Broad diversity of human gut bacteria accessible via a traceable strain deposition system

Thomas C. A. Hitch[1], Johannes M. Masson[1,#], Charlie Pauvert[1,#], Johanna Bosch[1], Selina Nüchtern[1], Nicole Treichel[1], Marko Baloh[1], Soheila Razavi[1], Afrizal Afrizal[1], Ntana Kousetzi[1], Andrea M. Aguirre[1], David Wylensek[1], Amy Coates[1], Susan A. V. Jennings[1], Atscharah Panyot[1], Alina Viehof[1], Matthias A. Schmitz[1], Maximilian Stuhrmann[1], Evelyn C. Deis[1], Kevin Bisdorf[1], Thorsten Cramer[2], Artur Lissin[3], Isabel Schober[3], Julius Witte[3], Thomas Riedel[3], Marie Wende[4], Katrin A. Winter[4], Alessandra Riva[5,6], Stefanie Trinh[7], Laura Mitchell[8], Jonathan Hartman[9], David Berry[5], Jochen Seitz[10], Lukas C. Bossert[9], Thorsten Allers[8], Till Strowig[4,11,12], Birte Abt[3,12], Lorenz C. Reimer[3], Jörg Overmann[3,12,13], Thomas Clavel[1*]

[1] Functional Microbiome Research Group, Institute of Medical Microbiology, University Hospital of RWTH Aachen, Germany

[2] Molecular Tumor Biology Research Group, Department of General, Visceral, Children and Transplantation Surgery, University Hospital of RWTH Aachen, Germany

[3] Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

[4] Department of Microbial Immune Regulation, Helmholtz Centre for Infection Research, Braunschweig, Germany

[5] University of Vienna, Center for Microbiology and Environmental Systems Science, Department of Microbiology and Ecosystem Science, Vienna, Austria

[6] Chair of Nutrition and Immunology, School of Life Sciences, Technische Universität München, Freising-Weihenstephan, Germany

[7] Institute of Neuroanatomy, University Hospital of RWTH Aachen, Germany

[8] School of Life Sciences, University of Nottingham, Nottingham, UK

[9] IT center, RWTH Aachen, Germany

[10] Clinic for Child and Adolescent Psychiatry, Psychosomatic Medicine and Psychotherapy, LVR-University Hospital Essen, University of Duisburg-Essen, Germany

[11] Centre for Individualised Infection Medicine (CiiM), a joint venture between the Helmholtz-Centre for Infection Research (HZI) and the Hannover Medical School (MHH), Hannover, Germany

[12] German Centre for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Braunschweig, Germany

[13] Technical University Braunschweig, Braunschweig, Germany

[#] Authors contributed equally to the manuscript

*Correspondence: tclavel@ukaachen.de

Abstract

Numerous bacteria in the human gut microbiome remains unknown and/or have yet to be cultured. While collections of human gut bacteria have been published, none are publicly accessible. This is partly due to issues with the deposition of strains to public culture collections. We address these key issues (microbial unknowns, lack of public access to isolates) by proposing a framework that facilitates large-scale submission of isolates, exemplified by a rich collection of human gut isolates that is taxonomically described, functionally novel, and made publicly available. An innovative bulk strain submission system was established by large-scale generation and curation of persistent strain identifiers (https://straininfo.dsmz.de). This was used to create the Human intestinal Bacterial Collection (www.hibc.rwth-aachen.de), which contains 340 strains representing 198 species within 29 families and 7 phyla, of which 32 previously unknown taxa are taxonomically described and named. These included three butyrate-producing species of *Faecalibacterium* and new dominant species associated with health and inflammatory bowel disease. Plasmids were identified to be prolific within the HiBC isolates, with almost half (46%) of strains containing one or more plasmids (up to 6). Megaplasmids were identified within two strains, one of them is globally present within multiple *Bacteroidales* species. The strain submission framework and public collection of easily searchable and available gut bacterial isolates presented here will facilitate the work of many by enabling functional studies.

Introduction

The cultivation of human gut bacteria has accelerated in the last years [1–4], providing valuable information on the presence of novel taxa within gut microbiomes. Yet while these published collections of human gut isolates note the novel taxa within their collections [2,5,6], rarely is this novelty made known by naming, describing, and validating the taxa [4]. Curating the taxonomic assignment of such collections is essential as often outdated names are used, or recently described taxa are ignored, causing greater confusion in the current taxonomic sphere [7–9]. Such curated taxonomy ensures strains are correctly assigned, allowing strain-level diversity to be studied [10]. Variation between strains of the same species can lead to functional shifts that alter the association with host dietary habits [11,12]. One way strains can vary is in the presence of mobile genetic elements, which are known to affect the phenotype of the isolates they inhabit [13]. The study of plasmids within the human gut has been limited until recently [14,15].

Accessibility to collections of strains from the human gut is essential for the mechanistic study of microbe-microbe and microbe-host interactions [16–18]. However, accessibility remains problematic, with few strains being deposited in public culture collections and even fewer of the proposed novel taxa being validated. For most of the strains deposited, genomes are available, but their quality is not always high and

metadata such as the source of isolation, cultivation requirements, and full taxonomy are often lacking. The small number of deposited strains limits future confirmatory or comparative studies, while the lack of curated metadata reduces the value of the genomic information. Only the professional acquisition of strains by public collections can ensure that the high-quality standards required for future work are maintained and that valuable strain-associated information is gathered and made freely accessible to researchers and users worldwide. This is particularly relevant for microbiota research to move beyond associations and towards a more mechanistic understanding, as exemplified by the study of *Akkermansia muciniphila* [19,20].

The deposit of strains in public collections is a time-demanding and expensive process, yet necessary to ensure quality. Once deposited, the preservation of strains becomes a highly cost-effective means of maintaining the value of microbial strains indefinitely [21]. However, the complete acquisition of single bacterial strains at international collections, which includes purity and identity confirmation, requires several months of time and for fastidious anaerobic strains such as those isolated from the human intestinal microbiome, can be accomplished for only about 20 strains in parallel by one laboratory technician. This has become a major bottleneck in the creation of publicly accessible collections emanating from current microbiome research that generates ever increasing numbers of novel isolates over comparatively short time intervals. A solution that enables rapid deposition is needed to enhance the time available to process strains.

To provide functional insights into the human gut microbiome based on isolates and facilitate public access to strains for the scientific community, we sought to create a bacterial collection made publicly available more rapidly through an innovative bulk submission system. In this way, all strains are easily traceable and all their associated metadata (cultivation conditions, taxonomy, genomic features) are provided. Studying this collection of isolates provided insight into the prevalence and diversity of plasmids within the human gut and the presence of megaplasmids. We also highlight health-associated variation within key genera, including novel species which require further *in vivo* study.

Results

A diverse range of key commensal species isolated from the human gut

The Human intestinal Bacterial Collection (HiBC) consists of 340 strains, representing 198 species, isolated from human faecal samples (**Figure 1**). It contains 29 families from across the seven dominant phyla in the human gut: *Bacillota* (n = 173 isolates), *Bacteroidota* (n = 95), *Actinomycetota* (n = 46), *Pseudomonadota* (n = 22), *Desulfobacterota* (n = 2), *Fusobacteriota* (n = 1), and *Verrucomicrobiota* (n = 1). Of the 198 species, 32 are novel taxa that we have described and named for validation according to both, the SeqCode [7] and the ICNP [22]. High quality genomes, defined as >90% completeness (99.22 ± 1.10%), <5% contamination (0.50 ± 0.77%), >10x

genome coverage (380.08 ± 293.97), were generated for all 340 strains. To improve access to the strains, their genomes, information on their cultivation, and isolation source material, we have created the HiBC web interface; www.hibc.rwth-aachen.de.

Genome size variation was observed between phyla, with *Actinomycetota* genomes being the smallest (2.38 ± 0.35 Mbp). *Bacteroidota* had the largest genomes (4.97 ± 1.02 Mbp), although not significantly larger than *Pseudomonadota* (4.84 ± 0.64 Mbp; adj. $p = 0.56$) (**Supplementary Figure 1a**; **Supplementary Table 1**). This variation in genome size between phyla was validated using >5,000 previously published genomes [23] (**Supplementary Table 1; Supplementary Figure 1b/c**). In addition to having the largest genomes, *Bacteroidota* also had the largest repertoire of CAZymes (**Supplementary Figure 2b**). To account for the variation in coding potential of each strain, and thereby examine their investment in carbohydrate utilisation, we normalised their CAZyme load based on their number of coding sequences (**Supplementary Figure 2c**). This analysis confirmed that the *Bacteroidota* have the highest investment in carbohydrate utilisation. However, they were followed by the *Actinomycetota* (**Supplementary Table 2**), indicating that, whilst members of this phylum have undergone host-associated genome size reduction [24], they have functionally invested in carbohydrate degradation.

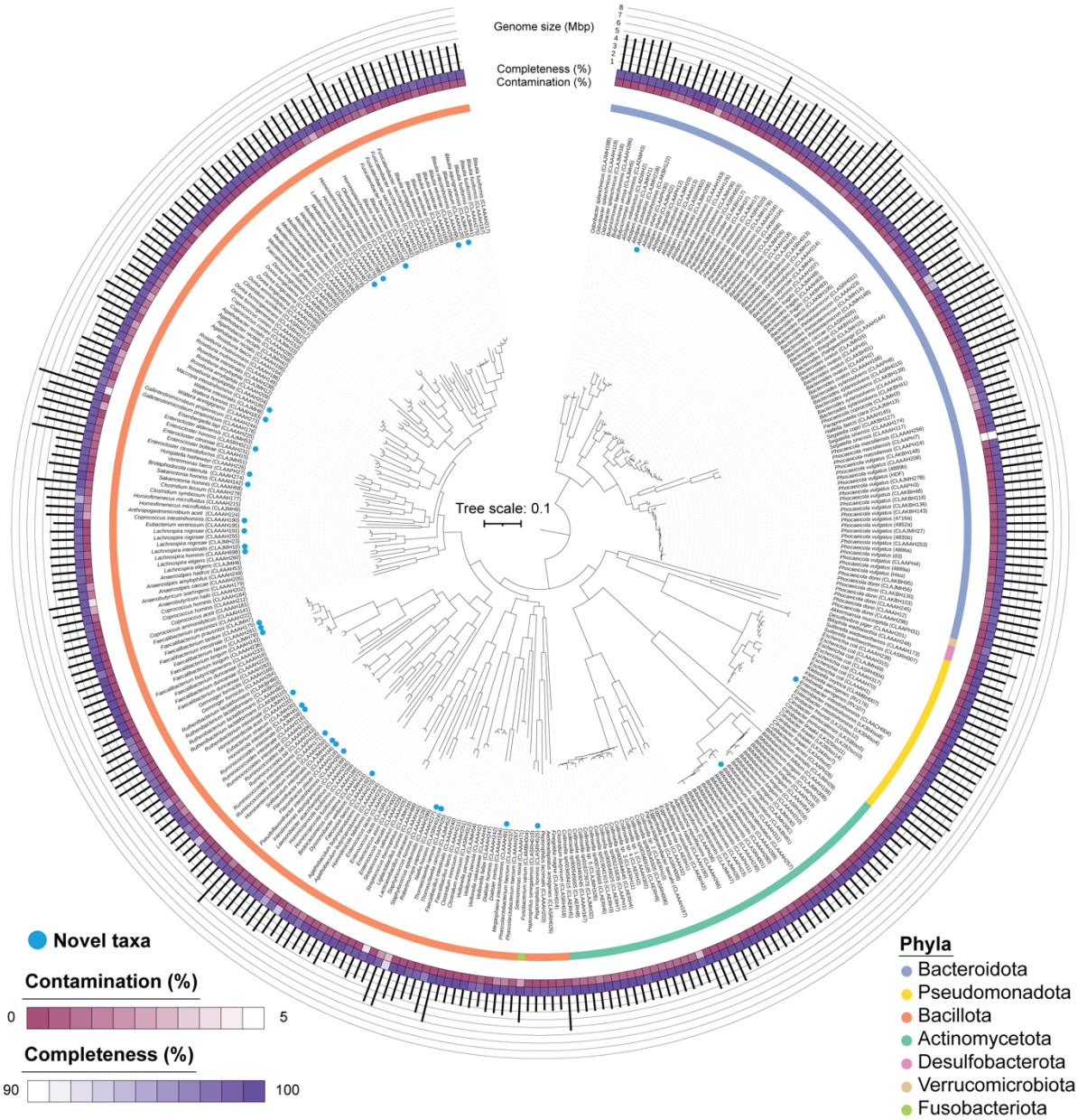


Figure 1: Phylogenomic diversity of isolates within HiBC. Tree based on all 340 genomes, generated using Phylophlan [25]. Phyla are indicated with colours. The *Bacillota* are split due to the placement of *Fusobacteriota,* which separated strains assigned to 'Bacillota_A' by GTDB, suggesting the need for the phylum *Bacillota* to be split. Blue circles identify strains belonging to the 32 novel species that are taxonomically described in this paper. Completeness and contamination of the genomes, as determined using CheckM, are indicated with colour gradients. The size of each genome (Mbp) is plotted as black bars on the outer ring.

<u>Plasmid landscape in the cultured strains and identification of prevalent megaplasmids</u>

Plasmids are known to be present in the human gut and vary between geographically distinct groups [13]. However, most analysis has been culture-independent, preventing accurate taxonomic assignment [14,15,26]. Even studies on horizontal gene transfer using isolates have not reconstructed plasmids and accounted for their impact [27]. As many

153 bacterial genome assembly workflows do not consider plasmids, we developed a
154 genome pipeline that first searches for the presence of plasmids using Recycler [28] and
155 plasmidSPAdes [29], assembles them, and then removes their reads from consideration
156 during genome assembly (see Methods). This resulted in the reconstruction of
157 plasmids from 46% of the strains (155 out of 340), with a total of 266 plasmids (**Figure**
158 **2a**). Plasmids were identified in all phyla except *Fusobacteriota*, for which the HiBC
159 contains only a single isolate. Almost half of the plasmid-containing strains contained
160 more than one plasmid (64 out of 155 strains), up to 6 plasmids in a strain of
161 *Phocaeicola vulgatus* and two strains of the novel species *Enterobacter*
162 *intestinihominis* (**Figure 2b**).

163 Given the number of plasmids a single strain can contain, and the size a single plasmid
164 can reach (see below), we hypothesised that they could affect the taxonomic
165 identification of an isolate, as most genomes lack delineation of plasmids. Given the
166 promiscuous nature of plasmids [30], it is possible that the taxonomic signal of a strain
167 may be masked by the presence of multiple large plasmids, distorting taxonomic
168 assignment. As average nucleotide identity (ANI) is a standard approach for
169 taxonomic assignment of genomes, we calculated the theoretical plasmid size
170 required to potentially cause a >1% change in ANI between two genomes of identical
171 size (**Supplementary Figure 3a**). Plasmids >80 kp should theoretically cause such a
172 difference in ANI within genomes ≤8 Mb. To assess this within the HiBC strains, we
173 selected all plasmid-containing strains that had a close relative (determined by GTDB-
174 Tk) (n = 152). We then calculated the ANI to these close relatives, both with and
175 without plasmid inclusion (**Supplementary Figure 3b**). The plasmids had a significant
176 (p = 0.01), but negligible (delta-ANI, 0.004 ± 0.01) effect on reducing the ANI to the
177 closest relative strains. The effect of plasmid inclusion on ANI was independent of
178 cumulative plasmid size, with some small plasmid sets having a greater effect on ANI
179 than the largest (**Figure 2d**). The only strain on which the removal of plasmids had a
180 significant effect was CLA-ER-H4, which was assigned to 'Collinsella sp900547855'
181 by GTDB-Tk when its plasmids were included, but was assigned to 'Collinsella
182 sp003487125' without the plasmids. This is indicative of the need for better taxonomic
183 study of this genus, as the plasmids had no effect on the assignment of CLA-ER-H4
184 to 'Collinsella sp003487125', but removing the plasmids increased ANI to 'Collinsella
185 sp900547855' by 0.05%. However, to avoid the creation of false species within
186 *Collinsella*, we have not described any new species within this genus, which needs to
187 be done in a comprehensive manner.

188 Sequence comparison of the HiBC plasmids identified multiple clusters, the two
189 largest of which were specific to the *Bacteroidales* (**Supplementary Figure 4a**). The
190 largest cluster shared similarities with 1,267 plasmids within the comprehensive
191 plasmid database, PLSDB [31], but lacked an assigned name and was therefore named
192 pBAC (plasmid *Bacteroidales*). pBAC was most frequently observed in the USA but
193 was also in Denmark, Japan, Ireland, and China (**Supplementary Figure 4b**). pBAC
194 occurred in both a shorter (~4,220 bp) and longer (~5,419 bp) form, with the two cluster
195 being connected by two 9,494 bp variants. The second largest plasmid cluster was

196 identified as the cryptic plasmid, pBI143, which is associated with IBD [14] and occurred
197 in 14 isolates across eight species. Given that pBAC matches 1,266 plasmids within
198 PLSDB, while pBI143 matched only 719, it may be the most prolific plasmid in the
199 human gut. Across the 266 plasmids identified in HiBC, 3,697 proteins were predicted,
200 although only 639 (17.28 %) could be functionally annotated. Of these, nine were
201 antibiotic resistance genes, four of which were copies of the *APH(2")-IIIa*
202 aminoglycoside resistance gene found on plasmids of identical size (7,686 bp) in four
203 strains of *Ruthenibacterium lactatiformans* (strains CLA-KB-H110, CLA-KB-H15, CLA-
204 KB-H80, CLA-AA-H80).

205 Megaplasmids (>100 kbp) have been described to occur in lactobacilli [32] and
206 bifidobacteria [33], but their occurrence in a broader range of commensals from the
207 human gut is unknown. We evaluated the length of the reconstructed plasmids and
208 found the average length was 11.74 ± 17.25 kbp (**Figure 2c**). This included the
209 identification of two megaplasmids, one within a strain of the recently described
210 species *Hominifimenecus microfluidus* (strain CLA-JM-H9; 125.6 kp) and the other in
211 a strain of *Phocaeicola vulgatus* (strain CLA-AA-H253; 103.4 kp). The *P. vulgatus*
212 megaplasmid matched 26 plasmids within PLSDB, including multiple designated as
213 pMMCAT, a recently proposed large plasmid from *Bacteriodales* which has yet to be
214 confirmed using sequencing-independent methods [34]. As such, the *P. vulgatus*
215 megaplasmid was designated pMMCAT_H253. Out of the 104 proteins encoded on
216 pMMCAT_H253, only 6 could be assigned to a functional category, highlighting the
217 need for further characterisation of these proteins. A locus containing four proteins
218 assigned to 'fimbrillin family proteins' was identified within a region of lower GC content
219 compared to the plasmids average, which may suggest this region is part of the
220 plasmids additional cargo and not the backbone (**Figure 2e**). While such
221 megaplasmids have previously been observed in sequencing data of human gut
222 commensals, their existence is rarely confirmed. To validate the existence of
223 pMMCAT_H253, we used pulsed field gel electrophoresis (PFGE) which showed a
224 band of ~100 kb after treatment using XbaI (**Figure 2f**), for which the plasmid had a
225 single restriction site allowing for its linearisation. The matching pMMCAT plasmids
226 were observed within species of *Bacteroides*, *Phocaeicola* and *Parabacteroides* from
227 the USA, China, Japan, and Israel (**Supplementary Figure 4c**). Network analysis of
228 the pMMCAT plasmids suggested those from *Phocaeicola* were more divergent than
229 the *Bacteroides* and *Parabacteroides* pMMCAT plasmids (**Supplementary Figure**
230 **4d**). MobMess[15] analysis assigned pMMCAT_H253 as a compound plasmid,
231 supporting the theory that it contains additional cargo, leading to the variable
232 associations with health conditions for the predicted proteins. A sequence within the
233 pMMCAT family was detected in 99.55% (1,774/1,782) of human gut metagenomes
234 [15]. However, only 52 metagenomes had detection values >95%, including two samples
235 from Fiji, with the highest coverage (135.82x) being from USA samples
236 (**Supplementary Figure 4e**). These results support the need for greater study of
237 pMMCAT as a potential source for phenotype variation within genera known to
238 contribute to host health [35,36].

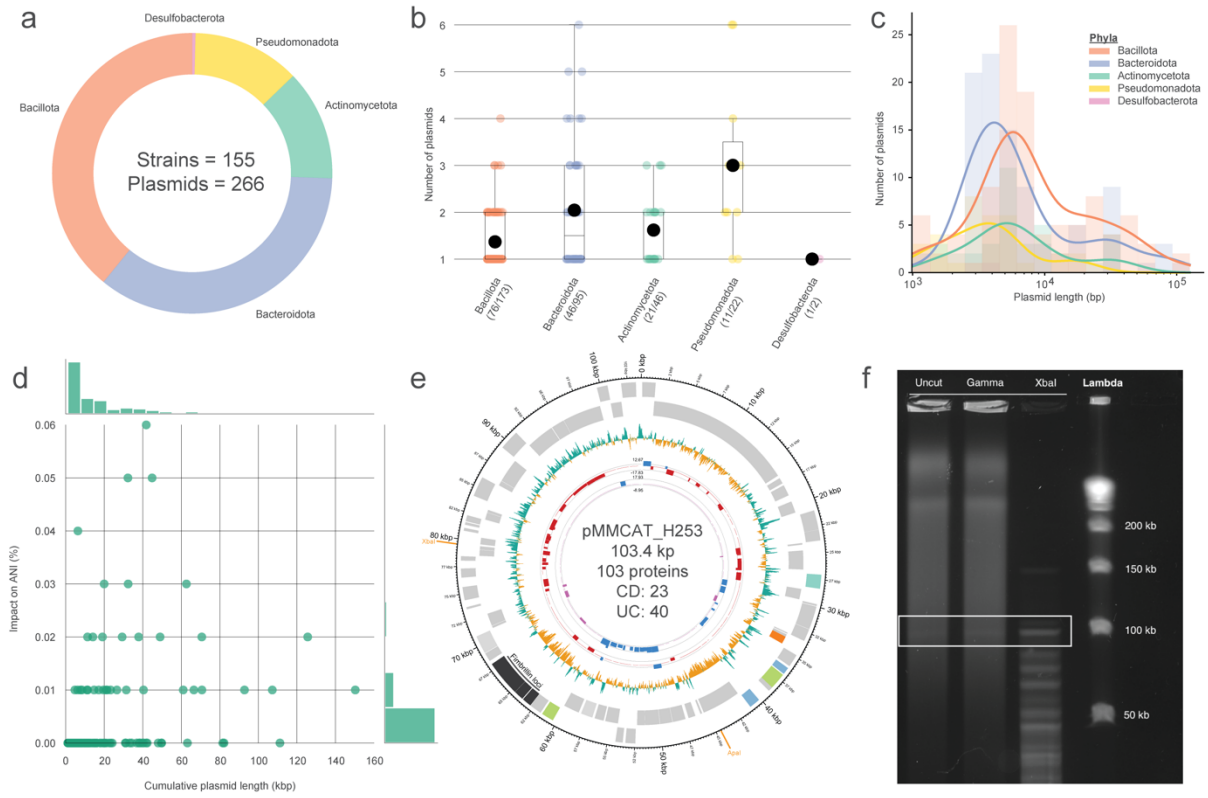


Figure 2: Plasmid repertoire of human gut bacterial isolates. a. Taxonomic diversity of reconstructed plasmids at the phylum level. **b.** Number of plasmids reconstructed within each isolate. **c.** Length of reconstructed plasmids; a bar represents the number; a line represents the distribution. **d.** Impact of plasmids on the ANI comparison to each strain's closest match, compared to the cumulative size of the plasmids within the strain. **e.** Plasmid map of pMMCAT_H253. The innermost rings represent the association of proteins identified to be differentially prevalent between Crohn's disease (CD) samples and healthy controls (first ring), or ulcerative colitis (UC) samples and healthy controls (second ring). Proteins enriched in CD are purple, those enriched in healthy samples are blue, and those enriched in UC are red. The third ring represents the GC content relative to the average of the entire plasmid. Boxes are used to represent genes identified on the plasmid in the forward (outer ring) and reverse (inner ring) strand, with those coloured being assigned a COG category, while grey boxes represent COG-unassigned proteins. The fimbriae loci proteins are indicated in dark grey. Enzymes with a single restriction site are indicated on the outer ring in orange. **f.** Pulsed-field gel electrophoresis of pMMCAT_H253 confirms its presence in the uncut/native form, as well as occurring after gamma irradiation (to linearise circular DNA), and cut with XbaI (cuts once in pMMCAT_H253). Lambda phage was used as a marker size ladder.

Description of dominant, as yet unknown taxa within the human gut

The environmental specificity of the isolates was assessed by examining their presence in 16S rRNA gene amplicon samples for different human body sites (gut, vagina, lung, skin; n = 1,000 each) (**Supplementary Figure 5**). The high prevalence of isolates in gut samples confirmed their specificity for this anaerobic environment. To assess the occurrence of HiBC strains in the human gut in greater detail, we looked

264 at their prevalence and relative abundance within 1,000 human gut samples (**Figure**
265 **3a**). This analysis was complemented by screening the genome of each isolate against
266 a collection of 42,927 MAGs [23,37]. *Bifidobacterium* spp. were some of the most
267 prevalent (~70% of samples) and abundant (~9% on average) species. It has
268 previously been described that cultivation and sequencing approaches capture
269 different fractions of the human gut; this is highlighted by our ability to isolate a strain
270 of *Selenomonas noxia*, which was only observed in a single sample, with a relative
271 abundance <0.1%, making it a sub-dominant member of the gut microbiota, likely of
272 oral origin [38]. Although only present at an average relative abundance of 2.7%, strains
273 of *Agathobacter rectalis* (synonym *Eubacterium rectale*) were identified as being the
274 most frequently reconstructed strain (n = 1,839 MAGs). In contrast, 19 strains did not
275 match any MAGs, including seven novel taxa. To further assess the importance of
276 HiBC species to the human gut ecosystem, we screened them against the original
277 Human Microbiome Projects "most wanted" list of taxa [39]. Of the 1,468 most wanted
278 sequences, 90 matched HiBC strains, including 4 high, 24 medium, and 62 low priority
279 sequences. Of these, 8 matching strains represented novel taxa within the phylum
280 *Bacillota* according to the current state of nomenclature, including two novel genera
281 (**Supplementary Figure 6**). These "most wanted" species, and all other novel taxa
282 discovered in this study are described below (all protologues in the methods).

283 Next, we assessed the coverage of the entire HiBC collection, as well as just the novel
284 taxa against a MAG collection for which corresponding relative abundance values
285 were available for 4,623 individuals. HiBC covered 56.51 ± 20.44 % of an individual's
286 microbiota. The 32 novel taxa described within this work enhanced this coverage by
287 4.55 ± 4.17% on average and 150 samples gained >15% relative abundance, which
288 peaked at more than >40% in two samples (**Figure 3b**). Of the 32 novel taxa, seven
289 species from five different genera were identified with mean relative abundances
290 >0.25% within positive metagenomic samples (**Figure 3c**). The most dominant of
291 these was *Ruminococcoides intestinale* (mean relative abundance = 2.84%; n =
292 3,023), followed by *Ruminococcoides intestinihominis* (1.19%; n = 285). Although only
293 recently described [40], these results suggest that *Ruminococcoides* is a dominant but
294 understudied genus within the human gut. A second genus of clear ecological
295 importance to the gut is *Lachnospira*, in which we describe two dominant novel
296 species. Interestingly, the two samples in which novel taxa accounted for >40%
297 relative abundance were dominated by either *R. intestinale*, or the novel species
298 *Lachnospira rogosae*. As *L. rogosae* was also identified as a medium priority, HMP
299 most wanted taxa and has been associated with human health conditions, such as
300 myalgic encephalomyelitis [41], further study of this species is needed to quantify its
301 impact on host health and identify the mechanisms, which is now enabled by this work.

302 To uncover associations of these taxa with human health conditions, we studied the
303 ecology of each protein within *R. intestinale* CLA-AA-H216^T (**Figure 3d**), as the most
304 abundant novel taxa, and *Blautia intestinihominis* CLA-AA-H95^T (**Figure 3e**), as
305 member of a genus associated with both health and disease conditions, using
306 InvestiGUT[42]. Out of the 2,167 proteins encoded by *R. intestinale* CLA-AA-H216^T,

307 comparison between Crohn's disease (CD) patients and healthy controls identified a
308 total of 1,902 differentially prevalent proteins, with 1,883 being significantly more
309 prevalent in healthy samples whilst 19 were more prevalent in CD samples. The same
310 pattern was observed with Ulcerative colitis (UC) where a total of 1,715 significantly
311 differentially prevalent proteins were identified, with 1,684 enriched in healthy
312 samples, and 31 enriched in UC samples. Pathway analysis of the health-associated
313 proteins identified multiple anti-inflammatory pathways, including the ABC transporter
314 for spermidine (PotB/C/D), an anti-inflammatory polyamine [43], and biosynthesis of
315 biotin, which has been shown to be immunomodulatory [44,45], ameliorating colitis [46].

316 The association of *Blautia* spp. with inflammatory bowel diseases is more complex, as
317 some species within this genus have been shown to ameliorate colitis [47], leading to
318 their inclusion within therapeutic products [48]. However, recent studies have suggested
319 some *Blautia* spp. may exacerbate colitis [49]. Given the complex interaction of this
320 genus with inflammatory bowel diseases, we studied the association of the novel
321 species, *B. intestinihominis* CLA-AA-H95^{T}, with both CD and UC. Out of the 3,906
322 proteins, comparison between CD patients and healthy controls identified 523
323 differentially prevalent proteins, with 488 being significantly more prevalent in healthy
324 samples whilst 35 were more prevalent in CD samples. In contrast, comparison
325 between UC patients and healthy controls identified 1466 proteins, with 16 being
326 significantly more prevalent in healthy samples whilst 1,450 were more prevalent in
327 UC samples. The functionality of the large number of proteins enriched within UC
328 samples was studied further, identifying ABC transporters for both branched-chain
329 amino acids (LivF/G/H/K/M) [50], phosphate (PstA/B/C/S) [51], and adenosine (NupA/B/C,
330 BmpA) [52], each of which has been linked to colitis.

331 These results highlight the relevance of describing novel isolates and provide insights
332 into the observed association of their proteins with health or disease.

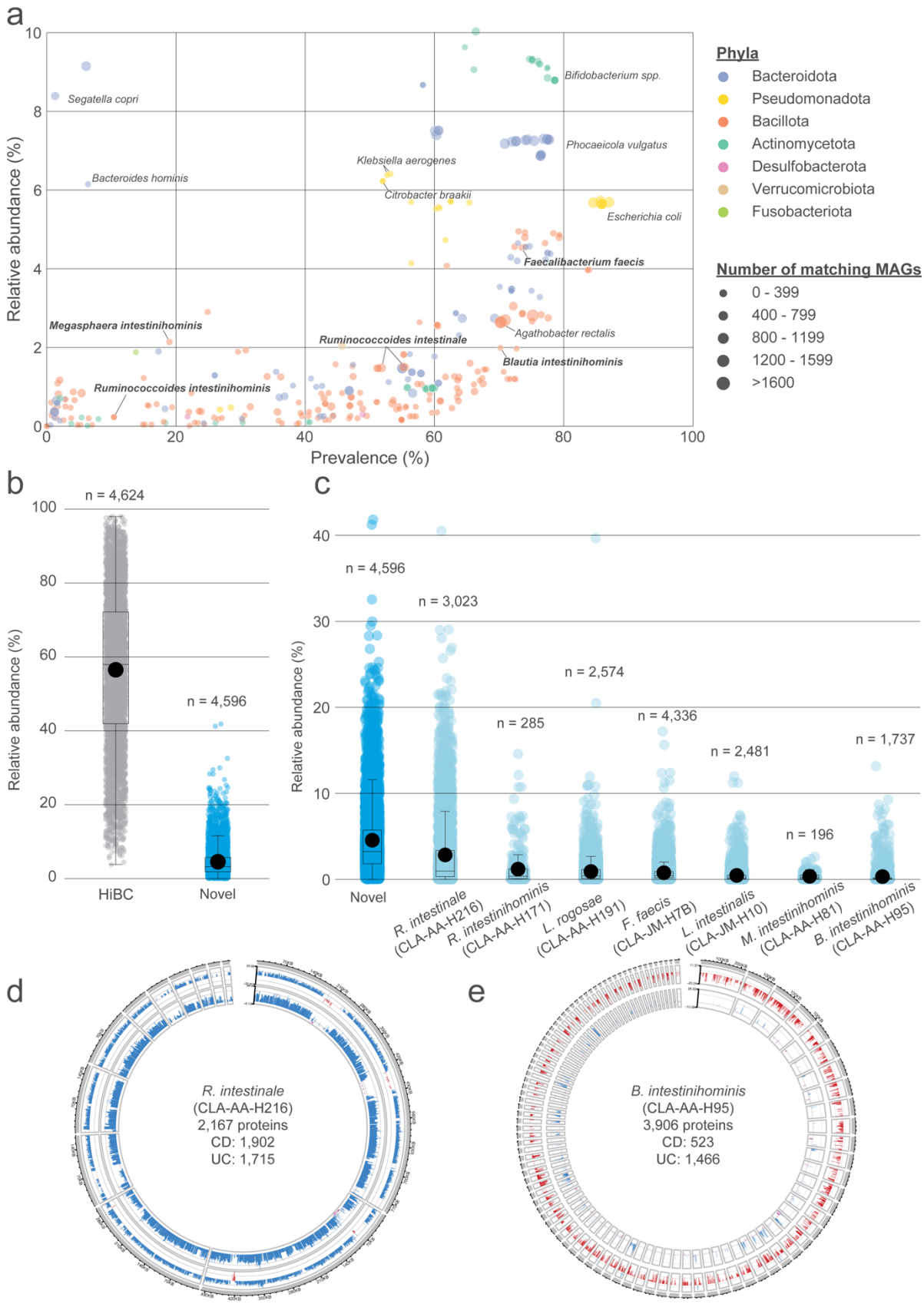


Figure 3: Ecology of isolates and novel taxa. a. The prevalence of each HiBC isolate across 16S rRNA amplicon datasets from the human gut (n = 1,000) was plotted against their relative abundance within positive samples. Each point represents a single isolate, coloured by phyla. The size of each point is proportional to the number of human gut MAGs matching the corresponding isolate's genome.

338 A selection of interesting isolates discussed in this study, or taxonomically coherent isolates, are
339 named. **b.** Cumulative relative abundance of gut metagenomes across 4,624 individuals from Leviatan
340 *et al* (2022) covered by the entire HiBC (left, grey) or the subset represented by novel taxa (right, blue).
341 **c.** Relative abundance of dominant (mean relative abundance >0.25%) novel taxa across 4,624
342 individuals, with the number of positive samples stated. **d-e.** Genomic location of proteins identified to
343 be significantly differentially prevalent between Crohn's disease (CD) samples and healthy controls
344 (inner ring), or ulcerative colitis (UC) samples and healthy controls (outer ring). The delta-prevalence
345 (prevalence in healthy donors – prevalence in corresponding patients) is shown in blue (more prevalent
346 in healthy controls), red (UC), or mauve (CD). The species, strain, number of proteins predicted within
347 the genome, and those significantly differentially between health conditions are shown within the circle.
348 Only contigs >10kp were plotted.

349

350 Enhanced diversity of *Faecalibacterium* species

351 The genus *Faecalibacterium* currently (March 2024) contains six validly published and
352 correctly named species. HiBC contains representatives for four of these, as well as
353 isolates representing three novel species. Initial 16S rRNA gene analysis assigned
354 these novel species as *Faecalibacterium prausnitzii*, a common commensal of the gut
355 microbiota that has been associated with human health conditions ranging from
356 reducing inflammation [53], to improving cognitive function in Alzheimer's disease [54].
357 Many of these beneficial observations have been attributed to *F. prausnitzii's* ability to
358 produce large amounts of the short-chain fatty acid butyrate [55].

359 Taxonomically, the genomic diversity of *F. prausnitzii* is proposed to be greater than
360 suggested by 16S rRNA gene sequence analysis [56]. This has led the GTDB to split
361 this species into 12 proposed species. HiBC contains representatives of the as yet
362 undescribed species defined by GTDB-Tk as *F. prausnitzii* A (strain CLA-AA-H175),
363 *F. prausnitzii* D (strain CLA-JM-H7B), and *F. prausnitzii* J (strain CLA-AA-H281). This
364 was confirmed using ANI and can be observed in a phylogenomic tree (**Figure 4a**).
365 We therefore propose to name these novel species *Faecalibacterium tardum* (CLA-
366 AA-H175), *Faecalibacterium intestinale* (CLA-AA-H281), and *Faecalibacterium faecis*
367 (CLA-JM-H7B). Protologues for these, and the other 29 novel taxa described in this
368 paper are provided at the end of the Methods section. Out of these three
369 *Faecalibacterium* species, *F. faecis* was the most prevalent (93.77% of samples
370 positive) and dominant (relative abundance: 0.77% ± 1.09) **Figure 4b**.

371 Because of the importance of butyrate production by *Faecalibacterium* spp., we
372 studied this function further (**Figure 4**). While all *Faecalibacterium* strains within HiBC
373 produced butyrate, none of the strains encoded for any butyrate production gut
374 metabolic modules (GMMs; MF0088, MF0089) (**Figure 4c**). To investigate this
375 inconsistency further, we searched a curated list of butyrate production pathways [57]
376 and identified all strains encoded for the acetyl-CoA pathway, but lacked annotations
377 for the terminal enzyme (**Figure 4c**). We extended this to all HiBC strains and
378 identified that 7.4 % of the isolates (n = 25) contained the complete acetyl-CoA
379 pathway (**Supplementary Figure 7**). Taxonomically, we observed that this pathway
380 was restricted to the phylum *Bacillota* (22/25 isolates), except for *Butyricimonas* spp.
381 within the *Bacteroidota*. These results highlight the limitations of current functional

annotation of genomes, even for well-known pathways of interest, such as butyrate production. Next, we measured the ability of all *Faecalibacterium* strains (n = 11) to produce butyrate (**Figure 4c**). Butyrate production varied widely, with *F. tardum* (strain CLA-AAH175) producing only 0.8 ± 0.16 mM, whereas strains of *F. longum* and *F. butyricigenerans* produced >10 mM. The butyrate production patterns were driven by the ability of each strain to grow in the test medium (YCFA). Due to this, we correlated the average butyrate production of each strain against its average growth, measured by OD600 (**Figure 4d**). A strong (r = 0.76) and significant (p = 0.01) correlation was identified. Interestingly, differences were observed between strains belong to the same species (*F. duncaniae* CLA-AA-H283, 8.67 ± 0.41 mM, OD600 1.31 ± 0.01 *vs. F. duncaniae* CLA-AA-H16, 4.25 ± 0.49 mM, OD600 0.52 ± 0.07). These results suggest that interpretations of the ability of different *Faecalibacterium* to produce butyrate *in vitro* should always consider both the exact strain used and its growth ability in the culture medium tested. *In vivo*, this may suggest that strains that grow at high abundance are also likely to produce the most butyrate.

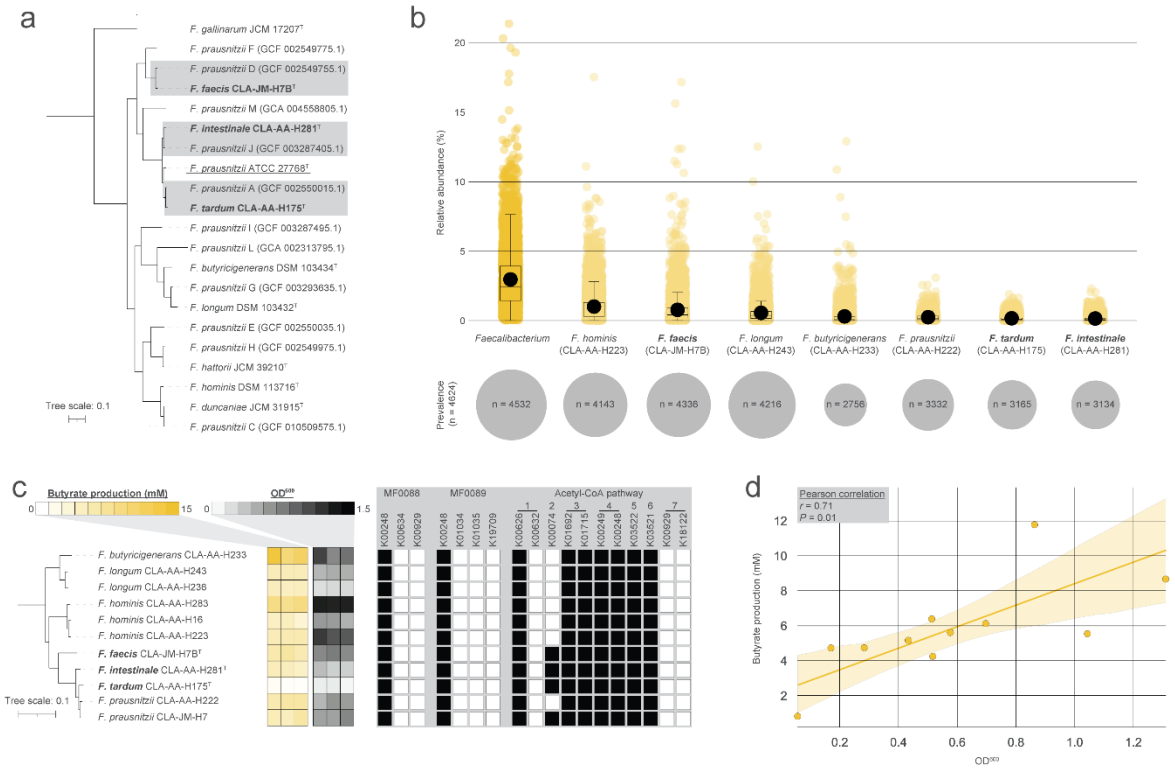


Figure 4: Novel diversity within *Faecalibacterium* and strain-dependent butyrate production. a. The three novel species of *Faecalibacteirum* described within this paper were placed within the current landscape of *Faecalibacterium* with a valid name, along with the type genomes for proposed divisions of *Faecalibacterium prausnitzii*, as determined by GTDB. The phylogenomic tree was rooted on *Ruminococcus bromii* ATCC 27255^T. Novel species are in bold, and the type strain for *Faecalibacterium prausnitzii* is underlined. **b.** Relative abundance, and prevalence, of the genus *Faecalibacterium*, and each *Faecalibacterium* species represented within HiBC across 4,624 metagenomic samples. **c.** Phylogenomic tree of the *Faecalibacterium* strains within HiBC, displaying the ability of each strain to produce butyrate over a 48h period, along with the OD600 that the strain achieved during the testing period (n = 3 independent batch cultures for each strain; the replicates are shown with individual boxes). The presence of KEGG orthologs (KO) relating to three different pathways for butyrate production is

409 displayed as a black square, whereas a white square indicates the lack of a KO. The enzymes required
410 for the acetyl-CoA pathway are: 1. *thl* (K00626, K00632), 2. *bhbd* (K00074), 3. *cro* (K01692, K18122),
411 4. *bcd* (K00249, K00248), 5. *etfA* (K03522), 6. *etfB* (K03521), 7. *but*/*buk* (K00929, K18122). The
412 phylogenomic tree was rooted on *Ruminococcus bromii* ATCC 27255^{T}. **d.** Correlation of the mean
413 OD600 against the mean butyrate production of each strain with a linear regression and its 95%
414 confidence interval.

415

416 Bulk submission system enables strain traceability at a large scale

417 Published isolates, such as those in HiBC, are often not easily accessible to
418 researchers, impeding them from working with the same quality-controlled strains and
419 being able to reproduce data. The main barriers to making isolates publicly available
420 are the time-demanding process of strain deposition and the lack of unique strain
421 identifiers for transparent referencing of strains while they are being processed.
422 Researchers frequently deposit only a small selection of strains, preventing access to
423 functionally relevant diversity, as highlighted above with *Faecalibacterium* (**Figure 4**).
424 Current strategies for publishing large collections of isolates vary widely, as shown in
425 **Figure 5** (upper box). Publication with a statement such as "available upon request"
426 without deposition of the strains (case 1) is quick. Still, it should not be accepted
427 anymore, as the strains and their data are effectively inaccessible and will ultimately
428 be lost over time. Depositing in an international culture collection (DSMZ, JCM, CCUG,
429 BCCM, KCCM, ATCC, etc.) ensures metadata quality is maintained and guarantees
430 the long-term accessibility of the strains for the scientific community. However, if a
431 publication is released before strains are deposited (case 2) or in parallel (case 3),
432 there is a risk that deposits will fail (e.g. strains do not grow, are not pure, or their
433 identity cannot be confirmed), and that different strain identifiers may be used in the
434 long term, compromising the traceability of the strains. Furthermore, once the work is
435 published, the lack of incentives often prevents researchers from taking the final and
436 important step of depositing in a public collection. Current best practice is to ensure
437 that all strains included in a manuscript are deposited and processed at an
438 international culture collection before publication, so that strain catalogue numbers
439 can be issued and included in the manuscript (case 4) [58,59]. The downside of this
440 approach is that deposition and strain processing for quality checks is time consuming,
441 and the time required grows linearly with the number of strains to be deposited. The
442 deposition of a collection such as HiBC is thus a significant undertaking.

443 We propose a bulk submission system to address these issues and facilitate large-
444 scale strain deposition, as shown in **Figure 5** (lower box). This new system relies on
445 stable identifiers provided by the database StrainInfo (https://straininfo.dsmz.de) early
446 in the process. After an initial check of the strains and curation of the metadata, the
447 strains data is published as a StrainInfo entry, a persistent StrainInfo (SI) identifiers is
448 assigned to each strain, and registered as Digital Object Identifiers (DOIs). This initial
449 step ensures that all data (e.g. publications and sequence data) for these strains can
450 be linked again later. The bulk submission of strains in one or several collections is
451 mediated though a new StrainRegistry system, which allows researchers and

collections to interact directly. Like a review system of a scientific journal, this new system accelerates the deposition process by providing all deposition data and communication about it in one place. After a culture collection has accepted a group of strains (or parts of it), it requests the strains from the researcher in batches which are deposited in an iterative process. The StrainInfo database first displays strain entries with the status 'Deposition in progress', which is changed to 'Published' once the strain is available from one or more culture collections. This new system significantly streamlines the deposition process, making it transparent to the outside and provides the persistent identifiers necessary to connect strain data at any time.

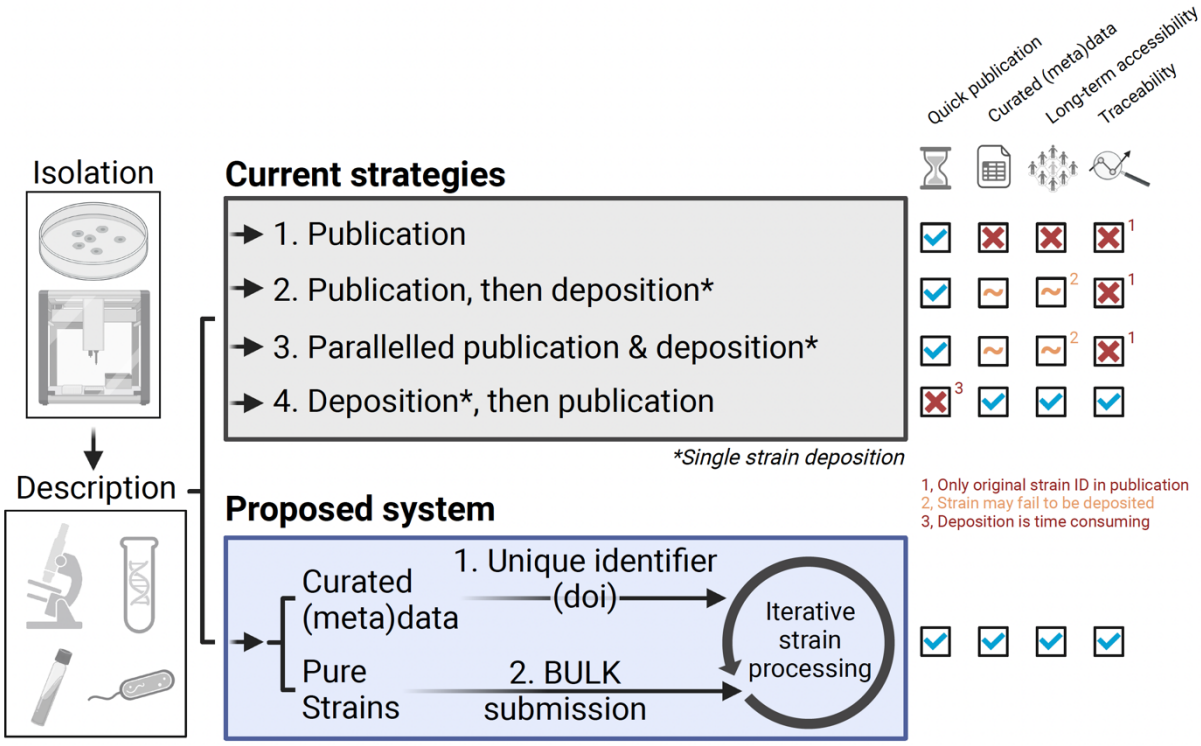


Figure 5: Schematic of the proposed submission system for large-scale strain deposition. Once bacteria have been isolated and described, there are multiple strategies for publishing the strains. In the upper box, we have four example strategies, each of which can be applied on a single strain basis, but are not designed for large-scale submission of strains. The benefits and negatives of each method is indicated on the right of the methods under four main categories, where a cross indicates a missing aspect to the strategy. The lower box pictures the main advantage of creating unique strain identifiers early in the process as a foundation for bulk submission of a large number of strains.

Discussion

Over the last decade, the renewed interest in cultivation has expanded the number of genomes available for taxa within the human gut. However, the corresponding strains are rarely accessible to the research community. Deposition of strains within international culture collections has been highlighted as a major limiting factor in making strains publicly available [60]. To overcome this, we established a new deposition system. Streamlined registration of strains before bulk submission ensures their traceability, facilitating publication while ensuring strains are consistently

478 referenced. Application of this system allowed for deposition of 340 human gut isolates
479 to the German Collection of Microorganisms and Cell Cultures, facilitating the naming
480 and description of 32 novel taxa, consisting of 27 novel species and five novel genera.
481 Based on this, the HiBC is the first publicly accessible collection of human gut isolates.

482 Large-scale cultivation of human gut bacteria in this study led to isolation of many
483 novel species and genera. Whilst previous cultivation studies report many isolates
484 represent novel diversity, these bacteria are never taxonomically described and
485 named. The discovery of prevalent and important novel species in this work will
486 facilitate future experiments on the role of bacteria in health and disease. This included
487 the isolation of multiple strains of *Faecalibacterium prausnitzii*, a species with the
488 potential to reduce gastrointestinal inflammation[53]. We confirmed that this species is
489 actually a diverse group of related species that show reduced 16S rRNA gene
490 sequence diversity [56]. By naming three of these species and describing the butyrate
491 production and growth characteristics of seven different species in this genus, we
492 observed the ability of these strains to produce butyrate was proportional to their ability
493 to grow. Given these results, further quantification of these strains ability to grow within
494 the human gut is required to understand their contribution to SCFA levels *in vivo*, and
495 therefore their potential impact on host health.

496 The reconstruction of plasmids for each of these strains allows for the direct link of
497 plasmids to strains of various species, key information previously lacking from studies
498 into plasmids within human gut [13–15]. Plasmids were identified to be highly prevalent
499 within human gut isolates, occurring within almost all phyla represented within the
500 HiBC, with as many as six occurring within a single strain. By making the reconstructed
501 plasmids, and strains publicly available, we believe this resource expands the toolbox
502 of methods for genetically modifying non-model commensal gut microbes.

503 Metagenomic studies of the human gut have provided various insights into the
504 taxonomic and functional diversity present, however, they are unable to provide
505 mechanistic insights. By providing researchers access to a collection of human gut
506 isolates, the functionality of taxa can be experimentally validated, facilitating
507 mechanistic study of host-microbe interactions.

508

509 **Methods**

510 Ethics

511 The Ethics Committee of the Medical Faculty of RWTH University Aachen permitted
512 bacterial isolation from human stool under ethical number EK 23-055, EK 316-16, and
513 EK 194/19. For strains originating from Vienna, isolation was approved by the
514 University of Vienna ethics committee under ethical number 00161. For strains
515 originating from Braunschweig, isolation was approved by the Ethics Committee of
516 Lower Saxony (MHH permit No. 6794, 8629, and 8750). Written informed consent was
517 signed by all enrolled participants.

518

519 Bacterial isolation and cultivation

520 Human stool was collected in sterile plastic bags and stored in tightly sealed plastic
521 buckets until further processing. An oxygen scavenger sachet (BD Biosciences; ref.
522 260683) was added to each bucket before sealing to reduce exposure to oxygen. All
523 samples were processed in the lab within 24 hours of collection. First the faecal
524 material was homogenized by manual kneading of the plastic bag. Stool samples were
525 either mixed 1:5 with anaerobic FMT media[61], distributed to 1 mL aliquots and stored
526 at -80 °C until further use, or the samples were further processed as described
527 previously[59]. Hence approximately 5 g of the samples were dissolved in 50 ml of
528 anaerobic PBS supplemented with peptone (0.05% w/v), L-cysteine (0.05% w/v) and
529 dithiothreitol (DTT) (0.02% w/v) by shaking the glass flask intensely. A syringe with
530 needle was used to transfer 5 ml of the slurry through a rubber stopper (previously
531 flamed using ethanol) into a second glass flask containing 45 ml of the same buffer in
532 an anaerobic atmosphere (89.3% N_2, 6% CO_2, 4.7% H_2) to create a 1:100 dilution of
533 the original sample. The flask was moved into an anaerobic workstation (MBraun
534 GmbH, Germany) to prepare 2 ml aliquots under anaerobic conditions, which were
535 mixed with 2 ml of 40% anaerobic glycerol to a final concentration of 20%, then stored
536 at -80 °C until use.

537 Two different approaches were used for bacterial isolation. For classical isolation, the
538 faecal aliquots were thawed inside the anaerobic workstation and diluted with anoxic
539 PBS (see above) in a tenfold dilution series (10^{-2} to 10^{-6}). 50 µl of each dilution step
540 were transferred onto different agar plates and spread using a L-spatula. Single
541 colonies were picked after 1-7 days of incubation at 37 °C under anaerobic conditions.
542 Bacterial cells were re-streaked at least three times to guarantee purity of the culture.
543 For the second isolation approach, the single-cell dispenser b.sight (Cytena GmbH,
544 Germany) was used as described before [62]. Details on culture media preparation and
545 ingredients are provided in the **Supplementary Methods**.

546 The isolated bacteria were first identified using MALDI-TOF MS (Bruker Daltonics,
547 Bremen, Germany). If a species was not yet present within HiBC, or could not be
548 reliably identified by MALDI-TOF MS, the bacteria were stored as cryo-stocks and
549 DNA extracted for genome sequencing. Cryo-stocks were generated by freezing in
550 glycerol media (end concentration 20%) at -80 °C. All HiBC strains have been
551 deposited at the Leibniz Institute DSMZ, German Collection of Microorganisms and
552 Cell Cultures. In addition, the strains representing novel taxa were deposited at either
553 the Belgian Coordinated Collections of Microorganisms (BCCM) or the Japanese
554 Collection of Microorganisms (JCM).

555

556 Metabolite production analysis

557 Concentrations of short-chain fatty acids (SCFAs) (acetate, butyrate, propionate,
558 valerate), branched chain fatty acids (isobutyrate, isovalerate), intermediate
559 metabolites (ethanol, formate, lactate, 1,2-propandiol, 1-propanol, succinate), as well

560 as mono- and disaccharides (galactose, glucose, lactose) were measured by high-
561 performance liquid chromatography (HPLC). The bacterial strains were grown
562 anaerobically, apart from *Robertmurraya yapensis* CLA-AA-H227 which was grown
563 under aerobic conditions, in YCFA broth (DSMZ Medium No. 1611) in Hungate tubes
564 for 48 hours at 37 °C. Triplicate cultures from each strain were measured. Baseline
565 controls were taken from each sterile Hungate tube before inoculation. Each taken
566 sample was centrifuged (10,000 x g, 10 min, 4 °C), supernatants were collected and
567 stored at -80 °C until HPLC measurement. Samples were prepared and measured
568 (including HPLC settings) as described previously [59]. External standards were used
569 for concentration determination by comparison of peak retention times (HPLC grade
570 compounds were purchased from Sigma-Aldrich). Peaks were integrated using the
571 Chromaster System Manager Software (Version 2.0, Hitachi High-Tech Science
572 Corporation 2013, 2017). Metabolite concentrations >0.2 mM (limit of detection (LOD)
573 for citrate, 1-propanol), >0.24 mM (LOD for butyrate, formate, galactose, glucose,
574 isobutyrate, isovalerate, lactose, valerate), >0.4 mM (ethanol, 1,2-propandiol) and
575 >0.8 mM (acetate, lactate, propionate, succinate) were considered for statistical
576 analysis if present in all three replicates. Production and consumption of metabolites
577 was calculated by subtracting the baseline values from the sample taken after 48
578 hours of growth.

579

580 Isolation of genomic DNA

581 DNA was isolated using a modified protocol according to Godon *et al*. [63]. Frozen cell
582 pellets were mixed with 600 µl stool DNA stabilizer (Stratec biomedical), thawed, and
583 transferred into autoclaved 2-ml screw-cap tubes containing 500 mg 0.1 mm-diameter
584 silica/zirconia beads. Next, 250 µL 4 M guanidine thiocyanate in 0.1 M Tris (pH 7.5)
585 and 500 µL 5 % N-lauroyl sarcosine in 0.1 M PBS (pH 8.0) were added. Samples were
586 incubated at 70 °C and 700 rpm for 60 min. A FastPrep® instrument (MP Biomedicals)
587 fitted with a 24 × 2 ml cooling adaptor filled with dry ice was used for cell disruption.
588 The program was run 3 times for 40 s at 6.5 M/s. After each run, the cooling adapter
589 was refilled with dry ice. An amount of 15 mg Polyvinylpyrrolidone (PVPP) was added
590 and samples were vortexed, followed by 3 min centrifugation at 15.000 x g and 4 °C.
591 Approximately 650 µl of the supernatant were transferred into a new 2 ml tube, which
592 was centrifuged again for 3 min at 15.000 x g and 4 °C. Subsequently, 500 µl of the
593 supernatant was transferred into a new 2 ml tube and 50 µg of RNase was added.
594 After 20 minutes at 37 °C and 700 rpm, gDNA was isolated using the NucleoSpin®
595 gDNA Clean-up Kit from Macherey-Nagel. Isolation was performed according to the
596 manufacturer's protocol. DNA was eluted from columns twice using 40 µl Elution buffer
597 and concentration was measured with NanoDrop® (Thermo Scientific). Samples were
598 stored at -20 °C.

599

600 Genome library preparation and Illumina Sequencing

601 Library preparation and sequencing were conducted using the NEBNext Ultra II FS
602 DNA Library Prep Kit for Illumina with dual index primers and ~300 ng of DNA (NEB
603 Inc.) on an automation platform (Biomeki5, Beckman Coulter) according to the
604 manufacturer's instructions. The time used for enzymatic shearing to ca. 500 bp to be
605 used on a MiSeq run was 10 minutes and to ca. 200 bp to be used on a NextSeq run
606 was 30 minutes. PCR enrichment of adaptor-ligated DNA was conducted with five
607 cycles using NEBNext Multiplex Oligos for Illumina (NEB, USA) for paired-end
608 barcoding. AMPure beads (Beckman Coulter, USA) were used for size selection and
609 clean-up of adaptor-ligated DNA. Sequencing was performed either at the IZKF Core
610 Facility Genomics (Uniklinik RWTH Aachen) on an NextSeq platform (Illumina) (PE
611 150 bp), or on a MiSeq (Illumina) (PE 300 bp) in house.

612

613 Sequencing data processing

614 Genomes were assembled from paired-end Illumina short-reads. Low-quality reads
615 with an expected average quality below 20 over a 5-bases window, containing
616 adapters, and shorter than 50 bases were discarded using Trimmomatic (v0.39) [64].
617 Reads with phiX sequences were removed using BBduk from the BBtools suite [65].
618 Plasmids sequences were reported only when inferred from plasmidSPades (v3.15.5)
619 [66] and then from Recycler (v0.7) [28]. Reads free from plasmid sequences were then
620 assembled using SPades (v3.15.5) [29] with default parameters.

621 Contigs above 500 bp in assemblies were kept for quality evaluation. Following the
622 MIxS specifications [67] as well as requirements from the SeqCode [7], genomes with a
623 coverage above 10X, a completion above 90%, and a contamination below 5% based
624 on CheckM (v1.2.2) [68] estimation using single-copy marker genes were flagged as
625 high-quality draft genomes. Additional quality flags on the assembly included: 1) <100
626 contigs, longest contig >100 kbp, and N50 >25 kbp using QUAST (v5.0.2) [69]; 2)
627 detectable 16S and 23S rRNA genes using metaxa2 (v2.2.3) [70]; 3) >18 unique
628 essential tRNAs genes and a detectable 5S rRNA gene sequence using the
629 annotations of bakta (v1.6.1) [71].

630 The genome assembly pipeline, including the reconstruction of plasmids and quality
631 checks steps, is available as a reproducible workflow using Snakemake (v7.9.0) [72]:
632 https://github.com/clavellab/genome-assembly.

633

634 Genome analysis

635 Taxonomic, functional, and ecological analysis for all genomes was conducted using
636 Protologger (v1.3) [73]. For this, 16S rRNA gene sequences were extracted from each
637 strains genome using barrnap v0.9 (default settings)
638 (https://github.com/tseemann/barrnap), and the longest 16S rRNA gene sequence
639 from each genome was used as input for Protologger. Phylogenomic trees were
640 generated using Phylophlan v3.0.60 [74] with proteomes predicted using Prodigal v2.6.3

641 (default settings) [75]. Gut metabolic modules presence was based on the identification
642 of all included KOs within a genome, detected by Kofamscan v1.3.0 [76].

643

644 Plasmid analysis

645 Sequence similarity between the isolates plasmids was determined using MobMess [15]
646 and clusters were visualised with Cytoscape [77]. Clustering was conducted firstly based
647 only on the isolate plasmids, and then including the background previously predicted
648 human gut plasmids.

649 The similarity of each plasmid to known sequences was determined with the 'mash
650 dist' option of PLSDB (v2023_11_03_v2) [31]. The ecology of each plasmid was inferred
651 from the geographical location assigned to each matching plasmid.

652

653 Pulsed Field Gel Electrophoresis

654 Cultures were grown overnight in 5 mL BHI media, centrifuged (2 mL for *P. vulgatus*
655 CLA-AA-H253, 5 mL for *H. microfluidus* CLA-JM-H9), resuspended in 100 µL sterile
656 PBS and incubated at 45 °C for 10 minutes. Of this, 80 µL was mixed with 320 µL 1%
657 agarose (Bio-Rad Certified Megabase Agarose) at 45 C, 80 µL transferred to PFGE
658 plug moulds (Bio-Rad CHEF Disposable Plug Molds, 50-Well) and cooled at 4 °C for
659 30 minutes. Solidified agarose plugs were transferred to lysis solution inside 2 mL
660 microcentrifuge tubes (20 mM Tris-HCl pH 8.8, 500 mM EDTA pH 8, 1% N-
661 laurylsarcosine, 1mg/mL Proteinase K) and incubated with shaking at 250 RPM at 45
662 °C for 4 hours. Lysis solution was replaced with fresh lysis solution + RNAse (10
663 µg/mL) and incubated with shaking at 45 °C overnight. Lysis solution was replaced
664 with wash solution (25 mM Tris-HCl pH 7.5, 100 mM EDTA pH 8) and incubated with
665 shaking at 45 °C. Plugs were then transferred to fresh 2 mL microcentrifuge tubes with
666 wash solution containing phenylmethanesulfonyl fluoride (1 mM) and incubated with
667 shaking at 45 °C for 30 minutes, twice. Plugs were then transferred to 1 mL wash
668 solution at stored at 4 °C until use in PFGE.

669

670 For analysis of intact genomic DNA, agarose plugs were subjected to 100 Gy of γ
671 radiation using a ^{137}Cs source (Gammacell 1000), to linearize circular chromosomes
672 [78]. PFGE was performed using a CHEF Mapper apparatus (Bio-Rad). Intact and XbaI-
673 digested DNA fragments were separated on a 1.2% agarose gel in 0.5 × TBE at 14 °C,
674 with a gradient voltage of 6 V/cm, linear ramping, an included angle of 120°, initial and
675 final switch times of 0.64 s and 1 min 13.22 s, respectively, and a run time of 20 h 46
676 min.

677

678 Website design

679 Access to taxonomic, cultivation, and genomic information about the HiBC strains is
680 available at: https://www.hibc.rwth-aachen.de. This website was created using the

681 Shiny package from R. The 16S rRNA gene sequence and genome for each strain
682 can be downloaded directly from this site, both individually for strains of interest via
683 the research data management platform Coscine, or for the entire collection via the
684 digital repository Zenodo. Further strain specific metadata provided includes: isolation
685 conditions, source, and risk group.

686

687 Description of novel taxa

688 The description of novel taxa was based on the analysis provided by Protologger v1.3
689 [73], and manually curated into the protologues below. For each isolate, taxonomy was
690 assigned using the following thresholds: <98.7% 16S rRNA gene sequence identity
691 (as indication for as yet undescribed species), <94.5% (undescribed genus), and
692 <86.5% (undescribed family) [79]. ANI values <95% to separate species [80]; POCP
693 values <50% for distinct genera [81]. Phylogenomic trees were also considered to make
694 decisions on genus- and family-level delineation [25]. All novel taxa have been
695 registered with the SeqCode and will be registered with the ICNP.

696

697 **Description of *Alistipes intestinihominis* sp. nov.**

698 *Alistipes intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, the intestine; L.
699 masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human intestine).

700 The genome size is 3.85 Mbp, G+C percentage is 58.29%, with 99.76% completeness
701 and 0.96% contamination. Strain CLA-KB-H122 was determined to be a new species
702 based on 16S rRNA gene analysis, with the closest validly named match being
703 *Alistipes timonensis* (98.17%). Separation from existing *Alistipes* species was
704 confirmed by ANI comparison, which gave a value of 91.78% to *A. timonensis*. GTDB-
705 Tk classified strain CLA-KB-H122 as 'Alistipes senegalensis'. However, the latter
706 name is as yet not valid. Functional analysis showed the strain has 83 transporters,
707 17 secretion genes, and predicted utilization of cellulose and starch along with
708 production of L-glutamate and folate. In total, 395 CAZymes were identified, with 59
709 different glycoside hydrolase families and 19 glycoside transferase families
710 represented. The type strain, CLA-KB-H122^T (phylum *Bacteroidota*, family
711 *Rikenellaceae*) (StrainInfo: 10.60712/SI-ID414389.1, genome: GCA_040095975.1),
712 was isolated from human faeces.

713

714 **Description of *Bifidobacterium hominis* sp. nov.**

715 *Bifidobacterium hominis* (ho'mi.nis. L. gen. n. *hominis*, of a human being, pertaining to
716 the human gut habitat, from where the type strain was isolated).

717 The genome size is 2.03 Mbp, G+C percentage is 55.98%, with 99.77% completeness
718 and 0.45% contamination. The closest relative to strain CLA-AA-H311 was
719 *Bifidobacterium pseudocatenulatum* (99.07%) based on 16S rRNA gene analysis.

However, ANI comparison identified strain CLA-AA-H311 as a novel species within the genus *Bifidobacterium*, with an ANI value of 93.18% against *B. pseudocatenulatum*. GTDB-Tk classification as 'Bifidobacterium sp002742445' confirmed the proposition of a novel species. Placement within the genus *Bifidobacterium* was confirmed by the presence of fructose-6-phosphate phosphoketolase (KO1621) [82]. Functional analysis showed the strain has 90 transporters, 20 secretion genes, and predicted utilization of starch and production of propionate, acetate, and folate. In total, 137 CAZymes were identified, with 28 different glycoside hydrolase families and 11 glycoside transferase families represented. The type strain, CLA-AA-H311^T (phylum *Actinomycetota*, family *Bifidobacteriaceae*) (=DSM 118068) (StrainInfo: 10.60712/SI-ID414317.1, genome: GCA_040095915.1), was isolated from human faeces.

Description of *Blautia aquisgranensis* sp. nov.

Blautia aquisgranensis (a.quis.gra.nen'sis. L. masc. adj. *aquisgranensis,* named after the German city of Aachen where it was isolated)

The genome size is 3.63 Mbp, G+C percentage is 43.76%, with 99.37% completeness and 0.32% contamination. It includes two plasmids (37,495 bp; 1,036 bp). The closest relative to strain CLA-JM-H16 was *Blautia intestinalis* (96.11%) based on 16S rRNA gene analysis. Placement of the strain within *Blautia* was confirmed based on POCP comparison as values above 50% to multiple *Blautia* species were obtained. However, comparison to the type species of the genus, *Blautia coccoides*, gave a value of 42.36%. This inconsistency was also highlighted by GTDB-Tk, which classified strain CLA-JM-H16 as 'Blautia_A sp900764225', suggesting *Blautia* may require splitting into multiple genera in future. As the separation of *Blautia* would require detailed analysis which is outside the scope of this manuscript, we propose strain CLA-JM-H16 as a novel species within *Blautia*. All three novel species of *Blautia* described within this manuscript were confirmed to represent distinct species based on ANI comparison (*B. aquisgranensis* Vs. *B. caccae*, 81.74%; *B. aquisgranensis* Vs. *B. intestinihominis*, 76.95%; *B. caccae* Vs. *B. intestinihominis*, 78.57%). Functional analysis revealed 157 transporters, 17 secretion genes, and predicted utilization of arbutin, salicin, sucrose, starch, and production of acetate, propionate, folate, L-glutamate, riboflavin, and cobalamin. In total, 205 CAZymes were identified, with 40 different glycoside hydrolase families and 12 glycoside transferase families represented. The type strain, CLA-JM-H16^T (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 114586, =LMG 33033) (StrainInfo: 10.60712/SI-ID414368.1, genome: GCA_040096615.1), was isolated from human faeces.

Description of *Blautia caccae* sp. nov.

759 *Blautia caccae* (cac'cae. N.L. fem. n. *cacca*, human ordure, feces; from Gr. fem. n.
760 *kakkê*, human ordure, feces; N.L. gen. n. *caccae*, of feces, referring to the source of
761 isolate).

762 The genome size is 5.83 Mbp, G+C percentage is 46.73%, with 98.73% completeness
763 and 0.63% contamination. The closest relative to strain CLA-SR-H028 was *Blautia*
764 *hominis* (98.66%) based on 16S rRNA gene analysis. ANI comparison identified CLA-
765 SR-H028 as a novel species within the genus *Blautia*, with all values being below the
766 species threshold. GTDB-Tk classification as 'Blautia sp001304935' confirmed the
767 proposition of a novel species within *Blautia*. Functional analysis showed the strain
768 has 158 transporters, 18 secretion genes, and predicted utilization of cellobiose,
769 sucrose, starch and production of propionate, acetate, cobalamin, and folate. In total,
770 353 CAZymes were identified, with 53 different glycoside hydrolase families and 15
771 glycoside transferase families represented. The type strain, CLA-SR-H028^T (phylum
772 *Bacillota*, family *Lachnospiraceae*) (Straininfo: 10.60712/SI-ID414428.1, genome:
773 GCA_040095955.1), was isolated from human faeces.

774

775 **Description of *Blautia intestinihominis* sp. nov.**

776 *Blautia intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, intestine; L. masc.
777 n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human intestine).

778 The genome size is 4.1 Mbp, G+C percentage is 43.49%, with 98.73% completeness
779 and 0.63% contamination. It includes a single plasmid of 22,629 bp. The isolate was
780 assigned to the species *Blautia obeum* (98.98%) based on 16S rRNA gene analysis.
781 However, ANI comparison to *B. obeum* clearly identified this isolate as being a
782 separate species (84.16%). This was confirmed by GTDB-Tk classification as
783 'Blautia_A sp000436615', recommending the creation of a novel species. Functional
784 analysis showed the strain has 155 transporters, 18 secretion genes, and predicted
785 utilization of sucrose and starch, along with production of L-glutamate, folate,
786 propionate, and cobalamin. In total, 177 CAZymes were identified. The type strain,
787 CLA-AA-H95^T (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 111354) (StrainInfo:
788 10.60712/SI-ID414326.1, genome: GCA_040096655.1), was isolated from human
789 faeces.

790

791 **Description of *Coprococcus intestinihominis* sp. nov.**

792 *Coprococcus intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, intestine; L.
793 masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human intestine).

794 The genome size is 3.6 Mbp, G+C percentage is 43.29%, with 98.43% completeness
795 and 2.52% contamination. A single plasmid of 20,255 bp was detected. The closest
796 relative to strain CLA-AA-H190 was *Coprococcus catus* (96.76%) based on 16S rRNA
797 gene analysis. ANI comparison identified CLA-AA-H190 as a novel species within the
798 genus *Coprococcus*, with an ANI value of 90.25% against the closest relative *C. catus*.

GTDB-Tk classification as 'Coprococcus_A catus_A' confirmed the proposition of a novel species, but also suggests that separation of *Coprococcus* into multiple genera could occur in future. Functional analysis showed the strain has 119 transporters, 18 secretion genes, and predicted utilization of starch and production of propionate, butyrate, acetate, cobalamin, and folate. In total, 122 CAZymes were identified, with 19 different glycoside hydrolase families and 13 glycoside transferase families represented. The type strain, CLA-AA-H190^T (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 114688, =LMG 33015) (StrainInfo: 10.60712/SI-ID414125.1, genome: GCA_040096555.1), was isolated from human faeces.

Description of *Enterobacter intestinihominis* sp. nov.

Enterobacter intestinihominis (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, the intestine; L. masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human intestine).

The genome size is 4.86 Mbp, G+C percentage is 54.88%, with 99.89% completeness and 0.12% contamination. It contains two plasmids (4,416 bp; 2,494 bp). Strain CLA-AC-H004 was determined to be a strain of *Enterobacter quasihormaechei* (99.80%) based on 16S rRNA gene analysis. Separation from existing *Enterobacter* species was confirmed by ANI comparison, which gave a value of 93.69% to *E. quasihormaechei*. GTDB-Tk classification of strain CLA-AC-H004 as 'Enterobacter hormaechei_A' supports the proposal of a novel species. Functional analysis showed the strain has 497 transporters, 98 secretion genes, and predicted utilization of arbutin, salicin, cellobiose, sucrose, and starch along with production of L-glutamate, biotin, riboflavin, acetate, propionate, and folate. In total, 316 CAZymes were identified, with 37 different glycoside hydrolase families and 19 glycoside transferase families represented. The type strain, CLA-AC-H004^T (phylum *Pseudomonadota*, family *Enterobacteriaceae*) (StrainInfo: 10.60712/SI-ID414328.1, genome: GCA_040096145.1), was isolated from human faeces.

Description of *Enterocloster hominis* sp. nov.

Enterocloster hominis (ho'mi.nis. L. gen. n. *hominis*, of a human being).

The genome size is 6.52 Mbp, G+C percentage is 50.14%, with 99.16% completeness and 2.53% contamination. It includes a single plasmid of 7,635 bp. The closest relative to strain CLA-SR-H021 was *Enterocloster aldenensis* (98.44%) based on 16S rRNA gene analysis. ANI comparison identified CLA-SR-H021 as a novel species within the genus *Enterocloster*, with all values being below the species threshold. GTDB-Tk classified CLA-SR-H021 as 'Enterocloster pacaense', a name derived from the proposed species 'Lachnoclostridium pacaense'. However, the fact that these two names are not valid supports the proposition of a novel species within *Enterocloster*. Functional analysis revealed 118 transporters, 14 secretion genes, and predicted utilization of cellobiose, starch and production of propionate, acetate, and folate. In

840 total, 307 CAZymes were identified, with 39 different glycoside hydrolase families and
841 14 glycoside transferase families represented. The type strain, CLA-SR-H021^T
842 (phylum *Bacillota*, family *Lachnospiraceae*) (StrainInfo: 10.60712/SI-ID414422.1,
843 genome: GCA_040096035.1), was isolated from human faeces.

844

845 **Description of *Faecalibacterium faecis* sp. nov.**

846 *Faecalibacterium faecis* (fae'cis. L. gen. fem. n. *faecis*, of dregs, pertaining to faeces,
847 from where the type strain was isolated).

848 The genome size is 2.71 Mbp, G+C percentage is 56.23%, with 100.0% completeness
849 and 0.0% contamination. The isolate was determined to be similar to *Faecalibacterium*
850 *butyricigenerans* (99.57%) based on 16S rRNA gene analysis. However, ANI similarity
851 to *F. butyricigenerans* was 84.61% and the closest relative based on ANI comparison
852 was *F. prausnitzii* (87.03%). These values are below the threshold for species
853 assignment, suggesting the creation of a novel species. This was supported via
854 GTDB-Tk classification of the isolate as 'Faecalibacterium prausnitzii_D'. Functional
855 analysis showed the strain has 130 transporters, 16 secretion genes, and predicted
856 utilization of cellobiose, starch, and production of L-glutamate, riboflavin, and
857 cobalamin. Production of butyrate (6.12 ± 0.53 mM) was confirmed for strain CLA-JM-
858 H7B when grown in YCFA broth (DSMZ Medium No. 1611) in Hungate tubes for 48
859 hours at 37 °C under anaerobic conditions (6% CO_2 and 4,7% H_2 in N_2). In total, 157
860 CAZymes were identified, with 25 different glycoside hydrolase families and 12
861 glycoside transferase families represented. The type strain, CLA-JM-H7-B^T (phylum
862 *Bacillota*, family *Oscillospiraceae*) (=DSM 114587) (StrainInfo: 10.60712/SI-
863 ID414382.1, genome: GCA_040096475.1), was isolated from human faeces.

864

865 **Description of *Faecalibacterium intestinale* sp. nov.**

866 *Faecalibacterium intestinale* (in.tes.ti.na'le. N.L. neut. adj. *intestinale*, pertaining to the
867 intestine, from where the type strain was isolated).

868 The genome size is 2.97 Mbp, G+C percentage is 56.43%, with 100.0% completeness
869 and 0.0% contamination. The isolate was determined to be related to *F. prausnitzii*
870 (98.08%) based on 16S rRNA gene analysis. ANI comparison to *F. prausnitzii* was
871 just below species level assignment (94.46%), and GTDB-Tk classification as
872 'Faecalibacterium prausnitzii_J' recommended the creation of a novel species.
873 Functional analysis showed the strain has 135 transporters, 18 secretion genes, and
874 predicted utilization of starch and production of L-glutamate, riboflavin, and cobalamin.
875 Production of butyrate (4.74 ± 0.30 mM) was confirmed for strain CLA-AA-H281 when
876 grown in YCFA broth (DSMZ Medium No. 1611) in Hungate tubes for 48 hours at 37
877 °C under anaerobic conditions (6% CO_2 and 4,7% H_2 in N_2). In total, 159 CAZymes
878 were identified, with 27 different glycoside hydrolase families and 12 glycoside
879 transferase families represented. The type strain, CLA-AA-H281^T (phylum *Bacillota*,

880 family *Oscillospiraceae*) (=DSM 116193, =LMG 33027) (StrainInfo: 10.60712/SI-
881 ID414306.1, genome: GCA_040096575.1), was isolated from human faeces.

882

883 **Description of *Faecalibacterium tardum* sp. nov.**

884 *Faecalibacterium tardum (tar'dum. L. neut. adj. tardum*, pertaining to its slow growth).

885 The genome size is 3.04 Mbp, G+C percentage is 56.28%, with 99.32% completeness
886 and 0.0% contamination. It contains one plasmid of 14,735 bp in size, which encodes
887 for vancomycin resistance via the *vanY* gene. The isolate was determined to be closely
888 related to *Faecalibacterium prausnitzii* (98.70%) based on 16S rRNA gene analysis.
889 ANI comparison to *F. prausnitzii* was on the border of species level assignment
890 (95.05%), however, GTDB-Tk classification as 'Faecalibacterium prausnitzii_A'
891 recommended the creation of a novel species. Strain CLA-AA-H175 was confirmed to
892 represent a distinct species to *Faecalibacterium intestinale* (CLA-AA-H281) (ANI:
893 94.39%) and *Faecalibacterium faecis* (CLA-JM-H7B) (ANI: 86.46%), also described in
894 this paper. Functional analysis showed the strain has 116 transporters, 18 secretion
895 genes, and predicted utilization of starch, and production of L-glutamate. Production
896 of butyrate (0.81 ± 0.16 mM) was confirmed for strain CLA-AA-H175 when grown in
897 YCFA broth (DSMZ Medium No. 1611) in Hungate tubes for 48 hours at 37 °C under
898 anaerobic conditions (6% CO_2 and 4,7% H_2 in N_2). In total, 177 CAZymes were
899 identified, with 28 different glycoside hydrolase families and 12 glycoside transferase
900 families represented. The type strain, CLA-AA-H175^T (phylum *Bacillota*, family
901 *Oscillospiraceae*) (=DSM 116192) (StrainInfo: 10.60712/SI-ID414281.1, genome:
902 GCA_040096515.1), was isolated from human faeces.

903

904 **Description of *Faecousia* gen. nov.**

905 Faecousia (Faec.ou'si.a. L. fem. n. *faex*, dregs; Gr. fem. n. *ousia*, an essence; N.L.
906 fem. n. *Faecousia*, a microbe associated with faeces).

907 Based on 16S rRNA gene sequence identity, the closest relatives are members of the
908 genus *Vescimonas* (*Vescimonas fastidiosa*, 93.48%). POCP analysis against
909 *Vescimonas coprocola* (44.75%), the type species of this genus, and *V. fastidiosa*
910 (41.06%) confirmed strain CLA-AA-H192 represents a distinct genus to *Vescimonas*.
911 GTDB-Tk supported the creation of a novel genus, placing strain CLA-AA-H192 within
912 the proposed genus of "*Candidatus Faecousia"*. The type species of this genus is
913 *Faecousia intestinalis*.

914

915 **Description of *Faecousia intestinalis* sp. nov.**

916 *Faecousia intestinalis* (in.tes.ti.na'lis. N.L. fem. adj. *intestinalis*, pertaining to the
917 intestines, from where the type strain was isolated).

The genome size is 2.98 Mbp, G+C percentage is 58.44%, with 93.29% completeness and 0.0% contamination. It contains two plasmids (23,094 bp; 3,448 bp). Functional analysis showed the strain has 144 transporters, 26 secretion genes, and predicted utilization of starch, and production of acetate, propionate, L-glutamate, and folate. In total, 116 CAZymes were identified, with 19 different glycoside hydrolase families and 11 glycoside transferase families represented. The type strain, CLA-AA-H192^T (phylum *Bacillota*, family *Oscillospiraceae*) (StrainInfo: 10.60712/SI-ID414286.1, genome: GCA_040096185.1), was isolated from human faeces.

Description of *Flavonifractor hominis* sp. nov.

Flavonifractor hominis (ho'mi.nis. L. gen. n. *hominis*, of a human being).

The genome size is 2.94 Mbp, G+C percentage is 58.64%, with 99.33% completeness and 0.0% contamination. Strain CLA-AP-H34 was determined to be a new species based on 16S rRNA gene sequence analysis, with the closest validly named match being *Flavonifractor plautii* (97.21%). Separation from existing *Flavonifractor* species was confirmed by ANI comparison, which gave a value of 82.25% to *F. plautii*. GTDB-Tk classification of strain CLA-AP-H34 as an unknown species within *Flavonifractor* supports the proposal of a novel species. Functional analysis showed the strain has 112 transporters, 14 secretion genes, and predicted utilization of starch and production of L-glutamate, riboflavin, butyrate, and folate. In total, 140 CAZymes were identified, with 19 different glycoside hydrolase families and 15 glycoside transferase families represented. The type strain, CLA-AP-H34^T (phylum *Bacillota*, family *Oscillospiraceae*) (StrainInfo: 10.60712/SI-ID414347.1, genome: GCA_040095835.1), was isolated from human faeces.

Description of *Hominicola* gen. nov.

Hominicola (Ho.mi.ni.co.la. L. masc. n. homo, human being, man L. fem. n. suff. -cola, inhabitant, deweller; N.L. fem. n. *Hominicola*, a microbe dwelling within human beings).

Based on 16S rRNA gene sequence identity, the closest relative is *Negativibacillus massiliensis* (94.53%), the type and singular species within the genus *Negativibacillus* within the *Bacillota*. POCP analysis confirmed strain CLA-JM-H45 represents a distinct genus to *Negativibacillus,* with 43.21% to *N. massiliensis*. GTDB-Tk supported the creation of a novel genus, placing strain CLA-AA- H45 within the genus '*Phocae'*, which is an invalid name. Therefore, creation of the genus *Hominicola* within family *Oscillospiraceae* is proposed to accommodate this isolate. The type species of this genus is *Hominicola intestinalis*.

Description of *Hominicola intestinalis* sp. nov.

957 *Hominicola intestinalis* (in.tes.ti.na'lis. N.L. fem. adj. *intestinalis*, pertaining to the
958 intestines, from where the type strain was isolated).

959 The genome size is 2.24 Mbp, G+C percentage is 56.25%, with 95.3% completeness
960 and 0.67% contamination. It contains a plasmid (3,541 bp). Functional analysis
961 showed the strain has 95 transporters, 13 secretion genes, and predicted utilization of
962 starch, and production of acetate, propionate, L-glutamate, and cobalamin. In total, 79
963 CAZymes were identified, with 15 different glycoside hydrolase families and 11
964 glycoside transferase families represented. The type strain, CLA-JM-H45^T (phylum
965 *Bacillota*, family *Oscillospiraceae*) (=DSM 114602, =LMG 33035) (StrainInfo:
966 10.60712/SI-ID414378.1, genome: GCA_040096325.1), was isolated from human
967 faeces.

968

969 **Description of *Hominiventricola aquisgranensis* sp. nov.**

970 *Hominiventricola aquisgranensis* (a.quis.gra.nen'sis. L. masc. adj. *aquisgranensis,*
971 named after the German city of Aachen where it was isolated).

972 The genome size is 3.17 Mbp, G+C percentage is 45.01%, with 99.37% completeness
973 and 0.27% contamination, including a single plasmid of 7,983 bp. The closest relative
974 to strain CLA-AA-H78B was *Hominiventricola filiformis* (96.24%) based on 16S rRNA
975 gene analysis. POCP analysis confirmed genus assignment to *Hominiventricola*, with
976 a POCP value of 69.06% to the type strain of the only current species within this genus,
977 *H. filiformis*. GTDB-Tk classified strain CLA-AA-H78B as 'Choladocola sp003480725'
978 within "*Candidatus* Choladocola". The latter is a heterosynonym of *Hominiventricola*,
979 a validly published name. ANI comparison confirmed that strain CLA-AA-H78B
980 represents a novel species as the ANI value to *H. filiformis* was 82.98%. Functional
981 analysis showed the strain has 134 transporters, 31 secretion genes, and predicted
982 utilization of cellobiose, starch, arbutin, salicin, and production of L-glutamate, folate,
983 acetate, propionate, and cobalamin. In total, 134 CAZymes were identified. The type
984 strain, CLA-AA-H78B^T (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 111355)
985 (StrainInfo: 10.60712/SI-ID414323.1, genome: GCA_040096225.1), was isolated
986 from human faeces.

987

988 **Description of *Lachnospira hominis* sp. nov.**

989 *Lachnospira hominis* (ho'mi.nis. L. gen. masc. n. *hominis*, of a human being).

990 The genome size is 3.1 Mbp, G+C percentage is 41.75%, with 99.33% completeness
991 and 0.0% contamination. Strain CLA-AA-H89B was determined to represent a
992 separate species to its closest relative, *Lachnospira pectinoschiza* (97.48%), based
993 on 16S rRNA gene analysis. This was confirmed based on ANI comparison to all close
994 relatives which were below the species threshold. GTDB-Tk classification of strain
995 CLA-AA-H89B as 'Lachnospira sp000437735' confirmed that this isolate represents a
996 novel species. Functional analysis showed the strain has 112 transporters, 29

997 secretion genes, and predicted utilization of starch and cellulose, along with
998 production of L-glutamate, folate, acetate, propionate, and riboflavin. Motility was
999 predicted based on the presence of the following genes: *FlhA, FlhB, FlgB, FlgC, FlgD,*
1000 *FlgE, FlgF, FlgJ, FlgK, FlgL, FliC, FliD, FliE, FliF, FliG, FliK, FliM, FliN, MotA, MotB*. In
1001 total, 162 CAZymes were identified, with 21 different glycoside hydrolase families and
1002 11 glycoside transferase families represented. The type strain, CLA-AA-H89B^T
1003 (phylum *Bacillota*, family *Lachnospiraceae*) (StrainInfo: 10.60712/SI-ID414325.1,
1004 genome: GCA_040095895.1), was isolated from human faeces.

1005

1006 **Description of *Lachnospira intestinalis* sp. nov.**

1007 *Lachnospira intestinalis* (in.tes.ti.na'lis. N.L. fem. adj. *intestinalis*, pertaining to the
1008 intestines, from where the type strain was isolated).

1009 The genome size is 3.15 Mbp, G+C percentage is 37.05%, with 98.66% completeness
1010 and 0.67% contamination. Strain CLA-JM-H10 was determined to represent a
1011 separate species to its closest relative, *Lachnospira rogosae* (96.37%), based on 16S
1012 rRNA gene analysis. This was confirmed by an ANI of 79.9% between *L. rogosa* (CLA-
1013 AA-H255), and strain CLA-JM-H10. GTDB-Tk classification of CLA-JM-H10 as
1014 'Lachnospira sp900316325' confirmed that it represents a novel species. Functional
1015 analysis showed the strain has 109 transporters, 27 secretion genes, and predicted
1016 utilization of starch, and production of L-glutamate, folate, and cobalamin. Motility was
1017 predicted based on detection of the following genes: *FlhA, FlhB, FlgB, FlgC, FlgD,*
1018 *FlgE, FlgF, FlgJ, FlgK, FlgL, FliC, FliD, FliE, FliF, FliG, FliK, FliM, FliN, MotA, MotB*.
1019 This is consistent with the type species of this genus being motile. In total, 165
1020 CAZymes were identified, with 22 different glycoside hydrolase families and 12
1021 glycoside transferase families represented. The type strain, CLA-JM-H10^T (phylum
1022 *Bacillota*, family *Lachnospiraceae*) (=DSM 114599) (StrainInfo: 10.60712/SI-
1023 ID414366.1, genome: GCA_040096395.1), was isolated from human faeces.

1024

1025 **Description of *Lachnospira rogosae* sp. nov.**

1026 *Lachnospira rogosae* (ro.go'sae. N.L. gen. masc. n. *rogosae*, of Rogosa).

1027 Strain CLA-AA-H255 was determined to be similar to *Lactobacillus rogosae* (99.87%)
1028 based on 16S rRNA gene analysis. However, the lack of a genome for the type strain
1029 of the latter species, along with the lack of the type strain at any established culture
1030 collection prevented further comparison (Tindall, 2014). GTDB-Tk classification of
1031 CLA-AA-H255 as 'Lachnospira rogosae_A' suggested that it represents a species
1032 within a genus distantly related to *Lactobacillus*. This was reconfirmed by the same
1033 placement of a second strain, CLA-AA-H191, which was also assigned to the same
1034 placeholder by GTDB-Tk (ANI of 98.86% between our two isolates). Based on these
1035 results, we propose *L. rogosae* was previously misassigned as a member of the genus
1036 *Lactobacillus*. To provide a type strain, and correct its placement, we propose the
1037 creation of the species *Lachnospira rogosae*. Functional analysis showed that strain

1038 CLA-AA-H255 has 100 transporters, 25 secretion genes, and predicted utilization of
1039 starch, and production of L-glutamate, folate, and riboflavin. The prediction of motility
1040 and acetate production based on genomic analysis is consisting with the observed
1041 phenotype of motility in the original type strain of *L. ragosae* as stated by Holdeman
1042 and Moore (Holdeman and Moore, 1974). In total, 142 CAZymes were identified,
1043 including 20 different glycoside hydrolase families and 12 glycoside transferase
1044 families. Ecological analysis of 1,000 human gut 16S rRNA gene amplicon samples
1045 identified this strain in 1.20% of samples with a relative abundance of 0.43 ± 0.78%.
1046 The type strain, CLA-AA-H255^T (phylum *Bacillota*, family *Lachnospiraceae*)
1047 (StrainInfo: 10.60712/SI-ID414300.1, genome: GCA_040096455.1), was isolated
1048 from human faeces.

1049

1050 **Description of *Laedolimicola intestinihominis* sp. nov.**

1051 *Laedolimicola intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, intestine; L.
1052 masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human intestine).

1053 The genome size is 3.45 Mbp, G+C percentage is 49.65%, with 99.37% completeness
1054 and 0.16% contamination. It includes a single plasmid of 5,131 bp. Strain CLA-AA-
1055 H132 was determined to represent a separate species to its closest relative,
1056 *Laedolimicola ammoniilytica* (98.38%), based on 16S rRNA gene analysis. POCP
1057 analysis confirmed that strain CLA-AA-H132 belongs to the recently named genus,
1058 *Laedolimicola*, with a POCP value of 75.19% to the type strain of the only current
1059 species within this genus, *L. ammoniilytica*. GTDDB-Tk placement as 'Merdisoma
1060 sp900553635' suggests that *Laedolimicola* and *Candidatus* Merdisoma are
1061 homonyms and require future reclassification. ANI comparison confirmed that CLA-
1062 AA-H132 represents a novel species, as the ANI value to *L. ammoniilytica* was only
1063 90.2%. Functional analysis showed the strain has 142 transporters, 21 secretion
1064 genes, and predicted utilization of cellobiose, sucrose, starch, and production of
1065 acetate, propionate, L-glutamate, cobalamin, and folate. In total, 153 CAZymes were
1066 identified, with 21 different glycoside hydrolase families and 16 glycoside transferase
1067 families represented. The type strain, CLA-AA-H132^T (phylum *Bacillota*, family
1068 *Lachnospiraceae*) (=DSM 117481) (StrainInfo: 10.60712/SI-ID414271.1, genome:
1069 GCA_040096155.1), was isolated from human faeces.

1070

1071 **Description of *Maccoya* gen. nov.**

1072 *Maccoya* (Mac.coy'a. N.L. fem. n. *Maccoya*, named after the immunologist Kathy
1073 McCoy, for her many scientific contributions in the field of microbe-host interactions).

1074 Based on 16S rRNA gene analysis, the closest species with a valid name are
1075 *Roseburia hominis* (94.9%) and *Eubacterium oxidoreducens* (94.23%). POCP values
1076 against all close relatives were below the genus delineation threshold of 50%,
1077 including many *Roseburia* spp. (*R. hominis*, *R. faecis*, *R. porci*, *R. intestinalis*), apart
1078 from *R. inulinivorans* (50.83%). GTDDB-Tk placement as "UBA11774 sp003507655"

confirmed that strain CLA-AA-H185 represents a novel genus within the family *Lachnospiraceae*. The type species is *Maccoya intestinihominis*.

Description of *Maccoya intestinihominis* sp. nov.

Maccoya intestinihominis (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, intestine; L. masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human intestine).

The genome size is 2.85 Mbp, G+C percentage is 41.17%, with 97.41% completeness and 0.45% contamination. It contains a single plasmid (2,341 bp). Functional analysis showed the strain has 99 transporters, 35 secretion genes, and predicted utilization of starch, and production of acetate, propionate, L-glutamate, cobalamin, folate, and riboflavin. In total, 110 CAZymes were identified, with 17 different glycoside hydrolase families and 13 glycoside transferase families represented. The type strain, CLA-AA-H185^T (phylum *Bacillota*, family *Lachnospiraceae*) (StrainInfo: 10.60712/SI-ID414284.1, genome: GCA_040096355.1), was isolated from human faeces.

Description of *Megasphaera intestinihominis* sp. nov.

Megasphaera intestinihominis (in.tes.ti.ni.ho'mi.nis. L. gen. neut. n. *intestini*, of the intestine; L. gen. masc. n. *hominis*, of a human being; N.L. gen. masc. n. *intestinihominis*, of the human intestine).

The genome size is 2.38 Mbp, G+C percentage is 53.59%, with 100.00% completeness and 0.00% contamination. The closest relative to strain CLA-AA-H81 was *Megasphaera indica* (99.09%) based on 16S rRNA gene analysis. ANI comparison to all close relatives were below species assignment threshold (highest to *Megasphaera elsdenii,* 90.78%). GTDB-Tk classification as 'Megasphaera sp000417505' supports the creation of a novel species. Functional analysis showed the strain has 133 transporters, 16 secretion genes, and predicted utilization of starch, and production of butyrate, propionate, L-glutamate, folate, riboflavin, and cobalamin. Ecological analysis identified 152 matching (MASH distance < 0.05) MAGs, of which 124 originate from the human gut, suggesting this species is most commonly observed within this environment. This is supported by ecological analysis using 16S rRNA gene amplicon datasets which identified it within 19.0% of 1,000 human gut samples, with a relative abundance of 2.14 ± 5.23%. The type strain, CLA-AA-H81^T (phylum *Bacillota*, family *Veillonellaceae*) (=DSM 118069) (StrainInfo: 10.60712/SI-ID414324.1, genome: GCA_040096415.1), was isolated from human faeces.

Description of *Niallia hominis* sp. nov.

Niallia hominis (ho'mi.nis. L. gen. n. *hominis*, of a human being).

The genome size is 4.9 Mbp, G+C percentage is 35.33%, with 92.24% completeness and 3.20% contamination. Strain CLA-SR-H024 was assigned to *Niallia circulans*

1118 (100%) based on 16S rRNA gene analysis. However, ANI suggested the isolate
1119 represents a novel species as comparison to all *Niallia* spp. were below 80%. GTDB-
1120 Tk classification of strain CLA-SR-H024 as 'Niallia sp001076885' confirmed that it
1121 represents a novel species. Functional analysis showed the strain has 244
1122 transporters, 43 secretion genes, and predicted utilization of arbutin, salicin,
1123 cellobiose, starch, and dextran, along with production of L-glutamate, folate, and
1124 cobalamin. In total, 294 CAZymes were identified, with 34 different glycoside
1125 hydrolase families and 14 glycoside transferase families represented. The type strain,
1126 CLA-SR-H024^T (phylum *Bacillota*, family *Bacillaceae*) (StrainInfo: 10.60712/SI-
1127 ID414424.1, genome: GCA_040095995.1), was isolated from human faeces.

1128

1129 **Description of *Peptoniphilus hominis* sp. nov.**

1130 *Peptoniphilus hominis* (ho'mi.nis. L. gen. n. *hominis*, of a human being).

1131 The genome size is 1.99 Mbp, G+C percentage is 33.8%, with 99.3% completeness
1132 and 0.0% contamination. Strain CLA-SR-H025 was assigned to *Peptoniphilus*
1133 *gorbachii* (99.11%) based on 16S rRNA gene analysis. However, ANI suggested the
1134 isolate represents a novel species as comparison to *P. gorbachii* gave a value of
1135 88.92%. GTDB-Tk classification of strain CLA-SR-H025 as 'Peptoniphilus_A
1136 grossensis' confirmed that it represents a novel species, as the name 'Peptoniphilus
1137 grossensis' was proposed but never validated [83]. Functional analysis showed the
1138 strain has 12 transporters, 1 secretion genes, and no gut metabolic models were
1139 identified within the genome. In total, 59 CAZymes were identified, with 10 different
1140 glycoside hydrolase families and 8 glycoside transferase families represented. The
1141 type strain, CLA-SR-H025^T (phylum *Bacillota*, family *Peptoniphilaceae*) (Straininfo:
1142 10.60712/SI-ID414425.1, genome: GCA_040096015.1), was isolated from human
1143 faeces.

1144

1145 **Description of *Pseudoflavonifractor intestinihominis* sp. nov.**

1146 *Pseudoflavonifractor intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, the
1147 intestine; L. masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the
1148 human intestine).

1149 The genome size is 3.69 Mbp, G+C percentage is 61.54%, with 99.33% completeness
1150 and 0.00% contamination. Strain CLA-AP-H29 was assigned to the species
1151 *Pseudoflavonifractor capillosus* (99.53%) based on 16S rRNA gene sequence
1152 analysis. However, ANI suggested the isolate represents a novel species as
1153 comparison to *P. capillosus* gave a value of 80.96%. GTDB-Tk classified strain CLA-
1154 AP-H29 as 'Pseudoflavonifractor sp944387275', which confirms that it represents a
1155 novel species. Functional analysis predicted the strain has 121 transporters, 14
1156 secretion genes, and predicted utilization of starch, and production of L-glutamate,
1157 folate, and cobalamin. In total, 180 CAZymes were identified, with 25 different
1158 glycoside hydrolase families and 15 glycoside transferase families represented. The

type strain, CLA-AP-H29^T (phylum *Bacillota*, family *Oscillospiraceae*) (=DSM 801173) (Straininfo: 10.60712/SI-ID414342.1, genome: GCA_040096055.1), was isolated from human faeces.

Description of *Robertmurraya yapensis* sp. nov.

Robertmurraya yapensis (yap'ensis. N.L. masc./fem. adj. *yapensis*, pertaining to Yap trench, which is the geographical position where the first isolate of this species was obtained).

The genome size is 4.74 Mbp, G+C percentage is 37.94%, with 98.85% completeness and 2.13% contamination. It contains a plasmid (2,470 bp). The closest relative to strain CLA-AA-H227 was *Robertmurraya spiralis* (99.23%) based on 16S rRNA gene analysis. The highest POCP scores were to members of the genus *Robertmurraya (Robertmurraya kyonggiensis,* 84.9%; *R. spiralis,* 65.33%). The placement of strain CLA-AA-H227 within *Robertmurraya* was confirmed by GTDB-Tk assignment as 'Robertmurraya yapensis', a reclassification of the species 'Bacillus yapensis', although neither name has been validated. ANI comparison confirmed that CLA-AA-H227 represents a novel species, as values to close relatives were below the species level threshold. Functional analysis showed the strain has 246 transporters, 53 secretion genes, and predicted utilization of arbutin, salicin, cellobiose, starch, and production of butyrate, acetate, propionate, folate, riboflavin, and cobalamin. In total, 238 CAZymes were identified. The type strain, CLA-AA-H227^T (phylum *Bacillota*, family *Bacillaceae*) (=DSM 113004, =LMG 33018) (Straininfo: 10.60712/SI-ID414150.1, genome: GCA_040096375.1), was isolated from human faeces.

Description of *Ruminococcoides intestinale* sp. nov.

Ruminococcoides intestinale (in.tes.ti.na'le. N.L. neut. adj. *intestinalis*, pertaining to the intestines, from where the type strain was isolated).

The genome size is 2.32 Mbp, G+C percentage is 40.88%, with 99.33% completeness and 1.01% contamination. The isolate was determined to be similar to *Ruminococcus bromii* (98.91%) and more distantly related to *Ruminococcoides bili* (96.76%) based on 16S rRNA gene analysis. While POCP comparison of strain CLA-JM-H38 to *R. bromii* was 59.79%, and 53.56% to *Ruminococcus bovis,* suggesting they belong to the same genus, all other comparisons to *Ruminococcus* species were below 50%, including to the type species, *Ruminococcus flavefaciens* (27.58%). POCP to *R. bili*, the type species of the genus *Ruminococcoides*, was 64.44%. GTDB-Tk classified strain CLA-JM-H38 as "Ruminococcus_E bromii_B", confirming it is not a member of the genus *Ruminococcus.* These results support GTDB assignment that both *R. bovis* and *R. bromii* should be reclassified as members of the genus *Ruminococcoides*. Strain CLA-JM-H38 was confirmed to represent a novel species as all ANI comparisons to close relatives were below 95%, and it represents a distinct novel species from *Ruminococcoides intestinihominis* described in this work (78.33%).

Functional analysis showed the strain has 81 transporters, 15 secretion genes, and predicted utilization of starch, and production of L-glutamate. In total, 108 CAZymes were identified, with 15 different glycoside hydrolase families and 12 glycoside transferase families represented. Ecological analysis based on 16S rRNA gene amplicons identified this species in 55.20% of 1,000 human gut samples with a relative abundance of 1.50 ± 2.49%, suggesting it is a prevalent and dominant bacterial species within the human gut. The type strain, CLA-JM-H38^T (phylum *Bacillota*, family *Oscillospiraceae*) (StrainInfo: 10.60712/SI-ID414376.1, genome: GCA_040096305.1), was isolated from human faeces.

Description of *Ruminococcoides intestinihominis* sp. nov.

Ruminococcoides intestinihominis (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, intestine; L. masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human intestine).

The genome size is 2.26 Mbp, G+C percentage is 34.26%, with 98.66% completeness and 0.00% contamination. It contains one plasmid (1,825 bp). Based on 16S rRNA gene sequence identity, the isolate was closely related to *Ruminococcus bovis* (99.3%) and more distant to *Ruminococcoides bili* (94.8%), the type species of this genus. ANI comparison confirmed the similarity of strain CLA-AA-H171 to *R. bovis* (95.4%), however classification by GTDB-Tk as 'Ruminococcus_E sp934476515' supported the creation of a novel species. Recently, 'Ruminococcus_E' has been validly named as *Ruminococcoides*, with the type species *R. bili* [40]. POCP comparison between the isolate and *R. bili* provided a value of 51.3%, suggesting that strain CLA-AA-H171 represents a novel species within the genus *Ruminococcoides*. Functional analysis showed the strain has 75 transporters, 14 secretion genes, and predicted utilization of starch, and production of acetate. In total, 124 CAZymes were identified, with 13 different glycoside hydrolase families and 12 glycoside transferase families represented. Ecological analysis based on 16S rRNA gene amplicons identified this species in 10.40% of 1,000 human gut samples with a relative abundance of 0.23 ± 0.68%. The type strain, CLA-AA-H171^T (phylum *Bacillota*, family *Oscillospiraceae*) (=DSM 114689) (StrainInfo: 10.60712/SI-ID414279.1, genome: GCA_040096285.1), was isolated from human faeces.

Description of *Ruthenibacterium intestinale* sp. nov.

Ruthenibacterium intestinale (in.tes.ti.na'le. N.L. neut. adj. intestinale, pertaining to the intestines, from where the type strain was isolated).

The genome size is 3.1 Mbp, G+C percentage is 54.69%, with 98.3% completeness and 0.00% contamination. It contains a plasmid (5,063 bp). The closest relative to strain CLA-JM-H11 is *Ruthenibacterium lactatiformans* (94.93%), the type species of this genus, based on 16S rRNA gene analysis. Placement of the isolate within the genus *Ruthenibacterium* was confirmed based on POCP comparison, with a value of

56.41% to the type species. Strain CLA-JM-H11 was confirmed to be distinct to *R. lactatiformans* based on ANI comparison (79.28%). GTDB-Tk classification confirmed this assignment as an unknown species within *Ruthenibacterium*. Functional analysis showed the strain has 143 transporters, 16 secretion genes, and predicted utilization of starch, and production of acetate, propionate, folate, and cobalamin. In total, 155 CAZymes were identified, with 27 different glycoside hydrolase families and 13 glycoside transferase families represented. The type strain, CLA-JM-H11^T (phylum *Bacillota*, family *Oscillospiraceae*) (=DSM 114604, =LMG 33032) (StrainInfo: 10.60712/SI-ID414367.1, genome: GCA_040096265.1), was isolated from human faeces.

Description of *Sakamotonia* gen. nov.

Sakamotonia (Sa.ka.mo.to.ni'a. N.L. fem. n. *Sakamotonia*, named after the Japanese microbiologist Mitsuo Sakamoto, for his many scientific contributions in the field of bacterial taxonomy).

Based on 16S rRNA gene sequence identity, the closest relatives are members of the genera *Clostridium* (*Clostridium indicum*, 94.0%) and *Lacrimispora* (*Lacrimispora amygdalina*, 94.0%; *Lacrimispora sphenoides*, 93.01%). POCP analysis confirmed that strain CLA-AA-H142 represents a distinct genus to *Clostridium* (*C. indicum*, 40.1%) and *Lacrimispora* (*L. sphenoides*, 37.2%) as all POCP values to close relatives were below 50%. GTDB-Tk supported the creation of a novel genus within the family *Lachnospiraceae*, placing strain CLA-AA-H142 within the proposed genus "*Candidatus* Copromonas". However, as the name "Copromonas" is a homonym of an existing genus, *Copromonas* (phylum: *Euglenozoa*) [84] within the Eukaryota, we have named this genus *Sakamotonia*. The type species is *Sakamotonia hominis*.

Description of *Sakamotonia hominis* sp. nov.

Sakamotonia hominis (ho'mi.nis. L. gen. masc. n. *hominis*, of a human being).

The genome size is 3.58 Mbp, G+C percentage is 48.07%, with 99.37% completeness and 0.16% contamination. It includes a single plasmid of 5,504 bp. GTDB-Tk placed CLA-AA-H142 within "*Candidatus* Copromonas" as "Copromonas sp900066535". This placement was consistent with strain CLA-AA-H58, a second strain assigned to this species (ANI of 97.4% between the two strains). Functional analysis showed that strain CLA-AA-H142 has 129 transporters, 30 secretion genes, and predicted utilization of starch, and production of acetate, propionate, L-glutamate, cobalamin, folate, and riboflavin. In total, 119 CAZymes were identified, with 17 different glycoside hydrolase families and 9 glycoside transferase families represented. The type strain, CLA-AA-H142^T (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 111146) (StrainInfo: 10.60712/SI-ID414272.1, genome: GCA_040096595.1), was isolated from human faeces.

1281

1282 **Description of *Solibaculum intestinale* sp. nov.**

1283 *Solibaculum intestinale (in.tes.ti.na'le. N.L. neut. adj. intestinale,* pertaining to the
1284 intestines, from where the type strain was isolated).

1285 The genome size is 2.81 Mbp, G+C percentage is 54.62 %, with 97.99% completeness
1286 and 0.67% contamination. The isolate was determined to be a new species based on
1287 16S rRNA gene analysis, with the closest validly named match being *Solibaculum*
1288 *mannosilyticum* (94.93%). Placement within *Solibaculum* was confirmed with POCP
1289 values above 50% to *S. mannosilyticum* (55.7%), the type species, and only member
1290 of this genus. ANI comparison confirmed the isolate represents a novel species, as no
1291 ANI values were above 80%. Functional analysis showed the strain has 82
1292 transporters, 16 secretion genes, and predicted utilization of starch, and production of
1293 acetate, propionate, and L-glutamate. In total, 169 CAZymes were identified, with 20
1294 different glycoside hydrolase families and 15 glycoside transferase families
1295 represented. The type strain, CLA-JM-H44^{T} (phylum *Bacillota*, family
1296 *Oscillospiraceae*) (=DSM 114601, =LMG 33034) (StrainInfo: 10.60712/SI-
1297 ID414377.1, genome: GCA_040096205.1), was isolated from human faeces.

1298

1299 **Description of *Ventrimonas* gen. nov.**

1300 *Ventrimonas* (Ven.tri.mo'nas. L. masc. n. *venter*, the belly; L. fem. n. *monas*, a monad;
1301 N.L. fem. n. *Ventrimonas*, a microbe associated with faeces).

1302 Based on 16S rRNA gene sequence identity, the closest relatives are members of the
1303 genus *Hungatella* (*Hungatella effluvii*, 95.15%; *Hungatella hathewayi*, 94.95%). POCP
1304 analysis against *H. effluvii* (40.38%) and *H. hathewayi* (39.86%) indicates that strain
1305 CLA-AP-H27 represents a distinct genus to *Hungatella*. GTDB-Tk supported the
1306 creation of a novel genus, placing strain CLA-AP-H27 within the proposed genus of
1307 "*Candidatus* Ventrimonas*"* in the family *Lachnospiraceae*. The type species of this
1308 genus is *Ventrimonas faecis*.

1309

1310 **Description of *Ventrimonas faecis* sp. nov.**

1311 *Ventrimonas faecis* (fae'cis. L. gen. n. *faecis*, of dregs, pertaining to faeces, from
1312 where the type strain was isolated).

1313 The genome size is 3.41 Mbp, G+C percentage is 49.27%, with 98.73% completeness
1314 and 0.63% contamination. Functional analysis showed the strain has 156 transporters,
1315 21 secretion genes, and predicted utilization of starch, and production of acetate,
1316 propionate, L-glutamate, cobalamin, and folate. In total, 139 CAZymes were identified,
1317 with 21 different glycoside hydrolase families and 12 glycoside transferase families
1318 represented. The type strain, CLA-AP-H27^{T} (phylum *Bacillota*, family

Lachnospiraceae) (=DSM 118072) (StrainInfo: 10.60712/SI-ID414341.1, genome: GCA_040096075.1), was isolated from human faeces.

Description of *Waltera hominis* sp. nov.

Waltera hominis (ho'mi.nis. L. gen. masc. n. *hominis*, of a human being, pertaining to the human gut habitat, from where the type strain was isolated).

The genome size is 3.88 Mbp, G+C percentage is 45.72%, with 99.52% completeness and 2.13% contamination. The isolate was assigned to the species *Waltera intestinalis* (100.0%) based on 16S rRNA gene analysis. However, ANI comparison identified strain CLA-AA-H183 as a novel species within the genus *Waltera*, with an ANI value of 91.97% against the type species *Waltera intestinalis*. GTDB-Tk currently lacks the inclusion of *Waltera*, causing misclassification as 'Acetatifactor sp003447295', however this confirmed the proposition of a novel species. Separation of *Waltera* from *Acetatifactor* was revalidated via both phylogenomic analysis (**Supplementary Figure 8**), which shows both genera form separate monophyletic groups, and POCP analysis, which shows clear similarity within each genus (*Waltera*: 68.86%, *Acetatifactor*: 59.25%), and separation between the genera (39.94 ± 1.63 %). Functional analysis showed the strain has 141 transporters, 36 secretion genes, and predicted utilization of starch, cellulose, and production of butyrate, propionate, acetate, and folate. In total, 228 CAZymes were identified, with 38 different glycoside hydrolase families and 14 glycoside transferase families represented. The type strain, CLA-AA-H183^T (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 114684) (StrainInfo: 10.60712/SI-ID414283.1, genome: GCA_040096245.1), was isolated from human faeces.

Data availability

The genomes for all strains have been deposited at NCBI under BioProject: PRJNA996881. Bulk download of HiBC resources is possible via Zenodo for the genomes (https://doi.org/10.5281/zenodo.12180083), plasmid sequences (https://doi.org/10.5281/zenodo.12187897), 16S rRNA gene sequences (https://doi.org/10.5281/zenodo.12180259) and the isolates metadata (https://doi.org/10.5281/zenodo.12180506).

Acknowledgments

We are thankful to the NGS Competence Center Tübingen (NCCT) and the IZKF Core Facility Genomics at the University Hospital of RWTH Aachen for their support with genome sequencing. We thank Patrick Buchta from the Audio-Visual department from the University Hospital of RWTH Aachen for designing the HiBC logo, Marzena Wyschkon from the German Collection of Microorganisms and Cell Cultures for help with the deposition of strains, and Maurice Heizer for managing the HiBC at the University Hospital of RWTH Aachen. We are also grateful to Peter Vandamme,

Claudine Vereecke, and Anneleen Wieme for handling the submission of strains at the BCCM/LMG Bacteria Collection.

Funding

TCAH received funding from the German Research Foundation (DFG) as part of SFB1382 Gut-liver axis and from the RWTH Aachen START program, project "LeakyGut". TC received funding from the German Research Foundation (DFG): project no. 513892404; project no. 395357507 – SFB1371 Microbiome signatures, subproject Z01; project no. 403224013 – SFB1382 Gut-liver axis, subproject Q02; project no. 460129525 – NFDI4Microbiota. The data used in this publication was managed using the research data management platform Coscine with storage space granted by the Research Data Storage (RDS) of the DFG and Ministry of Culture and Science of the State of North Rhine-Westphalia (DFG: INST222/1261-1 and MKW: 214-4.06.05.08 - 139057).

Conflicts of interest

The authors have no conflicts of interest

References

1. Diakite, A. *et al.* Extensive culturomics of 8 healthy samples enhances metagenomics efficiency. *PLoS One* **14**, 1–12 (2019).
2. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).
3. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053-2067.e18 (2021).
4. Liu, C. *et al.* Enlightening the taxonomy darkness of human gut microbiomes with a cultured biobank. *Microbiome* **9**, 1–29 (2021).
5. Browne, H. P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
6. Diop, K. *et al.* 'Lascolabacter vaginalis' strain KHD1, A new bacterial species cultivated from human female genital tract. *New Microbes New Infect.* **13**, 22–24 (2016).
7. Hedlund, B. P. *et al.* SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nat. Microbiol.* **7**, 1702–1708 (2022).
8. Göker, M., Moore, E. R. B., Oren, A. & Trujillo, M. E. Status of the SeqCode in the International Journal of Systematic and Evolutionary Microbiology. *Int. J. Syst. Evol. Microbiol.* **72**, 10–12 (2022).
9. Arahal, D. *et al.* The best of both worlds: a proposal for further integration of Candidatus names into the International Code of Nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* **74**, 1–21 (2024).

1399 10. Blanco-Míguez, A. *et al.* Extension of the Segatella copri complex to 13
1400 species with distinct large extrachromosomal elements and associations with
1401 host conditions. *Cell Host Microbe* **31**, 1804-1819.e9 (2023).

1402 11. De Filippis, F. *et al.* Distinct Genetic and Functional Traits of Human Intestinal
1403 Prevotella copri Strains Are Associated with Different Habitual Diets. *Cell Host*
1404 *Microbe* **25**, 444-453.e3 (2019).

1405 12. Fehlner-Peach, H. *et al.* Distinct Polysaccharide Utilization Profiles of Human
1406 Intestinal Prevotella copri Isolates. *Cell Host Microbe* **26**, 680-690.e5 (2019).

1407 13. Conteville, L. C. & Vicente, A. C. P. A plasmid network from the gut
1408 microbiome of semi-isolated human groups reveals unique and shared
1409 metabolic and virulence traits. *Sci. Rep.* **12**, 1–9 (2022).

1410 14. Fogarty, E. C. *et al.* Article A cryptic plasmid is among the most numerous
1411 genetic elements in the human gut ll ll Article A cryptic plasmid is among the
1412 most numerous genetic elements in the human gut. *Cell* **187**, 1206-1222.e16
1413 (2024).

1414 15. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their
1415 ecology across human gut metagenomes revealed by PlasX and MobMess.
1416 *Nat. Microbiol.* **9**, 830–847 (2024).

1417 16. Perez, M. *et al.* A synthetic consortium of 100 gut commensals modulates the
1418 composition and function in a colon model of the microbiome of elderly
1419 subjects. *Gut Microbes* **13**, 1–19 (2021).

1420 17. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and
1421 butyrate production. *Nat. Commun.* **12**, 1–16 (2021).

1422 18. Cheng, A. G. *et al.* Design, construction, and in vivo augmentation of a
1423 complex gut microbiome. *Cell* **185**, 3617-3636.e19 (2022).

1424 19. Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. Akkermansia
1425 municiphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium.
1426 *Int. J. Syst. Evol. Microbiol.* **54**, 1469–1476 (2004).

1427 20. Plovier, H. *et al.* A purified membrane protein from Akkermansia muciniphila or
1428 the pasteurized bacterium improves metabolism in obese and diabetic mice.
1429 *Nat. Med.* **23**, 107–113 (2017).

1430 21. Overmann, J. Significance and future role of microbial resource centers. *Syst.*
1431 *Appl. Microbiol.* **38**, 258–265 (2015).

1432 22. Parker, C. T., Tindall, B. J. & Garrity, G. M. International code of nomenclature
1433 of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* **69**, S1 (2019).

1434 23. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the
1435 human gut microbiome. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0603-
1436 3.

1437 24. Seshadri, R. *et al.* Expanding the genomic encyclopedia of Actinobacteria with
1438 824 isolate reference genomes. *Cell Genomics* **2**, (2022).

1439 25. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a
1440 new method for improved phylogenetic and taxonomic placement of microbes.
1441 *Nat. Commun.* **4**, 1–10 (2013).

1442 26. Jones, B. V, Sun, F. & Marchesi, J. R. Comparative metagenomic analysis of
1443 plasmid encoded functions in the human gut microbiome. *BMC Genomics* **11**,
1444 (2010).

1445 27. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the
1446 industrialized human microbiome. *Cell* **184**, 2053-2067.e18 (2021).

1447 28. Rozov, R. *et al.* Recycler: An algorithm for detecting plasmids from de novo
1448 assembly graphs. *Bioinformatics* **33**, 475–482 (2017).

1449 29. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its
1450 applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

1451 30. Pujol, C., Ehrlich, S. D. & Jannière, L. The promiscuous plasmids pIP501 and
1452 pAMβ1 from gram-positive bacteria encode complementary resolution
1453 functions. *Plasmid* vol. 31 100–105 (1994).

1454 31. Schmartz, G. P. *et al.* PLSDB: Advancing a comprehensive database of
1455 bacterial plasmids. *Nucleic Acids Res.* **50**, D273–D278 (2022).

1456 32. Li, Y. *et al.* Distribution of megaplasmids in Lactobacillus salivarius and other
1457 Lactobacilli. *J. Bacteriol.* **189**, 6128–6139 (2007).

1458 33. Bottacini, F. *et al.* Discovery of a conjugative megaplasmid in Bifidobacterium
1459 breve. *Appl. Environ. Microbiol.* **81**, 166–176 (2015).

1460 34. García-Bayona, L., Coyne, M. J. & Comstock, L. E. Mobile Type VI secretion
1461 system loci of the gut Bacteroidales display extensive intra-ecosystem transfer,
1462 multi-species spread and geographical clustering. *PLoS Genet.* **17**, 1–25
1463 (2021).

1464 35. Riepe, S. P., Goldstein, J. & Alpers, D. H. Effect of secreted Bacteroides
1465 proteases on human intestinal brush border hydrolases. *J. Clin. Invest.* **66**,
1466 314–322 (1980).

1467 36. Mills, R. H. *et al.* Multi-omics analyses of the ulcerative colitis gut microbiome
1468 link Bacteroides vulgatus proteases with disease severity. *Nat. Microbiol.* 1–15
1469 (2022) doi:10.1038/s41564-021-01050-3.

1470 37. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed
1471 by Over 150,000 Genomes from Metagenomes Spanning Age, Geography,
1472 and Lifestyle. *Cell* **176**, 649-662.e20 (2019).

1473 38. Cruz, P., Mehretu, A. M., Buttner, M. P., Trice, T. & Howard, K. M.
1474 Development of a polymerase chain reaction assay for the rapid detection of
1475 the oral pathogenic bacterium , Selenomonas noxia. *BMC Oral Health* 1–8
1476 (2015) doi:10.1186/s12903-015-0071-1.

1477 39. Fodor, A. A. *et al.* The 'most wanted' taxa from the human microbiome for
1478 whole genome sequencing. *PLoS One* **7**, (2012).

1479 40. Molinero, N. *et al.* Ruminococcoides bili gen. Nov., sp. nov., a bile-resistant
1480 bacterium from human bile with autolytic behavior. *Int. J. Syst. Evol. Microbiol.*
1481 **71**, 1–11 (2021).

1482 41. Guo, C. *et al.* Deficient butyrate-producing capacity in the gut microbiome is
1483 associated with bacterial network disturbances and fatigue symptoms in
1484 ME/CFS. *Cell Host Microbe* **31**, 288-304.e8 (2023).

1485 42. Schmitz, M. A., Dimonaco, N. J., Clavel, T. & Hitch, T. C. A. Lineage-specific
1486 microbial protein prediction enables large-scale exploration of protein ecology
1487 within the human gut. *bioRxiv* (2024) doi:10.1101/2024.05.29.596415.

1488 43. Ma, L. *et al.* Spermidine improves gut barrier integrity and gut microbiota
1489 function in diet-induced obese mice. *Gut Microbes* **12**, 1–19 (2020).

1490 44. Wiedmann, S., Eudy, J. D. & Zempleni, J. Biotin supplementation increases
1491 expression of genes encoding interferon-γ, interleukin-1β, and 3-
1492 methylcrotonyl-CoA carboxylase, and decreases expression of the gene
1493 encoding interleukin-4 in human peripheral blood mononuclear cells. *J. Nutr.*
1494 **133**, 716–719 (2003).

1495 45. Agrawal, S., Agrawal, A. & Said, H. M. Biotin deficiency enhances the
1496 inflammatory response of human dendritic cells. *Am. J. Physiol. - Cell Physiol.*
1497 **311**, C386–C391 (2016).

1498 46. Skupsky, J. *et al.* Biotin Supplementation Ameliorates Murine Colitis by
1499 Preventing NF-κB Activation. *Cmgh* **9**, 557–567 (2020).

1500 47. Mao, B. *et al.* Blautia producta displays potential probiotic properties against
1501 dextran sulfate sodium-induced colitis in mice. *Food Sci. Hum. Wellness* **13**,
1502 709–720 (2024).

1503 48. van der Lelie, D. *et al.* Rationally designed bacterial consortia to treat chronic
1504 immune-mediated colitis and restore intestinal homeostasis. *Nat. Commun.* **12**,
1505 1–17 (2021).

1506 49. Lee, C. *et al.* P926 Blautia Obeum Aggravates Colitis in a Murine Model. *J.*
1507 *Crohn's Colitis* **17**, i1035–i1035 (2023).

1508 50. Huang, T. Q. *et al.* Bergenin Alleviates Ulcerative Colitis By Decreasing Gut
1509 Commensal Bacteroides vulgatus-Mediated Elevated Branched-Chain Amino
1510 Acids. *J. Agric. Food Chem.* **72**, 3606–3621 (2024).

1511 51. Sugihara, K. *et al.* Dietary phosphate exacerbates intestinal inflammation in
1512 experimental colitis. *J. Clin. Biochem. Nutr.* **61**, 91–99 (2017).

1513 52. Kolachala, V. L. *et al.* A2B Adenosine Receptor Gene Deletion Attenuates
1514 Murine Colitis. *Gastroenterology* **135**, 861–870 (2008).

1515 53. Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal
1516 bacterium identified by gut microbiota analysis of Crohn disease patients.
1517 *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16731–16736 (2008).

1518 54. Ueda, A. *et al.* Identification of Faecalibacterium prausnitzii strains for gut
1519 microbiome-based intervention in Alzheimer's-type dementia. *Cell Reports*
1520 *Med.* **2**, 100398 (2021).

1521 55. D'hoe, K. *et al.* Integrated culturing, modeling and transcriptomics uncovers
1522 complex interactions and emergent behavior in a three-species synthetic gut
1523 community. *Elife* **7**, 1–30 (2018).

1524 56. Fitzgerald, C. B. *et al.* Comparative analysis of Faecalibacterium prausnitzii
1525 genomes shows a high level of genome plasticity and warrants separation into
1526 new species-level taxa. *BMC Genomics* **19**, 1–20 (2018).

1527 57. Vital, M., Karch, A. & Pieper, D. H. Colonic Butyrate-Producing Communities in
1528 Humans : an Overview Using Omics Data. *mSystems* **2**, 1–18 (2017).

1529 58. Afrizal, A. *et al.* Enhanced cultured diversity of the mouse gut microbiota
1530 enables custom-made synthetic communities. *Cell Host Microbe* 1–16 (2022)
1531 doi:10.1016/j.chom.2022.09.011.

1532 59. Wylensek, D. *et al.* A collection of bacterial isolates from the pig intestine
1533 reveals functional and taxonomic diversity. *Nat. Commun.* **in press**, 1–26
1534 (2020).

1535 60. Stackebrandt, E. *et al.* Deposit of microbial strains in public service collections
1536 as part of the publication process to underpin good practice in science.
1537 *Springerplus* **3**, 1–4 (2014).

1538 61. Burz, S. D. *et al.* A Guide for Ex Vivo Handling and Storage of Stool Samples
1539 Intended for Fecal Microbiota Transplantation. *Sci. Rep.* **9**, 1–16 (2019).

1540 62. Afrizal, A. *et al.* Anaerobic single-cell dispensing facilitates the cultivation of
1541 human gut bacteria. *Environ. Microbiol.* **00**, (2022).

1542 63. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular
1543 microbial diversity of an anaerobic digestor as determined by small-subunit
1544 rDNA sequence analysis. *Appl. Environ. Microbiol.* **63**, 2802–2813 (1997).

1545 64. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for
1546 Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

1547 65. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. in *9th Annual*
1548 *Genomics of Energy & Environment Meeting* (2014).

1549 66. Antipov, D. *et al.* PlasmidSPAdes: Assembling plasmids from whole genome
1550 sequencing data. *Bioinformatics* **32**, 3380–3387 (2016).

1551 67. Yilmaz, P. *et al.* Minimum information about a marker gene sequence
1552 (MIMARKS) and minimum information about any (x) sequence (MIxS)
1553 specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).

1554 68. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W.
1555 CheckM: assessing the quality of microbial genomes recovered from. *Cold*
1556 *Spring Harb. Lab. Press Method* **1**, 1–31 (2015).

1557 69. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile
1558 genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150
1559 (2018).

1560 70. Bengtsson-Palme, J. *et al.* metaxa2: Improved identification and taxonomic
1561 classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol.*
1562 *Resour.* **15**, 1403–1414 (2015).

1563 71. Schwengers, O. *et al.* Bakta: Rapid and standardized annotation of bacterial
1564 genomes via alignment-free sequence identification. *Microb. Genomics* **7**,
1565 (2021).

1566 72. Mölder, F. *et al.* Sustainable data analysis with Snakemake [version 2; peer
1567 review: 2 approved]. *F1000Research* **10**, 1–29 (2021).

1568 73. Hitch, T. C. A. *et al.* Automated analysis of genomic sequences facilitates high-
1569 throughput and comprehensive description of bacteria. *ISME Commun.* **1**,
1570 (2021).

1571 74. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and

genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 1–10 (2020).

75. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

76. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

77. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

78. Beverley, S. M. Estimation of circular DNA size using γ-irradiation and pulsed-field gel electrophoresis. *Anal. Biochem.* **177**, 110–114 (1989).

79. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**, 635–645 (2014).

80. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

81. Qin, Q. L. *et al.* A proposed genus boundary for the prokaryotes based on genomic insights. *J. Bacteriol.* **196**, 2210–2215 (2014).

82. Meile, L., Rohr, L. M., Geissmann, T. A., Herensperger, M. & Teuber, M. Characterization of the D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase gene (xfp) from Bifidobacterium lactis. *J. Bacteriol.* **183**, 2929–2936 (2001).

83. Mishra, A. K., Hugon, P., Robert, C., Raoult, D. & Fournier, P. E. Non contiguous-finished genome sequence and description of Peptoniphilus grossensis sp. nov. *Stand. Genomic Sci.* **7**, 320–330 (2012).

84. Dobell, C. C. The Structure and Life-History of Copromonas subtilis, nov. gen. et nov. spec.: a Contribution to our Knowledge of the Flagellata. *J. Cell Sci.* **S2-52**, 75–120 (1908).

1602 **Supplementary Tables**

1603

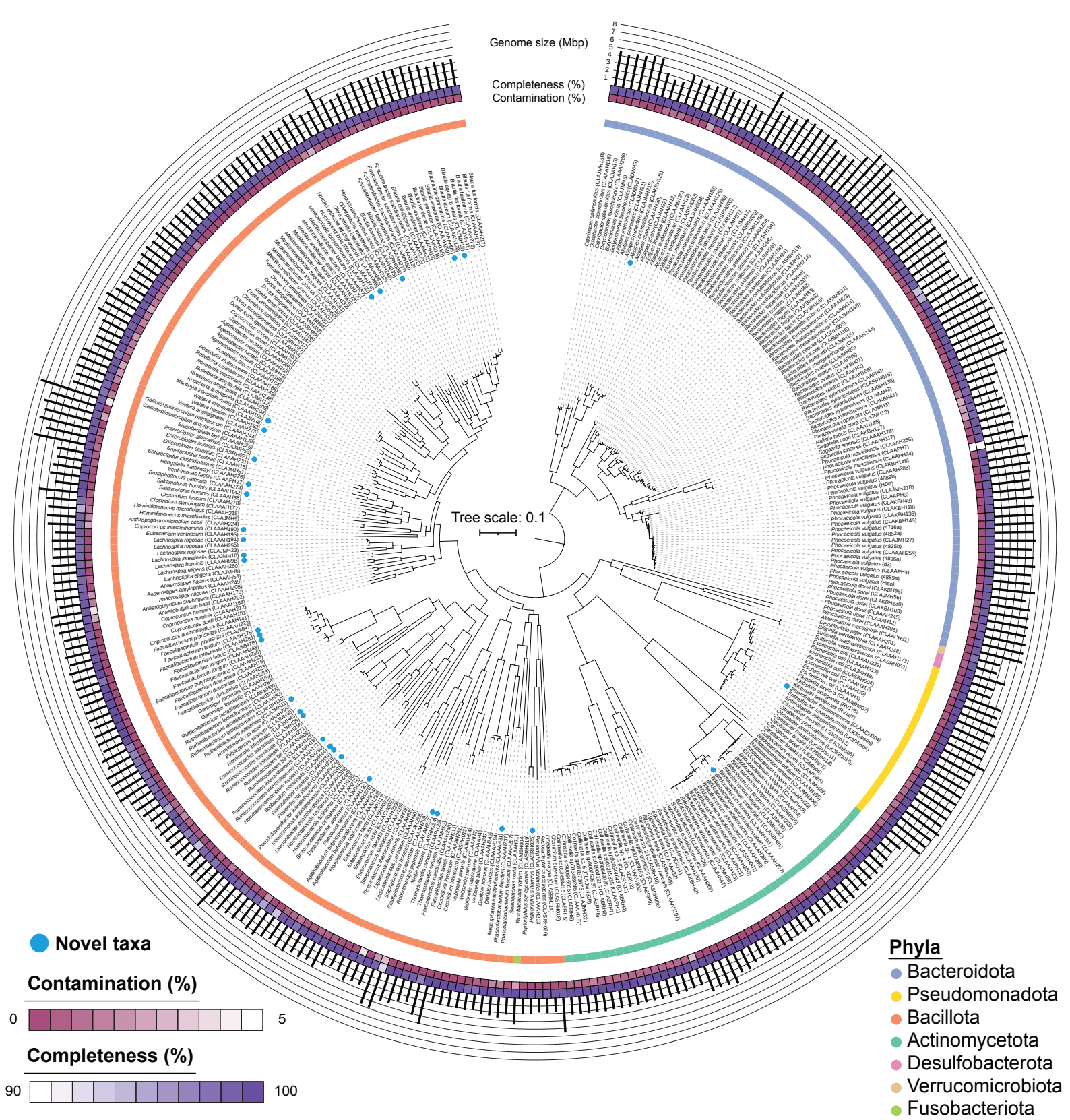
1604 **Supplementary Table 1:** Statistical comparison of genome sizes of phyla present
1605 within the HiBC, UHGG (isolates), and complete UHGG.

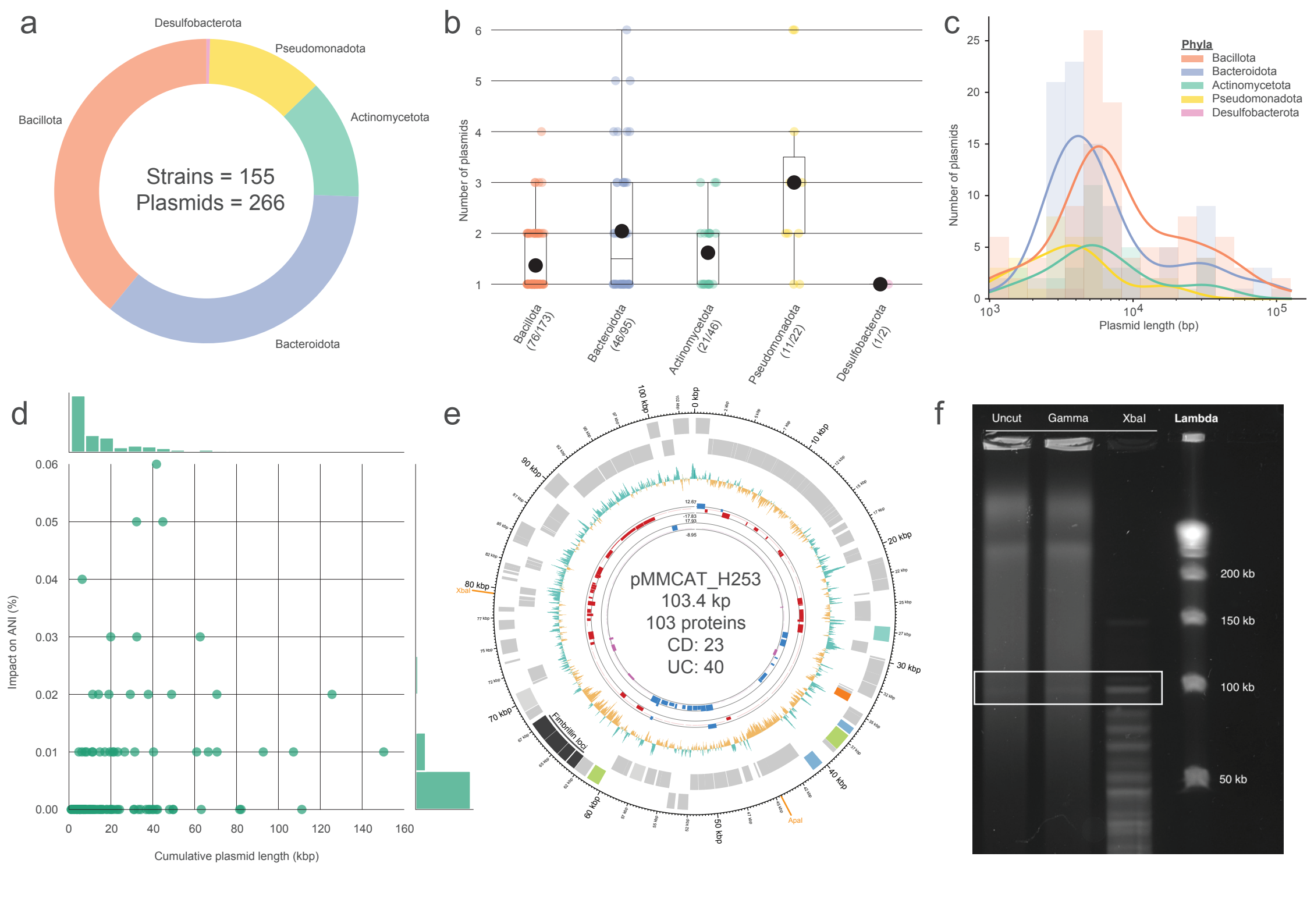
1606

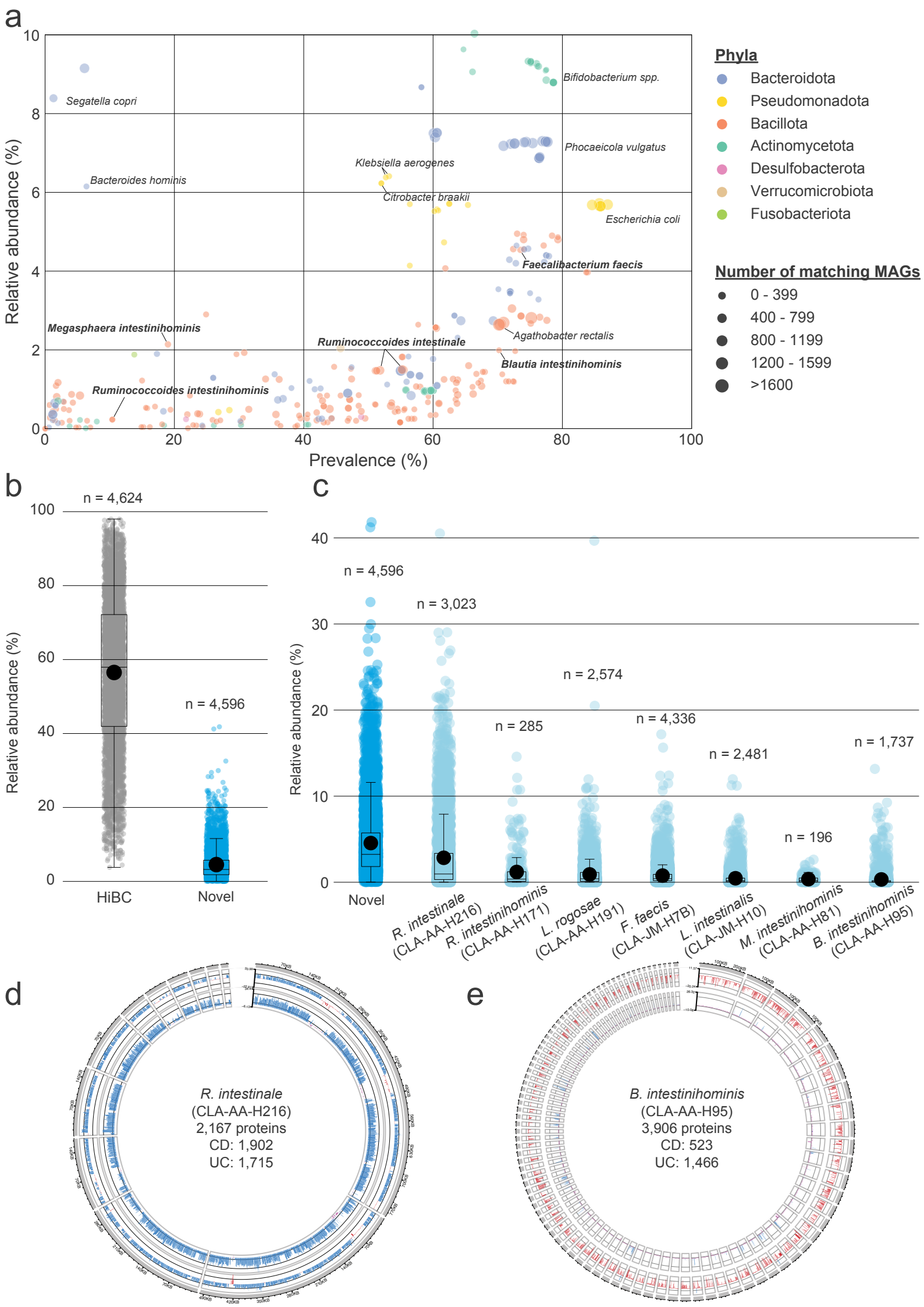
1607 **Supplementary Table 2:** Statistical comparison of the normalised carbohydrate
1608 active enzyme investment of phyla within the HiBC.

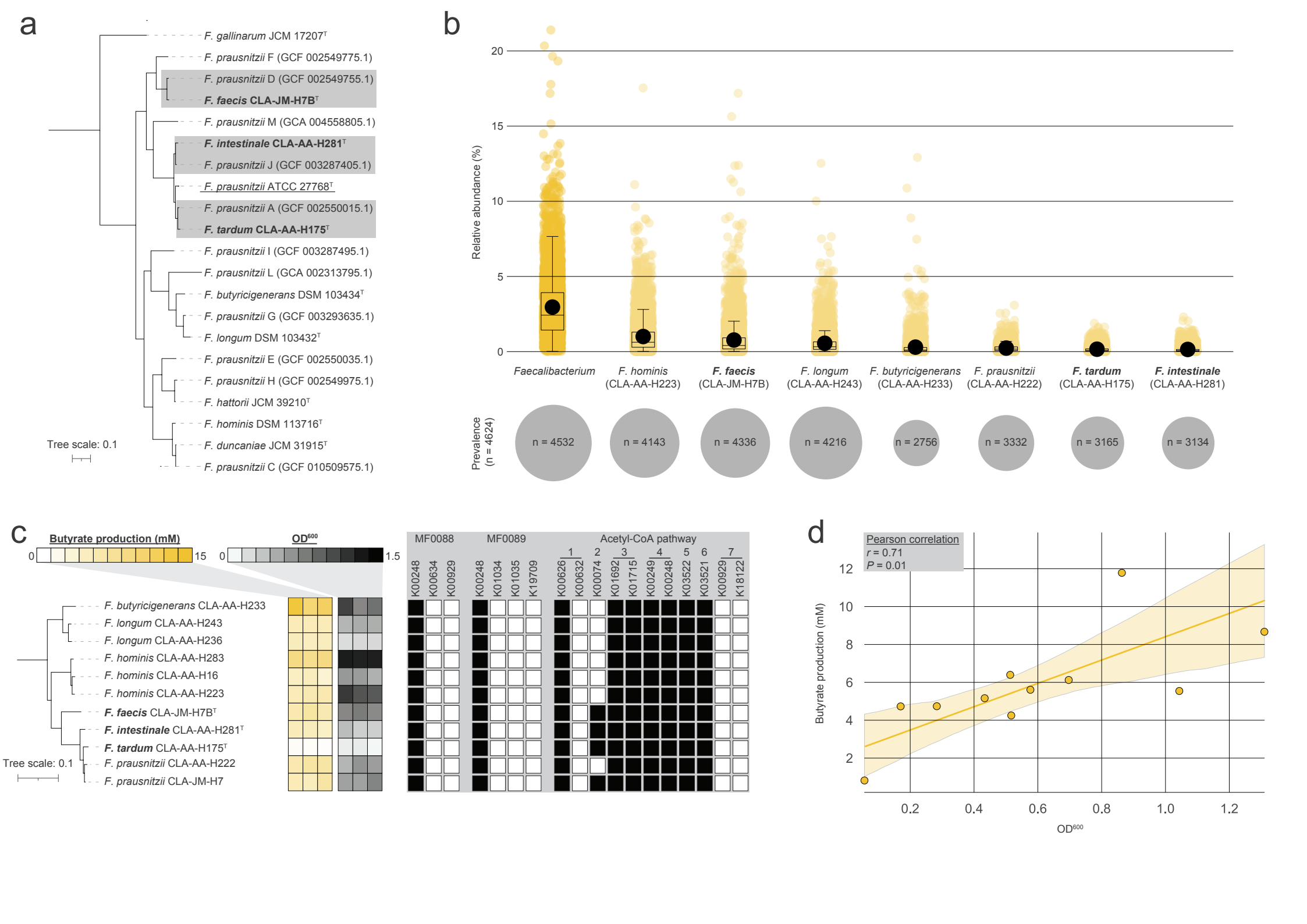
1609

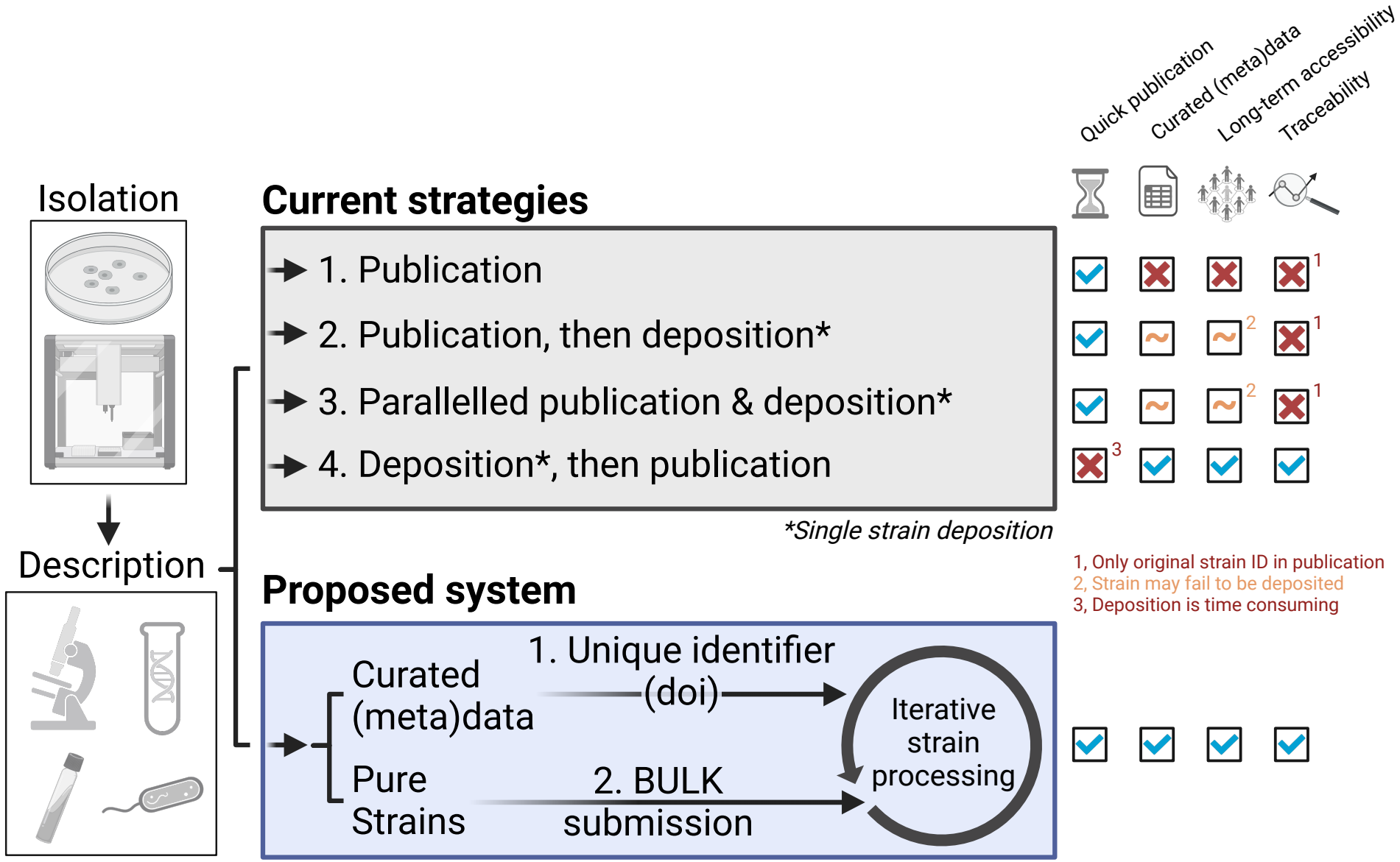
1610 **Supplementary Table 3:** HPLC measurement of the production and/or utilisation of
1611 multiple metabolites.

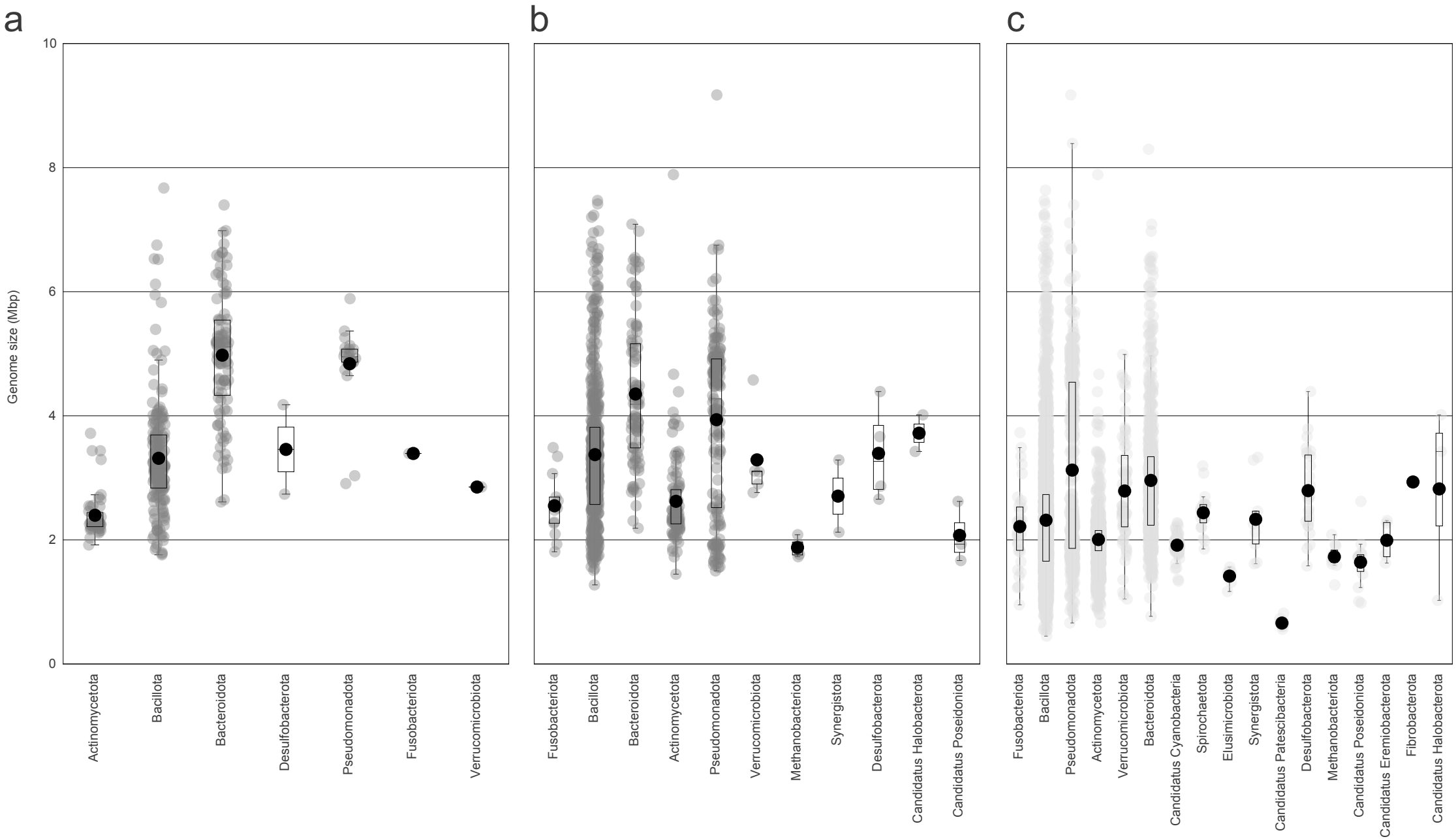

Genome size (Mbp)
Completeness (%)
Contamination (%)
Tree scale: 0.1
Novel taxa
Contamination (%)
0
5
Completeness (%)
90
100
Phyla
Bacteroidota
Pseudomonadota
Bacillota
Actinomycetota
Desulfobacterota
Verrucomicrobiota
Fusobacteriota


a
Desulfobacterota
Pseudomonadota
Actinomycetota
Bacillota
Bacteroidota
Strains = 155
Plasmids = 266
b
Number of plasmids
Bacillota (76/173)
Bacteroidota (46/95)
Actinomycetota (21/46)
Pseudomonadota (11/22)
Desulfobacterota (1/2)
c
Phyla
Bacillota
Bacteroidota
Actinomycetota
Pseudomonadota
Desulfobacterota
Number of plasmids
Plasmid length (bp)
d
Impact on ANI (%)
Cumulative plasmid length (kbp)
e
pMMCAT_H253
103.4 kp
103 proteins
CD: 23
UC: 40
Fimbrilin loci
XbaI
ApaI
f
Uncut
Gamma
XbaI
Lambda
200 kb
150 kb
100 kb
50 kb

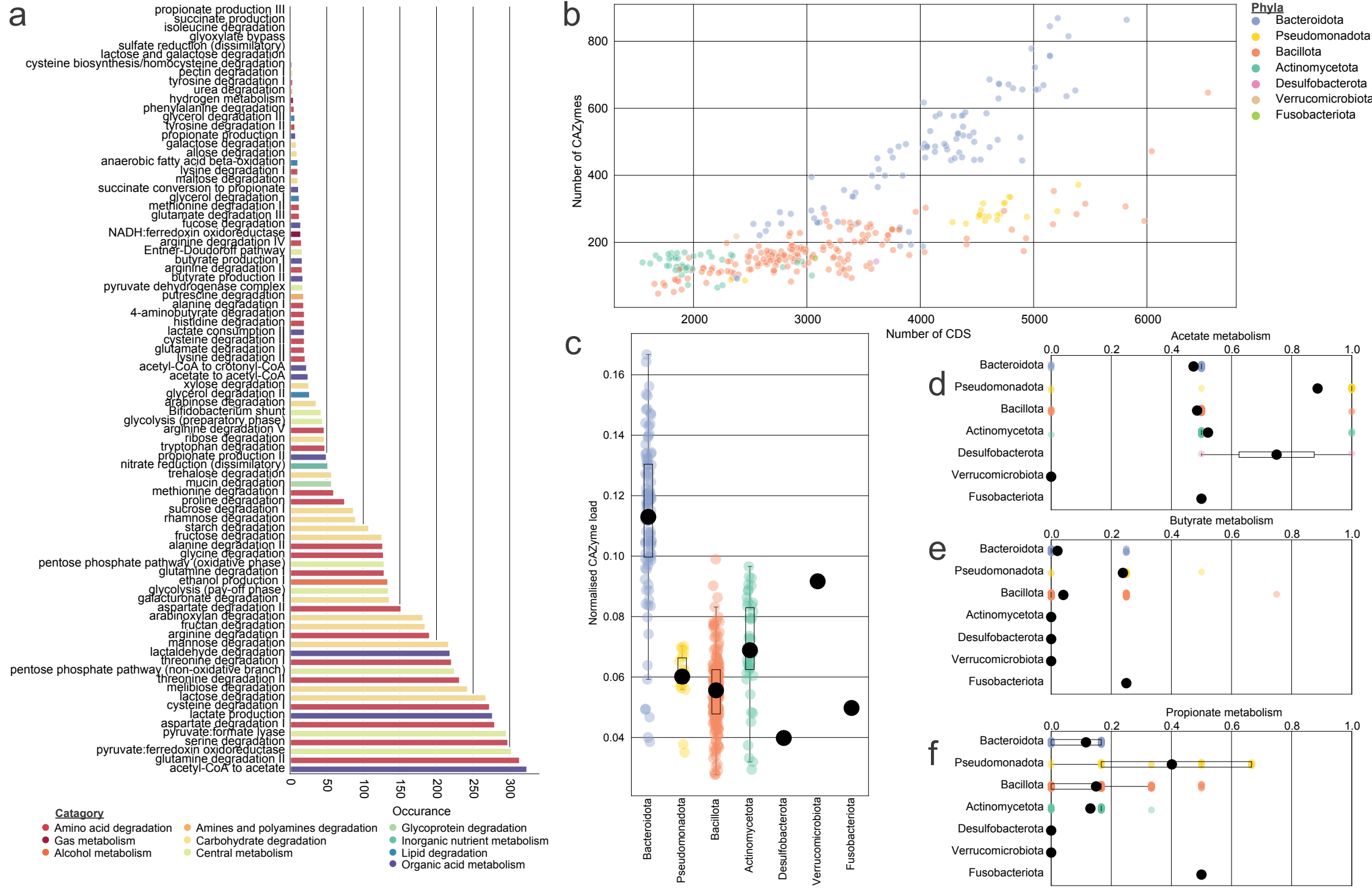

a
Relative abundance (%)
Prevalence (%)
Segatella copri
Bacteroides hominis
Bifidobacterium spp.
Phocaeicola vulgatus
Klebsiella aerogenes
Citrobacter braakii
Escherichia coli
Faecalibacterium faecis
Megasphaera intestinihominis
Ruminococcoides intestinale
Agathobacter rectalis
Blautia intestinihominis
Ruminococcoides intestinihominis
Phyla
Bacteroidota
Pseudomonadota
Bacillota
Actinomycetota
Desulfobacterota
Verrucomicrobiota
Fusobacteriota
Number of matching MAGs
0 - 399
400 - 799
800 - 1199
1200 - 1599
>1600
b
n = 4,624
n = 4,596
Relative abundance (%)
HiBC
Novel
c
Relative abundance (%)
n = 4,596
n = 3,023
n = 285
n = 2,574
n = 4,336
n = 2,481
n = 196
n = 1,737
Novel
R. intestinale (CLA-AA-H216)
R. intestinihominis (CLA-AA-H171)
L. rogosae (CLA-AA-H191)
F. faecis (CLA-JM-H7B)
L. intestinalis (CLA-JM-H10)
M. intestinihominis (CLA-AA-H81)
B. intestinihominis (CLA-AA-H95)
d
R. intestinale
(CLA-AA-H216)
2,167 proteins
CD: 1,902
UC: 1,715
e
B. intestinihominis
(CLA-AA-H95)
3,906 proteins
CD: 523
UC: 1,466


a
F. gallinarum JCM 17207T
F. prausnitzii F (GCF 002549775.1)
F. prausnitzii D (GCF 002549755.1)
F. faecis CLA-JM-H7BT
F. prausnitzii M (GCA 004558805.1)
F. intestinale CLA-AA-H281T
F. prausnitzii J (GCF 003287405.1)
F. prausnitzii ATCC 27768T
F. prausnitzii A (GCF 002550015.1)
F. tardum CLA-AA-H175T
F. prausnitzii I (GCF 003287495.1)
F. prausnitzii L (GCA 002313795.1)
F. butyricigenerans DSM 103434T
F. prausnitzii G (GCF 003293635.1)
F. longum DSM 103432T
F. prausnitzii E (GCF 002550035.1)
F. prausnitzii H (GCF 002549975.1)
F. hattorii JCM 39210T
F. hominis DSM 113716T
F. duncaniae JCM 31915T
F. prausnitzii C (GCF 010509575.1)
Tree scale: 0.1
b
Relative abundance (%)
Faecalibacterium
F. hominis (CLA-AA-H223)
F. faecis (CLA-JM-H7B)
F. longum (CLA-AA-H243)
F. butyricigenerans (CLA-AA-H233)
F. prausnitzii (CLA-AA-H222)
F. tardum (CLA-AA-H175)
F. intestinale (CLA-AA-H281)
Prevalence (n = 4624)
n = 4532
n = 4143
n = 4336
n = 4216
n = 2756
n = 3332
n = 3165
n = 3134
c
Butyrate production (mM)
OD600
MF0088
MF0089
Acetyl-CoA pathway
K00248
K00634
K00929
K01034
K01035
K19709
K00626
K00632
K00074
K01692
K01715
K00249
K03522
K03521
K18122
F. butyricigenerans CLA-AA-H233
F. longum CLA-AA-H243
F. longum CLA-AA-H236
F. hominis CLA-AA-H283
F. hominis CLA-AA-H16
F. hominis CLA-AA-H223
F. faecis CLA-JM-H7BT
F. intestinale CLA-AA-H281T
F. tardum CLA-AA-H175T
F. prausnitzii CLA-AA-H222
F. prausnitzii CLA-JM-H7
Tree scale: 0.1
d
Pearson correlation
r = 0.71
P = 0.01
Butyrate production (mM)
OD600

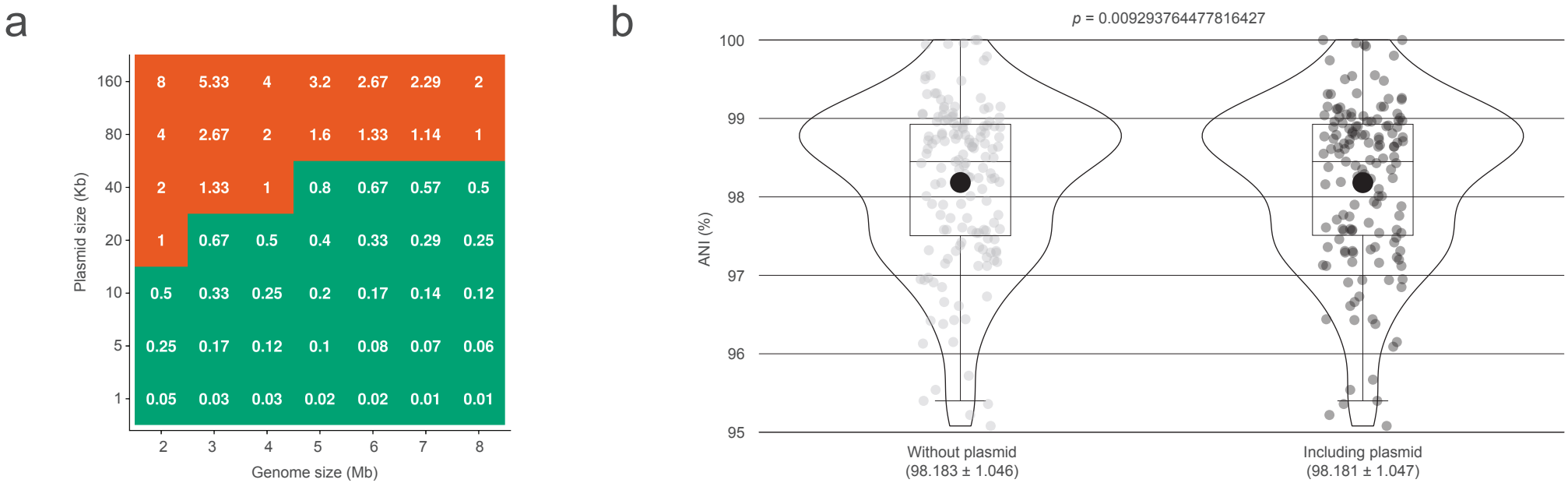

Quick publication
Curated (meta)data
Long-term accessibility
Traceability
Isolation
Description
Current strategies
1. Publication
2. Publication, then deposition*
3. Parallelled publication & deposition*
4. Deposition*, then publication
*Single strain deposition
1, Only original strain ID in publication
2, Strain may fail to be deposited
3, Deposition is time consuming
Proposed system
Curated (meta)data
Pure Strains
1. Unique identifier (doi)
2. BULK submission
Iterative strain processing



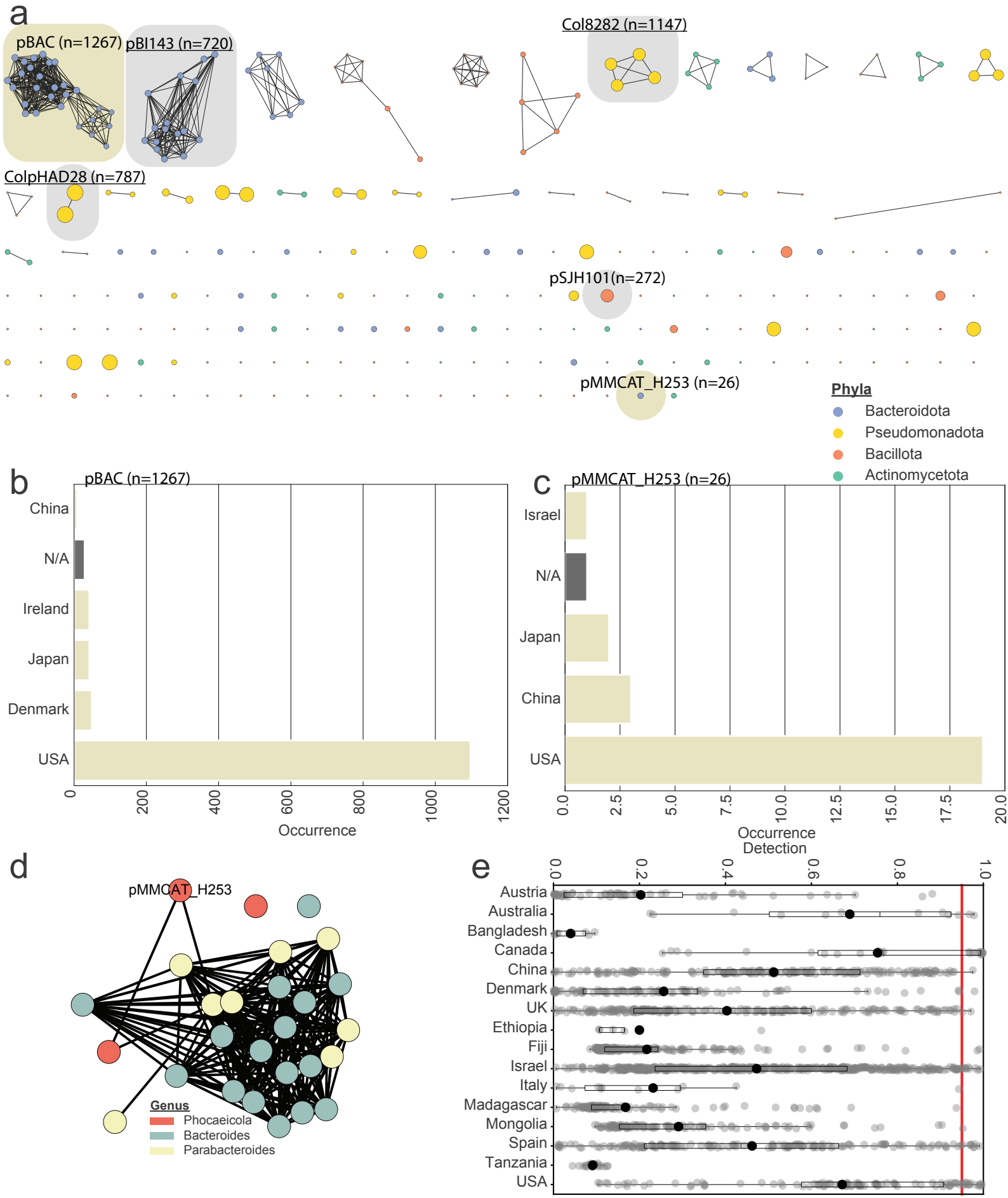
Supplementary Figure 1. a. The genome size of each HiBC strain (n = 340) was calculated in mega base pairs (Mbp). **b.** Genome size of all isolates (n = 891) within the UHGG collection. **c.** The genome size of all genomes (isolates and MAGs) within the UHGG collection (Almeida *et al*, 2021) (n = 4614). Black dots represent the mean value for each group. Statistical comparison of all groups is provided in Supplementary Table 1.



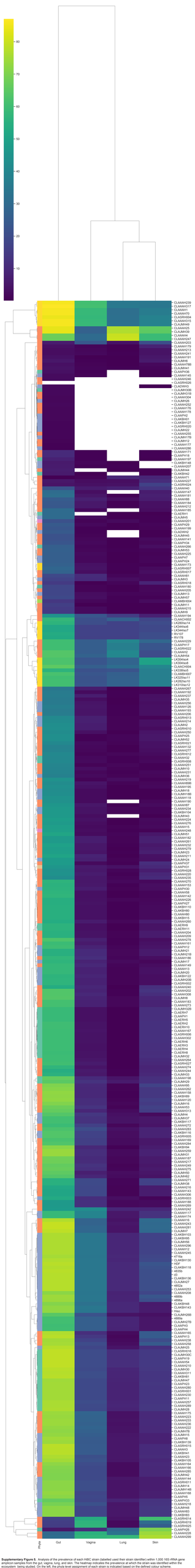
Supplementary Figure 2. a. Frequency of each gut metabolic modules detection across the 340 HiBC strains. Modules are coloured according to their broad functional assignments. **b.** Comparison of each strains CAZyme repertoire compared to its coding potential (total number of coding sequences), coloured based on phyla. **c.** The CAZyme repertoire of each strain normalised to their total coding potential, grouped by phyla. Black dots represent the mean value for each group. **d-f.** The functional investment of strains belonging to each phyla into the production of acetate, butyrate, and propionate.



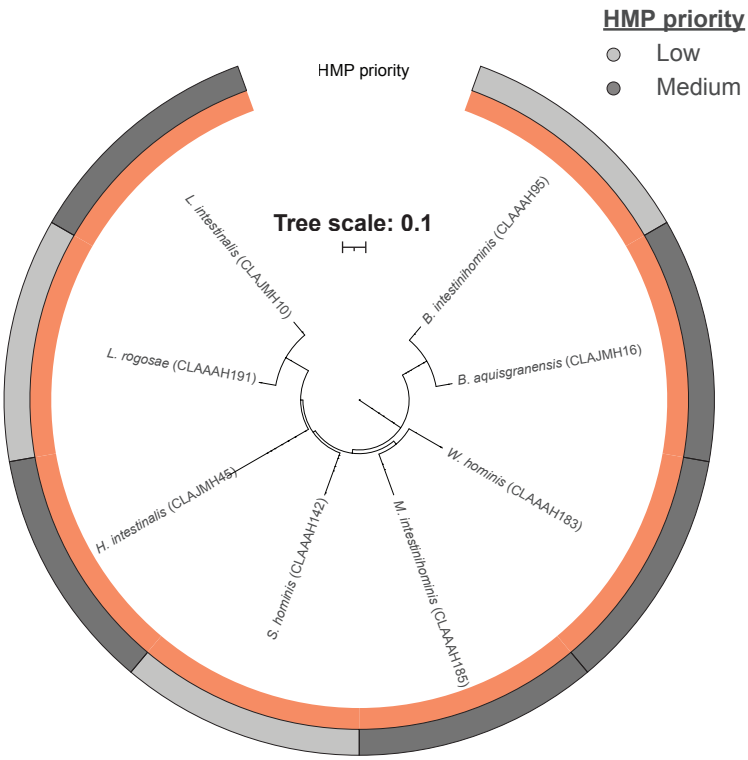
Supplementary Figure 3. a. The impact plasmids of specific sizes should theortically have on ANI comparison between two genomes of a set size was calculated and is stated in this matrix. In green are values less than 1%, when a plasmid has a >1% impact on ANI comparison, it is shown in orange. **b.** Comparison of each HiBC strain to its closest assignment, via GTDB-Tk, was calculated with and without plasmids (strains = 152). The presence of plasmids within the genomes significantly reduced the ANI score to the closest relatives, although the impact on ANI was 0.004 ± 0.01 and never altered the taxonomic assignment of a strain.



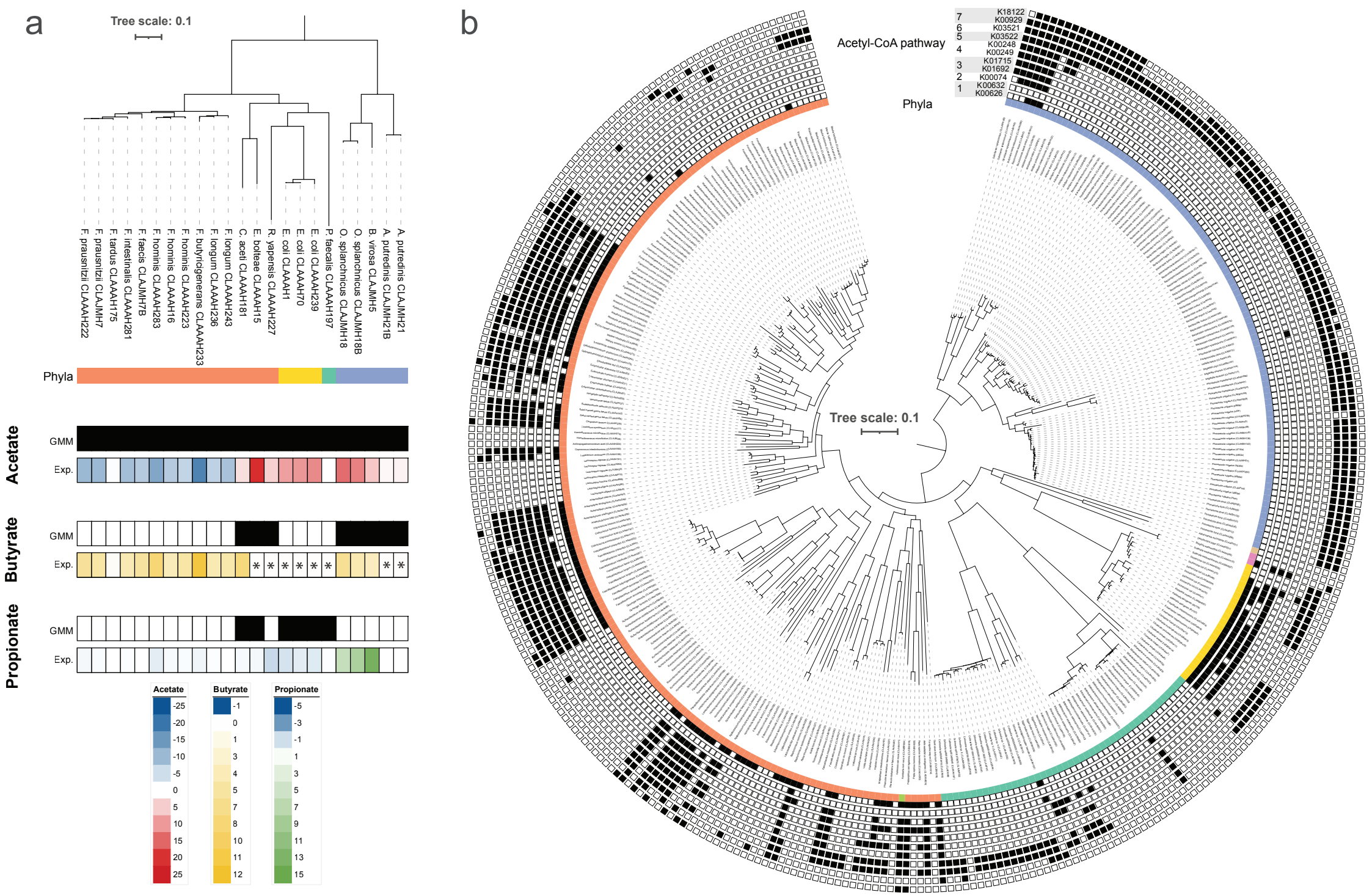
Supplementary Figure X. a. Network visualisation of plasmid sequence similarity. Plasmids (circles) with similar sequences are connected with edges, the distance of which is proportional to their similarity (shorter = similar; longer = dissimilar). Plasmids are coloured based on their phyla and their size is proportional to the number of matching plasmids within PLSDB. Plasmids / Clusters of interest are named (underlined for known) and highlighted; grey for known, brown for novel. **b.** Country occurrence of the pBAC plasmid. **c.** Country occurrence of the pMMCAT_H253 plasmid. **d.** Mobmess network comparison of sequence similarity between pMMCAT_H253 matches within PLSDB. Each node represents a plasmid sequence, coloured based on the taxonomy of their originating isolate. The length of edges between nodes is based on the pairwise ANI comparison of plasmids. **e.** Detection of pMMCAT_H253 across 1,782 metagenomes, grouped based on their country of origin. Detection is defined as the proportion of nucletodies within the pMMCAT reference sequence covered by atleast one read within the studied sample.



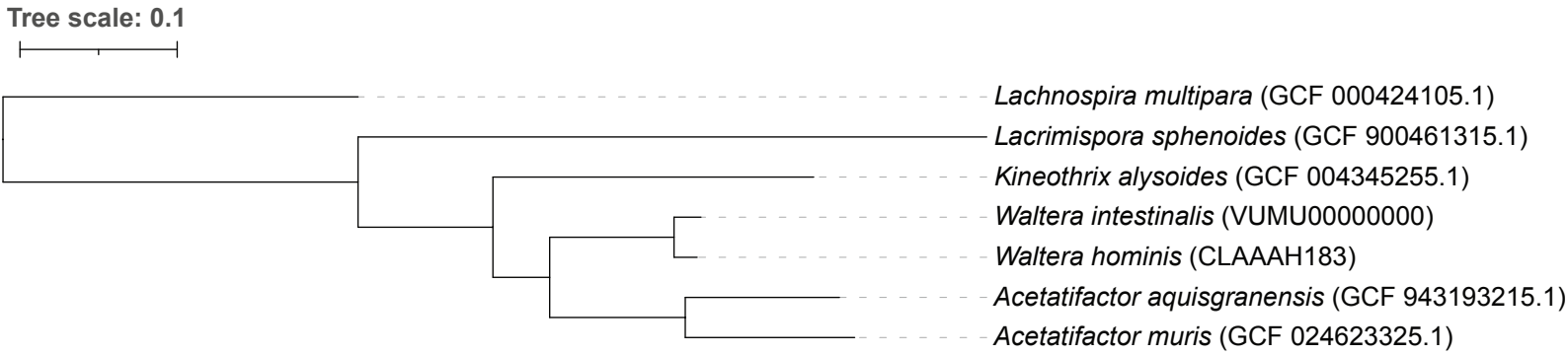
Supplementary Figure 5. Analysis of the prevalence of each HBC strain (labelled used their strain identifier) within 1,000 16S rRNA gene amplicon samples from the gut, vagina, lung, and skin. The heatmap indicates the prevalence at which the strain was identified within the population being studied. On the left, the phyla level assignment of each strain is indicated based on the defined colour scheme.



Supplementary Figure 6. Phylogenomic diversity of the eight novel species identified to be on the Human Microbiome Projects (HMP) most wanted listed. The inner ring is coloured according to the phyla assignment of the species, which are all *Bacillota*. The outer ring indicates the priority of each species in the most wanted list.



Supplementary Figure 7. a. Consistency between predicted short chain fatty acid production and experimental results were tested by studying 23 strains. Black boxes indicate the presence of a GMM for production of the SCFA of interest. Boxes marked with an asterik within the experimental data (Exp.) indicates values below the detection limit, or with high noise to signal ratio. **b.** Annotation of each HiBC strain for the KOs within the acetyl-CoA pathway for butyrate production.



Supplementary Figure 8. Phylogenomic tree of both validly named *Acetatifactor*, and the only currently validly named *Waltera* species, with the proposed *Waltera hominis*. The type species for both *Kineothrix* and *Lacrimispora* are also included to provide greater context. The tree is rooted on the outgroup of *Lachnospira multipara*.