AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

Draft Genome Sequences of 16 Strains of *Escherichia* Cryptic Clade II Isolated from Intertidal Sediment in Hong Kong

Zhiyong Shen,[a] Xiu Pei Koh,[b] Yanping Yu,[a] Chun Fai Woo,[b] Yigang Tong,[c] Stanley C. K. Lau[a,b]

[a]Department of Ocean Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
[b]Division of Environment and Sustainability, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
[c]State Key Laboratory of Pathogen and Biosecurity, Beijing, China

ABSTRACT The genus *Escherichia* includes several cryptic clades. Among them, the members of cryptic clade II have rarely been found, and their genome sequences remain largely uninvestigated. Here, we report the draft genome sequences of 16 strains of *Escherichia* cryptic clade II that were isolated from intertidal sediment in Hong Kong.

Escherichia is currently comprised of four validly published species (*E. albertii*, *E. coli*, *E. fergusonii*, and *E. marmotae*) and a number of genetically divergent yet taxonomically inconspicuous monophyletic lineages, commonly referred to as cryptic clades. The distribution, prevalence, ecological niches, and genomic features of the cryptic clades have been investigated in a number of studies (1–3). However, due to the scarcity of its isolates, the genomic composition of cryptic clade II remains largely uninvestigated, leaving a knowledge gap about the evolutionary history, ecological character, and evolution of the *Escherichia* genus as a whole. Here, we report the draft genome sequences of 16 strains of cryptic clade II, isolated from the intertidal sediment in the subtropical environment of Hong Kong.

The strains were isolated using the selective medium CHROMagar ECC (CHROMagar, France) and putatively identified as *E. coli* on the basis of their growth on the medium as blue colonies. However, in a maximum likelihood phylogenetic tree constructed using the concatenated DNA sequences of seven housekeeping genes, *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA* (4), the 16 strains occupied the same monophyletic lineage as that occupied by previously reported members of cryptic clade II (1) instead of being affiliated with *E. coli*.

To obtain the genome sequences of the 16 strains of cryptic clade II, genomic DNA extracted from overnight cultures in Luria-Bertani broth was sheared using Ion Shear Plus reagents, end repaired, and ligated to Ion Torrent adapters (Life Technologies, USA). Libraries containing fragments ca. 400 bp in length were sequenced on an Ion Torrent platform to generate single-ended reads. The sequence reads were processed and analyzed using software with default parameters, as described below.

Briefly, raw reads were filtered for quality using FastQC ($Q > 30$). Clean reads without adapter sequences were *de novo* assembled into contigs using MIRA v4.0.2 (5) and SPAdes v3.6.0 (6) on the SIMBA Workbench platform (7). Assembly quality was enhanced through (i) the mapping of contigs to reference genomes by using CONTIGuator (8) and the optical mapping reports generated by MapSolver (OpGen, Inc.), (ii) the determination of the origins in circular genomes by using the moveDNAA.py script, and (iii) the manual closure of gaps through the identification of repeats on the extremities of the contig using BLAST (9). After the removal of small contigs ($<$500 bp), the genome sequences obtained for the 16 strains were 4,687,583 to 5,244,655 bp (Table 1). N_{50} values were all greater than 100 kb. Except for strain E4742,

Citation Shen Z, Koh XP, Yu Y, Woo CF, Tong Y, Lau SCK. 2019. Draft genome sequences of 16 strains of *Escherichia* cryptic clade II isolated from intertidal sediment in Hong Kong. Microbiol Resour Announc 8:e00416-19. https://doi.org/10.1128/MRA.00416-19.

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

Copyright © 2019 Shen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Stanley C. K. Lau, scklau@ust.hk.

Received 9 April 2019
Accepted 27 June 2019
Published 18 July 2019

TABLE 1 Summary of draft genome assemblies

Strain	Total no. of reads	Read length (bp)	Coverage (×)	Assembly size (bp)	No. of contigs	N_{50} (bp)	G+C content (%)	No. of CDS[a]	No. of tRNA coding genes	GenBank accession no.	Sequence Read Archive accession no.
E1130	1,176,251	35–400	50	5,244,655	62	383,389	50.71	5,040	77	PDIL00000000	SRX3260321
E2562	1,280,272	35–400	50	4,940,096	49	332,841	50.47	4,724	85	PDIJ00000000	SRX3260343
E2586	1,102,521	35–400	45	5,094,594	78	364,248	50.47	5,232	77	PDIH00000000	SRX3260344
E2593	1,428,384	35–400	60	5,176,532	78	173,826	50.41	5,417	72	PDII00000000	SRX3260345
E2661	1,047,543	35–400	40	4,955,692	95	183,982	50.73	4,952	79	PDIG00000000	SRX3260346
E2748	1,158,520	35–400	45	4,837,679	60	314,245	50.61	4,788	76	PDIF00000000	SRX3260347
E3356	1,104,182	35–400	40	4,918,899	49	389,954	50.55	4,800	76	PDIE00000000	SRX3260348
E3659	973,907	35–400	40	5,033,702	76	136,855	50.60	4,886	73	PDID00000000	SRX3260350
E4208	1,041,529	35–400	45	4,687,583	57	135,714	50.81	4,762	81	PDIC00000000	SRX3260359
E4385	1,796,955	35–400	80	5,015,426	43	190,271	50.59	5,105	75	PDIK00000000	SRX5623432
E4694	1,179,794	35–400	50	4,891,728	36	361,728	50.62	4,575	76	PDIB00000000	SRX3260358
E4702	1,145,459	35–400	50	5,114,924	73	219,627	50.75	4,764	80	PDIA00000000	SRX3260360
E4736	1,879,325	35–400	100	4,844,105	43	310,513	50.50	4,453	67	PDHX00000000	SRX3200104
E4742	1,270,039	35–400	50	5,195,263	113	177,505	50.60	4,897	89	PDHY00000000	SRX3260361
E4930	1,886,454	35–400	100	5,119,612	78	247,834	50.60	4,767	79	PDHZ00000000	SRX3260362
E5028	1,787,838	35–400	100	5,050,713	91	147,572	50.50	4,596	84	PDHW00000000	SRX3268009

[a] CDS, coding DNA sequences.

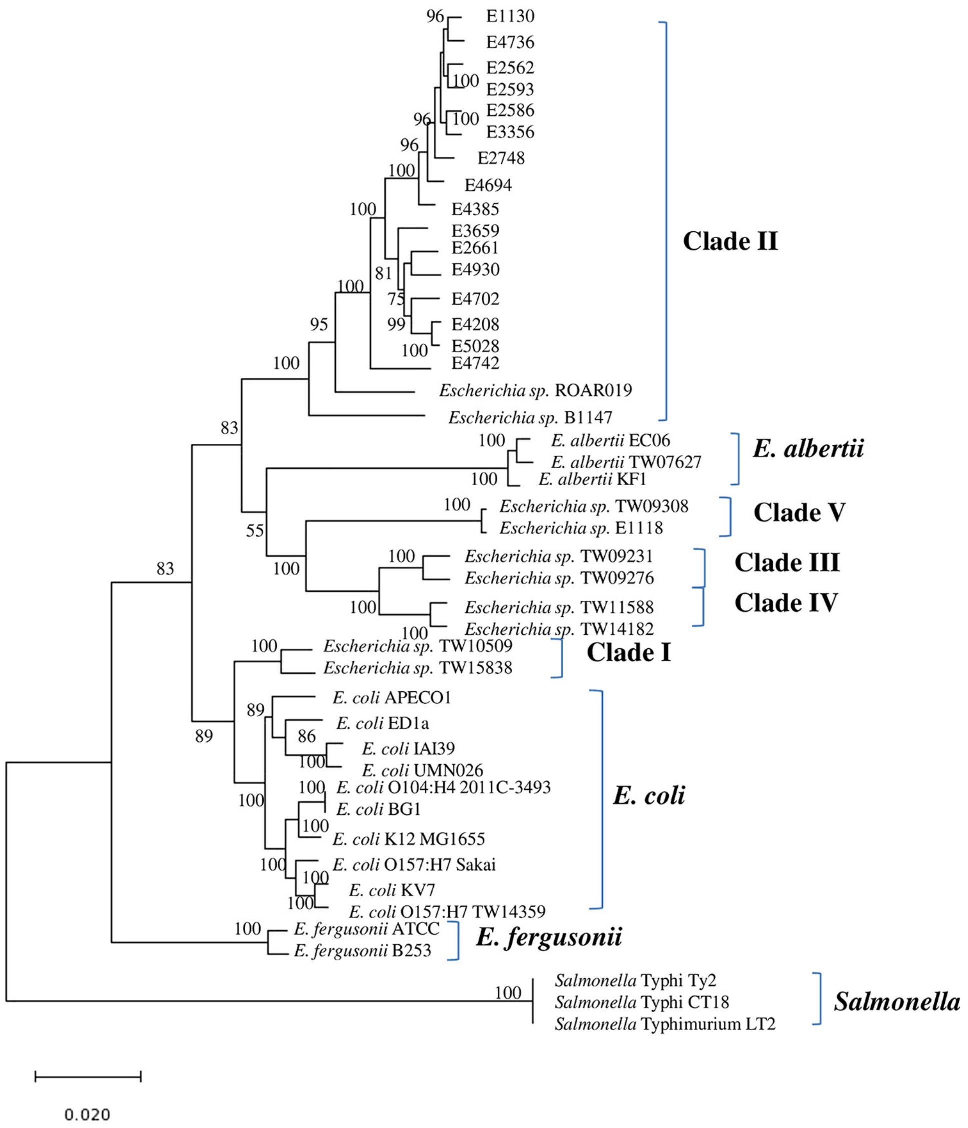

FIG 1 A maximum likelihood phylogenetic tree of 405 core genes extracted from 44 strains of *Escherichia* and *Salmonella*. The concentenated sequences of the core genes were aligned using MUSCLE (11). The tree was constructed by using MEGA7 (12) with the Jukes-Cantor subsitution model, the nearest-neighbor interchange topology search strategy, and 100-bootstrap replication. Numbers at the nodes indicate bootstrap values that are greater than 50. The scale bar indicates substitutions per nucleotide position.

the genome assembly of each strain contains less than 100 contigs. The genome sequences were annotated using the NCBI Prokaryotic Genome Annotation Pipeline. In a maximum likelihood phylogenetic tree constructed for the core genes in *Escherichia* spp. (Fig. 1), the 16 strains and the previously reported members of cryptic clade II (i.e., *Escherichia* sp. strains ROAR019 [10] and B1147 [1]) occupied a monophyletic lineage, congruent to the phylogeny inferred on the basis of seven housekeeping genes.

Data availability. The genome assemblies have been deposited in DDBJ/ENA/GenBank under BioProject number PRJNA412557 with accession numbers from PDHW00000000 to PDIL00000000 (Table 1).

ACKNOWLEDGMENTS

This research was supported by the Environment and Conservation Fund (ECF 35/2018), and the Theme-based Research Scheme (T21-602/16-R) and General Research Fund (16301018) of the Hong Kong Research Grants Council.

REFERENCES

1. Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. 2011. Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. Environ Microbiol 13:2468–2477. https://doi.org/10.1111/j.1462-2920.2011.02519.x.
2. Gangiredla J, Mammel MK, Barnaba TJ, Tartera C, Gebru ST, Patel IR, Leonard SR, Kotewicz ML, Lampel KA, Elkins CA, Lacher DW. 2018. Draft genome sequences of Escherichia albertii, Escherichia fergusonii, and strains belonging to six cryptic lineages of Escherichia spp. Genome Announc 6:e00271-18. https://doi.org/10.1128/genomeA.00271-18.
3. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A 108:7200–7205. https://doi.org/10.1073/pnas.1015622108.
4. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol 60:1136–1151. https://doi.org/10.1111/j.1365-2958.2006.05172.x.
5. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159. https://doi.org/10.1101/gr.1917404.
6. 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.
7. Mariano DC, Pereira FL, Aguiar EL, Oliveira LC, Benevides L, Guimaraes LC, Folador EL, Sousa TJ, Ghosh P, Barh D, Figueiredo HC, Silva A, Ramos RT, Azevedo VA. 2016. SIMBA: a Web tool for managing bacterial genome assembly generated by Ion PGM sequencing technology. BMC Bioinformatics 17:456. https://doi.org/10.1186/s12859-016-1344-7.
8. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med 6:11. https://doi.org/10.1186/1751-0473-6-11.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
10. von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm A-M, Sjöling Å, Dougan G. 2014. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. Nat Genet 46:1321–1326. https://doi.org/10.1038/ng.3145.
11. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
12. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874. https://doi.org/10.1093/molbev/msw054.