A highly conserved and globally prevalent cryptic plasmid is among the most numerous mobile genetic elements in the human gut

Emily C Fogarty[1,2,3,*], Matthew S Schechter[1,2,3], Karen Lolans[3], Madeline L. Sheahan[2,4], Iva Veseli[3,5], Ryan Moore[6], Evan Kiefl[3,5], Thomas Moody[7], Phoebe A Rice[1,8], Michael K Yu[9], Mark Mimee[1,4,10], Eugene B Chang[3], Sandra L Mclellan[11], Amy D Willis[12], Laurie E Comstock[2,4], A Murat Eren[13,14,15,16,*]

[1]Committee on Microbiology, University of Chicago, Chicago, IL 60637, USA; [2]Duchossois Family Institute, University of Chicago, Chicago, IL 60637, USA; [3]Department of Medicine, University of Chicago, Chicago, IL 60637, USA; [4]Department of Microbiology, University of Chicago, Chicago, IL, 60637, USA; [5]Graduate Program in Biophysical Sciences, University of Chicago, Chicago, IL 60637, USA; [6]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA; [7]Department of Systems Biology, Columbia University, New York, NY, 10032 USA; [8]Department of Biochemistry, University of Chicago, Chicago, IL, 60637, USA; [9]Toyota Technological Institute at Chicago; [10]Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL 60637, USA; [11]School of Freshwater Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI, 53204, USA; [12]Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA; [13]Marine Biological Laboratory, Woods Hole, MA, 02543, USA; [14]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, 27570 Bremerhaven, Germany. [15]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26129 Oldenburg, Germany; [16]Helmholtz Institute for Functional Marine Biodiversity, 26129 Oldenburg, Germany.

* Correspondence: meren@hifmb.de and efogarty@uchicago.edu

Running Title: pBI143 is ubiquitous across global human populations

ABSTRACT

Plasmids are extrachromosomal genetic elements that often encode fitness enhancing features. However, many bacteria carry 'cryptic' plasmids that do not confer clear beneficial functions. We identified one such cryptic plasmid, pBI143, which is ubiquitous across industrialized gut microbiomes, and is 14 times as numerous as crAssphage, currently established as the most abundant genetic element in the human gut. The majority of mutations in pBI143 accumulate in specific positions across thousands of metagenomes, indicating strong purifying selection. pBI143 is monoclonal in most individuals, likely due to the priority effect of the version first acquired, often from one's mother. pBI143 can transfer between Bacteroidales and although it does not appear to impact bacterial host fitness *in vivo*, can transiently acquire additional genetic content. We identified important practical applications of pBI143, including its use in identifying human fecal contamination and its potential as an inexpensive alternative for detecting human colonic inflammatory states.

INTRODUCTION

The tremendous density of microbes in the human gut provides a playground for the contact-dependent transfer of mobile genetic elements [1] including plasmids. Plasmids are typically defined as extrachromosomal elements that replicate autonomously from the host chromosome [1–4]. In addition to being a workhorse for molecular biology, plasmids have been extensively studied for their ability to expedite microbial evolution [5] and enhance host fitness by providing properties such as antibiotic resistance, heavy metal resistance, virulence factors, or metabolic functions [6–11].

Plasmids have been a major focus of microbiology not only for their biotechnological applications to molecular biology [12–15] but also for their role in the evolution and dissemination of genes for antibiotic resistance [16,17], which is a growing global public health concern [18]. However, outside the spotlight lie a group of plasmids that appear to lack genetic functions of interest and that do not contain genes encoding obvious beneficial functions to their hosts [19,20]. Such 'cryptic plasmids' are typically small and multi-copy [21], and are often difficult to study as they lack any measurable phenotypes or selectable markers [22,23], despite their presence in a broad range of microbial taxa [24–27]. In the absence of a clear advantage to their hosts, and the presumably non-zero cost of their maintenance, these plasmids are often described as selfish elements [28] or genetic parasites [29]. While they may provide unknown benefits to their hosts, a high transfer rate could also be a factor that enables cryptic plasmids to counteract the negative selection pressure of their maintenance [29–31].

Analyses of cryptic plasmids are often performed on monocultured bacteria, limiting insights into the ecology of cryptic plasmids in their host's natural environment. However, recent advances in shotgun metagenomics [30] and *de novo* plasmid prediction algorithms [31–40] offer a powerful means to bridge this gap. For instance, in a recent study we characterized over 68,000 plasmids from the human gut [40] and observed that the most prevalent known plasmid across geographically diverse human populations was a cryptic plasmid, called pBI143. Here we conduct an in-depth characterization of this cryptic plasmid through 'omics and experimental approaches to study its genetic diversity, host range, transmission routes, impact on the bacterial host, and associations with health and disease states. Our findings reveal the astonishing success of pBI143 in the human gut, where it occurs in up to 92% of individuals in industrialized countries with copy numbers 14

42   times higher on average than crAssphage, the most abundant phage in the human gut. We also

43   demonstrate the potential of pBI143 as a cost-effective biomarker to assess the extent of stress that

44   microbes experience in the human gut, and as a sensitive means to quantify the level of human

45   fecal contamination in environmental samples.

RESULTS
46

pBI143 is extremely prevalent across industrialized human gut microbiomes
47
48

49   pBI143 (accession ID U30316.1) is a 2,747 bp circular plasmid first identified in 1985 [41] in

50   *Bacteroides fragilis* [42], an important member of the human gut microbiome that is frequently

51   implicated in states of health [43–45] and disease [46,47]. pBI143 encodes only two annotated genes: a

52   mobilization protein (*mobA*) and a replication protein (*repA*) (Fig. 1A). Due to the desirable

53   features for cloning such as a high copy number and genetic stability, pBI143 has been primarily

54   used as a component of *E. coli-Bacteroides* shuttle vectors [42]. The absence of any ecological

55   studies of pBI143 prompted us to characterize it further beginning with a characterization of its

56   genetic diversity.

57   To comprehensively sample the diversity of pBI143, we screened 2,137 individually assembled

58   human gut metagenomes (Supplementary Table 1) for pBI143-like sequences. By surveying all

59   contigs using the known pBI143 sequence as reference, we found three distinct versions of pBI143

60   (Fig. 1A), all of which had over 95% nucleotide sequence identity to one another throughout their

61   entire length except at the *repA* gene, where the sequence identity was as low as 75% with a

62   maximum of 81% between Version 1 and Version 2 (Supplementary Table 2).
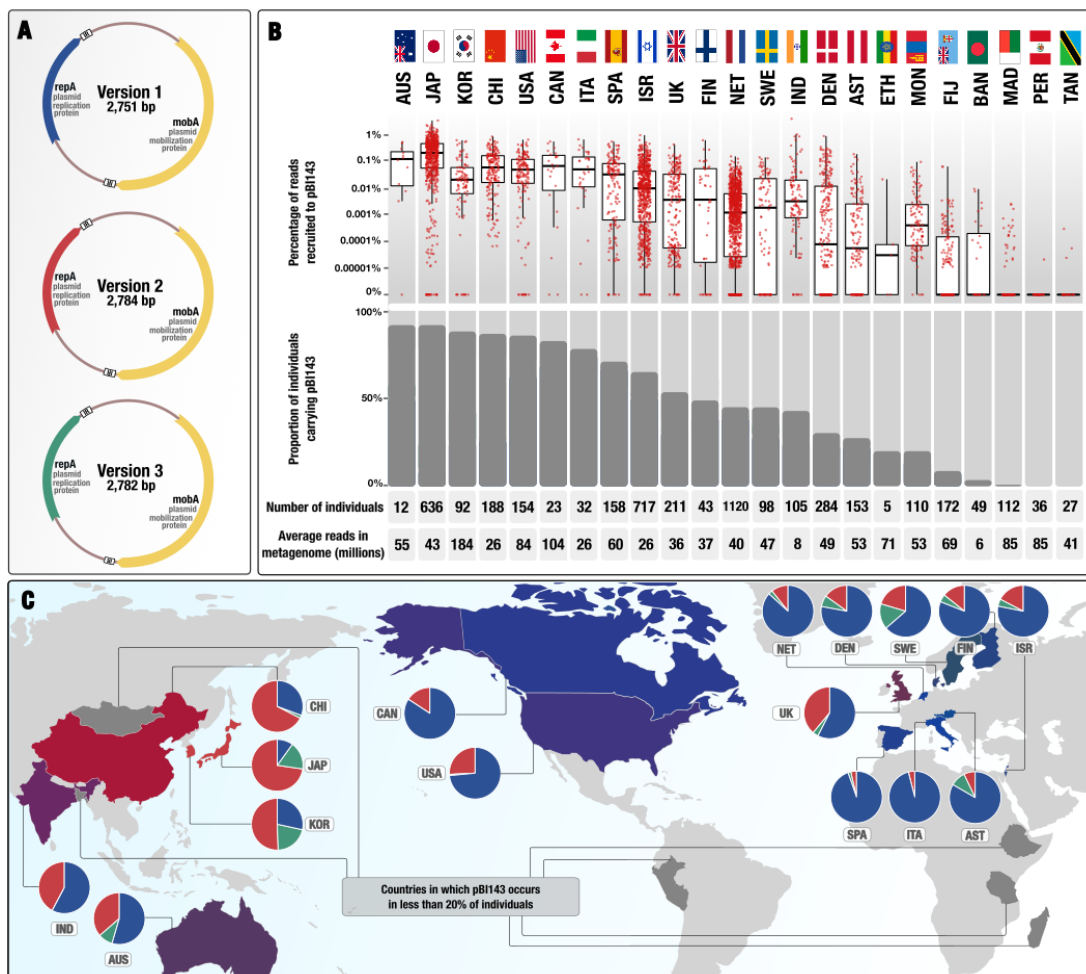
Fig. 1 pBI143 prevalence and abundance in globally distributed human populations. (A) Plasmid maps of the three distinct versions of pBI143, which differ primarily in the *repA* gene. IR = inverted repeat. The *repA* genes are colored according to Version 1 (blue), Version 2 (red) and Version 3 (green). (B) Read recruitment results from 4,516 metagenomes originating from 23 globally representative countries and mapped to pBI143. Top: The percentage of reads in each metagenome that mapped to pBI143 normalized by number of reads in the metagenome. Bottom: The proportion of individuals in a country that have pBI143 in their gut. Each red dot represents an individual metagenome. (C) Countries that are represented in our collection of 4,516 global adult gut metagenomes. Each country's pie chart is colored based on the version(s) of pBI143 that is most prevalent in that country (Version 1 = blue, Version 2 = red, Version 3 = green). Each country is colored based on the proportion of Version 1, 2 or 3 present in the population, or gray if fewer than 20% of individuals carry pBI143. Pie charts show the proportions of pBI143 versions in all individuals that carry it within a country.

77   We then sought to quantify the prevalence of pBI143 across global human populations using a

78   metagenomic read recruitment survey with an expanded set of 4,516 publicly available gut

79   metagenomes from 23 countries [48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,55,64,65,66,67,6866,69,70] (Supplementary

80   Table 1). Recruiting metagenomic short reads from each gut metagenome using each pBI143

81   version independently (Supplementary Fig. 1, Supplementary Table 3), we found that pBI143 was

82   present in 3,295 metagenomes, or 73% of all samples (Fig. 1B, see Methods for the 'detection'

83   criteria). However, the prevalence of pBI143 was not uniform across the globe (Fig. 1B): pBI143

84   occurred predominantly in metagenomes of individuals who lived in relatively industrialized

85   countries, such as Japan (92% of 636 individuals) and the United States (86% of 154 individuals).

86   We rarely detected pBI143 in individuals who lived in relatively non-industrialized countries such

87   as Madagascar (0.8% of 112 individuals) or Fiji (8.7% of 172 individuals). This differential

88   coverage is likely due to the non-uniform distribution of *Bacteroides* populations, which tend to

89   dominate individuals who live in relatively more industrialized countries [71]. Within each

90   individual, pBI143 was often highly abundant (Fig. 1B), and despite its small size, it often recruited

91   0.1% to 3.5% of all metagenomic reads with a median coverage of over 7,000X (Supplementary

92   Fig. 1, Supplementary Table 3). In one extreme example, pBI143 comprised an astonishing 7.5%

93   of all reads in an infant gut metagenome from Italy, with a metagenomic read coverage exceeding

94   54,000X (Supplementary Table 3).

95   The distribution of pBI143 versions across human populations was also not uniform as different

96   versions of pBI143 tended to be dominant in different geographic regions. pBI143 Version 1 (98%

97   identical to the original reference sequence for pBI143 [41]) dominated individuals in North America

98   and Europe, and occurred on average in 82.5% of all samples that carry pBI143 from Austria,

99   Canada, Denmark, England, Finland, Italy, Netherlands, Spain, Sweden and the USA (Fig. 1C,

100   Supplementary Table 3). In contrast, pBI143 Version 2 dominated countries in Asia and occurred

101   in 63.6% of all samples that carry pBI143 in China, Japan, and Korea (Fig. 1C, Supplementary

102   Table 3). pBI143 Version 3 was relatively rare, comprising only 7.4% of pBI143-positive samples,

103   and mostly occurred in individuals from Japan, Korea, Australia, Sweden, and Israel (Fig. 1C,

104   Supplementary Table 3).

105   The extremely high prevalence and coverage of pBI143 suggests that it is likely one of the most

106   numerous genetic elements in the gut microbiota of individuals from industrialized countries. We

5

107   compared the prevalence and relative abundance of pBI143 to crAssphage, a 97 kbp bacterial virus

108   that is widely recognized as the most abundant family of viruses in the human gut [72]. pBI143 was

109   more prevalent (73% vs 27%) in our metagenomes than crAssphage, although individual samples

110   differed widely with respect to the abundance of these two elements in a given individual

111   (Supplementary Table 3). The average percentage of metagenomic reads recruited by pBI143 and

112   crAssphage were 0.05% and 0.13%, respectively. However, taking into consideration that

113   crAssphage is approximately 36 times larger than pBI143, and assuming that average coverage is

114   an acceptable proxy to the abundance of genetic entities, these data suggest that on average pBI143

115   is 14 times more numerous than crAssphage in the human gut.

116   Overall, these data demonstrate that pBI143 is one of the most widely distributed and numerous

117   genetic elements in the gut microbiomes of industrialized human populations world-wide.

pBI143 is specific to the human gut and hosted by a wide range of Bacteroidales species

120   Interestingly, the detection patterns of pBI143 in metagenomes differed from the detection patterns

121   we observed for its *de facto* host *Bacteroides fragilis* in the same samples; *B. fragilis* and pBI143

122   co-occurred in only 41% of the metagenomes. Sequencing depth did not explain this observation,

123   as pBI143 was highly covered (i.e., >50X) in 25% of metagenomes where *B. fragilis* appeared to

124   be absent (Supplementary Table 11), suggesting that the host range of pBI143 extends beyond *B.*

125   *fragilis*.

126   To investigate the host range of pBI143, we employed a collection of bacterial isolates from the

127   human gut, which contained 717 genomes that represented 104 species in 54 genera

128   (Supplementary Table 4). We found pBI143 in a total of 82 isolates that resolved to 11 species

129   across 3 genera: *Bacteroides*, *Phocaeicola*, and *Parabacteroides*. Many of the pBI143-carrying

130   isolates of distinct species were from the same individuals, suggesting that pBI143 can be

131   mobilized between species. To confirm this, we inserted a tetracycline resistance gene, *tetQ*, into

132   pBI143 in the *Phocaeicola vulgatus* isolate MSK 17.67 (Supplementary Fig. 2, Supplementary

133   Table 4) and tested the ability of this engineered pBI143 to transfer to two strains of two different

134   families of Bacteroidales, *Bacteroides ovatus* D2 and *Parabacteroides johnsonii* CL02T12C29. In

135   these assays, we found that pBI143 was indeed transferred from the donor to the recipient strains

136   at a frequency of 5 x 10^{-7} and 3 x 10^{-6} transconjugants per recipient, respectively (Supplementary

137   Fig. 2).

138   Given the broad host range of pBI143, one interesting question is whether the ecological niche

139   boundaries of pBI143 hosts exceed a single biome, since the members of the order Bacteroidales

140   are not specific to the human gut and do occur in a wide range of other habitats from non-human

141   primate guts [73] to marine systems [74]. To investigate whether pBI143 might exist in non-human

142   environments, we searched for pBI143 in metagenomes from coastal and open ocean samples [75,76],

143   captive macaques [73], human-associated pets [77], and sewage samples from across the globe [78]. The

144   plasmid was absent from all non-human associated samples, but as expected, was present in

145   sewage (Supplementary Fig. 3, Supplementary Table 3, Supplementary Text). Given the absence

146   of pBI143 in non-human associated habitats, we also screened metagenomes from human skin and

147   oral cavity [70]. Unlike the extremely high presence of pBI143 in the human gut, pBI143 was poorly

148   detected both in samples from skin and the oral cavity (Supplementary Text). Finally, we designed

149   and tested a highly specific qPCR assay for pBI143 (Supplementary Table 5) to confirm its

150   specificity to the human gut. While there was a robust amplification of pBI143 from sewage

151   samples confirming our insights from metagenomic coverages (Fig. 2), pBI143 was virtually

152   absent in dog, alligator, raccoon, horse, pig, deer, cow, chicken, goose, cat, rabbit, deer, or gull

153   fecal samples (Supplementary Table 6). The only exception was the relatively low copy number

154   (i.e., 73-fold less than human fecal content of sewage) in three of the four cats tested.

155   The near-absolute exclusivity of pBI143 to the human gut presents practical opportunities, such as

156   the accurate detection of human fecal contamination outside the human gut. Using the same PCR

157   primers, we also amplified pBI143 from water and sewage samples and compared its sensitivity

158   to the gold standard markers currently used for detecting human fecal contamination in the

159   environment (16S rRNA gene amplification of human *Bacteroides* and Lachnospiraceae) [79,80].

160   pBI143 had higher amplification in all 41 samples where *Bacteroides* and Lachnospiraceae were

161   also detected (Fig. 2). pBI143 was also amplified in 6 samples with no *Bacteroides* or

162   Lachnospiraceae amplification, suggesting it is a highly sensitive marker for detecting the presence

163   of human-specific fecal material.

7

164 Overall, these data show that pBI143 has a broad range of Bacteroidales species, is highly specific

165 to the human gut environment, and can serve as a sensitive biomarker to detect human fecal

166 contamination.
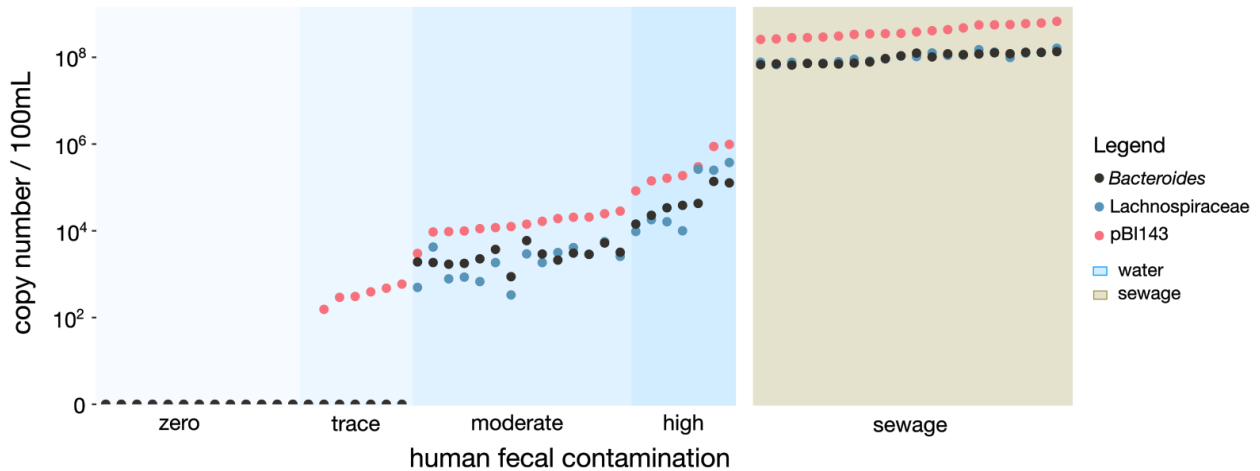
167

168 **Fig. 2. Detection of pBI143 and two established human fecal markers in water and sewage samples.**

169 Copy number of pBI143, human *Bacteroides* or Lachnospiraceae as measured by qPCR. Zero, trace,

170 moderate, high and sewage categories and sample order designations are determined based on pBI143 copy

171 number. Trace indicates one of the established markers was detected but was below the level of

172 quantification. The blue background indicates water samples and the beige background indicates samples

173 from sewage.

pBI143 is monoclonal within individuals, and its variants across

individuals are maintained by strong purifying selection

176 So far, our investigation of pBI143 has focused on its ecology. Next, we sought to understand the

177 evolutionary forces that have conserved the pBI143 sequence by quantifying the sequence

178 variation among the three distinct versions and examining the distribution of single nucleotide

179 variants (SNVs) within and across globally distributed individuals. Across the three versions, both

180 pBI143 genes had low dN/dS values (*mobA* = 0.11, *repA* = 0.04), suggesting the presence of strong

181 forces of purifying selection acting on *mobA* and *repA* resulting in primarily synonymous

182 substitutions. While the comparison of the three representative sequences provide some insights

183 into the conserved nature of pBI143, it is unlikely they capture its entire genetic diversity across

184 gut metagenomes.

185   To explore the pBI143 variation landscape, we analyzed metagenomic reads that matched the

186   Version 1 of *mobA* to gain insights into the population genetics of pBI143 in naturally occurring

187   habitats through single-nucleotide variants (SNVs). Since the *mobA* gene was more conserved

188   across distinct versions of the plasmid compared to the *repA* gene, focusing on *mobA* enabled

189   characterization of variation from all plasmid versions using a single read recruitment analysis.

190   Surprisingly, the vast majority (83.2%) of the nucleotide positions that varied in any metagenome

191   matched a nucleotide position that was variable between at least one pair of the three plasmid

192   versions (Fig. 3A, Supplementary Table 7). In other words, pBI143 variation across metagenomes

193   was predominantly localized to certain nucleotide positions that differed between the

194   representative sequences of pBI143 for Version 1, 2 and 3, indicating that the three representative

195   versions capture the majority of permissible pBI143 variation within our collection of gut

196   metagenomes. Indeed, only 24.5% of metagenomes had more than three novel SNVs that were not

197   present in at least one plasmid version, and 84.8% of metagenomes had pBI143 sequences that

198   were within 2-nucleotide distance of one of the three versions. In addition to the primarily localized

199   variation of pBI143, we also observed that the vast majority of SNVs were fixed within a

200   metagenome (i.e., a 'departure from consensus' value of ~0, see Methods), suggesting that most

201   humans carry a monoclonal population of pBI143 with little to no within-individual variation (Fig.

202   3C, Supplementary Table 7).

203
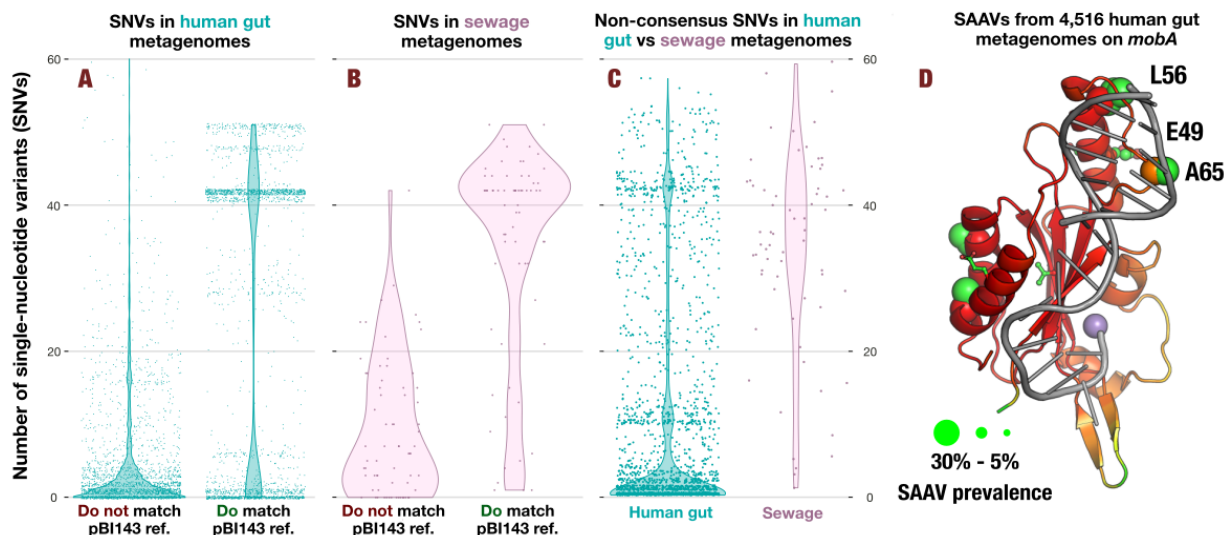
Fig. 3. The mutational landscape of pBI143 in sewage and the human gut. (A) The proportion of SNVs across 4,516 human gut metagenomes that are present in the same location (match) or different locations (do not match) as variation in one of the versions of pBI143 (turquoise). Each point is a single metagenome. **(B)** The proportion of SNVs across 68 sewage gut metagenomes that are present in the same location (match) or different locations (do not match) as variation in one of the versions of pBI143 (pink). **(C)** Non-consensus SNVs present in 4,516 human gut metagenomes and 68 sewage metagenomes. **(D)** AlphaFold 2 predicted structure of the catalytic domain of MobA with single amino acid variants from all 4,516 human gut metagenomes superimposed as ball-and-stick residues. oriT DNA (gray) and a Mn2+ ion marking the active site (purple) were modeled based on 4lvi.pdb [81]. The size of the ball-and-stick spheres indicate the proportion of samples carrying variation in that position (the larger the sphere, the more prevalent the variation at the residue) and the color is in CPK format. The color of the ribbon diagram indicates the pLDDT from AlphaFold 2 with red = very high (> 90 pLDDT) and orange = confident (80 pLDDT).

Next, we sought to investigate the functional context of non-synonymous environmental variants of MobA given its structure. For this, we employed single-amino acid variants [82] (SAAVs) we recovered from gut metagenomes and superimposed them on the AlphaFold 2 [82,83] predicted structure of MobA using anvi'o structure [84,82,83]. The predicted catalytic domain of pBI143 MobA was structurally similar to MobM of the MobV-family (Protein Data Bank accession: 4LVI) encoded by plasmid pMV158 [81]. We used the structurally similar catalytic domain in MobA to model the binding of the oriT of pBI143 to MobA. We found that there were only 21 SAAVs throughout MobA that were present in greater than 5% of the gut metagenomes (Fig. 3D). Interestingly, highly prevalent SAAVs occurred exclusively near the DNA binding site (L56, E49, and A64), suggesting that that the non-synonymous variants we observe in the context of MobA may be involved in altering the DNA binding specificity for the oriT sequence [81] demonstrating the coevolution of the oriT with the MobA protein between distinct pBI143 versions. Additionally, we find it likely that the cluster of high prevalence variation at residues V251, A246, V239, T238, I235, and L234 (Supplementary Fig. 4B) may be driven by interactions with different host conjugation machinery for plasmid transfer. The functional implications of prevalent SAAVs given the structural context of the MobA gene highlight the role of adaptive processes on the evolution of pBI143 versions.

Unlike the individual gut metagenomes, the pBI143 sequences did not occur in a monoclonal fashion in sewage metagenomes as expected (Supplementary Table 7). Sewage metagenomes had, on average, 35 SNVs with a departure from consensus value of lower than 0.9, revealing the polyclonal nature of pBI143 in sewage (Fig. 3C, Supplementary Table 7). Similar to the individual

237   gut metagenomes, most SNVs in sewage metagenomes (78.8%) occurred at a nucleotide position

238   that was variable across at least one pair of the three pBI143 versions (Fig. 3B, Supplementary

239   Table 7), suggesting that the majority of the variability in sewage is from the mixing of different

240   versions of pBI143. However, the number of novel SNVs was much higher in sewage: 61.8% of

241   sewage samples had greater than three SNVs that did not match a variable position in one of the

242   three reference plasmids (Fig. 3B). Given the marked increase in the number of novel SNVs in

243   sewage, it is likely there are additional but relatively rare versions of pBI143 in the human gut.

244   Overall, these results indicate that pBI143 has a highly restricted mutational landscape in natural

245   habitats, frequently occurs as a monoclonal element in individual gut metagenomes, and the non-

246   synonymous variants of MobA in the environment may be responsible for altering its DNA

247   binding.

pBI143 is vertically transmitted, its variants are more specific to

individuals than their host bacteria, and priority effects best explain its

monoclonality in most individuals

251   The largely monoclonal nature of pBI143 presents an interesting ecological question: how do

252   individuals acquire it, and what maintains its monoclonality? Multiple phenomena could explain

253   the monoclonality of pBI143 in individual gut metagenomes, including (1) low frequency of

254   exposure (i.e., most individuals are only ever exposed to one version), (2) bacterial host specificity

255   (i.e., some plasmid versions replicate more effectively in certain bacterial hosts), or (3) priority

256   effects (i.e., the first version of pBI143 establishes itself in the ecosystem and excludes others).

257   The sheer prevalence and abundance of pBI143 across industrialized populations renders the 'low

258   frequency of exposure' hypothesis an unlikely explanation. Yet the remaining two hypotheses

259   warrant further investigation.

260   Bacterial host specificity is a plausible driver for the presence of a singular pBI143 version within

261   an individual, given the interactions between plasmid replication genes and host replication

262   machinery [28,85]. However, our analysis of 82 bacterial cultures isolated from 10 donors shows that

263   the plasmid is more specific to individuals than it is to certain bacterial hosts (Fig. 4,

264   Supplementary Table 9). Indeed, identical pBI143 sequences often occurred in multiple distinct

265 taxa isolated from the same individual, in agreement with the monoclonality of pBI143 in gut

266 metagenomes and its ability to transfer within Bacteroidales. If pBI143 monoclonality is not driven

267 by rare exposure or host specificity, it could be driven by priority effects [86], where the initial

268 pBI143 version somehow prevents other pBI143 versions from establishing in the same gut

269 community.

270
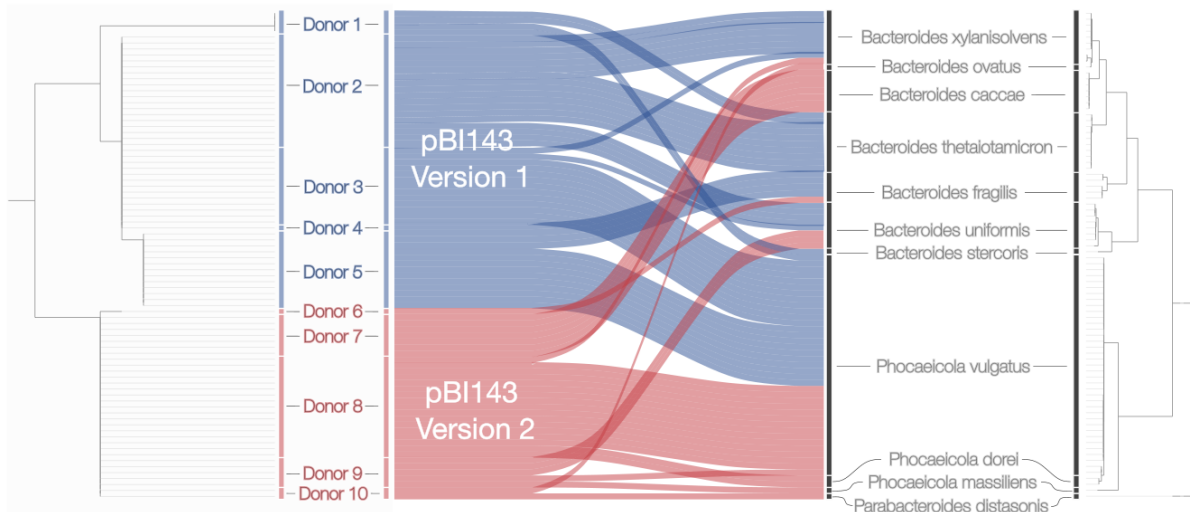
Fig. 4. Phylogeny of pBI143 in human donors versus the phylogeny of bacterial isolates recovered from the same individuals. pBI143 (left) and bacterial host (right) genome phylogenies. The pBI143 phylogeny was constructed using the MobA and RepA genes; the bacterial phylogeny was constructed using 38 ribosomal proteins (see Methods). Blue alluvial plots are isolates with Version 1 pBI143 and red alluvial plots are isolates with Version 2 pBI143. No isolates had the rarer Version 3.

276 To examine if priority effects play a role in pBI143 monoclonality, we aimed to determine how

277 pBI143 is acquired. Given that one established route of microbial acquisition is the vertical

278 transmission of microbes from mother to infant [87], we used our ability to track pBI143 SNVs

279 between environments to investigate if there is evidence for vertical transmission. We followed

280 the inheritance of identical SNV patterns in pBI143 using 154 mother and infant gut metagenomes

281 from four countries, Finland [57], Italy [60], Sweden [68], and the USA [69], where each study followed

282 participants from birth to 3 to 12 months of age. We recruited reads from each metagenome to

283 Version 1 pBI143 (Supplementary Table 1 and 3, Supplementary Fig. 5) and identified the location

284 of each SNV in *mobA* (Supplementary Table 10). These data revealed a large number of cases

285 where pBI143 had identical SNV patterns in mother-infant pairs (Fig. 5A, Supplementary Table

12

286   10). A network analysis of shared SNV positions across metagenomes appeared to cluster family

287   members more closely, indicating mother-infant pairs had more SNVs in common than they had

288   with unrelated individuals, which we could further confirm by quantifying the relative distance

289   between each sample to others (Supplementary Fig. 6, Supplementary Table 10, Methods).

290   Establishing that pBI143 is often vertically transferred, we next examined the impact of priority

291   effects on pBI143 maintenance over time. We assumed that if priority effects are driving

292   persistence of a single version of pBI143, the first version that enters the infant gut environment

293   should be maintained over time. Indeed, many phage populations are influenced by priority effects

294   where the presence of one phage provides a competitive advantage to the host [88] or host immunity

295   to infection with similar phages [89–91]. In our data, we found no instances where pBI143 acquired

296   from the mother was fully replaced in the infant during and up to the first year of life

297   (Supplementary Fig. 5, Supplementary Table 10). While 69% of infants maintained the version

298   received from the mother (Fig. 5B), we also observed other, less common genotypes. These less

299   common cases included a 'two versions' scenario where the mother possessed two versions of

300   pBI143, both of which were passed to the infant (21%), and a 'wilt' case, where the transferred

301   pBI143 was neither replaced nor persisted until the end of sampling (7%) (Fig. 5B). Although

302   these less prevalent phenotypes are not necessarily explained by priority effects, 69% maintenance

303   of the initial version of pBI143 suggests that priority effects have an important role in the

304   maintenance of pBI143 in the gut, despite many incoming populations colonizing the infant and

305   likely carrying other pBI143 versions.

306   Overall, by tracking SNV patterns between environments we established that pBI143 is vertically

307   transferred from mothers to infants and that priority effects likely play a role in maintaining the

308   predominantly monoclonal populations of pBI143.

13

Fig. 5. Transfer and maintenance of pBI143. (A) The network shows the degree of similarity between pBI143 SNVs across 154 mother and infant metagenomes from Finland, Italy, Sweden and the USA. Each node is an individual metagenome and nodes are colored based on family grouping. The surrounding coverage plots (colored) are visual representations of SNV patterns present in the indicated metagenomes. Nodes labeled with an "M" are mothers; nodes with no labels are infants. **(B)** Representative coverage plots showing different coverage patterns (maintained, two versions or wilt) observed in plasmids transferred from mothers to infants.

pBI143 is a highly efficient parasitic plasmid

An intuitive interpretation of the surprising levels of prevalence and abundance of pBI143 across the human population, in addition to its limited variation maintained by strong evolutionary forces, is that it provides some benefit to the bacterial host. However, the two annotated genes in pBI143 appear to serve only the purpose of ensuring its own replication and transfer, contradicting this

14

premise. The coverage of pBI143 and its *Bacteroides, Phocaeicola* and *Parabacteroides* hosts in gut metagenomes indeed show a significant positive correlation (R^2: 0.5, p-value < 0.001) (Fig. 6A, Supplementary Table 11), however, these data are not suitable to distinguish whether pBI143 provides a benefit to the bacterial host fitness, or acts as a genetic hitchhiker.
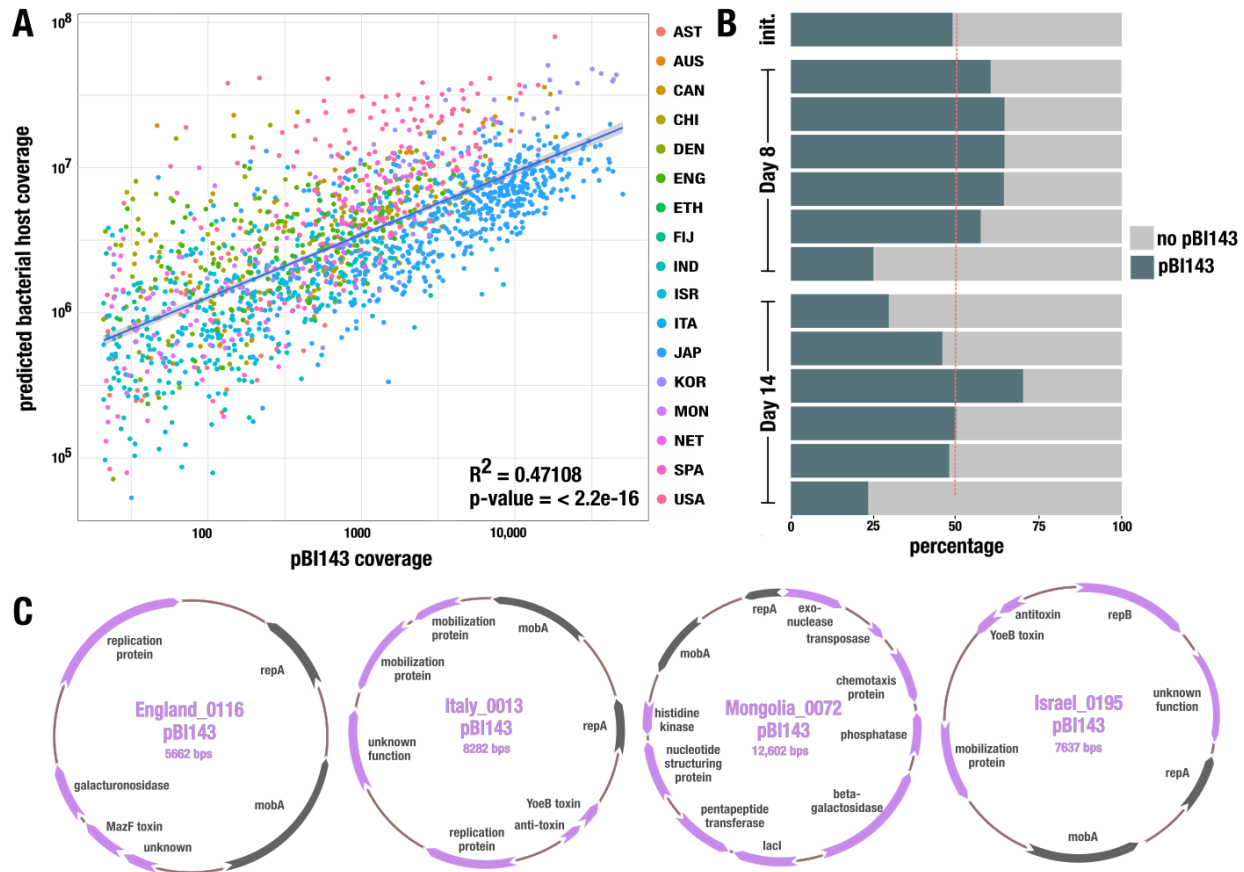
Fig. 6. The relationship between pBI143 and its bacterial hosts. (A) The average coverage of pBI143 and the corresponding coverage of predicted host genomes (*Bacteroides, Parabacteroides and Phocaeicola*) in 4,516 metagenomes. (B) Competition experiments in gnotobiotic mice between *B. fragilis* with and without pBI143. The proportion of pBI143-carrying cells in 6 mice in the initial inoculum, at Day 8 and at Day 14 are shown. (C) Four examples of pBI143 assembled from metagenomes that carry additional cargo genes. Gray genes are the canonical *repA* and *mobA* genes of naive pBI143; lilac genes are additional cargo.

15

334 To experimentally investigate if pBI143 is advantageous or parasitic, we constructed isogenic pairs

335 of *B. fragilis* 638R and *B. fragilis* 9343 with and without the native Version 1 sequence of pBI143

336 (Supplementary Methods). To determine if pBI143 is well-adapted to replication in a new

337 *Bacteroides* host, we tested its maintenance in culture. After 7 days of passaging, pBI143 was still

338 present in all colonies of *B. fragilis* 638R and *B. fragilis* 9343 (Supplementary Table 12). Next,

339 we competed the *B. fragilis* 638R (with and without pBI143) of *B. fragilis* 638R in gnotobiotic

340 mice for 2 weeks. At Day 8, 5/6 mice had more *B. fragilis* 638R with pBI143 than without;

341 however this trend did not continue into Day 14, where 4/6 mice had fewer cells with pBI143 (Fig.

342 6B, Supplementary Table 12). While we can speculate that these populations may continue to

343 fluctuate, the results at least suggest a negligible negative fitness impact of pBI143 on its bacterial

344 host.

345 One potential benefit that pBI143 could provide to its host is to act as a natural shuttle vector by

346 transiently acquiring additional genetic material and transferring it between cells in a community.

347 In fact, in our survey of assembled gut metagenomes we observed a few cases that may support

348 such a role for pBI143. In most individuals, we assembled pBI143 in its native form with 2 genes.

349 However, there were 10 instances where the assembled pBI143 sequence from a given

350 metagenome contained additional genes (Fig. 6C, Supplementary Table 2). Many of the additional

351 genes have no predicted function, but other cargo include toxin-antitoxin genes conferring plasmid

352 stability, as well as those that may confer beneficial functions to the bacterial host, such as

353 galacturonosidase, pentapeptide transferase, phosphatase, and histidine kinase genes. These

354 occasional larger versions of pBI143 share a common backbone of *repA* and *mobA* and thus form

355 a "plasmid system" [40], a common plasmid evolutionary pattern suggesting the possibility that

356 pBI143 may dynamically acquire different genes in different environments.

357 Overall, it does not appear that the native sequence of pBI143 provides a clear benefit to its host

358 cells, however it does appear to positively correlate with these hosts in metagenomic data, and is

359 maintained in the absence of selection in new hosts *in vitro*.

360 ## pBI143 responds to oxidative stress *in vitro*, and its copy number is

361 ## significantly higher in metagenomes from individuals who are diagnosed

362 ## with IBD

363 Mobile genetic elements rely on their hosts for replication machinery, but many have developed

364 mechanisms to increase their rates of replication and transfer during stressful conditions to increase

365 the likelihood of their survival if the host cell dies [92–95]. To investigate whether the copy number

366 of pBI143 changes as a function of stress, we first conducted an experiment with *B. fragilis* isolates

367 that naturally carry pBI143.

368 Given that oxygen exposure upregulates oxidative stress response pathways in the anaerobic *B.*

369 *fragilis* [96], we exposed two different *B. fragilis* cultures, *B. fragilis* R16 (which was isolated from

370 a healthy individual) and *B. fragilis* 214 (which was isolated from a pouchitis patient [97]) to 21%

371 oxygen for increasing periods of time (Fig. 7A, Supplementary Fig. 7, Supplementary Table 13).

372 To calculate the copy number of pBI143 in culture, we quantified the ratio between the total

373 number of plasmids and the total number of cells in culture using a qPCR with primers targeting

374 pBI143 and a *B. fragilis*-specific gene we identified through pangenomics (Methods). As the

375 length of oxygen exposure increased, the copy number of pBI143 per cell also increased. Notably,

376 the copy number was quickly reduced to control levels once the cultures were returned to anaerobic

377 conditions, indicating that copy number fluctuation is a rapid and transient process that is

378 dependent on host stress.

379 Oxidative stress is also a signature characteristic of inflammatory bowel disease (IBD), a group of

380 intestinal disorders that cause inflammation of the gastrointestinal tract [98]. The dysregulation of

381 the immune system during IBD typically leads to high levels of oxidative stress in the gut

382 environment [99]. We thus hypothesized that, if oxidative stress is among the factors that drive the

383 increased copy number of pBI143 in culture, one should expect a higher copy number of pBI143

384 in metagenomes from IBD patients compared to healthy controls.

385 To analyze the copy number of pBI143 in a given metagenome, we calculated the ratio of

386 metagenomic read coverage between pBI143 and its bacterial host in metagenomes where pBI143

387 could confidently be assigned to a single host. With these considerations, we developed an

17

approach to calculate an 'approximate copy number ratio' (ACNR) for pBI143 and its unambiguous bacterial host in a given metagenome using bacterial single-copy core genes (see Methods). We calculated the ACNR of pBI143 in 3,070 healthy and 1,350 IBD gut metagenomes (Supplementary Table 1, Supplementary Fig. 1 and 8). Our analyses showed that the geometric mean of the ACNR for pBI143 and its host was 3.72 times larger (robust-Wald 95% CI: 2.66x - 5.20x, p-value $< 10^{-13}$) in IBD compared to healthy metagenomes, indicating that the pBI143 ACNR was significantly higher in individuals with IBD compared to those who were healthy (Fig. 7B, Supplementary Table 14).
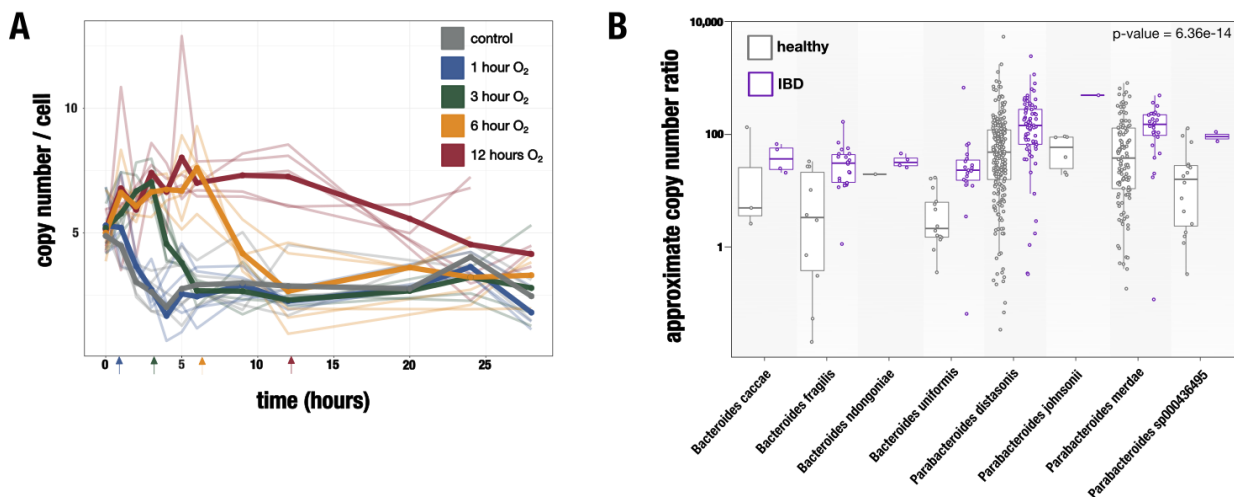
Fig. 7. pBI143 copy number increases in stressful environments. (A) Copy number of pBI143 in *B. fragilis* 214 cultures with increasing exposure to oxygen. Arrows indicate the time point at which the culture was returned to the anaerobic chamber. The control cultures (gray) were never exposed to oxygen. Opaque lines are the mean of 5 replicates (translucent lines). **(B)** Host-specific approximate copy number ratio (ACNR) of pBI143 in healthy individuals (gray) versus those with IBD (purple).

The copy number ratio of pBI143 to its *B. fragilis* host in culture calculated with qPCR primers was much lower (~5X on average) compared to its approximate copy number ratio in healthy metagenomes (~120X on average). Multiple factors can explain this difference, including biases associated with sequencing steps or the calculation of the coverage, or that the conditions naturally occuring communities experience vastly differ than those conditions encountered in culture media,

18

409   even in the presence of oxygen. Nevertheless, the marked increase of the relative coverages of

410   pBI143 and its host in IBD metagenomes suggest the potential utility of this cryptic plasmid for

411   unbiased measurements of stress. Overall, these results show that both in metagenomes and

412   experimental conditions, an increased copy number of pBI143 is a consistent phenotype in the

413   presence of host stress.

414   DISCUSSION

415   Our work shed lights on a mysterious corner of life in the human gut. Even though pBI143 is found

416   in greater than 90% of all individuals in some countries, the prevalence of this cryptic plasmid has

417   gone unnoticed for almost four decades since its discovery by Smith, Rollins, and Parker [41]. The

418   remarkable ecology, evolution, and potential practical applications of pBI143 that we

419   characterized here through 'omics analyses as well as *in vitro* and *in vivo* laboratory experiments

420   offer a glimpse of the world of understudied cryptic plasmids in the human gut, and elsewhere.

421   The application of population genetics principles to pBI143 through the recovery of single-

422   nucleotide variants (SNVs) and single-amino acid variants (SAAVs) from gut metagenomes

423   reveals not only the strong forces of purifying selection on the evolution of its sequence, but also

424   hints the presence of adaptive processes at localized amino acid positions that are variable in the

425   critical parts of the DNA-interacting residues of the catalytic domain of its mobilization protein.

426   The presence of pBI143 does not appear to systematically impact bacterial host fitness *in vivo*,

427   which makes this cryptic plasmid seem a mundane parasite, somewhat contradicting the strict

428   evolutionary pressures that maintain its environmental sequence variants.

429   That said, our observations from naturally occurring gut environments include cases where pBI143

430   carries additional genes, likely acting as a natural shuttle vector. Although traditionally mobile

431   genetic elements are classified as mutualistic or parasitic with respect to the bacterial host, the

432   fluidity of pBI143 to fluctuate between the cryptic 2-gene state and the larger 3 or more gene state

433   with potentially beneficial functions, suggests that the boundaries between parasitism and

434   mutualism for pBI143 are not clear cut. Instead, pBI143 may act as a 'discretionary parasite',

435   where it has a cryptic form for the majority of its existence in which it could be best described as

436   a parasite, while occasionally being found with additional functions that may be beneficial to its

437   host as a function of environmental pressures. Testing this hypothesis with future experimentation,
438   and if true, investigating to what extent discretionary parasitism applies to cryptic plasmids, may
439   lead to a deeper understanding of the role cryptic plasmids play in microbial fitness to changing
440   environmental conditions.

441   Our findings show that it has important potential practical applications beyond molecular biology.
442   The first and most straightforward of these applications relies on the prevalence and human
443   specificity of pBI143 to more sensitively detect human fecal contamination in water samples.
444   Human fecal pollution is a global public health problem, and accurate and sensitive indicators of
445   human fecal pollution are essential to identify and remediate contamination sources and to protect
446   public health [100]. While culture assays for *E. coli* or enterococci have historically been used to
447   detect human fecal contamination in environmental samples, the common occurrence of these
448   organisms in many different mammalian guts and the poor sensitivity of such assays motivated
449   researchers in the past two decades to utilize PCR amplification of 16S rRNA genes, specifically
450   those from human-specific *Bacteroides* and *Lachnospiraceae* populations, to detect human-
451   specific fecal contamination with minimal cross-reactivity with animal feces [79,80]. Our
452   benchmarking of pBI143 with qPCR revealed that pBI143 is an extremely sensitive and specific
453   marker of human fecal contamination that typically occurs in human fecal samples and sewage in
454   numbers that are several-fold higher than the state-of-the-art markers, which enabled the
455   quantification of fecal contamination in samples where it had previously gone undetected. Another
456   practical application of pBI143 takes advantage of its natural shuttle vector capabilities to
457   incorporate additional genetic material into its backbone. Our demonstration that pBI143 (1)
458   replicates in many abundant gut microbes, (2) can be stably introduced to new hosts, and (3)
459   naturally acquires genetic material makes this cryptic plasmid an ideal natural payload delivery
460   system for future therapeutics targeting the human gut microbiome. Indeed, our observations of
461   pBI143 with cargo genes in metagenomes indicates that this likely happens in nature. Yet another
462   practical implication of pBI143 is its utility to measure the level of stress in the human gut.
463   Surveying thousands of samples from individuals who are healthy or diagnosed with IBD, our
464   results show that across all bacterial hosts, the approximate copy number of pBI143 increases in
465   individuals with IBD.

466   From a more philosophical point of view, the prevalence and high conservancy of pBI143 across

467   globally distributed human populations questions the traditional definition of the 'core'

468   microbiome [101]. In its aim to define a core microbiome, the field of microbial ecology has primarily

469   focused on bacteria, although sometimes including prevalent archaea or fungi [102–105]. However,

470   our results indicate that there are mobile genetic elements that fit the standard criteria of prevalence

471   to be defined as core. Broadening the definition of a core microbiome beyond microbial taxa may

472   enable the recognition of other mobile genetic elements (eg. plasmids, phages, transposons) that

473   are prevalent across human populations and fill critical gaps in our understanding of gut microbial

474   ecology.

Materials and Methods

476   **Genomes and metagenomes.** We acquired the original pBI143 genome from the National Center for

477   Biotechnological Information (GenBank: U30316.1). We manually assembled the three reference versions

478   of pBI143 (Version 1, 2 and 3) from metagenomes samples USA0006, CHI0054 and ISR0084. We acquired

479   717 human gut isolate genomes from the Duchossois Family Institute collection (Supplementary Table 4).

480   We downloaded 4,516 healthy human adult gut metagenomes from the National Center for Biotechnology

481   Information (NCBI) from (Australia (Accession ID: PRJEB6092), Austria [48], Bangladesh [49], Canada [50],

482   China [51,52], Denmark [53], England [54], Ethiopia [55], Fiji [56], Finland [57], India [58], Israel [59], Italy [60,61], Japan [62],

483   Korea [63], Madagascar [55], Mongolia [55,64], Netherlands [65], Peru [66], Spain [67], Sweden [68], Tanzania [61], and the

484   USA [66,69,70]) (Supplementary Table 1). We acquired 1,096 gut metagenomes from infant-mother pairs from

485   Italy, Finland, Sweden and the USA from NCBI (Supplementary Table 1). We downloaded 935

486   metagenomes from non-human gut environments (marine ecosystems, pet dog guts, monkey guts, sewage,

487   human oral cavity, and human skin) (Supplementary Table 1).

488   **Metagenomic assembly, read recruitment, and the recovery of coverage and detection statistics.**

489   Unless otherwise specified, we performed all metagenomic analyses throughout the manuscript within the

490   open-source anvi'o v7 software ecosystem (https://anvio.org) [106]. We automated assembly and read

491   recruitment steps using the anvi'o metagenomics workflow [107] which used snakemake v5.10 [108]. To quality-

492   filter genomic and metagenomic raw paired-end reads we used illumina-utils v1.4.4 [109] program `iu-filter-

493   quality-minoche` with default parameters, and IDBA_UD v1.1.2 with the flag `--min_contig 1000` to

494   assemble the metagenomes [110]. We used Bowtie2 v2.4 [111] to recruit reads from the metagenomes to

495   reference sequences and samtools v1.9 [112] to convert resulting SAM files into sorted and indexed BAM

496   files. We generated anvi'o contigs databases (https://anvio.org/m/contigs-db) using the command `anvi-

497   gen-contigs-database`, during which Prodigal v2.6.3 [113] identifies open reading frames. We created anvi'o

498   profile databases of the mapping results for each metagenome using `anvi-profile`, which stores coverage

499   and detection statistics, and `anvi-merge` to combine all profiles together. To recover coverage and

500   detection statistics for a given merged profile database, we used the program `anvi-summarize` with `--init-

501   gene-coverages` flag.

502   **Criteria for detection of pBI143 and crAssphage in metagenomes**. Using mean coverage to assess the

503   occurrence of a given sequence in a given sample based on metagenomic read recruitment can yield

504   misleading insights due to non-specific read recruitment (i.e., recruitment of reads from metagenomes to a

505   reference sequence from non-target populations). Thus, we relied upon the detection statistic reported by

506   anvi'o, which is a measure of the proportion of the nucleotides in a given sequence that are covered by at

507   least one short read. We considered pBI143 was present in a metagenome only if its detection value was

508   0.5 or above. Values of detection in metagenomic read recruitment results often follow a bimodal

509   distribution for populations that are present and absent (see Supplementary Fig. 2 in ref. [114]). Thus, 0.5 is a

510   conservative cutoff to minimize a false-positive signal to assume presence.

511   **Distinguishing the presence of distinct pBI143 versions in a genome or metagenome.** We used the

512   results of individual read recruitments to each known version of pBI143 to measure the coverage of each

513   gene in pBI143 in samples that had a detection of greater than 0.9 and compared the ratio of the coverage

514   of each gene. The pBI143 version where the genes have the most even coverage ratio was considered the

515   predominant version in that genome or metagenome.

516   **Addition of tetQ to pIB143.** To study transfer of pBI143 from *Phocaeicola vulgatus* MSK 17.67 to other

517   Bacteroidales species, we added *tetQ* to pBI143. We PCR amplified tetQ from *Bacteroides caccae*

518   CL03T12C61 and inserted it at the site shown in Supplementary Fig. 2 (all primers are listed in

519   Supplementary Table 15). We PCR amplified the DNA regions flanking each side of this insertion site

520   and the three PCR products were cloned into BamHI-digested pLGB13 [115]. We conjugally transferred this

521   plasmid into *Phocaeicola vulgatus* MSK 17.67 and selected cointegrates on gentamycin 200 μg/ml and

522   erythromycin 10 μg/ml. We passaged the cointegrate in non-selective medium and selected the resolvents

523   by plating on anhydrotetracycline (75 ng/ml). We confirmed pIB143 contained *tetQ* by WGS the strain at

524   the DFI Microbiome Metagenomics Facility.

525   **Transfer assays.** The recipient strains that received pBI143-*tetQ* were *Parabacteroides johnsonii*

526   CL02T12C29 and *Bacteroides ovatus* D2, both erythromycin resistant and tetracycline sensitive. We grew

527   the donor strain *Phocaeicola vulgatus* MSK 17.67 pBI143-*tetQ* and recipient strains to an OD600 of ~ 0.7

528 and mixed them at a 10:1 ratio (v:v) donor to recipient, and spotted 10 µl onto BHIS plates and grew them

529 anaerobically for 20 h. We resuspended the co-culture spot in 1 mL basal media and cultured 10-fold serial

530 dilutions on plates with erythromycin (to calculate number of recipients) or erythromycin and tetracycline

531 (4.5 µg/ml) (to select for transconjugants). We performed multiplex PCR as described [116,117] to confirm that

532 TetR ErmR colonies were the recipient strain containing pBI143-tetQ (Supplementary Fig. 2).

533 **Calculations of purifying selection and characterization of single nucleotide variants across**

534 **metagenomes.** We calculated dN/dS ratios as described previously [84]; details of which can also be found

535 at   https://merenlab.org/data/anvio-structure/chapter-IV/#calculating-dndstextgene-for-1-gene.   To

536 determine the mutational landscape of pBI143 across metagenomes, we first identified all variable positions

537 present in the reference pBI143 sequences. We used the program `anvi-script-gen-short-reads` to generate

538 artificial short reads from the version 2 and version 3 pBI143 sequences and recruited these reads to the

539 version 1 pBI143 sequence to generate data similar to the read recruitment from metagenomes. Then, we

540 took read recruitment data from the global human gut metagenomes and sewage metagenomes mapped to

541 version 1 pBI143. We ran `anvi-gen-variability-profile` on the artificial read recruitment profile databases

542 as well as on all profile databases from metagenomes with greater than 10X Q2Q3 coverage to identify all

543 SNV positions. We compared the SNV positions in each gut or sewage metagenome to those present in our

544 reference sequences and calculated the number of SNVs in each metagenome that did and did not match

545 SNVs in the references. To calculate the number of non-consensus SNVs in a metagenome, we again ran

546 the command `anvi-gen-profile-database` on the same metagenomes, this time with the flags `--gene-caller-

547 ids 0`,  `--min-departure-from-consensus 0.1`, `--include-contig-names` and `--quince-mode`, which

548 produces a file that describes the variation in every single position across the reference and calculates the

549 departure from consensus for each SNV with a departure from consensus greater than 0.1.

550 **pBI143 structural and polymorphism analysis.** To explore the impact of SAAVs on the protein structure

551 of pBI143 MobA, we *de novo* predicted the monomer and dimer structures using AlphaFold 2 (AF) in

552 ColabFold with default settings [83]. AlphaFold 2 confidently predicted the structure of the catalytic domain

553 but had low pLDDT scores for the coil domains and the dimer interactions. However, we explored variants

554 across the whole dimer complex. Next, we integrated the pBI143 MobA AF structure into anvi'o structure

555 by running 'anvi-gen-structure-database' [82]. After that, we summarized SNV data as SAAVs from the

556 metagenomic read recruitment data using 'anvi-gen-variability-profile --engine AA' to create a variability

557 profile (https://anvio.org/m/variability-profile). Subsequently, we superimposed the SAAV data variability

558 profile on the structure with 'anvi-display-structure' which filtered for variants that had at least 0.05

559 departure from consensus (reducing our metagenomic samples size from 2221 to 1706). Finally, we

560 analyzed SAAVs that were prevalent in at least 5% of remaining samples. This left us with 21 SAAVs to

561   analyze on the monomer. Next, we explored the relationship between SAAVs, relative solvent accessibility
562   (RSA), and ligand binding residues in pBI143 MobA. To do this, we identified the homologous structure
563   PDB 4LVI (MobM) by searching the high pLDDT pBI143 AF domain against the structure database
564   PDB100 2201222 using Foldseek (https://search.foldseek.com/search). We next structurally aligned the
565   pBI143 MobA AF structure to PDB 4LVI (MobM) [118] using PyMol [119]. We chose the MobM structure
566   4LVI rather than a MobA because it had more structural and sequence homology to the pBI143 MobA
567   catalytic domain AF structure than any PDB MobA structures. Additionally, we leveraged residue
568   conservation values from the pre-calculated 4LVI ConSurf analysis to further explore ligand binding
569   residues [120,121].

570   **Phylogenetic tree construction.** To construct the pBI143 phylogeny, we identified pBI143 contigs from
571   the isolate genome assemblies (Supplementary Table 4) using BLAST [122]. We ran 'anvi-gen-contigs-
572   database' on each pBI143 contig followed by 'anvi-export-gene-calls' with the flag '--gene-caller prodigal'
573   and concatenated the resulting amino acid sequences. For the bacterial host phylogeny, we ran 'anvi-gen-
574   contigs-database' on each assembled genome. Then, we extracted ribosomal genes (see Supplementary
575   Methods for details), aligned them with MUSCLE v3.8.1551 [123], trimmed the alignments with trimAl [124]
576   using the flag '-gt 0.5', and computed the phylogeny with IQ-TREE 2.2.0-beta using the flags '-m MFP'
577   and '-bb 1000' [125]. We visualized the trees with 'anvi-interactive' in '--manual-mode', and used the
578   metadata provided by the Duchossois Family Institute to label the isolates to their corresponding donors.
579   We used the `geom_alluvium` function in ggplot2 to make the alluvial plots..

580   **Construction and analysis of the network that describes shared single-nucleotide variants across**
581   **mothers and infants.** To investigate whether single-nucleotide variants suggest a vertical transmission of
582   pBI143, we used metagenomic read recruitment results from four independent study that generated
583   metagenomic sequencing of fecal samples collected from mothers and their infants in Finland [57], Italy [60],
584   Sweden [68], and the USA [69], against the pBI143 Version 1 reference sequence. The URL
585   https://merenlab.org/data/pBI143 serves a fully reproducible workflow of this analysis. The primary input
586   for this investigation was the anvi'o variability data, which is calculated by the anvi'o program `anvi-
587   profile`, and reported by the anvi'o program `anvi-gen-variability-profile` (with the flag `--engine NT`).
588   The program `anvi-gen-variability-profile` (https://anvio.org/m/anvi-gen-variability-profile) offers a
589   comprehensive description of the single-nucleotide variants in metagenomes for downstream analyses.
590   Since the *mobA* gene was conserved enough to represent all three versions of pBI143, for downstream
591   analyses we limited the context to study variants to the mobA gene. The total number of samples in the
592   entire dataset with at least one variable nucleotide position was 309, which represented a total of 102
593   families (Supplementary Table 4) (Sweden: 52, USA: 24, Finland: 14, Italy: 11). We removed any sample that did not belong to a

594   minimal complete family (i.e., at least one sample for the mother, and at least one sample of her infant),

595   which reduced the number of families in which both members are represented to 57 families (Sweden: 36,

596   USA: 16, Finland: 3, Italy: 2). We further removed families if the coverage of the *mobA* gene was not 50X

597   or more in at least one mother and one infant sample in the family, which reduced the number of families

598   with both members represented and with a reliable coverage of mobA to 49 families (Sweden: 33, USA:

599   13, Finland: 2, Italy: 1), and from a given family, we only used the samples that had at least 50X for

600   downstream analyses. We subsampled the variability data in R to only include the variable nucleotide

601   position data for the final list of samples. We then used the list of single-nucleotide variants reported in this

602   file to generate a network description of these data using the program `anvi-gen-variability-network`, which

603   reports an 'edge' between any sample pairs that share a SNV with the same competing nucleotides. We then

604   used Gephi [126], an open-source network visualization program, with the ForceAtlas2 algorithm [127] to

605   visualize the network. To quantify the extent of similarity between family members based on single-

606   nucleotide patterns in the data, we generated a distance matrix from the same dataset using the `pdist`

607   function in Python's standard library with `cosine` distances. We calculated the average distance of each

608   sample to all other samples in its familial group (`within distance`), as well as the average distance from

609   each sample to all other samples not present in their familial group (`between distance`). We subtracted the

610   within distance from the between distance to get the `subtracted distance`.

611   **Metagenomic taxonomy estimation.** We used Kraken 2.0.8-beta with the flags `--output`, `--report`, `--

612   use-mpa-style`, `--quick`, `--use-names`, `--paired` and `--classified-out` to estimate taxonomic

613   composition of each metagenome [128]. For the genus-level taxonomic data, we filtered for metagenomes

614   where the total number of reads recruited to a *Bacteroides*, *Parabacteroides* or *Phocaeicola* genome was

615   >1000 and the mean coverage of pBI143 was >20X. For the species-level taxonomic data, we used a cutoff

616   of >0.1% percent of reads recruited to designate presence or absence of *B. fragilis* and >0.0001% for pBI143

617   based on the sizes of the genomes respectively (the *B. fragilis* genome is 3 orders of magnitude larger than

618   pBI143).

619   **Isogenic strain construction.** We constructed the plasmid vector pEF108 (as shown in Supp Fig.

620   pEF108_plasmid_map) by PCR amplifying the desired sections with primers vec_108F, vec_108R,

621   frag1_108F, frag1_108R, frag2_108R and frag2_108R (Supplementary Table 15) from existing

622   plasmids. We assembled the three fragments via Gibson assembly using standard conditions described for

623   NEB Gibson assembly mastermix. We selected for transconjugants on LB-carbenicillin (100ug/mL), then

624   conjugated pEF108 into B. fragilis 638R and selected on BHIS + erythromycin 25ug/mL. Then, we counter-

625   selected for recombination events in pEF108 to remove the markers and leave naive pBI143 by growing

626   cells on Bacteroides minimal media plates (BMM) with 10mM p-chlorophenylalanine. We screened

627 pBI143 positive, pheS-negative colonies via PCR and confirmed them by WGS. See Supplementary

628 Methods for details.

629 *In vitro* **competition experiments.** We grew each strain described above in BHIS to OD 0.6-0.8. We

630 combined equal volumes of cells and plated these cells on BHIS plates. We added 50µL of the combined

631 strains to 5mL BHIS and grew these cultures to OD 0.6-0.8, then plated again on BHIS plates. We replica

632 plated all colonies from BHIS to BHIS supplemented with cefoxitin (15ug/mL) or erythromycin (10 ug/mL)

633 and counted the resulting colonies to determine the starting and final ratios of each strain.

634 **Mouse competitive colonization assays.** All animal experimentation was approved by the Institutional

635 Animal Care and Use Committee at the University of Chicago. We gavaged three male and three female

636 10-15 week old germ-free C57BL/6J mice with a 1:1 inoculum of *B. fragilis* 638R:*B. fragilis* 638R pBI143.

637 Males and females were housed separately in isocages and remained gnotobiotic for the duration of the

638 experiment. We collected fecal pellets after eight and 14 days, diluted and plated on BHIS plates. We

639 performed PCR on 48 colonies per mouse using a mixture of four primers (Supplementary Table 15), one

640 set that amplifies a 1248-bp region of the 638R chromosome and a second set that amplifies a 662-bp

641 segment of pBI143. PCR amplicons from all colonies included the 1248-bp region of the 638R chromosome

642 and a subset also contained the amplicon for pBI143, allowing calculation of the ratio over time. The exact

643 starting ratio for gavage was also calculated using this same PCR.

644 **Approximate copy number ratio calculation in metagenomes.** The first challenge to use metagenomic

645 coverage values to study pBI143 copy number trends in human gut metagenomes is the unambiguous

646 identification of gut metagenomes that appear to have a single possible pBI143 bacterial host beyond

647 reasonable doubt. To establish insights into the taxonomic make up of the gut metagenomes we previously

648 assembled, we first ran the program `anvi-estimate-scg-taxonomy` (https://anvio.org/m/anvi-estimate-scg-

649 taxonomy) with the flags `--metagenome-mode` (to profile every single single-copy core gene (SCG)

650 independently) and `--compute-scg-coverages` (to compute coverages of each SCG from the read

651 recruitment results). We also used the flag `--scg-name-for-metagenome-mode` to limit the search space

652 for a single ribosomal protein. We used the following list of ribosomal proteins for this step as they are

653 included among the SCGs anvi'o assigns taxonomy using GTDB, and we merged resulting output files:

654 Ribosomal_S2, Ribosomal_S3_C, Ribosomal_S6, Ribosomal_S7, Ribosomal_S8, Ribosomal_S9,

655 Ribosomal_S11, Ribosomal_S20p, Ribosomal_L1, Ribosomal_L2, Ribosomal_L3, Ribosomal_L4,

656 Ribosomal_L6, Ribosomal_L9_C, Ribosomal_L13, Ribosomal_L16, Ribosomal_L17, Ribosomal_L20,

657 Ribosomal_L21p, Ribosomal_L22,ribosomal_L24, and Ribosomal_L27A. For our downstream analyses

658 that relied upon the merged SCG taxonomy and coverage output reported by anvi'o, we considered

659 *Bacteroides, Parabacteroides* and *Phocaeicola* as the genera for candidate pBI143 host 'species', and only

660 considered metagenomes in which a single species from these genera was present. Our determination of

661 whether or not a single species of these genera was present in a given metagenome relied on the coverage

662 of species-specific single-copy core genes (SCGs), where the taxonomic assignment to a given SCG

663 resolved all the way down to the level of species unambiguously. We excluded any metagenome from

664 further consideration if three or more candidate host species had positive coverage in any SCG in a

665 metagenome. Due to highly conserved nature of ribosomal proteins and bioinformatics artifacts, it is

666 possible that even when a single species is present in a metagenome, one of its ribosomal proteins may

667 match to a different species in the same genus given the limited representation of genomes in public

668 databases compared to the diversity of environmental populations. So, to minimize the removal of

669 metagenomes from our analysis, we took extra caution with metagenomes before discarding them if only

670 two candidate host species had positive coverage in any SCG. We kept such a metagenome in our

671 downstream analyses only if one species was detected with only a single SCG, and the other one was

672 detected by at least 8. In this case we assumed the large representation of one species (with 8 or more

673 ribosomal genes) suggests the presence of this organism in this habitat confidently, and assumed the single

674 hit to another species within the same genus was likely due to bioinformatics artifacts. It is the most

675 unambiguous case if only a single candidate host species was detected in a given metagenome, but we still

676 removed a given metagenome from further consideration if that single species had 3 or fewer SCGs in the

677 metagenome. These criteria deemed 584 of 2580 metagenomes to have an unambiguous pBI143 host that

678 resolved to 21 distinct species names. We further removed from our modeling the metagenomes where the

679 candidate host species did not occur in any other metagenome, which removed 5 of these candidate host

680 species from further consideration. Finally, we further removed any metagenome in which the pBI143

681 coverage was less than 5X. Our final dataset to calculate the "approximate copy number ratio" (ACNR) of

682 pBI143 in metagenomes through coverage ratios contained 579 metagenomes with one of 16 unambiguous

683 pBI143 hosts. We calculated the ACNR by dividing the observed coverage of pBI143 by the empirical

684 mean coverage of the host by averaging the coverage of all host SCGs found in the metagenome. To

685 estimate the multiplicative difference in the geometric mean ACNR, we fit a linear model for the expected

686 value of the logarithm of the ACRN, with disease status and bacterial host as predictors using rigr to

687 construct the interval and estimate [129] .

688 **Oxidative stress experiments.** We grew *B. fragilis* 214 in 5 mL BHIS for 15 hours in an anaerobic

689 chamber. We inoculated 750 µL of this culture into 30mL BHIS in quintuplicate, and grew them for 3

690 hours. We divided the 30 mL into a further 5 culture flasks of 5 mL BHIS, and exposed each to oxygen

691 with constant shaking for the appropriate time before returning the flask to the anaerobic chamber. At each

27

692   time point, we took an aliquot of culture to determine the copy number of pBI143 in that sample. We

693   extracted DNA from the cultures using a Thermal NaOH preparation [130] to prepare them for qPCR. Copy

694   number calculated can be found in Supplementary Table 13.

695   **Estimating the pBI143 Plasmid Copy Number by Real-time qPCR.** To evaluate plasmid copy number

696   (CN), we developed a real-time TaqMan probe multiplex PCR assay to amplify both pBI143 and a single-

697   copy *B. fragilis*-specific genomic reference gene (referred to as hsp [heat shock protein]) in the same

698   reaction (see Supplementary Information for details). We confirmed the primer and probe specificity to *B.*

699   *fragilis* with BLAST searches against the NCBI and Ensembl databases, and experimental validation on 45

700   common gut isolates. For absolute quantification, we constructed a standard curve for each gene of interest

701   by plotting the mean quantification cycle (Cq) values against log[quantity] of a dilution series of known

702   gene of interest amount (range: $3×10^0$ to $3×10^6$ copies/reaction). We calculated the CN of pBI143 per

703   genome equivalent (hsp), by dividing the absolute quantity of plasmid target by the absolute quantity of

704   chromosomal target in the sample using the standard-curve (SC) method of absolute quantification [131].

705   Standard curves were generated with every qPCR run for analysis and to confirm PCR efficiency.

706   Additional details for qPCR, including standard curves and controls, can be found in Supplemental

707   Information. Supplementary Table 5 and Supplementary Table 15 report the relevant data and all primers,

708   respectively.

709   **qPCR analysis of animal, untreated sewage and water samples.** Samples were tested with the pBI143

710   assay and two established assays for human fecal markers that included HF183 and Lachno3 [132]. For

711   screening of animal samples to assess the presence of this plasmid in non-human gut microbiomes, archived

712   DNA from a previous study [133] was analyzed and included 14 different animals encompassing 81 individual

713   fecal samples. For assessment of fecal contamination of surface waters, archived DNA from 40 samples of

714   river water [134–136] and freshwater beaches [137] were analyzed. These water samples were chosen from these

715   previous studies that represented a range of contamination based on HF183 and Lachno3 levels. A total of

716   20 archived untreated sewage samples as reported in Olds et al. [132] were also analyzed for comparison.

717   Since we were using archived samples from previous studies, we retested all the samples for the two human

718   markers to account for any degradation. Additional details for qPCR, including standard curves and

719   controls, can be found in the Supplemental Information.

720   **Visualizations**. We used ggplot2 [138] to generate all box and scatter plots. We generated coverage plots

721   using anvi'o, with the program `anvi-script-visualize-split-coverages`. We finalized the figures for

722   publication using Inkscape, an open-source vector graphics editor (available from http://inkscape.org/).

Data availability

All genomes and metagenomes are available via the NCBI Sequence Read Archive, and the accession numbers for metagenomes and genomes are reported in Supplementary Table 1 and Supplementary Table 4, respectively. The data object identifier (DOI) 10.6084/m9.figshare.22336666 gives access to Supplementary Table and Supplementary Information files. Additional DOIs for anvi'o data products that describe metagenomic read recruitment results as well as sequences for pBI143 versions and bioinformatics workflows are accessible at the URL https://merenlab.org/data/pBI143 to reproduce our findings. Bacterial cultures for host range investigations, which are listed in Supplementary Table 4, are courtesy of The Duchossois Family Institute (https://dfi.uchicago.edu/). *B. fragilis* strains with pBI143 are available upon request from the Comstock Lab collection (https://comstocklab.uchicago.edu/).

Acknowledgements

Author contributions

ECF and AME conceived the study. KL developed methodology. RM, EK, and AME developed computational analysis tools. ECF, MSS, PAR, ADW and AME performed formal analyses. ECF, KL, MS and LEC conducted investigations. TM, MKY, MM and SLM provided resources. ECF, MSS, ADW and AME curated data. ECF and AME prepared the figures. ECF and AME wrote the paper with critical input from all authors. LEC and AME supervised the project. EBC and AME acquired funding.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. *3*, 722–732.

2. Black, B.E. (2017). Centromeres and Kinetochores: Discovering the Molecular Mechanisms Underlying Chromosome Inheritance (Springer).

3. Kazlauskas, D., Varsani, A., Koonin, E.V., and Krupovic, M. (2019). Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. Nat. Commun. *10*, 3425.

4. Solar, G. del, del Solar, G., Giraldo, R., Ruiz-Echevarría, M.J., Espinosa, M., and Díaz-Orejas, R. (1998). Replication and Control of Circular Bacterial Plasmids. Microbiology and Molecular Biology Reviews *62*, 434–464. 10.1128/mmbr.62.2.434-464.1998.

5. Garoña, A., and Dagan, T. (2021). Darwinian individuality of extrachromosomal genetic elements calls for population genetics tinkering. Environ. Microbiol. Rep. *13*, 22–26.

6. Jacob, A.E., and Hobbs, S.J. (1974). Conjugal transfer of plasmid-borne multiple antibiotic resistance in Streptococcus faecalis var. zymogenes. J. Bacteriol. *117*, 360–372.

7. Moo-Young, M., Anderson, W.A., and Chakrabarty, A.M. (2013). Environmental Biotechnology: Principles and Applications (Springer Science & Business Media).

8. Endo, G., Ji, G., and Silver, S. (1995). Heavy Metal Resistance Plasmids and Use in Bioremediation. Environmental Biotechnology, 47–62. 10.1007/978-94-017-1435-8_5.

775   9.   Thouand, G., and Marks, R. (2016). Bioluminescence: Fundamentals and Applications in
776   Biotechnology - Volume 3 (Springer).

777   10.   Palomino, A., Gewurz, D., DeVine, L., Zajmi, U., Moralez, J., Abu-Rumman, F., Smith,
778   R.P., and Lopatkin, A.J. (2022). Metabolic genes on conjugative plasmids are highly
779   prevalent in Escherichia coli and can protect against antibiotic treatment. ISME J.
780   10.1038/s41396-022-01329-1.

781   11.   Al-Shayeb, B., Schoelmerich, M.C., West-Roberts, J., Valentin-Alvarado, L.E., Sachdeva,
782   R., Mullen, S., Crits-Christoph, A., Wilkins, M.J., Williams, K.H., Doudna, J.A., et al.
783   (2022). Borgs are giant genetic elements with potential to expand metabolic capacity.
784   Nature *610*, 731–736.

785   12.   Leonard, S.P., Perutka, J., Powell, J.E., Geng, P., Richhart, D.D., Byrom, M., Kar, S.,
786   Davies, B.W., Ellington, A.D., Moran, N.A., et al. (2018). Genetic Engineering of Bee Gut
787   Microbiome Bacteria with a Toolkit for Modular Assembly of Broad-Host-Range Plasmids.
788   ACS Synth. Biol. *7*, 1279–1290.

789   13.   Slattery, S.S., Diamond, A., Wang, H., Therrien, J.A., Lant, J.T., Jazey, T., Lee, K.,
790   Klassen, Z., Desgagné-Penix, I., Karas, B.J., et al. (2018). An Expanded Plasmid-Based
791   Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic
792   Pathways in Phaeodactylum tricornutum. ACS Synth. Biol. *7*, 328–338.

793   14.   Rihn, S.J., Merits, A., Bakshi, S., Turnbull, M.L., Wickenhagen, A., Alexander, A.J.T.,
794   Baillie, C., Brennan, B., Brown, F., Brunker, K., et al. (2021). A plasmid DNA-launched
795   SARS-CoV-2 reverse genetics system and coronavirus toolkit for COVID-19 research.
796   PLoS Biol. *19*, e3001091.

797   15.   Salvay, D.M., Zelivyanskaya, M., and Shea, L.D. (2010). Gene delivery by surface
798   immobilization of plasmid to tissue-engineering scaffolds. Gene Ther. *17*, 1134–1141.

799   16.   Mutuku, C., Gazdag, Z., and Melegh, S. (2022). Occurrence of antibiotics and bacterial
800   resistance genes in wastewater: resistance mechanisms and antimicrobial resistance control
801   approaches. World J. Microbiol. Biotechnol. *38*, 1–27.

802   17.   Dimitriu, T. (2022). Evolution of horizontal transmission in antimicrobial resistance
803   plasmids. Microbiology *168*, 001214.

804   18.   Prestinaci, F., Pezzotti, P., and Pantosti, A. (2015). Antimicrobial resistance: a global
805   multifaceted phenomenon. Pathog. Glob. Health *109*, 309–318.

806   19.   Kang, X., Li, C., and Luo, Y. (2020). Cloning of pAhX22, a small cryptic plasmid from
807   Aeromonas hydrophila, and construction of a pAhX22-derived shuttle vector. Plasmid *108*,
808   102490.

809   20.   Oliveira, V., Polónia, A.R.M., Cleary, D.F.R., Huang, Y.M., de Voogd, N.J., da Rocha,
810   U.N., and Gomes, N.C.M. (2021). Characterization of putative circular plasmids in sponge-
811   associated bacterial communities using a selective multiply-primed rolling circle

812   amplification. Mol. Ecol. Resour. *21*, 110–121.

813 21. Shareck, J., Choi, Y., Lee, B., and Miguez, C.B. (2004). Cloning vectors based on cryptic
814   plasmids isolated from lactic acid bacteria: their characteristics and potential applications in
815   biotechnology. Crit. Rev. Biotechnol. *24*, 155–208.

816 22. Attéré, S.A., Vincent, A.T., Paccaud, M., Frenette, M., and Charette, S.J. (2017). The Role
817   for the Small Cryptic Plasmids As Moldable Vectors for Genetic Innovation in Aeromonas
818   salmonicida subsp. salmonicida. Frontiers in Genetics *8*. 10.3389/fgene.2017.00211.

819 23. Challacombe, J.F., Pillai, S., and Kuske, C.R. (2017). Shared features of cryptic plasmids
820   from environmental and pathogenic Francisella species. PLoS One *12*, e0183554.

821 24. Roberts, M.C. (1989). Plasmids of Neisseria gonorrhoeae and other Neisseria species.
822   Clinical Microbiology Reviews *2*. 10.1128/cmr.2.suppl.s18.

823 25. Zillig, W., Prangishvilli, D., Schleper, C., Elferink, M., Holz, I., Albers, S., Janekovic, D.,
824   and Götz, D. (1996). Viruses, plasmids and other genetic elements of thermophilic and
825   hyperthermophilic Archaea. FEMS Microbiol. Rev. *18*, 225–236.

826 26. Heuer, H., and Smalla, K. (2012). Plasmids foster diversification and adaptation of bacterial
827   populations in soil. FEMS Microbiol. Rev. *36*, 1083–1104.

828 27. Vincent, A.T., Hosseini, N., and Charette, S.J. (2021). The *Aeromonas salmonicida*
829   plasmidome: a model of modular evolution and genetic diversity. Annals of the New York
830   Academy of Sciences *1488*, 16–32. 10.1111/nyas.14503.

831 28. Thomas, C.M. (2014). Evolution and Population Genetics of Bacterial Plasmids. Plasmid
832   Biology, 507–528. 10.1128/9781555817732.ch25.

833 29. Iranzo, J., Puigbò, P., Lobkovsky, A.E., Wolf, Y.I., and Koonin, E.V. (2016). Inevitability
834   of Genetic Parasites. Genome Biology and Evolution *8*, 2856–2869. 10.1093/gbe/evw193.

835 30. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun
836   metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844.

837 31. Andreopoulos, W.B., Geller, A.M., Lucke, M., Balewski, J., Clum, A., Ivanova, N.N., and
838   Levy, A. (2022). Deeplasmid: deep learning accurately separates plasmids from bacterial
839   chromosomes. Nucleic Acids Res. *50*, e17.

840 32. Krawczyk, P.S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid
841   sequences in metagenomic data using genome signatures. Nucleic Acids Res. *46*, e35.

842 33. Zhou, F., and Xu, Y. (2010). cBar: a computer program to distinguish plasmid-derived from
843   chromosome-derived sequence fragments in metagenomics data. Bioinformatics *26*, 2051–
844   2052.

845 34. Pellow, D., Mizrahi, I., and Shamir, R. (2020). PlasClass improves plasmid sequence

32

846   classification. PLoS Comput. Biol. *16*, e1007781.

847   35. Carattoli, A., Zankari, E., García-Fernández, A., Larsen, M.V., Lund, O., Villa, L.,
848   Aarestrup, F.M., and Hasman, H. (2014). *In Silico* Detection and Typing of Plasmids using
849   PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrobial Agents and
850   Chemotherapy *58*, 3895–3903. 10.1128/aac.02412-14.

851   36. Robertson, J., and Nash, J.H.E. (2018). MOB-suite: software tools for clustering,
852   reconstruction and typing of plasmids from draft assemblies. Microb Genom *4*.
853   10.1099/mgen.0.000206.

854   37. Garcillán-Barcia, M.P., Francia, M.V., and de la Cruz, F. (2009). The diversity of
855   conjugative relaxases and its application in plasmid classification. FEMS Microbiol. Rev.
856   *33*, 657–687.

857   38. Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., and Shamir,
858   R. (2017). Recycler: an algorithm for detecting plasmids from de novo assembly graphs.
859   Bioinformatics *33*, 475–482.

860   39. Pellow, D., Zorea, A., Probst, M., Furman, O., Segal, A., Mizrahi, I., and Shamir, R. (2021).
861   SCAPP: an algorithm for improved plasmid assembly in metagenomes. Microbiome *9*, 144.

862   40. Yu, M.K., Fogarty, E.C., and Murat Eren, A. The genetic and ecological landscape of
863   plasmids in the human gut. 10.1101/2020.11.01.361691.

864   41. Smith, C.J., Rollins, L.A., and Parker, A.C. (1995). Nucleotide sequence determination and
865   genetic analysis of the Bacteroides plasmid, pBI143. Plasmid *34*, 211–222.

866   42. Smith, C.J. (1985). Development and use of cloning systems for Bacteroides fragilis:
867   cloning of a plasmid-encoded clindamycin resistance determinant. J. Bacteriol. *164*, 294–
868   301.

869   43. Tan, H., Zhao, J., Zhang, H., Zhai, Q., and Chen, W. (2019). Novel strains of Bacteroides
870   fragilis and Bacteroides ovatus alleviate the LPS-induced inflammation in mice. Appl.
871   Microbiol. Biotechnol. *103*, 2353–2365.

872   44. Lee, Y.K., Mehrabian, P., Boyajian, S., Wu, W.-L., Selicha, J., Vonderfecht, S., and
873   Mazmanian, S.K. (2018). The Protective Role of *Bacteroides fragilis* in a Murine Model of
874   Colitis-Associated Colorectal Cancer. mSphere *3*. 10.1128/msphere.00587-18.

875   45. Ochoa-Repáraz, J., Mielcarz, D.W., Wang, Y., Begum-Haque, S., Dasgupta, S., Kasper,
876   D.L., and Kasper, L.H. (2010). A polysaccharide from the human commensal Bacteroides
877   fragilis protects against CNS demyelinating disease. Mucosal Immunol. *3*, 487–495.

878   46. Purcell, R.V., Pearson, J., Aitchison, A., Dixon, L., Frizelle, F.A., and Keenan, J.I. (2017).
879   Colonization with enterotoxigenic Bacteroides fragilis is associated with early-stage
880   colorectal neoplasia. PLoS One *12*, e0171602.

881   47.   Haghi, F., Goli, E., Mirzaei, B., and Zeighami, H. (2019). The association between fecal
882   enterotoxigenic B. fragilis with colorectal cancer. BMC Cancer *19*, 879.

883   48.   Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X.,
884   Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma–carcinoma
885   sequence. Nat. Commun. *6*, 1–13.

886   49.   David, L.A., Weil, A., Ryan, E.T., Calderwood, S.B., Harris, J.B., Chowdhury, F., Begum,
887   Y., Qadri, F., LaRocque, R.C., and Turnbaugh, P.J. (2015). Gut microbial succession
888   follows acute secretory diarrhea in humans. MBio *6*, e00381–15.

889   50.   Raymond, F., Ouameur, A.A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P.,
890   Plante, P.-L., Giroux, R., Bérubé, È., et al. (2015). The initial state of the human gut
891   microbiome determines its reshaping by antibiotics. ISME J. *10*, 707–720.

892   51.   Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D.,
893   et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes.
894   Nature *490*, 55–60.

895   52.   Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y.,
896   Zhang, L., et al. (2017). Quantitative metagenomics reveals unique gut microbiome
897   biomarkers in ankylosing spondylitis. Genome Biol. *18*, 1–13.

898   53.   Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M.,
899   Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut
900   microbiome correlates with metabolic markers. Nature *500*, 541–546.

901   54.   Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on
902   the Gut Microbiome (2016). Cell Systems *3*, 572–584.e3.

903   55.   Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes
904   from Metagenomes Spanning Age, Geography, and Lifestyle (2019). Cell *176*, 649–
905   662.e20.

906   56.   Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W.,
907   Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human
908   microbiome are structured from global to individual scales. Nature *535*, 435–439.

909   57.   Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few
910   Months of Life (2018). Cell Host Microbe *24*, 146–154.e4.

911   58.   Dhakan, D.B., Maji, A., Sharma, A.K., Saxena, R., Pulikkan, J., Grace, T., Gomez, A.,
912   Scaria, J., Amato, K.R., and Sharma, V.K. (2019). The unique composition of Indian gut
913   microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-
914   omics approaches. Gigascience *8*, giz004.

915   59.   Personalized Nutrition by Prediction of Glycemic Responses (2015). Cell *163*, 1079–1094.

916   60.   Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong,
917   D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from
918   Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe *24*,
919   133.

920   61.   Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota (2015). Curr. Biol.
921   *25*, 1682–1693.

922   62.   Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe,
923   H., Masuda, K., Nishimoto, Y., Kubo, M., et al. (2019). Metagenomic and metabolomic
924   analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer.
925   Nat. Med. *25*, 968–976.

926   63.   Kim, C.Y., Lee, M., Yang, S., Kim, K., Yong, D., Kim, H.R., and Lee, I. (2021). Human
927   reference gut microbiome catalog including newly assembled genomes from under-
928   represented Asian metagenomes. Genome Med. *13*, 134.

929   64.   Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou,
930   B., et al. (2016). Unique Features of Ethnic Mongolian Gut Microbiome revealed by
931   metagenomic analysis. Sci. Rep. *6*, 1–13.

932   65.   Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T.,
933   Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., et al. (2016). Population-based
934   metagenomics analysis reveals markers for gut microbiome composition and diversity.
935   Science *352*, 565–569.

936   66.   Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell,
937   L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence
938   strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 1–9.

939   67.   Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R.,
940   Prifti, E., Nielsen, T., et al. (2014). An integrated catalog of reference genes in the human
941   gut microbiome. Nat. Biotechnol. *32*, 834–841.

942   68.   Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life
943   (2015). Cell Host Microbe *17*, 690–703.

944   69.   Lou, Y.C., Olm, M.R., Diamond, S., Crits-Christoph, A., Firek, B.A., Baker, R., Morowitz,
945   M.J., and Banfield, J.F. (2021). Infant gut strain persistence is associated with maternal
946   origin, phylogeny, and traits including surface adhesion and iron acquisition. Cell Reports
947   Medicine *2*. 10.1016/j.xcrm.2021.100393.

948   70.   A framework for human microbiome research (2012). Nature *486*, 215–221.

949   71.   Gupta, V.K., Paul, S., and Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific
950   Variations in Human Microbiome Composition and Diversity. Frontiers in Microbiology *8*.
951   10.3389/fmicb.2017.01162.

72. Yutin, N., Makarova, K.S., Gussow, A.B., Krupovic, M., Segall, A., Edwards, R.A., and Koonin, E.V. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol *3*, 38–46.

73. Amato, K.R., Yeoman, C.J., Kent, A., Righini, N., Carbonero, F., Estrada, A., Rex Gaskins, H., Stumpf, R.M., Yildirim, S., Torralba, M., et al. (2013). Habitat degradation impacts black howler monkey (Alouatta pigra) gastrointestinal microbiomes. ISME J. *7*, 1344–1353.

74. Iino, T., Mori, K., Itoh, T., Kudo, T., Suzuki, K.-I., and Ohkuma, M. (2014). Description of Mariniphaga anaerophila gen. nov., sp. nov., a facultatively aerobic marine bacterium isolated from tidal flat sediment, reclassification of the Draconibacteriaceae as a later heterotypic synonym of the Prolixibacteraceae and description of the family Marinifilaceae fam. nov. Int. J. Syst. Evol. Microbiol. *64*, 3660–3667.

75. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science *348*, 1261359.

76. Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., Jeanthon, C., Rahav, E., Ullrich, M., et al. (2015). The ocean sampling day consortium. Gigascience *4*, 27.

77. Coelho, L.P., Kultima, J.R., Costea, P.I., Fournier, C., Pan, Y., Czarnecki-Maulden, G., Hayward, M.R., Forslund, S.K., Schmidt, T.S.B., Descombes, P., et al. (2018). Similarity of the dog and human gut microbiomes in gene content and response to diet. Microbiome *6*, 1–11.

78. Hendriksen, R.S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S.K., Kjeldgaard, J., et al. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat. Commun. *10*, 1–12.

79. Feng, S., Bootsma, M., and McLellan, S.L. (2018). Human-Associated Lachnospiraceae Genetic Markers Improve Detection of Fecal Pollution Sources in Urban Waters. Appl. Environ. Microbiol. *84*. 10.1128/AEM.00309-18.

80. Sauer, E.P., Vandewalle, J.L., Bootsma, M.J., and McLellan, S.L. (2011). Detection of the human specific Bacteroides genetic marker provides evidence of widespread sewage contamination of stormwater in the urban environment. Water Res. *45*, 4081–4091.

81. Pluta, R., Boer, D.R., Lorenzo-Díaz, F., Russi, S., Gómez, H., Fernández-López, C., Pérez-Luque, R., Orozco, M., Espinosa, M., and Coll, M. (2017). Structural basis of a histidine-DNA nicking/joining mechanism for gene transfer and promiscuous spread of antibiotic resistance. Proc. Natl. Acad. Sci. U. S. A. *114*, E6526–E6535.

82. Delmont, T.O., Kiefl, E., Kilinc, O., Esen, O.C., Uysal, I., Rappé, M.S., Giovannoni, S., and Eren, A.M. (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife *8*. 10.7554/eLife.46497.

83. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682.

84. Kiefl, E., Esen, O.C., Miller, S.E., Kroll, K.L., Willis, A.D., Rappé, M.S., Pan, T., and Eren, A.M. (2023). Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. Sci Adv *9*, eabq4632.

85. Lu, Y.B., Datta, H.J., and Bastia, D. (1998). Mechanistic studies of initiator-initiator interaction and replication initiation. EMBO J. *17*, 5192–5200.

86. Debray, R., Herbert, R.A., Jaffe, A.L., Crits-Christoph, A., Power, M.E., and Koskella, B. (2021). Priority effects in microbiome assembly. Nat. Rev. Microbiol. *20*, 109–121.

87. Vatanen, T., Jabbar, K.S., Ruohtula, T., Honkanen, J., Avila-Pacheco, J., Siljander, H., Stražar, M., Oikarinen, S., Hyöty, H., Ilonen, J., et al. (2022). Mobile genetic elements from the maternal microbiome shape infant gut microbial assembly and metabolism. Cell *185*, 4921–4936.e15.

88. Joo, J., Gunny, M., Cases, M., Hudson, P., Albert, R., and Harvill, E. (2006). Bacteriophage-mediated competition in Bordetella bacteria. Proc. Biol. Sci. *273*, 1843–1848.

89. Bondy-Denomy, J., Qian, J., Westra, E.R., Buckling, A., Guttman, D.S., Davidson, A.R., and Maxwell, K.L. (2016). Prophages mediate defense against phage infection through diverse mechanisms. ISME J. *10*, 2854–2866.

90. Mavrich, T.N., and Hatfull, G.F. (2019). Evolution of Superinfection Immunity in Cluster A Mycobacteriophages. mBio *10*. 10.1128/mbio.00971-19.

91. Chen, B., Chen, Z., Wang, Y., Gong, H., Sima, L., Wang, J., Ouyang, S., Gan, W., Krupovic, M., Chen, X., et al. (2020). ORF4 of the Temperate Archaeal Virus SNJ1 Governs the Lysis-Lysogeny Switch and Superinfection Immunity. J. Virol. *94*. 10.1128/JVI.00841-20.

92. Beaber, J.W., Hochhut, B., and Waldor, M.K. (2004). SOS response promotes horizontal dissemination of antibiotic resistance genes. Nature *427*, 72–74.

93. Comeau, A.M., Tétart, F., Trojet, S.N., Prère, M.-F., and Krisch, H.M. (2007). Phage-Antibiotic Synergy (PAS): β-Lactam and Quinolone Antibiotics Stimulate Virulent Phage Growth. PLoS One *2*, e799.

94. Ubeda, C., Maiques, E., Knecht, E., Lasa, I., Novick, R.P., and Penadés, J.R. (2005). Antibiotic-induced SOS response promotes horizontal dissemination of pathogenicity island-encoded virulence factors in staphylococci. Mol. Microbiol. *56*, 836–844.

95. Schumann, J.P., Jones, D.T., and Woods, D.R. (1984). Induction of proteins during phage reactivation induced by UV irradiation, oxygen and peroxide in Bacteroides fragilis. FEMS Microbiol. Lett. *23*, 131–135.

1026 96. Sund, C.J., Rocha, E.R., Tzianabos, A.O., Wells, W.G., Gee, J.M., Reott, M.A., O'Rourke,
1027   D.P., and Smith, C.J. (2008). The Bacteroides fragilis transcriptome response to oxygen and
1028   H2O2: the role of OxyR and its effect on survival and virulence. Mol. Microbiol. *67*, 129–
1029   142.

1030 97. Vineis, J.H., Ringus, D.L., Morrison, H.G., Delmont, T.O., Dalal, S., Raffals, L.H.,
1031   Antonopoulos, D.A., Rubin, D.T., Eren, A.M., Chang, E.B., et al. (2016). Patient-Specific
1032   Bacteroides Genome Variants in Pouchitis. MBio *7*. 10.1128/mBio.01713-16.

1033 98. Baumgart, D.C., and Carding, S.R. (2007). Inflammatory bowel disease: cause and
1034   immunobiology. Lancet *369*, 1627–1640.

1035 99. Graham, D.B., and Xavier, R.J. (2020). Pathway paradigms revealed from the genetics of
1036   inflammatory bowel disease. Nature *578*, 527–539.

1037 100. McLellan, S.L., and Eren, A.M. (2014). Discovering new indicators of fecal pollution.
1038   Trends Microbiol. *22*, 697–706.

1039 101. Neu, A.T., Allen, E.E., and Roy, K. (2021). Defining and quantifying the core microbiome:
1040   Challenges and prospects. Proc. Natl. Acad. Sci. U. S. A. *118*. 10.1073/pnas.2104429118.

1041 102. Aguirre de Cárcer, D. (2018). The human gut pan-microbiome presents a compositional
1042   core formed by discrete phylogenetic units. Sci. Rep. *8*, 14069.

1043 103. Mancabelli, L., Milani, C., Lugli, G.A., Turroni, F., Ferrario, C., van Sinderen, D., and
1044   Ventura, M. (2017). Meta-analysis of the human gut microbiome from urbanized and pre-
1045   agricultural populations. Environ. Microbiol. *19*, 1379–1390.

1046 104. Shetty, S.A., Kuipers, B., Atashgahi, S., Aalvink, S., Smidt, H., and de Vos, W.M. (2022).
1047   Inter-species Metabolic Interactions in an In-vitro Minimal Human Gut Microbiome of
1048   Core Bacteria. npj Biofilms and Microbiomes *8*. 10.1038/s41522-022-00275-2.

1049 105. Nash, A.K., Auchtung, T.A., Wong, M.C., Smith, D.P., Gesell, J.R., Ross, M.C., Stewart,
1050   C.J., Metcalf, G.A., Muzny, D.M., Gibbs, R.A., et al. (2017). The gut mycobiome of the
1051   Human Microbiome Project healthy cohort. Microbiome *5*, 153.

1052 106. Eren, A.M., Murat Eren, A., Kiefl, E., Shaiber, A., Veseli, I., Miller, S.E., Schechter, M.S.,
1053   Fink, I., Pan, J.N., Yousef, M., et al. (2020). Community-led, integrated, reproducible multi-
1054   omics with anvi'o. Nature Microbiology *6*, 3–6. 10.1038/s41564-020-00834-3.

1055 107. Shaiber, A., Willis, A.D., Delmont, T.O., Roux, S., Chen, L.-X., Schmid, A.C., Yousef, M.,
1056   Watson, A.R., Lolans, K., Esen, Ö.C., et al. (2020). Functional and genetic markers of niche
1057   partitioning among enigmatic members of the human oral microbiome. Genome Biol. *21*,
1058   1–35.

1059 108. Köster, J., and Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow
1060   engine. Bioinformatics *28*, 2520–2522.

1061 109. Eren, A.M., Vineis, J.H., Morrison, H.G., and Sogin, M.L. (2013). A filtering method to
1062   generate high quality short reads using illumina paired-end technology. PLoS One *8*,
1063   e66643.

1064 110. Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a *de novo*
1065   assembler for single-cell and metagenomic sequencing data with highly uneven depth.
1066   Bioinformatics *28*, 1420–1428. 10.1093/bioinformatics/bts174.

1067 111. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat.
1068   Methods *9*, 357–359.

1069 112. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
1070   G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence
1071   Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

1072 113. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010).
1073   Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC
1074   Bioinformatics *11*, 119.

1075 114. Utter, D.R., Borisy, G.G., Eren, A.M., Cavanaugh, C.M., and Mark Welch, J.L. (2020).
1076   Metapangenomics of the oral microbiome provides insights into habitat adaptation and
1077   cultivar diversity. Genome Biol. *21*, 293.

1078 115. García-Bayona, L., and Comstock, L.E. (2019). Streamlined Genetic Manipulation of
1079   Diverse Bacteroides and Parabacteroides Isolates from the Human Gut Microbiota. MBio
1080   *10*. 10.1128/mBio.01762-19.

1081 116. Zitomersky, N.L., Coyne, M.J., and Comstock, L.E. (2011). Longitudinal Analysis of the
1082   Prevalence, Maintenance, and IgA Response to Species of the Order Bacteroidales in the
1083   Human Gut. Infect. Immun. *79*, 2012.

1084 117. Evans, J.C., McEneany, V.L., Coyne, M.J., Caldwell, E.P., Sheahan, M.L., Von, S.S.,
1085   Coyne, E.M., Tweten, R.K., and Comstock, L.E. (2022). A proteolytically activated
1086   antimicrobial toxin encoded on a mobile plasmid of Bacteroidales induces a protective
1087   response. Nat. Commun. *13*. 10.1038/s41467-022-31925-w.

1088 118. Pluta, R., Boer, D.R., and Coll, M. (2014). MobM Relaxase Domain (MOBV; Mob_Pre)
1089   bound to plasmid pMV158 oriT DNA (22nt). Mn-bound crystal structure at pH 4.6.
1090   10.2210/pdb4lvi/pdb.

1091 119. Delano, W.L. (2002). The PyMOL molecular graphics system. http://www.pymol.org/.

1092 120. Ben Chorin, A., Masrati, G., Kessel, A., Narunsky, A., Sprinzak, J., Lahav, S., Ashkenazy,
1093   H., and Ben-Tal, N. (2020). ConSurf-DB: An accessible repository for the evolutionary
1094   conservation patterns of the majority of PDB proteins. Protein Sci. *29*, 258–267.

1095 121. Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009). The ConSurf-DB: pre-
1096   calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res. *37*,

1097   D323–D327.

1098 122. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local
1099   alignment search tool. J. Mol. Biol. *215*. 10.1016/S0022-2836(05)80360-2.

1100 123. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time
1101   and space complexity. BMC Bioinformatics *5*, 113.

1102 124. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for
1103   automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*,
1104   1972–1973.

1105 125. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast
1106   and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol.
1107   Biol. Evol. *32*, 268–274.

1108 126. Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for
1109   Exploring and Manipulating Networks. ICWSM *3*, 361–362.

1110 127. Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a continuous
1111   graph layout algorithm for handy network visualization designed for the Gephi software.
1112   PLoS One *9*, e98679.

1113 128. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken
1114   2. Genome Biol. *20*, 257.

1115 129. Chen, Y.T., Williamson, B.D., Okonek, T., Wolock, C.J., Spieker, A.J., Hee Wai, T.Y.,
1116   Hughes, J.P., Emerson, S.S., and Willis, A.D. (2022). rigr: Regression, Inference, and
1117   General Data Analysis Tools in R. Journal of Open Source Software *7*, 4847.
1118   10.21105/joss.04847.

1119 130. Conrad, S., Oethinger, M., Kaifel, K., Klotz, G., Marre, R., and Kern, W.V. (1996). gyrA
1120   mutations in high-level fluoroquinolone-resistant clinical isolates of Escherichia coli. J.
1121   Antimicrob. Chemother. *38*, 443–455.

1122 131. Lee, C., Kim, J., Shin, S.G., and Hwang, S. (2006). Absolute and relative QPCR
1123   quantification of plasmid copy number in Escherichia coli. Journal of Biotechnology *123*,
1124   273–280. 10.1016/j.jbiotec.2005.11.014.

1125 132. Olds, H.T., Corsi, S.R., Dila, D.K., Halmo, K.M., Bootsma, M.J., and McLellan, S.L.
1126   (2018). High levels of sewage contamination released from urban areas after storm events:
1127   A quantitative survey with sewage specific bacterial indicators. PLoS Med. *15*, e1002614.

1128 133. Feng, S., Ahmed, W., and McLellan, S.L. (2020). Ecological and Technical Mechanisms for
1129   Cross-Reaction of Human Fecal Indicators with Animal Hosts. Appl. Environ. Microbiol.
1130   *86*. 10.1128/AEM.02319-19.

1131 134. Lenaker, P.L., Corsi, S.R., McLellan, S.L., Borchardt, M.A., Olds, H.T., Dila, D.K.,

1132   Spencer, S.K., and Baldwin, A.K. (2018). Human-Associated Indicator Bacteria and
1133   Human-Specific Viruses in Surface Water: A Spatial Assessment with Implications on Fate
1134   and Transport. Environ. Sci. Technol. *52*, 12162–12171.

1135 135. Corsi, S.R., De Cicco, L.A., Hansen, A.M., Lenaker, P.L., Bergamaschi, B.A., Pellerin,
1136   B.A., Dila, D.K., Bootsma, M.J., Spencer, S.K., Borchardt, M.A., et al. (2021). Optical
1137   Properties of Water for Prediction of Wastewater Contamination, Human-Associated
1138   Bacteria, and Fecal Indicator Bacteria in Surface Water at Three Watershed Scales.
1139   Environ. Sci. Technol. *55*, 13770–13782.

1140 136. USGS water data for the nation https://waterdata.usgs.gov/nwis.

1141 137. Dila, D.K., Koster, E.R., McClary-Guterriez, J., Khazaei, B., Bravo, H.R., Bootsma, M.J.,
1142   and McLellan, S.L. (2022). Assessment of Regional and Local Sources of Contamination at
1143   Urban Beaches Using Hydrodynamic Models and Field-Based Monitoring. ACS EST Water
1144   *2*, 1715–1724.

1145 138. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).

1146

1147

Supplementary Tables (🔗)

Supplementary Table 1: The accompanying metadata for all publicly available metagenomes used in this study. This table contains 3 tabs. **(1)** healthy_gut: all healthy gut metagenomes. **(2)** IBD: all IBD gut metagenomes. **(3)** alternative_environment: all non-gut metagenomes.

Supplementary Table 2: The nucleotide sequence and average nucleotide identity (ANI) calculations for all pBI143 contigs. This table has 3 tabs. **(1)** pBI143_sequences: the nucleotide sequence for the 3 reference versions of pBI143 assembled from metagenomes. **(2)** ANI information for the 3 reference sequences of pBI143. **(3)** additional_genes: the nucleotide sequences for pBI143 assembled from metagenomes with additional genetic material.

Supplementary Table 3: Read recruitment data from metagenomes used in this study. This table has 4 tabs. **(1)** global adult gut metagenomes: coverage and detection data for the reference versions of pBI143 in global adult gut metagenomes **(2)** mother-infant metagenomes: coverage and detection data for the reference versions of pBI143 in mother and infant metagenomes **(3)** crassphage_comparison: coverage and detection data for crassphage in global adult gut metagenomes. **(4)** alternative_environments: coverage and detection data for the reference versions of pBI143 in non-human gut environments.

Supplementary Table 4: The metadata for the Duchossois Family Institute bacterial isolate genomes used in this study.

Supplementary Table 5: pBI143 copy number determination via qPCR. This table includes 2 tabs. **(1)** Seq DataSource: contains the Gen-Bank accession numbers and other data sources used in primer and probe development. **(2)** Hsp BLAST result: contains BLASTN results of the hsp nucleotide sequence against the 15 *Bacteroides fragilis* RefSeq complete genomes.

Supplementary Table 6: The data for pBI143 copy number for all animal, environmental and sewage samples as measured via qPCR. This table has 3 tabs. **(1)** animal_copy_number: contains the data showing sample and copy number of pBI143 in animal fecal samples. **(2)** environmental_copy_number: contains the data showing sample and copy number of pBI143 in water samples. **(3)** sewage_copy_number: contains the data showing sample and copy number of pBI143 in sewage samples.

Supplementary Table 7: All the data necessary for quantifying number and type of SNV in gut and sewage metagenomes. Variability profiles are generated by anvi'o to describe the variation found across all contigs of interest; for more information see https://merenlab.org/2015/07/20/analyzing-variability/. This table

has 9 tabs. **(1)** artificial_reads_var_profile: The variability profile generated following artificial short read generation and read recruitment of pBI143 Version 2 and 3 to Version 1 (see Methods). **(2)** global_mg_var_profiles: The variability profile generated following read recruitment of all global gut metagenomes to pBI143 Version 1. **(3)** sewage_mg_var_profiles: The variability profile generated following read recruitment of all global sewage metagenomes to pBI143. **(4)** matching_SNVs_gut: The number of SNVs that do or do not match one of the reference versions of pBI143 in global gut metagenomes. **(5)** matching_SNVs_sewage: The number of SNVs that do or do not match one of the reference versions of pBI143 in global sewage metagenomes. **(6)** gut_var_profile_quince_mode: This file does not fit in excel. Link to online data to regenerate single nucleotide variant data at every position of pBI143 across all global gut metagenomes (for more details on `quince-mode` see https://merenlab.org/2015/07/20/analyzing-variability/#parameters-to-refine-the-output). **(7)** sewage_var_profile_quince_mode: single nucleotide variant data at every position of pBI143 across all global sewage metagenomes. **(8)** gut_non-consensus_SNVs: Data about the plasmid version and number of non-consensus SNVs in each global gut metagenome. **(9)** sewage_non-consensus_SNVs: Data about the plasmid version and number of non-consensus SNVs in each global sewage metagenome.

Supplementary Table 8: This table contains SNV variability profiles for visualizing SAAVs on the pBI143 AF structure. This table has 3 tabs: **(1)** merged_variability: contains all SNV variability data calculated with 'anvi-gen-variability-profile --engine AA' which summarized metagenomic read recruitment results to pBI143; **(2)** merged_variability_filtered: filtered version of merged_variability that reflects the SAAV data visualized on the pBI143 structure in **Fig. 3D**; (3) most_prevelent_SAAVs: this tab contains a list of all SAAVs and their residue positions that are prevalent in at least 5% of samples.

Supplementary Table 9: The necessary data to generate pBI143 and isolate genome phylogenies. This table has 5 tabs. **(1)** amino_acid_repA_mobA_concat: the concatenated MobA and RepA sequences from all 82 isolate genomes. Concatenated genes are separated by `XXX`. **(2)** repA_mobA_treefile: the treefile generated from the concatenated mobA and repA sequences. **(3)** amino_acid_SCG_concat: the concatenated ribosomal protein sequences from all 82 isolate genomes. Concatenated genes are separated by `XXX`. **(4)** SCG_treefile: the treefile generated from the concatenated ribosomal protein sequences. **(5)** species_donor_information: the associated isolate data that matches the donor, species and pBI143 version.

Supplementary Table 10: The data necessary for generating and quantifying the mother-infant network based on single nucleotide variants. This table has 8 tabs. **(1)** Finalnd_variability_profile: data for all single nucleotide variants (SNVs) present in pBI143 in Finnish mother and infant metagenomes. **(2)** Sweden_variability_profile: data for all single nucleotide variants present in pBI143 in Swedish mother and

1209   infant metagenomes. **(3)** Italy_variability_profile: data for all single nucleotide variants present in pBI143

1210   in Italian mother and infant metagenomes. **(4)** USA_variability_profile: data for all single nucleotide

1211   variants present in pBI143 in American mother and infant metagenomes. **(5)** network_data: data used to

1212   generate the network. **(6)** distance_matrix_cosine: distance matrix calculated from network data used for

1213   quantification of distances between samples. **(7)** subtracted_distance_df: quantified distance between

1214   mother and infant samples based on cosine distance matrix. **(8)** summary of pBI143 version maintenance

1215   in infants over the sampling period.

1216   **Supplementary Table 11:** Kraken data. This table contains 2 tabs. **(1)** Kraken data for all *Bacteroides* (this

1217   includes *Phocaeicola* with old *Bacteroides* genus names) and *Parabacteroides* taxa in global gut

1218   metagenomes and the corresponding pBI143 coverage in each of these metagenomes. **(2)** Kraken data for

1219   the number of reads recruited to a *B. fragilis* compared to the coverage of pBI143 in the same metagenomes.

1220   **Supplementary Table 12**: This table contains 2 tabs. **(1)** The data for the mouse competition experiments.

1221   **(2)** The pBI143 maintenance in culture data.

1222   **Supplementary Table 13:** The data for pBI143 copy number for each timepoint and condition of the

1223   *Bacteroides fragilis* stress experiments in culture as measured via qPCR. This table has 2 tabs. **(1)**

1224   214_oxidative_stress_qPCR_data: contains data on the copy number for each test condition for the

1225   *Bacteroides fragilis* 214 strain. **(2)** R16_oxidative_stress_qPCR_data: contains data on the copy number

1226   for each test condition for the *Bacteroides fragilis* R16 strain.

1227   **Supplementary Table 14:** The calculated ACNR and necessary data for these calculations. This table has

1228   4 tabs. **(1)** Coverage_ratio_data: the final ACNR for all predicted single hosts of pBI143 in metagenomes.

1229   **(2)** pBI143_healthy: the coverage of pBI143 in healthy gut metagenomes. **(3)** pBI143_IBD: the coverage

1230   of pBI143 in IBD gut metagenomes. **(4)** SCG_taxonomy: Link to files containing SCG coverage data.

1231   **Supplementary Table 15:** The names and sequences of all primers and probes used in this study.
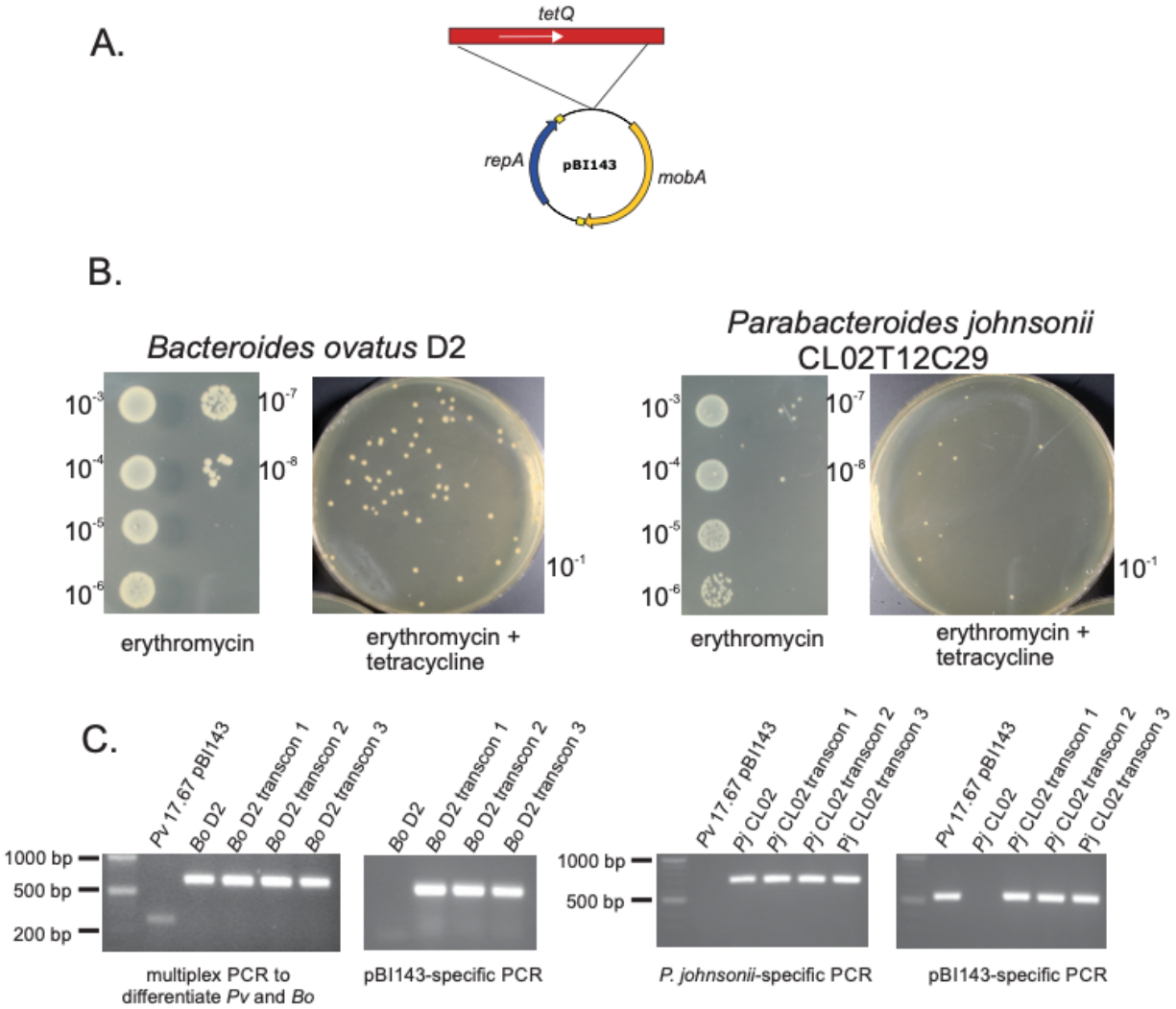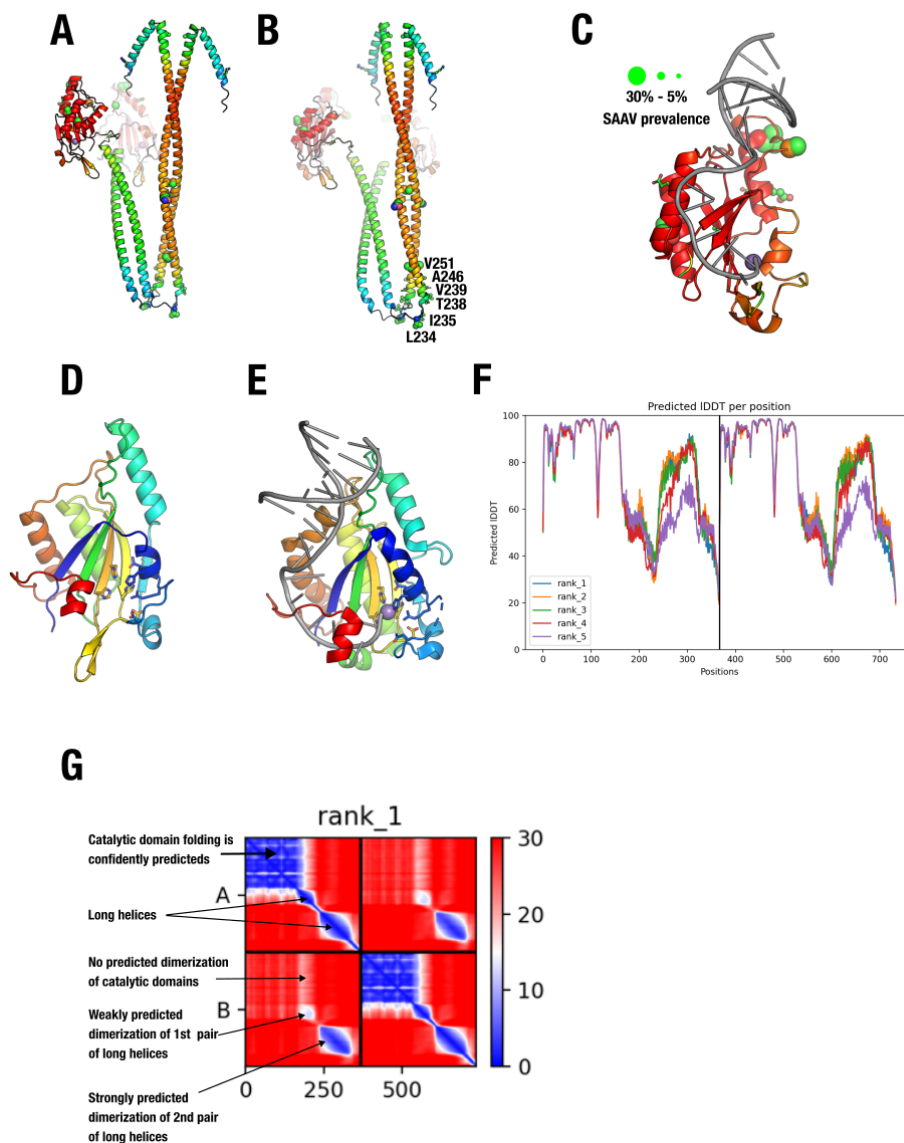
1232 # Supplementary Figures
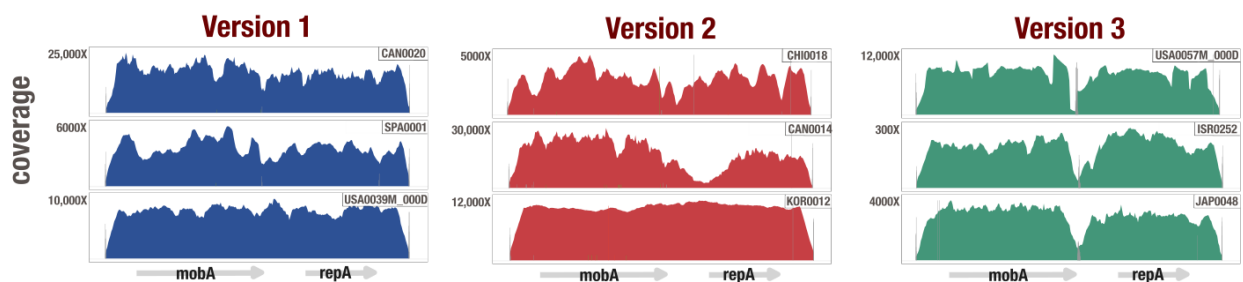
1233

1234 **Supplementary Fig. 1. Representative coverage plots of global metagenomes mapped to pBI143.** Each
1235 coverage plot shows the read recruitment results for an individual metagenome to a pBI143 Version 1 (blue),
1236 Version 2 (red) and Version 3 (green). Vertical bars show single nucleotide variants (red bar = variant in
1237 first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The
1238 x-axis is the pBI143 reference sequence. 3 coverage plots for each reference version of pBI143 are shown,
1239 the remaining 13,539 can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

1240

1241

Supplementary Fig. 2. pBI143 transfer to other Bacteroidales species. (A) Construct made to select for plasmid transfer. **(B)** Number of recipients (erythromycin) and number of transconjugants (erythromycin and tetracycline) for transfer of pBI143-tetQ to *Bacteroides ovatus* D2 and *Parabacteroides johnsonii* CL02T12C129. **(C)** PCR to confirm presence of pBI143-tetQ in recipient strain.
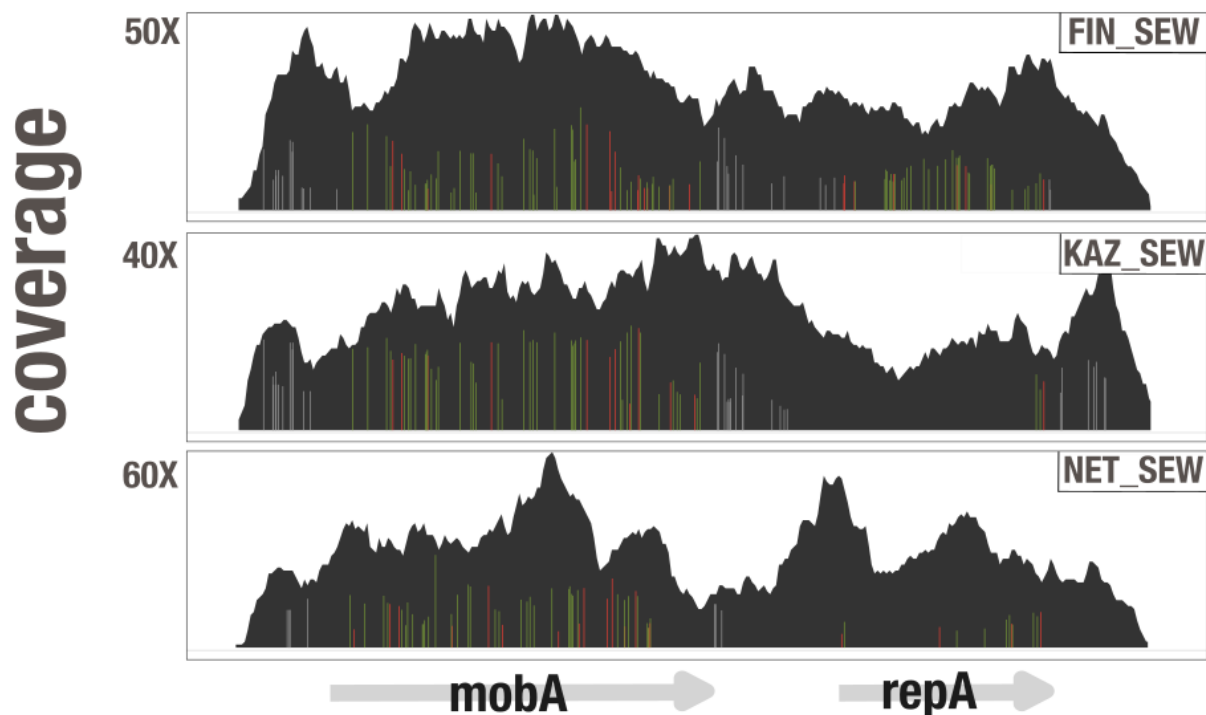
1248

Supplementary Fig. 3. Representative coverage plots of sewage metagenomes mapped to pBI143. Each coverage plot shows the read recruitment results for a sewage metagenome to the Version 1 pBI143 reference sequence. Vertical bars show single nucleotide variants (red bar = variant in first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The x-axis is the pBI143 reference sequence. 3 sewage coverage plots are shown, the other 435 coverage plots from all non-human environments can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

1255

Supplementary Fig. 4. Insights from conserved single amino acid variants. (A),(B) Different angles of the MobA AlphaFold 2 dimer prediction with single amino acid variants from all 4,516 human gut metagenomes superimposed as ball-and-stick residues. The size of the ball-and-stick spheres indicate the proportion of samples carrying variation in that position (the larger the sphere, the more prevalent the variation at the residue) and the color is in CPK format. The color of the ribbon diagram indicates the pLDDT from AlphaFold 2 (red > 90 pLDDT) and blue < 50 pLDDT. The purple sphere is the Mn++ ion that marks the protein active site (oriT DNA and Mn2+ from 4lvi.pdb; 10.1073/pnas.1702971114). **(C)** Catalytic domain with high pLDDT with single amino acid variants from all 4,516 human gut metagenomes superimposed as ball-and-stick residues. Size and coloring is the same as in A,B. **(D)** The catalytic domain of the AlphaFold 2 predicted MobA (residues 1-177) shown shaded from blue to red active site residues are shown as sticks. **(E)** MobM from pMV158 bound to oriT DNA (gray) and a catalytic Mn2+ ion (purple) (PDB id 4lvi [81]) shown shaded from blue to red and active site residues are shown as sticks. **(F)** AlphaFold 2 pLDDT score representing structural prediction accuracy of MobA. **(G)** AlphaFold 2 predicted aligned error plot (PAE) for MobA dimer prediction.

1272
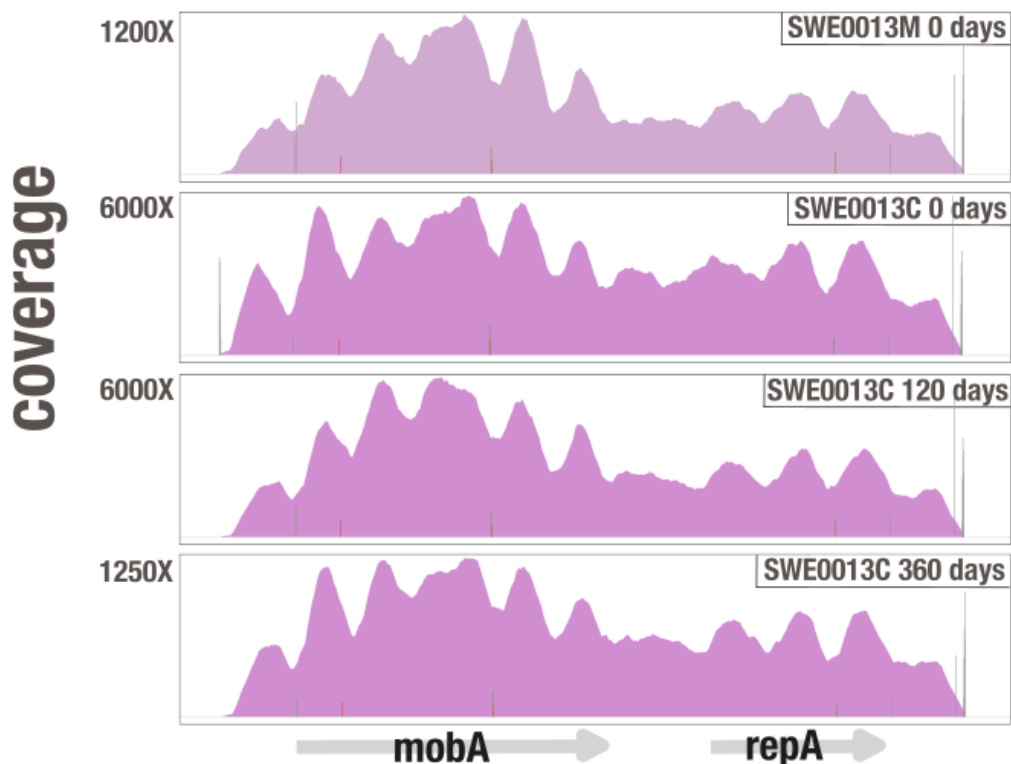
Supplementary Fig. 5. Representative mother-infant coverage plots. Each coverage plot shows the read recruitment results for an individual metagenome to the Version 1 pBI143 reference sequence. Vertical bars show single nucleotide variants (red bar = variant in first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The x-axis is the pBI143 reference sequence. 4 coverage plots are shown, the other 1,020 can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.
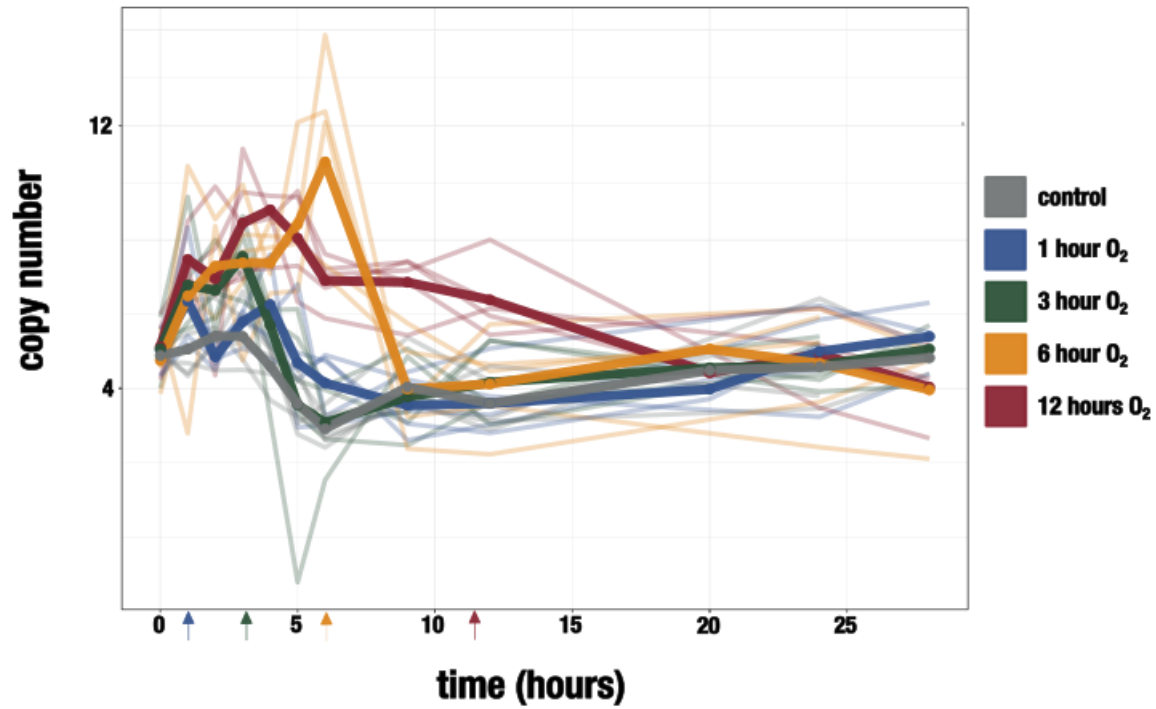
1279
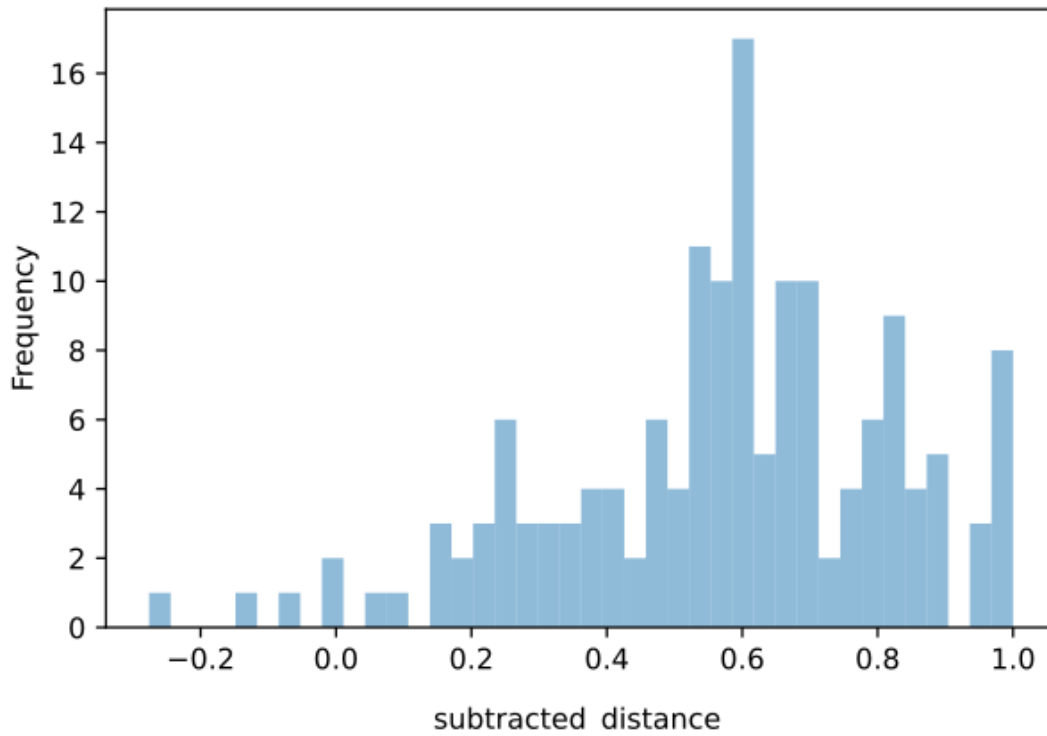
1280

Supplementary Fig. 6. Mother-infant network quantification. Quantification of distances between samples in the network, where distance is calculated by converting the network file to a distance matrix using the python `pdist` function with cosine distances. The "subtracted difference" shows the mean within-family distances subtracted from mean between-family distances for each sample in the mother infant pair network. See methods for more details.

1281
1282
1283
1284
1285

1286

1287

1288

Supplementary Fig. 7. R16 oxidative stress experiment. Copy number of pBI143 in *B. fragilis* R16 cultures with increasing exposure to oxygen. Arrows indicate the time point at which the culture was returned to the anaerobic chamber. The control cultures (gray) were never exposed to oxygen. Opaque lines are the mean of 5 replicates (translucent lines).

1289
1290
1291
1292

1293
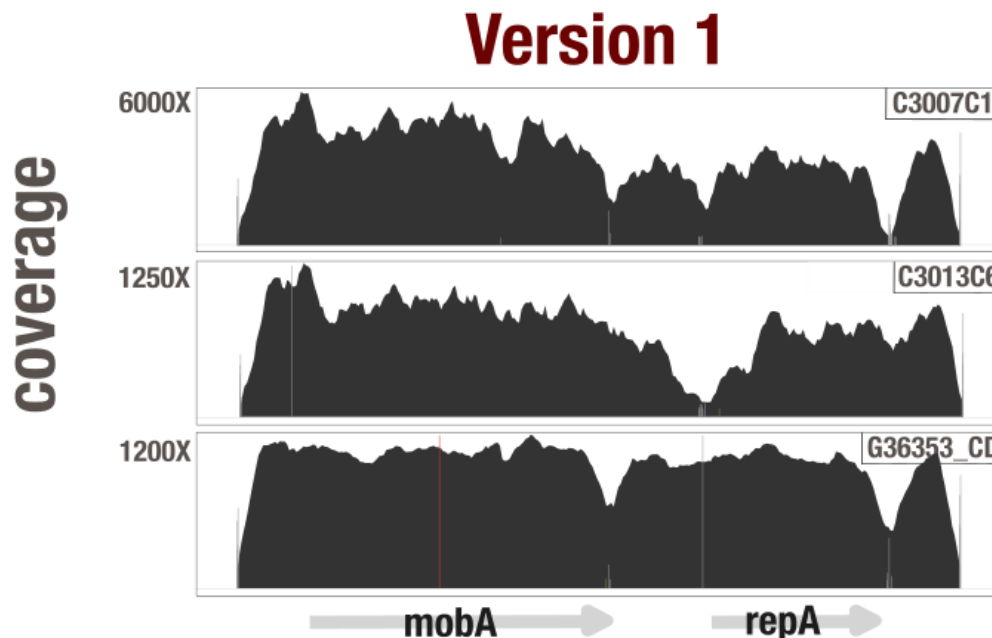
Supplementary Fig. 8. Representative IBD gut metagenome coverage plots. Each coverage plot shows the read recruitment results for an individual metagenome to a pBI143 Version 1. Vertical bars show single nucleotide variants (red bar = variant in first or second codon position, green bar = variant in third codon position, gray bar = intergenic variant). The x-axis is the pBI143 reference sequence. 3 coverage plots are shown, the other 3,087 can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

Supplementary Information

1301

1302 The supplementary information file (available at 10.6084/m9.figshare.22336666) includes additional

1303 information regarding the construction of the phylogenetic trees and plasmid constructs, the development

1304 of the qPCR assay for determining copy number of pBI143, and additional information about the read

1305 recruitment results of non-human gut environments to pBI143.