------------------|-------------|
| Total Yield (Gbp) | 13 | 35 | 14 | 15 |
| Read N50 (kbp) | 0.3 | 5.9 | 5.6 | 15.4 |
| Observed modal read accuracy (%) | 100 | 96.76 | 98.21 | 99.86 |
| Assembly size (Mbp) | 409 | 754 | 379 | 606 |
| Contigs (> 1kbp) | 145,976 | 24,680 | 21,585 | 8,989 |
| Circular contigs (> 0.5 Mbp) | 0 | 7 | 3 | 9 |
| Contig N50 (kbp) | 3.5 | 79.9 | 40.1 | 172.5 |
| Reads mapped to contigs (%) | 88.1 | 93.5 | 95.4 | 95.2 |
| HQ MAGs | 8 | 64/86 | 34/36 | 74 |
| MQ MAGs | 83 | 114/95 | 65/67 | 72 |
| Contigs pr. HQ MAG (median) | 184 | 15/16 | 21/21 | 9 |
| Mapped reads in HQ MAGs (%) | 16 | 46/49 | 39/40 | 48 |
| Costs (\$)* | 1,200 | 811/2,011 | 811/2,011 | 4,420 |
| Cost per HQ MAG (\$) | 150 | 13/23 | 24/56 | 60 |

85

86 All long-read methods produce high numbers of high-quality (HQ) MAGs, which capture 39-
87 49% of all reads (**Table 1**). Nanopore R9.4.1 is able to produce HQ MAGs as a standalone
88 technology, but Illumina polishing increases the number of HQ MAGs from 64 to 86. For
89 Nanopore R10.4, Illumina polishing increases the number of HQ MAGs from 34 to 36. Using
90 the IDEEL test (**Figure 2**), it can be seen that Illumina polishing results in minor
91 improvements for Nanopore R10.4 above a coverage of 40, and that the Nanopore R10.4 is
92 in the same IDEEL range as PacBio HiFi MAGs. As with sequencing of the Zymo mock, the
93 difference from R9.4.1 to R10.4 is largely due to significantly better accuracy in
94 homopolymers for lengths up to 10 (**Figure S7**).

95
96



97
98 **Figure 2:** IDEEL score vs. coverage for metagenome bins from the anaerobic digester
99 sample. The Nanopore bins are shown with and without Illumina polishing connected by a
100 line.

101

102 Since its introduction as an early access program in 2014 Oxford Nanopore
103 sequencing technology has democratized sequencing and enabled every laboratory and
104 classroom to engage in microbial genome sequencing. However, for the generation of high-
105 quality genomes, additional short-read polishing has been essential, as indels in
106 homopolymer regions cause fragmented gene calls. The additional sequencing requirements
107 have been one of the barriers to widespread uptake. Here we show that Oxford Nanopore
108 R10.4 enables the generation of near-perfect microbial genomes from pure cultures or

109 metagenomes at coverages of 40x without short-read polishing. While homopolymers of 10
110 or more bases will likely still be problematic, they constitute a minor part of microbial
111 genomes.

112 For genome-recovery from metagenomes, low-coverage bins (<40X) do need
113 Illumina polishing to attain quality comparable to PacBio HiFi. Hence, in some cases, the
114 most economic option could be Nanopore R9.4.1 supplemented with short-read sequencing,
115 as the throughput is currently at least 2 times higher on R9.4.1 compared to R10.4 and no
116 difference is seen between the methods after Illumina short-read polishing.

117

118 **Data availability**

119 Anaerobic digester sequencing data are available at the ENA with bio project ID
120 PRJEB48021, while the Zymo mock community sequencing data is available at
121 PRJEB48692. The code and datasets used to generate the figures and supplementary
122 material are available at <https://github.com/Serka-M/Digester-MultiSequencing>.

123

124 **Acknowledgments**

125 We would like to acknowledge the plant operators at Fredericia wastewater treatment plant
126 for supplying the sample material. The study was funded by research grants from VILLUM
127 FONDEN (15510) and the Poul Due Jensen Foundation (Microflora Danica).

128

129 **Author contributions**

130 MS and RHK performed DNA extraction, and sequencing of the anaerobic digester and
131 selected Zymo mock samples. RWO prepared and sequenced the Zymo mock using R9.4.1
132 and Illumina. MS, RHK, and MA wrote the first draft of the manuscript. SMK, TYM, RWO,
133 and EAS contributed to experiment design, result interpretation, and writing of the
134 manuscript. All authors reviewed the manuscript.

135

136 **Conflict of interest**

137 EAS, SMK, MA, RHK, and RWO are employed at DNAsense ApS that consults and
138 performs sequencing. The remaining authors declare no conflict of interest.

139

140

141

142 **Materials and methods**

143

144 **Sampling**

145 Sludge biomass was sampled from the anaerobic digester at Fredericia wastewater
146 treatment plant (Latitude 55.552219, Longitude 9.722003) at multiple time points and stored
147 as frozen 2 mL aliquots at -20°C. For the Zymo sample, the ZymoBIOMICS HMW DNA
148 Standard #D6322 (Zymo Research, USA) was used.

149

150 **DNA extraction**

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

155

156 **DNA QC**

157 DNA concentrations were determined using Qubit dsDNA HS kit and measured with a Qubit
158 3.0 fluorimeter (Thermo Fisher, USA). DNA size distribution was determined using an Agilent
159 2200 Tapestation system with genomic screentapes (Agilent Technologies, USA). DNA
160 purity was determined using a NanoDrop One Spectrophotometer (Thermo Fisher, USA).

161

162 **Oxford Nanopore DNA sequencing**

163 Library preparation was carried out using the ligation sequencing kits (Oxford Nanopore
164 Technologies, UK) SQK-LSK109 and SQK-LSK112 for sequencing on R.9.4.1 and the
165 R.10.4 flowcells, respectively. Anaerobic digester and Zymo R.9.4.1 datasets were
166 generated on a MinION Mk1B (Oxford Nanopore Technologies, UK) device, while Zymo
167 R10.4 dataset was produced on a PromethION and digester R10.4 read sequences were
168 generated on a GridION.

169

170 **Illumina DNA sequencing**

171 The anaerobic digester Illumina libraries were prepared using the Nextera DNA library
172 preparation kit (Illumina, USA), while the Zymo Mock sample was prepared with NEB Next
173 Ultra II DNA library prep kit for Illumina (New England Biolabs, USA) following the
174 manufacturer's protocols and sequenced using the Illumina MiSeq platform.

175

176

177 **PacBio HiFi**

178 A size-selected DNA sample was sent to the DNA Sequencing Center at Brigham Young
179 University, USA. The DNA sample was fragmented with Megaruptor (Diagenode, Belgium)
180 to 15 kb and size-selected using the Blue Pippin (Sage Science, USA) and prepared for
181 sequencing using SMRTbell Express Template Preparation Kit 1.0 (PacBio, USA) according
182 to manufacturers' instructions. Sequencing was performed on the Sequel II system (PacBio,
183 USA) using the Sequel II Sequencing Kit 1.0 (PacBio, USA) with the Sequel II SMRT Cell
184 8M (PacBio, USA) for a 30 hour data collection time.

185

186 **Read processing**

187 Illumina reads were trimmed for adapters using Cutadapt v. 1.16²⁴. The generated raw
188 Nanopore data was basecalled in super-accurate mode with using Guppy v. 5.0.16
189 (<https://community.nanoporetech.com/downloads>) with dna_r9.4.1_450bps_sup.cfg model
190 for R9.4.1 and dna_r10.4_e8.1_sup.cfg model for R10.4 chemistry. Concatemers in R10.4
191 data were split by using “split_on_adapter” command (5 iterations) of duplex-tools v. 0.2.5
192 (<https://github.com/nanoporetech/duplex-tools>). Adapters for Nanopore reads were removed
193 using Porechop v. 0.2.3²⁵ and reads with Phred quality scores below 7 and 10 for R9.4.1 and
194 R10.4 reads, respectively, were removed using NanoFilt v. 2.6.0²⁶. The CCS tool v. 6.0.0
195 (<https://ccs.how/>) was utilized with the sub-read data from PacBio CCS to produce HiFi
196 reads. Read statistics were acquired via NanoPlot v. 1.24.0²⁶. Zymo read datasets were
197 subsampled to custom coverage profiles using Rasusa v. 0.3.0
198 (<https://github.com/mbhall88/rasusa>). CounterR v. 0.1 (<https://github.com/dayzerodx/counterR>)
199 was used to assess homopolymer calling in reads.

200

201 **Read assembly and binning**

202 Long reads were assembled using Flye v. 2.9-b1768^{13,27} with the “--meta” setting enabled
203 and the “--nano-hq” option for assembling Nanopore reads, whereas “--pacbio-hifi” and “--
204 min-overlap 7500 --read-error 0.01” options were used for assembling PacBio CCS reads,
205 as it resulted in more HQ MAGs than using the default settings. Polishing tools for
206 Nanopore-based assemblies: Minimap2 v. 2.17²⁸, Racon v. 1.3.3 (used thrice)²⁹, and
207 Medaka v. 1.4.4 (used twice, <https://github.com/nanoporetech/medaka>). The trimmed
208 Illumina reads were assembled using Megahit v. 1.1.4³⁰.

209

210 Automated binning was carried out using MetaBAT2 v. 2.12.1³¹, with “-s 500000” settings,
211 MaxBin2 v. 2.2.7³² and Vamb v. 3.0.2³³ with “-o C --minfasta 500000” settings. Contig
212 coverage profiles from different sequencer data as well as 9 additional time-series Illumina
213 datasets of the same anaerobic digester were used for generating the bins. The binning
214 output of different tools was then integrated and refined using DAS Tool v. 1.1.2³⁴. CoverM

215 v. 0.6.1 (<https://github.com/wwood/CoverM>) was applied to calculate the bin coverage (“-m
216 mean” settings) and relative abundance (“-m relative_abundance”) values.

217

218 **Assembly processing**

219 The completeness and contamination of the genome bins were estimated using CheckM v.
220 1.1.2³⁵. The bins were classified using GDTB-Tk v. 1.5.0³⁶, R202 database. Protein
221 sequences were predicted using Prodigal v. 2.6.3³⁷ with “p meta” setting, while rRNA genes
222 were predicted using Barrnap v. 0.9 (<https://github.com/tseemann/barrnap>) and tRNAscan-
223 SE v. 2.0.5³⁸ was used for tRNA predictions. Bin quality was determined following the
224 Genomic Standards Consortium guidelines, wherein a MAG of high quality featured genome
225 completeness of more than 90 %, less than 5 % contamination, at least 18 distinct tRNA
226 genes and the 5S, 16S, 23S rRNA genes occurring at least once³⁹. MAGS with
227 completeness above 50 % and contamination below 10 % were classified as medium
228 quality, while low quality MAGs featured completeness below 50 % and contamination below
229 10 %. MAGs with contamination estimates higher than 10 % were classified as
230 contaminated.

231

232 Illumina reads were mapped to the assemblies using Bowtie2 v. 2.4.2⁴⁰ with the “--very-
233 sensitive-local” setting. The mapping was converted to BAM and sorted using SAMtools v.
234 1.9⁴¹. Single nucleotide polymorphism rate was then calculated using CMseq v. 1.0.3⁶ from
235 the mapping using poly.py script with “--mincov 10 --minqual 30” settings.

236

237 Bins were clustered using dRep v. 2.6.2⁴² with “-comp 50 -con 10 -sa 0.95” settings. Only the
238 bins that featured higher coverage than 10 in their respective sequencing platform and a
239 higher Illumina read coverage than 5 for bins from the hybrid approach were included in
240 downstream analysis. For IDEEL test^{17,23}, the predicted protein sequences from clustered
241 bins and Zymo assemblies were searched against the UniProt TrEMBL⁴³ database (release
242 2021_01) using Diamond v. 2.0.6⁴⁴. Query matches, which were not present in all datasets,
243 were omitted to reduce noise. The IDEEL scores were assigned as described by¹⁶.

244

245 QUAST v. 4.6.3⁴⁵ was applied on the Zymo assemblies and the clustered bins with less than
246 0.5 % SNP rate to acquire mismatch and indels metrics. Cases with Quast parameters
247 “Genome Fraction” of less than 75 % and “Unaligned length” of more than 250 kb were
248 omitted to reduce noise. For homopolymer analysis, the clustered bins were mapped to each
249 other using “asm5” mode of Minimap2 and Counterr was used on the mapping files to get
250 homopolymer calling errors. For QUAST and Counterr, PacBio CCS bins were used as
251 reference sequences. FastANI v. 1.33⁴⁶ was used to calculate identity scores between Zymo

252 assemblies and the Zymo reference sequences. The Zymo mock reference genome
253 sequences were obtained from a link in the accompanying instruction manual to the
254 ZymoBIOMICS HMW DNA Standard Catalog No. D6332 at
255 <https://s3.amazonaws.com/zymo-files/BioPool/D6322.refseq.zip>.

256 **References**

- 257 1. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5970–5975 (2016).
- 258 2. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
- 259 3. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* **23**, 260 111–120 (2013).
- 261 4. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by 262 differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 263 (2013).
- 264 5. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–265 509 (2020).
- 266 6. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 267 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 268 **176**, 649–662.e20 (2019).
- 269 7. Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes 270 from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **23**, 110–120 (2015).
- 271 8. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read 272 SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
- 273 9. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de 274 novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
- 275 10. Sharon, I. *et al.* Accurate, multi-kb reads resolve complex populations and detect rare 276 microorganisms. *Genome Res.* **25**, 534–543 (2015).
- 277 11. Frank, J. A. *et al.* Improved metagenome assemblies and taxonomic binning using long- 278 read circular consensus sequence data. *Sci. Rep.* **6**, 25373 (2016).
- 279 12. Singleton, C. M. *et al.* Connecting structure to function with the recovery of over 1000 280 281 282

- 283 high-quality metagenome-assembled genomes from activated sludge using long-read
284 sequencing. *Nat. Commun.* **12**, 2009 (2021).
- 285 13. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat
286 graphs. *Nat. Methods* **17**, 1103–1110 (2020).
- 287 14. Bickhart, D. M. *et al.* Generation of lineage-resolved complete metagenome-assembled
288 genomes by precision phasing. *bioRxiv* 2021.05.04.442591 (2021)
289 doi:10.1101/2021.05.04.442591.
- 290 15. Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long
291 reads with hifiasm-meta. *arXiv [q-bio.GN]* (2021).
- 292 16. Wick, R. R. *et al.* Trycycler: consensus long-read assemblies for bacterial genomes.
293 *Genome Biol.* **22**, 266 (2021).
- 294 17. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein
295 prediction. *Nature biotechnology* vol. 37 124–126 (2019).
- 296 18. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases.
297 *PLoS One* **16**, e0257521 (2021).
- 298 19. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial
299 genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**,
300 e1005595 (2017).
- 301 20. Hackl, T. *et al.* proovframe: frameshift-correction for long-read (meta)genomics. *bioRxiv*
302 2021.08.23.457338 (2021) doi:10.1101/2021.08.23.457338.
- 303 21. Arumugam, K. *et al.* Annotated bacterial chromosomes from frame-shift-corrected long-
304 read metagenomic data. *Microbiome* **7**, 61 (2019).
- 305 22. Huang, Y.-T., Liu, P.-Y. & Shih, P.-W. Homopolish: a method for the removal of
306 systematic errors in nanopore sequencing by homologous polishing. *Genome Biol.* **22**,
307 95 (2021).
- 308 23. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing
309 of the cow rumen. *Nat. Commun.* **9**, 870 (2018).
- 310 24. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing

- 311 reads. *EMBnet.journal* **17**, 10–12 (2011).
- 312 25. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome
313 assemblies with multiplex MinION sequencing. *Microb Genom* **3**, e000132 (2017).
- 314 26. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack:
315 visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669
316 (2018).
- 317 27. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads
318 using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
- 319 28. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
320 3100 (2018).
- 321 29. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome
322 assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
- 323 30. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-
324 node solution for large and complex metagenomics assembly via succinct de Bruijn
325 graph. *Bioinformatics* **31**, 1674–1676 (2015).
- 326 31. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient
327 genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
- 328 32. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm
329 to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607
330 (2016).
- 331 33. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational
332 autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
- 333 34. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication,
334 aggregation and scoring strategy. *Nat Microbiol* **3**, 836–843 (2018).
- 335 35. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
336 assessing the quality of microbial genomes recovered from isolates, single cells, and
337 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
- 338 36. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny

- 339 substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
- 340 37. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
341 identification. *BMC Bioinformatics* **11**, 119 (2010).
- 342 38. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic
343 Sequences. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 1–14
344 (Springer New York, 2019).
- 345 39. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG)
346 and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat.*
347 *Biotechnol.* **35**, 725–731 (2017).
- 348 40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*
349 *Methods* **9**, 357–359 (2012).
- 350 41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
351 2078–2079 (2009).
- 352 42. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate
353 genomic comparisons that enables improved genome recovery from metagenomes
354 through de-replication. *ISME J.* **11**, 2864–2868 (2017).
- 355 43. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids*
356 *Res.* **45**, D158–D169 (2017).
- 357 44. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
358 DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
- 359 45. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for
360 genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
- 361 46. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
362 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
363 *Nat. Commun.* **9**, 5114 (2018).

Supplementary information for

Oxford Nanopore R10.4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing

Mantas Sereika^{a*}, Rasmus Hansen Kirkegaard^{a,b*}, Søren Michael Karst^a, Thomas Yssing Michaelsen^a, Emil Aarre Sørensen^a, Rasmus Dam Wollenberg^c and Mads Albertsen^{a**}

^aCenter for microbial communities, Aalborg University, Denmark

^bJoint Microbiome Facility, University of Vienna, Austria

^cDNASense ApS, Denmark

*These authors contributed equally to the paper

**Corresponding author ma@bio.aau.dk

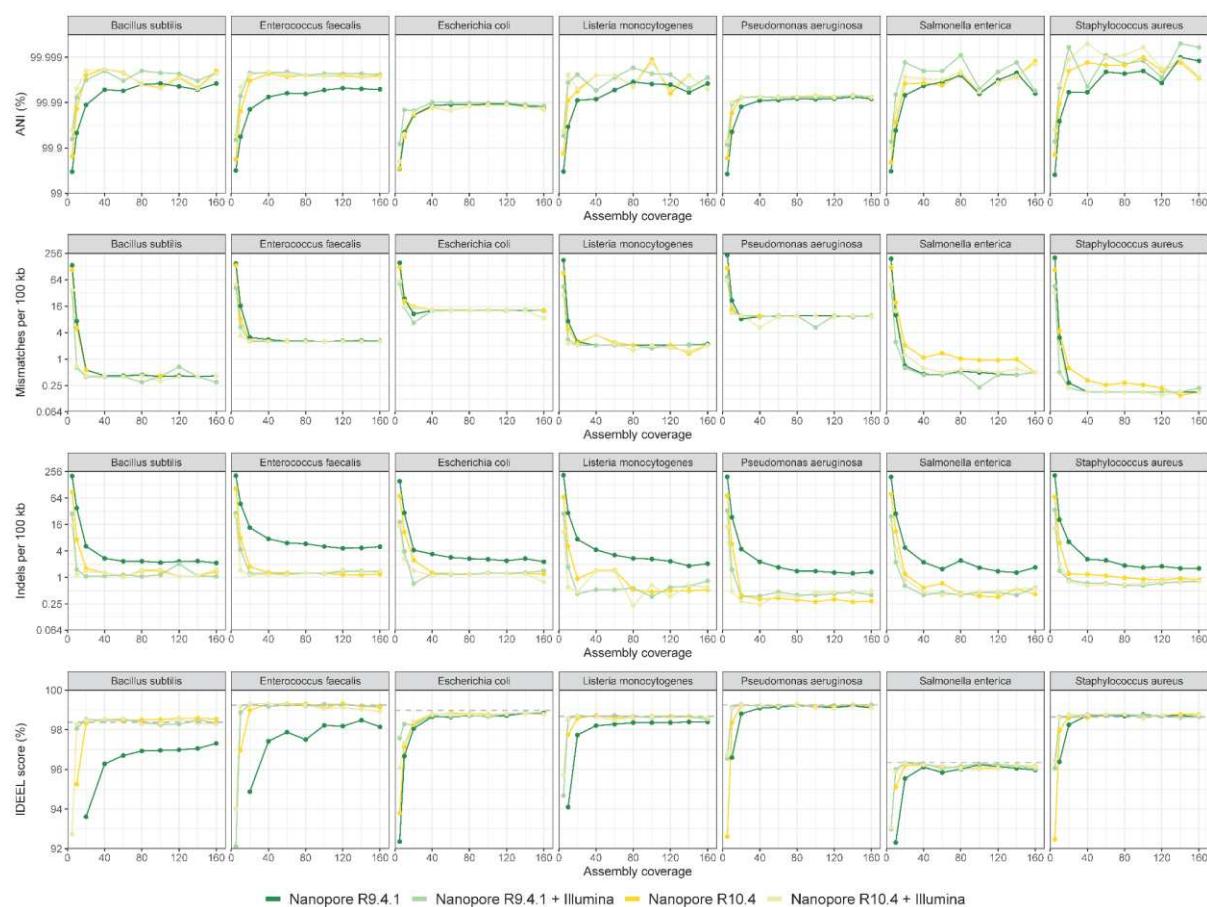


Figure S1: Assembly metrics for the ZYMO Mock HMW DNA.

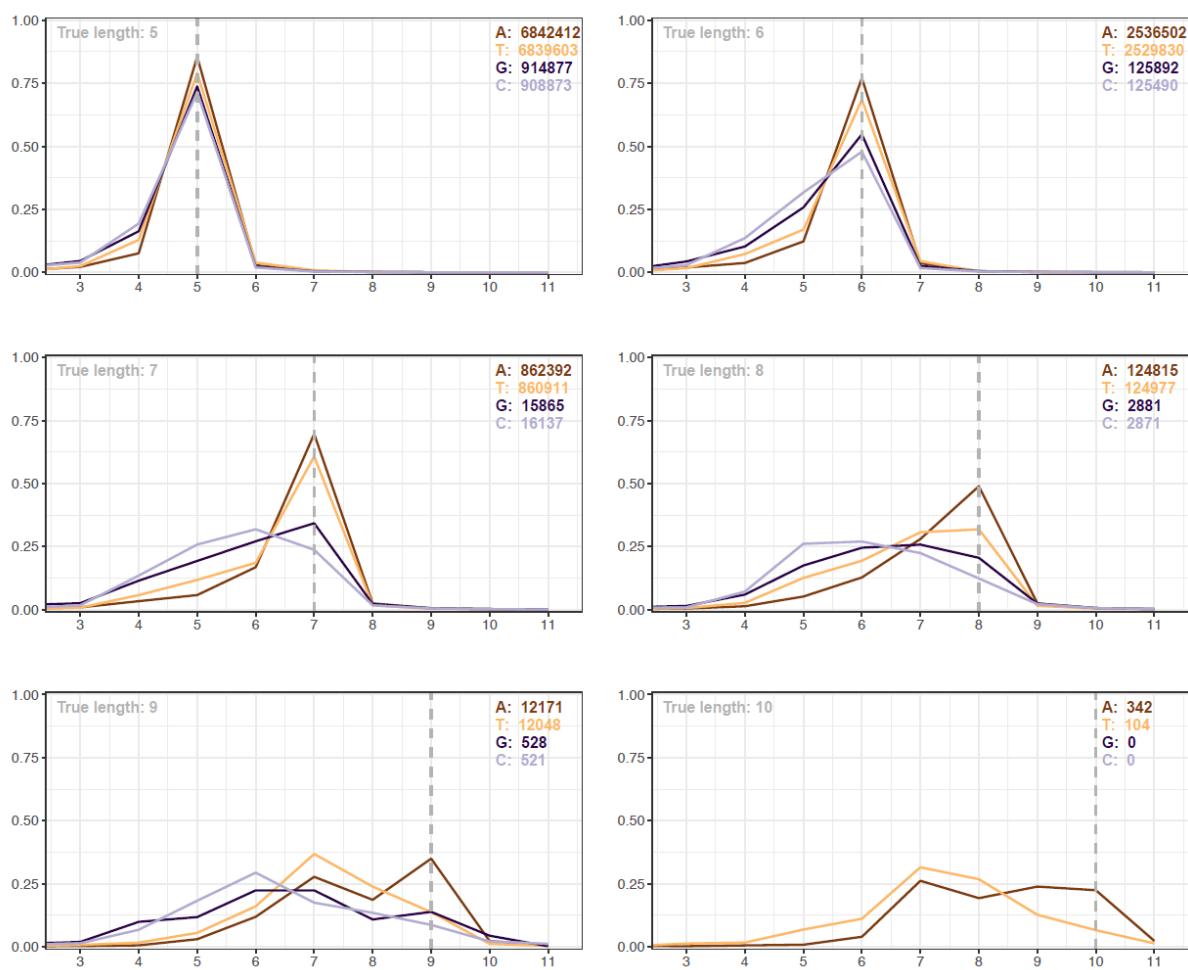


Figure S2: Counterr homopolymer plot for Nanopore R9.4.1 read data of the Zymo mock. Reads for each Zymo mock species, subsetted to a coverage of 160 were used for the analysis.

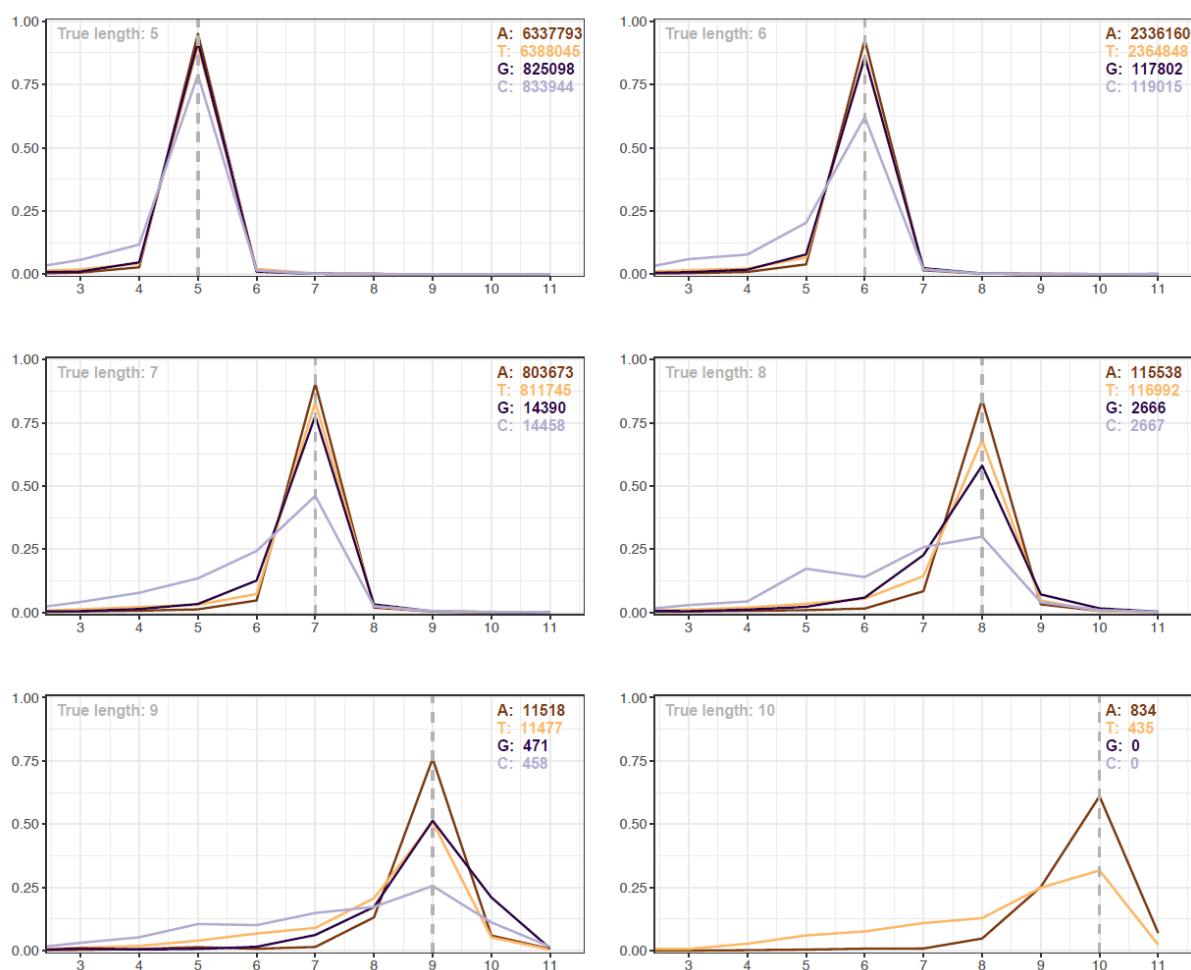


Figure S3: Counterr homopolymer plot for Nanopore R10.4 read data of the Zymo mock. Reads for each Zymo mock species, subsetted to a coverage of 160 were used for the analysis.

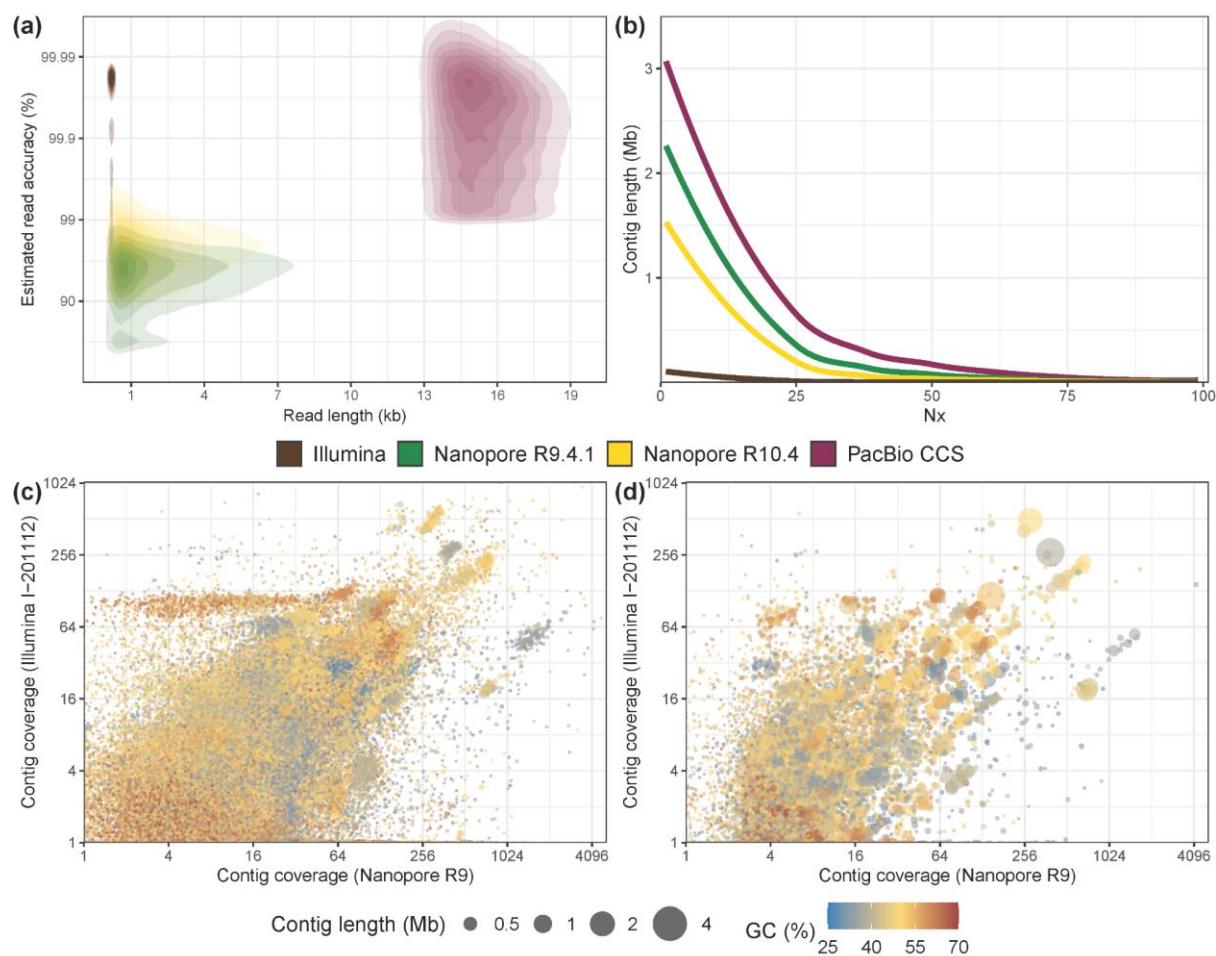


Figure S4: Sequencing and assembly overview for the anaerobic digester sample. **A)** Estimated read accuracy (from Q-scores) versus read length. Note that the PacBio HiFi sample underwent additional size selection prior to sequencing. **B)** Nx plot of the assemblies produced from different sequencing technologies. **C)** Differential coverage plot of the Illumina assembly. **D)** Differential coverage plot of the Nanopore R9.4.1 assembly.

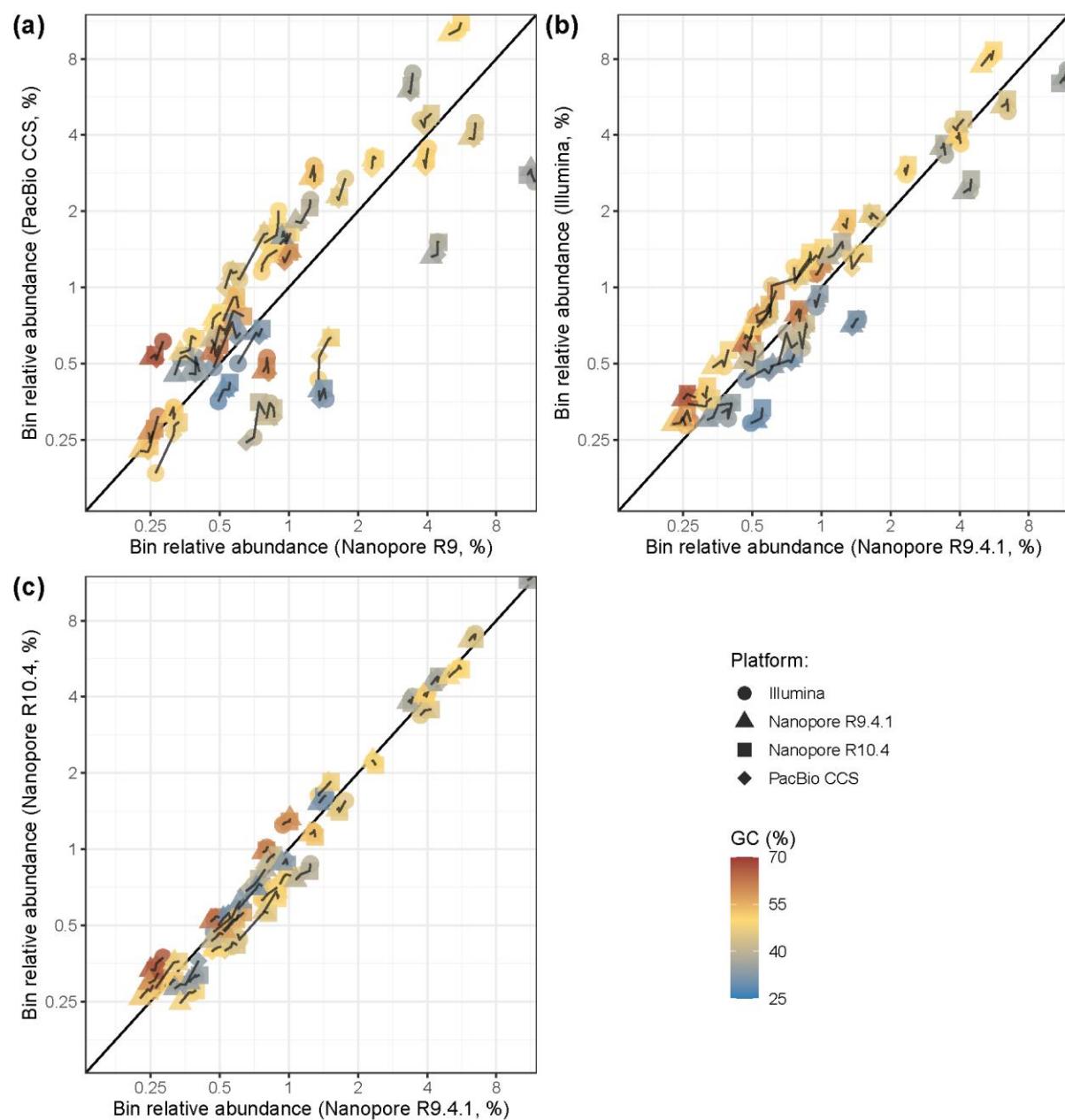


Figure S5: Comparison of bin relative abundances between different sequencing platforms. Relative abundance values (log-scaled) are presented between the Nanopore R9 data and **a**) PacBio CCS, **b**) Illumina, **c**) Nanopore R10. Only the bins that were clustered together between different platforms are presented in the plots and are interlinked.

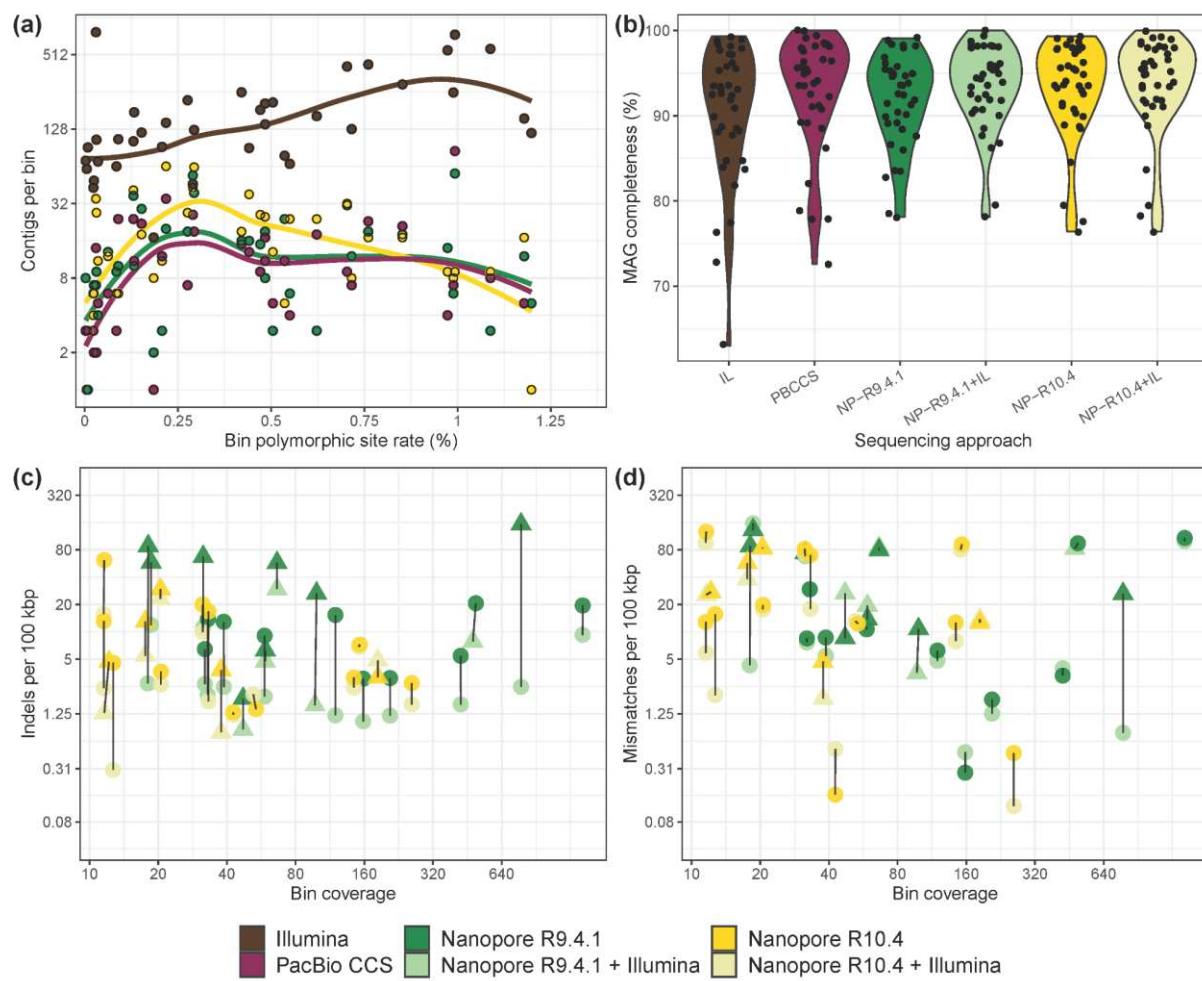


Figure S6: Comparison of bins from different sequencing approaches. **a)** MAG fragmentation (log-scaled) at different bin SNP rates in PacBio CCS MAGs. **b)** Genome bin completeness estimates for different sequencing platforms. IL — Illumina, NP — Nanopore, PBCCS — PacBio CCS. Bin **c)** indel and **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.

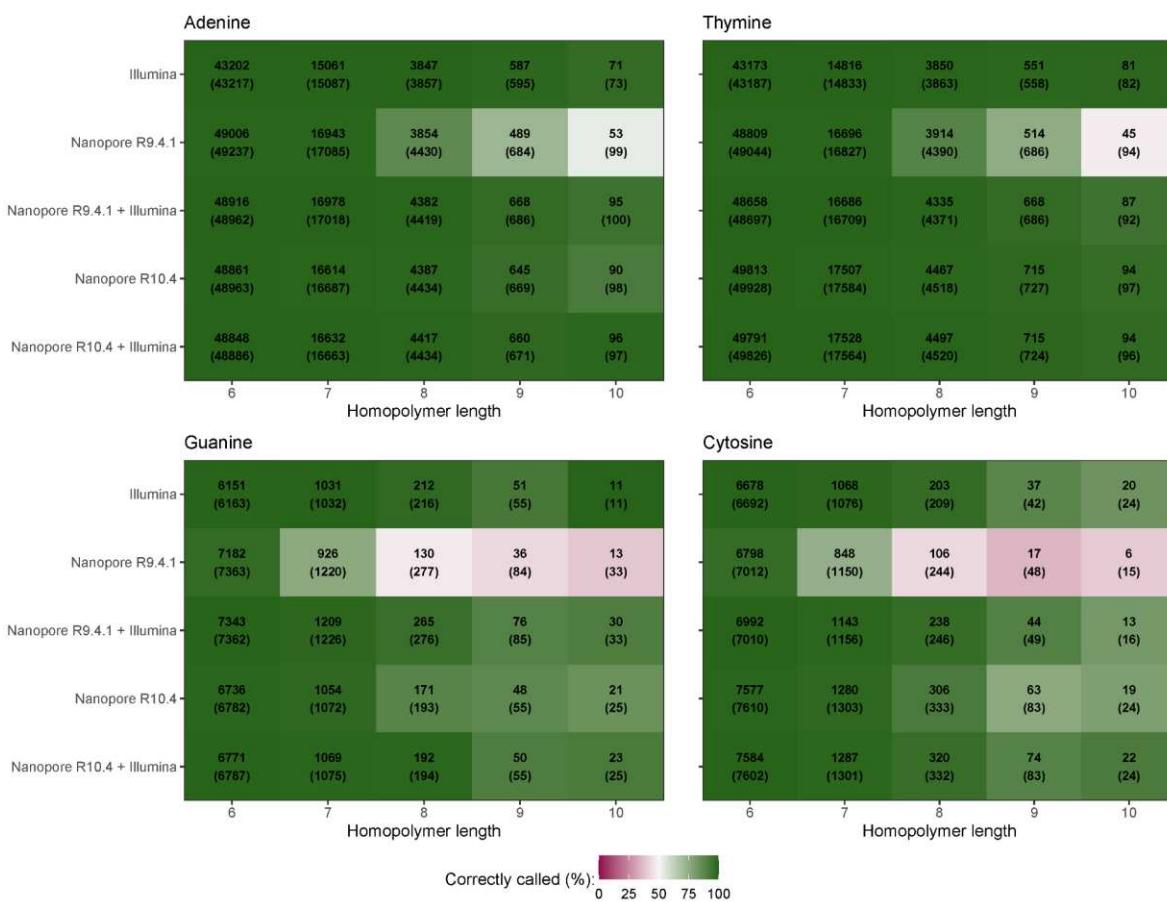


Figure S7: Homopolymer calling estimates in metagenomes (consensus sequences) from different sequencing platforms. Values in the heatmap show observed homopolymer counts estimated to be called correctly at a given sequence length. The total count of homopolymers (called correctly and incorrectly) are in brackets. Only the contigs for bins that were clustered together between different platforms were used to generate values for the plot.

Table S1: Sequence statistics for the Zymo HMW Mock using different sequencing platforms. Estimated modal read accuracy is measured using the reported Q-score for each read type. Observed modal read accuracy was measured by read-mapping to the reference genomes.

| | Illumina | Nanopore R9.4.1 | Nanopore R10.4 |
|--|------------|--------------------|-------------------|
| Total read count | 48,123,500 | 8,846,993 | 22,452,567 |
| Total yield (Gbp) | 7,2 | 31,6 | 52,3 |
| N50 (bp) | 151 | 14,018 | 5,992 |
| Estimated modal read accuracy (%) | 99.99 | 96.89 | 98.22 |
| Observed modal read accuracy (%) | 99.98 | 97.59 | 99.07 |

Table S2: Overview of read datasets used in the study.

| Read dataset | Instrument | Yield (Gb) | Read N50 (kb) | Read count | ENA sample ID | LOT# |
|--------------|------------------|------------|---------------|------------|---------------|-----------|
| 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 | |
| R104-202001 | GridION | 14.0 | 7.5 | 3,514,955 | ERS7672969 | |
| IL-ZYMO | Illumina MiSeq | 7.5 | 0.15 | 49,774,986 | ERS8296812 | ZRC195845 |
| R941-ZYMO | MinION Mk1B | 32.0 | 1.8 | 8,851,918 | ERS8296813 | ZRC195845 |
| R104-ZYMO | PromethION | 5.2 | 7.5 | 18,831,686 | ERS8296814 | |

Table S3: CMSeq SNP calling statistics for the Zymo mock reference sequences.

| | Covered bases (Mb) | Polymorphic bases (bp) | Polymorphic rate |
|-------------------------------|---------------------------|-------------------------------|-------------------------|
| Bacillus subtilis | 4.0 | 10 | 2.5e-06 |
| Enterococcus faecalis | 2.8 | 113 | 4.0e-05 |
| Escherichia coli | 4.8 | 1156 | 2.4e-04 |
| Listeria monocytogenes | 3.0 | 80 | 2.7e-05 |
| Pseudomonas aeruginosa | 6.8 | 1222 | 1.8e-04 |
| Salmonella enterica | 4.8 | 41 | 8.6e-06 |
| Staphylococcus aureus | 2.7 | 18 | 6.6e-06 |