

Genome sequence of *Escherichia coli* H1G2: a B1 lineage isolated from the gut of Hilsa shad (*Tenualosa ilisha*) of Bangladesh

Fahim Al Galib,¹ Abdullah Al Kafi,¹ Santu Biswas,¹ Soharth Hasnat, Dipali Rani Gupta,¹ M. Nazmul Hoque,² Md. Morshedur Rahman,³ Md. Mahbubur Rahman,¹ Tofazzal Islam¹

AUTHOR AFFILIATIONS See affiliation list on p. 3.

ABSTRACT We characterized the 4.8 Mbp genome of *Escherichia coli* H1G2, a B1 lineage isolated from the gut of Bangladesh's national fish, Hilsa (*Tenualosa ilisha*). The complete assembly reveals specialized genes for sugar metabolism and adhesion, facilitating niche-specific colonization, thereby providing insights into the microbial ecology of this commercially significant species' gastrointestinal tract.

KEYWORDS complete genome, gut microbiota, Hilsa gut, symbiotic, Oxford Nanopore sequencing

Escherichia coli is a vertebrate gut bacterium that supports host health (1, 2). While well-studied in humans, genomic insights into strains from aquatic hosts like Hilsa fish remain limited (3, 4). The *E. coli* was isolated from Hilsa shad (*Tenualosa ilisha*) gut captured from the Meghna River in Chandpur, Bangladesh (22.8333°N, 90.8333°E). Gut contents were aseptically removed, and a 1 g pooled sample was homogenized in 9 mL of phosphate-buffered saline (PBS, pH 7.0), enriched in MacConkey broth (37°C for 24 h at 180 rpm), and plated on eosin methylene blue (EMB) agar (Oxoid, UK) with incubation at 37°C for 24 h (3, 4). A single green metallic sheen colony was purified and designated H1G2 and was initially confirmed as *Escherichia* sp. by PCR amplification of the 16S rRNA gene (27F/1492R), followed by bidirectional Sanger sequencing (5). Genomic DNA from an overnight (37°C) EMB culture was extracted using the FavorPrep Tissue Genomic DNA Extraction HE Mini Kit (Favorgen, Taiwan) and quantified with a Qubit 4.0 fluorometer (Thermo Fisher Scientific). Libraries were prepared using the Oxford Nanopore Technologies (ONT) Ligation Sequencing Kit SQK-NBD114.24 with the NEBNext companion module (New England Biolabs) and Native Barcoding Kit 96 (v14) (no shearing and no size selection) and sequenced on an ONT MinION using an R10.4.1 flow cell (6). Basecalling with Dorado v1.2.0 (SUP model) yielded 62,359 reads totaling 391,812,180 bp (read N_{50} , 11,906). Read quality was assessed with NanoPlot v1.46.1 (7); adapters were trimmed with Porechop v0.2.4 (8); reads >1 kb were filtered with Filtrlong v0.2.1 (9); *de novo* assembly was performed with Flye v2.9.5-b1801 (10); and completeness and contamination were evaluated with BUSCO v6.0.0 (*-l bacteria_odb10, -m genome*) (11) and CheckM v1.2.4 (*lineage_wf, Escherichia coli* marker set) (12). The primary annotation present in the GenBank submission was generated via National Center for Biotechnology Information (NCBI) PGAP v6.10 (13). Furthermore, RAST v2.0 (14), PlasmidFinder v2.0.1 (15), and CRISPRCasFinder v2.0.3 (16) were utilized to identify specific metabolic features, plasmids, and CRISPR arrays, respectively. Manual curation was utilized to resolve differences in annotation. Default parameters were used for all software unless otherwise specified. The genome without rotating at specific base comprises a complete 4,779,014-bp circular chromosome (51% GC) and two plasmids (circular) with 4,516 protein-coding genes (Table 1). Phylogenomic analysis assigned H1G2 to *E. coli* phylogroup B1, sharing 99.24% average nucleotide identity (ANI) with the

Editor David Rasko, University of Maryland Baltimore, Baltimore, Maryland, USA

Address correspondence to Tofazzal Islam, tofazzalislam@gau.edu.bd.

Fahim Al Galib and Abdullah Al Kafi contributed equally to this article. Order was maintained based on alphabetic order.

The authors declare no conflict of interest.

Received 2 February 2026

Accepted 30 March 2026

Published 20 April 2026

Copyright © 2026 Galib et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

TABLE 1 Genomic features of *E. coli* strain H1G2 isolated from the Hilsa fish gut

Genome feature	H1G2
GenBank accession	JBSVSG000000000
Assembly accession	GCA_054403135.1
SRA	SRR37373213
Number of reads	62,359 (N_{50} 11,906)
Assembled genome size (bp)	4,779,014 bp
GC content (%)	51
Coverage (×)	153×
Genome completeness (%)	100% (BUSCO), 99.36% (CheckM)
Contamination (%)	0.23%
BioSample accession	SAMN53736888
16S GenBank	PZ053874
N_{50} value (bp)	4,734,525
Number of contigs	3 (1 chromosome, plasmid)
Genes (total)	4,631
Coding sequences (CDSs)	4,516
Genes (coding)	4,384
CDSs (with protein)	4,384
Genes (RNA)	115
rRNAs	8, 7, 7 (5S, 16S, 23S)
Complete rRNAs	8, 7, 7 (5S, 16S, 23S)
tRNAs	85
ncRNAs	8
Pseudo genes	132
CDSs (without protein)	132
Pseudo genes (ambiguous residues)	0 of 132
Pseudo genes (frameshifted)	57 of 132
Pseudo genes (incomplete)	68 of 132
Pseudo genes (internal stop)	31 of 132
Pseudo genes (multiple problems)	22 of 132
Pseudo Genes (short protein)	1 of 132
CRISPR arrays	1
Plasmid replicon	2
Number of subsystems (metabolic features)	377

commensal strain SE11 by GTDB-Tk v2.6.1 (17); *in silico* serotyping identified O40:H2 (18); BLAST against VFDB (19) identified it lacks major diarrheagenic toxins (e.g., Shiga toxin, *east1*) and pathogenic adhesins (*aggR*, *aatA*). However, it possesses genes for adhesion (*type-1 fimbriae*, *curlI*), iron acquisition, vitamin K₂ biosynthesis, and pathways for utilizing mucus-derived sugars—including fucose (*fucI*, *fucP*), galactose (*gal* and *mgl* operons), mannose (*manXYZ*), and N-acetylglucosamine (*nagE*, *nagA*), along with vitamin B₁₂ uptake and colicin V production. These genetic features suggest specialized adaptation as a potentially beneficial commensal bacterium within the Hilsa gastrointestinal tract.

While metagenomic surveys previously identified *E. coli* in the Hilsa gut (20, 21), this study for the first time reports the complete genome sequence of strain H1G2, a B1 phylogroup lineage indigenous to this host, offering critical insights for adaptation within aquatic environments.

ACKNOWLEDGMENTS

The authors are thankful to Maksudur Rahman Nayem of DNA Solution Ltd., Bangladesh for technical assistance for sequencing the DNA samples.

Fahim Al Galib, Investigation, Methodology, Data curation, Writing –original draft | Abdullah Al Kafi, Software, Investigation, Methodology, Writing – original draft | Santu

Biswas, Software, Investigation, Writing – original draft | Soharth Hasnat, Software, Validation, Writing – review and editing | Dipali Rani Gupta, Supervision, Resources, Validation, Writing – review and editing | M. Nazmul Hoque, Validation, Writing – review and editing | Md. Mahbubur Rahman, Validation, Writing – review and editing | Md. Morshedur Rahman, Validation, Writing – review and editing | Tofazzal Islam, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review and editing.

AUTHOR AFFILIATIONS

¹Institute of Biotechnology and Genetic Engineering (IBGE), Gazipur Agricultural University (GAU), Gazipur, Bangladesh

²Department of Gynecology, Obstetrics and Reproductive Health, Molecular Biology and Bioinformatics Laboratory, Gazipur Agricultural University, Gazipur, Bangladesh

³Department of Dairy and Poultry Science, Gazipur Agricultural University (GAU), Gazipur, Bangladesh

AUTHOR ORCIDs

Md. Mahbubur Rahman  <http://orcid.org/0000-0003-1181-6817>

Tofazzal Islam  <http://orcid.org/0000-0002-7613-0261>

AUTHOR CONTRIBUTIONS

Fahim Al Galib, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing | Santu Biswas, Formal analysis, Resources, Software, Writing – original draft | Soharth Hasnat, Formal analysis, Software, Validation, Writing – original draft | Dipali Rani Gupta, Data curation, Formal analysis, Funding acquisition, Project administration, Writing – original draft | M. Nazmul Hoque, Validation, Writing – review and editing | Md. Morshedur Rahman, Methodology, Supervision, Validation, Writing – review and editing | Md. Mahbubur Rahman, Validation, Writing – review and editing.

DATA AVAILABILITY

The complete genome sequence of *E. coli* strain H1G2 has been deposited in GenBank assembly under accession number [GCA_054403135.1](https://www.ncbi.nlm.nih.gov/assembly/GCA_054403135.1), within BioProject [PRJNA1055760](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1055760) (Mining Bangladesh Biogold), and Biosample [SAMN53736888](https://www.ncbi.nlm.nih.gov/biosample/SAMN53736888). The version described in this paper is [JBSVSG000000000](https://www.ncbi.nlm.nih.gov/biosample/SAMN53736888).

ETHICS APPROVAL

This study was reviewed and approved by the Animal Research Ethics Committee (AREC) of the Gazipur Agricultural University, Bangladesh (reference number: FVMAS/AREC/2025/95, date: 01/03/2026).

REFERENCES

1. Martinson JNV, Walk ST. 2020. *Escherichia coli* residency in the gut of healthy human adults. *EcoSal Plus* 9:ESP-0003. <https://doi.org/10.1128/ecosalplus.ESP-0003-2020>
2. Moreira de Gouveia MI, Bernalier-Donadille A, Jubelin G. 2024. Enterobacteriaceae in the human gut: dynamics and ecological roles in health and disease. *Biology (Basel)* 13:142. <https://doi.org/10.3390/biology13030142>
3. Oshima K, Toh H, Ogura Y, Sasamoto H, Morita H, Park S-H, Ooka T, Iyoda S, Taylor TD, Hayashi T, Itoh K, Hattori M. 2008. Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. *DNA Res* 15:375–386. <https://doi.org/10.1093/dnares/dsn026>
4. Segura A, Auffret P, Klopp C, Bertin Y, Forano E. 2017. Draft genome sequence and characterization of commensal *Escherichia coli* strain BG1 isolated from bovine gastro-intestinal tract. *Stand Genomic Sci* 12:61. <https://doi.org/10.1186/s40793-017-0272-0>
5. Hoque MN, Istiaq A, Clement RA, Gibson KM, Saha O, Islam OK, Abir RA, Sultana M, Siddiki AZ, Crandall KA, Hossain MA. 2020. Insights into the resistome of bovine clinical mastitis microbiome, a key factor in disease complication. *Front Microbiol* 11:860. <https://doi.org/10.3389/fmicb.2020.00860>
6. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M. 2022. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods* 19:823–826. <https://doi.org/10.1038/s41592-022-01539-7>

7. De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 39:btad311. <https://doi.org/10.1093/bioinformatics/btad311>
8. Bonenfant Q, Noé L, Touzet H. 2023. Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. *Bioinform Adv* 3:vbac085. <https://doi.org/10.1093/bioadv/vbac085>
9. Wick RR. 2025. Filtlong: a quality filtering tool for nanopore and PacBio reads. Available from: <https://github.com/rrwick/Filtlong>
10. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 17:1103–1110. <https://doi.org/10.1038/s41592-020-00971-x>
11. Tegenfeldt F, Kuznetsov D, Manni M, Berkeley M, Zdobnov EM, Kriventseva EV. 2025. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res* 53:D516–D522. <https://doi.org/10.1093/nar/gkae987>
12. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. <https://doi.org/10.1101/gr.186072.114>
13. 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>
14. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. 2014. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res* 42:D206–14. <https://doi.org/10.1093/nar/gkt1226>
15. Carattoli A, Hasman H. 2019. PlasmidFinder and *in silico* pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS), p 285–294. *In* Horizontal gene transfer: methods and protocols. Springer.
16. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCas-Finder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 46:W246–W251. <https://doi.org/10.1093/nar/gky425>
17. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2022. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38:5315–5316. <https://doi.org/10.1093/bioinformatics/bt ac672>
18. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheut F. 2015. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J Clin Microbiol* 53:2410–2426. <https://doi.org/10.1128/JCM.00008-15>
19. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 33:D325–D328. <https://doi.org/10.1093/nar/gki008>
20. Islam T, Hoque MN. 2023. Gut and flesh microbiome sequencing of the Bangladesh national fish hilsa (*Tenualosa ilisha*). *Microbiol Resour Announc* 12:e0044823. <https://doi.org/10.1128/MRA.00448-23>
21. Kawser A, Hoque MN, Rahman MS, Sakif TI, Coffey TJ, Islam T. 2024. Unveiling the gut bacteriome diversity and distribution in the national fish Hilsa (*Tenualosa ilisha*) of Bangladesh. *PLoS One* 19:e0303047. <https://doi.org/10.1371/journal.pone.0303047>