Assemblies of long-read metagenomes suffer from diverse errors

Florian Trigodet[1,2], Rohan Sachdeva[3], Jillian F. Banfield[3,4,5,6,7,*], A. Murat Eren[1,2,8,9,10,*]

[1]Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany; [2]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany; [3]Innovative Genomics Institute, University of California Berkeley, Berkeley, CA, USA; [4]Department of Earth and Planetary Sciences, University of California, Berkeley, CA, USA; [5]Department of Environmental Science Policy, and Management, University of California, Berkeley, CA, USA; [6]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [7]Biomedicine Discovery Institute, Monash University, Australia. [8]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany; [9]Marine Biological Laboratory, Woods Hole, MA, USA; [10]Max Planck Institute for Marine Microbiology, Bremen, Germany.

* Corresponding authors: meren@hifmb.de, jbanfield@berkeley.edu

Abstract

Genomes from metagenomes have revolutionised our understanding of microbial diversity, ecology, and evolution, propelling advances in basic science, biomedicine, and biotechnology. Assembly algorithms that take advantage of increasingly available long-read sequencing technologies bring the recovery of complete genomes directly from metagenomes within reach. However, assessing the accuracy of the assembled long reads, especially from complex environments that often include poorly studied organisms, poses remarkable challenges. Here we show that erroneous reporting is pervasive among long-read assemblers and can take many forms, including multi-domain chimeras, prematurely circularized sequences, haplotyping errors, excessive repeats, and phantom sequences. Our study highlights the need for rigorous evaluation of the algorithms while they are in development, and options for users who may opt for more accurate reads than shorter runtimes.

Introduction

By enabling massively parallel sequencing of DNA molecules, the second generation sequencing technologies that emerged in the early 2000s had a remarkable impact on all disciplines of life sciences (Shendure and Ji 2008; van Dijk et al. 2014). A crucial advancement brought by this technology in microbiology was the ability to reconstruct microbial genomes directly from environmental 'metagenomes' without cultivation (Tyson et al. 2004), which substantially enhanced our understanding of microbial diversity and function and continues to be an indispensable tool for biotechnology and biomedicine (Eren and Banfield 2024). Yet, the relatively short sequencing reads produced by the second generation sequencing posed significant limitations on assembly (Olson et al. 2019; Tørresen et al. 2019), a critical computational step in genome recovery workflows where reads are stitched together to rebuild contiguous segments of DNA (contigs) prior to binning, and often led to highly fragmented and sometimes highly contaminated genomes from metagenomes (Shaiber and Eren 2019; Chang et al. 2024; Chen et al. 2020).

The emergence of the third generation sequencing technologies, such as those implemented by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), marks another breakthrough in genome-resolved metagenomics through ultra-long (Jain et al. 2018) and increasingly accurate reads that can help solve complex genomic puzzles with unprecedented precision (Wenger et al. 2019). These new opportunities rely heavily on the advancement of long-read assembly algorithms, an area of active research with multiple successful software products that can assemble long-reads into complete chromosomes (Feng et al. 2022; Benoit et al. 2024; Kolmogorov et al. 2020; Nurk et al. 2020). However, assessing the accuracy of the assembled long reads poses remarkable challenges, especially from complex environments that often include poorly studied organisms but host substantial novel diversity.

The benchmarking of assembly algorithms typically relies on a few key principles, including the use of mock or simulated datasets, evaluation of contig length distributions, assessment of unassembled read fractions, genes, functions, and comparisons with other assemblers (Mikheenko, Saveliev, and Gurevich 2016; Meyer et al. 2022; Delgado and Andersson 2022). While these strategies are critical for algorithm development, benchmarking efforts rarely take into account how individual reads align to assemblies and quantify the rate, origins, and impact of mismatches between original reads and final contigs. Here we assessed the extent of agreement between individual high-fidelity PacBio reads and assembly results reported by four state-of-the-art long-read assemblers, and observed a wide range of issues, including haplotyping errors, chimerism, premature circularization, and regions of contigs that are not supported by an of the input sequences. Based on our prior experience (e.g., (Schoelmerich et al. 2024)), assemblies of ONT long reads exhibit similar errors, yet we limit our benchmarks to PacBio given the relatively higher base accuracy of PacBio reads (~99.95%, or 5 errors per 10 kbp) compared to the ONT reads (~99%, or 100 errors per 10 kbp). Overall, our survey offers reproducible means to identify long-read assembly errors and insights into their downstream implications for researchers who develop or use long-read assembly algorithms to consider.

Results

Our study benchmarks four state-of-the-art long-read assembly software, (1) HiCanu v2.2 (Nurk et al. 2020), (2) hifiasm-meta v0.3 (Feng et al. 2022), (3) metaFlye v2.9.5 (Kolmogorov et al. 2020) and (4) metaMDBG v1 (Benoit et al. 2024), based on their assembly performance of 21 PacBio HiFi metagenomes (Supplementary Table 1). Thirteen of these metagenomes have been used to evaluate one or more of the assembly algorithms, and represent mock communities and microbiomes from anaerobic digesters, and gut samples from humans, chickens, and sheep. The remaining 8 metagenomes are novel HiFi metagenomes from the surface ocean, a key biome

that was not included in the benchmarking of the long-read assembly algorithms and represents an extremely complex ecosystem compared to the human gut with little genomic representation (Schechter et al. 2025) (Supplementary Table 1).

Investigations of the agreement between individual long reads and assemblies they support require the alignment of individual reads to resulting contigs. For this task our study employs minmap2, a popular and high-performance mapping software designed to align long-reads to reference sequences (Li 2018). Unlike short-read alignment software such as Bowtie2 (Langmead and Salzberg 2012) or BWA (Li and Durbin 2009), minimap2 allows premature ends in read alignments and can map the remainder of the read to another reference location. This so-called 'read clipping' procedure is a critical tool to identify locations in the final assembly that is poorly supported by individual long-reads. To gain quantitative insights into assembly artifacts we developed 'anvi-script-find-misassemblies', a script that comprehensively summarizes locations and frequencies of read clipping events, where long reads are split systematically by the mapping software to maximize agreement between reads and assemblies, in addition to regions in contigs that are not supported by individual long reads.

Assembly errors are common to all long-read assemblers

Instances where the vast majority of long-reads are clipped at the same nucleotide position is strong evidence that the final assembled sequence contains an error. However, quantifying the number of errors in a given assembly is not as straightforward as counting the clipping events since a single assembly error may result in multiple clipping events associated with neighboring nucleotide positions in the reference sequence (Figure 1A). To minimize false positives in our results, and exclude clipping events due to within-population biological variation, our analyses below primarily focus on locations in long-read assemblies with 100% clipping rate.

97 The assembly metrics differed greatly between sample types and assembly algorithms. For
98 instance, all human gut metagenomes were assembled to a similar total size by all assemblers,
99 but the assembly size differed vastly between assemblers with surface seawater samples
100 (Supplementary Table 2), metaMDBG producing on average 610% more than HiCanu. We found
101 notable read clipping events, defined by at least 10 clipped long-reads, at least in one sample
102 from each assembler (Supplementary Table 2). However, the frequency of these events
103 normalized by the assembly size revealed remarkable differences between individual software as
104 metaFlye and metaMDBG generated up to 180 times more clipping events compared HiCanu and
105 hifiasm-meta for the same samples (Figure 1B). The number of clipping events were particularly
106 high in the surface seawater samples, where assemblies from metaMDBG had over three orders
107 of magnitude more clipping events compared to hifiasm-meta (Supplementary Table 2). Overall,
108 we found clipping events affected up to 5.6% of contigs longer than 10,000 nucleotides reported
109 by metaMDBG. We also computed the number of regions longer than 1,000 bp with no apparent
110 coverage by individual long reads and found that the occurrence of regions that are not supported
111 by any read were as pervasive as clipping events. While this issue also affected all assemblers
112 (Figure 1C, Supplementary Table 2), metaMDBG led the pack with up to 5.3% of all contigs longer
113 than 10,000 nucleotides with zero-coverage regions.

114 Reporting of contig circularity was a common feature to all assemblers, however, the number of
115 circular contigs in final assemblies varied vastly between different software (Figure 1H,
116 Supplementary Table 2). MetaMDBG generated substantially more circular contigs than the other
117 assemblers, notably in surface seawater metagenomes. Overall, we found at least one clipping
118 event for a large proportion of circular contigs reported by metaMDBG (Figure 1H), which, in some
119 cases represented up to 77% of the circular contigs in a sample (e.g., the metaMDBG assembly
120 of the sample HADS 013, Supplementary Table 2).

121 In addition to the clipping events, we reported the frequency of single-nucleotide variants (SNVs)
122 and insertion/deletion events (INDELs) in the assembled contigs that were either (1) represented
123 in only a minority of long reads. In these cases, the final assembly included a nucleotide or INDEL
124 that did not match to the most frequent nucleotide or INDEL in long reads (Figure 1A, D, F), or (2)
125 represented by none of the long reads, a more severe case where the final assembly included a
126 nucleotide or INDEL that did not occur in any of the long reads matching that context (Figure 1A,
127 E, F). After normalizing based on the assembly size, HiCanu and hifiasm-meta had the most
128 SNVs and INDELs that represented only a minority of long reads at a given position, while
129 metaFlye and metaMDBG had the most SNVs and INDELs that were not supported by any long
130 reads. The latter case affected thousands of genes in all assemblies (Supplementary Table 2),
131 leading to genes with incorrect amino acid sequences due to the impact of INDELs on open
132 reading frames. Our detailed investigation of individual clipping events associated their
133 occurrence with a few recurring classes of erroneous reporting of contigs, including chimeras,
134 premature circularization, haplotyping issues, false duplications, and non-existent sequences, for
135 which we offer an incomplete list of examples below.

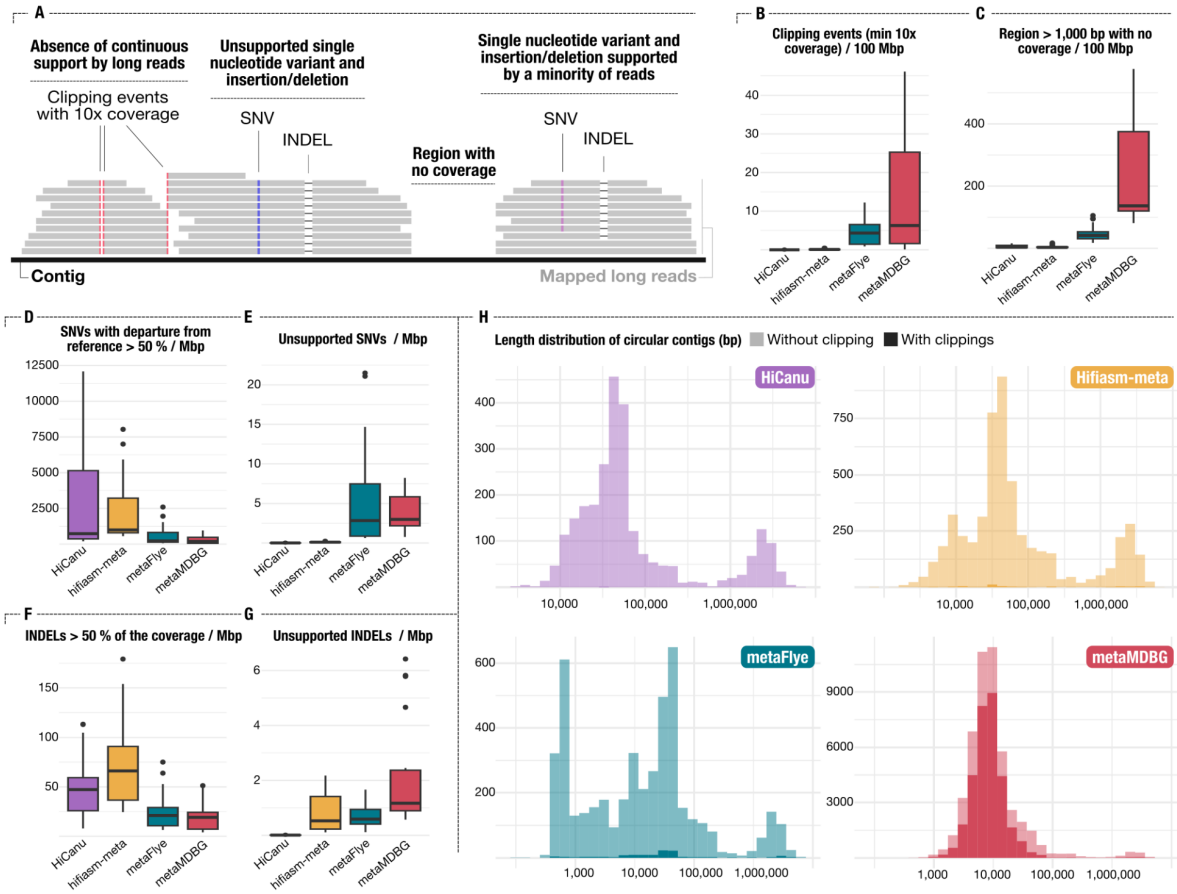


Figure 1. Assembly errors across assembler. (A) A schematic representation of long reads mapping to a contig with multiple types of read disagreement with the reference, including indel and single nucleotide variants representing more than half or all the coverage, and clipping events spanning the entire coverage. All metrics for B, C, D, E, F, G are normalized by assembly size and exclude the two mock community metagenomes. (B) Number of clipping events supported by at least 10 reads. (C) Number of regions over 1,000 bp with no apparent coverage. Number of single-nucleotide variants representing (D) > 50% or (E) all the coverage (G) at a given locus. Distribution of (F) indels > 50% of the coverage, or (G) all the coverage. (H) Length distribution of circular contigs by each assembler. The darker color represents the distribution of circular contigs with at least one clipping event.

Chimeric contigs

Our inspection of contigs with a high proportion of read clipping events revealed chimeric contigs. In some cases, chimeras brought together sequences from taxa that belonged to distinct phyla (Figure 2). In many examples, chimeras brought together sequences from two distinct taxa, but cases with sequences from more than three organisms were not uncommon (Figure 2). In other cases, chimeras combined sequences from distinct domains of life. In one extreme case, metaMDBG formed a sequence that included regions from organisms that originated from Euryarchaeota, Pseudomonadota, Bacteriodota, and Cyanobacteria (Figure 2). We acknowledge that the list of contigs we surveyed here for chimerism is far from exhaustive, and furthermore the reliance of clipping events alone can occasionally miss chimeric contigs. For instance, we stumbled upon a suspiciously long contig of 7.38 Mbp. Even though this contig was not flagged by the frequency of read clipping events, it was a result of metaMDBG bringing together two sequence-discrete Lachnospiraceae populations in the sample Human O1 (Supplementary Figure Fig 1).

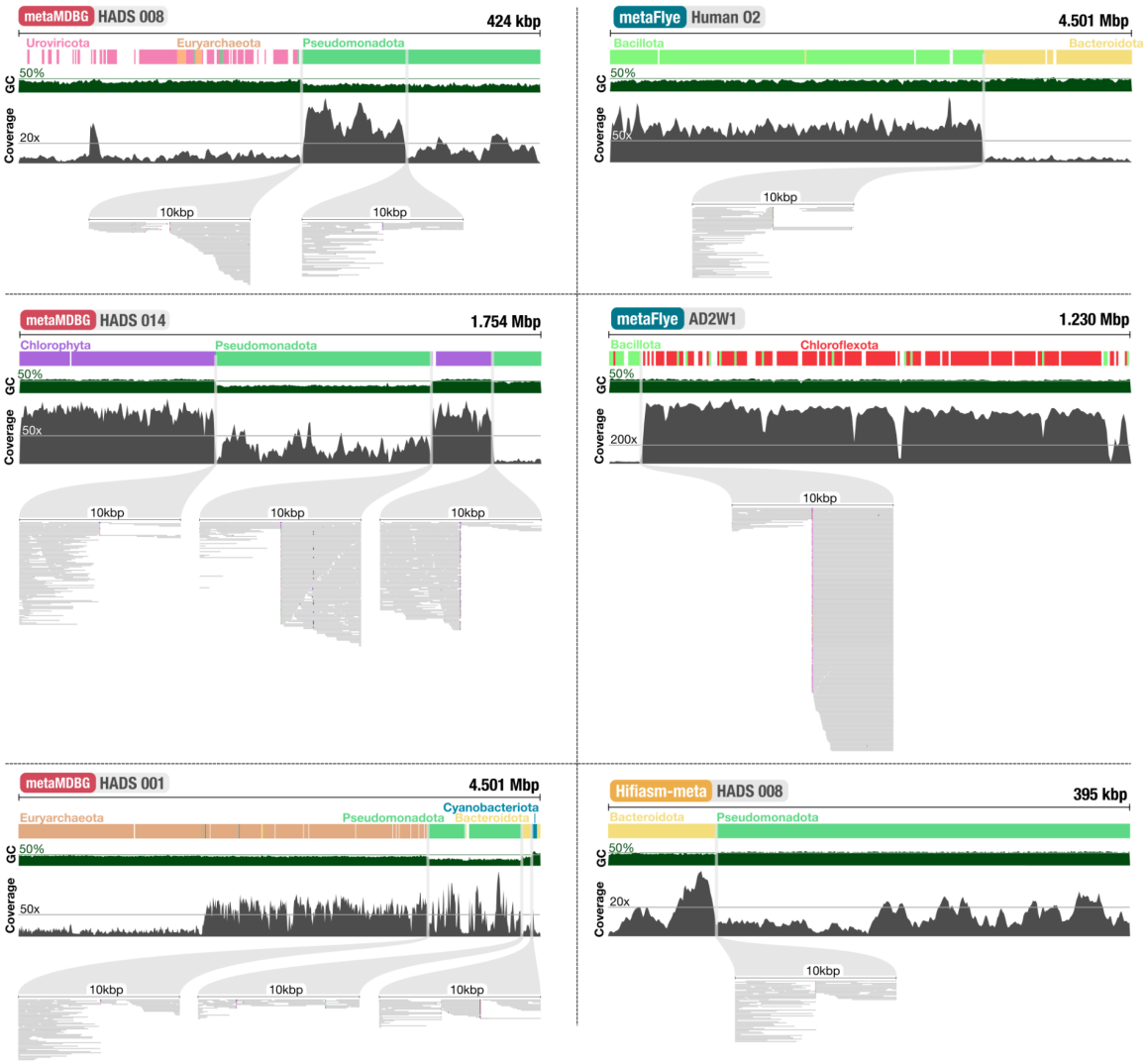


Figure 2. Multi-domain and multi-phyla contigs. Six contigs from metaMDBG, metaFlye and hifiasm-meta. For each contigs we displayed the GC content, coverage in the metagenomics reads used for their assembly, gene level taxonomy. For each assembly breakpoint, we display a zoomed-in detail of the read mapping from IGV. In these subplots, red arrows at the end of the mapped read indicate clipping and the coloring at the end of these reads indicates that the following portion of the read mapped to another contig and similar colors indicate that multiple reads continue to map on the same contig. Blue markers indicate large indels (> 150bp).

The highly dangerous nature of chimeric contigs for downstream analyses is dampened by the straightforward nature of their identification by anyone who carefully investigates their data: most chimeric contigs that erroneously connect genomic regions from two or more distinct populations

169 can be easily identified using a series of genomic markers such as a sudden shift in GC content,
170 read coverage, and gene-level taxonomy, or through unexpected inventories of single-copy core
171 genes (SCGs). However, with the increasing tendency of researchers to generate large
172 metagenome-derived genome compendiums (Pasolli et al. 2019; Parks et al. 2017; Ma et al.
173 2023; Nayfach et al. 2021; Almeida et al. 2021), such evaluations are rarely, if ever, conducted.
174 In an ideal world, the burden of resolving chimeras should never fall on the shoulders of the end-
175 users of assembly algorithms.

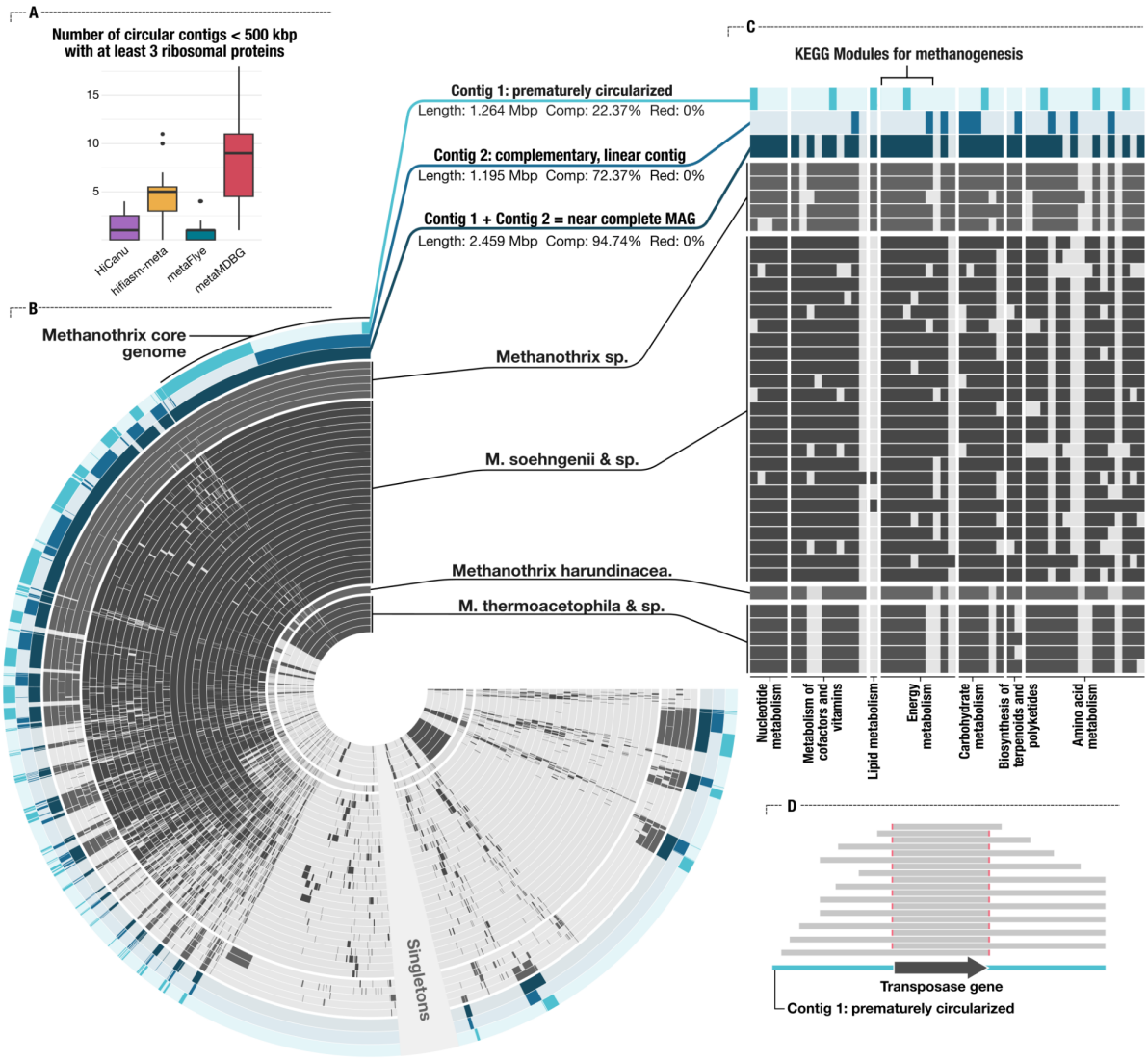
176 Premature circularization

177 To deliver one of the most sought after promises of long-read sequencing, most long-read
178 assemblers contain built-in features to circularize contigs and report potentially complete microbial
179 and mobile genetic element genomes. Premature circularization, the reporting of a contig as a
180 circular element when it omits parts of the genome it originates from, is a form of error with
181 dangerous downstream implications. Yet, our reanalyses showed that the algorithmic features
182 that reconstructed circular contigs in long-read assemblers, especially metaMDBG, and to a
183 lesser degree hifhasm-meta, were far from reliable.

184 For instance, an archaeal genome that belongs to the genus *Methanothrix* recovered from the AD
185 Sludge sample as a circular genome by hifiasm-meta represents a clear example of the nature
186 and implications of premature circularization. Through an analysis with additional genomes from
187 the NCBI's RefSeq collection we confirmed that the circular genome was missing a large fraction
188 of the core *Methanothrix* pangenome (Figure 3B, light blue), including key metabolic modules for
189 methanogenesis that were common to all *Methanothrix* genomes (Figure 3C). In this case, early
190 circularization occurred near a transposase (Figure 1D, Supplementary Figure 2). Nearly all the
191 reads clipping on either side of the transposase had supplementary mapping to another contig in
192 the assembly output (Figure 3B, C, medium blue) encoding the missing metabolic module. The

193 combination of the circular genome and the additional contig matched the other genomes in the
194 genus level pangenome (Figure 3B, C, dark blue), suggesting that both contigs belonged to the
195 same population, and neither of them are circular by themselves.

196



197

198 **Figure 3. Premature circularization of a *Methanothrix* genome**. (A) Frequency of circular contigs under 500kbp with
199 a minimum of 3 ribosomal proteins. Each point represents one assembly. (B) A pangenomic analysis of all publicly
200 available *Methanothrix* genomes from the NCBI's RefSeq database completed with the so-called circular genome of

201 *Methanothrix* assembled from the sample AD Sludge by hifiasm-meta (light blue), as well as a contig from the same
202 assembly which correspond to the rest of the missing Methanothrix genome (medium blue) and the combination of
203 these two contigs (dark blue). (C) KEGG metabolic module completion of all genomes and contigs in (B). (D) A
204 schematic representation of the reads mapping over a transposase gene in the prematurely circularized contigs (light
205 blue in B and C) showing the lack of reads support around the gene, the full figure is available in Supplementary Figure
206 2.

207 Comparing the overall accuracy of circularization across assemblers is difficult. For instance, low
208 completion estimates based on single-copy core genes can serve as a quick filter, but not all
209 prematurely circularized genomes will have low completion estimates since missing genomic
210 content will not always contain single-copy core genes. Length comparisons between circular
211 contigs and known genomes in databases could offer another means of scrutiny, but this strategy
212 will not be effective against poorly studied clades, or those that have no representation in public
213 databases. Nevertheless, here we conservatively assumed that circular contigs that were under
214 500 kbp and contained at least 3 ribosomal proteins were most likely circularized erroneously,
215 and the rates at which they are generated by each assembler could serve as a proxy for an overall
216 estimate of the rate of premature circularization events across assemblers. For this range,
217 metaMDBG reported about twice as many circular contigs than hifiasm-meta, and about four
218 times more circular contigs than HiCanu and metaFlye (Figure 3A). Assuming that easily
219 identifiable events of premature circularization errors are reasonable proxies for the rate of all
220 premature circularization errors, this result suggests that metaMDBG and hifiasm-meta are more
221 prone to premature circularization errors compared to other assemblers, such as HiCanu and
222 metaFlye. Interestingly, our estimates of the rate of false circularization come very close to the
223 benchmarking results shared by the authors in the original metaMDBG publication (Benoit et al.
224 2024), in which they reported two times more circular contigs than hifiasm-meta and four times
225 more circular contigs than metaFlye from the human gut. While Benoit et al. present the high rate
226 of circularization as a strength of metaMDBG in comparison to other assemblers (Benoit et al.

227 2024), our observations suggest that the higher rate of circularization may be a result of a higher
228 tendency to report non-circular elements as circular (Figure 1, Figure 3, Supplementary Figure 2,
229 Supplementary Table 2).

230 Given the vast difference in quality and perception between a draft genome and a circular
231 genome, it is essential for assembly algorithms that promise circular genomes to be conservative
232 in their decision making. The identification of prematurely circularized contigs in assembly results
233 will be a notoriously difficult task for end-users, especially when the missing genomic context is
234 relatively short, or circular contigs represent plasmids or viruses. Thus, it would be ideal if the
235 circular sequences are validated more rigorously by the assembler prior to the final reporting.

236 Haplotyping errors, false duplications, and non-existent sequences

237 Accurate reconstruction of genomic variation is essential to associate within-population structural
238 differences to ecological or evolutionary phenotypes. However, resolving genomic regions that
239 differ between otherwise very closely related subpopulations is a major challenge for de novo
240 assemblers (Liao et al. 2019). Assemblers may resolve such structural complexity (1) by reporting
241 separate contigs for variable sequences and conserved regions that flank them, (2) by reporting
242 the most prevalent variable region along with the flanking conserved regions in a single contig
243 and alternative regions in shorter contigs, or (3) by duplicating conserved flanking regions in
244 multiple contigs that describe each of the variable region and their surrounding as separate
245 contigs in a haplotype-aware fashion. Our survey of long-read assembly results revealed
246 unexpected haplotyping decisions in multiple recurrent forms. In some cases, assemblers
247 concatenated subpopulation-specific variable regions flanked by conserved loci, rather than
248 reporting only one of them accurately (Figure 4A). As a result, long-reads that map to these
249 regions are clipped at the end of their respective sub-population sequence. In other cases, the
250 final contig represented a variable region found in a minor subpopulation supported by a small

251 number of long-reads or a single one (Figure 4B), violating the logical expectation to recover a
252 consensus sequence that represents the most abundant sub-population.

253 We also identified cases where assemblers reported false genomic duplications in assembled
254 contigs that were not supported by any long-read. Such false repeats were often manifested by
255 high-frequency read clipping events and appeared as long direct repeats that had low likelihood
256 to be present in the target genomic context due to the sudden decrease in read coverage and/or
257 massive inserts in mapped reads (Figure 4C). Yet another anomaly was the reporting of
258 sequences that did not exist. BLAST-searching zero-coverage regions from contigs that are
259 longer than 500 bp against the database of raw long reads (Supplementary Table 3), we
260 confirmed that the assembly outputs by metaMDBG and metaFlye occasionally included up to
261 over 5,000 bp long regions that have no homology to any of the input long reads (Figure 4D).
262 False duplications and reporting of non-existent sequences in assembly results are unexpected
263 behaviours from any assembler and can lead to spurious open reading frames or the omission of
264 genuine ones.

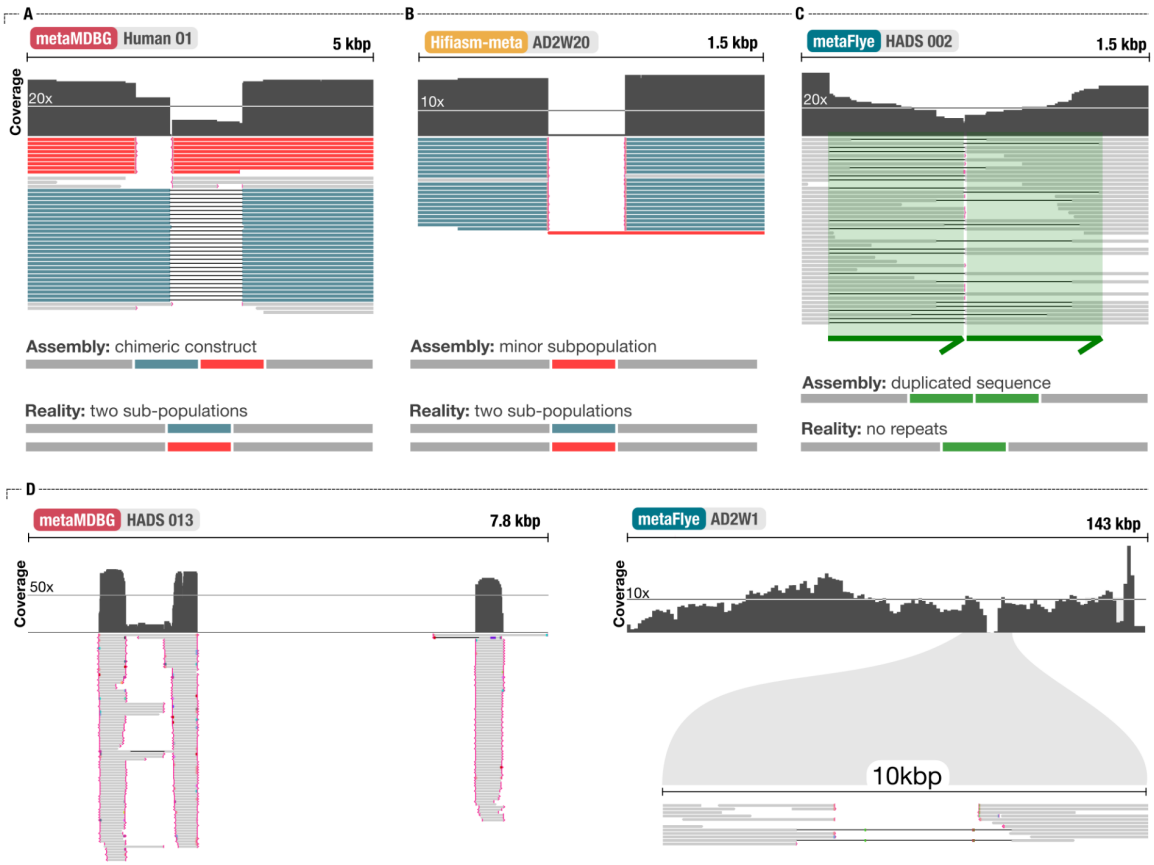


Figure 4. Prototypical mapping artifacts and their putative origin. (A) A chimeric sequence assembled from two subpopulations. At a conserved locus, two subpopulations existed with their own and distinct sequence. The assembled contig contains all or a part of each subpopulation specific sequence resulting in a chimeric construct. (B) Another example of a variable genomic site, but in this example the contig sequence contains the sequence of a very minor subpopulation, supported by only one read. (C) Duplicated sequence only found in the assembly, not supported by any long reads. (D) Two contigs assembled from metaMDBG (left) and metaFlye (left) presenting large regions with no coverage. We blasted these regions back to the long reads and found no hits. Coverage visualization was exported from the anvi'o interactive interface (left) or the IGV software (right) and the read mapping visualization was from IGV as well. Indel smaller than 150bp as well as mismatches are not displayed in the mapping. Red markers at the end of reads indicate read clipping.

Excessive repeats

A recurrent and puzzling observation throughout our manual inspections of assembly results was the astonishing number of repeats that occasionally made up the entirety of some contigs. Yet, these repeats were not caught by our survey of read clipping event frequencies to mark regions of concern as repeats rarely resulted in 100% read clipping events to pass our filter. Thus, we characterized repeats without any read mapping data but by aligning each contig to itself. We marked any region of a contig as a 'repeat' if it was longer than 200 bp and occurred multiple times in the same contig with at least 80% identity. Our survey of the frequency and distribution of such repeats across contigs under 50 kbp showed that each assembler reported contigs with repeats. Since repeats are common in nature, and the improved ability to resolve repeats is one of the strengths of long-read sequencing, this finding alone is not concerning. However, the nature of repeats in assemblies revealed by dot plots often displayed surprisingly intricate patterns suggested a high likelihood that they were generated by the assembly algorithms (Figure 5A). While naturally occurring tandem repeats, inversions, or palindromes could lead to similar dot plots, our manual inspection of individual cases revealed a variety of errors that spanned from duplicated reporting of known circular plasmids to contigs with multiple repeats that are not supported by any long read.

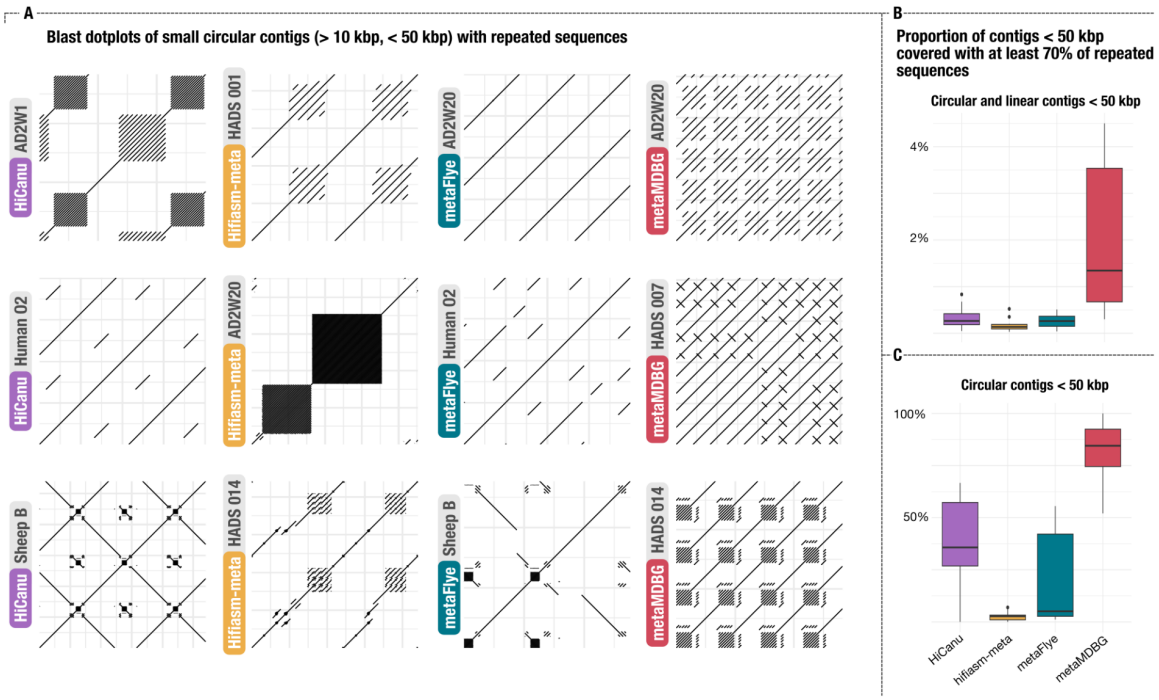


Figure 5. Repeats across short contigs. (A) Dotplots of reciprocal blast results from 12 short circular contigs showing high amounts of repeated sequences. (B,C) Proportion of contigs with at least 70% of their length duplicated at least once. Result computed for all short contigs under 50kbp (B) and only for circular contigs under 50kbp (C).

Repeats differed in their length, identity, and frequency across algorithms. The per sample average length and sequence identity of repeats varied between 600 bp to over 4,000 bp, and 88% to 96%, respectively. MetaMDBG generated more repeats than any other assembler (Figure 5B), up to over 300,000 repeats in a single assembly (Supplementary Table 2), exceeding the number of repeats reported by HiCanu for the same sample 280 times. The average length of repeats were relatively short, yet we found repeats that were up to 40,000 nucleotides long that were due to multiple tandem repeats forming near contig-long and overlapping repeats, and repeats occurred as much as over a 1,000 times in a single contig (Supplementary Table 2). To summarize the number and proportion of contigs reported by each assembler with an excessive number of repeats, we conservatively searched for assembled sequences in which at least 70% of the sequence was composed of repeats. This search revealed that metaMDBG generated on

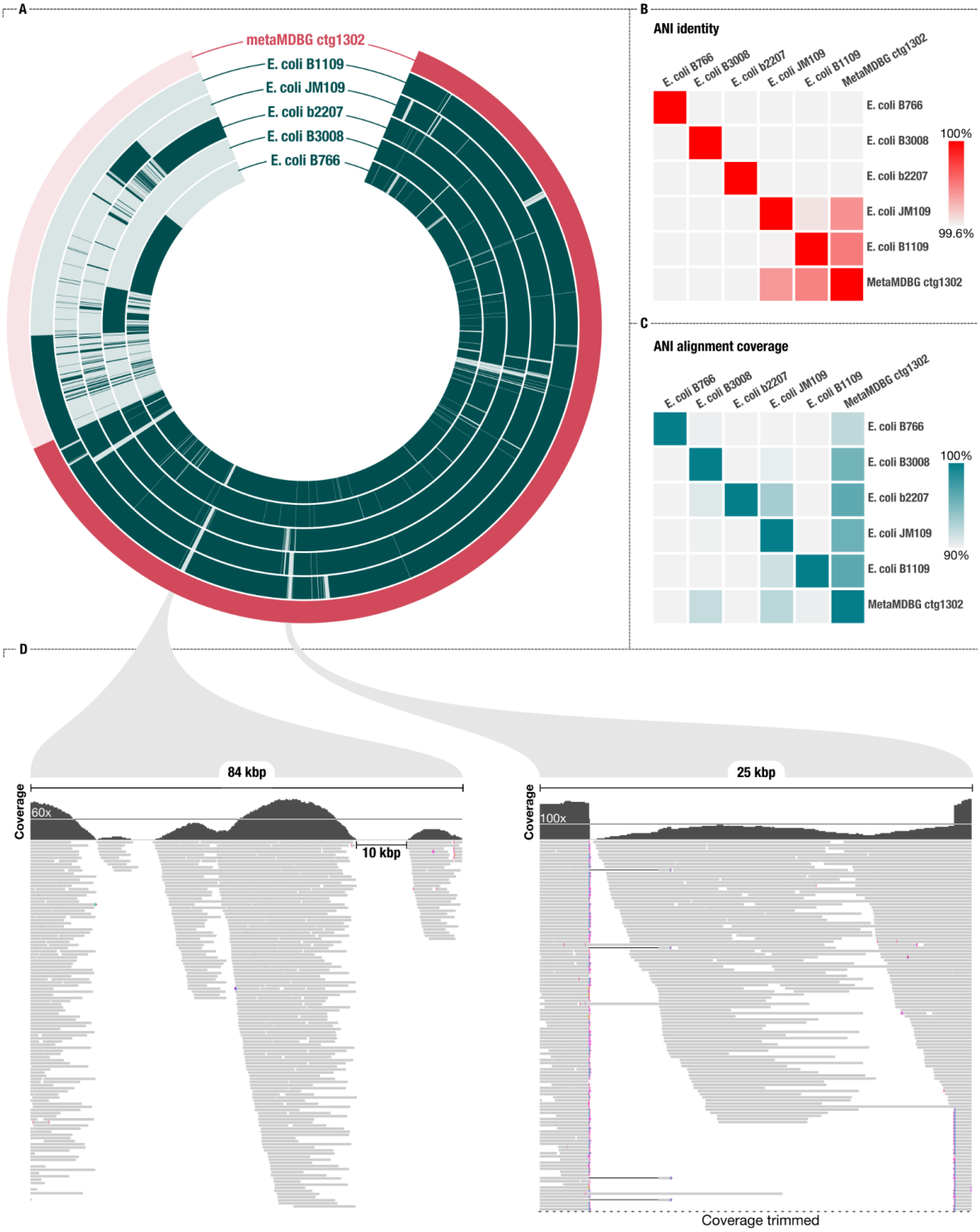
308 average 2.8% of contigs under 50,000 bp with a high amount of repeats, a rate that was over 6
309 times higher than its runner up, HiCanu (Figure 5B). When we limited our survey to circular
310 contigs, the proportion of repeat-rich contigs skyrocketed across all sample types for all
311 assemblers, with the exception of hifiasm-meta (Figure 5C, Supplementary Table 2). For instance,
312 87% of all circular contigs under 50 kbp reported by metaMDBG were largely composed of
313 repeats, and in some samples, such as Chicken, this number reached to 100%, suggesting that
314 artifactual repeats represent at least one of the factors that lead to false circularization (Figure
315 5C). Marine metagenomes were particularly difficult for all assemblers. In this sample type
316 metaMDBG reported the highest number of circular contigs (Supplementary Table 2), however,
317 92% of them under 50,000 bp were composed of repeats (Supplementary Table 2). While
318 metaFlye performed as well as hifiasm-meta in most other sample types, up to 55% of circular
319 contigs metaFlye generated from marine metagenomes were largely composed of repeats
320 (Supplementary Table 2).

321 These results show that the exciting prospects of recovering complete and circular plasmid and
322 virus genomes from long-read sequencing of complex metagenomes is still far from being
323 realized, and that current assemblers often fall short when handling short circular elements.

324 Mock datasets are useful, but can yield misleading insights into the 325 accuracy of algorithms in real-world applications

326 Popular mock datasets such as Zymo-HiFi D6331 and ATCC MSA-1003 are commonly used for
327 benchmarking long-read assemblers. While their known composition constitutes a reasonable
328 starting point, the mock datasets do not represent the complexity of the natural samples, as also
329 noted by the authors of metaFlye (Kolmogorov et al. 2020). Their utility to test assemblers is
330 further reduced when benchmarks that employ mock communities simply align contigs that
331 emerge from the assemblies to reference genomes without comprehensive reporting of other

assembly metrics (Benoit et al. 2024; Kolmogorov et al. 2020; Feng et al. 2022). For instance, while hifiasm-meta reports favorable outcomes given the reference genomes in mock datasets (Feng et al. 2022), in our tests the algorithm generated massive assemblies for each mock datasets, where the final assembly was 270 Mbp instead of the expected size of 93 Mbp for the Zymo-HiFi D6331, and it was 948 Mbp instead of the expected size of 66.44 Mbp for the ATCC MSA-1003, resulting in the lowest N50 values and the highest number of clipping events across all assemblers. Given its performance with the mock datasets, one may expect hifiasm-meta to perform poorly in its applications to complex metagenomes. Yet, in marine samples, hifiasm-meta was first in N50, and second to metaMDBG in assembly size with two orders of magnitude less clipping errors, suggesting that the performance of an assembler with mock communities may not predict its performance with real-world datasets.



343

344 **Figure 6. Assembled E. coli from a mock community does not match reference genomes**. (A) A pangenome
345 analysis of the *E. coli* reference genome used in the Zymo mock community and the *E. coli* circular genome assembled
346 by metaMDBG. Each layer corresponds to a genome and the coloring represents the presence/absence of a gene

347 cluster from each genome. The gene clusters are organized by the synteny of genes in the metaMDBG contig. (B)
348 Genome pairwise average nucleotide identity and (C) alignment coverage. (D) Details of the long read mapping over
349 two genomic regions highlighted in the first panel. These regions included genes not found in E. coli B1109 (closest
350 reference genome by ANI metrics), but rather found in other E. coli genomes.

351 While mock communities do not represent the diversity and complexity found in real-world
352 samples, they shine in one fundamental way: the known genomic makeup of input organisms to
353 identify glaring issues post-assembly. The Zymo-HiFi D6331 includes five different strains of *E.*
354 *coli*, which yields sequencing data with complex cases for assemblers due to the presence of
355 highly conserved and divergent genomic regions. HiCanu and hifiasm-meta were both successful
356 as reconstructing at least one of the five *E. coli* genomes (Supplementary Figure 3), while
357 metaMDBG reported a circular contig that corresponded to a chimeric genome (Figure 6). Based
358 on pairwise ANI comparisons, this genome appeared to be most similar to B1109, one of the *E.*
359 *coli* strains in the Zymo-HiFi D6331 mock dataset (Figure 5B, Supplementary Table 4). However,
360 the pangenome of the five *E. coli* genomes and the metaMDBG circular contig showed that not
361 only it was missing a portion of the B1109 genome, but also including genomic regions exclusive
362 to other *E. coli* strains and absent in B1109 (Figure 6A). Clipping events also captured chimeric
363 regions and revealed a ~10 kbp locus with no coverage (Figure 6C, D). That region without
364 coverage was duplicated in two other contigs, suggesting that long reads were preferentially
365 recruited there (Supplementary Table 5). Using ANI values that are calculated from local
366 alignments of assembled contigs from mock datasets to the collection of reference genomes they
367 contain risks masking glaring issues such as phantom sequences, chimeras, and unexpectedly
368 large assembly results, and invalidates perhaps the only useful aspect of mock datasets while
369 inflating the accuracy of assembly algorithms.

Conclusions

Our findings highlight the need for rigorous evaluation of long-read assembly algorithms beyond benchmarks that typically prioritize runtime, contig length, or the number of circular contigs. We demonstrate that identifying the most terrifying assembly errors can be as simple as mapping individual reads back to assembled contigs. Some popular long-read assemblers like metaFlye include a correction step that uses the long reads to polish the assembly, and could leverage mapping information like clipping events to generate more accurate assemblies. We believe assembly algorithms should provide better options for error correction using input reads, rather than offloading this burden onto end users who may lack the time, expertise, or computational resources to perform such refinements themselves. The biotechnology, biomedical, and basic research communities rely on high-quality assemblies, and trustworthy results are critical for progress in life sciences as well as the quality of public genome databases. Given the critical role of assembly in genomics, most researchers will likely prefer longer runtimes or fewer circular contigs over algorithms that are fast or produce erroneous sequences.

Materials and methods

The URL https://merenlab.org/data/benchmarking-long-read-assemblers/ serves our bioinformatics workflow to reproduce our findings or apply the same approaches to evaluate additional assemblers or datasets.

Datasets

We downloaded a set of publicly available HiFi PacBio metagenomes matching the ones used in long-read assembler publications. To complete and expand the set of biomes, we included eight surface seawater metagenomes. Supplementary Table 1 includes a comprehensive description of these data and their accession numbers.

Assembly algorithms

We used 4 different assemblers: HiCanu v2.2 (Nurk et al. 2020), hifiasm-meta v0.3 (Feng et al. 2022), metaFlye v2.9.5 and metaMDBG v0.3, v1 and v1.1 (Benoit et al. 2024). For HiCanu, we used the parameters maxInputCoverage=1000 genomeSize=100m batMemory=200 and -pacbio-hifi. We used hifiasm-meta with the default parameters. For metaFlye, we used the parameters --meta --pacbio-hifi. We included three versions of metaMDBG; v0.3, which is the original published version of the software, and v1 which was released August 2024, and v1.1 released in December 2025. The results for v0.3 and v.1.1 can be found in the supplementary material while the results from v1 are shown in this manuscript. For metaMDBG v1 and v1.1, we used the parameters --in-hifi.

Read mapping

We mapped the metagenomics long-reads back to their respective assembly by the four assemblers using minimap2 v2.28 (Li 2018) with the following parameters: -ax map-hifi -p1 --

407 secondary-seq. This set of parameters allows secondary mapping when the alignment score is
408 as good as the primary mapping score, i.e. multi mapping; as well as to keep the sequence for
409 secondary mapping in the output files so that secondary mapping is properly considered in
410 downstream analyses. We processed the resulting alimgent files using samtools v1.17 (Li et al.
411 2009).

412 Processing of assembled contigs and read mapping results

413 We used anvi'o development branch of v8.1 (Eren et al. 2021) to generate contigs databases for
414 each assembly with the command 'anvi-gen-contigs-database', which performed gene calls using
415 prodigal v2.6.3 (Hyatt et al. 2010). The anvi'o programs 'anvi-run-ncbi-cogs' and 'anvi-run-kegg-
416 kofams' annotated genes with function suing the NCBI's Clusters of Orthologous Genes (COGs)
417 (Galperin et al. 2015) database and KEGG KOfams (Aramaki et al. 2020), respectively, while
418 'anvi-run-hmms' identified single-copy core genes in these sequences which we associated with
419 taxonomy data from the GTDB (Parks et al. 2022) using 'anvi-run-scg-taxonomy'. We estimated
420 the completeness of KEGG KOfam modules using 'anvi-estimate-metabolism'. We finally used
421 'anvi-profile' to process the read mapping data to recover coverage values, single-nucleotide
422 variants (SNVs) and insertion/deletion events (INDELs).

423 Pangenomic analyses

424 To compute a pangenome for *Methanothrix* we first acquired publically available genomes from
425 NCBI's RefSeq database using the program 'ncbi-genome-download' (available from
426 https://github.com/kblin/ncbi-genome-download) with the parameters "--assembly-level all --
427 genera Methanothrix". For the *E. coli* pangenome, we downloaded the original genomes used to
428 create the mock dataset: https://s3.amazonaws.com/zymo-files/BioPool/D6331.refseq.zip. For
429 both pangenomics analysis, we used the program 'anvi-run-workflow' to run the anvi'o
430 pangenomics workflow implemented in Snakemake (Köster and Rahmann 2012), which used

431 DIMAOND v2.1.8 (Buchfink, Xie, and Huson 2015) to identify gene clusters as described
432 previously (Delmont and Eren 2018). We used the program 'anvi-display-pan' to visualize and
433 summarize the pangenomes.

434 Identification of assembly errors

435 To identify potential assembly errors based on clipping events, we developed a program within
436 the anvi'o platform, 'anvi-script-find-misassembly' (https://anvio.org/m/anvi-script-find-
437 misassemblies), which takes a single BAM file of long-read mapping results. The script searches
438 for premature end of alignments, i.e., clipping events, and reports positions where a proportion of
439 reads that are clipped exceeds a user-defined threshold. The script also reports regions in contigs
440 with no coverage. In addition, we used BLAST v2.16.0 (Camacho et al. 2009) to identify contigs
441 covered by at least 70% of repeated sequences. We blasted each contig against itself using blastn
442 default parameters, excluded the perfect reciprocal hit and transformed the remaining hits to a
443 bed file and we used bedtools v2.31.1 (Quinlan and Hall 2010) to compute the breath of coverage.
444 We computed additional statistics using a python script available in the link to the reproducible
445 workflow

446 Manual inspection of mapping results

447 We used IGV v2.17.4 (Robinson et al. 2017) and the anvi'o interactive interface to manually
448 inspect genomic regions of interest and generate figures.

Data availability

All metagenomes used in our study are publicly available through the NCBI, and Supplementary Table 1 lists their accession IDs. DOIs for intermediate data products are available at https://merenlab.org/data/benchmarking-long-read-assemblers/

Acknowledgements

Authors recognize that developing assembly algorithms, especially for metagenomes, is a notoriously complex and difficult task, and they would like to acknowledge their deep appreciation of those who invest their time and skills in creating and maintaining them.

Author contributions

FT, RS, JFB, and AME conceptualized the study. FT curated data, performed formal analyses, and developed software tools. FT, RS, JFB, and AME analyzed data and interpreted findings. FT and AME wrote the original draft of the study, all authors commented on and made suggestions, and approved the final manuscript.

Ethics declarations

Competing interests

Authors declare no competing interest.

References

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. “A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome.” *Nature Biotechnology* 39 (1): 105–14.

Aramaki, Takuya, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. 2020. “KofamKOALA: KEGG Ortholog Assignment Based on Profile HMM and Adaptive Score Threshold.” *Bioinformatics (Oxford, England)* 36 (7): 2251–52.

Benoit, Gaëtan, Sébastien Raguideau, Robert James, Adam M. Phillippy, Rayan Chikhi, and Christopher Quince. 2024. “High-Quality Metagenome Assembly from Long Accurate Reads with metaMDBG.” *Nature Biotechnology* 42 (9): 1378–83.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 421.

Chang, Tianyi, Gregory S. Gavelis, Julia M. Brown, and Ramunas Stepanauskas. 2024. “Genomic Representativeness and Chimerism in Large Collections of SAGs and MAGs of Marine Prokaryoplankton.” *Microbiome* 12 (1): 126.

Chen, Lin Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2020. “Accurate and Complete Genomes from Metagenomes.” *Genome Research*. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gr.258640.119.

Delgado, Luis Fernando, and Anders F. Andersson. 2022. “Evaluating Metagenomic Assembly Approaches for Biome-Specific Gene Catalogues.” *Microbiome* 10 (1): 72.

Delmont, Tom O., and A. Murat Eren. 2018. “Linking Pangenomes and Metagenomes: The Prochlorococcus Metapangenome.” *PeerJ* 6 (January):e4320.

Dijk, Erwin L. van, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. 2014. “Ten Years of next-Generation Sequencing Technology.” *Trends in Genetics: TIG* 30 (9): 418–26.

Eren, A. Murat, and Jillian F. Banfield. 2024. “Modern Microbiology: Embracing Complexity through Integration across Scales.” *Cell* 187 (19): 5151–70.

Eren, A. Murat, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E. Miller, Matthew S. Schechter, Isaac Fink, et al. 2021. “Community-Led, Integrated, Reproducible Multi-Omics with Anvi’o.” *Nature Microbiology* 6 (1): 3–6.

Feng, Xiaowen, Haoyu Cheng, Daniel Portik, and Heng Li. 2022. “Metagenome Assembly of High-Fidelity Long Reads with Hifiasm-Meta.” *Nature Methods* 19 (6): 671–74.

Galperin, Michael Y., Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. 2015. “Expanded Microbial Genome Coverage and Improved Protein Family Annotation in the COG Database.” *Nucleic Acids Research* 43 (Database issue): D261–69.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (1): 119.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology* 36 (4): 338–45.

Kolmogorov, Mikhail, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, et al. 2020. “metaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs.” *Nature Methods* 17 (11): 1103–10.

Köster, Johannes, and Sven Rahmann. 2012. “Snakemake—a Scalable Bioinformatics Workflow Engine.” *Bioinformatics* 28 (19): 2520–22.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Liao, Xingyu, Min Li, You Zou, Fang-Xiang Wu, Yi-Pan, and Jianxin Wang. 2019. “Current Challenges and Solutions of *de Novo* Assembly.” *Quantitative Biology (Beijing, China)* 7 (2): 90–109.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Ma, Bin, Caiyu Lu, Yiling Wang, Jingwen Yu, Kankan Zhao, Ran Xue, Hao Ren, et al. 2023. “A Genomic Catalogue of Soil Microbiomes Boosts Mining of Biodiversity and Genetic Resources.” *Nature Communications* 14 (1): 7318.

Meyer, Fernando, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, et al. 2022. “Critical Assessment of Metagenome Interpretation: The Second Round of Challenges.” *Nature Methods* 19 (4): 429–40.

Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich. 2016. “MetaQUAST: Evaluation of Metagenome Assemblies.” *Bioinformatics (Oxford, England)* 32 (7): 1088–90.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, et al. 2021. “A Genomic Catalog of Earth’s Microbiomes.” *Nature Biotechnology* 39 (4): 499–509.

Nurk, Sergey, Brian P. Walenz, Arang Rhie, Mitchell R. Vollger, Glennis A. Logsdon, Robert Grothe, Karen H. Miga, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. 2020. “HiCanu: Accurate Assembly of Segmental Duplications, Satellites, and Allelic Variants from High-Fidelity Long Reads.” *Genome Research* 30 (9): 1291–1305.

Olson, Nathan D., Todd J. Treangen, Christopher M. Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop. 2019. “Metagenomic Assembly through the Lens of Validation: Recent Advances in Assessing and Improving the Quality of Genomes Assembled from Metagenomes.” *Briefings in Bioinformatics* 20 (4): 1140–50.

Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. “GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy.” *Nucleic Acids Research* 50 (D1): D785–94.

549 Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J.
550 Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. "Recovery of
551 Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life."
552 *Nature Microbiology* 2 (11): 1533–42.

553 Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica
554 Armanini, Francesco Beghini, et al. 2019. "Extensive Unexplored Human Microbiome
555 Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age,
556 Geography, and Lifestyle." *Cell* 176 (3): 649–62.e20.

557 Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing
558 Genomic Features." *Bioinformatics (Oxford, England)* 26 (6): 841–42.

559 Robinson, James T., Helga Thorvaldsdóttir, Aaron M. Wenger, Ahmet Zehir, and Jill P.
560 Mesirov. 2017. "Variant Review with the Integrative Genomics Viewer." *Cancer Research*
561 77 (21): e31–34.

562 Schechter, Matthew S., Florian Trigodet, Iva A. Veseli, Samuel E. Miller, Matthew L. Klein,
563 Metehan Sever, Loïs Maignien, Tom O. Delmont, Samuel H. Light, and A. Murat Eren.
564 2025. "Ribosomal Protein Phylogeography Offers Quantitative Insights into the Efficacy of
565 Genome-Resolved Surveys of Microbial Communities." *bioRxiv*.
566 https://doi.org/10.1101/2025.01.15.633187.

567 Schoelmerich, Marie C., Lynn Ly, Jacob West-Roberts, Ling-Dong Shi, Cong Shen, Nikhil S.
568 Malvankar, Najwa Taib, et al. 2024. "Borg Extrachromosomal Elements of Methane-
569 Oxidizing Archaea Have Conserved and Expressed Genetic Repertoires." *Nature*
570 *Communications* 15 (1): 5414.

571 Shaiber, Alon, and A. Murat Eren. 2019. "Composite Metagenome-Assembled Genomes
572 Reduce the Quality of Public Genome Repositories." *mBio* 10 (3).
573 https://doi.org/10.1128/mBio.00725-19.

574 Shendure, Jay, and Hanlee Ji. 2008. "Next-Generation DNA Sequencing." *Nature*
575 *Biotechnology* 26 (10): 1135–45.

576 Tørresen, Ole K., Bastiaan Star, Pablo Mier, Miguel A. Andrade-Navarro, Alex Bateman, Patryk
577 Jarnot, Aleksandra Gruca, et al. 2019. "Tandem Repeats Lead to Sequence Assembly
578 Errors and Impose Multi-Level Challenges for Genome and Protein Databases." *Nucleic*
579 *Acids Research* 47 (21): 10994–6.

580 Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M.
581 Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F.
582 Banfield. 2004. "Community Structure and Metabolism through Reconstruction of
583 Microbial Genomes from the Environment." *Nature* 428 (6978): 37–43.

584 Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory
585 T. Concepcion, Jana Ebler, et al. 2019. "Accurate Circular Consensus Long-Read
586 Sequencing Improves Variant Detection and Assembly of a Human Genome." *Nature*
587 *Biotechnology* 37 (10): 1155–62.

Supplementary Information

Supplementary Tables

Supplementary Tables are also available via doi:10.6084/m9.figshare.28831904.

Supplementary Table 1: Sample information and metagenome accession numbers.

Supplementary Table 2: Assembly metrics. (a) General statistics about the assemblies. (b) Long-read clipping across samples and assemblers. (c) Metrics about circular contigs, including clipping events. (d) SNVs and INDELs that are either supported by a minority of long reads (called 'partial SNVs/INDELs') or not supported by any long reads. (e) Prematurely circularized contigs. Number of contigs less than 500 kbp with at least three ribosomal protein genes. (f) Frequency and metrics of repeats found in short contigs (less than 50 kbp).

Supplementary Table 3: Region of contigs with no apparent coverage and with no similarity (blast) with any long-reads.

Supplementary Table 4: Average Nucleotide Identity values for *E. coli* genomes. (a) ANI - Percent Identity. (b) ANI - Alignment coverage.

Supplementary Table 5: BLAST result of the *E. coli* sequence with no apparent coverage reported by the metaMDBG to the rest of the assembly.

Supplementary Figures

High-resolution figures are available via doi:10.6084/m9.figshare.28831904.

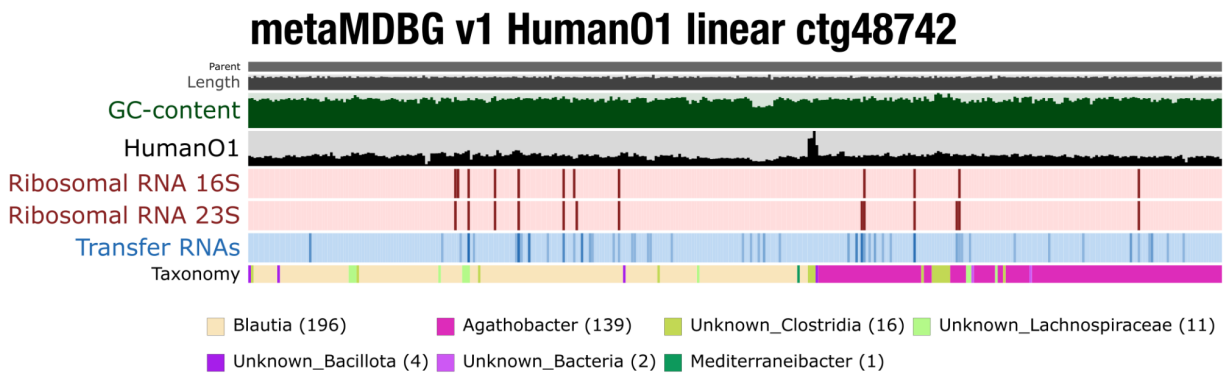
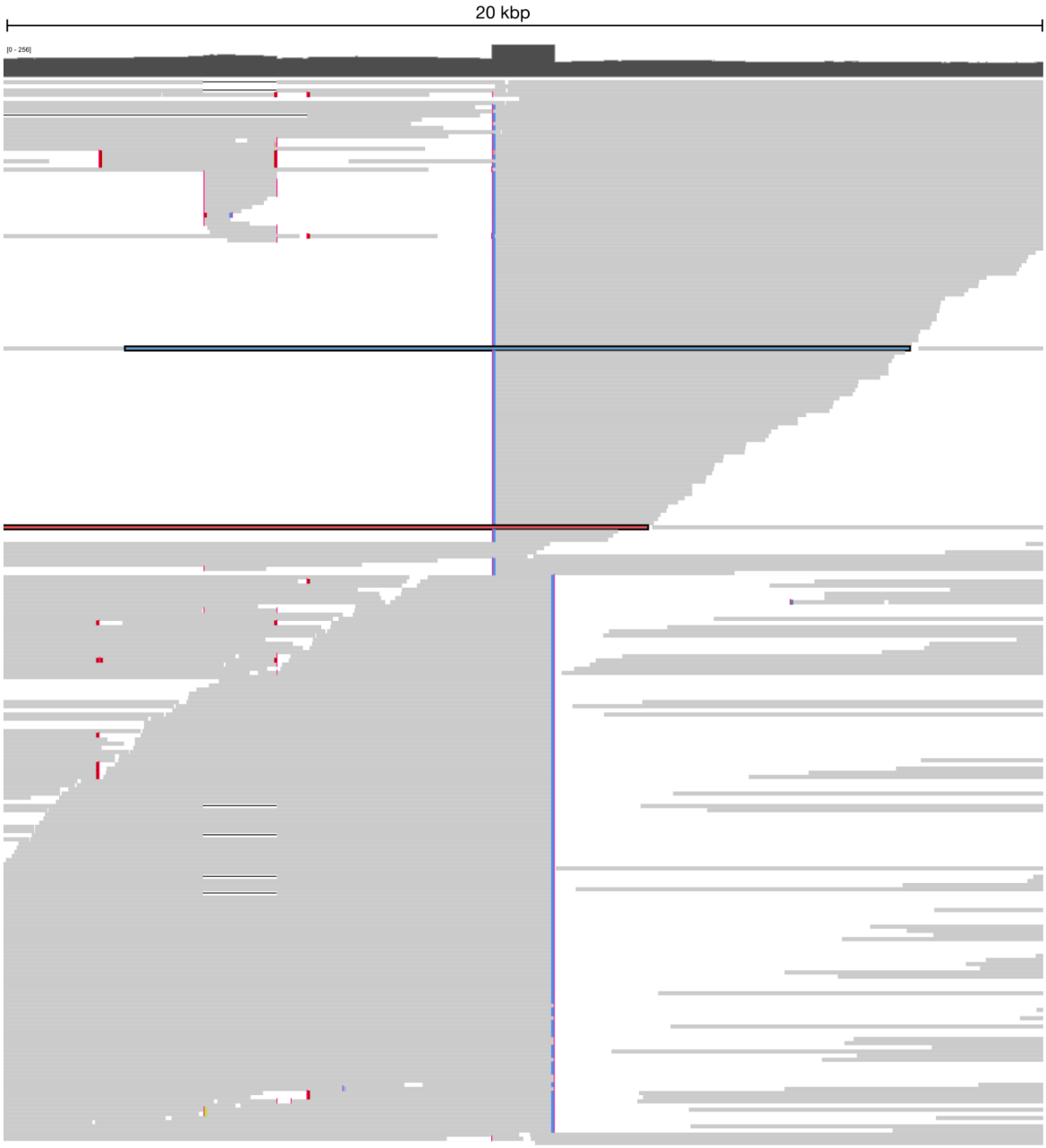


Figure S1: anvi'o interactive interface of a chimeric contig with no apparent clipping issues. The contig is 7.38 Mbp long and the estimated completion/redundancy are both 100%. Gene level taxonomy shows that two populations of Blautia and Agathobacter were merged into a single contig.

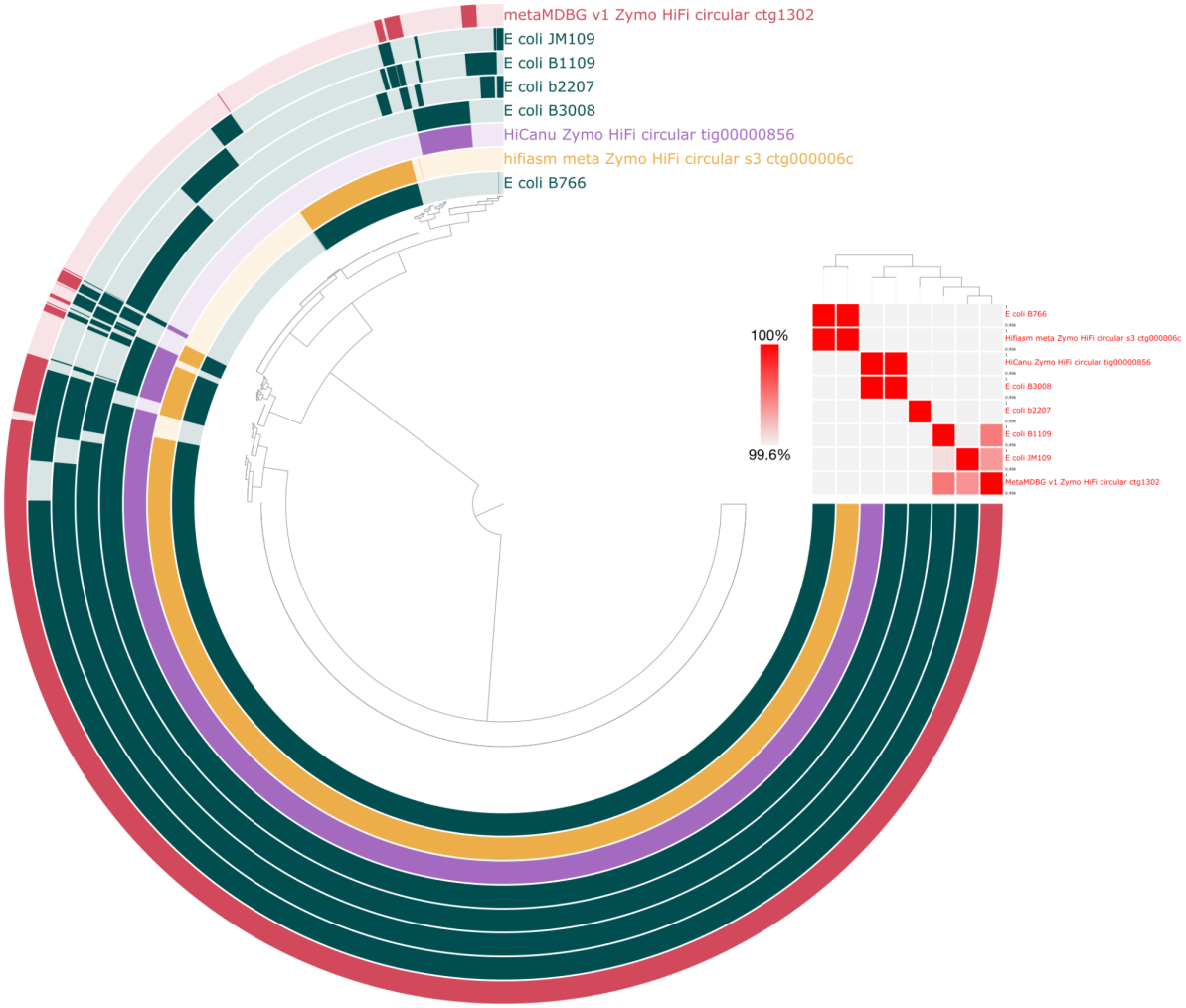


611

612 **Figure S2:** Long-read mapping of a 20kbp region in the prematurely circularized *Methanothrix*

613 genome from Figure 2. The long-read overlap in the middle of the figure over a genomic region

614 corresponding to a transposase gene. All long-read (except two, highlighted in red and blue),



615

616 **Figure S3:** Pangenome analysis of the five *E. coli* genomes from the Zymo-HiFi D6331 mock
617 community and the circular contigs generated by HiCanu, hifiasm-meta and metaMDBG matching
618 to *E. coli.* Both circular contigs generated by HiCanu and hifiasm-meta were perfectly matching
619 one of the five *E. coli* genomes while the contig from metaMDBG contained genes coming from
620 multiple *E. coli* genomes.