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 phylogenetically 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 genically 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.

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 (Jurú 18/11, Jurú 20/10, Envira_10/1, and Envira_8/11) obtained during ETEC outbreaks that caused severe diarrheal illness in two small villages, Jurú 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 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 phylogenetic 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 Mohakhhal 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-
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 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.
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). 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.
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. HHSN27220090009C, NIH grant no. RO1AI089894 and U19AI108820, 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
siah et al. gone bad: complete genome sequence of the prototypical enterotoxi-
dx.doi.org/10.1128/JB.00710-10.
ive analyses of phenotypic and genotypic methods for detection of
dx.doi.org/10.1126/science.287.5461.2196.
15. Haen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, Rasko DA. 2013. Refining the pathovar paradigm via phylogenomics of the at-
aching and effacing Escherichia coli. Proc Natl Acad Sci U S A 110: 
17. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-
dx.doi.org/10.1093/bioinformatics/btu033.
dx.doi.org/10.1186/1471-2105-6-2.
8893(02)00504-7.