

1 **Population dynamics of *Escherichia coli* in the gastrointestinal tracts of**
2 **Tanzanian children**

3
4
5 Authors: Taylor K. S. Richter^{1,2}, Tracy H. Hazen^{1,2}, Diana Lam⁶, Christian L. Coles⁴,
6 Jessica C. Seidman⁵, Yaqi You⁶, Ellen K. Silbergeld⁵, Claire M. Fraser^{1,3} and David A.
7 Rasko^{1,2*}

8
9 ¹ The Institute for Genome Sciences, University of Maryland School of Medicine,
10 Baltimore, MD, USA; ² Department of Microbiology and Immunology, University of
11 Maryland School of Medicine, Baltimore, MD, USA; ³ Department of Medicine,
12 University of Maryland School of Medicine, Baltimore, MD, USA; ⁴ Department of
13 International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD,
14 USA; ⁵ Division of International Epidemiology and Population Studies, Fogarty
15 International Center, National Institutes of Health, Bethesda, MD, USA; ⁶ Department of
16 Environmental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore,
17 MD, USA

18
19 Running title: Population dynamics of *E. coli*

20 * Corresponding Author

21 David A. Rasko

22 University of Maryland School of Medicine

23 Institute for Genome Sciences

24 Department of Microbiology and Immunology

25 801 W. Baltimore Street, Suite 619

26 Baltimore, MD 21201

27 Phone: 410.706.6774

28 E-mail: drasko@som.umaryland.edu
29

30
31 Keywords: *Escherichia coli*, microbial genomics, longitudinal samples, diversity

32

33 **Abstract.**

34 The stability of the *Escherichia coli* populations in the human gastrointestinal tract are
35 not fully appreciated, and represent a significant knowledge gap regarding
36 gastrointestinal community structure, as well as resistance to incoming pathogenic
37 bacterial species and antibiotic treatment. The current study examines the genomic
38 content of 240 *Escherichia coli* isolates from children 2 to 35 months old in Tanzania.
39 The *E. coli* strains were isolated from three time points spanning a six month time
40 period, with or without antibiotic treatment. The resulting isolates were sequenced, and
41 the genomes compared. The findings in this study highlight the transient nature of *E.*
42 *coli* strains in the gastrointestinal tract of children, as during a six-month interval, no one
43 individual contained phylogenomically related isolates at all three time points. While the
44 majority of the isolates at any one time point were phylogenomically similar, most
45 individuals did not contain phylogenomically similar isolates at more than two time
46 points. Examination of global genome content, canonical *E. coli* virulence factors,
47 multilocus sequence type, serotype, and antimicrobial resistance genes identified
48 diversity even among phylogenomically similar strains. There was no apparent increase
49 in the antimicrobial resistance gene content after antibiotic treatment. The examination
50 of the *E. coli* from longitudinal samples from multiple children in Tanzania provides
51 insight into the genomic diversity and population variability of resident *E. coli* within the
52 rapidly changing environment of the gastrointestinal tract.

53

54 **Importance.**

55 This study increases the number of resident *Escherichia coli* genome sequences, and
56 explores *E. coli* diversity through longitudinal sampling. We investigate the genomes of
57 *E. coli* isolated from human gastrointestinal tracts as part of an antibiotic treatment
58 program among rural Tanzanian children. Phylogenomics demonstrates that resident *E.*
59 *coli* are diverse, even within a single host. Though the *E. coli* isolates of the
60 gastrointestinal community tend to be phylogenomically similar at a given time, they
61 differed across the interrogated time points, demonstrating the variability of the
62 members of the *E. coli* community. Exposure to antibiotic treatment did not have an
63 apparent impact on the *E. coli* community or the presence of resistance and virulence
64 genes within *E. coli* genomes. The findings of this study highlight the variable nature of
65 bacterial members of the human gastrointestinal tract.

66

67

68

69 **Introduction**

70 *Escherichia coli* in the human gastrointestinal tract is often recognized as an important
71 source of disease (1, 2). As the causative agent of over 2 million deaths annually due to
72 diarrhea (3, 4), as well as millions of extraintestinal infections (5), its categorization as a
73 pathogen is not unwarranted. Particularly in developing countries, the consequences of
74 diarrheal *E. coli* is substantial among children under five years old, who incur the
75 majority of infections and deaths (3) and whose rapidly developing microbiomes can be
76 impacted by frequent bouts of disease and treatment (6, 7). Yet, *E. coli* are the
77 dominant aerobic organism in the human gastrointestinal tract, identified in greater than
78 90% of humans, and many other large mammals, often reaching concentrations up to
79 10^9 CFU per gram of feces (8) without causing disease. In this role as a resident
80 organism in healthy hosts, they are thought to have critical roles in digestion, nutrition,
81 metabolism, and protection against incoming enteric pathogens (9-12). Despite the
82 importance and involvement of *E. coli* in human health, studies of their role as native,
83 non-pathogenic members of the human gastrointestinal microbiome are poorly
84 represented among genome sequencing, comparative analysis efforts, and functional
85 characterization.

86

87 Most studies of bacterial genomics have focused on pathogenic isolates over a limited
88 time frame. *E. coli* genomic studies are no exception, having concentrated on
89 sequencing single isolates, from single time points, and on samples related to a clinical
90 presentation, such as diarrhea or urinary tract infection (10, 13-17). There have been
91 fewer than five genomes sequenced of non-pathogenic *E. coli*, in addition to a limited

92 number of isolates from the feces of individuals that do not have diarrhea (10, 17-20).
93 To date the genomic examination of longitudinal isolates is lacking, thus hindering the
94 ability to explore the diversity of *E. coli* isolates both within-host and across time. Most
95 studies of resident *E. coli* were completed prior to ready access to sequencing
96 technologies (11). An exception is Stoesser *et al.* (18), which identified multiple isolates
97 in single-host samples using single nucleotide polymorphism (SNP) level analyses,
98 leaving much to be learned about *E. coli* genomic diversity within and between human
99 hosts over longitudinal sampling.

100

101 A population-based longitudinal cohort study, PRET+ (Partnership for the Rapid
102 Elimination of Trachoma, January to July 2009), provided a unique opportunity to
103 examine both the diversity and dynamics of the *E. coli* isolates in the human
104 gastrointestinal tract among children in Tanzania (21, 22). In the PRET+ study, Seidman
105 *et al.* investigated the effects of mass distribution of azithromycin on antibiotic
106 resistance of resident *E. coli* (21, 22). *E. coli* were isolated from fecal swabs obtained
107 from children 2 to 35 months old living in rural Tanzania, half of whom were given a
108 single oral prophylactic azithromycin treatment for trachoma (an infection of the eye
109 caused by *Chlamydia trachomatis*). *E. coli* isolates from this cohort were selected for
110 genome sequencing and comparative analyses to investigate the within-subject and
111 longitudinal diversity of *E. coli* isolates in children (Table S1). Up to three isolates per
112 individual, from each of three time points spanning six months, were collected in the
113 PRET+ study, providing up to nine potential isolates from each subject for examination
114 (Figure 1).

115
116 Samples from the current study provide insight into *E. coli* diversity within a subject over
117 several time points. While other studies have examined resident *E. coli* in children in
118 developing countries, they limited their focus to using PCR and *in vitro* lab techniques to
119 identify a limited set of canonical virulence genes and determine resistance profiles of
120 the isolated strains (23-25). In addition to the virulence- and resistance-associated
121 gene content, the current study demonstrates previously uncharacterized diversity
122 among *E. coli* isolates from the human gastrointestinal tract on a whole genome level
123 within and across sampling periods. This work represents the most comprehensive
124 longitudinal genomic study of resident *E. coli* within the human gastrointestinal tract and
125 expands knowledge of the non-pathogen gut flora by increasing the available genome
126 sequences of resident *E. coli* and highlighting the dynamic nature of the *E. coli*
127 community.

128

129 **Results**

130 *Selection of E. coli strains for genome sequencing.*

131 A total of 247 *E. coli* isolates from 30 subjects (17 male and 13 female as shown in
132 Figure 2) in the study by Seidman et al. (21, 22) were selected for DNA extraction and
133 genome assembly, based on the criteria that these subjects contributed the most
134 complete longitudinal collection of isolates (i.e. the greatest number of subjects with the
135 greatest number of possible isolates). Of these, 240 isolates provided acceptable
136 sequence quality to generate genome assemblies with a genome size and GC-content
137 that is characteristic of *E. coli* to be analyzed using comparative genomics. The average

138 genome size was 5.17 Mb (range 4.46 to 5.81 Mb) with a 50.69% GC (range 50.21 to
139 51.04%), similar to other known *E. coli* genomes (Table S1). Of the 240 isolates, 120
140 isolates were from the subjects that received the antibiotic treatment of single oral dose
141 of prophylactic azithromycin, and 120 isolates from subjects in the non-treatment
142 (control) group (Table S1 and Fig.2).

143

144 *Subject clinical state and E. coli pathotype identification.*

145 There were 17 instances in which subjects had active diarrhea at the time of sample
146 collection (12 instances occurred at the baseline time point), yielding 46 isolates from
147 diarrheal conditions (21, 22), 23 each from the antibiotic treatment and control groups.
148 All cases of diarrhea were identified in children under the age of 2 at baseline. Only 16
149 of these isolates (34.8%) contained canonical virulence factors belonging to the EPEC,
150 ETEC, or EAEC pathotypes (Fig. 2), as determined by sequence homology searches of
151 canonical virulence genes in the assembled genomes. In most cases, observed
152 diarrhea could not be associated with a prototypically virulent *E. coli* in this data set.
153 Other sources of diarrhea were not investigated.

154

155 An additional 61 isolates from 19 individuals contained canonical *E. coli* virulence
156 factors, but were not obtained from samples taken during an active diarrheal event.
157 These data indicate that the presence of a potentially virulent *E. coli* does necessarily
158 result in clinical presentation of diarrhea. Overall, in our dataset there was no evidence
159 of an association between diarrheal cases and incidence of isolates containing
160 canonical *E. coli* virulence factors.

161
162 *Phylogenomic analysis.*
163 Phylogenomic analysis of the isolates identified a diverse population of *E. coli* within the
164 gastrointestinal community of these children. A phylogenetic tree of the 240 isolates
165 from this study plus 33 reference *E. coli* and *Shigella* genomes (Table S2) was used to
166 assess the genomic similarity of the isolates from a single subject both within and
167 across time points, as well as between subjects over the study period (Fig. 3). The
168 SNP-based phylogenomic analysis of the draft and reference genomes identified
169 304,497 polymorphic single nucleotide genomic sites. The isolates from the current
170 study were identified in the established *E. coli* phylogroups: A (132 isolates), B1 (62
171 isolates), B2 (24 isolates), D (17 isolates), and E (2 isolates) (Fig. 3, Table S1).
172 Additionally, three isolate genomes (isolates 1_176_05_S3_C2, 2_011_08_S1_C1, and
173 2_156_04_S3_C2) fell into cryptic clades located outside of the established *E. coli*
174 phylogroups. The distributions of the *E. coli* isolates in each of these phylogroups were
175 not associated with any of the clinical parameters associated with these isolates.
176
177 To further investigate the *E. coli* diversity of an individual subject at a given time, we
178 analyzed the phylogenetic groupings of isolates from each subject at each time point.
179 Most isolates from an individual at a single time point group together within a single
180 phylogenomic lineage, where a lineage is defined as a terminal grouping of isolates
181 (54.4%; 49 of the 90 same-subject time points). One third (35.5%; 32/90 of the same-
182 subject time point isolates) fell into two distinct lineages, and in 10% (9/90 time points),
183 all isolates belonged to a distinct lineage (Table 1). Overall, these data suggest that

184 while there is considerable diversity among the isolates from many subjects, in over half
185 of them, the population of *E. coli* at a given time point displays limited phylogenomic
186 variation.

187

188 These *E. coli* populations were variable over time, showing increased *E. coli* diversity in
189 each subject when observed over the multiple time points. Same-subject isolates from
190 different time points reside in distinct phylogenomic lineages in 93.3% (28/30) of
191 subjects. Only two subjects had isolates from multiple time points that occupied the
192 same clade. Subject 4_203_08 had one 3-month isolate that was most similar to the two
193 6-month isolates (Fig. 3). Additionally, subject 8_415_05 had all of the 3-month and 6-
194 month isolates belonging to the same phylogenomic lineage (Fig. 3). Similarly, for
195 subject 2_052_05, all 3-month isolates and one 6-month isolate are in neighboring
196 lineages, suggesting a close phylogenetic relationship.

197

198 Only three subjects (1_182_04, 1_250_04, 6_319_05) had a single phylogenomic clade
199 of isolates at each of the three time points (illustrated in Fig. 3 and detailed in Table S3),
200 suggesting colonization by a single dominant clone at any one time point, but dynamic
201 *E. coli* populations between each of the time points. In contrast, all isolates from subject
202 3_475_03 were phylogenomically distinct (Fig. 3). Additionally, the isolates from eight
203 subjects (26.7%) are represented in at least six distinct phylogenetic groups (Table 1).
204 To our knowledge, this level of phylogenomic diversity of *E. coli* in the human
205 gastrointestinal tract over relatively short time periods has not been previously reported.

206

207 *Multilocus sequence typing and molecular serotyping.*

208 The genomes in this study comprise a combined total of 87 sequence types (STs)
209 (Table S1). The most common ST was ST10, which was represented by 40 of the *E.*
210 *coli* genomes, while 40 additional STs occurred only once (Table S1). Only five isolates
211 were from ST131, which has been demonstrated to be associated with the spread of
212 antimicrobial resistance (PMID 24694052). There was, on average, 1.5 (range 1-3) STs
213 among isolates from a subject at a single time point, and an average of 4.4 (range 2-7)
214 STs per subject across all time points. Since the total number of available isolates per
215 subject varied, the values were normalized per the number of isolates, revealing an
216 average of 2 (range 1-4) isolates per sequence type and mimicking the diversity
217 observed in the phylogenetic analyses (Fig. 4, Table S3).

218
219 Similar to MLST, serotype analyses (26) reflect the diversity observed in the
220 phylogenomic analysis (Table S3). The 240 isolates represent a combined total of 106
221 O:H serotypes, with 54 of them only occurring once in the dataset, making serotype a
222 finer-scale measure of diversity than MLST. There are an average of 1.63 (range 1-3)
223 different serotypes in isolates from the same time point and 4.7 (range 2-7) serotypes in
224 a subject across all time points. The O, H, or either serotypes could not be predicted in
225 33 isolates (Table S1). *In silico* analyses were unable to distinguish between some
226 serotypes in an additional 58 isolates (Table S1). This left 149 isolates that could be
227 unambiguously assigned a single serotype (Table S1).

228

229 Nearly all isolates that shared a serotype also shared an MLST sequence type and
230 phylogroup (Table S1). There are five examples (excluding those isolates in which the
231 serotype could not be unambiguously differentiated) where MLST, serotype, and
232 phylogroup were not congruent (Table S4), suggesting molecular variation and strain
233 differentiation could not be detected by a single method alone. The combination of
234 these detailed molecular methods could add nuance to diversity measurements in
235 closely related strains.

236

237 *Genome content using LS-BSR*

238 Variations in genome content further demonstrated the diversity of the *E. coli* isolate
239 genomes both within and between time points. Using the LS-BSR analysis (27) and an
240 ergatis-based annotation pipeline, a gene content profile was determined which
241 identified 32,950 genes in the pangenome of the 240 isolate genomes. More than 3,000
242 genes in any single genome was comprised of genes that vary between genomes,
243 leaving only approximately 2000 genes in the conserved core, as has been previously
244 identified (10, 17). This level of variation is true even among the isolates from subject
245 8_415_05 in which the isolates from the 3-month and 6-month time points group
246 together phylogenetically, and are of the same MLST sequence type. In this case, each
247 isolate contains an average of 220 (range 95-259) variable genes. Given the level of
248 diversity suggested by the variability of the gene content, more detailed SNP analyses,
249 as previously preformed by Stoesser (18) were deemed unnecessary.

250

251 *Antibiotic resistance associated gene profiles*

252 The antibiotic treatment of half of the children in this study provided a unique
253 opportunity to investigate the impact of antibiotic treatment on the prevalence and
254 maintenance of antibiotic resistance genes in the *E. coli* community at 3 and 6 months
255 after administration. Antibiotic resistance genes were investigated in the isolate
256 genomes using 1,371 genes from the Comprehensive Antibiotic Resistance Database
257 (CARD) (28). The resistance gene profiles (assortment of present/absent genes) for
258 each isolate were used to create a cladogram to investigate the relationships among
259 isolates by time and by subject (Fig. S2). These relationships were then compared to
260 those in the phylogenetic groupings as well as in the cladogram of virulence gene
261 profiles (Table S5, Figure S3). Similar clustering patterns were identified between the
262 whole genome phylogeny or virulence gene presence and resistance gene-based
263 analysis 74% of the time at each time point, and 37% (phylogeny) or 27% (virulence) of
264 the time for each subject as a whole (Table 1).

265
266 There was no significant change in number or type of resistance-associated genes over
267 time, regardless of antibiotic treatment or isolation time point. As subjects were treated
268 with azithromycin, a macrolide, genes conferring resistance to macrolides were
269 investigated in greater detail (Table S6). Macrolide resistance genes were identified in
270 only 19% (46 of the 240) isolates (Table 2) and based on a logistic regression model,
271 there is no evidence to suggest that either time point or antibiotic treatment were
272 significantly associated with macrolide resistance genes ($p > 0.05$ for antibiotic treatment
273 adjusted for time point, for time point adjusted for antibiotic treatment, and overall
274 antibiotic treatment). Isolates from nearly half of the subjects had no known macrolide

275 resistance genes (46.67% antibiotic treatment, 40% control). Based on these results,
276 exposure to a single large dose of azithromycin did not lead to a significant change in
277 the number of known antimicrobial resistance genes or macrolide resistance genes
278 among these *E. coli* populations.

279

280 **Discussion**

281 This study represents a detailed examination of the genomic diversity of *Escherichia coli*
282 isolates obtained from longitudinal samples obtained from the gastrointestinal tract of
283 children. An overall trend identified in this study is that the identified *E. coli* from the
284 human gastrointestinal tract are diverse not just between subjects, but within the same
285 subject over time. The *E. coli* genomes sequenced in this study, were selected based
286 on the greatest number of longitudinal isolates per subject and include members of all
287 five of the traditional *E. coli* phylogroups, as well as 87 different MLST sequence types,
288 and 106 serotypes. The isolates in this study were most frequently of the A or B1
289 phylogroups, unlike a previous study by Gordon et al (29) in which greater than 70% of
290 the isolates obtained were either from phylogroup B2 or D. This observed difference
291 may be due to differences in sample acquisition (stool swab versus biopsy), or
292 differences in the study participants. The Gordon et al (29) study, obtained samples
293 from adults, the majority (72.5%, 50/69) of whom were diagnosed with either Crohn's
294 disease or ulcerative colitis, which would likely impact the immune status of the
295 gastrointestinal tract, and potentially alter the bacterial community structure. In contrast
296 our study participants were children under the age of 5, and, other than a few that
297 displayed diarrhea of an unknown source, were considered to be relatively healthy. This

298 study, by using a combination of molecular methods, including whole genome
299 sequencing, enhances the understanding that the *E. coli* in the human gastrointestinal
300 tract is variable and diverse.

301
302 Approximately half of *E. coli* isolates in an individual appear phylogenomically and
303 phenotypically similar at any given time point; however, even isolates that appeared
304 clonal based on MLST, phylogroup or serotype still contain unique genomic regions.
305 Gene content analyses revealed variation between isolates thought to be clonal by each
306 of the other methods. However, between time points, the prevalent *E. coli* clones from
307 individual subjects were variable. Only two subjects (4_203_08 and 8_415_05) had
308 isolates that were closely related based on the phylogenomic and other molecular data,
309 at more than one time point (Fig. 3, Fig.4). The more common observations were that
310 distinct and prevalent isolates were present at each 3-month sampling interval.

311
312 Previous studies of the variability of *E. coli*, using non-genome sequencing methods,
313 have also identified multiple isolates within a single host, reporting up to 4 *E. coli*
314 genotypes in adult human gastrointestinal studies (18, 29). The findings in this study are
315 similar in that it has identified a number of *E. coli* isolates that are genomically and
316 molecularly different in the subjects at each time, and between time points. While it is
317 possible, and likely, that in the current study less prevalent *E. coli* isolates were not
318 captured at some of the sampling time points, we assume that there was little bias in the
319 selection of the isolates, and that the relative isolate abundance in culture reflects the

320 relative abundance in the feces at the time of sampling. The current study likely still
321 underestimates the *E. coli* diversity in the examined subjects.

322

323 Dynamic populations within the human gastrointestinal tract have been previously
324 suggested as an explanation for observations of variable clones in *E. coli* diversity
325 studies (30), but the necessary longitudinal genomic studies were lacking. This study
326 begins to address that deficiency. The observed within-patient and longitudinal diversity
327 of *E. coli* isolates could be a function of age, as all of the subjects in this study were less
328 than three years of age, and thus the diversity could be a result of natural introduction of
329 new exposure to foods, as well as immune system and microbiome development (31,
330 32). It has been demonstrated that intra-host *E. coli* diversity is greatest in tropical
331 regions where hygiene may play a role and that *E. coli* density in the gastrointestinal
332 tract is altered most significantly in the first two years of a child's life (11, 33), therefore,
333 it is unclear how well these results correlate with *E. coli* diversity in adults or in other
334 geographic regions. It is thought that the infant microbiome is not established until
335 about three years of age (34), however the detailed longitudinal infant microbiome
336 studies are currently lacking. Future longitudinal studies that include sampling subjects
337 from multiple age groups will be necessary to fully appreciate levels of bacterial
338 population diversity and dynamics present across host populations of all age groups.

339

340 Virulence and resistance-associated gene analyses in this study confirm that genomic
341 analyses of single isolates are imperfect predictors of clinical phenotypes, as several
342 isolates harbored canonical *E. coli* virulence genes, classically identifying them as

343 enteric pathogens, but were present in subjects not displaying clinical symptoms, The
344 converse is also possible, in that *E. coli* strains may not contain traditional virulence
345 factors, but be obtained from a diarrheal sample, as has been highlighted in the recent
346 GEMS studies (35, Platts-Mills 2015). There are many potential explanations for these
347 observations which include: 1) the subjects have been previously exposed to these
348 bacteria, and thus, have an established immunity, 2) the organisms are not pathogenic
349 in the context of other host factors, including the host microbiota, 3) additional
350 necessary virulence factors are absent in these isolates, or 4) the virulence factors are
351 present but not expressed by the bacterium. Unfortunately, detailed immunological,
352 microbiota or transcriptional data are not available on the current samples, so the
353 impacts of these factors on pathogenicity cannot be determined conclusively. Whole
354 genome analyses have led to increasing recognition that virulence genes and
355 phylogeny are associated attributes in microbial pathogen genome and suggests that
356 there may be an optimal combination of chromosomal and virulence associated features
357 that results in maximal virulence, survival or transmission (36-39). This may also be true
358 of the success of a commensal isolate in the community (40).

359

360 This study adds significantly to the number of available *E. coli* genomes that were not
361 selected for based on pathogenic traits, a group that has been traditionally
362 underrepresented in the sequencing of this species. The scientific community is still in
363 the early stages of understanding gastrointestinal tract microbial ecology and the role
364 that the resident bacteria, including *E. coli*, play in microbiome stability and function.
365 The current study demonstrates that at the genomic level, the community of *E. coli* in

366 the human infant gastrointestinal tract is diverse and variable over time. Further studies
367 on human populations from different geographic areas, as well as other age groups, are
368 required to determine if *E. coli* communities would stabilize as a person approaches
369 adulthood, or whether the community diversity of *E. coli* regularly changes depending
370 on the development of the immune system, as well as many other exposures within the
371 gastrointestinal tract.

372

373 **Materials and Methods.**

374 *Isolate selection.*

375 *E. coli* isolates in this study were selected from isolates collected in Seidman *et al.* (21).
376 The PRET+ study was a 6-month, study designed to assess the ancillary effects on
377 pneumonia, diarrhea and malaria in children following mass distribution of azithromycin
378 for trachoma control. The study was conducted in 8 communities in the `Kongwa, a
379 district located in rural central Tanzania on a semiarid highland plateau with poor
380 access to drinking water. The district has a total population of approximately 248,656,
381 comprising mostly herders and subsistence farmers. The Tanzanian government
382 stipulates that villages with trachoma prevalence $\geq 10\%$ receive annual mass distribution
383 of azithromycin. On survey, 4 villages found eligible for antibiotic treatment became the
384 PRET+ treatment villages and 4 neighboring ineligible communities were included as
385 controls. The study methods and results detailing the impact of antibiotic treatment on
386 pneumonia and diarrhea morbidity, and antibiotic-resistant *Streptococcus pneumoniae*
387 carriage were published previously (41-43).

388

389 The selected *E. coli* isolates were chosen to represent individuals with the most
390 complete longitudinal sample sets from the PRET+ *E. coli* sub-study. Isolates were
391 obtained from 30 individuals between 2-35 months of age, living in 8 villages in the
392 same rural area of Tanzania. Half of these individuals received antibiotic treatment,
393 while the other half (control) received no antibiotic treatment. These isolates were
394 cultured from fecal samples collected at three time points (Figure 1 and Table S1): a
395 baseline prior to antibiotic treatment, three months post-treatment, and six months post-
396 treatment, with corresponding time points in the untreated controls. At each time point,
397 up to three *E. coli* colonies per individual were selected for sequencing and subsequent
398 comparative analyses.

399

400 *Bacterial growth and isolation*

401 *E. coli* colonies were obtained as described in Seidman *et al* (21, 22). Briefly, fecal
402 swabs were streaked on MacConkey agar (Difco) and grown overnight at 37°C. Three
403 lactose fermentation (LF) positive colonies were inoculated on nutrient agar stabs and
404 grown overnight at 37°C. *E. coli* isolates were identified as those colonies which were
405 LF-positive, indole-positive (DMACA Indole Reagent droppers, BD), and citrate-negative
406 (Simmons citrate agar slants). Isolates were transferred to Luria broth for overnight
407 growth at 37°C with shaking. *E. coli* cultures were frozen with 10% glycerol and stored
408 at -80°C.

409

410 *Genome sequencing and assembly.* Genomic DNA was extracted using standard
411 methods (16) and sequenced on the Illumina HiSeq 2000 platform at the Genome

412 Resource Center at the University of Maryland School of Medicine, Institute for Genome
413 Sciences. The resulting 100bp reads were assembled as previously described (36, 38).
414 The assembly details and corresponding GenBank accession numbers are provided in
415 Table S1.

416

417 *Identification of predicted pathogen isolates*

418 Isolate genomes were interrogated for the presence of pathotype-specific virulence
419 factor genes using LS-BSR and are derived from a similar *E. coli* typing schema used in
420 the MAL-ED studies (44). The nucleotide sequence for each factor or resistance gene
421 was aligned against all sequenced genomes with BLASTN (45) in conjunction with LS-
422 BSR (27). Genes with a BSR value ≥ 0.80 were considered highly conserved and
423 present in the isolate examined. The targeted virulence factors are as follows: ETEC
424 heat stable enterotoxin (estA147) or ETEC heat labile enterotoxin (eltb508) identifying
425 the isolate as being enterotoxigenic *E. coli* (ETEC); the *aggR*-activated island C
426 (*aic215*) or EAEC ABC transporter A (*aata650*) genes which are common diagnostic
427 markers for enteroaggregative *E. coli* (EAEC) (46, 47); and the major subunit of the
428 bundle-forming pilus (*bfpA*) (*bfpa300*) or intimin genes (*eae881*) which are indicative of
429 enteropathogenic *E. coli* (EPEC) (36).

430

431 *Phylogenomic analysis.*

432 A total of 273 genomes were used in the phylogenomic analyses: the 240 assembled in
433 this study, in addition to a collection of 33 *E. coli* and *Shigella* reference genomes from
434 GenBank (Table S2). Single nucleotide polymorphisms (SNPs) in all genomes were

435 detected relative to the completed genome sequence of commensal isolate *E. coli* HS
436 (phylogroup A) using the *In Silico* Genotyper (ISG) v.0.12.2 (48), which uses MUMmer
437 v.3.22 (49) for SNP detection. Analysis with ISG yielded 701,011 total SNP sites that
438 were filtered to a subset of 304,497 SNP sites present in all of the genomes analyzed.
439 These SNP sites were concatenated and used for phylogenetic analysis as previously
440 described (50). A maximum-likelihood phylogeny with 1000 bootstrap replicates was
441 generated using RAxML v.7.2.8 (51) and visualized using FigTree v.1.4.2
442 (<http://tree.bio.ed.ac.uk/software/figtree/>) and Interactive tree of life (52). Clades were
443 assigned based on visual determination of groupings. Three genome outliers
444 (1_176_05_S3_C2, 2_011_08_S1_C1, and 2_156_04_S3_C2) were removed from the
445 tree figures for visualization purposes.

446

447 *Serotype identification.*

448 *In silico* serotype identification was performed on the assembled genomes using the
449 online SerotypeFinder 1.1 (<https://cge.cbs.dtu.dk/services/SerotypeFinder/>) and an LS-
450 BSR analysis using the serotype sequences compiled for the SRS2 program
451 (<https://github.com/katholt/srst2/tree/master/data>) (15, 26).

452

453 *Multilocus sequence typing (MLST).*

454 *In silico* MLST was performed on the assembled genomes using the Achtman *E. coli*
455 MLST scheme (53). Gene sequences were identified in the isolate genomes using
456 BLASTn and MLST profiles were determined by querying the PubMLST database
457 (<http://pubmlst.org>).

458

459 *Variations in gene distributions.*

460 The gene content across all genomes was identified and compared using the large-
461 scale BLAST score ratio (LS-BSR) with default settings, as previously described (27).
462 Genes with a BSR value ≥ 0.80 are considered to be highly conserved and present in
463 the isolate examined at this level of homology. Those genes that are conserved in all
464 genomes were removed from further analyses. The predicted protein function of each
465 gene cluster was determined using an ergatis-based (54) in-house annotation pipeline
466 (55).

467

468 *Virulence factor and antibiotic resistance gene identification.*

469 The list of compiled common *E. coli* virulence factors genes was used for interrogation
470 of the study genomes (Table S2). Antibiotic resistance genes were compiled from the
471 Comprehensive Antibiotic Resistance Database (CARD; <http://arpcard.mcmaster.ca>,
472 downloaded June 24, 2015) (28). The nucleotide sequence for each factor or resistance
473 gene was aligned against all sequenced genomes with BLASTN (45) in conjunction with
474 LS-BSR (27). Genes with a BSR value ≥ 0.80 were considered highly conserved and
475 present in the isolate examined.

476

477 *Statistical analysis of macrolide resistance gene distributions*

478 A logistic regression on the probability of a macrolide gene being present in an *E coli*
479 isolate was run against 2 covariates: time point (excluding the baseline) or antibiotic
480 treatment. For each individual, the two to three isolates were considered replicates for

481 that time point, and the time points were far enough apart to be considered
482 independent. Therefore, gene presence was collapsed as presence in at least one of
483 the replicates at a given subject and time point. Each subject by time combination was
484 considered an independent observation. Genes in this analysis with p-values ≤ 0.05
485 were considered significant. If the covariate was dichotomous, then the Wald Chi-
486 Square test statistic was used to determine significance.

487

488 **FUNDING.**

489 The PRET+ study and isolate collection was funded by a grant from the Bill & Melinda
490 Gates Foundation, Seattle, WA, USA (#48027), an unrestricted grant from Research to
491 Prevent Blindness and a grant from the Johns Hopkins Global Water Program. The
492 sequencing and analysis component of the project was funded in part by federal funds
493 from the National Institute of Allergy and Infectious Diseases, National Institutes of
494 Health, Department of Health and Human Services under contract number
495 HHSN272200900009C, grant number U19AI110820 and National Institute of Diabetes
496 and Digestive and Kidney Diseases 2T32DK067872-11 (TKSR).

497

498

499 **REFERENCES.**

- 500
- 501 1. **Kaper JB, Nataro JP, Mobley HL.** 2004. Pathogenic *Escherichia coli*. *Nat Rev*
- 502 *Microbiol* **2**:123-140.
- 503 2. **Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB.** 2013.
- 504 Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin*
- 505 *Microbiol Rev* **26**:822-880.
- 506 3. **Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam**
- 507 **S, Wu Y, Sow SO, Sur D, Breiman RF, Faruque AS, Zaidi AK, Saha D,**
- 508 **Alonso PL, Tamboura B, Sanogo D, Onwuchekwa U, Manna B, Ramamurthy**
- 509 **T, Kanungo S, Ochieng JB, Omore R, Oundo JO, Hossain A, Das SK,**
- 510 **Ahmed S, Qureshi S, Quadri F, Adegbola RA, Antonio M, Hossain MJ,**
- 511 **Akinsola A, Mandomando I, Nhampossa T, Acacio S, Biswas K, O'Reilly CE,**
- 512 **Mintz ED, Berkeley LY, Muhsen K, Sommerfelt H, Robins-Browne RM,**
- 513 **Levine MM.** 2013. Burden and aetiology of diarrhoeal disease in infants and
- 514 young children in developing countries (the Global Enteric Multicenter Study,
- 515 GEMS): a prospective, case-control study. *Lancet* **382**:209-222.
- 516 4. **Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham**
- 517 **J, Adair T, Aggarwal R, Ahn SY, Alvarado M, Anderson HR, Anderson LM,**
- 518 **Andrews KG, Atkinson C, Baddour LM, Barker-Collo S, Bartels DH, Bell ML,**
- 519 **Benjamin EJ, Bennett D, Bhalla K, Bikbov B, Bin Abdulhak A, Birbeck G,**
- 520 **Blyth F, Bolliger I, Boufous S, Bucello C, Burch M, Burney P, Carapetis J,**
- 521 **Chen H, Chou D, Chugh SS, Coffeng LE, Colan SD, Colquhoun S, Colson**
- 522 **KE, Condon J, Connor MD, Cooper LT, Corriere M, Cortinovis M, de**
- 523 **Vaccaro KC, Couser W, Cowie BC, Criqui MH, Cross M, Dabhadkar KC, et**
- 524 **al.** 2012. Global and regional mortality from 235 causes of death for 20 age
- 525 groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease
- 526 Study 2010. *Lancet* **380**:2095-2128.
- 527 5. **Manges AR, Johnson JR.** 2015. Reservoirs of Extraintestinal Pathogenic
- 528 *Escherichia coli*. *Microbiol Spectr* **3**.
- 529 6. **Yassour M, Vatanen T, Siljander H, Hamalainen AM, Harkonen T, Ryhanen**
- 530 **SJ, Franzosa EA, Vlamakis H, Huttenhower C, Gevers D, Lander ES, Knip**

- 531 **M, Xavier RJ.** 2016. Natural history of the infant gut microbiome and impact of
532 antibiotic treatment on bacterial strain diversity and stability. *Sci Transl Med*
533 **8**:343ra381.
- 534 7. **Tamburini S, Shen N, Wu HC, Clemente JC.** 2016. The microbiome in early
535 life: implications for health outcomes. *Nat Med* **22**:713-722.
- 536 8. **Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI.** 2008. Worlds within
537 worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**:776-788.
- 538 9. **Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon**
539 **JI, Relman DA, Fraser-Liggett CM, Nelson KE.** 2006. Metagenomic Analysis of
540 the Human Distal Gut Microbiome. *Science* **312**:1355-1359.
- 541 10. **Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P,**
542 **Crabtree J, Sebahia M, Thomson NR, Chaudhuri R, Henderson IR,**
543 **Sperandio V, Ravel J.** 2008. The pangenome structure of *Escherichia coli*:
544 comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J*
545 *Bacteriol* **190**:6881-6893.
- 546 11. **Tenaillon O, Skurnik D, Picard B, Denamur E.** 2010. The population genetics
547 of commensal *Escherichia coli*. *Nat Rev Microbiol* **8**:207-217.
- 548 12. **Apperloo-Renkema HZ, Van der Waaij BD, Van der Waaij D.** 1990.
549 Determination of colonization resistance of the digestive tract by biotyping of
550 *Enterobacteriaceae*. *Epidemiol Infect* **105**:355-361.
- 551 13. **Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA.**
552 2016. Investigating the Relatedness of Enteroinvasive *Escherichia coli* to Other
553 *E. coli* and *Shigella* Isolates by Using Comparative Genomics. *Infect Immun*
554 **84**:2362-2371.
- 555 14. **Hazen TH, Donnenberg MS, Panchalingam S, Antonio M, Hossain A,**
556 **Mandomando I, Ochieng JB, Ramamurthy T, Tamboura B, Qureshi S,**
557 **Quadri F, Zaidi A, Kotloff KL, Levine MM, Barry EM, Kaper JB, Rasko DA,**
558 **Nataro JP.** 2016. Genomic diversity of EPEC associated with clinical
559 presentations of differing severity. *Nat Microbiol* **1**:15014.
- 560 15. **Ingle DJV, M.; Kuzevski, A.; Tauschek, M., Inouye, M.; Stinear, T.; Levine,**
561 **M. M.; Robins-Browne, R. M.; Holt, K. E.** 2016. In silico serotyping of *E. coli*

- 562 from short read data identifies limited novel O-loci but extensive diversity of O:H
563 serotype combinations within and between pathogenic lineages. *Microbial*
564 *Genomics* **2**.
- 565 16. **Sahl JW, Johnson JK, Harris AD, Phillippy AM, Hsiao WW, Thom KA, Rasko**
566 **DA**. 2011. Genomic comparison of multi-drug resistant invasive and colonizing
567 *Acinetobacter baumannii* isolated from diverse human body sites reveals
568 genomic plasticity. *BMC Genomics* **12**:291.
- 569 17. **Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E,**
570 **Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O,**
571 **Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L,**
572 **Ghigo JM, Gilles AM, Johnson J, Le Bouguenec C, Lescat M, Mangenot S,**
573 **Martinez-Jehanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z,**
574 **Ruf CS, Schneider D, Turret J, Vacherie B, Vallenet D, Medigue C, Rocha**
575 **EP, Denamur E**. 2009. Organised genome dynamics in the *Escherichia coli*
576 species results in highly diverse adaptive paths. *PLoS Genet* **5**:e1000344.
- 577 18. **Stoesser N, Sheppard AE, Moore CE, Golubchik T, Parry CM, Nget P,**
578 **Saroeun M, Day NP, Giess A, Johnson JR, Peto TE, Crook DW, Walker AS,**
579 **Modernizing Medical Microbiology Informatics G**. 2015. Extensive Within-
580 Host Diversity in Fecally Carried Extended-Spectrum-Beta-Lactamase-Producing
581 *Escherichia coli* Isolates: Implications for Transmission Analyses. *J Clin Microbiol*
582 **53**:2122-2131.
- 583 19. **Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, Park SH, Ooka T, Iyoda**
584 **S, Taylor TD, Hayashi T, Itoh K, Hattori M**. 2008. Complete genome sequence
585 and comparative analysis of the wild-type commensal *Escherichia coli* strain
586 SE11 isolated from a healthy adult. *DNA Res* **15**:375-386.
- 587 20. **Vejborg RM, Friis C, Hancock V, Schembri MA, Klemm P**. 2010. A virulent
588 parent with probiotic progeny: comparative genomics of *Escherichia coli* strains
589 CFT073, Nissle 1917 and ABU 83972. *Mol Genet Genomics* **283**:469-484.
- 590 21. **Seidman JC, Coles CL, Silbergeld EK, Levens J, Mkocho H, Johnson LB,**
591 **Munoz B, West SK**. 2014. Increased carriage of macrolide-resistant fecal *E. coli*

- 592 following mass distribution of azithromycin for trachoma control. *Int J Epidemiol*
593 **43**:1105-1113.
- 594 22. **Seidman JC, Johnson LB, Levens J, Mkocha H, Munoz B, Silbergeld EK,**
595 **West SK, Coles CL.** 2016. Longitudinal Comparison of Antibiotic Resistance in
596 Diarrheagenic and Non-pathogenic *Escherichia coli* from Young Tanzanian
597 Children. *Front Microbiol* **7**:1420.
- 598 23. **Calva JJ, Sifuentes-Osornio J, Ceron C.** 1996. Antimicrobial resistance in fecal
599 flora: longitudinal community-based surveillance of children from urban Mexico.
600 *Antimicrob Agents Chemother* **40**:1699-1702.
- 601 24. **Monira S, Shabnam SA, Ali SI, Sadique A, Johura FT, Rahman KZ, Alam NH,**
602 **Watanabe H, Alam M.** 2017. Multi-drug resistant pathogenic bacteria in the gut
603 of young children in Bangladesh. *Gut Pathog* **9**:19.
- 604 25. **Pons MJ, Mosquito S, Gomes C, Del Valle LJ, Ochoa TJ, Ruiz J.** 2014.
605 Analysis of quinolone-resistance in commensal and diarrheagenic *Escherichia*
606 *coli* isolates from infants in Lima, Peru. *Trans R Soc Trop Med Hyg* **108**:22-28.
- 607 26. **Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F.** 2015.
608 Rapid and Easy In Silico Serotyping of *Escherichia coli* Isolates by Use of Whole-
609 Genome Sequencing Data. *J Clin Microbiol* **53**:2410-2426.
- 610 27. **Sahl JW, Caporaso JG, Rasko DA, Keim P.** 2014. The large-scale blast score
611 ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between
612 bacterial genomes. *PeerJ* **2**:e332.
- 613 28. **McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar**
614 **K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M,**
615 **Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P,**
616 **Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright**
617 **GD.** 2013. The comprehensive antibiotic resistance database. *Antimicrob Agents*
618 *Chemother* **57**:3348-3357.
- 619 29. **Gordon DM, O'Brien CL, Pavli P.** 2015. *Escherichia coli* diversity in the lower
620 intestinal tract of humans. *Environ Microbiol Rep* **7**:642-648.
- 621 30. **Smati M, Clermont O, Le Gal F, Schichmanoff O, Jaureguy F, Eddi A,**
622 **Denamur E, Picard B.** 2013. Real-time PCR for quantitative analysis of human

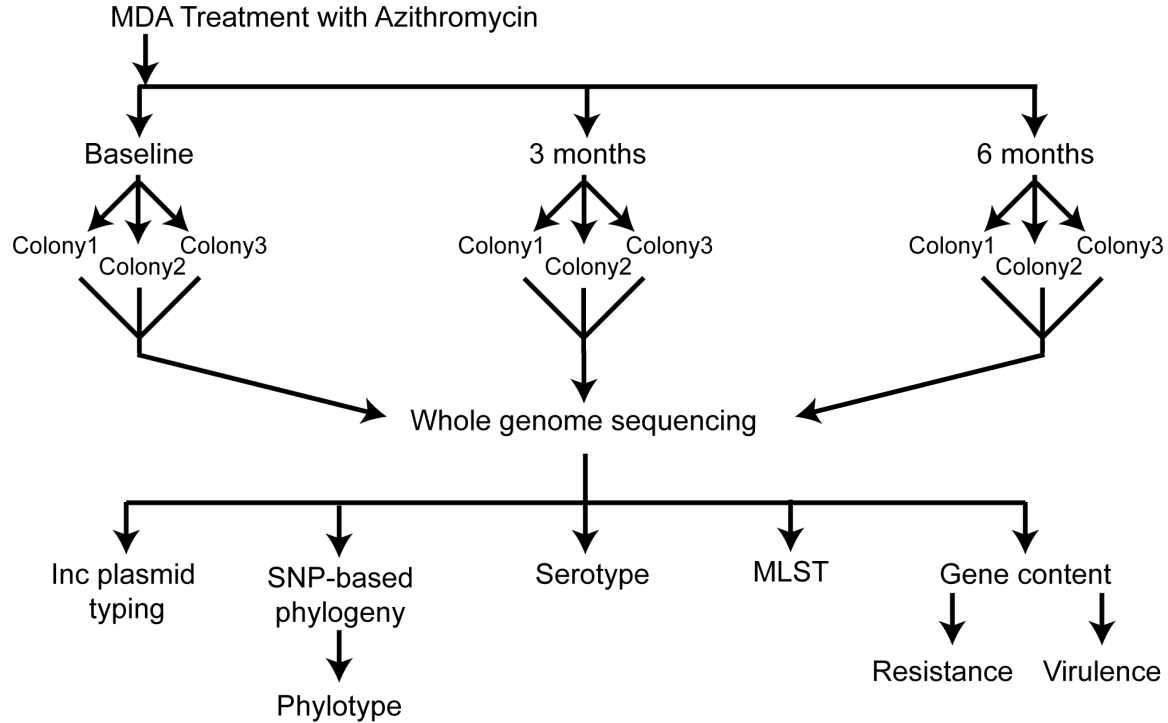
- 623 commensal *Escherichia coli* populations reveals a high frequency of
624 subdominant phylogroups. *Appl Environ Microbiol* **79**:5005-5012.
- 625 31. **van Best N, Hornef MW, Savelkoul PH, Penders J.** 2015. On the origin of
626 species: Factors shaping the establishment of infant's gut microbiota. *Birth*
627 *Defects Res C Embryo Today* **105**:240-251.
- 628 32. **Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO.** 2007. Development of
629 the human infant intestinal microbiota. *PLoS Biol* **5**:e177.
- 630 33. **Skurnik D, Bonnet D, Bernede-Bauduin C, Michel R, Guette C, Becker JM,**
631 **Balaire C, Chau F, Mohler J, Jarlier V, Boutin JP, Moreau B, Guillemot D,**
632 **Denamur E, Andreumont A, Ruimy R.** 2008. Characteristics of human intestinal
633 *Escherichia coli* with changing environments. *Environ Microbiol* **10**:2132-2137.
- 634 34. **Yatsunencko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG,**
635 **Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC,**
636 **Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C,**
637 **Clemente JC, Knights D, Knight R, Gordon JI.** 2012. Human gut microbiome
638 viewed across age and geography. *Nature* **486**:222-227.
- 639 35. **Lindsay B, Ochieng JB, Ikumapayi UN, Toure A, Ahmed D, Li S,**
640 **Panchalingam S, Levine MM, Kotloff K, Rasko DA, Morris CR, Juma J,**
641 **Fields BS, Dione M, Malle D, Becker SM, Houpt ER, Nataro JP, Sommerfelt**
642 **H, Pop M, Oundo J, Antonio M, Hossain A, Tamboura B, Stine OC.** 2013.
643 Quantitative PCR for detection of *Shigella* improves ascertainment of *Shigella*
644 burden in children with moderate-to-severe diarrhea in low-income countries. *J*
645 *Clin Microbiol* **51**:1740-1746.
- 646 36. **Hazen TH, Sahl JW, Fraser CM, Sonnenberg MS, Scheutz F, Rasko DA.**
647 2013. Refining the pathovar paradigm via phylogenomics of the attaching and
648 effacing *Escherichia coli*. *Proc Natl Acad Sci USA* **110**:12810-12815.
- 649 37. **von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR,**
650 **Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm AM,**
651 **Sjoling A, Dougan G.** 2014. Identification of enterotoxigenic *Escherichia coli*
652 (ETEC) clades with long-term global distribution. *Nat Genet* **46**:1321-1326.

- 653 38. **Donnenberg MS, Hazen TH, Farag TH, Panchalingam S, Antonio M, Hossain**
654 **A, Mandomando I, Ochieng JB, Ramamurthy T, Tamboura B, Zaidi A,**
655 **Levine MM, Kotloff K, Rasko DA, Nataro JP.** 2015. Bacterial factors associated
656 with lethal outcome of enteropathogenic *Escherichia coli* infection: genomic case-
657 control studies. PLoS Negl Trop Dis **9**:e0003791.
- 658 39. **Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, Kingsley**
659 **RA, Thomson NR, Keane JA, Weill FX, Edwards DJ, Hawkey J, Harris SR,**
660 **Mather AE, Cain AK, Hadfield J, Hart PJ, Thieu NT, Klemm EJ, Glinos DA,**
661 **Breiman RF, Watson CH, Kariuki S, Gordon MA, Heyderman RS, Okoro C,**
662 **Jacobs J, Lunguya O, Edmunds WJ, Msefula C, Chabalgoity JA, Kama M,**
663 **Jenkins K, Dutta S, Marks F, Campos J, Thompson C, Obaro S, MacLennan**
664 **CA, Dolecek C, Keddy KH, Smith AM, Parry CM, Karkey A, Mulholland EK,**
665 **Campbell JI, Dongol S, Basnyat B, Dufour M, Bandaranayake D, et al.** 2015.
666 Phylogeographical analysis of the dominant multidrug-resistant H58 clade of
667 *Salmonella* Typhi identifies inter- and intracontinental transmission events. Nat
668 Genet **47**:632-639.
- 669 40. **Blyton MD, Cornall SJ, Kennedy K, Colligon P, Gordon DM.** 2014. Sex-
670 dependent competitive dominance of phylogenetic group B2 *Escherichia coli*
671 strains within human hosts. Environ Microbiol Rep **6**:605-610.
- 672 41. **Coles CL, Mabula K, Seidman JC, Levens J, Mkocho H, Munoz B, Mfinanga**
673 **SG, West S.** 2013. Mass distribution of azithromycin for trachoma control is
674 associated with increased risk of azithromycin-resistant *Streptococcus*
675 *pneumoniae* carriage in young children 6 months after treatment. Clin Infect Dis
676 **56**:1519-1526.
- 677 42. **Coles CL, Levens J, Seidman JC, Mkocho H, Munoz B, West S.** 2012. Mass
678 distribution of azithromycin for trachoma control is associated with short-term
679 reduction in risk of acute lower respiratory infection in young children. Pediatr
680 Infect Dis J **31**:341-346.
- 681 43. **Coles CL, Seidman JC, Levens J, Mkocho H, Munoz B, West S.** 2011.
682 Association of mass treatment with azithromycin in trachoma-endemic

- 683 communities with short-term reduced risk of diarrhea in young children. *Am J*
684 *Trop Med Hyg* **85**:691-696.
- 685 44. **Haupt E, Gratz J, Kosek M, Zaidi AK, Qureshi S, Kang G, Babji S, Mason C,**
686 **Bodhidatta L, Samie A, Bessong P, Barrett L, Lima A, Havt A, Haque R,**
687 **Mondal D, Taniuchi M, Stroup S, McGrath M, Lang D, Investigators M-EN.**
688 2014. Microbiologic methods utilized in the MAL-ED cohort study. *Clin Infect Dis*
689 **59 Suppl 4**:S225-232.
- 690 45. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local
691 alignment search tool. *Journal of molecular biology* **215**:403-410.
- 692 46. **Lima IF, Boisen N, Quetz Jda S, Havt A, de Carvalho EB, Soares AM, Lima**
693 **NL, Mota RM, Nataro JP, Guerrant RL, Lima AA.** 2013. Prevalence of
694 enteroaggregative *Escherichia coli* and its virulence-related genes in a case-
695 control study among children from north-eastern Brazil. *J Med Microbiol* **62**:683-
696 693.
- 697 47. **Boisen N, Scheutz F, Rasko DA, Redman JC, Persson S, Simon J, Kotloff**
698 **KL, Levine MM, Sow S, Tamboura B, Toure A, Malle D, Panchalingam S,**
699 **Krogfelt KA, Nataro JP.** 2012. Genomic characterization of enteroaggregative
700 *Escherichia coli* from children in Mali. *J Infect Dis* **205**:431-444.
- 701 48. **Sahl JW, Beckstrom-Sternberg SM, Babic-Sternberg JS, Gillece JD, Hepp**
702 **CM, Auerbach RK, Tembe W, Wagner DM, Keim PS, Pearson T.** 2015. The *in*
703 *silico* genotyper (ISG): an open-source pipeline to rapidly identify and annotate
704 nucleotide variants for comparative genomics applications. bioRxiv
705 doi:<http://dx.doi.org/10.1101/015578>.
- 706 49. **Delcher AL, Salzberg SL, Phillippy AM.** 2003. Using MUMmer to identify
707 similar regions in large sequence sets. *Curr Protoc Bioinformatics* **Chapter**
708 **10**:Unit 10 13.
- 709 50. **Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, Fields B,**
710 **Breiman RF, Gilmour M, Nataro JP, Rasko DA.** 2015. Defining the
711 phylogenomics of *Shigella* species: A pathway to diagnostics. *J Clin Microbiol*
712 doi:10.1128/JCM.03527-14.

- 713 51. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic
714 analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688-
715 2690.
- 716 52. **Letunic I, Bork P.** 2016. Interactive tree of life (iTOL) v3: an online tool for the
717 display and annotation of phylogenetic and other trees. *Nucleic Acids Res*
718 **44**:W242-245.
- 719 53. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves**
720 **PR, Maiden MC, Ochman H, Achtman M.** 2006. Sex and virulence in
721 *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60**:1136-1151.
- 722 54. **Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S,**
723 **Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O,**
724 **Angiuoli SV.** 2010. Ergatis: a web interface and scalable software system for
725 bioinformatics workflows. *Bioinformatics* **26**:1488-1492.
- 726 55. **Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, White O, Wortman**
727 **J, Mahurkar A, Giglio MG.** 2011. The IGS standard operating procedure for
728 automated prokaryotic annotation. *Stand Genomic Sci* **4**:244-251.
729
730
731

732 **Figures.**



733

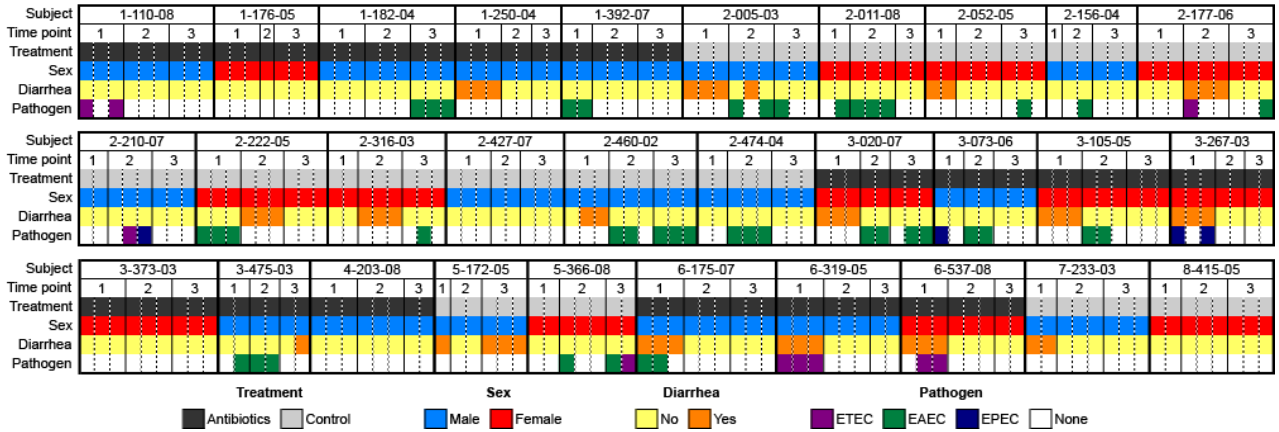
734

735 **Figure 1: Overall study design.** The overall design of the study highlighting the
736 sampling of up to three distinct colonies on three time points, one of which, termed the
737 baseline occurs prior to the administration of antibiotics in half of the subjects.

738

739

740



741

742

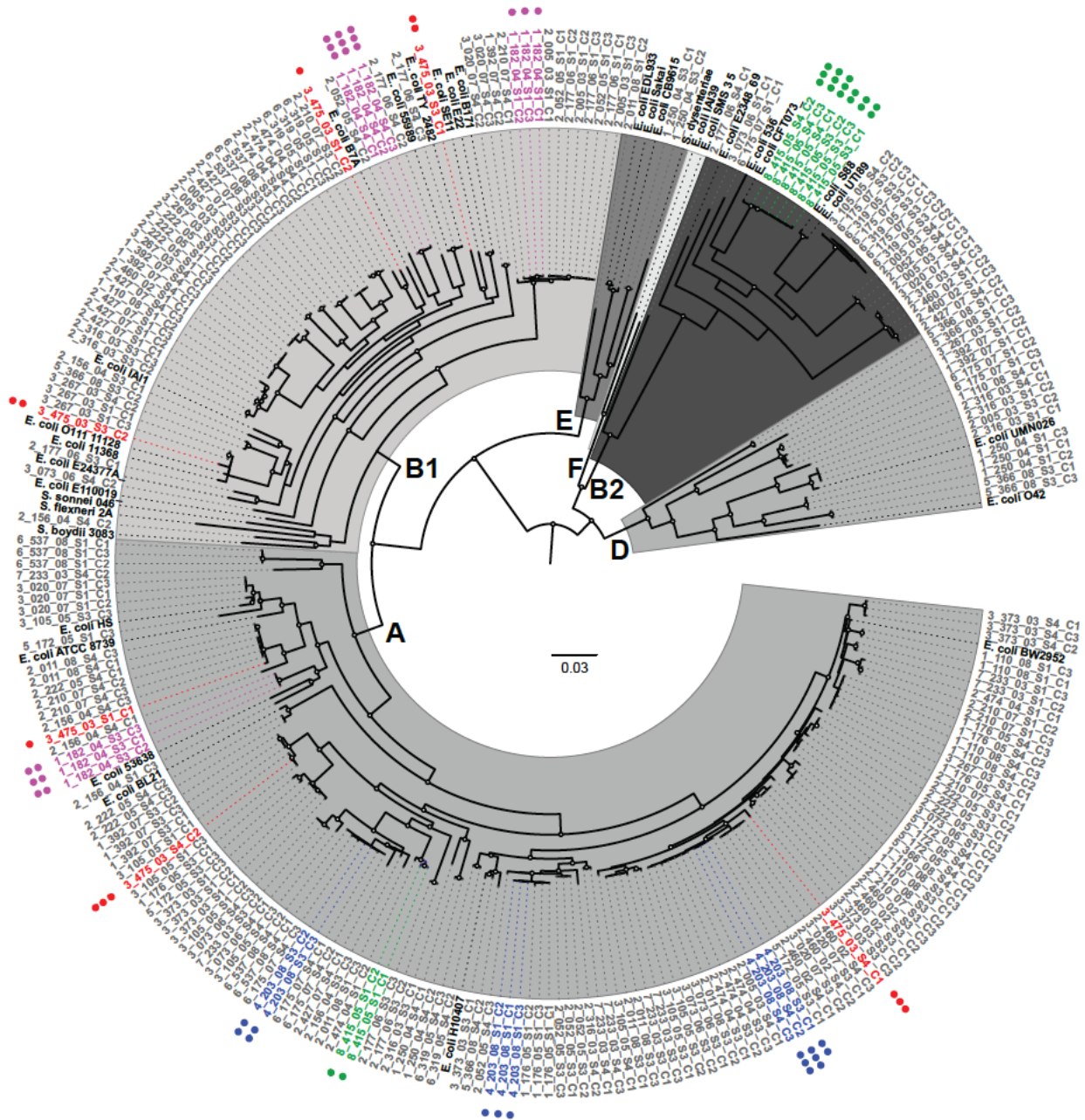
743 **Figure 2: Isolate metadata.** Summary of metadata showing time point of isolation,

744 treatment group, host sex, clinical presentation, and the identification of pathogenic

745 markers for ETEC, EAEC, or EPEC pathotypes for each isolate by subject. Further

746 details in Table S1

747



748

749

750

Figure 3. Phylogenomic analysis of *E. coli* isolates in study. A) A whole-genome

751 phylogeny of the isolate sequences and reference *E. coli* and *Shigella* genomes (shown

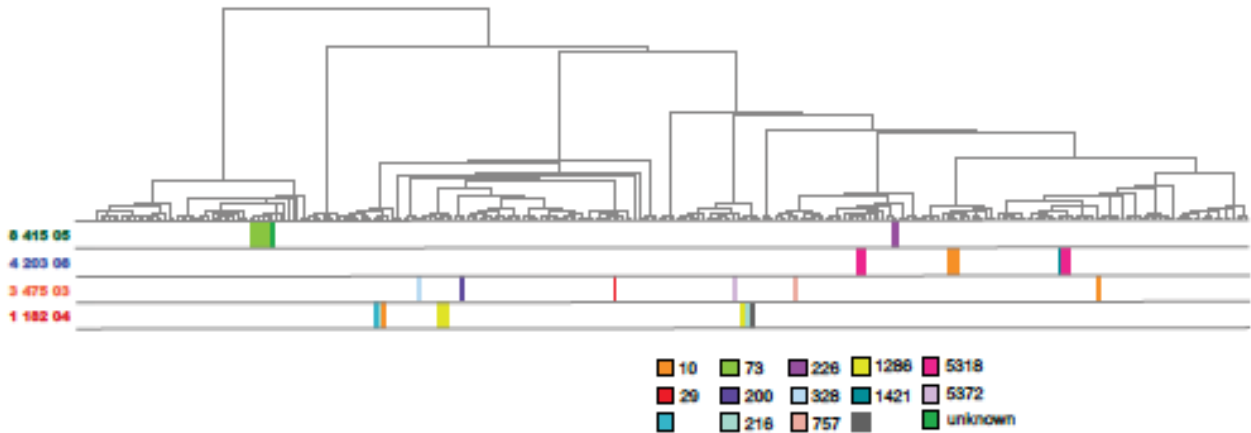
752 in black) highlighting examples of diversity among subject-specific isolates within and

753 across time points. The scale bar indicates the approximate distance of 0.03 nucleotide

754 substitutions per site. Nodes with bootstrap values of greater than 90 are marked with a

755 circle. Examples of isolates from subjects that demonstrate the greatest (3_475_03) and

756 least (4_203_08, 8_415_05, 1_182_04) amount of diversity are highlighted: 3_475_03
757 in red, 4_203_08 in blue, 8_415_05 in green, and 1_182_04 in purple. The number of
758 dots denote the sample number from which the isolate was obtained. *E. coli*
759 phylogroups are labeled. Full figure with all subjects is presented in Figure S1.
760



761

762

763 **Figure 4: Phylogenomic distribution of sequence types of isolates from select**
764 **subjects.** A cladogram of the phylogeny highlighting relative positions of genomes of
765 isolates from selected subjects with MLST sequence types shown in colored blocks
766 corresponding to the sequence type as shown in the legend. Selected example subjects
767 highlight low diversity within time points but high diversity across time (subject
768 1_182_04), high diversity within and across time (3_475_03), intermediate diversity
769 across time (4_203_08), and low diversity across time (8_415_05).

770

771

Table 1. Summary of isolate diversity within subject and within time point across several diversity measurements

Subject ID	Treatment	Isolate Phylogenomics		Resistance			Virulence			Phylogroup		MLST		Serotype	
		No. isolates from subject	No. of clades in subject*	No. resistance clades*	Isolates single resistance superclade	Similar distribution to phylogeny*	No. virulence gene clades*	Similar distribution to phylogeny*	Similar distribution to resistance genes*	No. phylogroups in subject*	Similar distribution to phylogeny*	No. sequence types in subject*	Similar distribution to phylogeny*	No. serotypes in subject*	Similar distribution to phylogeny*
1_110_08	MDA	9	5	5	No	No	5	No	Yes	3	No	3	No	5	Yes
1_176_05	MDA	8	4	4	No	Yes	4	Yes	Yes	2	No	3	No	4	Yes
1_182_04	MDA	9	3	5	No	No	3	Yes	No	2	No	3	Yes	3	Yes
1_250_04	MDA	7	3	2	Yes	No	3	Yes	No	3	Yes	3	Yes	3	Yes
1_392_07	MDA	8	4	5	No	No	4	Yes	No	3	No	4	Yes	4	Yes
3_020_07	MDA	8	4	4	No	Yes	4	Yes	Yes	5	Yes	5	Yes	4	No
3_073_06	MDA	7	5	5	No	Yes	4	No	No	3	No	6	No	7	No
3_105_05	MDA	9	7	6	No	Yes	7	Yes	No	3	No	5	Yes	5	Yes
3_267_03	MDA	7	6	6	No	Yes	6	Yes	Yes	2	No	6	No	7	Yes
3_373_03	MDA	9	4	4	No	Yes	3	No	No	3	No	6	Yes	6	Yes
3_475_03	MDA	6	6	5	No	No	6	Yes	No	2	No	6	Yes	6	Yes
4_203_08	MDA	8	3	5	No	No	3	No	No	2	No	4	Yes	4	Yes
6_175_07	MDA	9	4	5	No	Yes	5	Yes	Yes	3	No	6	Yes	5	No
6_319_05	MDA	8	3	5	No	No	3	Yes	No	3	No	5	No	7	No
6_537_08	MDA	8	3	5	No	No	3	Yes	No	3	No	4	Yes	5	No
2_005_03	No-MDA	9	5	7	No	No	6	Yes	No	2	No	4	Yes	4	Yes
2_011_08	No-MDA	8	6	5	No	No	6	Yes	No	3	No	4	Yes	4	Yes
2_052_05	No-MDA	8	5	4	No	Yes	5	No	No	3	No	5	No	5	No
2_156_04	No-MDA	7	7	5	No	Yes	6	No	No	2	No	7	Yes	7	Yes
2_177_06	No-MDA	9	6	5	No	No	7	No	No	3	No	6	Yes	6	Yes
2_210_07	No-MDA	8	6	6	No	Yes	5	No	No	1	No	3	No	4	Yes
2_222_05	No-MDA	9	4	5	No	No	4	Yes	No	2	No	6	Yes	6	Yes
2_316_03	No-MDA	8	6	7	No	No	5	No	No	1	No	2	No	3	Yes
2_427_07	No-MDA	8	5	4	No	Yes	6	No	No	1	No	4	Yes	4	Yes
2_460_02	No-MDA	9	4	4	No	Yes	4	Yes	Yes	3	No	6	No	5	Yes
2_474_04	No-MDA	8	4	3	No	No	3	No	Yes	3	No	4	Yes	4	Yes
5_172_05	No-MDA	6	4	3	No	No	4	Yes	No	3	Yes	4	No	3	Yes
5_366_08	No-MDA	7	5	5	No	Yes	4	No	No	2	No	3	Yes	3	Yes
7_233_03	No-MDA	8	5	5	Yes	Yes	5	Yes	Yes	1	No	4	No	6	No
8_415_05	No-MDA	8	2	3	No	No	2	Yes	No	2	Yes	3	No	2	Yes

*Further details are provided in Supplemental Table 3

772
773

774 Table 2: Summary of macrolide resistance gene presence by treatment group and time
 775 point. The proportion of isolates in which a macrolide resistance gene was identified is
 776 shown for each time point. Subjects are separated in to treatment groups and
 777 categorized based on the time points in which macrolide resistance genes are identified.
 778 Percentages reflect the proportion of subjects that fall in to each macrolide resistance
 779 gene category within treatment groups.

	Treatment					No treatment				
	Subject	Time point			%	Subject	Time point			%
		1	2	3			1	2	3	
<i>No macrolide resistance genes</i>	3_073_06	0	0	0	46.67% (7/15)	2_052_05	0	0	0	40% (6/15)
	3_373_03	0	0	0		2_156_04	0	0	0	
	3_475_03	0	0	0		2_177_06	0	0	0	
	4_203_08	0	0	0		2_222_05	0	0	0	
	6_175_07	0	0	0		2_474_04	0	0	0	
	6_319_05	0	0	0		8_415_05	0	0	0	
	6_537_08	0	0	0						
<i>Only in 3-month</i>	1_176_05	0	0.5	1	13.33% (2/15)	2_005_03	0	0.66	0	33.33% (5/15)
	1_182_04	0	0.66	0		2_011_08	0	0.66	0	
						2_210_07	0	0.33	0	
						5_366_08	0	0.66	0	
<i>Only in 6-month</i>	1_110_08	0	0	1	13.33%	2_316_03	0	0	0.66	13.33%
	1_392_07	0	0	0.66	(2/15)	2_427_07	0	0	0.66	(2/15)
<i>Pre- & post-treatment</i>	1_250_04	1	1	1	13.33%	2_460_02	0.66	0	1	6.67%
	3_105_05	0.33	0.33	0.33	(2/15)					(1/15)
<i>3- and 6-month</i>	3_020_07	0	1	0.66	13.33%					0.00%
	3_267_03	0	0.5	0.5	(2/15)					
<i>Only baseline</i>					0.00%	5_172_05	1	0	0	6.67% (1/15)

780

781