Examination of the Enterotoxigenic *Escherichia coli* Population Structure during Human Infection

Jason W. Sahl,[a,b] Jeticia R. Sistrunk,[a] Claire M. Fraser,[a] Erin Hine,[a] Nabilah Baby,[c] Yasmin Begum,[c] Qingwei Luo,[d] Alaullah Sheikh,[c,e] Firdausi Qadri,[c] James M. Fleckenstein,[d,e,f] David A. Rasko[a]

Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA[a]; Translational Genomics Research Institute, Flagstaff, Arizona, USA[b]; Centre for Vaccine Sciences, Immunology, Laboratory, International Centre Center for Diarrhoeal Disease Research, Mohakhali, Dhaka, Bangladesh[c]; Department of Medicine, Division of Infectious Diseases, Washington University in St. Louis, St. Louis, Missouri, USA[d]; The Molecular Microbiology and Microbial Pathogenesis Program, Division of Biology and Biomedical Sciences, Washington University in St. Louis, St. Louis, Missouri, USA[e]; Medicine Service, Veterans Affairs Medical Center, St. Louis, Missouri, USA[f]

ABSTRACT **Enterotoxigenic *E. coli* (ETEC) can cause severe diarrhea and death in children in developing countries; however, bacterial diversity in natural infection is uncharacterized. In this study, we explored the natural population variation of ETEC from individuals with cholera-like diarrhea. Genomic sequencing and comparative analysis of multiple ETEC isolates from twelve cases of severe diarrhea demonstrated clonal populations in the majority of subjects (10/12). In contrast, a minority of individuals (2/12) yielded phylogenomically divergent ETEC isolates. Detailed examination revealed that isolates also differed in virulence factor content. These genomic data suggest that severe, cholera-like ETEC infections are largely caused by a clonal population of organisms within individual patients. Additionally, the isolation of similar clones from geographically and temporally dispersed cases with similar clinical presentations suggests that some isolates are particularly suited for virulence. The identification of multiple genomically diverse isolates with variable virulence factor profiles from a single subject highlights the dynamic nature of ETEC, as well as a potential weakness in the examination of cultures obtained from a single colony in clinical settings. These findings have implications for vaccine design and provide a framework for the study of population variation in other human pathogens.**

IMPORTANCE **Enterotoxigenic *Escherichia coli* (ETEC) has been identified as one of the major causes of diarrheal diseases in children as well as travelers. It has been previously appreciated that this pathogenic variant of *E. coli* is diverse, both at the genomic level, as defined with multilocus sequence typing, and with regard to the presence or absence of virulence factors within clonal groups. Using whole-genome sequencing and comparative analysis, we identified and characterized diverse enterotoxigenic *E. coli* isolates from individual patients. In 17% of patients, we identified multiple distinct ETEC isolates, each with unique genomic features and in some cases diverse virulence factor profiles. These studies ascertained that any one person may be colonized by multiple pathogenic ETEC isolates, which may impact how we think about the development of vaccines and therapeutics against these organisms.**

Received 26 March 2015 **Accepted** 19 May 2015 **Published** 9 June 2015

Citation Sahl JW, Sistrunk JR, Fraser CM, Hine E, Baby N, Begum Y, Luo Q, Sheikh A, Qadri F, Fleckenstein JM, Rasko DA. 2015. Examination of the enterotoxigenic *Escherichia coli* population structure during human infection. mBio 6(3):e00501-15. doi:10.1128/mBio.00501-15.

Editor Michael S. Gilmore, Harvard Medical School

Address correspondence to David A. Rasko, drasko@som.umaryland.edu.

Enterotoxigenic *Escherichia coli* (ETEC) has been identified as one of the major causes of death due to diarrheal disease among children under the age of five in developing countries by the recent landmark publication of the Global Enteric Multisite Study (1). Although genetically diverse, the ETEC pathovar is molecularly defined by genes encoding heat-labile (LT) and/or heat-stable (ST) enterotoxins. For disease presentation, these toxins must be successfully delivered to cognate receptors on epithelial cells of the small intestine, where ensuing loss of salt and water in the lumen results in diarrhea (2). Studies of children in developing countries (3), as well as adults in clinical trials (4), demonstrate that prior infections with wild-type ETEC are protective. Nonetheless, despite considerable effort (5), no ETEC vaccine to date has afforded sustained broad-based protection, suggesting that vaccine preparations may need to incorporate additional antigens to achieve protective immunity. Similarly, the lack of an effective vaccine may in part relate to the considerable genetic variability exhibited by ETEC relative to other *E. coli* pathovars when gene-based typing systems (6, 7) or whole-genome scale analyses (8, 9) are used; this is a concept supported by early studies of prototype ETEC isolates (8, 10). While it is known that ETEC isolates can be genomically variable, detailed examination of that variability has not been financially or practically feasible prior to the advent of new sequencing technologies.

Interestingly, while many molecular epidemiology studies have been performed on collections of stored ETEC isolates, originally obtained from single colonies, archived over time, and later interrogated for potential virulence factors or putative ETEC vac-

cine targets, the diversity of the overall ETEC population from which these individual colonies are selected has not been examined. The advent of rapid, cost-effective automated DNA sequencing provides opportunities to examine in detail genetic variation within the population of bacteria from individual infections, as well as to complete comparative analyses to isolates from disparate sources.

This study examined the ETEC population variability of isolates recovered from individuals with severe cholera-like diarrhea using genomic comparison and detailed examination of virulence factors.

Bacterial strains. The ETEC bacterial strains analyzed in this study were isolated from liquid stool samples of individuals being treated for severe cholera-like diarrhea at the International Centre for Diarrhoeal Disease Research, Mohakhali, Dhaka (http://www.icddrb.org), Bangladesh, or the treatment center in the Mirpur district of Dhaka. Multiple lactose-fermenting colonies were selected from MacConkey agar culture plates and screened using multiplex PCR for genes encoding heat-labile toxin, as well as human and porcine heat-stable toxin (STh and STp), as previously described (11). Isolated colonies of ETEC were then grown overnight in Luria-Bertani (LB) medium at 37°C with shaking and preserved as glycerol stocks stored at −80°C. Included for comparison in these studies were isolates from geographically and temporally disparate sources, including the ThroopD strain, isolated from a patient with severe cholera-like diarrhea in Dallas, TX, in 1975 (12), and several isolates (Juruá_18/11, Juruá_20/10, Envira_10/1, and Envira_8/11) obtained during ETEC outbreaks that caused severe diarrheal illness in two small villages, Juruá and Envira, in the Amazonia region of Brazil in 1998 (13) (see Table S1 in the supplemental material). A total of 208 new ETEC isolates were included in this study.

Genome sequencing and assembly. Genomic DNA was isolated from bacterial stocks grown overnight in LB using the GenElute genomic kit (Sigma-Aldrich, St. Louis, MO). The genome sequence of each isolate was generated at the Institute for Genome Sciences, Genome Resource Center, on an Illumina HiSeq2000 instrument using paired-end libraries with 300-bp inserts. The draft genomes were assembled using Celera Assembler (14). The final assemblies were filtered to contain contigs of ≥500 bp. The average coverage of the genomes sequenced in this study was >200×. Information regarding the size of the assembled genomes, number of contigs, and GenBank numbers for each of the genomes sequenced in this study is available in Table S1 in the supplemental material.

Phylogenomic analysis. The ETEC genomes sequenced in this study were compared with a diverse collection of *E. coli* and *Shigella* genomes (15). Briefly, single nucleotide polymorphisms (SNPs) were detected relative to the completed genome sequence of the laboratory isolate *E. coli* K-12 W3110 with a direct mapping of sequence based on nucmer alignments (16). SNPs present in all genomes analyzed were concatenated. A maximum-likelihood phylogeny with 100 bootstrap replicates was generated using RAxML v8.0.16 (17), using the ASC_GTRGAMMA substitution model, and visualized using FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

LS-BSR analysis. The level of similarity of protein-encoding genes was compared across all 208 genomes in this study using a large-scale BLAST score ratio (LS-BSR) analysis (18). Genes were predicted for each genome sequence using Prodigal (19) with default settings. Predicted genes from all genomes were then concatenated into a single file. The genes were clustered based on similarity with USEARCH (20), using a nucleotide identity threshold of 90%. Following the clustering, a file was generated that contained a centroid sequence for each cluster. The consensus sequences were translated and compared to each genome using tBLASTn as described above. The maximum tBLASTn bit score value obtained for each cluster was used as the denominator to generate a ratio for the cluster compared to each genome.

BSR analysis. The presence or absence of known virulence-associated genes in the genome sequences generated in this study was determined using BLAST score ratio (BSR) analysis, performed as previously described (21). The predicted amino acid sequences encoded by genes of interest were compared to the genomes analyzed in this study using tBLASTn (22). The ratio of tBLASTn scores was calculated for each genome by dividing the tBLASTn score obtained for each amino acid sequence of interest by the score obtained by tBLASTn of the amino acid reference sequence to its source genome. The protein-encoding genes that were considered present but divergent had BSR values of ≥0.4 and <0.8, while those with BSR values of ≥0.8 were determined to be present with significant similarity.

Genomic analysis. The primary goal of this study was to examine the diversity of ETEC bacterial populations within individuals with severe diarrhea. Multiple ETEC colonies isolated from individual patients with severe diarrhea were cultured and then subjected to whole-genome sequencing to identify conserved and divergent genomic features from each population. Reference isolates from cases of severe diarrheal illness of geographically disparate origin, historical prototype isolates, and ETEC isolates in GenBank were also included for comparison. The basic strain characteristics regarding the genomic content of the sequenced strains are included in Table S1 in the supplemental material. Overall, the isolates assembled well, with the average number of contigs being 212 (range 36 to 687), resulting in genomes of approximately 5.1 Mbp (range, 4.7 to 6.3 Mbp) with a GC profile typical of *E. coli* (50.6% ± 0.001%). These features suggest that all isolates sequenced and included in further analysis were *E. coli* and, from the selection process, that they were ETEC.

Phylogenomic comparison. To determine the relatedness of the isolates to one another, a whole-genome phylogeny was performed by identifying all of the SNPs when the genomes were compared to the completed genome sequence of the laboratory isolate *E. coli* K-12 W3110 (23) (Fig. 1). This comparative analysis as previously described by our group (15, 24, 25) confirmed the significant diversity among the isolates of ETEC (6, 26) (Fig. 1). The phylogenetic analysis in Fig. 1 contains all of the isolates sequenced in this project, ETEC reference genomes, and other *E. coli* and *Shigella* reference genomes. The core *E. coli* genome consists of ~2.5 million bases. The SNP phylogeny separated the majority of isolates into the phylogroup B1 and A subgroups, with only two ETEC isolates being outside these phylogenetic groups, and only one isolate from this study. This phylogenic distribution is common among ETEC isolates and was previously identified via other typing methods (6, 7, 27). A phylogenetic distribution of the genomes based on the isolation in the ICDDR,B Mohakhal hospital in Bangladesh or from the Mirpur field site is not observed in this data set (isolates whose designations start with an MP are from Mirpur, and those whose designations start with P or a number are from the ICDDR,B hospital) (Fig. 1). While bacteria were ob-

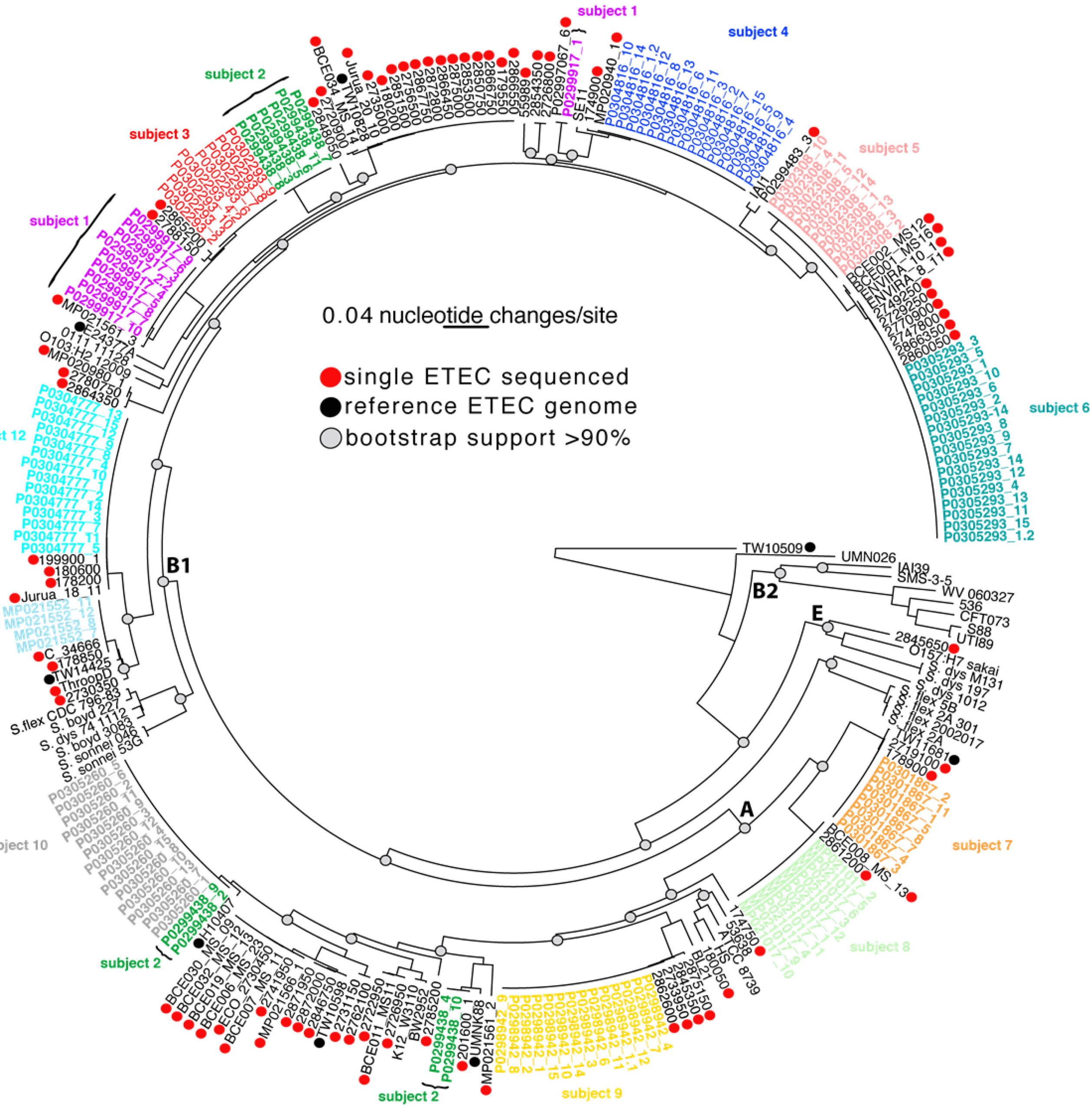

FIG 1 Global phylogeny identifying the distribution of multiple ETEC isolates from 12 individuals in Bangladesh. The global phylogeny includes 208 new ETEC isolates (labeled with red dots for single ETEC isolates sequenced, colored labels for multiple isolates per subject, and black dots for previously sequenced ETEC reference isolates; see Table S1 for subject distribution) and 32 reference *E. coli* isolates (in black with no additional label) representing each of the *E. coli* pathotypes and *Shigella* species. Colored isolate labels indicate the isolates that were obtained from the same individual during the single sampling period. Isolate designations beginning with "MP" are from the Mirpur treatment center in Dhaka, while those starting with "P" are from the ICDDR,B main hospital in Mohakhali, Dhaka. The isolates from two subjects are highlighted to indicate that phylogenetically distinct isolates were obtained from the same individual in two distinct cases, whereas 10 individuals appeared to contain isolates that were phylogenetically closely related.

tained from clinical sites located in two different districts within Dhaka, the ICDDR,B main hospital in Mohakhali, and the treatment center in Mirpur, phylogenetic distribution of the genomes did not segregate by treatment location.

In this work, we define a clonal population as isolates from any subject that are more related to each other than to any of the other isolates included in the analysis. Among the 12 patients from whom we sequenced multiple isolates, we identified a clonal ETEC population in 83.3% (10/12). Remarkably, some of the populations isolated in Dhaka, Bangladesh, in 2011 were nearly identical in core genome content to isolates obtained from geographically and temporally disparate cases of severe, cholera-like diarrheal illness. For instance, Juruá_18/11, isolated from a case of cholera-like diarrheal illness in Amazonia in 1998 (13), was nearly identical in core genome content to the isolates from subject P030477, a patient hospitalized at ICDDR,B with severe watery diarrhea in

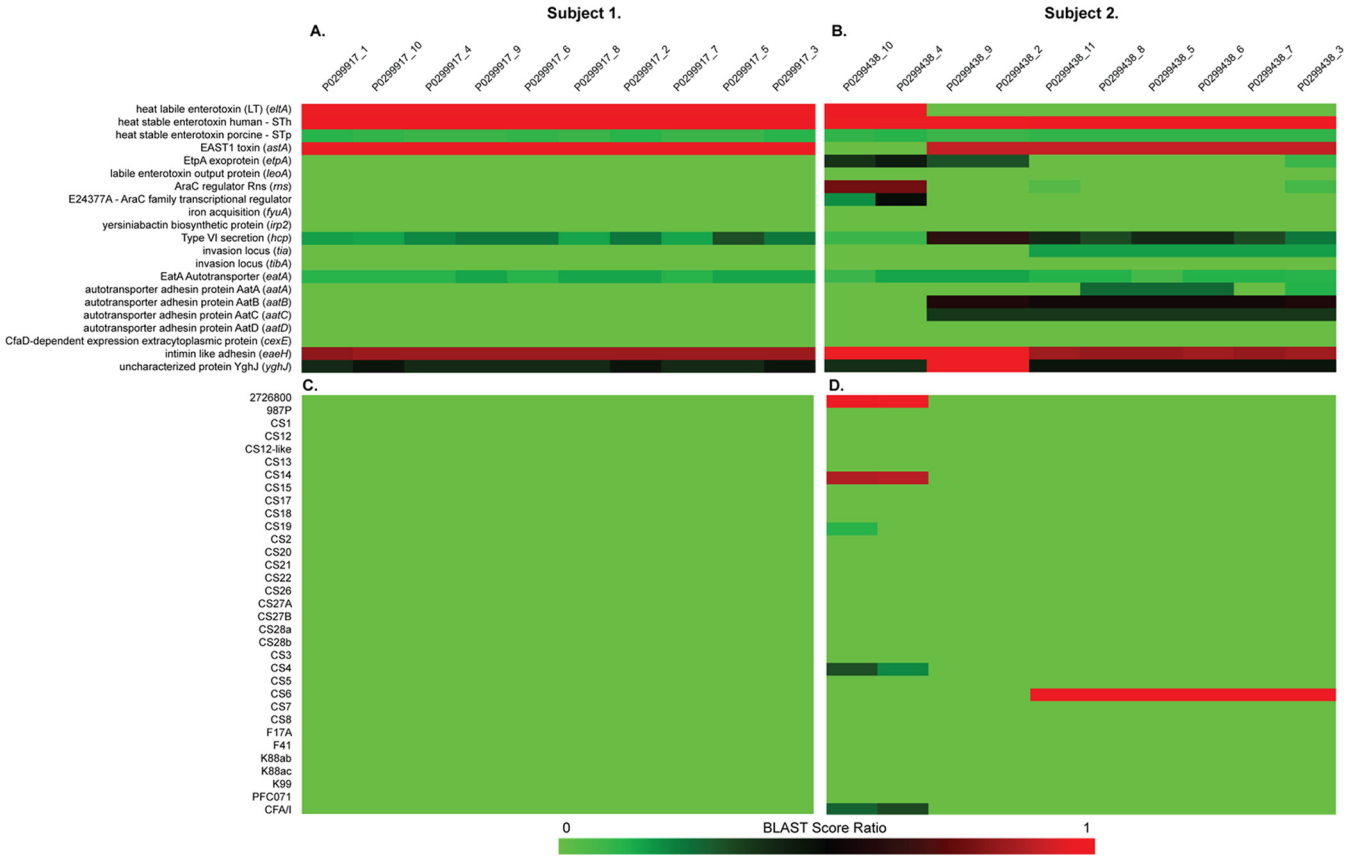

FIG 2 Comparison of known and putative virulence and colonization factors in phylogenetically divergent isolates obtained from the same individuals. The genomic contents of the isolates from subjects P0299917 and P0299438 represent the individuals that have differing virulence and colonization factor profiles among the isolates from that subject. (A) Subject 1 (P0299917), virulence factors; (B) subject 2 (P0299438), virulence factors; (C) subject 1 (P0299917), colonization factors; (D) subject 2 (P0299438), colonization factors). In the case of P0299917, similar profiles of virulence factor genes and colonization factors are observed, even though isolate P0299917_1 is genomically distinct (Fig. 1). In contrast P0299438 contains 3 different profiles of gene presence and absence that are congruent with the phylogenomic differences observed in Fig. 1. Isolates P0299438_4 and P0299438_10 form one group, P0299438_2 and P0299438_9 form a second group, and the remaining isolates from this subject with a similar profile form a third group.

2011, suggesting that some clones of ETEC could be particularly well equipped to cause severe disease in diverse human populations. In distinct contrast, isolates from the remaining two subjects contained a mixture of phylogenetically distinct isolates. In one subject, P0299917 (subject 1 in Fig. 1), all of the isolates (highlighted in purple in Fig. 1) belong to phylogenetic group B1, with a cluster of nine closely linked strains, and a single unique isolate. The number of genetic changes required for the observed differences in these isolates is relatively small and could potentially be explained by genetic diversification *in vivo* to result in two phylogroup B1 isolates from a single patient. The isolates from subject 2, P0299438, represent a different scenario, where six of the isolates are in phylogroup B1 and the remaining four isolates are in phylogroup A (Fig. 1). Additionally, within phylogroup A, these four isolates separate into two distinct phylogenetic groups. Such phylogenomic diversity, characterized by the identification of genomically distinct pathogenic *E. coli* isolates from an individual patient, is unprecedented.

Virulence factor profiles. ETEC virulence factors are typically encoded on plasmids and other mobile elements (2, 5, 8, 10). Therefore, we set out to examine determine whether these genomically distinct isolates exhibited unique virulence factor profiles. The BLAST score ratio (21) is plotted for each gene/feature listed on the left in each of the genomes from these two subjects in Fig. 2. The data are normalized between 0 and 1. Interestingly, virulence factor content did not strictly segregate by phylogenomic lineage. In the case of the isolates from P0299917, also identified as subject 1, we identified similar profiles of virulence factor genes and colonization factors (Fig. 2A and C), even though isolate P0299917_1 is genomically distinct (Fig. 1). Examination of other accessory genomic features of this isolate compared to the majority profile from this patient indicates that the gene content of this genomically distinct isolate, P0299917_1, is significantly different from the remainder of the isolates (data not shown). This indicates that there are two distinct populations within the individual at the time of culture and that these changes are not the result of *in vivo* alterations from a common ancestor. In contrast, the isolates from subject P0299438 display significant diversity in their virulence factor profiles, paralleling the observed genomic variation. P0299438 isolates segregated into three distinct phylogenomic clusters. Isolates P0299438_4 and P0299438_10 share an altered virulence and colonization factor profile compared to the majority of isolates, and potentially the most significant difference is the identification of the LT toxin

genes and a number of known regulators in these two isolates but in no other isolates from this subject (Fig. 2B). Isolates P0299438_2 and P0299438_9 have a similar virulence factor profile but an altered colonization factor profile compared to the majority isolates from this subject (Fig. 2B and D). However, these two isolates appear to lack the colonization factor antigens (Fig. 2D) (CS14, CS4, CS6 and a homolog to a novel CF cluster in isolate 2726800) shared by the majority of the isolates from this individual. These observed virulence factor differences are congruent with the phylogenomic differences observed in Fig. 1. Directionality in the evolution of these strains is difficult to ascertain, as we cannot tell from these studies if the virulence genes and colonization factors have been lost upon culture or were not present to start with. These types of studies can be examined only via long-term longitudinal studies on the ETEC populations. Nevertheless, the isolation of disparate strains from this particular patient highlights the potential for considerable diversity among ETEC isolates within a single individual during the course of clinical illness. Similar to the simultaneous isolation of multiple enteropathogens from patients with diarrheal illness (1), the diversity of isolates in this individual patient also significantly confounds determination of the strain(s) responsible for disease.

Conclusion. In these studies of bacterial diversity among patients with severe ETEC infections, we demonstrate that in the majority (83%) of individuals, ETEC isolates emerge in diarrheal stool essentially as clonal populations. These findings suggest that in general, severe cholera-like diarrhea from ETEC is largely the result of infection by a number of highly virulent clones. Remarkably, some of these diarrheal clones are genetically very similar to strains isolated from temporally and geographically dispersed cases of severe diarrhea, suggesting that despite the overall genetic diversity of these pathogens, some traits associated with cholera-like illness may be maintained over time. Additionally, this suggests that the genomic content and the virulence factors likely work in concert and that the acquisition of appropriate features, both virulence-related and non-virulence-related factors, results in a pathogen that is optimally equipped to infect susceptible hosts. Elucidation of the essential genetic features that dictate more severe forms of disease could be important for rational design of vaccines specifically targeted to prevent deaths from ETEC diarrhea.

The finding of distinct subpopulations of ETEC with divergent genomes and virulence factor content within an individual also has implications for vaccine design and testing. A vaccine that is too narrowly focused could select a population of bacteria possessing antigens that escape immunologic neutralization. While the impact of genomic diversity on ETEC vaccine performance has not been thoroughly assessed in field studies conducted to date, the present studies demonstrate the feasibility of incorporating high-throughput genome sequencing in assessment of vaccine outcomes. Collectively, the genomic approaches described here could serve as a template in future trials and could likewise permit targeting of conserved genomic elements relevant to rational design of vaccines to prevent deaths due to diarrhea.

Nucleotide sequence accession numbers. The genome sequences generated in this study are deposited in GenBank under the accession numbers listed in Table S1.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00501-15/-/DCSupplemental.

Table S1, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Richard Finkelstein and Ana Vicente for providing ETEC strains isolated from cases of severe diarrheal illness that were sequenced as part of this analysis.

This project was supported in part by the ICDDR,B, by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract no. HHSN272200900009C, NIH grant no. RO1AI089894 and U19AI110820, and by startup funds from the State of Maryland. Also, ICDDR,B is thankful to the governments of Australia, Bangladesh, Canada, Sweden, and the United Kingdom for providing core/unrestricted support.

REFERENCES

1. **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. http://dx.doi.org/10.1016/S0140-6736(13)60844-2.
2. **Fleckenstein JM, Hardwidge PR, Munson GP, Rasko DA, Sommerfelt H, Steinsland H.** 2010. Molecular mechanisms of enterotoxigenic Escherichia coli infection. Microbes Infect **12:**89–98. http://dx.doi.org/10.1016/j.micinf.2009.10.002.
3. **Qadri F, Saha A, Ahmed T, Al Tarique A, Begum YA, Svennerholm AM.** 2007. Disease burden due to enterotoxigenic Escherichia coli in the first 2 years of life in an urban community in Bangladesh. Infect Immun **75:**3961–3968. http://dx.doi.org/10.1128/IAI.00459-07.
4. **Harro C, Chakraborty S, Feller A, DeNearing B, Cage A, Ram M, Lundgren A, Svennerholm AM, Bourgeois AL, Walker RI, Sack DA.** 2011. Refinement of a human challenge model for evaluation of enterotoxigenic Escherichia coli vaccines. Clin Vaccine Immunol **18:**1719–1727. http://dx.doi.org/10.1128/CVI.05194-11.
5. **Svennerholm AM, Lundgren A.** 2012. Recent progress toward an enterotoxigenic Escherichia coli vaccine. Expert Rev Vaccines **11:**495–507. http://dx.doi.org/10.1586/erv.12.12.
6. **Steinsland H, Lacher DW, Sommerfelt H, Whittam TS.** 2010. Ancestral lineages of human enterotoxigenic Escherichia coli. J Clin Microbiol **48:**2916–2924. http://dx.doi.org/10.1128/JCM.02432-09.
7. **Steinsland H, Valentiner-Branth P, Aaby P, Mølbak K, Sommerfelt H.** 2004. Clonal relatedness of enterotoxigenic Escherichia coli strains isolated from a cohort of young children in Guinea-Bissau. J Clin Microbiol **42:**3100–3107. http://dx.doi.org/10.1128/JCM.42.7.3100-3107.2004.
8. **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. http://dx.doi.org/10.1128/JB.00619-08.
9. **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. http://dx.doi.org/10.1371/journal.pgen.1000344.
10. **Crossman LC, Chaudhuri RR, Beatson SA, Wells TJ, Desvaux M,**

Cunningham AF, Petty NK, Mahon V, Brinkley C, Hobman JL, Savarino SJ, Turner SM, Pallen MJ, Penn CW, Parkhill J, Turner AK, Johnson TJ, Thomson NR, Smith SG, Henderson IR. 2010. A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic Escherichia coli strain H10407. J Bacteriol **192:**5822–5831. http://dx.doi.org/10.1128/JB.00710-10.

11. **Sjöling A, Wiklund G, Savarino SJ, Cohen DI, Svennerholm AM**. 2007. Comparative analyses of phenotypic and genotypic methods for detection of enterotoxigenic Escherichia coli toxins and colonization factors. J Clin Microbiol **45:**3295–3301. http://dx.doi.org/10.1128/JCM.00471-07.
12. **Finkelstein RA, Vasil ML, Jones JR, Anderson RA, Barnard T**. 1976. Clinical cholera caused by enterotoxigenic Escherichia coli. J Clin Microbiol **3:**382–384.
13. **Vicente AC, Teixeira LF, Iniguez-Rojas L, Luna MG, Silva L, Andrade JR, Guth BE**. 2005. Outbreaks of cholera-like diarrhoea caused by enterotoxigenic Escherichia coli in the Brazilian Amazon rainforest. Trans R Soc Trop Med Hyg **99:**669–674. http://dx.doi.org/10.1016/j.trstmh.2005.03.007.
14. **Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC**. 2000. A whole-genome assembly of Drosophila. Science **287:**2196–2204. http://dx.doi.org/10.1126/science.287.5461.2196.
15. **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 U S A **110:**12810–12815. http://dx.doi.org/10.1073/pnas.1306836110.
16. **Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL**. 2004. Versatile and open software for comparing large genomes. Genome Biol **5:**R12. http://dx.doi.org/10.1186/gb-2004-5-2-r12.
17. **Stamatakis A**. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. BioInformatics **30:**1312–1313. http://dx.doi.org/10.1093/bioinformatics/btu033.
18. **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. http://dx.doi.org/10.7717/peerj.332.
19. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ**. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119. http://dx.doi.org/10.1186/1471-2105-11-119.
20. **Edgar RC**. 2010. Search and clustering orders of magnitude faster than BLAST. BioInformatics **26:**2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461.
21. **Rasko DA, Myers GS, Ravel J**. 2005. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics **6:**2. http://dx.doi.org/10.1186/1471-2105-6-2.
22. **Gertz EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF**. 2006. Composition-based statistics and translated nucleotide searches: improving the tBLASTn module of BLAST. BMC Biol **4:**41. http://dx.doi.org/10.1186/1741-7007-4-41.
23. **Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner BL, Mori H, Horiuchi T**. 2006. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. Mol Syst Biol **2:**2006.0007. http://dx.doi.org/10.1038/msb4100049.
24. **Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Moller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK**. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med **365:**709–717. http://dx.doi.org/10.1056/NEJMoa1106920.
25. **Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, Rasko DA**. 2011. A comparative genomic analysis of diverse clonal types of enterotoxigenic Escherichia coli reveals pathovar-specific conservation. Infect Immun **79:**950–960. http://dx.doi.org/10.1128/IAI.00932-10.
26. **Steinsland H, Valentiner-Branth P, Perch M, Dias F, Fischer TK, Aaby P, Mølbak K, Sommerfelt H**. 2002. Enterotoxigenic Escherichia coli infections and diarrhea in a cohort of young children in Guinea-Bissau. J Infect Dis **186:**1740–1747. http://dx.doi.org/10.1086/345817.
27. **Steinsland H, Valentiner-Branth P, Grewal HM, Gaastra W, Molbak K, Sommerfelt H**. 2003. Development and evaluation of genotypic assays for the detection and characterization of enterotoxigenic Escherichia coli. Diagn Microbiol Infect Dis **45:**97–105. http://dx.doi.org/10.1016/S0732-8893(02)00504-7.