

1 Improving recovery of member genomes from enrichment
2 reactor microbial communities using MinION-based long read
3 metagenomics

4 Krithika Arumugam^{1,*}, Irina Bessarab^{2,*}, Xianghui Liu¹, Gayathri Natarajan¹,
5 Daniela I. Drautz-Moses¹, Stefan Wuertz^{1,3}, Federico M. Lauro^{1,4}, Ying Yu Law¹,
6 Daniel H. Huson^{5,6}, and Rohan B. H. Williams^{2,**}

7 ¹Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological
8 University, 60 Nanyang Drive, SBS-01N-27, Singapore, 637551

9 ²Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore,
10 28 Medical Drive, Singapore, 117456

11 ³School of Civil and Environmental Engineering, Nanyang Technological University, N1-01a-29,
12 50 Nanyang Ave, Singapore, 639798

13 ⁴Asian School of the Environment, Nanyang Technological University, 50 Nanyang Avenue,
14 N2-01C-37, Singapore, 639798

15 ⁵Center for Bioinformatics, University of Tübingen, Sand 1, 72076, Tübingen, Germany

16 ⁶Life Sciences Institute, National University of Singapore, 28 Medical Drive, Singapore, 117456

17 *I.B and K.A contributed equally to this work

18 **Corresponding author: R.B.H.W.at lsirbhw@nus.edu.sg

19 November 8, 2018

Abstract

New long read sequencing technologies offer huge potential for effective recovery of complete, closed genomes. While much progress has been made on cultured isolates, the ability of these methods to recover genomes of member taxa in complex microbial communities is less clear. Here we examine the ability of long read data to recover genomes from enrichment reactor metagenomes. Such modified communities offer a moderate level of complexity compared to the source communities and so are realistic, yet tractable, systems to use for this problem. We sampled an enrichment bioreactor designed to target anaerobic ammonium-oxidising bacteria (AnAOB) and sequenced genomic DNA using both short read (Illumina 301bp PE) and long read data (MinION Mk1B) from the same extraction aliquot. The community contained 23 members, of which 16 had genome bins defined from an assembly of the short read data. Two distinct AnAOB species from genus *Candidatus Brocadia* were present and had complete genomes, of which one was the most abundant member species in the community. We can recover a 4Mb genome, in 2 contigs, of long read assembled sequence that is unambiguously associated with the most abundant AnAOB member genome. We conclude that obtaining near closed, complete genomes of members of low-medium microbial communities using MinION long read sequence is feasible.

The development of long read sequencing technologies, such as the Oxford Nanopore Technology MinION and Pacific Biosciences SMRT offer huge potential for effective recovery of complete, closed genomes [1, 2]. While much progress has been made towards this objective for single species isolates [3, 4], the ability of these methods to recover genomes of member taxa in complex microbial communities (microbiomes) is less clear.

To date there is limited long read data available in the metagenome context. Several recent studies have used MinION or PacBio sequencing on highly enriched reactor communities, for example a recent study using PacBio long read sequencing to recover a complete, closed genome of *Kuenenia stuttgartiensis* which was present at around 95% relative abundance [5]. Other studies have obtained long read data or combined long and short read data on more complex communities, for example from moderately enriched bioreactor communities [6], co-culture enrichments [7], marine holobionts [8] or from full scale anaerobic digester communities [9]. There are also several datasets which provide long and short read data from mock communities [10, 11, 12], and which constitute important new benchmarking data. New long read analysis methods [13] and binning algorithms designed for long read metagenome data [14] are becoming available, anticipating the future expansion of metagenome data from these new instruments.

Here we examine the ability of long read data to recover genomes from enrichment bioreactor metagenomes. Such modified communities offer a moderate level of complexity compared to the source communities and so are realistic, yet tractable, systems to use for this problem. Specifically, we sampled an enrichment bioreactor targeting anaerobic

62 ammonium-oxidising bacteria (AnAOB), extracting genomic DNA and obtaining both
63 short read (Illumina MiSeq 301bp PE) and long read data (MinION Mk1B) from the
64 same aliquot. We perform a comparative analysis of different long read assemblers and
65 assembly workflows, using the draft genomes extracted from the short read assembly as an
66 independent references to assess the ability of long read assemblies to capture genome of
67 member species of the community. Our findings show that the long read metagenomics can
68 recover a complete, near-closed genome of the most abundant community member, even
69 in the presence of multiple species of the same genus.

70 **Methods**

71 **Overall approach**

72 We studied a bioreactor community established to enrich for anaerobic ammonia-oxidizing
73 bacteria (AnAOB), all known species of which belong to the phylum *Planctomycetes*. Pre-
74 viously we used this enrichment reactor community to recover a draft genome of a member
75 of the genus *Candidatus Brocadia* [15]. We sampled the suspended biomass of the reactor,
76 extracted DNA and performed both Illumina and Nanopore DNA sequencing from the
77 same extracted aliquot. Such enrichment reactor communities have the advantage of being
78 only moderately complex relative to their source communities and therefore provide an
79 ideal tradeoff between realistic complexity and tractability in relation to member genome
80 recovery.

81 We constructed long read assemblies and then used draft genomes obtained from short
82 read assemblies for the purpose of assessment. This approach takes advantage of the
83 fact of our having obtained data from both sequencing modalities using the same DNA
84 aliquot, and leverages current understanding of short read metagenome assembly binning
85 and quality assessment procedures [16, 17]. We tested five different long read assemblers or
86 assembly workflows, namely *Canu* [18], *Miniasm* [19], *SMARTDenovo* [20], *Spectrassembler*
87 [21] and *Unicycler* [22]. We note that while these softwares are not designed or endorsed
88 for metagenome analysis, our experimental design explicitly tested their ability to be used
89 in this setting.

90 **Enrichment bioreactor and sampling**

91 As previously described [15], we (Y.Y.L, G.N, S.W) developed a protocol to enrich for
92 AnAOB organisms using a sequencing batch reactor seeded with return activated sludge
93 from a water reclamation plant (Public Utilities Board, Singapore). The reactor was fed
94 synthetic wastewater containing ammonium and nitrite and operated at 35 °C. The reactor
95 showed a characteristic reddish-brown biofilm on the walls of the reactor (characteristic
96 of AnAOB bacteria), and we observed simultaneous ammonium and nitrite depletion con-
97 sistent with the occurrence of Anammox. We confirmed the presence of AnAOB species

98 within the community using FISH, 16S amplicon sequencing and shotgun metagenomics
99 [15]. At the time of sampling for the present study (10/2016) the reactor community had
100 been established for approximately 9 months.

101 DNA extraction

102 DNA was extracted from sampled biomass with the FastDNATMSPIN kit (MP Biomedicals)
103 for Soil, using using 4× bead beating with a FastPrep homogeniser (MP Biomedicals)
104 followed by cleanup with Genomic DNA Clean & ConcentratorTM-10 (Zymo Research).

105 Short read sequencing

106 Genomic DNA Library preparation was performed using a modified version of the Illumina
107 TruSeq DNA Sample Preparation protocol. We then performed a MiSeq sequencing run
108 with a read length of 301 bp (paired-end). The raw short read data is publicly available
109 from the NCBI via BioProject accession PRJNA471614 or Sequence Read Archive accession
110 SRR8168439.

111 Long read sequencing

112 Library preparation was constructed from 1.7μg of genomic DNA using the SQK-LSK
113 108 Ligation Sequencing Kit (Oxford Nanopore Technologies). Sequencing was performed
114 on a MinION Mk1B instrument (Oxford Nanopore Technologies) using a SpotON FLO
115 MIN106 flowcell (FAF05082) and R 9.4 chemistry. Data acquisition was performed using
116 MinKNOW version 1.3.30 running on a HP ProDesk 600G2 computer (64-bit, 16Gb RAM,
117 2Tb SSD HD) running Windows10. The raw long read data is publicly available from
118 the NCBI via BioProject accession PRJNA471614 or and Sequence Read Archive accession
119 SRR7169548.

120 Analysis of short read sequence data

121 The raw FASTQ files were processed using cutadapt (version 1.14) [25] using the following
122 arguments: `--overlap 10 -m 30 -q 20,20 --quality-base 33`. We initially characterised
123 the community data using Ribotagger [26] selecting the V4 region and annotating
124 ribotags to the SILVA database [27]. We calculated relative abundance of each ribotag
125 by dividing their reads counts by the total read count from all ribotags. Single-sample
126 metagenome assembly of the short read sequence was constructed using Newbler v2.9.
127 The contigs generated from short read data are hereafter denoted as *short read assembled*
128 *contigs* (SRAC). We identified 16S genes within contigs using the `--search16` module of
129 USEARCH (version 10.0.240, 64 bit) [28], and annotated them using the SILVA::SINA web-
130 server (using default parameters) [29]. Genes (open reading frames) were predicted from
131 all contigs using MetaGeneMark [30] and annotated to RefSeq NR (January 2016 version)

132 [24] using DIAMOND v0.8.22.84 (running default parameters) [31]. We identified putative
133 member genomes using MetaBAT [32]. For each identified bin we performed taxonomic
134 level analysis using CheckM [33] and by applying the K -statistic methods of Williams *et*
135 *al.* [34].

136 Analysis of long read sequence data

137 Basecalling was performed using Albacore version 2.2.4. Adaptor trimming was performed
138 using Porechop [35] with default settings. We performed taxonomic analysis of long read
139 data using MEGAN-LR [13]. We assembled long read data using each of the following long
140 read assemblers: Canu [18], Miniasm [19], SMARTDenovo [20], Spectrassembler [21] and
141 Unicycler [22], using default setting unless otherwise stated. In Canu there is a required
142 parameter (`genomeSize`) that sets the expected genome length, and which controls the
143 number of reads used by assembler. Because this estimate is essentially meaningless in
144 the metagenome context, we computed multiple assemblies using the following values of
145 `genomeSize`: 3.4 Mbp, 5 Mbp, 12 Mbp, 25 Mbp, 50 Mbp and 120 Mbp. Contigs generated
146 from long read data are hereafter denoted as *long read assembled contigs* (LRAC). The
147 number of reads used in each assembly was estimated by mapping long read to LRAC
148 sequence with minimap2 [36] and using samtools-1.6 to calculate the number of aligned
149 reads [37].

150 Comparative analysis of long and short read assemblies

151 We used BLASTN (version 2.4.0+) [23] to examine the degree of sequence alignment between
152 LRAC and SRAC sequences. We treated the LRAC as the subject sequences and the SRAC
153 as the query sequences, using default BLASTN parameters and without the results outputted
154 in tabular format. To assess the extent to which SRAC sequence aligned to LRAC sequence,
155 we computed ratio of the alignment length to the length of the query (SRAC) sequence
156 (*al2ql*).

157 Results and Discussion

158 Community profiling and member genome recovery using short read data

159 Community profiling using RiboTagger [2] detected a total of 23 member taxa (defined
160 by 16S V4 region sequence captured in metagenome reads). Of these 7 taxa showed a
161 relative abundance of at least 5% (Fig. 1A). In order of decreasing relative abundance,
162 those seven taxa are *Candidatus* Brocadia (genus), a member of phylum *Armatimonadetes*,
163 an unknown taxa (no kingdom assigned), a member of genus *Denitratisoma*, a second *Ca.*
164 *Brocadia* (genus) taxon, a member of family *Anaerolineaceae* and another unknown taxon.
165 The short read data was subsequently assembled (Table S1) and automatically binned (see

166 Methods and Fig. 1B), which resulted in 16 bins being defined, of which 4 demonstrated an
167 estimated completeness of $> 95\%$ and contamination $< 5\%$ (Table S2). Taxonomic analysis
168 of each bin was based on presence of 16S sequence, marker gene analysis [33] and analysis of
169 ORF-level annotations [34]. The results were qualitatively consistent with the 16S analysis,
170 and two species of *Ca. Brocadia* identified (bin 2 and bin 6, with mean base coverage of 150
171 and 31, respectively), a member of phylum *Armatimonadetes* (bin 1, mean base coverage
172 180) and a member of phylum *Chloroflexi* (bin 7, mean base coverage, 150) (Fig. 1B,
173 Table S2–S3). We observed that bin 3, comprised of a set of 26 contigs (approximately
174 201k bp of sequence, with mean base coverage 142) that are strongly associated with *Ca.*
175 *Brocadia* (Table S2–S3), overlapped in the coverage–GC plane with bin 2 and we therefore
176 subsequently combined those two bins under the assumption that they were spuriously
177 separated by the binning procedure (referred to hereafter as bin 2).

178 Nanopore sequencing and initial taxonomic analysis of long read data

179 The MinION run generated a total of 115,383 basecalled reads, of which 95,571 held a
180 quality score > 7 (Table 1) and 93.4% were greater than 500 bp in length. The mean read
181 length was 4,677 bp with the longest read being 112,729 bp. There were 6,324 reads with
182 a length greater than 10,000 bp. A total of 539,410,340 bp of sequence was generated from
183 the run.

184 Taxonomic assignments using MEGAN–LR were obtained for 88,097 reads (76.4% of
185 all reads), representing a total of 326M aligned bases, of which 291Mbp were annotated
186 to bacterial phyla. The most prominent phyla as ranked by aligned base length was
187 Planctomycetes (157Mb or 54% of total sequence aligned to phylum level), *Proteobac-*
188 *teria* (69.7Mb, 24%), *Chloroflexi* (21.0 Mbp) and *Armatimonadetes* (1.79 Mbp, 6.14%)
189 (Table S4). A total of 162 Mbp of alignments were obtained at genus level, with 93.4%
190 of these attributable to *Ca. Brocadia*. Around 2% of aligned bases were attributable to
191 genus *Nitrospira*, with the remaining detected genera each accounting for less than 1% of
192 aligned bases (Table S5).

193 Long read assembly

194 Using the long read data, we constructed assemblies using five assemblers or assembly
195 workflows (Table 2). The total amount of assembled sequence varied between approxi-
196 mately 4.8Mb (*Canu* with `genomeSize` parameter set to 3.4 Mbp) and 11.2 Mbp (*Canu*
197 with `genomeSize` parameter set to 50 Mbp or 120Mbp) (Table 2). Examination of the
198 GC-length data showed a consistent tendency for longest contig to be clustered around a
199 GC of approximately 0.4, consistent with the high abundance of AnAOB (Fig. 2). The
200 distribution of LRAC sequence length varied across assemblers (Fig. 2), but all assem-
201 blers returned at least one LRAC sequence greater than 1Mb in length, with *Canu* and
202 *SMARTdenovo* returning 2 sequences longer than 1Mb. Between 33% and 45% of long reads

203 contributed to LRAC sequence, dependent on what assembler or assembly workflow was
204 used (Table 2).

205 Taxonomic and genomic analysis of long read assembled sequence

206 To begin taxonomic analysis of LRAC sequence, we next examined whether 16S gene
207 sequences were present using homology search, and if so, obtained their taxonomic anno-
208 tations. To do so we aligned all full length 16S genes identified in the short read assembly
209 to all LRAC sequences. In the short read assembly, we identified a full length 16S gene
210 in bin 2 (residing on contig00593) that was aligned to SILVA annotated sequence from
211 *Ca. Brocadia* with 95% identity. That sequence mapped with over 99% identity to LRAC
212 sequences generated by **Canu** and **Unicycler** and close to 99% identity in the case of
213 **Spectrassembler**. In the other two assemblers, the degree of alignment were lower at 88%
214 and 73% for **Miniasm** and **SMARTdenovo**, respectively. A second full length 16S gene anno-
215 tated to *Ca. Brocadia* in SILVA and recovered from bin 6 was aligned to the same LRAC
216 locations as the bin 2 sequence, but with lower percent identity in all instances (alignment
217 statistics for these two 16S sequences are provided in Supplementary Datafile 1).

218 We next examined whether these LRAC sequences constituted complete or near-complete
219 genomes of member species. Using BLASTN, we aligned SRAC sequences (treated as
220 queries) to the LRAC sequences (treated as subject) and then categorised the homology
221 search statistics by the bin of origin of the aligned SRAC sequence. We summarised the
222 bulk alignment statistics for each LRAC partitioned by SRAC-bin using the mean of the
223 median of the distribution of the percent identity (*pident*) and the median of the distri-
224 bution of the alignment length to query (SRAC) length (*al2ql*; expressed as a percentage
225 to be on the same scale as *pident*) (Fig. 3). With the exception of **Spectrassembler**
226 the longest LRAC sequences showed a strong degree of alignment with SRAC sequences
227 arising from the highest abundance *Ca. Brocadia* bin (bin 2) and to a lesser degree with
228 SRAC sequences from the lower abundance *Ca. Brocadia* taxon (bin 6). No other LRAC
229 sequence demonstrated a high degree of association with any other genome bin defined
230 in our analysis. We therefore supposed that these LRAC sequences capture substantial
231 sized fragments of the most abundant *Ca. Brocadia* genome present in the community,
232 and then developed further analyses to test this notion. We now focused attention of the
233 **Canu** assembly generated using `genomeSize=3.4Mb`, as this gave the longest single LRAC
234 sequence from any workflow. This LRAC sequence (hereafter referred to as `tig00000001`)
235 contained the full length 16S referred to above. Given this sequence also has a mean GC
236 close to 0.4, we can consider it a candidate for being a large genome fragment of *Ca. Bro-*
237 *cadia* present in the reactor community. This contig had a mean long read base coverage
238 of 48 (s.d: 12) (Fig. S1).

239 If this LRAC sequence represents large genomic fragments from a member species of
240 the community, then we would predict the SRAC sequences from the cognate short read
241 assembly bin will align in a spatially uniform fashion across it. We therefore examined the

242 spatial distribution of SRAC alignments across the entire the cohort of LRAC assembly
243 using dot plots, generating these for the following sets of SRAC sequence 1) all contigs
244 that are members of neither bin 2 nor bin 6; 2) contigs that are members of bin 6 and
245 3) contigs that are members of bin 2. We further subset for those alignments that were
246 full length (defined as alignments exhibiting $al2ql > 95\%$). The results of each of the
247 corresponding six analyses are shown in Figs 4–9. Each plot shows a set of four panels, the
248 left most of which is a dot plot showing the alignments of the SRAC sequence on the entire
249 LRAC assembly, the next rightmost panels showing the corresponding BLAST statistics
250 for SRAC alignments, namely $al2ql$ and $pident$, and the rightmost panel shows the length
251 of the aligned SRAC sequences.

252 Starting with the SRAC sequences that are neither members of bin 2 or bin 6, we
253 observe only a small number of alignments to `tig00000001` with low values of $al2ql$ (Fig.
254 4), and four alignments of full length, high percent identity to other short LRAC sequences
255 in the assembly (Fig. 5). SRAC sequences from each of the two *Ca. Brocadia* bins
256 showed markedly different patterns of alignment. Only SRAC sequences from bin 2 showed
257 spatially uniform tiling across `tig00000001` (Fig. 8), with a high degree of alignment and
258 percent identity (Fig. 9). SRAC sequences from bin 6 also showing a large number of
259 alignments, but at greatly reduced number when we considered only full length alignments
260 (Fig. 7). Based on this analysis, we can conclude that `tig00000001` appears to derive from
261 genome fragments associated with the *Ca. Brocadia* species captured in bin 2, but not bin
262 6, in the short read assembly. We note that the reduced degree of percent identity from
263 SRAC sequences from bin 6 is not a consequence low coverage (leading to a systematic
264 reduction in contig length), as a systematic relationship between SRAC length and coverage
265 is only evident at coverage values < 10 , not at the coverage ranges observed for either *Ca.*
266 *Brocadia* bin (Fig. S2A).

267 We observed that the second longest contig in the assembly, hereafter referred to as
268 `tig00000002` (length: 362k bp; mean base coverage of 42, s.d: 12, see Fig. S2B), showed a
269 similar pattern of alignments from the bin 2 SRAC sequence, suggesting that it is another,
270 shorter fragment of the *Ca. Brocadia* captured in `tig00000001`. The lack of repetition
271 of SRAC alignments relative to those aligned to `tig00000001` suggests this is a distinct
272 genome fragment, not a shorter sub-fragment of the `tig00000001` sequence.

273 Refining short read bin composition reduces alignment gaps in long read 274 assembly

275 Despite the spatially uniform tiling of SRAC sequences from bin 2 sequences across the
276 long read assembly, the dot plots suggest the presence of a number of gaps where no
277 alignments exist. We quantified the extent of this and observed that 21.3% and 44.6% of
278 `tig00000001` and `tig00000002` are not covered by bin 2 SRAC alignments, respectively.
279 Examination of the default bin assignments from MetaBAT (Fig. 1) suggest both bin
280 2 and bin 6 contain a proportion of outliers, which are be unlikely to derive from their

281 cognate genomes. Because neither the SRAC bin-genomes nor the LRAC sequences can
282 be considered a gold-standard assembly of given member species, we refined the short
283 assembly bin membership and re-examined the extent of tiling across the cognate LRAC
284 sequence, to determine if the occurrence of alignment gaps was determined largely by
285 bin membership, incomplete short read assembly and/or subsequences in the long read
286 assembly associated with sequencing or assembly errors. To expand the total amount of
287 candidate sequence derived from the bin 2 *Ca. Brocadia* genome we included all SRAC
288 contigs at least 1kbp in length and performed the following procedure. We started with
289 the bin 2 contig that contained the full length 16S sequence (contig00593), and extracted
290 contig sets defined by the k -nearest neighbours of that contig in the binning plane. We grew
291 these neighbourhoods in increments of 10 contigs, starting with 10 contigs and finishing
292 with 550 contigs (Fig. 10A). For each of these neighbourhoods, we calculated 1) **CheckM**
293 genome quality statistics (Fig. 10B); 2) total sequence length (Fig. 10C); 3) ORF-level
294 annotation statistics summarised at neighbourhood level for family and genus level for *Ca.*
295 *Brocadiaceae* (Fig. 10D) and *Ca. Brocadia* (Fig. 10E), respectively; 4) the proportion of
296 LRAC sequence of **tig00000001** and **tig00000002** that contain no aligned SRAC sequence
297 from bin 2 (Fig. 10F); 4) the number of bin 2 SRAC sequenced that were unaligned to
298 LRAC sequences (Fig. 10G). When calculating gaps in the alignments to LRAC sequence,
299 we only considered full length SRAC aligned sequences (*i.e.* those with $al2ql > 0.95$).

300 Examining this contig neighbourhood analysis, we observe that 90% completeness as
301 defined by **CheckM** is attained after the $k = 230$ neighbourhood, with non-zero values of
302 contamination and strain heterogeneity appearing at $k = 240$ and $k = 440$, respectively. We
303 note that contamination never exceeds 3.85% which is below the 5% threshold considered
304 to be a high quality draft metagenome assembled genome [17]. Visually, the occurrence of
305 contamination is associated with neighbourhoods starting to impinge on the region of the
306 binning plane associated with contigs in bin 6, as well as outlier contigs from other taxon
307 bins. Total sequence length increased in a near-linear fashion until $k = 220$, attaining
308 a length of 2.67Mb, after which a more gradual increase in length was observed. High
309 values of the K -statistic were observed across all neighbourhoods (with minimal signal
310 from other taxa), with a slight decrease towards higher values of k . Interestingly, for the
311 $k = 230$ neighbourhood, there remains 36% and 52% of **tig00000001** and **tig00000002**
312 that contain no aligned, full-length SRAC sequence. By $k = 400$ these proportions had
313 dropped to 15% and 33% respectively, with their nadir values of 10% and 27% being
314 observed at $k = 530$ and $k = 520$, respectively. We also note that all SRAC sequences from
315 the neighbourhood are aligned to an LRAC sequence until the $k = 320$ neighbourhood,
316 and across all neighbourhoods a total of 5 SRAC sequences never align to either LRAC
317 sequence. The combination of contamination and strain heterogeneity statistics, along with
318 the appearance of contigs that do not map to LRAC sequences, suggest that the short read
319 derived genome is most complete and uncontaminated between the $k = 310$ and $k = 430$
320 neighbourhoods, which at the latter contains a total of 3.6Mb of sequence.

321 Collectively these observations suggest that as defined by LRAC sequences, the AnAOB

322 genome arising from bin 2 of the short read assembly, is just over 4Mb in length and is cap-
323 tured by one long contig (`tig00000001` and one shorter contig (`tig00000002`) of lengths
324 3,698,683 bp and 362,151 bp, respectively. The gaps in the coverage of SRAC sequences
325 onto these LRAC draft genome is consistent with incomplete assembly of cognate genome
326 using the short read approach. The estimated length of 4Mb is consistent with previ-
327 ous AnAOB genome estimates, including a recent complete, closed assembly of *Kuenenia*
328 *stuttgartiensis* using PacBio sequencing [5]. Previous short read assemblies, obtained from
329 metagenome survey (MAG), are generally shorter than the estimated length by up to 1Mb,
330 consistent with the above findings.

331 Summary and Conclusions

332 In this paper we examine the capacity of long read data, generated by a Nanopore MinION
333 sequencer, to capture member genomes of a microbial community of low–medium complex-
334 ity. We show that we can recover a complete, near closed genome of the most abundant
335 community member, specifically an anaerobic ammonium oxidising bacterium (AnAOB)
336 belonging to genus *Ca. Brocadia*. We also show that we can unambiguously resolve the
337 genome of this species without contamination from a lower abundance community member
338 of the same genus. Our results also cast light on the limitations of short read metagenome
339 assembly, and the general challenges associated with accurate recovery of member genomes
340 from complex microbial communities.

341 Long read metagenomics will likely represent an area of substantive activity over the
342 next several years, particularly in regard to the recovery of complete genomes from mem-
343 ber species. Despite great progress in the last decade in developing procedures to obtain
344 short read metagenome–assembled genomes (MAG) [16, 17], substantial limitations of these
345 approaches have become evident, including problems related to the use of multi–sample
346 co–assemblies [38, 39], the challenges of resolving genomes to strain level [40], and the chal-
347 lenges of extracting MAGs from communities of moderate to high ecological complexity
348 [41, 42, 43]. Accordingly, several recent papers have combined short read with emerging
349 complementary techniques such as HiC metagenomics [42, 44], and it is likely to that the
350 combination of short and long read metagenomics will also yield considerable progress in
351 defining higher quality MAGs than is possible using short read metagenome assembly pro-
352 cedures alone. In the present paper, we seek to establish the feasibility of such hybrid
353 long–short metagenomic approaches, starting with communities of low–moderate complex-
354 ity.

355 Our analysis is based on the use of a single DNA aliquot co-assayed with both long and
356 short sequencing, thus avoiding the introduction of any difference between the two sets of
357 sequence that could be attributed to ecogenomic differences. Our analysis proceeds on the
358 basis that neither SRAC nor LRAC sequence is known to provide a true reference genome,
359 and so we seek to understand and characterise the degree of agreement between them

360 [45]. Although error prone MinION sequence can be corrected using higher quality short
361 read sequences, in this case, we have deliberately kept the two sources of data separate
362 so as to not introduce any positive bias in our SRAC to LRAC alignment analysis. One
363 consequence of this choice is that direct use of the LRAC assembly is limited in relation
364 to prediction of genes (open reading frames), due to the greatly increased frame-shift rate
365 in coding sequence [13]. Thus in this setting, we would consider our LRAC assembly to be
366 more akin to a genome-wide scaffold for higher quality short read genome fragments, rather
367 than an assembly *per se*. We note that we and others are actively working on correction
368 procedures for frame-shift errors at the time of writing, however, the issues of high error
369 rates in MinION sequence remains a substantial weakness of this technology.

370 We observe that most of the extant long read assembler or assembly workflows provide
371 broadly comparable results in relation to contig generation (Fig. 2 and Table 2) and
372 recruitment of bin-specific SRAC sequence (Fig. 3), although we note that some behaviours
373 are counter-intuitive. For example, in using *Canu*, increasing the `genomeSize` parameter
374 to ≥ 5 Mb results in splitting of the large contig observed in the 3.4Mb version, despite
375 the increased amount of sequence being assembled. We also note that less than 50% of
376 long reads are typically used in the assembly, for reasons that are unclear. Thus, further
377 development of metagenome assemblers and assembly workloads tailored for long read data,
378 or hybrid, short-long read data, is clearly needed. These data were obtained during a phase
379 of rapid development of MinION kits and protocols, only generating under 1Gb of sequence
380 from these complex environmental samples, and we anticipate more comprehensive results
381 using more recent datasets, given the improvement in yield, read length and quality now
382 available from the same technology.

383 The concept of genome completeness remains problematic in MAG-based analyses of
384 shotgun metagenome data. Current methods for estimating genome completeness from
385 short read data are based on marker gene sets whose membership is defined by the fact
386 of having a single genomic copy. Typically, these are of order of around 100 genes, and
387 by definition such sets cannot accurately estimate the completeness of a genome. In the
388 present case, we see that high level completeness in the SRAC sequence are occurring at
389 total sequence length that is around 1.4Mb less than the genome length inferred from LRAC
390 based analyses. In fact we would argue that given the complexity of how short read contigs
391 from many member species will be distributed in a given binning space, there must remain a
392 fundamental ambiguity in resolving genome content using these approaches, as highlighted
393 by our neighbourhood based reanalysis of bin 2. While the backbone of the genome is
394 certainly captured under these criteria, knowing when one has accurately captured a whole
395 genome remains challenging and reinforces the view that MAG-based analysis can likely
396 only resolve a working model of the genome (or pangenome) of a given community member
397 in most circumstances. In the present study we have also used completely automated
398 methods, with no recourse to manual or subjective decision making that appear to be
399 commonly deployed in the refinement of MAG draft genomes and represent a challenge to
400 the conduct of reproducible research in this area.

401 Further development of long read metagenomics may in fact obviate the need for com-
402 plex binning procedures, with methods refocusing on the evaluation of LRAC sequences
403 as putative draft genomes. Either way, it would appear as if conduct of hybrid short-
404 long read metagenome surveys will be essential for resolving these challenges for the fore-
405 seeable future. The engagement of other complementary approaches, namely single cell
406 metagenomics [46], HiC-metagenomics [44] or FACS-generated mini-metagenomes [43] in
407 combination with MAG approaches, will likely permit substantial progress.

408 List of Supplementary Materials

409 Supplementary items appear with this document unless otherwise stated.

- 410 • Table S1: Summary statistics for the short read assembly
- 411 • Table S2: CheckM statistics for short read assembly bins
- 412 • Table S3: *K*-statistics for short read assembly bins at phylum, family and genera
413 levels
- 414 • Table S4: Summary of phyla-level analysis of long read data using MEGAN-LR
- 415 • Table S5: Summary of genera-level analysis of long read data using MEGAN-LR
- 416 • Supplementary Datafile 1: Alignment statistics for SRAC derived 16S (xlsx file)
- 417 • Figure S1: Per-base long read coverage for AnAOB-associated LRAC sequences
- 418 • Figure S2: Relationship between length and coverage in the short-read metagenome
419 assembly

420 Author contributions

421 The study was conceptualised by RBHW and designed by RBHW and IB. GN, YYL and
422 SW setup and operated enrichment reactors, and obtained samples with IB. IB performed
423 DNA extractions and performed long read sequencing, with IB and FL designed long read
424 sequencing experiments. DID performed short read sequencing. KA, RBHW, XHL and
425 DHH designed and/or performed data analysis. All authors contributed to data interpre-
426 tation. RBHW wrote the manuscript with specific contributions from all other authors.

427 Acknowledgements

428 This research was supported by the Singapore National Research Foundation and Ministry
429 of Education under the Research Centre of Excellence Programme, and by program grant

430 1301–IRIS–59 from the National Research Foundation (NRF). We thank Gavin Huttley
431 (Australian National University) for critical feedback on sequence analysis. Early version
432 of this work were presented at APBC2018 (Yokohama, Japan, January 15–17 2018) and
433 ISME17 (Leipzig, Germany, August 12–17 2018)

434 References

- 435 [1] Loman, N.J., Quick, J., Simpson, J.T. (2015) A complete bacterial genome assembled
436 *de novo* using only Nanopore sequencing data. *Nat. Methods* **12** (8): 733-5
- 437 [2] Wick, R.R., Judd, L.M., Gorrie, C.L., Holt, K.E. (2017). Completing bacterial genome
438 assemblies with multiplex MinION sequencing, *Microb. Genom.* **3**(10): e000132
- 439 [3] Doyle, L.E., Williams, R.B.H., Rice, S.A., Marsili, E., Lauro, F.M. (2018). Draft
440 genome sequence of *Enterobacter* sp. Strain EA–1, an electrochemically active microor-
441 ganism isolated from tropical sediment, *Genome Announcements* **6**(9): e00111–18.
- 442 [4] Daebeler, A., Herbold C.W., Vierheilig, J., Sedlacek C.J., Pjevac, P., Albertsen, M.,
443 Kirkegaard, R.H., de la Torre, J.R., Daims, H., Wagner, M. (2018). Cultivation and
444 genomic analysis of “*Candidatus Nitrosocaldus islandicus*”, an obligately thermophilic,
445 ammonia–oxidizing Thaumarchaeon from a hot spring biofilm in Graendalur Valley,
446 Iceland. *Frontiers in Microbiology* **9**: 193.
- 447 [5] Frank, J., Lückner, S., Vossen, R.H.A.M., Jetten, M.S.M., Hall, R.J., Op den Camp,
448 H.J.M., Anvar, S.Y. (2018). Resolving the complete genome of *Kuenenia stuttgartiensis*
449 from a membrane bioreactor enrichment using Single–Molecule Real–Time sequencing.
450 *Scientific Reports.* **8**(1): 4580.
- 451 [6] Andersen, M.H., McIlroy, S.J., Nierychlo, M., Nielsen, P.H., Albertsen,
452 M. (2018). Genomic insights into *Candidatus Amarolinea aalborgensis* gen.
453 nov., sp. nov., associated with settleability problems in wastewater treatment
454 plants, *Systematic and Applied Microbiology*, available online 16 August 2018
455 <https://doi.org/10.1016/j.syapm.2018.08.001>
- 456 [7] Driscoll, C.B., Otten T.G., Brown, N.B., Dreher, T.W. (2017). Towards long–read
457 metagenomics: complete assembly of three novel genomes from bacteria dependent on
458 a diazotrophic cyanobacterium in a freshwater lake co–culture, *Standards in Genomic*
459 *Sciences.* **12**: 9
- 460 [8] Slaby, B.M., Hackl, T., Horn, H., Bayer, K., Hentschel, U. (2017). Metagenomic binning
461 of a marine sponge microbiome reveals unity in defense but metabolic specialization,
462 *ISME Journal*, **11**: 2465–2478.

- 463 [9] Frank, J.A., Pan, Y., Tooming–Klunderud, A., Eijsink, V.G.H., McHardy, A.C., Neder-
464 bragt, A.J. (2016). Improved metagenome assemblies and taxonomic binning using
465 long-read circular consensus sequence data, *Scientific Reports*, **6**: 25373.
- 466 [10] Singer, E., Andreopoulos, B., Bowers, R.M., Lee, J., Deshpande, S., Chiniquy, J.,
467 Ciobanu, D., Klenk, H-P., Zane, M., Daum, C., Clum, A., Cheng, J-F., Copeland,
468 A., Woyke, T. (2016) Next generation sequencing data of a defined microbial mock
469 community. *Scientific Data* **3**:160081.
- 470 [11] Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C. Franklin, R.B. (2017). Min-
471 ION nanopore sequencing of environmental metagenomes: a synthetic approach. *Giga-*
472 *Science* **6**: 1–10.
- 473 [12] Nanopore GridION and PromethION Mock Microbial Community Data Community
474 Release, Release 2 (2018-10-17). <https://github.com/LomanLab/mockcommunity>
- 475 [13] Huson, D.H., Albrecht, B., Bagci, C., Bessarab, I., Gorska, A., Jolic, D., Williams,
476 R.B.H (2018). MEGAN–LR: New algorithms allow accurate binning and easy interac-
477 tive exploration of metagenomic long reads and contigs, *Biology Direct* **13**: 6
- 478 [14] Laczny, C.C., Kiefer, C., Galata, V., Fehlmann, T., Backes, C., Keller, A. (2017).
479 BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and
480 annotation. *Nucleic Acids Res.* **45** (W1): W171–W179
- 481 [15] Liu, X.H., Arumugam, K., Natagaran, G., Seviour, T.W., Drautz-Moses, D.I., Wuertz,
482 S., Law, Y.Y., Williams, R.B.H. (2018). Draft genome sequence of an “*Candidatus Bro-*
483 *cadia*” bacterium enriched from tropical-climate activated sludge, *Genome Announce-*
484 *ments*, in press. Preprint [biorxiv.org/content/early/2017/04/24/123943](https://www.biorxiv.org/content/early/2017/04/24/123943)
- 485 [16] Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., Segata, N. (2017), Shotgun
486 metagenomics, from sampling to analysis, *Nature Biotechnology* **35**: 833–84.
- 487 [17] Bowers, R.M., Kyrpides, N.C. Stepanauskas, R., Harmon–Smith, M., Doud, D. et
488 al. (2017). Minimum information about a single amplified genome (MISAG) and a
489 metagenome–assembled genome (MIMAG) of bacteria and archaea, *Nature Biotech-*
490 *nology* **35**: 725–731.
- 491 [18] Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M.
492 (2017). Canu: scalable and accurate long–read assembly via adaptive *k*–mer weighting
493 and repeat separation, *Genome Research*, **27**(5): 722–736.
- 494 [19] Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy
495 long sequences, *Bioinformatics* **32**(14): 2103–2110.
- 496 [20] SMARTdenovo <https://github.com/ruanjue/smartdenovo>

- 497 [21] Recanati, A., Bruls, T., d’Aspremont, A. (2017). A spectral algorithm for fast *de novo*
498 layout of uncorrected long nanopore reads, *Bioinformatics* **33**(20): 3188–3194.
- 499 [22] Wick, R.R., Judd, L.M., Gorrie, C.L. Holt, K.E. (2017). Unicycler: Resolving bacterial
500 genome assemblies from short and long sequencing reads, *PLoS Computational Biology*
501 **13**(6): e1005595.
- 502 [23] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local
503 alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- 504 [24] O’Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D *et al.* (2016). Reference
505 sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and func-
506 tional annotation. *Nucleic Acids Res.* **44**(D1): D733–45
- 507 [25] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput se-
508 quencing reads. *EMBnet.journal*, **17**(1): 10–12.
- 509 [26] Xie, C., Goi, W.C., Huson, D.H., Little, P.F.R., Williams, R.B.H. (2016). RiboTagger:
510 fast and unbiased 16S/18S profiling using whole community shotgun metagenomic or
511 metatranscriptome surveys, *BMC Bioinformatics* **17**(Suppl 19): 1378
- 512 [27] Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glckner,
513 F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned
514 ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**: 7188–
515 7196.
- 516 [28] Edgar, R.C. (2017). SEARCH_16S: A new algorithm for iden-
517 tifying 16S ribosomal RNA genes in contigs and chromosomes.
518 <http://biorxiv.org/content/early/2017/04/04/124131>
- 519 [29] Pruesse, E., Peplies, J., Glckner, F.O. (2012) SINA: accurate high-throughput mul-
520 tiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829
- 521 [30] Zhu, W., Lomsadze, A., Borodovsky, M. (2010). *Ab initio* gene identification in
522 metagenomic sequences. *Nucleic Acids Research* **38**: e132.
- 523 [31] Buchfink, B., Xie, C., Huson, D.H. (2015). Fast and sensitive protein alignment using
524 DIAMOND, *Nature Methods* **12**(1): 59–60.
- 525 [32] Kang, D.D., Froula, J., Egan, R., Wang, Z. (2015). MetaBAT, an efficient tool for
526 accurately reconstructing single genomes from complex microbial communities, *PeerJ*,
527 **3**, e1165.
- 528 [33] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. (2015).
529 CheckM: assessing the quality of microbial genomes recovered from isolates, single cells,
530 and metagenomes” *Genome Research*, **25**, 1043–1055.

- 531 [34] Williams, R.B.H., Liu, X.H., Arumugam, K. (2018). Annotation statistics for ex-
532 ploratory analysis of metagenome assembly binning data, *unpublished manuscript*.
- 533 [35] Porechop. <https://github.com/rrwick/Porechop>
- 534 [36] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
535 **34** (18): 3094–3100.
- 536 [37] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
537 Abecasis, G., Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Se-
538 quence Alignment/Map format and SAMtools, *Bioinformatics*. **25**(16): 2078–2079.
- 539 [38] Olm, M.R., Brown, C.T., Brooks, B., Banfield, J.F. (2018). dRep: a tool for
540 fast and accurate genomic comparisons that enables improved genome recovery from
541 metagenomes through de-replication, *ISME J.* **11**(12): 2864–2868
- 542 [39] Miller, I.J., Rees, E.R., Ross, J., Miller, I., Baxa, J., Lopera, J., Kerby,
543 R.L., Rey, F.E., Kwan, J.C. (2018). Autometa: Automated extraction of mi-
544 crobial genomes from individual shotgun metagenomes, *bioRxiv* 251462; doi:
545 <https://doi.org/10.1101/251462>.
- 546 [40] Quince, C., Delmont, T.O., Raguideau, S., Alneberg, J., Darling, A.E., Collins,
547 G., Eren, A.M. (2017). DESMAN: a new tool for de novo extraction of strains from
548 metagenomes, *Genome Biol.* **18**(1): 181
- 549 [41] Delmont, D.O., Quince, C., Shaiber, A., Esen, O.E., Lee, S.T.M., Rappé, M.S., McLel-
550 lan, S.L., Lückner, S., Eren, A.M. Nitrogen-fixing populations of Planctomycetes and
551 Proteobacteria are abundant in surface ocean metagenomes, *Nature Microbiology* **3**:
552 804–813.
- 553 [42] Stewart, R.D., Auffret, M.D., Warr, A., Wisner, A.H., Press, M.O., Langford, K.W.,
554 Liachko, I., Snelling, T.J., Dewhurst, R.J., Walker, A.W., Roche, R., Watson, M.
555 (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow
556 rumen, *Nature Communications*, **9**: 870
- 557 [43] Ji, P., Zhang, Y.M., Wang, J.F., Zhao, F.Q. (2017). MetaSort untangles metagenome
558 assembly by reducing microbial community complexity. *Nature Communications*, **8**,
559 14306.
- 560 [44] Burton, J.N., Liachko, I., Dunham, M.J., Shendure, J. (2014). Species-level decon-
561 volution of metagenome assemblies with Hi-C-based contact probability maps. *G3*
562 (*Bethesda*), **4**(7): 1339–46
- 563 [45] Bland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement be-
564 tween two methods of clinical measurement. *Lancet* **327** (8476): 307–10.

- 565 [46] Lan, F., Demaree, B., Ahmed, N., Abate, A.R. (2017). Single-cell genome sequencing
566 at ultra-high-throughput with microfluidic droplet barcoding, *Nat. Biotechnol.* **35**(7):
567 640–646.

Table 1: Summary statistics for long read data

Measure	
Number of reads	115,383
Number of reads after adapter trimming	115,336
Total length of sequenced reads	539,410,340
Longest read (bp)	112,729
Shortest read (bp)	3
Number of reads > 500 bp	107,673
Number of reads > 1000 bp	98,574
Number of reads > 10k bp	6,324
Number of reads > 100k bp	1
Number of reads > 1M bp	0
Mean read length (bp)	4,677
N50 (bp)	6,502

Table 2: Summary statistics for long read assemblies

	Canu (genomeSize Mbp)										Unicycler	SMARTDenovo	Spectrassembler
	3.4	5	12	25	50	120	Miniasm						
Contigs (<i>N</i>)	32	143	252	298	165	159	135	130	289				
Total length (bp)	4,775,214	6,518,501	8,465,399	9,915,183	11,230,562	11,204,964	6,295,633	6,414,317	7,286,352	9,612,544			
Longest contig (bp)	3,698,683	2,169,090	1,922,826	1,906,315	2,359,014	1,288,835	1,866,536	2,602,251	2,476,359	396,957			
Shortest contig (bp)	8101	4694	1508	1121	3528	3528	895	1125	7815	9528			
Contigs>500bp (<i>N</i>)	32	143	252	298	165	159	135	130	130	289			
Contigs>1000bp (<i>N</i>)	32	143	252	298	165	159	134	130	130	289			
Contigs>10kbp (<i>N</i>)	28	107	170	214	147	142	126	123	128	288			
Contigs>100kbp (<i>N</i>)	3	5	4	7	21	22	7	4	3	15			
Contigs>1Mbp (<i>N</i>)	1	2	2	1	1	2	1	1	2	0			
Mean length (bp)	149,225	45,584	33,593	33,272	68,064	70,471	46,634	49,341	56,049	33,261			
N50	3,698,683	1,205,541	85,848	58,887	249,609	233,737	271,913	946,064	1,514,011	35,202			
Runtime (sec)	18,417	21,865	12,978	7,944	6,771	5,607	47	704	108	529			
#Cores	1	1	1	1	1	1	1	10	1	1			
#Reads used	37,952	41,868	47,097	50,220	52,005	52,396	41,570	43,277	43,354	47,812			

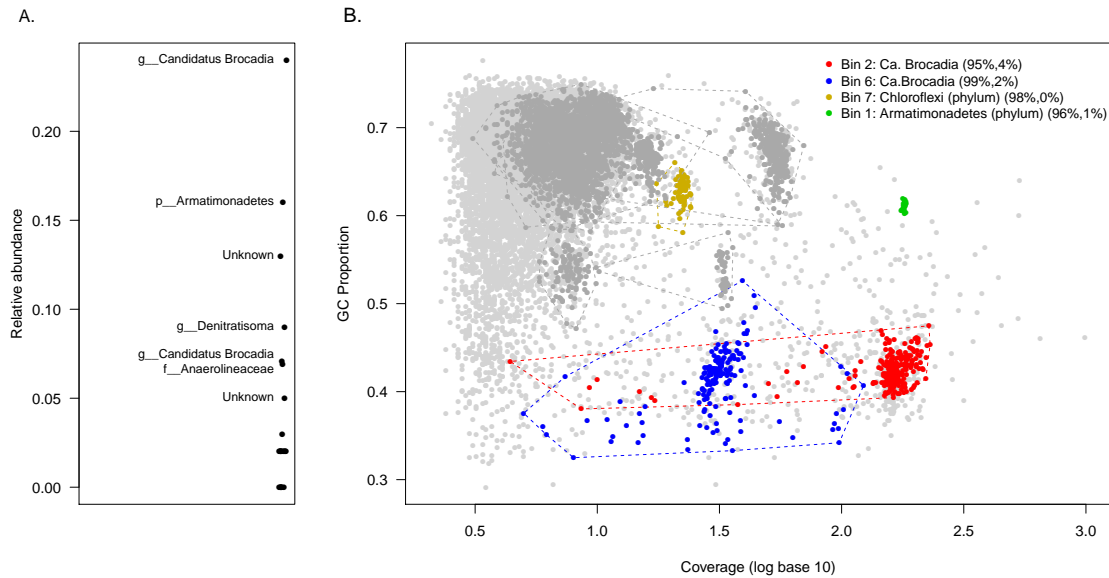


Figure 1: Analysis of short read sequencing data from AnAOB enrichment reactor community. *A*: RiboTagger analysis of short read data, showing taxon ranked by relative abundance. *Unknown* denotes cases in which RiboTagger could not assign any annotation to the corresponding ribotag sequence. A total of 23 taxa were detected, of which 7 had a relative abundance ≥ 0.05 (named). A member of genus *Candidatus Brocadia*, a known AnAOB, was the most abundant community member, with a second member of this genus being present at a lower abundance. *B*: Single-sample metgenome assembly of the short read sequence, constructed using *Newbler* v2.9. GC-proportion and log coverage are used to visualise contigs here. Contigs were binned using *MetaBAT*. A total of 16 bins were detected, of which 4 showed *CheckM* defined completeness $\geq 95\%$ and contamination $\leq 5\%$ (contigs shown in colours and described in legend: numbers in brackets show estimated contamination and completeness). Evidence supporting taxonomic assignments to bin is made on the basis of presence of full length, annotated 16S sequences, marker gene analysis in *CheckM* and ORF-level annotation statistics (see main text for details).

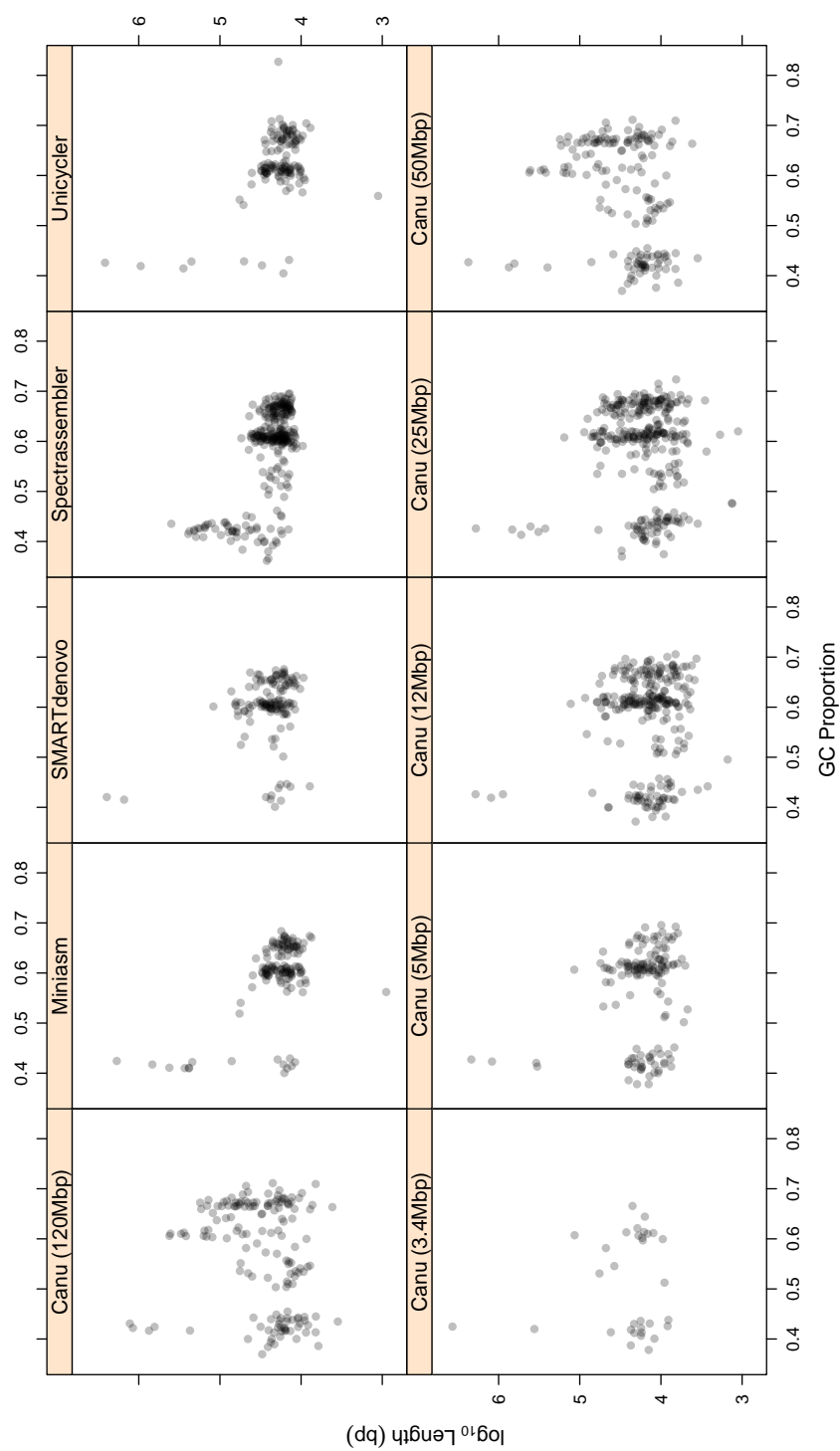


Figure 2: Length distribution of LRAC sequences from each of the five long read assemblers employed in this study

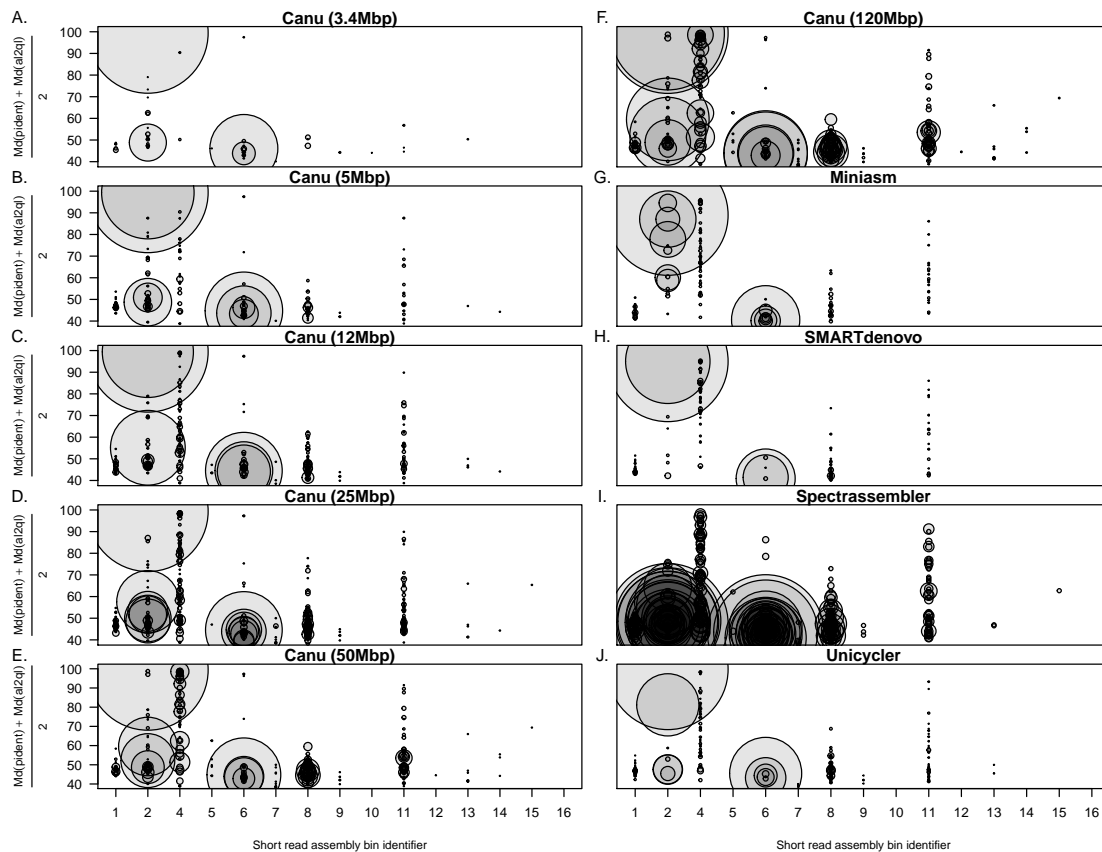


Figure 3: Summary of BLASTN alignment statistics between SRAC sequences and LRAC sequences categorised by bin membership of the SRAC sequences. In each panel, the horizontal axis lists SRAC bin identifiers (see **Figure 1** and **Table S1**). Using BLASTN, SRAC sequences were treated as queries and LRAC sequences as the subject (see **Methods: Comparative analysis of long and short read assemblies**). Each data point corresponds an LRAC sequence with diameter proportional the mean of the median of the distribution of the percent identity (*pident*) and the median of the distribution of the alignment length to query (SRAC) length (*al2ql*; expressed as a percentage to be on the same scale as *pident*). Note that the LRAC sequences show a large number of highly aligned SRAC sequences that arising from the *Ca. Brocadia* genome from bin 2, with a lesser degree of alignment evident from the second *Ca. Brocadia* genome captured by SRAC sequence from bin 6.

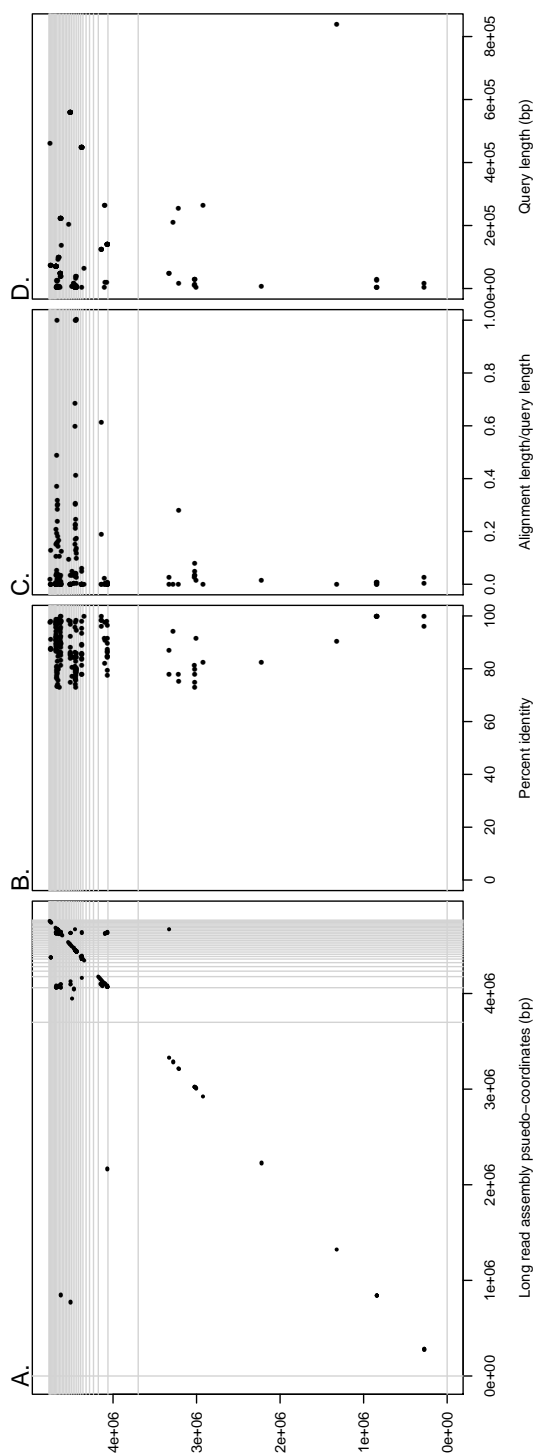


Figure 4: Alignment of SRAC and LTRAC sequence for LTRAC for SRAC contigs that are **not member of either AnAOB bin, using all alignments returned by BLASTN**. *A*). Dot-plot showing LTRAC assembly in pseudo-coordinates, ordered by LTRAC contig length in increasing order. Vertical and horizontal lines delineate individual contigs. The first two LTRAC contigs are `tig000000001` and `tig000000002`. Black segments denote individual alignments; *B*) percent identity statistics for aligned sequences. Magnitude of percent identity (*pident*) is plotted on *x*-axis with alignment location on the LTRAC assembly statistics for aligned sequences. Magnitude of query length is plotted on *x*-axis with alignment location on the LTRAC assembly plotted on the *y*-axis.

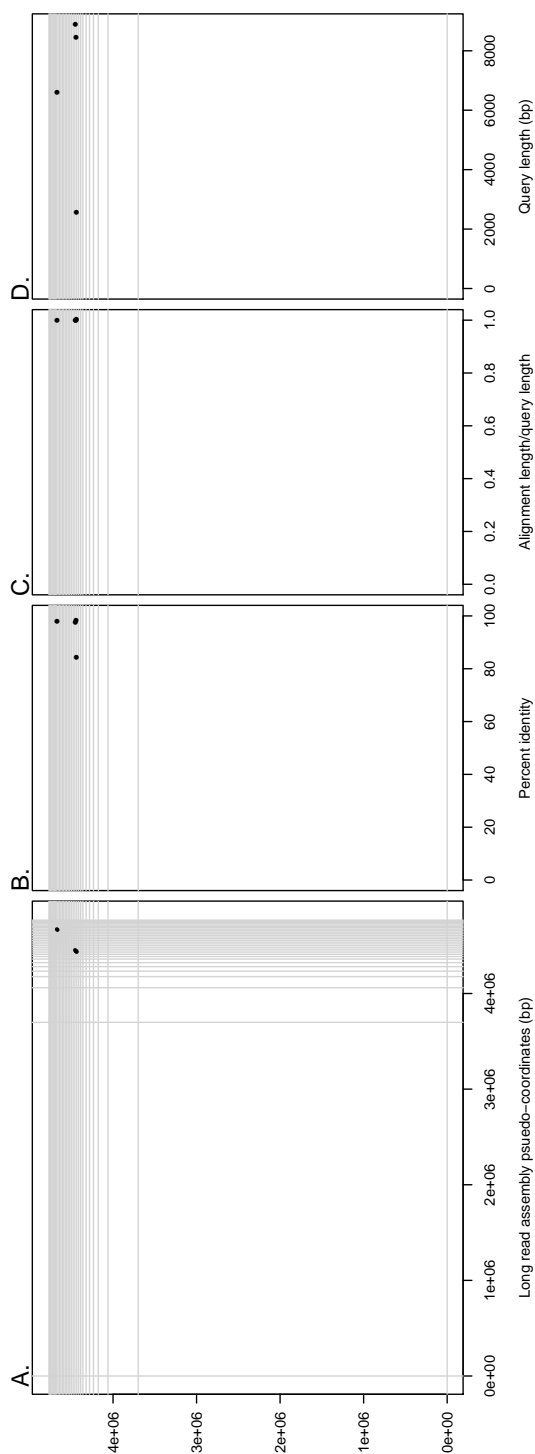


Figure 5: Alignment of SRAC and LRAC sequence for LRAC for SRAC contigs that are **not member of either AnAOB bin**, using **only full length alignments** ($al2ql > 95\%$). *A*). Dot-plot showing LRAC assembly in pseudo-coordinates, ordered by LRAC contig length in increasing order. Vertical and horizontal lines delineate individual contigs. The first two LRAC contigs are `ti000000001` and `ti000000002`. Black segments denote individual alignments; *B*) percent identity statistics for aligned sequences. Magnitude of percent identity (*pidcnt*) is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis; *C*) query length statistics for aligned sequences. Magnitude of query length is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis.

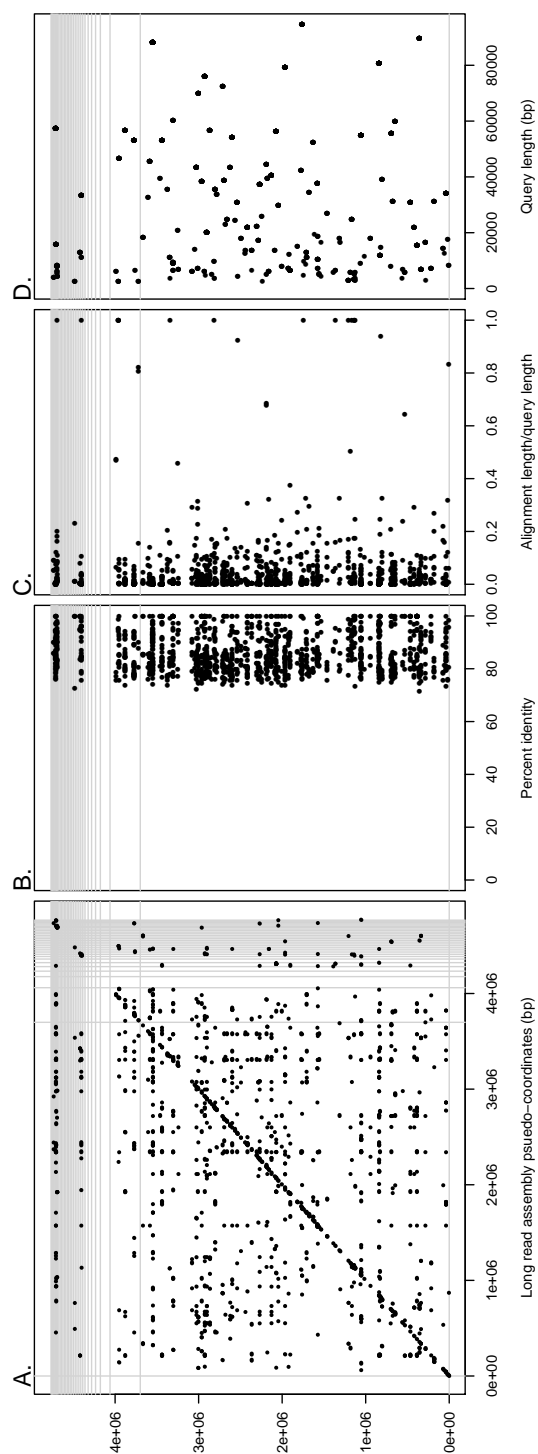


Figure 6: Alignment of SRAC and LTRAC sequence for SRAC for SRAC contigs that are members of AnAOB bin 6, using all alignments returned by BLASTN. A) Dot-plot showing LTRAC assembly in pseudo-coordinates, ordered by LTRAC contig length in increasing order. Vertical and horizontal lines delineate individual contigs. The first two LTRAC contigs are `tig00000001` and `tig00000002`. Black segments denote individual alignments; B) percent identity statistics for aligned sequences. Magnitude of percent identity (*percent*) is plotted on *x*-axis with alignment location on the LTRAC assembly plotted on the *y*-axis; C) query length statistics for aligned sequences. Magnitude of query length is plotted on *x*-axis with alignment location on the LTRAC assembly plotted on the *y*-axis.

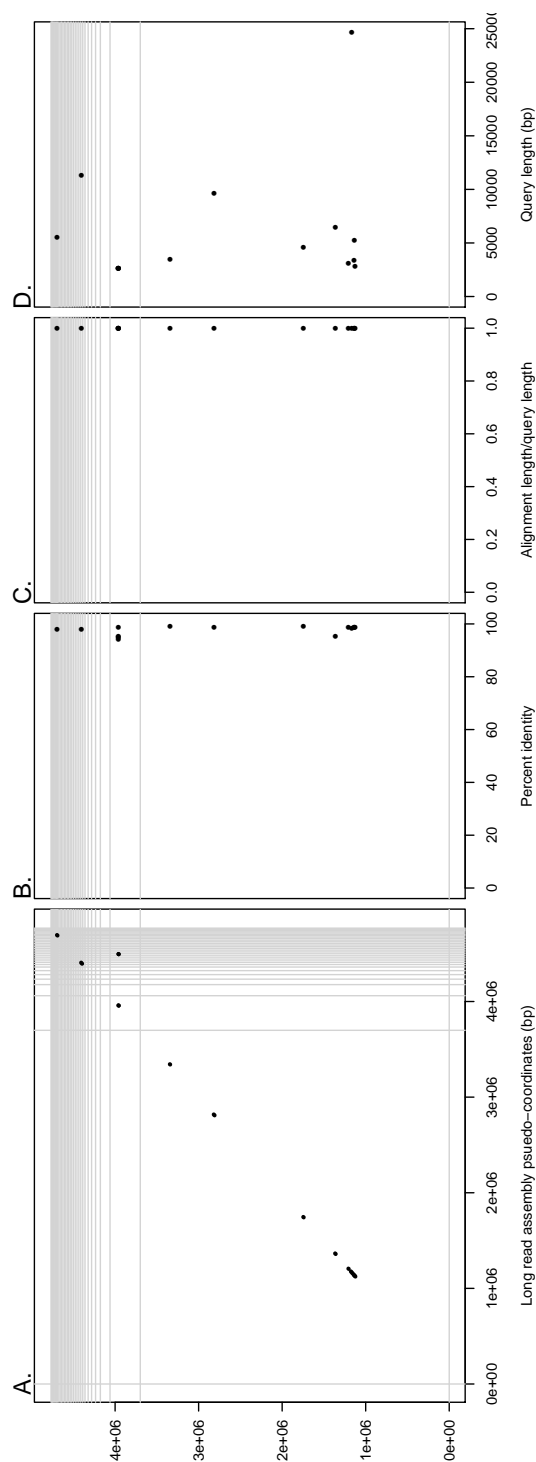


Figure 7: Alignment of SRAC and LRAC sequence for LRAC for SRAC contigs that are **members of AnAOB bin 6**, using **only full length alignments** ($al2ql > 95\%$). *A*). Dot-plot showing LRAC assembly in pseudo-coordinates, ordered by LRAC contig length in increasing order. Vertical and horizontal lines delineate individual contigs. The first two LRAC contigs are *tig000000001* and *tig000000002*. Black segments denote individual alignments; *B*) percent identity statistics for aligned sequences. Magnitude of percent identity (*pidcnt*) is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis; *C*) query length statistics for aligned sequences. Magnitude of query length is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis.

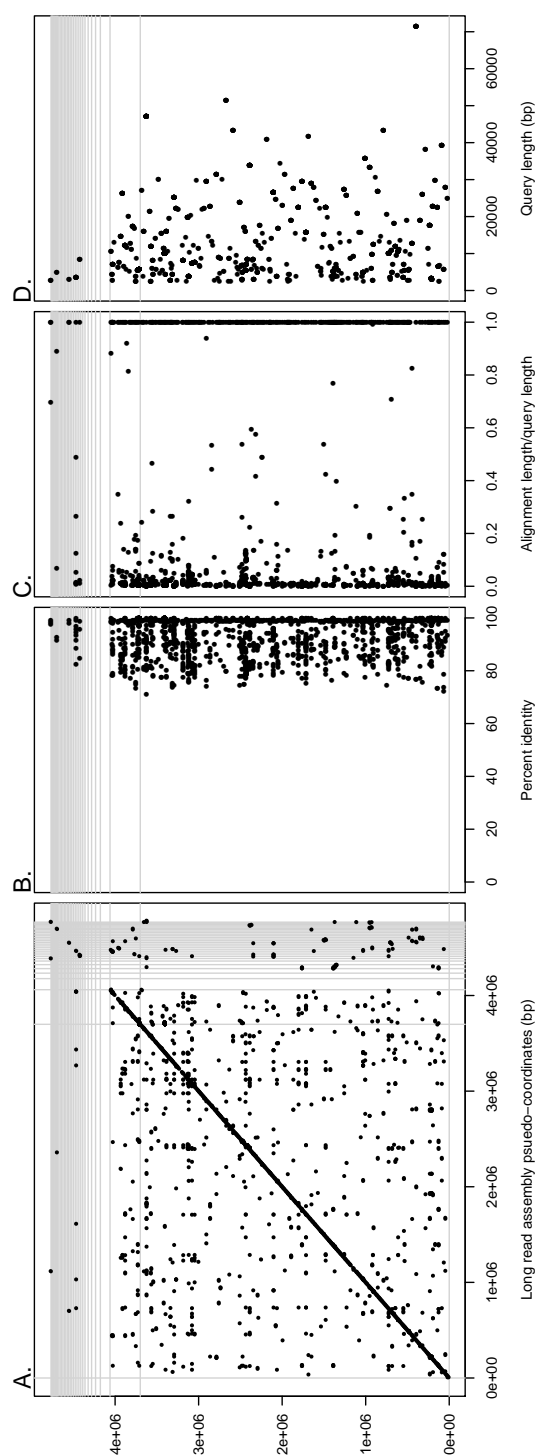


Figure 8: Alignment of SRAC and LRAC sequence for LRAC for SRAC contigs that are members of AnAOB bin 2, using all alignments returned by BLASTN. A). Dot-plot showing LRAC assembly in pseudo-coordinates, ordered by LRAC contig length in increasing order. Vertical and horizontal lines delineate individual contigs. The first two LRAC contigs are `tig00000001` and `tig00000002`. Black segments denote individual alignments; B) percent identity statistics for aligned sequences. Magnitude of percent identity (*percent*) is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis; C) query length statistics for aligned sequences. Magnitude of query length is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis.

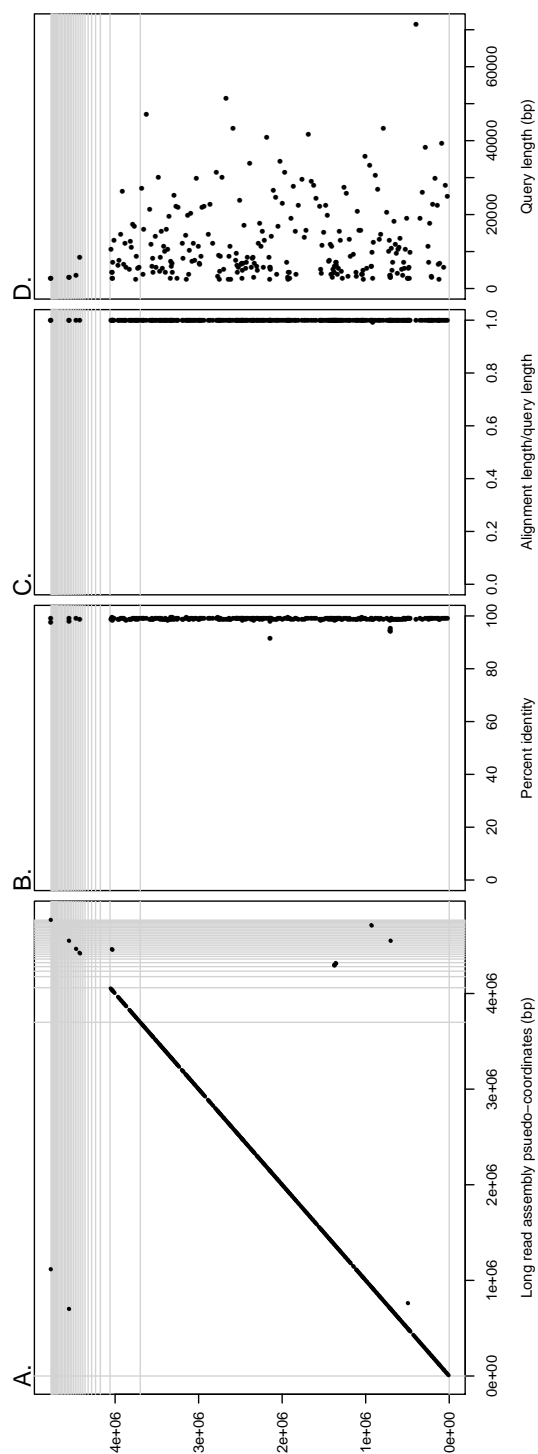


Figure 9: Alignment of SRAC and LRAC sequence for LRAC for SRAC contigs that are members of AnAOB bin 2, using only full length alignments ($al2ql > 95\%$). A) Dot-plot showing LRAC assembly in pseudo-coordinates, ordered by LRAC contig length in increasing order. Vertical and horizontal lines delineate individual contigs. The first two LRAC contigs are `tig000000001` and `tig000000002`. Black segments denote individual alignments; B) percent identity statistics for aligned sequences. Magnitude of percent identity (*percent*) is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis; C) query length statistics for aligned sequences. Magnitude of query length is plotted on *x*-axis with alignment location on the LRAC assembly plotted on the *y*-axis.

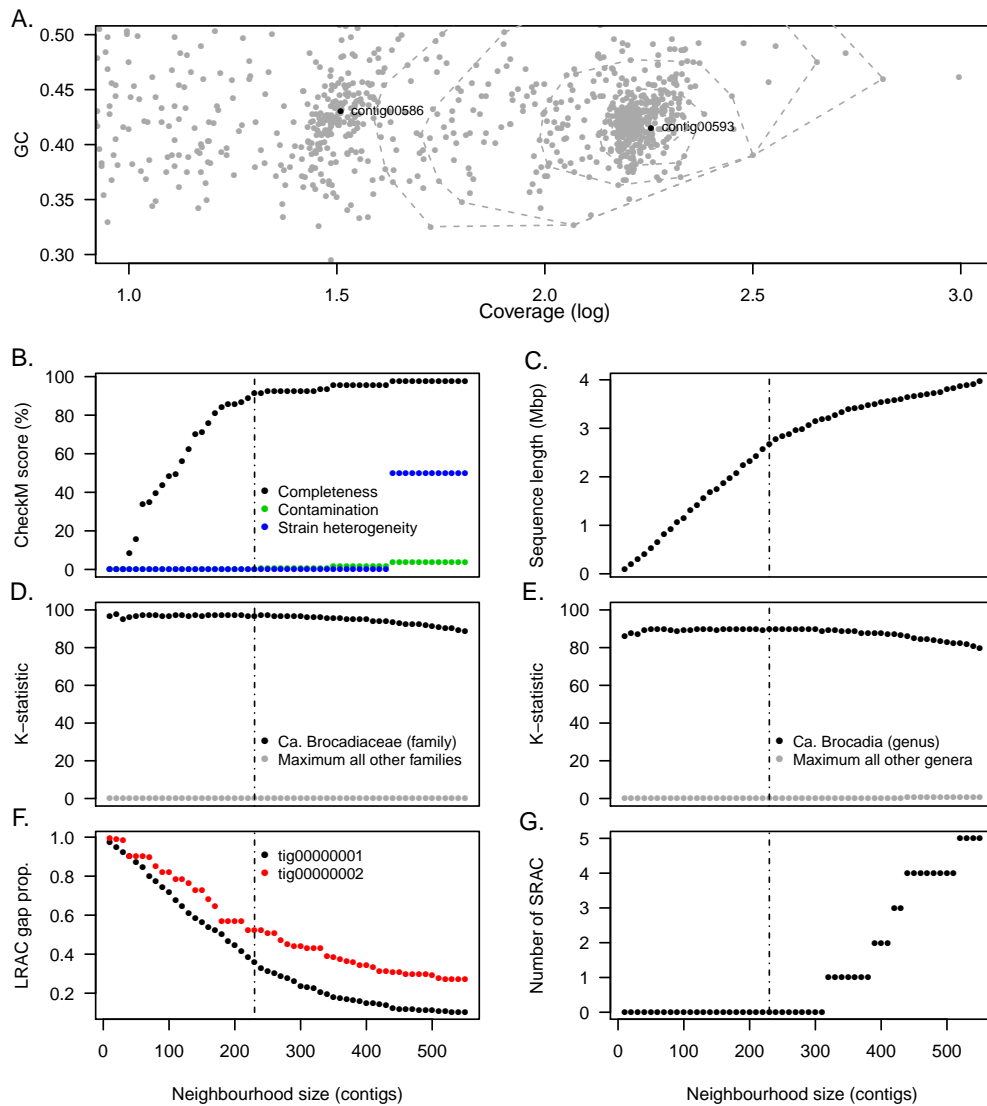


Figure 10: Re-analysis of AnAOB bin 2 using k -nearest neighbour contig sets (neighbourhoods) and alignments to LRAC sequence. A). Coverage (\log_{10})-GC plane showing all contigs with length $\geq 1\text{ kbp}$. Contig neighbourhoods (dashed convex hulls) are defined using the k -nearest neighbours of contig00593 which harbours a full length 16S gene annotated to *Ca. Brocadia*. In A)–G, the value of k is plotted on the x -axis with the following statistics on the y -axis: B) CheckM genome quality statistics (completeness, contamination and strain heterogeneity); C) total sequence length (bp); D) ORF-level annotation (K) statistics summarised at family level for *Ca. Brocadiaceae* (black) and maximum observed over all other families (grey) E) ORF-level annotation (K) statistics summarised at genus level for *Ca. Brocadia* (black) and maximum observed over all other genera (grey); F) the proportion of LRAC sequence of `tig00000001` and `tig00000002` that contain no aligned SRAC sequence from bin 2; G) the number of bin 2 SRAC sequenced that were unaligned to LRAC sequences. When calculating gaps in the alignments to LRAC sequence, we only considered full length SRAC aligned sequences (*i.e.* those with $al2ql > 0.95$).

Table S1: Summary statistics for the short read assembly

<u>Measure</u>	
Contigs (<i>N</i>)	38,988
Total length (bp)	75,821,912
Longest contig (bp)	838,453
Shortest contig (bp)	500
Contigs>500 bp (<i>N</i>)	38,910
Contigs>1000 bp (<i>N</i>)	14,450
Contigs>10k bp (<i>N</i>)	813
Contigs>100k bp (<i>N</i>)	37
Contigs>1M bp (<i>N</i>)	0
Mean length (bp)	1,945
N50	4,145

Table S2: Summary of phyla-level analysis of long read data using MEGAN-LR

Taxonomic path (domain, phylum)	Number of alignments	Proportion of total (%)
Bacteria;PVC group;Planctomycetes	1.57E+08	53.94
Bacteria;Proteobacteria	6.97E+07	23.97
Bacteria;Terrabacteria group;Chloroflexi;	2.10E+07	7.21
Bacteria;Terrabacteria group;Armatimonadetes;	1.79E+07	6.14
Bacteria;FCB group;Bacteroidetes/Chlorobi group;Bacteroidetes;	1.24E+07	4.26
Bacteria;Nitrospirae;	3.27E+06	1.12
Bacteria;Terrabacteria group;Actinobacteria <phylum>;	2.72E+06	0.93
Bacteria;Acidobacteria;	1.96E+06	0.67
Bacteria;Terrabacteria group;Cyanobacteria/Melainabacteria group;Cyanobacteria;	1.29E+06	0.44
Bacteria;FCB group;Bacteroidetes/Chlorobi group;Ignavibacteriae;	1.25E+06	0.43
Bacteria;PVC group;Candidatus Omnitrophica;	7.34E+05	0.25
Bacteria;unclassified Bacteria;unclassified Bacteria (miscellaneous);	5.99E+05	0.21
Bacteria;Terrabacteria group;Firmicutes;	4.77E+05	0.16
Bacteria;FCB group;Bacteroidetes/Chlorobi group;Chlorobi;	3.99E+05	0.14
Eukaryota;Opisthokonta;Metazoa;Eumetazoa;	1.75E+05	0.06
Bacteria;unclassified Bacteria;Bacteria candidate phyla;Patescibacteria group;	1.75E+05	0.06

Table S3: Summary of genera-level analysis of long read data using MEGAN-LR

Taxonomic path (phyla, family, genus)	Number of alignments	Proportion of total (%)
Planctomycetes; <i>Candidatus</i> Brocadiaceae; <i>Candidatus</i> Brocadia;	1.51E+08	93.29
Nitrospirae; Nitrospiraceae; Nitrospira;	3.20E+06	1.98
Chloroflexi; Thermoflexaceae; Thermoflexus;	1.14E+06	0.70
Proteobacteria; Nitrosomonadaceae; Nitrosomonas;	1.02E+06	0.63
Proteobacteria; unclassified Rhizobiales; Pseudorhodoplanae;	7.89E+05	0.49
Proteobacteria; Sandaracinaceae; Sandaracinus;	7.19E+05	0.44
Proteobacteria; Burkholderiaceae; Lautropia;	6.83E+05	0.42
Cyanobacteria/Melainabacteria group; Leptolyngbyaceae; Leptolyngbya;	5.75E+05	0.36
Chloroflexi; Caldilineaceae; Caldilinea;	4.28E+05	0.26
Proteobacteria; Phyllobacteriaceae; Mesorhizobium;	4.22E+05	0.26
Proteobacteria; Hyphomicrobiaceae; Hyphomicrobium;	3.45E+05	0.21
Chloroflexi; Ardentiacatenaceae; <i>Candidatus</i> Promineofilum;	2.79E+05	0.17
Proteobacteria; Bradyrhizobiaceae; Bradyrhizobium;	2.48E+05	0.15
Proteobacteria; Enterobacteriaceae; Escherichia;	2.24E+05	0.14
Proteobacteria; Burkholderiales Genera incertae sedis; Rubrivivax	2.07E+05	0.13
Chloroflexi; Roseiflexaceae; Roseiflexus;	2.00E+05	0.12
Planctomycetes; Phycisphaeraceae; Phycisphaera;	1.98E+05	0.12
Proteobacteria; unclassified Betaproteobacteria; <i>Candidatus</i> Accumulibacter;	1.73E+05	0.11

Table S4: CheckM genome quality statistics for short read assembly bins

Bin	Lineage	Completeness	Contamination	Strain heterogeneity	\widehat{Cov}	\widehat{GC}
8	k__Bacteria (UID203)	99.69	149.37	0.93	9	0.70
6	k__Bacteria (UID2565)	98.84	1.65	0.00	31	0.41
7	k__Bacteria (UID1452)	98.18	0.00	0.00	22	0.63
1	k__Bacteria (UID1452)	95.83	0.93	0.00	180	0.61
2	k__Bacteria (UID2565)	95.40	3.85	50.00	150	0.42
5	k__Bacteria (UID1452)	94.55	2.27	0.00	31	0.53
11	k__Bacteria (UID203)	85.34	26.72	28.30	9	0.68
4	k__Bacteria (UID203)	75.86	0.00	0.00	44	0.67
9	k__Bacteria (UID2565)	62.52	10.23	0.00	8	0.64
13	k__Bacteria (UID1452)	62.01	16.86	3.85	6	0.66
10	k__Bacteria (UID1452)	58.60	0.99	0.00	8	0.54
16	k__Bacteria (UID203)	3.26	0.00	0.00	7	0.65
3	root (UID1)	0.00	0.00	0.00	142	0.39
15	root (UID1)	0.00	0.00	0.00	8	0.59
14	root (UID1)	0.00	0.00	0.00	7	0.70
12	root (UID1)	0.00	0.00	0.00	7	0.65

\widehat{Cov} : mean coverage of member contigs
 \widehat{GC} : mean GC content of member contigs

Table S5: K -statistics for short read assembly bins at phylum, family and genera levels

Bin	K_{max} (phylum)	Phylum	K_{max} (family)	Family	K_{max} (genus)	Genus
8	70.50	Proteobacteria	2.0	Bradyrhizobiaceae	0.6	Bradyrhizobium
6	66.6	Planctomycetes	66.4	Candidatus Brocadiaceae	38.7	Candidatus Brocadia
7	94.0	Chloroflexi	0.2	Anaerolineaceae	0.1	Ruminiclostridium
1	83.2	Armatimonadetes	6.3	Fimbrimonadaceae	6.3	Fimbrimonas
2	96.2	Planctomycetes	96.2	Candidatus Brocadiaceae	87.3	Candidatus Brocadia
5	32.9	Chloroflexi	1.0	Anaerolineaceae	0.8	Ardenticatena
11	71.3	Proteobacteria	6.7	Rhodocyclaceae	1.2	Sulfuritalea
4	79.1	Proteobacteria	15.2	Rhodocyclaceae	3.5	Sulfuritalea
9	9.7	Proteobacteria	2.0	Planctomycetaceae	0.4	Plesiocystis
13	7.6	Proteobacteria	0.4	Polyangiaceae	0.4	Candidatus Entotheonella
10	3.3	Candidatus Ulrbacteria	0.5	Natrialbaceae	0.5	Thermodesulfator
16	50.0	Proteobacteria	2.5	Comamonadaceae	1.9	Candidatus Accumulibacter
3	94.0	Planctomycetes	94.0	Candidatus Brocadiaceae	88.5	Candidatus Brocadia
15	73.0	Proteobacteria	8.6	Bradyrhizobiaceae	3.0	Bradyrhizobium
14	81.3	Proteobacteria	3.4	Comamonadaceae	1.9	Methylibium
12	78.7	Proteobacteria	3.8	Comamonadaceae	3.8	Ideonella

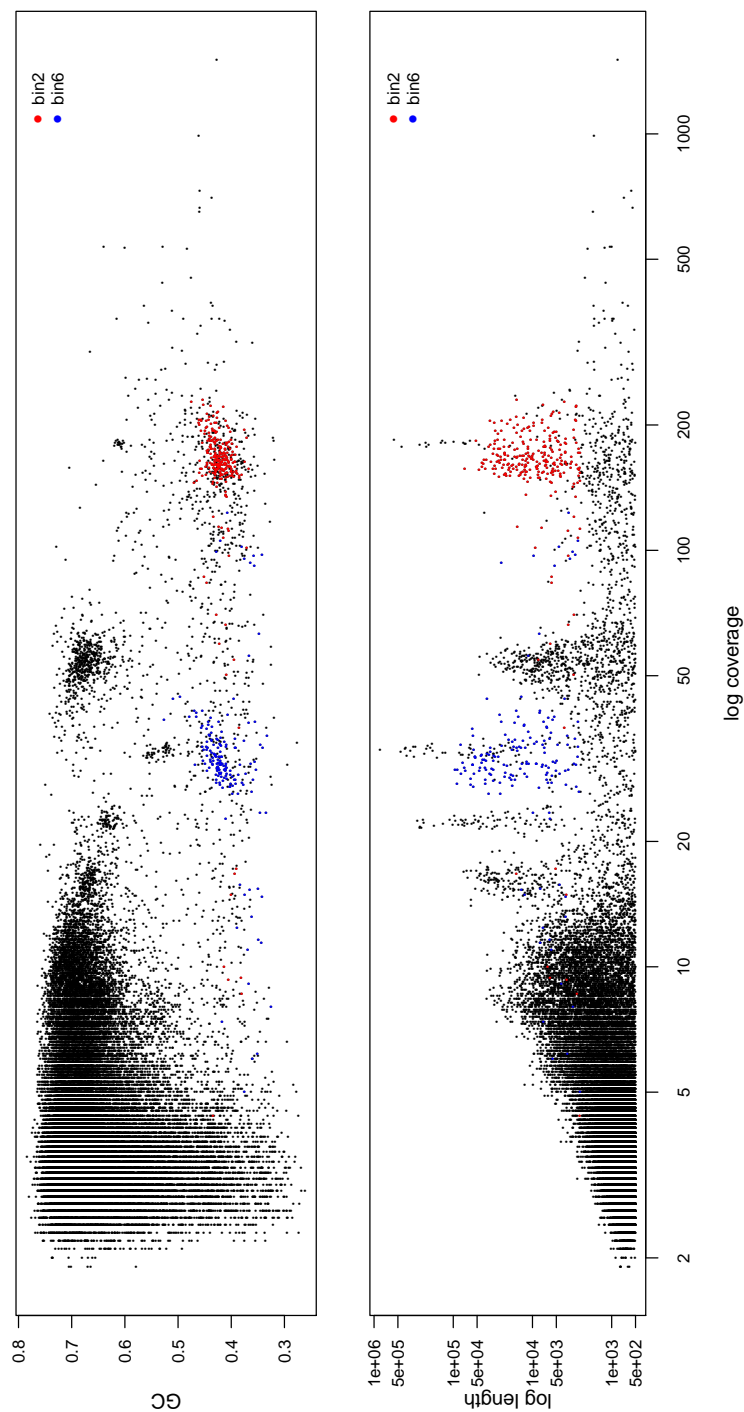


Figure S1: Inter-relationship between contig GC content, coverage and length in the short read assembly. Each datapoint is a contig. *Upper panel*: GC-proportion plotted against log coverage; *Lower panel*: Log length plotted against log coverage. In both plots, contigs that are members of bin 2 are highlighted in red and contigs that are members of bin 6 are highlighted in blue.

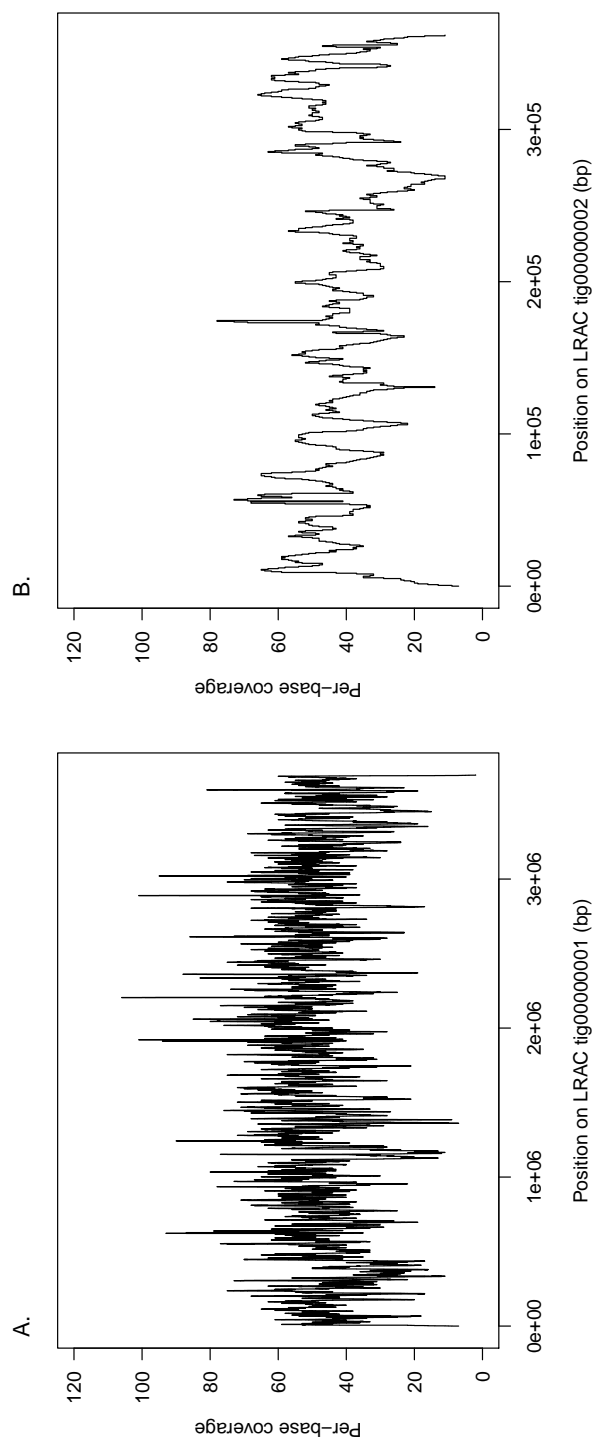


Figure S2: Smoothed per-base long read coverage for LTRAC sequences A *tig000000001*; and B: *tig000000002*. A running median has been applied to the raw per-base coverage data using a 1001 bp length window, and then the smoothed data has been subsampled at an interval of 500 basepairs