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

Authors: Taylor K. S. Richter[1,2], Tracy H. Hazen[1,2], Diana Lam[6], Christian L. Coles[4], Jessica C. Seidman[5], Yaqi You[6], Ellen K. Silbergeld[5], Claire M. Fraser[1,3] and David A. Rasko[1,2*]

[1] The Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA; [2] Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD, USA; [3] Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA; [4] Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; [5]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, MD, USA; [6] Department of Environmental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Running title: Population dynamics of *E. coli*

* Corresponding Author
David A. Rasko
University of Maryland School of Medicine
Institute for Genome Sciences
Department of Microbiology and Immunology
801 W. Baltimore Street, Suite 619
Baltimore, MD 21201
Phone: 410.706.6774
E-mail: drasko@som.umaryland.edu

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

**Abstract.**

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

**Importance.**

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

## Introduction

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

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

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

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

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

## Results

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

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

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

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

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

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

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

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

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

*Multilocus sequence typing and molecular serotyping.*

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

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

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

*Genome content using LS-BSR*

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

*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

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

## Discussion

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

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

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

Virulence and resistance-associated gene analyses in this study confirm that genomic analyses of single isolates are imperfect predictors of clinical phenotypes, as several 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 ≥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

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

*Bacterial growth and isolation*

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

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

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

*Identification of predicted pathogen isolates*

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

*Phylogenomic analysis*.

A total of 273 genomes were used in the phylogenomic analyses: the 240 assembled in this study, in addition to a collection of 33 *E. coli* and *Shigella* reference genomes from 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

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

**FUNDING.**

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

## REFERENCES.


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

**M, Xavier RJ.** 2016. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci Transl Med **8:**343ra381.

7. **Tamburini S, Shen N, Wu HC, Clemente JC.** 2016. The microbiome in early life: implications for health outcomes. Nat Med **22:**713-722.

8. **Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI.** 2008. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol **6:**776-788.

9. **Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE.** 2006. Metagenomic Analysis of the Human Distal Gut Microbiome. Science **312:**1355-1359.

10. **Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J.** 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol **190:**6881-6893.

11. **Tenaillon O, Skurnik D, Picard B, Denamur E.** 2010. The population genetics of commensal Escherichia coli. Nat Rev Microbiol **8:**207-217.

12. **Apperloo-Renkema HZ, Van der Waaij BD, Van der Waaij D.** 1990. Determination of colonization resistance of the digestive tract by biotyping of Enterobacteriaceae. Epidemiol Infect **105:**355-361.

13. **Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA.** 2016. Investigating the Relatedness of Enteroinvasive Escherichia coli to Other E. coli and Shigella Isolates by Using Comparative Genomics. Infect Immun **84:**2362-2371.

14. **Hazen TH, Donnenberg MS, Panchalingam S, Antonio M, Hossain A, Mandomando I, Ochieng JB, Ramamurthy T, Tamboura B, Qureshi S, Quadri F, Zaidi A, Kotloff KL, Levine MM, Barry EM, Kaper JB, Rasko DA, Nataro JP.** 2016. Genomic diversity of EPEC associated with clinical presentations of differing severity. Nat Microbiol **1:**15014.

15. **Ingle DJV, M.; Kuzevski, A.; Tauschek, M., Inouye, M.; Stinear, T.; Levine, M. M.; Robins-Browne, R. M.; Holt, K. E.** 2016. In silico serotyping of E. coli

from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. Microbial Genomics **2**.

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

following mass distribution of azithromycin for trachoma control. Int J Epidemiol **43:**1105-1113.

22. **Seidman JC, Johnson LB, Levens J, Mkocha H, Munoz B, Silbergeld EK, West SK, Coles CL.** 2016. Longitudinal Comparison of Antibiotic Resistance in Diarrheagenic and Non-pathogenic Escherichia coli from Young Tanzanian Children. Front Microbiol **7:**1420.

23. **Calva JJ, Sifuentes-Osornio J, Ceron C.** 1996. Antimicrobial resistance in fecal flora: longitudinal community-based surveillance of children from urban Mexico. Antimicrob Agents Chemother **40:**1699-1702.

24. **Monira S, Shabnam SA, Ali SI, Sadique A, Johura FT, Rahman KZ, Alam NH, Watanabe H, Alam M.** 2017. Multi-drug resistant pathogenic bacteria in the gut of young children in Bangladesh. Gut Pathog **9:**19.

25. **Pons MJ, Mosquito S, Gomes C, Del Valle LJ, Ochoa TJ, Ruiz J.** 2014. Analysis of quinolone-resistance in commensal and diarrheagenic Escherichia coli isolates from infants in Lima, Peru. Trans R Soc Trop Med Hyg **108:**22-28.

26. **Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F.** 2015. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol **53:**2410-2426.

27. **Sahl JW, Caporaso JG, Rasko DA, Keim P.** 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ **2:**e332.

28. **McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD.** 2013. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother **57:**3348-3357.

29. **Gordon DM, O'Brien CL, Pavli P.** 2015. Escherichia coli diversity in the lower intestinal tract of humans. Environ Microbiol Rep **7:**642-648.

30. **Smati M, Clermont O, Le Gal F, Schichmanoff O, Jaureguy F, Eddi A, Denamur E, Picard B.** 2013. Real-time PCR for quantitative analysis of human

commensal Escherichia coli populations reveals a high frequency of subdominant phylogroups. Appl Environ Microbiol **79:**5005-5012.

31. **van Best N, Hornef MW, Savelkoul PH, Penders J.** 2015. On the origin of species: Factors shaping the establishment of infant's gut microbiota. Birth Defects Res C Embryo Today **105:**240-251.

32. **Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO.** 2007. Development of the human infant intestinal microbiota. PLoS Biol **5:**e177.

33. **Skurnik D, Bonnet D, Bernede-Bauduin C, Michel R, Guette C, Becker JM, Balaire C, Chau F, Mohler J, Jarlier V, Boutin JP, Moreau B, Guillemot D, Denamur E, Andremont A, Ruimy R.** 2008. Characteristics of human intestinal Escherichia coli with changing environments. Environ Microbiol **10:**2132-2137.

34. **Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI.** 2012. Human gut microbiome viewed across age and geography. Nature **486:**222-227.

35. **Lindsay B, Ochieng JB, Ikumapayi UN, Toure A, Ahmed D, Li S, Panchalingam S, Levine MM, Kotloff K, Rasko DA, Morris CR, Juma J, Fields BS, Dione M, Malle D, Becker SM, Houpt ER, Nataro JP, Sommerfelt H, Pop M, Oundo J, Antonio M, Hossain A, Tamboura B, Stine OC.** 2013. Quantitative PCR for detection of Shigella improves ascertainment of Shigella burden in children with moderate-to-severe diarrhea in low-income countries. J Clin Microbiol **51:**1740-1746.

36. **Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, Rasko DA.** 2013. Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*. Proc Natl Acad Sci USA **110:**12810-12815.

37. **von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm AM, Sjoling A, Dougan G.** 2014. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. Nat Genet **46:**1321-1326.

38. **Donnenberg MS, Hazen TH, Farag TH, Panchalingam S, Antonio M, Hossain A, Mandomando I, Ochieng JB, Ramamurthy T, Tamboura B, Zaidi A, Levine MM, Kotloff K, Rasko DA, Nataro JP.** 2015. Bacterial factors associated with lethal outcome of enteropathogenic *Escherichia coli* infection: genomic case-control studies. PLoS Negl Trop Dis **9:**e0003791.

39. **Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, Kingsley RA, Thomson NR, Keane JA, Weill FX, Edwards DJ, Hawkey J, Harris SR, Mather AE, Cain AK, Hadfield J, Hart PJ, Thieu NT, Klemm EJ, Glinos DA, Breiman RF, Watson CH, Kariuki S, Gordon MA, Heyderman RS, Okoro C, Jacobs J, Lunguya O, Edmunds WJ, Msefula C, Chabalgoity JA, Kama M, Jenkins K, Dutta S, Marks F, Campos J, Thompson C, Obaro S, MacLennan CA, Dolecek C, Keddy KH, Smith AM, Parry CM, Karkey A, Mulholland EK, Campbell JI, Dongol S, Basnyat B, Dufour M, Bandaranayake D, et al.** 2015. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. Nat Genet **47:**632-639.

40. **Blyton MD, Cornall SJ, Kennedy K, Colligon P, Gordon DM.** 2014. Sex-dependent competitive dominance of phylogenetic group B2 Escherichia coli strains within human hosts. Environ Microbiol Rep **6:**605-610.

41. **Coles CL, Mabula K, Seidman JC, Levens J, Mkocha H, Munoz B, Mfinanga SG, West S.** 2013. Mass distribution of azithromycin for trachoma control is associated with increased risk of azithromycin-resistant Streptococcus pneumoniae carriage in young children 6 months after treatment. Clin Infect Dis **56:**1519-1526.

42. **Coles CL, Levens J, Seidman JC, Mkocha H, Munoz B, West S.** 2012. Mass distribution of azithromycin for trachoma control is associated with short-term reduction in risk of acute lower respiratory infection in young children. Pediatr Infect Dis J **31:**341-346.

43. **Coles CL, Seidman JC, Levens J, Mkocha H, Munoz B, West S.** 2011. 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. **Houpt 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.

51. **Stamatakis A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688-2690.

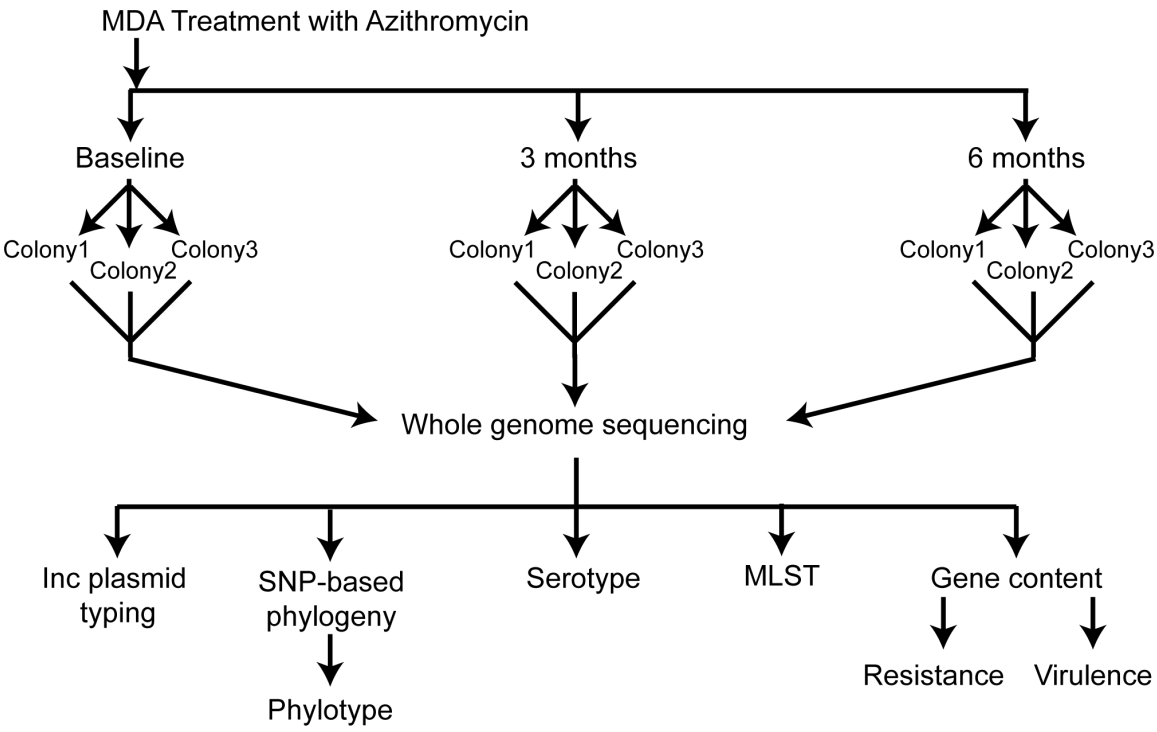
52. **Letunic I, Bork P.** 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res **44:**W242-245.

53. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M.** 2006. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol **60:**1136-1151.

54. **Orvis J, Crabtree J, Galens K, Gussman A, Inman JM, Lee E, Nampally S, Riley D, Sundaram JP, Felix V, Whitty B, Mahurkar A, Wortman J, White O, Angiuoli SV.** 2010. Ergatis: a web interface and scalable software system for bioinformatics workflows. Bioinformatics **26:**1488-1492.

55. **Galens K, Orvis J, Daugherty S, Creasy HH, Angiuoli S, White O, Wortman J, Mahurkar A, Giglio MG.** 2011. The IGS standard operating procedure for automated prokaryotic annotation. Stand Genomic Sci **4:**244-251.

732 **Figures.**




733

734

735 **Figure 1: Overall study design.** The overall design of the study highlighting the

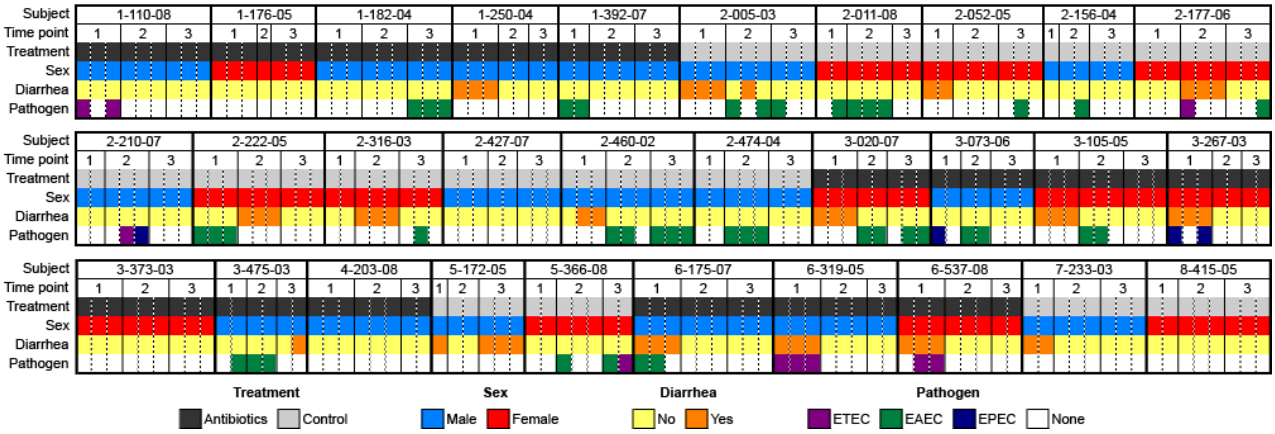
736 sampling of up to three distinct colonies on three time points, one of which, termed the

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

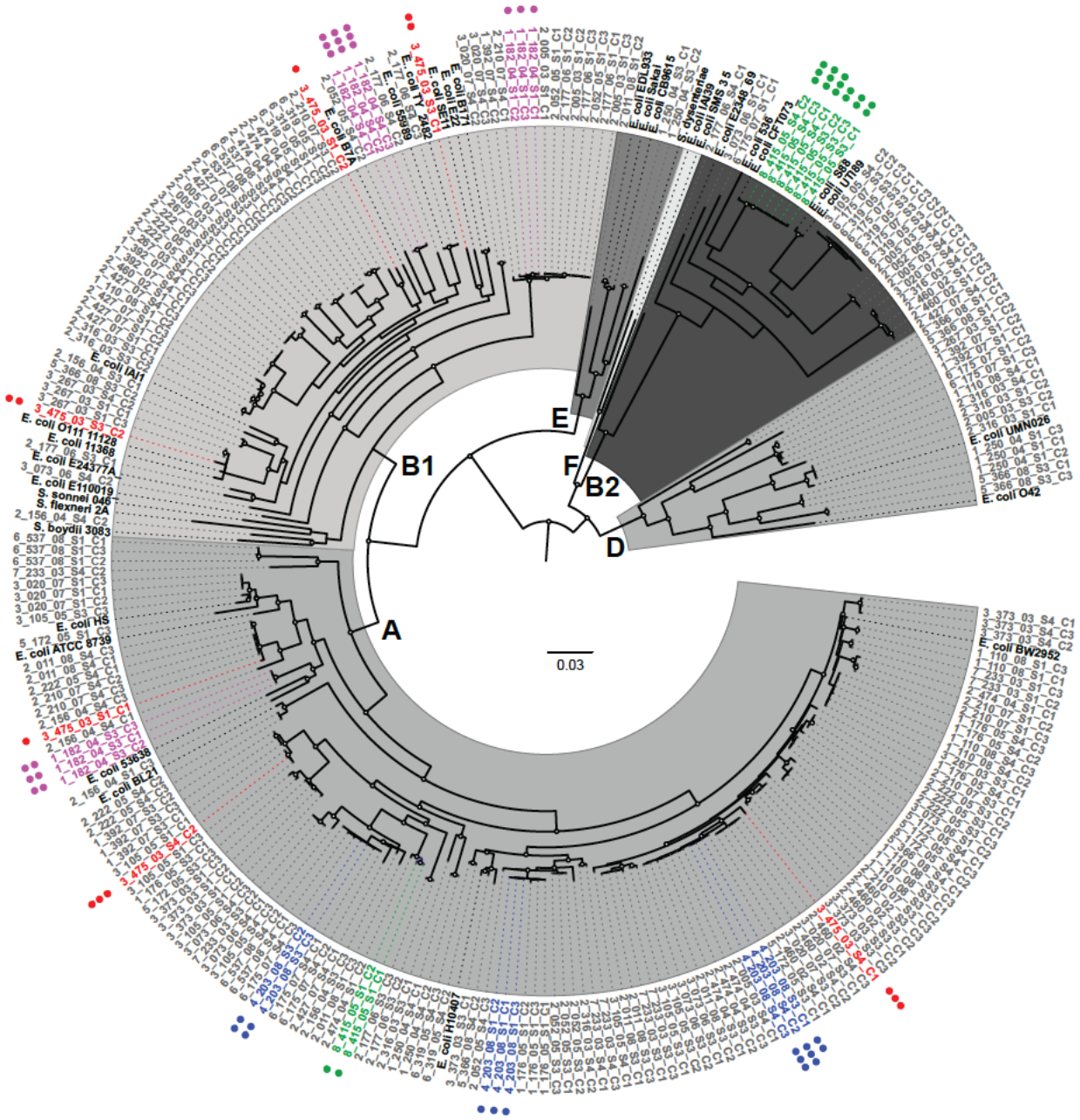
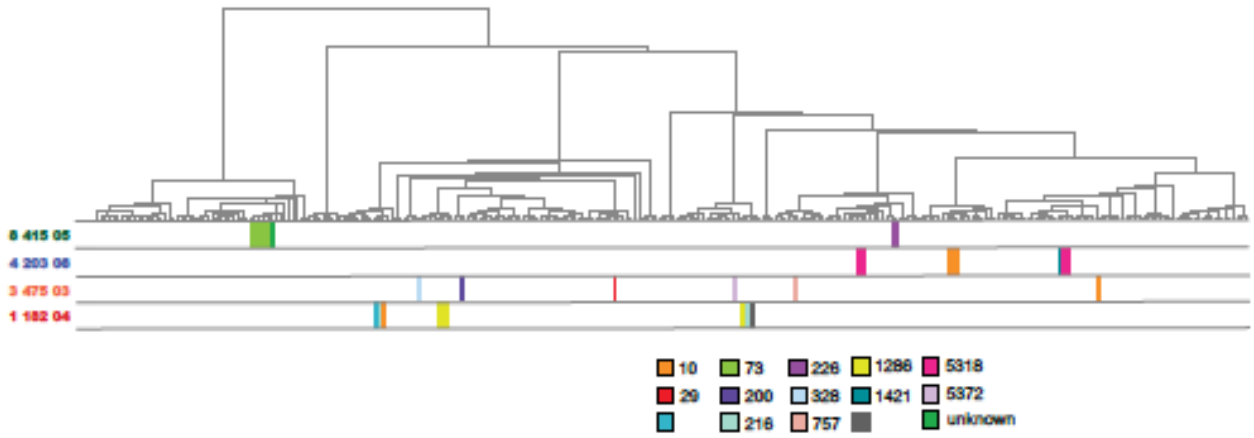



**Figure 3. Phylogenomic analysis of *E. coli* isolates in study. A)** A whole-genome phylogeny of the isolate sequences and reference *E. coli* and *Shigella* genomes (shown in black) highlighting examples of diversity among subject-specific isolates within and across time points. The scale bar indicates the approximate distance of 0.03 nucleotide substitutions per site Nodes with bootstrap values of greater than 90 are marked with a 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**

| | | Isolate Phylogenomics | | Resistance | | | Virulence | | | Phylogroup | | MLST | | Serotype | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Subject ID | Treatment | 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

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.

| Time points in which macrolide resistance genes are found | Treatment | | | | | No treatment | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Time point | | | | | Time point | | | |
| | *Subject* | *1* | *2* | *3* | *%* | *Subject* | *1* | *2* | *3* | *%* |
| *No macrolide resistance genes* | 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 | | | | | | |
| *Only in 3-month* | 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 | |
| | | | | | | 7_233_03 | 0 | 0.66 | 0 | |
| *Only in 6-month* | 1_110_08 | 0 | 0 | 1 | 13.33% (2/15) | 2_316_03 | 0 | 0 | 0.66 | 13.33% (2/15) |
| | 1_392_07 | 0 | 0 | 0.66 | | 2_427_07 | 0 | 0 | 0.66 | |
| *Pre- & post-treatment* | 1_250_04 | 1 | 1 | 1 | 13.33% (2/15) | 2_460_02 | 0.66 | 0 | 1 | 6.67% (1/15) |
| | 3_105_05 | 0.33 | 0.33 | 0.33 | | | | | | |
| *3- and 6-month* | 3_020_07 | 0 | 1 | 0.66 | 13.33% (2/15) | | | | | 0.00% |
| | 3_267_03 | 0 | 0.5 | 0.5 | | | | | | |
| *Only baseline* | | | | | 0.00% | 5_172_05 | 1 | 0 | 0 | 6.67% (1/15) |

780

781