JB Journals.ASM.org

GENOME ANNOUNCEMENT

Draft Genome Sequences of the Diarrheagenic *Escherichia coli* Collection

Tracy H. Hazen,[a,b] Jason W. Sahl,[a,b]* Julia C. Redman,[a,b] Carolyn R. Morris,[a,b] Sean C. Daugherty,[b] Marcus C. Chibucos,[a,b] Naomi A. Sengamalay,[b] Claire M. Fraser-Liggett,[b,c] Hans Steinsland,[e] Thomas S. Whittam,[d] Beth Whittam,[d] Shannon D. Manning,[d] and David A. Rasko[a,b]

Department of Microbiology and Immunology[a] and Institute for Genome Sciences,[b] University of Maryland School of Medicine, Baltimore, Maryland, USA; University of Maryland School of Medicine, Department of Medicine, Baltimore, Maryland, USA[c]; Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, USA[d]; and Centre for International Health and Department of Biomedicine, University of Bergen, Bergen, Norway[e]

We report the draft genome sequences of the collection referred to as the *Escherichia coli* DECA collection, which was assembled to contain representative isolates of the 15 most common diarrheagenic clones in humans (http://shigatox.net/new/). These genomes represent a valuable resource to the community of researchers who examine these enteric pathogens.

The most comprehensive diarrheal studies indicate that there are greater than 110 million cases of diarrhea in children under 5 each year (3) and approximately 2 million people die each year as a direct result of diarrheal disease; a large proportion of these are children. While many assume this a problem only for the developing world, the NIDDK indicates that the rate of diarrhea among the U.S. population is 100% per year; i.e., each person in the Unted States contracts diarrhea at least once each year (http://www.niddk.nih.gov/). The primary bacterial pathogens that contribute to diarrheal disease are *Escherichia coli* and *Shigella* species. Recent food-borne outbreaks attributable to both *E. coli* and *Shigella* illustrate that these diarrheal pathogens also constitute a significant public health problem in the developed world (1, 8, 11). While genome sequencing is entering a phase where rapid sequencing will become part of the normal clinical diagnostic paradigm, it has been demonstrated that adequate and reliable reference genomes are required for useful comparative studies (6, 9, 14).

The collection of isolates in this announcement represent the dominant clonal types of diarrheagenic *E. coli* and have been used in innumerable studies to highlight the diversity among *E. coli* isolates. Each of these isolates has been examined using multilocus sequencing typing schema (10), which was confirmed with each of the draft genome sequences generated in this project. The generation of these genomes allows the direct comparison of housekeeping gene typing schema with the large-scale genome phylogeny methods that are evolving (2, 4, 8, 12).

Genomic DNA was isolated from an overnight culture using the Sigma GenElute kit (Sigma-Aldrich) and was sequenced at the University of Maryland School of Medicine, Institute for Genome Sciences, Genome Resource Center (http://www.igs.umaryland.edu/). The genome sequence was generated using 3-kb insert paired-end libraries on the 454 Titanium FLX (Roche), and the draft genomes were assembled using the Celera assembler (5). The resulting genomes contained an average of 68 contigs per isolate (range, 31 to 148). The contig data were annotated using the Annotation pipeline at the Institute for Genome Sciences, Informatics Resource Center (http://www.igs.umaryland.edu/). The GC content of these genomes was ~50%, similar to those found in other large-scale *E. coli* genome sequencing projects (7, 13). The numbers of predicted genes from the draft genomes were also similar to those found in previously sequenced *E. coli* genomes, with 5,578 genes/genome (range, 4,758 to 6,505). Many of the genome projects contained contigs that mapped to plasmids known to be common in enteric pathogens.

The genomes of these isolates will be further examined in large-scale comparative genomic analyses that are under way, but they represent reference genomes for the at-large community to use.

Nucleotide sequence accession numbers. The genome data have been deposited in GenBank with accession numbers AIEV00000000 to AIEZ00000000, AIFA00000000 to AIFZ00000000, AIGA00000000 to AIGZ00000000, and AIHA00000000 to AIHS00000000. Please see http://gscid.igs.umaryland.edu/wp.php?wp=emerging_diarrheal_pathogens to match the GenBank accession number to the isolate of interest.

ACKNOWLEDGMENT

This 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.

REFERENCES

1. **Frank C, et al.** 2011. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. N. Engl. J. Med. **365**:1771–1780.
2. **He M, et al.** 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc. Natl. Acad. Sci. U. S. A. **107**:7527–7532.
3. **Kotloff KL, et al.** 1999. Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. Bull. World Health Organ. **77**:651–666.
4. **Mutreja A, et al.** 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature **477**:462–465.
5. **Myers EW, et al.** 2000. A whole-genome assembly of Drosophila. Science **287**:2196–2204.
6. **Price LB, et al.** 2012. *Staphylococcus aureus* CC398: host adaptation and emergence of methicillin resistance in livestock. mBio **3**(1):e00305-11.

Received 17 March 2012 **Accepted** 27 March 2012

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

* Present address: Translational Genomics Research Institute, Flagstaff, Arizona, USA.

doi:10.1128/JB.00426-12

7. **Rasko DA, et al.** 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. **190**:6881–6893.
8. **Rasko DA, et al.** 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N. Engl. J. Med. **365**:709–717.
9. **Rasko DA, et al.** 2011. *Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. Proc. Natl. Acad. Sci. U. S. A. **108**:5027–5032.
10. **Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS.** 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature **406**: 64–67.
11. **Rohde H, et al.** 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N. Engl. J. Med. **365**:718–724.
12. **Sahl JW, et al.** 2011. A comparative genomic analysis of diverse clonal types of enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation. Infect. Immun. **79**:950–960.
13. **Touchon M, et al.** 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. **5**:e1000344.
14. **Van Ert MN, et al.** 2007. Strain-specific single-nucleotide polymorphism assays for the *Bacillus anthracis* Ames strain. J. Clin. Microbiol. **45**:47–53.