AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Draft Genome Sequences of Isolates of Diverse Host Origin from the *E. coli* Reference Center at Penn State University

David W. Lacher,[a] Mark K. Mammel,[a] Jayanthi Gangiredla,[a] Solomon T. Gebru,[a] Tammy J. Barnaba,[a] Sydney A. Majowicz,[b] Edward G. Dudley[b,c]

[a]Division of Molecular Biology, Office of Applied Research and Safety Assessment, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, Laurel, Maryland, USA

[b]Department of Food Science, Penn State University, University Park, Pennsylvania, USA

[c]*E. coli* Reference Center, Penn State University, University Park, Pennsylvania, USA

ABSTRACT *Escherichia coli* strains present a vast genomic diversity. We report the draft genome sequences of 1,000 isolates from the *E. coli* Reference Center at Penn State University. These strains were originally isolated from multiple animal and environmental sources over the past 50 years.

Members of the genus *Escherichia*, specifically *Escherichia coli*, include pathogenic and nonpathogenic strains. The ability to differentiate these two groups of *E. coli* has an impact on food safety. As part of the U.S. Food and Drug Administration's efforts to expand state-of-the-art technology to identify pathogenic *E. coli* strains, we are developing an in-depth phylogenetic landscape of *E. coli* that parses these bacteria into different clades. In order to expand this landscape as well as provide further depth, whole-genome sequences are essential. Here, we report the draft genome sequences of 1,000 isolates from the culture collection housed at Penn State University's *E. coli* Reference Center. The diverse collection examined in this study contains isolates from animal, environmental, and food sources. *E. coli* is commonly found as a member of the gut microbiota of warm-blooded organisms and has been isolated from a wide range of animal hosts (1, 2). Phylogenetic analyses have shown that *E. coli* can be divided into several phylogroups (3, 4), with pathogenic and nonpathogenic strains seemingly randomly distributed among them. This project focuses on the whole-genome sequencing of *E. coli* isolates from nonhuman animal sources, as well as the environment, that may reveal lineages from nonpathogenic to pathogenic strains. Understanding this evolutionary path may provide molecular insight into the acquisition of virulence attributes from an environmental source.

Pure cultures for each strain were grown aerobically overnight in Luria-Bertani broth at 37°C. Total genomic DNA was extracted from 1 ml of overnight culture using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany). DNA extractions were performed with the Qiagen QIAcube instrument using the manufacturer's Gram-negative bacterium protocol. Sequencing libraries were prepared with 1 ng DNA using the Nextera XT DNA sample prep kit (Illumina, San Diego, CA, USA) and sequenced on either the Illumina MiSeq or NextSeq platform. The resulting paired-end reads (2×250 bp for MiSeq, 2×150 bp for NextSeq) were quality assessed by FastQC v0.11.8 (5). Low-quality reads were trimmed to a quality threshold of $Q > 30$, and adapter sequences were removed using the NexteraPE adapter file in Trimmomatic v0.38 (6). The genomes were *de novo* assembled with SPAdes v3.13.0 (7) using a k-mer size of 55, and assembly quality assessment was performed with QUAST v5.0 (8). The genomes were automatically annotated using the NCBI Prokaryotic Genome Annotation Pipeline (9). Default parameters were used for all software unless otherwise specified.

The depth of coverage for the draft genomes ranged from 17× to 161×, with the

Citation Lacher DW, Mammel MK, Gangiredla J, Gebru ST, Barnaba TJ, Majowicz SA, Dudley EG. 2020. Draft genome sequences of isolates of diverse host origin from the *E. coli* Reference Center at Penn State University. Microbiol Resour Announc 9:e01005-20. https://doi.org/10.1128/MRA.01005-20.

Editor Julie C. Dunning Hotopp, University of Maryland School of Medicine

Address correspondence to David W. Lacher, david.lacher@fda.hhs.gov.

Received 27 August 2020
Accepted 2 September 2020
Published 24 September 2020

TABLE 1 Summary of 1,000 genomes from the *E. coli* Reference Center

Category	No. of genomes	No. of source species or types	Phylogroup(s) observed	Cryptic lineage(s) observed
Avian	270	18	A, B1, B2, D, E, F	1, 3, 4, 5
Environmental	62	3	A, B1, B2, D, E, F	3, 4, 5
Food	37	6	A, B1, D	None
Mammal	629	41	A, B1, B2, D, E, F	1
Reptile	1	1	A	None
Unknown	1	1	B1	None

genomes ranging in size from 4,291,381 to 5,764,740 bp. The number of contigs ranged from 50 to 741, while the N_{50} values ranged from 16,761 to 315,275 bp. The genomes were placed into one of six categories according to their source, avian, environmental, food, mammal, reptile, or unknown (Table 1). Most ($n = 629$) of the strains are of mammalian origin, with bovine, porcine, and canine sources being the most common ($n = 203$, 168, and 92, respectively). Among the 270 isolates of avian origin, chicken and turkey were the most common sources ($n = 68$ and 60, respectively). Phylogroups were assigned based on the single nucleotide polymorphisms (SNPs) present within 45 genes found in *E. coli* K-12 MG1655 (GenBank accession number U00096.3). Briefly, the 45 genes were extracted from each assembly and aligned to the sequence from K-12 MG1655 using BLAST. A SNP profile of 45 concatenated sites was then used to assign the phylogroup. Each of the established *E. coli* phylogroups is represented among the 1,000 genomes, namely, phylogroups A ($n = 180$), B1 ($n = 438$), B2 ($n = 220$), D ($n = 69$), E ($n = 38$), and F ($n = 23$). Twenty isolates belong to one of the following four known "cryptic" lineages of *Escherichia* (10, 11): lineage 1 ($n = 3$), lineage 3 ($n = 4$), lineage 4 ($n = 2$), and lineage 5 ($n = 11$). The remaining 12 isolates were classified as undetermined, because their phylogroup could not be assigned using the panel of 45 SNP loci.

Data availability. The draft genome assemblies were deposited at DDBJ/ENA/GenBank through the FDA's GenomeTrakr pipeline under BioProject accession number PRJNA357722. The versions described in this announcement are the second versions. A full listing of the source and phylogroup information for the 1,000 genomes can be found at https://doi.org/10.6084/m9.figshare.12885527.v2 (12). A list of the 45 genes and diagnostic SNPs used for phylogroup assignment can be found at https://doi.org/10.6084/m9.figshare.12899765.v1 (13).

ACKNOWLEDGMENTS

The views expressed in this article are those of the authors and do not necessarily reflect the official policy of the Department of Health and Human Services, the U.S. Food and Drug Administration (FDA), or the U.S. Government. Reference to any commercial materials, equipment, or process does not in any way constitute approval, endorsement, or recommendation by the FDA.

REFERENCES

1. Belanger L, Garenaux A, Harel J, Boulianne M, Nadeau E, Dozois CM. 2011. *Escherichia coli* from animal reservoirs as a potential source of human extraintestinal pathogenic *E. coli*. FEMS Immunol Med Microbiol 62:1–10. https://doi.org/10.1111/j.1574-695X.2011.00797.x.
2. Kim J-S, Lee M-S, Kim JH. 2020. Recent updates on outbreaks of Shiga toxin-producing *Escherichia coli* and its potential reservoirs. Front Cell Infect Microbiol 10:273. https://doi.org/10.3389/fcimb.2020.00273.
3. Herzer PJ, Inouye S, Inouye M, Whittam TS. 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J Bacteriol 172:6175–6181. https://doi.org/10.1128/jb.172.11.6175-6181.1990.
4. Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep 5:58–65. https://doi.org/10.1111/1758-2229.12019.
5. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
6. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
7. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.

8. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
9. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI Prokaryotic Genome Annotation Pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.
10. Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, Whittam TS. 2009. Cryptic lineages of the genus *Escherichia*. Appl Environ Microbiol 75:6534–6544. https://doi.org/10.1128/AEM.01262-09.
11. Walk ST. 2015. The "Cryptic" *Escherichia*. EcoSal Plus 6. https://doi.org/10.1128/ecosalplus.ESP-0002-2015.
12. Lacher DW, Mammel MK, Gangiredla J, Gebru ST, Barnaba TJ, Majowicz SA, Dudley EG. 2020. Supplemental Table 1: assembly and strain information for 1000 genomes from the *E. coli* Reference Center at Penn State University. https://doi.org/10.6084/m9.figshare.12885527.v2.
13. Lacher DW, Mammel MK, Gangiredla J, Gebru ST, Barnaba TJ, Majowicz SA, Dudley EG. 2020. Supplemental Table 2: loci and diagnostic SNPs used for phylogroup determination. https://doi.org/10.6084/m9.figshare.12899765.v1.