# Benchmarking of shotgun sequencing depth highlights limitations of strain-level analysis and shallow metagenomics

Nicole S. Treichel,[1,*] Charlie Pauvert,[1] Joana Séneca,[2,3] Petra Pjevac,[2,3] David Berry,[2,3] John Penders,[4,5] Thomas C. A. Hitch,[1] Thomas Clavel[1,5,*]

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

2. Joint Microbiome Facility of the Medical University of Vienna and the University of Vienna, Djerassiplatz 1, 1030, Vienna, Austria.

3. Division of Microbial Ecology, Department of Microbiology and Ecosystem Science, Centre for Microbiology and Environmental Systems Science, University of Vienna, Djerassiplatz 1, 1030, Vienna, Austria

4. 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.

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

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

**Keywords:** 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 microbial communities. We used complex mixtures of DNA from cultured gut bacteria and analysed taxonomic composition, strain-level resolution, and functional profiles at up to eleven sequencing depths (0.1-50.0 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), and even high-quality MAGs were chimeric, with

54.5 to 81.8 % accurately representing the original strains, depending on the bioinformatic approach used. However, the issue of chimeric MAGs can be reduced by using strain-aware assembly methods or long-read sequencing. Functionally, 2 Gb provided reliable insights at the pathway level, but sufficient proteome coverage was only achieved at or above 10 Gb. Library preparation and host DNA contamination 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 accurately 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, the widespread use of this technology contrasts with the low number of studies that have tested the robustness and accuracy of conclusions derived from widely used analysis workflows.

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 generally 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, commonly referred to 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 information is accessible at what sequencing depth, and 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 exact 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[3,6,9]. However, these approaches do not include the impact of real sample processing and sequencing in the laboratory.

Given the need to better define the strengths and limitations of shallow metagenomics, we systematically evaluated the effects of shotgun sequencing depth on composition, strain-level diversity, and functional readouts (**Supplementary Fig. S1**). For this, we used 13 complex mixtures of bacterial DNA: an even distribution with 70 bacterial strains (Mock-even-70), a staggered distribution with 24 strains (Mock-stag-24), including 10 additional versions with different compositions (Mock-stag-24 v1-10), or a staggered distribution with 70 strains (Mock-stag-70) (**Supplementary Table S1**, **Supplementary Fig. S2**). As wet-lab factors can cofound results, DNA libraries were prepared in two laboratories and in the presence/absence of background DNA isolated from the gut of germfree mice.

## Methods

### DNA extraction and preparation of the Mock communities

For the creation of the mock communities (Mock-even-70, Mock-stag-24, Mock-stag-24 v1-10 for multi-coverage binning, and Mock-stag-70), DNA was obtained from isolates in our collections of mouse and human gut bacteria [10,11] (**Supplementary Fig. S2)**. DNA was extracted from freshly grown strains revived from frozen glycerol stocks based on the method of Godon et al. 1997 [12] modified as described in Afrizal et al [10]. DNA concentration was measured using a Qubit fluorometer (Thermo Fisher Scientist). The DNA of all 70 isolates was pooled in either equimolar amounts (Mock-even-70), or in a staggered distribution (Mock-stag-70). For 24 selected isolates, the DNA was pooled in a staggered distribution (Mock-stag-24). For testing multi-coverage binning to generate MAGs, ten additional mock communities with 24 DNA extracts in 10 different distribution patterns were created (Mock-stag-24 v1-10). The distribution of DNA from the different isolates in each Mock community is provided in **Supplementary Table S1**.

Non-bacterial background DNA was isolated, as described above, using gut content collected from germfree mice (LAVE ethical approval 81-02.04.2023.A253) in accordance with the German Animal Protection Law (TierSchG). All mice were fed a standard chow *ad libitum* (γ-irradiated ssniff Spezialdiäten ref. V1124-927). The germfree status of the mice was confirmed by microscopic observation after Gram-staining and by cultivation on both anaerobic and aerobic agar plates. Mice

96 were culled and caecal content was collected and stored immediately at -80°C. After DNA extraction,
97 the background DNA was mixed 1:1 (v/v) with Mock-even-70 and Mock-stag-24.

98

99 **Library preparation and short-read sequencing**

100 In sequencing facility 1 (UKA), DNA libraries of the mock communities were prepared using the
101 NEBNext Ultra II FS DNA Library Prep Kit for Illumina (NEB, USA) according to the manufacturer's
102 instructions, using 100 ng (Mock-even-70 and Mock-stag-24), 200ng (Mock-stag-70) or 40 ng (Mock-
103 stag-24 v1-10 for multi coverage binning) of input DNA and an automated platform (Beckman Coulter,
104 USA). Enzymatic shearing to approximate 250 bp was performed for 30 minutes. Adaptor-ligated DNA
105 was enriched using PCR (Mock-even-70, Mock-stag-24, Mock-stag-70: 5 cycles; Mock-stag-24- v1-10
106 multi-coverage binning: 7 cycles) and NEBNext Multiplex Oligos for Illumina (NEB, USA) for unique dual
107 barcoding. AMPure beads (Beckman Coulter, USA) were used for size selection and clean-up of adaptor
108 ligated DNA.

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

113 For cleaned DNA library from both facilities fragment size (~320 bp library fragment size ≙ ~200 bp
114 genomic DNA insert size) was determined on an Agilent D1000 Tapestation (Bioanalyzer System,
115 Agilent Technologies, USA) using High Sensitivity D1000 screentapes.

116 Quality check (Bioanalyzer System, Agilent Technologies, USA), DNA quantification (Quantus, Promega,
117 USA), and sequencing of the resulting libraries were conducted at the IZKF Core Facility Genomics (UKA,
118 RWTH Aachen University). The libraries for Mock-even-70, Mock-stag-70 and Mock-stag-24 were
119 pooled to reflect the nine or eleven different sequencing depths targeted, and sequenced on a
120 NovaSeq6000 (Illumina, USA) with NovaSeq 6000 Reagents v1.5 (2x150 cycles). Mock-stag-24 v1-10
121 for multi-coverage binning were sequenced at 10 Gb with the same chemistry at the NGS Competence
122 Center Tübingen (NCCT).

123

124 **Long-read sequencing**

125 The Mock-stag-70 DNA (200 fmol) was prepared for long read sequencing with the SQK-LSK114 kit
126 (Oxford Nanopore Technologies), following the manufacturer's protocol and the NEBNext Companion

127 module v2 (New England Biomedicals). About 80 fmol of a >20Kb library was loaded on a R10.4.1
128 flowcell (FLO-PRO114, Oxford Nanopore Technologies) and sequenced for 72h on a Promethion P2
129 solo (Oxford Nanopore Technologies) using Minknow v. 25.05.12. The flowcell was washed using the
130 EXP-WSH004 kit and leftover library was re-loaded and left to sequence until the flowcell end of life.
131 Reads were base called using Dorado (v. 1.0.0) using super accuracy mode (model:
132 r1041_e82_400bps_sup_v5.2.0). Long read sequencing was performed at the Joint Microbiome
133 Facility of the Medical University of Vienna and the University of Vienna under project ID JMF-2507-
134 23.

135

136 **Bioinformatic analysis**

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

139

140 **Reference genomes**

141 The genomes of all strains used in the Mock communities are hereon referred to as "reference
142 genomes". They have been deposited in a public repository and published previously[10,11]. A
143 phylogenetic tree was constructed from the genomes of the 70 isolates using PhyloPhlAn[13] v3.0.67
144 (options: *--diversity medium –f supermatrix_aa.cfg*) (**Supplementary Fig. S2**). Genome characteristics,
145 including size and GC content, were analysed using Biopython[14] v.1.79 and bioawk[15] v1.0
146 (**Supplementary Table S2**).

147

148 **Pre-processing of shotgun metagenomic data**

149 Raw reads of samples with the different empiric sequencing depths (**Supplementary Table S3**) were
150 further subsampled bioinformatically to the exact targeted sequencing depth (below). The raw FASTQ
151 files were subsampled using seqtk[16] v1.2 with default settings. The targeted number of paired-end
152 reads (2 x 150 bp) per sequencing depth was as follows: 333,333 read pairs (0.10 Gb), 833,333 (0.25
153 Gb), 1,666,667 (0.5 Gb), 2,500,000 (0.75 Gb), 3,333,333 (1.0 Gb), 5,000,000 (1.5 Gb), 6,666,667 (2.0
154 Gb), 16,666,667 (5.0 Gb), 33,333,333 (10.0 Gb). Adapters were removed and subsampled raw reads
155 were quality-filtered with Trimmomatic[17] v.0.39 (options: *TRAILING:3 LEADING:3*
156 *SLIDINGWINDOW:5:20 MINLEN:50 ILLUMINACLIP:{adapters.fa}:2:30:10*). The bbduk command
157 (options: *hdist=1 k=31*) in BBMap[18] v.38.84 was used for removal of phiX sequences. Quality-filtered

158 reads of all Mock communities were assembled into contigs using MEGAHIT[19] v1.2.9. Additionally, the
159 quality-filtered reads of Mock-stag-70 were assembled into contigs with metaSPAdes v.4.2.0[20]. Contigs
160 of metagenomes sequenced at 10 Gb, 20 Gb and 50 Gb assembled with either of the two assemblers
161 were aligned to the reference genomes with blastn (perc_identity, 97%; evalue, 1e-10; alignment
162 length, >150 bp). In addition to the 3 mock communities sequenced separately at all sequencing
163 depths, and to the 10 Mock-stag-24 communities with different compositions (v1-10) sequenced at 10
164 GB, a further validation step included in-silico subsampling of the 50 Gb Mock-stag-70 to 10 sequencing
165 depth (0.1 Gb – 20 Gb), each ten times, using seqtk[16] v1.2 (default settings with 10 different seeds).

166

167 **Taxonomic coverage**

168 Coverage of the reference genomes by the quality-filtered metagenomic reads was determined using
169 coverM[21] v0.6.1 (options: *coverm genome --mapper bwa-mem --methods covered_fraction --min-*
170 *covered-fraction 0 --coupled*). The relative abundance profiles were calculated using the read count
171 option of coverM (options: *coverm genome --methods count --min-covered-fraction 0 --coupled*), which
172 were converted to relative abundance for each sample.

173 Non-supervised taxonomic profiles were generated using MetaPhlAn v4.0.1[22] with the
174 mpa_vJun23_CHOCOPhlAnSGB_202403 database.

175 Average nucleotide identity (ANI) values between the reference genomes of four *Escherichia coli*
176 strains and four *Phocaeicola vulgatus* strains were calculated using FastANI[23] v1.34.

177

178 **Functional analysis**

179 Protein-coding genes were predicted in the pooled reference genomes and the assembled contigs of
180 the Mock samples (individually) using prodigal [24] v2.6.3.

181 Gene functions were annotated using KEGG Orthology and KofamScan[25] v1.3.0. Pathway completeness
182 was assessed by KEGG-Decoder[26] v1.3, as percentage of KOs covered, which are included in the
183 manually curated canonical pathways used by KEGG-Decoder.

184 To determine the ability of each sequencing depth to capture the protein-encoding potential in the
185 Mock communities, the genes predicted in the reference genomes were used to create a Diamond[27]
186 (v2.0.15) protein sequence database, with which the protein sequences predicted in the assemblies
187 (one assembly for each of the sequencing depths) were compared (options: *diamond blastp –sensitive*
188 *–query-cover 80 –id 90*).

189

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

191 Contigs <1,000 bp were removed prior to reconstructing MAGs. An index table was built from the size-
192 filtered contigs using bowtie2[28] v2.5.1 (*bowtie2-build*) with default options. The decontaminated
193 paired-end reads were aligned to the bowtie index of size-filtered contigs (*bowtie2 –S - --very-sensitive-*
194 *local –no-unal –p 30*). Bam files were sorted with samtools[29] v1.17 (*samtools view –bS).* Contigs were
195 binned using Metabat2[30] v2.12.1 and its algorithm for calculating coverage of each sequence in the
196 assembly (*jgi_summarize_bam_contig_depths*) before creating the bins (*metabat2 -m 1500 --maxP 95*
197 *--minS 60 --maxEdges 200 --unbinned --seed 0*).

198 The quality of the resulting bins was evaluated with checkM[31] v1.1.3 using the lineage workflow.
199 Accordingly, the MAGs were categorised as being of high quality (>90% completeness, <5%
200 contamination), medium quality (>70% completeness, <10% contamination), or low quality (all that
201 did not fulfil the previous criteria). They were then taxonomically classified using GTDB-Tk[32] v2.3.2
202 (*gtdbtk classify_wf*) with the Genome Taxonomy Database r207. The coverage of the reference
203 genomes by the different MAGs was determined using coverM[21] v0.6.1 (*coverm genome --mapper*
204 *bwa-mem --methods covered_fraction --min-covered-fraction 0 –single; no multiple read mapping*).
205 Reference genomes were aligned to MAGs by blastn v2.13.0 *(-evalue 1e-10 -perc_identity 90.0*) and
206 circular alignments plotted using shinyCircos[33].

207 For comparing sample-specific (approach above) to multi-coverage binning for the ability to
208 reconstruct high-quality, non-chimeric MAGs (i.e. one genome match only), the workflow of Mattock
209 and Watson[34] was used on Mock-stag-24 v1-10, sequenced at a depth of 10 Gb each.

210

211 **Long-read sequence analysis**

212 The raw reads were filtered to a minimum average read quality of Q20 and a minimum length of
213 1,000 bp using chopper (v0.10.0)[35]. The filtered reads were used to randomly subset 11 additional
214 datasets corresponding 0.1, 0.25, 0.5, 0.75, 1, 1.5,2, 5, 10, 20, and 50 Gb using rasusa (v2.1.1)[36]. Each
215 subset was subsequently assembled using flye (v2.9.5)[37] with "–nano-hq" and polished once with
216 medaka (v2.1.0, github.com/nanoporetech/medaka) using the –bacteria flag. Contigs <1,000bp were
217 removed. Assemblies were quality checked using QUAST[38] v5.3.0 and metagenomic binning was
218 performed using metabat2[30]. Bin coverage was estimated using the
219 *jgi_summarize_bam_contig_depths* function with a minimum percent identity of 90. Quality and
220 chimeric nature of MAGs was assessed as described above for the short-read data.

221

222 **Statistics and Plotting**

223 All statistical tests were performed in R[39] v4.4.1. using RStudio v2023.03.0+386 and the packages
224 vegan[40] , reshape[41], tidyverse[42] and dplyr[43]. For creation of the graphs, the R package ggplot2[44] v3.5.1,
225 ggpubr[45] v0.6.0, ggbreak[46] and ggfortify[47] v0.4.17 were used. The specific statistical tests used in each
226 analysis are stated in the results.

227

228 **Results**

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

230 We first investigated the impact of sequencing depth on taxonomic composition. For the three main
231 Mock communities (Mock-even-70, Mock-stag-24, Mock-stag-70), reads were detected for all
232 reference genomes already at 0.1 Gb. To obtain an overview of genome coverage, completeness
233 categories were compared between the sequencing depths. At 0.1 Gb, most of the genomes (63 to
234 91 %) had a low coverage (0-25 %) (**Fig. 1A**). Genome coverage increased substantially until 5 Gb of
235 sequencing, with a clear effect of mock community complexity. Most genomes reached >90 %
236 coverage at 5 Gb in Mock-even-70 and Mock-stag-24. However, in Mock-stag-70, coverage continued
237 to increase gradually from 5 Gb (36 % genomes at >90 % coverage) to 50 Gb (64 %), but 11 genomes
238 (16 %) added with 0.001 ng (0.00046‰) to 0.1 ng (0.0046 %) had still low coverage (0-25 %) at the
239 highest sequencing depth (50 Gb).

240 We then looked at the individual strains across the sequencing depths (**Supplementary Fig. S4**). In
241 Mock-even-70, the lowest coverage at 0.1 Gb was 3.34 % for *Phocaeicola sartorii* CLA-AV-12 and the
242 highest 58.7 % for *Bacteroides uniformis* CLA-AV-11, which reached 99 % coverage already at 0.75 Gb.
243 Full genome coverage (100 %) was only achieved at 10 Gb for V*eillonella intestinalis* CLA-AV-13. In
244 Mock-stag-24 and Mock-stag-70, higher coverage was associated with increasing DNA concentration.
245 The reference genomes with lowest concentration in Mock-stag-24 (0.01 %, 0.04 ng) had a minimum
246 coverage of 0.53 % at 0.1 Gb and a maximum coverage of 34.21 % at 10 Gb. In Mock-stag-70, the lowly
247 abundant genomes - 0.1ng (0.0046%) to 0.001 ng (0.00046‰) - reached a coverage of 0.8 % to 21.8 %
248 at highest sequencing depth (50 Gb), whilst the genomes of the five strains with highest DNA amounts
249 (>4.6 %) reached between 40.50 % and 98.25 % coverage already at 0.1 Gb.

250 Next, we assessed the relative abundance of strains and compared them to their theoretical values.
251 For Mock-even-70, relative abundances oscillated around the expected value of 1.43 % for most
252 species, without any significant effects of the sequencing depth on the overall relative abundance

253 profile (p-value = 0.78; Kruskall-Wallis test) (**Supplementary Fig. S5A**). Six species showed values >2.5
254 % for several sequencing depths, with highest values of 10-15 % for *B. uniformis* CLA-AV-11. In addition,
255 marked differences due to sequencing depth were detected in the case of *Collinsella* sp. CLA-ER-H5.
256 Sequencing depth had also no influence on the overall relative abundance profile in the case of Mock-
257 stag-24 and Mock-stag-70 (p-value = 1 in both cases; Kruskall-Wallis) (**Supplementary Fig. S5B**). In
258 addition to sequencing the mock communities separately, Mock-stag-70 sequenced at 50 Gb was
259 subsampled multiple times (10 replicates) bioinformatically for each sequencing depth ≤20 Gb to test
260 variability. Variations in the relative abundance of individual species was low in general (average
261 coefficients of variation <5 %), with increasing variability at lower sequencing depths (**Fig. 1B**; statistics
262 in **Supplementary Table S5**).

263 As reference genomes are not available for metagenomic analysis in most studies, the mock
264 communities were also taxonomically analysed using the commonly used profiler MetaPhlAn[22], hereon
265 referred to as non-supervised approach. Overall, the taxonomic assignment in this approach was less
266 sensitive, *i.e.*, fewer species were detected. In Mock-stag-24, one strain with low concentration (0.1 %,
267 0.4 ng, *Hominilimicola fabiformis* CLA-AA-H232) was not detected (**Supplementary Fig. S6**). In Mock-
268 even-70, 59 species could be assigned with MetaPhlAn4, including 47 matching a reference genome at
269 species level. In Mock-stag-70, 53 taxa were detected, 44 with a species-level match (**Supplementary**
270 **Fig. S6**). The non-supervised taxonomic assignment showed also higher variance from the targeted
271 relative abundances, especially at lower sequencing depths and with increased complexity of the mock
272 communities (**Fig. 1 C**; statistics in **Supplementary Table S5**).

273 In summary, reference-based detection of strains is possible with sequencing as little as 0.5 Gb, but
274 high coverage of reference genomes needs more data (>5 Gb) under the conditions tested. Relative
275 abundance profiles were not substantially affected by sequencing depths for most of the strains. In
276 contrast, non-supervised taxonomic assignment was less sensitive, with fewer species detected.

277

278 **Strain-level analysis: *De-novo* MAG reconstruction requires deep sequencing and generates**
279 **chimeras**

280 The ability of shallow metagenomics to resolve strain-level variation was investigated using four strains
281 of *E. coli* and *P. vulgatus* within the Mock-even-70 dataset (Supplementary **Fig. S2**). Coverage of their
282 reference genomes per sequencing depth, as determined during assignment of all reference genomes,
283 revealed that all strains were detected already at 0.1 Gb (Supplementary **Fig. S7A**). All four reference
284 genomes of both species reached >75 % coverage at 5 Gb and >98 % at 10 Gb. All *P. vulgatus* strains
285 showed a similar increase in coverage with increasing sequencing depth, mirroring ANI values between

286 the strains that were all close to one another (Supplementary **Fig. S7B**). In contrast, for the *E. coli*
287 strains, the reference genomes sharing the highest ANI value (*E. coli* Ec8850 and CR B11; ANI = 99.97 %)
288 showed a congruent coverage increase (Supplementary **Fig. S7B**; their coverage increase curves
289 overlaid), whilst *E. coli* CLA-AD-1, characterized by an ANI value more distant to the other strains (<
290 97 %), reached >99 % reference genome coverage already at the 2 Gb sequencing depth
291 (Supplementary **Fig. S7B**). Only one strain of *E. coli* and *P. vulgatus* was detected in Mock-even-70 and
292 Mock-stag-70 using MetaPhlAn4 at any of the sequencing depths (Supplementary **Figure S6**).

293 Reference datasets are rarely available during metagenomic analysis, hence *de-novo* reconstruction of
294 genomes is widely used to facilitate strain-level analysis. MAGs were thus assembled separately for
295 each metagenomic sample. Overall, the number of MAGs recovered increased with higher sequencing
296 depth for all mock communities (**Fig. 2A,** Supplementary **Fig. S8, left panels**). More MAGs than the
297 actual number of reference genomes (dashed line) were formed at the highest sequencing depths,
298 whilst not all reference genomes were represented by MAGs (**Fig. 2A,** Supplementary **Fig. S8, right**
299 **panels)**. The number of MAGs formed per reference genome increased with the amount of data
300 sequenced, with ≥2 MAGs observed for 22, 12 and 29 of the reference genomes across all sequencing
301 depths in Mock-even-70, Mock-stag-24, and Mock-stag-70, respectively (**Supplementary Fig. S9**). This
302 suggests that multiple MAGs are created per reference genome with greater sequencing depth, rather
303 than coalescing into a single high-quality MAG.

304 Given the unexpected number of MAGs and their multiplicity per genome, we next determined their
305 origin. MAGs were grouped by quality and their contigs were assigned to the reference genomes to
306 determine which isolates they were reconstructed from, with a threshold of >0.25 % coverage of a
307 reference genome. In Mock-even-70 at the highest sequencing depth (10 Gb), 10 of the 23 high-quality
308 MAGs (hqMAGs; completeness >90 %, contamination <5 %) were assigned to multiple genomes (red
309 bars) (**Fig. 2B, top**). The proportion of MAGs assigned to only one (blue) vs. multiple (red) genomes
310 decreased further with decreasing quality of the MAGs. Taxonomic assignment of hqMAGs to GTDB
311 showed a result congruent with the assignment to our own reference genomes (**Supplementary Table**
312 **S4**). Although four *E. coli* strains were included in Mock-even-70, only one MAG assigned to *E. coli* CLA-
313 AD-1 was reconstructed from each sequencing dataset > 1Gb, except 2 Gb (a second *E. coli* MAG of
314 low completeness and quality was obtained). In the case of *P. vulgatus* (also represented by four
315 strains), one MAG was found at 5 Gb and one at 10 Gb, both primarily assigned to *P. vulgatus* HDF,
316 indicating strain delineation was not possible with the method used (**Supplementary Table S4**).

317 In Mock-stag-24, a single hqMAG was assembled at 0.75 and 1 Gb (**Fig. 2B, bottom**). Most sequences
318 of this MAG were assigned to a genome with highest DNA concentration (40 ng, *T. ramosa* CLA-JM-
319 H52, ≥95 % genome coverage). With increasing sequencing depth, some hqMAGs assigned to genomes

320 with lower concentrations started to be reconstructed (**Supplementary Table S4**). Surprisingly, a
321 higher proportion of single-origin MAGs occurred with decreasing quality, while at 10 Gb half of the
322 hqMAGs in Mock-stag-24 were chimeric.

323 To illustrate the chimeric nature of MAGs, three exemplary hqMAGs (Mock-even-70, 10 Gb) were
324 aligned to the reference genomes they covered >0.25 % using blastn (**Fig 2C**). MAG 54 was a best-case
325 scenario with only one reference genome (*V. intestinalis* CLA−AV−13) matched with >98 % coverage.
326 MAG 16 showed fragments of a second reference genome. MAG 40 represented a highly chimeric
327 hqMAG, including 12 reference genomes (of which 10 are shown, ranked by decreasing coverage from
328 outer to inner circle). The mean/maximum number of reference genomes per hqMAG at 10 Gb was
329 2.3/12, 1.5/2 and 2.1/5 for Mock-even-70, Mock-stag-24 and Mock-stag-70, respectively
330 (**Supplementary Table S3**).

331 As multi-coverage binning was shown to increase the number and quality of MAGs[34], we sequenced
332 (10 Gb) ten additional Mock-stag-24 communities (v1-10) with varying reference genome distribution
333 (Supplementary **Table S1**). In comparison to single-coverage binning, significantly more hqMAGs per
334 mock community were assigned to only one reference genome using multi-coverage binning (**Fig. 2D,**
335 blue dots). Despite this improvement, chimeric MAGs were still detected in half of the communities
336 (**Fig. 2D,** orange dots).

337 To further assess ways to enhance MAG reconstruction, the most complex community (Mock-stag-70)
338 was used to test the effects of assemblers (MEGAHIT – as in all analyses above – Vs. metaSPAdes) and
339 the sequencing technology (short-read by Illumina- Vs. long-read by Nanopore-sequencing).
340 MetaSPAdes assemblies generated more hqMAGs assigned to only on reference genome, especially
341 ≥10 Gb (+1 at 10 Gb, +10 at 20 Gb, +5 at 50 Gb) (**Fig. 2E**). This was confirmed statistically by binning
342 MAGs from assemblies generated with either of the two methods (MEGAHIT Vs. MetaSPAdes) using
343 10 datasets at 10 Gb subsampled in silico from 50 Gb (**Fig. 2F**). Long-read sequencing provided even
344 more hqMAGs with a single reference match (14 of 15 at 10 Gb; 18 of 22 at 50 Gb) (**Fig. 2G**).

345 To know whether chimeric sequences are generated already during the assembly process, contigs were
346 assembled from the 10 Gb, 20 Gb, and 50 Gb data using either MEGAHIT or metaSPAdes and were
347 aligned to the reference genomes using blastn (Supplementary **Fig. S10**). With both assemblers, the
348 majority of the contigs were assigned to only one reference species (average of three sequencing
349 depths: MEGAHIT, 91.1 ± 1.6 %; metaSPAdes, 93.5 ± 1.0 %), suggesting that 6.5 to 8.9 % misassembled
350 contigs contributed to the reconstruction of chimeric MAGs.

351 In summary, when high-quality reference datasets are available, read mapping facilitates strain-level
352 analysis of shallow metagenomic data. *De-novo* analysis based on MAG reconstruction leads to

353 contaminated genomes (chimeras), even at supposed high-quality. Chimerism stems from both the
354 assembly and binning process, and was partially improved by multi-coverage binning for short reads,
355 strain-aware assembly, or long-read sequencing. These findings suggest that strain-level diversity
356 within MAG catalogues is artificially inflated due to chimerism. However, this issue can be reduced by
357 use of strain-aware assembly methods (metaSPAdes) and will be less common once long-read
358 sequencing of metagenomes becomes commonplace.

359

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

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

364 First, we investigated the impact of sequencing depth on pathway completeness, measured as
365 percentage of detected KOs involved in a pathway. Only those pathways occurring in the reference
366 genomes of the respective mock community were considered (Supplementary **Fig. S11**). Of the 178
367 pathways included in the analysis by KEGG-Decoder[26], predicted proteins for 121 (Mock-even-70,
368 Mock-stag-70) and 118 (Mock-stag-24) pathways were found in the reference genomes. A maximum
369 of 80 pathways were complete at 10 Gb for Mock-even-70, 77 for Mock-stag-24, and 81 for Mock-stag-
370 70 with the two assembly methods tested (Supplementary **Fig. S11**). One pathway was not detected
371 in the mock communities with staggered distributions despite being present in the reference genomes,
372 and 41, 40 and 39 pathways remained incomplete in Mock-even-70, Mock-stag-24 and Mock-stag-70
373 respectively. The average KEGG pathway completeness increased with sequencing depth (**Fig. 3A**).
374 Pathway completeness reached >80 % already at 2 Gb for Mock-stag-70 and at 5 Gb for the other two
375 mock communities. It then plateaued, with only a slight increase at maximal sequencing depth (Mock-
376 even-70, +7.79 %; Mock-stag-24, +3.69 %; Mock-stag-70, +3.95 %).

377 To account for functionally unassigned proteins, the protein sequences detected at each sequencing
378 depth were compared to the reference genomes protein repertoire (Mock-70: 227,543 protein
379 sequences; Mock-stag-24: 82,992). The coverage of the predicted protein sequences in the reference
380 genomes within a metagenome continued to increase up to 94.67 % for Mock-even-70 (+ 22.9 %
381 between 5 and 10 Gb), but started to plateau at 5 Gb sequencing depth for the less diverse Mock-stag-
382 24 (+ 12.9 % between 5 and 10 Gb) (**Fig. 3B**). In Mock-stag-70, the protein sequence coverage reached
383 55.5 % (MEGAHIT) and 58.3 % (metaSPAdes) at 10 Gb and increased only to 71.8 % (MEGAHIT) and
384 73.8 % (metaSPAdes) at highly deep sequencing (50 Gb). The protein sequence coverage obtained with
385 the three mock communities sequenced separately was confirmed statistically using ten Mock-stag-70
386 subsampled bioinformatically for each sequencing depth (0.1 – 20 Gb) from the 50 Gb dataset: the

387 recovered functions increased with sequencing step and variations between the replicates was low
388 (**Fig. 3C**) (p-values are provided in **Supplementary Table S5**).

389 In summary, while a moderate sequencing depth of 5 Gb was sufficient for functional analysis at the
390 pathway level for all mock communities, analysis at the level of single protein sequences requires
391 greater sequencing depth depending on the diversity and distribution of the microbial community
392 under investigation.

393

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

395 To account for the influence of wet-lab factors on shallow metagenomics results, the DNA libraries
396 were prepared in two facilities using different protocols (see methods). In addition, the two community
397 types (Mock-even-70 and Mock-stag-24) were spiked or not with background DNA isolated from the
398 gut content of germfree mice to simulate host DNA.

399 PCA plots based on relative abundance profiles showed distinct clustering due to both factors
400 (background DNA, facility), with more pronounced effects along PC1 linked to differences in library
401 preparation protocols between facilities (**Fig. 4A**). Within a condition (background DNA /facility pair),
402 the extreme sequencing depths (0.1 and 10 Gb) tended to be most distant from each other. The
403 relative abundance profiles of Mock-stag-24 prepared in facility 1, which used more template DNA
404 (100 vs. 1 ng) and a lower number of PCR cycles (5 vs. 12), were less sensitive to the effects of
405 sequencing depth when background DNA was present (**Fig. 4A, lower panel**).

406 The effect of wet-lab factors on reference genome coverage was then assessed by calculating delta-
407 values between the two facilities or the presence/absence of background DNA. For Mock-even-70,
408 consistent profiles were observed across nearly all species at the 5 and 10 Gb sequencing depths,
409 regardless of facility or background DNA (**Fig. 4B**). In contrast, variations in reference genome coverage
410 of up to 39 % for facilities, and up to 13 % for background DNA were observed at sequencing depths
411 ≤2 Gb. For Mock-stag-24, the variation between samples from the two facilities was additionally
412 influenced by the initial DNA amount indicated as coloured boxes in **Fig. 4C.** With a reference genome
413 input >1 ng and a sequencing depth >1 Gb, the total variance in reference genome coverage between
414 the two facilities dropped below 3.69%. A similar trend was observed for the impact of background
415 DNA on the coverage of reference genomes in Mock-stag-24, but with a more pronounced influence
416 of the amount of DNA. The effect of background DNA was negligible (total variance 1.33 %) with ≥10 ng
417 template DNA and a sequencing depth ≥1.5 Gb. For a DNA amount >1 ng per strain in Mock-stag, a
418 sequencing depth of 5 Gb or higher was necessary (total variance 3.23 %).

419 Analysed per strain, the relative abundance profiles in Mock-even-70 were consistent between the
420 two library preparation strategies, except for marked differences for *Collinsella* sp. CLA-ER-H5 due to
421 sequencing depth in facility 1 (**Supplementary Fig. S5A**). This indicates that the differences shown in
422 Figure 4A are due to changes in the relative abundance of a few strains. For Mock-stag-24, deviation
423 from the theoretical relative abundance values was different between the two facilities
424 (**Supplementary Fig. S5B**). The relative abundance of *E. clostridioformis*, for example, matched the
425 expected relative abundance in facility 1, while it had the highest relative abundance in facility 2. The
426 most abundant species in facility 1 was *T. ramosa*. This indicates that library preparation (specifically
427 the amount of template DNA and number of PCR cycles) influenced relative abundance profiles.
428 (**Supplementary Fig. S5B**)

429 We further analysed the effects of the wet-lab factors on the predicted protein profiles. Overall, facility
430 1 provided a more complete functional profile for both mock communities (**Supplementary Fig. S12**).
431 The addition of background DNA was associated with a decrease in coverage, which was attenuated
432 at high sequencing depths in both mock communities prepared in facility 1. In general, the influence
433 of background DNA was more pronounced with the protocol used in facility 2.

434 In summary, library preparation with a higher amount of template DNA and fewer PCR cycles is
435 recommended for more robust taxonomic and functional profiles. If necessary, a higher sequencing
436 depth can compensate for the effects of lower DNA amount or the presence of non-target DNA.

437

438 **Discussion**

439 In this work we investigated the potential and limitations of shallow metagenomic sequencing using
440 complex mock communities and sequencing depths from 0.1 Gb (0.3 million reads) to 50 Gb (166.7
441 million reads), with a specific focus on strain-level and functional analysis.

442 Previous studies reported that taxonomic profiles based on shallow shotgun sequencing (0.5-4 million
443 reads) can reflect those obtained at a higher sequencing depths[2,3,6,8]. Most of those studies relied on
444 deep sequencing of native samples then subsampled *in-silico*. Using defined input, we have confirmed
445 these results, showing that shallow metagenomics is suitable for reference-based taxonomic analysis:
446 reads for all strains were detected already at 0.1 Gb (0.3 million reads), even for lowly abundant taxa
447 in the staggered communities, enabling prevalence and relative abundance analyses despite overall
448 low genome coverage. This is in line with the study of Hillmann et al. and Xu et al. who reported that
449 0.5 million sequences (= 0.15 Gb) and 0.5 Gb, respectively, was sufficient for taxonomic analysis of the
450 human gut microbiome[2,3]. As most of the metagenomic costs is due to sequencing, being able to pool
451 more samples on the same run reduces the expenses, making shallow metagenomics at 0.5 Gb/sample

452 attractive for diversity and composition profiling in studies with large number of samples. However,
453 accurate analysis necessitates a high-quality, reference database of individual genomes.

454 Metagenomics is useful for high resolution taxonomic analysis, which is limited with 16S rRNA gene
455 amplicon sequencing. In a previous study using reference-based analysis, species profiles at 0.5 million
456 reads (0.15 Gb) showed an average correlation of 0.99 with ultra-deep sequencing (750 Gb)[3]. Using
457 mock communities, Xu et al. recovered 62 % of the 62 species at 1 Gb/sample using MetaPhlAn2 and
458 the Refseq database[2]. The analysis of our mock communities with MetaPhlAn4 at 1 Gb showed similar
459 result (Mock-stag-24: 75%, Mock-even-70: 67%, Mock-stag-70: 43%), without resolution at the strain
460 level (four strains of each *E. coli* and *P. vulgatus* in Mock-even-70). Using the reference genomes, we
461 detected reads for all 70 and 24 strains in the mock communities at the lowest sequencing depth of
462 0.1 Gb. The reference-based detection and discrimination of different strains with an ANI of up to
463 99.66 % was possible in our study. However, sequencing depths >5 Gb are required to obtain nearly
464 full genome coverage, which must be adjusted depending on the expected abundance and genome
465 size of the targeted strains in complex samples.

466 A sequencing depth ≥5 Gb was required to reconstruct sufficient MAGs, which agrees with previous
467 findings[3,6]. Importantly, at high sequencing depth (10 to 50 Gb), more MAGs were created than input
468 strains, yet not all strains were represented by a MAG. Inspecting this further revealed a substantial
469 number of hqMAGs that were chimeric due to containing fragments assigned to more than one
470 reference genome. This occurred despite excluding contigs smaller than 1,500 bp during binning to
471 prevent false inflation of chimerism. These results echo previous reports that ~5 % of genes within
472 MAGs differ from the dominant reconstructed taxa[48]. A recent study, which evaluated the quality of
473 short-read assemblies from a natural soil community by mapping them to the respective long-read
474 data, reported that assembly failures occurred in most genome bins[49], which agrees with our findings.

475 Our results show that deeper sequencing does not guarantee more complete MAG reconstruction and
476 does not eliminate chimerism. Data from the staggered Mock communities indicate that having more
477 reads for a reference genome does not assure its coverage by a MAG and does not prevent the risk of
478 multiple alignments. Instead, we observed that the creation of several MAGs for the same reference
479 genome was common. The aforementioned study of a natural soil community also found that
480 overabundance of reads can lead to high-coverage misassemblies[49]. Bioinformatic approaches that
481 reduced the occurrence of chimeric hqMAGs in our work were multi-coverage binning and strain aware
482 assembly, which is in line with previous studies[34,49]. Long-read sequencing also increased the number
483 of coherent hqMAGs. This highlights the need to consider long-read or hybrid sequencing for projects
484 where the goal is strain reconstruction, whilst simultaneously recognising the need to also enhance
485 such approaches further[50]. Nevertheless, none of the approaches tested could resolve the issue

486 completely and the data presented raises concerns about the prevalence of spurious sequences in
487 ever-growing MAG catalogues[51], as we used common bioinformatic workflows to create them.

488 To quantify the functional information lost with decreasing sequencing depth, we calculated the
489 average KEGG pathway completeness, showing that >2 Gb is required to reach >50 % completeness.
490 In contrast, one study with a mock community of 62 human gut bacteria reported that the functional
491 description of shallow metagenomic samples (1 Gb) was "quite similar" to the one obtained from the
492 reference genomes at KEGG level 1 and 2[2]. Another study based on in-silico subsampling of an ultra-
493 deep metagenome (2.5 billion reads; 750 Gb) to 0.5 million reads (0.15 Gb) showed a nearly full
494 recovery of KEGG Orthology groups[3]. This result might be influenced by differences in the complexity
495 of analysed communities and by the subsampling process being rarefied repeatedly and therefore
496 creating a higher chance of recovering functions. At the level of predicted protein sequences, our data
497 suggests that shallow metagenomics is not advisable, as a recovery of <75 % predicted proteins was
498 reached at sequencing depths ≤5 Gb. This agrees with a previous study of antimicrobial resistance
499 genes at different sequencing depths, in which at least 80 million reads (24 Gb) was needed to recover
500 the full richness of AMR gene families[52].

501 Regarding the influence of wet-lab factors, the performance of library preparation procedures has
502 previously been compared for traditional metagenomic sequencing of native samples and Mock
503 communities, revealing that the community diversity, amount of input DNA, and the sequencing
504 platform alter the results[53–55]. When analysing the effects of library preparation in two facilities and
505 the addition of background DNA from the gut content of germfree mice, both factors impacted
506 taxonomic and functional results at the sequencing depths <5 Gb. Strains with a higher relative
507 abundance were less sensitive to these effects, even at lower sequencing depths. If the use of different
508 protocols in large studies cannot be avoided (e.g., when handling low biomass samples), or a high
509 amount of background DNA is to be expected, higher sequencing depth may improve the robustness
510 of results.

511 This study has some limitations: (i) Metagenomics can be used to study various environments, which
512 may require different strategies; due to our own research interests, we focused on wet-lab parameters
513 related to gut microbiomes. Although we could show their overall effects on metagenomic data, the
514 study design was not appropriate to disentangle the effect of each parameter separately. Previous
515 studies have investigated other aspects such as low biomass, contamination, or additional library
516 preparation protocols (DNA input, PCR cycles, fragment/insert size)[53,56]; (ii) Whilst we used
517 MetaPhlAn4 for some of the analyses, most results were obtained using the exact strains (*i.e.*, their
518 genomes) as a reference, which represents an ideal case for taxonomic and functional readouts; (iii)
519 We have tested multiple approaches, including standard workflows used by many, to study the

520 chimeric MAGs. We do not exclude that additional strategies not tested in this work may perform
521 better. Future benchmarking of bioinformatic tools, as in Critical Assessment of Metagenome
522 Interpretation (CAMI), will be helpful to move forward with comprehensive testing of new approaches
523 to address chimeric MAGs at scales that go beyond our single study[57]; (iv) the mock communities
524 analysed included only bacteria; whilst bacteria are dominant members in many microbiomes, the
525 work excludes the influence of other types of microbes; (v) We used a higher number and more
526 complex mock communities (up to 70 strains, including different distributions) than in previous
527 studies; they nonetheless are only approximations of native bacterial populations in gut microbiomes,
528 which include ca. 300 species, with 150-200 usually detected by sequencing.

529 In summary, due to the lack of studies benchmarking shallow metagenomics using complex mock
530 communities, potential users may not be aware of the pitfalls of this approach. The sequencing depth
531 needs to match the main aims of a study. Factors to be considered include the expected microbial
532 diversity in the target samples, the evenness of distribution within the communities, and the amount
533 of DNA available. Shallow metagenomics is a valid choice for a large study aiming at database-guided
534 taxonomic profiles of well-characterised environments. However, it is not appropriate for high-
535 resolution functional analysis and *de-novo* strain resolution. Even at high depths of sequencing, MAG-
536 based approaches require careful planning and execution to minimise the generation and spread of
537 artificial bacterial diversity.

538 **Acknowledgement**

539 We are grateful to: (i) Evelyn Deis for sample processing, Alina Viehof-Beckmann for providing gut
540 content of germ-free mice, and Ntana Kousetzi for outstanding support with molecular work (all from
541 Functional Microbiome Research Group, University Hospital of RWTH Aachen); (ii) Christel Driessen for
542 genomic library preparation (Department of Medical Microbiology, Infectious Diseases and Infection
543 Prevention, Maastricht University Medical Centre). This work was supported by: (i) the Genomics
544 Facility of the Interdisciplinary Center for Clinical Research (IZKF) Aachen within the Faculty of Medicine
545 at RWTH Aachen University; (ii) the DFG-funded NGS Competence Center Tübingen (INST 37/1049-1)
546 and the Institute for Medical Microbiology and Hygiene at the University Hospital (Tübingen,
547 Germany), including help by the Quantitative Biology Center (QBiC) for raw data management and
548 storage; (iii) the Life Science Computer Cluster (LiSC) of the University of Vienna for the processing of
549 long-read sequencing data.

550

551 **Authors contributions**

552 NT: conceptualization, methods and software, investigation, formal analysis, validation, interpretation,
553 visualization, data curation, writing – original draft; CP: methods and software, validation,
554 interpretation, writing – review and editing; JS and PP: investigation, formal analysis, writing – review
555 and editing; DB and JP: resources, funding acquisition, writing – review and editing; TCAH:
556 conceptualization, interpretation, writing – original draft; TC: conceptualization, visualization,
557 interpretation, resources, supervision, project administration, funding acquisition, writing – original
558 draft.

559

560 **Material and data availability**

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

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

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

567

568 **Funding**

TC received funding from the German Research Foundation (DFG), project no. 403224013 (SFB1382, Q02), project no. 460129525 (NFDI4Microbiota), and project no. 445552570. DB received funding from the Austrian Science Fund (10.55776/DOC69; 10.55776/COE7).

**Ethics**

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

576 **Figure Legends**

577 **Fig. 1:** Reference-based taxonomic profiles

578 **A** Number of reference genomes per category of coverage by metagenomic reads (colour gradients)
579 for Mock-even-70, Mock-stag-24, and Mock-stag-70 (top to bottom) at up to eleven sequencing depths
580 (x-axis). **B** Coefficients of variation of relative abundances of the 70 strains in ten *in-silico* datasets for
581 each sequencing depth, subsampled from Mock-stag-70 sequenced at 50Gb. The coefficients of
582 variation between the sequencing depths were tested statistically using a Kruskal-Wallis rank-sum test
583 with Benjamini-Hochberg correction (p-values in **Supplementary Table S5**). **C** Difference (delta-values)
584 between measured relative abundances and theoretical values (x-axis) for the three mock
585 communities after taxonomic assignment using either the reference genomes (blue) or MetaPhlAn4
586 (red). In the MetaPhlAn analysis, only the taxa that matched reference genomes at the species level
587 were considered.

588

589 **Fig. 2:** Strain analysis

590 **A** Strain analysis based on the MAGs assembled in Mock-even-70 (top) and Mock-stag-24 (bottom) at
591 each sequencing depth individually. The left panels (dark blue) depict the number of bins assembled
592 from the shotgun data; the right panels (lighter blue) show the number of reference genomes matching
593 MAGs (coverM), indicating that multiple MAGs were reconstructed for some Mock species. The
594 reference genome with the highest coverage by MAG reads was chosen as representative of that MAG.
595 **B** Number of high-, medium- and low-quality MAGs assigned to either one (blue) or several (red)
596 reference genomes (with coverM using a cut-off of > 0.25% coverage of reference genomes). **C** Three
597 exemplary high-quality (> 90 % completeness, < 5% contamination) (hq) MAGs (dark grey, outer circle)
598 reconstructed from Mock-even-70 at 10 Gb. They were aligned to the reference genomes they covered
599 by more than 0.25% to illustrate different categories of chimerism. The predominantly covered
600 reference genome is depicted in blue, while chimeric sequences of further reference genomes (inner
601 circles) are shown in red. **D** Number of hqMAGs assigned to one reference genome (blue) or more
602 reference genomes (orange) for the ten different Mock-stag-24 (v1-10) with varying reference genome
603 abundance distribution. MAGs were constructed with either a single- (bright colours) or a multi-
604 coverage (darker colours) binning approach. Statistics: Kruskal-Wallis rank-sum test; *** p-value =
605 0.0003871. **E** Number of hqMAGs constructed from contigs assembled with either MEGAHIT or
606 metaSPAdes for Mock-stag-70. The MAGs were then assigned to one (green) or more (red) reference
607 genomes as above. **F** High-quality MAGs binned from assemblies generated with either MEGAHIT or
608 metaSPAdes using 10 datasets at 10 Gb subsampled *in-silico* from 50 Gb. Statistics: Wilcoxon rank-sum

609 test; * p-value = 0,031. **G** Number of hqMAGs binned from assemblies acquired by long-read
610 sequencing of Mock-stag-70 and categorised as in E.

611

612 **Fig. 3**: Functional coverage

613 **A** Average KEGG pathway completeness of Mock-even-70, Mock-stag-24, and Mock-stag-70 using
614 contigs assembled with MEGAHIT, and Mock-stag-70 using contigs assembled with metaSPAdes. **B**
615 Fraction of predicted proteins from the reference genomes covered by the metagenomic assemblies
616 (MEGAHIT) for three different mock communities at the respective sequencing depths. Reads of Mock-
617 stag-70 were additionally assembled using metaSPAdes (bottom). **C** Average coverage of functions of
618 the reference genomes by ten *in-silico* datasets, subsampled to ten different sequencing depths each
619 from Mock-stag-70 sequenced at 50Gb. Bars are mean values; whiskers are standard deviations.
620 Statistics: Kruskal-Wallis rank-sum test with Benjamini-Hochberg correction (p-values in
621 **Supplementary Table S5**).

622

623 **Fig. 4:** Effects of background DNA (bgDNA) and library preparation in two facilities

624 **A** PCA plot of relative abundance profiles in Mock-even-70 (top) and Mock-stag-24 (bottom). **B**
625 Difference in coverage of reference genomes between facility 1 and facility 2 (left) and samples with
626 or without bgDNA (right) for Mock-even-70. **C** Same as in panel B for Mock-stag. The reference
627 genomes were ranked (from top to bottom) according to increasing DNA amount in the mixture, as
628 indicated by the colour gradient (from blue to red; 0.04, 0.4, 1, 2, 4, 10, 20, 40 ng).

629

630 **Online Supplementary Files**

631 **Supplementary Fig. S1:** Schematic overview of the experimental design

632 The genomic DNA of 70 bacterial isolates (www.dsmz.de/miBC; https://www.hibc.rwth-aachen.de)
633 was used as input to create 13 different Mock communities. Two Mock communities contained all
634 members, one each with an even distribution (Mock-even-70) or staggered distribution (Mock-stag-
635 70). Additionally, mock communities with varying staggered distribution of 24 isolates were created:
636 Mock-stag 24, sequenced at nine different sequencing depths; Mock-stag 24 v1-10 sequenced at 10
637 Gb for testing multi-coverage binning for MAG construction. A second version of Mock-even-70 and
638 one Mock-stag-24 were created by spiking DNA isolated from the gut content of germfree mice.
639 Libraries for these four mock communities (Mock-even-70 and Mock-stag-24 ± bg) were prepared in
640 two different sequencing facilities. A library per sequencing depth was then sequenced using the
641 Illumina technology (short reads) at up to 11 sequencing depths. Mock-stag-70 was also sequenced

642 using Oxford Nanopore Sequencing (long reads) in a third facility. Bioinformatic analyses included: (i)
643 number and relative abundance of strains; (ii) coverage of reference genomes; (iii) number and
644 diversity of predicted proteins and completeness of functional pathways; (iv) strain-level resolution
645 using both a reference-based and metagenome-assembled genome (MAG) approach.

646

647 **Supplementary Fig. S2:** Phylogenetic tree of bacterial strains in the mock communities
648 A phylogenetic tree of high-quality, draft genomes of the 70 isolates was created using PhyloPhlAn
649 v3.0.67. The phyla to which the strains belong are represented by coloured boxes at the end of the
650 branches. Distribution of the strains in each mock community is provided in **Supplementary Table S1**.

651

652 **Supplementary Fig. S3:** Schematics of the bioinformatics workflow
653 The details of each analysis are described in the methods.

654

655 **Supplementary Fig. S4:** Coverage of individual reference genomes
656 Heatmaps showing the coverage of all reference genomes in Mock-even-70 (**A**), Mock-stag-24 (**B**), and
657 Mock-stag-70 (**C**) by metagenomic reads at up to 11 sequencing depths (x-axis). The grey boxes in panel
658 **A** indicate the multiple *E. coli* and *P. vulgatus* strains. The reference genomes in the mock communities
659 with staggered distribution in panel **B** and **C** were ranked (from top to bottom) according to increasing
660 DNA amount in the mixture, as indicated by the colour gradient in panel B (from blue to red; 0.04, 0.4,
661 1, 2, 4, 10, 20, 40 ng) or the grey gradient in panel C (concentrations are provided in **Supplementary**
662 **Table S1**).

663

664 **Supplementary Fig. S5:** Relative abundance of reads per reference genome
665 Relative abundances of each strain in (**A**) Mock-even-70 in facility 1 and 2, (**B**) Mock-stag-24 in facility
666 1 and 2, and (**C**) Mock-stag-70 in facility 1. Metagenomic reads were assigned to the reference
667 genomes using coverM. Sequencing depths are indicated with the coloured dots. The theoretical
668 relative abundances are indicated by vertical dashed lines. For Mock-stag-24 in panel B and Mock-stag-
669 70 in panel C, the reference genomes are ranked from top to bottom by increasing DNA amount in the
670 mixture either indicated by coloured boxes (from blue to red; 0.04, 0.4, 1, 2, 4, 10, 20, 40 ng) or the
671 grey gradient in panel C (concentrations are provided in **Supplementary Table S1**).

672

673 **Supplementary Fig. S6:** Relative abundance after "non-supervised" taxonomic assignment using
674 MetaPhlAn4

675 Relative abundances of each strain in (**A**) Mock-even-70, (**B**) Mock-stag-24, and (**C**) Mock-stag-70.
676 Metagenomic reads were assigned using MetaPhlAn4. Sequencing depths are indicated with the
677 coloured dots. The theoretical relative abundances are indicated by vertical dashed lines. For Mock-
678 stag-24 in panel B and Mock-stag-70 in panel C, the reference genomes are ranked from top to bottom
679 by increasing DNA amount in the mixture (see concentrations in **Supplementary Table S1**).

680

681 **Supplementary Fig. S7:** Strain level resolution based on reference genomes

682 **A** Reference genome coverage (y-axis) for the four strains of *E. coli* and *P. vulgatus* at the nine different
683 sequencing depths (x-axis). **B** ANI values between the strains within each species.

684

685 **Supplementary Fig. S8:** MAGs numbers in Mock-stag-70

686 Strain analysis based on the MAGs assembled in Mock-stag-70 using MEGAHIT for assembly at each
687 sequencing depth individually. The left panel (dark blue) depict the number of MAGs assembled from
688 the shotgun data; the right panel (lighter blue) show the number of reference genomes matching
689 MAGs (coverM), indicating that multiple MAGs were reconstructed for some Mock species. The
690 reference genome with the highest coverage by MAG reads was chosen as representative of that MAG.

691

692 **Supplementary Fig. S9:** Reference genomes per MAG for Mock-stag-24, Mock-even-70, and Mock-
693 stag-70.

694 Stacked bar charts of the number of MAGs per sequencing depth. Each box represents one reference
695 genome. The size of the box indicates the number of MAGs that were assigned to this reference
696 genome. The colours in the graph for Mock-stag-24 indicate the amount of input DNA for a given strain
697 in the pool. The colour of the graph for Mock-even-70 and Mock-stag-70 indicates the reference
698 genomes.

699

700 **Supplementary Fig. S10:** Contigs assigned to reference genomes

701 Contigs from metagenomes sequenced at 10 Gb, 20 Gb and 50 Gb either assembled with MEGAHIT or
702 metaSPAdes were aligned to the reference genomes with blastn (-perc_identity 97% -evalue 1e-10,
703 alignment length >150 bp). The number of contigs (y-axis) is plotted against the number of reference
704 genomes assigned to a contig (x-axis).

705

706 **Supplementary Fig. S11:** Heatmaps of functional coverage.

707 Completeness of all KEGG pathways present in the reference genomes in Mock-even-70, Mock-stag-
708 24 and Mock-stag-70 assembled with MEGAHIT, and Mock-stag-70 assembled with metaSPAdes (from
709 left to right) at the up to 11 sequencing depths (columns).

710

711 **Supplementary Fig. S12:** Bar charts of protein sequences recovered.

712 Protein sequenced predicted in the metagenomic datasets as percentage of the protein sequences
713 present in the reference genomes (y-axis). The colour gradient indicates the facility where the library
714 was prepared and the addition of background DNA (bgDNA) from the caecum of germfree mice.

715

716 **Supplementary Table S1**: Distribution of the reference genomes DNA in the different mock
717 communities.

718 **Supplementary Table S2**: Characteristics of the reference genomes.

719 **Supplementary Table S3**: Number of reads after sequencing and subsampling.

720 **Supplementary Table S4:** Assignment of the MAGs to reference genomes and using GTDB.

721 **Supplementary Table S5:** P-values for Fig. 1B, Fig. 1C and Fig. 3C.

722 **References**

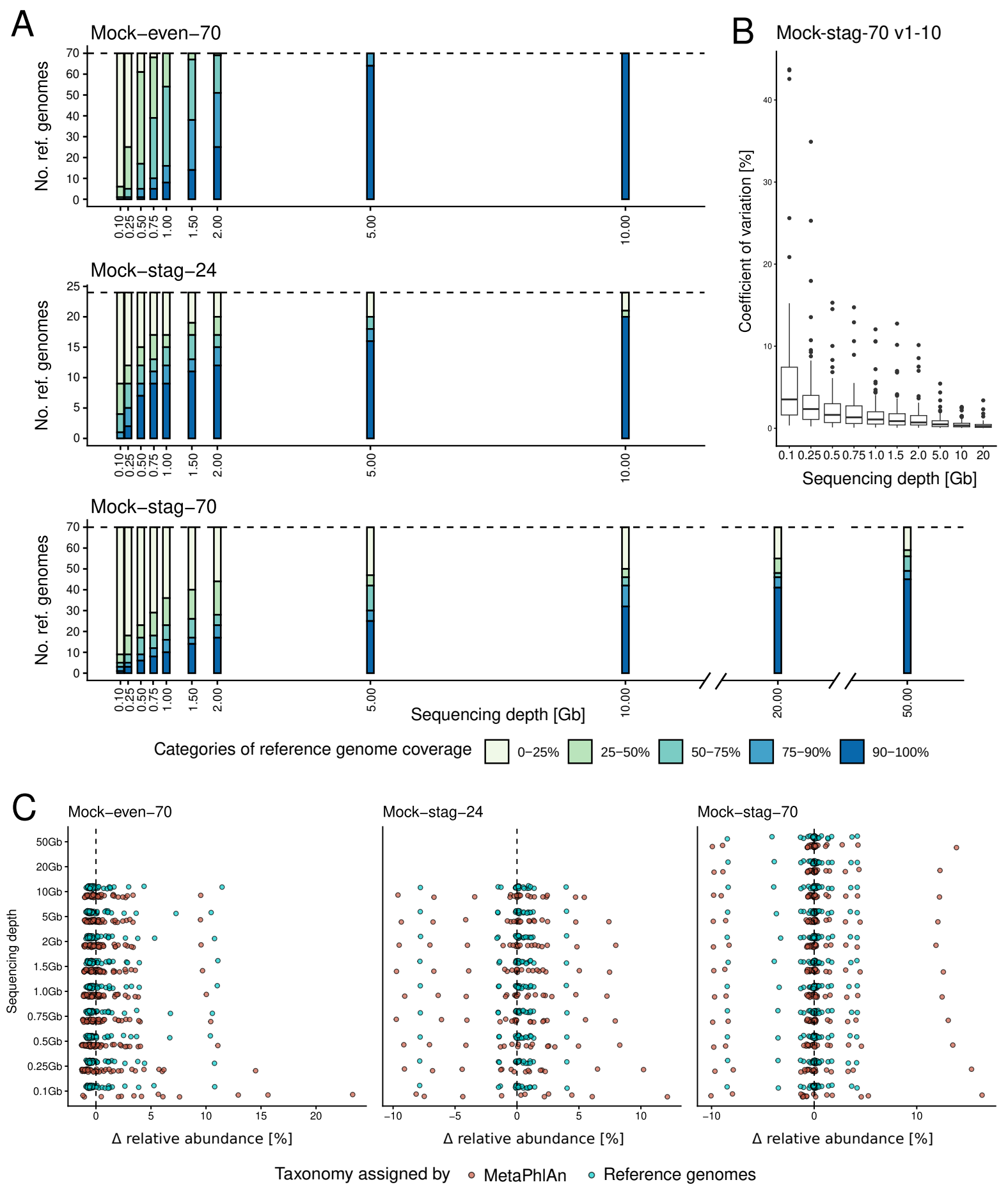

723 1. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and
724 next-generation sequencing-based diversity studies. *Nucleic Acids Research* **41**, e1–e1 (2013).
725 2. Xu, W. *et al.* Characterization of Shallow Whole-Metagenome Shotgun Sequencing as a High-
726 Accuracy and Low-Cost Method by Complicated Mock Microbiomes. *Front. Microbiol.* **12**, 678319
727 (2021).
728 3. Hillmann, B. *et al.* Evaluating the Information Content of Shallow Shotgun Metagenomics.
729 *mSystems* **3**, 10.1128/msystems.00069-18 (2018).
730 4. La Reau, A. J. *et al.* Shallow shotgun sequencing reduces technical variation in microbiome analysis.
731 *Sci Rep* **13**, 7668 (2023).
732 5. Stothart, M. R., McLoughlin, P. D. & Poissant, J. Shallow shotgun sequencing of the microbiome
733 recapitulates 16S amplicon results and provides functional insights. *Molecular Ecology Resources*
734 **23**, 549–564 (2023).
735 6. Tremblay, J., Schreiber, L. & Greer, C. W. High-resolution shotgun metagenomics: the more data,
736 the better? *Briefings in Bioinformatics* **23**, bbac443 (2022).
737 7. Usyk, M. *et al.* Comprehensive evaluation of shotgun metagenomics, amplicon sequencing, and
738 harmonization of these platforms for epidemiological studies. *Cell Reports Methods* **3**, 100391
739 (2023).
740 8. Cattonaro, F., Spadotto, A., Radovic, S. & Marroni, F. Do you cov me? Effect of coverage reduction
741 on metagenome shotgun sequencing studies. *F1000Res* **7**, 1767 (2020).
742 9. Fritz, A. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17
743 (2019).
744 10. Afrizal, A. *et al.* Enhanced cultured diversity of the mouse gut microbiota enables custom-made
745 synthetic communities. *Cell Host & Microbe* **30**, 1630-1645.e25 (2022).
746 11. Hitch, T. C. A. *et al.* Broad diversity of human gut bacteria accessible via a traceable strain
747 deposition system. Preprint at https://doi.org/10.1101/2024.06.20.599854 (2024).

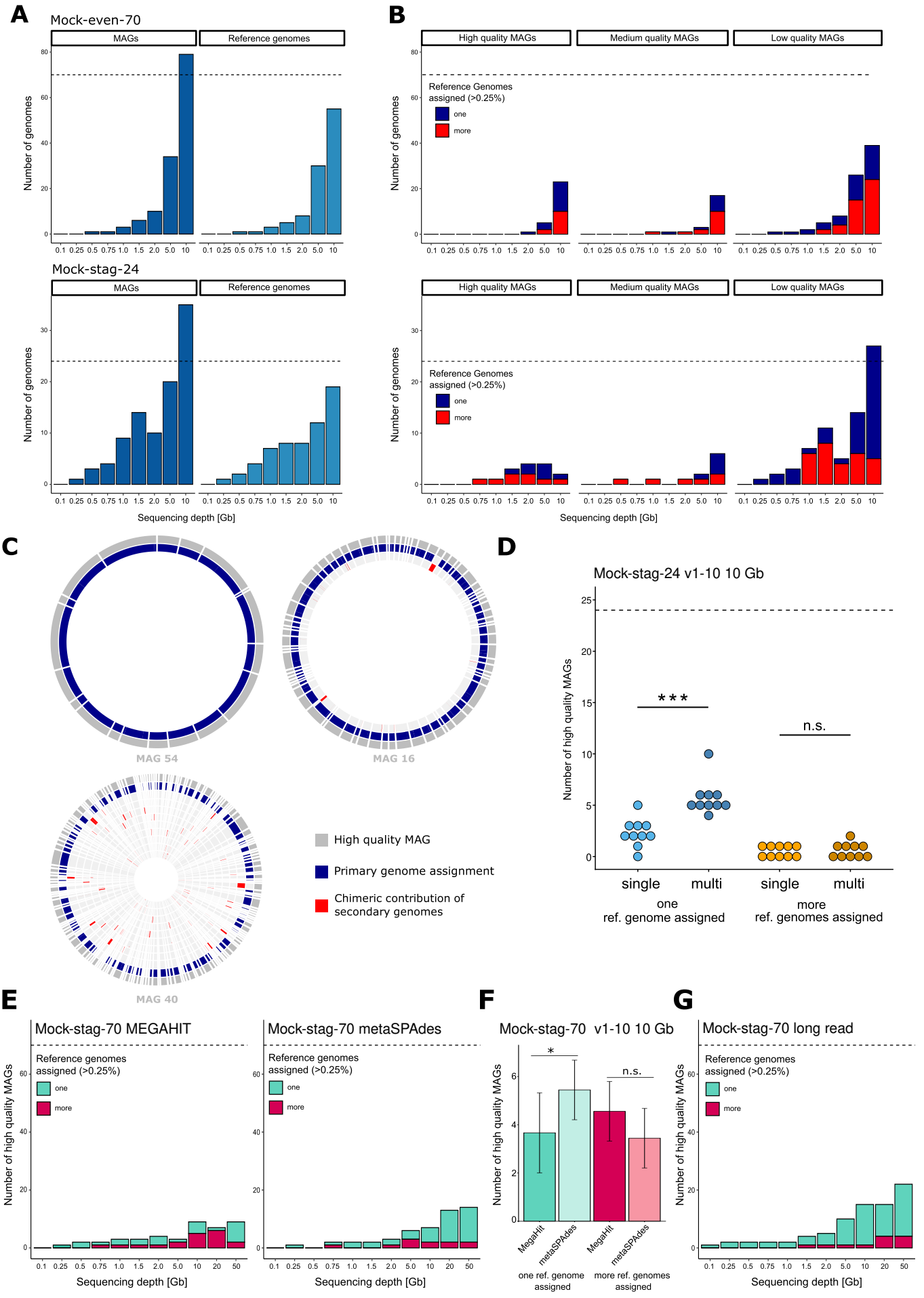
748 12. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an
749 anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl Environ*
750 *Microbiol* **63**, 2802–2813 (1997).
751 13. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from
752 metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**, 2500 (2020).
753 14. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology
754 and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
755 15. Li, H. bioawk, https://github.com/lh3/bioawk.
756 16. Li, H. seqtk, https://github.com/lh3/seqtk.
757 17. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data.
758 *Bioinformatics* **30**, 2114–2120 (2014).
759 18. Bushnell, B. BBMap, sourceforge.net/projects/bbmap/.
760 19. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution
761 for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**,
762 1674–1676 (2015).
763 20. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
764 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
765 21. Aroney, S. T. N. *et al.* CoverM: Read coverage calculator for metagenomics. Zenodo
766 https://doi.org/10.5281/ZENODO.10531253 (2024).
767 22. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with
768 uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, 1633–1644 (2023).
769 23. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI
770 analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114
771 (2018).
772 24. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification.
773 *BMC Bioinformatics* **11**, 119 (2010).

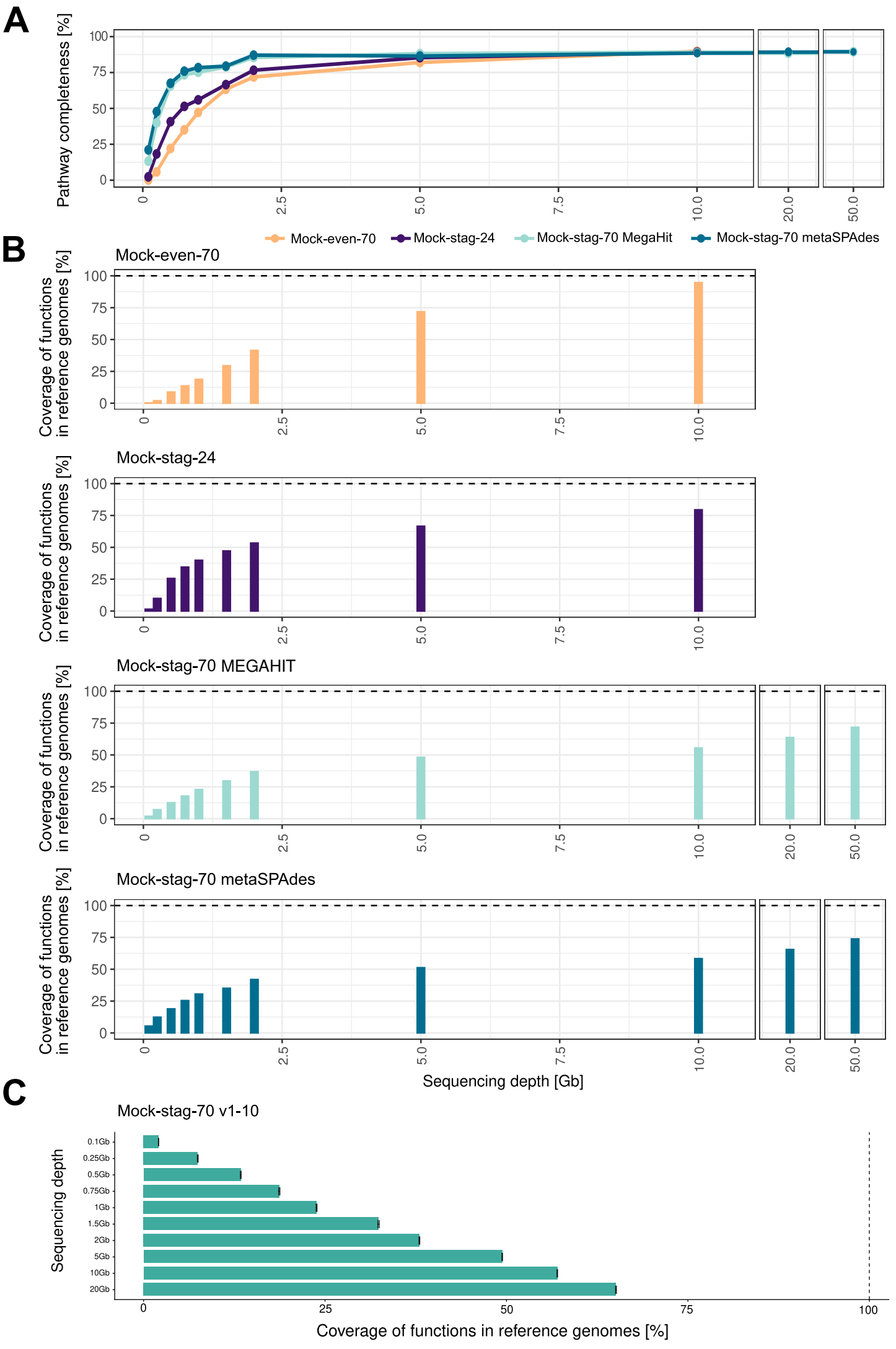
774 25. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive
775 score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
776 26. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-
777 distributed bacterial phototroph. *The ISME Journal* **12**, 1861–1866 (2018).
778 27. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat*
779 *Methods* **12**, 59–60 (2015).
780 28. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–
781 359 (2012).
782 29. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
783 30. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
784 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
785 31. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the
786 quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome*
787 *Res.* **25**, 1043–1055 (2015).
788 32. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes
789 with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
790 33. Yu, Y., Ouyang, Y. & Yao, W. shinyCircos: an R/Shiny application for interactive creation of Circos
791 plot. *Bioinformatics* **34**, 1229–1231 (2018).
792 34. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic
793 binning reveals extensive hidden contamination. *Nat Methods* **20**, 1170–1173 (2023).
794 35. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing
795 data. *Bioinformatics* **39**, btad311 (2023).
796 36. Hall, M. Rasusa: Randomly subsample sequencing reads to a specified coverage. *JOSS* **7**, 3941
797 (2022).
798 37. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat
799 graphs. *Nat Biotechnol* **37**, 540–546 (2019).

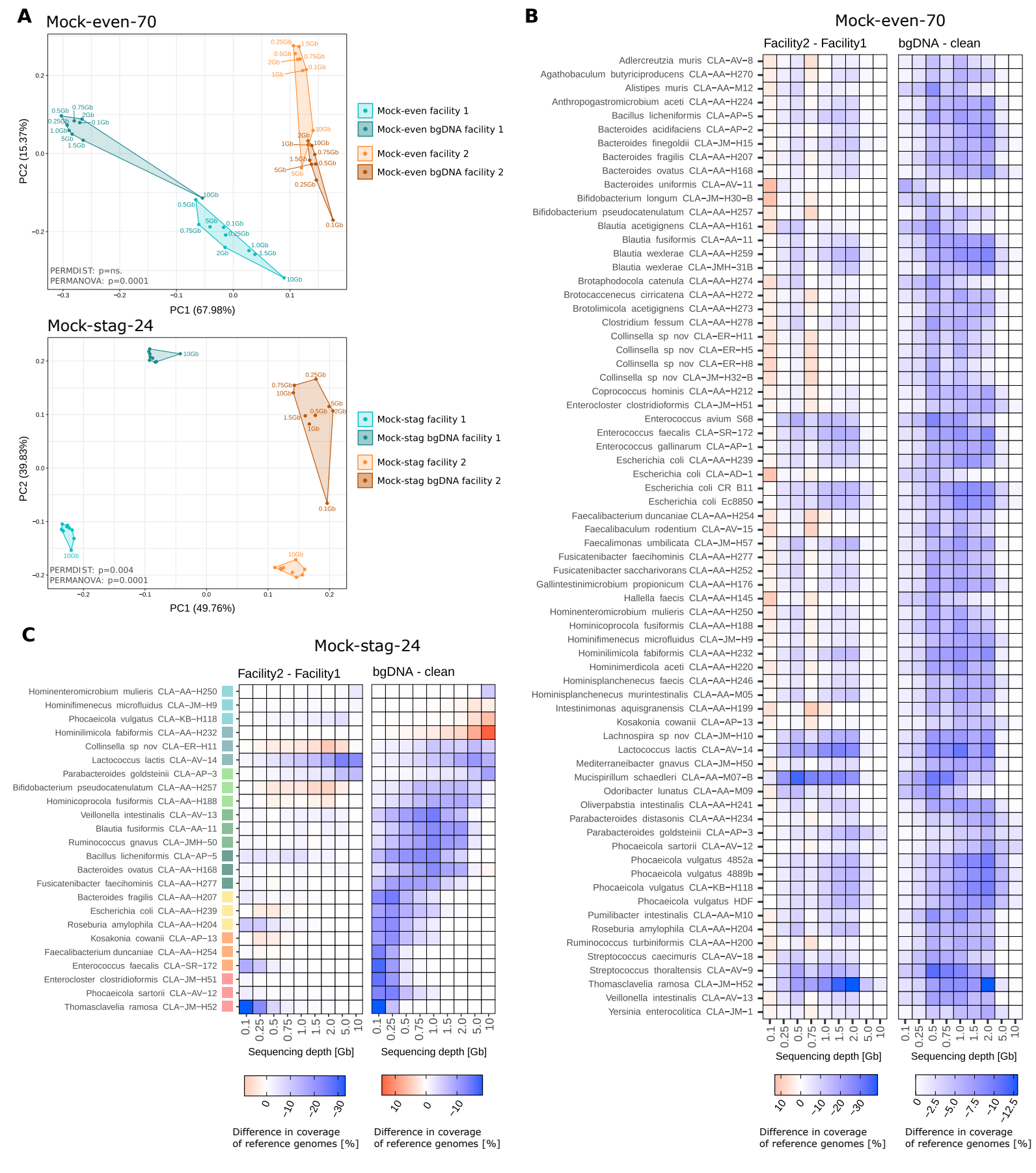
38. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
39. R Core Team. R: A Language and Environment for Statistical Computing, https://www.R-project.org/. *R Foundation for Statistical Computing, Vienna, Austria* https://www.R-project.org/ (2021).
40. Jari Oksanen and Gavin L. Simpson and F. Guillaume Blanchet and Roeland Kindt and Pierre Legendre and Peter R. Minchin and R.B. O'Hara and Peter Solymos and M. Henry H. Stevens and Eduard Szoecs and Helene Wagner and Matt Barbour and Michael Bedward and Ben Bolker and Daniel Borcard and Tuomas Borman and Gustavo Carvalho and Michael Chirico and Miquel {De Caceres} and Sebastien Durand and Heloisa Beatriz Antoniazi Evangelista and Rich FitzJohn and Michael Friendly and Brendan Furneaux and Geoffrey Hannigan and Mark O. Hill and Leo Lahti and Dan McGlinn and Marie-Helene Ouellette and Eduardo {Ribeiro Cunha} and Tyler Smith and Adrian Stier and Cajo J.F. {Ter Braak} and James Weedon. vegan: Community Ecology Package, https://vegandevs.github.io/vegan/. (2024).
41. Hadley Wickham. Flexibly Reshape Data, http://had.co.nz/reshape. (2022).
42. Wickham, H. *et al.* Welcome to the Tidyverse. *JOSS* **4**, 1686 (2019).
43. 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).
44. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer international publishing, Cham, 2016).
45. Alboukadel Kassambara. ggpubr: 'ggplot2' Based Publication Ready Plots, https://rpkgs.datanovia.com/ggpubr/. (2023).
46. Xu, S. *et al.* Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers. *Front. Genet.* **12**, 774846 (2021).
47. Yuan Tang and Masaaki Horikoshi. ggfortify: Data Visualization Tools for Statistical Analysis Results, https://CRAN.R-project.org/package=ggfortify. (2016).

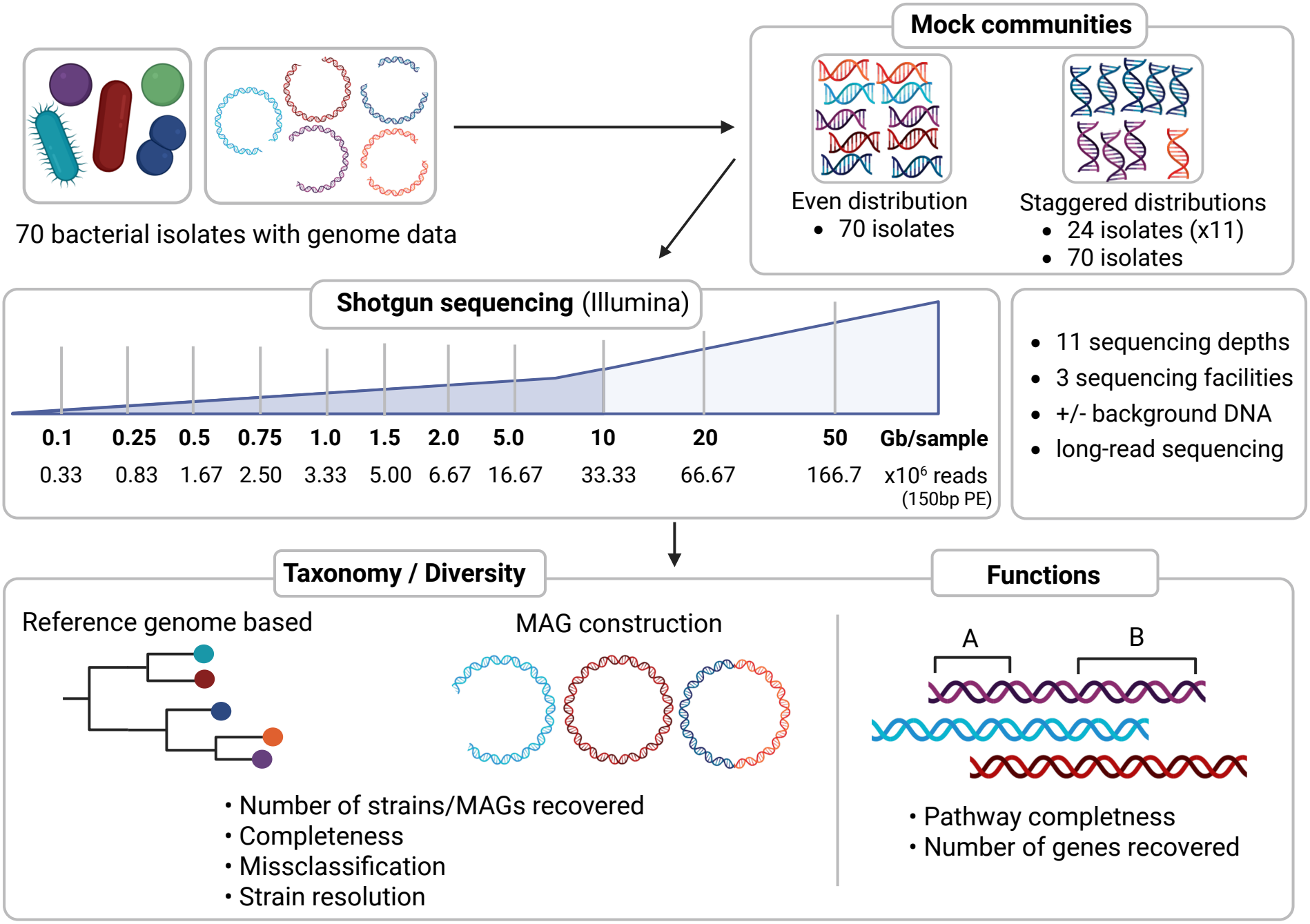
826 48. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing
827 Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the
828 Same Fecal Sample. *Appl Environ Microbiol* **87**, e02593-20 (2021).
829 49. Berg, M., Reiter, T., Emerson, J., Brown, C. T. & Roux, S. Comparison of short-read and long-read
830 metagenome assemblies in a natural soil community highlights systematic bias in recovery of high-
831 diversity populations. *NAR Genomics and Bioinformatics* **7**, lqaf163 (2025).
832 50. Trigodet, F., Sachdeva, R., Banfield, J. F. & Eren, A. M. Assemblies of long-read metagenomes suffer
833 from diverse errors. Preprint at https://doi.org/10.1101/2025.04.22.649783 (2025).
834 51. Ma, J. *et al.* A human gut metagenome-assembled genome catalogue spanning 41 countries
835 supports genome-scale metabolic models. *Nat Microbiol* https://doi.org/10.1038/s41564-025-
836 02206-1 (2025) doi:10.1038/s41564-025-02206-1.
837 52. Gweon, H. S. *et al.* The impact of sequencing depth on the inferred taxonomic composition and
838 AMR gene content of metagenomic samples. *Environmental Microbiome* **14**, 7 (2019).
839 53. Bowers, R. M. *et al.* Impact of library preparation protocols and template quantity on the
840 metagenomic reconstruction of a mock microbial community. *BMC Genomics* **16**, 856 (2015).
841 54. Poulsen, C. S., Ekstrøm, C. T., Aarestrup, F. M. & Pamp, S. J. Library Preparation and Sequencing
842 Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiol*
843 *Spectr* **10**, e00090-22 (2022).
844 55. Gaulke, C. A. *et al.* Evaluation of the Effects of Library Preparation Procedure and Sample
845 Characteristics on the Accuracy of Metagenomic Profiles. *mSystems* **6**, 10.1128/msystems.00440-
846 21 (2021).
847 56. Fierer, N. *et al.* Guidelines for preventing and reporting contamination in low-biomass microbiome
848 studies. *Nat Microbiol* **10**, 1570–1580 (2025).
849 57. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of
850 challenges. *Nat Methods* **19**, 429–440 (2022).

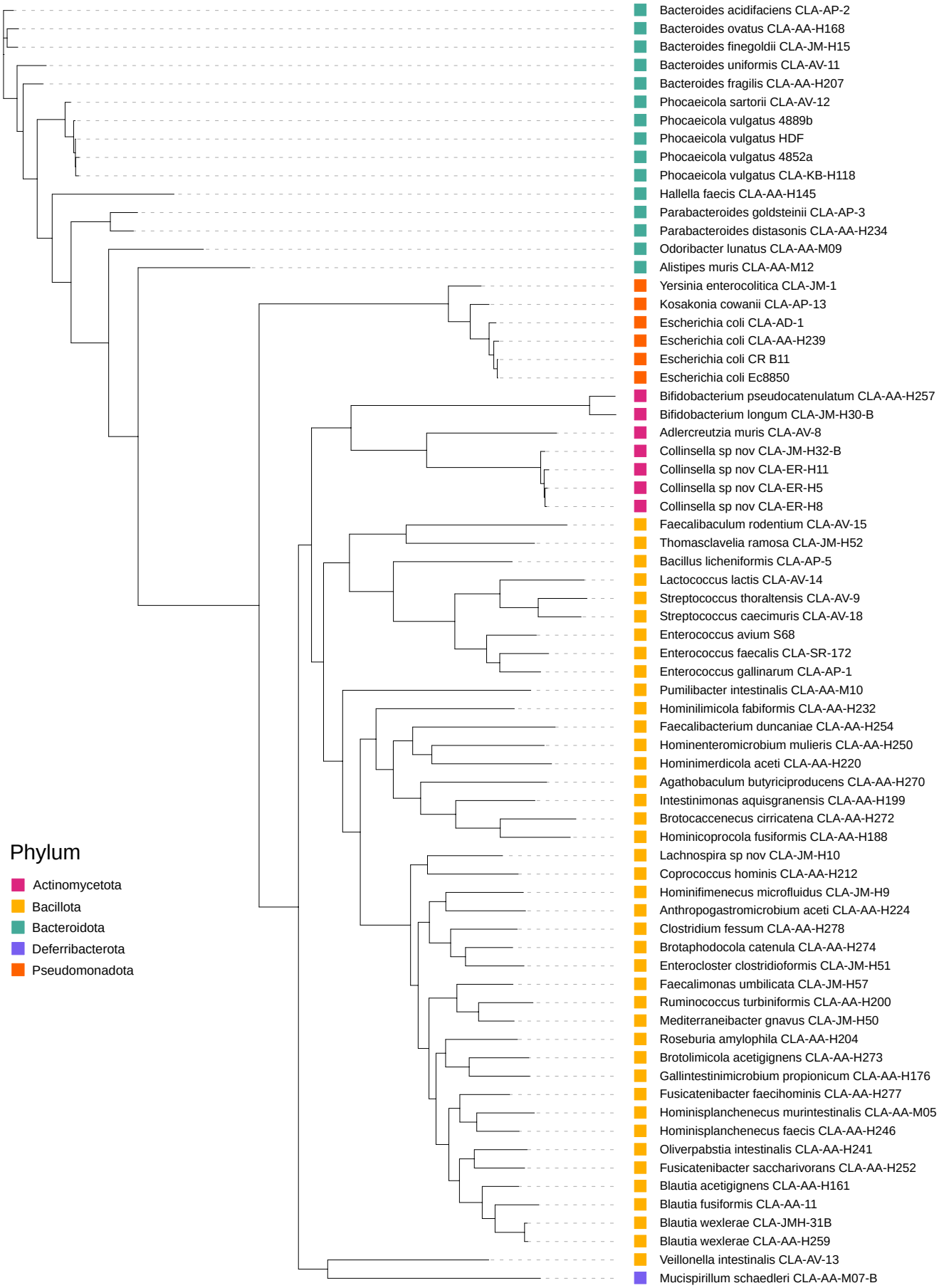

A
Mock−even−70
Mock−stag−24
Mock−stag−70
No. ref. genomes
Sequencing depth [Gb]
B
Mock-stag-70 v1-10
Coefficient of variation [%]
Sequencing depth [Gb]
Categories of reference genome coverage
0−25%
25−50%
50−75%
75−90%
90−100%
C
Mock−even−70
Mock−stag−24
Mock−stag−70
Sequencing depth
Δ relative abundance [%]
Taxonomy assigned by
MetaPhlAn
Reference genomes

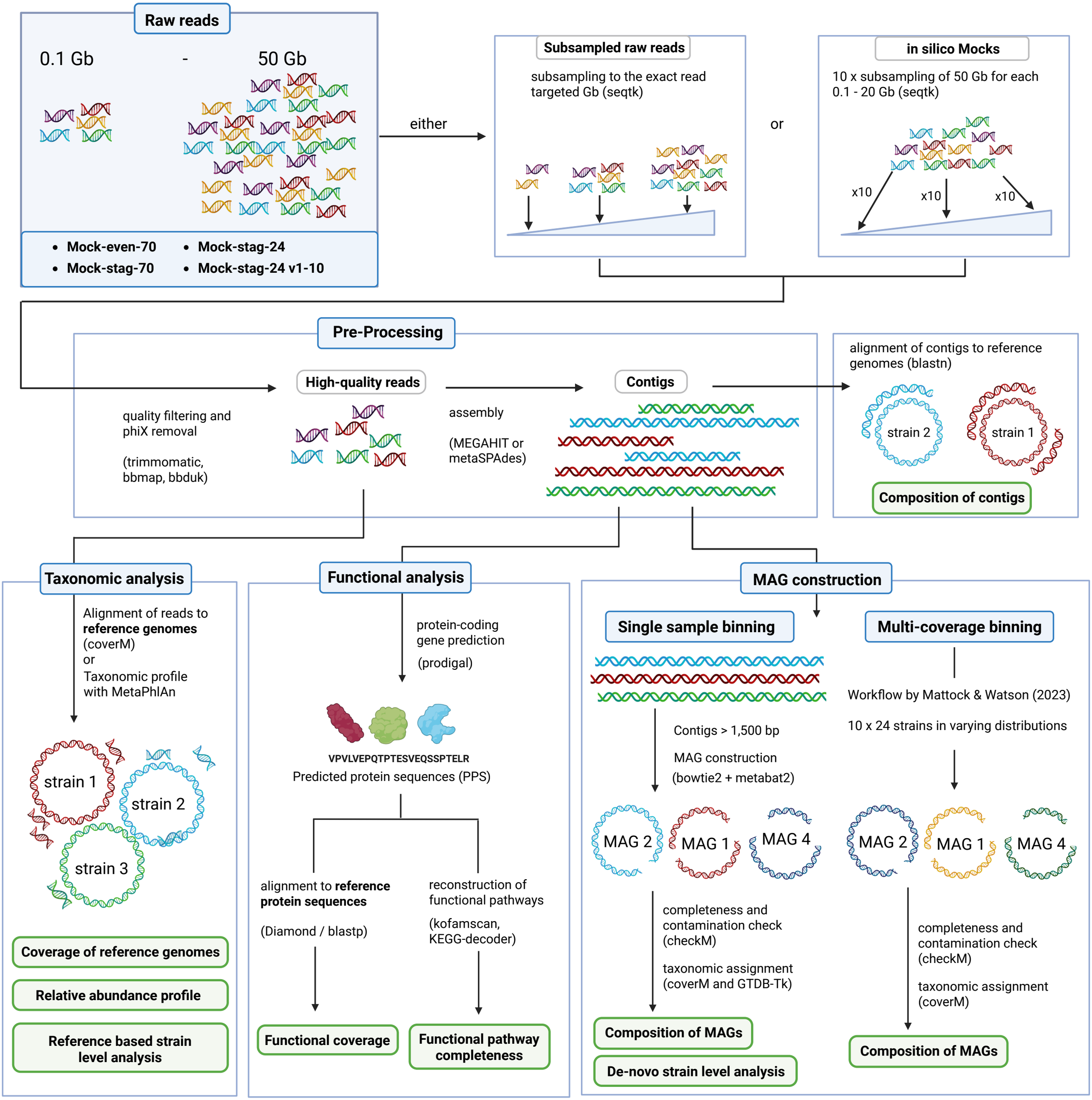

A
Mock-even-70
MAGs
Reference genomes
Number of genomes
Mock-stag-24
Sequencing depth [Gb]
B
High quality MAGs
Medium quality MAGs
Low quality MAGs
Reference Genomes assigned (>0.25%)
one
more
C
MAG 54
MAG 16
MAG 40
High quality MAG
Primary genome assignment
Chimeric contribution of secondary genomes
D
Mock-stag-24 v1-10 10 Gb
Number of high quality MAGs
***
n.s.
single
multi
one ref. genome assigned
more ref. genomes assigned
E
Mock-stag-70 MEGAHIT
Mock-stag-70 metaSPAdes
F
Mock-stag-70 v1-10 10 Gb
*
MegaHit
metaSPAdes
one ref. genome assigned
more ref. genomes assigned
G
Mock-stag-70 long read

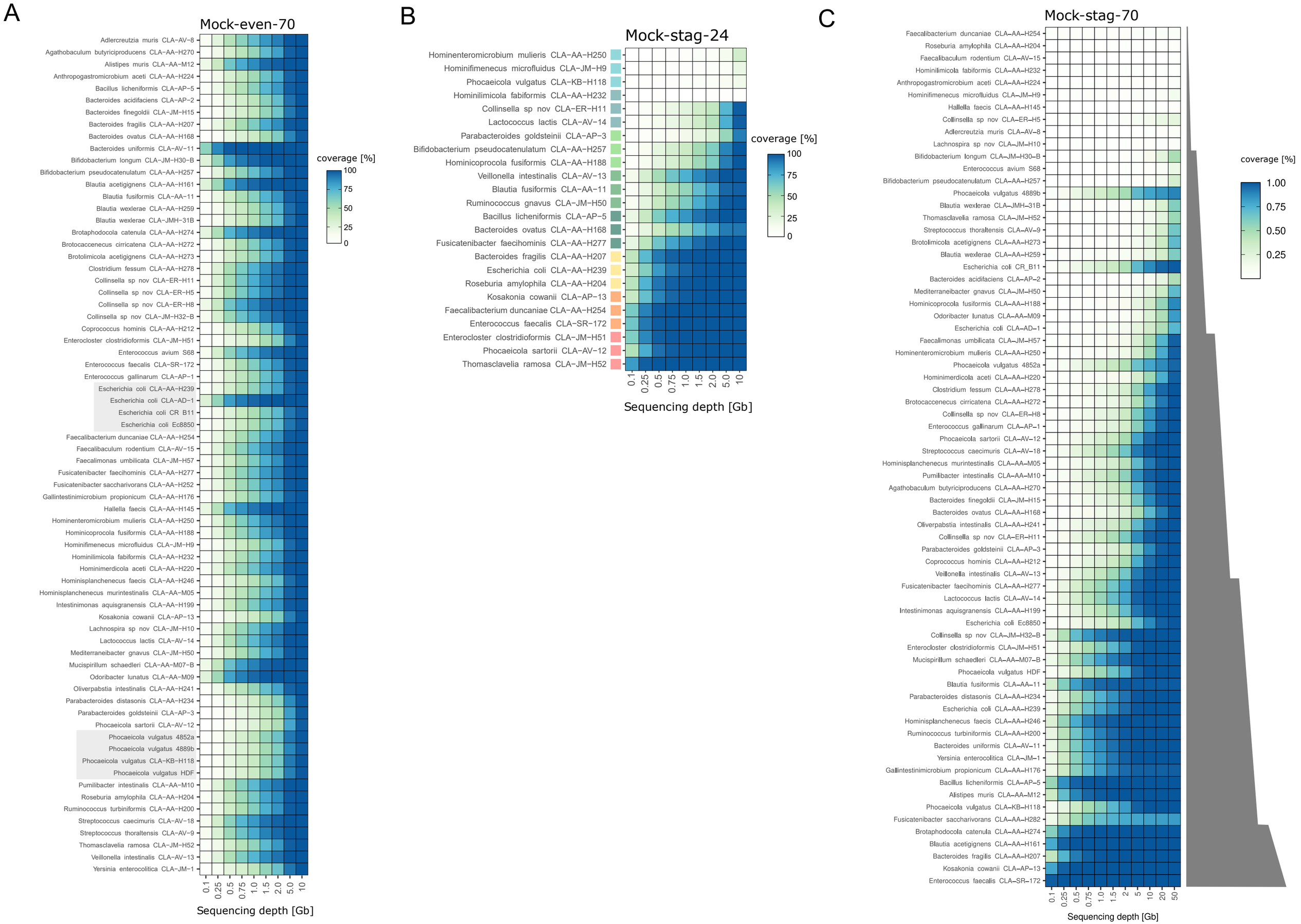

A
Pathway completeness [%]
Mock-even-70
Mock-stag-24
Mock-stag-70 MegaHit
Mock-stag-70 metaSPAdes
B
Mock-even-70
Mock-stag-24
Mock-stag-70 MEGAHIT
Mock-stag-70 metaSPAdes
Coverage of functions in reference genomes [%]
Sequencing depth [Gb]
C
Mock-stag-70 v1-10
Sequencing depth
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1Gb
1.5Gb
2Gb
5Gb
10Gb
20Gb
Coverage of functions in reference genomes [%]

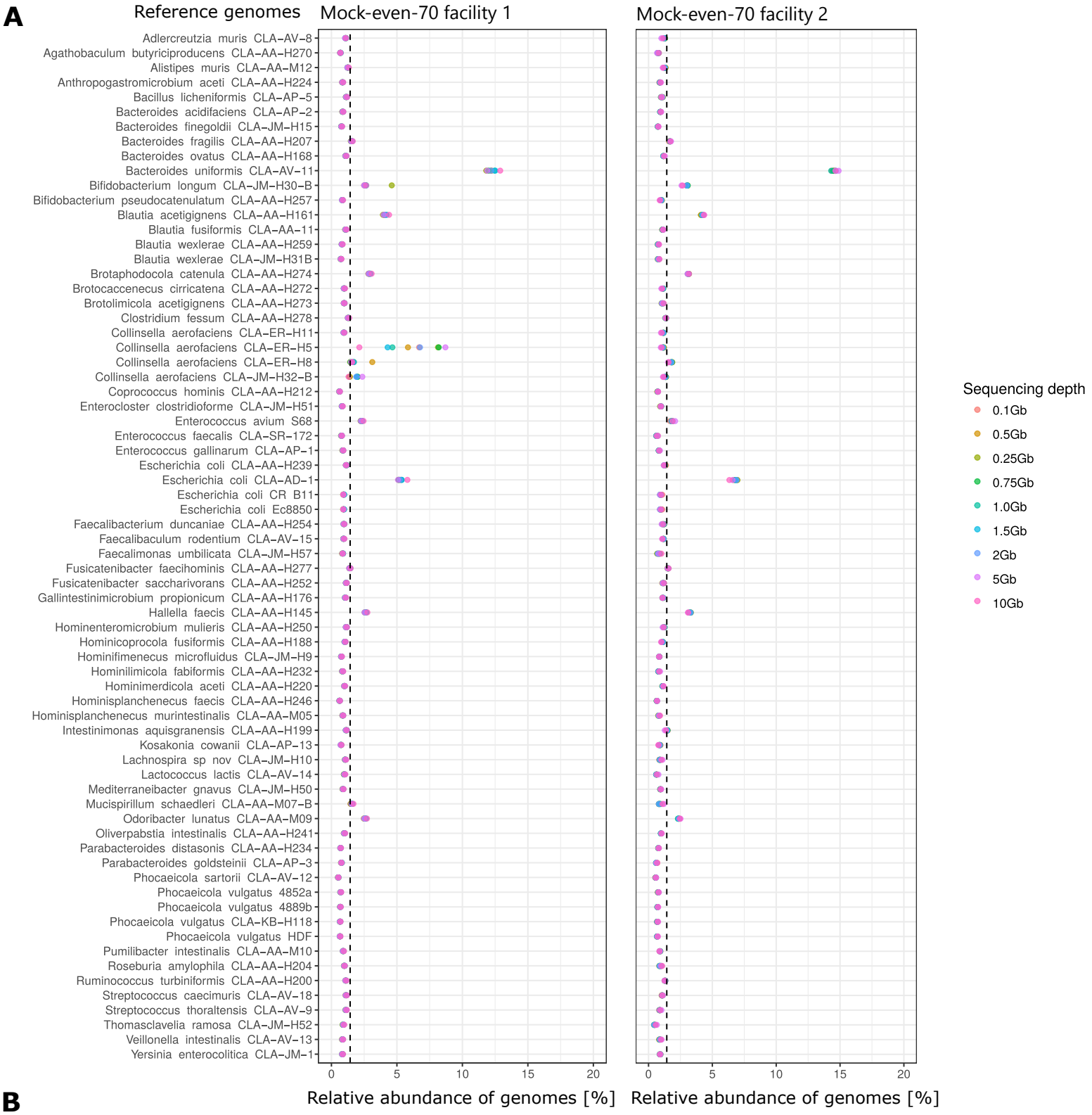

A
Mock-even-70
PC2 (15.37%)
PC1 (67.98%)
PERMDIST: p=ns.
PERMANOVA: p=0.0001
Mock-even facility 1
Mock-even bgDNA facility 1
Mock-even facility 2
Mock-even bgDNA facility 2
Mock-stag-24
PC2 (39.83%)
PC1 (49.76%)
PERMDIST: p=0.004
PERMANOVA: p=0.0001
Mock-stag facility 1
Mock-stag bgDNA facility 1
Mock-stag facility 2
Mock-stag bgDNA facility 2
B
Mock-even-70
Facility2 - Facility1
bgDNA - clean
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
Brotoccacenecus 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
Sequencing depth [Gb]
Difference in coverage of reference genomes [%]
C
Mock-stag-24
Facility2 - Facility1
bgDNA - clean
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-JMH-50
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
Sequencing depth [Gb]
Difference in coverage of reference genomes [%]

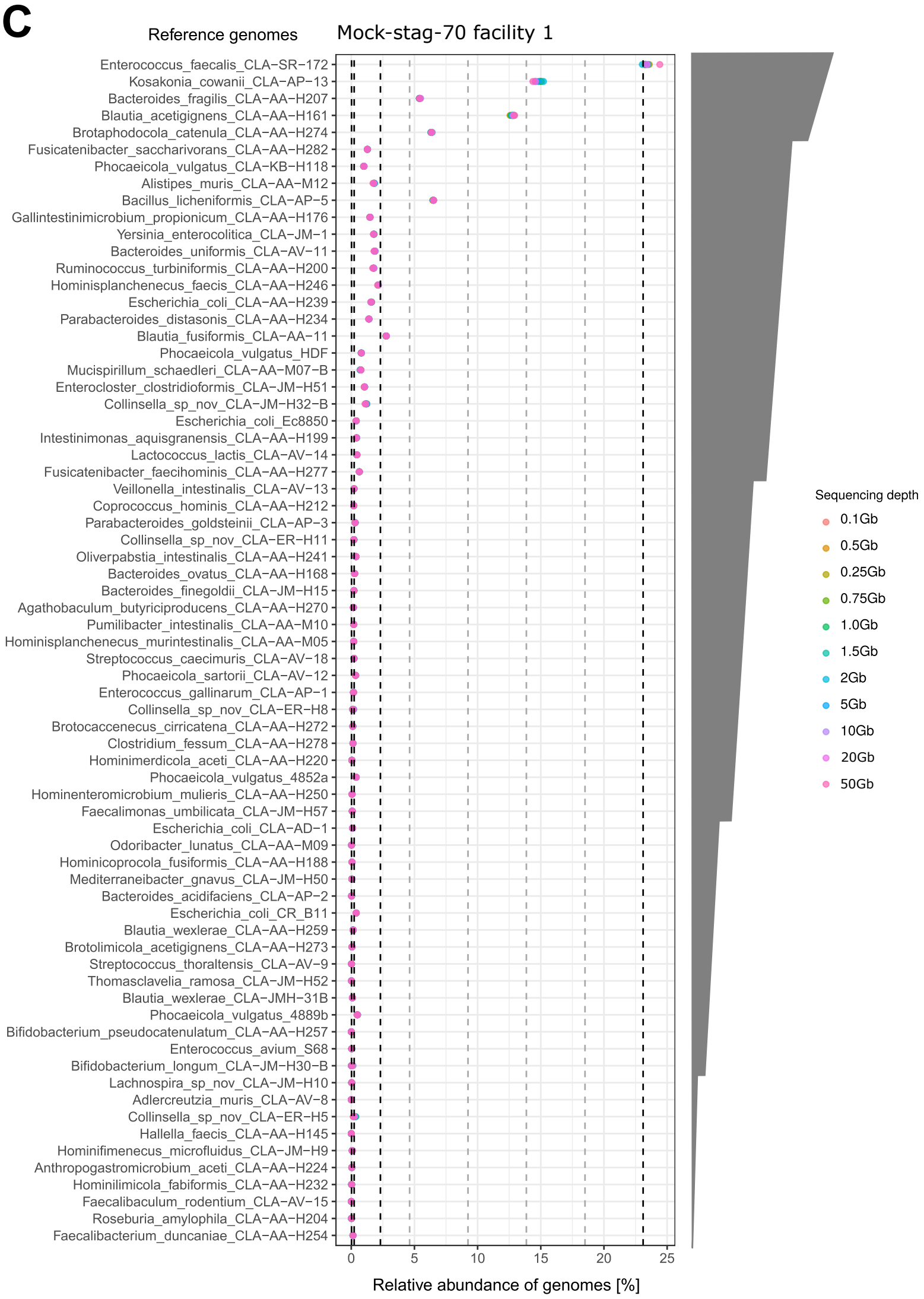

70 bacterial isolates with genome data
Mock communities
Even distribution
• 70 isolates
Staggered distributions
• 24 isolates (x11)
• 70 isolates
Shotgun sequencing (Illumina)
0.1 0.25 0.5 0.75 1.0 1.5 2.0 5.0 10 20 50 Gb/sample
0.33 0.83 1.67 2.50 3.33 5.00 6.67 16.67 33.33 66.67 166.7 x10⁶ reads
(150bp PE)
• 11 sequencing depths
• 3 sequencing facilities
• +/- background DNA
• long-read sequencing
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
Tree scale: 0.1

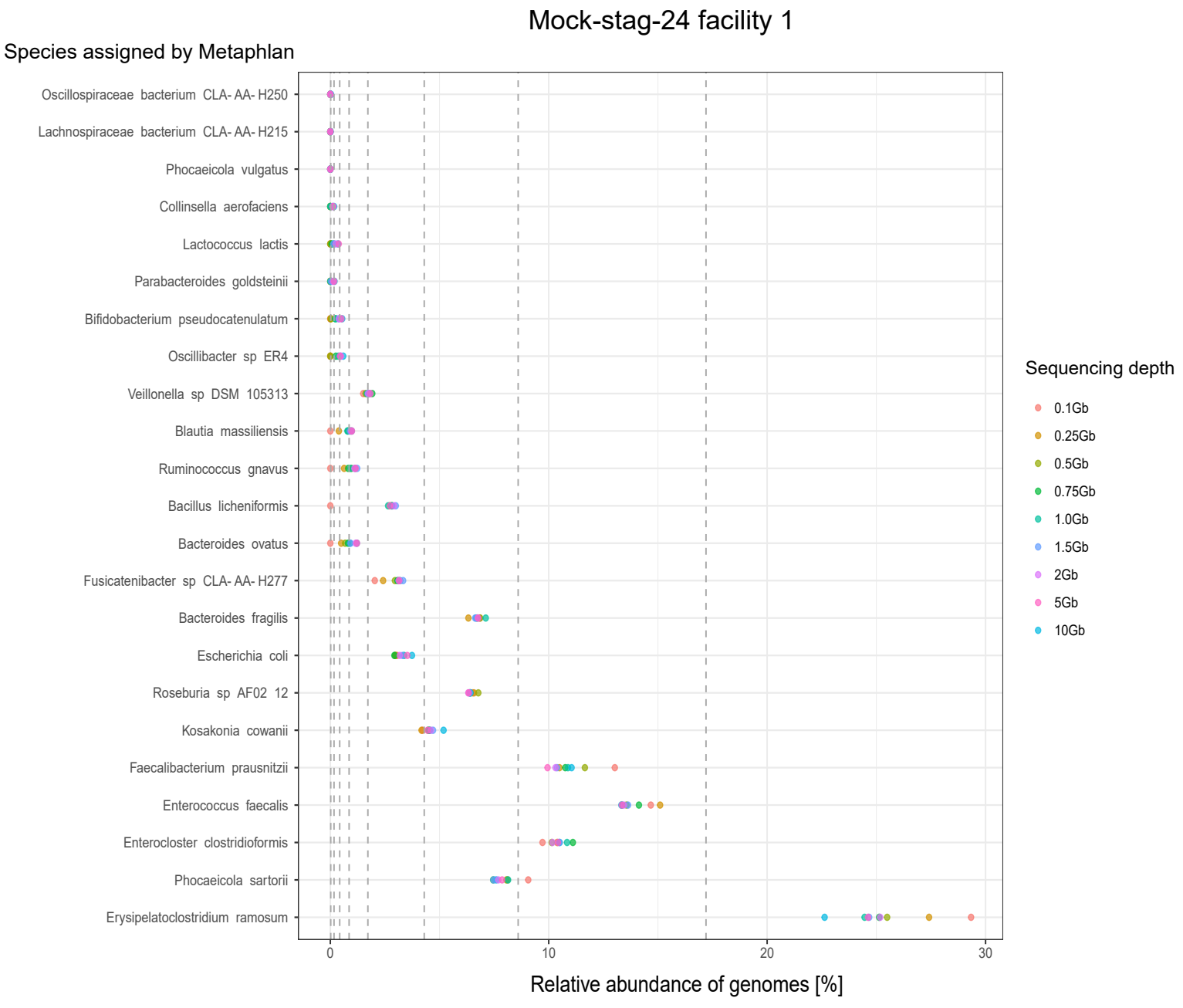

Raw reads
0.1 Gb
-
50 Gb
Mock-even-70
Mock-stag-70
Mock-stag-24
Mock-stag-24 v1-10
either
Subsampled raw reads
subsampling to the exact read targeted Gb (seqtk)
or
in silico Mocks
10 x subsampling of 50 Gb for each 0.1 - 20 Gb (seqtk)
x10
x10
x10
Pre-Processing
quality filtering and phiX removal
(trimmomatic, bbmap, bbduk)
High-quality reads
assembly
(MEGAHIT or metaSPAdes)
Contigs
alignment of contigs to reference genomes (blastn)
strain 2
strain 1
Composition of contigs
Taxonomic analysis
Alignment of reads to reference genomes (coverM)
or
Taxonomic profile with MetaPhlAn
strain 1
strain 2
strain 3
Coverage of reference genomes
Relative abundance profile
Reference based strain level analysis
Functional analysis
protein-coding gene prediction
(prodigal)
VPVLVEPQTPTESVEQSSPTELR
Predicted protein sequences (PPS)
alignment to reference protein sequences
(Diamond / blastp)
reconstruction of functional pathways
(kofamscan, KEGG-decoder)
Functional coverage
Functional pathway completeness
MAG construction
Single sample binning
Contigs > 1,500 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
Multi-coverage binning
Workflow by Mattock & Watson (2023)
10 x 24 strains in varying distributions
MAG 2
MAG 1
MAG 4
completeness and contamination check (checkM)
taxonomic assignment (coverM)
Composition of MAGs

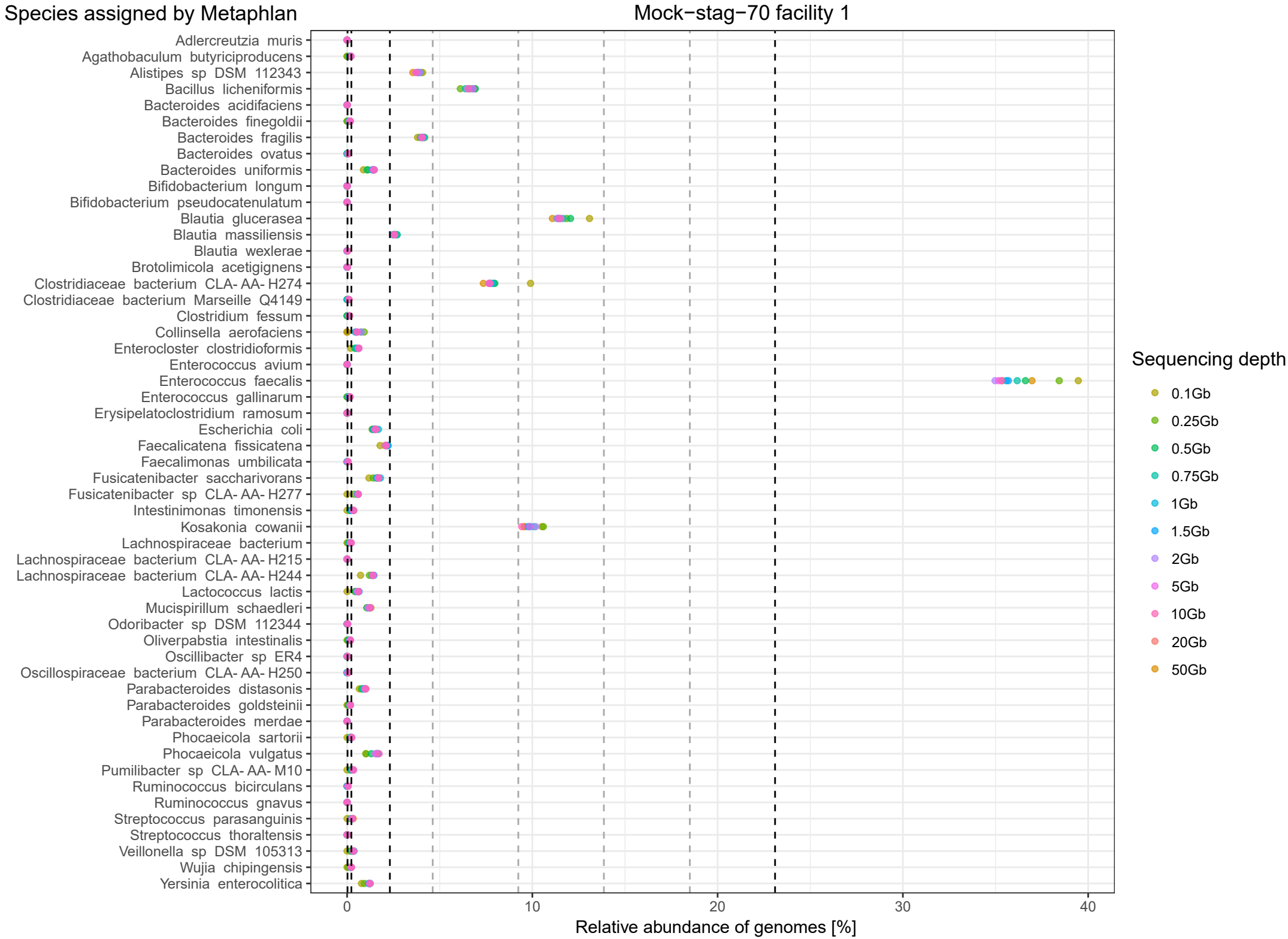

A
Mock-even-70
Adlercreutzia muris CLA-AV-8
Agathobaculum butyriciproducens CLA-AA-H270
Alistipes muris CLA-AA-M12
Anthropogastromicrobium aceti CLA-AA-H224
Bacillus licheniformis CLA-AP-5
Bacteroides acidifaciens CLA-AP-2
Bacteroides finegoldii CLA-JM-H15
Bacteroides fragilis CLA-AA-H207
Bacteroides 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-24
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-stag-70
Faecalibacterium duncaniae CLA-AA-H254
Roseburia amylophila CLA-AA-H204
Faecalibaculum rodentium CLA-AV-15
Hominilimicola fabiformis CLA-AA-H232
Anthropogastromicrobium aceti CLA-AA-H224
Hominifimenecus microfluidus CLA-JM-H9
Hallella faecis CLA-AA-H145
Collinsella sp nov CLA-ER-H5
Adlercreutzia muris CLA-AV-8
Lachnospira sp nov CLA-JM-H10
Bifidobacterium longum CLA-JM-H30-B
Enterococcus avium S68
Bifidobacterium pseudocatenulatum CLA-AA-H257
Phocaeicola vulgatus 4889b
Blautia wexlerae CLA-JMH-31B
Thomasclavelia ramosa CLA-JM-H52
Streptococcus thoraltensis CLA-AV-9
Brotolimicola acetigignens CLA-AA-H273
Blautia wexlerae CLA-AA-H259
Escherichia coli CR_B11
Bacteroides acidifaciens CLA-AP-2
Mediterraneibacter gnavus CLA-JM-H50
Hominicoprocola fusiformis CLA-AA-H188
Odoribacter lunatus CLA-AA-M09
Escherichia coli CLA-AD-1
Faecalimonas umbilicata CLA-JM-H57
Hominenteromicrobium mulieris CLA-AA-H250
Phocaeicola vulgatus 4852a
Hominimerdicola aceti CLA-AA-H220
Clostridium fessum CLA-AA-H278
Brotocaccenecus cirricatena CLA-AA-H272
Collinsella sp nov CLA-ER-H8
Enterococcus gallinarum CLA-AP-1
Phocaeicola sartorii CLA-AV-12
Streptococcus caecimuris CLA-AV-18
Hominisplanchenecus murintestinalis CLA-AA-M05
Pumilibacter intestinalis CLA-AA-M10
Agathobaculum butyriciproducens CLA-AA-H270
Bacteroides finegoldii CLA-JM-H15
Bacteroides ovatus CLA-AA-H168
Oliverpabstia intestinalis CLA-AA-H241
Collinsella sp nov CLA-ER-H11
Parabacteroides goldsteinii CLA-AP-3
Coprococcus hominis CLA-AA-H212
Veillonella intestinalis CLA-AV-13
Fusicatenibacter faecihominis CLA-AA-H277
Lactococcus lactis CLA-AV-14
Intestinimonas aquisgranensis CLA-AA-H199
Escherichia coli Ec8850
Collinsella sp nov CLA-JM-H32-B
Enterocloster clostridioformis CLA-JM-H51
Mucispirillum schaedleri CLA-AA-M07-B
Phocaeicola vulgatus HDF
Blautia fusiformis CLA-AA-11
Parabacteroides distasonis CLA-AA-H234
Escherichia coli CLA-AA-H239
Hominisplanchenecus faecis CLA-AA-H246
Ruminococcus turbiniformis CLA-AA-H200
Bacteroides uniformis CLA-AV-11
Yersinia enterocolitica CLA-JM-1
Gallintestinimicrobium propionicum CLA-AA-H176
Bacillus licheniformis CLA-AP-5
Alistipes muris CLA-AA-M12
Phocaeicola vulgatus CLA-KB-H118
Fusicatenibacter saccharivorans CLA-AA-H282
Brotaphodocola catenula CLA-AA-H274
Blautia acetigignens CLA-AA-H161
Bacteroides fragilis CLA-AA-H207
Kosakonia cowanii CLA-AP-13
Enterococcus faecalis CLA-SR-172
0.1 0.25 0.5 0.75 1.0 1.5 2 5 10 20 50
Sequencing depth [Gb]
coverage [%]
1.00 0.75 0.50 0.25


A
Reference genomes
Mock-even-70 facility 1
Mock-even-70 facility 2
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 genomes [%]
Relative abundance of genomes [%]
Sequencing depth
0.1Gb
0.5Gb
0.25Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb
B
Reference genomes
Mock-stag-24 facility 1
Mock-stag-24 facility 2
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 genomes [%]
Relative abundance of genomes [%]
Sequencing depth
0.1Gb
0.5Gb
0.25Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb

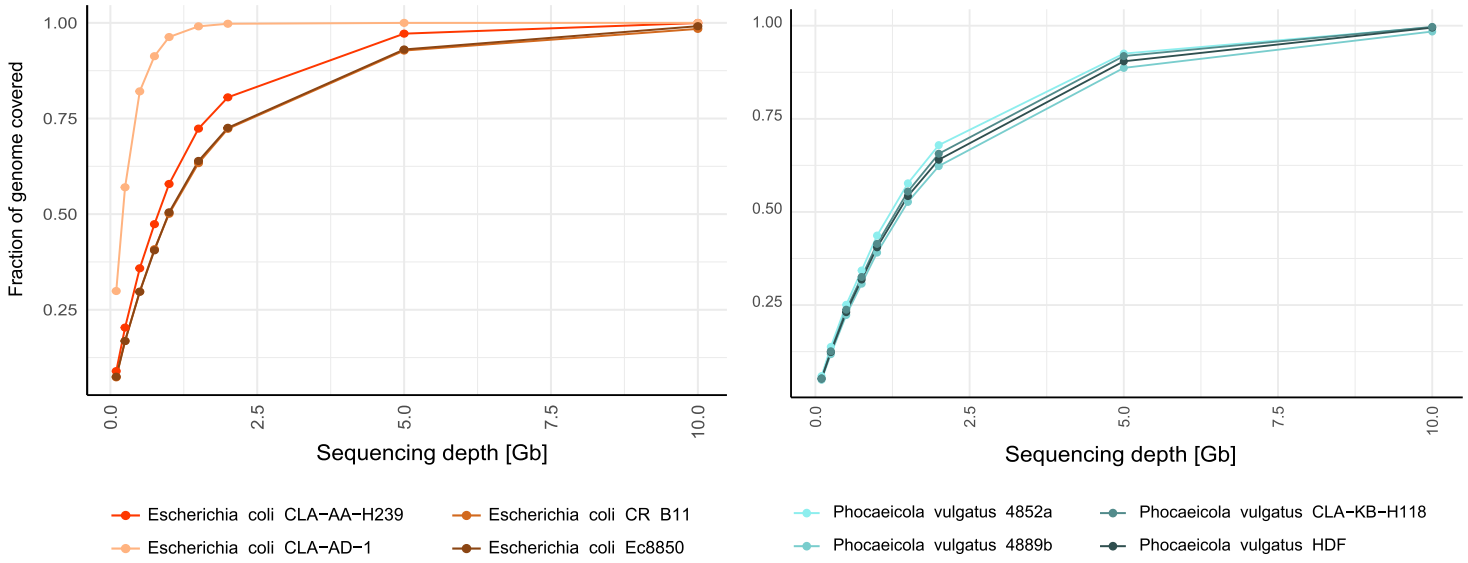

C
Reference genomes
Mock-stag-70 facility 1
Enterococcus_faecalis_CLA-SR-172
Kosakonia_cowanii_CLA-AP-13
Bacteroides_fragilis_CLA-AA-H207
Blautia_acetigignens_CLA-AA-H161
Brotaphodocola_catenula_CLA-AA-H274
Fusicatenibacter_saccharivorans_CLA-AA-H282
Phocaeicola_vulgatus_CLA-KB-H118
Alistipes_muris_CLA-AA-M12
Bacillus_licheniformis_CLA-AP-5
Gallintestinimicrobium_propionicum_CLA-AA-H176
Yersinia_enterocolitica_CLA-JM-1
Bacteroides_uniformis_CLA-AV-11
Ruminococcus_turbiniformis_CLA-AA-H200
Hominisplanchenecus_faecis_CLA-AA-H246
Escherichia_coli_CLA-AA-H239
Parabacteroides_distasonis_CLA-AA-H234
Blautia_fusiformis_CLA-AA-11
Phocaeicola_vulgatus_HDF
Mucispirillum_schaedleri_CLA-AA-M07-B
Enterocloster_clostridioformis_CLA-JM-H51
Collinsella_sp_nov_CLA-JM-H32-B
Escherichia_coli_Ec8850
Intestinimonas_aquisgranensis_CLA-AA-H199
Lactococcus_lactis_CLA-AV-14
Fusicatenibacter_faecihominis_CLA-AA-H277
Veillonella_intestinalis_CLA-AV-13
Coprococcus_hominis_CLA-AA-H212
Parabacteroides_goldsteinii_CLA-AP-3
Collinsella_sp_nov_CLA-ER-H11
Oliverpabstia_intestinalis_CLA-AA-H241
Bacteroides_ovatus_CLA-AA-H168
Bacteroides_finegoldii_CLA-JM-H15
Agathobaculum_butyriciproducens_CLA-AA-H270
Pumilibacter_intestinalis_CLA-AA-M10
Hominisplanchenecus_murintestinalis_CLA-AA-M05
Streptococcus_caecimuris_CLA-AV-18
Phocaeicola_sartorii_CLA-AV-12
Enterococcus_gallinarum_CLA-AP-1
Collinsella_sp_nov_CLA-ER-H8
Brotocaccenecus_cirricatena_CLA-AA-H272
Clostridium_fessum_CLA-AA-H278
Hominimerdicola_aceti_CLA-AA-H220
Phocaeicola_vulgatus_4852a
Hominenteromicrobium_mulieris_CLA-AA-H250
Faecalimonas_umbilicata_CLA-JM-H57
Escherichia_coli_CLA-AD-1
Odoribacter_lunatus_CLA-AA-M09
Hominicoprocola_fusiformis_CLA-AA-H188
Mediterraneibacter_gnavus_CLA-JM-H50
Bacteroides_acidifaciens_CLA-AP-2
Escherichia_coli_CR_B11
Blautia_wexlerae_CLA-AA-H259
Brotolimicola_acetigignens_CLA-AA-H273
Streptococcus_thoraltensis_CLA-AV-9
Thomasclavelia_ramosa_CLA-JM-H52
Blautia_wexlerae_CLA-JMH-31B
Phocaeicola_vulgatus_4889b
Bifidobacterium_pseudocatenulatum_CLA-AA-H257
Enterococcus_avium_S68
Bifidobacterium_longum_CLA-JM-H30-B
Lachnospira_sp_nov_CLA-JM-H10
Adlercreutzia_muris_CLA-AV-8
Collinsella_sp_nov_CLA-ER-H5
Hallella_faecis_CLA-AA-H145
Hominifimenecus_microfluidus_CLA-JM-H9
Anthropogastromicrobium_aceti_CLA-AA-H224
Hominilimicola_fabiformis_CLA-AA-H232
Faecalibaculum_rodentium_CLA-AV-15
Roseburia_amylophila_CLA-AA-H204
Faecalibacterium_duncaniae_CLA-AA-H254
0
5
10
15
20
25
Relative abundance of genomes [%]
Sequencing depth
0.1Gb
0.5Gb
0.25Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb
20Gb
50Gb

Mock-even-70 facility 1
Species assigned by Metaphlan
Adlercreutzia muris
Agathobaculum butyriciproducens
Alistipes sp DSM 112343
Bacillus licheniformis
Bacteroides acidifaciens
Bacteroides finegoldii
Bacteroides fragilis
Bacteroides ovatus
Bacteroides uniformis
Bifidobacterium longum
Bifidobacterium pseudocatenulatum
Blautia glucerasea
Blautia massiliensis
Blautia wexlerae
Brotolimicola acetigignens
Clostridiaceae bacterium
Clostridiaceae bacterium CLA- AA- H274
Clostridiaceae bacterium Marseille Q4149
Clostridiales bacterium KLE1615
Clostridium fessum
Collinsella aerofaciens
Enterocloster clostridioformis
Enterococcus avium
Enterococcus faecalis
Enterococcus gallinarum
Erysipelatoclostridium ramosum
Escherichia coli
Faecalibacterium prausnitzii
Faecalibaculum rodentium
Faecalicatena fissicatena
Faecalimonas umbilicata
Fusicatenibacter saccharivorans
Fusicatenibacter sp CLA- AA- H277
Hallella faecis
Intestinimonas timonensis
Kosakonia cowanii
Lachnospira sp NSJ 43
Lachnospiraceae bacterium
Lachnospiraceae bacterium CLA- AA- H215
Lachnospiraceae bacterium CLA- AA- H244
Lactococcus lactis
Mucispirillum schaedleri
Odoribacter sp DSM 112344
Oliverpabstia intestinalis
Oscillibacter sp ER4
Oscillospiraceae bacterium CLA- AA- H250
Parabacteroides distasonis
Parabacteroides goldsteinii
Phocaeicola sartorii
Phocaeicola vulgatus
Pumilibacter sp CLA- AA- M10
Roseburia sp AF02 12
Ruminococcus bicirculans
Ruminococcus gnavus
Streptococcus parasanguinis
Streptococcus thoraltensis
Veillonella sp DSM 105313
Wujia chipingensis
Yersinia enterocolitica
0
5
10
15
20
25
Relative abundance of genomes [%]
Sequencing depth
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb

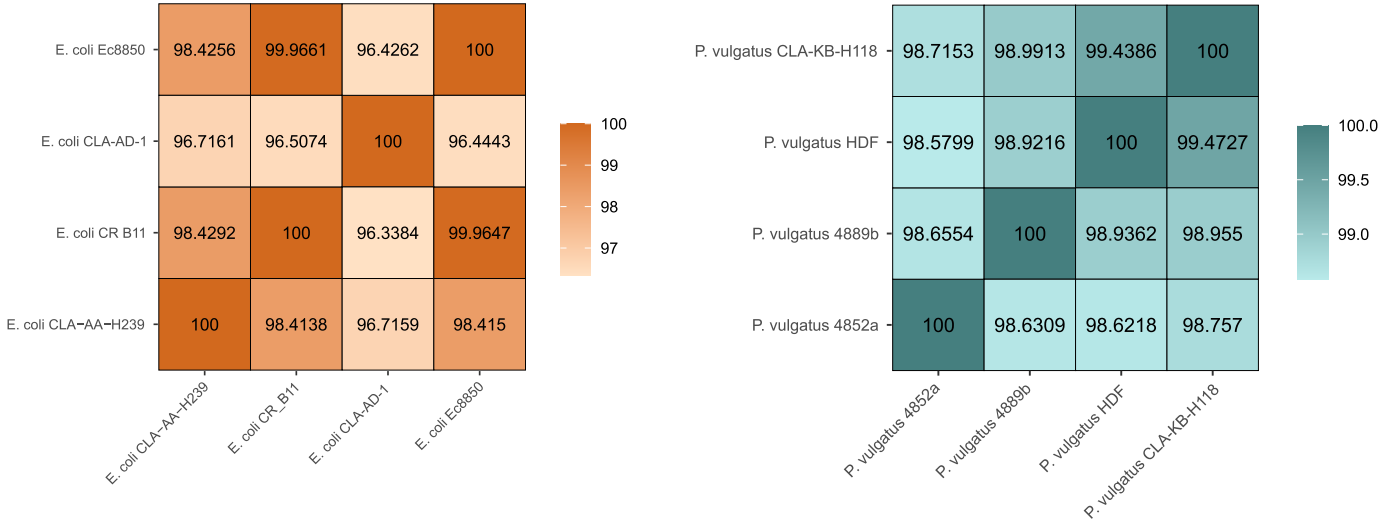

Mock-stag-24 facility 1
Species assigned by Metaphlan
Oscillospiraceae bacterium CLA- AA- H250
Lachnospiraceae bacterium CLA- AA- H215
Phocaeicola vulgatus
Collinsella aerofaciens
Lactococcus lactis
Parabacteroides goldsteinii
Bifidobacterium pseudocatenulatum
Oscillibacter sp ER4
Veillonella sp DSM 105313
Blautia massiliensis
Ruminococcus gnavus
Bacillus licheniformis
Bacteroides ovatus
Fusicatenibacter sp CLA- AA- H277
Bacteroides fragilis
Escherichia coli
Roseburia sp AF02 12
Kosakonia cowanii
Faecalibacterium prausnitzii
Enterococcus faecalis
Enterocloster clostridioformis
Phocaeicola sartorii
Erysipelatoclostridium ramosum
0
10
20
30
Relative abundance of genomes [%]
Sequencing depth
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1.0Gb
1.5Gb
2Gb
5Gb
10Gb

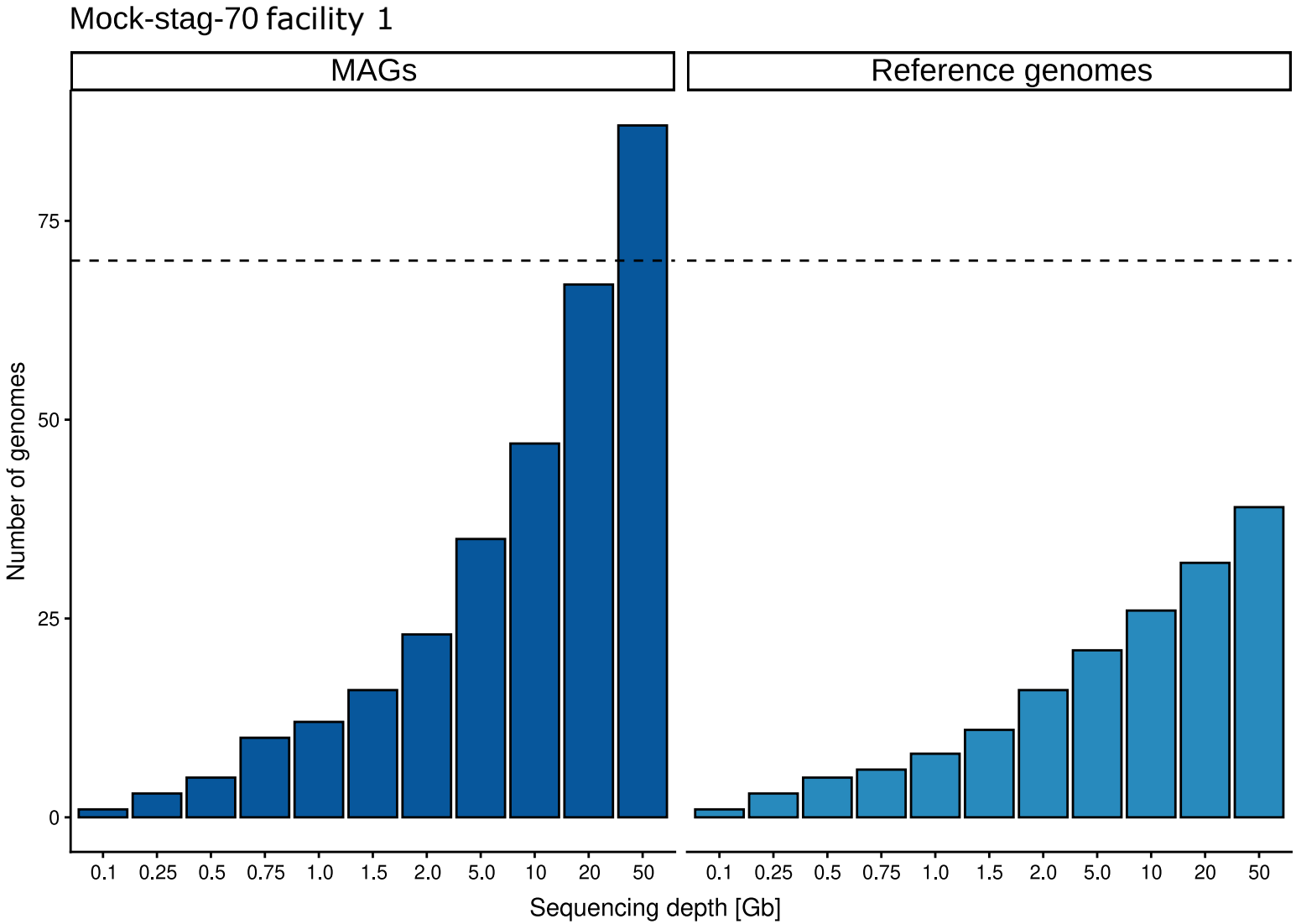

Species assigned by Metaphlan
Mock−stag−70 facility 1
Adlercreutzia muris
Agathobaculum butyriciproducens
Alistipes sp DSM 112343
Bacillus licheniformis
Bacteroides acidifaciens
Bacteroides finegoldii
Bacteroides fragilis
Bacteroides ovatus
Bacteroides uniformis
Bifidobacterium longum
Bifidobacterium pseudocatenulatum
Blautia glucerasea
Blautia massiliensis
Blautia wexlerae
Brotolimicola acetigignens
Clostridiaceae bacterium CLA-AA-H274
Clostridiaceae bacterium Marseille Q4149
Clostridium fessum
Collinsella aerofaciens
Enterocloster clostridioformis
Enterococcus avium
Enterococcus faecalis
Enterococcus gallinarum
Erysipelatoclostridium ramosum
Escherichia coli
Faecalicatena fissicatena
Faecalimonas umbilicata
Fusicatenibacter saccharivorans
Fusicatenibacter sp CLA-AA-H277
Intestinimonas timonensis
Kosakonia cowanii
Lachnospiraceae bacterium
Lachnospiraceae bacterium CLA-AA-H215
Lachnospiraceae bacterium CLA-AA-H244
Lactococcus lactis
Mucispirillum schaedleri
Odoribacter sp DSM 112344
Oliverpabstia intestinalis
Oscillibacter sp ER4
Oscillospiraceae bacterium CLA-AA-H250
Parabacteroides distasonis
Parabacteroides goldsteinii
Parabacteroides merdae
Phocaeicola sartorii
Phocaeicola vulgatus
Pumilibacter sp CLA-AA-M10
Ruminococcus bicirculans
Ruminococcus gnavus
Streptococcus parasanguinis
Streptococcus thoraltensis
Veillonella sp DSM 105313
Wujia chipingensis
Yersinia enterocolitica
0
10
20
30
40
Relative abundance of genomes [%]
Sequencing depth
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1Gb
1.5Gb
2Gb
5Gb
10Gb
20Gb
50Gb

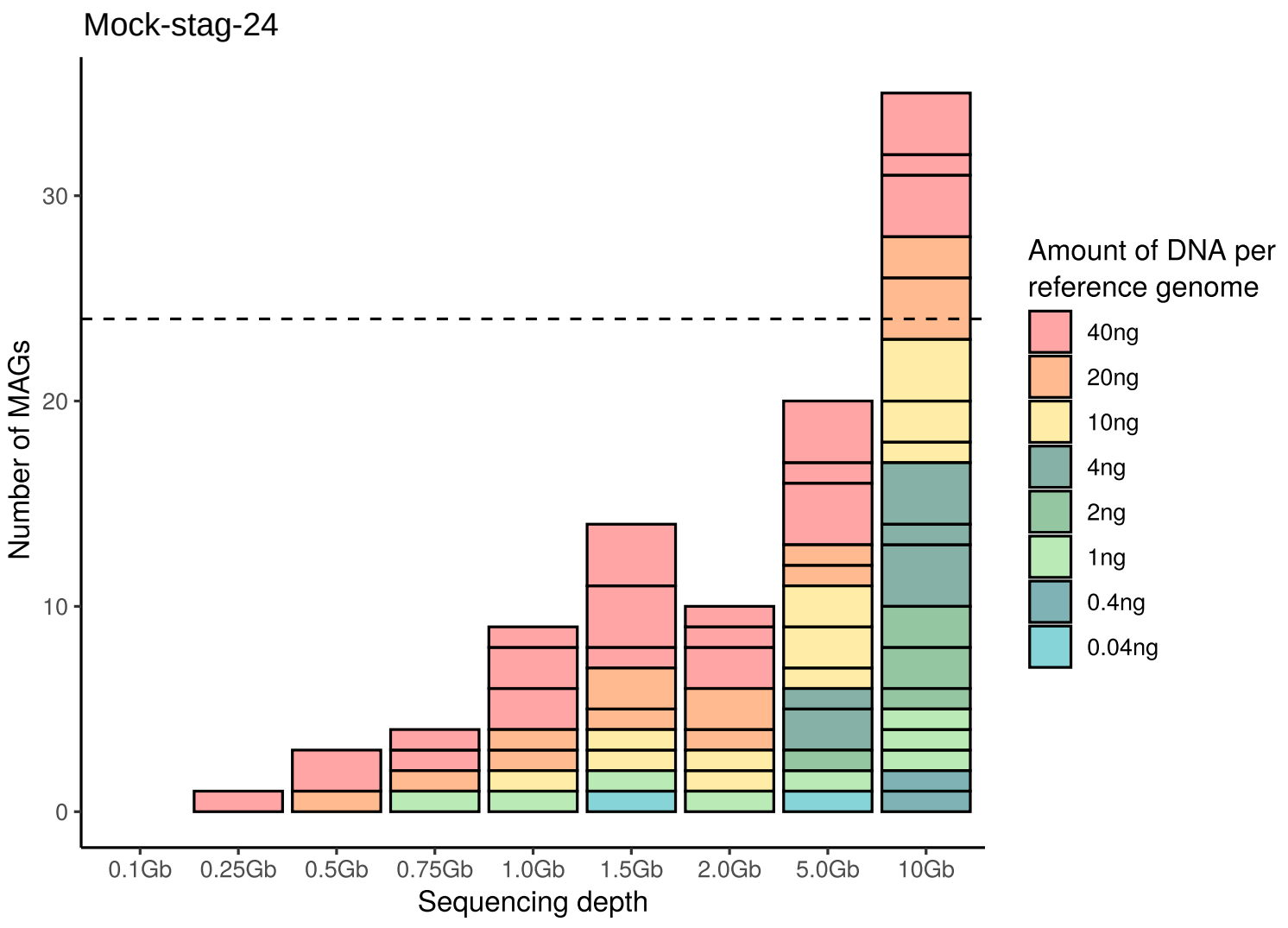
A

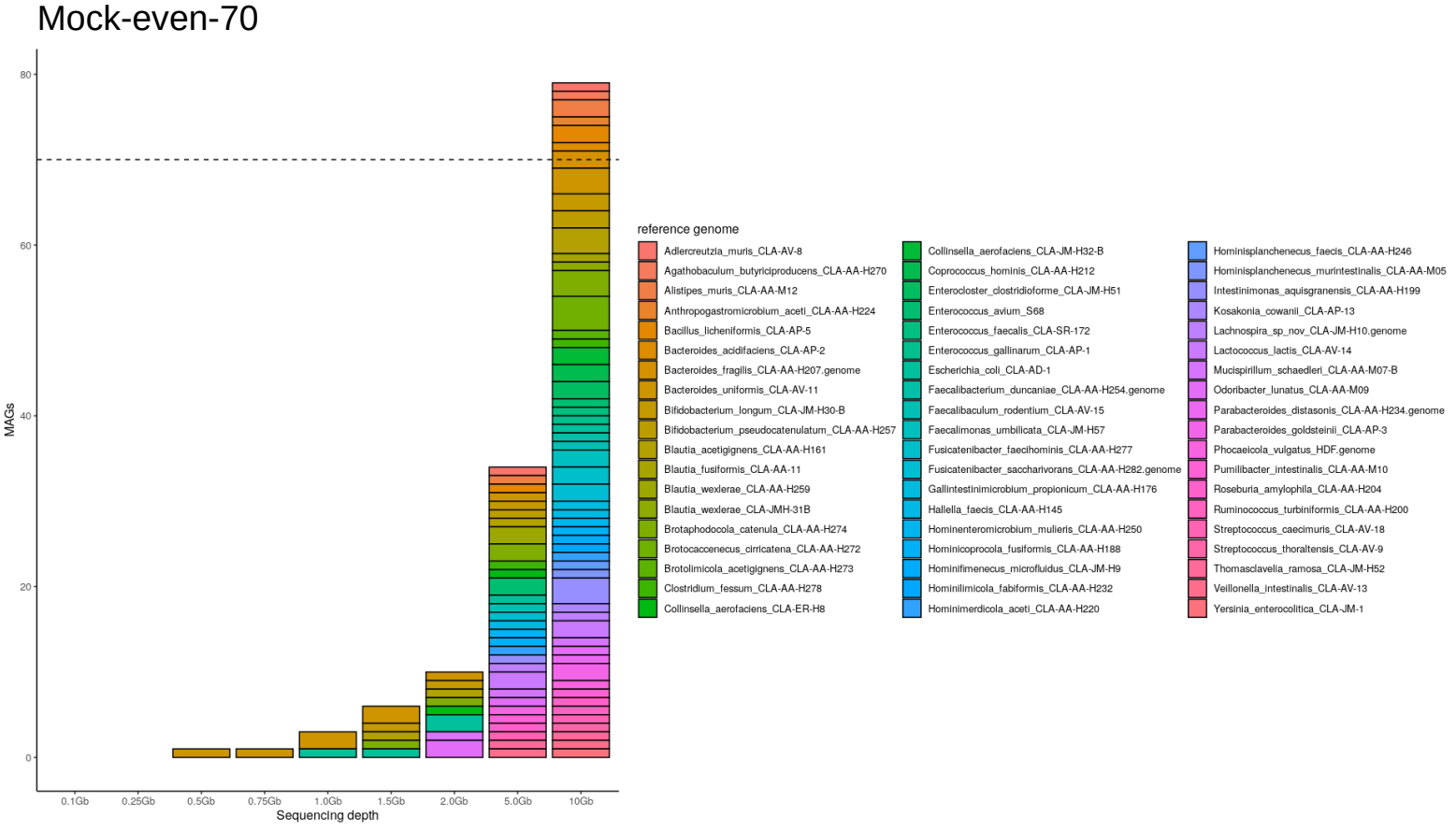

Fraction of genome covered
1.00
0.75
0.50
0.25
0.0
2.5
5.0
7.5
10.0
Sequencing depth [Gb]
Escherichia coli CLA−AA−H239
Escherichia coli CR B11
Escherichia coli CLA−AD−1
Escherichia coli Ec8850
Phocaeicola vulgatus 4852a
Phocaeicola vulgatus CLA−KB−H118
Phocaeicola vulgatus 4889b
Phocaeicola vulgatus HDF

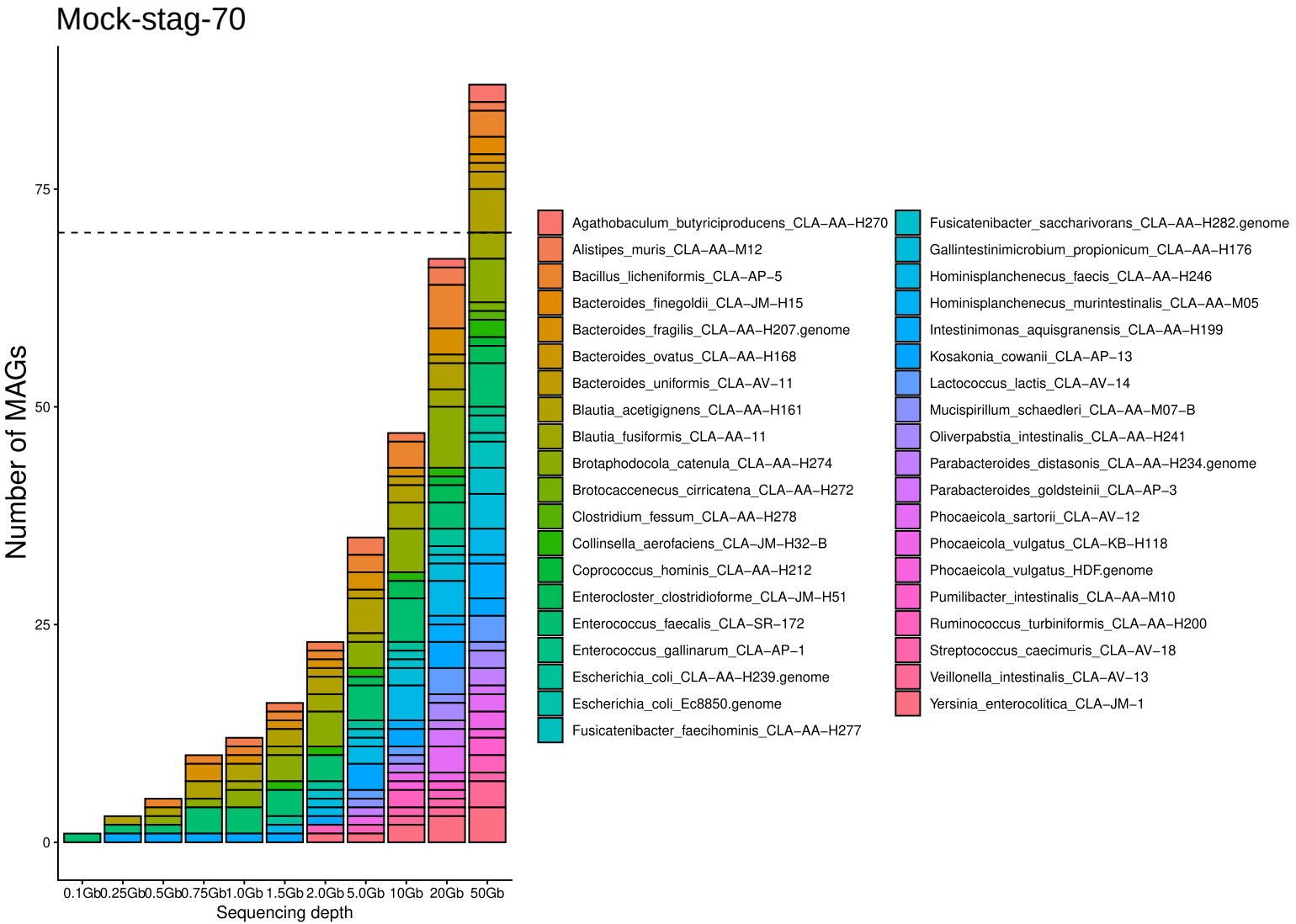
B

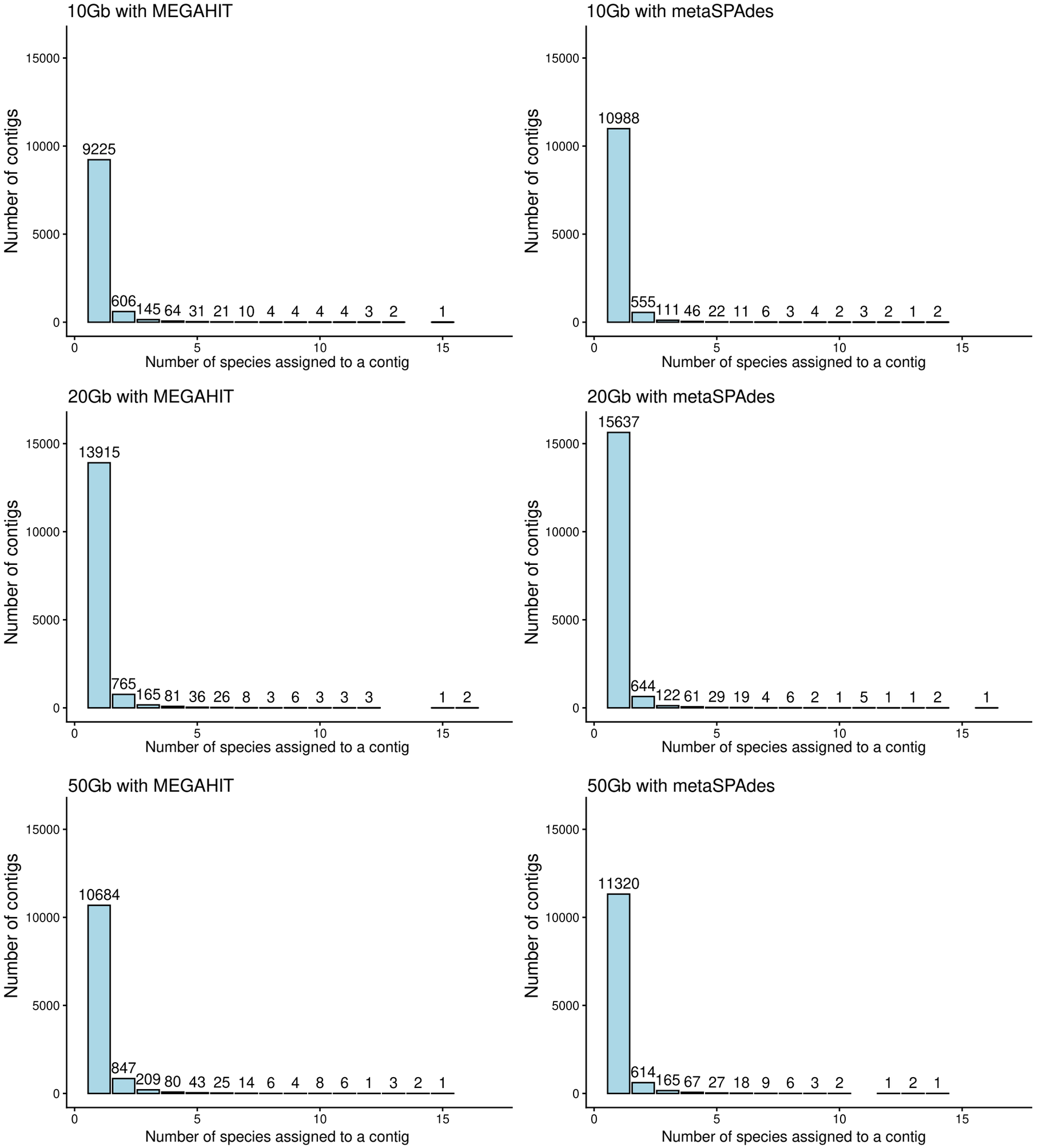

E. coli Ec8850
E. coli CLA-AD-1
E. coli CR B11
E. coli CLA−AA−H239
98.4256
99.9661
96.4262
100
96.7161
96.5074
100
96.4443
98.4292
100
96.3384
99.9647
100
98.4138
96.7159
98.415
E. coli CLA-AA-H239
E. coli CR_B11
E. coli CLA-AD-1
E. coli Ec8850
100
99
98
97
P. vulgatus CLA-KB-H118
P. vulgatus HDF
P. vulgatus 4889b
P. vulgatus 4852a
98.7153
98.9913
99.4386
100
98.5799
98.9216
100
99.4727
98.6554
100
98.9362
98.955
100
98.6309
98.6218
98.757
P. vulgatus 4852a
P. vulgatus 4889b
P. vulgatus HDF
P. vulgatus CLA-KB-H118
100.0
99.5
99.0


Mock-stag-70 facility 1
MAGs
Reference genomes
Number of genomes
0
25
50
75
0.1
0.25
0.5
0.75
1.0
1.5
2.0
5.0
10
20
50
Sequencing depth [Gb]

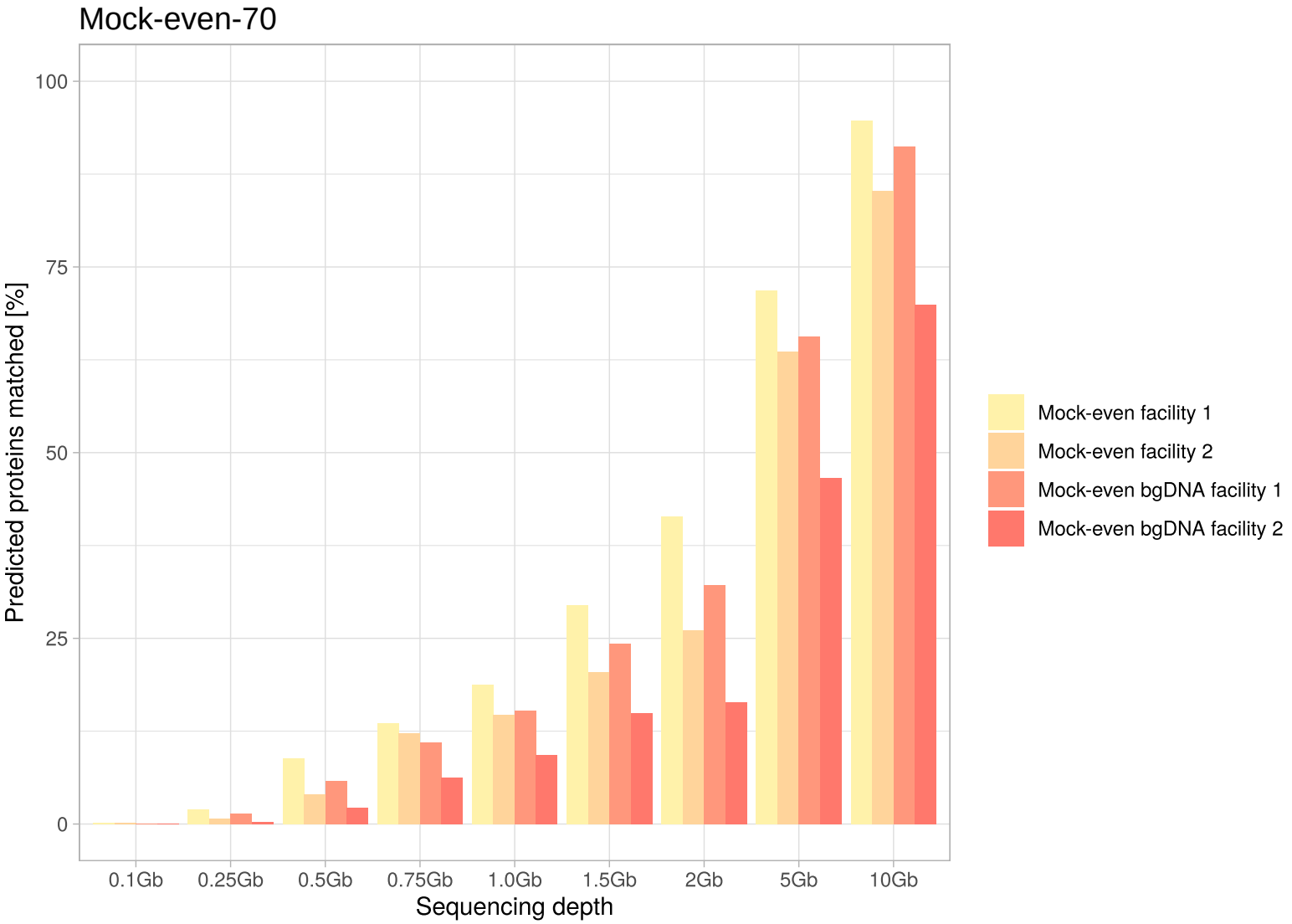

Mock-stag-24
Number of MAGs
0
10
20
30
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1.0Gb
1.5Gb
2.0Gb
5.0Gb
10Gb
Sequencing depth
Amount of DNA per reference genome
40ng
20ng
10ng
4ng
2ng
1ng
0.4ng
0.04ng

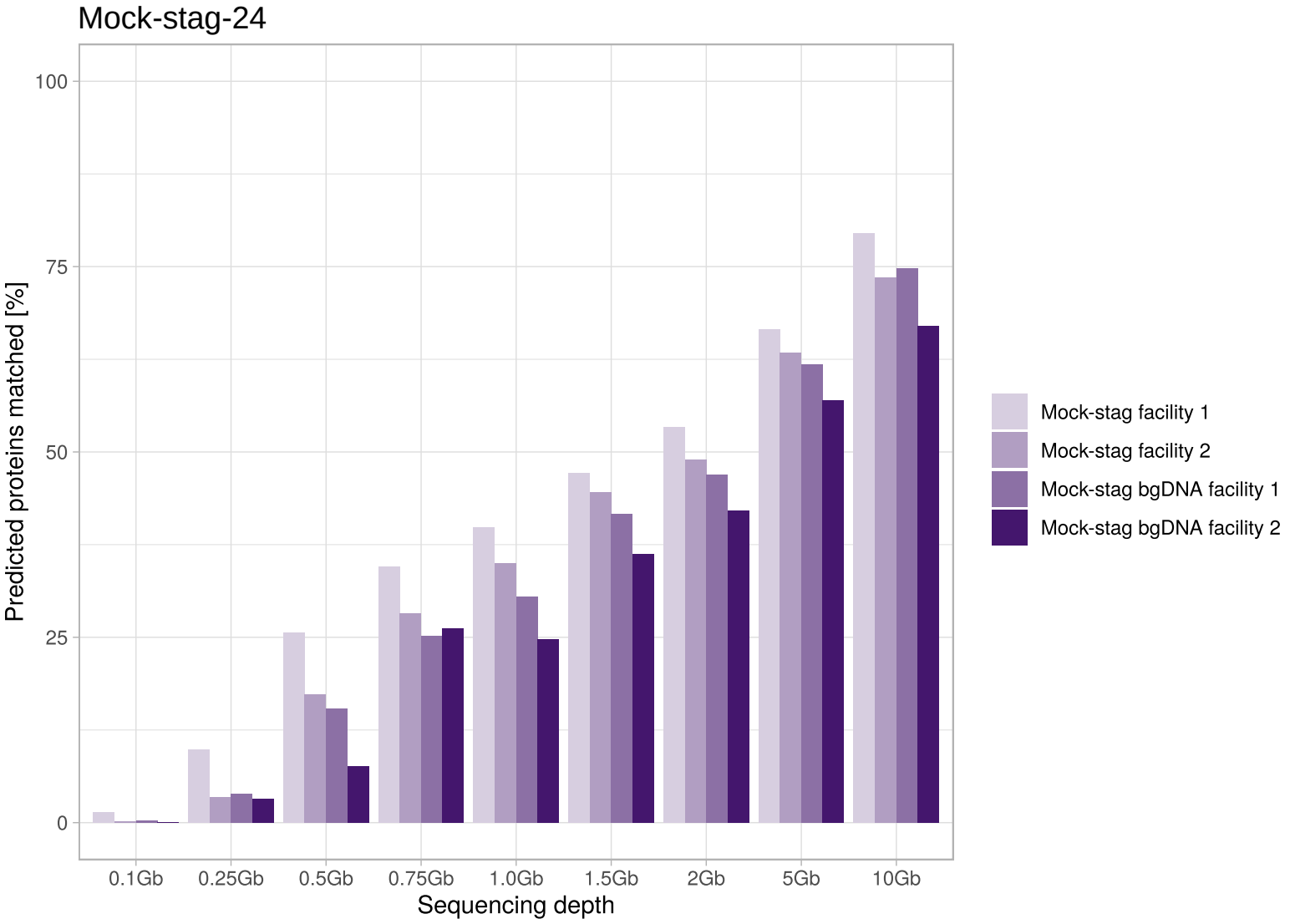

Mock-even-70
MAGs
0
20
40
60
80
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1.0Gb
1.5Gb
2.0Gb
5.0Gb
10Gb
Sequencing depth
reference genome
Adlercreutzia_muris_CLA-AV-8
Agathobaculum_butyriciproducens_CLA-AA-H270
Alistipes_muris_CLA-AA-M12
Anthropogastromicrobium_aceti_CLA-AA-H224
Bacillus_licheniformis_CLA-AP-5
Bacteroides_acidifaciens_CLA-AP-2
Bacteroides_fragilis_CLA-AA-H207.genome
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_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-AD-1
Faecalibacterium_duncaniae_CLA-AA-H254.genome
Faecalibaculum_rodentium_CLA-AV-15
Faecalimonas_umbilicata_CLA-JM-H57
Fusicatenibacter_faecihominis_CLA-AA-H277
Fusicatenibacter_saccharivorans_CLA-AA-H282.genome
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.genome
Lactococcus_lactis_CLA-AV-14
Mucispirillum_schaedleri_CLA-AA-M07-B
Odoribacter_lunatus_CLA-AA-M09
Parabacteroides_distasonis_CLA-AA-H234.genome
Parabacteroides_goldsteinii_CLA-AP-3
Phocaeicola_vulgatus_HDF.genome
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
Thomasclavella_ramosa_CLA-JM-H52
Veillonella_intestinalis_CLA-AV-13
Yersinia_enterocolitica_CLA-JM-1


Mock-stag-70
Number of MAGs
0
25
50
75
0.1Gb
0.25Gb
0.5Gb
0.75Gb
1.0Gb
1.5Gb
2.0Gb
5.0Gb
10Gb
20Gb
50Gb
Sequencing depth
Agathobaculum_butyriciproducens_CLA–AA–H270
Alistipes_muris_CLA–AA–M12
Bacillus_licheniformis_CLA–AP–5
Bacteroides_finegoldii_CLA–JM–H15
Bacteroides_fragilis_CLA–AA–H207.genome
Bacteroides_ovatus_CLA–AA–H168
Bacteroides_uniformis_CLA–AV–11
Blautia_acetigignens_CLA–AA–H161
Blautia_fusiformis_CLA–AA–11
Brotaphodocola_catenula_CLA–AA–H274
Brotocaccenecus_cirricatena_CLA–AA–H272
Clostridium_fessum_CLA–AA–H278
Collinsella_aerofaciens_CLA–JM–H32–B
Coprococcus_hominis_CLA–AA–H212
Enterocloster_clostridioforme_CLA–JM–H51
Enterococcus_faecalis_CLA–SR–172
Enterococcus_gallinarum_CLA–AP–1
Escherichia_coli_CLA–AA–H239.genome
Escherichia_coli_Ec8850.genome
Fusicatenibacter_faecihominis_CLA–AA–H277
Fusicatenibacter_saccharivorans_CLA–AA–H282.genome
Gallintestinimicrobium_propionicum_CLA–AA–H176
Hominisplanchenecus_faecis_CLA–AA–H246
Hominisplanchenecus_murintestinalis_CLA–AA–M05
Intestinimonas_aquisgranensis_CLA–AA–H199
Kosakonia_cowanii_CLA–AP–13
Lactococcus_lactis_CLA–AV–14
Mucispirillum_schaedleri_CLA–AA–M07–B
Oliverpabstia_intestinalis_CLA–AA–H241
Parabacteroides_distasonis_CLA–AA–H234.genome
Parabacteroides_goldsteinii_CLA–AP–3
Phocaeicola_sartorii_CLA–AV–12
Phocaeicola_vulgatus_CLA–KB–H118
Phocaeicola_vulgatus_HDF.genome
Pumilibacter_intestinalis_CLA–AA–M10
Ruminococcus_turbiniformis_CLA–AA–H200
Streptococcus_caecimuris_CLA–AV–18
Veillonella_intestinalis_CLA–AV–13
Yersinia_enterocolitica_CLA–JM–1


10Gb with MEGAHIT
Number of contigs
Number of species assigned to a contig
9225
606
145
64
31
21
10
4
4
4
4
3
2
1
10Gb with metaSPAdes
10988
555
111
46
22
11
6
3
4
2
3
2
1
2
20Gb with MEGAHIT
13915
765
165
81
36
26
8
3
6
3
3
3
1
2
20Gb with metaSPAdes
15637
644
122
61
29
19
4
6
2
1
5
1
1
2
1
50Gb with MEGAHIT
10684
847
209
80
43
25
14
6
4
8
6
1
3
2
1
50Gb with metaSPAdes
11320
614
165
67
27
18
9
6
3
2
1
2
1

KEGG pathway
MEGAHIT
metaSPAdes
Mock-even-70
Mock-stag-24
Mock-stag-70
Mock-stag-70
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


Mock-even-70
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-stag-24
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