# Megaplasmids associate with *Escherichia coli* and other *Enterobacteriaceae*

Allison K. Guitor[1,2], Shuai Wang[2], Owen T. Tuck[2,3], Brian Firek[4], Nadja Mostacci[5], April Juahal[6,7], Lin-Xing Chen[8], Agata H. Dziegiel[9,10], Stephen Baker[11], Vu Thuy Duong[12], Alison E. Mather[9,10,13], Jukka Corander[14], Anu Kantele[15], Liat Shenhav[6,7,16], Markus Hilty[5], Michael J. Morowitz[17], Rohan Sachdeva[2], Jillian F. Banfield[1,2,18,19*]

*Corresponding author: jbanfield@berkeley.edu

[1]Department of Earth and Planetary Sciences, University of California, Berkeley, CA, USA
[2]Innovative Genomics Institute, University of California, Berkeley, CA, USA
[3]Department of Chemistry, University of California Berkeley, Berkeley, CA, USA
[4]Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[5]Institute for Infectious Diseases, Faculty of Medicine, University of Bern, Bern, Switzerland
[6]Institute for Systems Genetics, New York Grossman School of Medicine, New York University, New York, NY, USA
[7]Department of Microbiology, New York Grossman School of Medicine, New York University, New York, NY, USA
[8]State Key Laboratory of Advanced Environmental Technology, the Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei, China, 230026
[9]Microbes and Food Safety, Quadram Institute Bioscience, Norwich, UK
[10]Centre for Microbial Interactions, Norwich Research Park, Norwich, UK
[11]A*STAR Infectious Diseases Labs (A*STAR IDL), Agency for Science, Technology and Research (A*STAR), Singapore 138648, Singapore
[12]Children's Hospital No. 1, Ho Chi Minh City, Vietnam.
[13]University of East Anglia, Norwich, UK.
[14]Department of Mathematics and Statistics University of Helsinki, Helsinki, Finland; Department of Biostatistics, University of Oslo, Oslo, Norway; Wellcome Sanger Institute, Hinxton, UK; Department of Genetics, University of Cambridge, Cambridge, UK.
[15]Meilahti Vaccine Research Center MeVac, Helsinki University Hospital and University of Helsinki, Helsinki, Finland
[16]Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

[17]Division of Pediatric General and Thoracic Surgery, University of Pittsburgh Children's Hospital, Pittsburgh, PA, USA
[18]Department of Environmental Science Policy, and Management, University of California, Berkeley, CA, USA
[19]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
Biomedicine Discovery Institute, Monash University, Australia

**Summary**

Humans and animals are ubiquitously colonized by *Enterobacteriaceae*, a bacterial family that contains both commensals and clinically significant pathogens. Here, we report *Enterobacteriaceae* megaplasmids of up to 1.58 Mbp in length in infant and adult guts, and other microbiomes. Of the 19 complete plasmid genomes, one was reconstructed from an *Escherichia coli* isolate; others were linked to species of *Citrobacter* and *Enterobacter* via analysis of genome modification patterns. The detection of related plasmids in different *Enterobacteriaceae*, conjugation machinery, and more diverse modified motifs in certain plasmids compared to hosts suggests that these elements are self-transmissible, with a broad host range. The plasmids encode multi-drug efflux systems and potential secreted effectors. Up to 208 tRNAs are encoded, and include sequence variants that may counter tRNA-centric defense mechanisms. Overall, the vast megaplasmid coding capacity may broaden host range, increase competitiveness, control invasion by other elements, and counter programmed cell death.

**Keywords**: Plasmids, megaplasmids, *Enterobacteriaceae*, human gut microbiome, *Escherichia coli*, antibiotic resistance, secretion systems, methylation, preterm infants, *Salmonella enterica*, virulence

58 **Introduction**

59 Plasmids are extrachromosomal elements that generally do not encapsulate and often can move
60 between hosts. They are interesting from the perspective of lateral gene transfer, including
61 medically relevant capacities such as antibiotic resistance and pathogenicity[1]. They may also
62 encode host-relevant genes and hold high technological significance, for example as genetic
63 tools[2]. More broadly, they are integral and often overlooked components of microbial communities
64 and evolution[3,4].

65 The sizes of plasmid genomes can range from a few thousand base pairs to well over a million
66 base pairs (i.e., megaplasmids - generally defined as plasmids >350 kb)[5]. Larger plasmids tend
67 to have low copy numbers (PCNs), often a 1:1 ratio with the chromosome, and more metabolic
68 capacities as compared to their smaller higher copy-number counterparts[6–8]. Chromids and
69 secondary chromosomes are large replicons with essential core genes. Secondary replicons in
70 *Burkholderiales* with genomes of 2.1 - 3.36 Mbp were initially proposed as megaplasmids;
71 however, an alternative classification as chromids or secondary chromosomes is now preferred[9–
72 12]. Chromids, common in species of *Burkholderia*, *Brucella*, *Agrobacterium*, and *Vibrio*, tend to
73 have plasmid-like replication and partitioning systems indicative of their origins from
74 megaplasmids[5,13–16]. Secondary chromosomes are believed to have evolved from fragments of
75 ancestral chromosomes[5,15]. Chromids and secondary chromosomes are often distinguished from
76 megaplasmids by their essentiality and compositional signatures (e.g., GC content, codon usage)
77 shared with the host's primary chromosome[5,14,15].

78 The best known megaplasmids are from *Rhizobium* (*Alphaproteobacteria*), with genomes up to
79 2.83 Mbp[17], and 1.8 Mb plasmids of *Streptomyces clavuligerus* (*Actinomycetes*) with linear
80 genomes[18]. In groundwater samples, the largest plasmids reported were circular and 2.96 and
81 1.74 Mb in length[19]. Surveys of plasmids from NCBI RefSeq and plasmid-specific databases
82 representing diverse environments suggest a generally low prevalence of plasmids-like entities

83 >1 Mb (0.03% - 1.17%) that mostly associate with *Alphaproteobacteria*[20–24]. Here, we report
84 *Enterobacteriaceae* megaplasmids, including some that associate with *E. coli*. *E. coli* plasmids
85 have been studied intensively, but the sizes of the best researched examples range from ~5.8
86 kbp (ColE1) to ~100 kbp (F (fertility) plasmid)[25,26]. In a collection of ~4,500 circular *E. coli* plasmid
87 genomes, the average size was 50 kb, although some clinically relevant IncF plasmids have ~150
88 kb genomes[27]. We demonstrate that *Enterobacteriaceae* megaplasmids are widely detected in
89 human and animal microbiomes, and notably, that they can be present in the microbiomes of
90 preterm infants, which are particularly susceptible to colonization by *E. coli, Klebsiella,*
91 *Enterobacter,* and *Citrobacter*[28–30]. Previous detection of these sometimes highly abundant
92 megaplasmids was likely prevented by assembly fragmentation and general focus on bacterial
93 host genomes. Our analyses employed long-read and short-read sequencing, along with manual
94 curation, to generate 20 confident, circularized genomes, of which 19 are fully validated and
95 complete. *In silico* protein structure prediction was used to improve functional assignments and
96 methylation pattern analysis thereby constraining plasmid-host associations. Overall, we report
97 new aspects of *Enterobacteriaceae* plasmid biology, with implications for understanding of host
98 bacteria that are especially relevant in human microbiomes.

99 **Results**

100 **Large extrachromosomal elements in human infant gut metagenomic data**

101 Long-read sequencing of a stool metagenomic sample from a preterm infant (INF1340011)
102 identified an unusual large linear contig that was curated using Illumina reads into a circular
103 genome of 1,583,849 bp in length (pMEGA1 - Table 1). As the genome does not encode rRNAs,
104 ribosomal proteins, or core bacterial metabolic functions, we inferred that pMEGA1 is an
105 extrachromosomal element (ECE).

106 Using long-read sequencing, we detected pMEGA1 in three other samples for infant 1340011 and
107 found a different (MASH distance 0.24) large ECE from infant 1330004, for which manual curation
108 generated a circular, complete genome of 1,334,733 bp (pMEGA2) (Tables 1, S2). Subsequently,
109 we identified another PacBio sequence for a circular 1.37 Mb element (pMEGA3) in INF1340011,
110 which lacked sufficient Illumina reads for curation to confirm all base calls (Table 1). The genome
111 shares 98.3% ANI (88.6% aligned fraction) with pMEGA2 (Table S2). Interestingly, throughout
112 the first 8 weeks of life, the gut microbiome of INF1340011 harbored pMEGA1 and pMEGA3
113 concurrently (Figure S1). In both infants, these elements appeared as the gut microbiome shifted
114 towards a community dominated by bacterial strains of the family *Enterobacteriaceae* (Figures
115 1A, S1). The maximum percentage of reads mapped to these ECEs in their respective infant gut
116 samples was 15.74% for pMEGA1, 3.73% for pMEGA2, and 1.30% for pMEGA3 (Figures 1A, S1;
117 Tables S3, S4).

118 **Large extrachromosomal elements in bacterial isolates and various host-associated**
119 **microbiomes**

120 Through BlastP searches against NCBI, we identified an *Escherichia coli* isolate (strain G269_1R)
121 from a rectal swab of a chicken from Bangkok, Thailand with assembled fragments similar to
122 pMEGA1[31]. After extensive manual curation of Oxford Nanopore long-read data with high-quality
123 Illumina reads, we reconstructed a complete circular 1,415,043 bp genome (pMEGA4) that shares
124 95.8% ANI with pMEGA1 (85.0% aligned fraction) (Tables 1, S2). The host *E. coli* strain is a ST10
125 clone and, in addition to pMEGA4, contains a 104 kb circular plasmid that we classified based on
126 the presence of plasmid markers (i.e., *oriT*, *oriV*, relaxase, replication initiation protein, and
127 partitioning proteins ParAB), and high similarity to other *E. coli* plasmids (NZ_CP070916.1,
128 NZ_CP050045.1)[31].

129 We surveyed more than 400 metagenomic and bacterial isolate sequencing projects and found
130 compelling evidence for nearly 100 high-quality ECEs related to pMEGA1 and pMEGA2 from
131 diverse sample types across multiple continents (Figures 1BCD, S2; Table S5). We detected
132 elements in 10 infants for which we had time series data; in three infants, these elements persisted
133 within the gut microbiome for months (Tables S6). Interestingly, these ECEs were also found in
134 at least 23 *Escherichia coli* and 12 *Salmonella spp.* isolates (Figure S2). We predicted the PCN
135 to range between 0.17 and 7.06, with *Salmonella spp*. showing a higher average PCN (2.79) than
136 *E. coli* isolates (0.59) (Table S5). On average, the pMEGA1 and pMEGA2 sequences represent
137 0.69% and 2.98% of reads in samples, respectively.

138 We manually curated and confirmed the circularity of an additional 16 elements (Table S7). Given
139 differences in size and GC content as well as ANI and MASH distance, we predict there are two
140 distinct clades: Clade 1 represented by pMEGA1 and 4, and Clade 2 by pMEGA2 and 3 (Figure
141 S3AB; Table S2). There is some synteny across the elements, but there is also evidence of
142 extensive rearrangements (Figure S3C).

143 **Identification of pMEGA elements as megaplasmids**

144 Given that ECEs are dominated by relatively novel proteins, we used *in silico* protein structure
145 prediction in addition to other annotation tools to predict functional annotations for pMEGAs 1-4
146 (Methods). We did not identify proteins that would suggest classification as bacteriophages. Only
147 a few genes encode proteins directly related to core bacterial metabolism, suggesting that these
148 elements are not secondary chromosomes/chromids (Table S8). Further support against
149 classification as a chromid or secondary chromosome is the difference in GC content between
150 pMEGA4 (25.98%) and its *E. coli* host genome (50.83%).

151 Consistent with the identification as megaplasmids, we identified genes for the partitioning
152 proteins (ParC,E (pMEGA4_1,2) components of Topoisomerase IV) and ParF (pMEGA4_1239),

153 a ParA family NTPase of the Type-Ib partitioning system typically co-located with ParG in low
154 copy number plasmids[32]. Although a ParG homolog was not identified, we identified PadC, a
155 protein with a ParB-like nuclease domain (pMEGA4_1030) known to interact with ParA.

156 We also identified most components of a Type IV secretion system within the first 80 kb of the
157 genome (e.g., (Vir/Dot/Icm proteins), two relaxases (pMEGA4_58, 1394), and two subfamilies of
158 replication initiation (Rep) proteins (Figure S4AB; Tables S8, S9). The putative relaxases are most
159 similar to the $MOB_F$ family, which includes many large and conjugative plasmids, and clade with
160 plasmids from diverse bacterial hosts (Figure S4B)[33]. An intergenic AT-rich region adjacent to a
161 relaxase (pMEGA4_1394) could serve as an origin of transfer (OriT), but this was not identified
162 as such by oriTFinder2[34]. The predicted Rep proteins (pMEGA4_629, 1423) share <31% amino
163 acid identity with bacterial sequences and the predicted structures are similar to those of RepC
164 from an IncQ plasmid in *Salmonella enteritidis* and TrfA from an IncP plasmid in *Escherichia coli*
165 (Figure S4C). However, the Rep proteins form distinct phylogenetic clades, and thus these
166 plasmids are only distantly related to IncQ and IncP plasmids (Figure S4A). One origin of
167 replication (OriV) with five 18 bp direct repeats was detected upstream of pMEGA4_1423, and
168 another possible iteron-containing locus with less conserved repeats occurs downstream of
169 pMEGA4_629 (Figure S4D)[35]. These features, and a type IV coupling protein (T4CP
170 pMEGA4_20), indicate that these are conjugative megaplasmids (i.e., self-transmissible)[33,36].

171 ***Enterobacteriaceae* host these megaplasmids**

172 pMEGA4 was reconstructed from an *E. coli* isolate and pMEGA17 and near identical pMEGA18,
173 were reconstructed from two impure *E. coli* isolates (Table S7). Incomplete megaplasmid
174 genomes were reconstructed from isolates of *Salmonella enterica* and *E. coli*. We attempted to
175 identify the pMEGA1, 2 and 3 hosts using DNA crosslinking applied to stool samples from two
176 infants, but the results were inconclusive. While most proteins (>66%) in pMEGA1-3 have an

177 unknown taxonomic classification, ~8% have sequence similarity to members of the order
178 *Enterobacterales* (Table S7).

179 In attempts to identify the pMEGA1-3 hosts, we used abundance-based correlation analysis. The
180 strongest correlation in abundance for pMEGA1 was with *Enterobacter cloacae* 1 (0.92), for
181 pMEGA2, *Citrobacter freundii* 1 (0.88), and for pMEGA3, *Klebsiella michiganensis* 1 (0.9) (Table
182 S10). We also investigated DNA modification patterns using PacBio HiFi reads and compared
183 potentially modified motifs shared between pMEGAs, coexisting bacterial genomes, phage and
184 plasmids (Figures 2, S5). Strong overlap in modified motifs between pMEGA2 and the *Citrobacter*
185 *freundii* 1 genome supports the correlation-based host prediction. For pMEGA1, which only
186 shares motif patterns with pMEGA3, the results are inconclusive (Figure S5). pMEGA3 has
187 overlapping motifs with a *Citrobacter freundii* and an *Enterobacter cloacae* strain. The pMEGAs
188 have more numerous and more diverse modified motifs compared to bacterial genomes within
189 the same microbiome sample. Paralleling this, pMEGA4 encodes 21 methyltransferases (MTs),
190 whereas the host *E. coli* genome encodes only 6 (Table S11). Overall, based on isolate source,
191 protein similarity, cooccurrence patterns, CRISPR targeting and methylation analysis, we
192 conclude that various *Enterobacteriaceae* host the megaplasmids.

193 pMEGA4 encodes a type-I-Fv CRISPR-Cas system with a complete Cas1-Cas2 integrase, a
194 Cas3 helicase-nuclease effector, and a compact Cascade surveillance complex likely encoding
195 only five subunits (Figure S6A-D)[37]. Components of these systems are not completely conserved
196 in all pMEGAs. Consistent with active adaptive immune activity, all pMEGAs contain CRISPR
197 arrays with between 27 and 220 spacers (Table S7). No CRISPR spacers from the pMEGAs
198 targeted coexisting bacterial genomes, and vice versa. However, pMEGAs share some CRISPR
199 spacer sequences with other pMEGAs from the same clade (but not between clades), and a
200 subset of plasmid CRISPR spacers target Cas genes of other pMEGAs (Figure S6E). Many

201 spacers of pMEGAs target plasmids from *Enterobacteriaceae*, consistent with the role of plasmid
202 CRISPR loci in plasmid-plasmid competition (Figure S6E).

203 We used CRISPR spacers from databases and metagenomes to search for targeting of pMEGAs
204 that may be indicative of megaplasmid host associations[38,39]. We identified 12 non-redundant
205 spacers (100% match) from 33 metagenomes and an isolate of *Proteus cibarius*
206 (*Enterobacteriaceae*) that target regions in at least one pMEGA (expectation for random match of
207 ~$e^{-11}$) (Table S12). The *Proteus* isolate spacer targets a $tRNA_{Met}$ locus that is conserved in
208 pMEGA1, 4, and 20[40]. Five of the 11 metagenome-derived spacers also target tRNAs.

209 To put these megabase-scale plasmids into context, we analyzed the PLSDB[21] and IMG/PR[22]
210 databases and estimated that of the more than 770,000 plasmids, there are 667 non-redundant
211 plasmids with a genome size > 1 Mb, all but four of which are from isolates. Only 30 were detected
212 in *Gammaproteobacteria* (Figure S7). Nine circular and five linear > 1 Mb secondary replicons of
213 *Enterobacteriaceae* genomes are listed in NCBI[23,41]. Given similar GC content compared to the
214 primary chromosome and genes for many core functions (including multiple rRNAs in some
215 cases), we suspected that they are not megaplasmids (Supplementary Information; Table S13).
216 We assayed hypothetical gene content of the public putative *Enterobacteriaceae* megaplasmids
217 and found essentially no differences compared to the main chromosome (24.9% vs 23.2%) (Table
218 S13). This is in stark contrast to the pMEGAs, which feature ~ 80% hypothetical genes, compared
219 to 21.1% for the *E. coli* host of pMEGA4 (Table S13). We conclude that these NCBI sequences
220 are secondary chromosomes or parts of the main chromosomes.

221 **Clades 1 and 2 differ in their multi-copy proteins**

222 Of 2,162 protein subfamilies (SF), 438 occur in all of the 20 pMEGAs (Table S9; Figure 3A). These
223 conserved proteins are involved in transcription, translation, replication, protein folding, defense
224 broadly defined, pilus formation, secretion systems, are annotated as proteases, phosphatases,

225 transport-related, methyltransferases, chemotaxis proteins and proteins involved in NAD
226 metabolism and acetylation/deacetylation (Table S9; Figure 3A). Of the 438 ubiquitous proteins,
227 336 occur only once per genome, typically in similar genomic locations within clades (although
228 the organization differs between clades; Figure 3C).

229 Predicated on the idea that multicopy proteins (here defined as 3 or more copies/genome) may
230 be important to the functioning of megaplasmids, we counted the instances of each subfamily in
231 the pMEGA1-4 genomes (Table S14; Figure 3B). Of the 10 subfamilies that are multicopy in all
232 four megaplasmids, the most prevalent are beta-solenoid structures (SF0950: 10 - 16 per
233 genome), possible bacterial shufflons/pilin proteins (SF2096: 7 - 16 per genome), TnpB proteins
234 (SF1483: up to 11 per genome), and protein phosphatases (SF1635). pMEGA1 and 4 (and to a
235 lesser extent, pMEGA2 and 3) have numerous proteins (from many subfamilies) with predicted
236 alpha-solenoid structures. The alpha-solenoids are of varying lengths and are encoded by >300
237 genes (~20% of the proteome) of pMEGA4 (Figure 3BC). Many had very low-scoring hits to
238 structures in the PDB, and those with higher-scoring matches were mostly to Designed Ankyrin
239 Repeat Proteins (DARPins), which are of interest as protein-based therapeutics[42]. Multicopy
240 proteins that are only highly prevalent in pMEGA1 and 4 are serine recombinases (SF2001: TnpA
241 - up to 23 copies) and their accompanying endonucleases, and DNA methyltransferases
242 (SF0095), whereas novel proteins, possible immunity-related proteins, and three distinct RNAse
243 H1 subfamilies are only highly prevalent in pMEGA2 and 3 (Figure 3B).

244 Some multi-copy protein subfamilies are clustered in the genomes (Table S14). Two to four copies
245 of proteins in the same subfamily occur sequentially in 58 and 33 instances in pMEGA1 and 4,
246 respectively. In pMEGA2 and 3, two to eleven genes of the same subfamily occur sequentially 92
247 and 110 times, and there are eight RNAse H1 subfamily proteins encoded almost consecutively.
248 In addition to the conserved RNAse H domain, these proteins have diverse additional

249 regions/domains, some of which are positively charged (possibly involved in DNA binding), but
250 for which functions could not be assigned.

251 Copies of SF0552, proteins similar to a toxin immunity protein, are clustered in pMEGA1, 2 and
252 3 (Figure S8A). Encoded immediately upstream is a protein with an AHH domain (SF1803),
253 commonly found in secreted nuclease toxin effectors (Figure S8BDE)[43]. When folded together,
254 the putative immunity protein is confidently predicted (chain_pair_iptm values 0.91-0.92) to
255 interact with the AHH domain (Figure S8C). The cluster is reminiscent of poly-immunity loci
256 associated with polymorphic toxins[44–46]. pMEGA2 and 3 have an additional region with the
257 putative toxin SF1803 encoded upstream of a cluster of possibly immunity proteins SF0823. To
258 screen for other toxin/poly-immunity loci, we analyzed proteins directly upstream of multicopy
259 clustered proteins to identify toxin domains. We found SF0417 as a possible ADP-
260 ribosyltransferase upstream of 3 and 5 copies of SF0062 in pMEGA2 and 3, and a possible
261 protease (SF0355) upstream of 3 copies of SF0744 in pMEGA2. Poly-immunity appears more
262 common in pMEGA2 and 3 as compared to pMEGA1 and 4.

263 **Functional overlap and differences between pMEGA4 and *E. coli* host**

264 Given that pMEGA4 originated from an *E. coli* isolate, we compared the predicted
265 proteomes/genetic repertoires of this megaplasmid with those of its host in detail (Table S15). Of
266 note, only 9 proteins shared >70% amino acid identity to host proteins (Table S16). pMEGA4
267 encodes numerous proteins involved in nucleic acid manipulation, including polymerases,
268 helicases, nucleases, proteins involved in recombination and nucleotide interconversion (e.g.,
269 GMP->GDP) and over 20 translation-related proteins (Table S8). These functions are also
270 encoded on the *E. coli* genome.

271 Possibly involved in homeostasis and quality control, the pMEGA4 genome encodes the protease
272 ClpP (pMEGA4_187) and associated proteins[47–50](Figure S9AB). The genome encodes three

273 copies of proteins with high structural similarity to FtsH (pMEGA4_57, 182, 1364) and three
274 proteins similar to HflK/C that multimerize into functional microdomains (Figure S9CD)[51,52]. The
275 *E. coli* genome encodes ClpP and associated proteins, HflK/C, as well as one copy of FtsH.

276 In addition to their role in conjugation, T4SSs may also be involved in effector protein delivery and
277 virulence (similar regions are found in pMEGAs 1, 2, and 3)[53]. The *E. coli* host genome encodes
278 components of Type II and Type VI secretion systems, including VgrG and Hcp, neither of which
279 were identified in pMEGA4[54]. Most interesting, and seemingly not found in the *E. coli* host, is a
280 protein with structural homology to InvG[55] (Figure 4ABC).

281 Also encoded in the pMEGA4 genome are at least four proteins (pMEGA4_1281, 90, 966, 977)
282 with a von Willebrand factor A–like domain (vWA), which is commonly associated with adhesins.
283 The *E. coli* genome encodes two proteins with vWA domains. Another protein seemingly
284 redundant with the host *E. coli* repertoire, pMEGA4_299 encoded *fliC,* multimerized nicely into a
285 typical flagellin structure (Figure 4DEFG). This appendage is involved in adhesion, motility,
286 virulence, and immune stimulation[56,57].

287 pMEGA4 encodes 105 manually verified tRNAs, including at least one for each canonical amino
288 acid, with 91 unique tRNA sequence types (87%), whereas the *E. coli* genome encodes 86, of
289 which 49 are unique (57%) (Table S17). The tRNAs are mostly arrayed in blocks, generally with
290 ~6 to ~ 25 nucleotide (nt) gaps between them (Figure 5A). Only two blocks (with 7 and 4 tRNAs)
291 occur in the *E. coli* genome. Within the set of complete pMEGA genomes, there are up to 208
292 tRNAs, with statistically significant enrichment in Clade 2 (average 199 ± 9) compared to Clade 1
293 (average 113 ± 8).

294 For the tRNAs for the most highly represented amino acids (Leu, Lys, Met), almost all have the
295 same codon. For Lys, 9/10 are encoded in one block and all have the same codon (AAA) despite
296 being quite divergent in the rest of their sequences (Figure 5B). In contrast, the six *E. coli* tRNA

297 Lys are all identical (Figure 5C). For 11 other amino acids, all the tRNAs for that amino acid are
298 in a single block. In contrast, in the *E. coli* host genome, there is only one tRNA (Gln) where all
299 instances occur within a block (two codon types). In some cases, different tRNA types alternate
300 pseudo-regularly (sometimes with three or more different tRNAs per cluster).

301 A commonly proposed explanation for mobile element tRNAs is compensation for differences in
302 their codon usage compared to that of their host. We find that some codons, which are much
303 preferred by pMEGA4 compared to the host *E. coli,* have more highly represented tRNAs (e.g.,
304 AAA, TTA); however, this is not the general rule (Figure 5D; Table S18). Some less commonly
305 used tRNA codons (e.g., ATG, AAC) are highly represented, and some much more highly used
306 codons (e.g., AAT, ATA) are not represented in pMEGA4. Notably, there is high sequence
307 diversity associated with all highly used codons (Figure 5D).

308 **Panda gut metatranscriptomic data reveals broad expression of pMEGAs**

309 We identified a public metatranscriptomic dataset from the gut microbiome of a panda, which
310 contains both a Clade 1 and a Clade 2 megaplasmid (SRR10902853). As there was no DNA data
311 available, we mapped the reads to pMEGAs 1-4 competitively and detected extensive read
312 recruitment to all four genomes. By far the highest recruitment was to pMEGA4 (almost one million
313 reads), but over 128,000 reads were recruited to pMEGA3. Most genes are expressed at relatively
314 even levels. Extreme peaks in RNA abundance are all associated with intergenic regions, two of
315 which are > 1 kbp in length. Many of the highly expressed protein-coding genes lack functional
316 assignments and are very small (<100 amino acids). However, some genes with high expression
317 have reasonably confident functional annotations. A region of pMEGA3 (Figure 6A) that features
318 a large cluster of proteins with RNAse H1 (as well as other) domains (Figure 6B) includes a highly
319 expressed protein annotated in part as TPP1, a possible Cas1-Cas4 fusion protein, and SlyB
320 (Figure 6C). Also expressed, albeit at lower levels, are two Clp proteases and a L, D-
321 transpeptidase, YcbB, which is possibly involved in resistance to β-lactams due to alternative

322 cross-linking of peptidoglycan[58]. Of the 170 tRNAs in pMEGA3 and 108 in pMEGA4, virtually all
323 recruited transcripts. Some abundant intergenic transcripts are adjacent to expressed TnpB
324 genes (likely the omega RNA). Some TnpBs are among the genes without expression, although
325 genes without transcription may be absent or divergent in the panda pMEGAs.

326 **pMEGA4 may confer stress protection and antibiotic resistance, and aid in host**
327 **competition**

328 pMEGA4 encodes a protein that forms a ferritin-like cage for DNA protection during starvation
329 (pMEGA4_661; also identified in *E. coli*) and two small alarmone hydrolases that regulate the
330 stringent response in a similar way to *gppA* found in the *E. coli*[59,60]. Further supporting defense
331 against external challenges are genes for redox homeostasis, nucleic acid repair, oxidative stress,
332 heat shock, and genes that form, maintain and repair disulfide bonds. Again, these functions
333 appear redundant with those encoded in the *E. coli* genome. pMEGA4 encodes the outer
334 membrane protein SlyB (also encoded in the host genome), which is a lipoprotein that forms a
335 13-mer pore (Figure S6C). In the panda pMEGA metatranscriptome, the SlyB gene was one of
336 the most highly expressed (Figure S6A). The PhoPQ regulon is responsible for the expression of
337 SlyB in response to external stress[61]. We also identified three near-consecutive genes for
338 tellurium resistance (TerD, not identified in *E. coli*).

339 Possibly in response to bacterial competition, pMEGA4 encodes a protein implicated in immunity
340 to effectors (pMEGA4_953)[62], an amidase antitoxin (pMEGA4_699), and a hydrolase against
341 ADP-ribosylating toxins (pMEGA4_1469[63]). These antitoxins appear without their toxin
342 counterparts. Conversely, there are at least two proteins annotated as the lipoprotein toxin
343 entericidin B that are separate on the plasmid and do not appear co-localized with an antitoxin
344 (SF1615)[64]. The *E. coli* host genome encodes multiple Type II toxin/antitoxin systems (ie. MazEF,
345 HipAB) and many possible secreted effectors/toxins[65–67].

346 In addition to the CRISPR-Cas system described earlier, pMEGA4 encodes putative defense
347 system proteins structurally similar to the Upf1 family of RNA helicases (pMEGA4_338,520;
348 detected by PADLOC as Mokosh TypeII systems), and a Ser/Thr kinase annotated as a PD-T4-
349 6 system (pMEGA4_91)[68–70]. The *E. coli* genome encodes 7 single-gene phage defense
350 candidate systems, 2 Hma (Helicase, Methylase, ATPase) defense systems, HEC, hhe, and a
351 CRISPR array[71].

352 The *E. coli* host of pMEGA4 encodes the extended-spectrum beta-lactamase (ESBL) $bla_{\text{CTX-M-55}}$
353 on the abovementioned 104 kb plasmid, along with genes for quinolone (*qnrS1*), chloramphenicol
354 (*catII*), and tetracycline resistance (*tetR* and *tetA*)[31] (Table S19). Few antibiotic resistance
355 mechanisms were detected in pMEGA4 (Table S19). However, this function may be conferred by
356 10 genes for transporters, including the MacAB- and AcrAB-TolC efflux systems, which were
357 identified based on sequences and predicted structures[72]. Possibly relating to antibiotic
358 resistance, pMEGA4 contains an outer membrane porin implicated in carbapenem import
359 (pMEGA4_670), and a methyltransferase that may activate a peptide antibiotic (pMEGA4_515).

360 **Discussion**

361 We report discovery and bioinformatic analysis of two clades of megaplasmids. The
362 megaplasmids are clearly distinguished from bacteriophages by absence of structural genes and
363 from chromids/chromosomes by the low fraction of metabolic genes, protein novelty, plasmid-
364 associated genes, replication style and difference in GC compared to host genomes. Members of
365 both pMEGA clades whose genomes were reconstructed from isolates are confidently linked to
366 *E. coli* and *Salmonella spp.* hosts. More broadly, protein similarity, co-occurrence, and DNA
367 modification pattern analyses indicate that the hosts are likely all *Enterobacteriaceae*. With
368 genomes of up to 1.58 Mb, these are the, by far, the largest megaplasmids in bacteria of this
369 family.

370 While initially identified in preterm infant gut microbiomes, the megaplasmids occur in
371 microbiomes of the human adult gut, animal gut and other environments, suggesting they are
372 widespread. All but one of 20 genomes is “complete”, rather than bins or putatively circularized
373 sequences. The exception is pMEGA3, which was reconstructed from PacBio reads but was too
374 low in abundance to verify 100% of the bases using Illumina reads. Complete genomes ensured
375 the accurate determination of genome lengths, provided information about genome organization
376 and replication style and facilitated accurate gene and proteome predictions. *In silico* structure
377 prediction and analysis helped identify proteins implicated in virulence, components of the T4SS,
378 proteins involved in host-cell adhesion, antibiotic resistance, and (phage) defense. Overall, the
379 findings suggest that these megaplasmids are self-transmissible. Extensive methylation motifs to
380 protect from diverse host restriction systems suggest that the pMEGAs may be able to replicate
381 in multiple hosts.

382 Plasmids must ensure that the burden of their existence does not compromise the host organism’s
383 viability[73]. The degree to which plasmids must offset the cost of their carriage is probably size-
384 dependent and may drive plasmid acquisition of genes that increase the host’s fitness.
385 Presumably, pMEGA genomes became large, in part, through the acquisition of gene variants
386 from prior hosts. This likely occurred in an ancestral plasmid, as all pMEGAs have very low GC
387 content compared to *Enterobacteriaceae* hosts (i.e., the GC content of host-derived genes has
388 been ameliorated).

389 Presumably reflecting acquisition from their host, there is substantial overlap in gene content of
390 pMEGA4 and *E. coli*. Some plasmid-borne variants may confer advantages by different specificity
391 or enzyme kinetics, by supplying genes for stress protection, or by extending growth over a wider
392 range of conditions. For example, plasmid-borne Clp and FtsH proteases exhibit different
393 thermotolerance[74–77] or specificities[48,49,78–80] that may expand substrate range. Associated
394 membrane microdomains (HflK/C) may result in compartmentalization of specific proteins

395 involved in signal transduction, membrane trafficking, and regulation of proteases, including
396 FtsH[51,52,81]. Other plasmid-derived benefits include expanding defense against invading
397 bacteriophages and other plasmids (e.g., by provision of additional CRISPR-Cas systems).
398 Plasmids may also increase antibiotic resistance by supplying efflux pumps with different
399 specificities, remodelling the cell wall, and expressing variants of thymidylate synthase, LpxC,
400 and ClpP that are insensitive to antibiotics[58,82–84].

401 Megaplasmids may acquire functions to counter host defenses. They encode genes for NAD
402 production, likely to offset the decreased availability of NAD when the host depletes its NAD pool
403 in response to infection[85]. Host tRNA degradation in response to invasion by ECEs may be
404 countered by megaplasmid tRNA ligases, as demonstrated with phages[86]. The pMEGAs encode
405 upwards of 100 tRNAs. Although some certainly offset differences in codon use between the
406 pMEGA and its host (Figure 5), structural diversity in tRNAs with the same anticodon may confer
407 a spectrum of resistances to host tRNA nucleases. This counter-defense is well established in
408 phages[87]. There is apparently strong selective pressure to preserve the codon and for sequence
409 divergence outside of the codon.

410 Megaplasmids encode functions that do not duplicate those of their hosts. Of potential medical
411 importance in the pMEGA4 plasmid in the ST10 *E. coli* strain is InvG. This protein forms the outer
412 membrane (OM) basal body in the Type III secretion system injectisome, which delivers effector
413 proteins into host cells, but is also found in Type II secretion systems, Type IV pili structures, and
414 phage assembly complexes[55,88,89] (Figure 4). Putative toxin and poly-immunity arrays may
415 enhance inter-bacterial competition and protect from diverse toxins, whereas diverse adhesion
416 factors (ie. vWA domains) may contribute to motility and virulence[90,91]. Highly prevalent multicopy
417 proteins have alpha- and beta-solenoid structures, which are often found in secreted effectors[92].
418 Effectors containing ankyrin repeats from intracellular pathogens such as *Legionella pneumophila*
419 and *Coxiella burnetii* are involved in the invasion of eukaryotic cells[93]. Alpha-solenoid folds are

420 more common in proteins of obligate intracellular pathogens as compared to other bacterial
421 genomes[94]. Ankyrin repeat proteins (a form of alpha-solenoids) often mediate protein-protein
422 interactions and may have functions analogous to those of DARPins, which are alpha-solenoids
423 that are engineered to bind to target molecules with high affinity and specificity for functions such
424 as antibodies and as anti-viral agents[95]. It is also possible that the megaplasmids accumulate
425 ankyrin proteins to modulate their host proteome[96].

426 Some megaplasmids associate with bacteria that cause gastroenteritis (i.e., *S. enterica*) and
427 sepsis (i.e., *E. coli*)[97,98]. The pMEGA4 host *E. coli* is an ESBL-producing ST10 clone. ESBL-
428 producing ST10 clones of *E. coli* have been increasing in prevalence globally[99] and are common
429 in animals, including chickens, associated food products, and humans[100–103]. Some other
430 megaplasmid-bearing *E. coli* and *S.* Enteriditis isolates originate from human samples, and one
431 *E. coli* isolate of serotype O157:H7 was associated with an outbreak due to contaminated lettuce
432 (Biosample: SAMN14113693)[104,105]. These observations motivate further investigation of the
433 medical significance of megaplasmids.

434 **Conclusions**

435 We identified the first megabase-scale plasmids in *Enterobacteriaceae*. They were detected in
436 infants, adults, and animals, as well as in hospital sinks where they likely serve as a reservoir for
437 human transmission[106]. The megaplasmid genomes encode numerous solenoid proteins that may
438 bind selectively to proteins or other molecules. We also identified mechanisms to extend host and
439 habitat ranges, increase competitiveness in microbiomes, counter host defense, and protect from
440 antibiotics. Much remains to be learned about their large and incompletely defined functional
441 potential; however, there are many indications that they hold medical significance. From the
442 perspective of future work, it is important that they occur in genetically tractable
443 *Enterobacteriaceae*, including *E. coli*.

## Acknowledgements

We thank Dr. Rani Phoolwanti for input regarding the RNAase H1 proteins and Colin Robinson for input regarding the metatranscriptome data. We thank the generators of the panda metatranscriptome data SRR10902853, whom, to date, we have not been able to identify. We thank Dr. Shufei Lei and Leylen Miloslavich for technical support. The Innovative Genomics Institute supported AKG, JFB, OTT, RS, and SW.

## Author Contributions:

AKG and JFB were responsible for the conceptualization and design of this study. Assembly was performed by AKG and genome curation by AKG and JFB. AKG, JFB, OTT, L-XC performed metagenome and metatranscriptome data analysis. SW, RS analysed the DNA modification patterns. OTT performed CRISPR analyses. NM and MH generated the nanopore sequencing of the *E. coli* isolate. AHD, SB, VTD, AEM, JC and AK provided data for megaplasmid detection. BF coordinated and submitted infant stool samples for sequencing. RS provided computational support. AJ, LS and MM provided input regarding clinical relevance. Figures were generated by AKG, SW, OTT, and JFB. The manuscript was written by JFB and AKG, with input from OTT. JFB and MJM provided the majority of the funding support. All authors read and approved the manuscript prior to submission.

## Declaration of interests

The authors declare no conflict of interest

## Declaration of generative AI and AI-assisted technologies

No generative AI and AI-assisted technologies were used in the writing process.

466 **Funding**

467 The primary funding for this project was provided by NIH award 5R01AI092531-14 to JFB and
468 MM, with additional support from the Chan Zuckerberg Foundation via the Innovative Genomics
469 Institute. Long-read PacBio sequencing was provided by a SMRT grant to AKG from PacBio and
470 SeqCenter. RS and SW were funded by Lyda Hill Philanthropies, Acton Family Giving, the
471 Valhalla Foundation, Hastings/Quillin Fund, the CH Foundation, Laura and Gary Lauder and
472 Family, the Sea Grape Foundation, the Emerson Collective, Mike Schroepfer and Erin Hoffman
473 Family Fund, the Anne Wojcicki Foundation. AHD and AEM were funded by the Biotechnology
474 and Biological Sciences Research Council (BBSRC) Institute projects BB/X011011/1 and
475 BBS/E/QU/230002A. AHD was funded by a UKRI BBSRC Doctoral Training Grant BB/T008717/1.
476 AK was funded by the Finnish Multidisciplinary Centre of Excellence in Antimicrobial Resistance
477 Research (FIMAR; 346127). MH received funding from the Swiss National Science Foundation
478 grant IZJFZ3-177614.

479 **Data and code availability**: Prior to publication, the genome sequences described in this study
480 can be accessed and downloaded from ggKbase:
481 https://ggkbase.berkeley.edu/project_groups/Megaplasmids_public . Raw sequencing data will
482 be made available on NCBI upon publication. Other data reported in this study can be accessed
483 at: 10.5281/zenodo.1723140**.**

484 **Methods**

485 **DNA extraction, metagenomic sequencing, and binning**

486 Stool samples from two preterm infants enrolled in the NICU Antibiotics and Outcomes (NANO)
487 trial (Clinical Trial ID: NCT03997266) were collected over the first eight weeks of life after
488 obtaining informed parental consent[107]. DNA was extracted from samples using the
489 ZymoBIOMICS™ DNA Miniprep Kit. Libraries prepared with the Illumina DNA prep kit were

490 sequenced on an Illumina NovaSeq 6000 (2 x 151 bp reads). Reads were trimmed using sickle
491 v1.33, and mapped against the human T2T reference using bowtie2 v2.5.4, resulting in an
492 average of 4.09 gigabases per sample (Table S1)[108,109]. Metagenomic assembly was performed
493 using metaSPAdes v4.0.0[110]. Scaffolds greater than 1000 bp were annotated using Prodigal
494 v2.6.3, and searched against KEGG, UniProt, and UniRef using USearch v10.0.240 (-ublast)[111–
495 115]. Ribosomal RNA sequences were predicted as previously described, and tRNA sequences
496 were identified using tRNAscan-SE v1.3.1[116–118]. For a subset of samples (Table S1), either the
497 same DNA extract used for Illumina sequencing, or a new DNA extract prepared as above but
498 with a modified lysis step, was sequenced on a PacBio Revio. HiFi reads were trimmed using
499 BBDuk v39.10 and assembled using hifiasm-meta v0.3-r073, resulting in, on average, 6.99
500 gigabases per sample[119,120]. Binning of Illumina short-read data to generate metagenomic-
501 assembled genomes (MAGs) was accomplished using a combination of CONCOCT v1.1.0,
502 MaxBin2 v2.2.7, MetaBAT2 v2.15 and Vamb v3.0.2, and the best bins were selected using DAS
503 Tool v1.1.6[121] with default settings[121–125]. The PacBio HiFi-MAG-Pipeline v3.3.2 was used with
504 slight modifications to detect putative complete circular genomes (93% completeness) and bin
505 other scaffolds into MAGs with fewer than 100 contigs[126]. Other settings included: a score
506 threshold of 0.2 for DAS Tool, a minimum completeness of 70%, and a maximum contamination
507 of 10%. Metagenomic assemblies were visualized and binned manually through an in-house
508 platform ggKbase (https://ggkbase.berkeley.edu/). MAGs from both Illumina and PacBio
509 assemblies for each infant were dereplicated using dRep v.3.4.5[127] (with completeness (-comp)
510 75, contamination (-con) 10, secondary cluster ANI (-sa) 0.98, and coverage threshold (-nc) 0.25).
511 This resulted in 27 MAGs for INF134001 (total 110) and 16 MAGs for 1330004 (total 72). The
512 taxonomy of each MAG was determined using PhyloPhlAn v3.0.36[128].

513 **Identification and manual curation of megaplasmids in the preterm infant gut**

514 Examination of a PacBio assembly of one infant sample (INF1340011_TP8) identified a 1.58 Mb
515 (pMEGA1) contig of unknown origin, with 3 bacterial single-copy genes, no ribosomal rRNA
516 sequences, and many gene annotations characteristic of a plasmid. Two additional putatively
517 complete plasmids were identified in other PacBio assemblies: a 1.33 Mb plasmid (pMEGA2) in
518 INF1330004 at TP4 and a 1.37 Mb plasmid (pMEGA3) in INF1340011 at 4 weeks. Trimmed and
519 filtered reads were mapped against the putative complete megaplasmids with BBMap v39.10[119]
520 (default settings with ambiguous=random). The alignment file was imported into Geneious Prime
521 v2025.1.1, and reads were re-mapped to allow 0% mismatches to identify regions of possible
522 variation or error. Reads were remapped to allow 4% mismatches to the reference to correct
523 possible single-nucleotide polymorphisms, insertions, deletions, or assembly errors. This process
524 was repeated until no errors were visible in the plasmid genome. Reads that surpassed the ends
525 of the sequence were used to extend the plasmid until a repeated sequence was present at the
526 beginning and end of the plasmid, suggesting circularity. Paired reads that mapped at both ends
527 of the sequence were also used as an indication of a circular plasmid, as described
528 previously[129,130]. CoverM v0.7.0 was used along with Illumina reads to determine the depth and
529 percentage of reads mapped to pMEGAs and the dereplicated bacterial MAGs for each infant
530 across all available timepoints[131].

531 **Identification and manual curation of a megaplasmid in *Escherichia coli***

532 To further assess the diversity and distribution of these megaplasmids in other datasets, we used
533 the bacterial single-copy gene histidyl-tRNA synthetase (*hisS*) as a marker. A BlastP search
534 against NCBI non-redundant protein sequences (executed December 16th, 2024) resulted in 1
535 highly similar hit for pMEGA1 (*Escherichia coli* – 93% cov, 97.63% aa ID, BioSample:
536 SAMN18191027, Assembly: GCA_024673755.1)[132–134]. The top BlastP match for pMEGA2 is

537 from a *Serratia sp.* with 98% coverage and 63.84% amino acid similarity (WP_282994782.1). The
538 *E. coli* sequence is from a genomic assembly of an isolate (G269_1R) from Bangkok,
539 Thailand[31,135]. We assembled the Illumina sequencing reads from this accession using the same
540 methods as described above. Scaffolds that were distinct in GC content from the host genome
541 (<30% for pMEGA vs ~51% for *E. coli*), similar in read coverage, and that had no consensus for
542 taxonomy based on the taxonomy of predicted gene sequences were binned. We also mapped
543 scaffolds from the assembly to the complete pMEGA1 and pMEGA2 sequences from above using
544 minimap2 v.2.28-r1209 (with the flag -x map-hifi)[136]. From these two approaches, we manually
545 selected 23 scaffolds encompassing ~1.39 Mb to represent a putative megaplasmid. Through this
546 approach, and to the best of our ability and knowledge, we believe that these scaffolds do not
547 belong to *E. coli* or other easily identifiable mobile genetic elements, such as other plasmids or
548 phages. Extensive manual curation and read mapping did not result in a complete circular
549 element, so Nanopore long-read sequencing was performed on the *E. coli* isolate. Briefly,
550 genomic DNA from the isolate was prepared with the Rapid sequencing DNA kit (SQK-
551 RBK114.24) and sequenced on a GridION MKI R10 flow cell. Reads were basecalled with the
552 high-accuracy model using Dorado v7.2.13[137]. Long-read sequencing yielded very low coverage
553 of the putative megaplasmid, but provided a complete 4.8 Mb genome for *E. coli* and a 104 kbp
554 plasmid. The few Nanopore long-reads that were recovered against the 1.39 Mb plasmid were
555 used to scaffold the contigs derived from Illumina sequencing. Through many iterative rounds of
556 read mapping and extension from the ends of Illumina contigs and into divergent gaps where no
557 direct read recruitment was possible, we reconstructed a complete 1.41 Mb plasmid (pMEGA4).

558 **Identification and manual curation of megaplasmids in other public datasets**

559 We employed two tools to search for similar sequences across read data deposited in NCBI's
560 Sequence Read Archive (SRA) and other collections. First, PebbleScout was used to search
561 against all available databases with the complete reference sequences of pMEGA1 and pMEGA2

562 elements[138]. This resulted in 55 hits with a PB score >60 for pMEGA1, and 319 hits with a PB
563 score >60 for pMEGA2. Next, we searched all available assembled data in Logan (11-20-2024)
564 using the *hisS* gene sequence (813 bp) from pMEGA1 and 1000 bp from the start of the *hisS*
565 gene sequence from pMEGA2[139]. This resulted in 61 hits for pMEGA1 and almost 500 hits for
566 pMEGA2 with a kmer coverage >=0.70. Many of the hits overlapped with the PebbleScout search
567 results; therefore, we selected a few studies and any bacterial isolate hits only for additional
568 investigation (Table S5). The Bioprojects PRJEB36775 and PRJNA938144 relate to hospital sink
569 metagenomes and infant gut metagenomes, respectively, and were selected given their high
570 number of hits and relevance to human health[140,141].

571 For all selected hits, reads from the SRA or European Nucleotide Archive (ENA) were downloaded
572 and processed as described above for trimming and human read removal, if needed. Each read
573 accession was treated as a separate sample regardless of whether more than one hit originated
574 from the same BioSample (eg. hits from PRJEB21259 and PRJEB36775). Trimmed reads were
575 mapped using bowtie2 and samtools to the pMEGA1 and pMEGA2 reference sequences to
576 determine the percent length coverage, depth of coverage and percentage of reads mapped to
577 the pMEGA in each sample[109,142]. We proceeded to assemble the hits using metaSPAdes v4.0.0
578 as described above and assessed the extent to which complete sequences could be recovered
579 by mapping scaffolds to pMEGA1 and pMEGA2 as references using minimap2[110,136]. Samples
580 were considered a high-quality hit for a pMEGA if 10 or fewer scaffolds provided coverage of
581 >=85% of the reference pMEGA. Samples were considered medium-quality hits if 11 to 50
582 scaffolds provided coverage of >=85% of a pMEGA, and all other hits were considered low-
583 quality. From the high-quality hits, a selection of samples was chosen for manual curation as
584 described above to attempt to complete the plasmid sequences and confirm circularity.

585 From the PebbleScout hits, SRR10902853 is listed as an RNA-seq dataset. To the best of our
586 knowledge, this sample and its corresponding Bioproject (PRJNA601687) are not affiliated with a

587 publication. We were unable to find an associated metagenomic dataset for this sample. We
588 mapped the trimmed reads to a concatenated reference of pMEGA1, 2, 3 and 4 using bowtie2
589 v2.5.4 and visualized the alignments in Geneious Prime v2025.1.1 to assess regions with high
590 coverage.

591 **Comparing the protein repertoire across 20 confident and circular megaplasmids**

592 After curation, we recovered 20 circular megaplasmid sequences (**Table S7**). These sequences
593 were re-oriented to start at a highly conserved protein sequence (ParE) and aligned using Mauve
594 v1.1.3 in Geneious Prime[143]. Skani v0.2.1 and MASH v2.2 dist were used to assess the average
595 nucleotide identity and distance between entire plasmid sequences[144,145]. The taxonomy of
596 predicted proteins within these plasmid genomes was summarized using Uniprot annotations and
597 tRep (https://github.com/MrOlm/tRep/tree/master/bin). All proteins across the 20 megaplasmids
598 were clustered into protein subfamilies using MMseqs2 and an all-vs-all search (cluster workflow
599 with E-value 0.001, sensitivity: 7.5, fraction of covered residues across target and query: 0.5, and
600 the Set-Cover (greedy) cluster mode)[146]. A multiple sequence alignment was generated for
601 subfamilies with at least 2 members using (mmseqs result2msa), and HMM profiles were
602 generated using hhmake from HH-suite3[147]. A HH-Suite database was generated from these
603 protein subfamilies, and subfamilies were compared against each other using their respective
604 HMMs and HHblits (with parameters -v 0 -p 50 -E 0.001 -z 1 -Z 32000 -B 0 -b 0 -n 2). Protein
605 families were generated from subfamilies with probability scores of at least 95% and coverage of
606 at least 0.70 using a Markov clustering algorithm[148]. Protein subfamilies were annotated based
607 on HMM-HMM comparisons against PFAM using HHsearch v.3.0.3 (parameters -p 50 -E 0.001 -
608 z 1 -Z 32000 -b 0 -B 0 -n 1)[149,150].

609 **Additional annotation of megaplasmid features and protein structure prediction**

610 Mobile genetic element prediction with geNomad v1.7.4 does not recognize the pMEGAs as
611 plasmids but predicts proviruses within the genome[151]. More extensive annotation was performed
612 for the pMEGA1-4 sequences, including prediction of secretion systems and conjugation
613 machinery using hmmscan against the CONJScan and TXSS hmm databases with an E value
614 cut-off of 1e-10[152,153]. Origins of transfer (*oriT*) and vegetative replication (*oriV*) were predicted by
615 oriTFinder2, oriTDB[34] and OriV-Finder[154]. Repeats were identified manually in Geneious Prime
616 using RepeatFinder. Virulence factors and toxins were predicted for pMEGA4 and the *E. coli*
617 genome using PathoFact v.2.0[155]. We used PADLOC v2.0.0 and structural comparisons to identify
618 antiviral defense systems and CRISPRCasFinder v.4.2.18 to detect CRISPR spacers[156,157].
619 Antibiotic resistance mechanisms in pMEGA4 were predicted from protein sequences using the
620 *main* function of the Resistance Gene Identifier (RGI) v.6.0.3 with CARD v4.0.0 (settings: --
621 include_loose, --include_nudge, --low_quality and reporting hits with a bitscore >100 regardless
622 of ARO cutoff)[158]. For the *E. coli* genome and its 104 kb plasmid, RGI was used in the default
623 mode with the --low_quality flag. Codon usage was calculated using coRdon v 1.26.0 in R v
624 4.4.3[159].

625 For pMEGAs 1-4, AlphaFold2 was used to predict the protein structure of all open reading frames
626 and compared to a database of folded proteins from PDB (retrieved 2023-04-20) using FoldSeek
627 (v427df8a6b5d0ef78bee0f98cd3e6faaca18f172d)[160,161]. Some structures, as well as multimers,
628 were predicted using AlphaFold3[160]. Multimer confidence was evaluated by consideration of ipTM
629 scores. The top-scoring hits to PDB structures (generally >200 bit score) were manually visualized
630 in ChimeraX (v1.9), and position-resolved pLDDT values were used to evaluate confidence.
631 Alignment was performed using matchmaker or TM-align to confirm structural similarity[162,163].
632 Foldseek matches to structures in the AlphaFold database were also taken into consideration.

633 Multi-copy protein subfamilies were defined as subfamilies with greater than three members within
634 at least one genome of pMEGA1-4. Multi-copy proteins and clusters of two or more proteins of
635 the same subfamily were identified in Microsoft Excel and with custom scripts in R v 4.4.3. The
636 predicted structures of members within a multi-copy protein subfamily were visualized to confirm
637 their predicted function and to identify structures that contain alpha-solenoid-like folds/domains.
638 Alpha-solenoid domains were identified by visual inspection, in some cases flagged by best
639 matches to DARPINs[42]. These proteins often had very low bit scores for best matches to proteins
640 in PDB.

641 **Phylogeny of plasmid relaxase and replication initiation proteins**

642 To generate a phylogeny of known $MOB_F$ relaxase proteins, sequences were gathered from two
643 comprehensive studies of conjugative plasmids and through a hmmscan for $MOB_F$ sequences in
644 PLSDB and IMG/PR, resulting in 1854 hits[33,153,164,165]. These were clustered at 99% identity using
645 vsearch v2.13.3 (--cluster_fast) and combined with the predicted $MOB_F$ relaxase sequences from
646 pMEGA1-4 (SF1156 and SF0489), resulting in 881 sequences[166]. These sequences were aligned
647 using MAFFT v7.505, then a phylogenetic tree was generated using FastTree v2.1.11, rooted at
648 the mid-point, and visualized using iTOL. A representative structure of SF0489 (pMEGA4_1394 -
649 median pLDDT 90.19) and SF1156 (pMEGA4_58 - median pLDDT 90.5) was aligned to RecD2
650 bound to ssDNA (PDB:3GP8)[167]. The TM alignment scores are 0.73 and 0.98, respectively.
651 Similarly, a phylogeny of 105 replication initiation proteins from plasmids of known incompatibility
652 groups (n = 65)[168] was generated with predicted Rep sequences from pMEGAs (SF1471 and
653 SF0640) after clustering at 90% identity. Rep sequences from megaplasmids identified in this
654 study were collapsed as clades. Select protein structures were compared to the pMEGA4
655 representative of SF640 and SF1471 using FoldSeek (bitscores as heatmaps on the right of the
656 tree). If not available in the AlphaFold database, the structures of Rep proteins were predicted
657 with AlphaFold3. For SF0640, pMEGA4_629 (median pLDDT 94.12) was aligned to RepC from

658 the *Salmonella enteritidis* IncQ plasmid pBLST (AFDB: A0A1D8X722) with TM-align (score 0.86).
659 For SF1471, pMEGA4_1423 (median pLDDT 91.5) was aligned to TrfA from the *Escherichia coli*
660 IncP plasmid pRK2 (AFDB: P07676) with TM-align (score 0.57). Finally, Rep protein structures
661 were compared to the Rep protein from E. coli F plasmid bound to an iteron sequence (PDB:
662 1REP). The TM-align score with 1REP for pMEGA4_629 was 0.68, and for pMEGA4_1423 was
663 0.70.

664 **Host prediction through abundance correlation, DNA modification analysis, and Hi-C**
665 **linkage**

666 From the SRA searches, we identified *E. coli* and *Salmonella spp.* isolates with evidence of a
667 megaplasmid. For one *E. coli* isolate, it was re-grown, DNA was re-isolated, and sequenced via
668 Nanopore, confirming the presence of the 1.4 Mb plasmid (pMEGA4). For infants with longitudinal
669 samples (pMEGA1, pMEGA2, pMEGA3), we measured the Pearson Correlation in R v. 4.4.3
670 between the percentage of reads mapped to the megaplasmid and all bacterial MAGs over
671 multiple sampling timepoints. For isolates with evidence of a medium- or high-quality
672 megaplasmid, plasmid and potential chromosomal contigs were separated and binned manually
673 based on GC content, coverage and similarity to pMEGA1 or 2. The megaplasmid copy number
674 (PCN) was estimated by comparing the mean depth of Illumina read coverage of the plasmid and
675 draft isolate genome as determined with CoverM v0.7.0 using BBMap v39.10[119,131] (Table S5).

676 To detect DNA methylation from PacBio data, we computed the mean inter-pulse duration (IPD)
677 at each base across the mapped reads. We empirically estimated baseline IPD values for each
678 k-mer by comparing to the mean IPD values of the kmer from the metagenome. Bases with
679 significantly elevated IPD ratios are inferred to be modified. Subsequently, motif discovery is
680 carried out using MotifMaker (SMRT Link v13.1.0.221970). We then scanned each contig to
681 evaluate the modified fraction of detected motifs, ultimately generating a genome-wide

682 methylation profile. In addition, to annotate restriction endonucleases (REases) and
683 methyltransferases (MTases) in the assemblies, we employed the annotate_rm module of
684 MicrobeMod with default parameters[169]. Finally, the ProxiMeta™ Metagenome Deconvolution
685 Platform (Hi-C) from Phase Genomics (Seattle, USA) was used to detect Hi-C linkages on the
686 week 4 stool samples from INF1330004 and INF1340011.

687 **Additional investigation into CRISPR spacers and possible hosts**

688 CRISPR spacers detected with CRISPRCasFinder v.4.2.18 from all 20 confident megaplasmid
689 genomes were searched against their respective metagenomes using blastn-short (-perc_identity
690 97, -qcov_hsp_perc 100)[132]. In addition, CRISPR spacers predicted within the 20 metagenomic
691 samples from which the plasmid genomes originated were searched against the megaplasmid
692 genomes to identify potential hosts. Finally, CRISPR spacers from the 20 megaplasmid genomes
693 were compared to PLSDB and IMG/PR and IMG/VR, representing over 772,529 plasmid and
694 15,722,824 viral sequences[21]. Spacer interaction networks were visualized in CytoScape v 3.10.3.
695 Finally, to infer potential hosts based on spacer acquisition against pMEGAs, we blasted over 322
696 million predicted spacers from 580,000 bacterial genomes[38] and ~450,000 public metagenomes[39]
697 against pMEGA1-20 and evaluated spacers with 100% identity hits.

698 **Comparing to other plasmid databases**

699 To ensure the megaplasmid elements had not been previously characterized, we searched for
700 similar elements in, to the best of our knowledge, all relevant and known plasmid and viral
701 sequence databases including PLSDB (v2024-05-31_v2 - 72,556 sequences), IMG/PR (699,973
702 sequences), IMG/VR (15,722,824 sequences), 226,194 plasmids predicted from the human gut
703 [129], as well as datasets from two recent studies that characterized plasmids (n = 265,476) and
704 viruses (n = 160,478) from infants[170,171] using MASH dist with pMEGA1 and pMEGA2[145]. To
705 ensure that no plasmids of similar size to the pMEGAs have been described in

706 *Enterobacteriaceae*, we analyzed all plasmid sequences >1Mb from PLSDB and IMG/PR, and
707 RefSeq *Enterobacteriaceae* genomes with secondary replicons >1Mb. Many of these plasmids
708 were derived from isolates. If a plasmid was derived from a metagenomic sample, we annotated
709 the taxonomy of the host based on ANI to other plasmids or the consensus taxonomy of predicted
710 proteins. Finally, to prevent inclusion of mis-annotated plasmids or secondary chromosomes, we
711 used CheckM2 to remove sequences with 50% or greater bacterial genome completeness[172].
712 Predicted proteins were annotated following the same basic pipeline for pMEGAs, and the
713 percentage of hypothetical, unknown, or uncharacterized proteins as determined by UniRef
714 annotations was determined for both the largest contig of each isolate assembly and the putative
715 >1Mb plasmid.

716 **References:**


717 1. Pilla, G. & Tang, C. M. Going around in circles: virulence plasmids in enteric pathogens.
718 *Nat. Rev. Microbiol.* **16**, 484–495 (2018).
719 2. Nora, L. C. *et al.* The art of vector engineering: towards the construction of next-generation
720 genetic tools. *Microb. Biotechnol.* **12**, 125–147 (2019).
721 3. Stockdale, S. R. & Hill, C. Incorporating plasmid biology and metagenomics into a holistic
722 model of the human gut microbiome. *Curr. Opin. Microbiol.* **73**, 102307 (2023).
723 4. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San Millán,
724 Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev.*
725 *Microbiol.* **19**, 347–359 (2021).
726 5. diCenzo, G. C. & Finan, T. M. The divided bacterial genome: Structure, function, and
727 evolution. *Microbiol. Mol. Biol. Rev.* **81**, (2017).
728 6. Maddamsetti, R. *et al.* Scaling laws of bacterial and archaeal plasmids. *Nat. Commun.* **16**,
729 (2025).
730 7. Ramiro-Martínez, P., de Quinto, I., Lanza, V. F., Gama, J. A. & Rodríguez-Beltrán, J.

731 Universal rules govern plasmid copy number. *Nat. Commun.* **16**, (2025).
732 8. Czarnecki, J. *et al.* Replication coordination marks the domestication of large
733 extrachromosomal replicons in bacteria. *bioRxiv* (2025) doi:10.1101/2025.03.15.643453.
734 9. Genome sequence of the plant pathogen Ralstonia solanacearum. *Nature* (2002)
735 doi:10.1038/news020128-8.
736 10. Janssen, P. J. *et al.* The complete genome sequence of Cupriavidus metallidurans strain
737 CH34, a master survivalist in harsh and anthropogenic environments. *PLoS One* **5**, e10433
738 (2010).
739 11. Mergeay, M. & Van Houdt, R. Cupriavidus metallidurans CH34, a historical perspective on
740 its discovery, characterization and metal resistance. *FEMS Microbiol. Ecol.* **97**, (2021).
741 12. diCenzo, G. C., Mengoni, A. & Perrin, E. Chromids aid genome expansion and functional
742 diversification in the family Burkholderiaceae. *Mol. Biol. Evol.* **36**, 562–574 (2019).
743 13. Fournes, F., Val, M.-E., Skovgaard, O. & Mazel, D. Replicate once per cell cycle:
744 Replication control of secondary chromosomes. *Front. Microbiol.* **9**, 1833 (2018).
745 14. Harrison, P. W., Lower, R. P. J., Kim, N. K. D. & Young, J. P. W. Introducing the bacterial
746 'chromid': not a chromosome, not a plasmid. *Trends Microbiol.* **18**, 141–148 (2010).
747 15. Hall, J. P. J., Botelho, J., Cazares, A. & Baltrus, D. A. What makes a megaplasmid? *Philos.*
748 *Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200472 (2022).
749 16. Liu, H. *et al.* Unexplored diversity and potential functions of extra-chromosomal elements.
750 *mSystems* e0017525 (2025).
751 17. Kohlmeier, M. G., O'Hara, G. W., Ramsay, J. P. & Terpolilli, J. J. Closed genomes of
752 commercial inoculant rhizobia provide a blueprint for management of legume inoculation.
753 *Appl. Environ. Microbiol.* **91**, e0221324 (2025).
754 18. Medema, M. H. *et al.* The sequence of a 1.8-mb bacterial linear plasmid reveals a rich
755 evolutionary reservoir of secondary metabolic pathways. *Genome Biol. Evol.* **2**, 212–224
756 (2010).

757 19. Kothari, A. *et al.* Large circular plasmids from groundwater plasmidomes span multiple
758 incompatibility groups and are enriched in multimetal resistance genes. *MBio* **10**, (2019).
759 20. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a
760 global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).
761 21. Schmartz, G. P. *et al.* PLSDB: advancing a comprehensive database of bacterial plasmids.
762 *Nucleic Acids Res.* **50**, D273–D278 (2022).
763 22. Camargo, A. P. *et al.* IMG/PR: a database of plasmids from genomes and metagenomes
764 with rich annotations and metadata. *Nucleic Acids Res.* **52**, D164–D173 (2024).
765 23. Li, Y. *et al.* PlasmidScope: a comprehensive plasmid database with rich annotations and
766 online analytical tools. *Nucleic Acids Res.* **53**, D179–D188 (2025).
767 24. Douarre, P.-E., Mallet, L., Radomski, N., Felten, A. & Mistou, M.-Y. Analysis of COMPASS,
768 a new comprehensive Plasmid database revealed prevalence of multireplicon and
769 extensive diversity of IncF plasmids. *Front. Microbiol.* **11**, 483 (2020).
770 25. Koraimann, G. Spread and persistence of virulence and antibiotic resistance genes: A ride
771 on the F Plasmid conjugation module. *EcoSal Plus* **8**, (2018).
772 26. Ares-Arroyo, M., Rocha, E. P. C. & Gonzalez-Zorn, B. Evolution of ColE1-like plasmids
773 across γ-Proteobacteria: From bacteriocin production to antimicrobial resistance. *PLoS*
774 *Genet.* **17**, (2021).
775 27. Arredondo-Alonso, S. *et al.* Plasmid-driven strategies for clone success in Escherichia coli.
776 *Nat. Commun.* **16**, 2921 (2025).
777 28. Brooks, B. *et al.* Strain-resolved analysis of hospital rooms and infants reveals overlap
778 between the human and room microbiome. *Nat. Commun.* **8**, 1814 (2017).
779 29. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-
780 section birth. *Nature* **574**, 117–121 (2019).
781 30. Chen, Y. *et al.* Draft Genome Sequences of Citrobacter freundii and Citrobacter murliniae
782 Strains Isolated from the Feces of Preterm Infants. *Microbiol. Resour. Announc.* **8**, (2019).

783 31. Sudatip, D. *et al.* The risk of pig and chicken farming for carriage and transmission of
784 Escherichia coli containing extended-spectrum beta-lactamase (ESBL) and mobile colistin
785 resistance (mcr) genes in Thailand. *Microb. Genom.* **9**, (2023).
786 32. Mishra, D. & Srinivasan, R. Catching a Walker in the act-DNA partitioning by ParA family of
787 proteins. *Front. Microbiol.* **13**, 856547 (2022).
788 33. Coluzzi, C., Garcillán-Barcia, M. P., de la Cruz, F. & Rocha, E. P. C. Evolution of Plasmid
789 mobility: Origin and fate of conjugative and nonconjugative plasmids. *Mol. Biol. Evol.* **39**,
790 (2022).
791 34. Liu, G. *et al.* oriTDB: a database of the origin-of-transfer regions of bacterial mobile genetic
792 elements. *Nucleic Acids Res.* **53**, D163–D168 (2025).
793 35. Konieczny, I., Bury, K., Wawrzycka, A. & Wegrzyn, K. Iteron plasmids. *Microbiol. Spectr.* **2**,
794 (2014).
795 36. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility
796 of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
797 37. Pausch, P. *et al.* Structural variation of type I-F CRISPR RNA guided DNA surveillance.
798 *Mol. Cell* **67**, 622–632.e4 (2017).
799 38. Dion, M. B. *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher
800 the viral dark matter. *Nucleic Acids Res.* **49**, 3127–3138 (2021).
801 39. Roux, S. *et al.* Planetary-scale metagenomic search reveals new patterns of CRISPR
802 targeting. *bioRxiv* (2025) doi:10.1101/2025.06.12.659409.
803 40. Li, R., Peng, K., Li, Y., Liu, Y. & Wang, Z. Exploring tet(X)-bearing tigecycline-resistant
804 bacteria of swine farming environments. *Sci. Total Environ.* **733**, 139306 (2020).
805 41. Riccardi, C. *et al.* Independent origins and evolution of the secondary replicons of the class
806 Gammaproteobacteria. *Microb. Genom.* **9**, (2023).
807 42. Plückthun, A. Designed ankyrin repeat proteins (DARPins): binding proteins for research,
808 diagnostics, and therapy. *Annu. Rev. Pharmacol. Toxicol.* **55**, 489–511 (2015).

809 43. Gong, Y. *et al.* A nuclease-toxin and immunity system for kin discrimination in Myxococcus
810 xanthus. *Environ. Microbiol.* **20**, 2552–2567 (2018).
811 44. Zhang, D., Iyer, L. M. & Aravind, L. A novel immunity system for bacterial nucleic acid
812 degrading toxins and its recruitment in various eukaryotic and DNA viral systems. *Nucleic*
813 *Acids Res.* **39**, 4532–4552 (2011).
814 45. Zhang, D., de Souza, R. F., Anantharaman, V., Iyer, L. M. & Aravind, L. Polymorphic toxin
815 systems: Comprehensive characterization of trafficking modes, processing, mechanisms of
816 action, immunity and ecology using comparative genomics. *Biol. Direct* **7**, 18 (2012).
817 46. Ruhe, Z. C., Low, D. A. & Hayes, C. S. Polymorphic toxins and their immunity proteins:
818 Diversity, evolution, and mechanisms of delivery. *Annu. Rev. Microbiol.* **74**, 497–520
819 (2020).
820 47. Liu, W., Schoonen, M., Wang, T., McSweeney, S. & Liu, Q. Cryo-EM structure of
821 transmembrane AAA+ protease FtsH in the ADP state. *Commun. Biol.* **5**, 257 (2022).
822 48. Bittner, L.-M., Arends, J. & Narberhaus, F. When, how and why? Regulated proteolysis by
823 the essential FtsH protease in Escherichia coli. *Biol. Chem.* **398**, 625–635 (2017).
824 49. Erbse, A. *et al.* ClpS is an essential component of the N-end rule pathway in Escherichia
825 coli. *Nature* **439**, 753–756 (2006).
826 50. Deville, C., Franke, K., Mogk, A., Bukau, B. & Saibil, H. R. Two-step activation mechanism
827 of the ClpB disaggregase for sequential substrate threading by the main ATPase motor.
828 *Cell Rep.* **27**, 3433–3446.e4 (2019).
829 51. Ma, C. *et al.* Structural insights into the membrane microdomain organization by SPFH
830 family proteins. *Cell Res.* **32**, 176–189 (2022).
831 52. Lopez, D. & Koch, G. Exploring functional membrane microdomains in bacteria: an
832 overview. *Curr. Opin. Microbiol.* **36**, 76–84 (2017).
833 53. Wallden, K., Rivera-Calzada, A. & Waksman, G. Type IV secretion systems: versatility and
834 diversity in function. *Cell. Microbiol.* **12**, 1203–1212 (2010).

835 54. Cascales, E. & Cambillau, C. Structural biology of type VI secretion systems. *Philos. Trans.*
836 *R. Soc. Lond. B Biol. Sci.* **367**, 1102–1111 (2012).
837 55. Hu, J. *et al.* Cryo-EM analysis of the T3S injectisome reveals the structure of the needle
838 and open secretin. *Nat. Commun.* **9**, 3840 (2018).
839 56. Hajam, I. A., Dar, P. A., Shahnawaz, I., Jaume, J. C. & Lee, J. H. Bacterial flagellin-a potent
840 immunomodulatory agent. *Exp. Mol. Med.* **49**, e373 (2017).
841 57. Haiko, J. & Westerlund-Wikström, B. The role of the bacterial flagellum in adhesion and
842 virulence. *Biology (Basel)* **2**, 1242–1267 (2013).
843 58. Caveney, N. A. *et al.* Structural insight into YcbB-mediated beta-lactam resistance in
844 Escherichia coli. *Nat. Commun.* **10**, 1849 (2019).
845 59. Ronneau, S. & Hallez, R. Make and break the alarmone: regulation of (p)ppGpp
846 synthetase/hydrolase enzymes in bacteria. *FEMS Microbiol. Rev.* **43**, 389–400 (2019).
847 60. Calhoun, L. N. & Kwon, Y. M. Structure, function and regulation of the DNA-binding protein
848 Dps and its role in acid and oxidative stress resistance in Escherichia coli: a review. *J. Appl.*
849 *Microbiol.* **110**, 375–386 (2011).
850 61. Janssens, A. *et al.* SlyB encapsulates outer membrane proteins in stress-induced lipid
851 nanodomains. *Nature* **626**, 617–625 (2024).
852 62. Srikannathasan, V. *et al.* Structural basis for type VI secreted peptidoglycan DL-
853 endopeptidase function, specificity and neutralization in Serratia marcescens. *Acta*
854 *Crystallogr. D Biol. Crystallogr.* **69**, 2468–2482 (2013).
855 63. Ting, S.-Y. *et al.* Bifunctional immunity proteins protect bacteria against FtsZ-targeting
856 ADP-ribosylating toxins. *Cell* **175**, 1380–1392.e14 (2018).
857 64. Bishop, R. E., Leskiw, B. K., Hodges, R. S., Kay, C. M. & Weiner, J. H. The entericidin
858 locus of Escherichia coli and its implications for programmed bacterial cell death. *J. Mol.*
859 *Biol.* **280**, 583–596 (1998).
860 65. Schmidt, O. *et al.* prlF and yhaV encode a new toxin-antitoxin system in Escherichia coli. *J.*

861 *Mol. Biol.* **372**, 894–905 (2007).
862 66. Hansen, S. *et al.* Regulation of the Escherichia coli HipBA toxin-antitoxin system by
863 proteolysis. *PLoS One* **7**, e39185 (2012).
864 67. Tripathi, A., Dewan, P. C., Ahmed, S. & Varadarajan, R. MazF-induced growth inhibition
865 and persister generation in Escherichia coli. *J. Biol. Chem.* **289**, 4191–4205 (2014).
866 68. May, J. P. & Simon, A. E. Targeting of viral RNAs by Upf1-mediated RNA decay pathways.
867 *Curr. Opin. Virol.* **47**, 1–8 (2021).
868 69. Millman, A. *et al.* An expanding arsenal of immune systems that protect bacteria from
869 phages. *bioRxiv* (2022) doi:10.1101/2022.05.11.491447.
870 70. Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C. & Laub, M. T. A
871 functional selection reveals previously undetected anti-phage defence systems in the E. coli
872 pangenome. *Nat. Microbiol.* **7**, 1568–1579 (2022).
873 71. Gao, L. *et al.* Diverse enzymatic activities mediate antiviral immunity in prokaryotes.
874 *Science* **369**, 1077–1084 (2020).
875 72. Fitzpatrick, A. W. P. *et al.* Structure of the MacAB–TolC ABC-type tripartite multidrug efflux
876 pump. *Nat. Microbiol.* **2**, 17070 (2017).
877 73. San Millan, A. & MacLean, R. C. Fitness costs of plasmids: A limit to Plasmid transmission.
878 *Microbiol. Spectr.* **5**, (2017).
879 74. Kamal, S. M. *et al.* A recently isolated human commensal Escherichia coli ST10 clone
880 member mediates enhanced thermotolerance and tetrathionate respiration on a P1 phage-
881 derived IncY plasmid. *Mol. Microbiol.* **115**, 255–271 (2021).
882 75. Lee, C. *et al.* Why? - Successful Pseudomonas aeruginosa clones with a focus on clone C.
883 *FEMS Microbiol. Rev.* **44**, 740–762 (2020).
884 76. Bojer, M. S., Struve, C., Ingmer, H., Hansen, D. S. & Krogfelt, K. A. Heat resistance
885 mediated by a new plasmid encoded Clp ATPase, ClpK, as a possible novel mechanism for
886 nosocomial persistence of Klebsiella pneumoniae. *PLoS One* **5**, e15467 (2010).

887 77. Motiwala, T. *et al.* Caseinolytic proteins (Clp) in the genus Klebsiella: Special focus on
888 ClpK. *Molecules* **27**, 200 (2021).
889 78. Kamal, S. M. *et al.* Two FtsH proteases contribute to fitness and adaptation of
890 Pseudomonas aeruginosa clone C strains. *Front. Microbiol.* **10**, 1372 (2019).
891 79. Mawla, G. D. *et al.* The membrane-cytoplasmic linker defines activity of FtsH proteases in
892 Pseudomonas aeruginosa clone C. *J. Biol. Chem.* **300**, 105622 (2024).
893 80. Stein, B. J., Grant, R. A., Sauer, R. T. & Baker, T. A. Structural basis of an N-degron
894 adaptor with more stringent specificity. *Structure* **24**, 232–242 (2016).
895 81. Qiao, Z. *et al.* Cryo-EM structure of the entire FtsH-HflKC AAA protease complex. *Cell Rep.*
896 **39**, 110890 (2022).
897 82. Choi, M., Karunaratne, K. & Kohen, A. Flavin-dependent thymidylate synthase as a new
898 antibiotic target. *Molecules* **21**, 654 (2016).
899 83. *Bacterial Enzymes as Targets for Drug Discovery*. (Academic Press, San Diego, CA, 2024).
900 84. Zhao, J. *et al.* Preclinical safety and efficacy characterization of an LpxC inhibitor against
901 Gram-negative pathogens. *Sci. Transl. Med.* **15**, eadf5668 (2023).
902 85. Osterman, I. *et al.* Phages reconstitute NAD+ to counter bacterial immunity. *Nature* **634**,
903 1160–1167 (2024).
904 86. Kaufmann, G. *et al.* RloC: A translation-disabling tRNase implicated in phage exclusion
905 during recovery from DNA damage. in *DNA Repair* (InTech, 2011).
906 87. van den Berg, D. F., van der Steen, B. A., Costa, A. R. & Brouns, S. J. J. Phage tRNAs
907 evade tRNA-targeting host defenses through anticodon loop mutations. *Elife* **12**, (2023).
908 88. Korotkov, K. V., Gonen, T. & Hol, W. G. J. Secretins: dynamic channels for protein
909 transport across membranes. *Trends Biochem. Sci.* **36**, 433–443 (2011).
910 89. Leighton, T. L., Buensuceso, R. N. C., Howell, P. L. & Burrows, L. L. Biogenesis of
911 Pseudomonas aeruginosa type IV pili and regulation of their function. *Environ. Microbiol.*
912 **17**, 4148–4163 (2015).

913 90. Hoffmann, L. *et al.* Structure and interactions of the archaeal motility repression module
914 ArnA-ArnB that modulates archaellum gene expression in Sulfolobus acidocaldarius. *J.*
915 *Biol. Chem.* **294**, 7460–7471 (2019).
916 91. Heidler, T. V., Ernits, K., Ziolkowska, A., Claesson, R. & Persson, K. Porphyromonas
917 gingivalis fimbrial protein Mfa5 contains a von Willebrand factor domain and an
918 intramolecular isopeptide. *Commun. Biol.* **4**, 106 (2021).
919 92. Patel, D. T. *et al.* Global atlas of predicted functional domains in Legionella pneumophila
920 Dot/Icm translocated effectors. *Mol. Syst. Biol.* **21**, 59–89 (2025).
921 93. Al-Khodor, S., Price, C. T., Habyarimana, F., Kalia, A. & Abu Kwaik, Y. A Dot/Icm-
922 translocated ankyrin protein of Legionella pneumophila is required for intracellular
923 proliferation within human macrophages and protozoa. *Mol. Microbiol.* **70**, 908–923 (2008).
924 94. Schmitz-Esser, S. *et al.* The genome of the amoeba symbiont 'Candidatus Amoebophilus
925 asiaticus' reveals common mechanisms for host cell interaction among amoeba-associated
926 bacteria. *J. Bacteriol.* **192**, 1045–1057 (2010).
927 95. Walser, M., Mayor, J. & Rothenberger, S. Designed ankyrin repeat proteins: A new class of
928 viral entry inhibitors. *Viruses* **14**, 2242 (2022).
929 96. Al-Khodor, S., Price, C. T., Kalia, A. & Abu Kwaik, Y. Functional diversity of ankyrin repeats
930 in microbial proteins. *Trends Microbiol.* **18**, 132–139 (2010).
931 97. Self, J. L. *et al.* Epidemiology of salmonellosis among infants in the United States: 1968-
932 2015. *Pediatrics* **151**, (2023).
933 98. Hoffman, A. *et al.* Predictors of mortality and severe illness from Escherichia coli sepsis in
934 neonates. *J. Perinatol.* **44**, 1816–1821 (2024).
935 99. Yang, J.-T., Zhang, L.-J., Lu, Y., Zhang, R.-M. & Jiang, H.-X. Genomic insights into global
936 blaCTX-M-55-positive Escherichia coli epidemiology and transmission characteristics.
937 *Microbiol. Spectr.* **11**, e0108923 (2023).
938 100. Day, M. J. *et al.* Extended-spectrum β-lactamase-producing Escherichia coli in human-

939 derived and foodchain-derived samples from England, Wales, and Scotland: an
940 epidemiological surveillance and typing study. *Lancet Infect. Dis.* **19**, 1325–1335 (2019).
941 101. Falgenhauer, L. *et al.* Detection and characterization of ESBL-producing Escherichia coli
942 from humans and poultry in Ghana. *Front. Microbiol.* **9**, 3358 (2018).
943 102. Lupo, A., Saras, E., Madec, J.-Y. & Haenni, M. Emergence of blaCTX-M-55 associated with
944 fosA, rmtB and mcr gene variants in Escherichia coli from various animal species in France.
945 *J. Antimicrob. Chemother.* **73**, 867–872 (2018).
946 103. Hu, Y.-Y. *et al.* Molecular typing of CTX-M-producing escherichia coli isolates from
947 environmental water, swine feces, specimens from healthy humans, and human patients.
948 *Appl. Environ. Microbiol.* **79**, 5988–5996 (2013).
949 104. Lim, J. Y., Yoon, J. & Hovde, C. J. A brief overview of Escherichia coli O157:H7 and its
950 plasmid O157. *J. Microbiol. Biotechnol.* **20**, 5–14 (2010).
951 105. Chen, J. C. *et al.* Reoccurring Escherichia coli O157:H7 strain linked to leafy greens-
952 associated outbreaks, 2016-2019. *Emerg. Infect. Dis.* **29**, 1895–1899 (2023).
953 106. Kizny Gordon, A. E. *et al.* The hospital water environment as a reservoir for carbapenem-
954 resistant organisms causing hospital-acquired infections-A systematic review of the
955 literature. *Clin. Infect. Dis.* **64**, 1435–1444 (2017).
956 107. Morowitz, M. J. *et al.* The NICU Antibiotics and Outcomes (NANO) trial: a randomized
957 multicenter clinical trial assessing empiric antibiotics and clinical outcomes in newborn
958 preterm infants. *Trials* **23**, 428 (2022).
959 108. Joshi NA, F. J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ
960 files (Version 1.33) [Software]. *https://github.com/najoshi/sickle* (2011).
961 109. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
962 **9**, 357–359 (2012).
963 110. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
964 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

965 111. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
966 identification. *BMC Bioinformatics* **11**, 119 (2010).
967 112. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
968 *Res.* **28**, 27–30 (2000).
969 113. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive
970 and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
971 114. UniProt Consortium. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids*
972 *Res.* **51**, D523–D531 (2023).
973 115. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*
974 **26**, 2460–2461 (2010).
975 116. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain
976 Bacteria. *Nature* **523**, 208–211 (2015).
977 117. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments.
978 *Bioinformatics* **25**, 1713–1713 (2009).
979 118. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA
980 genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
981 119. Bushnell, B. *BBMap*.
982 120. Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads
983 with hifiasm-meta. *Nat. Methods* **19**, 671–674 (2022).
984 121. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication,
985 aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
986 122. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to
987 recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
988 123. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat.*
989 *Methods* **11**, 1144–1146 (2014).
990 124. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational

991 autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
992 125. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient
993 genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
994 126. Portik, D. M. *et al.* Highly accurate metagenome-assembled genomes from human gut
995 microbiota using long-read assembly, binning, and consolidation methods. *bioRxiv* (2024)
996 doi:10.1101/2024.05.10.593587.
997 127. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate
998 genomic comparisons that enables improved genome recovery from metagenomes through
999 de-replication. *ISME J.* **11**, 2864–2868 (2017).
1000 128. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from
1001 metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
1002 129. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their ecology across
1003 human gut metagenomes revealed by PlasX and MobMess. *Nat. Microbiol.* **9**, 830–847
1004 (2024).
1005 130. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and
1006 complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
1007 131. Aroney, S. T. N. *et al.* CoverM: read alignment statistics for metagenomics. *Bioinformatics*
1008 **41**, (2025).
1009 132. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
1010 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
1011 133. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology
1012 Information. Protein [Internet]. (1988).
1013 134. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology
1014 Information in 2025. *Nucleic Acids Res.* **53**, D20–D29 (2025).
1015 135. Sudatip, D. *et al.* A One Health approach to assessing occupational exposure to
1016 antimicrobial resistance in Thailand: The FarmResist project. *PLoS One* **16**, e0245250

1017 (2021).
1018 136. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
1019 3100 (2018).
1020 137. Nanoporetech. *Dorado*. (2024).
1021 138. Shiryev, S. A. & Agarwala, R. Indexing and searching petabase-scale nucleotide resources.
1022 *Nat. Methods* **21**, 994–1002 (2024).
1023 139. Chikhi, R., Raffestin, B., Korobeynikov, A., Edgar, R. & Babaian, A. Logan: Planetary-scale
1024 genome assembly surveys life's diversity. *bioRxiv* (2024) doi:10.1101/2024.07.30.605881.
1025 140. Constantinides, B. *et al.* Genomic surveillance of Escherichia coli and Klebsiella spp. in
1026 hospital sink drains and patients. *Microb. Genom.* **6**, (2020).
1027 141. Zhang, A. *et al.* Infants exposed in utero to Hurricane Maria have gut microbiomes with
1028 reduced diversity and altered metabolic capacity. *mSphere* **8**, e0013423 (2023).
1029 142. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1030 2079 (2009).
1031 143. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of
1032 conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
1033 144. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse
1034 chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
1035 145. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using
1036 MinHash. *Genome Biol.* **17**, (2016).
1037 146. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
1038 analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
1039 147. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein
1040 annotation. *BMC Bioinformatics* **20**, 473 (2019).
1041 148. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale
1042 detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

1043 149. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–
1044 960 (2005).
1045 150. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–
1046 D419 (2021).
1047 151. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat.*
1048 *Biotechnol.* **42**, 1303–1312 (2024).
1049 152. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.*
1050 **6**, 23080 (2016).
1051 153. Cury, J., Abby, S. S., Doppelt-Azeroual, O., Néron, B. & Rocha, E. P. C. Identifying
1052 conjugative plasmids and integrative conjugative elements with CONJscan. *Methods Mol.*
1053 *Biol.* **2075**, 265–283 (2020).
1054 154. Li, Y. & Gao, F. OriV-Finder: a comprehensive web server for bacterial plasmid replication
1055 origin analysis. *Nucleic Acids Res.* **53**, W451–W456 (2025).
1056 155. Sunyer, J. O. *et al.* PathoFact 2.0: An integrative pipeline for predicting antimicrobial
1057 resistance genes, virulence factors, toxins and biosynthetic gene clusters in metagenomes.
1058 *bioRxiv* (2024) doi:10.1101/2024.12.09.627531.
1059 156. Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable
1060 version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.*
1061 **46**, W246–W251 (2018).
1062 157. Payne, L. J. *et al.* Identification and classification of antiviral defence systems in bacteria
1063 and archaea with PADLOC reveals new system types. *Nucleic Acids Res.* **49**, 10868–
1064 10878 (2021).
1065 158. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and
1066 resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids*
1067 *Res.* **51**, D690–D699 (2023).
1068 159. Vlahovicek, A. E. M. K. *coRdon*. (Bioconductor, 2018). doi:10.18129/B9.BIOC.CORDON.

1069 160. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with
1070 AlphaFold 3. *Nature* **630**, 493–500 (2024).
1071 161. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat.*
1072 *Biotechnol.* **42**, 243–246 (2024).
1073 162. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.*
1074 **32**, e4792 (2023).
1075 163. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-
1076 score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
1077 164. Garcillán-Barcia, M. P., Francia, M. V. & de la Cruz, F. The diversity of conjugative
1078 relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* **33**, 657–687
1079 (2009).
1080 165. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
1081 166. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open
1082 source tool for metagenomics. *PeerJ* (2016) doi:10.7287/peerj.preprints.2409.
1083 167. Saikrishnan, K., Powell, B., Cook, N. J., Webb, M. R. & Wigley, D. B. Mechanistic basis of
1084 5'-3' translocation in SF1B helicases. *Cell* **137**, 849–859 (2009).
1085 168. Suzuki, H., Yano, H., Brown, C. J. & Top, E. M. Predicting plasmid promiscuity based on
1086 genomic signature. *J. Bacteriol.* **192**, 6045–6055 (2010).
1087 169. Crits-Christoph, A., Kang, S. C., Lee, H. H. & Ostrov, N. MicrobeMod: A computational
1088 toolkit for identifying prokaryotic methylation and restriction-modification with nanopore
1089 sequencing. *bioRxiv* (2023) doi:10.1101/2023.11.13.566931.
1090 170. He, W., Russel, J., Klincke, F., Nesme, J. & Sørensen, S. J. Insights into the ecology of the
1091 infant gut plasmidome. *Nat. Commun.* **15**, 6924 (2024).
1092 171. Zeng, S. *et al.* A metagenomic catalog of the early-life human gut virome. *Nat. Commun.*
1093 **15**, 1864 (2024).
1094 172. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable

1095 and accurate tool for assessing microbial genome quality using machine learning. *Nat.*
1096 *Methods* **20**, 1203–1212 (2023).
1097 173. Kreutzberger, M. A. B. *et al.* Convergent evolution in the supercoiling of prokaryotic flagellar
1098 filaments. *Cell* **185**, 3487–3500.e14 (2022).

1099 ## TABLES AND FIGURES

1100 **Table 1**: Listing of 4 megaplasmids (name, size, GC content, source). Completion was
1101 determined by mapping reads to a single contig and ensuring complete coverage as well as
1102 overlapping read coverage at the ends of the contig.

| Name | Size | GC content | Source | ORFs | tRNAs | Accession (sequencing technology) | Status |
|---|---|---|---|---|---|---|---|
| pMEGA1 | 1,583,849 | 26.11 | Preterm infant 1340011 8 weeks (USA) | 1633 | 119 | This study (Illumina + PacBio) | Complete & circular |
| pMEGA2 | 1,334,733 | 29.36 | Preterm infant 1330004 4 weeks (USA) | 1681 | 204 | This study (Illumina + PacBio) | Complete & circular |
| pMEGA3 | 1,373,843 | 29.57 | Preterm infant 1340011 4 weeks (USA) | 1691 | 170 | This study (PacBio) | Circular & confident |
| pMEGA4 | 1,415,043 | 25.98 | *E. coli* isolate from chicken rectal swab (Thailand) | 1483 | 108 | This study (Illumina + Nanopore) | Complete & circular |

1103
1104

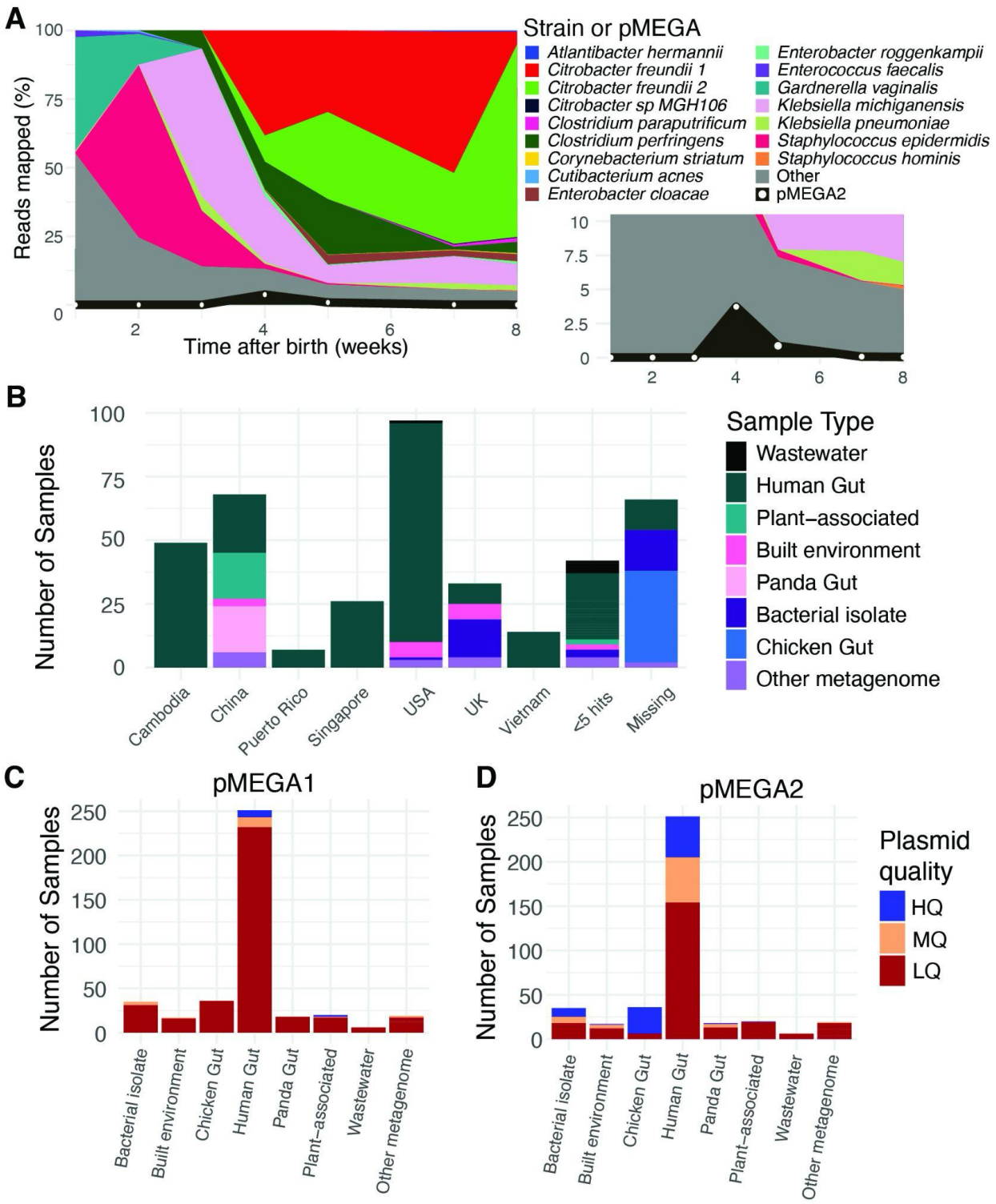



**Figure 1: Identification of megaplasmids in the preterm infant gut microbiome and public data. A)** The percentage of reads mapped to MAGs and pMEGA2 within INF1330004 was determined through Illumina read mapping. The "Other" category includes reads that did not map

to MAGs nor pMEGA2. The inset shows the same data, but with y-values ranging from 0 to 10. **B)** Sources of samples with evidence of related megaplasmids, grouped by source location and sample type (Missing indicates no information in the SRA). Source types with fewer than 5 hits were grouped together. **C), D)** The number and quality of pMEGA1(**C**)- and pMEGA2(**D**)-like plasmids across different sample types. Assembled plasmids were classified as high-quality (HQ - blue) if they had ≥85% coverage of a pMEGA and ≤10 contigs and medium-quality (MQ - orange) with ≥85% coverage and 11-50 contigs. The remaining samples were classified as low-quality (LQ - red).

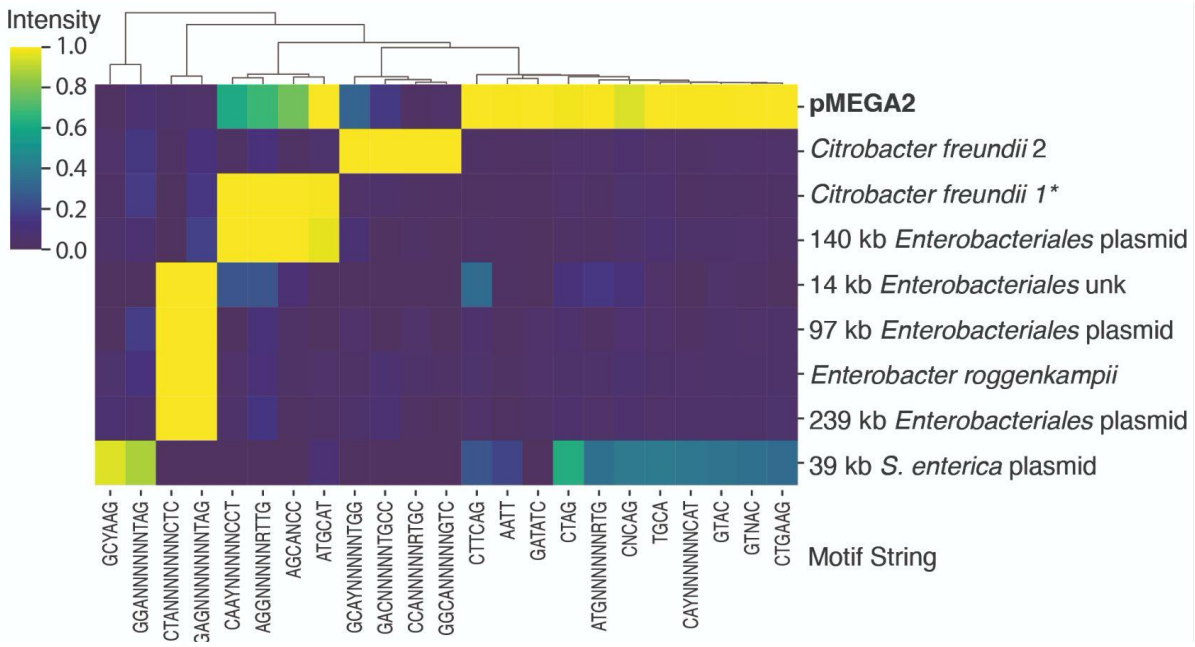



**Figure 2: A *Citrobacter freundii* is the host of pMEGA2.** Modified bases were predicted in the PacBio HiFi reads of infant 1330004 at week 4. The columns represent the motifs predicted to contain modified bases, and the rows represent genomes. *Likely host of pMEGA2 based on overlap in modification patterns.

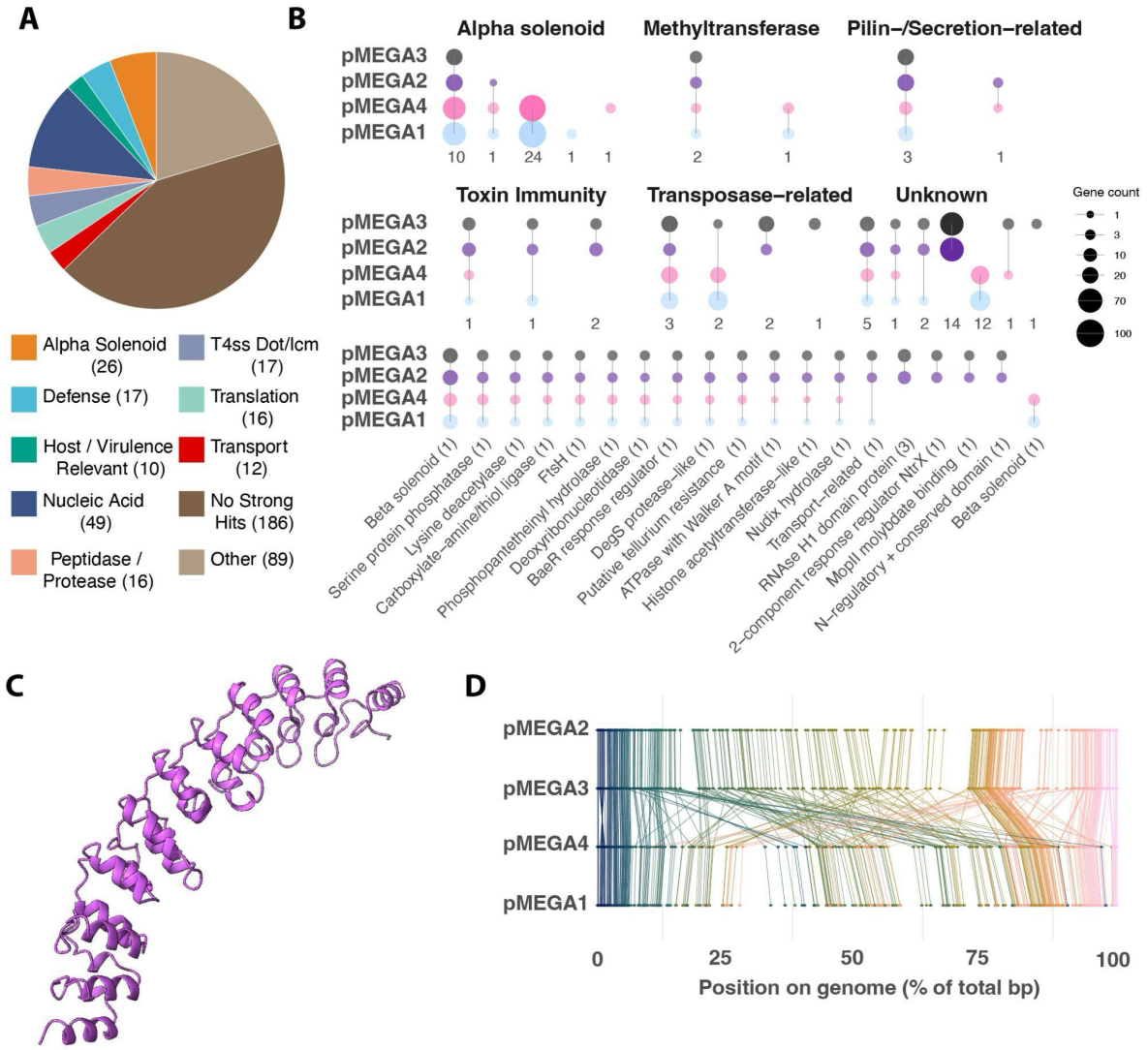



**Figure 3: Conserved protein functions across two clades of megaplasmids. A)** Distribution of functional annotations of the 438 conserved protein subfamilies across pMEGAs1-20. Annotations with fewer than 10 instances were grouped into the "Other" category. Numbers represent the number of protein families in each category. **B)** Multicopy protein subfamilies across pMEGA1-4. Circle size corresponds to the number of genes within each category. The number below each column or at the end of each annotation in the Other row indicates the number of subfamilies in that category. **C)** Examples of alpha solenoid-containing protein pMEGA4_362 (median pLDDT 95.06) whose domain is annotated as an Ankyrin repeat. **D)** Alignment of

pMEGA1-4 showing the locations of the 336 single-copy protein subfamilies conserved across pMEGA1-20 as a percentage of the total length of the genome. Each protein subfamily is colored based on its gene order in pMEGA2.

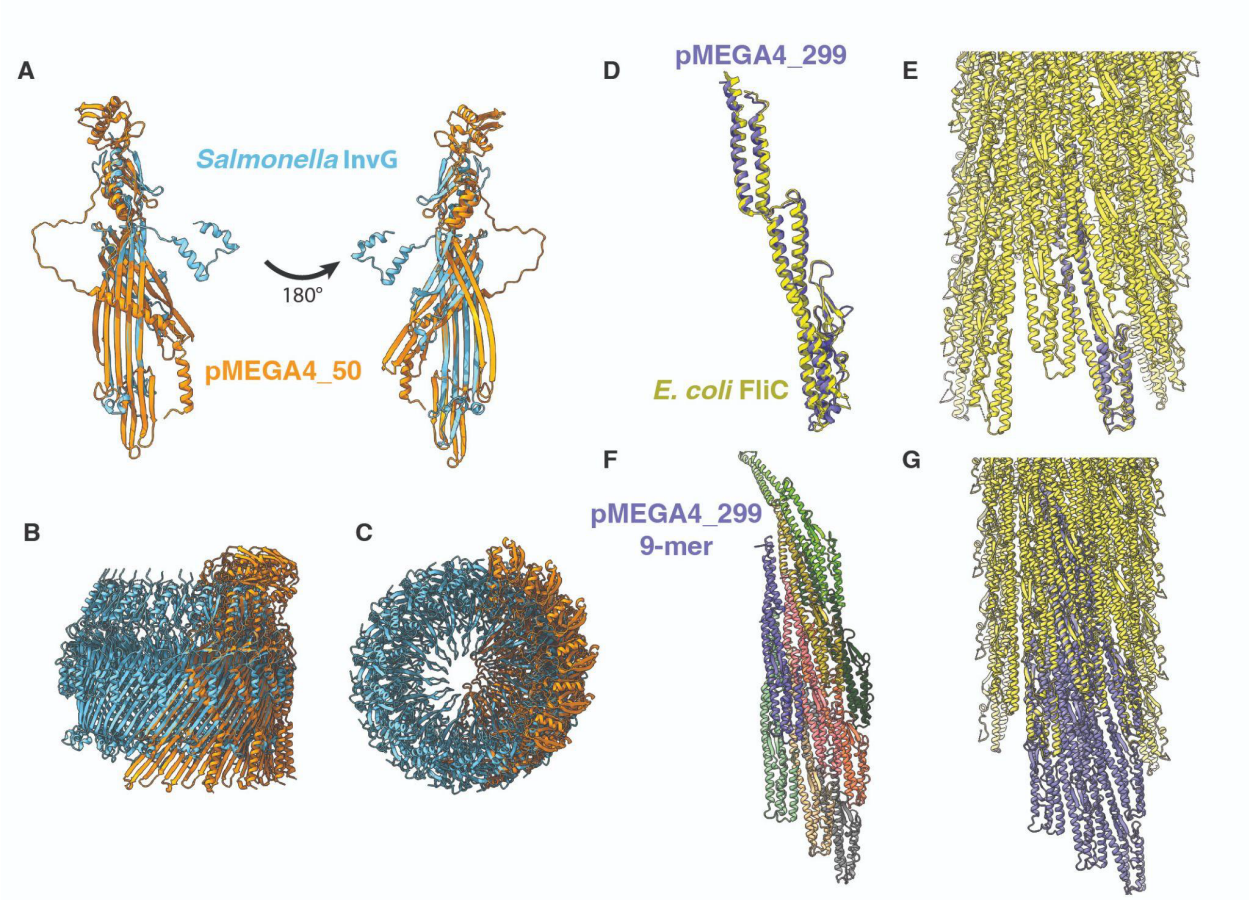



**Figure 4: pMEGA4 encodes mechanisms related to secretion and virulence. A)** Predicted structure of a monomer of pMEGA4_50 (median pLDDT 82.0) aligned to *Salmonella* pathogenicity island 1 injectisome protein InvG (PDB: 6PEE). Predicted hexamer of pMEGA4_50 aligned to the 15-mer complex of InvG shown from the side (**B**) and top (**C**). **D)** Predicted structure of monomer of pMEGA4_299 (median pLDDT 93.91) aligned to *E. coli* K12 flagellin monomer (PDB: 8XCM) and **E)** full filament core. **F)** Predicted homo-nonamer structure of pMEGA4_299. **G)** Alignment of predicted 9-mer of pMEGA4_299 with flagellar core of *E. coli* K12 (PDB:8XCM)[173].

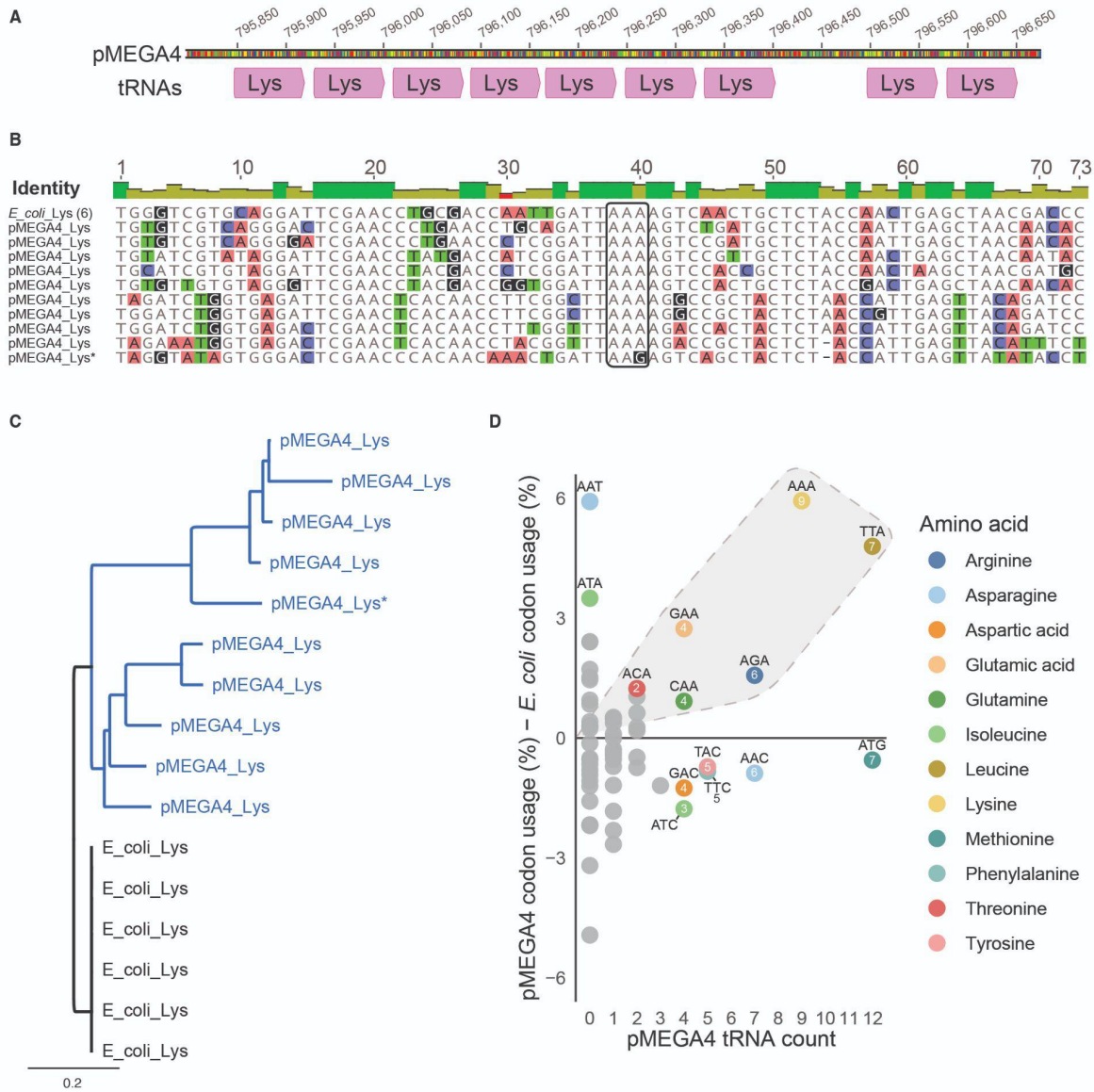



**Figure 5**: **tRNAs of pMEGA4 and *E. coli* host. A)** Nine of the ten pMEGA4 tRNAs for LYS are encoded in a single block. **B**) Nine of the ten pMEGA4 LYS tRNA genes have the same (AAA) codon, however the sequences differ. **C)** Phylogenetic tree constructed using LYS tRNAs for pMEGA4 and host *E. coli*, illustrating pMEGA4 sequence diversity yet conservation of the host sequences. The asterisk indicates the pMEGA4 AAG codon. **D**) Some high tRNA counts are predicted by more extensive use of that codon in pMEGA4 compared to *E. coli* (dots in dashed

outlined region). However, other highly represented codons do not follow this pattern. Numbers associated with each point (colored by amino acid) indicate the count of unique tRNA sequences.

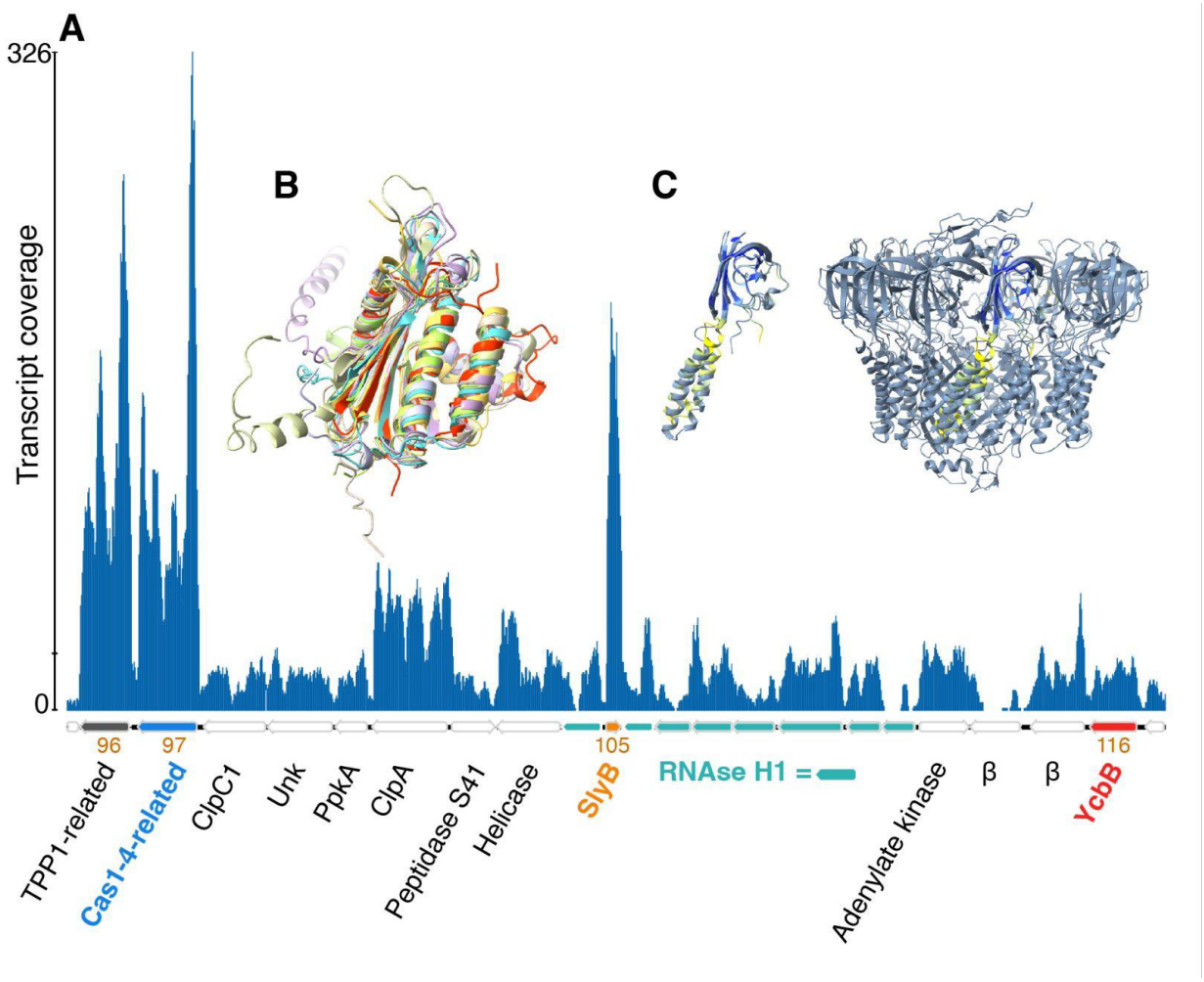
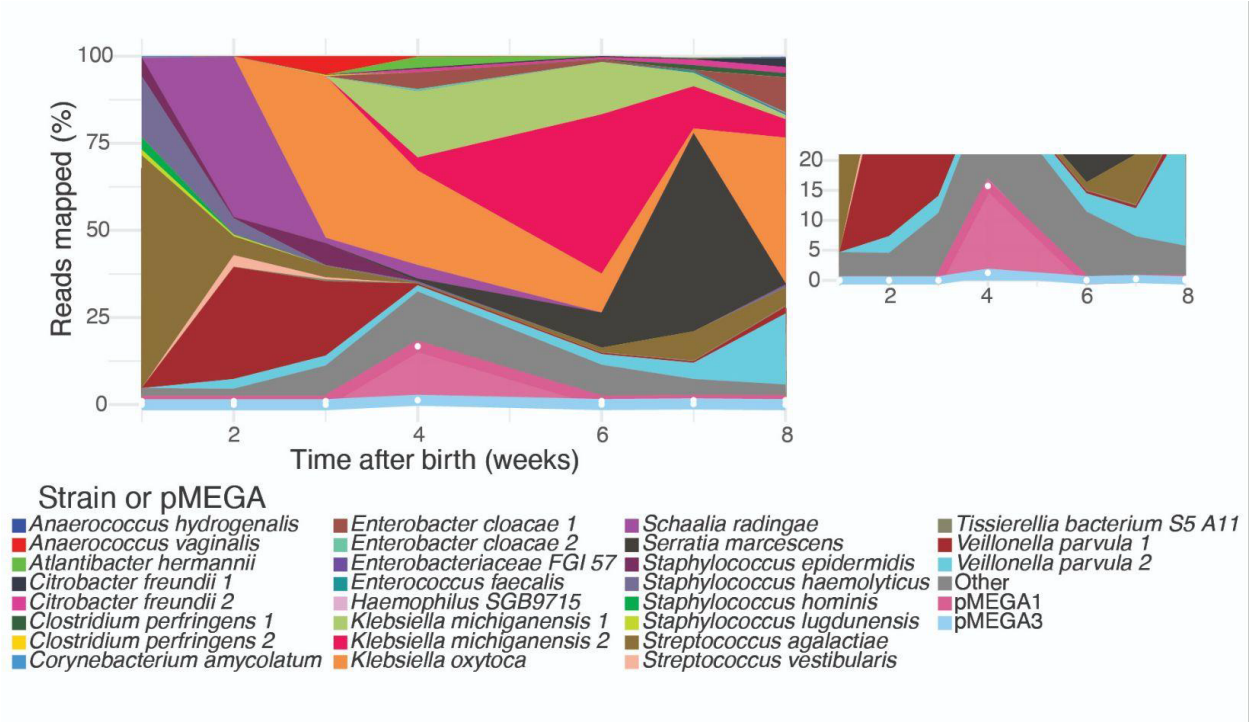



**Figure 6: Public data (SRR10902853) from a panda gut microbiome metatranscriptome mapped to pMEGA3. A)** Regions between pMEGA3_96 and pMEGA3_116 showing three of the 47 protein coding genes with > 100 x coverage. These are a protein related to telomerase-recruitment factor TPP1, a protein most similar to a Cas1-4 fusion and SlyB. Other proteins encoded in this region include two Clp proteases two beta-solenoid proteins and a protein related to YcbB (a L,D-transpeptidase involved in peptidoglycan cross linking to confer antibiotic resistance; red). Most proteins in the region have extra segments not found in the PDB reference.

1160 **B)** All of the 8 almost sequential proteins profiled as RNAse H1 in A have a domain that aligns
1161 well with PDB 1WSJ (red). **C)** Monomer of the protein encoded by pMEGA3_105 (median pLDDT
1162 90.06) and its alignment to PDB 7OJF.

1163 **SUPPLEMENTARY FIGURES**




1164

1165 **Figure S1: Infant gut microbiome community containing pMEGA1 and pMEGA3.** The
1166 percentage of reads mapped to MAGs, pMEGA1 and pMEGA3 within INF1340011 across 8
1167 weeks of life was determined through Illumina read mapping. The "Other" category includes reads
1168 that did not map to MAGs nor pMEGA1,3. Inset shows the same plot but with y values from 0 to
1169 10.

1170
1171
1172
1173
1174

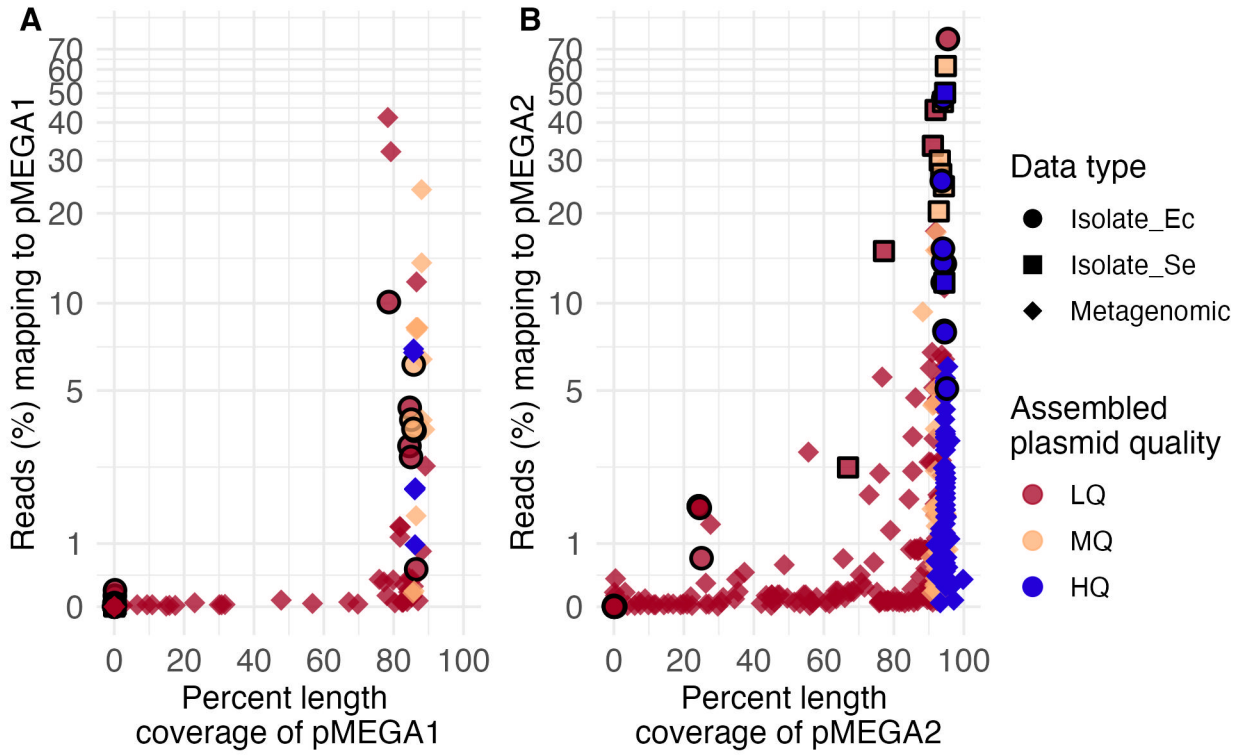



**Figure S2: Prevalence and completeness of megaplasmids in public data.** Sequencing reads from SRA accessions were mapped to pMEGA1 **(A)** and pMEGA2 **(B)** to determine the percent length coverage and percentage of reads mapped to each megaplasmid. For each accession, assembled plasmids were classified as high-quality (HQ - blue) if they had ≥85% coverage of a pMEGA and ≤10 contigs and medium-quality (MQ - yellow) with ≥85% coverage and 11-50 contigs. The remaining samples were classified as low-quality (LQ - red). Diamond: Metagenomic sample, circle: *Escherichia coli* Isolate (Ec), square: *Salmonella enteritidis* isolate (Se). Isolates are further highlighted by a black border.

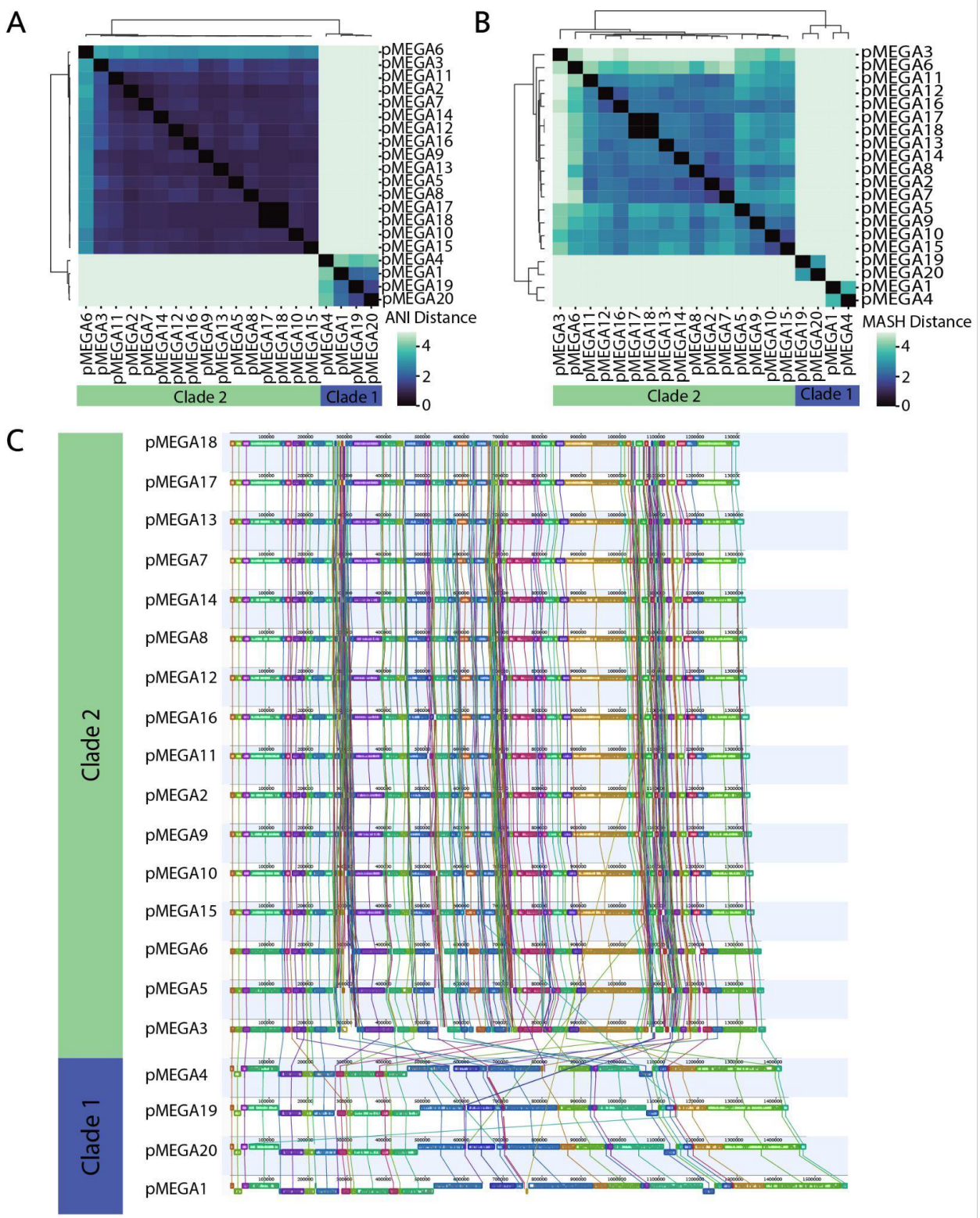
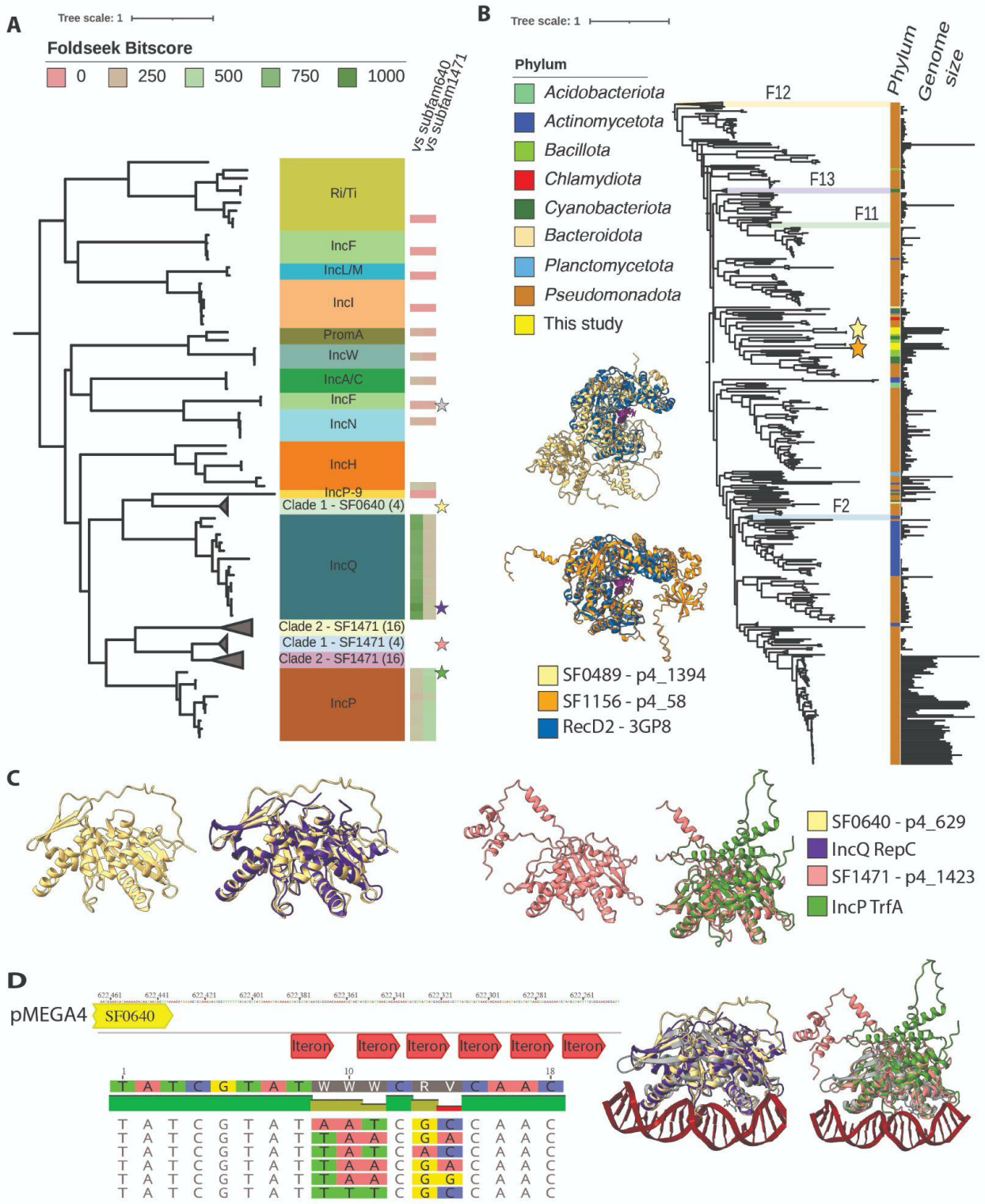



**Figure S3: Comparing 20 complete megaplasmids. A)** Whole genome alignment of complete plasmid sequences. Vertical lines connect blocks with similarity between plasmid genomes. The

sequences were ordered manually and grouped based on clades. **B)** Matrix of average nucleotide identity (ANI). **C)** Matrix of MASH distance.



1190 **Figure S4: Replication initiation proteins and relaxases from two clades of megaplasmids.**
1191 **A)** Phylogeny of 105 replication initiation (Rep) protein sequences. Reference Rep protein
1192 sequences are colored based on plasmid incompatibility group[168]. Bitscores from select FoldSeek
1193 hits are shown as heatmaps on the right of the tree. A star indicates that the protein structure is
1194 highlighted in panel C and D. **B)** Phylogeny of 881 $MOB_F$ protein sequences with clades (F12,
1195 F13, F11, F2) collapsed[164]. Megaplasmids identified in this study are highlighted with a star. The
1196 length of each plasmid is shown on the far-right panel. Predicted protein structures of relaxases
1197 aligned to PDB:3GP8 with ssDNA in purple are shown to the left of the tree. **C)** Representative
1198 structures of Rep proteins from pMEGA4 compared to RepC (purple) and TrfA (green). **D)** Region
1199 downstream of pMEGA4_629 with putative iterons and alignment of repeat sequences. To the
1200 right, alignment of structures from panel C to Rep protein from *E. coli* F plasmid (PDB: 1REP -
1201 grey) bound to the iteron sequence (red).

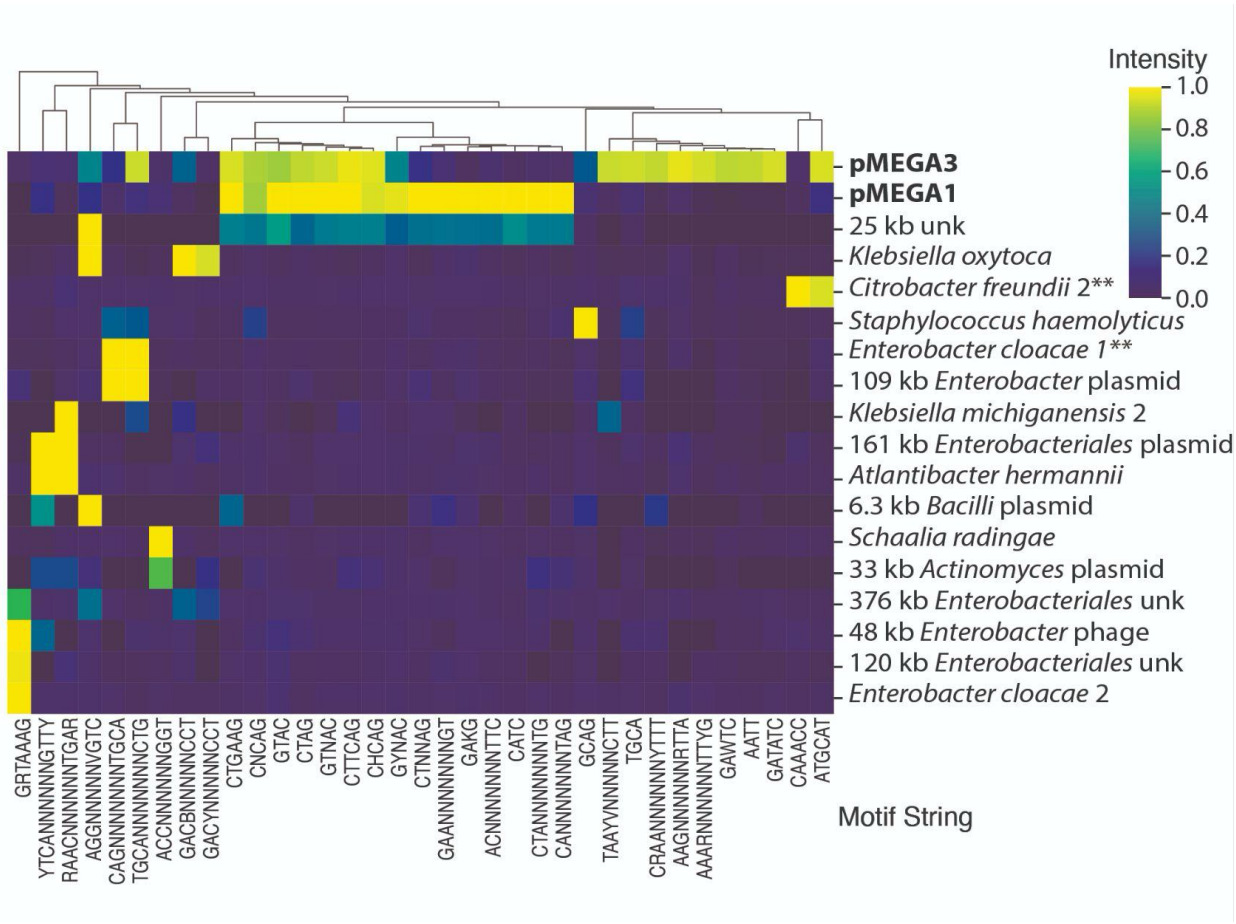
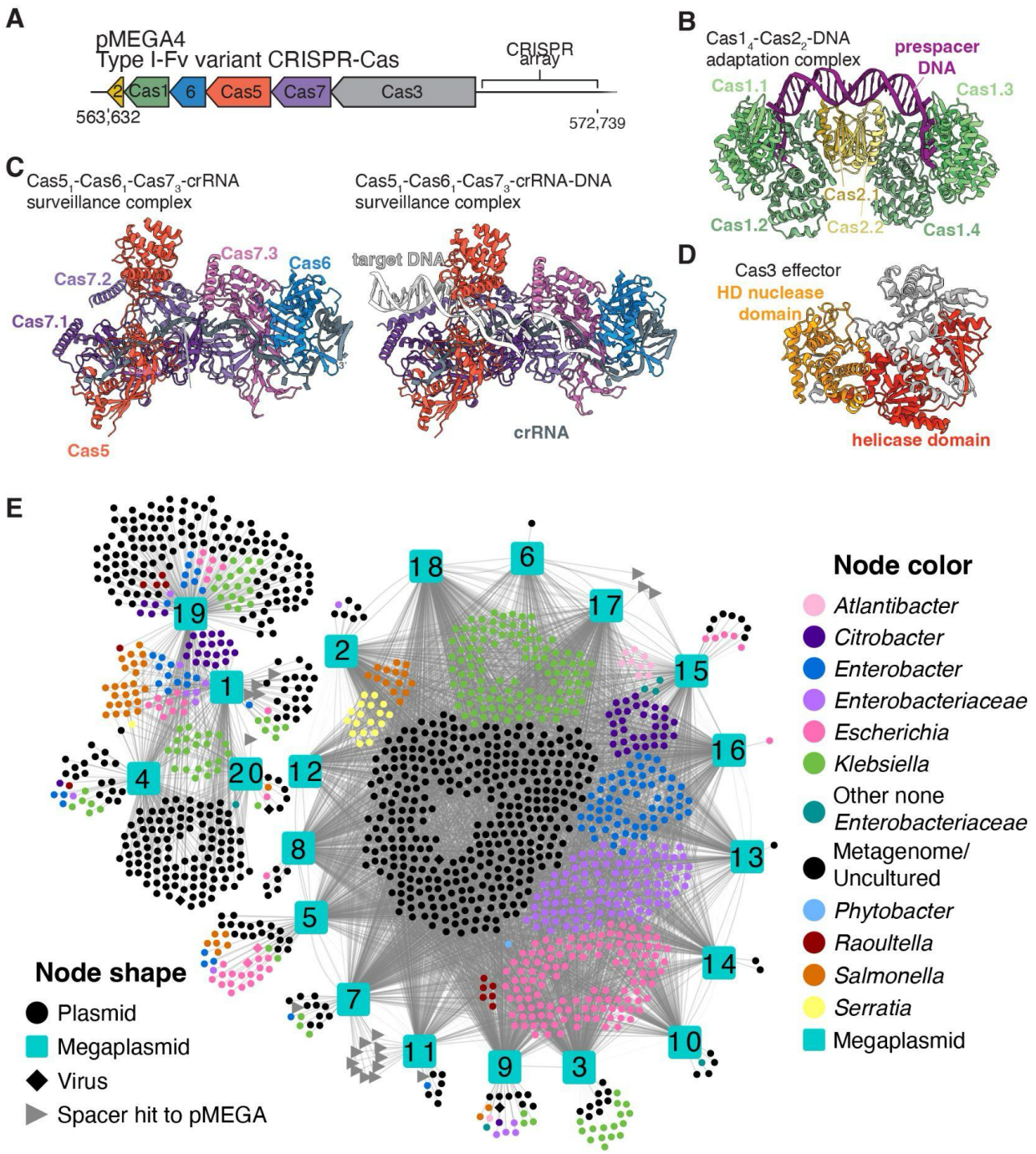
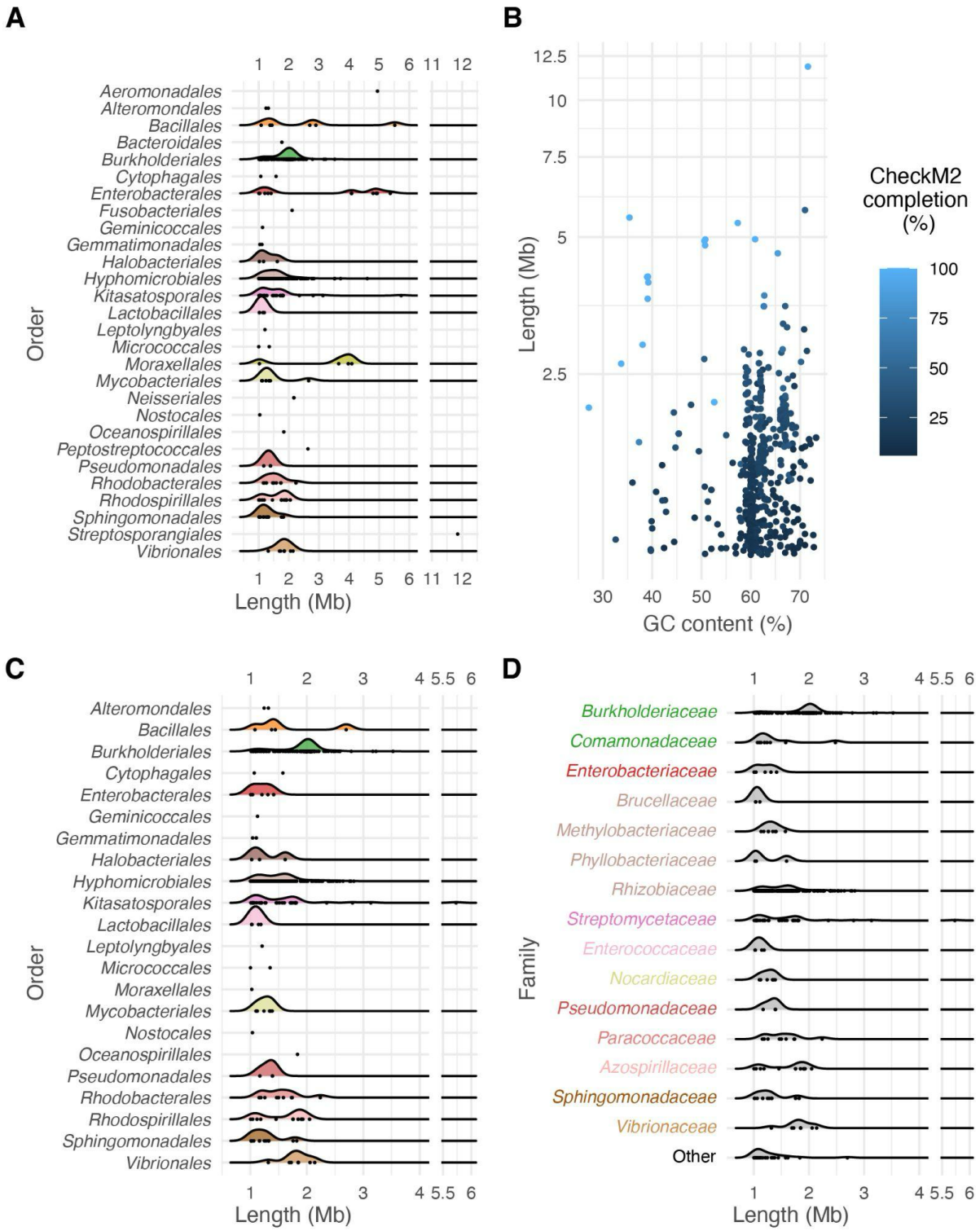



**Figure S5: Possible *Enterobacteriaceae* hosts of pMEGA1 and pMEGA3.** Modified bases were predicted in the PacBio HiFi reads of INF1340011 at week 4. The columns represent the motifs predicted to contain modified bases, and the rows represent genomes. ** Possible host(s) of pMEGA3 based on overlap in modified motif patterns.


A
pMEGA4
Type I-Fv variant CRISPR-Cas
CRISPR array
2
Cas1
6
Cas5
Cas7
Cas3
563,632
572,739
B
Cas1$_4$-Cas2$_2$-DNA adaptation complex
prespacer DNA
Cas1.1
Cas1.3
Cas2.1
Cas1.2
Cas2.2
Cas1.4
C
Cas5$_1$-Cas6$_1$-Cas7$_3$-crRNA surveillance complex
Cas7.3
Cas6
Cas7.2
Cas7.1
Cas5
Cas5$_1$-Cas6$_1$-Cas7$_3$-crRNA-DNA surveillance complex
target DNA
crRNA
D
Cas3 effector
HD nuclease domain
helicase domain
E
Node shape
Plasmid
Megaplasmid
Virus
Spacer hit to pMEGA
Node color
Atlantibacter
Citrobacter
Enterobacter
Enterobacteriaceae
Escherichia
Klebsiella
Other none Enterobacteriaceae
Metagenome/ Uncultured
Phytobacter
Raoultella
Salmonella
Serratia
Megaplasmid

1208

1209 **Figure S6: CRISPR-Cas system in pMEGA4 and targets of CRISPR spacers across**
1210 **pMEGA1-20. A)** Diagram of the type I-Fv CRISPR-Cas locus from pMEGA4. **B)** Structural
1211 prediction of the pMEGA4 heterohexameric $Cas1_4$-$Cas2_2$ integrase complex (pMEGA4_568 and
1212 569, median pLDDT 92.56 and 93.44) bound to a prespacer DNA. **C)** Structural predictions of the
1213 pMEGA4 Cascade surveillance complex in two states. Left, pMEGA4 $Cas5_1$-$Cas6_1$-$Cas7_3$ (ORFs
1214 570 - 572; median pLDDTs 85.19, 91.62, and 93.69) co-folded with a crRNA from an experimental
1215 structure (5O7H) of a similar type I-Fv system. Right, the same structure but co-folded with a
1216 target DNA. **D)** AlphaFold3 prediction of the pMEGA4 Cas3 effector (ORF 573 median pLDDT
1217 91.81), with the HD nuclease domain colored in orange and the helicase domain colored in red.
1218 **E)** CRISPR spacer targets from pMEGA1-20 within other pMEGAs, plasmids from IMG/PR and
1219 PLSDB, and viral sequences from IMG/VR. Each line indicates a spacer that targets a sequence
1220 in another element or, if the connection is between two pMEGA squares, the genomes share the
1221 same spacer with their CRISPR arrays. Grey triangles indicate a spacer from a pMEGA that
1222 targets another pMEGA.




1223

1224 **Figure S7: Plasmid databases are contaminated with complete bacterial genomes and**

1225 **possible secondary chromosomes. A)** Length distribution of 667 non-redundant plasmids >1

Mb from PLSDB and IMG/PR grouped by the Order of their bacterial host. **B)** Length and GC content distribution of plasmid colored by completeness as determined with CheckM2. **C)** Length distribution of 642 plasmids >1Mb after removing sequences with >50% completion, grouped by taxonomic Order. **D)** Length distribution of 642 plasmids >1Mb after removing sequences with >50% completion, grouped by taxonomic Family. Families with two or fewer plasmids were grouped into the 'Other' category.

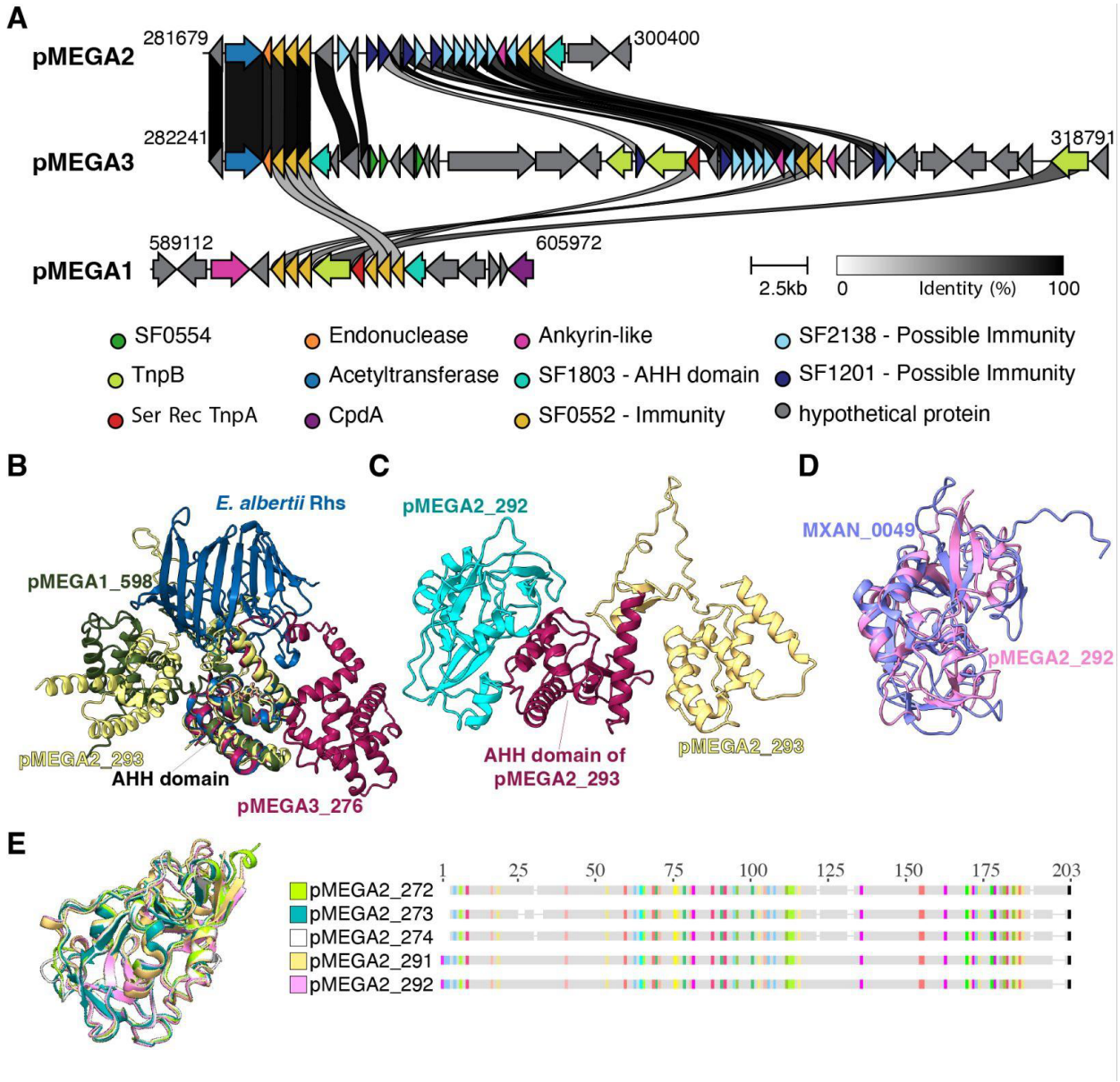



**Figure S8: Polymorphic toxin and poly-immunity loci diversity in megaplasmids. A)**

1234 Alignment of regions in pMEGAs1,2 and 3 with similar subfamilies, including a putative AHH
1235 nuclease toxin (SF1803) and tandem predicted immunity proteins (SF0552, 2138, 1201). **B)**
1236 Predicted polymorphic toxins of SF1803 (median pLDDTs: pMEGA1_598 71.88, pMEGA2_293
1237 78.25, pMEGA3_276 69.12) aligned to a predicted structure of a Rhs family protein from *E. albertii*
1238 A0A855STS0 (blue). The aligned region is a predicted nuclease domain of the HNH/ENDO VII
1239 superfamily (PF14412) and contains a conserved AHH motif. **C)** Result of co-folding the predicted
1240 toxin AHH nuclease toxin with the downstream encoded predicted immunity protein. **D)** Alignment
1241 of the structure of a known nuclease immunity protein MXAN_0049 (PDB: 6A4C) to the predicted
1242 immunity protein (pink: pMEGA2_292 median pLDDT 96.38). **E)** Structural and protein sequence
1243 alignment of 5 members of SF0552 from pMEGA2.

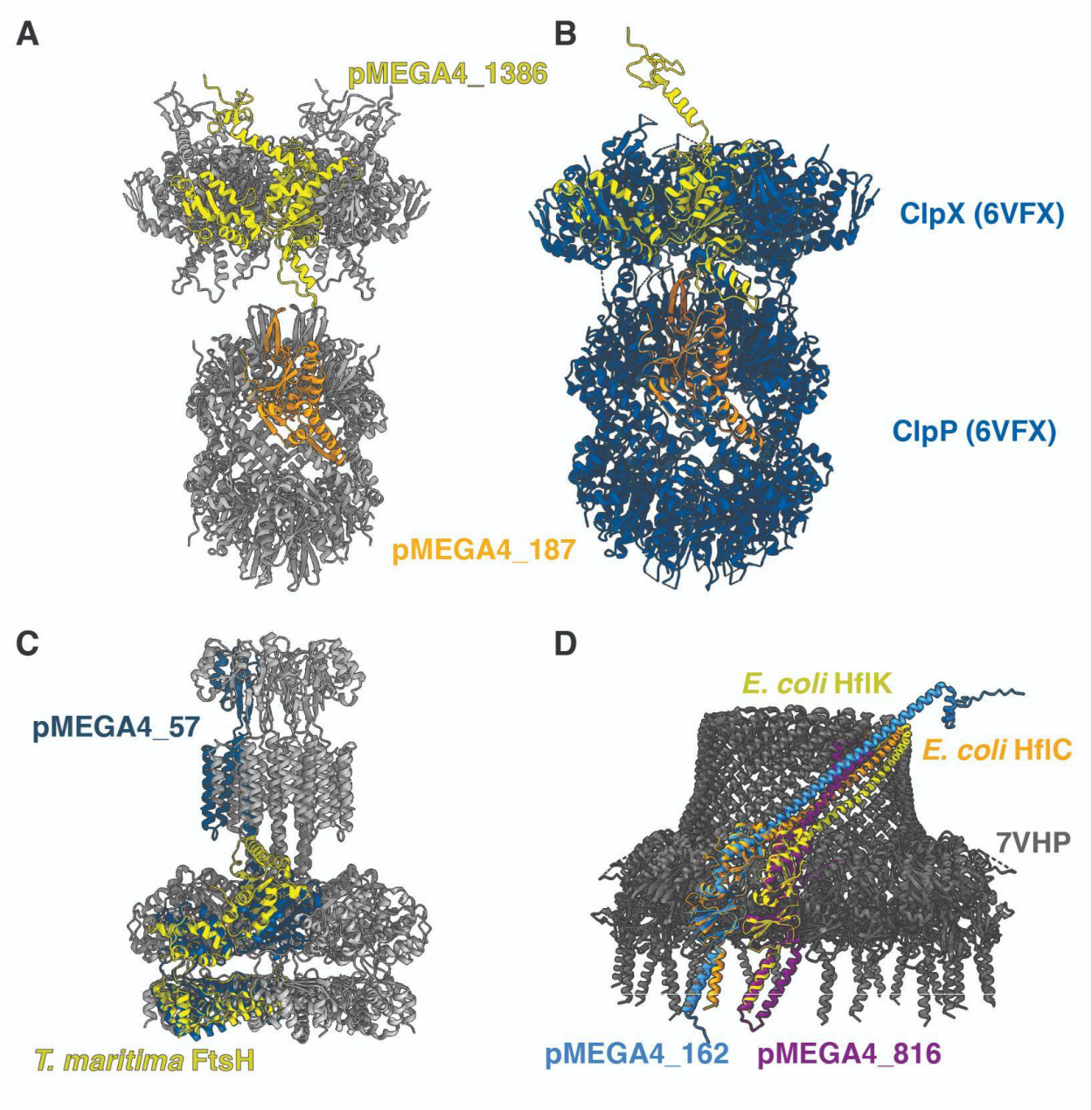



**Figure S9: pMEGA4 predicted protein structures of various functions. A)** Predicted hexamer of predicted ClpX (pMEGA4_1386) and dodecamer of ClpP (pMEGA4_187). **B)** Monomers of ClpX (median pLDDT 92.81) and ClpP (median pLDDT 98.56) from pMEGA4 aligned to the reference of ClpXP from *Neisseria meningitidis* (6VFX). **C)** Predicted hexamer of FtsH (pMEGA4_57) in grey with monomer shown in blue and aligned to *Thermotoga maritima* FtsH (PDB: 3KDS). **D)** Predicted membrane microdomain proteins (median pLDDTs: pMEGA4_162

1251 95.62, and pMEGA4_816 89.94) aligned to the structure of *E. coli* membrane microdomain (PDB:
1252 7VHP). *E. coli* HflC and HflK are colored individually in orange and yellow.

1253 **Supplementary Tables**

1254 **Table S1** List of NANO samples and assembly statistics
1255 **Table S2** ANI and MASH distance between 20 complete plasmids
1256 **Table S3** Relative abundance of MAGs and pMEGAs in INF1340011
1257 **Table S4** Relative abundance of MAGs and pMEGAs in INF1330004
1258 **Table S5** SRA read mapping and assembly from various studies
1259 **Table S6** Hits from other internal NIH samples
1260 **Table S7** List of 20 confident megaplasmids
1261 **Table S8** pMEGA4 confident protein annotations
1262 **Table S9** Core protein subfamilies identified across 20 megaplasmids
1263 **Table S10** Correlations between MAG abundance and pMEGAs
1264 **Table S11** Predicted modified motifs and methyltransferases within pMEGAs, other bacterial
1265 genomes, and MGEs in infant gut microbiome
1266 **Table S12** Spacer hits to pMEGAs from isolate genomes and metagenomes
1267 **Table S13** Putative Enterobacteriaceae 1Mb plasmids from NCBI
1268 **Table S14** Multi-copy protein subfamilies and frequency of occurrence in clusters in pMEGA1-4
1269 **Table S15** *E. coli* G269_1R ORF predictions
1270 **Table S16** Comparing protein sequences from pMEGA4 to the *E. coli* host genome
1271 **Table S17** tRNA sequence types in pMEGA4 and *E. coli* host
1272 **Table S18** tRNA counts and codon usage in pMEGA4 and *E. coli* host
1273 **Table S19** Antibiotic resistance genes predicted in *E. coli* G269_1R and pMEGA4