

1 Evaluating Metagenome Assembly on a Simple
2 Defined Community with Many Strain Variants

3 Sherine Awad¹, Luiz Irber¹, C. Titus Brown^{1*}
 ¹**Department of Population Health and Reproduction**
 University of California, Davis
 Davis, CA 95616 USA
 * E-mail: ctbrown@ucdavis.edu

4 July 2, 2017

5 **Abstract**

6 We evaluate the performance of three metagenome assemblers, IDBA,
7 MetaSPAdes, and MEGAHIT, on short-read sequencing of a defined
8 “mock” community containing 64 genomes (Shakya et al. (2013)). We
9 update the reference metagenome for this mock community and detect
10 several additional genomes in the read data set. We show that strain
11 confusion results in significant loss in assembly of reference genomes
12 that are otherwise completely present in the read data set. In agree-
13 ment with previous studies, we find that MEGAHIT performs best
14 computationally; we also show that MEGAHIT tends to recover larger
15 portions of the strain variants than the other assemblers.

16 Introduction

17 Metagenomics refers to sequencing of DNA from a mixture of organisms,
18 often from an environmental or uncultured sample. Unlike whole genome
19 sequencing, metagenomics targets a mixture of genomes, which introduces
20 metagenome-specific challenges in analysis [1]. Most approaches to analyz-
21 ing metagenomic data rely on mapping or comparing sequencing reads to
22 reference sequence collections. However, reference databases contain only a
23 small subset of microbial diversity [2], and much of the remaining diversity
24 is evolutionarily distant and search techniques may not recover it [3].

25 As sequencing capacity increases and sequence data is generated from
26 many more environmental samples, metagenomics is increasingly using *de*
27 *novo* assembly techniques to generate new reference genomes and metagenomes
28 [4]. There are a number of metagenome assemblers that are widely used -
29 see [5] for an overview of the available software, and [1] for a review of the
30 different assembler methodologies. However, evaluating the results of these
31 assemblers is challenging due to the general lack of good quality reference
32 metagenomes.

33 Moya et al. in [6] evaluated metagenome assembly using two simulated
34 454 viral metagenome and six assemblers. The assemblies were evaluated
35 based on several metrics including N50, percentages of reads assembled, ac-
36 curacy when compared to the reference genome. In addition to, chimeras per
37 contigs and the effect of assembly on taxonomic and functional annotations.

38 Mavromatis et al. in [7] provided a benchmark study to evaluate the
39 fidelity of metagenome processing methods. The study used simulated
40 metagenomic data sets constructed at different complexity levels. The datasets
41 were assembled using Phrap v3.57, Arachne v.2 [8] and JAZZ [9]. This study
42 evaluates assembly, gene prediction, and binning methods. However, the
43 study did not evaluate the assembly quality against a reference genome.

44 Rangwala et al. in [10] presented an evaluation study of metagenome
45 assembly. The study used a de Bruijn graph based assembler ABYSS [11] to
46 assemble simulated metagenome reads of 36 bp. The data set is classified at
47 different complexity levels. The study compared the quality of the assembly
48 of the data sets in terms of contig length and assembly accuracy. The
49 study also took into consideration the effect of kmer size and the degree of
50 chimericity. However, the study evaluated the assembly based on only one
51 assembler. Also, both previous studies used simulated data, which may lack
52 confounders of assembly such as sequencing artifacts and GC bias.

53 In a landmark study, Shakya et al. (2013) constructed a synthetic com-

54 munity of organisms by mixing DNA isolated from individual cultures of
55 64 bacteria and archaea, including a variety of strains across a range of
56 nucleotide distances [12]. In addition to performing 16s amplicon analy-
57 sis and doing 454 sequencing, the authors shotgun-sequenced the mixture
58 with Illumina. While the authors concluded that this metagenomic sequenc-
59 ing generally outperformed amplicon sequencing, they did not conduct an
60 assembly based analysis. This data set was also used in several other eval-
61 uation studies, including gbtools for binning [13] and benchmarking of the
62 MEGAHIT assembler [14].

63 More recently, several benchmark studies systematically evaluated metagenome
64 assembly of short reads. The Critical Assessment of Metagenome Interpre-
65 tation (CAMI) collaboration benchmarked a number of metagenome assem-
66 blers on several data sets of varying complexity, evaluating recovery of novel
67 genomes and multiple strain variants [3]. Notably, CAMI concluded that
68 “The resolution of strain-level diversity represents a substantial challenge
69 to all evaluated programs.” Another recent study evaluated eight assem-
70 blers on nine environmental metagenomes and three simulated data sets
71 and provided a workflow for choosing a metagenome assembler based on
72 the biological goal and computational resources available [15]. [5] explored
73 metagenome assembler performance on a pair of real data sets, again con-
74 cluding that the biological goal and computational resources defined the
75 choice of assembler. Also see [16] for an analysis of a previously generated
76 HMP benchmark data set; however, the Illumina reads used for this study
77 are much shorter than current sequencing and are arguably not relevant for
78 future studies.

79 In this study, we extend previous work by delving into questions of
80 chimeric misassembly and strain recovery in the Shakya et al. (2013) data
81 set. First, we update the list of reference genomes for Shakya et al. to in-
82 clude the latest GenBank assemblies along with plasmids. We then compare
83 IDBA [17], MetaSPAdes [18], and MEGAHIT [19] performance on assem-
84 bling this short-read data set, and explore concordance in recovery between
85 the three assemblers. We describe the effects of “strain confusion” between
86 multiple strains. We also detect and analyze several previously unreported
87 strains and genomes in the Shakya et al. data set. We find that in the ab-
88 sence of closely related genomes, all three metagenome assemblers recover
89 95% or more of known reference genomes. However, in the presence of
90 closely related genomes, these three metagenome assemblers vary widely in
91 their performance and, in extreme cases, can fail to recover the majority of
92 some genomes even when they are completely present in the reads. Our re-

port provides strong guidance on choice of assemblers and extends previous analyses of this low-complexity metagenome benchmarking data set.

Datasets

We used a diverse mock community data set constructed by pooling DNA from 64 species of bacteria and archaea and sequencing them with Illumina HiSeq. The raw data set consisted of 109,629,496 reads from Illumina HiSeq 101 bp paired-end sequencing (2x101) with an untrimmed total length of 11.07 Gbp and an estimated fragment size of 380 bp [12].

The original reads are available through the NCBI Sequence Read Archive at Accession SRX200676. We updated the 64 reference genomes sets from NCBI GenBank using the latest available assemblies with plasmid content (June 2017); the accession numbers are available as `accession-list-ref.txt` in the Zenodo repository, DOI: 10.5281/zenodo.818050. For convenience, the updated reference genome collection is available for download at the archival URL <https://osf.io/vbhy5/>.

Methods

The analysis code and run scripts for this paper are written in Python and bash, and are available at <https://github.com/dib-lab/2016-metagenome-assembly-eval/> (archived at Zenodo DOI: 10.5281/zenodo.818050). The scripts and overall pipeline were examined by the first and senior authors for correctness. In addition, the bespoke reference-based analysis scripts were tested by running them on a single-colony *E. coli* MG1655 data set with a high quality reference genome [20].

Quality Filtering

We removed adapters with Trimmomatic v0.30 in paired-end mode with the TruSeq adapters [21], using light quality score trimming (`LEADING:2 TRAILING:2 SLIDINGWINDOW:4:2 MINLEN:25`) as recommended in MacManes, 2014 [22].

Reference Coverage Profile

To evaluate how much of the reference metagenome was contained in the read data, we used `bwa aln` (v0.7.7.r441) to map reads to the reference

124 genome [23]. We then calculated how many reference bases were covered by
125 mapped reads (custom script `coverage-profile.py`).

126 Measuring k-mer inclusion and Jaccard similarity

127 We used MinHashing as implemented in sourmash to estimate k-mer inclu-
128 sion and Jaccard similarity between data sets [24]. MinHash signatures were
129 prepared with `sourmash compute` using `--scaled 10000`. K-mer inclusion
130 was computed by taking the ratio of the number of intersecting hashes with
131 the query over the total number of hashes in the subject MinHash. Jac-
132 card similarity was computed as in [25] by taking the ratio of the number
133 of intersecting hashes between the query and subject over the number of
134 hashes in the union. K-mer sizes for comparison were chosen at 21, 31, or
135 51, depending on the level of taxonomic specificity desired - genus, species,
136 or strain, respectively, as described in [26].

137 Where specified, high-abundance k-mers were selected for counting by
138 using the script `trim-low-abund.py` script with `-C 5` from khmer v2 [27,
139 28].

140 Assemblers

141 We assembled the quality-filtered reads using three different assemblers:
142 IDBA-UD [17], MetaSPAdes [18], and MEGAHIT [19]. For IDBA-UD v1.1.1
143 [17], we used `--pre-correction` to perform pre-correction before assembly
144 and `-r` for the pe files. IDBA could not ingest the single-ended files so they
145 were omitted from the assembly.

146 For MetaSPAdes v3.10.1 [18], we used `--meta --pe1-12 --pe1-s` where
147 `--meta` is used for metagenomic data sets, `--pe1-12` specifies the interlaced
148 reads for the first paired-end library, and `--pe1-s` provides the orphan reads
149 remaining from quality trimming.

150 For MEGAHIT v1.1.1-2-g02102e1 [19], we used `-l 101 -m 3e9 --cpu-only`
151 where `-l` is for maximum read length, `-m` is for max memory in bytes to
152 be used in constructing the graph, and `--cpu-only` to use only the CPU
153 and no GPUs. We also used `--presets meta-large` for large and complex
154 metagenomes, and `--12` and `-r` to specify the interleaved-paired-end and
155 single-end files respectively. MEGAHIT allows the specification of a memory
156 limit and we used `-M 1e+10` for 10 GB.

157 All three assemblies were executed on the same XSEDE Jetstream in-
158 stance (S1.Xxlarge) at Indiana University, running Ubuntu 16.04 (install

159 6/21/17, Ubuntu 16.04 LTS Development + GUI support + Docker; based
160 on Ubuntu cloud image for 16.04 LTS with basic dev tools, GUI/Xfce
161 added). Assemblers were limited to 16 threads. We recorded RAM and CPU
162 time for each assembly using `/usr/bin/time -v`. Install and execute de-
163 tails as well as output timings and logs are available in the `pipeline/runstats`
164 directory of the Zenodo release.

165 Unless otherwise mentioned, we eliminated all contigs less than 500 bp
166 from each assembly prior to further analysis.

167 Mapping

168 We aligned all quality-filtered reads to the reference metagenome with `bwa`
169 `aln` (v0.7.7.r441) [23]. We aligned paired-end and orphaned reads separately.
170 We then used `samtools` (v0.1.19) [29] to convert SAM files to BAM files for
171 both paired-end and orphaned reads. To count the unaligned reads, we
172 included only those records with the “4” flag in the SAM files [29].

173 Assembly analysis using NUCmer

174 We used the NUCmer tool from MUMmer3.23 [30] to align assemblies to the
175 reference genome with options `-coords -p`. Then we parsed the generated
176 “.coords” file using a custom script `analyze_assembly.py`, and calculated
177 several analysis metrics across all three assemblies at a 99% alignment iden-
178 tity.

179 Reference-based analysis of the assemblies

180 We conducted reference-based analysis of the assemblies under two condi-
181 tions. “Loose” alignment conditions used all available alignments, including
182 redundant and overlapping alignments. “Strict” alignment conditions took
183 only the longest alignment for any given contig, eliminating all other align-
184 ments.

185 The script `summarize-coords2.py` was used to calculate aligned cov-
186 erage from the loose alignment conditions: each base in the reference was
187 marked as “covered” if it was included in at least one alignment. The script
188 `analyze_ng50.py` was used to calculate NGA 50 for each individual refer-
189 ence genome.

190 **Analysis of chimeric misassemblies**

191 We analyzed each assembly for chimeric misassemblies by counting the num-
192 ber of contigs that contained matches to two distinct reference genomes. In
193 order to remove secondary alignments from consideration, we included only
194 the longest non-overlapping NUCmer alignments for each contig at a mini-
195 mum alignment identity of 99%. We then used the script `analyze_chimeric2.py`
196 to find individual contigs that matched more than one distinct reference
197 genome. As a negative control on our analysis, we verified that this ap-
198 proach yielded no positive results when applied to the alignments of the
199 reference metagenome against itself.

200 **Analysis of unmapped reads**

201 We conducted assembly and analysis of unmapped reads with MEGAHIT,
202 NUCmer, and sourmash as above. The new GenBank genomes are listed in
203 the Zenodo archive at the file `accession-list-unmapped.txt` and for con-
204 venience are available for download at the archival URL <https://osf.io/34ef8/>.

205 **Results**

206 **The raw data is high quality.**

207 The reads contains 11,072,579,096 bp (11.07 Gbp) in 109,629,496 reads with
208 101.0 average length (2x101bp Illumina HiSeq).

209 Trimming removed 686,735 reads (0.63%). After trimming, we retained
210 108,422,358 paired reads containing 10.94 Gbp with an average length of
211 100.9 bases. A total of 46.56 Mbp remained in 520,403 orphan reads with
212 an average length of 89.5 bases. In total, the quality trimmed data contained
213 10.98 Gbp in 108,942,761 reads. This quality trimmed (“QC”) data set was
214 used as the basis for all further analyses.

215 **The reference metagenome is not completely present in the** 216 **reads.**

217 We next evaluated the fraction of the reference genome covered by at least
218 one read (see Methods for details). Quality filtered reads cover 203,058,414
219 (98.76%) bases of the reference metagenome (205,603,715 bp total size). Fig-
220 ure 1 shows the cumulative coverage profile of the reference metagenome,

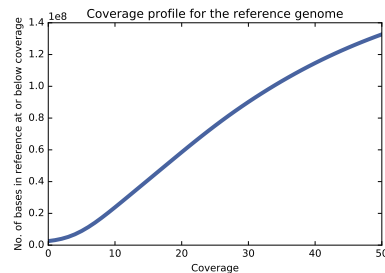


Figure 1: Cumulative coverage profile for the reference metagenome, based on read mapping.

Table 1: Jaccard containment of the reference in the reads

k-mer size	% reference in reads
21	96.8%
31	95.9%
41	94.9%
51	94.1%

and the percentage of bases with that coverage. Most of the reference metagenome was covered at least minimally; only 3.33% of the reference metagenome had mapping coverage <5 , and 1.24% of the bases in the reference were not covered by any reads in the QC data set.

In order to evaluate reconstructability with De Bruijn graph assemblers, we next examined k-mer containment of the reference in the reads for k of 21, 31, 41, and 51 (Table 1). The k-mer overlap decreases from 96.8% to 94.1% as the k-mer size increases. This could be caused by low coverage of some portions of the reference and/or variation between the reads and the reference.

Some individual reference genomes are poorly represented in the reads.

To see if specific reference genomes exhibited low coverage, we analyzed read mapping coverage for individual genomes. Of the 64 reference genomes used in the metagenome, 60 had a per-base mapping coverage above 95%. The remaining four varied significantly (Table 2), with *F. nucleatum* the lowest – only 47.6% of the bases in the reference genome are covered by one or more

Table 2: Top uncovered genomes

Genome	Read coverage
<i>Desulfovibrio vulgaris</i> DP4	93.2%
<i>Thermus thermophilus</i> HB27	91.1%
<i>Enterococcus faecalis</i> V583	74.6%
<i>Fusobacterium nucleatum</i>	47.6%

Table 3: Genomes removed from reference for low 51-mer presence

51-mers in reads	Genome
98.7	<i>Leptothrix cholodnii</i>
98.7	<i>Haloferax volcanii</i> DS2
98.6	<i>Salinispora tropica</i> CNB-440
97.4	<i>Deinococcus radiodurans</i>
97.2	<i>Zymomonas mobilis</i>
97.1	<i>Ruegeria pomeroyi</i>
96.8	<i>Shewanella baltica</i> OS223
95.5	<i>B. bronchiseptica</i> D989
94.5	<i>Burkholderia xenovorans</i>
72.0	<i>Desulfovibrio vulgaris</i> DP4
65.0	<i>Thermus thermophilus</i> HB27
53.4	<i>Enterococcus faecalis</i>
4.7	<i>Fusobacterium nucleatum</i> ATCC 25586

mapped reads.

We next did a 51-mer containment analysis of each reference genome in the reads; $k=51$ was chosen so as to be specific to strain content [26]. 99% or more of the constituent 51-mers for 51 of the 64 reference genomes were present in the reads, suggesting that each of the 51 genomes was entirely present at some minimal coverage.

We excluded the remaining 13 genomes (see Table 3) from any further reference-based analysis because interpreting recovery and misassembly statistics for these genomes would be confounding; also see the discussion of strain variants, below.

Table 4: Running Time and Memory Utilization

Assembler	CPU time	Wall time	RAM (Max RSS)
MEGAHIT	1191m	1h 33m	10 GB
IDBA-UD	2117m	2h 41m	21 GB
MetaSPAdes	2554m	4h 7m	28 GB

MEGAHIT is the fastest and lowest-memory assembler evaluated

We ran three commonly used metagenome assemblers on the QC data set: IDBA-UD, MetaSPAdes, and MEGAHIT. We recorded the time and memory usage of each (Table 4). In computational requirements, MEGAHIT outperformed both MetaSPAdes and IDBA-UD, producing an assembly in 1.5 hours (“wall time”) – 1.7 times faster than IDBA and 2.6 times faster than MetaSPAdes. MEGAHIT used only 10 GB of RAM as requested – half to almost a third of the memory used by IDBA and MetaSPAdes, respectively.

CPU time measurements (which include processing on multiple CPU cores) show that all three assemblers use multiple cores effectively.

The assemblies contain most of the raw data

Table 5: Read and high-abundance (> 5) k-mer exclusion from assemblies

Assembly	Unmapped Reads	51-mers omitted
IDBA	3,328,674 (3.05%)	2.4%
MetaSPAdes	3,844,123 (3.52%)	3.2%
MEGAHIT	2,737,640 (2.51%)	2.8%

We assessed read inclusion in assemblies by mapping the QC reads to the length-filtered assemblies and counting the remaining unmapped reads. Depending on the assembly, between 2.7 million and 3.9 million reads (2.5-3.5%) did not map to the assemblies (Table 5). All of the assemblies included the large majority of high-abundance 51-mers (more than 96.8% in all cases).

Table 6: Contig coverage of reference with loose alignment conditions.

Assembly	bases aligned	duplication	51-mers
MEGAHIT	94.8%	1.0%	96.7%
MetaSPAdes	93.1%	1.1%	96.2%
IDBA	93.6%	0.98%	97.2%

266 **Much of the reference is covered by the assemblies.**

267 We next evaluated the extent to which the assembled contigs recovered the
 268 “known/true” metagenome sequence by aligning each assembly to the ad-
 269 justed reference (Table 6). Each of the three assemblers generates contigs
 270 that cover more than 93.1% of the reference metagenome at high identity
 271 (99%) with little duplication (approximately 1%). All three assemblies con-
 272 tain between 96.2% and 97.2% of the 51-mers in the reference.

273 At 99% identity with the loose mapping approach, approximately 2.5% of
 274 the reference is missed by all three assemblers, while 1.7% is uniquely covered
 275 by MEGAHIT, 0.74% is uniquely covered by MetaSPAdes, and 0.64% is
 276 uniquely covered by IDBA.

277 **The generated contigs are broadly accurate.**

Table 7: Contig accuracy measured by reference coverage with strict align-
 ment.

Assembly	% covered
MEGAHIT	89.3%
IDBA	87.7%
MetaSPAdes	83.4%

278 When counting only the best (longest) alignment per contig at a 99%
 279 identity threshold, each of the three assemblies recovers more than 87.3% of
 280 the reference, with MEGAHIT recovering the most – 89.3% of the reference
 281 (Table 7).

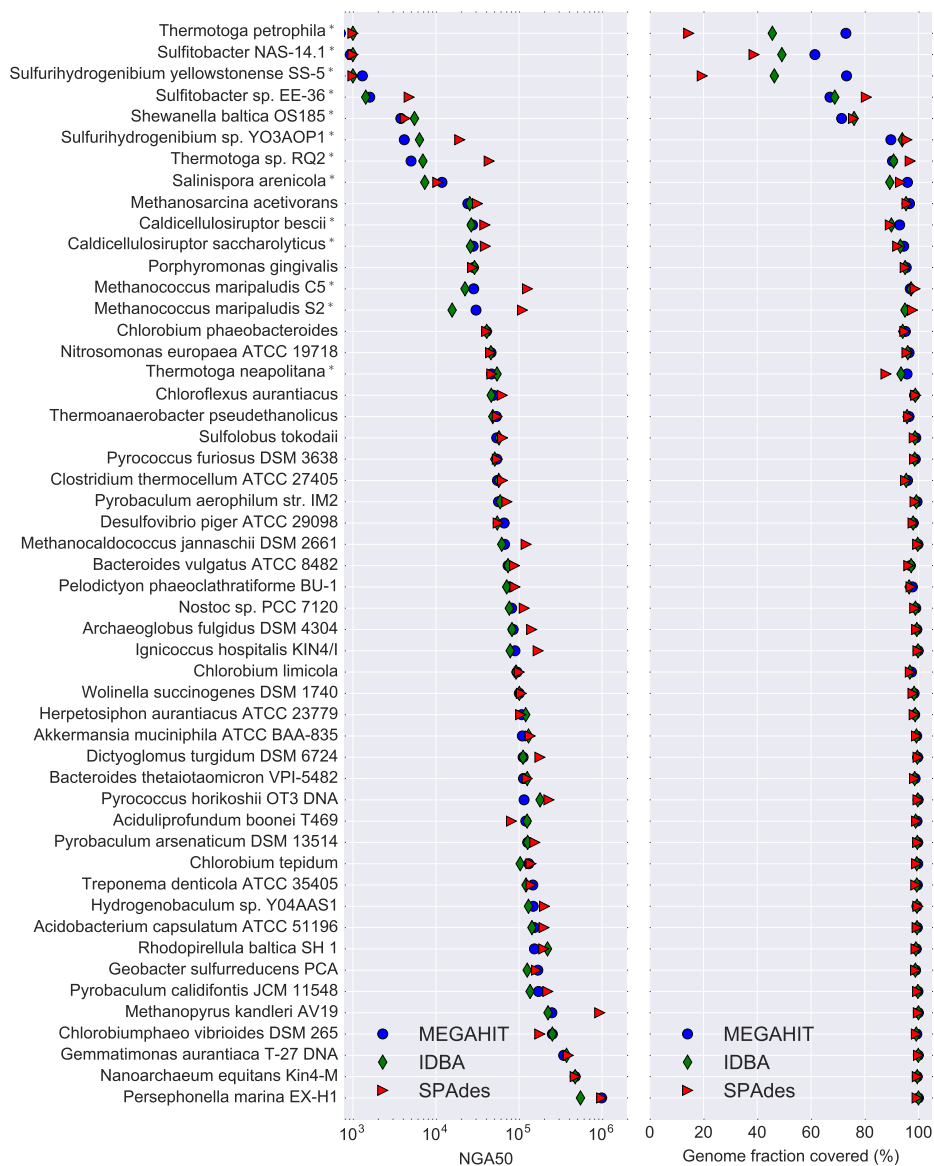


Figure 2: NGA50 and genome fraction covered, by genome and assembler. A '*' after the name indicates the presence of at least one other genome with > 2% Jaccard similarity at k=31 in the community.

282 **Individual genome statistics vary widely in the assemblies.**

283 We computed the NGA50 for each individual genome and assembly in order
284 to compare assembler performance on genome recovery (see left panel of Fig-
285 ure 2). The NGA50 statistics for individual genomes vary widely, but there
286 are consistent assembler-specific trends: IDBA yields the lowest NGA50 for
287 28 of the 51 genomes, while MetaSPAdes yields the highest NGA50 for 32
288 of the 51 genomes.

289 We also evaluated aligned coverage per genome for each of the three
290 assemblies (right panel, Figure 2). We found that 13 of the 51 genomes were
291 missing 5% or more of bases in at least one assembly, despite all 51 genomes
292 having 99% or higher read- and 51-mer coverage.

293 There are 12 genomes with k=31 Jaccard similarity greater than 2%
294 to other genomes in the community, and these (denoted by '*' after the
295 name) typically had lower NGA50 and aligned coverage numbers than other
296 genomes. In particular, these constituted 12 of the 13 genomes missing 5%
297 or more of their content, and the lowest eight NGA50 numbers.

298 **Longer contigs are less likely to be chimeric.**

Table 8: Chimeric contigs by contig length.

Assembly	> 50kb	> 5kb	> 500 bp
IDBA	0	1	7 (0.06%)
MEGAHIT	1	4	14 (0.13%)
MetaSPAdes	0	3	30 (0.48%)

299 Chimerism is the formation of contigs that include sequence from multi-
300 ple genomes. We evaluated the rate of chimerism in contigs at three different
301 contig length cutoffs: 500bp, 5kb, and 50kb (Table 8). We found that the
302 percentage of contigs that match to the genomes of two or more different
303 species drop as the minimum contig size increases, to the point where only
304 the MEGAHIT assembly had a single chimeric contig longer than 50kb.
305 Overall, chimeric misassemblies were rare, with no assembler generating
306 more than 30 chimeric contigs out of thousands of total contigs.

307 **The unmapped reads contain strain variants of reference genomes.**

308 Approximately 4.8 million reads (4.4%) from the QC data set did not map
309 anywhere in the reference provided by the authors of [12]. We extracted

Table 9: GenBank genomes detected in assembly of unmapped reads

match	GenBank genome
44.1%	<i>Fusobacterium</i> sp. OBRC1
23.0%	<i>P. ruminis</i> strain ML2
18.2%	<i>Thermus thermophilus</i> HB8
7.7%	<i>P. ruminis</i> strain CGMCC
8.2%	<i>Enterococcus faecalis</i> M7
7.3%	<i>F. nucleatum</i> 13_3C
3.7%	<i>F. nucleatum</i> subsp. <i>polymorphum</i>
2.9%	<i>Fusobacterium hwasookii</i>
1.0%	<i>E. coli</i> isolate YS
1.7%	<i>F. nucleatum</i> subsp. <i>polymorphum</i> , alt.
1.9%	<i>F. nucleatum</i> subsp. <i>vincentii</i>

and assembled these reads in isolation using MEGAHIT, yielding 6.5 Mbp of assembly in 1711 contigs > 500bp in length. We then did a k-mer inclusion analysis of this assembly against all of the GenBank genomes at k=31, and estimated the fraction of the k-mers that belonged to different species (Table 9). We find that 51.1% of the k-mer content of these contigs positively match to a genome present in GenBank but not in the reference metagenome.

To verify these assignments, we aligned the MEGAHIT assembly of unmapped reads to the GenBank genomes in Table 9 with NUCmer using “loose” alignment criteria. We found that 1.78 Mbp of the contigs aligned at 99% identity or better to these GenBank genomes. We also confirmed that, as expected, there are no matches in this assembly to the full updated reference metagenome.

We note that all but the two *P. ruminis* matches and the *E. coli* isolate YS are strain variants of species that are part of the defined community but are not completely present in the reads (see Table 2). For *Proteiniclasticum ruminis*, there is no closely related species in the mock community design, and very little of the MEGAHIT assembly aligns to known *P. ruminis* genomes at 99%. However, there are many alignments to *P. ruminis* at 94% or higher, for approximately 2.73 Mbp total. This suggests that the unmapped reads contain at least some data from a novel species of *Proteiniclasticum*; this matches the observation in [12] of a contaminating genome from an unknown *Clostridium* spp., as at the time there was no *P. ruminis* genome.

334 Discussion

335 Assembly recovers basic content sensitively and accurately.

336 All three assemblers performed well in assembling contigs from the con-
337 tent that was fully present in reads and k-mers. After length filtering,
338 all three assemblies contained more than 95% of the reference (Table 6);
339 even with removal of secondary alignments, more than 87% was recovered
340 by each assembler (Table 7). About half the constituent genomes had an
341 NGA50 of 50kb or higher (Figure 2), which, while low for current Illumina
342 single-genome sequencing, is sufficient to recover operon-level relationships
343 for many genes.

344 The presence of multiple closely related genomes confounds 345 assembly.

346 In agreement with CAMI, we also find that the presence of closely related
347 genomes in the metagenome causes loss of assembly [3]. This is clearly shown
348 by Figure 2, where 12 of the bottom 14 genomes by NGA50 (left panel)
349 also exhibit poor genome recovery by assembly (right panel). Interestingly,
350 different assemblers handle this quite differently, with e.g. MetaSPAdes
351 failing to recover essentially any of *Thermotoga petrophila*, while MEGAHIT
352 recovers 73%. The presence of nearby genomes is an almost perfect predictor
353 that one or more assembler will fail to recover 5% or more - of the 13/51
354 genomes for which less than 95% is recovered, 12 of them have close genomes
355 in the community. Interestingly, very little similarity is needed - all genomes
356 with Jaccard similarity of 2% or higher at k=31 exhibit these problems.

357 The *Shewanella baltica* OS185 genome is a good example: there are two
358 strain variants, OS185 and OS223, present in the defined community. Both
359 are present at more than 99% in the reads, and more than 98% in 51-mers,
360 but only 75% of *S. baltica* OS185 and 50% of *S. baltica* OS223 are recovered
361 by assemblers. This is a clear case of “strain confusion” where the assemblers
362 simply fail to output contigs for a substantial portion of the two genomes.

363 Another interest of this study was to examine cross-species chimeric as-
364 sembly, in which a single contig is formed from multiple genomes. In Table 8,
365 we show that there is relatively little cross-species chimerism. Surprisingly,
366 what little is present is length-dependent: longer contigs are less likely to
367 be chimeric. This might well be due to the same “strain confusion” effect
368 as above, where contigs that share paths in the assembly graphs are broken
369 in twain.

370 **MEGAHIT performs best by several metrics.**

371 MEGAHIT is clearly the most efficient computationally, outperforming both
372 MetaSPAdes and IDBA by 3-9x in memory and 12-23x in time (Table 4).
373 The MEGAHIT assembly also included more of the reads than either IDBA
374 or MetaSPAdes, and omitted only 0.4% more of the unique 51-mers from
375 the reads than IDBA. MEGAHIT covered more of the reference genome
376 with both loose and strict alignments (Table 6 and Table 7), with little
377 duplication. This is clearly because of MEGAHIT’s generally superior per-
378 formance in recovering the genomes of closely related strains (Figure 2, right
379 panel). The sum “fraction of genome recovered” is arguably the most im-
380 portant measure of a metagenome assembler (see [5] in particular) and here
381 MEGAHIT excels for individual genomes even in the presence of strain vari-
382 ation.

383 When comparing details of sequence recovery between the assemblers,
384 the assembly content differs by only a small amount when loose alignments
385 are allowed: all three assemblers miss more content (approximately 2.5% of
386 the reference) than they generate uniquely (1.7% or less). In addition to
387 preferring no one assembler over any other, this suggests that combining as-
388 semblies may have little value in terms of recovering additional metagenome
389 content.

390 **The missing reference may be present in strain variants of the**
391 **intended species.**

392 Several individual genomes are missing in measurable portion from the QC
393 reads (Table 2), and many QC reads (4.4% of 108m) did not map to the
394 full reference metagenome. These appear to be related issues: upon anal-
395 ysis of the unmapped reads against GenBank, we find that many of the
396 contigs assembled from the unmapped reads can be assigned to strain vari-
397 ants of the species in the mock community (Table 9). This suggests that
398 the constructors of the mock community may have unintentionally included
399 strain variants of *Fusobacterium nucleatum*, *Thermus thermophilus* HB27,
400 and *Enterococcus faecalis*; note that the microbes used were sourced from
401 the community rather than the ATCC (M. Podar, pers. communication). In
402 addition, we detect what may be portions of a novel member of the *Proteini-*
403 *clasticum* genus in the assembly of these reads - this is likely the *Clostridium*
404 spp. detected through amplicon sequencing in [12].

405 Without returning to the original DNA samples, it is impossible to con-
406 clusively confirm that unintended strains were used in the construction of the

407 mock community. In particular, our analysis is dependent on the genomes in
408 GenBank: the genomes we detect in the contigs are clearly closely related to
409 GenBank genomes not in the reference metagenome, based on k-mer anal-
410 ysis and contig alignment. However, GenBank is unlikely to contain the
411 exact genomes of the actually included strain variants, rendering conclusive
412 identification impossible.

413 Conclusions

414 Overall, assembly of this mock community works well, with good recovery
415 of known genomic sequence for the majority of genomes. All three assem-
416 blers that we evaluated recover similar amounts of most genomic sequence,
417 but (recapitulating several other studies [3, 5, 15]) MEGAHIT is compu-
418 tationally the most efficient of the three. We note that assembly resolves
419 substantial portions of several previously undetected strain variants, as well
420 as recovering a substantial portion of a novel *Proteiniclasticum* spp. that
421 was detected via amplicon analysis in [12], suggesting that assembly is a
422 useful complement to amplicon or reference-based analyses.

423 The presence of closely related strains is a major confounder of metagenome
424 assembly, and causes assemblers to drop considerable portions of genomes
425 that (based on read mapping and k-mer inclusion) are clearly present. In this
426 relatively simple community, this strain confusion is present but does not
427 dominate the assembly. However, real microbial communities are likely to
428 have many closely related strains and any resulting loss of assembly would
429 be hard to detect in the absence of good reference genomes. While high
430 polymorphism rates in e.g. animal genomes are known to cause duplication
431 or loss of assembly, some solutions have emerged that make use of assump-
432 tions of uniform coverage and diploidy [31]. These solutions cannot however
433 be transferred directly to metagenomes, which have unknown abundance
434 distributions and strain content.

435 An additional concern is that metagenome assemblies are often per-
436 formed after pooling data sets to increase coverage (e.g. [4, 32]); this pooled
437 data is more likely to contain multiple strains, which would then in turn
438 adversely affect assembly of strains. This may not be resolvable within the
439 current paradigm of assembly, which focuses on outputting linear assem-
440 blies that cannot properly represent strain variation. The human genomics
441 community is moving towards using *reference graphs*, which can represent
442 multiple incompatible variants in a single data structure [33]; this approach,
443 however, requires high-quality isolate reference genomes, which are generally

444 unavailable for environmental microbes.

445 Long read sequencing (and related technologies) will undoubtedly help
446 resolve strain variation in the future, but even with highly accurate long-
447 read sequencing, current sequencing depth is still too low to resolve deep
448 environmental metagenomes [34, 35]. It is unclear how well long error-
449 prone reads (such as those output by Pacific Biosciences SMRT [36] and
450 Oxford Nanopore instruments [37]) will perform on complex metagenomes:
451 with high error rates, deep coverage of each individual genome is required
452 to achieve accurate assembly, and this may not be easily obtainable for
453 complex communities. Single-molecule barcoding (e.g. 10X Genomics [38])
454 and HiC approaches [39] show promise but these remain untested on well-
455 defined complex communities and are still challenged by the complexity of
456 complex environmental metagenomes; see [40, 41, 42].

457 Much of our analysis above depends on having a high-quality “mock”
458 metagenome. While computationally constructed synthetic communities
459 and computational “spike-ins” to real data sets can provide valuable controls
460 (e.g. see [15] and [43]) we strongly believe that standardized communities
461 constructed *in vitro* and sequenced with the latest technologies are critical to
462 the evaluation of both canonical and emerging tools, e.g. efforts such as [44].
463 From the perspective of tool evaluation, we must disagree somewhat with
464 Vollmers et al. [5]: good metagenome tool evaluation necessarily depends on
465 mock communities that are as realistic as we can make them. Likewise, from
466 the perspective of bench biologists, actually sequencing real DNA is critical
467 because it can evaluate confounding effects such as kit contamination [45].
468 Large-scale studies of computational approaches systematically applied to
469 mock communities such as CAMI [3] can then provide fair comparisons of
470 entire toolchains (wet + dry) applied to these mock communities.

471 We omitted two important questions in this study: binning and choice
472 of parameters. We chose not to evaluate genome binning because most
473 binning strategies either operate post-assembly (see e.g. [46]), in which
474 case the challenges with assembly discussed above will apply; or require
475 multiple samples (e.g. [47]), which we do not have. We also chose to use
476 only default parameters with all three assemblers, for two reasons. First,
477 we are not aware of any widely used automated approaches for determining
478 the “best” set of parameters or evaluating the output, other than those
479 integrated into the assemblers themselves (e.g. choice of k-mer sizes), and
480 absent such guidance we do not feel comfortable blessing any particular set of
481 parameters; here the choice of default parameters is parsimonious. Second,
482 any parameter exploration pipeline would not only need to be automated

483 but would need to run multiple assemblies, whose time and resource usage
484 should be measured; in this case, any comparison based on runtime of the
485 parameter choice pipeline should naturally favor MEGAHIT because of its
486 substantial advantage in computational efficiency.

487 **Author contributions**

488 SA, LI and CTB developed, tested, and executed the analytical pipeline.
489 SA and CTB created the tables and figures and wrote the paper.

490 **Competing interests**

491 No competing interest to our knowledge.

492 **Grant information**

493 This work is funded by Gordon and Betty Moore Foundation Grant GBMF4551
494 and NIH NHGRI R01 grant HG007513-03, both to CTB.

495 **Acknowledgments**

496 We thank Michael R. Crusoe and Phillip T. Brooks for input on analysis and
497 pipeline development. We thank Migun Shakya, Mircea Podar, Jiarong Guo,
498 Harald R. Gruber-Vodicka, Juliane Wippler, Krista Ternus, and Stephen
499 Turner for valuable comments on drafts of this manuscript.

500 **References**

- 501 [1] Jay Ghurye, Victoria Cepeda-Espinoza, and Mihai Pop. Metagenomic assem-
502 bly: Overview, challenges and applications. *The Yale Journal of Biology and*
503 *Medicine*, 89(3):353–362, 2016.
- 504 [2] Nikos C. Kyrpides, Philip Hugenholtz, Jonathan A. Eisen, Tanja Woyke,
505 Markus Göker, Charles T. Parker, Rudolf Amann, Brian J. Beck, Patrick S. G.
506 Chain, Jongsik Chun, Rita R. Colwell, Antoine Danchin, Peter Dawyndt, Tom
507 Dedeurwaerdere, Edward F. DeLong, John C. Detter, Paul De Vos, Timothy J.
508 Donohue, Xiu-Zhu Dong, Dusko S. Ehrlich, Claire Fraser, Richard Gibbs, Jack
509 Gilbert, Paul Gilna, Frank Oliver Glöckner, Janet K. Jansson, Jay D. Keasling,
510 Rob Knight, David Labeda, Alla Lapidus, Jung-Sook Lee, Wen-Jun Li, Juncai
511 MA, Victor Markowitz, Edward R. B. Moore, Mark Morrison, Folker Meyer,
512 Karen E. Nelson, Moriya Ohkuma, Christos A. Ouzounis, Norman Pace, Julian

513 Parkhill, Nan Qin, Ramon Rossello-Mora, Johannes Sikorski, David Smith,
514 Mitch Sogin, Rick Stevens, Uli Stingl, Ken ichiro Suzuki, Dorothea Taylor,
515 Jim M. Tiedje, Brian Tindall, Michael Wagner, George Weinstock, Jean Weis-
516 senbach, Owen White, Jun Wang, Lixin Zhang, Yu-Guang Zhou, Dawn Field,
517 William B. Whitman, George M. Garrity, and Hans-Peter Klenk. Genomic
518 encyclopedia of bacteria and archaea: Sequencing a myriad of type strains.
519 *PLoS Biology*, 12(8):e1001920, aug 2014. doi: 10.1371/journal.pbio.1001920.
520 URL <https://doi.org/10.1371/journal.pbio.1001920>.

521 [3] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan
522 Janssen, Johannes Droege, Ivan Gregor, Stephan Majda, Jessika Fiedler,
523 Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue
524 Sparholt Jorgensen, Nicole Shapiro, Philip D Blood, Alexey Gurevich, Yang
525 Bai, Dmitrij Turaev, Matthew Z DeMaere, Rayan Chikhi, Niranjan Nagara-
526 jan, Christopher Quince, Lars Hestbjerg Hansen, Soren J Sorensen, Burton
527 K H Chia, Bertrand Denis, Jeff L Froula, Zhong Wang, Robert Egan, Dong-
528 wan Don Kang, Jeffrey J Cook, Charles Deltel, Michael Beckstette, Claire
529 Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu-Wei
530 Wu, Steven W Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter
531 Meinicke, Michael Barton, Thomas Lingner, Hsin-Hung Lin, Yu-Chieh Liao,
532 Genivaldo Gueiros Z. Silva, Daniel A Cuevas, Robert A Edwards, Surya Saha,
533 Vitor C Piro, Bernhard Y Renard, Mihai Pop, Hans-Peter Klenk, Markus
534 Goeker, Nikos Kyrpides, Tanja Woyke, Julia A Vorholt, Paul Schulze-Lefert,
535 Edward M Rubin, Aaron E Darling, Thomas Rattei, and Alice C McHardy.
536 Critical assessment of metagenome interpretation - a benchmark of compu-
537 tational metagenomics software. *bioRxiv*, 2017. doi: 10.1101/099127. URL
538 <http://biorxiv.org/content/early/2017/01/09/099127>.

539 [4] I. Sharon, M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman,
540 and J. F. Banfield. Time series community genomics analysis reveals rapid
541 shifts in bacterial species, strains, and phage during infant gut colonization.
542 *Genome Research*, 23(1):111–120, aug 2012. doi: 10.1101/gr.142315.112. URL
543 <https://doi.org/10.1101/gr.142315.112>.

544 [5] John Vollmers, Sandra Wiegand, and Anne-Kristin Kaster. Compar-
545 ing and evaluating metagenome assembly tools from a microbiol-
546 ogist’s perspective - not only size matters! *PLOS ONE*, 12
547 (1):e0169662, jan 2017. doi: 10.1371/journal.pone.0169662. URL
548 <https://doi.org/10.1371/journal.pone.0169662>.

549 [6] Jorge F Vázquez-Castellanos, Rodrigo García-López, Vicente Pérez-Brocal,
550 Miguel Pignatelli, and Andrés Moya. Comparison of different assembly and
551 annotation tools on analysis of simulated viral metagenomic communities in
552 the gut. *BMC genomics*, 15(1):1, 2014.

553 [7] Konstantinos Mavromatis, Natalia Ivanova, Kerrie Barry, Harris Shapiro, Eu-
554 gene Goltsman, Alice C McHardy, Isidore Rigoutsos, Asaf Salamov, Frank

555 Korzeniewski, Miriam Land, et al. Use of simulated data sets to evaluate the
556 fidelity of metagenomic processing methods. *Nature methods*, 4(6):495–500,
557 2007.

558 [8] David B Jaffe, Jonathan Butler, Sante Gnerre, Evan Mauceli, Kerstin
559 Lindblad-Toh, Jill P Mesirov, Michael C Zody, and Eric S Lander. Whole-
560 genome sequence assembly for mammalian genomes: Arachne 2. *Genome*
561 *research*, 13(1):91–96, 2003.

562 [9] Samuel Aparicio, Jarrod Chapman, Elia Stupka, Nik Putnam, Jer-ming Chia,
563 Paramvir Dehal, Alan Christoffels, Sam Rash, Shawn Hoon, Arian Smit, et al.
564 Whole-genome shotgun assembly and analysis of the genome of *fugu rubripes*.
565 *Science*, 297(5585):1301–1310, 2002.

566 [10] Anveshi Charuvaka and Huzefa Rangwala. Evaluation of short read metage-
567 nomic assembly. *BMC genomics*, 12(2):1, 2011.

568 [11] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein,
569 Steven JM Jones, and Inanç Birol. Abyss: a parallel assembler for short read
570 sequence data. *Genome research*, 19(6):1117–1123, 2009.

571 [12] Shakya Migun, Christopher Quince, James Campbell, Zamin Yang, Christo-
572 pher Schadt, and Mircea Podar. Comparative metagenomic and rrna microbial
573 diversity characterization using archaeal and bacterial synthetic communities.
574 *Enivromental Microbiology*, 15(6):1882–1899, 2013.

575 [13] Brandon K. B. Seah and Harald R. Gruber-Vodicka. gbtools: In-
576 teractive visualization of metagenome bins in r. *Frontiers in Mi-
577 crobiology*, 6, dec 2015. doi: 10.3389/fmicb.2015.01451. URL
578 <https://doi.org/10.3389/fmicb.2015.01451>.

579 [14] Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-
580 Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah
581 Lam. MEGAHIT v1.0: A fast and scalable metagenome assembler
582 driven by advanced methodologies and community practices. *Meth-
583 ods*, 102:3–11, jun 2016. doi: 10.1016/j.ymeth.2016.02.020. URL
584 <https://doi.org/10.1016/j.ymeth.2016.02.020>.

585 [15] Andries Johannes van der Walt, Marc Warwick Van Goethem,
586 Jean-Baptiste Ramond, Thulani Peter Makhalanyane, Oleg Reva,
587 and Don Arthur Cowan. Assembling metagenomes, one com-
588 munity at a time. *bioRxiv*, 2017. doi: 10.1101/120154. URL
589 <http://biorxiv.org/content/early/2017/06/06/120154>.

590 [16] William W. Greenwald, Niels Klitgord, Victor Seguritan, Shibu Yooseph,
591 J. Craig Venter, Chad Garner, Karen E. Nelson, and Weizhong Li. Utilization
592 of defined microbial communities enables effective evaluation of meta-genomic

- assemblies. *BMC Genomics*, 18(1), apr 2017. doi: 10.1186/s12864-017-3679-5. URL <https://doi.org/10.1186/s12864-017-3679-5>.
- [17] Yu Peng, Henry C.M. Leung, S.M. Yiu, and Francis Y.L. Chin. Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28:1420–1428, 2012.
- [18] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, mar 2017. doi: 10.1101/gr.213959.116. URL <https://doi.org/10.1101/gr.213959.116>.
- [19] Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. Megahit v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102:3–11, 2016.
- [20] H Chitsaz, JL Yee-Greenbaum, G Tesler, MJ Lombardo, CL Dupont, JH Badger, M Novotny, DB Rusch, LJ Fraser, NA Gormley, O Schulz-Trieglaff, GP Smith, DJ Evers, PA Pevzner, and RS Lasken. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol*, 29(10):915–21, 2011.
- [21] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.
- [22] Matthew D MacManes. On the optimal trimming of high-throughput mrna sequence data. *Frontiers in genetics*, 5:13, 2014.
- [23] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.
- [24] C. Titus Brown and Luiz Irber. sourmash: a library for MinHash sketching of DNA. *The Journal of Open Source Software*, 1(5), sep 2016. doi: 10.21105/joss.00027. URL <https://doi.org/10.21105/joss.00027>.
- [25] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1), jun 2016. doi: 10.1186/s13059-016-0997-x. URL <https://doi.org/10.1186/s13059-016-0997-x>.
- [26] David Koslicki and Daniel Falush. Metapalette: a k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. *mSystems*, 1(3), 2016. doi: 10.1128/mSystems.00020-16. URL <http://msystems.asm.org/content/1/3/e00020-16>.

- [27] Zhang Qingpeng, Awad Sherine, and Brown Titus. Crossing the streams: a framework for streaming analysis of short dna sequencing reads. *PeerJ PrePrints* 3:e1100 <https://dx.doi.org/10.7287/peerj.preprints.890v1>, 2015.
- [28] MR Crusoe, HF Alameldin, S Awad, E Boucher, A Caldwell, R Cartwright, A Charbonneau, B Constantinides, G Edverson, S Fay, J Fenton, T Fenzl, J Fish, L Garcia-Gutierrez, P Garland, J Gluck, I Gonzlez, S Guermond, J Guo, A Gupta, JR Herr, A Howe, A Hyer, A Hrpfer, L Irber, R Kidd, D Lin, J Lippi, T Mansour, P McA’Nulty, E McDonald, J Mizzi, KD Murray, JR Nahum, K Nanlohy, AJ Nederbragt, H Ortiz-Zuazaga, J Ory, J Pell, C Pepe-Ranne, ZN Russ, E Schwarz, C Scott, J Seaman, S Sievert, J Simpson, CT Skennerton, J Spencer, R Srinivasan, D Standage, JA Stapleton, SR Steinman, J Stein, B Taylor, W Trimble, HL Wiencko, M Wright, B Wyss, Q Zhang, e zyme, and CT Brown. The khmer software package: enabling efficient nucleotide sequence analysis [version 1; referees: 2 approved, 1 approved with reservations]. *F1000Research*, 4(900), 2015. doi: 10.12688/f1000research.6924.1.
- [29] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [30] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):1, 2004.
- [31] J. H. Kim, M. S. Waterman, and L. M. Li. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Research*, 17(7):1101–1110, jun 2007. doi: 10.1101/gr.5894107. URL <https://doi.org/10.1101/gr.5894107>.
- [32] Ping Hu, Lauren Tom, Andrea Singh, Brian C. Thomas, Brett J. Baker, Yvette M. Piceno, Gary L. Andersen, and Jillian F. Banfield. Genome-resolved metagenomic analysis reveals roles for candidate phyla and other microbial community members in biogeochemical transformations in oil reservoirs. *mBio*, 7(1):e01669–15, jan 2016. doi: 10.1128/mbio.01669-15. URL <https://doi.org/10.1128/mbio.01669-15>.
- [33] Benedict Paten, Adam M Novak, Jordan M Eizenga, and Erik Garrison. Genome graphs and the evolution of genome inference. *Genome research*, 27(5):665–676, 2017.
- [34] Itai Sharon, Michael Kertesz, Laura A. Hug, Dmitry Pushkarev, Timothy A. Blauwkamp, Cindy J. Castelle, Mojgan Amirebrahimi, Brian C. Thomas, David Burstein, Susannah G. Tringe, Kenneth H. Williams, and Jillian F. Banfield. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Research*, 25(4):534–543, feb 2015. doi: 10.1101/gr.183012.114. URL <https://doi.org/10.1101/gr.183012.114>.

- [35] Richard Allen White, Eric M. Bottos, Taniya Roy Chowdhury, Jeremy D. Zucker, Colin J. Brislawn, Carrie D. Nicora, Sarah J. Fansler, Kurt R. Glaesemann, Kevin Glass, and Janet K. Jansson. Molecule long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. *mSystems*, 1(3):e00045–16, jun 2016. doi: 10.1128/msystems.00045-16. URL <https://doi.org/10.1128/msystems.00045-16>.
- [36] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, jan 2009. doi: 10.1126/science.1162986. URL <https://doi.org/10.1126/science.1162986>.
- [37] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of DNA in a nanopore at 5-AA precision. *Nature Biotechnology*, 30(4):344–348, feb 2012. doi: 10.1038/nbt.2147. URL <https://doi.org/10.1038/nbt.2147>.
- [38] Eli Moss, Alex Bishara, Ekaterina Tkachenko, Joyce B Kang, Tessa M Andermann, Christina Wood, Christine Handy, Hanlee Ji, Serafim Batzoglou, and Ami S Bhatt. De novo assembly of microbial genomes from human gut metagenomes using barcoded short read sequences. *bioRxiv*, 2017. doi: 10.1101/125211. URL <http://biorxiv.org/content/early/2017/04/07/125211>.
- [39] Caiti Smukowski Heil, Joshua N. Burton, Ivan Liachko, Anne Friedrich, Noah A. Hanson, Cody L. Morris, Joseph Schacherer, Jay Shendure, James H. Thomas, and Maitreya J. Dunham. Identification of a novel interspecific hybrid yeast from a metagenomic open fermentation sample using hi-c. *bioRxiv*, 2017. doi: 10.1101/150722. URL <http://biorxiv.org/content/early/2017/06/15/150722>.
- [40] Joshua N. Burton, Ivan Liachko, Maitreya J. Dunham, and Jay Shendure. Species-level deconvolution of metagenome assemblies with hi-c-based contact probability maps. *G3*, 4(7):1339–1346, may 2014. doi: 10.1534/g3.114.011825. URL <https://doi.org/10.1534/g3.114.011825>.
- [41] Martial Marbouty, Axel Cournac, Jean-François Flot, Hervé Marie-Nelly, Julien Mozziconacci, and Romain Koszul. Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organiza-

- tion in microorganisms. *eLife*, 3, dec 2014. doi: 10.7554/elife.03318. URL <https://doi.org/10.7554/elife.03318>.
- [42] Christopher W. Beitel, Lutz Froenicke, Jenna M. Lang, Ian F. Korf, Richard W. Micheltmore, Jonathan A. Eisen, and Aaron E. Darling. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*, 2:e415, may 2014. doi: 10.7717/peerj.415. URL <https://doi.org/10.7717/peerj.415>.
- [43] Adina Chuang Howe, Janet K Jansson, Stephanie A Malfatti, Susannah G Tringe, James M Tiedje, and C Titus Brown. Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences*, 111(13):4904–4909, 2014.
- [44] Bonnie L. Brown, Mick Watson, Samuel S. Minot, Maria C. Rivera, and Rima B. Franklin. MinIONTM nanopore sequencing of environmental metagenomes: a synthetic approach. *GigaScience*, 6(3):1–10, feb 2017. doi: 10.1093/gigascience/gix007. URL <https://doi.org/10.1093/gigascience/gix007>.
- [45] Susannah J Salter, Michael J Cox, Elena M Turek, Szymon T Calus, William O Cookson, Miriam F Moffatt, Paul Turner, Julian Parkhill, Nicholas J Loman, and Alan W Walker. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12(1), nov 2014. doi: 10.1186/s12915-014-0087-z. URL <https://doi.org/10.1186/s12915-014-0087-z>.
- [46] Cedric C Laczny, Christina Kiefer, Valentina Galata, Tobias Fehlmann, Christina Backes, and Andreas Keller. Busybee web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Research*, page gkx348, 2017.
- [47] Brian Cleary, Ilana Lauren Brito, Katherine Huang, Dirk Gevers, Terrance Shea, Sarah Young, and Eric J Alm. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology*, 33(10):1053–1060, sep 2015. doi: 10.1038/nbt.3329. URL <https://doi.org/10.1038/nbt.3329>.