AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

PROKARYOTES

High-Quality Whole-Genome Sequences for 77 Shiga Toxin-Producing *Escherichia coli* Strains Generated with PacBio Sequencing

Pooja N. Patel,[a,b] Rebecca L. Lindsey,[a] Lisley Garcia-Toledo,[a,b] Lori A. Rowe,[c] Dhwani Batra,[c] Samuel W. Whitley,[a,b] Daniel Drapeau,[a,b] Devon Stoneburg,[a] Haley Martin,[a] Phalasy Jueing,[c] Vladimir N. Loparev,[c] Nancy Strockbine[a]

[a]Enteric Diseases Laboratory Branch, Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

[b]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA

[c]Biotechnology Core Facility Branch, Division of Scientific Resources, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

ABSTRACT Shiga toxin-producing *Escherichia coli* (STEC) is an enteric foodborne pathogen that can cause mild to severe illness. Here, we report the availability of high-quality whole-genome sequences for 77 STEC strains generated using the PacBio sequencing platform.

Shiga toxin-producing *Escherichia coli* (STEC) is a major foodborne pathogen responsible for outbreaks and sporadic cases of diarrheal illness (1). Although the majority of reported STEC infections in the United States are caused by *E. coli* O157:H7, non-O157 serotypes have grown to be a public health concern both in the United States and internationally, as they can cause severe illness comparable to that caused by STEC O157 (2, 3). Non-O157 STEC has been linked to a range of clinical illnesses, from asymptomatic shedding and mild diarrhea to hemorrhagic colitis and potentially fatal hemolytic-uremic syndrome (HUS); more than 100 STEC serotypes have been linked to such human disease (4). Many of these non-O157 serotypes do not have publicly available PacBio-sequenced genomes.

Here, we report whole-genome sequences for 77 STEC strains representing 43 serotypes. The STEC cultures were grown overnight on blood agar plates at 37°C, and genomic DNA was extracted according to the manufacturer's protocol (ArchivePure; 5 Prime, Gaithersburg, MD). The DNA was sheared to 20 kb using needle shearing, and the prepared libraries were further size selected using BluePippin (Sage Scientific, Beverly, MA). The large SMRTbell libraries were generated using standard library protocols of the Pacific Biosciences DNA template preparation kit (Pacific Biosciences, Menlo Park, CA). Each strain was sequenced using one, two, or three single-molecule real-time (SMRT) cells. The finished libraries were bound to proprietary P6 version 2 polymerase and sequenced on a PacBio RS II platform using C4 chemistry for 360-min movies. The sequence reads were then filtered and assembled *de novo* using Falcon, Canu, or the PacBio Hierarchical Genome Assembly Process version 3 (5–7). For 30 strains, whole-genome optical maps were generated using the Argus platform (OpGen, Gaithersburg, MD), and the sequence order was verified using corresponding AflII and NcoI whole-genome maps.

The detected serotypes, accession numbers, and assembly metrics for each genome are listed in Table 1. The average G+C content for all 77 chromosomal sequences was 50.6%. The average coverage ranged from 39.5× to 230.8×, with an average coverage of 109×. All but nine chromosomal sequences were circularized and found to have overlapping ends. Of the nine genomes that could not be circularized due to collapsed

Received 30 March 2018 **Accepted** 30 March 2018 **Published** 10 May 2018

Citation Patel PN, Lindsey RL, Garcia-Toledo L, Rowe LA, Batra D, Whitley SW, Drapeau D, Stoneburg D, Martin H, Jueing P, Loparev VN, Strockbine N. 2018. High-quality whole-genome sequences for 77 Shiga toxin-producing *Escherichia coli* strains generated with PacBio sequencing. Genome Announc 6:e00391-18. https://doi.org/10.1128/genomeA.00391-18.

Address correspondence to Rebecca L. Lindsey, rebecca.lindsey@cdc.hhs.gov.

TABLE 1 Accession numbers and assembly metrics of 77 STEC whole-genome sequences

Escherichia coli strain ID[a]	Serotype	Chromosomal GenBank accession no.	No. of contigs	Chromosome size (bp)	Associated plasmid size(s) (bp) (GenBank accession no.)
2015C-3163	O103:H2	CP027219	2	5,500,189	94,104 (CP027220)
2015C-3101	O111:H8	CP027221	3	5,313,278	48,390 (CP027222), 72,543[c] (CP027223)
2015C-3108	O111:H8	CP027307	3	5,364,442	63,664[c] (CP027308), 93,724 (CP027309)
2014C-4135	O113:H21	CP027310	2	4,949,048	133,438 (CP027311)
2013C-3181	O113:H21	CP027312	1	5,167,951	No plasmids
2014C-3550	O118:H16	CP027313	4	5,549,395	59,928 (CP027314), 88,840 (CP027315), 179,514 (CP027316)
2015C-3107	O121:H19	CP027317	2	5,388,260	81,954 (CP027318)
2014C-3084	O145:H28	CP027319	4	4,717,123	78,854[c] (CP027320), 84,276[c] (CP027321), 706,680 (CP027322)
2013C-3033	O146:H21	CP027323	2	5,426,201	127,667 (CP027324)
2013C-4830	O165:H25	CP027325	3	5,135,675	74,671 (CP027326), 93,170 (CP027327)
2014C-3741	O174:H8	CP027328	3	5,394,679[c]	128,345[c] (CP027329), 102,897[c] (CP027330)
2013C-3277	O26:H11	CP027331	4	5,438,694	20,839[c] (CP027332), 46,866[c] (CP027333), 181,066[c] (CP027334)
2014C-3716	O26:H11	CP027335	3	5,568,215[c]	144,060[c] (CP027336), 62,870[c] (CP027337)
2013C-3925	O5:H9	PVMF00000000	6	331,062,[c] 4,413,627,[c] 496,511[c]	77,933,[c] 170,283,[c] 99,102[c]
2014C-3051	O71:H11	CP027338	2	5,597,475	92,644 (CP027339)
2015C-3121	O91:H14	CP027340	2	5,366,577	104,198[c] (CP027341)
2014C-4587	OUND:H19	CP027342	2	5,040,163	131,410 (CP027343)
2014C-3946	O111:H8	CP027344	3	5,264,938	22,197[c] (CP027345), 18,123[c] (CP027346)
2013C-4361	O111:H8	CP027347	2	5,317,846	70,613[c] (CP027348)
2014C-3655	O121:H19	CP027351	2	5,442,537	97,117[c] (CP027350)
2012C-4606	O26:H11	CP027352	3	5,647,195	20,881[c] (CP027353), 57,720[c] (CP027354)
2013C-4390	O76:H19	CP027484	2	5,353,719	147,394 (CP027485)
2013C-4991	O80:H2	CP027355	4	5,367,251	71,714[c] (CP027356), 131,463[c] (CP027357), 110,001 (CP027358)
2014C-4639	O26:H11	CP027361	3	5,325,246	54,873[c] (CP027359), 329,873 (CP027360)
95-3192	O145:H28	CP027362	1	5,385,516	No plasmids
88-3001	O165:H25	CP027363	2	5,195,753	74,659 (CP027364)
89-3156	O174:H21	CP027366	2	5,065,883	125,561 (CP027367)
03-3375	O145:H25	PVMG00000000	4	5,199,239,[c] 40,965[c]	30,901,[c] 83,963[c]
2014C-3307	O178:H19	CP027368	3	4,965,987	109,641 (CP027369), 176,149[c] (CP027370)
2015C-3905	O181:H49	CP027371	2	4,901,620	175,427[c] (CP027372)
2014C-4638	O26:H11	PVMH00000000	4	261,681,[c] 2,112,842,[c] 3,317,231[c]	88,223[c]
05-3629	O8:H16	CP027373	3	4,904,151	91,648 (CP027374), 118,863[c] (CP027375)
2013C-4404	O91:H14	CP027376	4	5,009,822	70,152 (CP027377), 113,102[c] (CP027378), 104,889 (CP027379)
2013C-3250	O111:H8	CP027380	6	5,401,672	24,547[c] (CP027381), 36,491[c] (CP027382), 73,784[c] (CP027383), 27,224[c] (CP027384), 118,259 (CP027385)
2014C-3057	O26:H11	CP027387	2	5,645,983	54,452[c] (CP027386)
2011C-4251	O45:H2	CP027388	2	5,440,026	68,062[c] (CP027389)
2015C-4944	O26:H11	CP027390	2	5,802,748	98,724 (CP027391)
97-3250	O26:H11	CP027599	3	5,942,969	120,604 (CP027600), 92,590[c] (CP027601)
2014C-3599	O121:H19	CP027435	2	5,400,138	83,611 (CP027436)
2012C-4221[b]	O101:H6	CP027437	3	5,012,557	74,904 (CP027438), 107,188 (CP027439)
2012C-4502	O185:H28	CP027440	2	4,892,666	173,714 (CP027441)
2013C-3252	O69:H11	CP027442	3	5,636,732	95,157 (CP027443), 91,399 (CP027444)
2013C-3492[b]	O172:H25	CP027445	2	5,196,105	74,269 (CP027446)
2014C-3075	O36:H42	CP027447	2	5,168,620	170,848 (CP027448)
2014C-3097[b]	O181:H49	CP027449	3	5,077,228	34,867 (CP027450), 173,649 (CP027451)
2014C-3338[b]	O183:H18	CP027452	2	4,799,014	159,611 (CP027453)
2014C-4423[b]	O121:H19	CP027454	3	5,338,915	73,262 (CP027455), 79,682 (CP027456)
88-3493[b]	O137:H41	CP027457	2	5,001,754	107,796 (CP027458)
90-3040[b]	O172:H25	CP027459	2	5,253,712	74,247 (CP027460)
95-3322[b]	O22:H5	CP027461	1	5,095,223	No plasmids
07-4299[b]	O130:H11	CP027462	2	4,847,172	125,059[c] (CP027463)
2013C-4248	O186:H2	CP027464	8	5,243,827	113,063 (CP027465), 10,950[c] (CP027466), 62,602 (CP027467), 97,439[c] (CP027468), 62,881[c] (CP027469), 80,206 (CP027470), 243,267 (CP027471)

(Continued on next page)

TABLE 1 (Continued)

Escherichia coli strain ID[a]	Serotype	Chromosomal GenBank accession no.	No. of contigs	Chromosome size (bp)	Associated plasmid size(s) (bp) (GenBank accession no.)
2014C-3050[b]	O118:H16	CP027472	2	5,671,594	81,624[c] (CP027473)
89-3506[b]	O126:H27	CP027520	3	5,178,386[c]	160,231[c] (CP027521), 93,253[c] (CP027522)
2013C-3264[b]	O103:H25	CP027544	2	5,486,407	101,089 (CP027545)
2013C-4187[b]	O71:H11	CP027546	2	5,509,931	95,367 (CP027547)
2014C-3061[b]	O156:H25	CP027548	2	5,303,935	94,116 (CP027549)
2014C-4705[b]	O112:H21	CP027640	2	5,329,029	126,957 (CP027641)
2015C-4136CT1[b]	O145:H34	CP027550	2	4,836,918	162,810 (CP027551)
2015C-4498[b]	O117:H8	CP027552	2	5,434,442	67,055 (CP027553)
2013C-3513[b]	O186:H11	CP027555	3	5,584,939	70,129[c] (CP027554), 91,046[c] (CP027556)
2013C-3996	O26:H11	CP027572	2	5,858,766[c]	96,937 (CP027571)
2013C-4081[b]	O111:H8	CP027573	4	5,411,943	48,183 (CP027574), 95,952[c] (CP027575), 78,427 (CP027576)
2013C-4225[b]	O103:H11	CP027577	2	5,646,446	87,714[c] (CP027578)
2013C-4282[b]	O77:H45	CP027579	3	5,030,044	54,544[c] (CP027580), 118,822 (CP027581)
2013C-4538[b]	O118:H16	CP027582	2	5,680,428	88,339[c] (CP027583)
2014C-3003[b]	O76:H19	CP027672	3	5,234,640	88,529[c] (CP027673), 133,420 (CP027674)
2015C-3125[b]	O145:H28	CP027763	3	5,471,132	66,944[c] (CP027764), 66,388 (CP027765)
00-3076[b]	O113:H21	CP027584	2	4,997,979	160,576 (CP027585)
2012C-4196	O145:H25	PVZZ00000000	5	3,847,435,[c] 1,375,699[c]	26,290,[c] 111,344,[c] 65,126[c]
2012EL-2448[b]	O91:H14	CP027586	1	5,272,286	No plasmids
2013C-4974[b]	O5:H9	CP027587	2	5,235,560	58,109[c] (CP027588)
2014C-3011[b]	O177:H25	CP027591	4	5,168,350	75,065[c] (CP027589), 92,449[c] (CP027590), 17,880[c] (CP027592)
2013C-3304	O71:H8	CP027593	4	5,309,950[c]	14,119[c] (CP027594), 36,845[c] (CP027595), 87,855 (CP027596)
86-3153[b]	O5:H9	CP027597	2	5,342,528	74,505[c] (CP027598)
88-3510[b]	O172:H25	CP027675	2	5,140,386	65,738[c] (CP027676)
2013C-3342	O117:H8	CP027766	2	5,489,451	66,545 (CP027767)

[a]ID, identification.
[b]Strain for which an optical map was generated and used to confirm the sequence order.
[c]A linear sequence that could not be circularized due to unresolved or collapsed repeats.

or unresolved repeats, a single chromosomal sequence was obtained for 2014C-3741, 2014C-3716, 89-3506, 2013C-3996, and 2013C-3304. The remaining four genomes (2013C-3925, 03-3375, 2014C-4638, and 2012C-4196) had two or more chromosomal contigs. The average genome size of the 73 isolates with a single chromosomal sequence was 5,287,902 bp, ranging from 4,717,123 to 5,858,766 bp. Each genome contained between one and seven plasmids.

Accession number(s). The whole-genome sequences have been deposited in the DDBJ/ENA/GenBank under the accession numbers listed in Table 1. The versions described in this paper are first versions.

ACKNOWLEDGMENTS

This work was funded by federal appropriations to the Centers for Disease Control and Prevention, through the Advanced Molecular Detection Initiative line item.

The findings and conclusions of this article are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. The use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention or by the U.S. Department of Health and Human Services.

REFERENCES

1. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV. 1999. Food-related illness and death in the United States. Emerg Infect Dis 5:607–625. https://doi.org/10.3201/eid0505.990502.
2. Bettelheim KA. 2007. The non-O157 Shiga-toxigenic (verocytotoxigenic) *Escherichia coli*; under-rated pathogens. Crit Rev Microbiol 33:67–87. https://doi.org/10.1080/10408410601172172.
3. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM, Hoekstra RM, Strockbine NA. 2005. Non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States, 1983–2002. J Infect Dis 192:1422–1429. https://doi.org/10.1086/466536.
4. Hughes JM, Wilson ME, Johnson KE, Thorpe CM, Sears CL. 2006. The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*. Clin Infect Dis 43:1587–1595. https://doi.org/10.1086/509573.

5. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Non-hybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563–569. https://doi.org/10.1038/nmeth.2474.
6. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054. https://doi.org/10.1038/nmeth.4035.
7. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.