

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 October 25, 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
23 a small subset of microbial diversity [2], and much of the remaining diver-
24 sity is evolutionarily distant and reference-based search techniques may not
25 recover it [3].

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

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

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

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

54 confounders of assembly such as sequencing artifacts and GC bias.

55 In a landmark study, Shakya et al. (2013) constructed a synthetic com-
56 munity of organisms by mixing DNA isolated from individual cultures of 64
57 bacteria and archaea, including a variety of strains across a range of average
58 nucleotide distances [12]. In addition to performing 16s amplicon analy-
59 sis and doing 454 sequencing, the authors shotgun-sequenced the mixture
60 with Illumina. While the authors concluded that this metagenomic sequenc-
61 ing generally outperformed amplicon sequencing, they did not conduct an
62 assembly based analysis. This data set was also used in several other eval-
63 uation studies, including gbtools for binning [13] and benchmarking of the
64 MEGAHIT and metaSPAdes assemblers [14, 15]. Importantly, the authors
65 of the MEGAHIT benchmarking paper noted the presence of unexpected
66 sequence in this data set.

67 More recently, several benchmark studies systematically evaluated metagenome
68 assembly of short reads. The Critical Assessment of Metagenome Interpre-
69 tation (CAMI) collaboration benchmarked a number of metagenome as-
70 semblers on several simulated data sets of varying complexity, evaluating
71 recovery of novel genomes and multiple strain variants [3]. Notably, CAMI
72 concluded that “The resolution of strain-level diversity represents a substan-
73 tial challenge to all evaluated programs.” Another recent study evaluated
74 eight assemblers on nine environmental metagenomes and three simulated
75 data sets and provided a workflow for choosing a metagenome assembler
76 based on the biological goal and computational resources available [16]. [5]
77 explored metagenome assembler performance on a pair of real data sets,
78 again concluding that the biological goal and computational resources de-
79 fined the choice of assembler. Also see [17] for an analysis of a previously
80 generated HMP benchmark data set; however, the Illumina reads used for
81 this study are much shorter than current sequencing and are arguably not
82 relevant to future studies.

83 In this study, we extend previous work by delving into questions of
84 chimeric misassembly and strain recovery in the Shakya et al. (2013) data
85 set. This data set is the most complex synthetic community for which bulk
86 sequencing data is available, and has been used for several independent as-
87 sembly benchmarking efforts [13, 14, 15]. However, while previous efforts
88 have noted the presence of unexpected sequence data in the data set [14],
89 no further analysis has been done to characterize this sequence or its likely
90 origins.

91 Below, we first update the list of reference genomes for Shakya et al.
92 to include the latest GenBank assemblies along with plasmids. We then

93 compare IDBA [18], MetaSPAdes [19], and MEGAHIT [20] performance on
94 assembling this short-read data set, and explore concordance in recovery
95 between the three assemblers. We describe the effects of “strain confusion”
96 between multiple strains. We also detect and analyze several previously
97 unreported strains and genomes in the Shakya et al. data set. We find that
98 in the absence of closely related genomes, all three metagenome assemblers
99 recover 95% or more of known reference genomes. However, in the presence
100 of closely related genomes, these three metagenome assemblers vary widely
101 in their performance and, in extreme cases, can fail to recover the majority
102 of some genomes even when they are completely present in the reads. Our
103 report looks specifically at the most poorly recovered genomes, provides
104 strong guidance on choice of assemblers, and extends previous analyses of
105 this low-complexity metagenome benchmarking data set.

106 Datasets

107 We used a diverse mock community data set constructed by pooling DNA
108 from 64 species of bacteria and archaea and sequencing them with Illumina
109 HiSeq. The raw data set consisted of 109,629,496 reads from Illumina HiSeq
110 101 bp paired-end sequencing (2x101) with an untrimmed total length of
111 11.07 Gbp and an estimated fragment size of 380 bp [12].

112 The original reads are available through the NCBI Sequence Read Archive
113 at Accession SRX200676. We updated the 64 reference genomes sets from
114 NCBI GenBank using the latest available assemblies with plasmid content
115 (June 2017); the accession numbers are available as `accession-list-ref.txt`
116 in the Zenodo repository, DOI: 10.5281/zenodo.821919. For convenience, the
117 updated reference genome collection is available for download at the archival
118 URL <https://osf.io/vbhy5/>.

119 Methods

120 The analysis code and run scripts for this paper are written in Python and
121 bash, and are available at [https://github.com/dib-lab/2016-metagenome-](https://github.com/dib-lab/2016-metagenome-assembly-eval/)
122 `assembly-eval/` (archived at Zenodo DOI: 10.5281/zenodo.821919). The
123 scripts and overall pipeline were examined by the first and senior authors for
124 correctness. In addition, the bespoke reference-based analysis scripts were
125 tested by running them on a single-colony *E. coli* MG1655 data set with a
126 high quality reference genome [21].

127 **Quality Filtering**

128 We removed adapters with Trimmomatic v0.30 in paired-end mode with
129 the TruSeq adapters [22], using light quality score trimming (`LEADING:2`
130 `TRAILING:2 SLIDINGWINDOW:4:2 MINLEN:25`) as recommended in MacManes,
131 2014 [23].

132 **Reference Coverage Profile**

133 To evaluate how much of the reference metagenome was contained in the
134 read data, we used `bwa aln` (v0.7.7.r441) to map reads to the reference
135 genome [24]. We then calculated how many reference bases were covered by
136 mapped reads (custom script `coverage-profile.py`).

137 **Measuring k-mer inclusion and Jaccard similarity**

138 We used MinHashing as implemented in sourmash to estimate k-mer inclu-
139 sion and Jaccard similarity between data sets [25]. MinHash signatures were
140 prepared with `sourmash compute` using `--scaled 10000`. K-mer inclusion
141 was computed by taking the ratio of the number of intersecting hashes with
142 the query over the total number of hashes in the subject MinHash. Jac-
143 card similarity was computed as in [26] by taking the ratio of the number
144 of intersecting hashes between the query and subject over the number of
145 hashes in the union. K-mer sizes for comparison were chosen at 21, 31, or
146 51, depending on the level of taxonomic specificity desired - genus, species,
147 or strain, respectively, as described in [27].

148 Where specified, high-abundance k-mers were selected for counting by
149 using the script `trim-low-abund.py` script with `-C 5` from khmer v2 [28,
150 29].

151 **Assemblers**

152 We assembled the quality-filtered reads using three different assemblers:
153 IDBA-UD [18], MetaSPAdes [19], and MEGAHIT [20]. For IDBA-UD v1.1.3
154 [18], we used `--pre_correction` to perform pre-correction before assembly
155 and `-r` for the pe files. IDBA could not ingest orphan sequences so singleton
156 reads were omitted from this assembly.

157 For MetaSPAdes v3.10.1 [19], we used `--meta --pe1-12 --pe1-s` where
158 `--meta` is used for metagenomic data sets, `--pe1-12` specifies the interlaced

reads for the first paired-end library, and `--pe1-s` provides the orphan reads remaining from quality trimming.

For MEGAHIT v1.1.1-2-g02102e1 [20], we used `-l 101 -m 3e9 --cpu-only` where `-l` is for maximum read length, `-m` is for max memory in bytes to be used in constructing the graph, and `--cpu-only` uses only the CPU and no GPUs. We also used `--presets meta-large` for large and complex metagenomes, and `--12` and `-r` to specify the interleaved-paired-end and single-end files respectively. MEGAHIT allows the specification of a memory limit and we used `-M 1e+10` for 10 GB.

All three assemblies were executed on the same XSEDE Jetstream instance (S1.Xxlarge) at Indiana University, running Ubuntu 16.04 (install 6/21/17, Ubuntu 16.04 LTS Development + GUI support + Docker; based on Ubuntu cloud image for 16.04 LTS with basic dev tools, GUI/Xfce added). Assemblers were limited to 16 threads. We recorded RAM and CPU time for each assembly using `/usr/bin/time -v`. Install and execute details as well as output timings and logs are available in the `pipeline/runstats` directory of the Zenodo archive.

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

Mapping

We aligned all quality-filtered reads to the reference metagenome with `bwa aln` (v0.7.7.r441) [24]. We aligned paired-end and orphaned reads separately. We then used `samtools` (v0.1.19) [30] to convert SAM files to BAM files for both paired-end and orphaned reads. To count the unaligned reads, we included only those records with the “4” flag in the SAM files [30].

Assembly analysis using NUCmer

We used the NUCmer tool from MUMmer3.23 [31] to align assemblies to the reference genome with options `-coords -p`. Then we parsed the generated “coords” file using a custom script `analyze_assembly.py`, and calculated several analysis metrics across all three assemblies at a 99% alignment identity.

190 **Reference-based analysis of the assemblies**

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

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

201 **Analysis of chimeric misassemblies**

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

211 **Analysis of unmapped reads**

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

216 **Results**

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

218 The reads contain 11,072,579,096 bp (11.07 Gbp) in 109,629,496 reads with
219 101.0 average length (2x101bp Illumina HiSeq).

220 Trimming removed 686,735 reads (0.63%). After trimming, we retained

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%

221 108,422,358 paired reads containing 10.94 Gbp with an average length of
 222 100.9 bases. A total of 46.56 Mbp remained in 520,403 orphan reads with
 223 an average length of 89.5 bases. In total, the quality trimmed data contained
 224 10.98 Gbp in 108,942,761 reads. This quality trimmed (“QC”) data set was
 225 used as the basis for all further analyses.

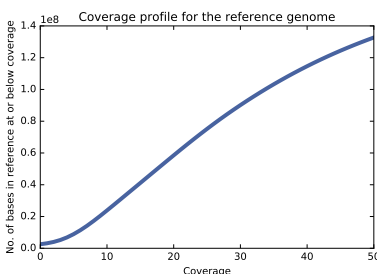


Figure 1: Cumulative coverage profile for the reference metagenome, based on read mapping.

226 **The reference metagenome is not completely present in the**
 227 **reads.**

228 We next evaluated the fraction of the reference genome covered by at least
 229 one read (see Methods for details). Quality filtered reads cover 203,058,414
 230 (98.76%) bases of the reference metagenome (205,603,715 bp total size). Fig-
 231 ure 1 shows the cumulative coverage profile of the reference metagenome,
 232 and the percentage of bases with that coverage. Most of the reference
 233 metagenome was covered at least minimally; only 3.33% of the reference
 234 metagenome had mapping coverage <5, and 1.24% of the bases in the ref-
 235 erence were not covered by any reads in the QC data set.

236 In order to evaluate reconstructability with De Bruijn graph assemblers,
 237 we next examined k-mer containment of the reference in the reads for k of

238 21, 31, 41, and 51 (Table 1). The k-mer overlap decreases from 96.8% to
 239 94.1% as the k-mer size increases. This could be caused by low coverage of
 240 some portions of the reference and/or variation between the reads and the
 241 reference.

242 **Some individual reference genomes are poorly represented in**
 243 **the reads.**

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%

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

250 We next did a 51-mer containment analysis of each reference genome in
 251 the reads; k=51 was chosen so as to be specific to strain content [27]. 99%
 252 or more of the constituent 51-mers for 51 of the 64 reference genomes were
 253 present in the reads, suggesting that each of the 51 genomes was entirely
 254 present at some minimal coverage.

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

259 **MEGAHIT is the fastest and lowest-memory assembler eval-**
 260 **uated**

261 We ran three commonly used metagenome assemblers on the QC data set:
 262 IDBA-UD, MetaSPAdes, and MEGAHIT. We recorded the time and mem-
 263 ory usage of each (Table 4). In computational requirements, MEGAHIT
 264 outperformed both MetaSPAdes and IDBA-UD, , producing an assembly in

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

Table 4: Running Time and Memory Utilization

Assembler	CPU time	Wall time	RAM (Max RSS)
MEGAHIT	1191m	1h 33m	10 GB
IDBA-UD	1904m	2h 27m	17 GB
MetaSPAdes	2554m	4h 7m	28 GB

265 1.5 hours (“wall time”) – 1.6 times faster than IDBA and 2.6 times faster
 266 than MetaSPAdes. MEGAHIT used only 10 GB of RAM as requested –
 267 about 60% of the memory used by IDBA and a third of the memory used by
 268 MetaSPAdes. CPU time measurements (which include processing on multi-
 269 ple CPU cores) show that all three assemblers use multiple cores effectively.

270 **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%

271 We assessed read inclusion in assemblies by mapping the QC reads to
 272 the length-filtered assemblies and counting the remaining unmapped reads.
 273 Depending on the assembly, between 2.7 million and 3.9 million reads (2.5-
 274 3.5%) did not map to the assemblies (Table 5). All of the assemblies included
 275 the large majority of high-abundance 51-mers (more than 96.8% in all cases).

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

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%

277 We next evaluated the extent to which the assembled contigs recovered
 278 the “known/true” metagenome sequence by aligning each assembly to the
 279 adjusted reference (Table 6). Each of the three assemblers generates contigs
 280 that cover more than 93.1% of the reference metagenome at high identity
 281 (99%) with little duplication (approximately 1%). All three assemblies con-
 282 tain between 96.2% and 97.2% of the 51-mers in the reference.

283 At 99% identity with the loose mapping approach, approximately 2.5% of
 284 the reference is missed by all three assemblers, while 1.7% is uniquely covered
 285 by MEGAHIT, 0.74% is uniquely covered by MetaSPAdes, and 0.64% is
 286 uniquely covered by IDBA.

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

Table 7: Contig accuracy measured by reference coverage with strict alignment.

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

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

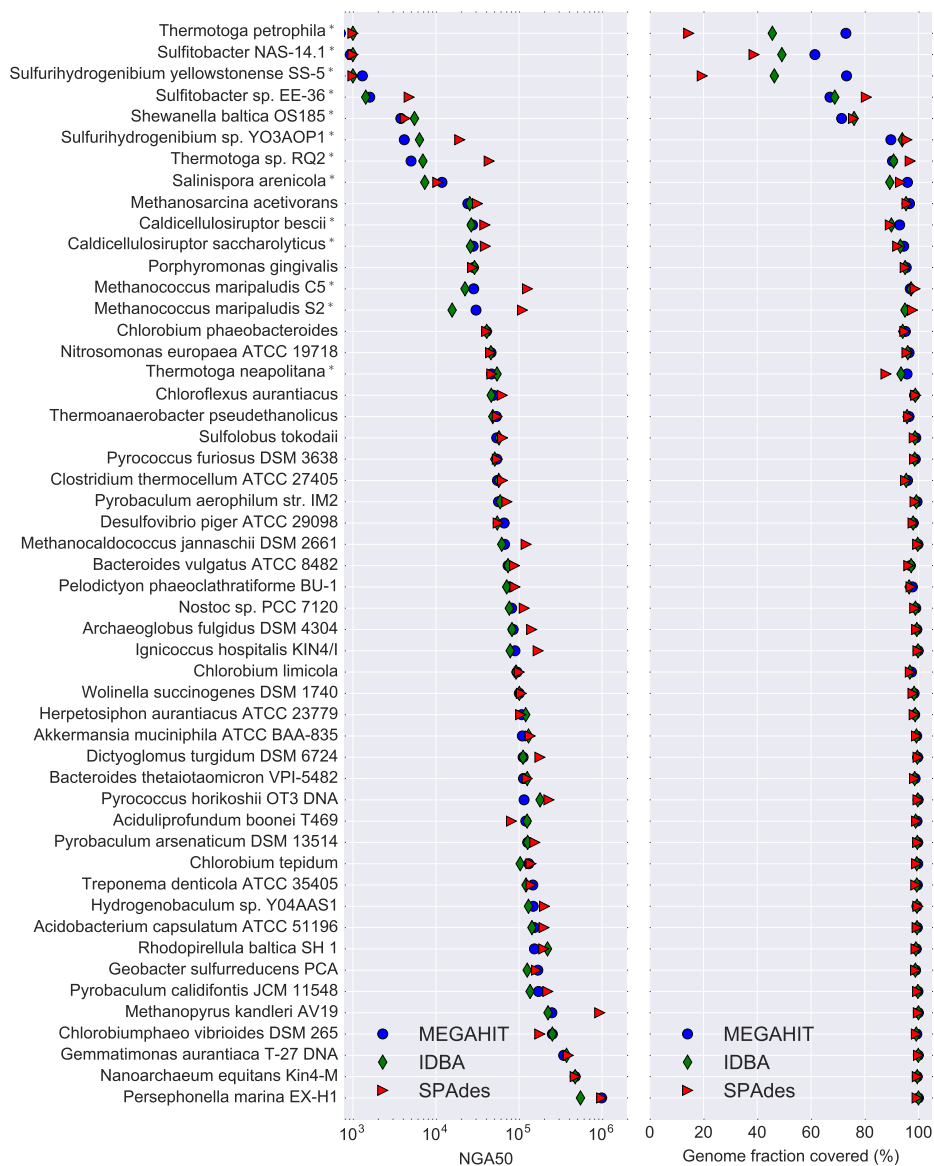


Figure 2: NGA50 and genome fraction covered, by genome and assembler. A '*' after the name indicates the presence of at least one other genome with > 2% Jaccard similarity at k=31 in the community. Where NGA50 cannot be calculated due to poor coverage, a marker is placed at 1kb.

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

293 We computed the NGA50 for each individual genome and assembly in order
 294 to compare assembler performance on genome recovery (see left panel of Fig-
 295 ure 2). The NGA50 statistics for individual genomes vary widely, but there
 296 are consistent assembler-specific trends: IDBA yields the lowest NGA50 for
 297 28 of the 51 genomes, while MetaSPAdes yields the highest NGA50 for 32
 298 of the 51 genomes.

299 We also evaluated aligned coverage per genome for each of the three
 300 assemblies (right panel, Figure 2). We found that 13 of the 51 genomes
 301 were missing 5% or more of bases in at least one assembly, despite all 51
 302 genomes having 99% or higher read- and 51-mer coverage. While some of
 303 these missing bases may be in the assembled contigs that are less than 500
 304 bp in length, contigs shorter than 500 bp are unlikely to contain more than
 305 half of a typical bacterial gene [32].

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

311 **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%)

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

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>

320 The unmapped reads contain strain variants of reference genomes.

321 Approximately 4.8 million reads (4.4%) from the QC data set did not map
322 anywhere in the reference provided by the authors of [12]. We extracted
323 and assembled these reads in isolation using MEGAHIT, yielding 6.5 Mbp
324 of assembly in 1711 contigs > 500bp in length. We then did a k-mer in-
325 clusion analysis of this assembly against all of the GenBank genomes at
326 k=31, and estimated the fraction of the k-mers that belonged to different
327 species (Table 9). We find that 51.1% of the k-mer content of these contigs
328 positively match to a genome present in GenBank but not in the reference
329 metagenome.

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

336 We note that all but the two *P. ruminis* matches and the *E. coli* isolate
337 YS are strain variants of species that are part of the defined community
338 but are not completely present in the reads (see Table 2). For *Proteiniclas-*
339 *ticum ruminis*, there is no closely related species in the mock community
340 design, and very little of the MEGAHIT assembly aligns to known *P. ru-*
341 *minis* genomes at 99%. However, there are many alignments to *P. ruminis*
342 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 *Proteini-
clasticum*; 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.

Discussion

Assembly recovers basic content sensitively and accurately.

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

The presence of multiple closely related genomes confounds assembly.

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

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

Another interest of this study was to examine cross-species chimeric as-

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

MEGAHIT performs best by several metrics.

MEGAHIT is clearly the most efficient computationally, outperforming both MetaSPAdes and IDBA in memory and time (Table 4). The MEGAHIT assembly also included more of the reads than either IDBA or MetaSPAdes, and omitted only 0.4% more of the unique 51-mers from the reads than IDBA. MEGAHIT covered more of the reference genome with both loose and strict alignments (Table 6 and Table 7), with little duplication. This is clearly because of MEGAHIT’s generally superior performance in recovering the genomes of closely related strains (Figure 2, right panel). The sum “fraction of genome recovered” is arguably the most important measure of a metagenome assembler (see [5] in particular) and here MEGAHIT excels for individual genomes even in the presence of strain variation.

In general other studies have found that MEGAHIT excels in recovery of sequence through assembly [3, 17] and is considerably more computationally efficient than most other assemblers [3, 16]. However, studies have also shown that MEGAHIT produces more misassemblies than other assemblers [3] and performs poorly on high coverage portions of the data set [5]. Thus while we can recommend MEGAHIT as a good first assembler, we can also not unambiguously recommend it as the only assembler to use.

When comparing details of sequence recovery between the assemblers, the assembly content differs by only a small amount when loose alignments are allowed: all three assemblers miss more content (approximately 2.5% of the reference) than they generate uniquely (1.7% or less). In addition to preferring no one assembler over any other, this suggests that combining assemblies may have little value in terms of recovering additional metagenome content. The genome alignment statistics in Figure 2 suggest that much of this differential assembly content is due to the impact of strains.

410 **The missing reference may be present in strain variants of the**
411 **intended species.**

412 Several individual genomes are missing in measurable portion from the QC
413 reads (Table 2), and many QC reads (4.4% of 108m) did not map to the full
414 reference metagenome. These appear to be related issues: upon analysis of
415 the unmapped reads against GenBank, we find that many of the contigs as-
416 sembled from the unmapped reads can be assigned to strain variants of the
417 species in the mock community (Table 9) and align closely to the identified
418 genomes. This suggests that the constructors of the mock community may
419 have unintentionally included strain variants of *Fusobacterium nucleatum*,
420 *Thermus thermophilus* HB27, and *Enterococcus faecalis*; note that the mi-
421 crobes used were sourced from the community rather than the ATCC (M.
422 Podar, pers. communication). In addition, we detect what may be por-
423 tions of a novel member of the *Proteiniclasticum* genus in the assembly of
424 these reads - this is likely the *Clostridium* spp. detected through amplicon
425 sequencing in [12].

426 Without returning to the original DNA samples, it is impossible to con-
427 clusively confirm that unintended strains were used in the construction of the
428 mock community. In particular, our analysis is dependent on the genomes in
429 GenBank: the genomes we detect in the contigs are clearly closely related to
430 GenBank genomes not in the reference metagenome, based on k-mer anal-
431 ysis and contig alignment. However, GenBank is unlikely to contain the
432 exact genomes of the actually included strain variants, rendering conclusive
433 identification impossible.

434 **Omissions in this study: binning and parameter sweeps**

435 We omitted two important questions in this study: binning and choice of
436 parameters. We chose not to evaluate genome binning because most bin-
437 ning strategies either operate post-assembly (see e.g. [33]), in which case
438 the challenges with assembly discussed above will apply; or require multi-
439 ple samples (e.g. [34]), which we do not have. We also chose to use only
440 default parameters with all three assemblers, for two reasons. First, we
441 are not aware of any effective automated approaches for determining the
442 “best” set of parameters or evaluating the output for metagenome assem-
443 blers, other than those integrated into the assemblers themselves (e.g. the
444 choice of k-mer sizes by MEGAHIT and MetaSPAdes), and absent such
445 guidance we do not feel comfortable blessing any particular set of param-
446 eters; here the choice of default parameters is parsimonious (and also see [35])

447 for the dangers of poorly chosen objective functions). Second, any param-
448 eter exploration pipeline would not only need to be automated but would
449 need to run multiple assemblies, whose time and resource usage should be
450 measured; in this case, any comparison based on runtime of the parameter
451 choice pipeline should naturally favor MEGAHIT because of its advantage
452 in computational efficiency.

453 Conclusions

454 Overall, assembly of this mock community works well, with good recovery
455 of known genomic sequence for the majority of genomes. All three assem-
456 blers that we evaluated recover similar amounts of most genomic sequence,
457 but (recapitulating several other studies [3, 5, 16]) MEGAHIT is compu-
458 tationally the most efficient of the three. We note that assembly resolves
459 substantial portions of several previously undetected strain variants, as well
460 as recovering a substantial portion of a novel *Proteiniclasticum* spp. that
461 was detected via amplicon analysis in [12], suggesting that assembly is a
462 useful complement to amplicon or reference-based analyses.

463 The presence of closely related strains is a major confounder of metagenome
464 assembly, and causes assemblers to drop considerable portions of genomes
465 that (based on read mapping and k-mer inclusion) are clearly present. In this
466 relatively simple community, this strain confusion is present but does not
467 dominate the assembly. However, real microbial communities are likely to
468 have many closely related strains and any resulting loss of assembly would
469 be hard to detect in the absence of good reference genomes. While high
470 polymorphism rates in e.g. animal genomes are known to cause duplication
471 or loss of assembly, some solutions have emerged that make use of assump-
472 tions of uniform coverage and diploidy [36]. These solutions cannot however
473 be transferred directly to metagenomes, which have unknown abundance
474 distributions and strain content.

475 An additional concern is that metagenome assemblies are often per-
476 formed after pooling data sets to increase coverage (e.g. [4, 37]); this pooled
477 data is more likely to contain multiple strains, which would then in turn
478 adversely affect assembly of strains. This may not be resolvable within the
479 current paradigm of assembly, which focuses on outputting linear assem-
480 blies that cannot properly represent strain variation. The human genomics
481 community is moving towards using *reference graphs*, which can represent
482 multiple incompatible variants in a single data structure [38]; this approach,
483 however, requires high-quality isolate reference genomes, which are generally

484 unavailable for environmental microbes.

485 Long read sequencing (and related technologies) will undoubtedly help
486 resolve strain variation in the future, but even with highly accurate long-
487 read sequencing, current sequencing depth is still too low to resolve deep
488 environmental metagenomes [39, 40]. It is unclear how well long error-prone
489 reads (such as those output by Pacific Biosciences SMRT [41] and Oxford
490 Nanopore instruments [42]) will perform on complex metagenomes: with
491 high error rates, deep coverage of each individual genome is required to
492 achieve accurate assembly, and this may not be easily obtainable for com-
493 plex communities. Single-molecule barcoding (e.g. 10X Genomics [43]),
494 HiC approaches [44], and cell-sorting cells into “minimetagenomes” [45]
495 show promise but these remain untested on well-defined complex commu-
496 nities and are still challenged by the complexity of complex environmental
497 metagenomes; see [46, 47, 48].

498 Much of our analysis above depends on having a high-quality “mock”
499 metagenome. While computationally constructed synthetic communities
500 and computational “spike-ins” to real data sets can provide valuable controls
501 (e.g. see [16] and [49]) we strongly believe that standardized communities
502 constructed *in vitro* and sequenced with the latest technologies are critical
503 to the evaluation of both canonical and emerging tools, e.g. efforts such as
504 [50]. From the perspective of tool evaluation, we disagree somewhat with
505 Vollmers et al. [5]: good metagenome tool evaluation necessarily depends on
506 mock communities that are as realistic as we can make them. Likewise, from
507 the perspective of bench biologists, actually sequencing real DNA is critical
508 because it can evaluate confounding effects such as kit contamination [51].
509 Large-scale studies of computational approaches systematically applied to
510 mock communities such as CAMI [3] can then provide fair comparisons of
511 entire toolchains (wet and dry combined).

512 **Author contributions**

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

515 **Competing interests**

516 No competing interest to our knowledge.

517 Grant information

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

520 Acknowledgments

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

525 References

- 526 [1] Jay Ghurye, Victoria Cepeda-Espinoza, and Mihai Pop. Metagenomic assem-
527 bly: Overview, challenges and applications. *The Yale Journal of Biology and*
528 *Medicine*, 89(3):353–362, 2016.
- 529 [2] Nikos C. Kyrpides, Philip Hugenholtz, Jonathan A. Eisen, Tanja Woyke,
530 Markus Göker, Charles T. Parker, Rudolf Amann, Brian J. Beck, Patrick S. G.
531 Chain, Jongsik Chun, Rita R. Colwell, Antoine Danchin, Peter Dawyndt, Tom
532 Dedeurwaerdere, Edward F. DeLong, John C. Detter, Paul De Vos, Timothy J.
533 Donohue, Xiu-Zhu Dong, Dusko S. Ehrlich, Claire Fraser, Richard Gibbs, Jack
534 Gilbert, Paul Gilna, Frank Oliver Glöckner, Janet K. Jansson, Jay D. Keasling,
535 Rob Knight, David Labeda, Alla Lapidus, Jung-Sook Lee, Wen-Jun Li, Juncai
536 MA, Victor Markowitz, Edward R. B. Moore, Mark Morrison, Folker Meyer,
537 Karen E. Nelson, Moriya Ohkuma, Christos A. Ouzounis, Norman Pace, Julian
538 Parkhill, Nan Qin, Ramon Rossello-Mora, Johannes Sikorski, David Smith,
539 Mitch Sogin, Rick Stevens, Uli Stengl, Ken ichiro Suzuki, Dorothea Taylor,
540 Jim M. Tiedje, Brian Tindall, Michael Wagner, George Weinstock, Jean Weis-
541 senbach, Owen White, Jun Wang, Lixin Zhang, Yu-Guang Zhou, Dawn Field,
542 William B. Whitman, George M. Garrity, and Hans-Peter Klenk. Genomic
543 encyclopedia of bacteria and archaea: Sequencing a myriad of type strains.
544 *PLoS Biology*, 12(8):e1001920, aug 2014. doi: 10.1371/journal.pbio.1001920.
545 URL <https://doi.org/10.1371/journal.pbio.1001920>.
- 546 [3] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan
547 Janssen, Johannes Droege, Ivan Gregor, Stephan Majda, Jessika Fiedler,
548 Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue
549 Sparholt Jorgensen, Nicole Shapiro, Philip D Blood, Alexey Gurevich, Yang
550 Bai, Dmitrij Turaev, Matthew Z DeMaere, Rayan Chikhi, Niranjan Nagara-
551 jan, Christopher Quince, Lars Hestbjerg Hansen, Soren J Sorensen, Burton
552 K H Chia, Bertrand Denis, Jeff L Froula, Zhong Wang, Robert Egan, Dong-
553 wan Don Kang, Jeffrey J Cook, Charles Deltel, Michael Beckstette, Claire

554 Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu-Wei
 555 Wu, Steven W Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter
 556 Meinicke, Michael Barton, Thomas Lingner, Hsin-Hung Lin, Yu-Chieh Liao,
 557 Genivaldo Gueiros Z. Silva, Daniel A Cuevas, Robert A Edwards, Surya Saha,
 558 Vitor C Piro, Bernhard Y Renard, Mihai Pop, Hans-Peter Klenk, Markus
 559 Goeker, Nikos Kyrpides, Tanja Woyke, Julia A Vorholt, Paul Schulze-Lefert,
 560 Edward M Rubin, Aaron E Darling, Thomas Rattei, and Alice C McHardy.
 561 Critical assessment of metagenome interpretation - a benchmark of compu-
 562 tational metagenomics software. *bioRxiv*, 2017. doi: 10.1101/099127. URL
 563 <http://biorxiv.org/content/early/2017/01/09/099127>.

564 [4] I. Sharon, M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman,
 565 and J. F. Banfield. Time series community genomics analysis reveals rapid
 566 shifts in bacterial species, strains, and phage during infant gut colonization.
 567 *Genome Research*, 23(1):111–120, aug 2012. doi: 10.1101/gr.142315.112. URL
 568 <https://doi.org/10.1101/gr.142315.112>.

569 [5] John Vollmers, Sandra Wiegand, and Anne-Kristin Kaster. Compar-
 570 ing and evaluating metagenome assembly tools from a microbiol-
 571 ogist’s perspective - not only size matters! *PLOS ONE*, 12
 572 (1):e0169662, jan 2017. doi: 10.1371/journal.pone.0169662. URL
 573 <https://doi.org/10.1371/journal.pone.0169662>.

574 [6] Jorge F Vázquez-Castellanos, Rodrigo García-López, Vicente Pérez-Brocal,
 575 Miguel Pignatelli, and Andrés Moya. Comparison of different assembly and
 576 annotation tools on analysis of simulated viral metagenomic communities in
 577 the gut. *BMC genomics*, 15(1):1, 2014.

578 [7] Konstantinos Mavromatis, Natalia Ivanova, Kerrie Barry, Harris Shapiro, Eu-
 579 gene Goltsman, Alice C McHardy, Isidore Rigoutsos, Asaf Salamov, Frank
 580 Korzeniewski, Miriam Land, et al. Use of simulated data sets to evaluate the
 581 fidelity of metagenomic processing methods. *Nature methods*, 4(6):495–500,
 582 2007.

583 [8] David B Jaffe, Jonathan Butler, Sante Gnerre, Evan Mauceli, Kerstin
 584 Lindblad-Toh, Jill P Mesirov, Michael C Zody, and Eric S Lander. Whole-
 585 genome sequence assembly for mammalian genomes: Arachne 2. *Genome*
 586 *research*, 13(1):91–96, 2003.

587 [9] Samuel Aparicio, Jarrod Chapman, Elia Stupka, Nik Putnam, Jer-ming Chia,
 588 Paramvir Dehal, Alan Christoffels, Sam Rash, Shawn Hoon, Arian Smit, et al.
 589 Whole-genome shotgun assembly and analysis of the genome of *fugu rubripes*.
 590 *Science*, 297(5585):1301–1310, 2002.

591 [10] Anveshi Charuvaka and Huzefa Rangwala. Evaluation of short read metage-
 592 nomic assembly. *BMC genomics*, 12(2):1, 2011.

- 593 [11] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein,
594 Steven JM Jones, and Inanç Birol. Abyss: a parallel assembler for short read
595 sequence data. *Genome research*, 19(6):1117–1123, 2009.
- 596 [12] Shakya Migun, Christopher Quince, James Campbell, Zamin Yang, Christo-
597 pher Schadt, and Mircea Podar. Comparative metagenomic and rrna microbial
598 diversity characterization using archaeal and bacterial synthetic communities.
599 *Enivromental Microbiology*, 15(6):1882–1899, 2013.
- 600 [13] Brandon K. B. Seah and Harald R. Gruber-Vodicka. gbtools: In-
601 teractive visualization of metagenome bins in r. *Frontiers in Mi-*
602 *crobiology*, 6, dec 2015. doi: 10.3389/fmicb.2015.01451. URL
603 <https://doi.org/10.3389/fmicb.2015.01451>.
- 604 [14] Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-
605 Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah
606 Lam. MEGAHIT v1.0: A fast and scalable metagenome assembler
607 driven by advanced methodologies and community practices. *Meth-*
608 *ods*, 102:3–11, jun 2016. doi: 10.1016/j.ymeth.2016.02.020. URL
609 <https://doi.org/10.1016/j.ymeth.2016.02.020>.
- 610 [15] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner.
611 metaspades: a new versatile metagenomic assembler. *Genome Research*, 27
612 (5):824–834, 2017.
- 613 [16] Andries Johannes van der Walt, Marc Warwick Van Goethem,
614 Jean-Baptiste Ramond, Thulani Peter Makhanyane, Oleg Reva,
615 and Don Arthur Cowan. Assembling metagenomes, one com-
616 munity at a time. *bioRxiv*, 2017. doi: 10.1101/120154. URL
617 <http://biorxiv.org/content/early/2017/06/06/120154>.
- 618 [17] William W. Greenwald, Niels Klitgord, Victor Seguritan, Shibu Yooseph,
619 J. Craig Venter, Chad Garner, Karen E. Nelson, and Weizhong Li. Utilization
620 of defined microbial communities enables effective evaluation of meta-genomic
621 assemblies. *BMC Genomics*, 18(1), apr 2017. doi: 10.1186/s12864-017-3679-5.
622 URL <https://doi.org/10.1186/s12864-017-3679-5>.
- 623 [18] Yu Peng, Henry C.M. Leung, S.M. Yiu, and Francis Y.L. Chin. Idba-ud: a de
624 novo assembler for single-cell and metagenomic sequencing data with highly
625 uneven depth. *Bioinformatics*, 28:1420–1428, 2012.
- 626 [19] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner.
627 metaSPAdes: a new versatile metagenomic assembler. *Genome Re-*
628 *search*, 27(5):824–834, mar 2017. doi: 10.1101/gr.213959.116. URL
629 <https://doi.org/10.1101/gr.213959.116>.
- 630 [20] Dinghua Li, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting,
631 Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. Megahit v1. 0:

- 632 A fast and scalable metagenome assembler driven by advanced methodologies
633 and community practices. *Methods*, 102:3–11, 2016.
- 634 [21] H Chitsaz, JL Yee-Greenbaum, G Tesler, MJ Lombardo, CL Dupont, JH Bad-
635 ger, M Novotny, DB Rusch, LJ Fraser, NA Gormley, O Schulz-Trieglaff,
636 GP Smith, DJ Evers, PA Pevzner, and RS Lasken. Efficient de novo assembly
637 of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol*, 29
638 (10):915–21, 2011.
- 639 [22] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible
640 trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.
- 641 [23] Matthew D MacManes. On the optimal trimming of high-throughput mrna
642 sequence data. *Frontiers in genetics*, 5:13, 2014.
- 643 [24] Heng Li and Richard Durbin. Fast and accurate short read alignment with
644 burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.
- 645 [25] C. Titus Brown and Luiz Irber. sourmash: a library for MinHash sketch-
646 ing of DNA. *The Journal of Open Source Software*, 1(5), sep 2016. doi:
647 10.21105/joss.00027. URL <https://doi.org/10.21105/joss.00027>.
- 648 [26] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mal-
649 lonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy.
650 Mash: fast genome and metagenome distance estimation using MinHash.
651 *Genome Biology*, 17(1), jun 2016. doi: 10.1186/s13059-016-0997-x. URL
652 <https://doi.org/10.1186/s13059-016-0997-x>.
- 653 [27] David Koslicki and Daniel Falush. Metapalette: a k-mer painting approach
654 for metagenomic taxonomic profiling and quantification of novel strain vari-
655 ation. *mSystems*, 1(3), 2016. doi: 10.1128/mSystems.00020-16. URL
656 <http://msystems.asm.org/content/1/3/e00020-16>.
- 657 [28] Zhang Qingpeng, Awad Sherine, and Brown C. Titus. Crossing the streams:
658 a framework for streaming analysis of short dna sequencing reads. *PeerJ*
659 *PrePrints* 3:e1100 <https://dx.doi.org/10.7287/peerj.preprints.890v1>, 2015.
- 660 [29] MR Crusoe, HF Alameldin, S Awad, E Boucher, A Caldwell, R Cartwright,
661 A Charbonneau, B Constantinides, G Edverson, S Fay, J Fenton, T Fenzl,
662 J Fish, L Garcia-Gutierrez, P Garland, J Gluck, I Gonzlez, S Guermond,
663 J Guo, A Gupta, JR Herr, A Howe, A Hyer, A Hrpfer, L Irber, R Kidd,
664 D Lin, J Lippi, T Mansour, P McA’Nulty, E McDonald, J Mizzi, KD Mur-
665 ray, JR Nahum, K Nanlohy, AJ Nederbragt, H Ortiz-Zuazaga, J Ory, J Pell,
666 C Pepe-Ranne, ZN Russ, E Schwarz, C Scott, J Seaman, S Sievert, J Simp-
667 son, CT Skennerton, J Spencer, R Srinivasan, D Standage, JA Stapleton,
668 SR Steinman, J Stein, B Taylor, W Trimble, HL Wiencko, M Wright,

- 669 B Wyss, Q Zhang, e zyme, and CT Brown. The khmer software pack-
 670 age: enabling efficient nucleotide sequence analysis [version 1; referees: 2 ap-
 671 proved, 1 approved with reservations]. *F1000Research*, 4(900), 2015. doi:
 672 10.12688/f1000research.6924.1.
- 673 [30] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer,
 674 Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence align-
 675 ment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- 676 [31] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin
 677 Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open soft-
 678 ware for comparing large genomes. *Genome biology*, 5(2):1, 2004.
- 679 [32] Lin Xu, Hong Chen, Xiaohua Hu, Rongmei Zhang, Ze Zhang, and ZW Luo.
 680 Average gene length is highly conserved in prokaryotes and eukaryotes and
 681 diverges only between the two kingdoms. *Molecular biology and evolution*, 23
 682 (6):1107–1108, 2006.
- 683 [33] Cedric C Laczny, Christina Kiefer, Valentina Galata, Tobias Fehlmann,
 684 Christina Backes, and Andreas Keller. Busybee web: metagenomic data anal-
 685 ysis by bootstrapped supervised binning and annotation. *Nucleic Acids Re-*
 686 *search*, page gkx348, 2017.
- 687 [34] Brian Cleary, Ilana Lauren Brito, Katherine Huang, Dirk Gevers, Terrance
 688 Shea, Sarah Young, and Eric J Alm. Detection of low-abundance bacte-
 689 rial strains in metagenomic datasets by eigengenome partitioning. *Nature*
 690 *Biotechnology*, 33(10):1053–1060, sep 2015. doi: 10.1038/nbt.3329. URL
 691 <https://doi.org/10.1038/nbt.3329>.
- 692 [35] Bastian Greshake, Simonida Zehr, Francesco Dal Grande, Anjuli Meiser,
 693 Imke Schmitt, and Ingo Ebersberger. Potential and pitfalls of eukaryotic
 694 metagenome skimming: a test case for lichens. *Molecular Ecology Re-*
 695 *sources*, 16(2):511–523, sep 2015. doi: 10.1111/1755-0998.12463. URL
 696 <https://doi.org/10.1111/1755-0998.12463>.
- 697 [36] J. H. Kim, M. S. Waterman, and L. M. Li. Diploid genome reconstruc-
 698 tion of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*.
 699 *Genome Research*, 17(7):1101–1110, jun 2007. doi: 10.1101/gr.5894107. URL
 700 <https://doi.org/10.1101/gr.5894107>.
- 701 [37] Ping Hu, Lauren Tom, Andrea Singh, Brian C. Thomas, Brett J. Baker,
 702 Yvette M. Piceno, Gary L. Andersen, and Jillian F. Banfield. Genome-
 703 resolved metagenomic analysis reveals roles for candidate phyla and other
 704 microbial community members in biogeochemical transformations in oil reser-
 705 voirs. *mBio*, 7(1):e01669–15, jan 2016. doi: 10.1128/mbio.01669-15. URL
 706 <https://doi.org/10.1128/mbio.01669-15>.

- [38] 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.
- [39] 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>.
- [40] 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>.
- [41] 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>.
- [42] 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>.
- [43] 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>.
- [44] 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>.

[45] Peifeng Ji, Yanming Zhang, Jinfeng Wang, and Fangqing Zhao. Metasort untangles metagenome assembly by reducing microbial community complexity. *Nature communications*, 8:14306, 2017.

[46] 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>.

[47] 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 organization in microorganisms. *eLife*, 3, dec 2014. doi: 10.7554/elife.03318. URL <https://doi.org/10.7554/elife.03318>.

[48] Christopher W. Beitel, Lutz Froenicke, Jenna M. Lang, Ian F. Korf, Richard W. Michelmore, 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>.

[49] 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.

[50] 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>.

[51] 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>.