**Title: Shallow sequencing (Shall-seq) procedure: A method for determining the genetic lineages and detecting antimicrobial resistance genes of antimicrobial-resistant bacteria using minimum Nanopore sequencing data**

**Authors:** Nobuyoshi Yagi[1#], Nanase Miyagi[2], Ayumi Uechi[3] and Itaru Hirai[2]

**Author Affiliations:**
[1]Laboratory of Clinical physiology, School of Health Sciences, Faculty of Medicine, University of the Ryukyus, Okinawa, Japan
[2]Laboratory of Microbiology, School of Health Sciences, Faculty of Medicine, University of the Ryukyus, Okinawa, Japan
[3]Division of Clinical Laboratory and Blood Transfusion, University of the Ryukyus Hospital, Okinawa, Japan.

**Co-author details including name, affiliation, email address**:
Nanase Miyagi; Laboratory of Microbiology, School of Health Sciences, Faculty of Medicine, University of the Ryukyus, 207 Uehara, Nishihara, Okinawa 903-0215, JAPAN, miyagina12@gmail.com
Ayumi Uechi; Division of Clinical Laboratory and Blood Transfusion, University of the Ryukyus Hospital, 207 Uehara, Nishihara, Okinawa 903-0215, JAPAN, au03193@med.u-ryukyu.ac.jp
Itaru Hirai; Laboratory of Microbiology, School of Health Sciences, Faculty of Medicine, University of the Ryukyus, 207 Uehara, Nishihara, Okinawa 903-0215, JAPAN, hiraii@med.u-ryukyu.ac.jp

**Corresponding author:** Nobuyoshi Yagi, Ph. D., Laboratory of Clinical physiology, School of health Sciences, Faculty of Medicine, University of the Ryukyus., 207 Uehara, Nishihara, Okinawa 903-0215, JAPAN, TEL: +81-98-895-3331, E-mail address: yagin@med.u-ryukyu.ac.jp

**Authorship statement:** NY and IH conceived and designed the study. NY, NM and AU collected data. IH supervised this study. All authors contributed to the writing of the manuscript. All authors approved the final manuscript.

**Abstract (199 words)**

Antimicrobial-resistant bacteria could cause nosocomial infections and outbreaks in healthcare facilities. Phylogenetic analyses based on whole-genome sequencing (WGS) could become the gold-standard method for understanding the route of antimicrobial-resistant bacterial spreading. However, generally, the WGS needs to analyze much amount of data. Therefore, sufficient resources such as budget and data analysis system are needed and it is a burden for introduction of the WGS in the routine clinical examination of pathogenic bacterial isolates. In this study, we used *Escherichia coli* as a model and evaluated whether determination of the genetic background and detection of antimicrobial-resistance genes of 29 *E. coli* clinical isolates were achieved by searching databases using sequence reads output by the Nanopore sequencer as the search keys. Consequently, only 66.2 MB data was sufficient to search for a genome sequence with ≥90% range of coverage rate. Importantly, AMR genes and plasmid replicon types were also detected with minimum data, and the detected AMR genes and phenotypes of the *E. coli* isolates did not present any discrepancy. Taken together, this shallow sequencing (Shall-seq) procedure consists of “shallow of coverage” sequencing using the Nanopore sequencer and data search using minimum data could be used to analyze bacterial isolates cost-effectively.

**Keywords; Antimicrobial-resistant bacteria; Whole-genome sequencing; genetic characterization**

## Introduction

The emergence of various antimicrobial-resistant bacteria has become an important and serious public health issues of this century [1, 2]. Many previous studies have indicated the spread of antimicrobial-resistant bacteria such as extended-spectrum beta-lactamase (ESBL)-producing Enterobacterales and colistin-resistant Enterobacteriaceae in in many countries and regions [3, 4].

Nosocomial infections caused by antimicrobial-resistant bacteria and antimicrobial-resistant pathogenic bacterial outbreaks occur occasionally. To understand the distribution mechanisms of these causative bacterial isolates, genetic relatedness among the bacterial isolates was evaluated using the conventional gold-standard method, that is, pulsed-field gel electrophoresis. Bacterial whole-genome sequencing (WGS) has recently become a major method used to study the molecular epidemiology of antimicrobial-resistant bacteria. WGS can not only determine the antimicrobial-resistant bacterial species and genetic lineage, but also detect antimicrobial resistance (AMR) genes that are contained in plasmid(s) [5, 6]. The core-genome multilocus sequence typing (MLST) and core-genome single nucleotide polymorphism (SNP) analyses have been applied to study the molecular epidemiology of antimicrobial-resistant bacteria [7,

8]. Therefore, if the introduction of WGS into routine clinical examinations could enable to monitor causative bacterial isolates, which might enable rapid management of nosocomial infections and outbreaks. However, high cost is a bottleneck to utilize the WGS for a routine clinical examination. Thus, a novel labor-, power-, and cost-effective analytical method is required to address AMR.

Meanwhile, bacterial whole-genome sequences have been accumulating rapidly in public genetic databases, such as GenBank and RefSeq [9, 10]. Therefore, many genome sequence records of bacterial species commonly detected in healthcare-associated facilities, such as *E. coli* and ESKAPE, have possibly been deposited in the public databases.

We aimed to establish a novel labor-, power-, and cost-effective method for analyzing antimicrobial-resistant bacteria. Accordingly, we determined the minimum amount of sequence data required to select bacterial genome sequences from the public database and detect AMR gene sequences in the bacterial isolate genomes. In this study, we propose a shallow sequencing (Shall-seq) procedure that determines bacterial genetic lineages and detects the AMR genes harbored by antimicrobial-resistant bacteria.

## Materials and Methods

**Bacterial isolates**

A total of 29 cefotaxime-resistant *E. coli* isolates obtained from two hospitals in Okinawa Prefecture, Japan, between 2013 and 2023 were used in this study.

***E. coli* reference genome sequences**

The complete *E. coli* genome sequences were retrieved from nucleotide database, "RefSeq" [10], using the keyword "complete sequence," and filtered by taxonomy, "*Escherichia coli*." Short sequences such as less than 4,000,000 bp and inaccurate sequences containing "N" nucleotide were removed, and the remaining 3,186 *E. coli* genome sequences were used as the reference genome sequences (**Table S1**).

Phylogenetic groups and sequence types (STs) of the 3,186 *E. coli* genome sequences were determined using EzClermont [11] and Staramr [12] with the MLST [13] database, respectively (**Table S1**).

**Library preparation for the Nanopore sequencing**

The whole DNA of the *E. coli* isolates was purified using a Monarch DNA purification kit (New England Biolabs Japan Inc., Tokyo, Japan). DNA concentrations were measured using Qubit™ 4 fluorometer with Qubit™ 1× dsDNA High Sensitivity (HS) kit (Thermo Fisher Scientific K.K., Tokyo, Japan). The library was prepared using 250 ng extracted DNA with the Ligation Sequencing Kit SQK-LSK109 (Oxford

Nanopore Technologies, Oxford, UK) and the native barcoding expansion kit EXP-NBD104 (Oxford Nanopore Technologies) according to the manufacturer's protocol. The library containing eight barcode-ligated DNA was loaded onto an R9.4.1 Flongle flow cell (Oxford Nanopore Technologies).

**Reference genome mapping**

Sequence reads were aligned with 3,186 *E. coli* genome sequences using Minimap2 [14]. The coverage range was calculated using SAMtools with a coverage command [15], as listed in **Table S2**. The assembled genome sequence of *the E. coli* isolate was constructed using Unicycler [16].

**Collection of AMR genes and plasmid replicons**

To detect AMR genes and plasmid replicon types from sequence reads, AMR genes and plasmid replicon types were collected from the AMRFinderPlus [17] and PlasmidFinder [18] nucleotide databases, respectively. Two or more >90% homologous sequences were clustered into one sequence using VSEARCH [19].

**Construction of phylogenetic tree by core-genome analysis**

Among the 3,186 reference genome sequences, 260 were selected based on the phylogenetic group and ST. After adding **NZ_CP051719.1** and the draft N75 genome sequence, we phylogenetically analyzed these 260 genomes based on the core-genome

using Roary [20]. The phylogenetic tree was drawn using R and RStudio with “ggtree” package.

**Shall-seq procedure**

**(1) Shall-seq:** The libraries for Nanopore sequencing were prepared using 250 ng DNA. Up to 16 samples could be analyzed simultaneously using a single Flongle flow cell. A total of 250 ng (approximately 16–30 ng per bacterial isolate) of mixed DNA samples from up to 16 bacterial isolates were used for library preparation.

**(2) Indicator genome sequence (IGS) selection:** The resulting Nanopore sequence reads with >9 quality score and >500 bp read length were used to select an IGS. The selected sequence reads obtained from each bacterial isolate were aligned with each *E. coli* genome sequence, as described above. In this study, the range of coverage was calculated using SAMtools, and the genome sequence showing the highest range of coverage rate, which should be >90%, was selected as an IGS representing the phylogenetic group and ST.

**(3) AMR gene and plasmid replicon type detection:** The obtained sequence reads were aligned with sequences from the collection of AMR genes and plasmid replicons using KMA [21]. The sequences of AMR genes and plasmid replicon types with ≥60% “Template coverage” and ≥90% “Template identity” were counted.

## Result and Discussion

### Evaluating the data amount required to determine the genetic lineage of antimicrobial-resistant bacteria

To determine the minimum data required to identify the most identical *E. coli* genome sequence from the 3,186 *E. coli* reference genome sequences, we performed whole-genome Nanopore sequencing of *E. coli* clinical isolate N75. The entire sequence read output (962.8 MB, equivalent to 471.4 M bp) from a single Flongle flow cell were divided 1/2 to 1/512 with the seqkit [22] using command listed in **Table S2**. The divided sequence reads were aligned with each *E. coli* genome sequence. The range of coverage rates decreased depending on the sequence data amount (**Fig. 1**). However, the range of coverage rate for 66.4 MB (32.6 M bp) was not substantially different from that for >66.4 MB sequence data. Moreover, the *E. coli* isolate N75 sequence data selected **NZ_CP051719.1** from the 3,186 *E. coli* reference genome sequences, but even if only 8.5 MB (4.2 Mbp) sequence data were used for the alignment, the same **NZ_CP051719.1** was selected.

We then evaluated the location of **NZ_CP051719.1** and the assembled *E. coli* isolate N75 genome sequence in the phylogenetic tree constructed using the 260 *E. coli* genome sequences. **NZ_CP051719.1** and the assembled *E. coli* isolate N75 genome

sequence were in very close branches of the phylogenetic group B2 (**Fig. 2**). Consistently, BLAST indicated a cover % of between **NZ_CP051719.1** and the assembled genome sequence of the *E. coli* isolate N75 was more than 99%.

These results suggested that 66.4 MB (32.6 M bp) was the minimum data amount required to identify reference genome sequences homologous to clinical isolate genome sequences from the database.

**Examination of the 29 antimicrobial-resistant *E. coli* clinical isolates by the Shall-seq procedure**

Our results suggest that 16 samples can be analyzed simultaneously using a single Flongle flow cell. However, we simultaneously analyzed eight samples in this study, considering the fluctuations in the amount of sequence data obtained.

For further evaluation, we analyzed total of 29 *E. coli* clinical isolates obtained from hospitals in Okinawa Prefecture, Japan. The obtained sequence reads were used to select IGSs and detect the *AMR* gene and plasmid replicon types (**Table 1**). Consequently, 11 IGSs were selected for this study (**Table 1**). Among them, 11 (81.8%), 1 (9.1%), and 1 (9.1%) IGSs were classified into phylogenetic groups B2, D, and F, respectively. A total of 174 AMR genes were detected in the 29 *E. coli* clinical isolates. On average, six *AMR* genes were detected in each antimicrobial-resistant *E. coli* clinical

isolate, and *mph*(A) was the most prevalent gene (**Table 1**). Importantly, the phenotypes of the 29 antimicrobial-resistant *E. coli* isolates and *AMR* genes detected in their sequence data did not have major discrepancies.

Interestingly, **NZ_AP022362.1**, **NZ_LT632320.1**, and **NZ_CP082831.1**, were selected based on sequence data obtained from 11 (37.9%), 4 (13.8%), and 3 (10.3%) of the 29 examined *E. coli* clinical isolates, respectively. Since the number and species of the detected *AMR* genes in these clinical isolates were not completely matched, it is not highly possible that the three IGSs were selected due to antimicrobial-resistant bacterial contamination and sample mixing. This means that the Shall-seq procedure can classify antimicrobial-resistant bacterial isolates, at least at the ST level.

**Conclusion**

The selection of the IGS homologous to clinical isolates from public nucleotide database using the Shall-seq procedure could be a cost-effective WGS, which might facilitate introducing WGS as a routine clinical examination. Consequently, it might facilitate rapid managements and interventions in the case of nosocomial infections and outbreaks caused by antimicrobial-resistant bacteria.

Although a high error rate is considered a disadvantage of Nanopore sequencing, its accuracy has recently improved [23, 24]. This improved accuracy

encourages using the Shall-seq procedure to monitor antimicrobial-resistant bacteria in niches other than healthcare-associated facilities, such as communities and natural environments.

**Data availability**

The sequence reads used in this study are available under BioProject accession number **PRJDB17430**.

**Declaration of competing interest**

None.

**Funding**

This work was supported by Institute for Fermentation, Osaka (IFO) (Grant number **G-2022-1-003**). In addition, this work was partly supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number **23K09672** and **24K20197**.

**Reference**

[1] Antimicrobial Resistance C. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet 2022;399:629-55. https://doi.org/10.1016/S0140-6736(21)02724-0.

[2] Chereau F, Opatowski L, Tourdjman M, Vong S. Risk assessment for antibiotic resistance in South East Asia. BMJ 2017;358:j3393. https://doi.org/10.1136/bmj.j3393.

[3] Yamaguchi T, Kawahara R, Hamamoto K, Hirai I, Khong DT, Nguyen TN, et al. High Prevalence of Colistin-Resistant Escherichia coli with Chromosomally

Carried mcr-1 in Healthy Residents in Vietnam. mSphere 2020;5. https://doi.org/10.1128/mSphere.00117-20.

[4] Rosantia S, Higa T, Yagi N, Tokunaga T, Higa S, Yakabi Y, et al. Characterization of CTX-M-type-extended-spectrum beta-lactamase (ESBL)-producing Enterobacteriaceae isolated from Indonesian undergraduate medical students of a university in Surabaya, Indonesia. J Infect Chemother 2020;26:575-81. https://doi.org/10.1016/j.jiac.2020.01.010.

[5] Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra B, et al. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. J Antimicrob Chemother 2018;73:1121-37. https://doi.org/10.1093/jac/dkx488.

[6] Gao H, Liu Y, Wang R, Wang Q, Jin L, Wang H. The transferability and evolution of NDM-1 and KPC-2 co-producing Klebsiella pneumoniae from clinical settings. EBioMedicine 2020;51:102599. https://doi.org/10.1016/j.ebiom.2019.102599.

[7] Mesa V, Delannoy J, Ferraris L, Diancourt L, Mazuet C, Barbut F, et al. Core-genome multilocus sequence typing and core-SNP analysis of Clostridium neonatale strains isolated in different spatio-temporal settings. Microbiol Spectr 2023;11:e0276623. https://doi.org/10.1128/spectrum.02766-23.

[8] Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. Int J Food Microbiol 2018;274:1-11. https://doi.org/10.1016/j.ijfoodmicro.2018.02.023.

[9] Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2016;44:D67-72. https://doi.org/10.1093/nar/gkv1276.

[10] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733-45. https://doi.org/10.1093/nar/gkv1189.

[11] Waters NR, Abram F, Brennan F, Holmes A, Pritchard L. Easy phylotyping of Escherichia coli via the EzClermont web app and command-line tool. Access Microbiol 2020;2:acmi000143. https://doi.org/10.1099/acmi.0.000143.

[12] Bharat A, Petkau A, Avery BP, Chen JC, Folster JP, Carson CA, et al. Correlation between Phenotypic and In Silico Detection of Antimicrobial Resistance in Salmonella enterica in Canada Using Staramr. Microorganisms 2022;10. https://doi.org/10.3390/microorganisms10020292.

[13] Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 2018;3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.

[14] Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics 2021;37:4572-4. https://doi.org/10.1093/bioinformatics/btab705.

[15] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10. https://doi.org/10.1093/gigascience/giab008.

[16] Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

[17] Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 2021;11:12728. https://doi.org/10.1038/s41598-021-91456-0.

[18] Carattoli A, Hasman H. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). Methods Mol Biol 2020;2075:285-94. https://doi.org/10.1007/978-1-4939-9877-7_20.

[19] Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016;4:e2584. https://doi.org/10.7717/peerj.2584.

[20] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691-3. https://doi.org/10.1093/bioinformatics/btv421.

[21] Clausen P, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics 2018;19:307. https://doi.org/10.1186/s12859-018-2336-6.

[22] Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One 2016;11:e0163962. https://doi.org/10.1371/journal.pone.0163962.

[23] van Dijk EL, Naquin D, Gorrichon K, Jaszczyszyn Y, Ouazahrou R, Thermes C, et al. Genomics in the long-read sequencing era. Trends Genet 2023;39:649-71. https://doi.org/10.1016/j.tig.2023.04.006.

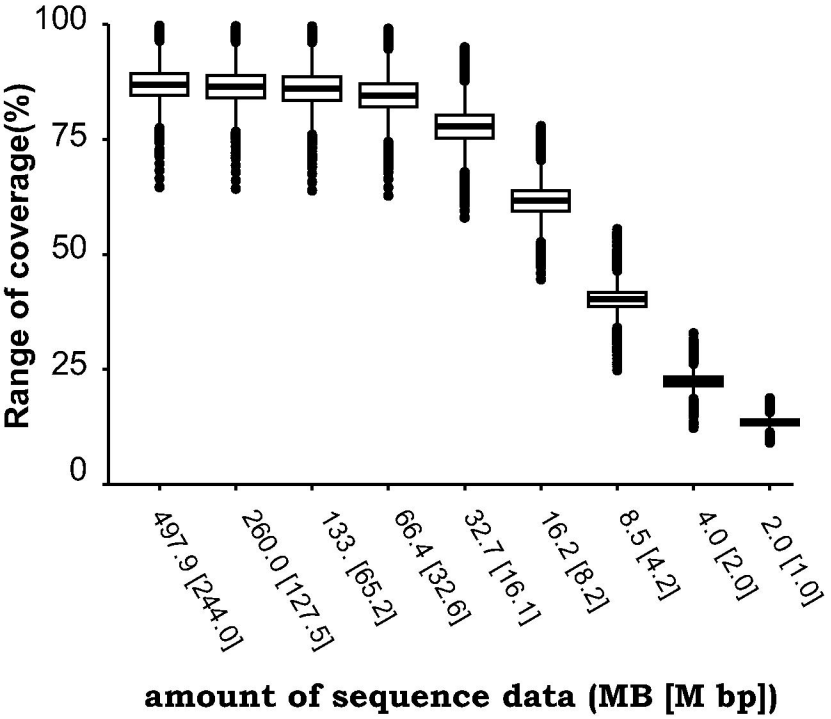
[24] Wagner GE, Dabernig-Heinz J, Lipp M, Cabal A, Simantzik J, Kohl M, et al. Real-Time Nanopore Q20+ Sequencing Enables Extremely Fast and Accurate Core Genome MLST Typing and Democratizes Access to High-Resolution

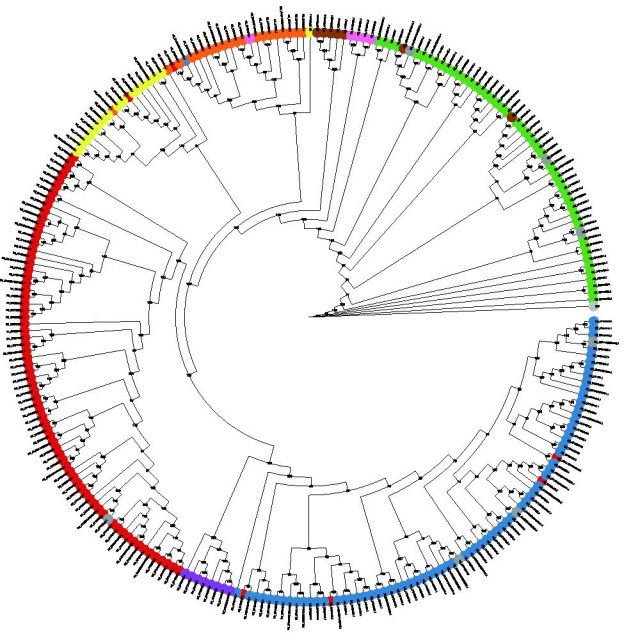
Bacterial Pathogen Surveillance. J Clin Microbiol 2023;61:e0163122. https://doi.org/10.1128/jcm.01631-22.

**Figure Legends**

**Fig. 1 The effect of analysis using less sequence data.** Boxplots show the impact of sequence data reduction on the coverage range analyzed by reference mapping against 3,186 *Escherichia coli* genomes individually. X-axis represents the sequence data used for alignment with 3,186 reference genome sequences, and y-axis represents the coverage range against *E. coli* reference genome sequences.

**Fig. 2 The phylogenetic tree constructed by core-genome analysis.** To compare the genome sequences of **NZ_CP051719.1** and **N75**, the phylogenetic tree having 262 *Escherichia coli* genome sequences including **NZ_CP051719.1**, the assembled genome sequence of the *E. coli* clinical isolate **N75**, and 260 reference genome sequences selected based on the differences of their phylogenetic group and ST among the 3,186 *E. coli* reference genome sequences was constructed using Roary.


100
75
50
25
0
Range of coverage(%)
497.9 [244.0]
260.0 [127.5]
133. [65.2]
66.4 [32.6]
32.7 [16.1]
16.2 [8.2]
8.5 [4.2]
4.0 [2.0]
2.0 [1.0]
amount of sequence data (MB [M bp])



**Phylogenetic Group**

A B1 B2 C D

E F G U NA