AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Draft Genome Sequences of the *Escherichia coli* Reference (ECOR) Collection

Isha R. Patel,[a] Jayanthi Gangiredla,[a] Mark K. Mammel,[a] Keith A. Lampel,[a] Christopher A. Elkins,[a] David W. Lacher[a]

[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

ABSTRACT Here, we report the genomes of all 72 isolates belonging to the *Escherichia coli* reference (ECOR) collection. Strains in this collection were isolated from diverse hosts and geographic locations and have been used for more than 30 years to represent the phylogenetic diversity of *E. coli*.

Escherichia coli has been used as a model species to analyze the processes involved in bacterial genome evolution. More than 30 years ago, a set of strains known as the *E. coli* reference (ECOR) collection was assembled to represent the known genetic diversity of the species (1). Subsequent phylogenetic studies have shown that pathogenic and nonpathogenic strains of *E. coli* are randomly distributed when classified into the ECOR phylogroups (2, 3). PCR-based methods were used later to reassign strains ECOR 35, 36, 38, 39, 40, and 41 from phylogroup D to the newly described phylogroup F (4, 5). This finding was also supported by whole-genome-based microarray data (6). Since its creation, the ECOR collection has been widely used by scientists around the world. Unfortunately, during this time, several discrepancies from the original collection have been reported (7). Other researchers have made genome sequence data available for the entire ECOR collection (8). However, we caution against use of these data since nearly half of the strains appear to be contaminated, as evidenced by the presence of multiple molecular serotyping loci within the affected assemblies. For example, the assembly for strain ECOR 46 (GenBank accession number LYCC00000000) contains the wzx_{O1}, wzx_{O7}, wzy_{O1}, wzy_{O7}, $fliC_{H6}$ and $fliC_{H45}$ alleles. Here, we report our version of the ECOR collection so that others may use the data to better understand the nature of their differences with the original collection (Table 1).

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 protocol for Gram-negative bacteria. Sequencing libraries were prepared with 1 ng of 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 controlled using FastQC (Q score, $>$30) and *de novo* assembled using SPAdes 3.8.2 (9) or CLC Genomics Workbench 8.2.1 (CLC bio, Aarhus, Denmark).

Depth of coverage for the draft genomes ranged from 23× to 229×, with the genome sizes ranging from 4,506,698 to 5,591,744 bp. The number of contigs ranged from 47 to 354, while the N_{50} values ranged from 58,260 to 467,104 bp. Preliminary phylogenetic analysis utilizing polymorphisms present within conserved core genes identified two strains as belonging to a phylogroup inconsistent with their expected ECOR designation. Phylogroup A strains ECOR 7 and ECOR 23 were found to cluster within phylogroups B1 and B2, respectively. The phylogroup F status of strains ECOR 35, 36, 38, 39, 40, and 41 was confirmed by our phylogenetic analysis.

Received 13 August 2018 **Accepted** 14 September 2018 **Published** 11 October 2018

Citation Patel IR, Gangiredla J, Mammel MK, Lampel KA, Elkins CA, Lacher DW. 2018. Draft genome sequences of the *Escherichia coli* reference (ECOR) collection. Microbiol Resour Announc 7:e01133-18. https://doi.org/10.1128/MRA.01133-18.

Editor Irene L. G. Newton, Indiana University Bloomington

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

TABLE 1 Accession numbers and assembly metrics for the 72 ECOR strains

ECOR strain	SRA run no.	No. of reads	GenBank accession no.	Average coverage (×)	No. of contigs	Genome size (bp)	N_{50} values (bp)	G+C content (%)	Messerer et al. (8) GenBank accession no.
1	SRR3989531	2,464,522	QOWM00000000	32.5	79	4,715,492	430,669	50.6	LYBJ00000000[a]
2	SRR7819145	6,102,266	QOWN00000000	78.9	234	5,239,177	104,055	50.7	LYBI00000000[a]
3	SRR3951465	6,911,676	QOWO00000000	81.7	143	4,926,647	124,670	50.7	LYBH00000000
4	SRR3951466	5,536,830	QOWP00000000	72.0	121	4,587,238	129,060	50.8	LYBG00000000[a]
5	SRR3951467	4,299,776	QOWQ00000000	50.8	246	5,065,618	121,007	50.7	LYBF00000000[a]
6	SRR3951468	8,044,114	QOWR00000000	125.0	144	4,556,862	69,692	50.9	LYBE00000000[a]
7	SRR3951469	6,530,276	QOWS00000000	80.2	90	4,896,746	169,935	50.7	LYBD00000000
8	SRR3951470	10,733,010	QOWT00000000	130.6	150	4,896,359	146,113	50.5	LYBC00000000[a]
9	SRR3951471	6,022,232	QOWU00000000	68.6	287	5,177,520	67,394	50.9	LYBB00000000
10	SRR3951472	5,570,968	QOWV00000000	68.6	98	4,751,057	230,922	50.6	LYBA00000000[a]
11	SRR3951473	12,657,330	QOWW00000000	145.5	280	5,182,586	144,669	50.7	LYAZ00000000
12	SRR3951474	9,606,228	QOWX00000000	114.4	218	5,078,970	128,904	50.7	LYAY00000000[a]
13	SRR3951475	9,627,106	QOWY00000000	127.8	165	4,619,031	70,958	50.8	LYAX00000000
14	SRR3951476	6,464,928	QOWZ00000000	79.1	150	4,941,866	112,447	50.7	LYAW00000000[a]
15	SRR3951517	4,867,598	QOXA00000000	60.5	128	4,911,383	121,074	50.7	LYAV00000000
16	SRR3951477	4,046,036	QOXB00000000	52.9	126	4,636,981	93,940	50.8	LYAU00000000
17	SRR3951478	4,786,784	QOXC00000000	66.1	110	4,506,698	88,853	50.7	LYAT00000000
18	SRR3951479	9,042,146	QOXD00000000	121.2	112	4,634,531	104,351	50.8	LYAS00000000
19	SRR3951480	9,318,330	QOXE00000000	142.7	164	4,535,554	82,278	50.8	LYAR00000000
20	SRR3951481	10,805,000	QOXF00000000	155.2	154	4,595,721	84,141	50.8	LYAQ00000000
21	SRR3951482	6,045,112	QOXG00000000	85.1	159	4,568,662	85,429	50.9	LYAP00000000
22	SRR3951484	9,380,404	QOXH00000000	118.8	84	4,514,994	211,903	50.8	LYAO00000000
23	SRR3951485	4,819,930	QOXI00000000	56.6	161	5,093,876	233,304	50.4	LYAN00000000
24	SRR3951486	10,917,604	QOXJ00000000	119.9	146	5,227,547	157,129	50.7	LYAM00000000
25	SRR3951488	7,017,374	QOXK00000000	86.9	125	4,752,894	184,763	50.5	LYAL00000000
26	SRR3951489	12,549,182	QOXL00000000	153.3	93	4,678,648	236,086	50.7	LYAK00000000
27	SRR3951490	10,793,212	QOXM00000000	128.6	109	4,867,073	190,031	50.5	LYAJ00000000
28	SRR3951491	6,645,234	QOXN00000000	81.3	109	4,925,046	187,237	50.7	LYAI00000000
29	SRR3951492	6,383,664	QOXO00000000	79.1	112	4,928,564	177,525	50.6	LYAH00000000
30	SRR3951493	6,835,550	QOXP00000000	85.1	120	4,825,526	193,320	50.6	LYAG00000000[a]
31	SRR3951494	11,223,998	QOXQ00000000	126.0	120	5,302,667	135,887	50.7	LYAF00000000
32	SRR3951518	10,348,842	QOXR00000000	129.2	129	4,794,190	185,245	50.7	LYAE00000000
33	SRR3951496	11,691,094	QOXS00000000	152.7	129	4,795,454	185,228	50.7	LYAD00000000[a]
34	SRR3951497	12,857,174	QOXT00000000	165.3	128	4,908,743	154,834	50.7	LYAC00000000[a]
35	SRR3951498	9,789,342	QOXU00000000	132.4	220	5,104,518	79,466	50.6	LYAB00000000
36	SRR3951499	10,108,776	QOXV00000000	134.8	279	5,231,499	58,260	50.5	LYBO00000000
37	SRR3951500	14,011,950	QOXW00000000	149.3	313	5,589,959	97,745	50.3	LYAA00000000
38	SRR3951501	19,252,412	QOXX00000000	211.7	206	5,240,321	109,902	50.5	LXZZ00000000[a]
39	SRR3951503	16,096,112	QOXY00000000	170.4	211	5,284,758	109,902	50.4	LYCJ00000000[a]
40	SRR3951504	5,498,306	QOXZ00000000	62.4	190	5,201,125	109,345	50.5	LYCI00000000[a]
41	SRR3951505	12,179,258	QOYA00000000	131.7	204	5,242,084	105,640	50.4	LYCH00000000
42	SRR3951506	10,293,022	QOYB00000000	114.7	111	5,189,763	467,104	50.5	LYCG00000000[a]
43	SRR3951508	10,938,788	QOYC00000000	127.3	226	5,272,828	107,735	50.6	LYCF00000000[a]
44	SRR3951509	9,842,910	QOYD00000000	112.4	171	5,240,115	179,757	50.6	LYCE00000000[a]
45	SRR3951510	17,558,230	QOYE00000000	226.1	95	4,726,888	210,969	50.7	LYCD00000000[a]
46	SRR3987677	2,932,548	QOYF00000000	107.7	151	5,259,340	90,485	50.5	LYCC00000000[a]
47	SRR3951512	18,191,986	QOYG00000000	228.5	77	4,920,788	214,558	50.6	LYCB00000000[a]
48	SRR3951513	18,794,936	QOYH00000000	218.1	131	5,333,881	171,963	50.5	LYCA00000000[a]
49	SRR3989514	4,399,522	QOYI00000000	187.7	277	5,278,098	118,118	50.6	LYBZ00000000
50	SRR3989532	1,990,960	QOYJ00000000	22.7	354	5,591,744	82,352	50.5	LYBY00000000
51	SRR3989533	3,562,062	QOYK00000000	44.5	135	5,169,898	234,505	50.5	LYYB00000000[a]
52	SRR7819144	5,273,668	QOYL00000000	72.6	172	5,097,032	219,132	50.4	LYCT00000000
53	SRR7819143	3,163,954	QOYM00000000	43.3	141	5,131,093	244,875	50.4	LYCU00000000
54	SRR3989534	5,331,188	QOYN00000000	70.4	113	5,035,502	332,014	50.4	LYCV00000000
55	SRR3989535	4,680,452	QOYO00000000	60.0	122	5,049,489	284,057	50.6	LYCW00000000
56	SRR7819142	2,047,664	QOYP00000000	28.8	110	4,956,973	196,112	50.5	LYDK00000000
57	SRR3989515	3,149,298	QOYQ00000000	137.2	128	5,277,375	185,269	50.5	LYCX00000000[a]
58	SRR3989521	1,955,120	QOYR00000000	91.7	103	4,902,658	120,850	50.6	LYCY00000000[a]
59	SRR3989507	2,122,200	QOZF00000000	99.6	84	4,764,104	194,584	50.4	LYCZ00000000[a]
60	SRR3989522	1,995,476	QOYS00000000	90.8	135	5,068,027	305,372	50.7	LYDA00000000[a]
61	SRR3987678	2,362,538	QOYT00000000	99.1	118	4,885,105	129,207	50.6	LYDB00000000[a]
62	SRR3989523	2,518,590	QOYU00000000	114.4	115	5,155,092	161,545	50.4	LYDL00000000[a]
63	SRR3989529	1,790,086	QOYV00000000	78.0	162	5,113,907	113,045	50.5	LYDM00000000
64	SRR7819148	10,979,330	QOYW00000000	143.2	169	5,105,087	196,230	50.8	LYDC00000000
65	SRR3989537	3,338,208	QOYX00000000	42.2	102	4,946,845	237,525	50.7	LYDD00000000

(Continued on next page)

TABLE 1 (Continued)

ECOR strain	SRA run no.	No. of reads	GenBank accession no.	Average coverage (×)	No. of contigs	Genome size (bp)	N_{50} values (bp)	G+C content (%)	Messerer et al. (8) GenBank accession no.
66	SRR3989538	2,878,266	QOYY00000000	39.5	47	4,739,108	239,490	50.9	LYDE00000000[a]
67	SRR3989530	1,916,842	QOYZ00000000	170.2	82	4,725,509	198,152	50.7	LYDN00000000
68	SRR7819147	7,341,922	QOZA00000000	99.6	156	5,037,255	191,327	50.7	LYDF00000000
69	SRR3989539	2,363,422	QOZB00000000	33.7	74	4,641,530	141,371	50.6	LYDG00000000
70	SRR3989540	4,864,768	QOZC00000000	61.8	135	5,123,298	225,733	50.8	LYDH00000000[a]
71	SRR3989541	4,321,280	QOZD00000000	56.8	138	4,875,196	113,172	50.8	LYDI00000000[a]
72	SRR7819146	2,653,970	QOZE00000000	39.2	75	4,725,669	288,629	50.6	LYDJ00000000[a]

[a] Possible contamination identified by the presence of multiple O and/or H molecular serotyping loci.

Data availability. The draft genome assemblies were deposited at DDBJ/ENA/GenBank under the accession numbers QOWM00000000 to QOZF00000000 and under BioProject accession number PRJNA230969. The versions described in this announcement are the first versions.

ACKNOWLEDGMENTS

We thank Shannon Manning at Michigan State University for providing replacements for some of the ECOR strains.

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. Ochman H, Selander RK. 1984. Standard reference strains of *Escherichia coli* from natural populations. J Bacteriol 157:690–693.
2. Gordon DM, Clermont O, Tolley H, Denamur E. 2008. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. Environ Microbiol 10:2484–2496. https://doi.org/10.1111/j.1462-2920.2008.01669.x.
3. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol 8:207–217. https://doi.org/10.1038/nrmicro2298.
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. Clermont O, Gordon D, Denamur E. 2015. Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. Microbiology 161:980–988. https://doi.org/10.1099/mic.0.000063.
6. Jackson SA, Patel IR, Barnaba T, LeClerc JE, Cebula TA. 2011. Investigating the global genomic diversity of *Escherichia coli* using a multi-genome DNA microarray platform with novel gene prediction strategies. BMC Genomics 12:349. https://doi.org/10.1186/1471-2164-12-349.
7. Johnson JR, Delavari P, Stell AL, Prats G, Carlino U, Russo TA. 2001. On the integrity of archival strain collections, including the ECOR collection. ASM News 67:288–289.
8. Messerer M, Fischer W, Schubert S. 2017. Investigation of horizontal gene transfer of pathogenicity islands in *Escherichia coli* using next-generation sequencing. PLoS One 12:e0179880. https://doi.org/10.1371/journal.pone.0179880.
9. 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.