Data in Brief 43 (2022) 108407

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# Genome sequencing data of extended-spectrum beta-lactamase-producing *Escherichia coli* INF191/17/A isolates of nosocomial infection

Nik Siti Hanifah Nik Ahmad[a], Khor Bee Yin[b], Nik Yusnoraini Yusof[c,*]

[a] *School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian, Kelantan 16150, Malaysia*
[b] *BioEasy Sdn Bhd, Setia Avenue, 33A-3 Jalan Setia Prima S, U13/S, Setia Alam, Seksyen U13, Shah Alam, Selangor Darul Ehsan 40170, Malaysia*
[c] *Institute for Research in Molecular Medicine (INFORMM), Health Campus, Universiti Sains Malaysia, Kubang Kerian, Kelantan 16150, Malaysia*

ARTICLE INFO

*Article history:*
Received 25 August 2021
Revised 10 June 2022
Accepted 17 June 2022
Available online 23 June 2022

*Keywords:*
*Escherichia coli*
Genome sequencing
Extended-spectrum beta-lactamase
Antimicrobial resistant gene

ABSTRACT

The infection with extended-spectrum beta-lactamase-producing *Escherichia coli* is associated with higher mortality, longer length of hospital-stay and increased costs compared to infection with antibiotic-susceptible E. coli. Here, the draft genome of ESBL-producing E. coli circulating at local hospital is reported. The strain was detected as containing the genes of antibiotic resistance TEM, CTX-M-1, and CTX-M-9. The 5,136,548-bp genome, with a GC content of 50.59%, comprised 4987 protein-coding genes, four ribosomal RNA, and 66 transfer RNA. The ResFinder was successfully predicted fourteen antimicrobial genes in the *E. coli* INF191/17/A genome. Sequence data has been deposited in the GenBank database under the accession number JAIEXV000000000. The BioProject ID in the GenBank database is PRJNA752944. The raw data was sequenced using Ilumina MiSeq and submitted to the NCBI SRA database (SRX11797310), which is publicly available.

* Corresponding author.
*E-mail address:* nikyus@usm.my (N.Y. Yusof).

https://doi.org/10.1016/j.dib.2022.108407
2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

© 2022 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license
(http://creativecommons.org/licenses/by/4.0/)

## Specifications Table

| | |
|---|---|
| Subject | Health and medical sciences |
| Specific subject area | Microbiology and genomics.<br>Genome sequencing of pathogenic bacteria by using next generation sequencing approach. |
| Type of data | Table<br>Sequencing raw reads in FASTQ format text file<br>Assembled draft genome of *E. coli* strain INF191/17/A in FASTA format text file<br>Genome sequence data in FASTA and FASTQ format |
| How data were acquired | The Illumina MiSeq platform was used to generate paired-end reads of extended spectrum beta lactamase (ESBL)-producing *E. coli* strain INF191/17/A genome. |
| Data format | Raw data in FASTQ format<br>Assembled data in FASTA format: GenBank assembly accession: GCA_019599325.1 (https://www.ncbi.nlm.nih.gov/assembly/ GCA_019599325.1). |
| Parameters for data collection | Bacterial genomic DNA was extracted from a pure culture of ESBL-producing *E. coli* INF191/17/A . Nextera XT DNA library preparation kit was used for the whole-genome sequencing library preparation to generate 2 × 251 paired end reads data. |
| Description of data collection | Whole genome sequencing was performed using Illumina MiSeq system (Illumina®,USA). BBDuk (BBTools v36) was used to trim raw reads, and SPAdes v3.9.0 was used to assemble clean reads. Genome scaffolding was performed with Medusa v1.6. ResFinder software predicted the putative antimicrobial resistant genes. |
| Data source location | Institution: Institute for Research in Molecular Medicine (INFORMM)<br>City/Town/Region: Kubang Kerian, Kelantan<br>Country: Malaysia<br>Latitude and longitude for collected samples/data: 6.10 N 102.28 E |
| Data accessibility | The data is hosted on a public repository.<br>Bioproject: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA752944<br>Biosample: https://www.ncbi.nlm.nih.gov/biosample/SAMN20668118<br>NCBI GenBank Accession Number: JAIEXV000000000<br>https://www.ncbi.nlm.nih.gov/nuccore/JAIEXV000000000<br>Repository name: NCBI SRA database<br>Data identification number: SRR15497613<br>Direct URL to data: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR15497613 |

## Value of the Data

- The whole genome sequencing data provides insight into genomic determinants of the ESBL-producing *E. coli* strains INF191/17/A and antimicrobial resistance (AMR) genes.
- This data should be used by researchers and public health officers to keep up surveillance and control of ESBL-producing gram negative organisms in order to prevent the emergence of highly resistant strain, which is one of serious problem in the world.
- The genome data of *E. coli* strain INF191/17/A accelerates knowledge for pathogenic microbial research in the context of comparative studies, pan-genome, and evolution of non-ESBL and ESBL strains within different epidemiology.
- Furthermore, prior to biomarker discovery, drug or vaccine development, the comprehensive understanding of the whole genome of this pathogen is critically important.

## 1. Data Description

The *Escherichia coli* INF191/17/A was discovered as an extended-spectrum beta-lactamase (ESBL) strain carrying the antibiotic resistance genes TEM, CTX-M-1, and CTX-M-9 via polymerase chain reaction using ESBL specific primers [1]. The 251 base-pair paired-end (2 × 251 bp) sequencing raw reads of the *E. coli* strain INF191/17/A genome were obtained from the Illumina MiSeq system (Illumina, CA, USA) [2]. The raw reads were pre-processed before the genome assembly and annotation. Antimicrobial resistant genes were predicted using curated public database. Genomic DNA was extracted from *E. coli* strain INF191/17/A and sequenced to generate a total of 1,368,224 reads in a 500-cycle run. The total reads from a paired-end dataset (191-17-A_R1.fastq and 191-17-A_R2.fastq) have resulted in 329,238,355 total bases (Table 1). The pre-processed of raw reads including trimming adapter sequences, low-quality and short reads, resulting 46.9% of clean readings. *De novo* assembly of the clean reads was performed and generated 314 contigs with a total size of 5.12 Mbp. Scaffolding resulted in 74 scaffolds with the longest scaffold is 2,520,446 and N50 scaffold length of 1,733,129 bases (Table 2). The average coverage of assembled sequence is 66x with 50.59% of G+C content. Using PGAP, a total of 4987 coding sequences (CDS), four ribosomal RNA, and 66 transfer RNA (Table 3) were predicted. Fur-

**Table 1**
Statistics of the raw and clean reads data including forward (191-17-A_R1.fastq) and reverse (191-17-A_R2.fastq) reads.

| 191-17-A | R1 | R2 | Total |
|---|---|---|---|
| Total Raw Reads | 684,112 | 684,112 | 1,368,224 |
| Total Raw Reads Bases | 164,465,730 | 164,772,625 | 329,238,355 |
| Total Clean Reads | 320,871 | 320,871 | 641,742 |
| Total Clean Reads Bases | 54,470,383 | 40,781,248 | 95,251,631 |
| Clean Reads (%) | 46.90 | 46.90 | 46.90 |

**Table 2**
The statistics of the assembled draft genome of *E. coli* strain INF191/17/A.

| Attributes | Value |
|---|---|
| Number of scaffolds | 74 |
| Total size of scaffolds | 5,136,548 |
| Longest scaffold | 2,520,446 |
| Shortest scaffold | 204 |
| Number of scaffolds > 1 K nt | 51 (68.9%) |
| Number of scaffolds > 10 K nt | 21 (28.4%) |
| Number of scaffolds > 100 K nt | 3 (4.1%) |
| Number of scaffolds > 1 M nt | 2 (2.7%) |
| Number of scaffolds > 10 M nt | 0 (0.0%) |
| Mean scaffold size | 69,413 |
| Median scaffold size | 2736 |
| N50 scaffold length | 1,733,129 |
| L50 scaffold count | 2 |

**Table 3**
The annotation of draft genome of *E. coli* INF191/17/A.

| Attributes | Value |
|---|---|
| Total number of genes | 5062 |
| Number of coding sequences | 4987 |
| Number of genes (coding) | 4736 |
| Total number of RNAs | 75 |
| Number of rRNAs | 4 |
| Number of tRNAs | 66 |
| Number of ncRNAs | 5 |
| Number of pseudogenes | 251 |

**Table 4**
Antimicrobial resistance genes and their corresponding antibiotics detected in the *E. coli* INF191/17/A.

| AMR gene | Description | Resistance |
|---|---|---|
| *mdf(A)* | Multidrug transporter MdfA | Fluoroquinolone, Aminoglycoside, Tetracycline, Macrolide, Rifamycin, Phenicol |
| *aph(3'')-Ib* | Aminoglycoside resistance protein B | Streptomycin |
| *aac(3)-IId* | Aminoglycoside-(3)-N-acetyl-transferase (aacC2) gene | Apramycin, Gentamicin, Tobramycin, Dibekacin, Netilmicin, Sisomicin |
| *aph(6)-Id* | Inosamine-phosphate amidinotransferase | Streptomycin |
| *aadA5* | Streptomycin and spectinomycin resistance aminoglycoside adenyltransferase | Spectinomycin, Streptomycin |
| *tet(A)* | Trimethoprim resistant dihydrofolate reductase | Doxycycline, Tetracycline |
| *mph(A)* | Macrolide 2′-phosphotransferase I | Erythromycin, Azithromycin, Spiramycin, Telithromycin |
| *sitABCD* | Periplasmic binding protein (sitA), ATP-binding component (sitB), inner membrane component (sitC), inner membrane component (sitD) | Hydrogen peroxide |
| *blaTEM-1B* | Bet-lactamase TEM-1 | Amoxicillin, Ampicillin, Cephalothin, Piperacillin, Ticarcillin |
| *blaCTX-M-27* | Beta-lactamase CTX-M-27 | Amoxicillin, Ampicillin, Aztreonam, Cefepime, Cefotaxime, Ceftazidime, Ceftriaxone, Piperacillin, Ticarcillin |
| *sul2* | Dihydropteroate synthase type-2 | Sulfamethoxazole |
| *sul1* | Dihydropteroate synthase type-1 | Sulfamethoxazole |
| *dfrA17* | Dihydrofolate reductase | Trimethoprim |
| *qacE* | Quaternary ammonium compound-resistance protein QacE | Benzylkonium chloride, Ethidium bromide, Chlorhexidine, Cetylpyridinium chloride |

thermore, ResFinder predicted that *E. coli* INF191/17/A will develop fourteen antibiotic resistance genes (Table 4).

## 2. Experimental Design, Materials and Methods

### 2.1. Sample Collection and Isolation of ESBL E. coli Strain INF191/17/A

*E. coli* strain INF191/17/A was isolated from a 45-year-old male patient who was suffering from a high fever at a local hospital. In brief, the sample was cultured in the Bactec 9240 blood culture system (Becton, Dickinson, USA) before proceeding with the biochemical testing and gram staining [3]. The ESBL screening and disk confirmation tests were measured according to Clinical and Laboratory Standards Institute (CLSI) [4]. The 16S rRNA sequences for this strain were validated using specific primers of *E. coli* [5]. Then, the PCR was conducted using ESBL-primers for the confirmation of ESBL-type [1].

### 2.2. DNA Isolation, Genome Sequencing, Assembly, and Annotation

Genomic DNA was isolated using NucleoSpin tissue DNA, RNA, and protein purification kit according to manufacturer's instructions (Macherey-Nagel). The purified DNA was processed using Nextera XT DNA library preparation kit following the manufacturer's instructions (Illumina, USA). A whole-genome sequence was performed using the Miseq platform (Illumina, USA) (2 × 251 bp). The adapter trimming, quality trimming, contaminant filtering and read length filtering were performed using BBDuk (BBTools version 36) (http://jgi.doe.gov/data-and-tools/bbtools/). The low-quality bases (<Q30) and short reads (<50 bp) were trimmed to produce

clean reads with a high quality read dataset. The clean reads were assembled *de novo* using SPAdes v3.9.0 [6] to obtain contigs. These assembled contigs were subjected to scaffolding against the closest reference genomes [3] to produce a draft genome using Medusa (Multi-Draft based Scaffolder) software [7]. The genome was annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v4.10 [8].

### *2.3. Antimicrobial Resistant Genes Analysis*

ResFinder (v4.1) [9] was used to screen for antimicrobial resistance genes. The assembled genome was searched against the curated *Escherichia coli* database using the default parameters. The prediction of the genes was confirmed if the assembled sequence had at least 95% nucleotide matching identity and 80% coverage with candidate genes in the database.

## Ethics Statement

The study protocol was approved by the ethics committee of the Universiti Sains Malaysia (USM/JEPeM/20030152).

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

## CRediT Author Statement

**Nik Siti Hanifah Nik Ahmad:** Software, Formal analysis, Writing – review & editing, Funding acquisition; **Khor Bee Yin:** Conceptualization, Software, Formal analysis, Data curation, Writing – original draft; **Nik Yusnoraini Yusof:** Conceptualization, Software, Methodology, Resources, Writing – review & editing, Supervision, Funding acquisition.

## Acknowledgment

This work was supported by a USM Short Term grant (304.CIPPM.6315337). We would like to thank the Hospital Universiti Sains Malaysia and Department of Microbiology and Parasitology, School of Medical Sciences, Universiti Sains Malaysia (USM), for providing the isolate.

## References

[1] E.J. Alyamani, A.M. Khiyami, R.Y. Booq, et al., The occurrence of ESBL-producing *Escherichia coli* carrying aminoglycoside resistance genes in urinary tract infections in Saudi Arabia, Ann. Clin. Microbiol. Antimicrob. 16 (1) (2017), doi:10.1186/s12941-016-0177-6.

[2] H.M. Kim, S. Jeon, O. Chung, J.H. Jun, H.S. Kim, A. Blazyte, H.Y. Lee, Y. Yu, Y.S. Cho, D.M. Bolser, J. Bhak, Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing, Gigascience 10 (3) (2021) giab014, doi:10.1093/gigascience/giab014.

[3] W.M.W. Ratmaazila, M.M. Azlan, N.H. Hassan, I. Aziah, N.H. Samsurizal, N.Y. Yusof, Draft genome sequence of the extended-spectrum $\beta$-lactamase-producing *Escherichia coli* isolate INF13/18/A, recovered from Kelantan, Malaysia, Microbiol. Resour. Announc. 9 (33) (2020), doi:10.1128/MRA.01497-19 e01497-19.

[4] R. Sari, P. Apridamayanti, I.D. Puspita, Sensitivity of *Escherichia coli* bacteria towards antibiotics in patient with diabetic foot ulcer, Pharm. Sci. Res. 5 (2018) 19–24, doi:10.7454/psr.v5i1.3649.

[5] S.A. Al-Jamei, A.Y. Albsoul, F.G. Bakri, A.G. Al-Bakri, Extended spectrum beta-lactamase-producing *E. coli* in urinary tract infections: a two-center, cross-sectional study of prevalence, genotypes and risk factors in Amaan, Jordan, J. Infect. Public Health 12 (2019) 21–25, doi:10.1016/j.jiph.2018.07.011.8.

[6] A. Bankevich, N. Sergey, A. Dmitry, A.G. Alexey, D. Mikhail, S.K. Alexander, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (5) (2012) 455–477, doi:10.1089/cmb.2012.0021.
[7] E. Bosi, B. Donati, M. Galardini, S. Brunetti, M.F. Sagot, P. Lió, et al., MeDuSa: a multi-draft based scaffolder, Bioinformatics 31 (15) (2015) 2443–2451, doi:10.1093/bioinformatics/btv171.
[8] W. Li, K.R. O'Neill, D.H. Haft, M. DiCuccio, V. Chetvernin, A. Badretdin, G. Coulouris, F. Chitsaz, M.K. Derbyshire, A.S. Durkin, N.R. Gonzales, M. Gwadz, C.J. Lanczycki, J.S. Song, N. Thanki, J. Wang, R.A. Yamashita, M. Yang, C. Zheng, A. Marchler-Bauer, F. Thibaud-Nissen, RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation, Nucl. Acids Res. 49 (D1) (2021) D1020–D1028, doi:10.1093/nar/gkaa1105.
[9] V. Bortolaia, R.F. Kaas, E. Ruppe, M.C. Roberts, S. Schwarz, V. Cattoir, A. Philippon, R.L. Allesoe, A.R. Rebelo, A.R. Florensa, L. Fagelhauer, T. Chakraborty, B. Neumann, G. Werner, J.K. Bender, K. Stingl, M. Nguyen, J. Coppens, B.B. Xavier, S. Malhotra-Kumar, H. Westh, M. Pinholt, M.F. Anjum, N.A. Duggett, I. Kempf, S. NykÃ¤senoja, S. Olkkola, K. Wieczorek, A. Amaro, L. Clemente, J.S. Losch, C. Ragimbeau, O. Lund, F.M. Aarestrup, ResFinder 4.0 for predictions of phenotypes from genotypes, J. Antimicrob. Chemother. 75 (12) (2020) 3491–3500.