Benchmarking of shotgun sequencing depth highlights strain-level limitations of metagenomic analysis

Nicole S. Treichel,[1,*] Thomas C. A. Hitch,[1] John Penders,[2,3] Thomas Clavel[1,3,*]

1. Functional Microbiome Research Group, Institute of Medical Microbiology, University Hospital of RWTH Aachen, Pauwelsstrasse 30, Aachen, Germany.

2. Department of Medical Microbiology, Infectious Diseases and Infection Prevention, NUTRIM Institute for Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre+, Universiteitssingel 40, Maastricht, The Netherlands.

3. Euregional Microbiome Center, P. Debyelaan 25, 6229 HX Maastricht, The Netherlands

*** Correspondence:** ntreichel@ukaachen.de; tclavel@ukaachen.de

Key words: shotgun sequencing, shallow metagenomics, complex mock communities, benchmarking, strain diversity, metagenome-assembled genomes (MAGs)

Abstract

Shallow metagenomics promises taxonomic and functional insights into samples at an affordable price. To determine the depth of sequencing required for specific analysis, benchmarking is required using defined communities. We used complex mixtures of DNA from cultured gut bacteria and analysed taxonomic composition, strain-level resolution, and functional profiles at nine sequencing depths (0.1-10.0 Gb). Reference-based analysis provided accurate taxonomic, and strain-level insights at 0.5-1.0 Gb. In contrast, *de-novo* metagenome-assembled genome (MAG) reconstruction required deep sequencing (>10 Gb). Furthermore, it was found that even high-quality MAGs were chimeric, with few (54.5 %) accurately representing the original strains. Functionally, 2 Gb provided reliable insights at the pathway level, but sufficient proteome coverage was only achieved at 10 Gb. Library preparation and host-DNA were identified as confounders in shallow metagenomic analysis. This comprehensive analysis using complex mock communities provides guidance to an increasing community of scientists interested in using shallow metagenomics, and highlights the limitations of MAGs in capturing strain-level diversity.

Introduction

Next-generation sequencing is essential for microbiome research, enabling comprehensive analysis of complex microbial communities at an ever-increasing sample size. However, their widespread use contrasts the low number of studies that tested their robustness and accuracy.

The most frequent used method for microbiota analysis is 16S rRNA gene amplicon sequencing due to its ease of implementation, both experimentally and bioinformatically, and low cost compared to metagenomics. However, it is limited to structural analysis of the microbiota (diversity and composition) and taxonomic resolution is limited to the genus level (strain analysis may be possible for certain taxonomic groups when using amplicon sequence variants). Experimentally it is also prone to amplification bias[1]. Whilst deep metagenomics offers a solution to these shortcomings, it is challenging to perform with low biomass samples or those with a high fraction of non-microbial DNA (e.g., human tissue samples) and is also more cost intensive.

Shallow metagenomics, defined as shotgun sequencing at a depth ≤1 Gb (ca. 3 million reads), has been proposed to overcome the limitations of both, high-throughput 16S rRNA gene amplicon and deep metagenomic sequencing[2–4]. Few studies have attempted to benchmark shallow metagenomics and thereby define which metagenomic information is accessible at what sequencing depth, and also how reliably it can be obtained. A common approach used by most benchmarking studies is to compare sequencing techniques and depths using native samples, such as human stool[3–7]. The main disadvantage of this is that the original composition of the input is unknown, preventing accurate benchmarking. Even fewer studies investigated the impact of sequencing depth using samples with defined microbial composition, hereon referred to as Mock communities[2,8]. Cattonaro et al. used a commercially available Mock community, containing 20 strains, of which the abundance of 5 was below their detection threshold[8]. Xu et al. used Mock communities with two different distributions, including up to 69 human gut bacterial species[2]. However, their analysis relied on a limited set of sequencing depths and focused on comparing shallow metagenomics with 16S rRNA gene amplicon sequencing. Other common practices are to bioinformatically subsample deeply sequenced datasets to mimic different sequencing depths or to generate artificial Mock communities in-silico. However, these approaches do not reflect the impact of real sample processing and sequencing in the laboratory.

Given the urgent need to better define the strengths and limitations of shallow metagenomics, we systematically evaluated the effects of shotgun sequencing depth on the analysis resolution of composition, strain-level diversity, and functional readouts (**Supplementary Fig. S1**). For this, we used two complex mixtures of bacterial DNA: one with an even distribution of DNA from 70 bacterial strains (Mock-even) and one with a staggered distribution of DNA from 24 strains (Mock-stag)

64 (**Supplementary Fig. S2**). As wet-lab factors can cofound results, DNA libraries were prepared in two
65 laboratories and in the presence/absence of background DNA isolated from the gut of germfree mice.

66

67 ## Methods

68 ### DNA extraction and preparation of the Mock communities

69 For creation of the two mock communities (Mock-even and Mock-stag), DNA was obtained from
70 isolates in our collections of mouse and human gut bacteria [9,10]. DNA was extracted from freshly grown
71 strains revived from frozen glycerol stocks based on the method of Godon et al. 1997 [11] modified as
72 described in Afrizal et al [9]. DNA concentration was measured using a Qubit fluorometer (Thermo Fisher
73 Scientist). The DNA was pooled in equimolar amounts for all 70 isolates or with a staggered distribution
74 for 24 selected isolates (**Supplementary Figure S2**).

75 Non-bacterial background DNA was isolated, as described above, using gut content collected from
76 germfree mice (LANUV ethical approval 81-02.04.2023.A253) in accordance with the German Animal
77 Protection Law (TierSchG). All mice were fed a standard chow *ad libitum* (γ-irradiated ssniff
78 Spezialdiäten ref. V1124-927). The germfree status of the mice was confirmed by microscopic
79 observation after Gram-staining and by cultivation on both, anaerobic and aerobic agar plates. Mice
80 were culled and caecal content was collected and stored immediately at -80°C. After DNA extraction,
81 the background DNA was mixed 1:1 (v/v) with each of the two mock communities.

82

83 ### Library preparation and sequencing

84 In sequencing facility 1 (UKA), DNA libraries of the Mock communities were prepared using the
85 NEBNext Ultra II FS DNA Library Prep Kit for Illumina (NEB, USA) according to the manufacturer's
86 instructions, using 100 ng of DNA and an automated platform (Beckman Coulter, USA). Enzymatic
87 shearing to approximate 250 bp was performed for 30 minutes. Adaptor-ligated DNA was enriched
88 using PCR (5 cycles) and NEBNext Multiplex Oligos for Illumina (NEB, USA) for unique dual barcoding.
89 AMPure beads (Beckman Coulter, USA) were used for size selection and clean-up of adaptor ligated
90 DNA.

91 Sequencing facility 2 (UMC) used the Nextera XT DNA Library preparation kit (Illumina, USA) with
92 Nextera XT indexes (Illumina, USA) and 1 ng template DNA according to manufacturer's protocol.
93 Twelve cycles of PCR were used for indexing. AMPure beads (Beckman Coulter, USA) were used for
94 double-sided selection and clean-up of adaptor ligated DNA. Fragment size of cleaned DNA was

determined on an Agilent D1000 Tapestation (Bioanalyzer System, Agilent Technologies, USA) using High Sensitivity D1000 screentapes.

Quality check (Bioanalyzer System, Agilent Technologies, USA), DNA quantification (Quantus, Promega, USA), and sequencing of the resulting libraries were conducted at the IZKF Core Facility Genomics (UKA, RWTH Aachen University). The libraries were pooled to reflect the nine different sequencing depths targeted, and sequenced on a NovaSeq6000 (Illumina, USA) with a NovaSeq 6000 S1 Reagent Kit v1.5 (2x150 cycles).

Bioinformatic analysis

An overview of the bioinformatic workflow is provided in **Supplementary Figure S3**. All steps are described in detail in the following sections.

Reference genomes

The genomes of all strains used in the Mock communities are hereon referred to as "reference genomes". They have been deposited in a public repository and published previously[9,10]. A phylogenetic tree was constructed from the genomes of the 70 isolates using PhyloPhlAn[12] v3.0.67 (options: *--diversity medium –f supermatrix_aa.cfg*) (**Supplementary Figure S2**). Genome characteristics, including size and GC content, were analysed using biobython[13] v.1.79 and bioawk[14] v1.0 (**Supplementary Table S1**).

Pre-processing of shotgun metagenomic data

Raw reads of samples with the different sequencing depths were further subsampled to the exact targeted sequencing depth, calculated as:

$$\frac{Sequencing\ depth\ in\ Gb\ \times\ 1{,}000{,}000{,}000}{150bp\ \times 2}$$

The raw FASTQ files were subsampled using seqtk[15] v1.2 with default settings. The targeted number of paired-end reads (150 bp) per sequencing depth was as follows: 333,333 reads (0.10 Gb), 833,333 (0.25 Gb), 1,666,667 (0.5 Gb), 2,500,000 (0.75 Gb), 3,333,333 (1.0Gb), 5,000,000 (1,5Gb), 6,666,667 (2.0Gb), 16,666,667 (5.0 Gb), 33,333,333 (10.0Gb). Adapters were removed and subsampled raw reads were quality-filtered with Trimmomatic[16] v.0.39 (options: *TRAILING:3 LEADING:3 SLIDINGWINDOW:5:20 MINLEN:50 ILLUMINACLIP:{adapters.fa}:2:30:10*). The bbduk command (options: *hdist=1 k=31*) in

125 BBMap[17] v.38.84 was used for removal of phiX sequences. Quality-filtered reads were assembled into
126 contigs using MegaHit[18] v1.2.9.

127

128 **Taxonomic coverage**

129 Coverage of the reference genomes by the quality-filtered metagenomic reads was determined using
130 coverM[19] v0.6.1 (options: *coverm genome --mapper bwa-mem --methods covered_fraction --min-*
131 *covered-fraction 0 --coupled*). The relative abundance profiles were calculated using the read count
132 option of coverM (options: *coverm genome --methods count --min-covered-fraction 0 --coupled*), which
133 were converted to relative abundance for each sample.

134 Average nucleotide identity (ANI) values between the reference genomes of four *Escherichia coli*
135 strains and four *Phocaeicola vulgatus* strains were calculated using FastANI[20] v1.34.

136

137 **Functional analysis**

138 Protein-coding genes were predicted in the pooled reference genomes and the assembled contigs of
139 the Mock samples (individually) using prodigal [21] v2.6.3.

140 Gene functions were annotated using KEGG Orthology and KofamScan[22] v1.3.0. Pathway completeness
141 was assessed by KEGG-Decoder[23] v1.3, as percentage of KOs covered, which are included in the
142 manually curated canonical pathways used by KEGG-Decoder.

143 To determine the ability of each sequencing depth to capture the protein-encoding potential in the
144 Mock communities, the genes predicted in the reference genomes were used to create a Diamond[24]
145 (v2.0.15) protein sequence database, with which the protein sequences predicted in the assemblies
146 (one assembly for each of the sequencing depths) were compared (options: *diamond blastp –sensitive*
147 *–query-cover 80 –id 90*).

148

149 **Metagenome-assembled genomes (MAGs)**

150 Contigs <1,000 bp were removed prior to reconstructing MAGs. An index table was built from the size-
151 filtered contigs using bowtie2[25] v2.5.1 (*bowtie2-build*) with default options. The decontaminated
152 paired-end reads were aligned to the bowtie index of size-filtered contigs (*bowtie2 –S - --very-sensitive-*
153 *local –no-unal –p 30*). Bam files were sorted with samtools[26] v1.17 (*samtools view –bS).* Contigs were
154 binned using Metabat2[27] v2.12.1 and its algorithm for calculating coverage of each sequence in the

155 assembly (*jgi_summarize_bam_contig_depths*) before creating the bins (*metabat2 -m 1500 --maxP 95*
156 *--minS 60 --maxEdges 200 --unbinned --seed 0*). The quality of the resulting bins was evaluated with
157 checkM[28] v1.1.3 using the lineage workflow. Accordingly, the MAGs were categorised as being of high
158 quality (>90% completeness, <5% contamination), medium quality (>70% completeness, <10%
159 contamination), or low quality (all that did not fulfil the previous criteria). They were then
160 taxonomically classified using GTDB-Tk[29] v2.3.2 (*gtdbtk classify_wf*) with the Genome Taxonomy
161 Database r207. The coverage of the reference genomes by the different MAGs was determined using
162 coverM[19] v0.6.1 (*coverm genome --mapper bwa-mem --methods covered_fraction --min-covered-*
163 *fraction 0 --single*). Reference genomes were aligned to MAGs by blastn v2.13.0 *(-evalue 1e-10 -*
164 *perc_identity 90.0*) and circular alignments plotted using shinyCircos[30].

165

166 **Statistics and Plotting**

167 All statistical tests were performed in R[31] v4.4.1. using RStudio v2023.03.0+386 and the packages
168 vegan[32] , reshape[33], tidyverse[34] and dplyr[35]. For creation of the graphs, the R package ggplot2[36] v3.5.1,
169 ggpubr[37] v0.6.0, and ggfortify[38] v0.4.17 were used. The specific statistical tests used in each analysis
170 are stated in the results.

171

172 **Results**

173 **Shallow metagenomics enables reliable reference-based taxonomic composition**

174 To investigate the impact of sequencing depth on taxonomic composition, recovery of the reference
175 genomes in the metagenomic data of the Mock communities was determined.

176 For the two types of Mock communities (even and staggered), reads were detected for all reference
177 genomes (N = 70 and 24, respectively) at 0.1 Gb sequencing depth. For Mock-even, the lowest
178 coverage at 0.1 Gb was 3.34 % for *Phocaeicola sartorii* CLA-AV-12 and the highest 58.7 % for
179 *Bacteroides uniformis* CLA-AV-11 (**Fig. 1A**). The threshold of 99 % coverage was first reached at 0.75 Gb
180 for *B. uniformis* CLA-AV-11. At the highest sequencing depth, all reference genomes reached >98 %
181 coverage. Full reference genome coverage (100 %) was only achieved at 10 Gb for V*eillonella*
182 *intestinalis* CLA-AV-13.

183 For Mock-stag, there was a clear trend of higher coverage with increasing DNA concentration of the
184 reference genomes at low sequencing depths (**Fig. 1B**). The reference genomes with lowest
185 concentration (0.01 %, 0.04 ng) had a minimum coverage of 0.53 % at 0.1 Gb and a maximum coverage

186 of 34.21 % at 10 Gb sequencing depth. At 10 Gb, three genomes reached 100% coverage:
187 *Thomasclavelia ramosa* CLA-JM-H52 and *Enterocloster clostridioformis* CLA-JM-H51, which were added
188 to the Mock with highest DNA concentration (17.2 %; 40 ng), and *V. intestinalis* CLA-AV-13, added at a
189 low concentration (0.86 %, 2 ng).

190 To obtain an overview of genome coverage, completeness categories were compared between the
191 sequencing depths (**Fig. 1C)**. At 0.1 Gb, 91 % of the genomes in Mock-even and 63 % in Mock-stag had
192 the lowest coverage (0-25 %). At 2 Gb, coverage in Mock-even was >50 % for 69 of the reference
193 genomes (99 %). All reference genomes reached >75 % coverage at 5 Gb and >90 % at 10 Gb. In Mock-
194 stag, four genomes (17 %), which were added at low concentrations (0.04 and 0.4 ng), did not reach
195 >50 %, even with the highest sequencing depth of 10 Gb. All remaining genomes reached >50 %
196 coverage already at 5 Gb.

197 We then compared relative abundance profiles for the two Mock community types. For Mock-even,
198 relative abundances oscillated around the expected value of 1.43 % for most of the 70 species, without
199 any significant effects of the sequencing depth on the overall abundance profile (p = 0.78; Kruskall-
200 Wallis test) (**Supplementary Fig. S4A**). Six species showed values >2.5 % for several sequencing depths,
201 with highest values of 10-15 % for *B. uniformis* CLA-AV-11. In addition, marked differences due to
202 sequencing depth were detected in the case of *Collinsella* sp. CLA-ER-H5. Sequencing depth had also
203 no influence on the overall relative abundance profile in the case of Mock-stag (p = 1; Kruskall-Wallis
204 test) (**Supplementary Fig. S4B**). The relative abundances generally followed the expected trend. The
205 experimental values deviated from theoretical relative abundances to a higher extent with increasing
206 amount of DNA.

207 In summary, reference-based detection of strains is possible with sequencing as little as 0.5 Gb, but
208 high coverage of reference genomes needs more data (>5 Gb). Relative abundance profiles were not
209 substantially affected by sequencing depths for most of the strains.

210

211 ***De-novo* MAG reconstruction for strain-level analysis requires deep sequencing and generates**
212 **chimeras**

213 The ability of shallow metagenomics to resolve strain-level variation was investigated using four strains
214 of *E. coli* and *P. vulgatus* within the Mock-even dataset (Supplementary **Fig. S2**). Coverage of their
215 reference genomes per sequencing depth, as determined during assignment of all reference genomes,
216 revealed that all strains were detected already at 0.1 Gb (**Fig. 2A**). All four reference genomes of both
217 species reached >75 % coverage at 5 Gb and >98 % at 10 Gb. All *P. vulgatus* strains showed a similar
218 increase in coverage with increasing sequencing depth, mirroring ANI values between the strains that

219 were all close to one another (**Fig. 2B**). In contrast, for the *E. coli* strains, the reference genomes sharing
220 the highest ANI value (*E. coli* Ec8850 and CR B11; ANI = 99.97 %) showed a congruent coverage increase
221 (**Fig. 2B**, curves overlay), whilst *E. coli* CLA-AD-1, characterized by an ANI value more distant to the
222 other strains (< 97 %), reached >99 % reference genome coverage already at the 2 Gb sequencing
223 depth (**Fig. 2B**).

224 Reference datasets are rarely available during metagenomic analysis, hence *de-novo* reconstruction of
225 genomes is widely used to facilitate strain-level analysis. MAGs were thus assembled separately for
226 each metagenomic samples. Overall, the number of bins increased with higher sequencing depth for
227 both Mock types (**Fig. 2C**, **left panels**). For Mock-even (N = 70 strains), only a low number of bins per
228 sequencing depth (≤10) was obtained below 5 Gb. For both, Mock-even and Mock-stag, more MAGs
229 than the actual number of reference genomes (dashed line) were formed at 10 Gb. This means several
230 reference genomes were represented by multiple fragmented MAGs, whilst not all reference genomes
231 were represented by MAGs (**Fig. 2C, right panels)**. The number of MAGs formed per reference genome
232 increased with the amount of data sequenced, with ≥2 MAGs observed for 22 and 12 reference
233 genomes across all sequencing depths in Mock-even and Mock-stag, respectively (**Supplementary Fig.
234 S5**). This suggests MAGs become increasingly fragmented with greater sequencing depth, rather than
235 coalescing into a single high-quality MAG.

236 Given that fragmented MAGs were obtained even at high depth of sequencing, we aimed to determine
237 the frequency of misassembly. MAGs were grouped by quality and their contigs were assigned to the
238 reference genomes to determine which isolates they were reconstructed from. In Mock-even, at
239 10 Gb, 10 of the 23 high-quality MAGs (hqMAGs; completeness >90 %, contamination <5 %) were
240 assigned to multiple reference genomes (red bars) with a threshold of >0.25 % coverage of a reference
241 genome (**Fig. 2D, top**). The proportion of MAGs assigned to only one (blue) vs. more (red) reference
242 genomes decreased further with decreasing quality of the MAGs. Taxonomic assignment of hqMAGs
243 to GTDB showed a result congruent with the assignment to our own reference genomes
244 (**Supplementary Tables S3**). Although four *E. coli* strains were included in Mock-even only one MAG
245 assigned to *E. coli* was reconstructed from each sequencing dataset > 1Gb, except 2 Gb. At 2 Gb, a
246 second *E. coli* MAG of very low completeness and quality occurred. All *E. coli* MAGs were
247 predominantly assigned to *E. coli* CLA-AD-1. In the case of *P. vulgatus* (also represented by four strains),
248 one MAG was found at 5 Gb and one at 10 Gb, both primarily assigned to *P. vulgatus* HDF, indicating
249 strain delineation was not possible with the method used (**Supplementary Table S3**).

250 For Mock-stag, a single hqMAG was assembled at 0.75 and 1 Gb (**Fig. 2D, bottom**). At both sequencing
251 depths, most sequences of this MAG were assigned to a reference genome with highest DNA
252 concentration (40ng, *T. ramosa* CLA-JM-H52, ≥95 % reference genome coverage). With increasing

253 sequencing depth, some hqMAGs assigned to reference genomes that were added with lower
254 concentrations started to be reconstructed **(Supplementary Table S3**). Surprisingly, a higher
255 proportion of single-origin MAGs occurred with decreasing quality, while at 10 Gb half of the hqMAGs
256 in Mock-stag were misassembled. To illustrate the misassembly of MAGs, three exemplary hqMAGs
257 (Mock-even, 10 Gb) were aligned to the reference genomes they covered >0.25 % using blastn (**Fig 2E**).
258 MAG 54 was a best-case scenario with only one reference genome (*V. intestinalis* CLA−AV−13) matched
259 with >98 % coverage. MAG 16 showed fragments of a second reference genome. MAG 40 represented
260 a highly chimeric hqMAG, including 12 reference genomes (of which 10 are shown, ranked by
261 decreasing coverage from outer to inner circle). The mean/maximum number of reference genomes
262 per hqMAG at 10 Gb was 2.3/12 and 1.5/2 for Mock-even and Mock-stag, respectively (**Supplementary**
263 **Tables S2**).

264 In summary, when high-quality reference datasets are available, read mapping facilitates strain-level
265 analysis of shallow metagenomic data. However, *de-novo* analysis based on MAG reconstruction leads
266 to contaminated genomes, even at supposed high-quality. Together with the creation of multiple
267 MAGs for individual strains, these findings suggest that strain-level diversity within MAG catalogues is
268 artificially inflated.

269

270 **Functional coverage is limited by shallow sequencing**

271 Analysing the functional potential of bacterial communities is a common reason to prefer
272 metagenomics instead of 16S rRNA gene amplicon sequencing. Therefore, we investigated the
273 sequencing depth required for functional analyses at the level of both, pathways and genes.

274 First, we investigated the impact of sequencing depth on pathway completeness, measured as
275 percentage of detected KOs involved in a pathway. Only those pathways occurring in the reference
276 genomes of the respective Mock were considered (**Fig. 3A**). Of the 178 pathways included in the
277 analysis by KEEG-Decoder[23], predicted proteins for 121 and 118 pathways were found in the reference
278 genomes for the Mock-even and Mock-stag, respectively. Pathway completeness increased with
279 sequencing depth, with a maximum of 80 pathways being complete at 10 Gb for Mock-even and 77 for
280 Mock-stag (**Fig. 3A**). For both types of community, 40 pathways remained incomplete and one was not
281 detected despite being present in the reference genomes. The average KEEG pathway completeness
282 increased with sequencing depth (**Fig. 3B**). At 5 Gb, pathway completeness reached >80 % for both
283 types of Mocks, and only a slight increase was observed after sequencing twice as deep (even, +7.79 %;
284 log, +3.69 %).

285 To account for functionally unassigned proteins, the protein sequences detected at each sequencing
286 depth were compared to the reference genomes protein repertoire (even, 227,543 protein sequences;
287 stag, 82,992). The number of protein sequences detected within a metagenome continued to increase
288 up to 94.67 % for Mock-even (+ 22.9 % between 5 and 10 Gb), but started to plateau at 5 Gb sequencing
289 depth for the less diverse Mock-stag (+ 12.9 % between 5 and 10 Gb) (**Fig. 3C**).

290 In summary, while a moderate sequencing depth of 5 Gb was sufficient for functional analysis at the
291 pathway level for both mock communities, analysis at the level of single protein sequences requires
292 greater sequencing depth depending on the diversity of the microbial community under investigation.

293

294 **Sample processing protocol and background DNA impact shallow sequencing results**

295 To account for the influence of wet-lab factors on shallow metagenomics results, the DNA libraries
296 were prepared in two facilities using different protocols (see methods). In addition, the two community
297 types (Mock-even and Mock-stag) were spiked or not with background DNA isolated from the gut
298 content of germfree mice to simulate host DNA.

299 PCA plots based on relative abundance profiles showed distinct clustering due to both factors
300 (background DNA, facility), with more pronounced effects along PC1 linked to differences in library
301 preparation protocols between facilities (**Fig. 4A**). Within a condition (background DNA /facility pair),
302 the extreme sequencing depths (0.1 and 10 Gb) tended to be most distant from each other. The
303 relative abundance profiles of Mock-stag prepared in facility 1, which used more template DNA (100
304 vs. 1 ng) and a lower number of PCR cycles (5 vs. 12), were less sensitive to the effects of sequencing
305 depth when background DNA was present (**Fig. 4A, lower panel**).

306 The effect of wet-lab factors on reference genome coverage was then assessed by calculating delta-
307 values between the two facilities or the presence/absence of background DNA. For Mock-even,
308 consistent profiles were observed across nearly all species at the 5 and 10 Gb sequencing depths,
309 regardless of facility or background DNA (**Fig. 4B**). In contrast, variations in reference genome coverage
310 of up to 39 % for facilities, and up to 13 % for background DNA were observed at sequencing depths
311 ≤2 Gb. For Mock-stag, the variation between samples from the two facilities was additionally
312 influenced by the initial DNA amount indicated as coloured boxes in **Fig. 4C.** With a reference genome
313 input >1 ng and a sequencing depth >1 Gb, the total variance in reference genome coverage between
314 the two facilities dropped below 3.69%. A similar trend was observed for the impact of background
315 DNA on the coverage of reference genomes in Mock-stag, but with a more pronounced influence of
316 the amount of DNA. The effect of background DNA was negligible (total variance 1.33 %) with ≥10 ng

317 template DNA and a sequencing depth ≥1.5 Gb. For a DNA amount >1 ng per strain in Mock-stag, a
318 sequencing depth of 5 Gb or higher was necessary (total variance 3.23 %).

319 Analysed per strain, the relative abundance profiles in Mock-even were consistent between the two
320 library preparation strategies, except for marked differences for *Collinsella* sp. CLA-ER-H5 due to
321 sequencing depth in facility 1 (**Supplementary Fig. S4A**). This indicates that the differences shown in
322 Figure 4A are due to changes in the relative abundance of a few strains. For Mock-stag, deviation from
323 the theoretical relative abundance values was different between the two facilities (**Supplementary Fig.**
324 **S4B**). The relative abundance of *E. clostridioformis*, for example, matched the expected relative
325 abundance in facility 1, while it had the highest relative abundance in facility 2. The most abundant
326 species in facility 1 was *T. ramosa*. This indicates that library preparation (specifically the amount of
327 template DNA and number of PCR cycles) influenced relative abundance profiles. (**Supplementary Fig.**
328 **S4B**)

329 We further analysed the effects of the wet-lab factors on the predicted protein profiles. Overall, facility
330 1 provided a more complete functional profile for both mock communities (**Supplementary Fig. S6**).
331 The addition of background DNA was associated with a decrease in coverage, which was attenuated
332 at high sequencing depths in both mock communities prepared in facility 1. In general, the influence
333 of background DNA was more pronounced with the protocol used in facility 2.

334 In summary, library preparation with higher amount of template DNA and less PCR cycles is
335 recommended for more robust taxonomic and functional profiles. If necessary, a higher sequencing
336 depth can compensate the effects of lower DNA amount or environmental DNA.

337

338 **Discussion**

339 In this work we investigated the potential and limitations of shallow metagenomic sequencing using
340 complex Mock communities and sequencing depths from 0.1 Gb (0.3 million reads) to 10 Gb (33 million
341 reads), with a specific focus on strain-level and functional analysis.

342 Previous studies reported that taxonomic profiles based on shallow shotgun sequencing (0.5-4 million
343 reads) can reflect those obtained at a higher sequencing depths[2,3,6,8]. Most of those studies relied on
344 deep sequencing of native samples then subsampled *in-silico*. Using defined input, we have confirmed
345 these results, showing that shallow metagenomics is suitable for reference-based taxonomic analysis:
346 reads for all strains were detected already at 0.1 Gb (0.3 million reads), even for low abundant taxa in
347 the staggered community, enabling prevalence and relative abundance analyses despite overall low
348 genome coverage. This is in line with the study of Hillmann et al. and Xu et al. who reported that 0.5
349 million sequences (= 0.15 Gb) and 0.5Gb, respectively, was sufficient for taxonomic analysis of the

350 human gut microbiome [2,3]. As most of the metagenomic costs is due to sequencing, being able to pool
351 more samples on the same run reduces the expenses, making shallow metagenomics at 0.5 Gb/sample
352 attractive for diversity and composition profiling in studies with large number of samples. However,
353 accurate analysis necessitates a high-quality, reference database of individual genomes.

354 Metagenomics is useful for high resolution taxonomic analysis, which is limited with 16S rRNA gene
355 amplicon sequencing. In a previous study using reference-based analysis, species profiles at 0.5 million
356 reads (0.15 Gb) showed an average correlation of 0.99 with ultra-deep sequencing (750 Gb)[3]. Using
357 mock communities, Xu et al. recovered only 62 % of the 62 species at 1 Gb/sample using MetaPhlan 2
358 and the Refseq database[2]. In contrast, we detected reads for all 70 and 24 strains in the Mock
359 communities at the lowest sequencing depth of 0.1 Gb, possibly due to enhanced reference databases.
360 Reference-based detection and discrimination of different strains with an ANI of up to 99.66% was
361 possible in our study. However, sequencing depths >5 Gb are required to obtain nearly full genome
362 coverage, which must be adjusted depending on the expected abundance of the targeted strains in
363 complex samples.

364 A high sequencing depth was required for high-quality MAG reconstruction, which agrees with
365 previous findings[3,6]. Importantly, at high sequencing depth (10 Gb), more MAGs were created than
366 input strains, yet not all strains were represented by a MAG. Inspecting this further revealed a
367 substantial number of hqMAGs that were chimeric due to containing fragments assigned to more than
368 one reference genome. This was despite excluding contigs smaller than 1,500 bp during binning to
369 prevent false inflation of chimerism. These results echo previous reports that ~5 % of genes within
370 MAGs differ from the dominant reconstructed taxa[39].

371 Our study demonstrates chimeric MAGs occur independent of sequencing depth, and deeper
372 sequencing does not guarantee more complete MAG reconstruction. Data from the staggered Mock
373 community indicates that having more reads for a reference genome does not increase its coverage
374 by a MAG and does not prevent the risk of multiple alignments. Instead, we observed that the creation
375 of several MAGs for the same reference genome was common. The use of multiple samples with
376 varying occurrence of taxa has been shown to limit the risk of creating spurious MAGs by enhancing
377 co-occurrence analysis[40]. In addition, consensus binning may help refining the MAGs by using multiple
378 binning algorithms and combining their output (e.g. metawrap[41]). Nevertheless, the data presented
379 raises concerns about the prevalence of spurious sequences in ever-growing MAG catalogues, as we
380 used a common bioinformatic workflow to create the MAGs.

381 To quantify the functional information lost with decreasing sequencing depth, we calculated the
382 average KEGG pathway completeness, showing that >2 Gb is required to reach >50 % completeness.

383 In contrast, one study with a Mock community of 62 human gut bacteria reported that the functional
384 description of shallow metagenomic samples (1 Gb) was "quite similar" to the one obtained from the
385 reference genomes at KEGG level 1 and 2[2]. Another study based on in-silico subsampling of an ultra-
386 deep metagenome (2.5 billion reads; 750 Gb) to 0.5 million reads (0.15 Gb) showed a nearly full
387 recovery of KEGG Orthology groups[3]. This result might be influenced by the subsampling process being
388 rarefied repeatedly and therefore creating a higher chance of recovering functions. At the level of
389 predicted protein sequences, our data suggests that shallow metagenomics is not advisable, as a
390 recovery of <75 % predicted proteins was reached at sequencing depths ≤5 Gb. This concords with a
391 previous study of antimicrobial resistance genes at different sequencing depths, in which at least 80
392 million reads (24 Gb) was needed to recover the full richness of AMR gene families[42].

393 Regarding the influence of wet-lab factors, the performance of library preparation procedures has
394 previously been compared for traditional metagenomic sequencing of native samples and Mock
395 communities, revealing that the community diversity, amount of input DNA, and the sequencing
396 platform alter the results[43–45]. When analysing the effects of library preparation in two facilities and
397 the addition of background DNA from the gut content of germfree mice, both factors impacted
398 taxonomic and functional results at the sequencing depths <5 Gb. Strains with a higher relative
399 abundance were less sensitive to these effects, even at lower sequencing depths. If the use of different
400 protocols in large studies cannot be avoided (e.g., when handling low biomass samples), or a high
401 amount of background DNA is to be expected, higher sequencing depth may improve the robustness
402 of results.

403 In summary, due to the lack of studies benchmarking shallow metagenomics using complex mock
404 communities, potential users may not be aware of the pitfalls of this approach. The sequencing depth
405 needs to match the main aims of a study. Factors to be considered include the expected microbial
406 diversity in the target samples, the evenness of distribution within the communities, and the amount
407 of DNA available. One limitation of this work is the focus on bacterial members of microbiomes,
408 although this is also the focus of most studies in the literature. Shallow metagenomics is a valid choice
409 for a large study aiming at database-guided taxonomic profiles. However, it is not appropriate for high-
410 resolution functional analysis and *de-novo* strain resolution. Even at high depths of sequencing, MAG-
411 based approaches require careful planning and execution to avoid the generation and spread of
412 artificial bacterial diversity.

413

414 **Acknowledgement**: We are grateful to: (i) Charlie Pauvert for helpful discussions, Evelyn Deis for
415 sample processing, Alina Viehof-Beckmann for providing gut content of germ-free mice, and Ntana
416 Kousetzi for outstanding support with molecular work (all from Functional Microbiome Research
417 Group, University Hospital of RWTH Aachen); (ii) Christe Driessen for genomic library preparation
418 (Department of Medical Microbiology, Infectious Diseases and Infection Prevention, Maastricht
419 University Medical Centre). This work was supported by the Genomics Facility, a core facility of the
420 Interdisciplinary Center for Clinical Research (IZKF) Aachen within the Faculty of Medicine at RWTH
421 Aachen University.

422

423 **Material and data availability**

424 The raw metagenomic sequencing data was deposited at the European Nucleotide Archive/NCBI and
425 is accessible under Project no. PRJEB83573.

426 The code used for bioinformatic analyses is available on GitHub:
427 https://github.com/ClavelLab/Benchmarking-shallow-Metagenomics

428 All strains (and their DNA) used in this study are available at the Leibniz Institute DSMZ (German
429 Collection of Microorganisms and Cell Cultures): www.dsmz.de/miBC; https://hibc.rwth-aachen.de/.

430

431 **Funding**

432 TC received funding from the German Research Foundation (DFG), project no. 403224013 (SFB1382,
433 Q02), project no. 460129525 (NFDI4Microbiota), and project no. 445552570.

434

435 **Ethics**

436 All mice experiments were performed under Ethical Approval (LANUV no. 81-02.04.2023.A253) in
437 accordance with the German Animal Protection Law (TierSchG).

438

Figure Legends

Figure 1: Reference-based taxonomic profiles

A Heatmap showing the coverage of all reference genomes in Mock-even by metagenomic reads at nine different sequencing depths (x-axis). The grey boxes indicate the multiple *E. coli* and *P. vulgatus* strains. **B** Same as in panel A for Mock-stag. The reference genomes were ranked (from top to bottom) according to increasing DNA amount in the mixture, as indicated by the colour gradient (from blue to red; 0.04, 0.4, 1, 2, 4, 10, 20, 40 ng). **C** Number of reference genomes per coverage category (colour gradients).

Figure 2: Strain analysis

A Reference genome coverage (y-axis) for the four strains of *E. coli* and *P. vulgatus* at the nine different sequencing depths (x-axis). **B** ANI values between the strains within each species. **C** Strain analysis based on the MAGs assembled in Mock-even (top) and Mock-stag (bottom) at each sequencing depth individually. The left panels (dark blue) depict the number of bins assembled from the shotgun data; the right panels (lighter blue) show the number of reference genomes matching MAGs (coverM), indicating that multiple MAGs were reconstructed for some Mock species. The reference genome with the highest coverage by MAG reads was chosen as representative of that MAG. The average coverage of representative genomes is shown in Supplementary Table S2. **D** Number of high-, medium- and low-quality MAGs assigned to either one (blue) or several (red) reference genomes (with coverM using a cut-off of > 0.25% coverage of reference genomes). **E** Three exemplary high-quality (> 90 % completeness, < 5% contamination) MAGs (dark grey, outer circle) reconstructed from Mock-even at the 10 Gb sequencing depth. They were aligned to the reference genomes they covered by more than 0.25% to illustrate different categories of misassembly. The predominantly covered reference genome is depicted in blue, while chimeric sequences of further reference genomes (inner circles) are shown in red.

Figure 3: Functional coverage

A Completeness of all KEGG pathways present in the reference genomes in Mock-even (left) and Mock-stag (right) at the nine different sequencing depths (columns). **B** Average KEGG pathway completeness. **C** Fraction of predicted proteins from the reference genomes covered by the metagenomic assemblies at the respective sequencing depths.

Figure 4: Effects of background DNA (bgDNA) and library preparation in two facilities

472 **A** PCA plot of relative abundance profiles in Mock-even (top) and Mock-stag (bottom). **B** Difference in
473 coverage of reference genomes between facility 1 and facility 2 (left) and samples with or without
474 background DNA (right) for Mock-even. **C** Same as in panel B for Mock-stag. The reference genomes
475 were ranked (from top to bottom) according to increasing DNA amount in the mixture, as indicated by
476 the colour gradient (from blue to red; 0.04, 0.4, 1, 2, 4, 10, 20, 40 ng).

477

478 **Supplementary Figure S1:** Experimental design
479 The genomic DNA of 70 bacterial isolates (www.dsmz.de/miBC; https://www.hibc.rwth-aachen.de/)
480 was used as input to create a mock community with all members evenly distributed (Mock-even) or a
481 community with a subset of 24 isolates with staggered distribution (Mock-stag). A second version of
482 each Mock community was created by spiking DNA isolated from the gut content of germfree mice.
483 Libraries for all four mock communities were prepared in two different sequencing facilities. The
484 libraries were all sequenced in Facility 1, including the nine sequencing depths shown in the figure.
485 Bioinformatic analyses included: (i) number and relative abundance of strains; (ii) coverage of
486 reference genomes; (iii) number and diversity of predicted proteins and completeness of functional
487 pathways; (iv) strain-level resolution using both, a reference-based and metagenome-assembled
488 genome (MAG) approach.

489

490 **Supplementary Figure S2:** Phylogenetic tree of bacterial strains in the mock communities
491 A phylogenetic tree of high-quality, draft genomes of the 70 isolates was created using PhyloPhlAn
492 v3.0.67. The phyla to which the strains belong are represented by coloured boxes at the end of the
493 branches. The strains used for Mock-stag are highlighted with colours indicating their respective DNA
494 concentration in the mixture.

495

496 **Supplementary Figure S3:** Schematics of the bioinformatics workflow
497 The details of each analysis are described in the methods section.

498

499 **Supplementary Figure S4:** Relative abundance of reads per reference genome
500 Relative abundances of each strain in (**A**) Mock-even, and (**B**) Mock-stag. Metagenomic reads were
501 assigned to the reference genomes using coverM. Sequencing depths are indicated with the coloured
502 dots. The theoretical relative abundances are indicated by vertical dashed lines. For Mock-stag in panel

503 B, the reference genomes are ranked from top to bottom by increasing DNA amount in the mixture as
504 indicated by coloured boxes (from blue to red; 0.04, 0.4, 1, 2, 4, 10, 20, 40 ng).

505

506 **Supplementary Figure S5:** Reference genomes per MAG for Mock-stag and Mock-even.
507 Stacked bar charts of the number of MAGs per sequencing depth. Each box represents one reference
508 genome. The size of the box indicates the number of MAGs that were assigned to this reference
509 genome. The colours in the graph for Mock-stag (top) indicate the amount of input DNA for a given
510 strain in the pool. The colour of the graph for Mock-even (bottom) indicate the reference genomes.

511

512 **Supplementary Figure S6:** Bar charts of protein sequences recovered.
513 Protein sequenced predicted in the metagenomic datasets as percentage of the protein sequences
514 present in the reference genomes (y-axis). The colour gradient indicates the facility where the library
515 was prepared and the addition of background DNA (bgDNA) from the caecum of germfree mice.

516

517 **Supplementary Tabel S1**: Characteristics of the reference genomes
518 **Supplementary Tabel S2:** Assignment of MAGs to reference genomes and GTDB
519
520

521 **References**

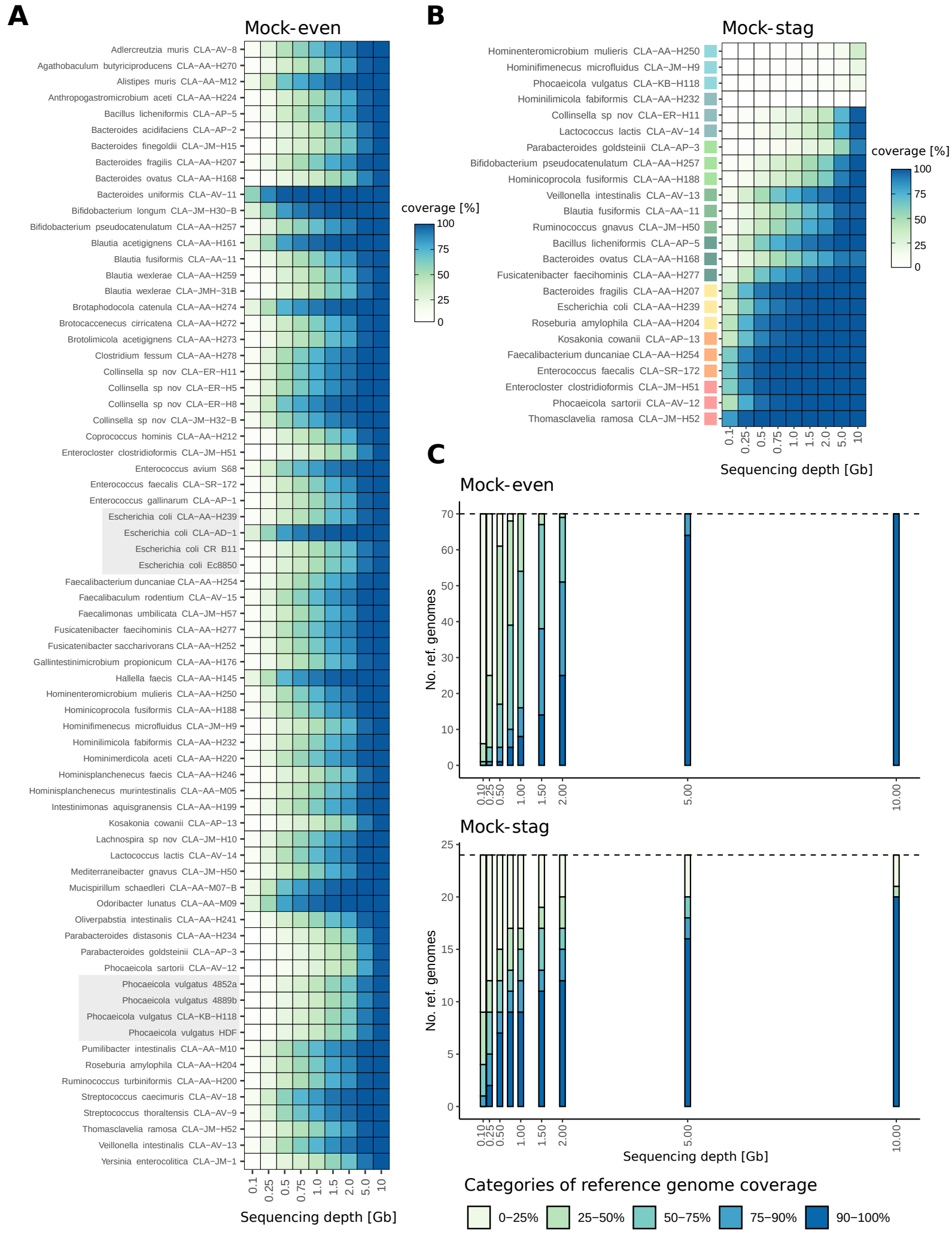
522 1. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and
523 next-generation sequencing-based diversity studies. *Nucleic Acids Research* **41**, e1–e1 (2013).
524 2. Xu, W. *et al.* Characterization of Shallow Whole-Metagenome Shotgun Sequencing as a High-
525 Accuracy and Low-Cost Method by Complicated Mock Microbiomes. *Front. Microbiol.* **12**, 678319
526 (2021).
527 3. Hillmann, B. *et al.* Evaluating the Information Content of Shallow Shotgun Metagenomics.
528 *mSystems* **3**, 10.1128/msystems.00069-18 (2018).
529 4. La Reau, A. J. *et al.* Shallow shotgun sequencing reduces technical variation in microbiome
530 analysis. *Sci Rep* **13**, 7668 (2023).
531 5. Stothart, M. R., McLoughlin, P. D. & Poissant, J. Shallow shotgun sequencing of the microbiome
532 recapitulates 16S amplicon results and provides functional insights. *Molecular Ecology Resources*
533 **23**, 549–564 (2023).
534 6. Tremblay, J., Schreiber, L. & Greer, C. W. High-resolution shotgun metagenomics: the more data,
535 the better? *Briefings in Bioinformatics* **23**, bbac443 (2022).
536 7. Usyk, M. *et al.* Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and
537 harmonization of these platforms for epidemiological studies. *Cell Reports Methods* **3**, 100391
538 (2023).
539 8. Cattonaro, F., Spadotto, A., Radovic, S. & Marroni, F. Do you cov me? Effect of coverage
540 reduction on metagenome shotgun sequencing studies. *F1000Res* **7**, 1767 (2020).
541 9. Afrizal, A. *et al.* Enhanced cultured diversity of the mouse gut microbiota enables custom-made
542 synthetic communities. *Cell Host & Microbe* **30**, 1630-1645.e25 (2022).
543 10. Hitch, T. C. A. *et al.* Broad diversity of human gut bacteria accessible via a traceable strain
544 deposition system. Preprint at https://doi.org/10.1101/2024.06.20.599854 (2024).

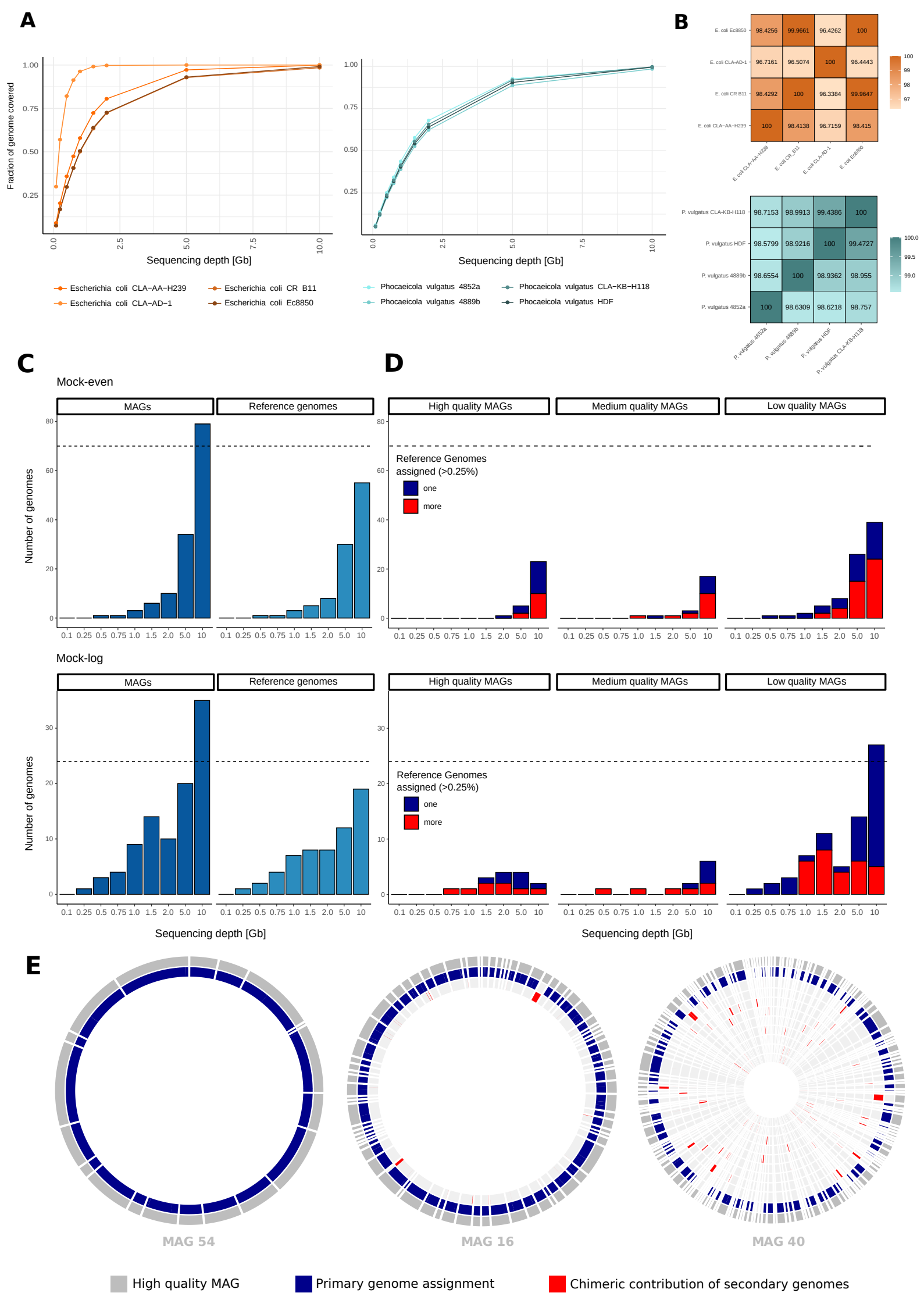
545 11. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of
546 an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl Environ*
547 *Microbiol* **63**, 2802–2813 (1997).
548 12. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from
549 metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**, 2500 (2020).
550 13. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology
551 and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
552 14. Li, H. bioawk, https://github.com/lh3/bioawk.
553 15. Li, H. seqtk, https://github.com/lh3/seqtk.
554 16. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
555 data. *Bioinformatics* **30**, 2114–2120 (2014).
556 17. Bushnell, B. BBMap, sourceforge.net/projects/bbmap/.
557 18. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution
558 for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**,
559 1674–1676 (2015).
560 19. Aroney, S. T. N. *et al.* CoverM: Read coverage calculator for metagenomics. Zenodo
561 https://doi.org/10.5281/ZENODO.10531253 (2024).
562 20. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI
563 analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114
564 (2018).
565 21. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification.
566 *BMC Bioinformatics* **11**, 119 (2010).
567 22. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive
568 score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
569 23. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-
570 distributed bacterial phototroph. *The ISME Journal* **12**, 1861–1866 (2018).

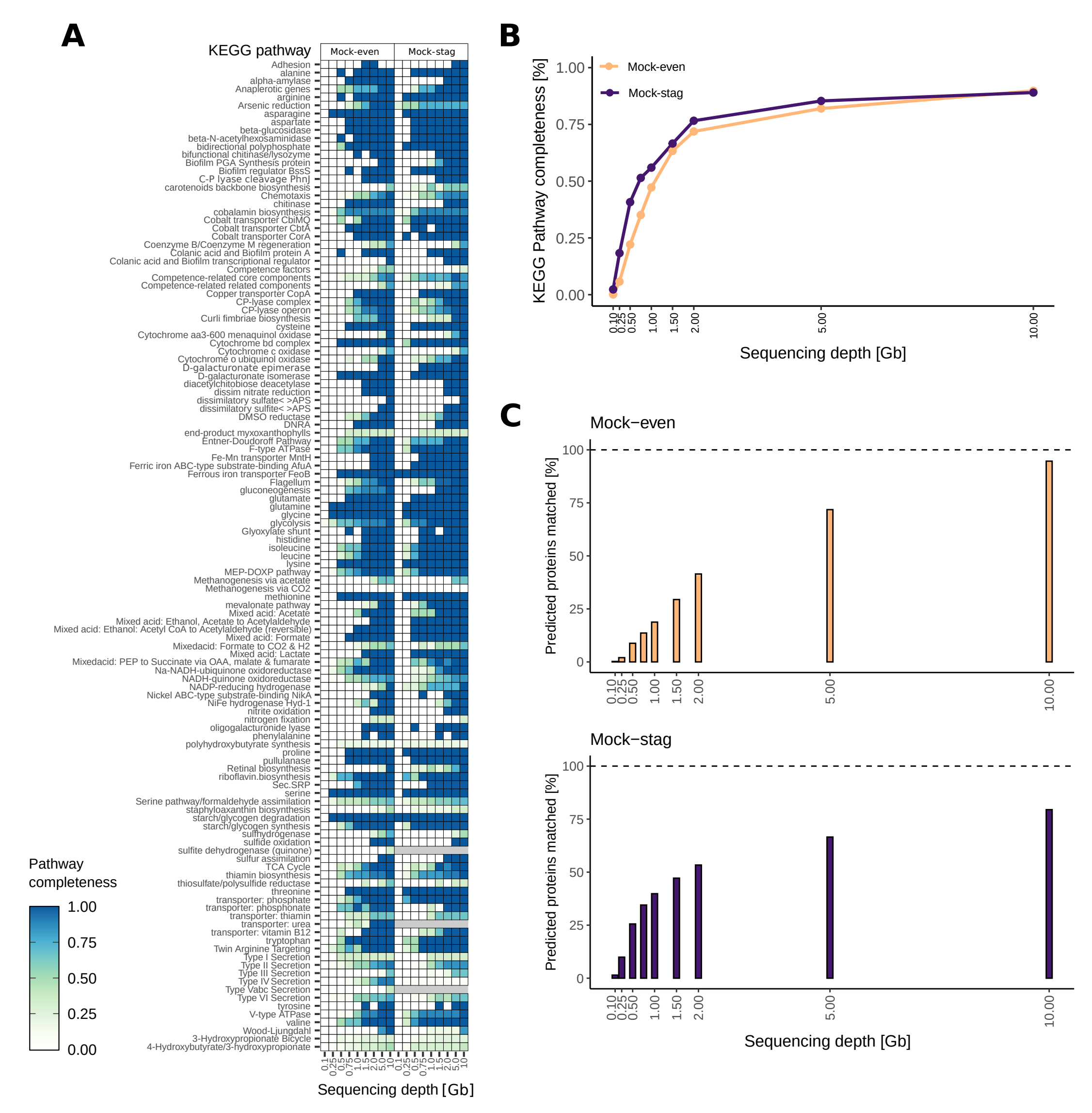
571 24. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat*
572 *Methods* **12**, 59–60 (2015).
573 25. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–
574 359 (2012).
575 26. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
576 27. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
577 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
578 28. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the
579 quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome*
580 *Res.* **25**, 1043–1055 (2015).
581 29. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify
582 genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
583 30. Yu, Y., Ouyang, Y. & Yao, W. shinyCircos: an R/Shiny application for interactive creation of Circos
584 plot. *Bioinformatics* **34**, 1229–1231 (2018).
585 31. R Core Team. R: A Language and Environment for Statistical Computing, https://www.R-
586 project.org/. *R Foundation for Statistical Computing, Vienna, Austria* (2021).
587 32. Jari Oksanen and Gavin L. Simpson and F. Guillaume Blanchet and Roeland Kindt and Pierre
588 Legendre and Peter R. Minchin and R.B. O'Hara and Peter Solymos and M. Henry H. Stevens and
589 Eduard Szoecs and Helene Wagner and Matt Barbour and Michael Bedward and Ben Bolker and
590 Daniel Borcard and Tuomas Borman and Gustavo Carvalho and Michael Chirico and Miquel {De
591 Caceres} and Sebastien Durand and Heloisa Beatriz Antoniazi Evangelista and Rich FitzJohn and
592 Michael Friendly and Brendan Furneaux and Geoffrey Hannigan and Mark O. Hill and Leo Lahti
593 and Dan McGlinn and Marie-Helene Ouellette and Eduardo {Ribeiro Cunha} and Tyler Smith and
594 Adrian Stier and Cajo J.F. {Ter Braak} and James Weedon. vegan: Community Ecology Package,
595 https://vegandevs.github.io/vegan/. (2024).
596 33. Hadley Wickham. Flexibly Reshape Data, http://had.co.nz/reshape. (2022).

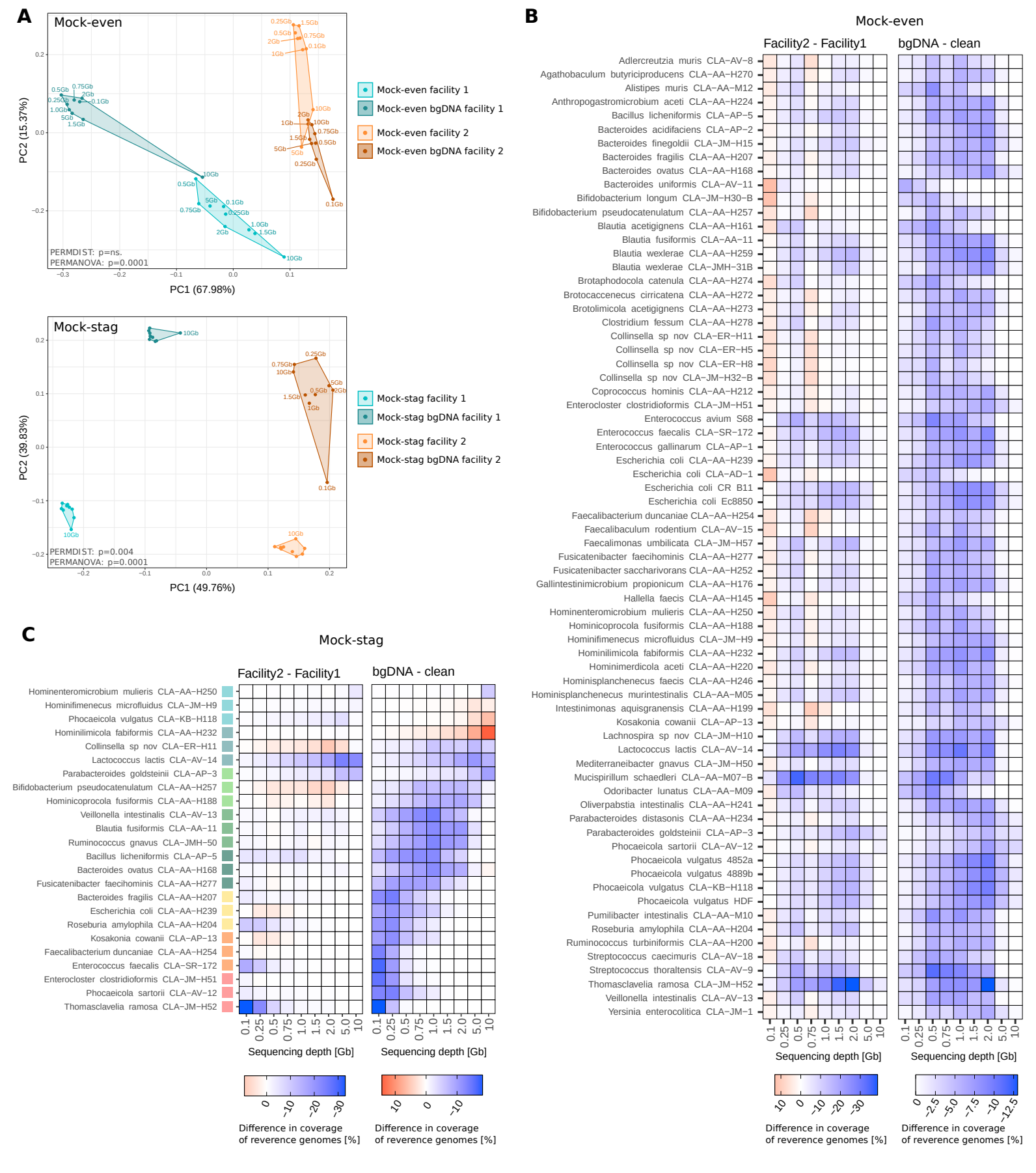
34. Wickham, H. *et al.* Welcome to the Tidyverse. *JOSS* **4**, 1686 (2019).
35. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. 1.1.4 https://doi.org/10.32614/CRAN.package.dplyr (2014).
36. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer international publishing, Cham, 2016).
37. Alboukadel Kassambara. ggpubr: 'ggplot2' Based Publication Ready Plots, https://rpkgs.datanovia.com/ggpubr/. (2023).
38. Yuan Tang and Masaaki Horikoshi. ggfortify: Data Visualization Tools for Statistical Analysis Results, https://CRAN.R-project.org/package=ggfortify. (2016).
39. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl Environ Microbiol* **87**, e02593-20 (2021).
40. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat Methods* **20**, 1170–1173 (2023).
41. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
42. Gweon, H. S. *et al.* The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environmental Microbiome* **14**, 7 (2019).
43. Bowers, R. M. *et al.* Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* **16**, 856 (2015).
44. Poulsen, C. S., Ekstrøm, C. T., Aarestrup, F. M. & Pamp, S. J. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiol Spectr* **10**, e00090-22 (2022).
45. Gaulke, C. A. *et al.* Evaluation of the Effects of Library Preparation Procedure and Sample Characteristics on the Accuracy of Metagenomic Profiles. *mSystems* **6**, 10.1128/msystems.00440-21 (2021).

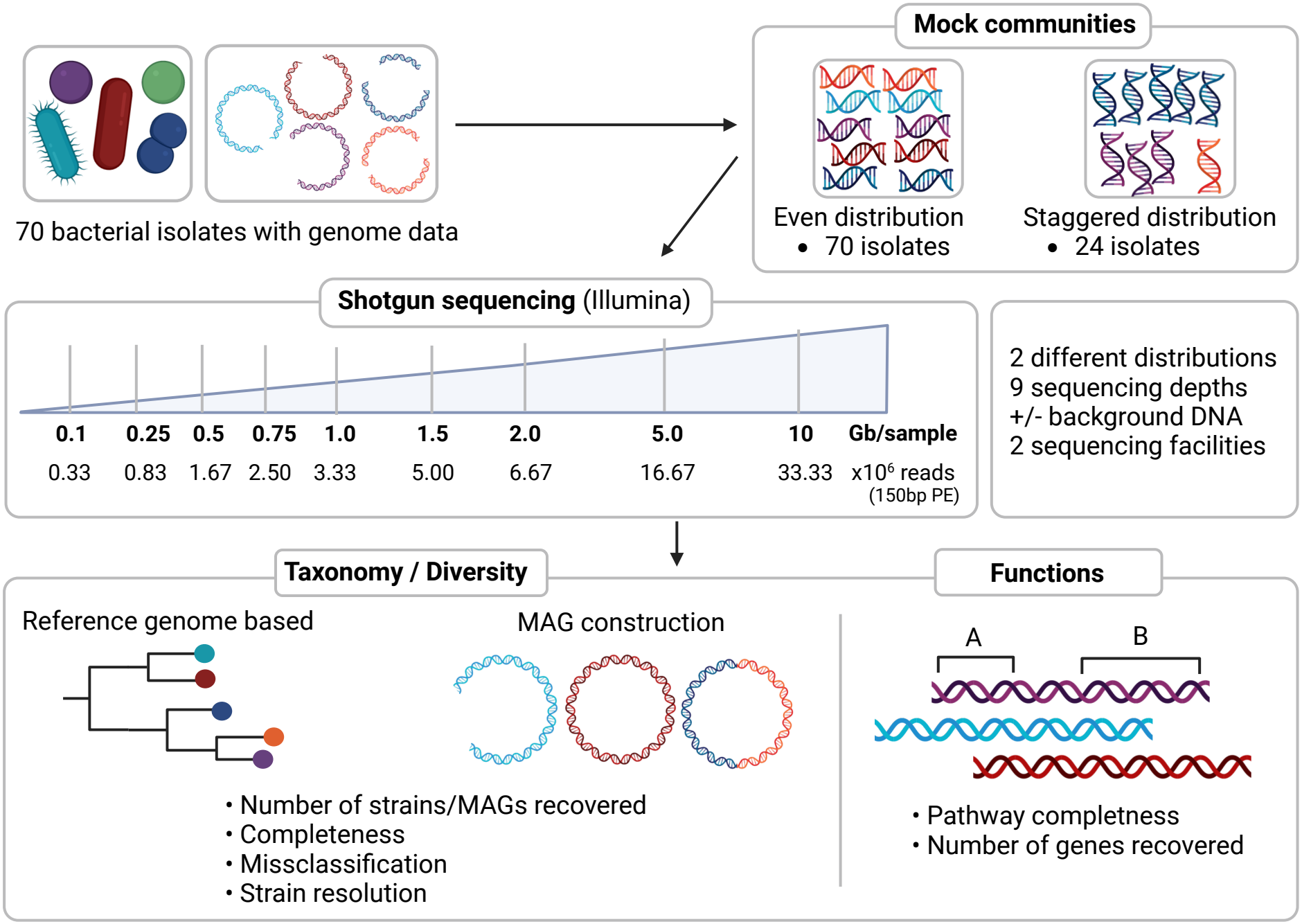
623


A
Mock-even
Adlercreutzia muris CLA−AV−8
Agathobaculum butyriciproducens CLA−AA−H270
Alistipes muris CLA−AA−M12
Anthropogastromicrobium aceti CLA−AA−H224
Bacillus licheniformis CLA−AP−5
Bacteroides acidifaciens CLA−AP−2
Bacteroides finegoldii CLA−JM−H15
Bacteroides fragilis CLA−AA−H207
Bacteroides ovatus CLA−AA−H168
Bacteroides uniformis CLA−AV−11
Bifidobacterium longum CLA−JM−H30−B
Bifidobacterium pseudocatenulatum CLA−AA−H257
Blautia acetigignens CLA−AA−H161
Blautia fusiformis CLA−AA−11
Blautia wexlerae CLA−AA−H259
Blautia wexlerae CLA−JMH−31B
Brotaphodocola catenula CLA−AA−H274
Brotocaccenecus cirricatena CLA−AA−H272
Brotolimicola acetigignens CLA−AA−H273
Clostridium fessum CLA−AA−H278
Collinsella sp nov CLA−ER−H11
Collinsella sp nov CLA−ER−H5
Collinsella sp nov CLA−ER−H8
Collinsella sp nov CLA−JM−H32−B
Coprococcus hominis CLA−AA−H212
Enterocloster clostridioformis CLA−JM−H51
Enterococcus avium S68
Enterococcus faecalis CLA−SR−172
Enterococcus gallinarum CLA−AP−1
Escherichia coli CLA−AA−H239
Escherichia coli CLA−AD−1
Escherichia coli CR B11
Escherichia coli Ec8850
Faecalibacterium duncaniae CLA−AA−H254
Faecalibaculum rodentium CLA−AV−15
Faecalimonas umbilicata CLA−JM−H57
Fusicatenibacter faecihominis CLA−AA−H277
Fusicatenibacter saccharivorans CLA−AA−H252
Gallintestinimicrobium propionicum CLA−AA−H176
Hallella faecis CLA−AA−H145
Hominenteromicrobium mulieris CLA−AA−H250
Hominicoprocola fusiformis CLA−AA−H188
Hominifimenecus microfluidus CLA−JM−H9
Hominilimicola fabiformis CLA−AA−H232
Hominimerdicola aceti CLA−AA−H220
Hominisplanchenecus faecis CLA−AA−H246
Hominisplanchenecus murintestinalis CLA−AA−M05
Intestinimonas aquisgranensis CLA−AA−H199
Kosakonia cowanii CLA−AP−13
Lachnospira sp nov CLA−JM−H10
Lactococcus lactis CLA−AV−14
Mediterraneibacter gnavus CLA−JM−H50
Mucispirillum schaedleri CLA−AA−M07−B
Odoribacter lunatus CLA−AA−M09
Oliverpabstia intestinalis CLA−AA−H241
Parabacteroides distasonis CLA−AA−H234
Parabacteroides goldsteinii CLA−AP−3
Phocaeicola sartorii CLA−AV−12
Phocaeicola vulgatus 4852a
Phocaeicola vulgatus 4889b
Phocaeicola vulgatus CLA−KB−H118
Phocaeicola vulgatus HDF
Pumilibacter intestinalis CLA−AA−M10
Roseburia amylophila CLA−AA−H204
Ruminococcus turbiniformis CLA−AA−H200
Streptococcus caecimuris CLA−AV−18
Streptococcus thoraltensis CLA−AV−9
Thomasclavelia ramosa CLA−JM−H52
Veillonella intestinalis CLA−AV−13
Yersinia enterocolitica CLA−JM−1
0.1
0.25
0.5
0.75
1.0
1.5
2.0
5.0
10
Sequencing depth [Gb]
coverage [%]
100
75
50
25
0
B
Mock-stag
Hominenteromicrobium mulieris CLA−AA−H250
Hominifimenecus microfluidus CLA−JM−H9
Phocaeicola vulgatus CLA−KB−H118
Hominilimicola fabiformis CLA−AA−H232
Collinsella sp nov CLA−ER−H11
Lactococcus lactis CLA−AV−14
Parabacteroides goldsteinii CLA−AP−3
Bifidobacterium pseudocatenulatum CLA−AA−H257
Hominicoprocola fusiformis CLA−AA−H188
Veillonella intestinalis CLA−AV−13
Blautia fusiformis CLA−AA−11
Ruminococcus gnavus CLA−JM−H50
Bacillus licheniformis CLA−AP−5
Bacteroides ovatus CLA−AA−H168
Fusicatenibacter faecihominis CLA−AA−H277
Bacteroides fragilis CLA−AA−H207
Escherichia coli CLA−AA−H239
Roseburia amylophila CLA−AA−H204
Kosakonia cowanii CLA−AP−13
Faecalibacterium duncaniae CLA−AA−H254
Enterococcus faecalis CLA−SR−172
Enterocloster clostridioformis CLA−JM−H51
Phocaeicola sartorii CLA−AV−12
Thomasclavelia ramosa CLA−JM−H52
0.1
0.25
0.5
0.75
1.0
1.5
2.0
5.0
10
Sequencing depth [Gb]
coverage [%]
100
75
50
25
0
C
Mock-even
No. ref. genomes
70
60
50
40
30
20
10
0
0.10
0.25
0.50
1.00
1.50
2.00
5.00
10.00
Mock-stag
No. ref. genomes
25
20
15
10
5
0
0.10
0.25
0.50
1.00
1.50
2.00
5.00
10.00
Sequencing depth [Gb]
Categories of reference genome coverage
0−25%
25−50%
50−75%
75−90%
90−100%

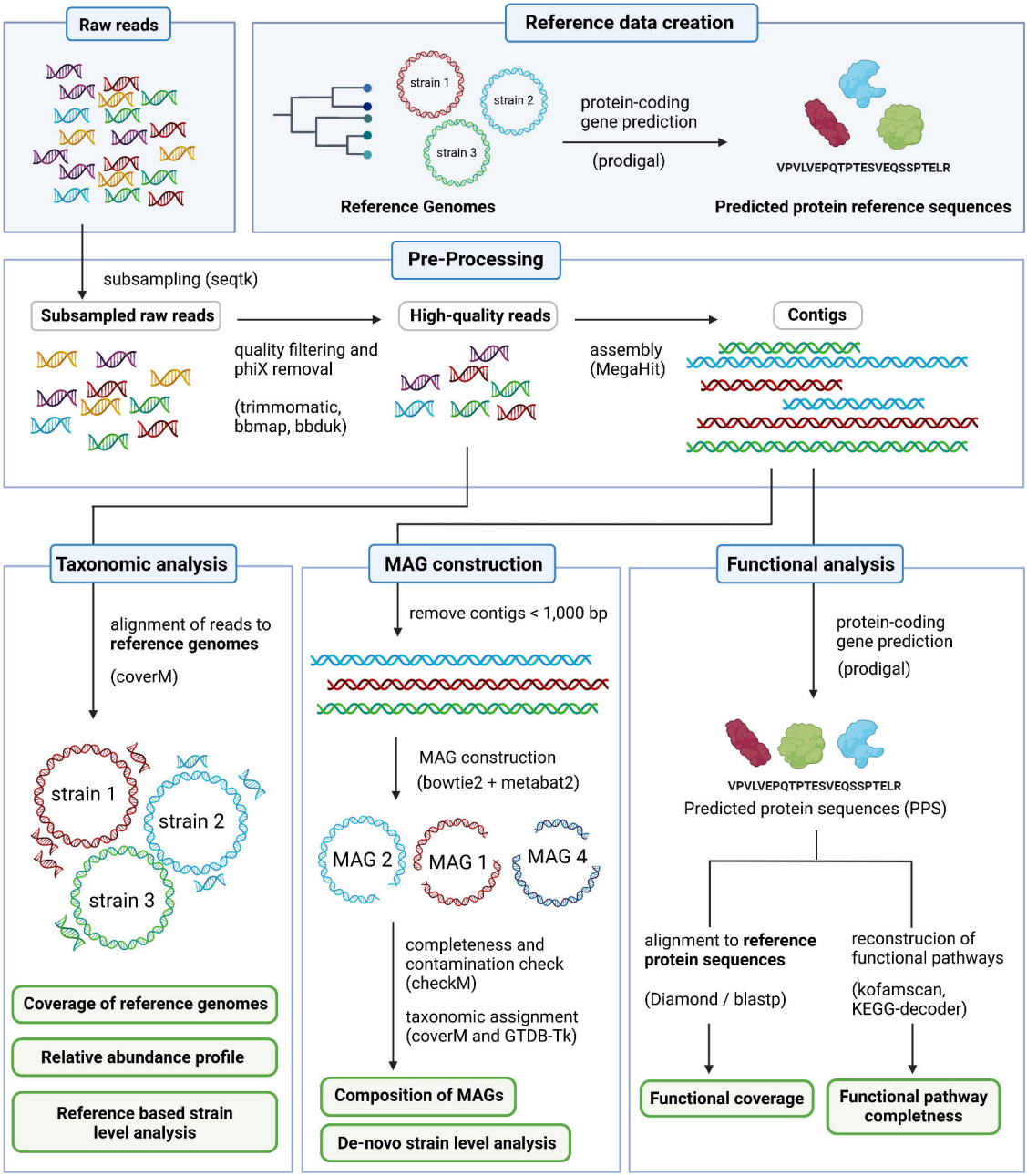

A
Fraction of genome covered
Sequencing depth [Gb]
Escherichia coli CLA−AA−H239
Escherichia coli CLA−AD−1
Escherichia coli CR B11
Escherichia coli Ec8850
Phocaeicola vulgatus 4852a
Phocaeicola vulgatus 4889b
Phocaeicola vulgatus CLA−KB−H118
Phocaeicola vulgatus HDF
B
E. coli Ec8850 98.4256 99.9661 96.4262 100
E. coli CLA-AD-1 96.7161 96.5074 100 96.4443
E. coli CR B11 98.4292 100 96.3384 99.9647
E. coli CLA−AA−H239 100 98.4138 96.7159 98.415
P. vulgatus CLA-KB-H118 98.7153 98.9913 99.4386 100
P. vulgatus HDF 98.5799 98.9216 100 99.4727
P. vulgatus 4889b 98.6554 100 98.9362 98.955
P. vulgatus 4852a 100 98.6309 98.6218 98.757
C
Mock-even
MAGs
Reference genomes
Number of genomes
Sequencing depth [Gb]
Mock-log
D
High quality MAGs
Medium quality MAGs
Low quality MAGs
Reference Genomes assigned (>0.25%)
one
more
E
MAG 54
MAG 16
MAG 40
High quality MAG
Primary genome assignment
Chimeric contribution of secondary genomes

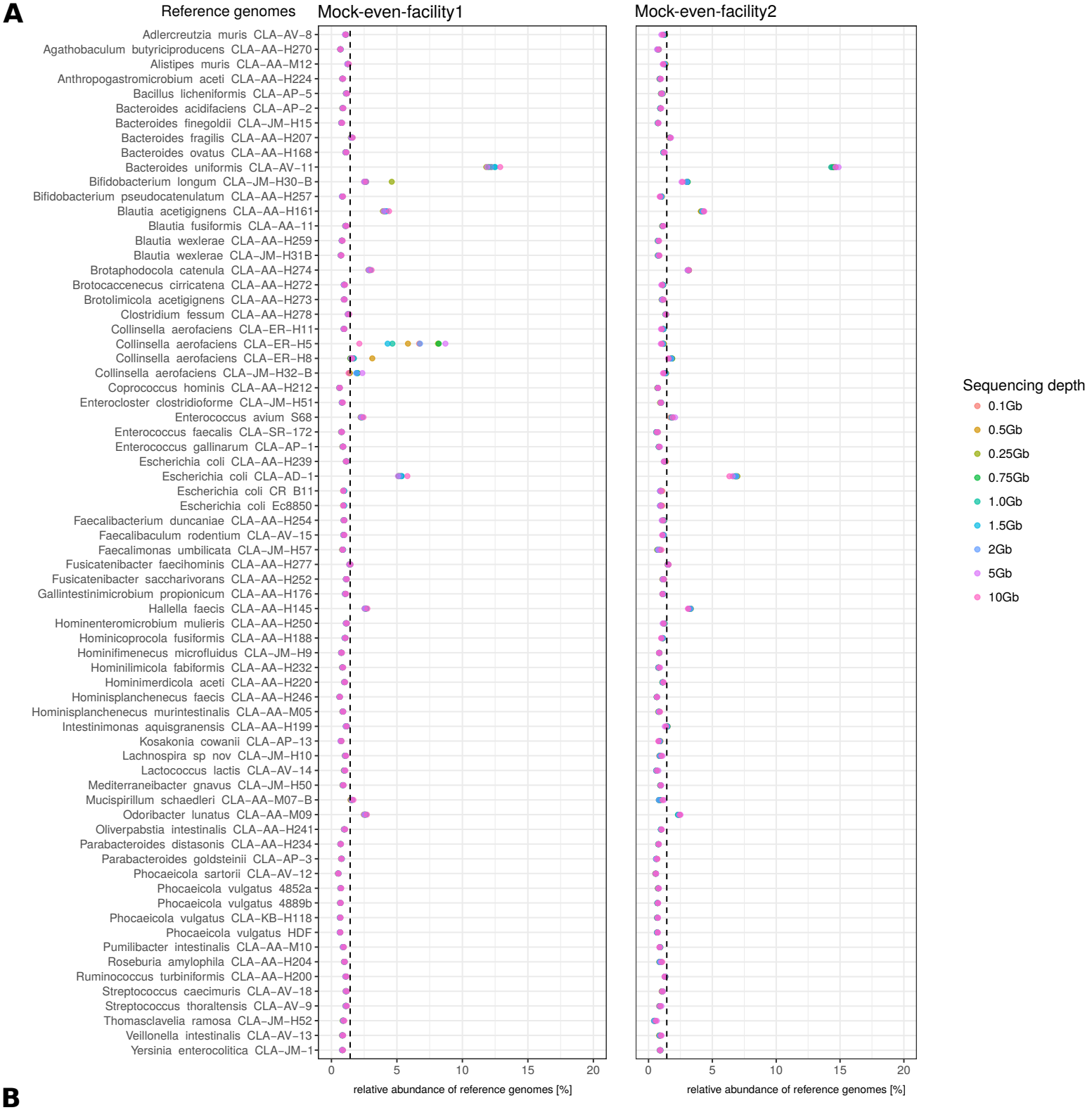

A
KEGG pathway
Mock-even
Mock-stag
Adhesion
alanine
alpha-amylase
Anaplerotic genes
arginine
Arsenic reduction
asparagine
aspartate
beta-glucosidase
beta-N-acetylhexosaminidase
bidirectional polyphosphate
bifunctional chitinase/lysozyme
Biofilm PGA Synthesis protein
Biofilm regulator BssS
C-P lyase cleavage PhnJ
carotenoids backbone biosynthesis
Chemotaxis
chitinase
cobalamin biosynthesis
Cobalt transporter CbiMQ
Cobalt transporter CbtA
Cobalt transporter CorA
Coenzyme B/Coenzyme M regeneration
Colanic acid and Biofilm protein A
Colanic acid and Biofilm transcriptional regulator
Competence factors
Competence-related core components
Competence-related related components
Copper transporter CopA
CP-lyase complex
CP-lyase operon
Curli fimbriae biosynthesis
cysteine
Cytochrome aa3-600 menaquinol oxidase
Cytochrome bd complex
Cytochrome c oxidase
Cytochrome o ubiquinol oxidase
D-galacturonate epimerase
D-galacturonate isomerase
diacetylchitobiose deacetylase
dissim nitrate reduction
dissimilatory sulfate< >APS
dissimilatory sulfite< >APS
DMSO reductase
DNRA
end-product myxoxanthophylls
Entner-Doudoroff Pathway
F-type ATPase
Fe-Mn transporter MntH
Ferric iron ABC-type substrate-binding AfuA
Ferrous iron transporter FeoB
Flagellum
gluconeogenesis
glutamate
glutamine
glycine
glycolysis
Glyoxylate shunt
histidine
isoleucine
leucine
lysine
MEP-DOXP pathway
Methanogenesis via acetate
Methanogenesis via CO2
methionine
mevalonate pathway
Mixed acid: Acetate
Mixed acid: Ethanol, Acetate to Acetylaldehyde
Mixed acid: Ethanol: Acetyl CoA to Acetylaldehyde (reversible)
Mixed acid: Formate
Mixedacid: Formate to CO2 & H2
Mixed acid: Lactate
Mixedacid: PEP to Succinate via OAA, malate & fumarate
Na-NADH-ubiquinone oxidoreductase
NADH-quinone oxidoreductase
NADP-reducing hydrogenase
Nickel ABC-type substrate-binding NikA
NiFe hydrogenase Hyd-1
nitrite oxidation
nitrogen fixation
oligogalacturonide lyase
phenylalanine
polyhydroxybutyrate synthesis
proline
pullulanase
Retinal biosynthesis
riboflavin.biosynthesis
Sec.SRP
serine
Serine pathway/formaldehyde assimilation
staphyloaxanthin biosynthesis
starch/glycogen degradation
starch/glycogen synthesis
sulfhydrogenase
sulfide oxidation
sulfite dehydrogenase (quinone)
sulfur assimilation
TCA Cycle
thiamin biosynthesis
thiosulfate/polysulfide reductase
threonine
transporter: phosphate
transporter: phosphonate
transporter: thiamin
transporter: urea
transporter: vitamin B12
tryptophan
Twin Arginine Targeting
Type I Secretion
Type II Secretion
Type III Secretion
Type IV Secretion
Type Vabc Secretion
Type VI Secretion
tyrosine
V-type ATPase
valine
Wood-Ljungdahl
3-Hydroxypropionate Bicycle
4-Hydroxybutyrate/3-hydroxypropionate
0.1 0.25 0.5 0.75 1.0 1.5 2.0 5.0 10
Sequencing depth [Gb]
Pathway completeness
1.00
0.75
0.50
0.25
0.00
B
Mock-even
Mock-stag
KEGG Pathway completeness [%]
1.00
0.75
0.50
0.25
0.00
0.10 0.25 0.50 1.00 1.50 2.00 5.00 10.00
Sequencing depth [Gb]
C
Mock−even
Predicted proteins matched [%]
100
75
50
25
0
0.10 0.25 0.50 1.00 1.50 2.00 5.00 10.00
Mock−stag
Predicted proteins matched [%]
100
75
50
25
0
0.10 0.25 0.50 1.00 1.50 2.00 5.00 10.00
Sequencing depth [Gb]


A
Mock-even
PC2 (15.37%)
PC1 (67.98%)
PERMDIST: p=ns.
PERMANOVA: p=0.0001
Mock-even facility 1
Mock-even bgDNA facility 1
Mock-even facility 2
Mock-even bgDNA facility 2
Mock-stag
PC2 (39.83%)
PC1 (49.76%)
PERMDIST: p=0.004
PERMANOVA: p=0.0001
Mock-stag facility 1
Mock-stag bgDNA facility 1
Mock-stag facility 2
Mock-stag bgDNA facility 2
B
Mock-even
Facility2 - Facility1
bgDNA - clean
Sequencing depth [Gb]
Difference in coverage of reverence genomes [%]
C
Mock-stag
Facility2 - Facility1
bgDNA - clean
Sequencing depth [Gb]
Difference in coverage of reverence genomes [%]

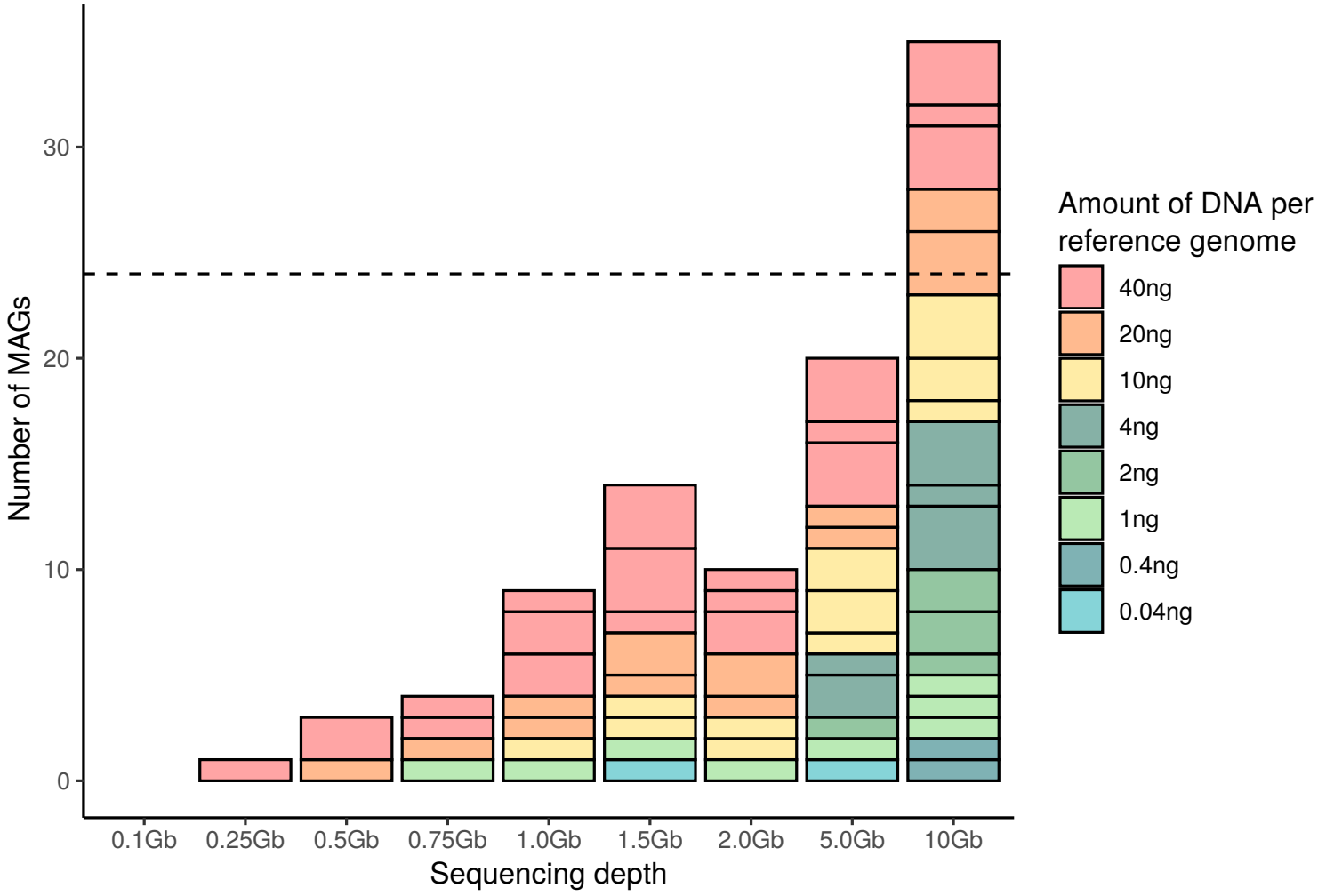

70 bacterial isolates with genome data
Mock communities
Even distribution
• 70 isolates
Staggered distribution
• 24 isolates
Shotgun sequencing (Illumina)
0.1 0.25 0.5 0.75 1.0 1.5 2.0 5.0 10 Gb/sample
0.33 0.83 1.67 2.50 3.33 5.00 6.67 16.67 33.33 x10⁶ reads
(150bp PE)
2 different distributions
9 sequencing depths
+/- background DNA
2 sequencing facilities
Taxonomy / Diversity
Reference genome based
MAG construction
• Number of strains/MAGs recovered
• Completeness
• Missclassification
• Strain resolution
Functions
A
B
• Pathway completness
• Number of genes recovered

Bacteroides acidifaciens CLA-AP-2
Bacteroides ovatus CLA-AA-H168
Bacteroides finegoldii CLA-JM-H15
Bacteroides uniformis CLA-AV-11
Bacteroides fragilis CLA-AA-H207
Phocaeicola sartorii CLA-AV-12
Phocaeicola vulgatus 4889b
Phocaeicola vulgatus HDF
Phocaeicola vulgatus 4852a
Phocaeicola vulgatus CLA-KB-H118
Hallella faecis CLA-AA-H145
Parabacteroides goldsteinii CLA-AP-3
Parabacteroides distasonis CLA-AA-H234
Odoribacter lunatus CLA-AA-M09
Alistipes muris CLA-AA-M12
Yersinia enterocolitica CLA-JM-1
Kosakonia cowanii CLA-AP-13
Escherichia coli CLA-AD-1
Escherichia coli CLA-AA-H239
Escherichia coli CR B11
Escherichia coli Ec8850
Bifidobacterium pseudocatenulatum CLA-AA-H257
Bifidobacterium longum CLA-JM-H30-B
Adlercreutzia muris CLA-AV-8
Collinsella sp nov CLA-JM-H32-B
Collinsella sp nov CLA-ER-H11
Collinsella sp nov CLA-ER-H5
Collinsella sp nov CLA-ER-H8
Faecalibaculum rodentium CLA-AV-15
Thomasclavelia ramosa CLA-JM-H52
Bacillus licheniformis CLA-AP-5
Lactococcus lactis CLA-AV-14
Streptococcus thoraltensis CLA-AV-9
Streptococcus caecimuris CLA-AV-18
Enterococcus avium S68
Enterococcus faecalis CLA-SR-172
Enterococcus gallinarum CLA-AP-1
Pumilibacter intestinalis CLA-AA-M10
Hominilimicola fabiformis CLA-AA-H232
Faecalibacterium duncaniae CLA-AA-H254
Hominenteromicrobium mulieris CLA-AA-H250
Hominimerdicola aceti CLA-AA-H220
Agathobaculum butyriciproducens CLA-AA-H270
Intestinimonas aquisgranensis CLA-AA-H199
Brotocaccenecus cirricatena CLA-AA-H272
Hominicoprocola fusiformis CLA-AA-H188
Lachnospira sp nov CLA-JM-H10
Coprococcus hominis CLA-AA-H212
Hominifimenecus microfluidus CLA-JM-H9
Anthropogastromicrobium aceti CLA-AA-H224
Clostridium fessum CLA-AA-H278
Brotaphodocola catenula CLA-AA-H274
Enterocloster clostridioformis CLA-JM-H51
Faecalimonas umbilicata CLA-JM-H57
Ruminococcus turbiniformis CLA-AA-H200
Mediterraneibacter gnavus CLA-JM-H50
Roseburia amylophila CLA-AA-H204
Brotolimicola acetigignens CLA-AA-H273
Gallintestinimicrobium propionicum CLA-AA-H176
Fusicatenibacter faecihominis CLA-AA-H277
Hominisplanchenecus murintestinalis CLA-AA-M05
Hominisplanchenecus faecis CLA-AA-H246
Oliverpabstia intestinalis CLA-AA-H241
Fusicatenibacter saccharivorans CLA-AA-H252
Blautia acetigignens CLA-AA-H161
Blautia fusiformis CLA-AA-11
Blautia wexlerae CLA-JMH-31B
Blautia wexlerae CLA-AA-H259
Veillonella intestinalis CLA-AV-13
Mucispirillum schaedleri CLA-AA-M07-B
Phylum
Actinomycetota
Bacillota
Bacteroidota
Deferribacterota
Pseudomonadota
Distribution in log SynCom
40 ng (17.2%)
20 ng (8.6 %)
10 ng (4.3 %)
4.0 ng (1.72 %)
2.0 ng (0.86%)
1.0 ng (0.43%)
0.4 ng (0.17%)
0.04 ng (0.017%)
Tree scale: 0.1

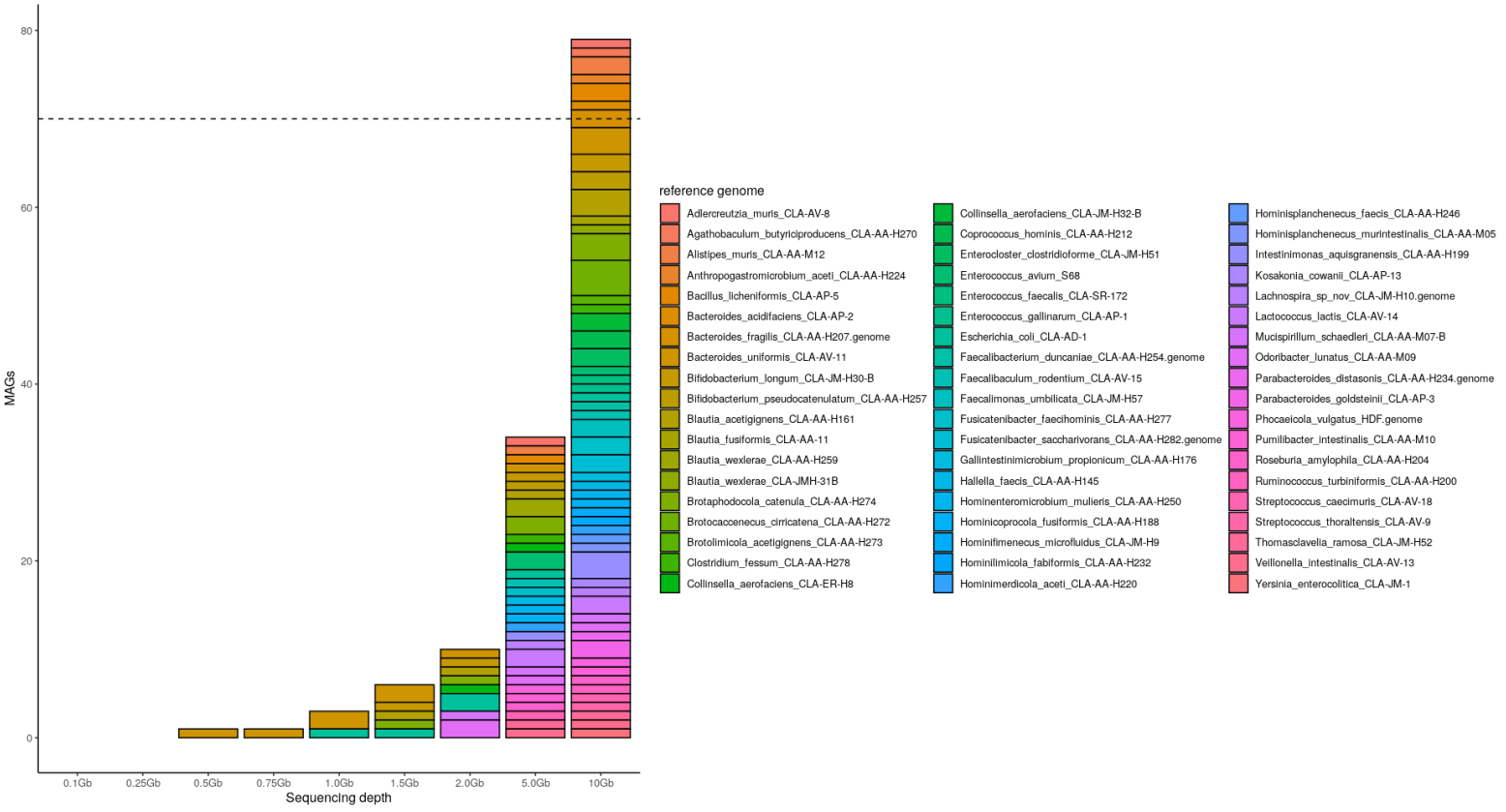

Raw reads
Reference data creation
strain 1
strain 2
strain 3
protein-coding gene prediction
(prodigal)
VPVLVEPQTPTESVEQSSPTELR
Reference Genomes
Predicted protein reference sequences
Pre-Processing
subsampling (seqtk)
Subsampled raw reads
quality filtering and phiX removal
(trimmomatic, bbmap, bbduk)
High-quality reads
assembly (MegaHit)
Contigs
Taxonomic analysis
alignment of reads to reference genomes
(coverM)
strain 1
strain 2
strain 3
Coverage of reference genomes
Relative abundance profile
Reference based strain level analysis
MAG construction
remove contigs < 1,000 bp
MAG construction
(bowtie2 + metabat2)
MAG 2
MAG 1
MAG 4
completeness and contamination check
(checkM)
taxonomic assignment
(coverM and GTDB-Tk)
Composition of MAGs
De-novo strain level analysis
Functional analysis
protein-coding gene prediction
(prodigal)
VPVLVEPQTPTESVEQSSPTELR
Predicted protein sequences (PPS)
alignment to reference protein sequences
(Diamond / blastp)
reconstrucion of functional pathways
(kofamscan, KEGG-decoder)
Functional coverage
Functional pathway completness

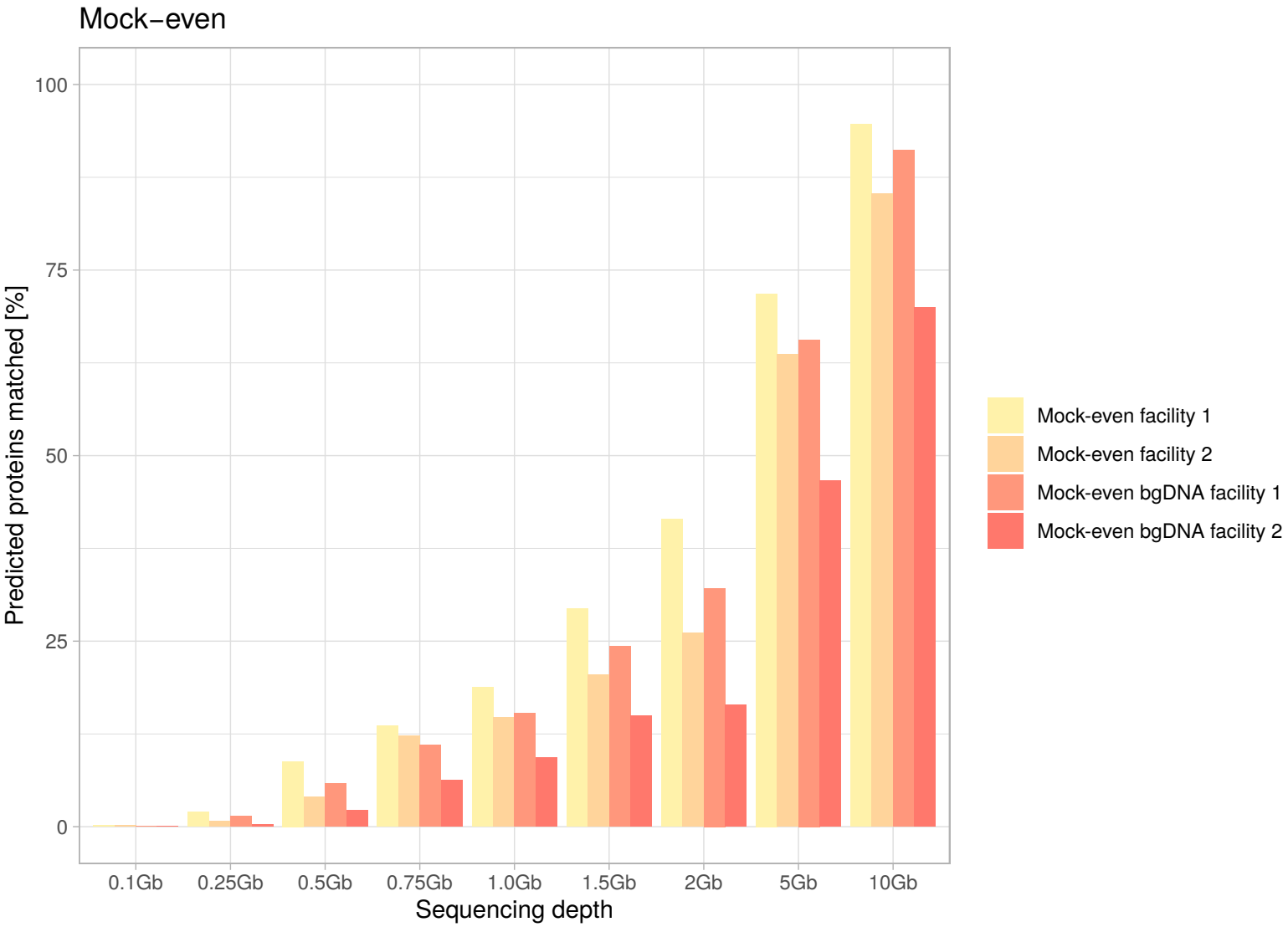

A
Reference genomes
Mock-even-facility1
Mock-even-facility2
Adlercreutzia muris CLA-AV-8
Agathobaculum butyriciproducens CLA-AA-H270
Alistipes muris CLA-AA-M12
Anthropogastromicrobium aceti CLA-AA-H224
Bacillus licheniformis CLA-AP-5
Bacteroides acidifaciens CLA-AP-2
Bacteroides finegoldii CLA-JM-H15
Bacteroides fragilis CLA-AA-H207
Bacteroides ovatus CLA-AA-H168
Bacteroides uniformis CLA-AV-11
Bifidobacterium longum CLA-JM-H30-B
Bifidobacterium pseudocatenulatum CLA-AA-H257
Blautia acetigignens CLA-AA-H161
Blautia fusiformis CLA-AA-11
Blautia wexlerae CLA-AA-H259
Blautia wexlerae CLA-JM-H31B
Brotaphodocola catenula CLA-AA-H274
Brotocaccenecus cirricatena CLA-AA-H272
Brotolimicola acetigignens CLA-AA-H273
Clostridium fessum CLA-AA-H278
Collinsella aerofaciens CLA-ER-H11
Collinsella aerofaciens CLA-ER-H5
Collinsella aerofaciens CLA-ER-H8
Collinsella aerofaciens CLA-JM-H32-B
Coprococcus hominis CLA-AA-H212
Enterocloster clostridioforme CLA-JM-H51
Enterococcus avium S68
Enterococcus faecalis CLA-SR-172
Enterococcus gallinarum CLA-AP-1
Escherichia coli CLA-AA-H239
Escherichia coli CLA-AD-1
Escherichia coli CR B11
Escherichia coli Ec8850
Faecalibacterium duncaniae CLA-AA-H254
Faecalibaculum rodentium CLA-AV-15
Faecalimonas umbilicata CLA-JM-H57
Fusicatenibacter faecihominis CLA-AA-H277
Fusicatenibacter saccharivorans CLA-AA-H252
Gallintestinimicrobium propionicum CLA-AA-H176
Hallella faecis CLA-AA-H145
Hominenteromicrobium mulieris CLA-AA-H250
Hominicoprocola fusiformis CLA-AA-H188
Hominifimenecus microfluidus CLA-JM-H9
Hominilimicola fabiformis CLA-AA-H232
Hominimerdicola aceti CLA-AA-H220
Hominisplanchenecus faecis CLA-AA-H246
Hominisplanchenecus murintestinalis CLA-AA-M05
Intestinimonas aquisgranensis CLA-AA-H199
Kosakonia cowanii CLA-AP-13
Lachnospira sp nov CLA-JM-H10
Lactococcus lactis CLA-AV-14
Mediterraneibacter gnavus CLA-JM-H50
Mucispirillum schaedleri CLA-AA-M07-B
Odoribacter lunatus CLA-AA-M09
Oliverpabstia intestinalis CLA-AA-H241
Parabacteroides distasonis CLA-AA-H234
Parabacteroides goldsteinii CLA-AP-3
Phocaeicola sartorii CLA-AV-12
Phocaeicola vulgatus 4852a
Phocaeicola vulgatus 4889b
Phocaeicola vulgatus CLA-KB-H118
Phocaeicola vulgatus HDF
Pumilibacter intestinalis CLA-AA-M10
Roseburia amylophila CLA-AA-H204
Ruminococcus turbiniformis CLA-AA-H200
Streptococcus caecimuris CLA-AV-18
Streptococcus thoraltensis CLA-AV-9
Thomasclavelia ramosa CLA-JM-H52
Veillonella intestinalis CLA-AV-13
Yersinia enterocolitica CLA-JM-1
0
5
10
15
20
relative abundance of reference genomes [%]
Sequencing depth
0.1Gb
0.5Gb
0.25Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb
B
Reference genomes
Mock-stag-facility1
Mock-stag-facility2
Hominenteromicrobium mulieris CLA-AA-H250
Hominifimenecus microfluidus CLA-JM-H9
Phocaeicola vulgatus CLA-KB-H118
Hominilimicola fabiformis CLA-AA-H232
Collinsella sp nov CLA-ER-H11
Lactococcus lactis CLA-AV-14
Parabacteroides goldsteinii CLA-AP-3
Bifidobacterium pseudocatenulatum CLA-AA-H257
Hominicoprocola fusiformis CLA-AA-H188
Veillonella intestinalis CLA-AV-13
Blautia fusiformis CLA-AA-11
Ruminococcus gnavus CLA-JM-H50
Bacillus licheniformis CLA-AP-5
Bacteroides ovatus CLA-AA-H168
Fusicatenibacter faecihominis CLA-AA-H277
Bacteroides fragilis CLA-AA-H207
Escherichia coli CLA-AA-H239
Roseburia amylophila CLA-AA-H204
Kosakonia cowanii CLA-AP-13
Faecalibacterium duncaniae CLA-AA-H254
Enterococcus faecalis CLA-SR-172
Enterocloster clostridioformis CLA-JM-H51
Phocaeicola sartorii CLA-AV-12
Thomasclavelia ramosa CLA-JM-H52
0
5
10
15
20
relative abundance of reference genomes [%]
Sequencing depth
0.1Gb
0.5Gb
0.25Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb

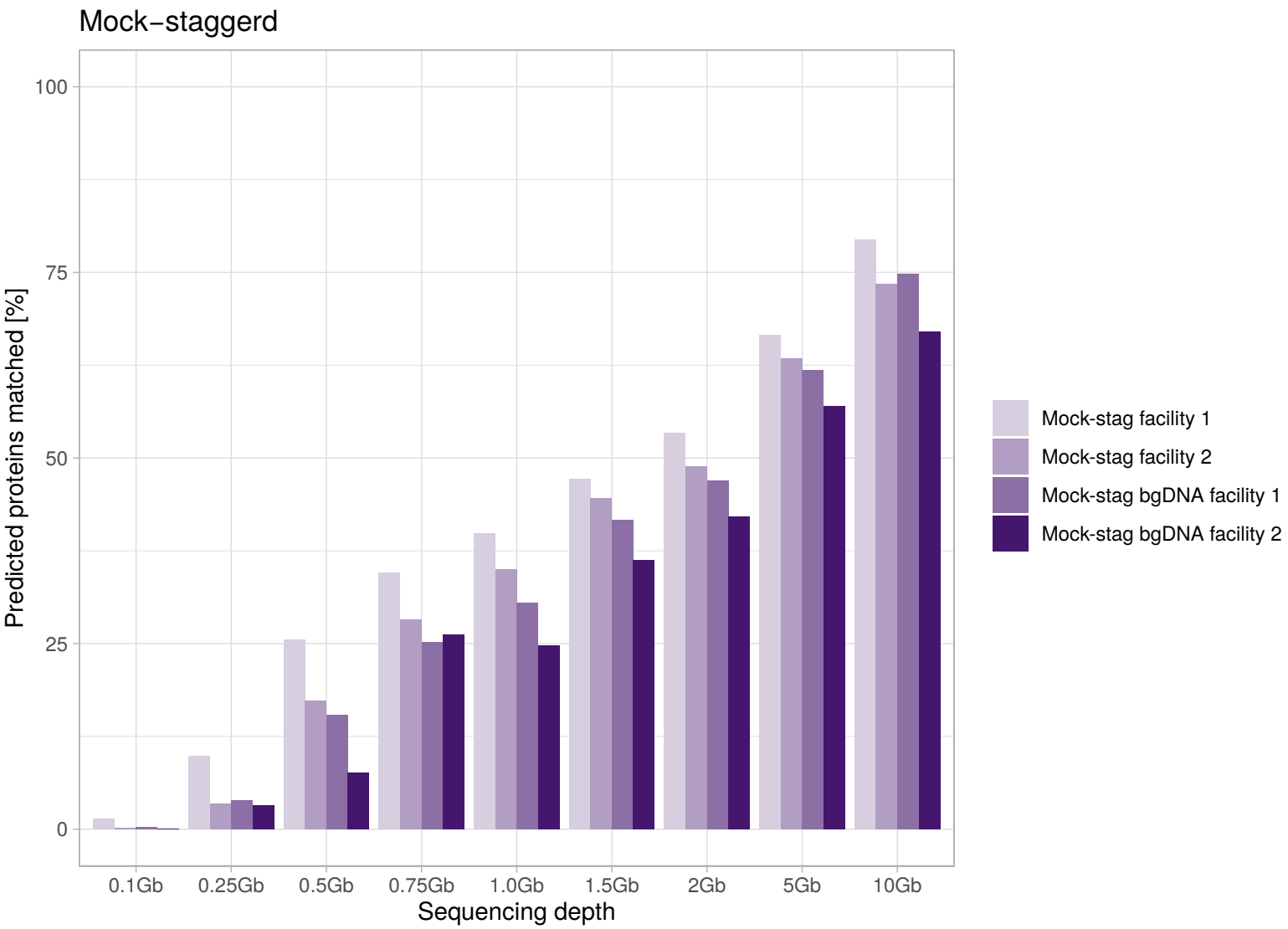
Mock−stag



Mock−even




Mock−even
Predicted proteins matched [%]
100
75
50
25
0
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb
Sequencing depth
Mock-even facility 1
Mock-even facility 2
Mock-even bgDNA facility 1
Mock-even bgDNA facility 2


Mock−staggerd
Predicted proteins matched [%]
100
75
50
25
0
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb
Sequencing depth
Mock-stag facility 1
Mock-stag facility 2
Mock-stag bgDNA facility 1
Mock-stag bgDNA facility 2