

1 Broad diversity of human gut bacteria with traceable strain identifiers to facilitate bulk  
2 deposition

3

4 Thomas C. A. Hitch<sup>1</sup>, Johannes M. Masson<sup>1,#</sup>, Charlie Pauvert<sup>1,#</sup>, Johanna Bosch<sup>1</sup>,  
5 Selina Nüchtern<sup>1</sup>, Nicole Treichel<sup>1</sup>, Marko Baloh<sup>1</sup>, Soheila Razavi<sup>1</sup>, Afrizal Afrizal<sup>1</sup>,  
6 Ntana Kousetzi<sup>1</sup>, Andrea M. Aguirre<sup>1</sup>, David Wylensek<sup>1</sup>, Amy Coates<sup>1</sup>, Susan A. V.  
7 Jennings<sup>1</sup>, Atscharah Panyot<sup>1</sup>, Alina Viehof<sup>1</sup>, Matthias A. Schmitz<sup>1</sup>, Maximilian  
8 Stuhmann<sup>1</sup>, Evelyn C. Deis<sup>1</sup>, Kevin Bisdorf<sup>1</sup>, Maria D. Chiotelli<sup>2</sup>, Artur Lissin<sup>3</sup>, Isabel  
9 Schober<sup>3</sup>, Julius Witte<sup>3</sup>, Thorsten Cramer<sup>4</sup>, Thomas Riedel<sup>3,5</sup>, Marie Wende<sup>6</sup>, Katrin  
10 A. Winter<sup>6</sup>, Lena Amend<sup>6</sup>, Alessandra Riva<sup>7,8</sup>, Stefanie Trinh<sup>9</sup>, Laura Mitchell<sup>10</sup>,  
11 Jonathan Hartman<sup>11</sup>, David Berry<sup>7</sup>, Jochen Seitz<sup>12</sup>, Lukas C. Bossert<sup>11</sup>, Marianne  
12 Grognot<sup>2</sup>, Thorsten Allers<sup>10</sup>, Till Strowig<sup>5,6,13</sup>, Michael Pester<sup>3,14</sup>, Birte Abt<sup>3,5</sup>, Lorenz  
13 C. Reimer<sup>3</sup>, Jörg Overmann<sup>3,5,14</sup>, Thomas Clavel<sup>1\*</sup>

14

15 <sup>1</sup> Functional Microbiome Research Group, Institute of Medical Microbiology, University  
16 Hospital of RWTH Aachen, Germany

17 <sup>2</sup> Biophysics of Host-Microbe Interactions Research Group, Institute of Medical Microbiology,  
18 University Hospital of RWTH Aachen, Germany

19 <sup>3</sup> Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures,  
20 Braunschweig, Germany

21 <sup>4</sup> Molecular Tumor Biology Research Group, Department of General, Visceral, Children and  
22 Transplantation Surgery, University Hospital of RWTH Aachen, Germany

23 <sup>5</sup> German Centre for Infection Research (DZIF), Partner Site Hannover-Braunschweig,  
24 Braunschweig, Germany

25 <sup>6</sup> Department of Microbial Immune Regulation, Helmholtz Centre for Infection Research,  
26 Braunschweig, Germany

27 <sup>7</sup> University of Vienna, Center for Microbiology and Environmental Systems Science,  
28 Department of Microbiology and Ecosystem Science, Vienna, Austria

29 <sup>8</sup> Chair of Nutrition and Immunology, School of Life Sciences, Technical University Munich,  
30 Freising-Weihenstephan, Germany

31 <sup>9</sup> Institute of Neuroanatomy, University Hospital of RWTH Aachen, Germany

32 <sup>10</sup> School of Life Sciences, University of Nottingham, Nottingham, United Kingdom

33 <sup>11</sup> IT centre, RWTH Aachen, Germany

34 <sup>12</sup> Clinic for Child and Adolescent Psychiatry, Psychosomatic Medicine and Psychotherapy,  
35 LVR-University Hospital Essen, University of Duisburg-Essen, Germany

36 <sup>13</sup> Centre for Individualised Infection Medicine (CiiM), a joint venture between the Helmholtz-  
37 Centre for Infection Research (HZI) and the Hannover Medical School (MHH), Hannover,  
38 Germany

39 <sup>14</sup> Technical University Braunschweig, Braunschweig, Germany

40 # Authors contributed equally to the manuscript

41 \* Correspondence: [tclavel@ukaachen.de](mailto:tclavel@ukaachen.de)

42

## 43 **Abstract**

44 Numerous bacteria in the human gut microbiome remain unknown and/or have yet to  
45 be cultured. While collections of human gut bacteria have been published, few  
46 strains have been made publicly available. A major hurdle in making strains publicly  
47 available is their deposition to public culture collections. We propose a framework for  
48 the bulk-deposition of strains to culture collections, which removes many of the  
49 barriers previously identified ([www.dsmz.de/bulk-deposit](http://www.dsmz.de/bulk-deposit)). Using this bulk-deposition  
50 system we have created a publicly available collection of human gut isolates. The  
51 Human intestinal Bacteria Collection (HiBC) ([www.hibc.rwth-aachen.de](http://www.hibc.rwth-aachen.de)) contains  
52 340 strains representing 198 species within 29 families and 7 phyla, of which 29  
53 previously unknown species are taxonomically described and named. These  
54 included two butyrate-producing species of *Faecalibacterium* and new dominant  
55 species associated with health and inflammatory bowel disease, *Ruminococcoides*  
56 *intestinale* and *Blautia intestinihominis*, respectively. Plasmids were prolific within the  
57 HiBC isolates, with almost half (46%) of strains containing plasmids, with a maximum  
58 of six within a strain. This included a broadly occurring plasmid (pBAC) that exists in  
59 three diverse forms across *Bacteroidales* species. Megaplasmids were identified  
60 within two strains, the pMMCAT megaplasmid is globally present within multiple  
61 *Bacteroidales* species. This collection of easily searchable and publicly available gut  
62 bacterial isolates will facilitate functional studies of the gut microbiome.

63

## 64 **Introduction**

65 The cultivation of human gut bacteria has accelerated in the last years<sup>1-4</sup>, providing  
66 valuable information on the presence of novel taxa within gut microbiomes. Yet while  
67 these published collections of human gut isolates note the novel taxa within their  
68 collections<sup>1,5</sup>, rarely is this novelty made known by describing and validly naming the  
69 taxa<sup>3</sup>. Curating the taxonomic assignment of such collections is essential as often  
70 outdated names are used, or recently described taxa are ignored, causing greater  
71 confusion in the current taxonomic sphere<sup>6-9</sup>. Such curated taxonomy ensures  
72 strains are correctly assigned, allowing strain-level diversity to be studied<sup>10</sup>.  
73 Variation between strains of the same species can lead to functional shifts that alter  
74 the association with host dietary habits<sup>11,12</sup>. One-way strains can vary is due to the  
75 presence of mobile genetic elements, which are known to affect the phenotype of the  
76 isolates in which they occur<sup>13</sup>. The study of plasmids within the human gut has been  
77 limited until recently<sup>14,15</sup>.

78 Accessibility to collections of strains from the human gut is essential for the  
79 mechanistic study of microbe-microbe and microbe-host interactions<sup>16-18</sup>. However,  
80 accessibility remains problematic, with few strains being deposited in public culture

81 collections and even fewer of the proposed novel taxa being validated. For most of  
82 the strains deposited, genomes are available, but their quality is not always high and  
83 metadata such as the source of isolation, cultivation requirements, and full taxonomy  
84 are often lacking. The small number of deposited strains limits future confirmatory or  
85 comparative studies, while the lack of curated metadata reduces the value of the  
86 genomic information. Only the professional acquisition of strains by public collections  
87 can ensure that the high-quality standards required for future work are maintained  
88 and that valuable strain-associated information is gathered and made freely  
89 accessible to researchers and users worldwide. This is particularly relevant for  
90 microbiota research to move beyond associations and towards a more mechanistic  
91 understanding, as exemplified by the study of *Akkermansia muciniphila*<sup>19,20</sup>.

92 The deposit of strains in public collections is a time-demanding and expensive  
93 process, yet necessary to ensure access to pure strains. Once deposited, the  
94 preservation of strains becomes a highly cost-effective means of maintaining the  
95 value of microbial strains over long periods of time<sup>21</sup>. The complete acquisition of  
96 single bacterial strains at culture collections, which includes purity and identity  
97 confirmation, requires several months of time and for fastidious anaerobic strains  
98 such as those isolated from the human intestinal microbiome, is currently  
99 accomplished for only about 20 strains in parallel by one laboratory technician. This  
100 has become a major bottleneck in the creation of publicly accessible collections  
101 emanating from current microbiome research that generates ever increasing  
102 numbers of novel isolates over comparatively short time intervals. A solution that  
103 enables rapid deposition is needed to enhance the time available to process strains.

104 To provide functional insights into the human gut microbiome based on isolates and  
105 facilitate public access to strains for the scientific community, we sought to create a  
106 bacterial collection and implement unique strain identifiers to facilitate bulk  
107 deposition at an international collection. In this way, all strains are easily traceable  
108 and made publicly available more rapidly. Studying this collection of isolates  
109 provided insight into the prevalence and diversity of plasmids within the human gut  
110 and the presence of megaplasmids. We also highlight health-associated variation  
111 within key genera, including novel species which require further *in vivo* study.

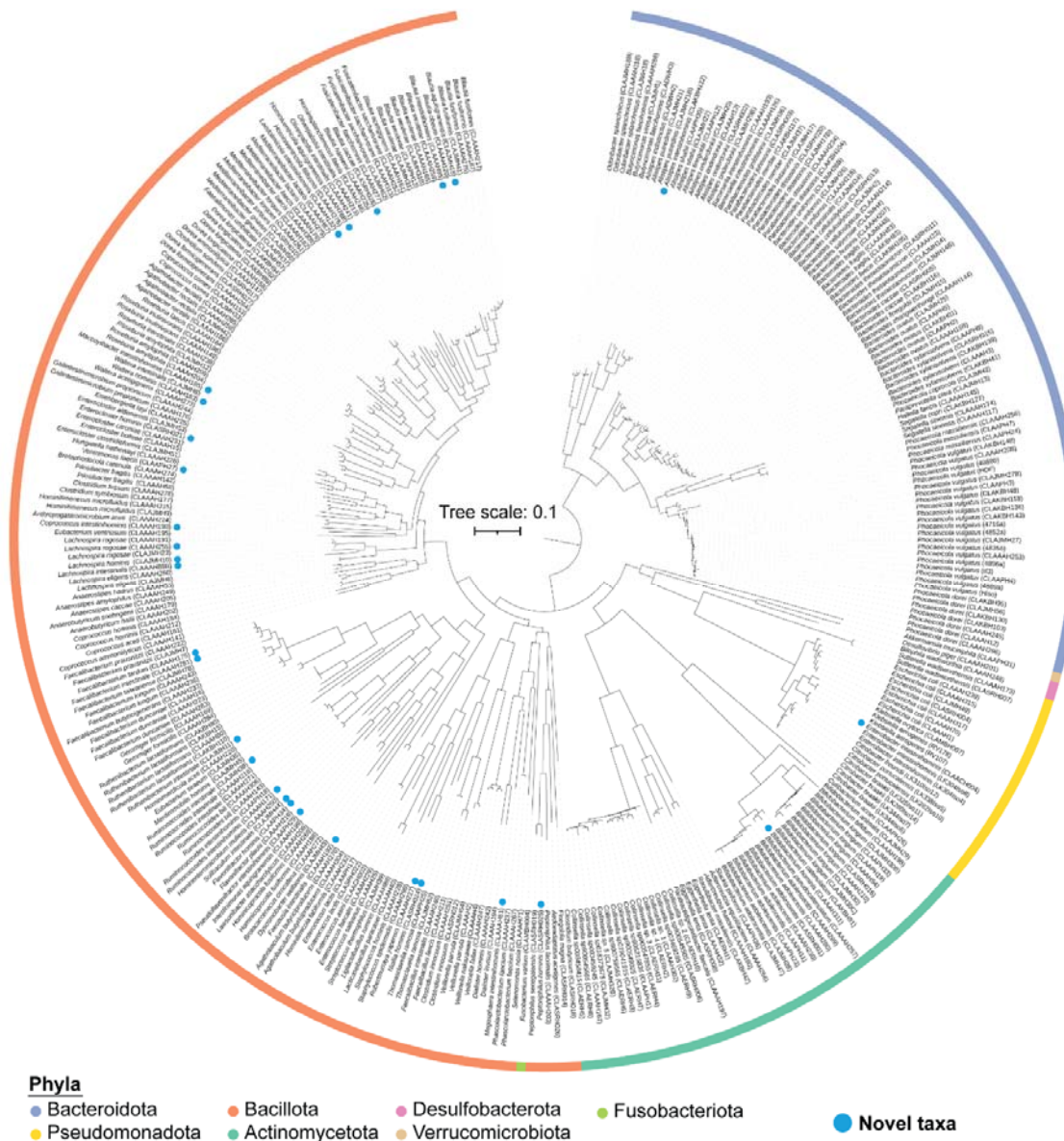
112

## 113 **Results**

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

115 The Human intestinal Bacteria Collection (HiBC) consists of 340 strains,  
116 representing 198 species, isolated from human faecal samples (**Figure 1**). It  
117 contains 29 families from across the seven dominant phyla in the human gut:  
118 *Bacillota* (n = 173 isolates), *Bacteroidota* (n = 95), *Actinomycetota* (n = 46),  
119 *Pseudomonadota* (n = 22), *Desulfobacterota* (n = 2), *Fusobacteriota* (n = 1), and  
120 *Verrucomicrobiota* (n = 1). Of the 198 species, 29 are novel taxa that we have  
121 described and named for validation according to both, the SeqCode<sup>6</sup> and the ICNP

122 <sup>22</sup>. High-quality genomes defined as >90% completeness ( $99.22 \pm 1.10\%$ ), <5%  
123 contamination ( $0.50 \pm 0.77\%$ ), >10x genome coverage ( $380.08 \pm 293.97$ ), were  
124 generated for all 340 strains. To improve access to the strains, their genomes,  
125 information on their cultivation, and isolation source material, we have created the  
126 HiBC web interface; [www.hibc.rwth-aachen.de](http://www.hibc.rwth-aachen.de). The website provides access to the  
127 complete collections of 16S rRNA gene sequences, genomes, plasmids, and  
128 metadata.



129  
130 **Figure 1: Phylogenomic diversity of isolates within HiBC.** Tree based on all 340 genomes,  
131 generated using PhyloPhlan <sup>23</sup>. Phyla are indicated with colours. The *Bacillota* are split due to the  
132 placement of *Fusobacteriota*, which separated strains assigned to 'Bacillota\_A' by GTDB. The  
133 potential need for splitting the phylum *Bacillota* is therefore independently supported by the  
134 PhyloPhlan tree, and GTDB. Blue circles identify strains belonging to the 29 novel species that are  
135 taxonomically described in this paper.

136

137 Description of dominant novel taxa within the human gut

138 There has been a renaissance in isolation of strains from the human gut in the last  
139 decade<sup>24–32</sup>. To place the HiBC isolates in the context of prior studies, we analysed  
140 previously published work in which phylogenetically diverse isolates were cultured.  
141 Across eight large-scale isolation studies and repositories, a total of 12,565 strains  
142 from the human gut were reported (**Figure 2a, Supplementary Table 1**). Of these,  
143 8,019 (63.8%) isolates claim to be requestable in the original publications, but only  
144 1,539 (12.2%) have been deposited to a culture collection to ensure long-term  
145 availability. However, this includes 1,063 isolates that have only been deposited to  
146 China General Microbiological Culture Collection Centre (CGMCC), which prevents  
147 the accessibility of risk group 2 organisms, which many of these are, to researchers  
148 outside China. This means only 476 bacterial isolates (3.8%) from the human gut are  
149 currently accessible to the community. While access to many of the isolates is  
150 limited, most isolates have been genomes sequenced (11,498, 91.5%) with most  
151 being high-quality (10,893, 86.7%). To understand the diversity captured by each  
152 study, we dereplicated the genomes to estimate the number of species cultured by  
153 each study (**Figure 2b**). A 33-fold reduction in diversity was found in some studies.  
154 Despite the value of capturing variability at the strain level, the limited number of  
155 samples used for isolation suggests redundancy within these collections. Therefore,  
156 while many studies claim to have a large collection of isolates, this is an  
157 overestimate of the true diversity captured.

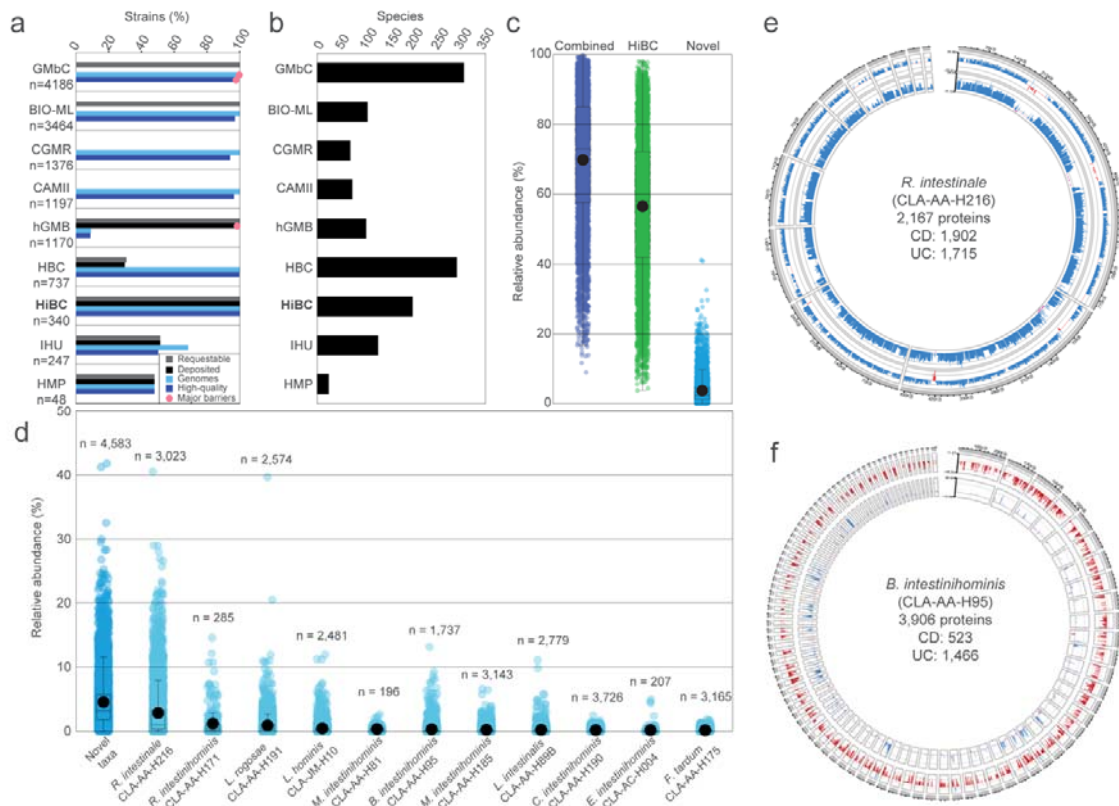
158 Next, we assessed the ability of all isolated strains to capture the diversity present in  
159 the human gut microbiota (**Figure 2c**). This was achieved by mapping the isolates  
160 genomes against a MAG collection for which corresponding relative abundance  
161 values were available<sup>33</sup>. The entire landscape of bacteria isolated from the human  
162 gut, both from the literature and this study, captured  $69.83 \pm 18.62$  % of an  
163 individual's microbiota, while the HiBC alone covered  $56.51 \pm 20.44$  %. The 29 novel  
164 taxa described within this work enhanced this coverage by  $3.75 \pm 3.99$  % on  
165 average, and represented >10% of the microbiota from 380 people, which peaked at  
166 >40% in two samples. Of the 29 novel taxa, 11 species were identified to have a  
167 mean relative abundance >0.15% within positive metagenomic samples (**Figure 2d**).  
168 The most dominant of these was *Ruminococcoides intestinale* (mean relative  
169 abundance = 2.84%; n = 3,023), followed by *R. intestinalihominis* (1.19%; n = 285).  
170 Although only recently described<sup>34</sup>, these results suggest that *Ruminococcoides* is a  
171 dominant but understudied genus within the human gut. A second genus of clear  
172 ecological importance to the gut is *Lachnospira*, in which we describe two dominant  
173 novel species. Interestingly, the two samples in which novel taxa accounted for  
174 >40% relative abundance were dominated by either *R. intestinale*, or the novel  
175 species *L. rogosae*, identified as a medium priority, HMP most wanted taxa  
176 (**Supplementary Results**). Further study of these species is needed to understand

177 their impact on host health and identify the mechanisms, which is now enabled by  
178 access to isolated strains.

179 To uncover associations of the novel taxa with human health conditions, we studied  
180 the ecological occurrence of each protein within *R. intestinale* CLA-AA-H216 (=DSM  
181 117897) (**Figure 2e**), as the most abundant novel taxa, and *Blautia intestinihominis*  
182 CLA-AA-H95<sup>T</sup> (=DSM 111354) (**Figure 2f**), as member of an important genus  
183 associated with both health and disease conditions. The association of individual  
184 proteins with disease states were determined using InvestiGUT<sup>35</sup>. This involves  
185 studying the prevalence of each individual protein encoded by the strain's genome,  
186 or plasmid, across thousands of metagenomes, then statistically determining if they  
187 occur significantly more or less frequently in the gut metagenomes of people with a  
188 specific condition or healthy controls. Out of the 2,167 proteins encoded by  
189 *R. intestinale* CLA-AA-H216, comparison between Crohn's disease (CD) patients  
190 and healthy controls identified a total of 1,902 differentially prevalent proteins, with  
191 1,883 being significantly more prevalent in healthy samples whilst 19 were more  
192 prevalent in CD samples. The same pattern was observed with Ulcerative colitis  
193 (UC) where a total of 1,715 significantly differentially prevalent proteins were  
194 identified, with 1,684 enriched in healthy samples, and 31 enriched in UC samples.  
195 Pathway analysis of the health-associated proteins identified multiple anti-  
196 inflammatory pathways, including the ABC transporter for spermidine (PotB/C/D), an  
197 anti-inflammatory polyamine<sup>36</sup>, and biosynthesis of biotin, which has been shown to  
198 be immunomodulatory<sup>37,38</sup>, ameliorating colitis<sup>39</sup>.

199 The association of *Blautia* spp. with inflammatory bowel diseases is more complex,  
200 as some species within this genus have been shown to ameliorate colitis<sup>40</sup>, leading  
201 to their inclusion within therapeutic products<sup>41</sup>. However, recent studies have  
202 suggested some *Blautia* spp. may exacerbate colitis<sup>42</sup>. Given the complex  
203 interaction of this genus with inflammatory bowel diseases, we studied the  
204 association of the novel species, *B. intestinihominis* CLA-AA-H95<sup>T</sup>, with both CD and  
205 UC. Out of the 3,906 proteins, comparison between CD patients and healthy controls  
206 identified 523 differentially prevalent proteins, with 488 being significantly more  
207 prevalent in healthy samples whilst 35 were more prevalent in CD samples. In  
208 contrast, comparison between UC patients and healthy controls identified 1,466  
209 proteins, with 16 being significantly more prevalent in healthy samples whilst 1,450  
210 were more prevalent in UC samples. The functionality of the large number of  
211 proteins enriched within UC samples was studied further, identifying ABC  
212 transporters for branched-chain amino acids (LivF/G/H/K/M)<sup>43</sup>, phosphate  
213 (PstA/B/C/S)<sup>44</sup>, and adenosine (NupA/B/C, BmpA)<sup>45</sup>, each of which has been linked  
214 to colitis.

215 These results highlight the importance of describing novel isolates and provide  
216 insights into the observed association of their proteins with health or disease.



217

218 **Figure 2: Ecology of isolated human gut bacteria and their proteins.** **a.** The number of strains  
 219 and genomes produced by eight major isolation projects, along with HiBC, were compared. Strains  
 220 were deemed requestable if it was stated as such in the original publication. They were deemed  
 221 deposited if culture collection identifiers were included in the original paper and were confirmed to  
 222 exist. Genomes were deemed high-quality if they were >90% complete, and <5% contaminated. The  
 223 number of strains within each study are stated, while the percentage meeting each criterion are  
 224 plotted. Red dots highlight datasets which have barriers to their accessibility, *i.e.*, data available upon  
 225 request or access limited to specific countries. Strain collections: GMbC, Global Microbiome  
 226 Conservancy<sup>28</sup>; BIO-ML, Broad Institute-OpenBiome Microbiome Library<sup>27</sup>; CGMR, Chinese Gut  
 227 Microbial Reference<sup>30</sup>; CAMII, Culturomics by Automated Microbiome Imaging and Isolation<sup>31</sup>;  
 228 hGMB, Human Gut Microbial BioBank<sup>29</sup>; HBC, Human Gastrointestinal Bacterial Collection<sup>26</sup>, HiBC,  
 229 Human Intestinal Bacteria Collection (this study); IHU, collection of the Institut Hospitalier  
 230 Universitaire Méditerranée Infection<sup>32</sup>; HMP, Human Microbiome Project at ATCC. **b.** Number of  
 231 species per isolate collection, either via manual curation (HiBC) or dereplication of the available  
 232 genomes (ANI values >95 % indicated identical species). **c.** The cumulative relative abundance of gut  
 233 metagenomes across 4,624 individuals from Leviatan *et al* (2022) covered by all isolated bacteria  
 234 across studies including the HiBC (Global isolates, dark blue), HiBC alone (green), or the subset  
 235 represented by the 29 novel taxa described in this work (light blue), which had matches within 4,583  
 236 of the samples. **d.** Relative abundance of dominant (mean relative abundance >0.25%) novel taxa  
 237 across 4,624 individuals, with the number of positive samples stated. Each strain represents a distinct  
 238 novel species, described in detail in in the protologues at the end of the methods section. **e-f.**  
 239 Genomic location of proteins significantly differentially prevalent between Crohn's disease (CD)  
 240 samples and healthy controls (inner ring), or ulcerative colitis (UC) samples and healthy controls  
 241 (outer ring). The delta-prevalence (prevalence in healthy donors – prevalence in corresponding  
 242 patients) is shown in blue (more prevalent in healthy controls), red (UC), or mauve (CD). The species,  
 243 strain, number of proteins predicted within the genome, and those significantly differentially between  
 244 health conditions are shown within the circle. Only contigs >10 kp were plotted.

245

246 Plasmid landscape in the cultured strains and identification of prevalent  
247 megaplasmids Plasmids are present in the human gut and vary between  
248 geographically distinct populations<sup>13</sup>. However, most analyses have been culture-  
249 independent, preventing accurate taxonomic assignment<sup>14,15,46</sup>. Even studies on  
250 horizontal gene transfer using isolates have not reconstructed plasmids and  
251 accounted for their impact<sup>47</sup>. Analysis of bacterial isolates has uncovered many  
252 novel plasmids from non-human primates<sup>48</sup>. As many bacterial genome assembly  
253 workflows do not consider plasmids, we developed a genome pipeline that first  
254 searches for the presence of plasmids using Recycler<sup>49</sup> and plasmidSPAdes<sup>50</sup>,  
255 assembles them, and then removes their reads from consideration during genome  
256 assembly (see Methods). This resulted in the reconstruction of plasmids from 46% of  
257 the strains (155 out of 340), with a total of 266 plasmids (**Figure 3a**). Plasmids were  
258 identified in all phyla except *Fusobacteriota*, for which the HiBC includes only a  
259 single isolate. Almost half of the plasmid-positive strains contained more than one  
260 plasmid (64 out of 155 strains), up to 6 plasmids in a strain of *Phocaeicola vulgatus*  
261 and two strains of the novel species *Enterobacter intestinhominis* (**Figure 3b**).  
262 Across the 266 plasmids identified in HiBC, 3,697 proteins were predicted, although  
263 only 639 (17.28 %) could be functionally annotated. Of these, nine were antibiotic  
264 resistance genes, four of which were copies of the *APH(2'')-IIIa* aminoglycoside  
265 resistance gene found on plasmids of identical size (7,686 bp) in four strains of  
266 *Ruthenibacterium lactatiformans* (strains CLA-KB-H110, CLA-KB-H15, CLA-KB-H80,  
267 CLA-AA-H80).

268 Megaplasmids (>100 kp) have been described to occur in *Lactobacillaceae*<sup>51</sup> and  
269 *Bifidobacterium breve*<sup>52</sup>, but their occurrence in a broader range of commensals  
270 from the human gut is unknown. We evaluated the length of the reconstructed  
271 plasmids and found the average length was  $11.74 \pm 17.25$  kp (**Figure 3c**). This  
272 included the identification of two megaplasmids, one within a strain of the recently  
273 described species *Hominifimenecus microfluidus* (strain CLA-JM-H9, = DSM  
274 114605; 125.6 kp) and the other in a strain of *Phocaeicola vulgatus* (strain CLA-AA-  
275 H253, = DSM 118718; 103.4 kp). Given the number and size of plasmids a single  
276 strain can contain, we assessed their impact on species assignments based on  
277 average nucleotide identity, but confirmed that plasmids had only a minor effect on  
278 the taxonomic assignment of a genome (**Supplementary Results**).

279



289 reconstructed pBAC plasmids. **g.** Proteins encoded on pBAC plasmids from different strains with  
290 significantly different prevalence between Crohn's disease (CD) patients and healthy controls. Each  
291 point represents a single differentially prevalent protein, coloured based on the pBAC cluster the  
292 protein was encoded on (panel e). **h.** Plasmid map of pMNCAT\_H253. The innermost rings represent  
293 the association of proteins differentially prevalent between CD samples and healthy controls (first  
294 ring), or ulcerative colitis (UC) samples and healthy controls (second ring) via InvestiGUT<sup>35</sup>. Proteins  
295 enriched in CD are purple, those enriched in healthy samples are blue, and those enriched in UC are  
296 red. The third ring represents the GC content relative to the average of the entire plasmid. Boxes are  
297 used to represent genes identified on the plasmid in the forward (outer ring) and reverse (inner ring)  
298 strand, with coloured boxes being assigned a COG category, while grey boxes represent COG-  
299 unassigned proteins. The fimbriae loci proteins are indicated in dark grey. Enzymes with a single  
300 restriction site are indicated on the outer ring in orange. **i.** Quantification of cell adhesion from a  
301 pMNCAT-containing strain of *P. vulgatus* (CLA-AA-H253) and its closest relative strain (H-iso) without  
302 the megaplasmid (see methods). Statistics: paired, one-tailed t-tests. **j.** Representative images of cell  
303 adhesion of the two selected strains. The green scale bars represent 50  $\mu\text{m}$ .

304

305 Sequence comparison with MobMess<sup>53</sup> identified 168 plasmid clusters, the two  
306 largest of which were specific to the *Bacteroidales* (**Supplementary Figure 1a**). The  
307 largest cluster shared similarities with 1,266 plasmids within the comprehensive  
308 plasmid database, PLSDB<sup>54</sup>, but lacked an assigned name, hence we named it  
309 pBAC (plasmid *Bacteroidales*). The second largest plasmid cluster was identified as  
310 the cryptic plasmid, pBI143, which was recently described to be associated with IBD  
311<sup>14</sup> and occurred in 14 isolates across eight species. Given that pBAC matches 1,266  
312 plasmids within PLSDB, while pBI143 matched only 719, it may be the most prolific  
313 plasmid in the human gut. We therefore investigated pBAC further, observing it has  
314 been most frequently found in the USA but has also been detected in Denmark,  
315 Japan, Ireland, and China (**Supplementary Figure 1b**). Investigation of the  
316 sequence similarity network identified that pBAC occurred in both a shorter (pBAC-1;  
317 ~4,220 bp) and longer (pBAC-2; ~5,419 bp) form, with the two clusters being  
318 connected by two 9,494 bp variants (pBAC-3) (**Figure 3d**). Comparison of  
319 representative sequences from each of these three clusters uncovered that both,  
320 pBAC-1 and pBAC-2 shared no similarity, but were identical to regions of pBAC-3  
321 (**Figure 3e**). These results suggest that the different forms of pBAC share a common  
322 origin, implying they can co-occur within the same strain. We identified CLA-KB-  
323 H139 as a strain which was predicted to contain both pBAC-1 and pBAC-2 and  
324 confirmed their existence within the isolate by extracting plasmids from a diverse  
325 range of species predicted to contain different forms of pBAC (**Figure 3f**). This  
326 analysis confirmed that pBAC occurs across the *Bacteroidales*. Given the prevalence  
327 of pBAC within common denizens of the human gut, we aimed to study their  
328 functional potential. All three pBAC clusters encode mobilisation machinery  
329 (MbpA/B/C)<sup>55</sup>. Interestingly, both pBAC-1 and pBAC-2 encode different toxin-  
330 antitoxin pairs, with that in pBAC-1 resembling YoeB and YefM<sup>56</sup>, while the pBAC-2  
331 encoded system shared similarity with RelE and Phd<sup>57</sup>. Given their differing protein  
332 content, we hypothesised their association with host health may differ, hence studied  
333 the prevalence of their encoded proteins across disease cohorts (**Figure 3g**). pBAC-

334 1 was observed to encode 2-3 proteins that were significantly more prevalent within  
335 CD patients, with the most enriched protein being a replication protein, followed by  
336 YefM-like and YoeB-like proteins. Conversely, pBAC-2 plasmids encoded 5-6  
337 proteins within significantly lower prevalence in CD patients, with the most being the  
338 RelE-like and Phd-like proteins. Toxin-antitoxin systems within *Bacteroidota*  
339 plasmids have previously been reported, although pBAC shared no similarity with  
340 these described plasmids<sup>58</sup>. These results suggest the pBAC encoded toxin-  
341 antitoxin pairs are associated with human health, potentially by altering the fitness of  
342 their host strain.

343 In addition to the highly prevalent pBAC plasmid, we studied the megaplasmid  
344 present in *P. vulgatus* (CLA-AA-H253), a prevalent and dominant species in the  
345 human gut. The megaplasmid matched 26 plasmids within PLSDB, including multiple  
346 designated as pMMCAT (**Supplementary Figure 1a**), a recently proposed large  
347 plasmid from *Bacteriodales*, which has yet to be confirmed using sequencing-  
348 independent methods<sup>59</sup>, but has been shown to impact its host ability to form  
349 biofilms<sup>60</sup>. As such, the *P. vulgatus* megaplasmid was designated pMMCAT\_H253.  
350 While megaplasmids have previously been observed in sequencing data of human  
351 gut commensals, their existence is rarely confirmed. PacBio sequencing generated a  
352 complete genome for strain CLA-AA-H253, which confirmed the reconstruction of  
353 pMMCAT\_H253. To provide sequencing-independent validation of pMMCAT\_H253,  
354 we used pulsed field gel electrophoresis which showed a band of ~100 kb after  
355 treatment using XbaI, for which the plasmid had a single restriction site allowing for  
356 its linearisation (**Supplementary Figure 2a**). The bands' identity was confirmed  
357 using southern blot with primers designed based on the predicted pMMCAT\_H253  
358 sequence (**Supplementary Figure 2b**). Out of the 104 proteins encoded on  
359 pMMCAT\_H253, only 6 could be assigned to a functional category, highlighting the  
360 need for further characterisation of these proteins (**Figure 3h**). A locus containing  
361 four proteins assigned to 'fimbrillin family proteins' was identified within a region of  
362 lower GC content compared to the plasmids average, which may suggest this region  
363 represents variable cargo and not the backbone. The presence of 'fimbrillin family  
364 proteins' on pMMCAT\_H253 was further investigated as fimbriae can facilitate  
365 adhesion of cells<sup>61</sup>, including to the host epithelium, to enhance colonisation<sup>62</sup>. We  
366 therefore studied the ability of CLA-AA-H253, the pMMCAT containing strain, and its  
367 closest related strain (H-iso, 98.9% ANI), which lacks pMMCAT, to adhere to plastic  
368 wells. CLA-AA-H253 adhered to the plastic significantly more than H-iso after only 4  
369 hours, with 2-fold more bacteria adherent after 30 hours of growth (**Figure 3i**). The  
370 adherence of strain CLA-AA-H253 was observed in the wells, while H-iso was  
371 adhered in patches across the wells (**Figure 3j**). These results support the need for  
372 greater study of pMMCAT as a potential source for phenotype variation within  
373 genera known to contribute to host health<sup>64,65</sup>.

374

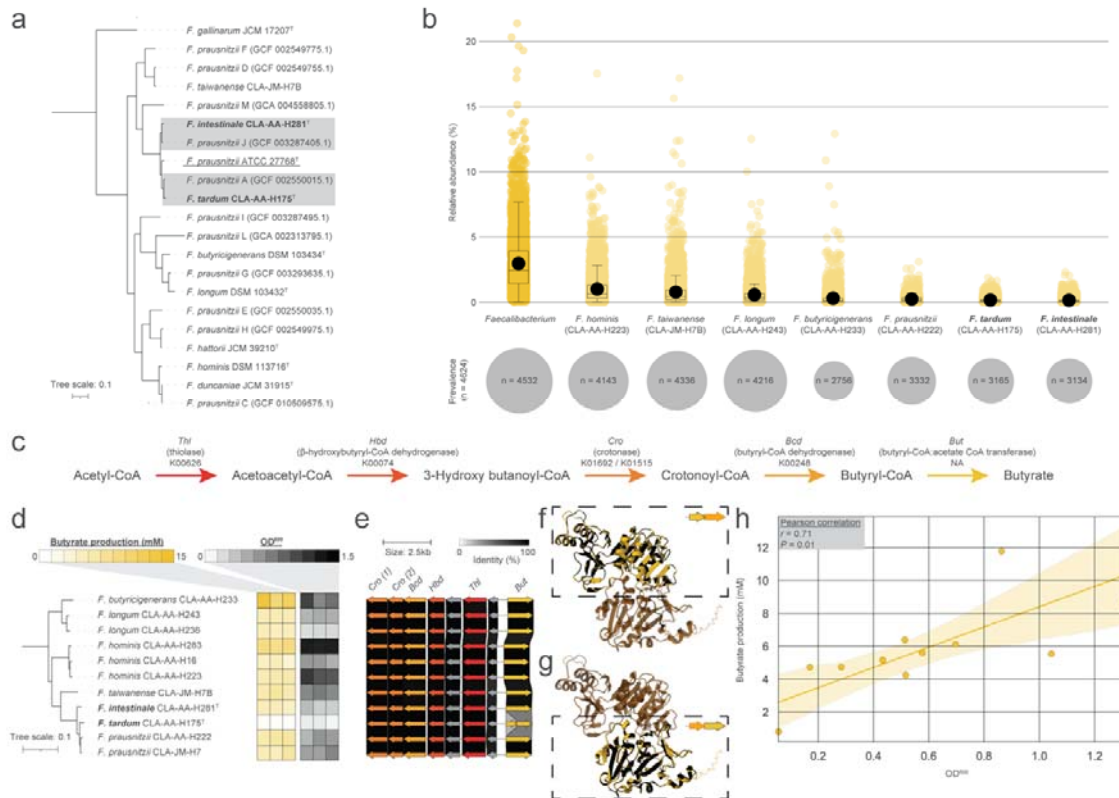
375 Enhanced diversity of *Faecalibacterium* species

376 The genus *Faecalibacterium* currently (January 2025) contains seven validly  
377 published and correctly named species. HiBC contains representatives for five of  
378 these, and isolates representing two novel species. Initial 16S rRNA gene analysis  
379 assigned these novel species as *Faecalibacterium prausnitzii*, a common  
380 commensal of the gut microbiota that has been associated with human health  
381 conditions ranging from reducing inflammation<sup>66</sup> to improving cognitive function in  
382 Alzheimer's disease<sup>67</sup>. Many of these beneficial observations have been attributed  
383 to *F. prausnitzii*'s ability to produce large amounts of the short-chain fatty acid  
384 butyrate<sup>68</sup>.

385 Taxonomically, the genomic diversity of *F. prausnitzii* is proposed to be greater than  
386 suggested by 16S rRNA gene sequence analysis<sup>69</sup>. This has led the GTDB to split  
387 this species into 12 proposed species. HiBC contains representatives of the not yet  
388 described species defined by GTDB-Tk as *F. prausnitzii* A (strain CLA-AA-H175) and  
389 *F. prausnitzii* J (strain CLA-AA-H281). This was confirmed using ANI and can be  
390 observed in a phylogenomic tree (**Figure 4a**). We therefore propose to name these  
391 novel species *Faecalibacterium tardum* (CLA-AA-H175) and *Faecalibacterium*  
392 *intestinale* (CLA-AA-H281). Protologues for these, and the other 27 novel taxa  
393 described in this paper are provided at the end of the Methods section. Both novel  
394 species were observed in the majority of microbiota studied at similar relative  
395 abundances as *F. prausnitzii* (*F. prausnitzii* =  $0.24 \pm 0.24$  %; *F. tardum* =  $0.16 \pm 0.20$   
396 %; *F. intestinale* =  $0.14 \pm 0.21$  %) (**Figure 4b**).

397 Butyrate production by *Faecalibacterium* spp. is critical for host health, hence we  
398 studied variability between the strains in their ability to produce butyrate (**Figure 4**;  
399 **Supplementary Table 2**). Butyrate production by *Faecalibacterium* is achieved via  
400 the acetyl-CoA pathway, which is also the dominant pathway for butyrate production  
401 in the human gut (**Figure 4c**)<sup>70,71</sup>. While all *Faecalibacterium* strains within HiBC  
402 produced butyrate, the amounts varied greatly, with *F. tardum* (strain CLA-AA-H175)  
403 producing only  $0.8 \pm 0.16$  mM, whereas strains of *F. longum* and *F. butyricigenans*  
404 produced >10 mM (**Figure 4d**). Given that butyrate production is a conserved  
405 phenotype of this genus, we investigated if variation in the acetyl-CoA pathway loci  
406 was responsible for the observed variation between isolates (**Figure 4e**). While little  
407 variation was observed between most of the strains, *F. tardum* CLA-AA-H175  
408 encoded two truncated copies of the butyryl-CoA:acetate-CoA transferase gene. To  
409 understand the cause of these truncated genes, we studied their placement in  
410 relation to each other and identified they occurred on the same strand but in different  
411 frames, with an overlap by 55 bp. Interestingly, the second truncated gene is  
412 encoded with 'GTG' as an alternative start codon, which may lead to lower  
413 transcription and hence alter butyrate production further<sup>72</sup>. Structural modelling of  
414 these proteins using AlphaFold3 identified that the truncated proteins form a complex  
415 that share 100% (**Figure 4f**) and 90% (**Figure 4g**) identity with the full protein from  
416 *F. prausnitzii* CLA-AA-H222, respectively. Acetate utilisation and butyrate production  
417 have been linked to the growth of *Faecalibacterium* spp.<sup>73</sup>, hence we considered if  
418 the ability of the isolate to grow under the testing conditions alters its metabolism,

419 and eventually the ability to produce butyrate. We therefore correlated the average  
 420 butyrate production of each strain against its average growth, measured by OD600  
 421 (**Figure 4h**). A strong significant correlation ( $r = 0.76$ ;  $p = 0.01$ ) was identified  
 422 between the growth of a strain and the amount of butyrate it produced. These results  
 423 suggest that strains of *Faecalibacterium* vary greatly in their ability to produce  
 424 butyrate, either as a result of genetic modification, as in the case of *F. tardum*, or due  
 425 to the strains ability to grow *in vitro*. This further implies that strains that grow at high  
 426 abundance *in vivo* are likely to produce the most butyrate.



427

428 **Figure 4: Novel diversity within *Faecalibacterium* and strain-dependent butyrate production. a.**  
 429 The two novel species of *Faecalibacterium* described within this paper placed within the current  
 430 landscape of *Faecalibacterium* spp. with a valid name, along with the type genomes for proposed  
 431 divisions of *F. prausnitzii*, as determined by GTDB. The phylogenomic tree was rooted on  
 432 *Ruminococcus bromii* ATCC 27255<sup>T</sup>. Novel species are in bold, and the type strain of *F. prausnitzii*  
 433 is underlined. **b.** Relative abundance, and prevalence, of the genus *Faecalibacterium*, and each  
 434 *Faecalibacterium* species represented within HiBC across 4,624 metagenomic samples. **c.** Butyrate  
 435 production pathway in *Faecalibacterium* with gene names and KEGG ortholog identifiers when  
 436 possible. **d.** Phylogenomic tree of the *Faecalibacterium* strains within HiBC, displaying the ability of  
 437 each strain to produce butyrate over a 48 h-period, along with the OD600 that the strain achieved  
 438 during the testing period ( $n = 3$  independent batch cultures for each strain; the replicates are shown  
 439 with individual boxes). The phylogenomic tree was rooted on *R. bromii* ATCC 27255<sup>T</sup>. **e.** Sequence  
 440 comparison of the butyrate production loci across the *Faecalibacterium* strains. Genes are coloured  
 441 based on their assignment to each step in the butyrate production pathway in panel c. **f.** AlphaFold3  
 442 model of the *But* complex in *F. tardum* CLA-AA-H175 against the full *But* protein in *F. prausnitzii*  
 443 CLA-AA-H222. The first CLA-AA-H175 *But* gene is highlighted in yellow in the dashed box, while the  
 444 second gene is shown in faded orange. The highlighted protein is indicated in the top right of the  
 445 dashed box. **g.** Same as in panel f, but this time the second CLA-AA-H175 *But* gene is highlighted in

446 yellow in the dashed box, while the first gene in faded orange. The highlighted protein is indicated in  
447 the top right of the dashed box. **h.** Correlation of the mean OD600 against the mean butyrate  
448 production of each strain with a linear regression and its 95% confidence interval.

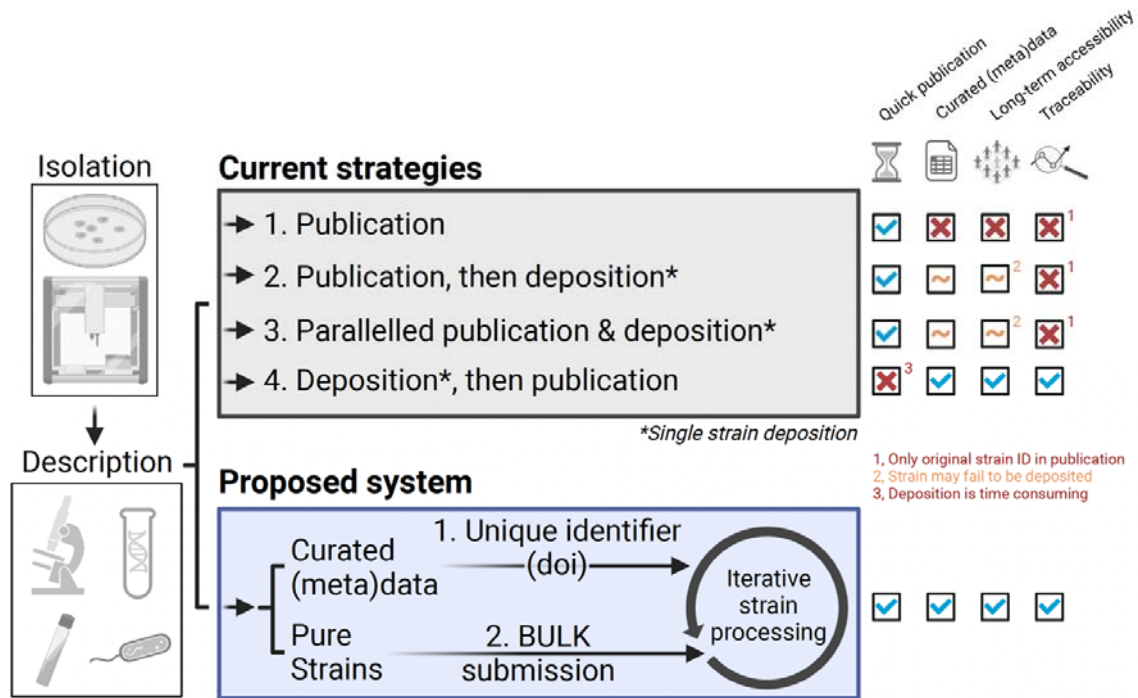
449

#### 450 Traceable strain identifiers facilitate bulk isolates deposition

451 Published isolates are often not easily accessible to researchers, impeding them  
452 from working with the same quality-controlled strains that are needed to compare  
453 and reproduce data. The main barriers to making isolates publicly available is the  
454 time-demanding process of strain deposition and the lack of unique strain identifiers  
455 for reliable referencing of strains while they are being processed. Researchers  
456 frequently deposit only a small selection of strains (**Figure 2a**), preventing access to  
457 an entire collection that may contain functionally relevant diversity, as highlighted  
458 above with *Faecalibacterium* (**Figure 4**). Current strategies for publishing large  
459 collections of isolates vary, as shown in **Figure 5** (upper box). Publication with a  
460 statement such as "available upon request" without deposition of the strains (case 1)  
461 appears simple and quick. Yet, it should not be accepted anymore, as the strains  
462 and their data are effectively inaccessible and ultimately will be lost over time <sup>74</sup>.  
463 Depositing in a culture collection (DSMZ, JCM, CCUG, BCCM, KCCM, etc.) ensures  
464 that metadata quality is maintained and guarantees the long-term accessibility of the  
465 strains for the scientific community. Current best practice is to ensure that all strains  
466 included in a manuscript are deposited and processed at a culture collection before  
467 publication, so that strain catalogue numbers can be issued and included in the  
468 manuscript (case 4) <sup>75,76</sup>. The shortcoming of the latter approach is that deposition  
469 and strain processing for quality checks is time consuming, and the time required  
470 grows linearly with the number of strains to be deposited. The deposition of a  
471 collection such as HiBC is thus a significant undertaking.

472 Accordingly, a bulk submission system ([www.dsmz.de/bulk-deposit](http://www.dsmz.de/bulk-deposit)) has been  
473 developed by the Leibniz Institute-German Collection of Microorganisms and Cell  
474 Cultures (DSMZ) to address these issues and facilitate large-scale strain deposition,  
475 as shown in **Figure 5** (lower box) (**Supplementary Results; Supplementary Figure**  
476 **9**). The HiBC presented in this article has been used as a blueprint to develop this  
477 system. After the culture collection has accepted the submission of a researchers  
478 strain collection, it requests the strains in large batches which are deposited in an  
479 iterative process. Once cultures have been received by the DSMZ and passed initial  
480 checks, the system allows for the allocation of stable identifiers in the form of Digital  
481 Object Identifiers (DOIs) provided by the database StrainInfo  
482 ([www.straininfo.dsmz.de](http://www.straininfo.dsmz.de)), where each strain is permanently registered with its  
483 metadata. This step ensures that all data (e.g., publications and sequence data) can  
484 be linked to the corresponding strain entry when the strains become available. Strain  
485 entries are initially displayed with the status 'Deposition in progress'. When the  
486 deposition process is completed (which may take several months) and the strains  
487 become available in the DSMZ catalogue, the status of the strains are updated to  
488 'Published' in StrainInfo. Therefore, this system would accommodate fast publication

489 without compromising metadata quality, ensuring adherence to good scientific  
 490 practices.



491

492 **Figure 5: Schematic of the proposed system for large-scale strain deposition.** Once bacteria  
 493 have been isolated and described, there are multiple strategies for publishing the strains. In the upper  
 494 box, we have four example strategies, each of which can be applied on a single strain basis, but are  
 495 not designed for large-scale submission of strains. The benefits and negatives of each method is  
 496 indicated on the right of the methods under four main categories, where a cross indicates a missing  
 497 aspect to the strategy. The lower box pictures the main advantage of creating unique strain identifiers  
 498 early in the process as a foundation for bulk deposition of many strains ([www.dsmz.de/bulk-deposit](http://www.dsmz.de/bulk-deposit)).  
 499 Figure designed in BioRender.

500

## 501 Discussion

502 Over the last decade, the renewed interest in cultivation has expanded the number  
 503 of genomes available for taxa within the human gut. However, the corresponding  
 504 strains are rarely accessible to the research community, as highlighted by only 3.8%  
 505 of human gut isolates being publicly deposited within a culture collection. The  
 506 process of strain deposition within culture collections has previously been highlighted  
 507 as a major limiting factor in making strains publicly available<sup>77</sup>. To overcome this, we  
 508 initiated a system for bulk deposition of strains. Streamlined registration of strains  
 509 before bulk submission ensures their traceability, facilitating publication while  
 510 ensuring strains are consistently referenced. Application of this system allowed for  
 511 deposition of 340 human gut isolates in the DSMZ, facilitating the naming and  
 512 description of 29 novel species, including the type species of three novel genera.  
 513 Based on this, the HiBC is the first collection of human gut isolates that is entirely  
 514 publicly accessible.

515 While the new bulk deposition process described here currently relies on manual  
516 steps and separate communication between multiple partners, DSMZ is currently  
517 developing a deposition management system that will largely automate this process.  
518 Similar to the review systems of scientific journals, it will expedite the deposition  
519 process by facilitating the submission of deposition data and the communication  
520 between submitter and curator in a web portal named StrainRegistry  
521 (<https://straininfo.dsmz.de/strainregistry>). The StrainRegistry will be closely  
522 connected to the StrainInfo database described in this manuscript for automated DOI  
523 assignment to identify strains throughout literature and databases. We will also  
524 encourage participation of other culture collections, to allow researchers to submit  
525 their strains to multiple collections at the same time. This will significantly streamline  
526 the deposition process and make it more transparent, while providing the persistent  
527 identifiers necessary to connect strain data at any time.

528 Large-scale cultivation of human gut bacteria in this study led to the isolation of  
529 many novel species and genera. Whilst previous cultivation studies report many  
530 isolates representing novel diversity, these bacteria are rarely taxonomically  
531 described and named. The discovery of prevalent and important novel species in this  
532 work will facilitate future experiments on the role of bacteria in health and disease.  
533 This included *Ruminococcoides intestinale*, *Blautia intestinihominis*, and the isolation  
534 of multiple strains of *Faecalibacterium*, including species with the potential to reduce  
535 gastrointestinal inflammation<sup>66</sup>. We confirmed that *F. prausnitzii* represents a  
536 diverse group of related species that show reduced 16S rRNA gene sequence  
537 diversity<sup>69</sup>. By naming two of these species and describing the butyrate production  
538 and growth characteristics of seven different species in this genus, we observed the  
539 ability of these strains to produce butyrate was proportional to their ability to grow.  
540 Given these results, further quantification of these strains ability to grow within the  
541 human gut is required to understand their contribution to SCFA levels *in vivo*, and  
542 therefore their potential impact on host health.

543 The reconstruction of plasmids for almost half of HiBC isolates allows the direct link  
544 of plasmids to strains of various species, information previously lacking from studies  
545 on plasmids from the human gut<sup>13-15</sup>. Of note, *Bacteroidales* strains often contain  
546 multiple plasmids, particularly copies of pBI143 or pBAC. The pBAC plasmid was the  
547 most frequently reconstructed plasmid, and was observed to occur in three forms.  
548 Investigation of these forms uncovered both major forms have differing associations  
549 with host health. The megaplasmid pMMCAT was also identified within a strain of  
550 *P. vulgatus*, leading to enhanced adhesion to surfaces compared to strains lacking  
551 pMMCAT. By making the reconstructed plasmids, and strains publicly available, we  
552 believe this resource expands the toolbox of methods for genetically modifying non-  
553 model commensal gut microbes.

554 Metagenomic studies have provided important insights into the taxonomic and  
555 functional diversity present in the human gut, however, they are unable to provide  
556 mechanistic insights. By enabling broad access to human gut isolates in this work,

557 the functionality of taxa can be experimentally validated, facilitating mechanistic  
558 studies of microbe-host interactions.

559

## 560 **Methods**

### 561 Ethics

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

569

### 570 Bacterial isolation and cultivation

571 Human stool was collected in sterile plastic bags and stored in tightly sealed plastic  
572 buckets until further processing. An oxygen scavenger sachet (BD Biosciences; ref.  
573 260683) was added to each bucket before sealing to reduce exposure to oxygen. All  
574 samples were processed in the lab within 24 hours of collection. First the faecal  
575 material was homogenised by manual kneading of the plastic bag. Stool samples  
576 were either mixed 1:5 with anaerobic FMT media <sup>78</sup>, distributed to 1 mL aliquots and  
577 stored at -80 °C until further use, or the samples were further processed as  
578 described previously <sup>76</sup>. Hence approximately 5 g of the samples were dissolved in  
579 50 ml of anaerobic PBS supplemented with peptone (0.05% w/v), L-cysteine (0.05%  
580 w/v) and dithiothreitol (DTT) (0.02% w/v) by shaking the glass flask intensely. A  
581 syringe with needle was used to transfer 5 ml of the slurry through a rubber stopper  
582 (previously flamed using ethanol) into a second glass flask containing 45 ml of the  
583 same buffer in an anaerobic atmosphere (89.3% N<sub>2</sub>, 6% CO<sub>2</sub>, 4.7% H<sub>2</sub>) to create a  
584 1:100 dilution of the original sample. The flask was moved into an anaerobic  
585 workstation (MBraun GmbH, Germany) to prepare 2 ml aliquots under anaerobic  
586 conditions, which were mixed with 2 ml of 40% anaerobic glycerol to a final  
587 concentration of 20%, then stored at -80 °C until use.

588 Two different approaches were used for bacterial isolation. For classical isolation,  
589 the faecal aliquots were thawed inside the anaerobic workstation and diluted with  
590 anoxic PBS (see above) in a tenfold dilution series (10<sup>-2</sup> to 10<sup>-6</sup>). 50 µl of each  
591 dilution step were transferred onto different agar plates and spread using a L-  
592 spatula. Single colonies were picked after 1-7 days of incubation at 37 °C under  
593 anaerobic conditions. Bacterial cells were re-streaked at least three times to  
594 guarantee purity of the culture. For the second isolation approach, the single-cell  
595 dispenser b.sight (Cytena GmbH, Germany) was used as described before <sup>79</sup>.  
596 Details on culture media preparation and ingredients are provided in the  
597 **Supplementary Methods**.

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

607

#### 608 Metabolite production analysis

609 Concentrations of short-chain fatty acids (SCFAs) (acetate, butyrate, propionate,  
610 valerate), branched chain fatty acids (isobutyrate, isovalerate), intermediate  
611 metabolites (ethanol, formate, lactate, 1,2-propandiol, 1-propanol, succinate), as well  
612 as mono- and disaccharides (galactose, glucose, lactose) were measured by high-  
613 performance liquid chromatography (HPLC). The bacterial strains were grown  
614 anaerobically, apart from *Robertmurraya yapensis* CLA-AA-H227 which was grown  
615 under aerobic conditions, in YCFA broth (DSMZ Medium No. 1611) in Hungate tubes  
616 for 48 hours at 37 °C. Triplicate cultures from each strain were measured. Baseline  
617 controls were taken from each sterile Hungate tube before inoculation. Each taken  
618 sample was centrifuged (10,000 x g, 10 min, 4 °C), supernatants were collected and  
619 stored at -80 °C until HPLC measurement. Samples were prepared and measured  
620 (including HPLC settings) as described previously<sup>76</sup>. External standards were used  
621 for concentration determination by comparison of peak retention times (HPLC grade  
622 compounds were purchased from Sigma-Aldrich). Peaks were integrated using the  
623 Chromaster System Manager Software (Version 2.0, Hitachi High-Tech Science  
624 Corporation 2013, 2017). Metabolite concentrations >0.2 mM (limit of detection  
625 (LOD) for citrate, 1-propanol), >0.24 mM (LOD for butyrate, formate, galactose,  
626 glucose, isobutyrate, isovalerate, lactose, valerate), >0.4 mM (ethanol, 1,2-  
627 propandiol) and >0.8 mM (acetate, lactate, propionate, succinate) were considered  
628 for statistical analysis if present in all three replicates. Production and consumption  
629 of metabolites was calculated by subtracting the baseline values from the sample  
630 taken after 48 hours of growth.

631

#### 632 Cellular fatty acids (CFAs) determination

633 Cellular fatty acids were measured at the Leibniz Institute-DSMZ. The strains  
634 representing novel taxa were grown under the conditions indicated in the respective  
635 protologues. Approximately 100 mg (wet weight) of cell biomass was extracted  
636 according to the standard protocol of the Microbial Identification System (MIDI Inc.,  
637 version 6.1; technical note #101). CFAs were analysed by converting them into fatty  
638 acid methyl esters (FAMES) through saponification, methylation, and extraction. The

639 resulting FAME mixtures were separated using gas chromatography (GC) and  
640 detected with a flame ionization detector (FID). Subsequent analysis involved the  
641 identification of fatty acids using a GC-MS system (Agilent GC-MS 7000D) as  
642 described by Vieira et al. (2021)<sup>80</sup>. Further derivatization methods were used for  
643 structural elucidation of unidentified compounds. For branched-chain fatty acids,  
644 cyclo-positions, and multiple double bonds, 4,4-dimethyloxazoline (DMOX)  
645 derivatives were analysed<sup>81</sup>.

646

#### 647 Isolation of genomic DNA

648 DNA was isolated using a modified protocol according to Godon *et al.*<sup>82</sup>. Frozen cell  
649 pellets were mixed with 600 µl stool DNA stabilizer (Strattec biomedical), thawed, and  
650 transferred into autoclaved 2-ml screw-cap tubes containing 500 mg 0.1 mm-  
651 diameter silica/zirconia beads. Next, 250 µL 4 M guanidine thiocyanate in 0.1 M Tris  
652 (pH 7.5) and 500 µL 5 % N-lauroyl sarcosine in 0.1 M PBS (pH 8.0) were added.  
653 Samples were incubated at 70 °C and 700 rpm for 60 min. A FastPrep® instrument  
654 (MP Biomedicals) fitted with a 24 × 2 ml cooling adaptor filled with dry ice was used  
655 for cell disruption. The program was run 3 times for 40 s at 6.5 M/s. After each run,  
656 the cooling adapter was refilled with dry ice. An amount of 15 mg  
657 Polyvinylpyrrolidone (PVPP) was added and samples were vortexed, followed by 3  
658 min centrifugation at 15.000 x g and 4 °C. Approximately 650 µl of the supernatant  
659 were transferred into a new 2 ml tube, which was centrifuged again for 3 min at  
660 15.000 x g and 4 °C. Subsequently, 500 µl of the supernatant was transferred into a  
661 new 2 ml tube and 50 µg of RNase was added. After 20 minutes at 37 °C and 700  
662 rpm, gDNA was isolated using the NucleoSpin® gDNA Clean-up Kit from Macherey-  
663 Nagel. Isolation was performed according to the manufacturer's protocol. DNA was  
664 eluted from columns twice using 40 µl Elution buffer and concentration was  
665 measured with NanoDrop® (Thermo Scientific). Samples were stored at -20 °C.

666

#### 667 Genome library preparation and Illumina Sequencing

668 Library preparation and sequencing were conducted using the NEBNext Ultra II FS  
669 DNA Library Prep Kit for Illumina with dual index primers and ~300 ng of DNA (NEB  
670 Inc.) on an automation platform (Biomeki5, Beckman Coulter) according to the  
671 manufacturer's instructions. The time used for enzymatic shearing to ca. 500 bp to  
672 be used on a MiSeq run was 10 minutes and to ca. 200 bp to be used on a NextSeq  
673 run was 30 minutes. PCR enrichment of adaptor-ligated DNA was conducted with  
674 five cycles using NEBNext Multiplex Oligos for Illumina (NEB, USA) for paired-end  
675 barcoding. AMPure beads (Beckman Coulter, USA) were used for size selection and  
676 clean-up of adaptor-ligated DNA. Sequencing was performed either at the IZKF Core  
677 Facility Genomics (Uniklinik RWTH Aachen) on a NextSeq platform (Illumina) (PE  
678 150 bp), or on a MiSeq (Illumina) (PE 300 bp) in house.

679

## 680 Sequencing data processing

681 Genomes were assembled from paired-end Illumina short-reads. Low-quality reads  
682 with an expected average quality below 20 over a 5-bases window, containing  
683 adapters, and shorter than 50 bases were discarded using Trimmomatic (v0.39)<sup>83</sup>.  
684 Reads with phiX sequences were removed using BBduk from the BBtools suite<sup>84</sup>.  
685 Plasmids sequences were reported only when inferred from plasmidSPades  
686 (v3.15.5)<sup>85</sup> and then from Recycler (v0.7)<sup>49</sup>. Reads free from plasmid sequences  
687 were then assembled using SPades (v3.15.5)<sup>50</sup> with default parameters.

688 Contigs above 500 bp in assemblies were kept for quality evaluation. Following the  
689 MlxS specifications<sup>86</sup> as well as requirements from the SeqCode<sup>6</sup>, genomes with a  
690 coverage above 10X, a completion above 90%, and a contamination below 5%  
691 based on CheckM (v1.2.2)<sup>87</sup> estimation using single-copy marker genes were  
692 flagged as high-quality draft genomes. Additional quality flags on the assembly  
693 included: 1) <100 contigs, longest contig >100 kp, and N50 >25 kp using QUAST  
694 (v5.0.2)<sup>88</sup>; 2) detectable 16S and 23S rRNA genes using metaxa2 (v2.2.3)<sup>89</sup>; 3) >18  
695 unique essential tRNAs genes and a detectable 5S rRNA gene sequence using the  
696 annotations of bakta (v1.6.1)<sup>90</sup>.

697 The genome assembly pipeline, including the reconstruction of plasmids and quality  
698 checks steps, is available as a reproducible workflow using Snakemake (v7.9.0)<sup>91</sup>:  
699 [www.github.com/clavellab/genome-assembly](http://www.github.com/clavellab/genome-assembly).

700

## 701 Genome analysis

702 Taxonomic, functional, and ecological analysis for all genomes was conducted using  
703 Protologger (v1.3)<sup>92</sup>. For this, 16S rRNA gene sequences were extracted from each  
704 strain's genome using barrnap v0.9 (default settings)  
705 ([www.github.com/tseemann/barrnap](http://www.github.com/tseemann/barrnap)), and the longest 16S rRNA gene sequence  
706 from each genome was used as input for Protologger. Phylogenomic trees were  
707 generated using PhyloPhlan v3.0.60<sup>93</sup> with proteomes predicted using Prodigal  
708 v2.6.3 (default settings)<sup>94</sup>. Gut metabolic modules presence was based on the  
709 identification of all included KOs within a genome, detected by Kofamscan v1.3.0<sup>95</sup>.

710

## 711 Comparison to published isolate collections

712 Metadata for previously published collection of microbial isolates from the human gut  
713 were obtained from screening literature. To prevent redundancy, we have excluded  
714 articles that are a subset of large collections<sup>25</sup>, and those that lack strain metadata  
715 or genome access<sup>96</sup>. In the case of isolate collections from multiple body sites, the  
716 isolation source was checked to be related to the gastrointestinal tract. Of the 12,565  
717 strains identified, 11,498 had genomes and hence could be analysed. Of these,  
718 10,893 had a high-quality genome and were studied. High-quality was determined by  
719 completion above 90%, and less than 5% contamination based on CheckM (v1.2.2).

720 Strains were determined to be ‘publicly available’ if a genome is available and the  
721 strain has been deposited within a culture collection, *i.e.*, DSMZ, CGMCC etc. The  
722 validity of each published culture accession number was checked to ensure strain  
723 availability. Genomes within the collections were dereplicated at the species level  
724 based on >95% ANI values via FastANI <sup>101</sup>.

725

#### 726 Ecological analysis

727 The occurrence of each strain across 16S rRNA gene amplicon samples was  
728 conducted using Protologger <sup>97</sup>. In essence, 1,000 16S rRNA amplicon samples for  
729 each body site (gut, vagina, skin, lung) were processed by IMNGS <sup>98</sup>. Next, their  
730 operational taxonomic unit (OTU) sequences were compared against the 16S rRNA  
731 gene sequences of the HiBC strains using BLASTN (>97% identity, >80% coverage).  
732 Comparison of HiBC genomes against metagenome assembled genomes (MAGs)  
733 was also conducted using Protologger and its collection of 42,927 MAGs with  
734 curated metadata. The 16S rRNA gene sequences of the Human Microbiome  
735 Projects “most wanted” taxa were obtained from Fodor *et al* (2012) and their priority  
736 taken from the original publication <sup>99</sup>. The most wanted sequences were compared  
737 to the 16S rRNA sequences of the HiBC strains using BLASTN (>97% identity,  
738 >80% coverage) <sup>100</sup>. The prevalence and relative abundance of the HiBC strains  
739 within shotgun metagenomic samples was determined by comparison of the HiBC  
740 strain genomes to the representative genomes from Leviatan *et al* (2022) via  
741 FastANI (>95% ANI) <sup>33,101</sup>. HiBC strains matching to representative genomes were  
742 then connected to the pre-calculated relative abundance of the representative  
743 genomes across 4,623 individuals, published in Leviatan *et al* (2022).

744

#### 745 Plasmid analysis

746 Sequence similarity between the isolates plasmids was determined using MobMess  
747 <sup>15</sup> and clusters were visualised with Cytoscape <sup>102</sup>. Clustering was conducted firstly  
748 based only on the isolate plasmids, and then including the background previously  
749 predicted human gut plasmids. The similarity of each plasmid to known sequences  
750 was determined with the ‘mash dist’ option of PLSDB (v2023\_11\_03\_v2) <sup>54</sup>. The  
751 ecology of each plasmid was inferred from the geographical location assigned to  
752 each matching plasmid. The sequences of pBAC plasmids were rotating using  
753 Rotate, and then aligned with Easyfig <sup>103</sup>. Only regions of similarity with an identity  
754 >90% were studied.

755

#### 756 Pulsed Field Gel Electrophoresis

757 Cultures were grown overnight in 5 mL BHI media, centrifuged (2 mL for *P. vulgatus*  
758 CLA-AA-H253, 5 mL for *H. microfluidus* CLA-JM-H9), resuspended in 100 µL sterile  
759 PBS and incubated at 45 °C for 10 minutes. Of this, 80 µL was mixed with 320 µL

760 1% agarose (Bio-Rad Certified Megabase Agarose) at 45 °C, 80 µL transferred to  
761 PFGE plug moulds (Bio-Rad CHEF Disposable Plug Molds, 50-Well) and cooled at 4  
762 °C for 30 minutes. Solidified agarose plugs were transferred to lysis solution inside 2  
763 mL microcentrifuge tubes (20 mM Tris-HCl pH 8.8, 500 mM EDTA pH 8, 1% N-  
764 laurylsarcosine, 1 mg/mL Proteinase K) and incubated with shaking at 250 rpm at 45  
765 °C for 4 h. Lysis solution was replaced with fresh lysis solution + RNase (10 µg/mL)  
766 and incubated with shaking at 45 °C overnight. Lysis solution was replaced with  
767 wash solution (25 mM Tris-HCl pH 7.5, 100 mM EDTA pH 8) and incubated with  
768 shaking at 45 °C. Plugs were then transferred to fresh 2 mL microcentrifuge tubes  
769 with wash solution containing phenylmethanesulfonyl fluoride (1 mM) and incubated  
770 with shaking at 45 °C for 30 min, twice. Plugs were then transferred to 1 mL wash  
771 solution at stored at 4 °C until use in PFGE. For analysis of intact genomic DNA,  
772 agarose plugs were subjected to 100 µGy of γ radiation using a <sup>137</sup>Cs source  
773 (Gammacell 1000), to linearize circular chromosomes <sup>104</sup>. PFGE was performed  
774 using a CHEF Mapper apparatus (Bio-Rad). Intact and XbaI-digested DNA  
775 fragments were separated on a 1.2% agarose gel in 0.5 × TBE at 14 °C, with a  
776 gradient voltage of 6 V/cm, linear ramping, an included angle of 120°, initial and  
777 final switch times of 0.64 s and 1 min 13.22 s, respectively, and a run time of 20 h  
778 46 min.

779

#### 780 Southern blot

781 The DNA probe was designed by identifying a 1,428 bp DNA sequence unique to the  
782 pMMCAT\_H258 plasmid sequence. Primers sequences were checked against the  
783 *P. vulgatus* genome sequence using SnapGene ([www.snapgene.com](http://www.snapgene.com)) to confirm  
784 specificity to only plasmid DNA. PCR primers were then designed based on the  
785 regions flanking the unique region. *P. vulgatus* total genomic and plasmid DNA was  
786 extracted using the Mericon DNA Bacteria Plus Kit following the manufacturer's  
787 instructions (Qiagen). This DNA was used as a template to amplify the target DNA  
788 sequence using ThermoFisher Scientific Phusion Hot Start 2 DNA polymerase and  
789 the above primer set. The amplified DNA was visualized by gel electrophoresis to  
790 confirm the expected DNA size and the DNA extracted and purified using  
791 ThermoFisher Scientific GeneJET gel extraction kit.

792

793 The Southern blot gel was stained with 0.5 µg/ml ethidium bromide and visualised,  
794 then acid-nicked in 0.25 M HCl, and subsequently denatured in 1.5 M NaCl, 0.5 M  
795 NaOH. The DNA was transferred onto a GE healthcare Amersham Hybond XL  
796 membrane by vacuum transfer using a Vacugene XL gel blotter (Pharmacia Biotech)  
797 for 1 hour at 40 mBar. The membrane was briefly neutralised in 2x SSPE (20x  
798 SSPE: 3 M NaCl, 230 mM NaH<sub>2</sub>PO<sub>4</sub>, 32 mM EDTA, pH 7.4) and the DNA then  
799 crosslinked with 120 mJ/cm<sup>2</sup> UV. The membrane was pre-hybridised for 4 h at 65°C  
800 in 6x SSPE, 1% SDS, 5x Denhardt's solution (100x Denhardt's solution: 2% Ficoll  
801 400, 2 % polyvinyl pyrrolidone 360, 2% bovine serum albumin), 200 µg/ml salmon  
802 sperm DNA (Roche, boiled). The DNA probe used 50 ng of DNA template and was  
803 radiolabelled using 0.74 MBq of [α-<sup>32</sup>P] dCTP (Perkin Elmer) and HiPrime random

804 priming mix (Roche), then purified on a BioRad P-30 column. The membrane was  
805 hybridised with the radiolabelled DNA probe overnight at 65 °C in 6x SSPE, 1%  
806 SDS, 5% dextran sulphate, 200 µg/ml salmon sperm DNA (Roche, boiled). The  
807 membrane was then washed twice for 30 min at 65 °C in 2x SSPE, 0.5 % SDS, and  
808 twice for 30 min at 65 °C in 0.2x SSPE, 0.5% SDS, before being exposed to a  
809 phosphorimager screen (Fujifilm) for 24 h and then scanned on a GE Healthcare  
810 Typhoon.

811

#### 812 Plasmid extraction and visualisation

813 Frozen cryostocks were plated on mGAM plates and incubated anaerobically at 37  
814 °C for 48 h. A single colony was inoculated into 5 mL of BHI media within a Hungate  
815 tube and incubated for 48 h. 2 mL of culture was transferred into 2 mL  
816 microcentrifuge tubes and centrifuged at 18,000 rcf for 5 min. The medium was  
817 removed and the pellet processed using QIAprep Spin Miniprep kit, following the  
818 manufacturer's recommendations (Qiagen, Cat. No. 27106). Depending on DNA  
819 concentration, 10-20 µL of extracted DNA was run on a 0.5 % agarose gel (Sigma-  
820 Aldrich A9539) at 80 V for 120 min (Bio-Rad Biometra) along with GeneRuler 1 kb  
821 DNA ladder (ThermoFisherScientific SM0312) and stained with a dye (Midori Green  
822 Advance, BulldogBio). The gel was imaged with Bio-Rad GelDoc Go imaging  
823 system.

824

#### 825 Crystal violet staining to quantify strain adhesion

826 Strains of interest were tested in triplicates. For each, one colony was picked and  
827 grown to saturation overnight in BHI. Cultures were diluted back to an OD<sub>600nm</sub> of  
828 0.1 in fresh BHI, 3 ml were placed per well in 6-well plates (Nunc cell-culture,  
829 treated, flat clear polystyrene plates, Thermo Scientific) and the plates were  
830 incubated in static conditions at 37 °C. For each timepoint, a plate containing the two  
831 strains to be compared were removed from the anaerobic cultivation chamber, rinsed  
832 with water, and dried at 60 °C for 2 h. Each well was incubated with 1 % (w/v) crystal  
833 violet solution then rinsed 3 times with deionised water and air dried. Once all plates  
834 were collected, wells were destained by adding 1 ml of a 1:4 acetone:ethanol  
835 solution. After gentle mixing, at least three technical replicates of 200 µL per well  
836 were placed in wells of a 96-well plate. The optical density of crystal violet staining  
837 present in the destaining solution was measured at 590 nm. Wells filled with 200 µL  
838 of pure destaining solution were used as blank reference. All wells from the same  
839 biological replicate at a given time point were averaged to provide a single point and  
840 used for statistical analysis. Cell counting in Neubauer chambers ensured an initial  
841 cell concentration varying by less than 5 %, and OD<sub>600nm</sub> measurements in the  
842 supernatant at all timepoints of incubation showed similar OD values over all  
843 biological replicates, confirming that growth rates were not driving the observed 2-  
844 fold difference in attached biomass after 30 h. Representative images of attachment

845 were obtained after 24 h following the above protocol until the first drying stage. Two  
846 wells were imaged in their centre using phase contrast (Nikon Ti-E inverted  
847 microscope, 40x objective lens).

848

#### 849 Protein modelling

850 Protein sequences were entered into the AlphaFoldServer<sup>105</sup>. Pairwise structure  
851 alignment was conducted using the RCSB PDB webserver with the TM-align method  
852 <sup>106,107</sup>.

853

#### 854 Website design

855 Access to taxonomic, cultivation, and genomic information about the HiBC strains is  
856 available at: [www.hibc.rwth-aachen.de](http://www.hibc.rwth-aachen.de). This website was created using the Shiny  
857 package from R and the code is available at: [https://git.rwth-](https://git.rwth-aachen.de/clavellab/hibc)  
858 [aachen.de/clavellab/hibc](https://git.rwth-aachen.de/clavellab/hibc). The 16S rRNA gene sequence and genome for each strain  
859 can be downloaded directly from this site, both individually for strains of interest via  
860 the research data management platform Coscine, or for the entire collection via the  
861 digital repository Zenodo. Further strain specific metadata provided includes:  
862 isolation conditions, source, and risk group.

863

#### 864 Description of novel taxa

865 The description of novel taxa was based on the analysis provided by Protologger  
866 v1.3<sup>92</sup>, and manually curated into the protologues below. For each isolate, taxonomy  
867 was assigned using the following thresholds: <98.7% 16S rRNA gene sequence  
868 identity (as indication for as-yet undescribed species), <94.5% (undescribed genus),  
869 and <86.5% (undescribed family)<sup>108</sup>. ANI values <95% to separate species<sup>109</sup>;  
870 POCP values <50% for distinct genera<sup>110</sup>. Phylogenomic trees were also considered  
871 to make decisions on genus- and family-level delineation<sup>23</sup>. All novel taxa have been  
872 registered with the SeqCode and will be registered with the ICNP.

873

#### 874 **Description of *Alistipes intestinhominis* sp. nov.**

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

877 The genome size is 3.85 Mbp, G+C percentage is 58.29%, with 99.76%  
878 completeness and 0.96% contamination. Strain CLA-KB-H122 was determined to be  
879 a new species based on 16S rRNA gene analysis, with the closest validly named  
880 match being *Alistipes timonensis* (98.17%). Separation from existing *Alistipes*  
881 species was confirmed by ANI comparison, which gave a value of 91.78% to *A.*  
882 *timonensis*. GTDB-Tk classified strain CLA-KB-H122 as '*Alistipes senegalensis*'.  
883 However, the latter name is as yet not valid. Functional analysis showed the strain

884 has 83 transporters, 17 secretion genes, and predicted utilization of cellulose and  
885 starch along with production of L-glutamate and folate. In total, 395 CAZymes were  
886 identified, with 59 different glycoside hydrolase families and 19 glycoside transferase  
887 families represented. Major ( $\geq 5$  %) cellular fatty acids after 72 h of growth in DSMZ  
888 medium 1611 included 15:0 ISO FAME (19.4 %), 15:0 FAME (18.9 %), 16:0 ISO  
889 FAME (10.6 %), 17:0 ISO 3OH FAME (8.8 %), 17:0 FAME (7.2 %), and 17:0 3OH  
890 FAME (6.9 %). The type strain, CLA-KB-H122<sup>T</sup> (phylum *Bacteroidota*, family  
891 *Rikenellaceae*) (=DSM 118481) (StrainInfo: 10.60712/SI-ID414389.1, genome:  
892 GCA\_040095975.1), was isolated from human faeces.

893

#### 894 **Description of *Bifidobacterium hominis* sp. nov.**

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

897 The genome size is 2.03 Mbp, G+C percentage is 55.98%, with 99.77%  
898 completeness and 0.45% contamination. The closest relative to strain CLA-AA-H311  
899 was *Bifidobacterium pseudocatenulatum* (99.07%) based on 16S rRNA gene  
900 analysis. However, ANI comparison identified strain CLA-AA-H311 as a novel  
901 species within the genus *Bifidobacterium*, with an ANI value of 93.18% against  
902 *B. pseudocatenulatum*. GTDB-Tk classification as 'Bifidobacterium sp002742445'  
903 confirmed the proposition of a novel species. Placement within the genus  
904 *Bifidobacterium* was confirmed by the presence of fructose-6-phosphate  
905 phosphoketolase (KO1621) <sup>111</sup>. Functional analysis showed the strain has 90  
906 transporters, 20 secretion genes, and predicted utilization of starch and production of  
907 propionate, acetate, and folate. In total, 137 CAZymes were identified, with 28  
908 different glycoside hydrolase families and 11 glycoside transferase families  
909 represented. Major ( $\geq 5$  %) cellular fatty acids after 24 h of growth in DSMZ medium  
910 1203a included 16:0 FAME (28.8 %), 18:0 FAME (15.3 %), 18:1 CIS 9 DMA (15.1  
911 %), 18:1 CIS 9 FAME (14.3 %), and 14:0 FAME (7.8 %). The type strain, CLA-AA-  
912 H311<sup>T</sup> (phylum *Actinomycetota*, family *Bifidobacteriaceae*) (=DSM 118068, =LMG  
913 33596) (StrainInfo: 10.60712/SI-ID414317.1, genome: GCA\_040095915.1), was  
914 isolated from human faeces.

915

#### 916 **Description of *Blautia aquisgranensis* sp. nov.**

917 *Blautia aquisgranensis* (a.quis.gra.nen'sis. L. fem. adj. *aquisgranensis*, named after  
918 the German city of Aachen (Latin name *Aquisgranum*) where it was isolated)

919 The genome size is 3.63 Mbp, G+C percentage is 43.76%, with 99.37%  
920 completeness and 0.32% contamination. It includes two plasmids (37,495 bp; 1,036  
921 bp). The closest relative to strain CLA-JM-H16 was *Blautia intestinalis* (96.11%)  
922 based on 16S rRNA gene analysis. Placement of the strain within *Blautia* was  
923 confirmed based on POCP comparison as values above 50% to multiple *Blautia*  
924 species were obtained. However, comparison to the type species of the genus,

925 *Blautia coccoides*, gave a value of 42.36%. This inconsistency was also highlighted  
926 by GTDB-Tk, which classified strain CLA-JM-H16 as 'Blautia\_A sp900764225',  
927 suggesting *Blautia* may require splitting into multiple genera in future. As the  
928 separation of *Blautia* would require detailed analysis, which is outside the scope of  
929 this manuscript, we propose strain CLA-JM-H16 as a novel species within *Blautia*. All  
930 three novel species of *Blautia* described within this manuscript were confirmed to  
931 represent distinct species based on ANI comparison (*B. aquisgranensis* Vs. *B.*  
932 *caccae*, 81.74%; *B. aquisgranensis* Vs. *B. intestinhominis*, 76.95%; *B. caccae* Vs. *B.*  
933 *intestinhominis*, 78.57%). Functional analysis revealed 157 transporters, 17  
934 secretion genes, and predicted utilization of arbutin, salicin, sucrose, starch, and  
935 production of acetate, propionate, folate, L-glutamate, riboflavin, and cobalamin. In  
936 total, 205 CAZymes were identified, with 40 different glycoside hydrolase families  
937 and 12 glycoside transferase families represented. The type strain, CLA-JM-H16<sup>T</sup>  
938 (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 114586, =LMG 33033)  
939 (StrainInfo: 10.60712/SI-ID414368.1, genome: GCA\_040096615.1), was isolated  
940 from human faeces.

941

#### 942 **Description of *Blautia caccae* sp. nov.**

943 *Blautia caccae* (cac'cae. N.L. fem. n. *cacca*, human ordure, faeces; from Gr. fem. n.  
944 *kakkê*, human ordure, faeces; N.L. gen. n. *caccae*, of faeces, referring to the source  
945 of isolate).

946 The genome size is 5.83 Mbp, G+C percentage is 46.73%, with 98.73%  
947 completeness and 0.63% contamination. The closest relative to strain CLA-SR-H028  
948 was *Blautia hominis* (98.66%) based on 16S rRNA gene analysis. ANI comparison  
949 identified CLA-SR-H028 as a novel species within the genus *Blautia*, with all values  
950 being below the species threshold. GTDB-Tk classification as 'Blautia sp001304935'  
951 confirmed the proposition of a novel species within *Blautia*. Functional analysis  
952 showed the strain has 158 transporters, 18 secretion genes, and predicted utilization  
953 of cellobiose, sucrose, starch and production of propionate, acetate, cobalamin, and  
954 folate. In total, 353 CAZymes were identified, with 53 different glycoside hydrolase  
955 families and 15 glycoside transferase families represented. Major ( $\geq 5$  %) cellular  
956 fatty acids after 24 h of growth in DSMZ medium 1203a included 16:0 FAME (20.2  
957 %), 14:0 FAME (19.8 %), 16:0 DMA (17.5 %), 18:1 CIS 11 DMA (6.8 %), and 14:0  
958 DMA (5.9 %). The type strain, CLA-SR-H028<sup>T</sup> (phylum *Bacillota*, family  
959 *Lachnospiraceae*) (=DSM 118556, =LMG 33609) (StrainInfo: 10.60712/SI-  
960 ID414428.1, genome: GCA\_040095955.1), was isolated from human faeces.

961

#### 962 **Description of *Blautia intestinhominis* sp. nov.**

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

965 The genome size is 4.1 Mbp, G+C percentage is 43.49%, with 98.73%  
966 completeness and 0.63% contamination. It includes a single plasmid of 22,629 bp.  
967 The isolate was assigned to the species *Blautia obeum* (98.98%) based on 16S  
968 rRNA gene analysis. However, ANI comparison to *B. obeum* clearly identified this  
969 isolate as being a separate species (84.16%). This was confirmed by GTDB-Tk  
970 classification as 'Blautia\_A sp000436615', recommending the creation of a novel  
971 species. Functional analysis showed the strain has 155 transporters, 18 secretion  
972 genes, and predicted utilization of sucrose and starch, along with production of L-  
973 glutamate, folate, propionate, and cobalamin. In total, 177 CAZymes were identified.  
974 The type strain, CLA-AA-H95<sup>T</sup> (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM  
975 111354, =LMG 33582) (StrainInfo: 10.60712/SI-ID414326.1, genome:  
976 GCA\_040096655.1), was isolated from human faeces.

977

#### 978 **Description of *Coprococcus intestinihominis* sp. nov.**

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

981 The genome size is 3.6 Mbp, G+C percentage is 43.29%, with 98.43%  
982 completeness and 2.52% contamination. A single plasmid of 20,255 bp was  
983 detected. The closest relative to strain CLA-AA-H190 was *Coprococcus catus*  
984 (96.76%) based on 16S rRNA gene analysis. ANI comparison identified CLA-AA-  
985 H190 as a novel species within the genus *Coprococcus*, with an ANI value of  
986 90.25% against the closest relative *C. catus*. GTDB-Tk classification as  
987 'Coprococcus\_A catus\_A' confirmed the proposition of a novel species, but also  
988 suggests that separation of *Coprococcus* into multiple genera could occur in future.  
989 Functional analysis showed the strain has 119 transporters, 18 secretion genes, and  
990 predicted utilization of starch and production of propionate, butyrate, acetate,  
991 cobalamin, and folate. In total, 122 CAZymes were identified, with 19 different  
992 glycoside hydrolase families and 13 glycoside transferase families represented. The  
993 type strain, CLA-AA-H190<sup>T</sup> (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM  
994 114688, =LMG 33015) (StrainInfo: 10.60712/SI-ID414125.1, genome:  
995 GCA\_040096555.1), was isolated from human faeces.

996

#### 997 **Description of *Enterobacter intestinihominis* sp. nov.**

998 *Enterobacter intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, the  
999 intestine; L. masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the  
1000 human intestine).

1001 The genome size is 4.86 Mbp, G+C percentage is 54.88%, with 99.89%  
1002 completeness and 0.12% contamination. It contains two plasmids (4,416 bp; 2,494  
1003 bp). Strain CLA-AC-H004 was determined to be a strain of *Enterobacter*  
1004 *quasihormaechei* (99.80%) based on 16S rRNA gene analysis. Separation from  
1005 existing *Enterobacter* species was confirmed by ANI comparison, which gave a value

1006 of 93.69% to *E. quasihormaechei*. GTDB-Tk classification of strain CLA-AC-H004 as  
1007 'Enterobacter hormaechei\_A' supports the proposal of a novel species. An ANI value  
1008 of 99.01% was obtained when compared to *Enterobacter hormaechei* subsp.  
1009 *hoffmannii*. However, given the separation from *E. hormaechei* we propose it these  
1010 strains represent a separate species and not only a sub-species. ANI comparison  
1011 also highlighted similarity with *Pedobacter himalayensis* (95.89%), which has been  
1012 classified as 'Enterobacter hormaechei\_B' within GTDB. However, this suggests  
1013 reclassification of *Pedobacter* may be required in future. Functional analysis showed  
1014 the strain has 497 transporters, 98 secretion genes, and predicted utilization of  
1015 arbutin, salicin, cellobiose, sucrose, and starch along with production of L-glutamate,  
1016 biotin, riboflavin, acetate, propionate, and folate. In total, 316 CAZymes were  
1017 identified, with 37 different glycoside hydrolase families and 19 glycoside transferase  
1018 families represented. Major ( $\geq 5$  %) cellular fatty acids after 24 h of growth in DSMZ  
1019 medium 1203a included 16:0 FAME (38.3 %), 17:0 CYC FAME (18.6 %), 18:1 CIS  
1020 11 FAME (9.7 %), 14:0 FAME (9.0 %), 19:0 CYCLO CIS 11,12 FAME (9.0 %), and  
1021 14:0 3OH (8.8 %). The type strain, CLA-AC-H004<sup>T</sup> (phylum *Pseudomonadota*, family  
1022 *Enterobacteriaceae*) (=DSM 118557, =LMG 33610) (StrainInfo: 10.60712/SI-  
1023 ID414328.1, genome: GCA\_040096145.1), was isolated from human faeces.

1024

#### 1025 **Description of *Enterocloster hominis* sp. nov.**

1026 *Enterocloster hominis* (ho'mi.nis. L. gen. n. *hominis*, of a human being).

1027 The genome size is 6.52 Mbp, G+C percentage is 50.14%, with 99.16%  
1028 completeness and 2.53% contamination. It includes a single plasmid of 7,635 bp.  
1029 The closest relative to strain CLA-SR-H021 was *Enterocloster aldenensis* (98.44%)  
1030 based on 16S rRNA gene analysis. ANI comparison identified CLA-SR-H021 as a  
1031 novel species within the genus *Enterocloster*, with all values being below the species  
1032 threshold. GTDB-Tk classified CLA-SR-H021 as 'Enterocloster pacaense', a name  
1033 derived from the proposed species 'Lachnoclostridium pacaense'. However, the fact  
1034 that these two names are not valid supports the proposition of a novel species within  
1035 *Enterocloster*. Functional analysis revealed 118 transporters, 14 secretion genes,  
1036 and predicted utilization of cellobiose, starch and production of propionate, acetate,  
1037 and folate. In total, 307 CAZymes were identified, with 39 different glycoside  
1038 hydrolase families and 14 glycoside transferase families represented. Major ( $\geq 5$  %)  
1039 cellular fatty acids after 24 h of growth in DSMZ medium 1203a included 16:0 FAME  
1040 (30.3 %), 16:1 CIS 9 DMA (13.7 %), 16:1 CIS 9 FAME (10.7 %), 14:0 FAME (10.4  
1041 %), 16:0 DMA (9.6 %), 18:1 CIS 11 DMA (5.6 %), and 18:1 CIS 11 FAME (5.0 %).  
1042 The type strain, CLA-SR-H021<sup>T</sup> (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM  
1043 118482, =LMG 33606) (StrainInfo: 10.60712/SI-ID414422.1, genome:  
1044 GCA\_040096035.1), was isolated from human faeces.

1045

#### 1046 **Description of *Faecalibacterium intestinale* sp. nov.**

1047 *Faecalibacterium intestinale* (in.tes.ti.na'le. N.L. neut. adj. *intestinale*, pertaining to  
1048 the intestine, from where the type strain was isolated).

1049 The genome size is 2.97 Mbp, G+C percentage is 56.43%, with 100.0%  
1050 completeness and 0.0% contamination. The isolate was determined to be related to  
1051 *F. prausnitzii* (98.08%) based on 16S rRNA gene analysis. ANI comparison to *F.*  
1052 *prausnitzii* was just below species level assignment (94.46%), and GTDB-Tk  
1053 classification as 'Faecalibacterium prausnitzii\_J' recommended the creation of a  
1054 novel species. Functional analysis showed the strain has 135 transporters, 18  
1055 secretion genes, and predicted utilization of starch and production of L-glutamate,  
1056 riboflavin, and cobalamin. In total, 159 CAZymes were identified, with 27 different  
1057 glycoside hydrolase families and 12 glycoside transferase families represented.  
1058 Production of butyrate ( $4.74 \pm 0.30$  mM) was confirmed for strain CLA-AA-H281  
1059 when grown in YCFA broth (DSMZ Medium No. 1611) in Hungate tubes for 48 h at  
1060 37 °C under anaerobic conditions (6% CO<sub>2</sub> and 4,7% H<sub>2</sub> in N<sub>2</sub>). Major ( $\geq 5$  %) cellular  
1061 fatty acids after 48 h of growth in DSMZ medium 215 included 16:0 FAME (33.2 %),  
1062 18:1 CIS 11 DMA (11.7 %), 18:1 CIS 11 FAME (11.4 %), 14:0 FAME (9.3 %), 16:0  
1063 DMA (8.4 %), and 16:1 CIS 9 FAME (6.7 %). The type strain, CLA-AA-H281<sup>T</sup>  
1064 (phylum *Bacillota*, family *Oscillospiraceae*) (=DSM 116193, =LMG 33027)  
1065 (StrainInfo: 10.60712/SI-ID414306.1, genome: GCA\_040096575.1), was isolated  
1066 from human faeces.

1067

#### 1068 **Description of *Faecalibacterium tardum* sp. nov.**

1069 *Faecalibacterium tardum* (tar'dum. L. neut. adj. *tardum*, pertaining to its slow  
1070 growth).

1071 The genome size is 3.04 Mbp, G+C percentage is 56.28%, with 99.32%  
1072 completeness and 0.0% contamination. It contains one plasmid of 14,735 bp in size,  
1073 which encodes for vancomycin resistance via the *vanY* gene. The isolate was  
1074 determined to be closely related to *Faecalibacterium prausnitzii* (98.70%) based on  
1075 16S rRNA gene analysis. ANI comparison to *F. prausnitzii* was on the border of  
1076 species level assignment (95.05%), however, GTDB-Tk classification as  
1077 'Faecalibacterium prausnitzii\_A' recommended the creation of a novel species.  
1078 Strain CLA-AA-H175 was confirmed to represent a distinct species to  
1079 *Faecalibacterium intestinale* (CLA-AA-H281, =DSM 116193) (ANI: 94.39%) and  
1080 *Faecalibacterium faecis* (CLA-JM-H7B, =DSM 114587) (ANI: 86.46%), also  
1081 described in this paper. Functional analysis showed the strain has 116 transporters,  
1082 18 secretion genes, and predicted utilization of starch, and production of L-  
1083 glutamate. In total, 177 CAZymes were identified, with 28 different glycoside  
1084 hydrolase families and 12 glycoside transferase families represented. Production of  
1085 butyrate ( $0.81 \pm 0.16$  mM) was confirmed for strain CLA-AA-H175 when grown in  
1086 YCFA broth (DSMZ Medium No. 1611) in Hungate tubes for 48 h at 37 °C under  
1087 anaerobic conditions (6% CO<sub>2</sub> and 4,7% H<sub>2</sub> in N<sub>2</sub>). The type strain, CLA-AA-H175<sup>T</sup>

1088 (phylum *Bacillota*, family *Oscillospiraceae*) (=DSM 116192) (StrainInfo: 10.60712/SI-  
1089 ID414281.1, genome: GCA\_040096515.1), was isolated from human faeces.

1090

#### 1091 **Description of *Faecousia* gen. nov.**

1092 *Faecousia* (Faec.ou'si.a. L. fem. n. *faex*, dregs; N.L. fem. n. *ousia*, an essence; N.L.  
1093 fem. n. *Faecousia*, a microbe associated with faeces).

1094 Based on 16S rRNA gene sequence identity, the closest relatives are members of  
1095 the genus *Vescimonas* (*Vescimonas fastidiosa*, 93.48%). POCP analysis against  
1096 *Vescimonas coprocola* (44.75%), the type species of this genus, and *V. fastidiosa*  
1097 (41.06%) confirmed strain CLA-AA-H192 represents a distinct genus to *Vescimonas*.  
1098 GTDB-Tk supported the creation of a novel genus, placing strain CLA-AA-H192  
1099 within the proposed genus of "*Candidatus Faecousia*". The type species of this  
1100 genus is *Faecousia intestinalis*.

1101

#### 1102 **Description of *Faecousia intestinalis* sp. nov.**

1103 *Faecousia intestinalis* (in.tes.ti.na'lis. N.L. fem. adj. *intestinalis*, pertaining to the  
1104 intestines, from where the type strain was isolated).

1105 The genome size is 2.98 Mbp, G+C percentage is 58.44%, with 93.29%  
1106 completeness and 0.0% contamination. It contains two plasmids (23,094 bp; 3,448  
1107 bp). Functional analysis showed the strain has 144 transporters, 26 secretion genes,  
1108 and predicted utilization of starch, and production of acetate, propionate, L-  
1109 glutamate, and folate. In total, 116 CAZymes were identified, with 19 different  
1110 glycoside hydrolase families and 11 glycoside transferase families represented. The  
1111 type strain, CLA-AA-H192<sup>T</sup> (phylum *Bacillota*, family *Oscillospiraceae*) (StrainInfo:  
1112 10.60712/SI-ID414286.1, genome: GCA\_040096185.1), was isolated from human  
1113 faeces.

1114

#### 1115 **Description of *Flavonifractor hominis* sp. nov.**

1116 *Flavonifractor hominis* (ho'mi.nis. L. gen. n. *hominis*, of a human being).

1117 The genome size is 2.94 Mbp, G+C percentage is 58.64%, with 99.33%  
1118 completeness and 0.0% contamination. Strain CLA-AP-H34 was determined to be a  
1119 new species based on 16S rRNA gene sequence analysis, with the closest validly  
1120 named match being *Flavonifractor plautii* (97.21%). Separation from existing  
1121 *Flavonifractor* species was confirmed by ANI comparison, which gave a value of  
1122 82.25% to *F. plautii*. GTDB-Tk classification of strain CLA-AP-H34 as an unknown  
1123 species within *Flavonifractor* supports the proposal of a novel species. Functional  
1124 analysis showed the strain has 112 transporters, 14 secretion genes, and predicted  
1125 utilization of starch and production of L-glutamate, riboflavin, butyrate, and folate. In  
1126 total, 140 CAZymes were identified, with 19 different glycoside hydrolase families

1127 and 15 glycoside transferase families represented. Major ( $\geq 5$  %) cellular fatty acids  
1128 after 72 h of growth in DSMZ medium 1611 included 16:0 DMA (25.2 %), 15:0 FAME  
1129 (15.7 %), 15:0 DMA (11.4 %), 14:0 FAME (8.9 %), 15:0 ISO FAME (8.4 %), 17:0  
1130 DMA (7.5 %), and 16:0 FAME (6.4 %). The type strain, CLA-AP-H34<sup>T</sup> (phylum  
1131 *Bacillota*, family *Oscillospiraceae*) (=DSM 118484, =LMG 33602) (StrainInfo:  
1132 10.60712/SI-ID414347.1, genome: GCA\_040095835.1), was isolated from human  
1133 faeces.

1134

### 1135 **Description of *Hominiventricola aquisgranensis* sp. nov.**

1136 *Hominiventricola aquisgranensis* (a.quis.gra.nen'sis. L. masc. adj. *aquisgranensis*,  
1137 named after the German city of Aachen, where it was isolated).

1138 The genome size is 3.17 Mbp, G+C percentage is 45.01%, with 99.37%  
1139 completeness and 0.27% contamination, including a single plasmid of 7,983 bp. The  
1140 closest relative to strain CLA-AA-H78B was *Hominiventricola filiformis* (96.24%)  
1141 based on 16S rRNA gene analysis. POCP analysis confirmed genus assignment to  
1142 *Hominiventricola*, with a POCP value of 69.06% to the type strain of the only current  
1143 species within this genus, *H. filiformis*. GTDB-Tk classified strain CLA-AA-H78B as  
1144 'Choladocola sp003480725' within "*Candidatus Choladocola*". The latter is a  
1145 heterosynonym of *Hominiventricola*, a validly published name. ANI comparison  
1146 confirmed that strain CLA-AA-H78B represents a novel species as the ANI value to  
1147 *H. filiformis* was 82.98%. Functional analysis showed the strain has 134 transporters,  
1148 31 secretion genes, and predicted utilization of cellobiose, starch, arbutin, salicin,  
1149 and production of L-glutamate, folate, acetate, propionate, and cobalamin. In total,  
1150 134 CAZymes were identified. The type strain, CLA-AA-H78B<sup>T</sup> (phylum *Bacillota*,  
1151 family *Lachnospiraceae*) (=DSM 111355, =LMG 33583) (StrainInfo: 10.60712/SI-  
1152 ID414323.1, genome: GCA\_040096225.1), was isolated from human faeces.

1153

### 1154 **Description of *Lachnospira intestinalis* sp. nov.**

1155 *Lachnospira intestinalis* (in.tes.ti.na'lis. N.L. fem. adj. *intestinalis*, pertaining to the  
1156 intestines, from where the type strain was isolated).

1157 The genome size is 3.1 Mbp, G+C percentage is 41.75%, with 99.33%  
1158 completeness and 0.0% contamination. Strain CLA-AA-H89B was determined to  
1159 represent a separate species to its closest relative, *Lachnospira pectinoschiza*  
1160 (97.48%), based on 16S rRNA gene analysis. This was confirmed based on ANI  
1161 comparison to all close relatives which were below the species threshold. GTDB-Tk  
1162 classification of strain CLA-AA-H89B as 'Lachnospira sp000437735' confirmed that  
1163 this isolate represents a novel species. Functional analysis showed the strain has  
1164 112 transporters, 29 secretion genes, and predicted utilization of starch and  
1165 cellulose, along with production of L-glutamate, folate, acetate, propionate, and  
1166 riboflavin. Motility was predicted based on the presence of the following genes: *FliA*,  
1167 *FliB*, *FlgB*, *FlgC*, *FlgD*, *FlgE*, *FlgF*, *FlgJ*, *FlgK*, *FlgL*, *FliC*, *FliD*, *FliE*, *FliF*, *FliG*, *FliK*,

1168 *FliM*, *FliN*, *MotA*, *MotB*. In total, 162 CAZymes were identified, with 21 different  
1169 glycoside hydrolase families and 11 glycoside transferase families represented.  
1170 Major ( $\geq 5\%$ ) cellular fatty acids after 48 h of growth in DSMZ medium 1611 included  
1171 16:0 FAME (32.4 %), 18:1 CIS 11 DMA (20.6 %), 14:0 FAME (7.8 %), 16:0 DMA (6.3  
1172 %), 18:1 CIS 11 aldehyde (6.2 %), and 16:1 CIS 9 DMA (5.5 %). The type strain,  
1173 CLA-AA-H89B<sup>T</sup> (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 118070)  
1174 (StrainInfo: 10.60712/SI-ID414325.1, genome: GCA\_040095895.1), was isolated  
1175 from human faeces.

1176

#### 1177 **Description of *Lachnospira hominis* sp. nov.**

1178 *Lachnospira hominis* (ho'mi.nis. L. gen. masc. n. *hominis*, of a human being).

1179 The genome size is 3.15 Mbp, G+C percentage is 37.05%, with 98.66%  
1180 completeness and 0.67% contamination. Strain CLA-JM-H10 was determined to  
1181 represent a separate species to its closest relative, *Lachnospira rogosae* (96.37%),  
1182 based on 16S rRNA gene analysis. This was confirmed by an ANI of 79.9% between  
1183 *L. rogosae* (CLA-AA-H255), and strain CLA-JM-H10. GTDB-Tk classification of CLA-  
1184 JM-H10 as 'Lachnospira sp900316325' confirmed that it represents a novel species.  
1185 Functional analysis showed the strain has 109 transporters, 27 secretion genes, and  
1186 predicted utilization of starch, and production of L-glutamate, folate, and cobalamin.  
1187 Motility was predicted based on detection of the following genes: *FliA*, *FliB*, *FlgB*,  
1188 *FlgC*, *FlgD*, *FlgE*, *FlgF*, *FlgJ*, *FlgK*, *FlgL*, *FliC*, *FliD*, *FliE*, *FliF*, *FliG*, *FliK*, *FliM*, *FliN*,  
1189 *MotA*, *MotB*. This is consistent with the type species of this genus being motile. In  
1190 total, 165 CAZymes were identified, with 22 different glycoside hydrolase families  
1191 and 12 glycoside transferase families represented. The type strain, CLA-JM-H10<sup>T</sup>  
1192 (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 114599, =LMG 33585)  
1193 (StrainInfo: 10.60712/SI-ID414366.1, genome: GCA\_040096395.1), was isolated  
1194 from human faeces.

1195

#### 1196 **Description of *Lachnospira rogosae* sp. nov.**

1197 *Lachnospira rogosae* (ro.go'sae. N.L. gen. masc. n. *rogosae*, of Rogosa).

1198 Strain CLA-AA-H255 was determined to be similar to *Lactobacillus rogosae*  
1199 (99.87%) based on 16S rRNA gene analysis. However, the lack of a genome for the  
1200 type strain of the latter species, along with the lack of the type strain at any  
1201 established culture collection prevented further comparison (Tindall, 2014). GTDB-Tk  
1202 classification of CLA-AA-H255 as 'Lachnospira rogosae\_A' suggested that it  
1203 represents a species within a genus distantly related to *Lactobacillus*. This was  
1204 reconfirmed by the same placement of a second strain, CLA-AA-H191, which was  
1205 also assigned to the same placeholder by GTDB-Tk (ANI of 98.86% between our two  
1206 isolates). Based on these results, we propose *L. rogosae* was previously  
1207 misassigned as a member of the genus *Lactobacillus*. To provide a type strain, and  
1208 correct its placement, we propose the creation of the species *Lachnospira rogosae*.

1209 Functional analysis showed that strain CLA-AA-H255 has 100 transporters, 25  
1210 secretion genes, and predicted utilization of starch, and production of L-glutamate,  
1211 folate, and riboflavin. The prediction of motility and acetate production based on  
1212 genomic analysis is consistent with the observed phenotype of motility in the original  
1213 type strain of *L. ramosae* as stated by Holdeman and Moore (Holdeman and Moore,  
1214 1974). In total, 142 CAZymes were identified, including 20 different glycoside  
1215 hydrolase families and 12 glycoside transferase families. Ecological analysis of  
1216 1,000 human gut 16S rRNA gene amplicon samples identified this strain in 1.20% of  
1217 samples with a relative abundance of  $0.43 \pm 0.78\%$ . The type strain, CLA-AA-H255<sup>T</sup>  
1218 (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 118602, =LMG 33594)  
1219 (StrainInfo: 10.60712/SI-ID414300.1, genome: GCA\_040096455.1), was isolated  
1220 from human faeces.

1221

#### 1222 **Description of *Laedolimicola intestinhominis* sp. nov.**

1223 *Laedolimicola intestinhominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, intestine;  
1224 L. masc. n. *homo*, a human being; N.L. gen. n. *intestinhominis*, of the human  
1225 intestine).

1226 The genome size is 3.45 Mbp, G+C percentage is 49.65%, with 99.37%  
1227 completeness and 0.16% contamination. It includes a single plasmid of 5,131 bp.  
1228 Strain CLA-AA-H132 was determined to represent a separate species to its closest  
1229 relative, *Laedolimicola ammoniilytica* (98.38%), based on 16S rRNA gene analysis.  
1230 POCP analysis confirmed that strain CLA-AA-H132 belongs to the recently named  
1231 genus, *Laedolimicola*, with a POCP value of 75.19% to the type strain of the only  
1232 current species within this genus, *L. ammoniilytica*. GTDB-Tk placement as  
1233 'Merdisoma sp900553635' suggests that *Laedolimicola* and *Candidatus Merdisoma*  
1234 are homonyms and require future reclassification. ANI comparison confirmed that  
1235 CLA-AA-H132 represents a novel species, as the ANI value to *L. ammoniilytica* was  
1236 only 90.2%. Functional analysis showed the strain has 142 transporters, 21 secretion  
1237 genes, and predicted utilization of cellobiose, sucrose, starch, and production of  
1238 acetate, propionate, L-glutamate, cobalamin, and folate. In total, 153 CAZymes were  
1239 identified, with 21 different glycoside hydrolase families and 16 glycoside transferase  
1240 families represented. Major ( $\geq 5\%$ ) cellular fatty acids after 24 h of growth in DSMZ  
1241 medium 339 included 16:0 FAME (24.8%), 18:1 CIS 11 DMA (12.8%), 14:0 FAME  
1242 (10.0%), 16:0 DMA (8.2%), and 16:0 CIS 9 DMA (8.2%). The type strain, CLA-AA-  
1243 H132<sup>T</sup> (phylum *Bacillota*, family *Lachnospiraceae*) (=DSM 117481, =LMG 33588)  
1244 (StrainInfo: 10.60712/SI-ID414271.1, genome: GCA\_040096155.1), was isolated  
1245 from human faeces.

1246

#### 1247 **Description of *Maccoyibacter* gen. nov.**

1248 *Maccoyibacter* (Mac.co.y.i.bac'ter N.L. fem. gen. n. *Maccoyae*, referring to the  
1249 immunologist Kathy McCoy; N.L. masc. n. *bacter*, a rod; N.L. masc.

1250 n. *Maccoyibacter*, a rod-shaped bacterium named after the immunologist Kathy  
1251 McCoy, for her many scientific contributions in the field of microbe-host interactions).

1252 Based on 16S rRNA gene analysis, the closest species with a valid name are  
1253 *Roseburia hominis* (94.9%) and *Eubacterium oxidoreducens* (94.23%). POCP values  
1254 against all close relatives were below the genus delineation threshold of 50%,  
1255 including many *Roseburia* spp. (*R. hominis*, *R. faecis*, *R. porci*, *R. intestinalis*), apart  
1256 from *R. inulinivorans* (50.83%). GTDB-Tk placement as “UBA11774 sp003507655”  
1257 confirmed that strain CLA-AA-H185 represents a novel genus within the family  
1258 *Lachnospiraceae*. The type species is *Maccoyibacter intestinihominis*.

1259

#### 1260 **Description of *Maccoyibacter intestinihominis* sp. nov.**

1261 *Maccoyibacter intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, intestine;  
1262 L. masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the human  
1263 intestine).

1264 The genome size is 2.85 Mbp, G+C percentage is 41.17%, with 97.41%  
1265 completeness and 0.45% contamination. It contains a single plasmid (2,341 bp).  
1266 Functional analysis showed the strain has 99 transporters, 35 secretion genes, and  
1267 predicted utilization of starch, and production of acetate, propionate, L-glutamate,  
1268 cobalamin, folate, and riboflavin. In total, 110 CAZymes were identified, with 17  
1269 different glycoside hydrolase families and 13 glycoside transferase families  
1270 represented. The type strain, CLA-AA-H185<sup>T</sup> (phylum *Bacillota*, family  
1271 *Lachnospiraceae*) (=DSM 118601) (StrainInfo: 10.60712/SI-ID414284.1, genome:  
1272 GCA\_040096355.1), was isolated from human faeces.

1273

#### 1274 **Description of *Megasphaera intestinihominis* sp. nov.**

1275 *Megasphaera intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. gen. neut. n. *intestini*, of the  
1276 intestine; L. gen. masc. n. *hominis*, of a human being; N.L. gen. masc.  
1277 n. *intestinihominis*, of the human intestine).

1278 The genome size is 2.38 Mbp, G+C percentage is 53.59%, with 100.00%  
1279 completeness and 0.00% contamination. The closest relative to strain CLA-AA-H81  
1280 was *Megasphaera indica* (99.09%) based on 16S rRNA gene analysis. ANI  
1281 comparison to all close relatives were below species assignment threshold (highest  
1282 to *Megasphaera elsdenii*, 90.78%). GTDB-Tk classification as ‘*Megasphaera*  
1283 sp000417505’ supports the creation of a novel species. Functional analysis showed  
1284 the strain has 133 transporters, 16 secretion genes, and predicted utilization of  
1285 starch, and production of butyrate, propionate, L-glutamate, folate, riboflavin, and  
1286 cobalamin. Ecological analysis identified 152 matching (MASH distance < 0.05)  
1287 MAGs, of which 124 originate from the human gut, suggesting this species is most  
1288 commonly observed within this environment. This is supported by ecological analysis  
1289 using 16S rRNA gene amplicon datasets which identified it within 19.0% of 1,000

1290 human gut samples, with a relative abundance of  $2.14 \pm 5.23\%$ . Major ( $\geq 5\%$ )  
1291 cellular fatty acids after 24 h of growth in DSMZ medium 215 included 12:0 FAME  
1292 (17.1 %), 14:0 3OH (14.8 %), 18:1 CIS 9 FAME (9.6 %), 18:1 CIS 9 DMA (9.3 %),  
1293 16:0 FAME (8.0 %), and 16:1 CIS 7 FAME (5.0 %). The type strain, CLA-AA-H81<sup>T</sup>  
1294 (phylum *Bacillota*, family *Veillonellaceae*) (=DSM 118069, =LMG 33597) (StrainInfo:  
1295 10.60712/SI-ID414324.1, genome: GCA\_040096415.1), was isolated from human  
1296 faeces.

1297

#### 1298 **Description of *Niallia hominis* sp. nov.**

1299 *Niallia hominis* (ho'mi.nis. L. gen. n. *hominis*, of a human being).

1300 The genome size is 4.9 Mbp, G+C percentage is 35.33%, with 92.24%  
1301 completeness and 3.20% contamination. Strain CLA-SR-H024 was assigned to  
1302 *Niallia circulans* (100%) based on 16S rRNA gene analysis. However, ANI  
1303 suggested the isolate represents a novel species as comparison to all *Niallia* spp.  
1304 were below 80%. GTDB-Tk classification of strain CLA-SR-H024 as '*Niallia*  
1305 sp001076885' confirmed that it represents a novel species. Functional analysis  
1306 showed the strain has 244 transporters, 43 secretion genes, and predicted utilization  
1307 of arbutin, salicin, cellobiose, starch, and dextran, along with production of L-  
1308 glutamate, folate, and cobalamin. In total, 294 CAZymes were identified, with 34  
1309 different glycoside hydrolase families and 14 glycoside transferase families  
1310 represented. Major ( $\geq 5\%$ ) cellular fatty acids after 72 h of growth in DSMZ medium  
1311 1203a included 15:0 ANTEISO FAME (29.7 %), 16:0 FAME (16.8 %), 15:0 ISO  
1312 FAME (14.9 %), 16:0 ISO FAME (9.8 %), 14:0 FAME (8.4 %), 17:0 ANTEISO FAME  
1313 (6.9 %), and 14:0 ISO FAME (5.2 %). The type strain, CLA-SR-H024<sup>T</sup> (phylum  
1314 *Bacillota*, family *Bacillaceae*) (=DSM 118483) (StrainInfo: 10.60712/SI-ID414424.1,  
1315 genome: GCA\_040095995.1), was isolated from human faeces.

1316

#### 1317 **Description of *Peptoniphilus hominis* sp. nov.**

1318 *Peptoniphilus hominis* (ho'mi.nis. L. gen. n. *hominis*, of a human being).

1319 The genome size is 1.99 Mbp, G+C percentage is 33.8%, with 99.3% completeness  
1320 and 0.0% contamination. Strain CLA-SR-H025 was assigned to *Peptoniphilus*  
1321 *gorbachii* (99.11%) based on 16S rRNA gene analysis. However, ANI suggested the  
1322 isolate represents a novel species as comparison to *P. gorbachii* gave a value of  
1323 88.92%. GTDB-Tk classification of strain CLA-SR-H025 as '*Peptoniphilus\_A*  
1324 *grossensis*' confirmed that it represents a novel species, as the name '*Peptoniphilus*  
1325 *grossensis*' was proposed but never validated<sup>112</sup>. Functional analysis showed the  
1326 strain has 12 transporters, 1 secretion genes, and no gut metabolic models were  
1327 identified within the genome. In total, 59 CAZymes were identified, with 10 different  
1328 glycoside hydrolase families and 8 glycoside transferase families represented. The  
1329 type strain, CLA-SR-H025<sup>T</sup> (phylum *Bacillota*, family *Peptoniphilaceae*) (=DSM

1330 118555, =LMG 33608) (Straininfo: 10.60712/SI-ID414425.1, genome:  
1331 GCA\_040096015.1), was isolated from human faeces.

1332

1333 **Description of *Pseudoflavonifractor intestinihominis* sp. nov.**

1334 *Pseudoflavonifractor intestinihominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*, the  
1335 intestine; L. masc. n. *homo*, a human being; N.L. gen. n. *intestinihominis*, of the  
1336 human intestine).

1337 The genome size is 3.69 Mbp, G+C percentage is 61.54%, with 99.33%  
1338 completeness and 0.00% contamination. Strain CLA-AP-H29 was assigned to the  
1339 species *Pseudoflavonifractor capillosus* (99.53%) based on 16S rRNA gene  
1340 sequence analysis. However, ANI suggested the isolate represents a novel species  
1341 as comparison to *P. capillosus* gave a value of 80.96%. GTDB-Tk classified strain  
1342 CLA-AP-H29 as 'Pseudoflavonifractor sp944387275', which confirms that it  
1343 represents a novel species. Functional analysis predicted the strain has 121  
1344 transporters, 14 secretion genes, and predicted utilization of starch, and production  
1345 of L-glutamate, folate, and cobalamin. In total, 180 CAZymes were identified, with 25  
1346 different glycoside hydrolase families and 15 glycoside transferase families  
1347 represented. Major ( $\geq 5$  %) cellular fatty acids after 48 h of growth in DSMZ medium  
1348 215 included 14:0 FAME (33.9 %), 16:0 DMA (33.8 %), 16:0 FAME (9.8 %), and 16:0  
1349 ALDE (8.0 %). The type strain, CLA-AP-H29<sup>T</sup> (phylum *Bacillota*, family  
1350 *Oscillospiraceae*) (=DSM 118073, =LMG 33601) (Straininfo: 10.60712/SI-  
1351 ID414342.1, genome: GCA\_040096055.1), was isolated from human faeces.

1352

1353 **Description of *Robertmurraya yapensis* sp. nov.**

1354 *Robertmurraya yapensis* (yap'ensis. N.L. fem. adj. *yapensis*, pertaining to Yap  
1355 trench, which is the geographical position where the first isolate of this species was  
1356 obtained).

1357 The genome size is 4.74 Mbp, G+C percentage is 37.94%, with 98.85%  
1358 completeness and 2.13% contamination. It contains a plasmid (2,470 bp). The  
1359 closest relative to strain CLA-AA-H227 was *Robertmurraya spiralis* (99.23%) based  
1360 on 16S rRNA gene analysis. The highest POCP scores were to members of the  
1361 genus *Robertmurraya* (*Robertmurraya kyonggiensis*, 84.9%; *R. spiralis*, 65.33%).  
1362 The placement of strain CLA-AA-H227 within *Robertmurraya* was confirmed by  
1363 GTDB-Tk assignment as 'Robertmurraya yapensis', a reclassification of the species  
1364 'Bacillus yapensis', although neither name has been validated. ANI comparison  
1365 confirmed that CLA-AA-H227 represents a novel species, as values to close  
1366 relatives were below the species level threshold. Functional analysis showed the  
1367 strain has 246 transporters, 53 secretion genes, and predicted utilization of arbutin,  
1368 salicin, cellobiose, starch, and production of butyrate, acetate, propionate, folate,  
1369 riboflavin, and cobalamin. In total, 238 CAZymes were identified. The type strain,  
1370 CLA-AA-H227<sup>T</sup> (phylum *Bacillota*, family *Bacillaceae*) (=DSM 113004, =LMG 33018)

1371 (Straininfo: 10.60712/SI-ID414150.1, genome: GCA\_040096375.1), was isolated  
1372 from human faeces.

1373

#### 1374 **Description of *Ruminococcoides intestinale* sp. nov.**

1375 *Ruminococcoides intestinale* (in.tes.ti.na'le. N.L. neut. adj. *intestinalis*, pertaining to  
1376 the intestines, from where the type strain was isolated).

1377 The genome size is 2.32 Mbp, G+C percentage is 40.88%, with 99.33%  
1378 completeness and 1.01% contamination. The isolate was determined to be similar to  
1379 *Ruminococcus bromii* (98.91%) and more distantly related to *Ruminococcoides bili*  
1380 (96.76%) based on 16S rRNA gene analysis. While POCP comparison of strain  
1381 CLA-JM-H38 to *R. bromii* was 59.79%, and 53.56% to *Ruminococcus bovis*,  
1382 suggesting they belong to the same genus, all other comparisons to *Ruminococcus*  
1383 species were below 50%, including to the type species, *Ruminococcus flavefaciens*  
1384 (27.58%). POCP to *R. bili*, the type species of the genus *Ruminococcoides*, was  
1385 64.44%. GTDB-Tk classified strain CLA-JM-H38 as “*Ruminococcus\_E bromii\_B*”,  
1386 confirming it is not a member of the genus *Ruminococcus*. These results support  
1387 GTDB assignment that both *R. bovis* and *R. bromii* should be reclassified as  
1388 members of the genus *Ruminococcoides*. Strain CLA-JM-H38 was confirmed to  
1389 represent a novel species as all ANI comparisons to close relatives were below 95%,  
1390 and it represents a distinct novel species from *Ruminococcoides intestinhominis*  
1391 described in this work (78.33%). Functional analysis showed the strain has 81  
1392 transporters, 15 secretion genes, and predicted utilization of starch, and production  
1393 of L-glutamate. In total, 108 CAZymes were identified, with 15 different glycoside  
1394 hydrolase families and 12 glycoside transferase families represented. Ecological  
1395 analysis based on 16S rRNA gene amplicons identified this species in 55.20% of  
1396 1,000 human gut samples with a relative abundance of  $1.50 \pm 2.49\%$ , suggesting it is  
1397 a prevalent and dominant bacterial species within the human gut. Major ( $\geq 5\%$ )  
1398 cellular fatty acids after 72 h of growth in DSMZ medium 1611 included 16:0 ISO  
1399 FAME (49.5 %) and 16:0 ISO DMA (23.8 %). The type strain, CLA-JM-H38<sup>T</sup> (phylum  
1400 *Bacillota*, family *Oscillospiraceae*) (=DSM 118486, =LMG 33604) (StrainInfo:  
1401 10.60712/SI-ID414376.1, genome: GCA\_040096305.1), was isolated from human  
1402 faeces.

1403

#### 1404 **Description of *Ruminococcoides intestinhominis* sp. nov.**

1405 *Ruminococcoides intestinhominis* (in.tes.ti.ni.ho'mi.nis. L. neut. n. *intestinum*,  
1406 intestine; L. masc. n. *homo*, a human being; N.L. gen. n. *intestinhominis*, of the  
1407 human intestine).

1408 The genome size is 2.26 Mbp, G+C percentage is 34.26%, with 98.66%  
1409 completeness and 0.00% contamination. It contains one plasmid (1,825 bp). Based  
1410 on 16S rRNA gene sequence identity, the isolate was closely related to  
1411 *Ruminococcus bovis* (99.3%) and more distant to *Ruminococcoides bili* (94.8%), the

1412 type species of this genus. ANI comparison confirmed the similarity of strain CLA-  
1413 AA-H171 to *R. bovis* (95.4%), however classification by GTDB-Tk as  
1414 'Ruminococcus\_E sp934476515' supported the creation of a novel species.  
1415 Recently, 'Ruminococcus\_E' has been validly named as *Ruminococcoides*, with the  
1416 type species *R. bili*<sup>34</sup>. POCP comparison between the isolate and *R. bili* provided a  
1417 value of 51.3%, suggesting that strain CLA-AA-H171 represents a novel species  
1418 within the genus *Ruminococcoides*. Functional analysis showed the strain has 75  
1419 transporters, 14 secretion genes, and predicted utilization of starch, and production  
1420 of acetate. In total, 124 CAZymes were identified, with 13 different glycoside  
1421 hydrolase families and 12 glycoside transferase families represented. Ecological  
1422 analysis based on 16S rRNA gene amplicons identified this species in 10.40% of  
1423 1,000 human gut samples with a relative abundance of  $0.23 \pm 0.68\%$ . The type  
1424 strain, CLA-AA-H171<sup>T</sup> (phylum *Bacillota*, family *Oscillospiraceae*) (=DSM 114689,  
1425 =LMG 33587) (StrainInfo: 10.60712/SI-ID414279.1, genome: GCA\_040096285.1),  
1426 was isolated from human faeces.

1427

#### 1428 **Description of *Ruthenibacterium intestinale* sp. nov.**

1429 *Ruthenibacterium intestinale* (*in.tes.ti.na'le*. *N.L. neut. adj. intestinale*, pertaining to  
1430 the intestines, from where the type strain was isolated).

1431 The genome size is 3.1 Mbp, G+C percentage is 54.69%, with 98.3% completeness  
1432 and 0.00% contamination. It contains a plasmid (5,063 bp). The closest relative to  
1433 strain CLA-JM-H11 is *Ruthenibacterium lactatiformans* (94.93%), the type species of  
1434 this genus, based on 16S rRNA gene analysis. Placement of the isolate within the  
1435 genus *Ruthenibacterium* was confirmed based on POCP comparison, with a value of  
1436 56.41% to the type species. Strain CLA-JM-H11 was confirmed to be distinct to *R.*  
1437 *lactatiformans* based on ANI comparison (79.28%). GTDB-Tk classification  
1438 confirmed this assignment as an unknown species within *Ruthenibacterium*.  
1439 Functional analysis showed the strain has 143 transporters, 16 secretion genes, and  
1440 predicted utilization of starch, and production of acetate, propionate, folate, and  
1441 cobalamin. In total, 155 CAZymes were identified, with 27 different glycoside  
1442 hydrolase families and 13 glycoside transferase families represented. The type  
1443 strain, CLA-JM-H11<sup>T</sup> (phylum *Bacillota*, family *Oscillospiraceae*) (=DSM 114604,  
1444 =LMG 33032) (StrainInfo: 10.60712/SI-ID414367.1, genome: GCA\_040096265.1),  
1445 was isolated from human faeces.

1446

#### 1447 **Description of *Solibaculum intestinale* sp. nov.**

1448 *Solibaculum intestinale* (*in.tes.ti.na'le*. *N.L. neut. adj. intestinale*, pertaining to the  
1449 intestines, from where the type strain was isolated).

1450 The genome size is 2.81 Mbp, G+C percentage is 54.62 %, with 97.99%  
1451 completeness and 0.67% contamination. The isolate was determined to be a new  
1452 species based on 16S rRNA gene analysis, with the closest validly named match

1453 being *Solibaculum mannosilyticum* (94.93%). Placement within *Solibaculum* was  
1454 confirmed with POCP values above 50% to *S. mannosilyticum* (55.7%), the type  
1455 species, and only member of this genus. ANI comparison confirmed the isolate  
1456 represents a novel species, as no ANI values were above 80%. Functional analysis  
1457 showed the strain has 82 transporters, 16 secretion genes, and predicted utilization  
1458 of starch, and production of acetate, propionate, and L-glutamate. In total, 169  
1459 CAZymes were identified, with 20 different glycoside hydrolase families and 15  
1460 glycoside transferase families represented. The type strain, CLA-JM-H44<sup>T</sup> (phylum  
1461 *Bacillota*, family *Oscillospiraceae*) (=DSM 114601, =LMG 33034) (StrainInfo:  
1462 10.60712/SI-ID414377.1, genome: GCA\_040096205.1), was isolated from human  
1463 faeces.

1464

#### 1465 **Description of *Ventrimonas* gen. nov.**

1466 *Ventrimonas* (Ven.tri.mo'nas. L. masc. n. *venter*, the belly; L. fem. n. *monas*, a  
1467 monad; N.L. fem. n. *Ventrimonas*, a microbe associated with the belly  
1468 (intestines/faeces)).

1469 Based on 16S rRNA gene sequence identity, the closest relatives are members of  
1470 the genus *Hungatella* (*Hungatella effluvii*, 95.15%; *Hungatella hathewayi*, 94.95%).  
1471 POCP analysis against *H. effluvii* (40.38%) and *H. hathewayi* (39.86%) indicates that  
1472 strain CLA-AP-H27 represents a distinct genus to *Hungatella*. GTDB-Tk supported  
1473 the creation of a novel genus, placing strain CLA-AP-H27 within the proposed genus  
1474 of "*Candidatus Ventrimonas*" in the family *Lachnospiraceae*. The type species of this  
1475 genus is *Ventrimonas faecis*.

1476

#### 1477 **Description of *Ventrimonas faecis* sp. nov.**

1478 *Ventrimonas faecis* (fae'cis. L. gen. n. *faecis*, of dregs, pertaining to faeces, from  
1479 where the type strain was isolated).

1480 The genome size is 3.41 Mbp, G+C percentage is 49.27%, with 98.73%  
1481 completeness and 0.63% contamination. Functional analysis showed the strain has  
1482 156 transporters, 21 secretion genes, and predicted utilization of starch, and  
1483 production of acetate, propionate, L-glutamate, cobalamin, and folate. In total, 139  
1484 CAZymes were identified, with 21 different glycoside hydrolase families and 12  
1485 glycoside transferase families represented. Major ( $\geq 10$  %) cellular fatty acids after 24  
1486 h of growth in DSMZ medium 1203a included 16:0 FAME (35.5 %), 18:1 CIS 11  
1487 DMA (10.0 %), 18:1 CIS 11 FAME (9.6 %), 18:1 CIS 9 DMA (5.9 %), and 16:0 DMA  
1488 (5.5 %). The type strain, CLA-AP-H27<sup>T</sup> (phylum *Bacillota*, family *Lachnospiraceae*)  
1489 (=DSM 118072, =LMG 33600) (StrainInfo: 10.60712/SI-ID414341.1, genome:  
1490 GCA\_040096075.1), was isolated from human faeces.

1491

#### 1492 **Description of *Waltera hominis* sp. nov.**

1493 *Waltera hominis* (ho'mi.nis. L. gen. masc. n. *hominis*, of a human being, pertaining to  
1494 the human gut habitat, from where the type strain was isolated).

1495 The genome size is 3.88 Mbp, G+C percentage is 45.72%, with 99.52%  
1496 completeness and 2.13% contamination. The isolate was assigned to the species  
1497 *Waltera intestinalis* (100.0%) based on 16S rRNA gene analysis. However, ANI  
1498 comparison identified strain CLA-AA-H183 as a novel species within the genus  
1499 *Waltera*, with an ANI value of 91.97% against the type species *Waltera intestinalis*.  
1500 GTDB-Tk currently lacks the inclusion of *Waltera*, causing misclassification as  
1501 'Acetatifactor sp003447295', however this confirmed the proposition of a novel  
1502 species. Separation of *Waltera* from *Acetatifactor* was revalidated via both  
1503 phylogenomic analysis (**Supplementary Figure 3**), which shows both genera form  
1504 separate monophyletic groups, and POCP analysis, which shows clear similarity  
1505 within each genus (*Waltera*: 68.86%, *Acetatifactor*: 59.25%), and separation  
1506 between the genera (39.94 ± 1.63 %). Functional analysis showed the strain has 141  
1507 transporters, 36 secretion genes, and predicted utilization of starch, cellulose, and  
1508 production of butyrate, propionate, acetate, and folate. In total, 228 CAZymes were  
1509 identified, with 38 different glycoside hydrolase families and 14 glycoside transferase  
1510 families represented. The type strain, CLA-AA-H183<sup>T</sup> (phylum *Bacillota*, family  
1511 *Lachnospiraceae*) (=DSM 114684, =LMG 33586) (StrainInfo: 10.60712/SI-  
1512 ID414283.1, genome: GCA\_040096245.1), was isolated from human faeces.

1513

#### 1514 **Data availability**

1515 The genomes for all strains have been deposited at NCBI under BioProject:  
1516 PRJNA996881. Bulk download of HiBC resources is possible via Zenodo for the  
1517 genomes (<https://doi.org/10.5281/zenodo.12755497>), plasmid sequences  
1518 (<https://doi.org/10.5281/zenodo.12187897>), 16S rRNA gene sequences  
1519 (<https://doi.org/10.5281/zenodo.12180259>) and the isolates metadata  
1520 (<https://doi.org/10.5281/zenodo.14592301>). The PacBio genome for *P. vulgatus*  
1521 CLA-AA-H253 (=DSM 118718) has been deposited at ENA (PRJEB80480) and is  
1522 transiently available via Zenodo: <https://doi.org/10.5281/zenodo.14674027>. The code  
1523 for the creation of the HiBC website has been made available at; [https://git.rwth-  
1524 aachen.de/clavellab/hibc](https://git.rwth-aachen.de/clavellab/hibc).

1525

#### 1526 **Acknowledgments**

1527 Sequencing was performed with the support of: (i) the DFG-funded NGS  
1528 Competence Center Tübingen (INST 37/1049-1) and the Institute for Medical  
1529 Microbiology and Hygiene at the University Hospital (Tübingen, Germany), including  
1530 help by the Quantitative Biology Center (QBiC) for raw data management and  
1531 storage; (ii) Nassos Typas and Carlos Geert Pieter Voogdt (EMBL, Heidelberg), for  
1532 long-read sequencing; (iii) the Genomics Facility, a core facility of the  
1533 Interdisciplinary Center for Clinical Research (IZKF) Aachen within the Faculty of

1534 Medicine at RWTH Aachen University. We are also thankful to: (iv) Patrick Buchta  
1535 from the Audio-Visual department at the University Hospital of RWTH Aachen for  
1536 designing the HiBC logo; (v) Marzena Wyschkon and Meina Neumann-Schaal from  
1537 the Leibniz Institute DSMZ for their help with the deposition of strains and CFA  
1538 analysis, respectively; (vi) Catherine Gonzalez and Maurice Heizer from the  
1539 Functional Microbiome Research Group (Institute of Medical Microbiology; University  
1540 Hospital of RWTH Aachen) for data management with Coscine and for the curation  
1541 of HiBC, respectively; (vii) Peter Vandamme, Claudine Vereecke, and Anneleen  
1542 Wieme for handling the deposition of strains at the BCCM/LMG Bacteria Collection.

1543

## 1544 **Funding**

1545 TCAH received funding from the German Research Foundation (DFG) as part of  
1546 SFB1382 Gut-liver axis and from the RWTH Aachen START program, project  
1547 “LeakyGut”. TC received funding from the German Research Foundation (DFG):  
1548 project no. 513892404, no. 445552570, no. 395357507 – SFB1371 Microbiome  
1549 signatures, no. 403224013 – SFB1382 Gut-liver axis, and no. 460129525 –  
1550 NFDI4Microbiota. JO received funding from the DFG, project no. 6270054 NFDI 28/1  
1551 “NFDI4Microbiota” and 6270048 NFDI 5/1 “NFDI4Biodiversity”, the BMBF, projects  
1552 no. 8005512901 and 8005512001 of DZIF, and EU/Horizon IRA project MICROBE  
1553 no. 101094353. MDC and MG received funding from the DFG, project no.  
1554 403224013. The data used in this publication was managed using the research data  
1555 management platform Coscine with storage space granted by the Research Data  
1556 Storage (RDS) of the DFG and Ministry of Culture and Science of the State of North  
1557 Rhine-Westphalia (DFG: INST222/1261-1 and MKW: 214-4.06.05.08 - 139057). LM  
1558 and TA received funding from The Leverhulme Trust: project no. RF-2023-286\2.

1559

## 1560 **Conflicts of interest**

1561 The authors have no conflicts of interest

1562

## 1563 **References**

- 1564 1. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for  
1565 improved metagenomic analyses. *Nat Biotechnol* **37**, 186–192 (2019).
- 1566 2. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the  
1567 industrialized human microbiome. *Cell* **184**, 2053-2067.e18 (2021).
- 1568 3. Liu, C. *et al.* Enlightening the taxonomy darkness of human gut microbiomes  
1569 with a cultured biobank. *Microbiome* **9**, 1–29 (2021).
- 1570 4. Lagier, J. *et al.* Culture of previously uncultured members of the human gut  
1571 microbiota by culturomics. **1**, (2016).
- 1572 5. Browne, H. P. *et al.* Culturing of ‘unculturable’ human microbiota reveals novel  
1573 taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
- 1574 6. Hedlund, B. P. *et al.* SeqCode: a nomenclatural code for prokaryotes  
1575 described from sequence data. *Nat Microbiol* **7**, 1702–1708 (2022).

- 1576 7. Göker, M., Moore, E. R. B., Oren, A. & Trujillo, M. E. Status of the SeqCode in  
1577 the International Journal of Systematic and Evolutionary Microbiology. *Int J*  
1578 *Syst Evol Microbiol* **72**, 10–12 (2022).
- 1579 8. Arahall, D. *et al.* The best of both worlds: a proposal for further integration of  
1580 Candidatus names into the International Code of Nomenclature of  
1581 Prokaryotes. *Int J Syst Evol Microbiol* **74**, 1–21 (2024).
- 1582 9. Hitch, T. C. A. *et al.* Harmonious naming across nomenclature codes  
1583 exemplified by the description of bacterial isolates from the mammalian gut.  
1584 *Syst Appl Microbiol* **47**, (2024).
- 1585 10. Blanco-Míguez, A. *et al.* Extension of the Segatella copri complex to 13  
1586 species with distinct large extrachromosomal elements and associations with  
1587 host conditions. *Cell Host Microbe* **31**, 1804-1819.e9 (2023).
- 1588 11. De Filippis, F. *et al.* Distinct Genetic and Functional Traits of Human Intestinal  
1589 Prevotella copri Strains Are Associated with Different Habitual Diets. *Cell Host*  
1590 *Microbe* **25**, 444-453.e3 (2019).
- 1591 12. Fehlner-Peach, H. *et al.* Distinct Polysaccharide Utilization Profiles of Human  
1592 Intestinal Prevotella copri Isolates. *Cell Host Microbe* **26**, 680-690.e5 (2019).
- 1593 13. Contevelle, L. C. & Vicente, A. C. P. A plasmid network from the gut  
1594 microbiome of semi-isolated human groups reveals unique and shared  
1595 metabolic and virulence traits. *Sci Rep* **12**, 1–9 (2022).
- 1596 14. Fogarty, E. C. *et al.* Article A cryptic plasmid is among the most numerous  
1597 genetic elements in the human gut II Article A cryptic plasmid is among the  
1598 most numerous genetic elements in the human gut. *Cell* **187**, 1206-1222.e16  
1599 (2024).
- 1600 15. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their  
1601 ecology across human gut metagenomes revealed by PlasX and MobMess.  
1602 *Nat Microbiol* **9**, 830–847 (2024).
- 1603 16. Perez, M. *et al.* A synthetic consortium of 100 gut commensals modulates the  
1604 composition and function in a colon model of the microbiome of elderly  
1605 subjects. *Gut Microbes* **13**, 1–19 (2021).
- 1606 17. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and  
1607 butyrate production. *Nat Commun* **12**, 1–16 (2021).
- 1608 18. Cheng, A. G. *et al.* Design, construction, and in vivo augmentation of a  
1609 complex gut microbiome. *Cell* **185**, 3617-3636.e19 (2022).
- 1610 19. Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. Akkermansia  
1611 muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium.  
1612 *Int J Syst Evol Microbiol* **54**, 1469–1476 (2004).
- 1613 20. Plovier, H. *et al.* A purified membrane protein from Akkermansia muciniphila or  
1614 the pasteurized bacterium improves metabolism in obese and diabetic mice.  
1615 *Nat Med* **23**, 107–113 (2017).
- 1616 21. Overmann, J. Significance and future role of microbial resource centers. *Syst*  
1617 *Appl Microbiol* **38**, 258–265 (2015).
- 1618 22. Parker, C. T., Tindall, B. J. & Garrity, G. M. International code of nomenclature  
1619 of Prokaryotes. *Int J Syst Evol Microbiol* **69**, S1 (2019).
- 1620 23. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a  
1621 new method for improved phylogenetic and taxonomic placement of microbes.  
1622 *Nat Commun* **4**, 1–10 (2013).
- 1623 24. Hitch, T. C. A. *et al.* Recent advances in culture-based gut microbiome  
1624 research. *International Journal of Medical Microbiology* **311**, (2021).

- 1625 25. Browne, H. P. *et al.* Culturing of ‘unculturable’ human microbiota reveals novel  
1626 taxa and extensive sporulation. *Nature* **533**, 543–546 (2016).
- 1627 26. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for  
1628 improved metagenomic analyses. *Nat Biotechnol* **37**, 186–192 (2019).
- 1629 27. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal  
1630 multiomics data enables mechanistic microbiome research. *Nat Med* **25**,  
1631 1442–1452 (2019).
- 1632 28. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the  
1633 industrialized human microbiome. *Cell* **184**, 2053-2067.e18 (2021).
- 1634 29. Liu, C. *et al.* Enlightening the Taxonomy Darkness of Human Gut Microbiomes  
1635 With Cultured Biobank. *Microbiome* **9**, (2021).
- 1636 30. Huang, P. *et al.* Gut microbial genomes with paired isolates from China  
1637 illustrate probiotic and cardiometabolic effects. *Cell Genomics* **4**, (2024).
- 1638 31. Huang, Y. *et al.* High-throughput microbial culturomics using automation and  
1639 machine learning. *Nat Biotechnol* (2023) doi:10.1038/s41587-023-01674-2.
- 1640 32. Lagier, J. C. *et al.* Culture of previously uncultured members of the human gut  
1641 microbiota by culturomics. *Nat Microbiol* **1**, (2016).
- 1642 33. Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M. & Segal, E. An  
1643 expanded reference map of the human gut microbiome reveals hundreds of  
1644 previously unknown species. *Nat Commun* **13**, 1–14 (2022).
- 1645 34. Molinero, N. *et al.* *Ruminococcoides bili* gen. Nov., sp. nov., a bile-resistant  
1646 bacterium from human bile with autolytic behavior. *Int J Syst Evol Microbiol* **71**,  
1647 1–11 (2021).
- 1648 35. Schmitz, M. A., Dimonaco, N. J., Clavel, T. & Hitch, T. C. A. Lineage-specific  
1649 microbial protein prediction enables large-scale exploration of protein ecology  
1650 within the human gut. *bioRxiv* (2024) doi:10.1101/2024.05.29.596415.
- 1651 36. Ma, L. *et al.* Spermidine improves gut barrier integrity and gut microbiota  
1652 function in diet-induced obese mice. *Gut Microbes* **12**, 1–19 (2020).
- 1653 37. Wiedmann, S., Eudy, J. D. & Zemleni, J. Biotin supplementation increases  
1654 expression of genes encoding interferon- $\gamma$ , interleukin-1 $\beta$ , and 3-  
1655 methylcrotonyl-CoA carboxylase, and decreases expression of the gene  
1656 encoding interleukin-4 in human peripheral blood mononuclear cells. *Journal of*  
1657 *Nutrition* **133**, 716–719 (2003).
- 1658 38. Agrawal, S., Agrawal, A. & Said, H. M. Biotin deficiency enhances the  
1659 inflammatory response of human dendritic cells. *Am J Physiol Cell Physiol*  
1660 **311**, C386–C391 (2016).
- 1661 39. Skupsky, J. *et al.* Biotin Supplementation Ameliorates Murine Colitis by  
1662 Preventing NF- $\kappa$ B Activation. *Cmgh* **9**, 557–567 (2020).
- 1663 40. Mao, B. *et al.* *Blautia producta* displays potential probiotic properties against  
1664 dextran sulfate sodium-induced colitis in mice. *Food Science and Human*  
1665 *Wellness* **13**, 709–720 (2024).
- 1666 41. van der Lelie, D. *et al.* Rationally designed bacterial consortia to treat chronic  
1667 immune-mediated colitis and restore intestinal homeostasis. *Nat Commun* **12**,  
1668 1–17 (2021).
- 1669 42. Lee, C. *et al.* P926 *Blautia Obeum* Aggravates Colitis in a Murine Model. *J*  
1670 *Crohns Colitis* **17**, i1035–i1035 (2023).
- 1671 43. Huang, T. Q. *et al.* Bergenin Alleviates Ulcerative Colitis By Decreasing Gut  
1672 Commensal *Bacteroides vulgatus*-Mediated Elevated Branched-Chain Amino  
1673 Acids. *J Agric Food Chem* **72**, 3606–3621 (2024).

- 1674 44. Sugihara, K. *et al.* Dietary phosphate exacerbates intestinal inflammation in  
1675 experimental colitis. *J Clin Biochem Nutr* **61**, 91–99 (2017).
- 1676 45. Kolachala, V. L. *et al.* A2B Adenosine Receptor Gene Deletion Attenuates  
1677 Murine Colitis. *Gastroenterology* **135**, 861–870 (2008).
- 1678 46. Jones, B. V, Sun, F. & Marchesi, J. R. Comparative metagenomic analysis of  
1679 plasmid encoded functions in the human gut microbiome. *BMC Genomics* **11**,  
1680 (2010).
- 1681 47. Groussin, M. *et al.* Elevated rates of horizontal gene transfer in the  
1682 industrialized human microbiome. *Cell* **184**, 2053-2067.e18 (2021).
- 1683 48. Sanders, J. G. *et al.* A low-cost genomics workflow enables isolate screening  
1684 and strain-level analyses within microbiomes. *Genome Biol* **23**, (2022).
- 1685 49. Rozov, R. *et al.* Recycler: An algorithm for detecting plasmids from de novo  
1686 assembly graphs. *Bioinformatics* **33**, 475–482 (2017).
- 1687 50. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its  
1688 applications to single-cell sequencing. *Journal of Computational Biology* **19**,  
1689 455–477 (2012).
- 1690 51. Li, Y. *et al.* Distribution of megaplasmids in *Lactobacillus salivarius* and other  
1691 *Lactobacilli*. *J Bacteriol* **189**, 6128–6139 (2007).
- 1692 52. Bottacini, F. *et al.* Discovery of a conjugative megaplasmid in *Bifidobacterium*  
1693 *breve*. *Appl Environ Microbiol* **81**, 166–176 (2015).
- 1694 53. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their  
1695 ecology across human gut metagenomes revealed by PlasX and MobMess.  
1696 *Nat Microbiol* **9**, 830–847 (2024).
- 1697 54. Schmartz, G. P. *et al.* PLSDB: Advancing a comprehensive database of  
1698 bacterial plasmids. *Nucleic Acids Res* **50**, D273–D278 (2022).
- 1699 55. Novicki, T. J. & Hecht, D. W. *Characterization and DNA Sequence of the*  
1700 *Mobilization Region of PLV22a from Bacteroides Fragilis*. *JOURNAL OF*  
1701 *BACTERIOLOGY* vol. 177 (1995).
- 1702 56. Nieto, C. *et al.* The yefM-yoeB toxin-antitoxin systems of *Escherichia coli* and  
1703 *Streptococcus pneumoniae*: Functional and structural correlation. *J Bacteriol*  
1704 **189**, 1266–1278 (2007).
- 1705 57. Smith, J. A. & Magnuson, R. D. Modular Organization of the Phd  
1706 Repressor/Antitoxin Protein. *J Bacteriol* **186**, 2692–2698 (2004).
- 1707 58. Evans, J. C. *et al.* A proteolytically activated antimicrobial toxin encoded on a  
1708 mobile plasmid of Bacteroidales induces a protective response. *Nat Commun*  
1709 **13**, (2022).
- 1710 59. García-Bayona, L., Coyne, M. J. & Comstock, L. E. Mobile Type VI secretion  
1711 system loci of the gut Bacteroidales display extensive intra-ecosystem transfer,  
1712 multi-species spread and geographical clustering. *PLoS Genet* **17**, 1–25  
1713 (2021).
- 1714 60. García-Bayona, L. *et al.* A pervasive large conjugative plasmid mediates  
1715 multispecies biofilm formation in the intestinal microbiota increasing resilience  
1716 to perturbations. Preprint at <https://doi.org/10.1101/2024.04.29.590671> (2024).
- 1717 61. Sokurenko, E. V, Chesnokova, V., Doyle, R. J. & Hasty, D. L. *Diversity of the*  
1718 *Escherichia Coli Type 1 Fimbrial Lectin DIFFERENTIAL BINDING TO*  
1719 *MANNOSIDES AND UROEPITHELIAL CELLS\**. and the iResearch Service  
1720 <http://www.jbc.org>.
- 1721 62. Jønsson, R. *et al.* Aggregative adherence fimbriae form compact structures as  
1722 seen by SAXS. *Sci Rep* **13**, (2023).

- 1723 63. García-Bayona, L. *et al.* A pervasive large conjugative plasmid mediates  
1724 multispecies biofilm formation in the intestinal microbiota increasing resilience  
1725 to perturbations. *bioRxiv* 2024.04.29.590671 (2024).
- 1726 64. Riepe, S. P., Goldstein, J. & Alpers, D. H. Effect of secreted *Bacteroides*  
1727 proteases on human intestinal brush border hydrolases. *Journal of Clinical*  
1728 *Investigation* **66**, 314–322 (1980).
- 1729 65. Mills, R. H. *et al.* Multi-omics analyses of the ulcerative colitis gut microbiome  
1730 link *Bacteroides vulgatus* proteases with disease severity. *Nat Microbiol* 1–15  
1731 (2022) doi:10.1038/s41564-021-01050-3.
- 1732 66. Sokol, H. *et al.* *Faecalibacterium prausnitzii* is an anti-inflammatory commensal  
1733 bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc*  
1734 *Natl Acad Sci U S A* **105**, 16731–16736 (2008).
- 1735 67. Ueda, A. *et al.* Identification of *Faecalibacterium prausnitzii* strains for gut  
1736 microbiome-based intervention in Alzheimer's-type dementia. *Cell Rep Med* **2**,  
1737 100398 (2021).
- 1738 68. D'hoë, K. *et al.* Integrated culturing, modeling and transcriptomics uncovers  
1739 complex interactions and emergent behavior in a three-species synthetic gut  
1740 community. *Elife* **7**, 1–30 (2018).
- 1741 69. Fitzgerald, C. B. *et al.* Comparative analysis of *Faecalibacterium prausnitzii*  
1742 genomes shows a high level of genome plasticity and warrants separation into  
1743 new species-level taxa. *BMC Genomics* **19**, 1–20 (2018).
- 1744 70. Vital, M., Karch, A. & Pieper, D. H. Colonic Butyrate-Producing Communities in  
1745 Humans: an Overview Using Omics Data. *mSystems* **2**, 1–18 (2017).
- 1746 71. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human  
1747 colonic microbiota. *Environmental Microbiology* vol. 19 29–41 Preprint at  
1748 <https://doi.org/10.1111/1462-2920.13589> (2017).
- 1749 72. Panicker, I. S., Browning, G. F. & Markham, P. F. The effect of an alternate  
1750 start codon on heterologous expression of a PhoA fusion protein in  
1751 *Mycoplasma gallisepticum*. *PLoS One* **10**, 1–10 (2015).
- 1752 73. Duncan, S. H., Barcenilla, A., Stewart, C. S., Pryde, S. E. & Flint, H. J. Acetate  
1753 utilization and butyryl coenzyme A (CoA): Acetate-CoA transferase in butyrate-  
1754 producing bacteria from the human large intestine. *Appl Environ Microbiol* **68**,  
1755 5186–5190 (2002).
- 1756 74. Stackebrandt, E. Diversification and focusing: Strategies of microbial culture  
1757 collections. *Trends Microbiol* **18**, 283–287 (2010).
- 1758 75. Afrizal, A. *et al.* Enhanced cultured diversity of the mouse gut microbiota  
1759 enables custom-made synthetic communities. *Cell Host Microbe* 1–16 (2022)  
1760 doi:10.1016/j.chom.2022.09.011.
- 1761 76. Wylensek, D. *et al.* A collection of bacterial isolates from the pig intestine  
1762 reveals functional and taxonomic diversity. *Nat Commun in press*, 1–26  
1763 (2020).
- 1764 77. Stackebrandt, E. *et al.* Deposit of microbial strains in public service collections  
1765 as part of the publication process to underpin good practice in science.  
1766 *Springerplus* **3**, 1–4 (2014).
- 1767 78. Burz, S. D. *et al.* A Guide for Ex Vivo Handling and Storage of Stool Samples  
1768 Intended for Fecal Microbiota Transplantation. *Sci Rep* **9**, 1–16 (2019).
- 1769 79. Afrizal, A. *et al.* Anaerobic single-cell dispensing facilitates the cultivation of  
1770 human gut bacteria. *Environ Microbiol* **00**, (2022).

- 1771 80. Vieira, S. *et al.* Usitatibacter rugosus gen. nov., sp. nov. and Usitatibacter  
1772 palustris sp. nov., novel members of Usitatibacteraceae fam. nov. within the  
1773 order Nitrosomonadales isolated from soil. *Int J Syst Evol Microbiol* (2021).  
1774 81. Spitzer, V. Structure analysis of fatty acids by gas chromatography - Low  
1775 resolution electron impact mass spectrometry of their 4,4-dimethyloxazoline  
1776 derivatives - A review. *Prog Lipid Res* **35**, 387–408 (1996).  
1777 82. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular  
1778 microbial diversity of an anaerobic digester as determined by small-subunit  
1779 rDNA sequence analysis. *Appl Environ Microbiol* **63**, 2802–2813 (1997).  
1780 83. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for  
1781 Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).  
1782 84. Bushnell, B. BMap: A Fast, Accurate, Splice-Aware Aligner. in *9th Annual*  
1783 *Genomics of Energy & Environment Meeting* (2014).  
1784 85. Antipov, D. *et al.* PlasmidSPAdes: Assembling plasmids from whole genome  
1785 sequencing data. *Bioinformatics* **32**, 3380–3387 (2016).  
1786 86. Yilmaz, P. *et al.* Minimum information about a marker gene sequence  
1787 (MIMARKS) and minimum information about any (x) sequence (MIxS)  
1788 specifications. *Nat Biotechnol* **29**, 415–420 (2011).  
1789 87. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W.  
1790 CheckM: assessing the quality of microbial genomes recovered from. *Cold*  
1791 *Spring Harbor Laboratory Press Method* **1**, 1–31 (2015).  
1792 88. Mikheenko, A., Prijibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile  
1793 genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150  
1794 (2018).  
1795 89. Bengtsson-Palme, J. *et al.* metaxa2: Improved identification and taxonomic  
1796 classification of small and large subunit rRNA in metagenomic data. *Mol Ecol*  
1797 *Resour* **15**, 1403–1414 (2015).  
1798 90. Schwengers, O. *et al.* Bakta: Rapid and standardized annotation of bacterial  
1799 genomes via alignment-free sequence identification. *Microb Genom* **7**, (2021).  
1800 91. Mölder, F. *et al.* Sustainable data analysis with Snakemake [version 2; peer  
1801 review: 2 approved]. *F1000Res* **10**, 1–29 (2021).  
1802 92. Hitch, T. C. A. *et al.* Automated analysis of genomic sequences facilitates high-  
1803 throughput and comprehensive description of bacteria. *ISME Communications*  
1804 **1**, (2021).  
1805 93. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and  
1806 genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**, 1–10  
1807 (2020).  
1808 94. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation  
1809 site identification. *BMC Bioinformatics* **11**, 119 (2010).  
1810 95. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile  
1811 HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).  
1812 96. Nelson, K. E. *et al.* A catalog of reference genomes from the human  
1813 microbiome. *Science (1979)* **328**, 994–999 (2010).  
1814 97. Hitch, T. C. A. *et al.* Automated analysis of genomic sequences facilitates high-  
1815 throughput and comprehensive description of bacteria. *ISME Communications*  
1816 **1**, (2021).  
1817 98. Lagkouvardos, I. *et al.* IMNGS: A comprehensive open resource of processed  
1818 16S rRNA microbial profiles for ecology and diversity studies. *Sci Rep* **6**, 1–9  
1819 (2016).

- 1820 99. Fodor, A. A. *et al.* The ‘most wanted’ taxa from the human microbiome for  
1821 whole genome sequencing. *PLoS One* **7**, (2012).
- 1822 100. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local  
1823 alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
- 1824 101. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S.  
1825 High throughput ANI analysis of 90K prokaryotic genomes reveals clear  
1826 species boundaries. *Nat Commun* **9**, 5114 (2018).
- 1827 102. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of  
1828 biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
- 1829 103. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: A genome comparison  
1830 visualizer. *Bioinformatics* **27**, 1009–1010 (2011).
- 1831 104. Beverley, S. M. Estimation of circular DNA size using  $\gamma$ -irradiation and pulsed-  
1832 field gel electrophoresis. *Anal Biochem* **177**, 110–114 (1989).
- 1833 105. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold.  
1834 *Nature* **596**, 583–589 (2021).
- 1835 106. Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K. & Rose, Y. RCSB protein  
1836 Data Bank: exploring protein 3D similarities via comprehensive structural  
1837 alignments. *Bioinformatics* **40**, (2024).
- 1838 107. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm  
1839 based on the TM-score. *Nucleic Acids Res* **33**, 2302–2309 (2005).
- 1840 108. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria  
1841 and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**, 635–645  
1842 (2014).
- 1843 109. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S.  
1844 High throughput ANI analysis of 90K prokaryotic genomes reveals clear  
1845 species boundaries. *Nat Commun* **9**, 5114 (2018).
- 1846 110. Qin, Q. L. *et al.* A proposed genus boundary for the prokaryotes based on  
1847 genomic insights. *J Bacteriol* **196**, 2210–2215 (2014).
- 1848 111. Meile, L., Rohr, L. M., Geissmann, T. A., Herensperger, M. & Teuber, M.  
1849 Characterization of the D-xylulose 5-phosphate/D-fructose 6-phosphate  
1850 phosphoketolase gene (*xfp*) from *Bifidobacterium lactis*. *J Bacteriol* **183**, 2929–  
1851 2936 (2001).
- 1852 112. Mishra, A. K., Hugon, P., Robert, C., Raoult, D. & Fournier, P. E. Non  
1853 contiguous-finished genome sequence and description of *Peptoniphilus*  
1854 *grossensis* sp. nov. *Stand Genomic Sci* **7**, 320–330 (2012).
- 1855  
1856

1857 **Supplementary Tables**

1858 **Supplementary Table 1:** List of existing isolate collections from the human gut. For  
1859 each isolate collection we detail the accessibility of the published strains. Additional  
1860 sheets provide specific information on strain collections

1861

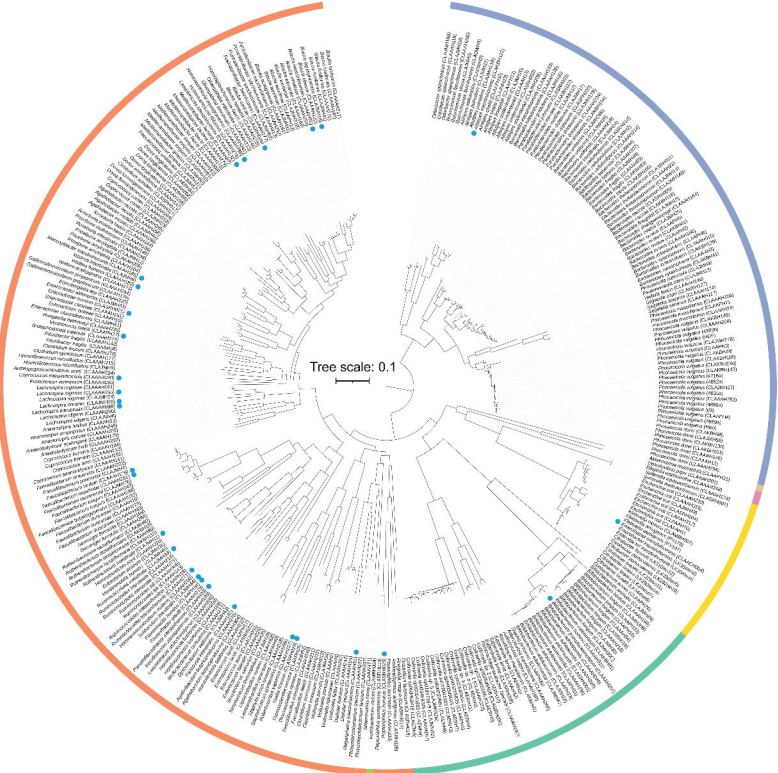
1862 **Supplementary Table 2:** HPLC measurement of the production and/or utilisation of  
1863 multiple metabolites.

1864

1865 **Supplementary Table 3:** Statistical comparison of genome sizes of phyla present  
1866 within the HiBC, UHGG (isolates), and complete UHGG.

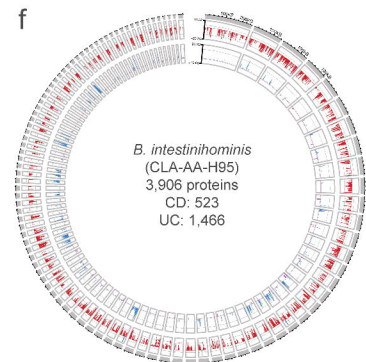
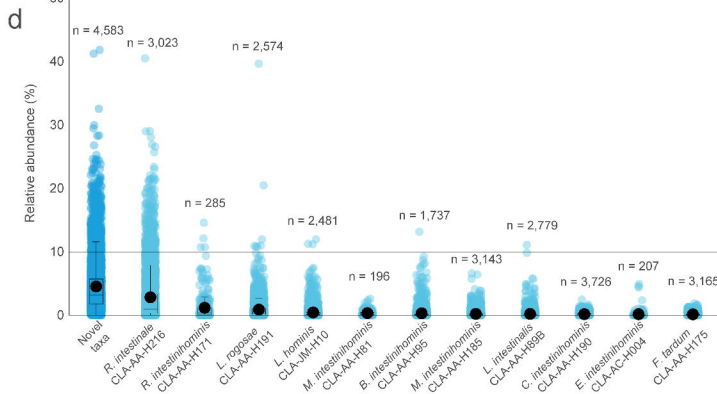
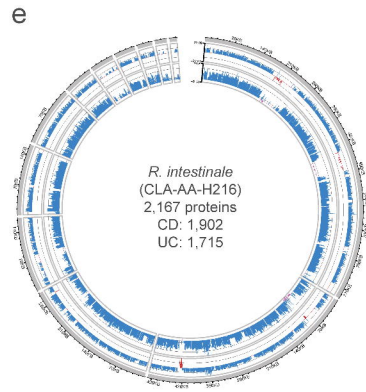
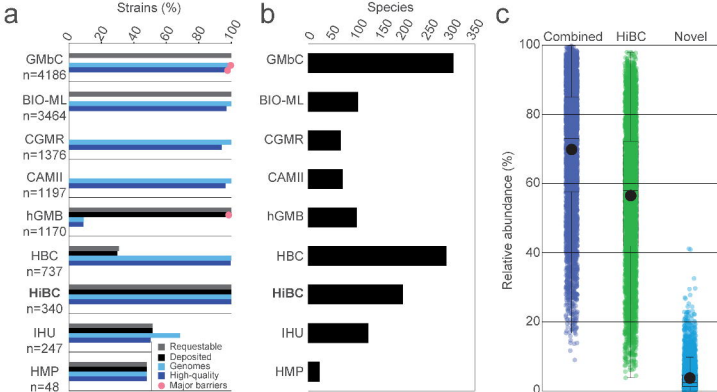
1867

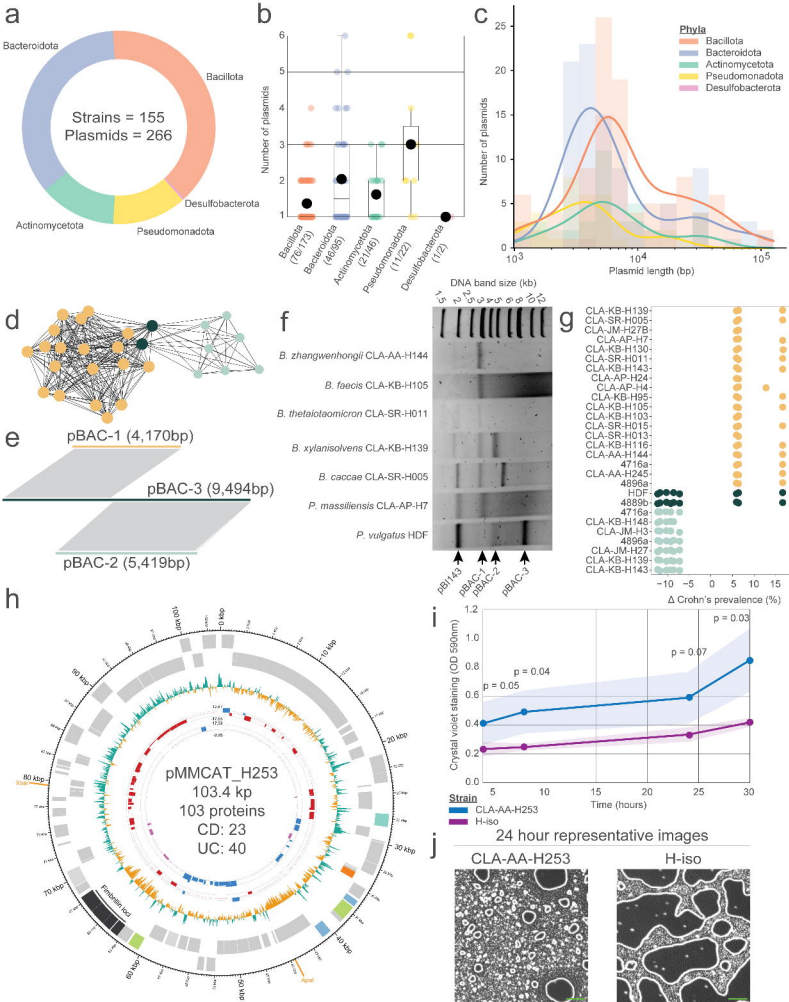
1868 **Supplementary Table 4:** Statistical comparison of the normalised carbohydrate  
1869 active enzyme investment of phyla within the HiBC.



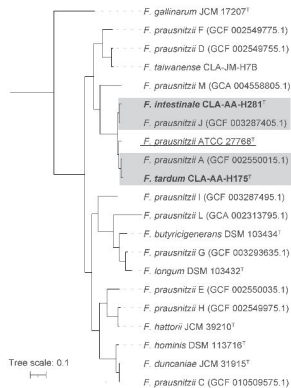
## Phyla

- Bacteroidota
- Bacillota
- Desulfobacterota
- Fusobacteriota
- Pseudomonadota
- Actinomycetota
- Verrucomicrobiota
- Novel taxa

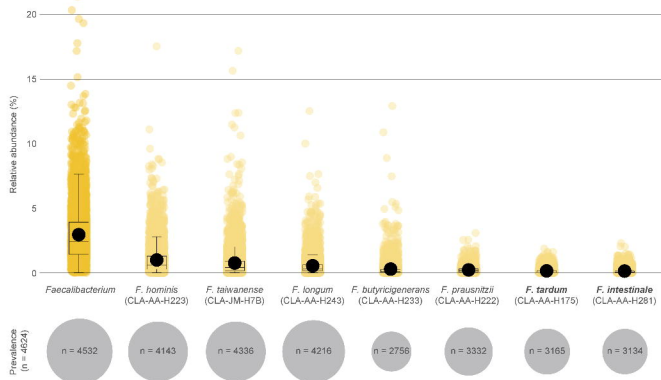




a



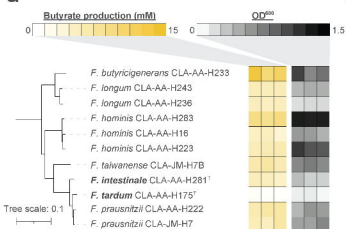
b



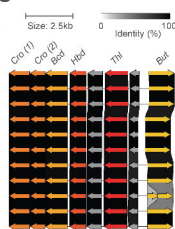
c



d



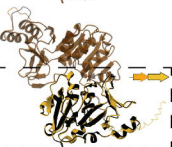
e



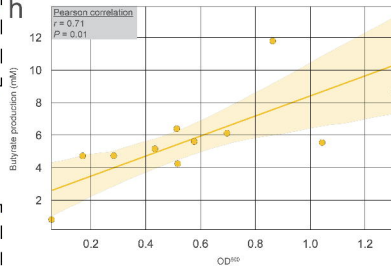
f



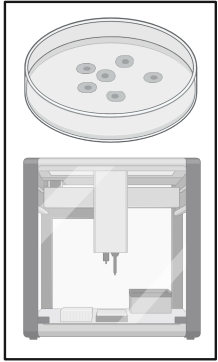
g



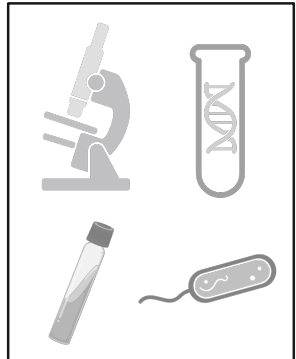
h



## Isolation



## Description

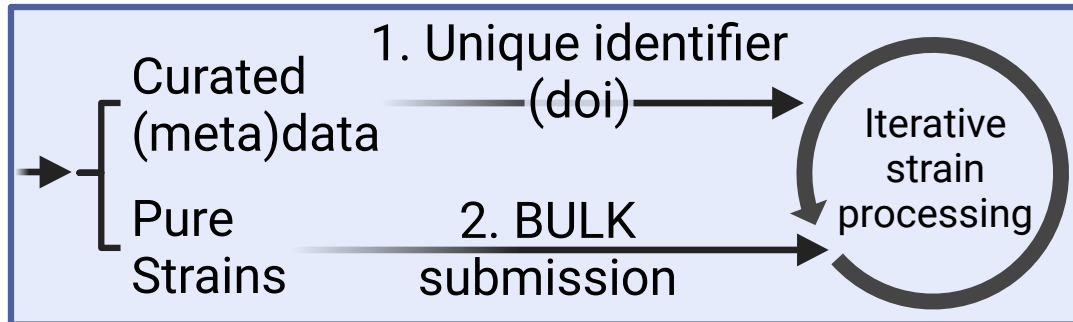


## Current strategies

- ➔ 1. Publication
- ➔ 2. Publication, then deposition\*
- ➔ 3. Parallelled publication & deposition\*
- ➔ 4. Deposition\*, then publication

\*Single strain deposition

## Proposed system



Quick publication  
Curated (meta)data  
Long-term accessibility  
Traceability



<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

1, Only original strain ID in publication  
2, Strain may fail to be deposited  
3, Deposition is time consuming

<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
-------------------------------------	-------------------------------------	-------------------------------------	-------------------------------------