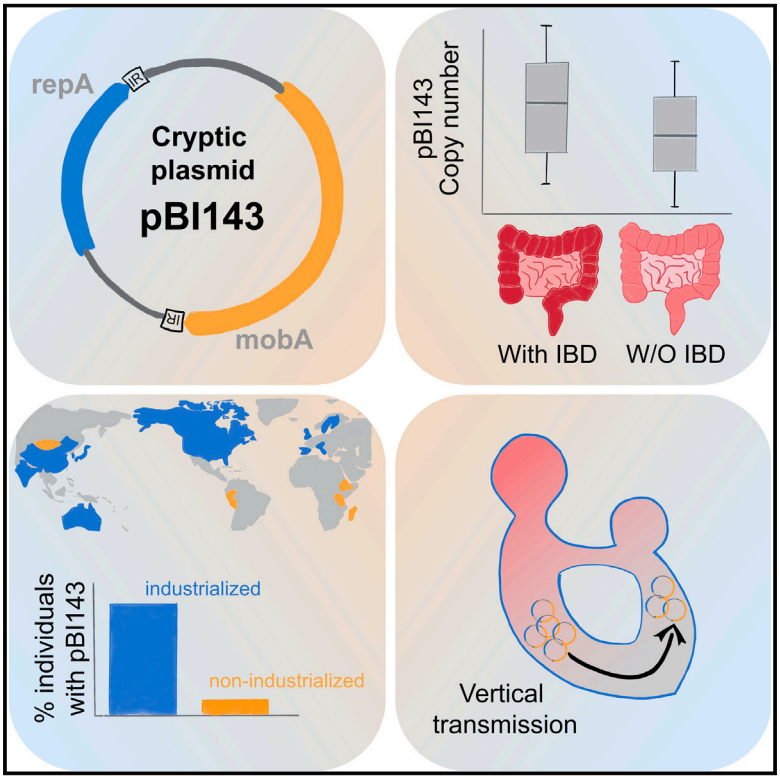
Cell

Article

A cryptic plasmid is among the most numerous genetic elements in the human gut

Graphical abstract



Highlights

- pBI143 is a cryptic plasmid that is prevalent in industrialized human populations
- pBI143 is only ~2.7 kb, yet it regularly makes up over 0.1% of gut metagenomic reads
- The naive 2-gene form appears parasitic, but pBI143 can carry additional genes
- The relative copy number of pBI143 increases during stress, including in IBD

Authors

Emily C. Fogarty, Matthew S. Schechter, Karen Lolans, ..., Amy D. Willis, Laurie E. Comstock, A. Murat Eren

Correspondence

efogarty@uchicago.edu (E.C.F.), lecomstock@bsd.uchicago.edu (L.E.C.), meren@hifmb.de (A.M.E.)

In brief

A widely distributed and highly conserved human gut plasmid is extremely numerous in industrialized human gut metagenomes and increases its copy number in response to stress.

Fogarty et al., 2024, Cell *187*, 1206–1222
February 29, 2024
https://doi.org/10.1016/j.cell.2024.01.039

CellPress

CellPress
OPEN ACCESS

Cell

Article

A cryptic plasmid is among the most numerous 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 M. 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] Hans-Joachim Ruscheweyh,[11] Shinichi Sunagawa,[11] Sandra L. Mclellan,[12] Amy D. Willis,[13] Laurie E. Comstock,[1,2,4,*] and A. Murat Eren[3,14,15,16,17,18,19,*]

[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, Chicago, IL 60637, USA
[10]Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL 60637, USA
[11]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich 8093, Switzerland
[12]School of Freshwater Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53204, USA
[13]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
[14]Marine Biological Laboratory, Woods Hole, MA 02543, USA
[15]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, 27570 Bremerhaven, Germany
[16]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26129 Oldenburg, Germany
[17]Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany
[18]Helmholtz Institute for Functional Marine Biodiversity, 26129 Oldenburg, Germany
[19]Lead contact
*Correspondence: efogarty@uchicago.edu (E.C.F.), lecomstock@bsd.uchicago.edu (L.E.C.), meren@hifmb.de (A.M.E.)
https://doi.org/10.1016/j.cell.2024.01.039

SUMMARY

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 extrachromosomal 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*, it 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 alternative approach to track human colonic inflammatory states.

INTRODUCTION

The tremendous density of microorganisms 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 lies a group of plasmids that appear to lack genetic functions of interest and that do not contain genes encoding obvious beneficial functions for 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] and their distribution across many different environments.[20,28] 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[29] or genetic parasites.[30] Although 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.[30–32]

Analyses of cryptic plasmids are often performed on monocultured bacteria, limiting insights into their ecology in naturally occurring microbial habitats. However, recent advances in shotgun metagenomics[33] and *de novo* plasmid prediction algorithms[34–43] offer a powerful means to bridge this gap. For instance, in a recent study, we characterized over 68,000 plasmids from the human gut[43] and observed that one of the most prevalent reference plasmids across our dataset of geographically diverse human populations was a cryptic plasmid called pBI143.[44] 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 human health and disease states. Our findings reveal the astonishing success of pBI143 in the human gut, where we were able to detect it in up to 92% of individuals in industrialized countries with copy numbers at least 14 times higher on average than crAssphage, one of the most abundant phages in the human gut. We also demonstrate the potential of pBI143 as a cost-effective biomarker to assess the extent of stress that microbes experience in the human gut and as a sensitive means to quantify the level of human fecal contamination in environmental samples.

RESULTS

pBI143 is extremely prevalent across industrialized human gut microbiomes

pBI143 (Genbank: U30316.1) is a 2,747 bp circular plasmid first identified in 1985 in *Bacteroides fragilis*,[44,45] a member of the human gut microbiome that is frequently implicated in states of health[46–48] and disease.[49,50] pBI143 encodes only two annotated genes: a mobilization protein (*mobA*) and a replication protein (*repA*) (Figure 1A). Due to the desirable features for cloning such as a high copy number and genetic stability, cryptic plasmids have often been primarily used as components of *E. coli-Bacteroides* shuttle vectors.[45,51] The absence of any ecological studies of pBI143 prompted us to characterize it further, beginning with a characterization of its genetic diversity.

To comprehensively sample the diversity of pBI143, we screened 2,137 individually assembled human gut metagenomes (Table S1) for pBI143-like sequences. By surveying all contigs using the known pBI143 sequence as reference, we found three distinct versions of pBI143 (Figure 1A), all of which had over 95% nucleotide sequence identity to one another throughout their entire length except at the *repA* gene, where the sequence identity was as low as 75% with a maximum of 81% between version 1 and version 2 (Table S1).

We then sought to quantify the prevalence of pBI143 across global human populations using a metagenomic read recruitment survey with an expanded set of 4,516 publicly available gut metagenomes from 23 countries[52–74] (Table S1). Recruiting metagenomic short reads from each gut metagenome using each pBI143 version independently (Figure 1; Table S1), we found that pBI143 was present in 3,295 metagenomes, or 73% of all samples at a detection threshold over 0.5 (Figure 1B; see STAR Methods for details). However, the prevalence of pBI143 was not uniform across the globe (Figure 1B): pBI143 occurred predominantly in metagenomes of individuals who lived in relatively industrialized countries, such as Japan (92% of 636 individuals) and the United States (86% of 154 individuals). We rarely detected pBI143 in individuals who lived in relatively non-industrialized countries such as Madagascar (0.8% of 112 individuals) or Fiji (8.7% of 172 individuals). This difference is likely due to the non-dominant presence of taxa that harbor pBI143 in the gut microbiomes of individuals from relatively less industrialized countries, whose microbiomes typically differ from those who live in industrialized countries.[75] Within individuals who carried it, pBI143 was often highly abundant (Figure 1B), and despite its small size, it often recruited 0.1% to 3.5% of all metagenomic reads with a median coverage of over 7,000× (Figure S1; Table S1).

The distribution of pBI143 versions across industrialized human populations was also not uniform as different versions of pBI143 tended to be dominant in different geographic regions. pBI143 version 1 (98% identical to the original reference sequence for pBI143[44]) dominated individuals in North America and Europe and occurred on average in 82.5% of all samples that carry pBI143 from Austria, Canada, Denmark, England, Finland, Italy, Netherlands, Spain, Sweden, and the USA (Figure 1C; Table S1). By contrast, pBI143 version 2 dominated countries in Asia and occurred in 63.6% of all samples that carry pBI143 in China, Japan, and Korea (Figure 1C; Table S1). pBI143 version 3 was relatively rare, comprising only 7.4% of pBI143-positive samples, and mostly occurred in individuals from Japan, Korea, Australia, Sweden, and Israel (Figure 1C; Table S1).

The extremely high prevalence and coverage of pBI143 suggest that it is likely one of the most numerous genetic elements in the gut microbiota of individuals from industrialized countries. We compared the prevalence and relative abundance of pBI143 to crAssphage,[76,77] a 97 kbp bacterial virus that is widely recognized as the most abundant family of viruses in the human gut.[78] pBI143 was more prevalent (73% vs. 27%) in our metagenomes than all 21 crAssphage genomes we analyzed, although individual samples differed widely with respect to the abundance of pBI143 and crAssphage (Table S1). The average percentage of metagenomic reads recruited by pBI143 and the most abundant crAssphage were 0.05% and 0.13%, respectively. However, taking into consideration that crAssphage is approximately 36 times larger than pBI143 and assuming that average coverage is an acceptable proxy to the abundance of genetic entities, these data suggest that on average, pBI143 is at least 14 times more numerous than crAssphage in the human gut (Table S2).

Overall, these data demonstrate that pBI143 is one of the most widely distributed and numerous genetic elements in the gut microbiomes of industrialized human populations worldwide.

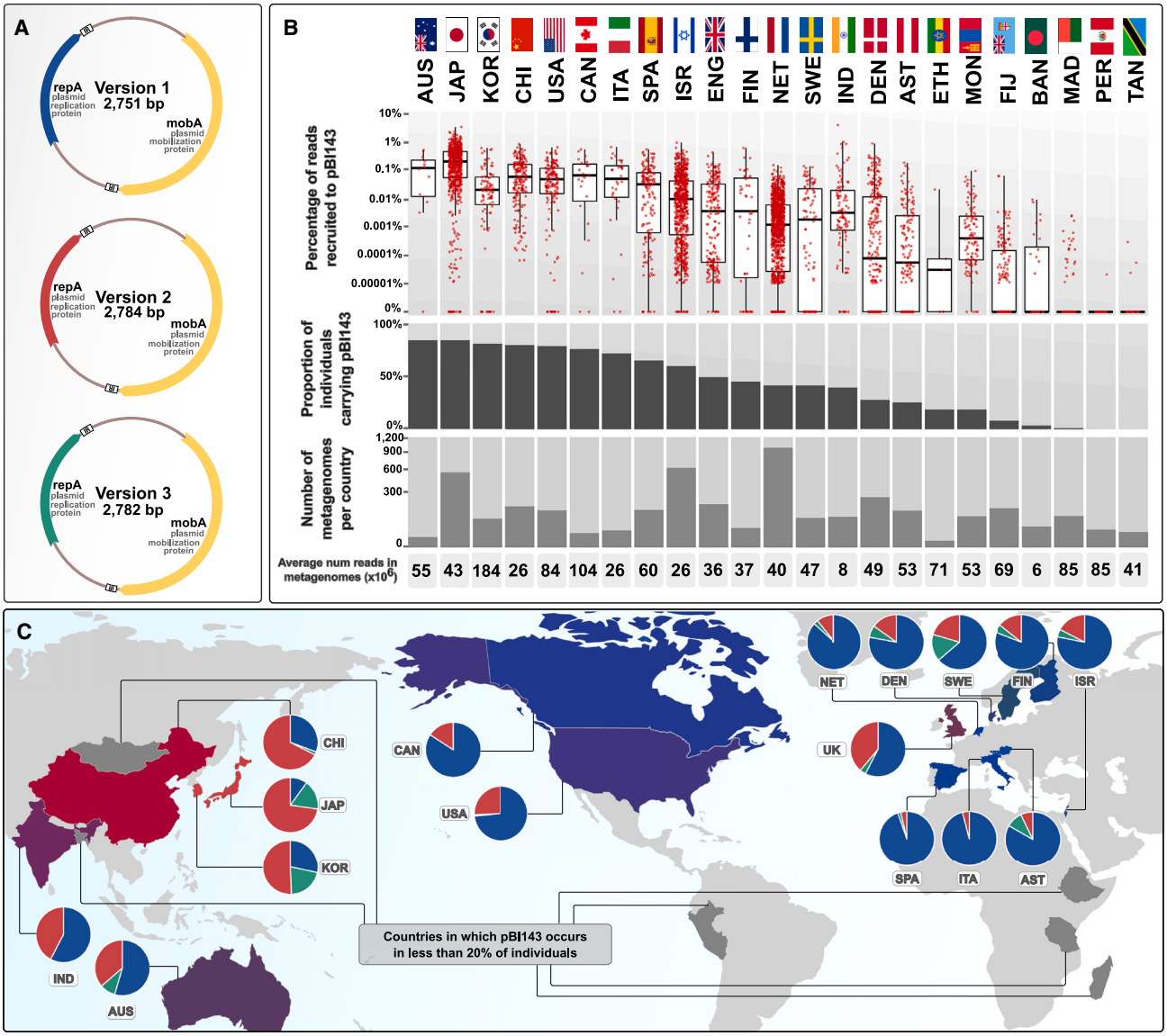


Figure 1. pBI143 prevalence and abundance in globally distributed human populations

(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 versions 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.

See also Table S1 and Figure S1.

Multiple taxa within the order Bacteroidales carry pBI143

Interestingly, the detection patterns of pBI143 in metagenomes differed from the detection patterns we observed for its *de facto* host *Bacteroides fragilis* in the same samples; *B. fragilis* and pBI143 co-occurred in only 41% of the metagenomes. Sequencing depth did not explain this observation, as pBI143 was highly covered (i.e., >50×) in 25% of metagenomes where *B. fragilis* appeared to be absent (Table S2), suggesting that the host range of pBI143 extends beyond *B. fragilis*.

To investigate the host range of pBI143, we employed a collection of bacterial isolates from the human gut, which

contained 717 genomes that represented 104 species in 54 genera (Table S1). We found pBI143 in a total of 82 isolates that resolved to 11 species across 3 genera: *Bacteroides*, *Phocaeicola*, and *Parabacteroides*. Many of the pBI143-carrying isolates of distinct species were from the same individuals, suggesting that pBI143 can be mobilized between species. To confirm this, we inserted a tetracycline resistance gene, *tetQ*, into pBI143 in the *Phocaeicola vulgatus* isolate MSK 17.67 (Figure S2; Table S1) and tested the ability of this engineered pBI143 to transfer to two strains of two different families of Bacteroidales, *Bacteroides ovatus* D2 and *Parabacteroides johnsonii* CL02T12C29. In these assays, we found that pBI143 was indeed transferred from the donor to the recipient strains at a frequency of 5×10^{-7} and 3×10^{-6} transconjugants per recipient, respectively (Figure S2).

pBI143 is primarily restricted to the human gut

Given the host range of pBI143, one interesting question is whether the ecological niche boundaries of pBI143 hosts exceed a single biome since the members of *Bacteroides*, *Phocaeicola*, and *Parabacteroides* are not specific to the human gut and do occur in a range of other habitats, including non-human primate guts.[79] For a comprehensive survey, we searched for pBI143 in over 100,000 metagenomes from 88 diverse environments, including ocean, wastewater, soil, plants, hospital surfaces, and animal guts (buffalo, cat, chicken, cow, deer, dog, elk, fish, goat, human, insect, macaques, mouse, panda, pig, rat, sheep, termite, vole, whale, boar, yak, and zebu). The results of read recruitment from these metagenomes indicated pBI143 is largely specific to the human gut (Figures 2A and S3). In an extreme example, pBI143 comprised an astonishing 20.1% of all reads (33 million) in an individual person in the United States, with a metagenomic read coverage of 377,600× (Table S1). But this was not a singular example: the average coverage of pBI143 exceeded 100,000× in over 40 individuals in the dataset and was higher than 10,000× in over 1,500 humans (Table S1). By contrast, the detection of pBI143 in non-host-associated environments was virtually zero, except in human-impacted habitats such as sewage and hospital surfaces, where pBI143 was systematically detected in very low abundances (Figure S3; Table S1). Across the gut metagenomes of 30 animal species, we consistently detected pBI143 only in rats housed in laboratories and pet cats (Figure S3). However, pBI143 represented only 0.003% and 0.001% (∼50× and ∼25× coverage) of all reads in cat and rat metagenomes (Figure S1) on average in contrast to 0.1% (∼1,600× coverage) of all reads in human gut metagenomes (Table S1). As pBI143 appeared to primarily flourish in humans, we also screened metagenomes from various human body sites, including the human skin, oral cavity, respiratory tract, nose, and vagina,[74] and found that pBI143 was poorly detected on the human body outside of the gut environment (Figure 2A; Table S1).

Finally, to confirm the results of our metagenomic screen, we designed and tested a highly specific qPCR assay for pBI143 (Table S3). Although there was a robust amplification of pBI143 from sewage samples (Figure 2B), pBI143 was virtually absent in fecal samples from dogs, alligators, raccoons, horses, pigs, deer, cows, chickens, geese, cats, rabbits, or gulls (Table S3). Our qPCR data did show low levels of amplification in three of the four cats tested, however, at a copy number that was 73-fold less than human fecal content of sewage.

The near-absolute exclusivity of pBI143 to the human gut presents practical opportunities, such as the accurate detection of human fecal contamination outside the human gut. Using the same PCR primers, we also amplified pBI143 from water and sewage samples and compared its sensitivity to the gold standard markers currently used for detecting human fecal contamination in the environment (16S rRNA gene amplification of human *Bacteroides* and Lachnospiraceae).[80,81] pBI143 had higher amplification in all 41 samples where *Bacteroides* and Lachnospiraceae were also detected (Figure 2). pBI143 was also amplified in 6 samples with no *Bacteroides* or Lachnospiraceae amplification, suggesting it is a highly sensitive marker for detecting the presence of human-specific fecal material.

Overall, these data show that pBI143 thrives specifically in the human gut environment, and can serve as a sensitive biomarker to detect human fecal contamination.

pBI143 is monoclonal within individuals, and its variants across individuals are maintained by the strong purifying selection

So far, our investigation of pBI143 has focused on its ecology. Next, we sought to understand the evolutionary forces that have conserved the pBI143 sequence by quantifying the sequence variation among the three distinct versions and examining the distribution of single-nucleotide variants (SNVs) within and across globally distributed individuals. Across the three versions, both pBI143 genes had low dN/dS values (*mobA* = 0.11, *repA* = 0.04) (Table S4), suggesting the presence of strong forces of purifying selection acting on *mobA* and *repA*, resulting in primarily synonymous substitutions. Although the comparison of the three representative sequences provides some insights into the conserved nature of pBI143, it is unlikely that they capture its entire genetic diversity across gut metagenomes.

To explore the pBI143 variation landscape, we analyzed metagenomic reads that matched the version 1 of *mobA* to gain insights into the population genetics of pBI143 in naturally occurring habitats through SNVs. Since the *mobA* gene was more conserved across distinct versions of the plasmid compared with the *repA* gene, focusing on *mobA* enabled characterization of variation from all plasmid versions using a single-read recruitment analysis. Surprisingly, the vast majority (83.2%) of mutation hotspots that varied in any metagenome matched a nucleotide position that differed between at least one pair of the three plasmid versions (Figure 3A; Table S4). In other words, pBI143 variation across metagenomes was predominantly localized to certain nucleotide positions that differed between the representative sequences of pBI143 for versions 1, 2, and 3, indicating that the three representative versions capture the majority of permissible pBI143 variation within our collection of gut metagenomes. Indeed, only 24.5% of metagenomes had more than three additional SNVs that were not present in at least one plasmid version, and 84.8% of metagenomes had a pBI143 population that was within 2-nucleotide distance (i.e., over 99.93% sequence identity) of one of the three versions. In addition to the primarily localized variation of pBI143, we also observed

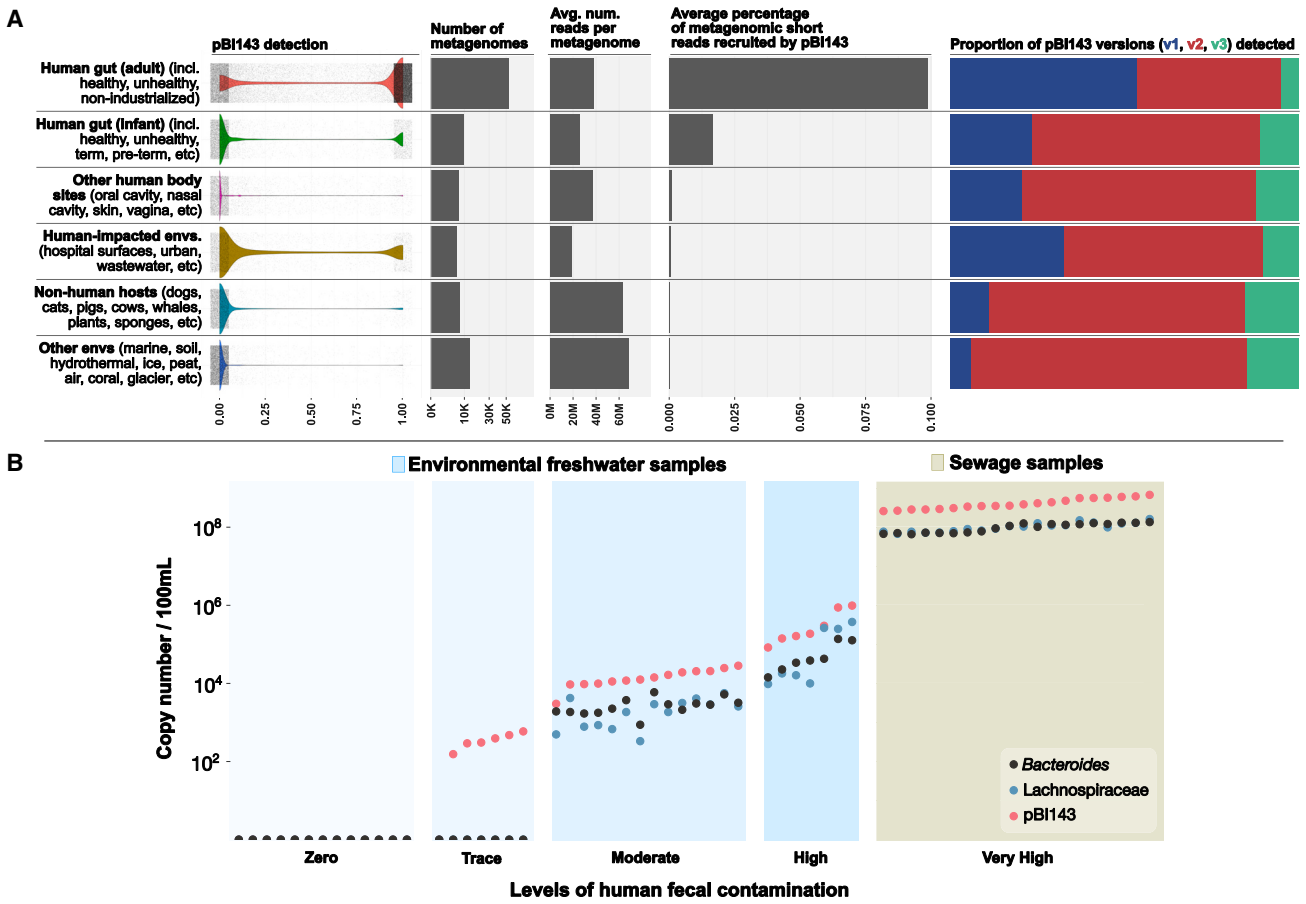


Figure 2. Presence or absence of pBI143 in non-human gut environments

(A) Survey of pBI143 across more than 100,000 metagenomes from diverse environments.

(B) Copy number of pBI143 compared with two established human fecal markers *Bacteroides* or Lachnospiraceae, as measured by qPCR. Zero, trace, moderate, high, and sewage categories and sample order designations are determined based on pBI143 copy number relative to the established markers.

See also Tables S1 and S2 and Figures S2 and S3.

that the vast majority of metagenomes had zero non-consensus SNVs (Figure 3C, teal). In other words, pBI143 populations in most metagenomes either had no SNVs and were identical to one of three pBI143 versions, or any SNVs found in a given metagenome were fixed in the population (for more details, see STAR Methods). A similar analysis with the *repA* gene also showed similar patterns (Figure S4; Table S4). Overall, these data suggest that most humans carry a monoclonal population of pBI143 with little to no within-individual variation (Figure 3C; Table S4).

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[83] (SAAVs) we recovered from gut metagenomes and superimposed them on the AlphaFold 2[83,84] predicted structure of MobA using anvi'o structure.[85] The predicted catalytic domain of pBI143 MobA was structurally similar to MobM of the MobV-family (PDB: 4LVI) encoded by plasmid pMV158.[82] 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 (Figure 3D; Table S4). Interestingly, highly prevalent SAAVs occurred exclusively near the DNA binding site (L56, E49, and A64), leading us to hypothesize 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[82] 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 (Figure S4) could 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 suggest a likely role for adaptive processes on the evolution of pBI143 versions.

In contrast to the individual gut metagenomes, the pBI143 populations did not occur in a monoclonal fashion in sewage metagenomes (Table S4). 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

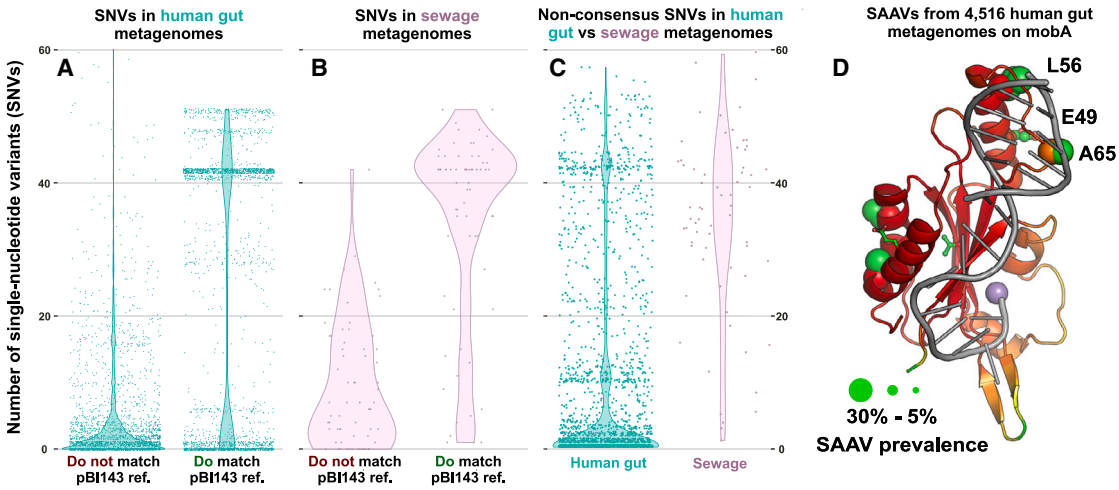


Figure 3. The mutational landscape of pBI143 in sewage and the human gut

(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 Mn^{2+} ion marking the active site (purple) were modeled based on 4lvi.pdb.[82] The size of the ball-and-stick spheres indicates 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 displays the values of the AlphaFold2-generated per-residue confidence metric, predicted local distance difference test (pLDDT), where the color red indicates very high confidence (>90 pLDDT) and the color orange indicates high confidence (>80 pLDDT).

See also Tables S3 and S4.

sewage (Figure 3C; Table S4). Similar to the individual gut metagenomes, most SNVs in sewage metagenomes (78.8%) occurred at a nucleotide position that was variable between at least one pair of the three pBI143 versions (Figure 3B; Table S4), suggesting that the majority of the variability in sewage is from the mixing of different versions of pBI143. However, the number of additional SNVs was much higher in sewage: 61.8% of sewage samples had greater than three SNVs that did not match a variable position in one of the three reference plasmids (Figure 3B). Given the marked increase in the number of additional SNVs in sewage, it is likely there are alternate but relatively rare versions of pBI143 in the human gut.

Overall, these results indicate that pBI143 has a highly restricted mutational landscape in natural habitats, frequently occurs as a monoclonal element in individual gut metagenomes, and the non-synonymous variants of MobA in the environment may be responsible for altering its DNA 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

The largely monoclonal nature of pBI143 presents an interesting ecological question: how do individuals acquire it, and what maintains its monoclonality? Multiple phenomena could explain the monoclonality of pBI143 in individual gut metagenomes, including (1) low frequency of exposure (i.e., most individuals are only ever exposed to one version), (2) bacterial host specificity (i.e., some plasmid versions replicate more effectively in certain bacterial hosts), or (3) priority effects (i.e., the first version of pBI143 establishes itself in the ecosystem and excludes others). The sheer prevalence and abundance of pBI143 across industrialized populations renders the "low frequency of exposure" hypothesis an unlikely explanation. Yet the remaining two hypotheses warrant further investigation.

Bacterial host specificity is a plausible driver for the presence of a singular pBI143 version within an individual, given the interactions between plasmid replication genes and host replication machinery.[29,86] However, our analysis of 82 bacterial cultures isolated from 10 donors shows that the plasmid is more specific to individuals than it is to certain bacterial hosts (Figure 4; Table S5). Indeed, identical pBI143 sequences often occurred in multiple distinct taxa isolated from the same individual, in agreement with the monoclonality of pBI143 in gut metagenomes and its ability to transfer within Bacteroidales. If pBI143 monoclonality is not driven by rare exposure or host specificity, it could be driven by priority effects,[87] where the initial pBI143 version somehow prevents other pBI143 versions from establishing in the same gut community.

To examine if priority effects play a role in pBI143 monoclonality, we aimed to determine how pBI143 is acquired. The vertical transmission of microbes from mother to infant[64] is a well-understood mechanism that transfers not only microbial populations[64] but also their mobile genetic elements, such as phages and transposons.[88] We investigated evidence for the vertical transmission of pBI143 using our ability to track pBI143 SNVs between

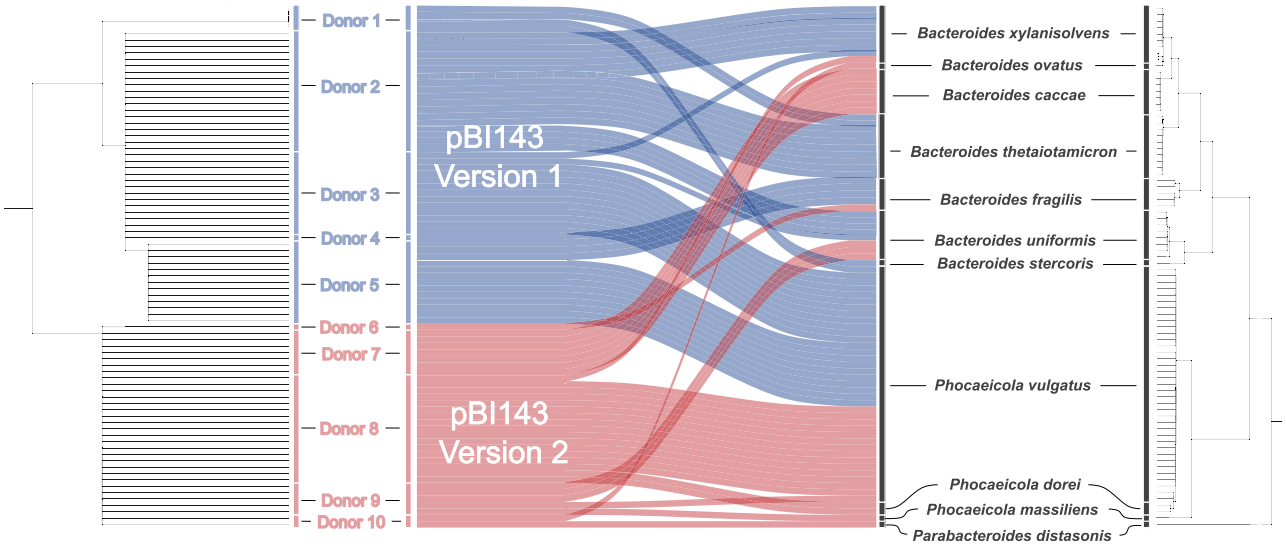


Figure 4. Phylogeny of pBI143 in human donors versus the phylogeny of bacterial isolates recovered from the same individuals

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 STAR 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.

See also Tables S1 and S4.

environments and followed the inheritance of identical pBI143 SNV patterns among 154 mother and infant gut metagenomes from four countries, Finland,[61] Italy,[64] Sweden,[72] and the USA,[73] where each study followed participants from birth to 3 to 12 months of age. It is reasonable to assume that in some cases, infants may acquire pBI143 from other caregivers aside from the mother, but these data were not appropriate to test other transmission routes. We recruited reads from each metagenome to version 1 pBI143 (Table S1) and identified the location of each SNV in *mobA* (Table S6). These data revealed a large number of cases where pBI143 had identical SNV patterns in mother-infant pairs (Figure 5A; Table S6). A network analysis of shared SNV positions across metagenomes appeared to cluster family members more closely, indicating mother-infant pairs had more SNVs in common than they had with unrelated individuals, which we could further confirm by quantifying the relative distance between each sample to others (Figure S5; Table S6; STAR Methods).

Establishing that pBI143 is often vertically transferred, we next examined the impact of priority effects on pBI143 maintenance over time. We assumed that if priority effects are driving persistence of a single version of pBI143, the first version that enters the infant gut environment should be maintained over time. Indeed, many phage populations are influenced by priority effects where the presence of one phage provides a competitive advantage to the bacterial host[89] or bacterial host immunity to infection with similar phages.[90–92] In our data, we found no instances where pBI143 acquired from the mother was fully replaced in the infant during and up to the first year of life (Table S6). Although 69% of infants maintained the version received from the mother (Figure 5B), we also observed other, less common genotypes. These less common cases included a "two versions" scenario where the mother possessed two versions of pBI143, both of which were passed to the infant (21%), and a "wilt" case, where the transferred pBI143 was neither replaced nor persisted until the end of sampling (7%) (Figure 5B). Of the five total wilt cases where pBI143 was lost, only one did not show a corresponding drop in *Bacteroides*/*Phocaeicola* abundance (Figure S5). Although these less prevalent phenotypes are not necessarily explained by priority effects, 69% maintenance of the initial version of pBI143 suggests that priority effects have an important role in the maintenance of pBI143 in the gut, despite many incoming populations colonizing the infant and likely carrying other pBI143 versions.

Overall, by tracking SNV patterns between environments, we established that pBI143 is vertically transferred from mothers to infants and that priority effects likely play a role in maintaining the predominantly monoclonal populations of pBI143.

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 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) (Figure 6A; Table S2); however, these data are not suitable to distinguish whether pBI143 provides a benefit to the bacterial host fitness or acts as a genetic hitchhiker.

To experimentally investigate if pBI143 is advantageous or parasitic, we constructed isogenic pairs of *B. fragilis* 638R and *B. fragilis* 9343 with and without the native version 1 sequence

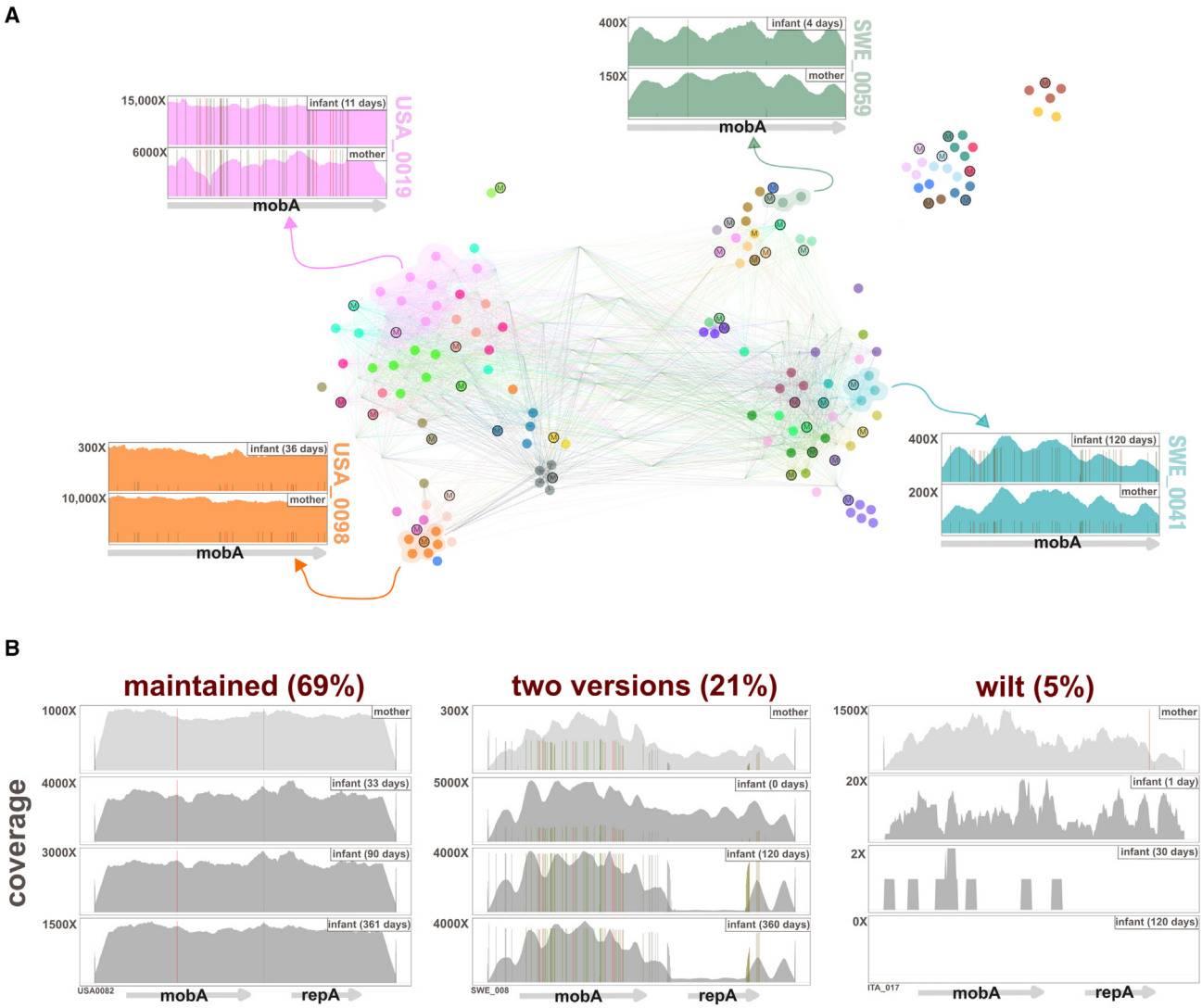


Figure 5. Transfer and maintenance of pBI143

(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. See also Table S5 and Figure S5.

of pBI143. To ensure pBI143 is maintained in these new *Bacteroides* hosts, we passaged them in culture for 7 days and found that the plasmid was still present in all colonies in both strains (Table S2) showing that it is faithfully replicated. Next, we competed the *B. fragilis* 638R (with and without pBI143) in gnotobiotic mice for 40 days to determine if pBI143 affects the fitness of this bacterial host. In contrast to the stable maintenance of pBI143 we observed in human populations (Figures 6A and S1), here we observed a gradual decline in the ratio of pBI143-containing cells to plasmid-free cells over time (Figure 6B; Table S2). The slow decrease in pBI143-containing cells suggests that pBI143 has a small but negative impact on *B. fragilis* 638R fitness in this *in vivo* model. However, it is worth noting that pBI143 may have different fitness effects on different bacterial hosts, as has been shown for other plasmids.[93]

If pBI143 exerts a cost to its hosts, how then is it maintained in Bacteroidales populations in the human gut? This falls under the umbrella of a more general question: Why are plasmids maintained in cells at all? The "plasmid paradox" states that, in theory, plasmids should not even exist given that their conjugal transfer rate is too low to allow them to persist in populations by infectious transmission and that, over time, the cost of their maintenance outweighs any benefits they may confer as those beneficial traits are eventually captured by the chromosome.[94,95] However, more recent work has suggested that plasmid conjugation rates are high enough to allow for infectious transfer

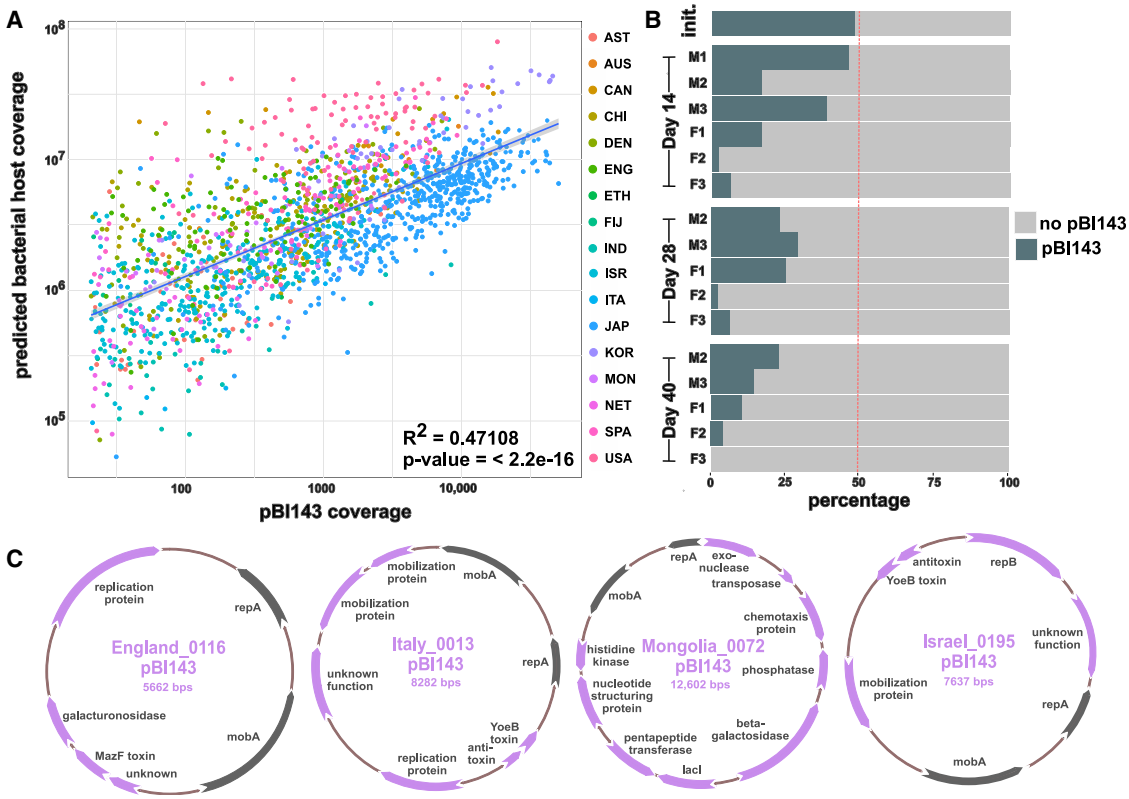


Figure 6. The relationship between pBI143 and its bacterial hosts

(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 male (M) and female (F) mice in the initial inoculum at days 14, 28, and 40.

(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.

See also Tables S2 and S6 and Figure S6.

and maintenance.[30,96,97] As a mobilizable yet non-conjugative plasmid, pBI143 relies on the conjugation machinery of other elements in the bacterial genome to transfer between cells, and most likely uses infectious transfer to maintain itself in Bacteroidales populations. In the absence of the opportunity for infectious transfer, it is conceivable to expect a decline of pBI143 in a population as our experiments demonstrated. Overall, it is likely that in the "wild" environment of the human gut, pBI143 at least in part relies on transferring between organisms (Figure S2) to overcome the costs it exerts on its bacterial host.

Another strategy for pBI143 to maintain itself in the population and provide a benefit to its bacterial hosts is to act as a natural shuttle vector by transiently acquiring additional genetic material and transferring it between cells in a community. In fact, in our survey of assembled gut metagenomes, we observed a few cases that may support such a role for pBI143. In most individuals, we assembled pBI143 in its native form with two genes. However, there were 10 instances where the assembled pBI143 sequence from a given metagenome contained additional genes (Figures 6C and S6; Table S1). Many of the additional genes had no predicted function, but other cargo include predicted toxin-antitoxin genes conferring plasmid stability, as well as those that may confer beneficial functions to the bacterial host, such as galacturonosidase, pentapeptide transferase, phosphatase, and histidine kinase genes. A further examination of these larger plasmids suggested that the additional genetic material was likely acquired from both other extrachromosomal mobile elements and chromosomal DNA (Table S1) in *Bacteroides* and *Eubacterium* species (Table S3). There did not appear to be site specificity, as additional material recombined into both of pBI143's intergenic regions (Figure S6). These occasional larger versions of pBI143 share a common backbone of *repA* and *mobA* and thus form a "plasmid system,"[43] a common plasmid evolutionary pattern suggesting the possibility that pBI143 may dynamically acquire different genes in different environments.

Overall, it appears that the native pBI143 can be mildly detrimental to cells under controlled environments, maintains itself in natural populations through infectious transfer, and is also capable of acquiring additional genes into its backbone, which may provide benefits to the host cells.

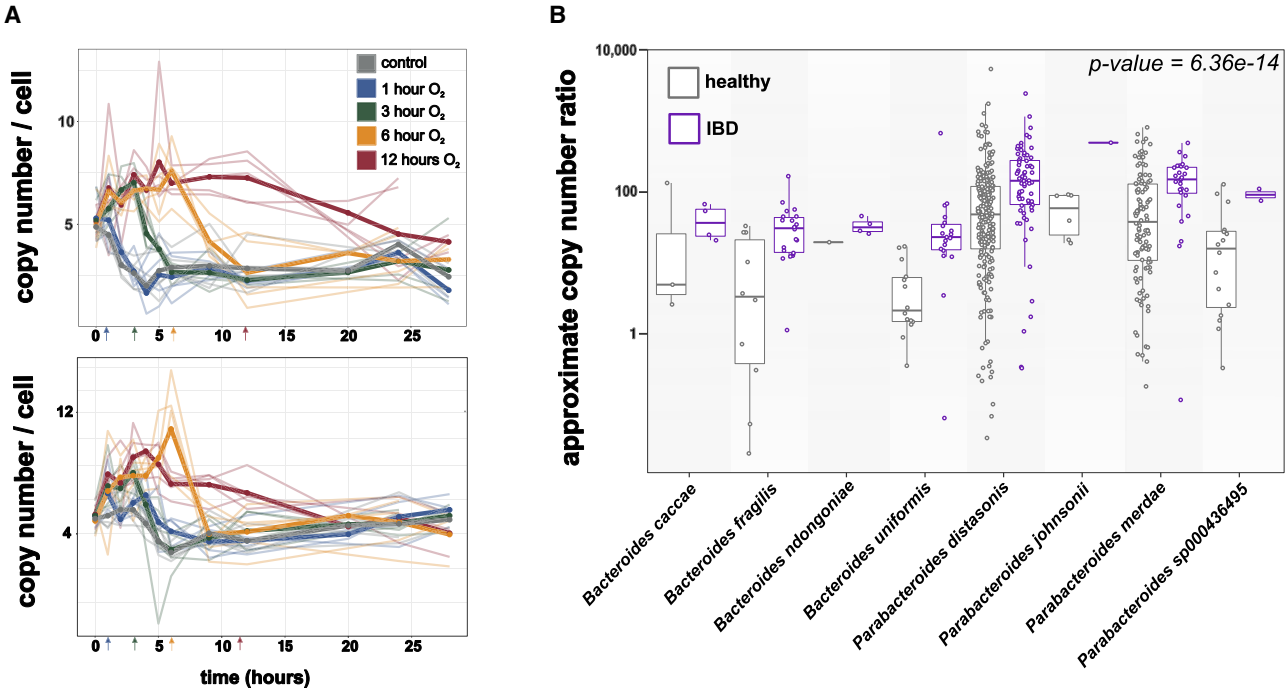


Figure 7. pBI143 copy number increases in stressful environments

(A) Copy number of pBI143 in *B. fragilis* cultures with increasing exposure to oxygen. Top: *B. fragilis* 214. Bottom: *B. fragilis* RI6. 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).

See also Tables S2 and S7.

pBI143 responds to oxidative stress *in vitro*, and its copy number is significantly higher in metagenomes from individuals who are diagnosed with IBD

Mobile genetic elements rely on their hosts for replication machinery, but many have developed mechanisms to increase their rates of replication and transfer during stressful conditions to increase the likelihood of their survival if the host cell dies.[98–101] To investigate whether the copy number of pBI143 changes as a function of stress, we first conducted an experiment with *B. fragilis* isolates that naturally carry pBI143.

Given that oxygen exposure upregulates oxidative stress response pathways in the anaerobic *B. fragilis*,[102] we exposed two different *B. fragilis* cultures, *B. fragilis* RI6 (which was isolated from a healthy individual) and *B. fragilis* 214 (which was isolated from a pouchitis patient[103]) to 21% oxygen for increasing periods of time (Figure 7A; Table S7). To calculate the copy number of pBI143 in culture, we quantified the ratio between the total number of plasmids and the total number of cells in culture using a qPCR with primers targeting pBI143 and a *B. fragilis*-specific gene we identified through pangenomics. As the length of oxygen exposure increased, the copy number of pBI143 per cell also increased. Notably, the copy number was quickly reduced to control levels once the cultures were returned to anaerobic conditions, indicating that copy-number fluctuation is a rapid and transient process that is dependent on host stress.

Oxidative stress is also a signature characteristic of inflammatory bowel disease (IBD), a group of intestinal disorders that cause inflammation of the gastrointestinal tract.[104] The dysregulation of the immune system during IBD typically leads to high levels of oxidative stress in the gut environment.[105] We thus hypothesized that, if oxidative stress is among the factors that drive the increased copy number of pBI143 in culture, one should expect a higher copy number of pBI143 in metagenomes from IBD patients compared with healthy controls.

To analyze the copy number of pBI143 in a given metagenome, we calculated the ratio of metagenomic read coverage between pBI143 and its bacterial host in metagenomes where pBI143 could confidently be assigned to a single host. With these considerations, we developed an 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 STAR Methods). We calculated the ACNR of pBI143 in 3,070 healthy and 1,350 IBD gut metagenomes (Table S1). Our analyses showed that the geometric mean of the ACNR for pBI143 and its host was 3.72 times larger (robust Wald 95% confidence interval [CI]: 2.66×–5.20×, p value < 10^{-13}) in IBD compared with healthy metagenomes, indicating that the pBI143 ACNR was significantly higher in individuals with IBD compared with those who were healthy (Figure 7B; Table S7).

The copy-number ratio of pBI143 to its *B. fragilis* host in culture calculated with qPCR primers was much lower (~5× on average) compared with its approximate copy-number ratio in healthy metagenomes (~120× 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 occurring communities experience vastly differ than those conditions encountered in culture media, even in the presence of oxygen. Nevertheless, the marked increase of the relative coverages of pBI143 and its host in IBD metagenomes suggest the potential utility of this cryptic plasmid for unbiased measurements of stress. Overall, these results show that both in metagenomes and experimental conditions, an increased copy number of pBI143 is a consistent feature in the presence of host stress.

DISCUSSION

Our work sheds light on a mysterious corner of life in the human gut. Even though pBI143 is found in greater than 90% of all individuals in some countries, the prevalence of this cryptic plasmid has gone unnoticed for almost four decades since its discovery by Smith, Rollins, and Parker.[44] The remarkable ecology, evolution, and potential practical applications of pBI143 that we characterized here through 'omics analyses as well as *in vitro* and *in vivo* experiments offer a glimpse of the world of understudied cryptic plasmids in the human gut and elsewhere.

The application of population genetics principles to pBI143 through the recovery of SNVs and SAAVs from gut metagenomes reveals not only the strong forces of purifying selection on the evolution of its sequence but also hints the presence of adaptive processes at localized amino acid positions that are variable in the critical parts of the DNA-interacting residues of the catalytic domain of its mobilization protein. With our current measurements of fitness, the presence of pBI143 appears to be slightly detrimental to bacterial host fitness *in vivo*, which makes this cryptic plasmid seem a mundane parasite using host machinery for replication without providing a benefit, and somewhat contradicting the strict evolutionary pressures that maintain its environmental sequence variants.

That said, our observations from naturally occurring gut environments include cases where pBI143 carries additional genes, likely acting as a natural shuttle vector. Although traditionally mobile genetic elements are classified as mutualistic or parasitic with respect to the bacterial host, the fluidity of pBI143 to fluctuate between the cryptic 2-gene state and the larger 3 or more gene state with potentially beneficial functions suggests that the boundaries between parasitism and mutualism for pBI143 are not clear cut. Instead, pBI143 may act as a "discretionary parasite," where it has a cryptic form for the majority of its existence in which it could be best described as a parasite, while occasionally being found with additional functions that may be beneficial to its host as a function of environmental pressures. Testing this hypothesis with future experimentation, and if true, investigating to what extent discretionary parasitism applies to other cryptic plasmids, may lead to a deeper understanding of the role of this enigmatic group of mobile genetic elements in microbial fitness under changing environmental conditions.

Our findings show that pBI143 has important potential practical applications beyond molecular biology. The first and most straightforward of these applications relies on the prevalence and human specificity of pBI143 to more sensitively detect human fecal contamination in water samples. Human fecal pollution is a global public health problem, and accurate and sensitive indicators of human fecal pollution are essential to identify and remediate contamination sources and to protect public health.[106] Although culture assays for *E. coli* or enterococci have historically been used to detect human fecal contamination in environmental samples, the common occurrence of these organisms in many different mammalian guts and the poor sensitivity of such assays motivated researchers in the past two decades to utilize PCR amplification of 16S rRNA genes, specifically those from human-specific *Bacteroides* and *Lachnospiraceae* populations, to detect human-specific fecal contamination with minimal cross-reactivity with animal feces.[80,81] Our benchmarking of pBI143 with qPCR revealed that pBI143 is an extremely sensitive and specific marker of human fecal contamination that typically occurs in human fecal samples and sewage in numbers that are several-fold higher than the state-of-the-art markers, which enabled the quantification of fecal contamination in samples where it had previously gone undetected. Another practical application of pBI143 takes advantage of its natural shuttle vector capabilities to incorporate additional genetic material into its backbone. Our demonstration that pBI143 (1) replicates in many abundant gut microbes, (2) can be stably introduced to new hosts, and (3) naturally acquires genetic material makes this cryptic plasmid an ideal natural payload delivery system for future therapeutics targeting the human gut microbiome. Indeed, our observations of pBI143 with cargo genes in metagenomes indicate that this likely happens in nature. Yet another practical implication of pBI143 is its potential to measure the level of stress in the human gut. Surveying thousands of samples from individuals who are healthy or diagnosed with IBD, our results show that across all bacterial hosts, the approximate copy number of pBI143 increases in individuals with IBD.

From a more philosophical point of view, the prevalence and high conservancy of pBI143 across globally distributed human populations questions the traditional definition of the "core" microbiome.[107] In its aim to define a core microbiome, the field of microbial ecology has primarily focused on bacteria, although sometimes including prevalent archaea or fungi.[108–111] However, our results indicate that there are mobile genetic elements that fit the standard criteria of prevalence to be defined as core. Broadening the definition of a core microbiome beyond microbial taxa may enable the recognition of other mobile genetic elements (e.g., plasmids, phages, and transposons) that are prevalent across human populations and fill critical gaps in our understanding of gut microbial ecology and evolution.

Limitations of the study

A precise understanding of the drivers of the ecology, evolution, and function of this cryptic plasmid in the human gut demands additional research. Given the complexity of generating bacterial strains with plasmids containing no antibiotic markers, we were limited in our ability to test the fitness of different pBI143 versions or its impact on the fitness of different taxa. These experiments would more definitively show if different versions have competitive advantages over others or if pBI143 differentially impacts fitness in different hosts. Our fitness experiments relied on

growth measurements as the sole indicator of pBI143 impact on the host. Future experiments could explore if there are transcriptional changes that occur as a result of pBI143 carriage.

In our structure-informed interpretations of the genetic variants of the *mobA* gene, we point to residues that are involved in different MobA versions binding to the oriT sequence. Although a deeper investigation into the functional role of observed variants was beyond the scope of this work, *in vivo* or *in vitro* experiments that explicitly test SAAVs of MobA are necessary to establish robust insights into the DNA binding properties of this gene.

Our observation that pBI143 occasionally occurs in human gut metagenomes with a larger number of genes show the likelihood of cryptic plasmids to uptake additional DNA into their backbones. Our study was limited to computational characterizations of these larger, cargo-containing plasmids. Establishing precise insights into the functional impact of the additional genes in pBI143 and whether they serve as fitness determinants for host bacterial populations as a function of environmental requirements represent an exciting future direction to investigate additional roles of cryptic plasmids in microbial adaptation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.01.039.

ACKNOWLEDGMENTS

We thank the members of the Meren Lab (https://merenlab.org) and Comstock Lab (https://comstocklab.uchicago.edu/) for helpful discussions, Jason Koval for help procuring bacterial cultures, and the Duchossois Family Institute WGS facility for sequencing constructs and providing access to their strain bank. We thank Melinda Bootsma for help with the qPCRs on water and sewage samples and Jessika Füssel for designing the graphical abstract. E.C.F. acknowledges support from the University of Chicago International Student Fellowship, A.D.W. acknowledges support from NIGMS R35 GM133420, and L.E.C. acknowledges support from the Duchossois Family Institute. I.V. acknowledges support from the National Science Foundation Graduate Research Fellowship under grant number 1746045. S.S. acknowledges funding from the Swiss National Science Foundation (NCCR Microbiomes - 51NF40_180575) and support from the ETH IT services for calculations that were carried out on the ETH Euler cluster. Additional support for E.C.F. came from an NIH NIDDK grant (RC2 DK122394) to E.B.C. The authors thank the University of Chicago Center for Data and Computing for their support. This project was funded by University of Chicago start-up funds to A.M.E.

AUTHOR CONTRIBUTIONS

E.C.F. and A.M.E. conceived the study. K.L. developed methodology. R.M.M., E.K., and A.M.E. developed computational analysis tools. E.C.F., M.S.S., P.A.R., S.L.M., A.D.W., and A.M.E. performed formal analyses. E.C.F., K.L., M.L.S., and L.E.C. conducted investigations. T.M., M.K.Y., M.M., E.B.C., H.-J.R., S.S., and S.L.M. provided resources. E.C.F., M.S.S., H.-J.R., S.S., A.D.W., and A.M.E. curated data. E.C.F., M.S.S., P.A.R., and A.M.E. prepared the figures. E.B.C. and A.M.E. acquired funding. E.C.F. and A.M.E. wrote the paper with critical input from all authors. L.E.C. and A.M.E. supervised the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 29, 2023
Revised: October 3, 2023
Accepted: January 25, 2024
Published: February 29, 2024

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. del Solar, G., Giraldo, R., Ruiz-Echevarría, M.J., Espinosa, M., and Díaz-Orejas, R. (1998). Replication and Control of Circular Bacterial Plasmids. Microbiol. Mol. Biol. Rev. *62*, 434–464.
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. Environ. Biotechnol., 47–62.
9. Thouand, G., and Marks, R. (2016). Bioluminescence: Fundamentals and Applications in Biotechnology*3* (Springer).
10. Palomino, A., Gewurz, D., DeVine, L., Zajmi, U., Moralez, J., Abu-Rumman, F., Smith, R.P., and Lopatkin, A.J. (2023). Metabolic genes on conjugative plasmids are highly prevalent in Escherichia coli and can protect against antibiotic treatment. ISME J. *17*, 151–162.
11. Al-Shayeb, B., Schoelmerich, M.C., West-Roberts, J., Valentin-Alvarado, L.E., Sachdeva, R., Mullen, S., Crits-Christoph, A., Wilkins, M.J., Williams, K.H., Doudna, J.A., et al. (2022). Borgs are giant genetic elements with potential to expand metabolic capacity. Nature *610*, 731–736.
12. Leonard, S.P., Perutka, J., Powell, J.E., Geng, P., Richhart, D.D., Byrom, M., Kar, S., Davies, B.W., Ellington, A.D., Moran, N.A., et al. (2018). Genetic Engineering of Bee Gut Microbiome Bacteria with a Toolkit for Modular Assembly of Broad-Host-Range Plasmids. ACS Synth. Biol. *7*, 1279–1290.
13. Slattery, S.S., Diamond, A., Wang, H., Therrien, J.A., Lant, J.T., Jazey, T., Lee, K., Klassen, Z., Desgagné-Penix, I., Karas, B.J., et al. (2018). An Expanded Plasmid-Based Genetic Toolbox Enables Cas9 Genome Editing and Stable Maintenance of Synthetic Pathways in Phaeodactylum tricornutum. ACS Synth. Biol. *7*, 328–338.
14. Rihn, S.J., Merits, A., Bakshi, S., Turnbull, M.L., Wickenhagen, A., Alexander, A.J.T., Baillie, C., Brennan, B., Brown, F., Brunker, K., et al. (2021). A plasmid DNA-launched SARS-CoV-2 reverse genetics system and coronavirus toolkit for COVID-19 research. PLoS Biol. *19*, e3001091.
15. Salvay, D.M., Zelivyanskaya, M., and Shea, L.D. (2010). Gene delivery by surface immobilization of plasmid to tissue-engineering scaffolds. Gene Ther. *17*, 1134–1141.
16. Mutuku, C., Gazdag, Z., and Melegh, S. (2022). Occurrence of antibiotics and bacterial resistance genes in wastewater: resistance mechanisms and antimicrobial resistance control approaches. World J. Microbiol. Biotechnol. *38*, 152.
17. Dimitriu, T. (2022). Evolution of horizontal transmission in antimicrobial resistance plasmids. Microbiology (Reading), 168. https://doi.org/10.1099/mic.0.001214.
18. Prestinaci, F., Pezzotti, P., and Pantosti, A. (2015). Antimicrobial resistance: a global multifaceted phenomenon. Pathog. Glob. Health *109*, 309–318.
19. Kang, X., Li, C., and Luo, Y. (2020). Cloning of pAhX22, a small cryptic plasmid from Aeromonas hydrophila, and construction of a pAhX22-derived shuttle vector. Plasmid *108*, 102490.
20. Oliveira, V., Polónia, A.R.M., Cleary, D.F.R., Huang, Y.M., de Voogd, N.J., da Rocha, U.N., and Gomes, N.C.M. (2021). Characterization of putative circular plasmids in sponge-associated bacterial communities using a selective multiply-primed rolling circle amplification. Mol. Ecol. Resour. *21*, 110–121.
21. Shareck, J., Choi, Y., Lee, B., and Miguez, C.B. (2004). Cloning vectors based on cryptic plasmids isolated from lactic acid bacteria: their characteristics and potential applications in biotechnology. Crit. Rev. Biotechnol. *24*, 155–208.
22. Attéré, S.A., Vincent, A.T., Paccaud, M., Frenette, M., and Charette, S.J. (2017). The Role for the Small Cryptic Plasmids As Moldable Vectors for Genetic Innovation in Aeromonas salmonicida subsp. salmonicida. Front. Genet. *8*, 211.
23. Challacombe, J.F., Pillai, S., and Kuske, C.R. (2017). Shared features of cryptic plasmids from environmental and pathogenic Francisella species. PLoS One *12*, e0183554.
24. Roberts, M.C. (1989). Plasmids of Neisseria gonorrhoeae and other Neisseria species. Clin. Microbiol. Rev. *2* (*Suppl*), S18–S23.
25. Zillig, W., Prangishvilli, D., Schleper, C., Elferink, M., Holz, I., Albers, S., Janekovic, D., and Götz, D. (1996). Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. FEMS Microbiol. Rev. *18*, 225–236.
26. Heuer, H., and Smalla, K. (2012). Plasmids foster diversification and adaptation of bacterial populations in soil. FEMS Microbiol. Rev. *36*, 1083–1104.
27. Vincent, A.T., Hosseini, N., and Charette, S.J. (2021). The *Aeromonas salmonicida* plasmidome: a model of modular evolution and genetic diversity. Ann. N. Y. Acad. Sci. *1488*, 16–32.
28. Kothari, A., Wu, Y.W., Chandonia, J.M., Charrier, M., Rajeev, L., Rocha, A.M., Joyner, D.C., Hazen, T.C., Singer, S.W., and Mukhopadhyay, A. (2019). Large Circular Plasmids from Groundwater Plasmidomes Span Multiple Incompatibility Groups and Are Enriched in Multimetal Resistance Genes. mBio *10*.
29. Thomas, C.M. (2014). Evolution and Population Genetics of Bacterial Plasmids. Plasmid Biol., 507–528.
30. Iranzo, J., Puigbò, P., Lobkovsky, A.E., Wolf, Y.I., and Koonin, E.V. (2016). Inevitability of Genetic Parasites. Genome Biol. Evol. *8*, 2856–2869.
31. Levin, B.R., and Stewart, F.M. (1980). The population biology of bacterial plasmids: a priori conditions for the existence of mobilizable nonconjugative factors. Genetics *94*, 425–443.
32. Simonsen, L. (1991). The existence conditions for bacterial plasmids: Theory and reality. Microb. Ecol. *22*, 187–205.
33. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844.
34. Andreopoulos, W.B., Geller, A.M., Lucke, M., Balewski, J., Clum, A., Ivanova, N.N., and Levy, A. (2022). Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. Nucleic Acids Res. *50*, e17.
35. Krawczyk, P.S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. *46*, e35.
36. Zhou, F., and Xu, Y. (2010). cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics *26*, 2051–2052.
37. Pellow, D., Mizrahi, I., and Shamir, R. (2020). PlasClass improves plasmid sequence classification. PLoS Comput. Biol. *16*, e1007781.
38. Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., and Hasman, H. (2014). *In Silico* Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrob. Agents Chemother. *58*, 3895–3903.
39. Robertson, J., and Nash, J.H.E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb. Genom. *4*, e000206.
40. Garcillán-Barcia, M.P., Francia, M.V., and de la Cruz, F. (2009). The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol. Rev. *33*, 657–687.
41. Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., and Shamir, R. (2017). Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics *33*, 475–482.

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

43. Yu, M.K., Fogarty, E.C., and Eren, A.M. (2024). Diverse plasmid systems and their ecology across human gut metagenomes revealed by PlasX and MobMess. Nature Microbiol. https://doi.org/10.1038/s41564-024-01610-3.

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

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

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

47. Lee, Y.K., Mehrabian, P., Boyajian, S., Wu, W.L., Selicha, J., Vonderfecht, S., and Mazmanian, S.K. (2018). The Protective Role of *Bacteroides fragilis* in a Murine Model of Colitis-Associated Colorectal Cancer. mSphere *3*, e00587-18.

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

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

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

51. Thompson, J.S., and Malamy, M.H. (1990). Sequencing the gene for an imipenem-cefoxitin-hydrolyzing enzyme (CfiA) from Bacteroides fragilis TAL2480 reveals strong similarity between CfiA and Bacillus cereus beta-lactamase II. J. Bacteriol. *172*, 2584–2593.

52. Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., et al. (2015). Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat. Commun. *6*, 6528.

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

54. Raymond, F., Ouameur, A.A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.L., Giroux, R., Bérubé, È., et al. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. ISME J. *10*, 707–720.

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

56. Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y., Zhang, L., et al. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. Genome Biol. *18*, 142.

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

58. Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., Ward, K.J., Jackson, M.A., Xia, Y., Chen, X., et al. (2016). Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. Cell Syst. *3*, 572–584.e3.

59. Pasolli, E., Asnicar, F., Serena Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20.

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

61. Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S.M., et al. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe *24*, 146–154.e4.

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

63. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., et al. (2015). Personalized Nutrition by Prediction of Glycemic Responses. Cell *163*, 1079–1094.

64. Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe *24*, 133–145.e5.

65. Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr. Biol. *25*, 1682–1693.

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

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

68. Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou, B., et al. (2016). Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. Sci. Rep. *6*, 34826.

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

70. Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 6505.

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

72. Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 690–703.

73. Lou, Y.C., Olm, M.R., Diamond, S., Crits-Christoph, A., Firek, B.A., Baker, R., Morowitz, M.J., and Banfield, J.F. (2021). Infant gut strain persistence is associated with maternal origin, phylogeny, and traits including surface adhesion and iron acquisition. Cell Rep. Med. *2*, 100393.

74. Human Microbiome Project Consortium (2012). A framework for human microbiome research. Nature *486*, 215–221.

75. Gupta, V.K., Paul, S., and Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Front. Microbiol. *8*, 1162.

76. Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat. Commun. *5*, 4498.

77. Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S., Draper, L.A., Gonzalez-Tortuero, E., Ross, R.P., and Hill, C. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host Microbe *24*, 653–664.e6.

78. 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.

79. Amato, K.R., Yeoman, C.J., Kent, A., Righini, N., Carbonero, F., Estrada, A., Gaskins, H.R., 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.

80. 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*, e00309-18.

81. 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.

82. 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. USA *114*, E6526–E6535.

83. 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*, e46497.

84. 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.

85. 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.

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

87. Debray, R., Herbert, R.A., Jaffe, A.L., Crits-Christoph, A., Power, M.E., and Koskella, B. (2022). Priority effects in microbiome assembly. Nat. Rev. Microbiol. *20*, 109–121.

88. 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.

89. 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.

90. 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.

91. Mavrich, T.N., and Hatfull, G.F. (2019). Evolution of Superinfection Immunity in Cluster A Mycobacteriophages. mBio *10*, e00971-19.

92. 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*, e00841-20.

93. Alonso-Del Valle, A., León-Sampedro, R., Rodríguez-Beltrán, J., DelaFuente, J., Hernández-García, M., Ruiz-Garbajosa, P., Cantón, R., Peña-Miller, R., and San Millán, A. (2021). Variability of plasmid fitness effects contributes to plasmid persistence in bacterial communities. Nat. Commun. *12*, 2653.

94. Brockhurst, M.A., and Harrison, E. (2022). Ecological and evolutionary solutions to the plasmid paradox. Trends Microbiol. *30*, 534–543.

95. Bergstrom, C.T., Lipsitch, M., and Levin, B.R. (2000). Natural selection, infectious transfer and the existence conditions for bacterial plasmids. Genetics *155*, 1505–1519.

96. Lopatkin, A.J., Meredith, H.R., Srimani, J.K., Pfeiffer, C., Durrett, R., and You, L. (2017). Persistence and reversal of plasmid-mediated antibiotic resistance. Nat. Commun. *8*, 1689.

97. Stevenson, C., Hall, J.P., Harrison, E., Wood, A., and Brockhurst, M.A. (2017). Gene mobility promotes the spread of resistance in bacterial populations. ISME J. *11*, 1930–1932.

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

99. 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.

100. 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.

101. 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.

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

103. Vineis, J.H., Ringus, D.L., Morrison, H.G., Delmont, T.O., Dalal, S., Raffals, L.H., Antonopoulos, D.A., Rubin, D.T., Eren, A.M., Chang, E.B., et al. (2016). Patient-Specific Bacteroides Genome Variants in Pouchitis. mBio *7*, e01713-16.

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

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

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

107. Neu, A.T., Allen, E.E., and Roy, K. (2021). Defining and quantifying the core microbiome: Challenges and prospects. Proc. Natl. Acad. Sci. USA *118*, e2104429118.

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

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

110. Shetty, S.A., Kuipers, B., Atashgahi, S., Aalvink, S., Smidt, H., and de Vos, W.M. (2022). Inter-species Metabolic Interactions in an In-vitro Minimal Human Gut Microbiome of Core Bacteria. NPJ Biofilms Microbiomes *8*, 21.

111. Nash, A.K., Auchtung, T.A., Wong, M.C., Smith, D.P., Gesell, J.R., Ross, M.C., Stewart, C.J., Metcalf, G.A., Muzny, D.M., Gibbs, R.A., et al. (2017).

The gut mycobiome of the Human Microbiome Project healthy cohort. Microbiome *5*, 153.

112. Privitera, G., Dublanchet, A., and Sebald, M. (1979). Transfer of multiple antibiotic resistance between subspecies of Bacteroides fragilis. J. Infect. Dis. *139*, 97–101.
113. Ben Chorin, A., Masrati, G., Kessel, A., Narunsky, A., Sprinzak, J., Lahav, S., Ashkenazy, H., and Ben-Tal, N. (2020). ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. Protein Sci. *29*, 258–267.
114. Eren, A.M., Vineis, J.H., Morrison, H.G., and Sogin, M.L. (2013). A filtering method to generate high quality short reads using illumina paired-end technology. PLoS One *8*, e66643.
115. Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420–1428.
116. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
117. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
118. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.
119. Delano, W.L. (2002). The PyMOL molecular graphics system. http://www.pymol.org/.
120. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
121. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.
122. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.
123. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.
124. Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. ICWSM *3*, 361–362.
125. Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One *9*, e98679.
126. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257.
127. Chen, Y.T., Williamson, B.D., Okonek, T., Wolock, C.J., Spieker, A.J., Hee Wai, T.Y., Hughes, J.P., Emerson, S.S., and Willis, A.D. (2022). rigr: Regression, Inference, and General Data Analysis Tools in R. J. Open Source Softw. *7*, 4847.
128. Lassmann, T. (2019). Kalign 3: multiple sequence alignment of large data sets. Bioinformatics *36*, 1928–1929.
129. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).
130. Eren, A.M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S.E., Schechter, M.S., Fink, I., Pan, J.N., Yousef, M., et al. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. Nat. Microbiol. *6*, 3–6.
131. Shaiber, A., Willis, A.D., Delmont, T.O., Roux, S., Chen, L.X., Schmid, A.C., Yousef, M., Watson, A.R., Lolans, K., Esen, Ö.C., et al. (2020). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol. *21*, 292.
132. Köster, J., and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.
133. Ruscheweyh, H.J., Milanese, A., Paoli, L., Karcher, N., Clayssen, Q., Keller, M.I., Wirbel, J., Bork, P., Mende, D.R., Zeller, G., et al. (2022). Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. Microbiome *10*, 212.
134. Utter, D.R., Borisy, G.G., Eren, A.M., Cavanaugh, C.M., and Mark Welch, J.L. (2020). Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. Genome Biol. *21*, 293.
135. García-Bayona, L., and Comstock, L.E. (2019). Streamlined Genetic Manipulation of Diverse Bacteroides and Parabacteroides Isolates from the Human Gut Microbiota. mBio *10*, e01762-19.
136. Zitomersky, N.L., Coyne, M.J., and Comstock, L.E. (2011). Longitudinal Analysis of the Prevalence, Maintenance, and IgA Response to Species of the Order Bacteroidales in the Human Gut. Infect. Immun. *79*, 2012–2020.
137. Evans, J.C., McEneany, V.L., Coyne, M.J., Caldwell, E.P., Sheahan, M.L., Von, S.S., Coyne, E.M., Tweten, R.K., and Comstock, L.E. (2022). A proteolytically activated antimicrobial toxin encoded on a mobile plasmid of Bacteroidales induces a protective response. Nat. Commun. *13*, 4258.
138. Pluta, R., Boer, D.R., and Coll, M. (2014). MobM Relaxase Domain (MOBV; Mob_Pre) bound to plasmid pMV158 oriT DNA (22nt). Mn-bound crystal structure at pH 4.6. https://www.wwpdb.org/pdb?id=pdb_00004lvi.
139. Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res. *37*, D323–D327.
140. Comstock, L.E., Pantosti, A., and Kasper, D.L. (2000). Genetic diversity of the capsular polysaccharide C biosynthesis region of Bacteroides fragilis. Infect. Immun. *68*, 6182–6188.
141. Conrad, S., Oethinger, M., Kaifel, K., Klotz, G., Marre, R., and Kern, W.V. (1996). gyrA mutations in high-level fluoroquinolone-resistant clinical isolates of Escherichia coli. J. Antimicrob. Chemother. *38*, 443–455.
142. Wawrzyniak, P., Płucienniczak, G., and Bartosik, D. (2017). The Different Faces of Rolling-Circle Replication and Its Multifunctional Initiator Proteins. Front Microbiol *8*, 2353. https://doi.org/10.3389/fmicb.2017.02353.
143. Delmont, T.O., and Eren, A.M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ *6*, e4320. https://doi.org/10.7717/peerj.4320.
144. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res *30*, 1575–1584. https://doi.org/10.1093/nar/30.7.1575.
145. Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J. Comput. Biol. *7*, 203–214.
146. Sakamoto, M., and Ohkuma, M. (2010). Usefulness of the hsp60 gene for the identification and classification of Gram-negative anaerobic rods. J. Med. Microbiol. *59*, 1293–1302.
147. Gallie, D.R., Fortner, D., Peng, J., and Puthoff, D. (2002). ATP-dependent hexameric assembly of the heat shock protein Hsp101 involves multiple interaction domains and a functional C-proximal nucleotide-binding domain. J. Biol. Chem. *277*, 39617–39626.
148. Whelan, J.A., Russell, N.B., and Whelan, M.A. (2003). A method for the absolute quantification of cDNA using real-time PCR. J. Immunol. Methods *278*, 261–269.
149. Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. *55*, 611–622.
150. Olds, H.T., Corsi, S.R., Dila, D.K., Halmo, K.M., Bootsma, M.J., and McLellan, S.L. (2018). High levels of sewage contamination released from urban areas after storm events: A quantitative survey with sewage specific bacterial indicators. PLoS Med. *15*, e1002614.
151. Feng, S., Ahmed, W., and McLellan, S.L. (2020). Ecological and Technical Mechanisms for Cross-Reaction of Human Fecal Indicators with Animal Hosts. Appl. Environ. Microbiol. *86*, e02319-19.

152. Lenaker, P.L., Corsi, S.R., McLellan, S.L., Borchardt, M.A., Olds, H.T., Dila, D.K., Spencer, S.K., and Baldwin, A.K. (2018). Human-Associated Indicator Bacteria and Human-Specific Viruses in Surface Water: A Spatial Assessment with Implications on Fate and Transport. Environ. Sci. Technol. *52*, 12162–12171.

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

154. USGS. USGS water data for the nation. 2023. https://waterdata.usgs.gov/nwis.

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

156. Poyet, M., Groussin, M., Gibbons, S.M., Avila-Pacheco, J., Jiang, X., Kearney, S.M., Perrotta, A.R., Berdy, B., Zhao, S., Lieberman, T.D., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat. Med. *25*, 1442–1452.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Bacteroides fragilis 214	Vineis et al.[103]	Bacteroides fragilis 214
Bacteroides fragilis RI6	This study.	Bacteroides fragilis RI6
Bacteroides fragilis 638R	Privitera et al.[112]	Bacteroides fragilis 638R (also available from the Comstock Lab).
Bacteroides fragilis 9343	https://www.atcc.org/products/25285	Bacteroides fragilis (Veillon and Zuber) Castellani and Chalmers 25285
EC100	https://www.fishersci.com/shop/products/transformax-ec100-elec-ecoli/NC9768848	Catalog No.NC9768848
E. coli S17 λpir	https://www.fishersci.com/shop/products/e-coli-s17-1pir/NC1526716	Catalog No.NC1526716
Critical commercial assays		
Gibson Assembly® Master Mix	New England Biosciences	E2611S
Qiagen miniprep kit	Qiagen	Cat. No. 27104
Deposited data		
Duchossois Family Institute bacterial genomes	Duchossois Family Institute at the University of Chicago	https://dfi.uchicago.edu/node/566
Raw sequencing data for 100,029 metagenomes	See Table S1 for accession information.	N/A
pBI143	GenBank	GenBank: U30316.1
Genome Taxonomy Database	GTDB	https://gtdb.ecogenomic.org/
Experimental models: Organisms/strains		
Germ-free C57BL/6J mice	The Jackson Laboratory	https://www.jax.org/strain/000664
Oligonucleotides		
See Table S2 for primer sequences	This study	N/A
Software and algorithms		
anvi'o v7	Eren et al.[113]	https://anvio.org/
snakemake v5.10	Köster and Rahmann[95]	https://snakemake.readthedocs.io/en/stable/
illumina-utils v1.4.4	Eren et al.[114]	https://github.com/merenlab/illumina-utils
IDBA_UD v1.1.2	Peng et al.[115]	https://github.com/loneknightpy/idba
Bowtie2 v2.4	Langmead and Salzberg[116]	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml
samtools v1.9	Li et al.[117]	https://www.htslib.org/
Prodigal v2.6.3	Hyatt et al.[118]	https://github.com/hyattpd/Prodigal
AlphaFold 2	Mirdita et al.[84]	https://deepmind.google/technologies/alphafold/
PyMol	Delano[119]	https://pymol.org/2/
ConSurf	Ben Chorin et al.[113]	https://consurf.tau.ac.il/
BLAST	Altschul et al.[120]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
MUSCLE v3.8.1551	Edgar[121]	https://www.ebi.ac.uk/Tools/msa/muscle/
trimAl	Capella-Gutiérrez et al.[122]	http://trimal.cgenomics.org/trimal
IQ-TREE 2.2.0-beta	Nguyen et al.[123]	http://www.iqtree.org/
Gephi	Bastian et al.[124]	https://gephi.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ForceAtlas2	Jacomy et al.[125]	https://github.com/gephi/gephi/wiki/Force-Atlas-2
Kraken 2.0.8-beta	Wood et al.[126]	https://bio.tools/kraken2
PlasX	Yu et al.[43]	https://github.com/michaelkyu/PlasX
	Chen et al.[127]	https://github.com/statdivlab/rigr
Kalign	Lassmann[128]	https://www.ebi.ac.uk/Tools/msa/kalign
ggplot2	Wickham[129]	https://cran.r-project.org/web/packages/ggplot2/index.html
Inkscape	N/A	http://inkscape.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, A. Murat Eren (meren@hifmb.de).

Materials availability

Bacterial cultures for host range investigations, which are listed in Table S1, are courtesy of The Duchossois Family Institute (https://dfi.uchicago.edu/). Primer sequences can be found in Table S3. *B. fragilis* strains with pBI143 are available upon request from the Comstock Lab collection (https://comstocklab.uchicago.edu/).

Data and code availability

- All molecular data used in this study are publicly available via the NCBI Sequence Read Archive, and Table S1 reports the accession numbers for all genomes and metagenomes.
- Original code accompanying this paper, anvi'o data products that describe metagenomic read recruitment results, and sequences for pBI143 versions and bioinformatics workflows is available at https://merenlab.org/data/pBI143.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODELS AND STUDY PARTICIPANT DETAILS

Husbandry and housing conditions of experimental animals

Maintenance and handling of all mice were carried out in the Gnotobiotic Mouse Facility at the University of Chicago. All experimentation was approved by the Institutional Animal Care and Use Committee at the University of Chicago in compliance with federal regulations. Mice were housed in individually vented sterile cages at 22°C and 30-70% humidity and provided water and food ad libitum. The mouse chow pellets were LabDiet® JL Rat and Mouse/Auto 6F 5K67 complete life-cycle diet that was autoclaved to sterilize. Three male and three female 10-15 week old germ-free C57BL/6J mice were gavaged with a 1:1 inoculum of *B. fragilis* 638R[112]: *B. fragilis* 638R pBI143. Males and females were housed separately in isocages and remained gnotobiotic for the duration of the experiment.

Bacterial cell culture

All experiments involving Bacteroidales organisms were performed in a Coy Anaerobic Chamber unless otherwise specified. The chamber atmosphere is H_2(7.5%), CO_2(10%), N_2(82.5%). We streaked the strain of interest onto Brain-Heart Infusion agar from Fisher Scientific supplemented with yeast extract, hemin and Vitamin K (BHIS media). We incubated the plates for 48 hours at 37°C. Appropriate antibiotics were added to plates or media as described in method details. Unless otherwise specified, we inoculated BHIS broth with colonies from the agar plates, and grew cultures for up to 24 hours at 37°C before performing downstream experiments. The *Escherichia coli* strains we used in this study were grown aerobically. We streaked the strain of interest onto Luria-Bertani (LB) agar plates with appropriate antibiotics and incubated at 37°C for 24 hours. When needed, we inoculated LB broth with strain of interest and grew at 37°C for 12-15 hours before performing downstream experiments.

METHOD DETAILS

Genomes and metagenomes

We acquired the original pBI143 genome from the National Center for Biotechnological Information (GenBank: U30316.1). We manually assembled the three reference versions of pBI143 (Version 1, 2 and 3) from metagenomes samples USA0006, CHI0054 and

ISR0084. We acquired 717 human gut isolate genomes from the Duchossois Family Institute collection (Table S1). For our in-depth metagenomic analyses, we downloaded (1) 4,516 healthy human adult gut metagenomes from the National Center for Biotechnology Information (NCBI) from (Australia (Accession ID: PRJEB6092), Austria,[52] Bangladesh,[53] Canada,[54] China,[55,56] Denmark,[57] England,[58] Ethiopia,[59] Fiji,[60] Finland,[61] India,[62] Israel,[63] Italy,[64,65] Japan,[66] Korea,[67] Madagascar,[59] Mongolia,[59,68] Netherlands,[69] Peru,[70] Spain,[71] Sweden,[72] Tanzania,[65] and the USA[70,73,74]) (Table S1), (2) 1,096 gut metagenomes from infant-mother pairs from Italy, Finland, Sweden and the USA from NCBI (Table S1), and (3) 77 globally distributed sewage metagenomes (Table S1). We also conducted a large-scale survey of pBI143 in 100,029 metagenomes found on public databases (Table S1, we discarded 85 metagenomes which we could not reliably assign to a biome).

Metagenomic assembly, read recruitment, coverage and detection statistics

Unless otherwise specified, we performed all metagenomic analyses throughout the manuscript within the open-source anvi'o v7 software ecosystem (https://anvio.org).[130] We automated assembly and read recruitment steps using the anvi'o metagenomics workflow[131] which used snakemake v5.10.[132] To quality-filter genomic and metagenomic raw paired-end reads (with the exception of the dataset of 100,029 metagenomes) we used illumina-utils v1.4.4[114] program 'iu-filter-quality-minoche' with default parameters, and IDBA_UD v1.1.2 with the flag '–min_contig 1000' to assemble the metagenomes.[115] The 100,029 metagenomes were quality screened to remove adapters and spike-in control sequences, as well reads that matched to a reference human genome and those that were low quality based on the quality scores assigned by the sequencer.[133] We used Bowtie2 v2.4[116] to recruit reads from the metagenomes to reference sequences and samtools v1.9[117] to convert resulting SAM files into sorted and indexed BAM files. We generated anvi'o contigs databases (https://anvio.org/m/contigs-db) using the program 'anvi-gen-contigs-database', during which Prodigal v2.6.3[118] identified open reading frames. We created anvi'o profile databases (https://anvio.org/m/profile-db) from the mapping results for each metagenome using 'anvi-profile', which stores coverage and detection statistics, and 'anvi-merge' to combine all profiles together. To recover coverage and detection statistics for a given merged profile database, we used 'anvi-summarize' with '–init-gene-coverages' flag. To profile the distribution pattern of pBI143 across the larger set of 100K metagenomes we used the 'anvi-profile-blitz', which rapidly profiles read recruitment results to only report essential statistics (https://anvio.org/m/bam-stats-txt).

Detection of pBI143 and crAssphage in metagenomes

Using mean coverage to assess the occurrence of a given sequence in a given sample based on metagenomic read recruitment can yield misleading insights due to non-specific read recruitment (i.e., recruitment of reads from metagenomes to a reference sequence from non-target populations). Thus, we relied upon the detection statistic reported by anvi'o, which is a measure of the proportion of the nucleotides in a given sequence that are covered by at least one short read. We considered pBI143 was present in a metagenome only if its detection value was 0.5 or above. Values of detection in metagenomic read recruitment results often follow a bimodal distribution for populations that are present and absent (see Supplementary Figure 2 in ref. Utter et al.[134]). Thus, 0.5 is a conservative cutoff to minimize a false-positive signal to assume presence.

Presence of distinct pBI143 versions in a genome or metagenome

We used the results of individual read recruitments to each known version of pBI143 to measure the coverage of each gene in pBI143 in samples that had a detection of greater than 0.9 and compared the ratio of the coverage of each gene. The pBI143 version where the genes have the most even coverage ratio was considered the predominant version in that genome or metagenome.

Addition of *tetQ* to pIB143

To study transfer of pBI143 from *Phocaeicola vulgatus* MSK 17.67 to other Bacteroidales species, we added *tetQ* to pBI143. We PCR amplified *tetQ* from *Bacteroides caccae* CL03T12C61 and inserted it at the site shown in Figure S2 (all primers are listed in Table S3). We PCR amplified the DNA regions flanking each side of this insertion site and the three PCR products were cloned into BamHI-digested pLGB13.[135] We conjugally transferred this plasmid into *Phocaeicola vulgatus* MSK 17.67 and selected cointegrates on gentamycin 200 μg/ml and erythromycin 10 μg/ml. We passaged the cointegrate in non-selective media and selected the resolvents by plating on anhydrotetracycline (75 ng/ml). We confirmed pIB143 contained *tetQ* by WGS of the strain at the DFI Microbiome Metagenomics Facility.

Transfer assays

The strains tested as recipients in the pBI143-*tetQ* transfer assays were *Parabacteroides johnsonii* CL02T12C29 and *Bacteroides ovatus* D2, both erythromycin resistant and tetracycline sensitive. We grew the donor strain *Phocaeicola vulgatus* MSK 17.67 pBI143-*tetQ* and recipient strains to an OD600 of ~ 0.7 and mixed them at a 10:1 ratio (v:v) donor to recipient, and spotted 10 μl onto BHIS plates and grew them anaerobically for 20 h. We resuspended the co-culture spot in 1 mL basal media and cultured 10-fold serial dilutions on plates with erythromycin (to calculate number of recipients) or erythromycin and tetracycline (4.5 μg/ml) (to select for transconjugants). We performed multiplex PCR[136,137] to confirm that the tetracycline and erythromycin resistant colonies were the recipient strain containing pBI143-*tetQ* (Figure S2).

Purifying selection and single nucleotide variant characterization

We calculated dN/dS ratios in anvi'o[85] (see https://merenlab.org/data/anvio-structure/chapter-IV/#calculating-dndstextgene-for-1-gene for details). To determine the mutational landscape of pBI143 across metagenomes, we first identified all variable positions present in the reference pBI143 sequences. We used 'anvi-script-gen-short-reads' to generate artificial short reads from the version 2 and version 3 pBI143 sequences and recruited these reads to the pBI143 version 1 sequence to generate data similar to the read recruitment from metagenomes. Then, we combined these data with the read recruitment results from the global human gut metagenomes and sewage metagenomes against the same pBI143 version. We used 'anvi-gen-variability-profile' to recover the location of each single-nucleotide variant (SNV) from both artificial read recruitment results and the global human gut and sewage metagenomes in which the Q2Q3 coverage of pBI143 exceeded 10X Q2Q3. We then compared the SNV positions in each gut or sewage metagenome to those positions that varied between the three versions of pBI143 and for each metagenome we calculated the number of SNVs that occurred in a location that was also variable in pBI143 versions, and the number of SNVs in metagenomes that did not match to a known variant between pBI143 versions. To calculate the number of 'non-consensus SNVs' in a metagenome, we ran 'anvi-gen-profile-database' on the same metagenomes, with the flags '–gene-caller-ids 0' (i.e., the gene call in the anvi'o contigs-db that matched to the mobA gene in pBI143), '–min-departure-from-consensus 0.1' (to minimize noise), '–include-contig-names' (for a verbose output) and '–quince-mode' (to include coverage information for SNV positions even from metagenomes in which they do not vary). The resulting file described the variation in every single SNV position across metagenomes, and gave access to the 'departure from consensus' statistic to identify positions that are variable in the environment.

pBI143 structural and polymorphism analysis

To explore the impact of single-amino acid variants (SAAVs) on the protein structure of pBI143 MobA, we *de novo* predicted the monomer and dimer structures using AlphaFold 2 (AF) in ColabFold with default settings.[84] AlphaFold 2 confidently predicted the structure of the catalytic domain but had low pLDDT scores for the coil domains and the dimer interactions. However, we explored variants across the whole dimer complex. Next, we integrated the pBI143 MobA AF structure into *anvi'o structure* by running 'anvi-gen-structure-database'[83]. After that, we summarized SNV data as SAAVs from the metagenomic read recruitment data using 'anvi-gen-variability-profile' with the '–engine AA' flag to recover a variability profile (https://anvio.org/m/variability-profile). Subsequently, we superimposed the SAAV data variability profile on the structure with 'anvi-display-structure' which filtered for variants that had at least 0.05 departure from consensus (reducing our metagenomic samples size from 2221 to 1706). Finally, we analyzed SAAVs that were prevalent in at least 5% of remaining samples. This left us with 21 SAAVs to analyze on the monomer. Next, we explored the relationship between SAAVs, relative solvent accessibility (RSA), and ligand binding residues in pBI143 MobA. To do this, we identified the homologous structure PDB 4LVI (MobM) by searching the high pLDDT pBI143 AF domain against the structure database PDB100 2201222 using Foldseek (https://search.foldseek.com/search). We next structurally aligned the pBI143 MobA AF structure to PDB 4LVI (MobM)[138] using PyMol.[119] We chose the MobM structure 4LVI rather than a MobA because it had more structural and sequence homology to the pBI143 MobA catalytic domain AF structure than any PDB MobA structures. Additionally, we leveraged residue conservation values from the pre-calculated 4LVI ConSurf analysis to further explore ligand binding residues.[113,139]

Phylogenetic tree construction

To construct the pBI143 phylogeny, we identified pBI143 contigs from the assemblies of bacterial isolates (Table S1) using BLAST.[120] We ran 'anvi-gen-contigs-database' on each pBI143 contig followed by 'anvi-export-gene-calls' with the flag '–gene-caller prodigal' and concatenated the resulting amino acid sequences. For the bacterial host phylogeny, we ran 'anvi-gen-contigs-database' on each assembled genome, then extracted ribosomal genes (Ribosomal_L1, Ribosomal_L13, Ribosomal_L14, Ribosomal_L16, Ribosomal_L17, Ribosomal_L18p, Ribosomal_L19, Ribosomal_L2, Ribosomal_L20, Ribosomal_L21p, Ribosomal_L22, Ribosomal_L23, Ribosomal_L27, Ribosomal_L27A, Ribosomal_L28, Ribosomal_L29, Ribosomal_L3, Ribosomal_L32p, Ribosomal_L35p, Ribosomal_L4, Ribosomal_L5, Ribosomal_L6, Ribosomal_L9_C, Ribosomal_S10, Ribosomal_S11, Ribosomal_S13, Ribosomal_S15, Ribosomal_S16, Ribosomal_S17, Ribosomal_S19, Ribosomal_S2, Ribosomal_S20p, Ribosomal_S3_C, Ribosomal_S6, Ribosomal_S7, Ribosomal_S8, Ribosomal_S9, ribosomal_L24) using the command 'anvi-get-sequences-for-hmm-hits' with the flags '–return-best-hit', '–get-aa-sequences', '–concatenate' and '–min-num-bins-gene-occurs 82' and '–hmm-source Bacteria_71'. For both phylogenies, we aligned the genes with MUSCLE v3.8.1551,[121] trimmed the alignments with trimAl[122] using the flag '-gt 0.5', and computed the phylogeny with IQ-TREE 2.2.0-beta using the flags '-m MFP' and '-bb 1000'[123]. We visualized the trees with 'anvi-interactive' in '–manual-mode', and used the metadata provided by the Duchossois Family Institute to label the isolates to their corresponding donors. We used the 'geom_alluvium' function in ggplot2 to make the alluvial plots.

Mother-infant single nucleotide variant network

To investigate whether single-nucleotide variants (SNVs) suggest a vertical transmission of pBI143, we used metagenomic read recruitment results from four independent study that generated metagenomic sequencing of fecal samples collected from mothers and their infants in Finland,[61] Italy,[64] Sweden,[72] and the USA,[73] against the pBI143 Version 1 reference sequence. The primary input for this investigation was the anvi'o variability data, which is calculated by the anvi'o program 'anvi-profile', and reported by 'anvi-gen-variability-profile' (with the flag '–engine NT'). The program 'anvi-gen-variability-profile' (https://anvio.org/m/anvi-gen-variability-profile) offers a comprehensive description of the single-nucleotide variants in metagenomes for downstream analyses.

Since the *mobA* gene was conserved enough to represent all three versions of pBI143, for downstream analyses we limited the context to study variants to the mobA gene. The total number of samples in the entire dataset with at least one variable nucleotide position was 309, which represented a total of 102 families (Sweden: 52, USA: 24, Finland: 14, Italy: 11). We removed any sample that did not belong to a minimal complete family (i.e., at least one sample for the mother, and at least one sample of her infant), which reduced the number of families in which both members are represented to 57 families (Sweden: 36, USA: 16, Finland: 3, Italy: 2). We further removed families if the coverage of the *mobA* gene was not 50X or more in at least one mother and one infant sample in the family, which reduced the number of families with both members represented and with a reliable coverage of mobA to 49 families (Sweden: 33, USA: 13, Finland: 2, Italy: 1), and from a given family, we only used the samples that had at least 50X for downstream analyses. We subsampled the variability data in R to only include the variable nucleotide position data for the final list of samples. We then used the list of single-nucleotide variants reported in this file to generate a network description of these data using the program 'anvi-gen-variability-network', which reports an 'edge' between any sample pairs that share a SNV with the same competing nucleotides. We then used Gephi,[124] an open-source network visualization program, with the ForceAtlas2 algorithm[125] to visualize the network. To quantify the extent of similarity between family members based on single-nucleotide patterns in the data, we generated a distance matrix from the same dataset using the 'pdist' function in Python's standard library with 'cosine' distances. We calculated the average distance of each sample to all other samples in its familial group ('within distance'), as well as the average distance from each sample to all other samples not present in their familial group ('between distance'). We subtracted the within distance from the between distance to get the 'subtracted distance'.

Metagenomic taxonomy estimation

We used Kraken 2.0.8-beta with the flags '–output', '–report', '–use-mpa-style', '–quick', '–use-names', '–paired' and '–classified-out' to estimate taxonomic composition of each metagenome.[126] For the genus-level taxonomic data, we filtered for metagenomes where the total number of reads recruited to a *Bacteroides*, *Parabacteroides* or *Phocaeicola* genome was >1000 and the mean coverage of pBI143 was >20X. For the species-level taxonomic data, we used a cutoff of >0.1% percent of reads recruited to designate presence or absence of *B. fragilis* and >0.0001% for pBI143 based on the sizes of the genomes respectively (the *B. fragilis* genome is 3 orders of magnitude larger than pBI143).

Isogenic strain construction

We constructed the plasmid vector pEF108 (Figure S6A) by PCR amplifying the desired sections with primers vec_108F, vec_108R, frag1_108F, frag1_108R, frag2_108R and frag2_108R (Table S3) from existing plasmids. We assembled the three fragments via Gibson assembly using standard conditions described for NEB Gibson assembly mastermix (https://www.neb.com/protocols/2012/12/11/gibson-assembly-protocol-e5510). See pEF108 plasmid map below. We transformed the construct into *E. coli* S17 λ*pir* via electroporation with a BioRad micropulser using 0.1cm cuvettes and selected on LB-carbenicillin 100ug/mL agar plates. We conjugally transferred pEF108 from *E. coli* S17 λ*pir* into *B. fragilis* 638R or 9343.[140] Briefly, we grew the donor and recipient strains in LB-carbenicillin 100ug/mL broth and vitamin K supplemented brain-heart infusion media (BHIS) broth respectively for 12-15 hours. We spun down the cultures and resuspended in BHIS and combined at a 1:5 ratio of donor to recipient. We spotted the donor and recipient mixture onto BHIS plates and incubated for 12 hours aerobically. We scraped the cells off the plate, resuspended in BHIS, then plated on BHIS + erythromycin 25ug/mL. We restreaked the colonies and validated the presence of the construct via PCR and sanger sequencing. Next, we wanted to select for cells where a recombination event had removed the vector containing pheS, ampicillin and erythromycin resistance and left pBI143 in its native form. Colonies with the full sized pEF108 construct were grown in Bacteroides minimal media (BMM) with 10mM p-chlorophenylalanine (PCPA) broth for 24 hours, and plated onto BMM + 10mM PCPA. PCPA prevents the growth of cells that are expressing the pheS gene. We used PCR and WGS to screen for pBI143-positive, *pheS*-negative colonies that grew on BMM + 10mM PCPA and used these isolates for downstream experiments.

Mouse competitive colonization assays

All animal experimentation was approved by the Institutional Animal Care and Use Committee at the University of Chicago. We gavaged three male and three female 10-15 week old germ-free C57BL/6J mice with a 1:1 inoculum of *B. fragilis* 638R:*B. fragilis* 638R pBI143. Males and females were housed separately in isocages and remained gnotobiotic for the duration of the experiment. We collected fecal pellets after 14, 28, and 40 days, diluted and plated on BHIS plates. One mouse was lost during fecal collection. We performed PCR on 48 colonies per mouse using a mixture of four primers (Table S3), one set that amplifies a 1248-bp region of the 638R chromosome and a second set that amplifies a 662-bp segment of pBI143. All colonies produced PCR amplicons for the 1248-bp region of the 638R chromosome and a subset also contained the amplicon for pBI143, allowing calculation of the ratio over time. The exact starting ratio for gavage was also calculated using this same PCR.

PlasX prediction of pBI143 additional gene origin

To determine if the additional genes acquired by pBI143 are of plasmid origin, we ran the plasmid prediction software PlasX[43] (https://github.com/michaelkyu/PlasX). To identify genes and annotate COGs and Pfams we used the anvi'o programs 'anvi-gen-contigs-database', 'anvi-export-gene-calls', 'anvi-run-ncbi-cogs', 'anvi-run-pfams, and 'anvi-export-functions'. To annotate *de novo* gene

families, we ran 'plasx search_de_novo_families'. With the outputs from anvi'o and the de novo gene families file, we used PlasX to classify the sequences as plasmid or non-plasmid sequences using the command 'plasx predict'.

'Approximate copy number ratio' calculation

The first challenge to use metagenomic coverage values to study pBI143 copy number trends in human gut metagenomes is the unambiguous identification of gut metagenomes that appear to have a single possible pBI143 bacterial host beyond reasonable doubt. To establish insights into the taxonomic make up of the gut metagenomes we previously assembled, we first ran the program 'anvi-estimate-scg-taxonomy' (https://anvio.org/m/anvi-estimate-scg-taxonomy) with the flags '–metagenome-mode' (to profile every single single-copy core gene (SCG) independently) and '–compute-scg-coverages' (to compute coverages of each SCG from the read recruitment results). We also used the flag '–scg-name-for-metagenome-mode' to limit the search space for a single ribosomal protein. We used the following list of ribosomal proteins for this step as they are included among the SCGs anvi'o assigns taxonomy using GTDB, and we merged resulting output files: Ribosomal_S2, Ribosomal_S3_C, Ribosomal_S6, Ribosomal_S7, Ribosomal_S8, Ribosomal_S9, Ribosomal_S11, Ribosomal_S20p, Ribosomal_L1, Ribosomal_L2, Ribosomal_L3, Ribosomal_L4, Ribosomal_L6, Ribosomal_L9_C, Ribosomal_L13, Ribosomal_L16, Ribosomal_L17, Ribosomal_L20, Ribosomal_L21p, Ribosomal_L22,ribosomal_L24, and Ribosomal_L27A. For our downstream analyses that relied upon the merged SCG taxonomy and coverage output reported by anvi'o, we considered *Bacteroides*, *Parabacteroides* and *Phocaeicola* as the genera for candidate pBI143 host 'species', and only considered metagenomes in which a single species from these genera was present. Our determination of whether or not a single species of these genera was present in a given metagenome relied on the coverage of species-specific single-copy core genes (SCGs), where the taxonomic assignment to a given SCG resolved all the way down to the level of species unambiguously. We excluded any metagenome from further consideration if three or more candidate host species had positive coverage in any SCG in a metagenome. Due to highly conserved nature of ribosomal proteins and bioinformatics artifacts, it is possible that even when a single species is present in a metagenome, one of its ribosomal proteins may match to a different species in the same genus given the limited representation of genomes in public databases compared to the diversity of environmental populations. So, to minimize the removal of metagenomes from our analysis, we took extra caution with metagenomes before discarding them if only two candidate host species had positive coverage in any SCG. We kept such a metagenome in our downstream analyses only if one species was detected with only a single SCG, and the other one was detected by at least 8. In this case we assumed the large representation of one species (with 8 or more ribosomal genes) suggests the presence of this organism in this habitat confidently, and assumed the single hit to another species within the same genus was likely due to bioinformatics artifacts. It is the most unambiguous case if only a single candidate host species was detected in a given metagenome, but we still removed a given metagenome from further consideration if that single species had 3 or fewer SCGs in the metagenome. These criteria deemed 584 of 2580 metagenomes to have an unambiguous pBI143 host that resolved to 21 distinct species names. We further removed from our modeling the metagenomes where the candidate host species did not occur in any other metagenome, which removed 5 of these candidate host species from further consideration. Finally, we further removed any metagenome in which the pBI143 coverage was less than 5X. Our final dataset to calculate the ''approximate copy number ratio'' (ACNR) of pBI143 in metagenomes through coverage ratios contained 579 metagenomes with one of 16 unambiguous pBI143 hosts. We calculated the ACNR by dividing the observed coverage of pBI143 by the empirical mean coverage of the host by averaging the coverage of all host SCGs found in the metagenome. To estimate the multiplicative difference in the geometric mean ACNR, we fit a linear model for the expected value of the logarithm of the ACRN, with disease status and bacterial host as predictors using rigr to construct the interval and estimate.[127]

Oxidative stress experiments

We grew *B. fragilis* in 5 mL BHIS for 15 hours in an anaerobic chamber. We inoculated 750 μL of this culture into 30 mL BHIS in quintuplicate, and grew them for 3 hours. We divided the 30 mL into a further 5 culture flasks of 5 mL BHIS, and exposed each to oxygen with constant shaking for the appropriate time before returning the flask to the anaerobic chamber. At each time point, we took an aliquot of culture to determine the copy number of pBI143 in that sample. We extracted DNA from the cultures using a Thermal NaOH preparation[141] to prepare them for qPCR. Copy number calculated can be found in Table S7.

pBI143 copy number qPCR

To evaluate plasmid copy number (CN), we developed a real-time TaqMan probe multiplex PCR assay to amplify both pBI143 and a single-copy *B. fragilis*-specific genomic reference gene (referred to as hsp [heat shock protein]) in the same reaction.

Primer design for hsp and pBI143

We aligned the canonical pBI143 plasmid DNA sequence from GenBank, whole genome assemblies and metagenome-assembled genomes (MAGs) as outlined in Table S3. The two known pBI143 genes, *rep* and *mob*, are common plasmid features across the bacterial kingdom (DelSolar et al.[4]; Wawrzyniak et al.[142]) and use of either gene alone had high potential for cross-amplification from other mobile genetic elements. To ensure pBI143 specificity, we designed our primer set so that the forward primer was located within the 3' region of the *rep* gene (Table S3) while the reverse primer was located in the intergenic region (Table S3). This required that two conditions would have to be met for amplification to occur: (1) presence of the gene of interest and (2) homology to the pBI143 plasmid backbone. Despite the existence of plasmid variants differing across the rep gene, the 3' region used in the forward primer

design is conserved across the source sequences. The 38-yr old canonical pBI143 sequence (U30316.1) demonstrated greater sequence variation in intergenic regions than more contemporary sequences as determined by existing publicly-available metagenomic data. In designing the reverse primer, we strategically excluded U30316.1 in favor of using the more recent pBI143 sequences. The FAM-labeled hydrolysis probe was designed within a conserved plasmid feature, the 56-bp inverted repeat (IR) region and in concert with the designed primers, amplified/detected a 145-bp product (Table S3). The choice to use *rep*, over *mob*, as our target was based on (1) its conservancy in the 3' gene region and (2) technical difficulties in optimizing a *mob*-based assay. To perform relative quantification experiments and to normalize bacterial cell numbers between samples for the purposes of determining the copy number of pBI143 per genome equivalent required identifying a suitable genomic reference gene. A key prerequisite was identifying a single copy gene present in the genus *Bacteroides*, but absent in other common gastrointestinal (GI) tract organisms. We employed a pangenomic analysis of 12 *Bacteroides* and 15 other human commensal gut microbe genomes to determine potential candidates and used the program 'anvi-run-workflow' with '–workflow pangenomics'. Anvi'o pangenomics workflow is detailed elsewhere (Delmont and Eren[143]). Briefly, the pangenomic analysis used the NCBI's BLAST[120] to quantify similarity between each pair of genes, and the Markov Cluster algorithm (MCL) (Enright et al.[144]) (with inflation parameter of 2) to resolve clusters of homologous genes. The program 'anvi-summarize' created summary tables for pangenomes and 'anvi-display-pan' provided interactive visualizations of pangenomes. Using the criteria of (1) maximum functional homogeneity of 0.99 and (2) maximum geometric homogeneity of 0.99, we identified 35 gene clusters for further interrogation. The corresponding DNA sequences were gathered using the program 'anvi-get-sequences-for-gene-cluster' and aligned using Kalign[128] (https://www.ebi.ac.uk/Tools/msa/kalign) for multiple sequences.

Despite our initial desire to identify a target that could serve as a reference gene across all *Bacteroides* spp., we found that within these 35 gene clusters, the percent sequence identity dropped from >98% in *B. fragilis* to ~50-85% in non-*B. fragilis* sequences. Therefore, we focused on finding a *B. fragilis*-specific target by further requiring 100% coverage and 99.8% - 100% percent identity across all *B. fragilis* genomes. Five gene clusters qualified; a single gene cluster demonstrated 100% identity across all seven *B. fragilis* genomes used in the pangenome. Using this gene clusters' 177-bp nucleotide sequence, we performed BLASTN[145] on the NCBI Reference Sequence (RefSeq) Database (release 99, 3/2/2020), using Megablast (optimize for highly similar sequences) to conduct a systematic and thorough *in-silico* assessment of *B. fragilis* specificity. A list of the 17,785 complete genomes was downloaded from RefSeq (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/, accessed 3/31/2020) and found to contain 39 Bacteroides genomes; of which, 15 were cataloged as *B. fragilis*. 6 organisms labeled as *B. fragilis* in this collection appeared to be a different species, given their overall ANI to B. fragilis genomes was only 84-85% and we disregarded these organisms. The only significant BLAST alignments of the hsp gene were to the nine true *B. fragilis* genomes. These genomes annotated the gene cluster protein product as hypothetical or conserved hypothetical protein (n=2); DUF4250 domain-containing protein (n=5); or heat shock protein (n=2) (Table S3).

Based on its specificity to *B. fragilis* only, the previous use of heat shock proteins to discriminate amongst anaerobes,[146] and the documented conservancy of these molecules,[147] this gene cluster was chosen as our candidate reference gene and is hereafter referred to as hsp. The primers and Cy5-labeled hydrolysis probe were designed to amplify/detect a 101-bp product (Table S3).

qPCR analytical specificity

We assessed the *in vitro* analytical specificity of the hsp qPCR assay using DNA templates extracted from a collection of 41 bacterial isolates (13 aerobes, 28 anaerobes; representing 16 commonly encountered commensal gastrointestinal tract genera). *hsp* was not detected in any aerobic or anaerobic microorganisms, except for the collections' four *B. fragilis* isolates. The lack of amplification in other *Bacteroides spp.*, including *B. ovatus* (n=3), *B. thetaiotamicron* (n=2), *B. uniformis* (n=1) and *B. vulgatus* (n=3) corroborated the previous *in silico* results.

qPCR experimental conditions

We performed real-time PCR amplification on a LightCycler 480 II system (Roche Diagnostics), using 10 microliter reactions consisting of 2X PrimeTime Gene Expression master mix (Integrated DNA Technologies, Coralville, IA), 0.8 μM pBI143_R, and 0.4 μM of pBI143_F, B.fragilis_hsp_F, and B.fragilis_hsp_R primers. We used optimized probe concentrations of 0.2 μM HSP and 0.4 μM pBI143 probe. Probe and primer sequences are outlined in Table S3. We assessed triplicate PCR reactions using genomic DNA templates (2-l volume per reaction) and the optimal cycling conditions of an initial denaturation step of 95°C for 3-min, followed by 40 cycles of 95°C for 15-s (denaturation) and 60°C for 60-s (annealing and extension).

qPCR assay performance characteristics

We constructed a single plasmid, by standard recombinant DNA methods, containing both the entire pBI143 plasmid and the reference gene (hsp) DNA and then transformed the plasmid into *E. coli* EC100D. The DNA concentration of the recombinant plasmid was converted to the number of template copies using the mass of the plasmid molecule.[148] Using a 10-fold serial dilution series of the plasmid DNA standard (ranging from 3×10^0 to 3×10^6 copies/reaction), we constructed standard curves for both chromosomal reference gene and the target plasmid.

Each targets' lower limit of detection (LOD) was determined to be 30 copies per reaction, as defined by the first dilution that detects 95% of positive samples.[149] We validated a linear dynamic range of six orders of magnitude for each target, and this range was then used in further assay performance metric calculations.

The primer amplification efficiencies were determined by standard procedure[149] that includes (1) making a log10 dilution series of target DNA, (2) calculating a linear regression based on the targets' mean Cq data points and (3) inferring the efficiency from the slope of the line. Over 11 experiments, mean Cq values were derived and PCR efficiencies were calculated as 97.8% and 98.98% for pBI143 and hsp, respectively. We demonstrate less than a <=5.2% difference when comparing same run target and reference gene efficiency, demonstrating the two genes amplify similarly.

qPCR analysis of animal, untreated sewage and water samples

Samples were tested with the pBI143 assay and two established assays for human fecal markers that included HF183 and Lachno3.[150] Standard curves were generated based on a minimum of 16 runs (in triplicate) and consisted of linearized plasmids containing the HF183, Lachno3, and pBI143 target sequences. The plasmids used for the standard curves were purified using a Qiagen mini plasmid prep kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. Standard curves were run with DNA serially diluted from 1.5 x 106 to 1.5 x 101 copies/reaction with resulting linear equations and efficiencies as follows:

HF183: Slope: -3.37, Y-intercept 39.363, R2 0.998, Eff% 98.18

Lachno3: Slope: -3.42, Y-intercept 38.13, R2 0.999, Eff% 95.92

pBI143: Slope: -3.43, Y-intercept 39.363, R2 0.999, Eff% 95.90

For each run, two of the standard concentrations (as quality assurance for the standard curve, sterile water (as negative control) and each sample was run in duplicate in a final volume of 25 μL with a final concentration of 1μM for each primer, 80 nM for the probe, 5 μL of sample DNA, and 12.5 μL of 2X Taqman® Gene Expression Master Mix Kit (Applied Biosystems; Foster City, CA). DNA template was added as undiluted sample for surface water and animal samples, and 1:100 dilutions of sewage samples. Amplification conditions consisted of the following cycles: 1 cycle at 50° C for 2 minutes to activate the uracil-N-glycosylase (UNG); 1 cycle at 95° C for 10 minutes to inactivate the UNG and activate the Taq polymerase; 40 cycles of 95° C for 15 seconds; and 1 minute at 60o C for HF183 or 1 minute at 64o for Lachno3 using a StepOne Plus™ instrument (Applied Biosystems; Foster City, CA).

Water samples that amplify after 35 cycles were considered below the standard curve limit of 15 CN/reaction and were therefore considered below limit of quantification. For water samples where 400 ml was filter for extraction, this value is 113 CN/100 ml. All no template controls (water) showed no amplification.

For screening of animal samples to assess the presence of this plasmid in non-human gut microbiomes, archived DNA from a previous study[151] was analyzed and included 14 different animals encompassing 81 individual fecal samples. For assessment of fecal contamination of surface waters, archived DNA from 40 samples of river water[152–154] and freshwater beaches[155] were analyzed. These water samples were chosen from these previous studies that represented a range of contamination based on HF183 and Lachno3 levels. A total of 20 archived untreated sewage samples as reported in Olds et al.[150] were also analyzed for comparison. Since we were using archived samples from previous studies, we retested all the samples for the two human markers to account for any degradation.

Visualizations

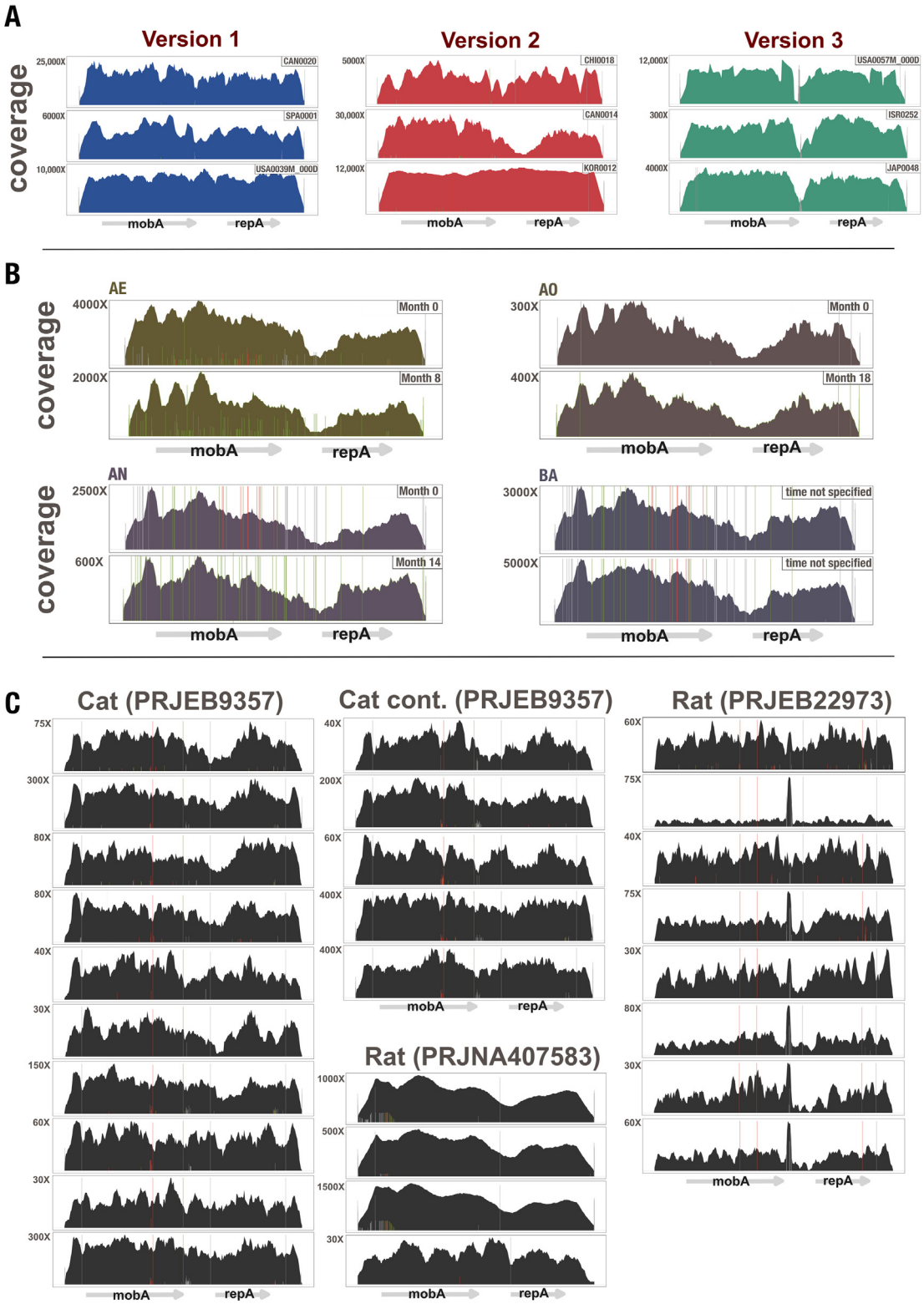
We used ggplot2[129] to generate all box and scatter plots. We generated coverage plots using anvi'o, with the program 'anvi-script-visualize-split-coverages'. We finalized the figures for publication using Inkscape, an open-source vector graphics editor (available from http://inkscape.org/).

QUANTIFICATION AND STATISTICAL ANALYSIS

The section "Approximate copy number ratio calculation in metagenomes" above describes the details for the regression model behind ACNR. The function 'regress' from the R package 'rigr' was used with 'fnctl = mean' and 'robustSE=TRUE' to estimate effect sizes and calculate model-robust confidence intervals. A robust Wald test was used to test the null hypothesis of no change in the true log-mean ACNR (Figure 7B, upper right). n = 579 metagenome-host pairs were used to fit the regression model, which had 17 parameters (16 host and 1 disease coefficients). The use of model-robust testing and confidence intervals, as well as the large sample size, alleviated any need to investigate heteroskedasticity and normality for valid error rate control.

CellPress
OPEN ACCESS

Supplemental figures



(legend on next page)

Figure S1. Representative coverage plots of metagenomes mapped to pBI143, related to Figure 1

Each coverage plot shows the read recruitment results for an individual metagenome to pBI143. 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.

(A) Global human gut metagenomes. Versions 1 (blue), 2 (red), and 3 (green). 3 coverage plots for each reference version of pBI143 are shown, the remaining 13,539 can be generated from the anvi'o databases at https://merenlab.org/data/pBI143.

(B) pBI143 coverage plots from individuals sampled across time. Each coverage plot shows the coverage of pBI143 first and last sample collected for each subject possessing pBI143 this dataset.[156]

(C) The same pBI143 population is found within individual cat and rat cohorts from 3 separate studies.

See also Table S1.

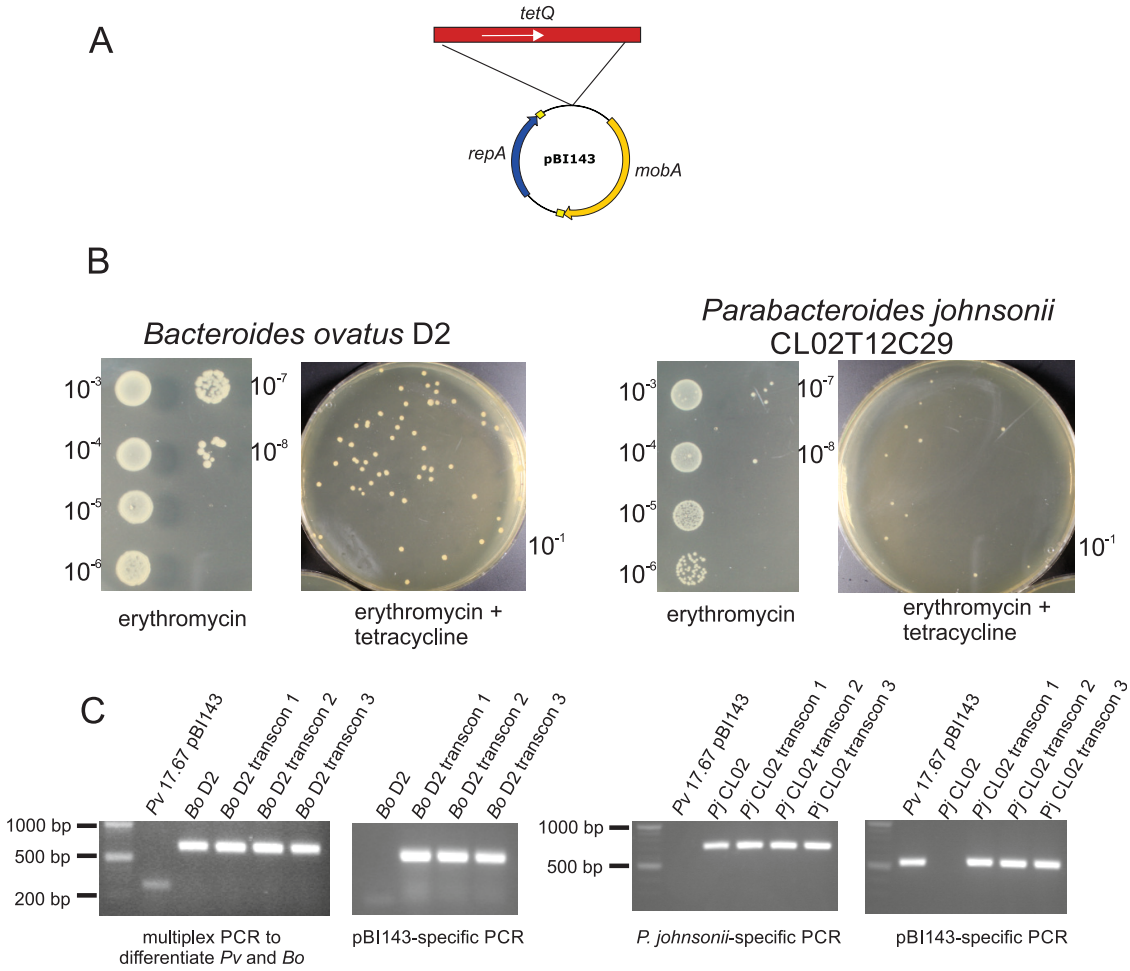


Figure S2. pBI143 transfers to other Bacteroidales species, related to Figure 2

(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.

See also Tables S1 and S2.

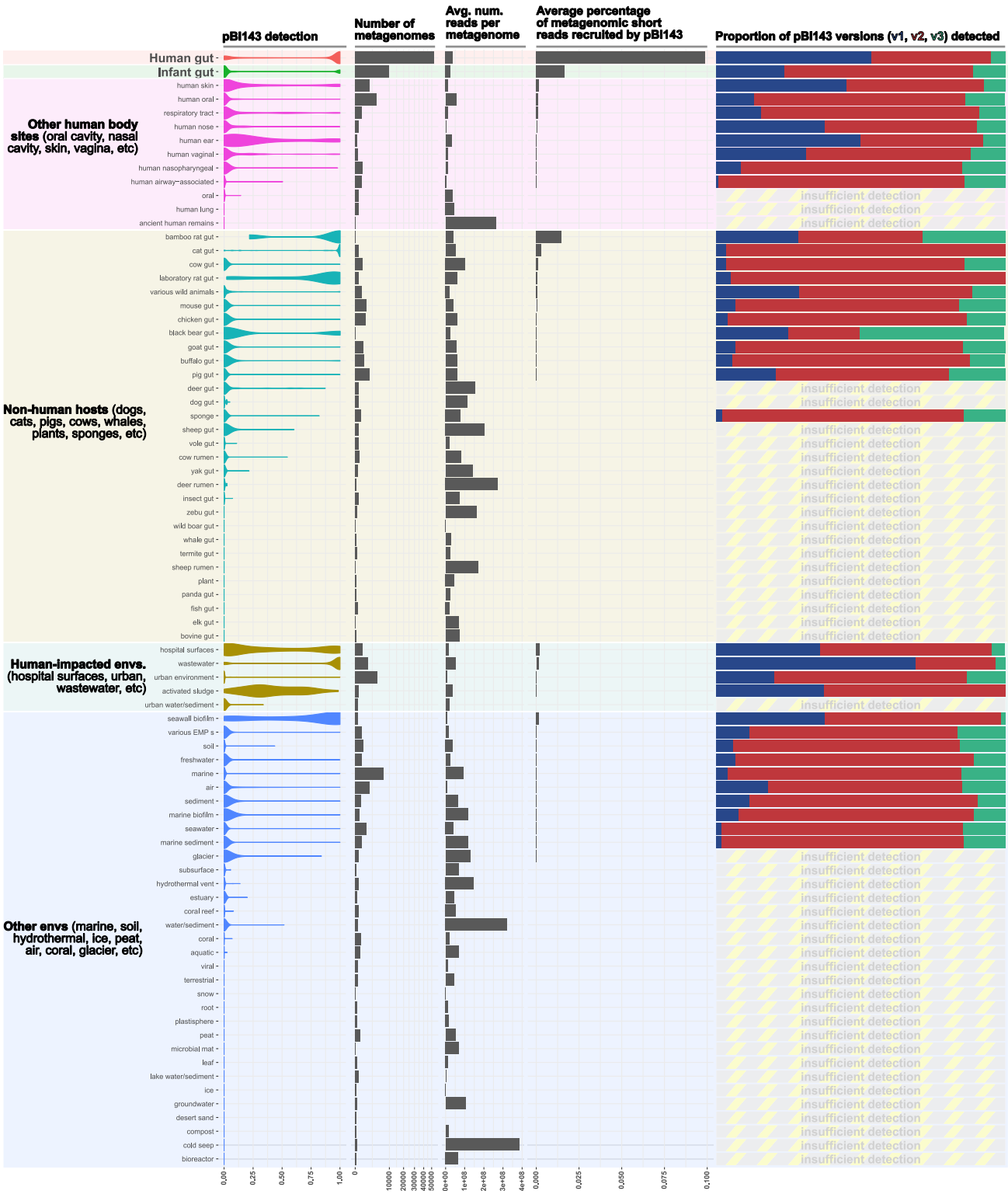


Figure S3. Presence or absence of pBI143 across 100,000 metagenomes, separated by environment, related to Figure 2

See also Table S1.

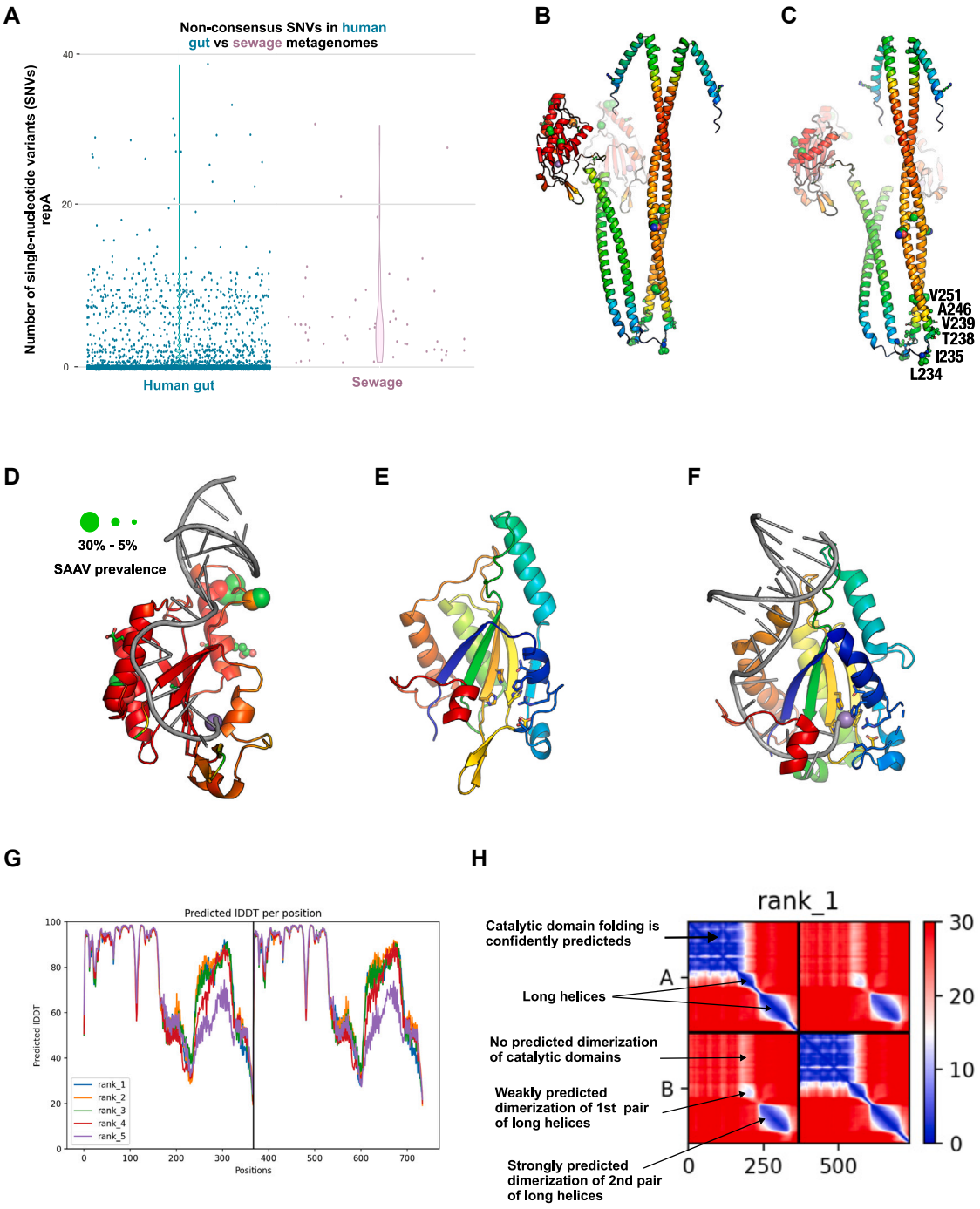


Figure S4. pBI143 SNV and SAAV distribution across globally distributed genomes and metagenomes, related to Figure 3

(A) Non-consensus SNVs present in 4,516 human gut metagenomes and 68 sewage metagenomes.

(B and C) 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 indicates 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 Mn^{2+} from 4lvi.pdb; https://doi.org/10.1073/pnas.1702971114).

(D) 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 are the same as in (A) and (B).

(E) 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.

(legend continued on next page)

(F) MobM from pMV158 bound to oriT DNA (gray) and a catalytic Mn^{2+} ion (purple) (PDB: 4lvi [82]) shown shaded from blue to red, and active site residues are shown as sticks.
(G) AlphaFold 2 pLDDT score representing structural prediction accuracy of MobA.
(H) AlphaFold 2 predicted aligned error plot (PAE) for MobA dimer prediction.
See also Table S3.

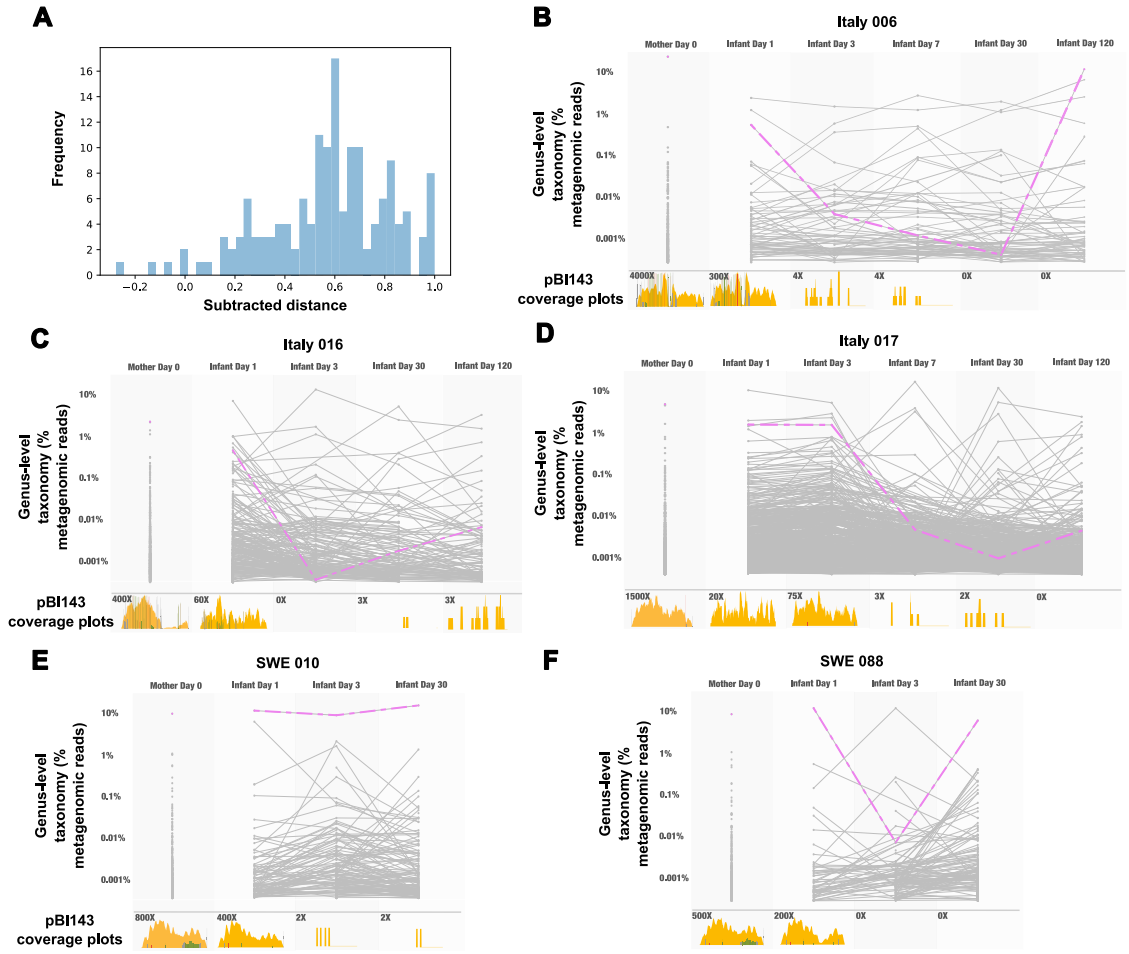


Figure S5. Mother-infant network quantification and kraken wilt data, related to Figure 5

(A) 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 STAR Methods for more details.

(B–F) Genus-level taxonomy of mother-infant gut metagenomes for pairs with a pBI143 wilt phenotype. Line plots show the abundance of individual genera as estimated by kraken2. Bacteroides is highlighted in pink. Yellow coverage plots below show the coverage of pBI143 at corresponding time points.

See also Table S5.

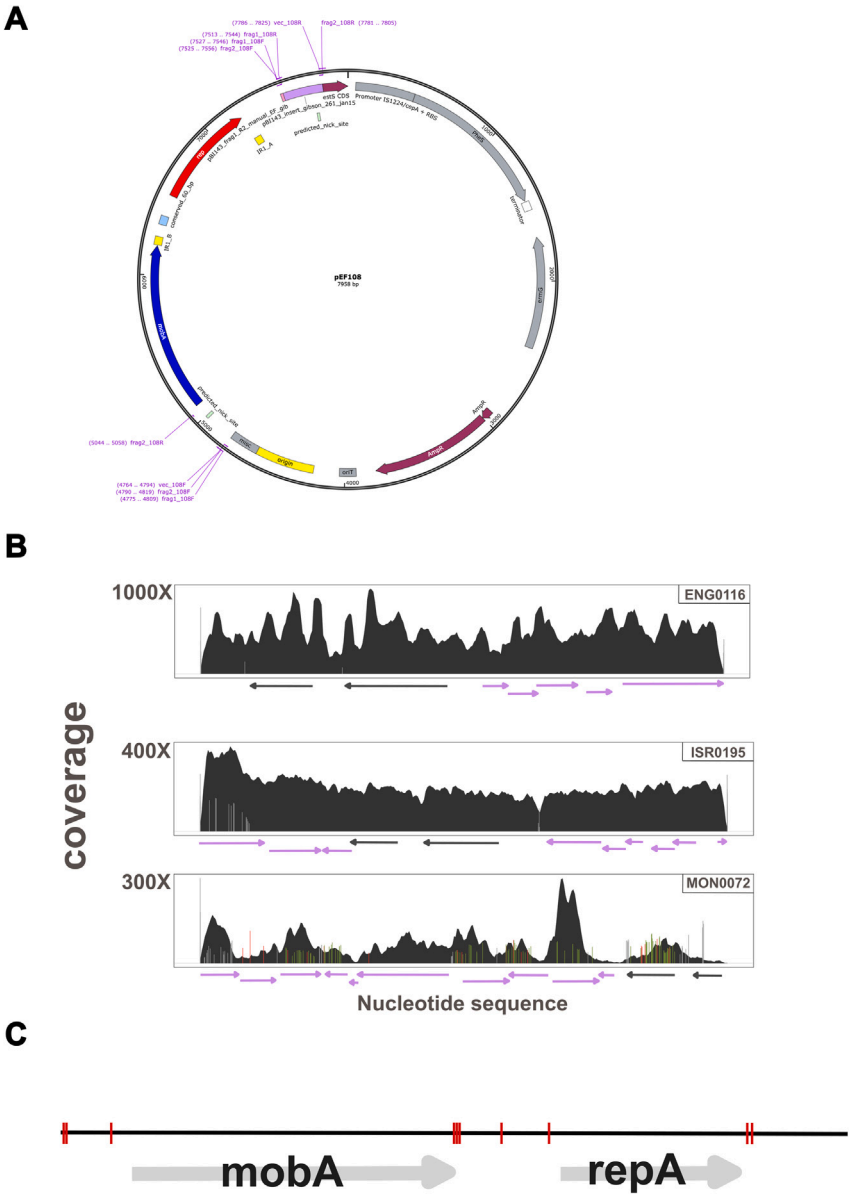


Figure S6. Additional genetic material in pBI143, related to Figure 6

(A) Plasmid map of pEF108 construct used to create isogenic strain set of cells ± pBI143 for mouse competition experiments (for details on construct recombination to form naive pBI143, see method details in STAR Methods).

(B) Coverage plots of naturally occurring pBI143 containing additional genetic material assembled from metagenomes and confirmed to be present in isolate genomes. Data comes from the metagenomes as labeled. pBI143 is dark gray, additional genes are in pink as in Figure 6.

(C) Regions of additional gene acquisition on pBI143.

See also Tables S2 and S6.