Data in Brief 38 (2021) 107311

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Whole genome sequencing data of *Klebsiella aerogenes* isolated from agricultural soil of Haryana, India

Avantika Mann, Shikha Malik, J.S. Rana, Kiran Nehra*

Department of Biotechnology, Deenbandhu Chottu Ram University of Science and Technology, Murthal, Sonipat, Haryana 131039, India

ARTICLE INFO

Article history:
Received 7 May 2021
Revised 1 July 2021
Accepted 18 August 2021
Available online 24 August 2021

Keywords:
Whole genome sequencing
Antibiotic resistance
Antibiotic resistance genes
Genomic data
Genome annotation

ABSTRACT

Klebsiella aerogenes, is a Gram-negative bacterium, which was previously known as *Enterobacter aerogenes*. It is present in all environments such as water, soil, air and hospitals; and is an opportunistic pathogen that causes several types of infections. As compared to other clinically important pathogens included in the ESKAPE category (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter species*), the pangenome and population structure of *Klebsiella aerogenes* is still poorly understood. For the present study, the bacterial sample was isolated from agricultural soils of Haryana, India. With an aim to identify the occurrence of multi-drug resistance genes in the agricultural field soil bacterial isolate, whole genome sequencing (WGS) of the bacteria was performed; and the antibiotic resistance causing genes, along with the genes responsible for other major functions of the cell; and the different Single Nuceotide Polymorphisms (SNPs) and Insertions and deletions (InDels) were identified. The data presented in this manuscript can be reused by researchers as a reference for determining the antibiotic resistance genes that could be present in different bacterial isolates, and it would also help in determination of functions of various other genes present in other genomes of *Klebsiella* species.

* Corresponding author.
E-mail addresses: nehrakiran@gmail.com, kirannehra.bt@dcrustm.org (K. Nehra).

https://doi.org/10.1016/j.dib.2021.107311

Specifications Table

Subject	Microbial antibiotic resistance, Genomics & Genome annotation, Biotechnology & Molecular biology
Specific subject area	Monitoring antibiotic resistance genes in a soil bacterial isolate using next generation sequencing
Type of data	Table(s) Figure(s) Excel sheet(s)
How data were acquired	For Whole Genome Sequencing: Illumina Next Seq Softwares used: Trimmomatic, BWA (Burrows Wheeler transform), SAM tools, BED tools, Antibiotic resistance genes database (ARDB), RASTtk server, SEED server
Data format	Raw Analyzed
Parameters for data collection	The soil samples for isolation of antibiotic resistant bacteria were collected from Haryana state in India. The state is located between 27°39′ to 30°35′ N latitude and between 74°28′ and 77°36′ E longitude in India; and its total geographical area is 4.42 m ha, among which 86.2% area is used for cultivation. The Haryana state in India consists of 22 districts [1] and the soil samples were collected from different agricultural farms of each district. The agricultural farms under this study were divided into two major categories: (i) organic farms (with application of cattle manure), and (ii) inorganic farms (with conventional farm practices including application of chemical fertilizers). During sample collection, it was taken under consideration that at least three samples from both types of fields from each district were collected. Hence, the study involved a total of 132 soil samples representing six samples from two types of farms (organic and inorganic), and a total of three samples from each farm were collected i.e. six samples from each district. The sampling pattern was unbiased random sampling. The soil samples were collected in order to analyze the antibiotic resistance profile of the bacterial isolates present in the cultivated agricultural soils in Haryana State of India; however, more particularly to analyze the over-all genetic information of the ESKAPE isolates (*Klebsiella species* being one of the most prominent organism of this category) present in the agricultural field soils [2].
Description of data collection	The debris was removed and an approximate amount of 10g debris-free agricultural field soil was collected in sterile tubes from six-inches below the top layer. The collected soil samples were then transported to the laboratory and stored at 4°C. The bacteria isolated from these agricultural field soils were subjected to antibiotic susceptibility screening tests. The isolates showing resistance to a set of four commonly used antibiotics (penicillin, streptomycin, tetracycline, and erythromycin) were subjected to further molecular studies. The isolate exhibiting maximum resistance to all the four antibiotics, and identified as *Klebsiella aerogenes*, was finally selected for detailed genetic analysis using whole genome sequencing.
Data source location	Institution: Deenbandhu Chhotu Ram University of Science and Technology City/Town/Region: Murthal, Sonipat, Haryana Country: India GPS co-ordinates: 29.06°N, 76.08°E
Data accessibility	Analyzed data are provided in this report. Raw data are deposited on public repository. Repository name: Klebsiella aerogenes strain G3_AM chromosome deposited in NCBI Data identification number: CP072327 Direct URL to data: https://www.ncbi.nlm.nih.gov/nuccore/CP072327 Database link: Bio Project: PRJNA715319, Bio Sample: SAMN18346029

Value of the Data

- This data is useful for the scientific community as it provides insights into the genome of a pathogenic bacterial isolate, mainly the type of genes present in it at particular positions. This data has its uniqueness, however, this genome could be used as a reference sequence to determine the presence of antibiotic resistance genes in an unknown or a new isolate whose genomic information is yet to be determined.
- The presented data is useful as it was generated specifically to determine the presence of multi- drug-resistant genes present in the bacterial isolates prevalent in agricultural field soils. This data provides information about such genes, including- the name of the gene, the drug class to which they infer resistance to, their mechanism of action and sequence data; which could be used further to analyse resistance genes in other isolates.
- Researchers involved with the work related to genomics data could also benefit from this data. Also, the researchers who are working in a wet lab, and who are interested in confirming the presence of a particular multi-drug-resistant gene in their isolate can use this data in order to design primers for their sequence.

1. Data Description

The data in the present study is a part of a larger study where in the authors are interested in evaluating the effect of application of cattle manure used in the organic farms on the prevalence of the antibiotic-resistance-genes (ant^r) harbouring microorganisms in such soils. The rationale behind this study is the fact that most of the farm animals (cattle, pigs, poultry etc.) are being given antibiotic doses at continuous intervals (either to treat or prevent diseases in them, or, simply to increase their produce). Since antibiotics are not completely digested by these animals, they get excreted in their faecal matter; which in turn when used as fertilizers in organic farms, increases their presence in the soil environment. This also leads to a higher possibility of making the soil environment favourable for better survival of microorganisms having multiple antibiotic resistance genes. However, the present manuscript focuses on the data generated as a result of WGS of a MDR bacteria isolated from such soils.

This article consists of raw and analysed data. The data provided in this article shares the information of complete genome sequencing conducted using Illumina Nextseq, and also several other details, such as the presence of multi-drug-resistance genes, SNPs and InDels. The bacterial isolate which exhibited a high resistance to a set of four commonly used antibiotics was selected for whole genome sequencing, in order to determine the genes responsible for the occurrence of resistance in them. The sequence of this isolate was further analysed using bioinformatics tool to determine the presence of different types of genes and to find out their functions.

Fig. 1 represents the pathway followed while carrying-out the full experiment. First of all, raw reads were obtained, and then with the help of trimmomatic tool [3], these raw reads were filtered and High Quality (HQ) reads were obtained. With the help of Burrows Wheeler transform (BWA) tool [4], mapping of the contigs to a reference genome (*Klebsiella aerogenes* KCTC 2190) was carried out. The Sequence Alignment or Map formation was done using SAM tools mpileup [5], and consensus sequence was obtained. GC content was identified and by using Bedtools [6], gene identification was carried out. The sequence was subjected to annotation and then submitted to NCBI portal. The antibiotic resistance genes were identified using ARDB database [7].

The sequenced raw data was processed to obtain high quality clean reads using Trimmomatic (v 0.38) to remove adapter sequences, ambiguous reads (reads with unknown nucleotides- "N" larger than 5%), and low-quality sequences (reads with more than 10% quality threshold (QV) $<$ 20 phred score). Parameters considered for filtration included: (a) adapter trimming; (b) minimum length threshold of 100 bp; (c) sliding window trimming of 10 bp, cutting once if the average quality within the window fell below a threshold of 20; (d) leading: cutting bases off

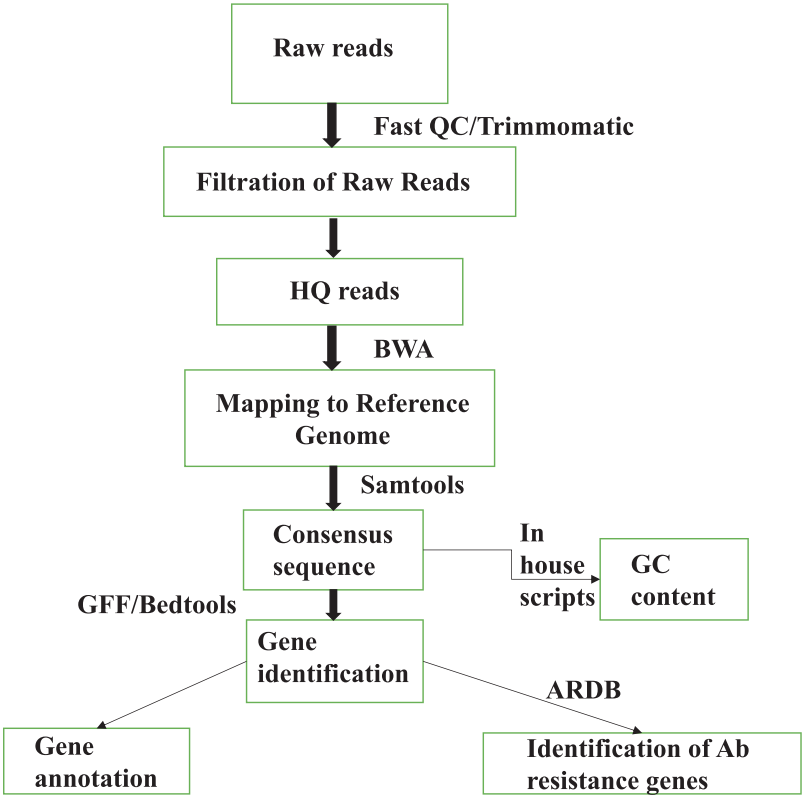

Fig. 1. Schematic representation of the steps involved in the whole genome sequencing and bioinformatic analysis of *Klebsiella aerogenes* genome.

Table 1
High quality data statistics of *Klebsiella aerogenes* genome.

Sample Name/Consensus sequence	Klebsiella aerogenes
Number of paired end (PE) reads	5,071,737
Total number of bases	1,505,053,235
Length of the consensus sequence	5.27 Mb
GC content	50.64%
Mapping percentage	91.04%
Genome coverage	90.47%
Number of sequence	1
Reference organism	*Klebsiella aerogenes* KCTC 2190
Total genome length of the reference genome	5.28 Mb
Reference genome link	https://www.ncbi.nlm.nih.gov/genome/3417?genome_assembly_id=173101

the start of a read, if below a threshold quality of 20; (e) trailing: cutting bases off the end of a read, if below a threshold quality of 20. The high quality reads of the sample were aligned to the reference genome using BWA MEM (version 0.7.17). Respective consensus sequence was extracted using SAM tools mpileup.

Table 1 shows that the genome of the isolate "*Klebsiella aerogenes*" consists of 5,071,737 paired end reads, and the chromosomal DNA has 1,505,053,235 number of bases. The length of consensus sequence was found to be 5.27 Mb, with a GC content of 50.64%. After mapping to

Table 2
Identification of SNP variants, InDels, and annotation summary.

Sample		*Klebsiella aerogenes*
Total number of SNPs		53,604
SNP count	#Homozygous SNP	52,092
	#Heterozygous SNP	1,512
Annotated SNP count	#Genic SNP	48,654
	#Intergenic SNP	4,950
Total number of InDels		222
InDels Count	#Homozygous InDels	195
	#Heterozygous InDels	27
Annotated InDels Count	#Genic InDels	50
	#IntergenicInDels	172

the reference genome, *Klebsiella aerogenes* KCTC 2190 (5.28 Mb length), the mapping percentage was found to be 91.04% when 90.47% genome was covered.

Table 2 shows the details of the Single Nucleotide polymorphisms (SNPs) and Insertion-Deletions (InDels) obtained upon comparison of the genome of *Klebsiella aerogenes* isolated in the present study with the reference genome. The mpileup utility of SAM tools (v 0.1.18) was used to identify the SNPs and InDels from the sorted BAM file of each of the mappings (Excel sheet 1). The variants were filtered based on a minimum read depth of 15, and a quality threshold of 25. The identified variants were annotated using BED tools, an intersect tool. A total of 53,604 SNPs were obtained, among which 52,092 were observed to be homozygous SNPs, whereas 1,512 were heterozygous SNPs. Among the annotated SNPs, 48,654 were genic SNPs and 4,950 were intergenic SNPs. The total number of InDels identified was 222, among which 195 were homozygous InDels, and 27 were heterozygous InDels. Among the annotated InDels, 50 were genic InDels and 172 were intergenic InDels.

Fig. 2 shows the presence of all the important genes found in this genome. The pie chart was developed by rapid annotation using subsystem technology toolkit (RASTtk) [8] (rast.nmpdr.org) and has been viewed in SEED server (https://pubseed.theseed.org/) [8]. A total number of 4749 coding sequences were observed to be present in the subsystem. The bar chart on the left-hand side of the figure shows the subsystems coverage in percentage (blue bar corresponds to the percentage of proteins not in the subsystem, and the green one represents the percentage of proteins in the subsystem). The pie chart to the right shows the distribution of 27 most abundant subsystem categories out of a total of 2118 subsystem categories.

Excel sheet 1 consists of four sheets, the first and the second sheet contain details about all SNPs, about all InDels, respectively. The third excel sheet details information about all genic SNPs, and the fourth one details information about all genic InDels.

Excel sheet 2 represents the details of all the genes monitored by RASTtk server including the type of gene, location of gene, start and stop codon, type of strand, function of the gene, aliases, figfam, evidence codes, the nucleotide and the amino acid sequence.

Excel sheet 3 shows the presence of identified Antibiotic Resistance genes using ARDB. The genes identified in the sample were screened against ARDB (https://ardb.cbcb.umd.edu/blast/genome.shtml), and a total of 28 antibiotic resistant genes were identified (Excel sheet 3).

2. Experimental Design, Materials and Methods

2.1. Soil sample collection sites

Haryana state is considered as the land of agriculture and most of its area is used for agricultural purposes. Soil samples were collected from different agricultural fields representing each

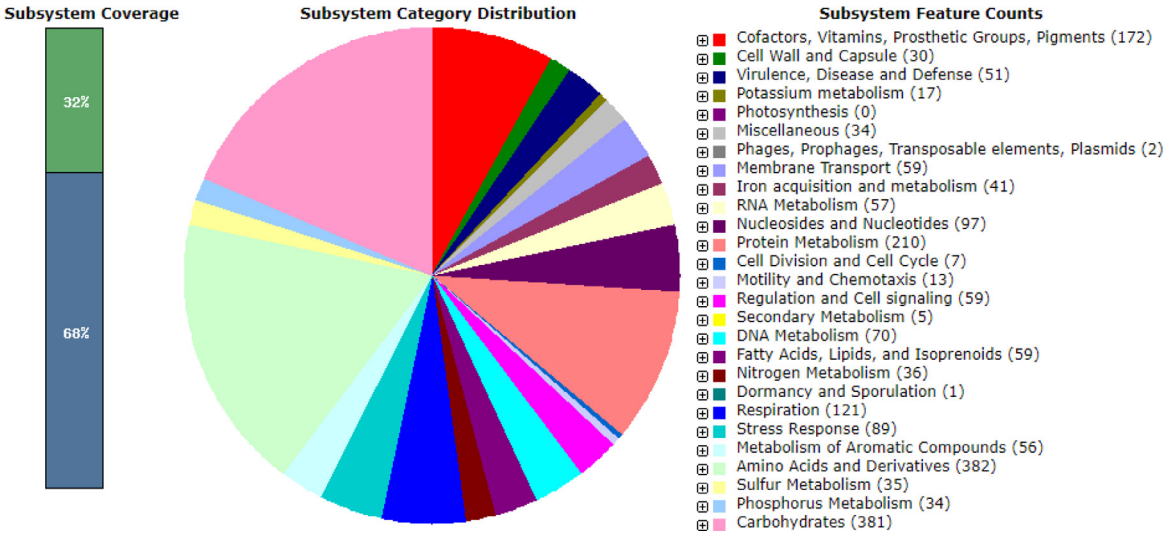

Fig 2. Genes connected to subsystems and their distribution in different categories.

of its 22 districts. From each district two types of agricultural farms were selected i.e. organic farm and inorganic farm; and from each type of farm, three soil samples were collected from different locations [9]. In this way, the study consisted of a total of 132 samples, consisting of six samples from each district (three of each type).

2.2. Isolation and purification of bacterial isolates

Bacterial population was isolated from the collected soil samples by using standard serial dilution plate technique. From the collected soil, 5g soil sample was mixed in 45ml distilled water and volume was made up to 50ml (to prepare soil stock solution) and vortexed to mix properly. An aliquot of 1ml of the soil stock solution was taken and mixed in 9 ml of sterile physiological saline and vortexed to mix properly. The sample was then serially diluted up to 10^{-8} dilution, and an aliquot of 100 µl each from 10^{-4}, 10^{-5}, $10^{-6,}$ 10^{-7}, 10^{-8} dilutions of the soil samples was used for spread plating onto nutrient agar plates, which were then subjected to incubation for 24–48h. Any plates with fungal growth were discarded. However, the nutrient agar plates having colonies between 30–100 were finally selected [10]. The colonies exhibiting differences in morphology were picked-up and streaked on nutrient agar plates (quadrant streaking method), and re-streaked again several times until pure colonies were obtained. Pure cultures were stored in agar slants at 4°C for subsequent studies. A total of 135 pure bacteria were isolated from the collected soil samples. Gram's staining was performed to classify the isolated bacteria into Gram-positive and Gram-negative strain.

2.3. Antibiotic susceptibility screening of the soil bacterial isolates

All 135 pure bacterial colonies were streaked in triplicate on nutrient agar medium, which was supplemented with 5µg/ml of four antibiotics (penicillin, streptomycin, tetracycline and erythromycin), separately on different petri plates. The plates were incubated at room temperature from 24 to 72h. This approach for evaluation of antibiotic susceptibility of microorganisms has been reported earlier as well [11]. Out of these 135 pure colonies, 122 exhibited resistance to at least one antibiotic (72, 4, 24 and 22 bacteria resistant to penicillin, tetracycline, streptomycin and erythromycin, respectively). These 122 strains were further subjected to antibiotic resistance screening tests at higher concentrations of antibiotics i.e. 10, 15, 20, 25µg/ml. Bacterial strains showing maximum growth up to 25µg/ml for at least 3 or all 4 antibiotics were picked and were stored at 4°C on antibiotic agar plates, and in nutrient broth supplemented with 10% glycerol at −80°C for further use in characterization studies. Among a total of 13 such isolates, ten isolates were resistant to at least three antibiotics, and only three isolates exhibited resistance to all four antibiotics. These 13 isolates were selected for further molecular characterization studies.

2.4. Molecular characterization and identification of bacterial strain

The selected 13 isolates were subjected to16SrRNA based molecular identification. The results revealed that nine isolates belonged to different *Bacillus* strains, one isolate was identified as *Ochrobactrum intermedium*, and three isolates which were resistant to all four selected antibiotics (up to 25µg/ml conc. for penicillin, streptomycin, erythromycin and 15µg/ml for tetracycline) were *Klebsiella aerogenes*. Based on the resistance profile of these three *Klebsiella* isolates, one isolate, i.e. G3, was selected for whole genome sequence analysis.

2.5. Selection of the soil bacterial isolate for whole genome sequencing

The bacterial isolate G3 which was found to exhibit maximum resistance to all the four antibiotics, was selected for further WGS and annotation. The whole genome of this bacterium (identified as *Klebsiella aerogenes*) was further subjected to sequencing using Illumina NextSeq technique and results were analyzed using several bioinformatics tools. The whole genome sequence of this isolate was annotated by mapping to a reference *Klebsiella aerogenes* (KCTC 2190) genome to study the differences and similarities between the bacteria isolated in the present study and similar such isolates already reported in the NCBI database.

Ethics Statement

None.

CRediT Author Statement

Avantika Mann: Conceptualization, methodology, investigation, resources, data curation, writing original draft, visualization; **Kiran Nehra:** Conceptualization, validation, review and editing, supervision; **J.S. Rana:** Validation, review and editing, supervision; **Shikha Malik:** Conceptualization

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgement

The authors wish to acknowledge National Project Implementation Unit (NPIU), a unit of Ministry of Human Resource Development, Government of India, for the financial assistantship awarded to Ms. Avantika Mann through TEQIP-III project at Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonipat, Haryana.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107311.

References

[1] Districts, (n.d.). https://haryana.gov.in/districts/.
[2] S. Zhao, G.H. Tyson, Y. Chen, C. Li, S. Mukherjee, S. Young, C. Lam, J.P. Folster, J.M. Whichard, P.F. McDermott, Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in Campylobacter spp, Appl. Environ. Microbiol. 82 (2016) 459–466, doi:10.1128/AEM.02873-15.
[3] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.
[4] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 1754–1760, doi:10.1093/bioinformatics/btp324.
[5] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, doi:10.1093/bioinformatics/btp352.

[6] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842, doi:10.1093/bioinformatics/btq033.
[7] B. Liu, M. Pop, ARDB—antibiotic resistance genes database, Nucl. Acids Res. 37 (2009) D443–D447, doi:10.1093/nar/gkn656.
[8] R.K. Aziz, D. Bartels, A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, E.M. Glass, M. Kubal, F. Meyer, G.J. Olsen, R. Olson, A.L. Osterman, R.A. Overbeek, L.K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G.D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, The RAST server: rapid annotations using subsystems technology, BMC Genom. 9 (2008) 1–15, doi:10.1186/1471-2164-9-75.
[9] S. Ghosh, T.M. LaPara, The effects of subtherapeutic antibiotic use in farm animals on the proliferation and persistence of antibiotic resistance among soil bacteria, ISME J. 1 (2007) 191–203, doi:10.1038/ismej.2007.31.
[10] T.H.E. Procedures, 11: Bacterial Numbers, (15819) 1–7. https://hyp.is/go?url=https%3A%2F%2Fbio.libretexts.org%2FBookshelves%2FAncillary_Materials%2FLaboratory_Experiments%2FMicrobiology_Labs%2FMicrobiology_Labs_I%2F11%253A_Bacterial_Numbers&group=__world__.
[11] N. Esiobu, L. Armenta, J. Ike, Antibiotic resistance in soil and water environments, Int. J. Environ. Health Res. 12 (2002) 133–144, doi:10.1080/09603120220129292.