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 Jauhal[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[1,2,18-20*]

*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

32 [17]Division of Pediatric General and Thoracic Surgery, University of Pittsburgh Children's
33 Hospital, Pittsburgh, PA, USA
34 [18]Department of Environmental Science Policy, and Management, University of California,
35 Berkeley, CA, USA
36 [19]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA,
37 USA
38 [20]Biomedicine Discovery Institute, Monash University, Australia

39 **Summary**

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

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

Introduction

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

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

The best known megaplasmids are from *Rhizobium* (*Alphaproteobacteria*), with genomes up to 2.83 Mbp[17], and 1.8 Mb plasmids of *Streptomyces clavuligerus* (*Actinomycetes*) with linear

77 genomes[18]. In groundwater samples, the largest plasmids reported were circular and 2.96 and
78 1.74 Mb in length[19]. Surveys of plasmids from NCBI RefSeq and plasmid-specific databases
79 representing diverse environments suggest a generally low prevalence of plasmids-like entities >1
80 Mb (0.03% - 1.17%) that mostly associate with *Alphaproteobacteria*[20–24]. Here, we report
81 *Enterobacteriaceae* megaplasmids, including some that associate with *E. coli*. *E. coli* plasmids
82 have been studied intensively, but the sizes of the best researched examples range from ~5.8 kb
83 (ColE1) to ~100 kb (F (fertility) plasmid)[25,26]. In a collection of ~4,500 circular *E. coli* plasmid
84 genomes, the average size was 50 kb, although some clinically relevant IncF plasmids have ~150
85 kb genomes[27].

86 We show that *Enterobacteriaceae* megaplasmids are widely detected in human and animal
87 microbiomes, and notably, that they can be present in the microbiomes of preterm infants, which
88 are especially prone to colonization by *E. coli, Klebsiella, Enterobacter* and *Citrobacter*[28–30].
89 Previous detection of these sometimes highly abundant megaplasmids was likely prevented by
90 assembly fragmentation and general focus on bacterial host genomes. Our analyses employed
91 long-read and short-read sequencing along with manual curation to generate 20 confident,
92 circularized genomes, of which 19 are fully validated and complete. *In silico* protein structure
93 prediction was used to improve functional assignments, and methylation pattern analysis
94 uncovered plasmid-host associations. Overall, we report new aspects of *Enterobacteriaceae*
95 plasmid biology, with implications for the understanding of host bacteria that are especially
96 relevant in human microbiomes.

97 **Results**

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

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

104 Using long-read sequencing, we detected pMEGA1 in three other samples for infant
105 1340011 and found a different (MASH distance 0.24) large ECE from infant 1330004, for which
106 manual curation generated a circular, complete genome of 1,334,733 bp (pMEGA2) (**Tables 1,**
107 **S2**). Subsequently, we identified another PacBio sequence for a circular 1.37 Mb element
108 (pMEGA3) in INF1340011, which did not have sufficient Illumina reads for curation to confirm
109 all base calls (**Table 1**). The genome shares 98.3% ANI (88.6% aligned fraction) with pMEGA2
110 (**Table S2**). Interestingly, throughout the first 8 weeks of life, the gut microbiome of INF1340011
111 harbored pMEGA1 and pMEGA3 concurrently (**Figure S1)**. In both infants, these elements
112 appeared as the gut microbiome shifted towards a community dominated by bacterial strains of the
113 family *Enterobacteriaceae* (**Figures 1A, S1**). The maximum percentage of reads mapped to these
114 ECEs in their respective infant gut samples was 15.74% for pMEGA1, 3.73% for pMEGA2, and
115 1.30% for pMEGA3 (**Figures 1A, S1; Tables S3, S4**).

116 **Large extrachromosomal elements in bacterial isolates and various host-associated**
117 **microbiomes**

118 Through BlastP searches against NCBI, we identified an *Escherichia coli* isolate (strain
119 G269_1R) from a rectal swab of a chicken from Bangkok, Thailand, with assembled fragments
120 similar to pMEGA1[31]. After extensive manual curation of Oxford Nanopore long-read data with
121 high-quality Illumina reads, we reconstructed a complete circular 1,415,043 bp genome
122 (pMEGA4) that shares 95.8% ANI with pMEGA1 (85.0% aligned fraction) (**Tables 1, S2**). The
123 host *E. coli* strain is a ST10 clone with a complete and polished 4,800,525 Mb genome and, in
124 addition to pMEGA4, contains a 104 kb circular plasmid that we classified based on the presence
125 of plasmid markers (i.e., *oriT*, *oriV*, relaxase, replication initiation protein, and partitioning
126 proteins ParAB), and high similarity to other *E. coli* plasmids (NZ_CP070916.1,
127 NZ_CP050045.1)[31].

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

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

142 **Identification of pMEGA elements as megaplasmids**

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

150 Consistent with the identification as megaplasmids, we identified genes for the partitioning
151 proteins (ParC,E (pMEGA4_1,2) components of Topoisomerase IV) and ParF (pMEGA4_1239),
152 a ParA family NTPase of the Type-Ib partitioning system typically co-located with ParG in low
153 copy number plasmids[32]. Although a ParG homolog was not identified, we identified PadC, a
154 protein with a ParB-like nuclease domain (pMEGA4_1030) known to interact with ParA.

155 We also identified most components of a Type IV secretion system within the first 80 kb
156 of the genome (e.g., (Vir/Dot/Icm proteins), two relaxases (pMEGA4_58, 1394), and two
157 subfamilies of replication initiation (Rep) proteins (**Figure S4AB; Tables S8, S9**). The putative
158 relaxases are most similar to the MOB_F family, which includes many large and conjugative

159 plasmids, and clade with plasmids from diverse bacterial hosts (**Figure S4B**)[33]. An intergenic AT-
160 rich region adjacent to a relaxase (pMEGA4_1394) could serve as an origin of transfer (OriT), but
161 this was not identified as such by oriTFinder2[34]. The predicted Rep proteins (pMEGA4_629, 1423)
162 share <31% amino acid identity with bacterial sequences and the predicted structures are similar
163 to those of RepC from an IncQ plasmid in *Salmonella enteritidis* and TrfA from an IncP plasmid
164 in *Escherichia coli* (**Figure S4C**). However, the Rep proteins form distinct phylogenetic clades,
165 and thus these plasmids are only distantly related to IncQ and IncP plasmids (**Figure S4A**). One
166 origin of replication (OriV) with five 18 bp direct repeats was detected upstream of
167 pMEGA4_1423, and another possible iteron-containing locus with less conserved repeats occurs
168 downstream of pMEGA4_629 (**Figure S4D**)[35]. These features, and a type IV coupling protein
169 (T4CP pMEGA4_20), indicate that these are self-transmissible conjugative megaplasmids[33,36].

170 ***Enterobacteriaceae* host these megaplasmids**

171 pMEGA4 was reconstructed from an *E. coli* isolate, and pMEGA17 and near-identical
172 pMEGA18 were reconstructed from two impure *E. coli* isolates (**Table S7**). Incomplete
173 megaplasmid genomes were reconstructed from isolates of *Salmonella enterica* and *E. coli*. We
174 attempted to identify the pMEGA1, 2, and 3 hosts using DNA crosslinking applied to stool samples
175 from two infants, but the results were inconclusive. While most proteins (>66%) in pMEGA1-3
176 have an unknown taxonomic classification, ~8% have sequence similarity to members of the order
177 *Enterobacterales* (**Table S7**).

178 To further identify the pMEGA1-3 hosts, we used abundance-based correlation analysis.
179 The strongest correlation in abundance for pMEGA1 was with *Enterobacter cloacae* 1 (0.92), for
180 pMEGA2, *Citrobacter freundii* 1 (0.88), and for pMEGA3, *Klebsiella michiganensis* 1 (0.9)

181 (**Table S10**). We also investigated DNA modification patterns using PacBio HiFi reads and
182 compared potentially modified motifs shared between pMEGAs, coexisting bacterial genomes,
183 phage, and plasmids (**Figures 2, S5**). Strong overlap in modified motifs between pMEGA2 and
184 the *Citrobacter freundii* 1 genome supports the correlation-based host prediction. For pMEGA1,
185 which only shares motif patterns with pMEGA3, the results are inconclusive (**Figure S5**).
186 pMEGA3 has overlapping motifs with a *Citrobacter freundii* and an *Enterobacter cloacae* strain.
187 The pMEGAs have a greater number and diversity of modified motifs compared to bacterial
188 genomes within the same microbiome sample. Paralleling this, pMEGA4 encodes 21
189 methyltransferases (MTs), whereas the host *E. coli* genome encodes only 6 (**Table S11**).

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

200 We used CRISPR spacers from databases and metagenomes to search for targeting of
201 pMEGAs that may be indicative of megaplasmid host associations[38,39]. We identified 12 non-
202 redundant spacers (100% match) from 33 metagenomes and an isolate of *Proteus cibarius*
203 (*Enterobacteriaceae*) that target regions in at least one pMEGA (expectation for random match of

204 ~e^{-11}) (**Table S12**). The *Proteus* isolate spacer targets a $tRNA_{Met}$ locus that is conserved in
205 pMEGA1, 4, and 20[40]. Five of the 11 metagenome-derived spacers also target tRNAs. Overall,
206 based on isolate source, protein similarity, co-occurrence patterns, CRISPR targeting and
207 methylation analysis, we conclude that various *Enterobacteriaceae* host the megaplasmids.

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

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

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

226 NAD metabolism and acetylation/deacetylation (**Table S9; Figure 3A**). 336 of the 438 ubiquitous
227 proteins occur only once per genome, typically in similar genomic locations within clades (but
228 organization differs between clades; **Figure 3C**).

229 Predicated on the idea that multicopy proteins (here defined as 3 or more copies/genome)
230 may be important to the functioning of megaplasmids, we counted the instances of each subfamily
231 in the pMEGA1-4 genomes (**Table S14**; **Figure 3B**). Of the 10 subfamilies that are multicopy in
232 all 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) that are of interest as protein-based therapeutics[42]. Multicopy proteins
240 that are only highly prevalent in pMEGA1 and 4 are serine recombinases (SF2001: TnpA - up to
241 23 copies) and their accompanying endonucleases, and DNA methyltransferases (SF0095),
242 whereas novel proteins, possible immunity-related proteins, and three distinct RNAse H1
243 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
245 four copies of proteins in the same subfamily occur sequentially in 58 and 33 instances in pMEGA1
246 and 4, respectively. In pMEGA2 and 3, two to eleven genes of the same subfamily occur
247 sequentially 92 and 110 times, and there are eight RNAse H1 subfamily proteins encoded almost
248 consecutively. In addition to the conserved RNAse H domain, these proteins have diverse

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

251 Copies of SF0552, proteins similar to a toxin immunity protein, are clustered in pMEGA1,
252 2 and 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 protease
261 (SF0355) upstream of 3 copies of SF0744 in pMEGA2. Poly-immunity appears more common in
262 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 genetic
265 repertoires and proteomes of this megaplasmid to its host in detail (**Table S15**). Of note, only 9
266 proteins shared >70% amino acid identity to host proteins (**Table S16**). pMEGA4 encodes
267 numerous proteins involved in nucleic acid manipulation, including polymerases, helicases,
268 nucleases, proteins involved in recombination and nucleotide interconversion (e.g., GMP->GDP)
269 and over 20 translation-related proteins (**Table S8**). These functions are also encoded on the *E.*
270 *coli* genome.

271 Possibly involved in homeostasis and quality control, the pMEGA4 genome encodes the
272 protease ClpP (pMEGA4_187) and associated proteins[47–50](**Figure S9AB**). The genome encodes
273 three 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 a role in conjugation, T4SSs may be involved in effector protein delivery
277 and virulence (similar regions occur in pMEGAs1,2 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 are 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,
282 977) with a von Willebrand factor A–like domain (vWA), which is commonly associated with
283 adhesins. 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. This appendage is involved in adhesion, motility, virulence, and immune
286 stimulation[56,57] (**Figure 4DEFG**).

287 pMEGA4 encodes 105 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 which
289 49 are unique (57%) (**Table S17**). The tRNAs are mostly arrayed in blocks, generally with ~6 to
290 ~ 25 nucleotide (nt) gaps between them (**Figure 5A**). Only two blocks (with 7 and 4 tRNAs) occur
291 in the *E. coli* genome. Within the set of complete pMEGA genomes, there are up to 208 tRNAs,

292 with statistically significant enrichment in Clade 2 (average 199 ± 9) compared to Clade 1 (average
293 113 ± 8).

294 For the tRNAs for the most highly represented amino acids (Leu, Lys, (Met)), almost all
295 have the same codon. For Lys, 9/10 are encoded in one block and all have the same codon (AAA)
296 despite being quite divergent in the rest of their sequences (**Figure 5B**). In contrast, the six *E. coli*
297 tRNA Lys are all identical (**Figure 5C**). For 11 other amino acids, all tRNAs for that amino acid
298 are in a single block. In contrast, in the *E. coli* host genome, there is only one tRNA (Gln) where
299 all 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
302 differences in their codon usage compared to that of their host. We find that some codons much
303 preferred by pMEGA4 compared to the host *E. coli* have more highly represented tRNAs (e.g.,
304 AAA, TTA), but this is not the rule (**Figure 5D; Table S18**). Some less commonly used tRNA
305 codons (e.g., ATG, AAC) are highly represented, and some much more highly used codons (e.g.,
306 AAT, ATA) are not represented in pMEGA4. Notably, there is high sequence diversity associated
307 with all highly used codons (**Figure 5D**).

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

309 pMEGA4 encodes a protein that forms a ferritin-like cage for DNA protection during
310 starvation (pMEGA4_661; also identified in *E. coli*) and two small alarmone hydrolases that
311 regulate the stringent response in a similar way to *gppA* found in the *E. coli*[58,59]. Further supporting
312 defense against external challenges are genes for redox homeostasis, nucleic acid repair, oxidative
313 stress, heat shock, and genes that form, maintain and repair disulfide bonds. Again, these functions

314 appear redundant with those encoded in the *E. coli* genome. pMEGA4 encodes the outer membrane
315 protein SlyB (also encoded in the host genome), which is a lipoprotein that forms a 13-mer pore
316 (**Figure S6C**). The PhoPQ regulon is responsible for the expression of SlyB in response to external
317 stress[60]. We also identified three near-consecutive genes for tellurium resistance (TerD, not
318 identified in *E. coli*).

319 Possibly in response to bacterial competition, pMEGA4 encodes a protein implicated in
320 immunity to effectors (pMEGA4_953)[61], an amidase antitoxin (pMEGA4_699), and a hydrolase
321 against ADP-ribosylating toxins (pMEGA4_1469[62]). These antitoxins appear without their toxin
322 counterparts. Conversely, there are at least two proteins annotated as the lipoprotein toxin
323 entericidin B that are separate on the plasmid and do not appear co-localized with an antitoxin
324 (SF1615)[63]. The *E. coli* host genome encodes multiple Type II toxin/antitoxin systems (ie. MazEF,
325 HipAB) and many possible secreted effectors/toxins [64–66].

326 In addition to the CRISPR-Cas system described earlier, pMEGA4 encodes putative
327 defense system proteins structurally similar to the Upf1 family of RNA helicases
328 (pMEGA4_338,520; detected by PADLOC as Mokosh TypeII systems), and a Ser/Thr kinase
329 annotated as a PD-T4-6 system (pMEGA4_91)[67–69]. The *E. coli* genome encodes 7 single-gene
330 phage defense candidate systems, 2 Hma (Helicase, Methylase, ATPase) defense systems, HEC,
331 hhe, and a CRISPR array[70].

332 The *E. coli* host of pMEGA4 encodes the extended-spectrum beta-lactamase (ESBL)
333 $bla_{\text{CTX-M-55}}$ on the abovementioned 104 kb plasmid, along with genes for quinolone (*qnrS1*),
334 chloramphenicol (*catII*), and tetracycline resistance (*tetR* and *tetA*)[31] (**Table S19**). Few antibiotic
335 resistance mechanisms were detected in pMEGA4 (**Table S19**). However, this function may be

336 conferred by 10 genes for transporters, including MacAB- and AcrAB-TolC efflux systems that
337 were identified based on sequences and predicted structures[71]. Possibly relating to antibiotic
338 resistance, pMEGA4 contains an outer membrane protein/porin implicated in carbapenem import
339 (pMEGA4_670), and a methyltransferase that may activate a peptide antibiotic (pMEGA4_515).

340 **Panda gut metatranscriptomic data reveals broad gene expression**

341 We identified a public metatranscriptomic dataset from a panda gut microbiome with both
342 a clade 1 and a clade 2 megaplasmid (SRR10902853). As there was no DNA data, we mapped the
343 reads to pMEGAs 1 - 4 competitively and detected extensive read recruitment to all four genomes.
344 By far the highest recruitment was to pMEGA4 with almost one million reads, but over 128,000
345 reads were recruited to pMEGA3. From the transcripts, we assembled 1,613 contigs > 5 kb in
346 length, with a maximum length of 145 kb (4 of the 5 longest scaffolds were from *E. coli* and
347 *Klebsiella* phages), and a 109 kb scaffold with similarity to pMEGA4. 75 contigs > 5 kb mapped
348 uniquely to either clade 1 (pMEGA4) or clade 2 (pMEGA4) genomes with no overlap of contigs
349 between clades. For the clade 1-like plasmid, we recovered 940,590 bp with ANI of 99.52% across
350 63.81% of pMEGA4. We used the transcripts to curate the 109 kb sequence from the clade 1
351 genome, primarily closing two scaffolding gaps (**Figure 6A**). For the clade 2-like plasmid, we
352 recovered 563,270 bp with ANI of 98.70% across 37.93% of pMEGA3.

353 Most genes are expressed at relatively even levels across both clade representatives. The
354 most extreme peaks in RNA abundance are all associated with intergenic regions, two of which in
355 the clade 2 genome are > 1 kb in length. Many of the highly expressed protein-coding genes lack
356 functional assignments (49% and 55% of highly expressed proteins in clades 1 and 2, respectively),
357 and some are very small (<100 amino acids; 23% in both; **Tables S20, S21**). However, proteins

358 predicted to be involved with the T4SS and pilus are highly expressed in the clade 1-like genome,
359 including two copies of a major pilin subunit (**Figure 6AB**). A region of pMEGA3 (**Figure 6C**)
360 that features a large cluster of proteins with RNAse H1 (as well as other) domains (**Figure 6D**)
361 includes a highly expressed protein annotated in part as TPP1, a possible Cas1-Cas4 fusion protein,
362 and SlyB (**Figure 6E**). Also expressed but at lower levels are two Clp proteases and a L,D-
363 transpeptidase YcbB, possibly involved in resistance to β-lactams due to alternative cross-linking
364 of peptidoglycan[72].

365 Members from 10 protein subfamilies are highly expressed by megaplasmids of both
366 Clades, including SlyB, a protein involved in lipid biosynthesis, the ribosome hibernation factor
367 YhbH, IcmD/DotP, tellurium resistance protein TerD, oxygen damage protein YfiD, TPP1, a Type
368 IV pilin subunit, and two unknown proteins (**Tables S20, S21**). Of the 170 tRNAs in pMEGA3
369 and 108 in pMEGA4, virtually all recruited transcripts.

370 From the assembled transcripts, ribosomal protein S3 diversity revealed a highly abundant
371 *E. coli* (~1615X), another potential *E. coli* with lower coverage (~33X), as well as multiple
372 *Bacillota* (ie. *Clostridiales* and *Bacillales*) and a potential *Bacteroidota* member. Binning of
373 transcript-assembled contigs to the *E. coli* OP50 genome (CP059137.1) resulted in 525 contigs
374 representing 4.02 Mb, with 99.15% ANI over 79.04% of the OP50 genome. Given the presence of
375 two megaplasmids and two members of *Enterobacteriaceae*, the high coverage Clade 1-like
376 plasmid is likely hosted by the highly abundant *E. coli*, while the Clade 2-like plasmid may
377 associate with the less abundant *E. coli* in the panda gut community.

378 **Discussion**

379 We report the discovery and bioinformatic analysis of two clades of megaplasmids. The
380 megaplasmids are clearly distinguished from bacteriophages by the absence of structural genes
381 and from chromids/chromosomes by their low fraction of metabolic genes, protein novelty,
382 plasmid-associated genes, replication style, and differences in GC content compared to host
383 genomes. Members of both pMEGA clades whose genomes were reconstructed from isolates are
384 confidently linked to *E. coli* and *Salmonella spp.* hosts. More broadly, analyses of protein
385 similarity, co-occurrence, and DNA modification patterns indicate that the hosts are likely all
386 *Enterobacteriaceae*. With genomes of up to 1.58 Mb, these are, by far, the largest megaplasmids
387 in bacteria of this family.

388 While initially identified in preterm infant gut microbiomes, the megaplasmids occur in
389 microbiomes of the human adult gut, animal gut, and other environments, suggesting they are
390 widespread. All but one of the 20 genomes is "complete", rather than bins or putatively circularized
391 sequences. The exception is pMEGA3, which was reconstructed from PacBio reads but was too
392 low in abundance to very 100% of the bases using Illumina reads. Complete genomes ensured
393 accurate determination of genome lengths, provided information about genome organization and
394 replication style, and facilitated accurate gene and proteome predictions. *In silico* structure
395 prediction and analysis helped identify proteins implicated in virulence, components of the T4SS,
396 proteins involved in host-cell adhesion, antibiotic resistance, and phage defense. Overall, the
397 findings suggest that these megaplasmids are self-transmissible. Extensive methylation motifs to
398 protect from diverse host restriction systems suggest that the pMEGAs may be able to replicate in
399 multiple hosts.

400 Plasmids must ensure that the burden of their existence does not compromise the host
401 organism's viability[73]. The degree to which plasmids must offset the cost of their carriage is
402 probably size-dependent and may drive plasmid acquisition of genes that increase the host's
403 fitness. Presumably, pMEGA genomes became large partly via the acquisition of gene variants
404 from prior hosts. This likely occurred in an ancestral plasmid, as all pMEGAs have very low GC
405 content compared to *Enterobacteriaceae* hosts (i.e., the GC content of host-derived genes has been
406 ameliorated).

407 Presumably reflecting the acquisition of host functions, there is substantial overlap in gene
408 content of pMEGA4 and its *E. coli* host. Some plasmid-borne variants may confer advantages by
409 different specificity or enzyme kinetics, by supplying genes for stress protection, or by extending
410 growth over a wider range of conditions. For example, plasmid-borne Clp and FtsH proteases
411 exhibit different thermotolerance[74–77] or specificities[48,49,78–80] that may expand substrate range.
412 Associated membrane microdomains (HflK/C) may result in compartmentalization of specific
413 proteins involved in signal transduction, membrane trafficking, and regulation of proteases,
414 including FtsH[51,52,81]. Other plasmid-derived benefits include expanding defense against invading
415 bacteriophages and other plasmids (e.g., by provision of additional CRISPR-Cas systems).
416 Plasmids may also increase antibiotic resistance by supplying efflux pumps with different
417 specificities, remodelling the cell wall, and expressing variants of thymidylate synthase, LpxC,
418 and ClpP that are insensitive to antibiotics[72,82–84].

419 Megaplasmids may acquire functions to counter host defenses. They encode genes for
420 NAD production, probably to offset decreased availability of NAD when the host depletes its NAD
421 pool in response to infection[85]. Host tRNA degradation in response to invasion by ECEs may be
422 countered by megaplasmid tRNA ligases, as has been shown for phages[86]. The pMEGAs encode

423 upwards of 100 tRNAs. Although some certainly offset differences in codon use between the
424 pMEGA and its host (**Figure 5**), structural diversity in tRNAs with the same anticodon may confer
425 a spectrum of resistances to host tRNA nucleases. This counter-defense is well established in
426 phages[87]. There is apparently strong selective pressure to preserve the codon and for sequence
427 divergence outside of the codon.

428 Megaplasmids encode functions that do not duplicate those of their hosts. Of potential
429 medical importance in the pMEGA4 plasmid in the ST10 *E. coli* strain is InvG. This protein forms
430 the outer membrane (OM) basal body in the Type III secretion system injectisome that delivers
431 effector proteins into host cells, but is also found in Type II secretion systems, Type IV pili
432 structures, and phage assembly complexes[55,88,89] (**Figure 4**). Putative toxin and poly-immunity
433 arrays may enhance inter-bacterial competition and protect from diverse toxins, whereas diverse
434 adhesion factors (ie. vWA domains) may contribute to motility and virulence[90,91]. Highly prevalent
435 multicopy proteins have alpha- and beta-solenoid structures, which are often found in secreted
436 effectors[92]. Effectors containing ankyrin repeats from intracellular pathogens such as *Legionella*
437 *pneumophila* and *Coxiella burnetii* are involved in the invasion of eukaryotic cells[93]. Alpha-
438 solenoid folds are more common in proteins of obligate intracellular pathogens as compared to
439 other bacterial genomes[94]. Ankyrin repeat proteins (a form of alpha-solenoids) often mediate
440 protein-protein interactions and may have functions analogous to those of DARPins, which are
441 alpha-solenoids that are engineered to bind to target molecules with high affinity and specificity
442 for functions such as antibodies and as anti-viral agents[95]. It is also possible that the megaplasmids
443 accumulate ankyrin proteins to modulate their bacterial host proteome, or modulate eukaryotic
444 immune response, as has been demonstrated in phage[96,97].

445 Some megaplasmids associate with bacteria that cause gastroenteritis (i.e., *S. enterica*) and
446 sepsis (i.e., *E. coli*)[98,99]. The pMEGA4 host *E. coli* is an ESBL-producing ST10 clone. ESBL-
447 producing ST10 clones of *E. coli* have been increasing in prevalence globally[100] and are common
448 in animals, including chickens, associated food products, and humans[101–104]. Some other
449 megaplasmid-bearing *E. coli* and *S.* Enteriditis isolates originate from human samples and one *E.*
450 *coli* isolate of serotype O157:H7 was associated with an outbreak due to contaminated lettuce
451 (Biosample: SAMN14113693)[105,106]. These observations motivate the investigation of the medical
452 significance of megaplasmids.

453 **Conclusions**

454 We identified the first megabase-scale plasmids in *Enterobacteriaceae*. They were detected
455 in infants, adults and animals, as well as in hospital sinks where they likely serve as a reservoir for
456 human transmission[107]. The megaplasmid genomes encode many solenoid proteins that may bind
457 selectively to proteins or other molecules. We also identified mechanisms to extend host and
458 habitat ranges, increase competitiveness in microbiomes, counter host defense, and protect from
459 antibiotics. Much remains to be learned about their large and incompletely defined functional
460 potential, but there are many indications that they hold medical significance. From the perspective
461 of future work, it is important that they occur in genetically tractable *Enterobacteriaceae*,
462 including *E. coli*.

463 **Acknowledgements**

464 We thank Dr. Rani Phoolwanti for input regarding the RNAase H1 proteins and Colin Robinson
465 for input regarding the metatranscriptome data. We thank the generators of the panda
466 metatranscriptome data SRR10902853, whom, to date, we have not been able to identify. We thank

467 Dr. Shufei Lei and Leylen Miloslavich for technical support. AKG, JFB, OTT, RS, and SW were
468 supported by the Innovative Genomics Institute.

469 **Author Contributions:**

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

480 **Declaration of interests**

481 The authors declare no conflicts of interest.

482 **Declaration of generative AI and AI-assisted technologies**

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

484 **Funding**

485 The primary funding for this project was provided by NIH award 5R01AI092531-14 to JFB, MM,
486 with some support from the Chan Zuckerberg Foundation (CZIF2022-007203) via the Innovative
487 Genomics Institute. Long-read PacBio sequencing was provided by a SMRT grant to AKG from

488 PacBio and SeqCenter. RS and SW were funded by Lyda Hill Philanthropies, Acton Family
489 Giving, the Valhalla Foundation, Hastings/Quillin Fund, the CH Foundation, Laura and Gary
490 Lauder and Family, the Sea Grape Foundation, the Emerson Collective, Mike Schroepfer and Erin
491 Hoffman Family Fund, the Anne Wojcicki Foundation. AHD and AEM were funded by the
492 Biotechnology and Biological Sciences Research Council (BBSRC) Institute projects
493 BB/X011011/1 and BBS/E/QU/230002A. AHD was funded by a UKRI BBSRC Doctoral Training
494 Grant BB/T008717/1. AK was funded by the Finnish Multidisciplinary Centre of Excellence in
495 Antimicrobial Resistance Research (FIMAR; 346127). MH received funding from the Swiss
496 National Science Foundation grant IZJFZ3-177614.

497 **Data and code availability**

498 Prior to publication, the genome sequences described in this study can be accessed and downloaded
499 from ggKbase: https://ggkbase.berkeley.edu/project_groups/Megaplasmids_public. Raw
500 sequencing data will be made available on NCBI upon publication. Other data reported in this
501 study can be accessed at: 10.5281/zenodo.1723140.

502

503 **Methods**

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

505 Stool samples from two preterm infants enrolled in the NICU Antibiotics and Outcomes
506 (NANO) trial (Clinical Trial ID: NCT03997266) were collected over the first eight weeks of life
507 after obtaining informed parental consent[108]. DNA was extracted from samples using the
508 ZymoBIOMICS™ DNA Miniprep Kit. Libraries prepared with the Illumina DNA prep kit were
509 sequenced on an Illumina NovaSeq 6000 (2 x 151 bp reads). Reads were trimmed using sickle
510 v1.33, and mapped against the human T2T reference using bowtie2 v2.5.4, resulting in an average

511 of 4.09 gigabases per sample (**Table S1**)[109,110]. Metagenomic assembly was performed using
512 metaSPAdes v4.0.0[111]. Scaffolds greater than 1000 bp were annotated using Prodigal v2.6.3, and
513 searched against KEGG, UniProt, and UniRef using USearch v10.0.240 (-ublast)[112–116]. Ribosomal
514 RNA sequences were predicted as previously described, and tRNA sequences were identified
515 using tRNAscan-SE v1.3.1[117–119].

516 For a subset of samples (**Table S1**), either the same DNA extract used for Illumina
517 sequencing, or a new DNA extract prepared as above but with a modified lysis step was sequenced
518 on a PacBio Revio. HiFi reads were trimmed using BBDuk v39.10 and assembled using hifiasm-
519 meta v0.3-r073, resulting in, on average, 6.99 gigabases per sample[120,121]. Binning of Illumina
520 short-read data to generate metagenomic-assembled genomes (MAGs) was accomplished using a
521 combination of CONCOCT v1.1.0, MaxBin2 v2.2.7, MetaBAT2 v2.15 and Vamb v3.0.2, and the
522 best bins were selected using DAS Tool v1.1.6[122] with default settings[122–126]. The PacBio HiFi-
523 MAG-Pipeline v3.3.2 was used with slight modifications to detect putative complete circular
524 genomes (93% completeness) and bin other scaffolds into MAGs with less than 100 contigs[127].
525 Other settings included: a score threshold of 0.2 for DAS Tool, a minimum completeness of 70%,
526 and a maximum contamination of 10%. Metagenomic assemblies were visualized and binned
527 manually through an in-house platform ggKbase (https://ggkbase.berkeley.edu/). MAGs from both
528 Illumina and PacBio assemblies for each infant were dereplicated using dRep v.3.4.5[128] (with
529 completeness (-comp) 75, contamination (-con) 10, secondary cluster ANI (-sa) 0.98, and coverage
530 threshold (-nc) 0.25). This resulted in 27 MAGs for INF134001 (total 110) and 16 MAGs for
531 1330004 (total 72). The taxonomy of each MAG was determined using PhyloPhlAn v3.0.36[129].

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

533 Examination of a PacBio assembly of one infant sample (INF1340011_TP8) identified a
534 1.58 Mb (pMEGA1) contig of unknown origin, with 3 bacterial single-copy genes, no ribosomal
535 rRNA sequences, and many gene annotations characteristic of a plasmid. Two additional
536 putatively complete plasmids were identified in other PacBio assemblies; a 1.33 Mb plasmid
537 (pMEGA2) in INF1330004 at TP4 and a 1.37 Mb plasmid (pMEGA3) in INF1340011 at 4 weeks.
538 Trimmed and filtered reads were mapped against the putative complete megaplasmids with
539 BBMap v39.10[120] (default settings with ambiguous=random). The alignment file was imported
540 into Geneious Prime v2025.1.1, and reads were re-mapped to allow 0% mismatches to identify
541 regions of possible variation or error. Reads were mapped again to allow 4% mismatches to the
542 reference to correct possible single-nucleotide polymorphisms, insertions, deletions, or assembly
543 errors. This process was repeated until no errors were visible in the plasmid genome. Reads that
544 surpassed the ends of the sequence were used to extend the plasmid until a repeated sequence was
545 present at the beginning and end of the plasmid, suggesting circularity. Paired reads that mapped
546 at both ends of the sequence were also used as an indication of a circular plasmid, as described
547 previously[130,131]. CoverM v0.7.0 was used along with Illumina reads to determine the depth and
548 percentage of reads mapped to pMEGAs and the dereplicated bacterial MAGs for each infant
549 across all available timepoints[132].

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

551 To further assess the diversity and distribution of these megaplasmids in other datasets, we
552 used the bacterial single-copy gene histidyl-tRNA synthetase (hisS) as a marker. A BlastP search
553 against NCBI non-redundant protein sequences (executed December 16th, 2024) resulted in 1
554 highly similar hit for pMEGA1 (*Escherichia coli* – 93% cov, 97.63% aa ID, BioSample:

555 SAMN18191027, Assembly: GCA_024673755.1)[133–135]. The top BlastP match for pMEGA2 is
556 from a *Serratia sp.* with 98% coverage and 63.84% amino acid similarity (WP_282994782.1). The
557 *E. coli* sequence is from a genomic assembly of an isolate (G269_1R) from Bangkok,
558 Thailand[31,136]. We assembled the Illumina sequencing reads from this accession using the same
559 methods as described above. Scaffolds that were distinct in GC content from the host genome
560 (<30% for pMEGA vs ~51% for *E. coli*), similar in read coverage, and that had no consensus for
561 taxonomy based on the taxonomy of predicted gene sequences were binned. We also mapped
562 scaffolds from the assembly to the complete pMEGA1 and pMEGA2 sequences from above using
563 minimap2 v.2.28-r1209 (with the flag -x map-hifi)[137]. From these two approaches, we manually
564 selected 23 scaffolds encompassing ~1.39 Mb to represent a putative megaplasmid. Through this
565 approach, and to the best of our ability and knowledge, we believe these scaffolds do not belong
566 to *E. coli* or other easily identified mobile genetic elements such as other plasmids or phages.

567 Extensive manual curation and read mapping did not result in a complete circular element,
568 so Nanopore long-read sequencing was performed on the *E. coli* isolate. Briefly, genomic DNA
569 from the isolate was prepared with the Rapid sequencing DNA kit (SQK-RBK114.24) and
570 sequenced on a GridION MKI R10 flow cell. Reads were basecalled with the high-accuracy model
571 using Dorado v7.2.13[138]. Long-read sequencing resulted in very low coverage of the putative
572 megaplasmid, but a complete 4.8 Mb genome for *E. coli* and a 104 kbp plasmid. The few Nanopore
573 long-reads that were recovered against the 1.39 Mb plasmid were used to scaffold the contigs
574 derived from Illumina sequencing. Through many iterative rounds of read mapping and extension
575 from the ends of Illumina contigs and into divergent gaps where no direct read recruitment was
576 possible, we reconstructed a complete 1.41 Mb plasmid (pMEGA4).

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

578 We employed two tools to search for similar sequences across read data deposited in
579 NCBI's Sequence Read Archive (SRA) and other collections. First, PebbleScout was used to
580 search against all available databases with the complete reference sequences of pMEGA1 and
581 pMEGA2 elements[139]. This resulted in 55 hits with a PB score >60 for pMEGA1, and 319 hits
582 with a PB score >60 for pMEGA2. Next, we searched all available assembled data in Logan (11-
583 20-2024) using the *hisS* gene sequence (813 bp) from pMEGA1 and 1000 bp from the start of the
584 *hisS* gene sequence from pMEGA2[140]. This resulted in 61 hits for pMEGA1 and almost 500 hits
585 for pMEGA2 with a kmer coverage >=0.70. Many of the hits overlapped with the PebbleScout
586 search results, therefore, we selected a few studies and any bacterial isolate hits only for additional
587 investigation (**Table S5**). The bioprojects PRJEB36775 and PRJNA938144 relate to hospital sink
588 metagenomes and infant gut metagenomes, respectively, and were selected given their high
589 number of hits and relevance to human health[141,142].

590 For all selected hits, reads from the SRA or European Nucleotide Archive (ENA) were
591 downloaded and processed as described above for trimming and human read removal, if needed.
592 Each read accession was treated as a separate sample regardless of whether more than one hit
593 originated from the same BioSample (eg. hits from PRJEB21259 and PRJEB36775). Trimmed
594 reads were mapped using bowtie2 and samtools to the pMEGA1 and pMEGA2 reference
595 sequences to determine the percent length coverage, depth of coverage and percentage of reads
596 mapped to the pMEGA in each sample[110,143]. We proceeded to assemble the hits using metaSPAdes
597 v4.0.0 as described above and assessed to what extent complete sequences could be recovered by
598 mapping scaffolds to pMEGA1 and pMEGA2 as references using minimap2[111,137]. Samples were
599 considered a high-quality hit for a pMEGA if 10 or fewer scaffolds provided coverage of >=85%

600 of the reference pMEGA. Samples were considered medium-quality hits if 11 to 50 scaffolds
601 provided coverage of >=85% of a pMEGA, and all other hits were considered low-quality. From
602 the high-quality hits, a selection of samples was chosen for manual curation as described above to
603 attempt to complete the plasmid sequences and confirm circularity.

604 From the PebbleScout hits, SRR10902853 is listed as an RNA-seq dataset. To the best of
605 our knowledge, this sample and its corresponding Bioproject (PRJNA601687) are not affiliated
606 with a publication. We were unable to find an associated metagenomic dataset for this sample. We
607 mapped the trimmed reads to a concatenated reference of pMEGA1,2,3 and 4 using bowtie2 v2.5.4
608 and visualized the alignments in Geneious Prime v2025.1.1 to assess regions with high coverage.
609 For the clade 1-like plasmid, we assembled and manually curated a 109 kb fragment with >98%
610 nucleotide identity to pMEGA4. A cut-off of 100X coverage was chosen to classify ORFs with
611 high coverage for the clade 1-like genome, based on transcripts mapping to pMEGA3, and 500X
612 coverage was chosen as a comparable level of high expression in the clade 2-like genome, based
613 on transcripts mapping to pMEGA4 and the curated fragments. The clade 1-like plasmid had
614 approximately 5X as many transcripts mapped as the clade 2-like plasmid. From the assembly, we
615 detected ribosomal protein S3 sequences using hmmsearch with PF00189[144]. Contigs were mapped
616 to the *E. coli* OP50 genome (CP059137.1) with minimap2, and ANI was calculated with
617 skani[137,145].

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

619 After curation, we recovered 20 circular megaplasmid sequences (**Table S7**). These
620 sequences were re-oriented to start at a highly conserved protein sequence (ParE) and aligned using
621 Mauve v1.1.3 in Geneious Prime[146]. Skani v0.2.1 and MASH v2.2 dist were used to assess the

622 average nucleotide identity and distance between entire plasmid sequences[145,147]. The taxonomy
623 of predicted proteins within these plasmid genomes was summarized using Uniprot annotations
624 and tRep (https://github.com/MrOlm/tRep/tree/master/bin). All proteins across the 20
625 megaplasmids were clustered into protein subfamilies using MMseqs2 and an all-vs-all search
626 (cluster workflow with E-value 0.001, sensitivity: 7.5, fraction of covered residues across target
627 and query: 0.5, and the Set-Cover (greedy) cluster mode)[148]. A multiple sequence alignment was
628 generated for subfamilies with at least 2 members using (mmseqs result2msa) and HMM profiles
629 were generated using hhmake from HH-suite3[149]. A HH-Suite database was generated from these
630 protein subfamilies and subfamilies were compared against each other using their respective
631 HMMs and HHblits (with parameters -v 0 -p 50 -E 0.001 -z 1 -Z 32000 -B 0 -b 0 -n 2). Protein
632 families were generated from subfamilies with probability scores of at least 95% and coverage of
633 at least 0.70 using a Markov clustering algorithm[150]. Protein subfamilies were annotated based on
634 HMM-HMM comparisons against PFAM using HHsearch v.3.0.3 (parameters -p 50 -E 0.001 -z 1
635 -Z 32000 -b 0 -B 0 -n 1)[151,152].

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

637 Mobile genetic element prediction with geNomad v1.7.4 does not recognize the pMEGAs
638 as plasmids but predicts proviruses within the genome[153]. More extensive annotation was
639 performed for the pMEGA1-4 sequences including prediction of secretion systems and
640 conjugation machinery using hmmscan against the CONJScan and TXSS hmm databases with an
641 E value cut-off of 1e-10[154,155]. Origins of transfer (*oriT*) and vegetative replication (*oriV*) were
642 predicted by oriTFinder2, oriTDB[34] and OriV-Finder[156]. Repeats were identified manually in
643 Geneious Prime using RepeatFinder. Virulence factors and toxins were predicted for pMEGA4
644 and the *E. coli* genome using PathoFact v.2.0[157]. We used PADLOC v2.0.0 and structural

645 comparisons to identify antiviral defense systems and CRISPRCasFinder v.4.2.18 to detect
646 CRISPR spacers[158,159]. Antibiotic resistance mechanisms in pMEGA4 were predicted from protein
647 sequences using the *main* function of the Resistance Gene Identifier (RGI) v.6.0.3 with CARD
648 v4.0.0 (settings: --include_loose, --include_nudge, --low_quality and reporting hits with a bitscore
649 >100 regardless of ARO cutoff)[160]. For the *E. coli* genome and its 104 kb plasmid, RGI was used
650 in the default mode with the --low_quality flag. Codon usage was calculated using coRdon v 1.26.0
651 in R v 4.4.3[161].

652 For pMEGAs 1-4, AlphaFold2 was used to predict the protein structure of all open reading
653 frames and compared to a database of folded proteins from PDB (retrieved 2023-04-20) using
654 FoldSeek (v427df8a6b5d0ef78bee0f98cd3e6faaca18f172d)[162,163]. Some structures, as well as
655 multimers, were predicted using AlphaFold3[162]. Multimer confidence was evaluated by
656 consideration of ipTM scores. The top-scoring hits to PDB structures (generally >200 bit score)
657 were manually visualized in ChimeraX (v1.9) and position-resolved pLDDT values were used to
658 evaluate confidence. Alignment was performed using matchmaker or TM-align to confirm
659 structural similarity[164,165]. Foldseek matches to structures in the AlphaFold database were also
660 considered.

661 Multi-copy protein subfamilies were defined as subfamilies with greater than three
662 members within at least one genome of pMEGA1-4. Multi-copy proteins and clusters of two or
663 more proteins of the same subfamily were identified in Microsoft Excel and with custom scripts
664 in R v 4.4.3. The predicted structures of members within a multi-copy protein subfamily were
665 visualized to confirm their predicted function and to identify structures that contain alpha solenoid-
666 like folds/domains. Alpha-solenoid domains were identified by visual inspection, in some cases

667 flagged by best matches to DARPINs[42]. These proteins often had very low bit scores for best
668 matches to proteins in PDB.

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

670 To generate a phylogeny of known MOB_F relaxase proteins, sequences were gathered from
671 two comprehensive studies of conjugative plasmids and through an hmmscan for MOB_F sequences
672 in PLSDB and IMG/PR, resulting in 1854 hits[33,155,166,167]. These were clustered at 99% identity
673 using vsearch v2.13.3 (--cluster_fast) and combined with the predicted MOB_F relaxase sequences
674 from pMEGA1-4 (SF1156 and SF0489), resulting in 881 sequences[168]. These sequences were
675 aligned using MAFFT v7.505, then a phylogenetic tree was generated using FastTree v2.1.11,
676 rooted at the mid-point, and visualized using iTOL. A representative structure of SF0489
677 (pMEGA4_1394 - median pLDDT 90.19) and SF1156 (pMEGA4_58 - median pLDDT 90.5) was
678 aligned to RecD2 bound to ssDNA (PDB:3GP8)[169]. The TM alignment scores are 0.73 and 0.98,
679 respectively. Similarly, a phylogeny of 105 replication initiation proteins from plasmids of known
680 incompatibility groups (n=65)[170] was generated with predicted Rep sequences from pMEGAs
681 (SF1471 and SF0640) after clustering at 90% identity. Rep sequences from megaplasmids
682 identified in this study were collapsed as clades. Select protein structures were compared to the
683 pMEGA4 representative of SF640 and SF1471 using FoldSeek (bitscores as heatmaps on the right
684 of the tree). If not available in the AlphaFold database, the structures of Rep proteins were
685 predicted with AlphaFold3. For SF0640, pMEGA4_629 (median pLDDT 94.12) was aligned to
686 RepC from the *Salmonella enteritidis* IncQ plasmid pBLST (AFDB: A0A1D8X722) with TM-
687 align (score 0.86). For SF1471, pMEGA4_1423 (median pLDDT 91.5) was aligned to TrfA from
688 the *Escherichia coli* IncP plasmid pRK2 (AFDB: P07676) with TM-align (score 0.57). Finally,
689 Rep protein structures were compared to the Rep protein from E. coli F plasmid bound to an iteron

690 sequence (PDB: 1REP). The TM-align score with 1REP for pMEGA4_629 was 0.68 and for
691 pMEGA4_1423 was 0.70.

692 **Host prediction through abundance correlation, DNA modification analysis, and Hi-C**
693 **linkage**

694 From the searches of the SRA, we identified *E. coli* and *Salmonella spp.* isolates with
695 evidence of a megaplasmid. For one *E. coli* isolate, it was re-grown, DNA was re-isolated, and
696 sequenced via Nanopore confirming the presence of the 1.4 Mb plasmid (pMEGA4). For infants
697 with longitudinal samples (pMEGA1, pMEGA2, pMEGA3), we measured the Pearson Correlation
698 in R v. 4.4.3 between the percentage of reads mapped to the megaplasmid and all bacterial MAGs
699 over multiple sampling timepoints. For isolates with evidence of a medium- or high-quality
700 megaplasmid, plasmid and potential chromosomal contigs were separated and binned manually
701 based on GC content, coverage and similarity to pMEGA1 or 2. The megaplasmid copy number
702 (PCN) was estimated by comparing the mean depth of Illumina read coverage of the plasmid and
703 draft isolate genome as determined with CoverM v0.7.0 using BBMap v39.10[120,132] (**Table S5**).

704 To detect DNA methylation from PacBio data, we computed the mean inter-pulse duration
705 (IPD) at each base across the mapped reads. We empirically estimated baseline IPD values for
706 each k-mer by comparing to the mean IPD values of the kmer from the metagenome. Bases with
707 significantly elevated IPD ratios are inferred to be modified. Subsequently, motif discovery is
708 carried out using MotifMaker (SMRT Link v13.1.0.221970). We then scanned each contig to
709 evaluate the modified fraction of detected motifs, ultimately generating a genome-wide
710 methylation profile. In addition, to annotate restriction endonucleases (REases) and
711 methyltransferases (MTases) in the assemblies, we employed the annotate_rm module of
712 MicrobeMod with default parameters[171]. Finally, the ProxiMeta™ Metagenome Deconvolution

713 Platform (Hi-C) from Phase Genomics (Seattle, USA) was used to detect Hi-C linkages on the
714 week 4 stool samples from INF1330004 and INF1340011.

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

716 CRISPR spacers detected with CRISPRCasFinder v.4.2.18 from all 20 confident
717 megaplasmid genomes were searched against their respective metagenomes using blastn-short (-
718 perc_identity 97, -qcov_hsp_perc 100)[133]. In addition, CRISPR spacers predicted within the 20
719 metagenomic samples from which the plasmid genomes originated were searched against the
720 megaplasmid genomes to identify potential hosts. Finally, CRISPR spacers from the 20
721 megaplasmid genomes were compared to PLSDB and IMG/PR and IMG/VR representing over
722 772,529 plasmid and 15,722,824 viral sequences[21]. Spacer interaction networks were visualized in
723 CytoScape v 3.10.3. Finally, to infer potential hosts based on spacer acquisition against pMEGAs,
724 we blasted over 322 million predicted spacers from 580,000 bacterial genomes[38] and ~450,000
725 public metagenomes[39] against pMEGA1-20 and evaluated spacers with 100% identity hits.

726 **Comparing to other plasmid databases**

727 To ensure our megaplasmid elements had not been previously characterized, we searched
728 for similar elements in, to the best of our knowledge, all relevant and known plasmid and viral
729 sequence databases including PLSDB (v2024-05-31_v2 - 72,556 sequences), IMG/PR (699,973
730 sequences), IMG/VR (15,722,824 sequences), 226,194 plasmids predicted from the human gut [130],
731 as well as datasets from two recent studies that characterized plasmids (n = 265,476) and viruses
732 (n = 160,478) from infants[172,173] using MASH dist with pMEGA1 and pMEGA2[147]. To ensure that
733 no plasmids of similar size to the pMEGAs have been described in *Enterobacteriaceae*, we
734 analyzed all plasmid sequences >1Mb from PLSDB and IMG/PR, and RefSeq *Enterobacteriaceae*

735 genomes with secondary replicons >1Mb. Many of these plasmids were derived from isolates. If
736 a plasmid was derived from a metagenomic sample, we annotated the taxonomy of the host based
737 on ANI to other plasmids or consensus taxonomy of predicted proteins. Finally, to prevent
738 inclusion of mis-annotated plasmids or secondary chromosomes, we used CheckM2 to remove
739 sequences with 50% or greater bacterial genome completeness[174]. Predicted proteins were
740 annotated following the same basic pipeline for pMEGAs, and the percent of hypothetical,
741 unknown, or uncharacterized proteins as determined by UniRef annotations was counted for both
742 the largest contig of each isolate assembly and the putative >1Mb plasmid.

743 **References:**

744 1. Pilla, G. & Tang, C. M. Going around in circles: virulence plasmids in enteric pathogens. *Nat.*
745 *Rev. Microbiol.* **16**, 484–495 (2018).
746 2. Nora, L. C. *et al.* The art of vector engineering: towards the construction of next-generation
747 genetic tools. *Microb. Biotechnol.* **12**, 125–147 (2019).
748 3. Stockdale, S. R. & Hill, C. Incorporating plasmid biology and metagenomics into a holistic
749 model of the human gut microbiome. *Curr. Opin. Microbiol.* **73**, 102307 (2023).
750 4. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San Millán, Á.
751 Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat. Rev.*
752 *Microbiol.* **19**, 347–359 (2021).
753 5. diCenzo, G. C. & Finan, T. M. The divided bacterial genome: Structure, function, and
754 evolution. *Microbiol. Mol. Biol. Rev.* **81**, (2017).
755 6. Maddamsetti, R. *et al.* Scaling laws of bacterial and archaeal plasmids. *Nat. Commun.* **16**,
756 (2025).
757 7. Ramiro-Martínez, P., de Quinto, I., Lanza, V. F., Gama, J. A. & Rodríguez-Beltrán, J.

758 Universal rules govern plasmid copy number. *Nat. Commun.* **16**, (2025).

759 8. Czarnecki, J. *et al.* Replication coordination marks the domestication of large
760 extrachromosomal replicons in bacteria. *bioRxiv* (2025) doi:10.1101/2025.03.15.643453.

761 9. Genome sequence of the plant pathogen Ralstonia solanacearum. *Nature* (2002)
762 doi:10.1038/news020128-8.

763 10. Janssen, P. J. *et al.* The complete genome sequence of Cupriavidus metallidurans strain CH34,
764 a master survivalist in harsh and anthropogenic environments. *PLoS One* **5**, e10433 (2010).

765 11. Mergeay, M. & Van Houdt, R. Cupriavidus metallidurans CH34, a historical perspective on
766 its discovery, characterization and metal resistance. *FEMS Microbiol. Ecol.* **97**, (2021).

767 12. diCenzo, G. C., Mengoni, A. & Perrin, E. Chromids aid genome expansion and functional
768 diversification in the family Burkholderiaceae. *Mol. Biol. Evol.* **36**, 562–574 (2019).

769 13. Fournes, F., Val, M.-E., Skovgaard, O. & Mazel, D. Replicate once per cell cycle: Replication
770 control of secondary chromosomes. *Front. Microbiol.* **9**, 1833 (2018).

771 14. Harrison, P. W., Lower, R. P. J., Kim, N. K. D. & Young, J. P. W. Introducing the bacterial
772 "chromid": not a chromosome, not a plasmid. *Trends Microbiol.* **18**, 141–148 (2010).

773 15. Hall, J. P. J., Botelho, J., Cazares, A. & Baltrus, D. A. What makes a megaplasmid? *Philos.*
774 *Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200472 (2022).

775 16. Liu, H. *et al.* Unexplored diversity and potential functions of extra-chromosomal elements.
776 *mSystems* e0017525 (2025).

777 17. Kohlmeier, M. G., O'Hara, G. W., Ramsay, J. P. & Terpolilli, J. J. Closed genomes of
778 commercial inoculant rhizobia provide a blueprint for management of legume inoculation.
779 *Appl. Environ. Microbiol.* **91**, e0221324 (2025).

780 18. Medema, M. H. *et al.* The sequence of a 1.8-mb bacterial linear plasmid reveals a rich

781 evolutionary reservoir of secondary metabolic pathways. *Genome Biol. Evol.* **2**, 212–224
782 (2010).
783 19. Kothari, A. *et al.* Large circular plasmids from groundwater plasmidomes span multiple
784 incompatibility groups and are enriched in multimetal resistance genes. *MBio* **10**, (2019).
785 20. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global
786 map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).
787 21. Schmartz, G. P. *et al.* PLSDB: advancing a comprehensive database of bacterial plasmids.
788 *Nucleic Acids Res.* **50**, D273–D278 (2022).
789 22. Camargo, A. P. *et al.* IMG/PR: a database of plasmids from genomes and metagenomes with
790 rich annotations and metadata. *Nucleic Acids Res.* **52**, D164–D173 (2024).
791 23. Li, Y. *et al.* PlasmidScope: a comprehensive plasmid database with rich annotations and
792 online analytical tools. *Nucleic Acids Res.* **53**, D179–D188 (2025).
793 24. Douarre, P.-E., Mallet, L., Radomski, N., Felten, A. & Mistou, M.-Y. Analysis of COMPASS,
794 a new comprehensive Plasmid database revealed prevalence of multireplicon and extensive
795 diversity of IncF plasmids. *Front. Microbiol.* **11**, 483 (2020).
796 25. Koraimann, G. Spread and persistence of virulence and antibiotic resistance genes: A ride on
797 the F Plasmid conjugation module. *EcoSal Plus* **8**, (2018).
798 26. Ares-Arroyo, M., Rocha, E. P. C. & Gonzalez-Zorn, B. Evolution of ColE1-like plasmids
799 across γ-Proteobacteria: From bacteriocin production to antimicrobial resistance. *PLoS Genet.*
800 **17**, (2021).
801 27. Arredondo-Alonso, S. *et al.* Plasmid-driven strategies for clone success in Escherichia coli.
802 *Nat. Commun.* **16**, 2921 (2025).
803 28. Brooks, B. *et al.* Strain-resolved analysis of hospital rooms and infants reveals overlap

804 between the human and room microbiome. *Nat. Commun.* **8**, 1814 (2017).

805 29. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-
806 section birth. *Nature* **574**, 117–121 (2019).

807 30. Chen, Y. *et al.* Draft Genome Sequences of Citrobacter freundii and Citrobacter murliniae
808 Strains Isolated from the Feces of Preterm Infants. *Microbiol. Resour. Announc.* **8**, (2019).

809 31. Sudatip, D. *et al.* The risk of pig and chicken farming for carriage and transmission of
810 Escherichia coli containing extended-spectrum beta-lactamase (ESBL) and mobile colistin
811 resistance (mcr) genes in Thailand. *Microb. Genom.* **9**, (2023).

812 32. Mishra, D. & Srinivasan, R. Catching a Walker in the act-DNA partitioning by ParA family
813 of proteins. *Front. Microbiol.* **13**, 856547 (2022).

814 33. Coluzzi, C., Garcillán-Barcia, M. P., de la Cruz, F. & Rocha, E. P. C. Evolution of Plasmid
815 mobility: Origin and fate of conjugative and nonconjugative plasmids. *Mol. Biol. Evol.* **39**,
816 (2022).

817 34. Liu, G. *et al.* oriTDB: a database of the origin-of-transfer regions of bacterial mobile genetic
818 elements. *Nucleic Acids Res.* **53**, D163–D168 (2025).

819 35. Konieczny, I., Bury, K., Wawrzycka, A. & Wegrzyn, K. Iteron plasmids. *Microbiol. Spectr.*
820 **2**, (2014).

821 36. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility
822 of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).

823 37. Pausch, P. *et al.* Structural variation of type I-F CRISPR RNA guided DNA surveillance. *Mol.*
824 *Cell* **67**, 622-632.e4 (2017).

825 38. Dion, M. B. *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher
826 the viral dark matter. *Nucleic Acids Res.* **49**, 3127–3138 (2021).

827 39. Roux, S. *et al.* Planetary-scale metagenomic search reveals new patterns of CRISPR targeting.
828 *bioRxiv* (2025) doi:10.1101/2025.06.12.659409.
829 40. Li, R., Peng, K., Li, Y., Liu, Y. & Wang, Z. Exploring tet(X)-bearing tigecycline-resistant
830 bacteria of swine farming environments. *Sci. Total Environ.* **733**, 139306 (2020).
831 41. Riccardi, C. *et al.* Independent origins and evolution of the secondary replicons of the class
832 Gammaproteobacteria. *Microb. Genom.* **9**, (2023).
833 42. Plückthun, A. Designed ankyrin repeat proteins (DARPins): binding proteins for research,
834 diagnostics, and therapy. *Annu. Rev. Pharmacol. Toxicol.* **55**, 489–511 (2015).
835 43. Gong, Y. *et al.* A nuclease-toxin and immunity system for kin discrimination in Myxococcus
836 xanthus. *Environ. Microbiol.* **20**, 2552–2567 (2018).
837 44. Zhang, D., Iyer, L. M. & Aravind, L. A novel immunity system for bacterial nucleic acid
838 degrading toxins and its recruitment in various eukaryotic and DNA viral systems. *Nucleic*
839 *Acids Res.* **39**, 4532–4552 (2011).
840 45. Zhang, D., de Souza, R. F., Anantharaman, V., Iyer, L. M. & Aravind, L. Polymorphic toxin
841 systems: Comprehensive characterization of trafficking modes, processing, mechanisms of
842 action, immunity and ecology using comparative genomics. *Biol. Direct* **7**, 18 (2012).
843 46. Ruhe, Z. C., Low, D. A. & Hayes, C. S. Polymorphic toxins and their immunity proteins:
844 Diversity, evolution, and mechanisms of delivery. *Annu. Rev. Microbiol.* **74**, 497–520 (2020).
845 47. Liu, W., Schoonen, M., Wang, T., McSweeney, S. & Liu, Q. Cryo-EM structure of
846 transmembrane AAA+ protease FtsH in the ADP state. *Commun. Biol.* **5**, 257 (2022).
847 48. Bittner, L.-M., Arends, J. & Narberhaus, F. When, how and why? Regulated proteolysis by
848 the essential FtsH protease in Escherichia coli. *Biol. Chem.* **398**, 625–635 (2017).
849 49. Erbse, A. *et al.* ClpS is an essential component of the N-end rule pathway in Escherichia coli.

850 *Nature* **439**, 753–756 (2006).

851 50. Deville, C., Franke, K., Mogk, A., Bukau, B. & Saibil, H. R. Two-step activation mechanism
852 of the ClpB disaggregase for sequential substrate threading by the main ATPase motor. *Cell*
853 *Rep.* **27**, 3433-3446.e4 (2019).

854 51. Ma, C. *et al.* Structural insights into the membrane microdomain organization by SPFH family
855 proteins. *Cell Res.* **32**, 176–189 (2022).

856 52. Lopez, D. & Koch, G. Exploring functional membrane microdomains in bacteria: an
857 overview. *Curr. Opin. Microbiol.* **36**, 76–84 (2017).

858 53. Wallden, K., Rivera-Calzada, A. & Waksman, G. Type IV secretion systems: versatility and
859 diversity in function. *Cell. Microbiol.* **12**, 1203–1212 (2010).

860 54. Cascales, E. & Cambillau, C. Structural biology of type VI secretion systems. *Philos. Trans.*
861 *R. Soc. Lond. B Biol. Sci.* **367**, 1102–1111 (2012).

862 55. Hu, J. *et al.* Cryo-EM analysis of the T3S injectisome reveals the structure of the needle and
863 open secretin. *Nat. Commun.* **9**, 3840 (2018).

864 56. Hajam, I. A., Dar, P. A., Shahnawaz, I., Jaume, J. C. & Lee, J. H. Bacterial flagellin-a potent
865 immunomodulatory agent. *Exp. Mol. Med.* **49**, e373 (2017).

866 57. Haiko, J. & Westerlund-Wikström, B. The role of the bacterial flagellum in adhesion and
867 virulence. *Biology (Basel)* **2**, 1242–1267 (2013).

868 58. Ronneau, S. & Hallez, R. Make and break the alarmone: regulation of (p)ppGpp
869 synthetase/hydrolase enzymes in bacteria. *FEMS Microbiol. Rev.* **43**, 389–400 (2019).

870 59. Calhoun, L. N. & Kwon, Y. M. Structure, function and regulation of the DNA-binding protein
871 Dps and its role in acid and oxidative stress resistance in Escherichia coli: a review. *J. Appl.*
872 *Microbiol.* **110**, 375–386 (2011).

873 60. Janssens, A. *et al.* SlyB encapsulates outer membrane proteins in stress-induced lipid
874 nanodomains. *Nature* **626**, 617–625 (2024).
875 61. Srikannathasan, V. *et al.* Structural basis for type VI secreted peptidoglycan DL-
876 endopeptidase function, specificity and neutralization in Serratia marcescens. *Acta*
877 *Crystallogr. D Biol. Crystallogr.* **69**, 2468–2482 (2013).
878 62. Ting, S.-Y. *et al.* Bifunctional immunity proteins protect bacteria against FtsZ-targeting ADP-
879 ribosylating toxins. *Cell* **175**, 1380-1392.e14 (2018).
880 63. Bishop, R. E., Leskiw, B. K., Hodges, R. S., Kay, C. M. & Weiner, J. H. The entericidin locus
881 of Escherichia coli and its implications for programmed bacterial cell death. *J. Mol. Biol.* **280**,
882 583–596 (1998).
883 64. Schmidt, O. *et al.* prlF and yhaV encode a new toxin-antitoxin system in Escherichia coli. *J.*
884 *Mol. Biol.* **372**, 894–905 (2007).
885 65. Hansen, S. *et al.* Regulation of the Escherichia coli HipBA toxin-antitoxin system by
886 proteolysis. *PLoS One* **7**, e39185 (2012).
887 66. Tripathi, A., Dewan, P. C., Ahmed, S. & Varadarajan, R. MazF-induced growth inhibition
888 and persister generation in Escherichia coli. *J. Biol. Chem.* **289**, 4191–4205 (2014).
889 67. May, J. P. & Simon, A. E. Targeting of viral RNAs by Upf1-mediated RNA decay pathways.
890 *Curr. Opin. Virol.* **47**, 1–8 (2021).
891 68. Millman, A. *et al.* An expanding arsenal of immune systems that protect bacteria from phages.
892 *bioRxiv* (2022) doi:10.1101/2022.05.11.491447.
893 69. Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C. & Laub, M. T. A functional
894 selection reveals previously undetected anti-phage defence systems in the E. coli pangenome.
895 *Nat. Microbiol.* **7**, 1568–1579 (2022).

896 70. Gao, L. *et al.* Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*
897 **369**, 1077–1084 (2020).
898 71. Fitzpatrick, A. W. P. *et al.* Structure of the MacAB–TolC ABC-type tripartite multidrug efflux
899 pump. *Nat. Microbiol.* **2**, 17070 (2017).
900 72. Caveney, N. A. *et al.* Structural insight into YcbB-mediated beta-lactam resistance in
901 Escherichia coli. *Nat. Commun.* **10**, 1849 (2019).
902 73. San Millan, A. & MacLean, R. C. Fitness costs of plasmids: A limit to Plasmid transmission.
903 *Microbiol. Spectr.* **5**, (2017).
904 74. Kamal, S. M. *et al.* A recently isolated human commensal Escherichia coli ST10 clone
905 member mediates enhanced thermotolerance and tetrathionate respiration on a P1 phage-
906 derived IncY plasmid. *Mol. Microbiol.* **115**, 255–271 (2021).
907 75. Lee, C. *et al.* Why? - Successful Pseudomonas aeruginosa clones with a focus on clone C.
908 *FEMS Microbiol. Rev.* **44**, 740–762 (2020).
909 76. Bojer, M. S., Struve, C., Ingmer, H., Hansen, D. S. & Krogfelt, K. A. Heat resistance mediated
910 by a new plasmid encoded Clp ATPase, ClpK, as a possible novel mechanism for nosocomial
911 persistence of Klebsiella pneumoniae. *PLoS One* **5**, e15467 (2010).
912 77. Motiwala, T. *et al.* Caseinolytic proteins (Clp) in the genus Klebsiella: Special focus on ClpK.
913 *Molecules* **27**, 200 (2021).
914 78. Kamal, S. M. *et al.* Two FtsH proteases contribute to fitness and adaptation of Pseudomonas
915 aeruginosa clone C strains. *Front. Microbiol.* **10**, 1372 (2019).
916 79. Mawla, G. D. *et al.* The membrane-cytoplasmic linker defines activity of FtsH proteases in
917 Pseudomonas aeruginosa clone C. *J. Biol. Chem.* **300**, 105622 (2024).
918 80. Stein, B. J., Grant, R. A., Sauer, R. T. & Baker, T. A. Structural basis of an N-degron adaptor

919 with more stringent specificity. *Structure* **24**, 232–242 (2016).

920 81. Qiao, Z. *et al.* Cryo-EM structure of the entire FtsH-HflKC AAA protease complex. *Cell Rep.*
921 **39**, 110890 (2022).

922 82. Choi, M., Karunaratne, K. & Kohen, A. Flavin-dependent thymidylate synthase as a new
923 antibiotic target. *Molecules* **21**, 654 (2016).

924 83. *Bacterial Enzymes as Targets for Drug Discovery*. (Academic Press, San Diego, CA, 2024).

925 84. Zhao, J. *et al.* Preclinical safety and efficacy characterization of an LpxC inhibitor against
926 Gram-negative pathogens. *Sci. Transl. Med.* **15**, eadf5668 (2023).

927 85. Osterman, I. *et al.* Phages reconstitute NAD+ to counter bacterial immunity. *Nature* **634**,
928 1160–1167 (2024).

929 86. Kaufmann, G. *et al.* RloC: A translation-disabling tRNase implicated in phage exclusion
930 during recovery from DNA damage. in *DNA Repair* (InTech, 2011).

931 87. van den Berg, D. F., van der Steen, B. A., Costa, A. R. & Brouns, S. J. J. Phage tRNAs evade
932 tRNA-targeting host defenses through anticodon loop mutations. *Elife* **12**, (2023).

933 88. Korotkov, K. V., Gonen, T. & Hol, W. G. J. Secretins: dynamic channels for protein transport
934 across membranes. *Trends Biochem. Sci.* **36**, 433–443 (2011).

935 89. Leighton, T. L., Buensuceso, R. N. C., Howell, P. L. & Burrows, L. L. Biogenesis of
936 Pseudomonas aeruginosa type IV pili and regulation of their function. *Environ. Microbiol.* **17**,
937 4148–4163 (2015).

938 90. Hoffmann, L. *et al.* Structure and interactions of the archaeal motility repression module
939 ArnA-ArnB that modulates archaellum gene expression in Sulfolobus acidocaldarius. *J. Biol.*
940 *Chem.* **294**, 7460–7471 (2019).

941 91. Heidler, T. V., Ernits, K., Ziolkowska, A., Claesson, R. & Persson, K. Porphyromonas

942 gingivalis fimbrial protein Mfa5 contains a von Willebrand factor domain and an
943 intramolecular isopeptide. *Commun. Biol.* **4**, 106 (2021).
944 92. Patel, D. T. *et al.* Global atlas of predicted functional domains in Legionella pneumophila
945 Dot/Icm translocated effectors. *Mol. Syst. Biol.* **21**, 59–89 (2025).
946 93. Al-Khodor, S., Price, C. T., Habyarimana, F., Kalia, A. & Abu Kwaik, Y. A Dot/Icm-
947 translocated ankyrin protein of Legionella pneumophila is required for intracellular
948 proliferation within human macrophages and protozoa. *Mol. Microbiol.* **70**, 908–923 (2008).
949 94. Schmitz-Esser, S. *et al.* The genome of the amoeba symbiont "Candidatus Amoebophilus
950 asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated
951 bacteria. *J. Bacteriol.* **192**, 1045–1057 (2010).
952 95. Walser, M., Mayor, J. & Rothenberger, S. Designed ankyrin repeat proteins: A new class of
953 viral entry inhibitors. *Viruses* **14**, 2242 (2022).
954 96. Al-Khodor, S., Price, C. T., Kalia, A. & Abu Kwaik, Y. Functional diversity of ankyrin repeats
955 in microbial proteins. *Trends Microbiol.* **18**, 132–139 (2010).
956 97. Jahn, M. T. *et al.* A phage protein aids bacterial symbionts in eukaryote immune evasion. *Cell*
957 *Host Microbe* **26**, 542-550.e5 (2019).
958 98. Self, J. L. *et al.* Epidemiology of salmonellosis among infants in the United States: 1968-
959 2015. *Pediatrics* **151**, (2023).
960 99. Hoffman, A. *et al.* Predictors of mortality and severe illness from Escherichia coli sepsis in
961 neonates. *J. Perinatol.* **44**, 1816–1821 (2024).
962 100. Yang, J.-T., Zhang, L.-J., Lu, Y., Zhang, R.-M. & Jiang, H.-X. Genomic insights into global
963 blaCTX-M-55-positive Escherichia coli epidemiology and transmission characteristics.
964 *Microbiol. Spectr.* **11**, e0108923 (2023).

965 101. Day, M. J. *et al.* Extended-spectrum β-lactamase-producing Escherichia coli in human-
966 derived and foodchain-derived samples from England, Wales, and Scotland: an
967 epidemiological surveillance and typing study. *Lancet Infect. Dis.* **19**, 1325–1335 (2019).
968 102. Falgenhauer, L. *et al.* Detection and characterization of ESBL-producing Escherichia coli
969 from humans and poultry in Ghana. *Front. Microbiol.* **9**, 3358 (2018).
970 103. Lupo, A., Saras, E., Madec, J.-Y. & Haenni, M. Emergence of blaCTX-M-55 associated with
971 fosA, rmtB and mcr gene variants in Escherichia coli from various animal species in France.
972 *J. Antimicrob. Chemother.* **73**, 867–872 (2018).
973 104. Hu, Y.-Y. *et al.* Molecular typing of CTX-M-producing escherichia coli isolates from
974 environmental water, swine feces, specimens from healthy humans, and human patients. *Appl.*
975 *Environ. Microbiol.* **79**, 5988–5996 (2013).
976 105. Lim, J. Y., Yoon, J. & Hovde, C. J. A brief overview of Escherichia coli O157:H7 and its
977 plasmid O157. *J. Microbiol. Biotechnol.* **20**, 5–14 (2010).
978 106. Chen, J. C. *et al.* Reoccurring Escherichia coli O157:H7 strain linked to leafy greens-
979 associated outbreaks, 2016-2019. *Emerg. Infect. Dis.* **29**, 1895–1899 (2023).
980 107. Kizny Gordon, A. E. *et al.* The hospital water environment as a reservoir for carbapenem-
981 resistant organisms causing hospital-acquired infections-A systematic review of the literature.
982 *Clin. Infect. Dis.* **64**, 1435–1444 (2017).
983 108. Morowitz, M. J. *et al.* The NICU Antibiotics and Outcomes (NANO) trial: a randomized
984 multicenter clinical trial assessing empiric antibiotics and clinical outcomes in newborn
985 preterm infants. *Trials* **23**, 428 (2022).
986 109. Joshi NA, F. J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ
987 files (Version 1.33) [Software]. *https://github.com/najoshi/sickle* (2011).

988 110. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
989 357–359 (2012).
990 111. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
991 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
992 112. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
993 identification. *BMC Bioinformatics* **11**, 119 (2010).
994 113. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
995 *Res.* **28**, 27–30 (2000).
996 114. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive
997 and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
998 115. UniProt Consortium. UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids*
999 *Res.* **51**, D523–D531 (2023).
1000 116. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*
1001 **26**, 2460–2461 (2010).
1002 117. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain
1003 Bacteria. *Nature* **523**, 208–211 (2015).
1004 118. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments.
1005 *Bioinformatics* **25**, 1713–1713 (2009).
1006 119. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA
1007 genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
1008 120. Bushnell, B. *BBMap*.
1009 121. Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads
1010 with hifiasm-meta. *Nat. Methods* **19**, 671–674 (2022).

1011 122. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication,
1012 aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
1013 123. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to
1014 recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
1015 124. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods*
1016 **11**, 1144–1146 (2014).
1017 125. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational
1018 autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
1019 126. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
1020 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
1021 127. Portik, D. M. *et al.* Highly accurate metagenome-assembled genomes from human gut
1022 microbiota using long-read assembly, binning, and consolidation methods. *bioRxiv* (2024)
1023 doi:10.1101/2024.05.10.593587.
1024 128. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate
1025 genomic comparisons that enables improved genome recovery from metagenomes through
1026 de-replication. *ISME J.* **11**, 2864–2868 (2017).
1027 129. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from
1028 metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
1029 130. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their ecology across
1030 human gut metagenomes revealed by PlasX and MobMess. *Nat. Microbiol.* **9**, 830–847
1031 (2024).
1032 131. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and
1033 complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).

1034 132. Aroney, S. T. N. *et al.* CoverM: read alignment statistics for metagenomics. *Bioinformatics*
1035 **41**, (2025).
1036 133. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
1037 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
1038 134. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology
1039 Information. Protein [Internet]. (1988).
1040 135. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information
1041 in 2025. *Nucleic Acids Res.* **53**, D20–D29 (2025).
1042 136. Sudatip, D. *et al.* A One Health approach to assessing occupational exposure to antimicrobial
1043 resistance in Thailand: The FarmResist project. *PLoS One* **16**, e0245250 (2021).
1044 137. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100
1045 (2018).
1046 138. Nanoporetech. *Dorado*. (2024).
1047 139. Shiryev, S. A. & Agarwala, R. Indexing and searching petabase-scale nucleotide resources.
1048 *Nat. Methods* **21**, 994–1002 (2024).
1049 140. Chikhi, R., Raffestin, B., Korobeynikov, A., Edgar, R. & Babaian, A. Logan: Planetary-scale
1050 genome assembly surveys life's diversity. *bioRxiv* (2024) doi:10.1101/2024.07.30.605881.
1051 141. Constantinides, B. *et al.* Genomic surveillance of Escherichia coli and Klebsiella spp. in
1052 hospital sink drains and patients. *Microb. Genom.* **6**, (2020).
1053 142. Zhang, A. *et al.* Infants exposed in utero to Hurricane Maria have gut microbiomes with
1054 reduced diversity and altered metabolic capacity. *mSphere* **8**, e0013423 (2023).
1055 143. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1056 2079 (2009).

1057 144. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204
1058 (2018).
1059 145. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse
1060 chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
1061 146. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of
1062 conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
1063 147. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash.
1064 *Genome Biol.* **17**, (2016).
1065 148. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
1066 analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
1067 149. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein
1068 annotation. *BMC Bioinformatics* **20**, 473 (2019).
1069 150. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale
1070 detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
1071 151. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–
1072 960 (2005).
1073 152. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–
1074 D419 (2021).
1075 153. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat.*
1076 *Biotechnol.* **42**, 1303–1312 (2024).
1077 154. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**,
1078 23080 (2016).
1079 155. Cury, J., Abby, S. S., Doppelt-Azeroual, O., Néron, B. & Rocha, E. P. C. Identifying

1080 conjugative plasmids and integrative conjugative elements with CONJscan. *Methods Mol.*
1081 *Biol.* **2075**, 265–283 (2020).
1082 156. Li, Y. & Gao, F. OriV-Finder: a comprehensive web server for bacterial plasmid replication
1083 origin analysis. *Nucleic Acids Res.* **53**, W451–W456 (2025).
1084 157. Sunyer, J. O. *et al.* PathoFact 2.0: An integrative pipeline for predicting antimicrobial
1085 resistance genes, virulence factors, toxins and biosynthetic gene clusters in metagenomes.
1086 *bioRxiv* (2024) doi:10.1101/2024.12.09.627531.
1087 158. Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version,
1088 enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–
1089 W251 (2018).
1090 159. Payne, L. J. *et al.* Identification and classification of antiviral defence systems in bacteria and
1091 archaea with PADLOC reveals new system types. *Nucleic Acids Res.* **49**, 10868–10878
1092 (2021).
1093 160. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and
1094 resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids*
1095 *Res.* **51**, D690–D699 (2023).
1096 161. Vlahovicek, A. E. M. K. *CoRdon*. (Bioconductor, 2018). doi:10.18129/B9.BIOC.CORDON.
1097 162. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold
1098 3. *Nature* **630**, 493–500 (2024).
1099 163. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat.*
1100 *Biotechnol.* **42**, 243–246 (2024).
1101 164. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.*
1102 **32**, e4792 (2023).

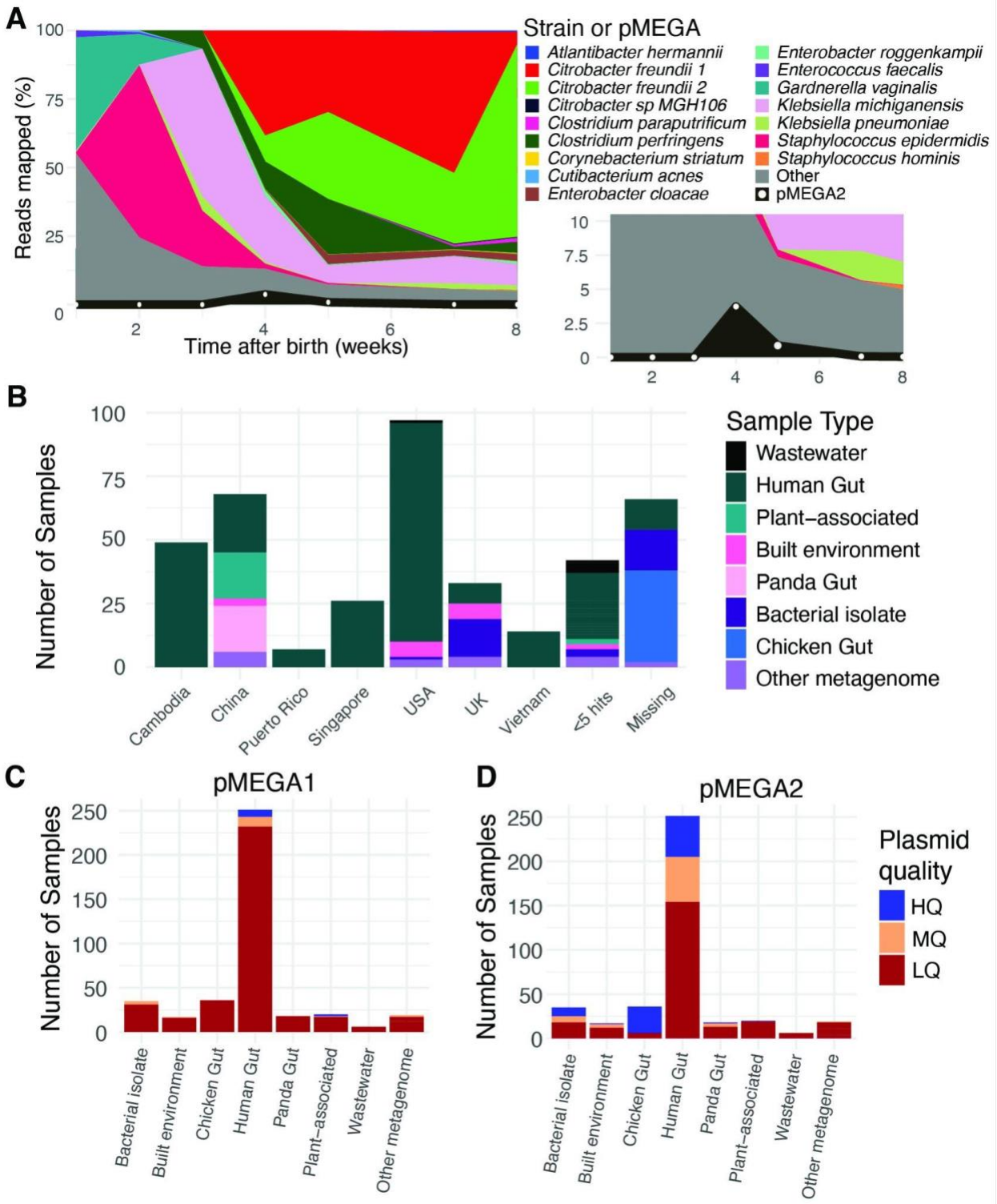
1103 165. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-
1104 score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
1105 166. Garcillán-Barcia, M. P., Francia, M. V. & de la Cruz, F. The diversity of conjugative relaxases
1106 and its application in plasmid classification. *FEMS Microbiol. Rev.* **33**, 657–687 (2009).
1107 167. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
1108 168. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source
1109 tool for metagenomics. *PeerJ* (2016) doi:10.7287/peerj.preprints.2409.
1110 169. Saikrishnan, K., Powell, B., Cook, N. J., Webb, M. R. & Wigley, D. B. Mechanistic basis of
1111 5'-3' translocation in SF1B helicases. *Cell* **137**, 849–859 (2009).
1112 170. Suzuki, H., Yano, H., Brown, C. J. & Top, E. M. Predicting plasmid promiscuity based on
1113 genomic signature. *J. Bacteriol.* **192**, 6045–6055 (2010).
1114 171. Crits-Christoph, A., Kang, S. C., Lee, H. H. & Ostrov, N. MicrobeMod: A computational
1115 toolkit for identifying prokaryotic methylation and restriction-modification with nanopore
1116 sequencing. *bioRxiv* (2023) doi:10.1101/2023.11.13.566931.
1117 172. He, W., Russel, J., Klincke, F., Nesme, J. & Sørensen, S. J. Insights into the ecology of the
1118 infant gut plasmidome. *Nat. Commun.* **15**, 6924 (2024).
1119 173. Zeng, S. *et al.* A metagenomic catalog of the early-life human gut virome. *Nat. Commun.* **15**,
1120 1864 (2024).
1121 174. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and
1122 accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods*
1123 **20**, 1203–1212 (2023).
1124 175. Kreutzberger, M. A. B. *et al.* Convergent evolution in the supercoiling of prokaryotic flagellar
1125 filaments. *Cell* **185**, 3487-3500.e14 (2022).

1126 **TABLES AND FIGURES**

1127 **Table 1**: Listing of 4 complete plasmids (name, size, GC content, source). Completion was
1128 determined by mapping reads to a single contig and ensuring complete coverage as well as
1129 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

1130

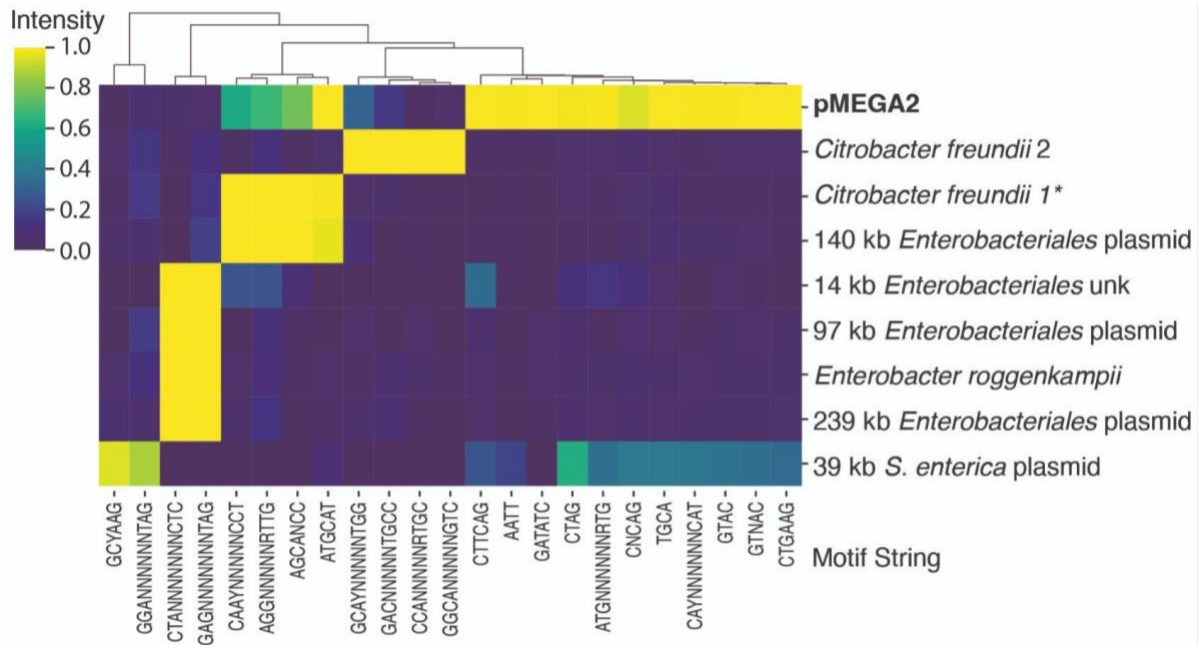


1131

1132 **Figure 1: Identification of megaplasmids in the preterm infant gut microbiome and public**

1133 **data. A)** The percentage of reads mapped to MAGs and pMEGA2 within INF1330004 was

1134 determined through Illumina read mapping. The "Other" category includes reads that did not map
1135 to MAGs nor pMEGA2. Inset shows the same plot but with y values from 0 to 10. **B)** Sources of
1136 samples with evidence of related megaplasmids, grouped by source location and sample type
1137 (Missing indicates no information in the SRA). Source types with fewer than 5 hits were grouped
1138 together. **C), D)** The number and quality of pMEGA1(**C**)- and pMEGA2(**D**)-like plasmids across
1139 different sample types. Assembled plasmids were classified as high-quality (HQ - blue) if they had
1140 ≥85% coverage of a pMEGA and ≤10 contigs and medium-quality (MQ - orange) with ≥85%
1141 coverage and 11-50 contigs. The remaining samples were classified as low-quality (LQ - red).

1142 

1143 **Figure 2: A *Citrobacter freundii* is the host of pMEGA2.** Modified bases were predicted in the
1144 PacBio HiFi reads of infant 1330004 at week 4. The columns represent the motifs predicted to
1145 contain modified bases, and the rows represent genomes. *Likely host of pMEGA2 based on
1146 overlap in modification patterns.

1147

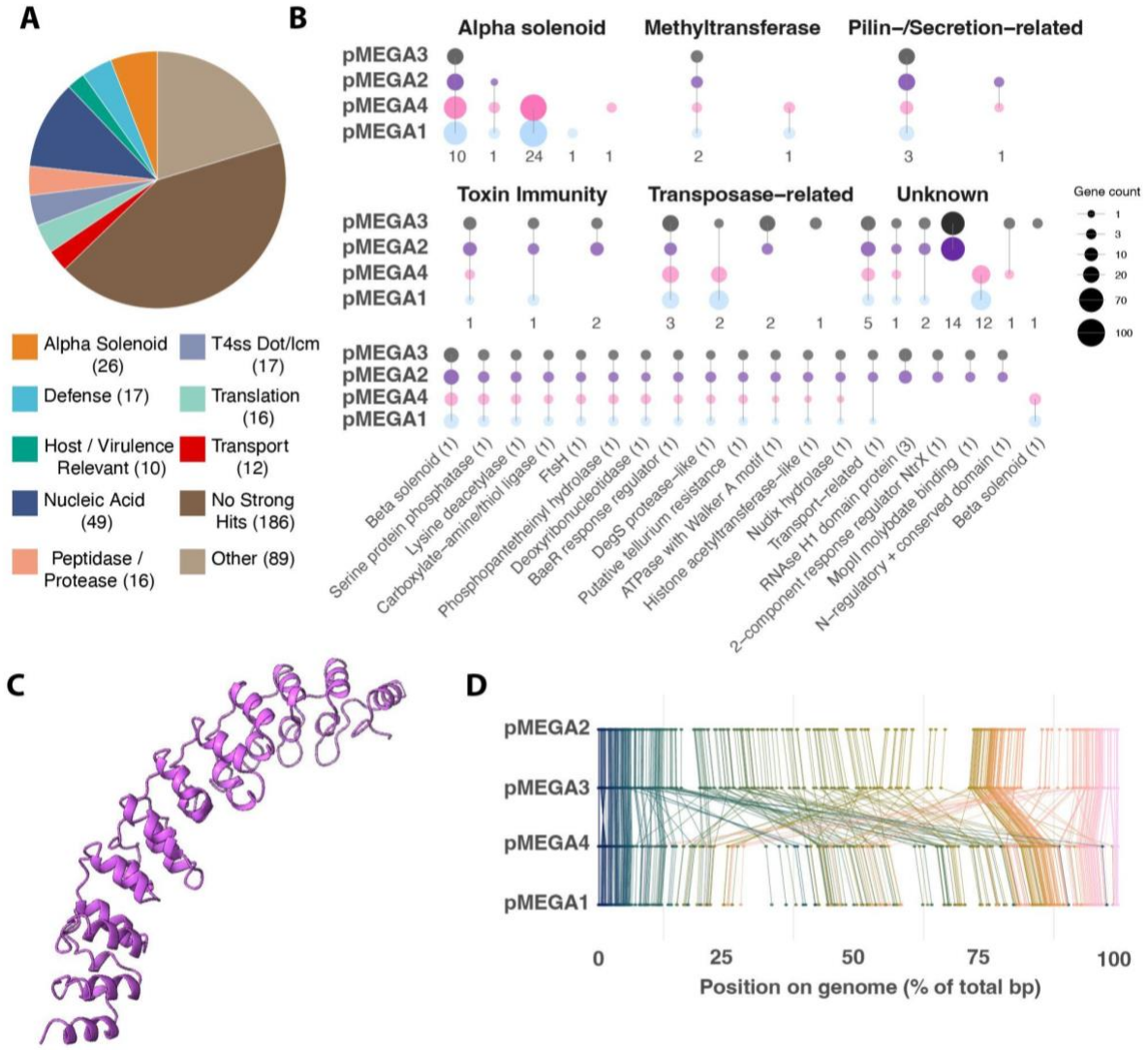
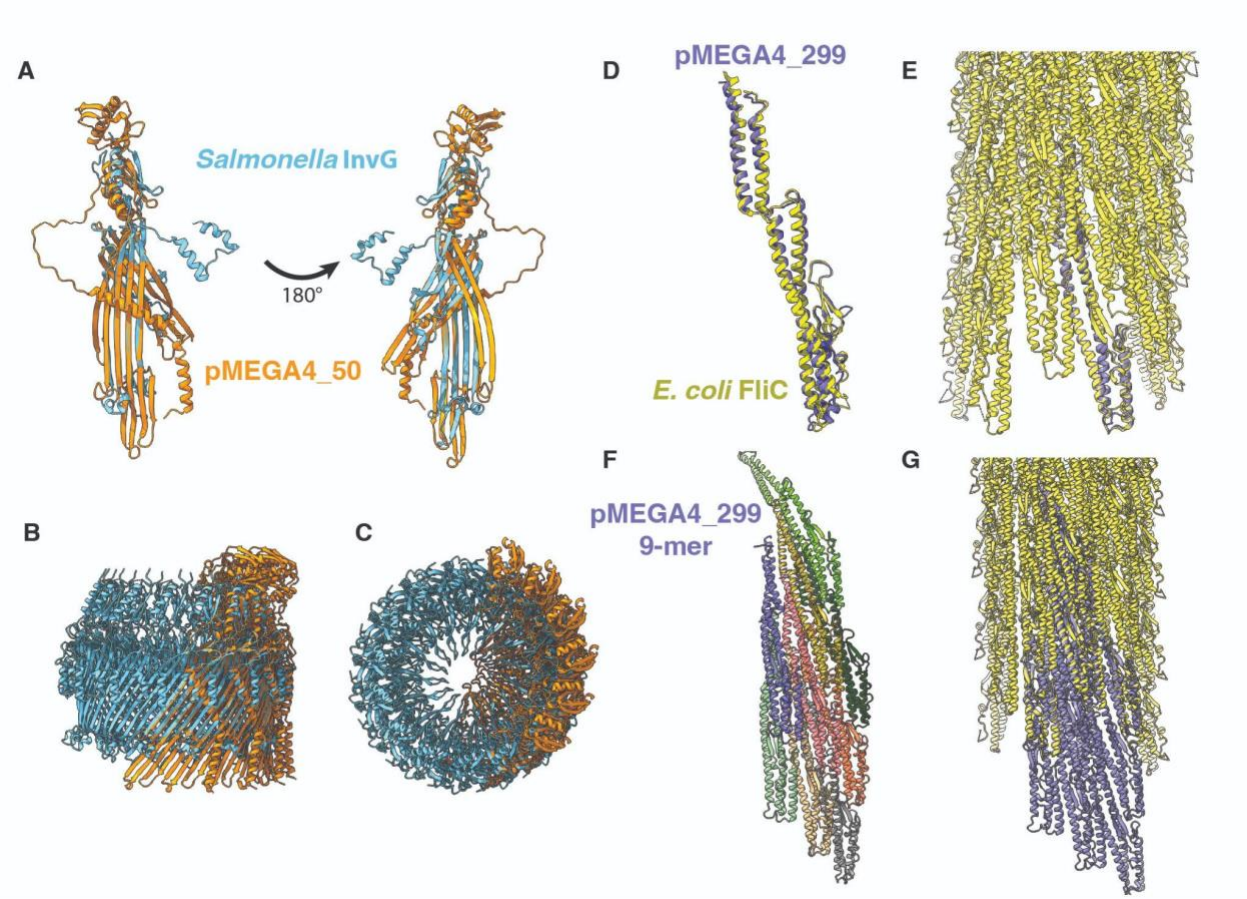


Figure 3: Conserved protein functions across two clades of megaplasmids. A) Distribution of functional annotations of the 438 conserved protein subfamilies across pMEGAs1-20. Annotations with less than 10 instances were grouped into the "Other" category. Numbers represent the number of protein families 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

1156 domain is annotated as an Ankyrin repeat. **D)** Alignment of pMEGA1-4 showing the locations of
1157 the 336 single copy protein subfamilies conserved across pMEGA1-20 as a percentage of the total
1158 length of the genome. Each protein subfamily is coloured based on their gene order in pMEGA2.



1159

1160 **Figure 4: pMEGA4 encodes mechanisms related to secretion and virulence. A)** Predicted
1161 structure of a monomer of pMEGA4_50 (median pLDDT 82.0) aligned to *Salmonella*
1162 pathogenicity island 1 injectisome protein InvG (PDB: 6PEE). Predicted hexamer of pMEGA4_50
1163 aligned to the 15-mer complex of InvG shown from the side (**B**) and top (**C**). **D)** Predicted structure
1164 of monomer of pMEGA4_299 (median pLDDT 93.91) aligned to *E. coli* K12 flagellin monomer
1165 (PDB: 8XCM) and **E)** full filament core. **F)** Predicted homo-nonamer structure of pMEGA4_299.
1166 **G)** Alignment of predicted 9-mer of pMEGA4_299 with flagellar core of *E. coli* K12
1167 (PDB:8XCM)[175].

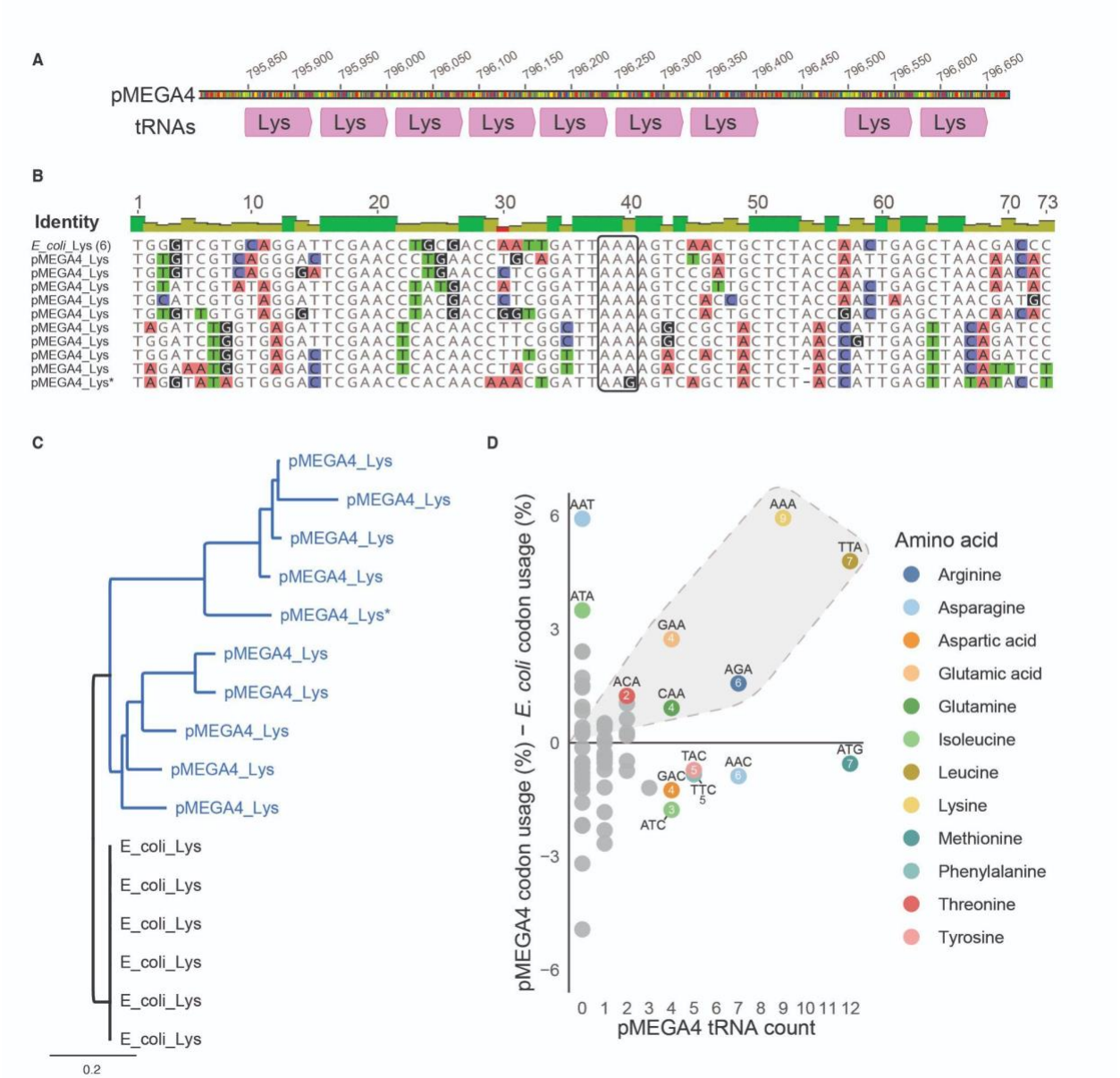
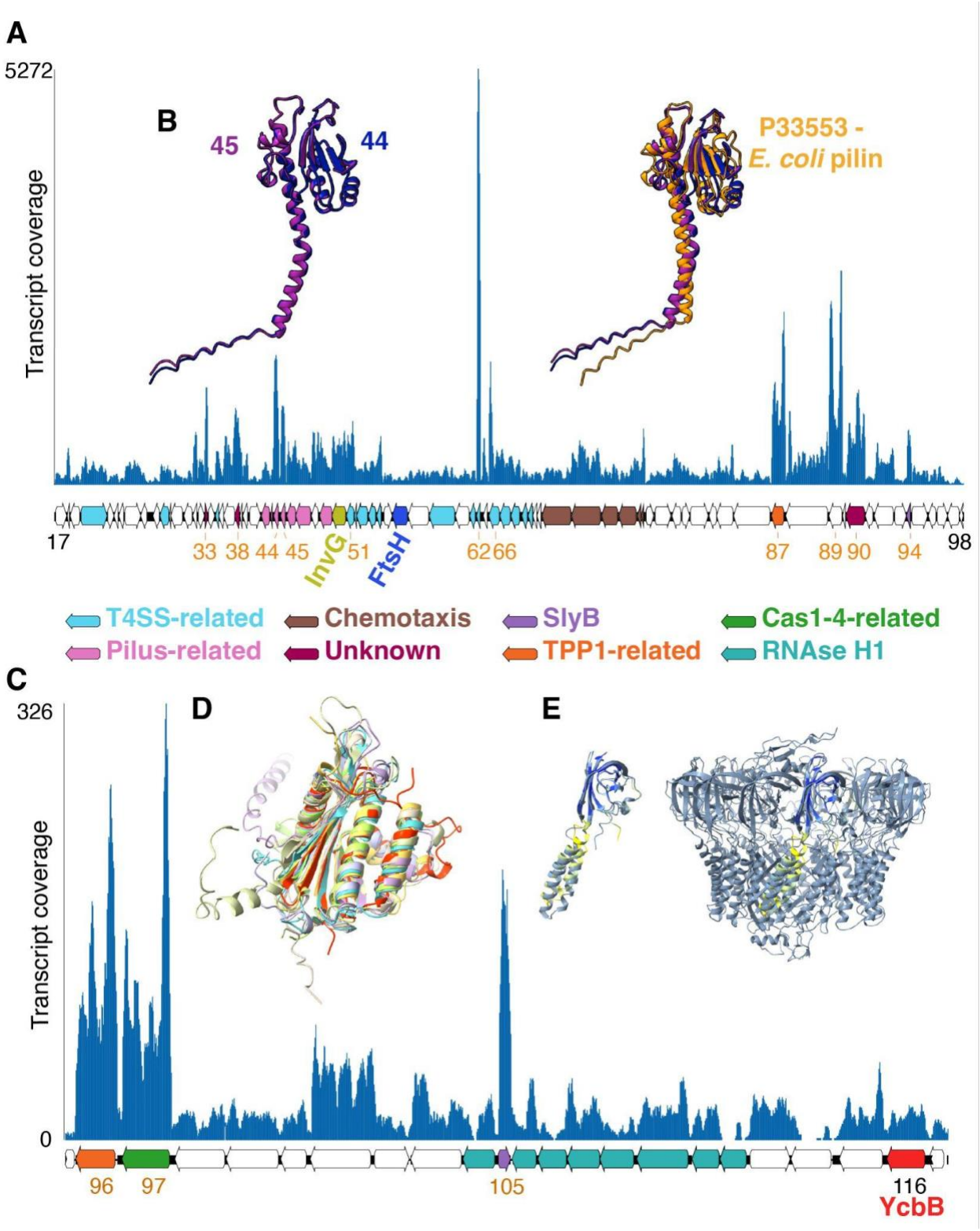


Figure 5: **tRNAs of pMEGA4 and *E. coli* host. A)** Nine of the ten pMEGA4 tRNAs for LYS are encoded in a single block. **B**) Nine of the ten pMEGA4 LYS tRNA genes have the same (AAA) codon, however the sequences differ. **C)** Phylogenetic tree constructed using LYS tRNAs for pMEGA4 and host *E. coli*, 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

1175 outlined region). However, other highly represented codons do not follow this pattern. Numbers

1176 associated with each point (colored by amino acid) indicate the count of unique tRNA sequences.



1177

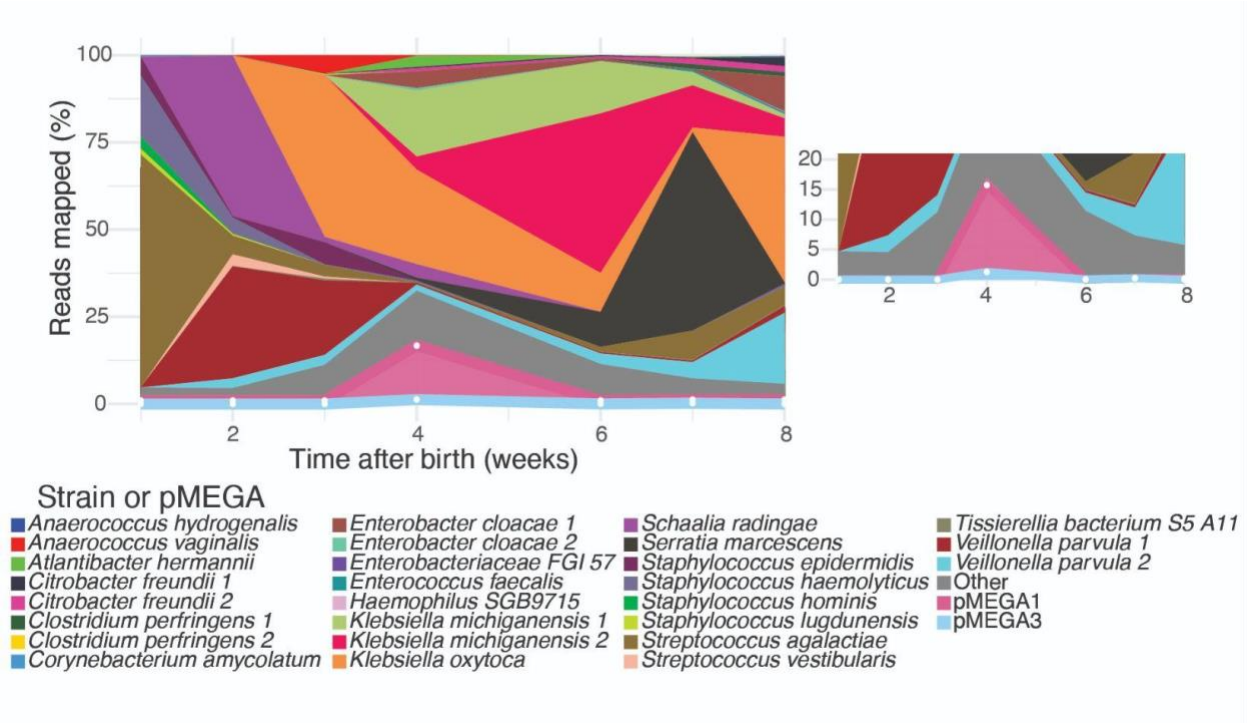
1178 **Figure 6: Public data (SRR10902853) from a panda gut metatranscriptome reveals extensive**
1179 **megaplasmid gene expression. A)** A 109 kb region of the clade 1-like plasmid assembled and
1180 manually curated using transcripts encodes 11 protein coding genes with >500X transcript
1181 coverage. These include two pilin subunits (44, 45), VirB7 (51), IcmD/DotP (62), DotF (66), TPP1
1182 (87), RuvB (89), and SlyB (94). Three highly expressed proteins have no functional predictions
1183 (33, 38, 90). The region also includes proteins of interest InvG (50) and FtsH (57). For the panda
1184 metagenome-derived reference, gene numbering is based on pMEGA4 homologs (genes 63 and
1185 64, encoding transposon genes, are absent in the panda megaplasmid genome). **B**) Left - The
1186 predicted structure of two pilin subunit genes from pMEGA4 (44, 45), which share 92.4% amino
1187 acid identity. Right - The two structures aligned to *E. coli* major structural subunit of bundle-
1188 forming pilus BfpA (AF-P33553). **C)** An example region of transcript mapping to pMEGA3
1189 showing protein coding genes 96, 97 and 105 with > 100X coverage. These are a protein related
1190 to telomerase-recruitment factor TPP1, a protein most similar to a Cas1-4 fusion, and SlyB. **D)** All
1191 of the 8 almost sequential proteins profiled as RNAse H1 in **B**have a domain that aligns well with
1192 PDB 1WSJ (red). **E)** Monomer of SlyB (pMEGA3_105; median pLDDT 90.06) and its alignment
1193 to PDB 7OJF.

1194

1195

1196

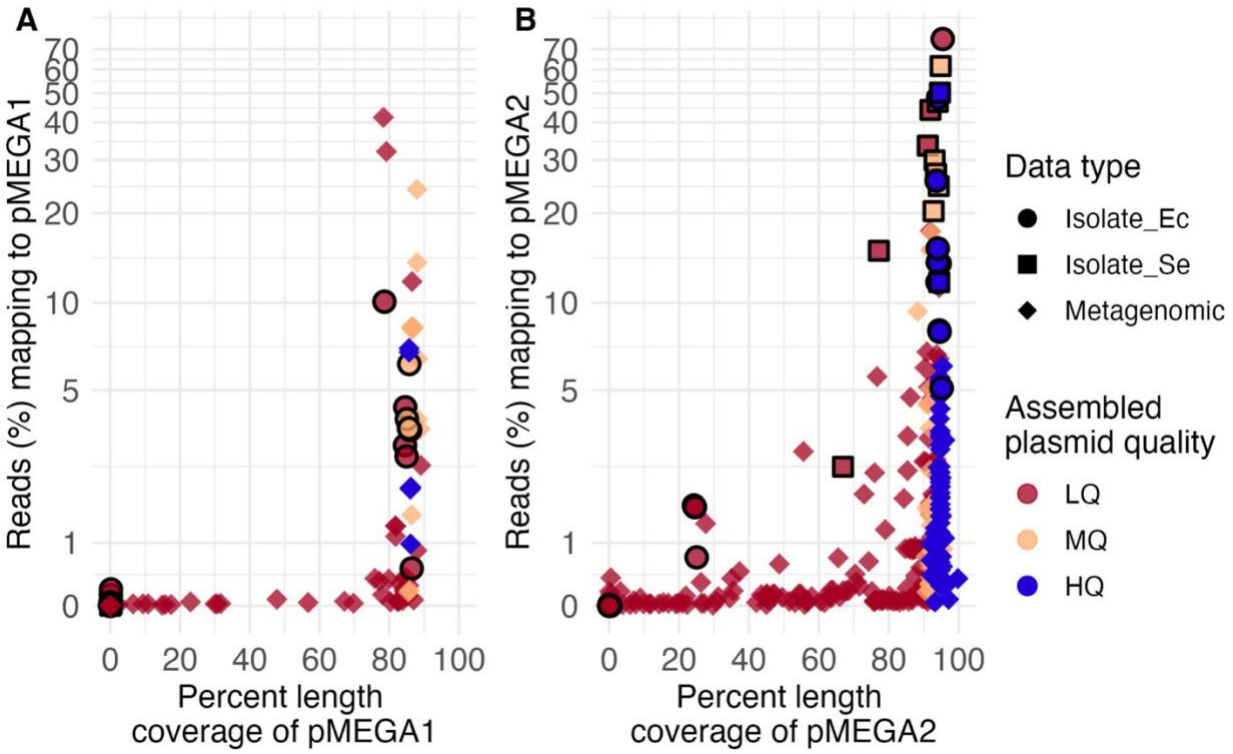
1197 **SUPPLEMENTARY FIGURES**



1198

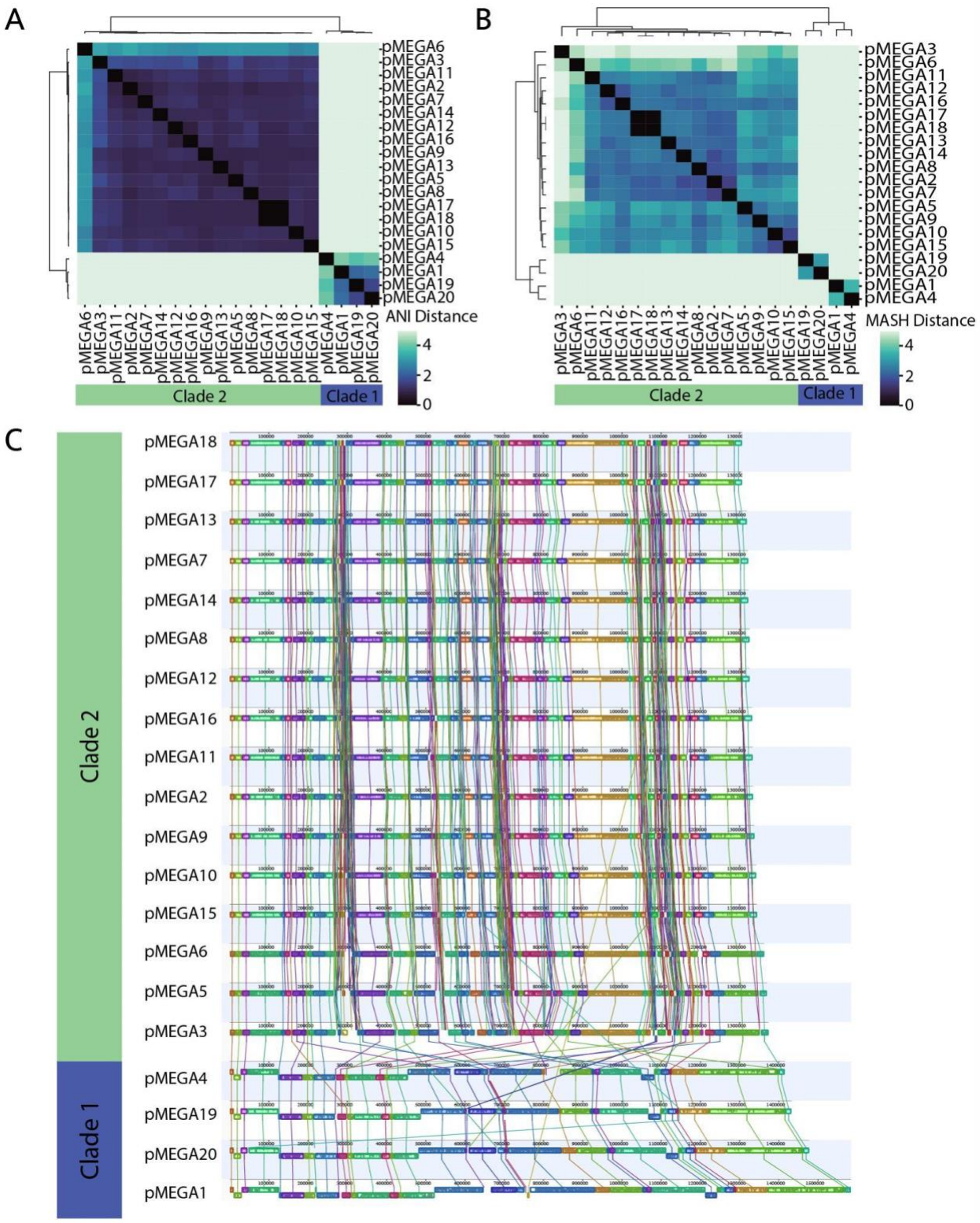
1199 **Figure S1: Infant gut microbiome community containing pMEGA1 and pMEGA3.** The
1200 percentage of reads mapped to MAGs, pMEGA1 and pMEGA3 within INF1340011 across 8
1201 weeks of life was determined through Illumina read mapping. The "Other" category includes reads
1202 that did not map to MAGs nor pMEGA1,3. Inset shows the same plot but with y values from 0 to
1203 10.

1204



1205

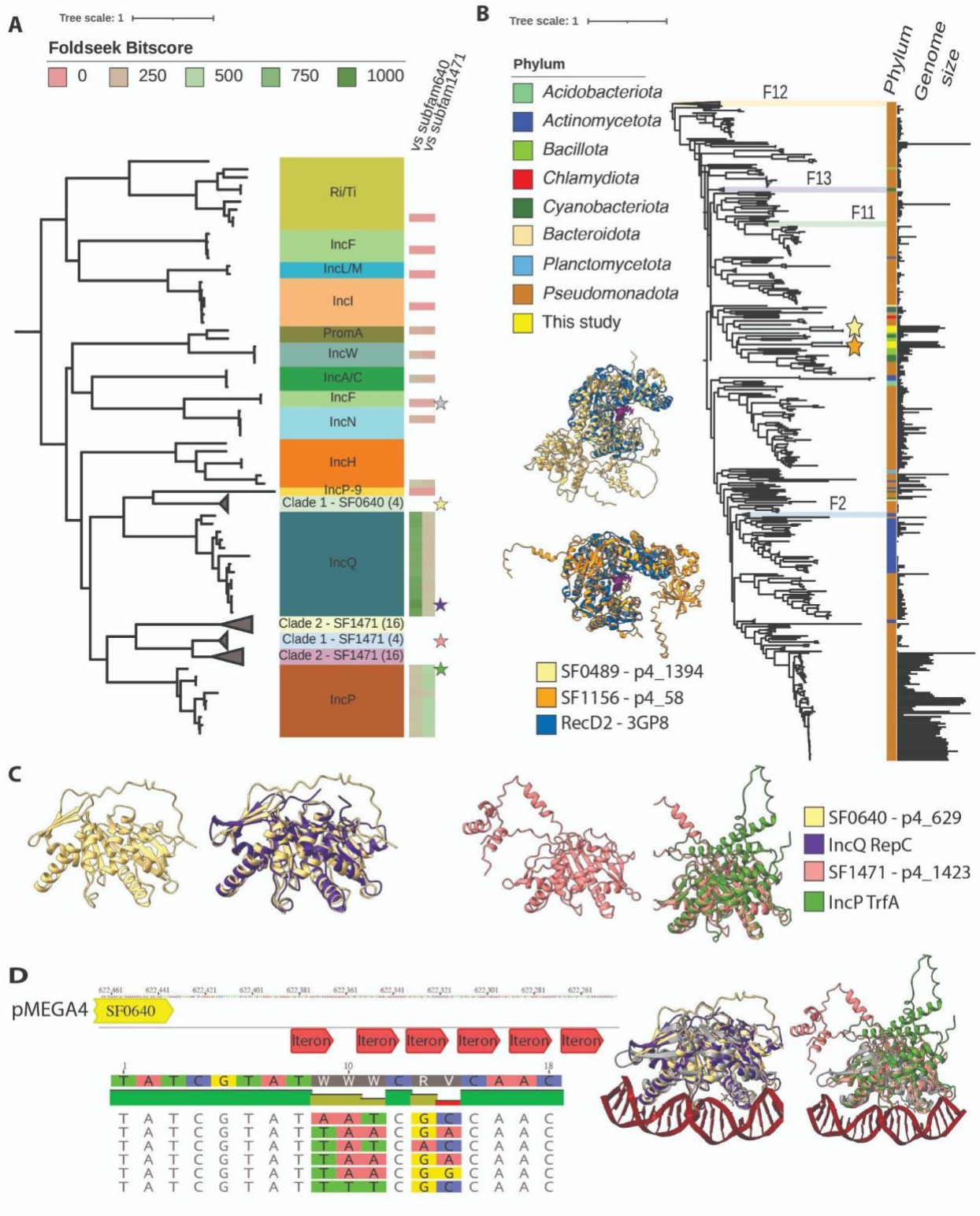
1206 **Figure S2: Prevalence and completeness of megaplasmids in public data.** Sequencing reads
1207 from SRA accessions were mapped to pMEGA1 **(A)** and pMEGA2 **(B)** to determine the percent
1208 length coverage and percentage of reads mapped to each megaplasmid. For each accession,
1209 assembled plasmids were classified as high-quality (HQ - blue) if they had ≥85% coverage of a
1210 pMEGA and ≤10 contigs and medium-quality (MQ - yellow) with ≥85% coverage and 11-50
1211 contigs. The remaining samples were classified as low-quality (LQ - red). Diamond: Metagenomic
1212 sample, circle: *Escherichia coli* Isolate (Ec), square: *Salmonella enteritidis* isolate (Se). Isolates
1213 are further highlighted by a black border.


A
pMEGA6
pMEGA3
pMEGA11
pMEGA2
pMEGA7
pMEGA14
pMEGA12
pMEGA16
pMEGA9
pMEGA13
pMEGA5
pMEGA8
pMEGA17
pMEGA18
pMEGA10
pMEGA15
pMEGA4
pMEGA1
pMEGA19
pMEGA20
ANI Distance
4
2
0
Clade 2
Clade 1
B
pMEGA3
pMEGA6
pMEGA11
pMEGA12
pMEGA16
pMEGA17
pMEGA18
pMEGA13
pMEGA14
pMEGA8
pMEGA2
pMEGA7
pMEGA5
pMEGA9
pMEGA10
pMEGA15
pMEGA19
pMEGA20
pMEGA1
pMEGA4
MASH Distance
4
2
0
Clade 2
Clade 1
C
Clade 2
Clade 1
pMEGA18
pMEGA17
pMEGA13
pMEGA7
pMEGA14
pMEGA8
pMEGA12
pMEGA16
pMEGA11
pMEGA2
pMEGA9
pMEGA10
pMEGA15
pMEGA6
pMEGA5
pMEGA3
pMEGA4
pMEGA19
pMEGA20
pMEGA1

1214

1215 **Figure S3: Comparing 20 complete megaplasmids. A)** Whole genome alignment of complete
1216 plasmid sequences. Vertical lines connect blocks with similarity between plasmid genomes. The
1217 sequences were ordered manually and grouped based on clades. **B)** Matrix of average nucleotide
1218 identity (ANI). **C)** Matrix of MASH distance.

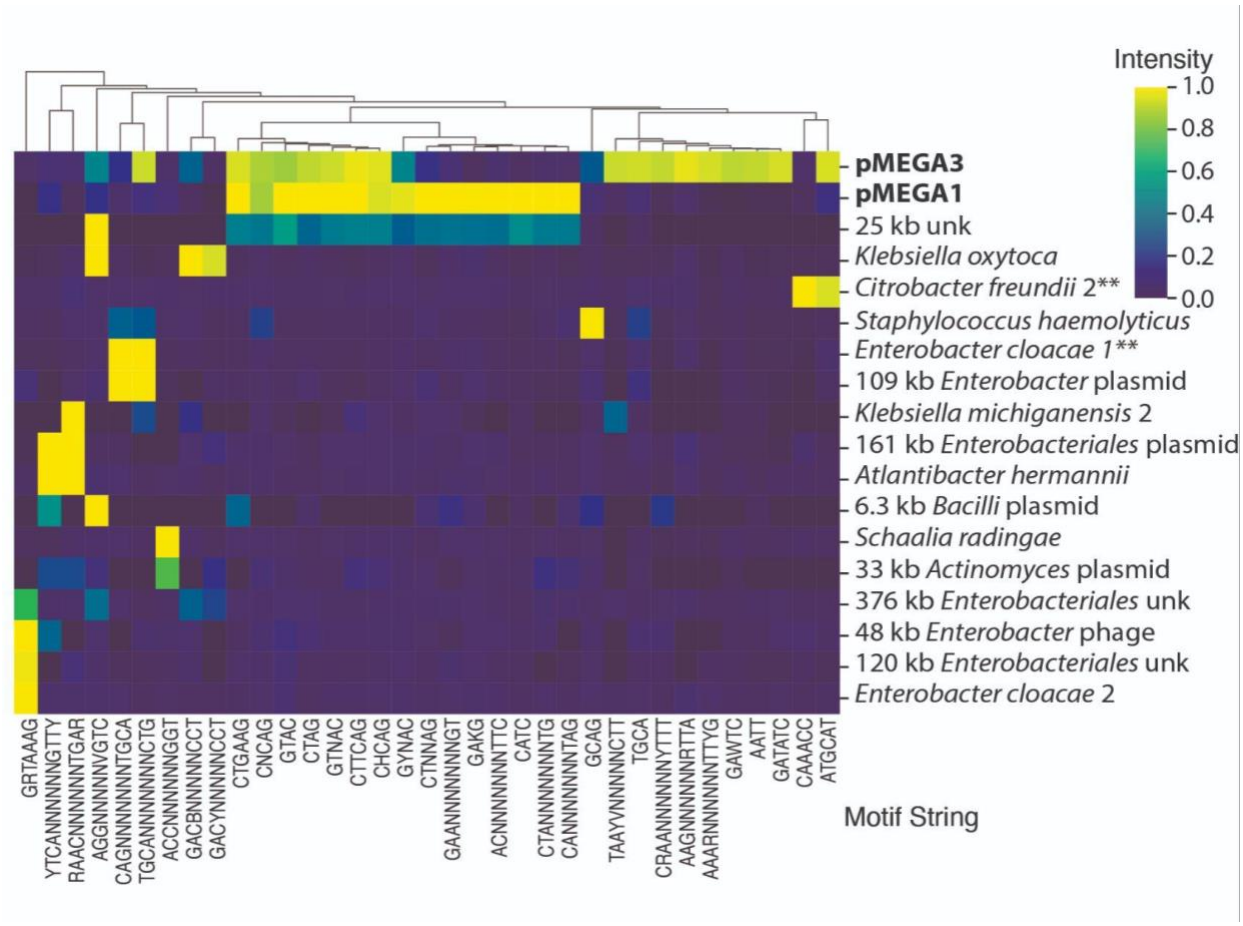
1219


A
Tree scale: 1
Foldseek Bitscore
0
250
500
750
1000
vs subfam640
vs subfam1471
Ri/Ti
IncF
IncL/M
IncI
PromA
IncW
IncA/C
IncF
IncN
IncH
IncP-9
Clade 1 - SF0640 (4)
IncQ
Clade 2 - SF1471 (16)
Clade 1 - SF1471 (4)
Clade 2 - SF1471 (16)
IncP
B
Tree scale: 1
Phylum
Acidobacteriota
Actinomycetota
Bacillota
Chlamydiota
Cyanobacteriota
Bacteroidota
Planctomycetota
Pseudomonadota
This study
F12
F13
F11
F2
Phylum
Genome size
SF0489 - p4_1394
SF1156 - p4_58
RecD2 - 3GP8
C
SF0640 - p4_629
IncQ RepC
SF1471 - p4_1423
IncP TrfA
D
pMEGA4
SF0640
Iteron
Iteron
Iteron
Iteron
Iteron
Iteron

1220

1221 **Figure S4: Replication initiation proteins and relaxases from two clades of megaplasmids.**
1222 **A)** Phylogeny of 105 replication initiation (Rep) protein sequences. Reference Rep protein
1223 sequences are colored based on plasmid incompatibility group[170]. Bitscores from select FoldSeek
1224 hits are shown as heatmaps on the right of the tree. A star indicates that the protein structure is
1225 highlighted in panel C and D. **B)** Phylogeny of 881 MOB_F protein sequences with clades (F12,
1226 F13, F11, F2) collapsed[166]. Megaplasmids identified in this study are highlighted with a star. The
1227 length of each plasmid is shown on the far right panel. Predicted protein structures of relaxases
1228 aligned to PDB:3GP8 with ssDNA in purple are shown to the left of the tree. **C)** Representative
1229 structures of Rep proteins from pMEGA4 compared to RepC (purple) and TrfA (green). **D)** Region
1230 downstream of pMEGA4_629 with putative iterons and alignment of repeat sequences. To the
1231 right, alignment of structures from panel C to Rep protein from *E. coli* F plasmid (PDB: 1REP -
1232 grey) bound to the iteron sequence (red).

1233
1234



1235

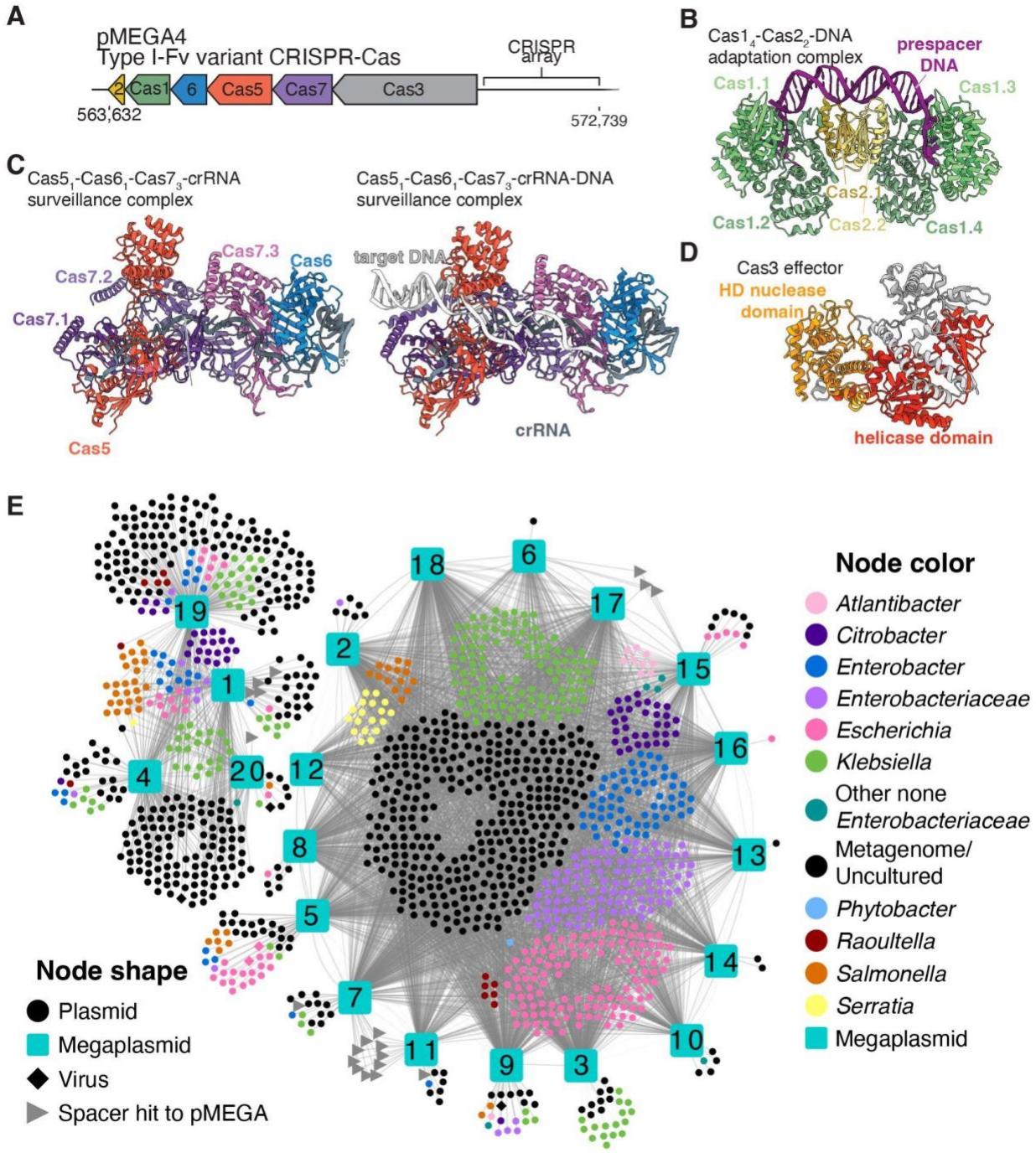
1236 **Figure S5: Possible *Enterobacteriaceae* hosts of pMEGA1 and pMEGA3.** Modified bases were

1237 predicted in the PacBio HiFi reads of INF1340011 at week 4. The columns represent the motifs

1238 predicted to contain modified bases, and the rows represent genomes. ** Possible host(s) of

1239 pMEGA3 based on overlap in modified motif patterns.

1240

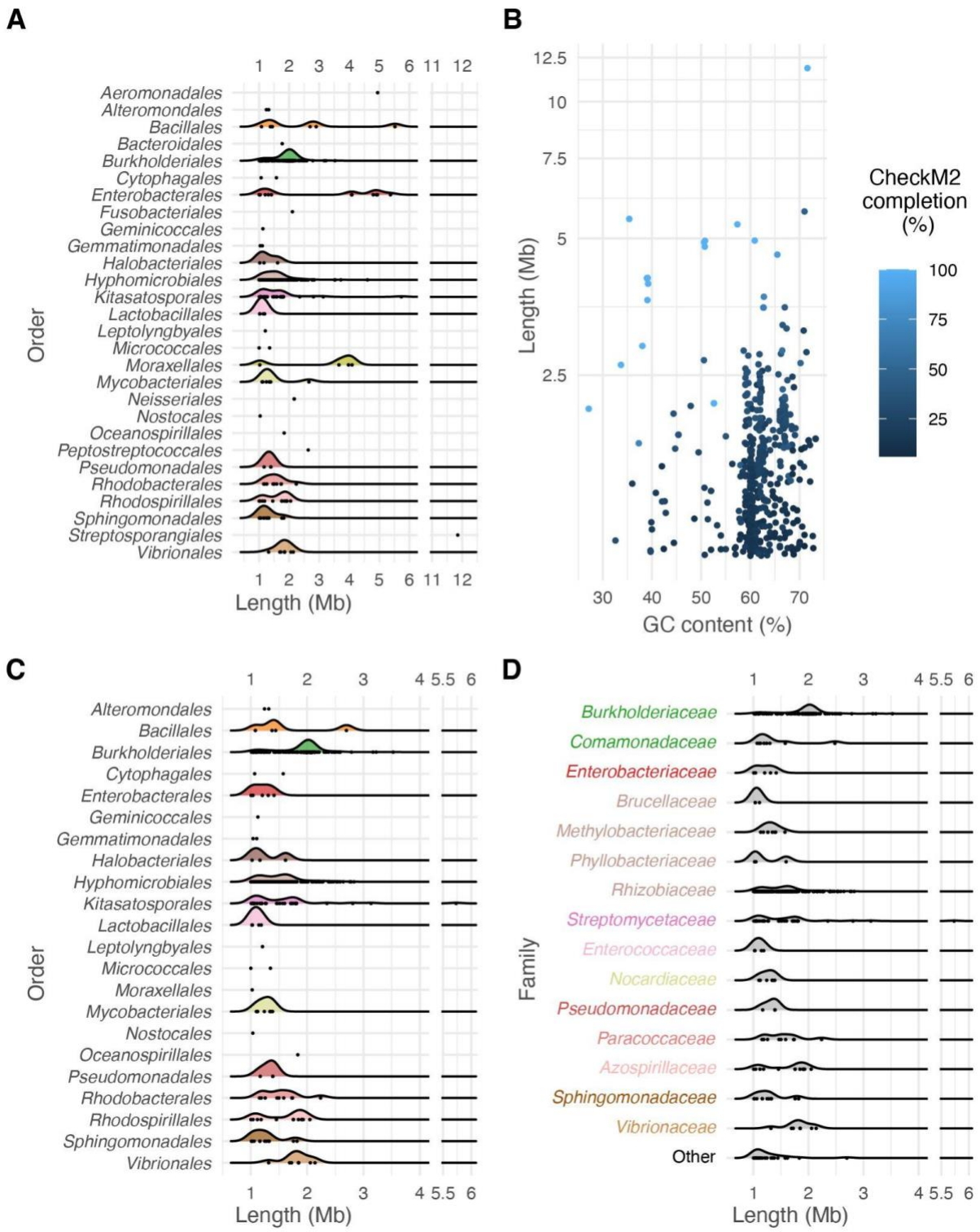


1241

1242

1243

1244 **Figure S6: CRISPR-Cas system in pMEGA4 and targets of CRISPR spacers across**
1245 **pMEGA1-20. A)** Diagram of the type I-Fv CRISPR-Cas locus from pMEGA4. **B)** Structural
1246 prediction of the pMEGA4 heterohexameric $Cas1_4$-$Cas2_2$ integrase complex (pMEGA4_568 and
1247 569, median pLDDT 92.56 and 93.44) bound to a prespacer DNA. **C)** Structural predictions of the
1248 pMEGA4 Cascade surveillance complex in two states. Left, pMEGA4 $Cas5_1$-$Cas6_1$-$Cas7_3$ (ORFs
1249 570 - 572; median pLDDTs 85.19, 91.62, and 93.69) co-folded with a crRNA from an experimental
1250 structure (5O7H) of a similar type I-Fv system. Right, the same structure but co-folded with a
1251 target DNA. **D)** AlphaFold3 prediction of the pMEGA4 Cas3 effector (ORF 573 median pLDDT
1252 91.81), with the HD nuclease domain colored in orange and the helicase domain colored in red. **E)**
1253 CRISPR spacer targets from pMEGA1-20 within other pMEGAs, plasmids from IMG/PR and
1254 PLSDB, and viral sequences from IMG/VR. Each line indicates a spacer targets a sequence in
1255 another element or, if the connection is between two pMEGA squares, the genomes share the same
1256 spacer with their CRISPR arrays. Grey triangles indicate a spacer from a pMEGA that targets
1257 another pMEGA.


A
Order
Aeromonadales
Alteromondales
Bacillales
Bacteroidales
Burkholderiales
Cytophagales
Enterobacterales
Fusobacteriales
Geminicoccales
Gemmatimonadales
Halobacteriales
Hyphomicrobiales
Kitasatosporales
Lactobacillales
Leptolyngbyales
Micrococcales
Moraxellales
Mycobacteriales
Neisseriales
Nostocales
Oceanospirillales
Peptostreptococcales
Pseudomonadales
Rhodobacterales
Rhodospirillales
Sphingomonadales
Streptosporangiales
Vibrionales
1 2 3 4 5 6 11 12
Length (Mb)
B
Length (Mb)
12.5
10
7.5
5
2.5
30 40 50 60 70
GC content (%)
CheckM2 completion (%)
100
75
50
25
C
Order
Alteromondales
Bacillales
Burkholderiales
Cytophagales
Enterobacterales
Geminicoccales
Gemmatimonadales
Halobacteriales
Hyphomicrobiales
Kitasatosporales
Lactobacillales
Leptolyngbyales
Micrococcales
Moraxellales
Mycobacteriales
Nostocales
Oceanospirillales
Pseudomonadales
Rhodobacterales
Rhodospirillales
Sphingomonadales
Vibrionales
1 2 3 4 5.5 6
Length (Mb)
D
Family
Burkholderiaceae
Comamonadaceae
Enterobacteriaceae
Brucellaceae
Methylobacteriaceae
Phyllobacteriaceae
Rhizobiaceae
Streptomycetaceae
Enterococcaceae
Nocardiaceae
Pseudomonadaceae
Paracoccaceae
Azospirillaceae
Sphingomonadaceae
Vibrionaceae
Other
1 2 3 4 5.5 6
Length (Mb)

1258

1259 **Figure S7: Plasmid databases are contaminated with complete bacterial genomes and**
1260 **possible secondary chromosomes. A)** Length distribution of 667 non-redundant plasmids >1 Mb
1261 from PLSDB and IMG/PR grouped by the Order of their bacterial host. **B)** Length and GC content
1262 distribution of plasmid colored by completeness as determined with CheckM2. **C)** Length
1263 distribution of 642 plasmids >1Mb after removing sequences with >50% completion, grouped by
1264 taxonomic Order. **D)** Length distribution of 642 plasmids >1Mb after removing sequences with
1265 >50% completion, grouped by taxonomic Family. Families with two or fewer plasmids were
1266 grouped into Other.

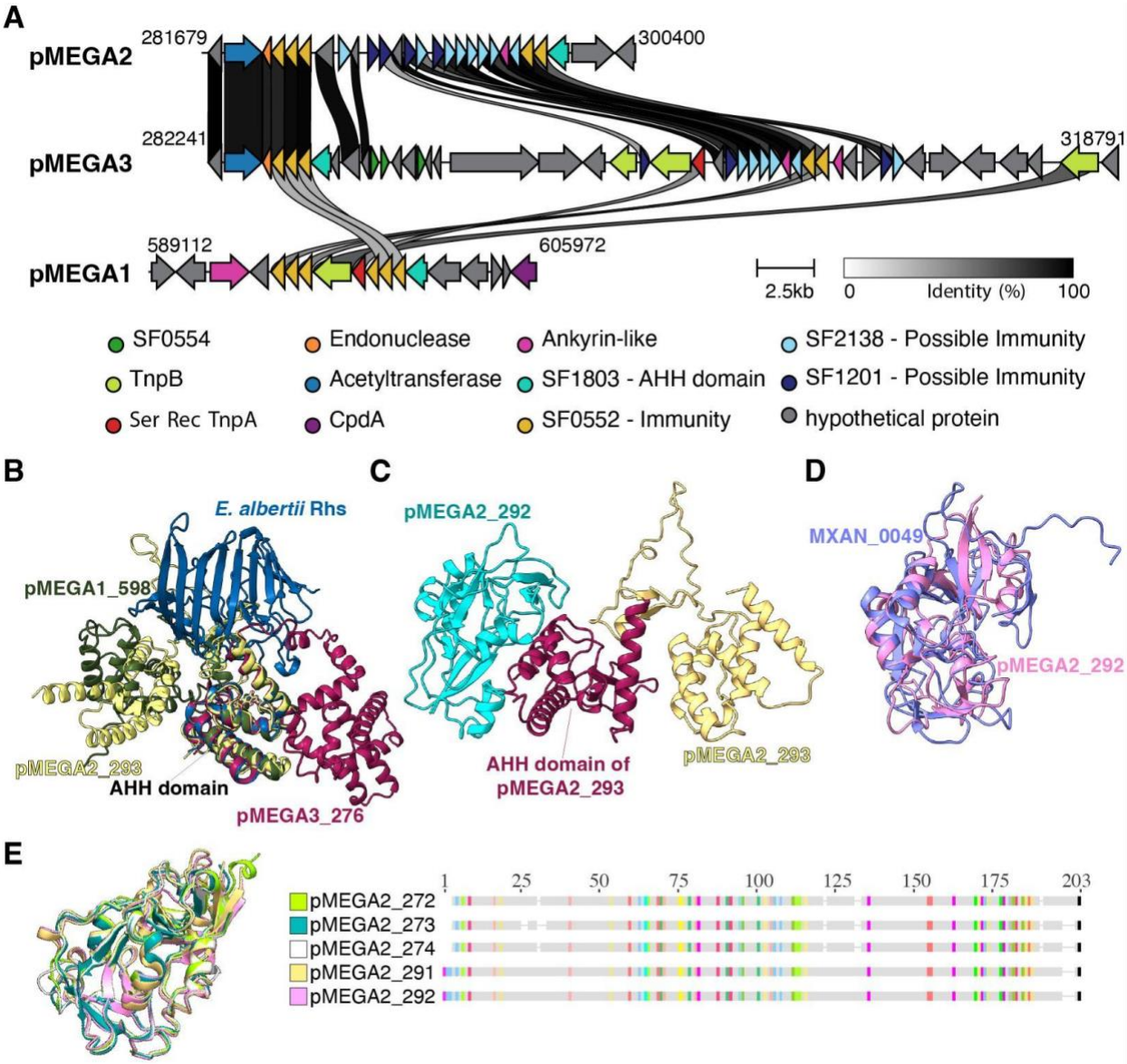


Figure S8: Polymorphic toxin and poly-immunity loci diversity in megaplasmids. A) Alignment of regions in pMEGAs1,2 and 3 with similar subfamilies, including a putative AHH nuclease toxin (SF1803) and tandem predicted immunity proteins (SF0552, 2138, 1201). **B)** Predicted polymorphic toxins of SF1803 (median pLDDTs: pMEGA1_598 71.88, pMEGA2_293 78.25, pMEGA3_276 69.12) aligned to a predicted structure of a Rhs family protein from *E. albertii* A0A855STS0 (blue). The aligned region is a predicted nuclease domain of the

1274 HNH/ENDO VII superfamily (PF14412) and contains a conserved AHH motif. **C)** Result of co-
1275 folding the predicted toxin AHH nuclease toxin with the downstream encoded predicted immunity
1276 protein. **D)** Alignment of the structure of a known nuclease immunity protein MXAN_0049 (PDB:
1277 6A4C) to the predicted immunity protein (pink: pMEGA2_292 median pLDDT 96.38). **E)**
1278 Structural and protein sequence alignment of 5 members of SF0552 from pMEGA2.

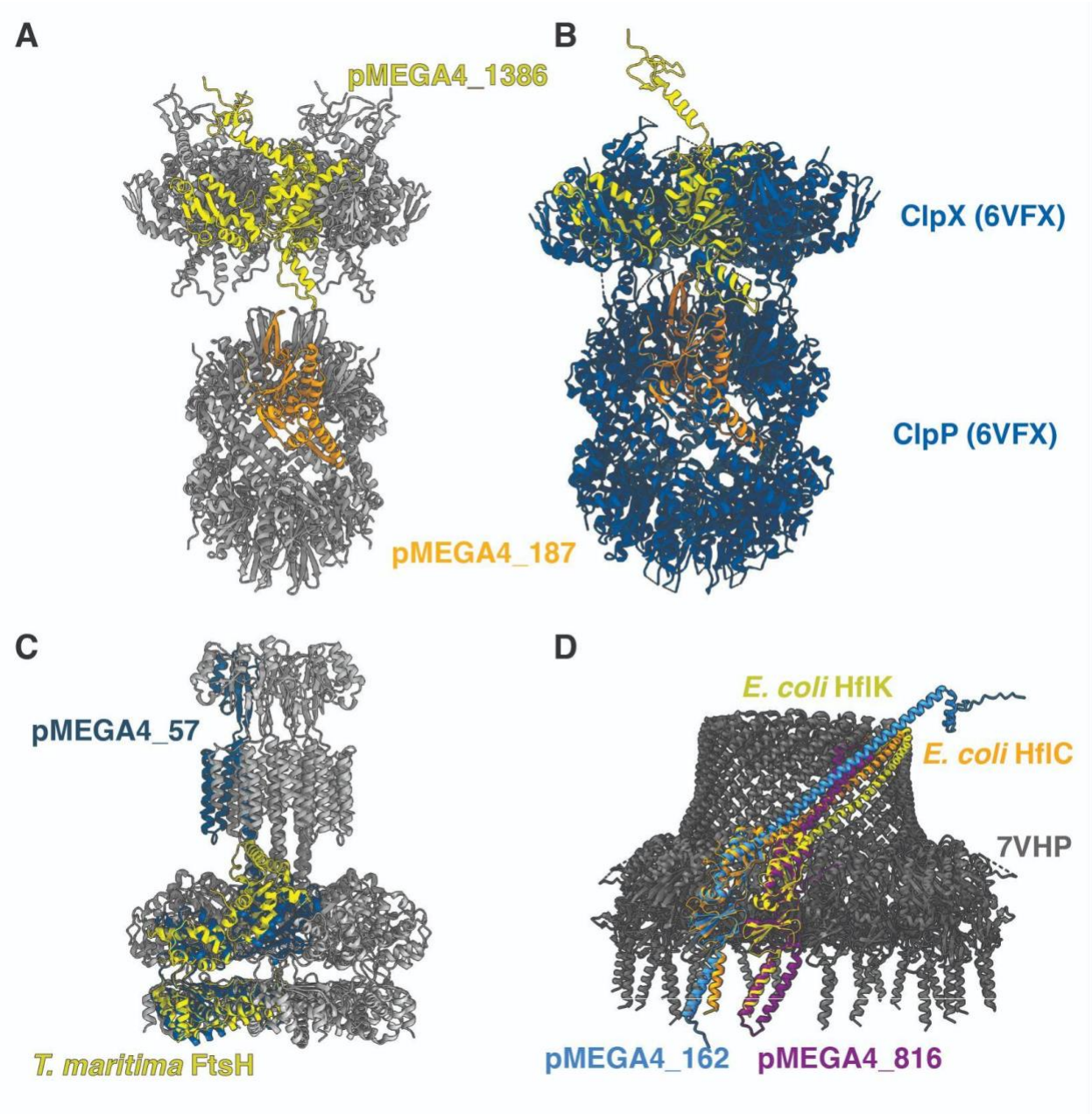


Figure S9: pMEGA4 predicted protein structures of various functions. A) Predicted hexamer of predicted ClpX (pMEGA4_1386) and dodecamer of ClpP (pMEGA4_187). **B)** Monomers of ClpX (median pLDDT 92.81) and ClpP (median pLDDT 98.56) from pMEGA4 aligned to reference of ClpXP from *Neisseria meningitidis* (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

1286 95.62, and pMEGA4_816 89.94) aligned to the structure of *E. coli* membrane microdomain (PDB:
1287 7VHP). *E. coli* HflC and HflK are colored individually in orange and yellow.

1288 **Supplementary Tables**

1289 **Table S1** List of NANO samples and assembly statistics
1290 **Table S2** ANI and MASH distance between 20 complete plasmids
1291 **Table S3** Relative abundance of MAGs and pMEGAs in INF1340011
1292 **Table S4** Relative abundance of MAGs and pMEGAs in INF1330004
1293 **Table S5** SRA read mapping and assembly from various studies
1294 **Table S6** Hits from other internal NIH samples
1295 **Table S7** List of 20 confident megaplasmids
1296 **Table S8** pMEGA4 confident protein annotations
1297 **Table S9** Core protein subfamilies identified across 20 megaplasmids
1298 **Table S10** Correlations between MAG abundance and pMEGAs
1299 **Table S11** Predicted modified motifs and methyltransferases within pMEGAs, other bacterial
1300 genomes, and MGEs in infant gut microbiome
1301 **Table S12** Spacer hits to pMEGAs from isolate genomes and metagenomes
1302 **Table S13** Putative Enterobacteriaceae 1Mb plasmids from NCBI
1303 **Table S14** Multi-copy protein subfamilies and frequency of occurrence in clusters in pMEGA1-4
1304 **Table S15** *E. coli* G269_1R ORF predictions
1305 **Table S16** Comparing protein sequences from pMEGA4 to the *E. coli* host genome
1306 **Table S17** tRNA sequence types in pMEGA4 and *E. coli* host
1307 **Table S18** tRNA counts and codon usage in pMEGA4 and *E. coli* host
1308 **Table S19** Antibiotic resistance genes predicted in *E. coli* G269_1R and pMEGA4
1309 **Table S20** Highly expressed genes in clade 1-like megaplasmid from panda gut
1310 metatranscriptome dataset
1311 **Table S21** Highly expressed genes in clade 2-like megaplasmid from panda gut
1312 metatranscriptome dataset