Title: A rapid method to determine the genetic lineage of *Escherichia coli* using open reading frame composition in the shallow sequencing

Authors: Nobuyoshi Yagi[1], Ayumi Uechi[2] and Itaru Hirai[3]

Author Affiliations:
[1]Laboratory of Clinical physiology, School of Health Sciences, Faculty of Medicine, University of the Ryukyus, 207 Uehara, Nishihara, Okinawa 903-0215, JAPAN
[2]Division of Clinical Laboratory and Blood Transfusion, University of the Ryukyus Hospital, 207 Uehara, Nishihara, Okinawa 903-0215, JAPAN
[3]Laboratory of Microbiology, School of Health Sciences, Faculty of Medicine, University of the Ryukyus, 207 Uehara, Nishihara, Okinawa 903-0215, JAPAN

Co-author details including name, affiliation, email address:
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 and AU collected data. IH supervised this study. All authors contributed to the writing of the manuscript. All authors approved the final manuscript.

Abstract (192 words)

Determining the genetic background of bacterial isolates and evaluating the genetic relatedness among these isolates in a short time period are important to identify the spreading route(s) in cases of healthcare-associated infections and outbreaks caused by antimicrobial-resistant bacteria. Previously, we proposed a shallow sequencing (Shall-seq) procedure to determine the genetic backgrounds of clinical isolates using the minimum amount of sequence data. However, it took a longer time, such as longer than 10 h, to determine the genetic background of one clinical isolate. In this study, we developed a search procedure using open reading frame (ORF) composition to select the reference genome sequence with the highest matching ratio (>90%), the indicator genome sequence (IGS), for the examined bacterial isolate. Consequently, IGSs were selected for 28 (96.6%) of the examined 29 isolates and selection was performed within 30 min for each bacterial isolate. More importantly, the comparison of IGSs indicated that the IGSs, determined by ORF composition, of the examined bacterial isolates were closely related to the genome sequences determined using the Shall-seq procedure. Taken together, these results suggest that our newly developed search procedure can quickly determine the genetic background of bacterial isolates.

Keywords: Antimicrobial-resistant bacteria, Whole-genome sequencing, genetic characterization, Open reading frame

Introduction

Healthcare-associated infections caused by antimicrobial-resistant bacteria and outbreaks of pathogenic and/or antimicrobial-resistant bacteria occasionally occur. In these cases, determining the genetic background of the causative bacterial isolates and evaluating their genetic relatedness are essential to trace the spreading route of these causative bacteria and to take effective countermeasures.

In our previous study, we established a Shall-seq procedure consisting of nanopore sequencing and data analysis with a minimum amount of sequence data of bacterial isolates[1]. The Shall-seq procedure could be used to select a bacterial genome sequence that was genetically homologous to the sequenced isolate (indicator genome sequence, IGS) from the nucleotide database and detect antimicrobial-resistance genes matched with the isolate's antimicrobial resistance phenotype. Although the Shall-seq procedure would provide information related to bacterial genetic backgrounds, it took a longer time, such as longer than over 10 h, to select IGS from the nucleotide database. This could be a disadvantage for applying the Shall-seq procedure in clinical practice.

Molecular epidemiology classifying bacterial isolates by patterning the presence or absence of ORFs in their genomes has been proposed [2]. This implies that a homology search using ORF composition could shorten the time required to select IGS for the

examined bacterial isolates. In this study, we established a data analysis procedure to select IGS using the minimum sequence data of bacterial isolates obtained from the shall-seq procedure in a shorter time.

Materials and Methods

Sequence data

Nanopore sequence data from 29 *Escherichia coli* clinical isolates were obtained from a previous study. The characteristics of the 29 *E. coli* clinical isolates were previously described[1].

E. coli reference genome sequences

A total of 3,186 *E. coli* genome sequences were used, as previously described[1].

Constructing an *E. coli* ORF database

We annotated 3,186 reference sequences by Prokka [3] using the commands listed in **Table S1**. Consequently, a total of 145,742 sequences annotated as "gene" were collected from the 3,186 reference genome sequences by PanTA [4]. The collected "gene" sequences with sequence identity of 90% or more were sorted into the same ORF using the VSEARCH [5]. Finally, 33,606 "gene" sequences were identified and a database containing these "gene" sequences was designated as an "*E. coli* ORF database".

Constructing the ORF datasets of 3,186 reference genome sequences

The ORFs of the 3,186 reference genome sequences were detected using the BLASTN and ORF databases under conditions of over 60% query coverage and over 90% identity. A dataset containing the ORF composition of the reference genome sequences was created and designated as the ORF dataset of the reference genome sequences.

Selecting indicator genome sequences using the *E. coli* ORF database

The ORFs in the sequence reads of the clinical isolates were detected under conditions where query coverage and percent identity were greater than 90% using KMA [6] and the ORF database. We calculated the logical conjunction between the ORF composition of the clinical isolate and each of the 3,186 reference genome sequences. The matching ratio was then calculated:

$$\text{Matching ratio}=\frac{\text{a total of logical } conjunction}{\text{the number of ORF detected from clinical isolates}}*100$$

Among the 3,186 reference genome sequences, the reference genome sequence(s) with the highest matching ratio (>90%) was designated as the IGS for clinical isolates.

Comparison between IGS_{SS} and IGS_{ORF} by constructing phylogenetic tree and drawing a dot plot

To evaluate the difference between the IGS_{ORF} and IGS_{SS}, we compared representative 281 *E. coli* genome sequences, containing 23 IGSs determined by ORF

composition and Shall-seq procedure. We phylogenetically analyzed these 281 genomes based on the core-genome using Roary[7]. then, phylogenetic tree was constructed by IQTREE2[8] and drawn by R and RStudio. Moreover, we compared IGS_{SS} and IGS_{ORF} and drew dot plots using D-GENIES [9] for clinical isolates wherein the IGS_{ORF} was different from the IGS_{SS}.

Result and Discussion

The ORF composition of the examined *E. coli* clinical isolates was used to search for the genome sequence(s) with the highest matching rate, exceeding 90%, from the ORF dataset of reference genome sequences. Suitable genome sequences were obtained for 28 of the 29 clinical isolates examined. The obtained sequences were designated as indicator genome sequences and IGS_{ORF} of the examined clinical isolates (**Table 1**).

The IGS_{ORF}s differed in 15 (53.6%) of the 28 clinical isolates from the IGS designated by the shall-seq procedure (IGS_{SS}), even though there was no discrepancy between them in the phylogenetic groups (**Table 1**). Therefore, we evaluated the difference between IGS_{ORF} and IGS_{SS} by phylogenetic tree (**Fig. 1**). The IGSs determined by ORF composition was closely related to the genome sequences determined using the Shall-seq procedure (**Table 1, Fig. 1**). Moreover, we compared the IGS_{ORF} and IGS_{SS} by using D-genies (**Fig. 2**). Even if there were inversions in certain region of these genomes,

the concordance rates between IGS_{ORF} and IGS_{SS} was 92.0% or higher (**Table 1**). Taken together, these results indicated that there was only small difference between IGS_{ORF} and IGS_{SS}.

It took 10 h per isolate to determine IGS_{SS} when the shall-seq procedure was performed. The application of the ORF composition reduced the time from 10 h to 30 min per isolate. It was suggested that IGS_{ORF} would be obtained within one hour after obtaining sequence reads of the Nanopore sequencing even considering the time required for quality check of output sequence reads (approximately 15 min).

As mentioned above, our data indicated that the IGS_{ORF} was obtained by a combination of the shall-seq procedure and a search using the ORF composition. Many bacterial isolates should be analyzed and genetic relatedness among them should be evaluated as soon as possible when healthcare-associated infections or outbreaks of pathogenic bacteria occur. In this regard, the genetic relatedness among bacterial isolates can be evaluated and visualized by drawing a phylogenetic tree using IGS_{ORF}. Until recently, higher error rates were reported as a disadvantage of nanopore sequencing[10]. However, the accuracy of nanopore sequencing has recently improved [11]. Therefore, the improved accuracy of nanopore sequencing might encourage the application of the Shall-Seq procedure and searches using ORF composition.

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**). This work was partly supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number **23K09672** and **24K20197**.

Reference

[1] Yagi N, Miyagi N, Uechi A, Hirai I. 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. bioRxiv 2024:2024.06.25.600723. https://doi.org/10.1101/2024.06.25.600723.

[2] Hayashi K, Doi Y, Suzuki M. Rapid phylogenetic analysis using open reading frame content patterns acquired by Oxford nanopore sequencing. J Appl Microbiol 2022;133:3699-707. https://doi.org/10.1111/jam.15807.

[3] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068-9. https://doi.org/10.1093/bioinformatics/btu153.

[4] Le DQ, Nguyen TA, Nguyen TT, Nguyen SH, Do VH, Nguyen CH, et al. PanTA: An ultra-fast method for constructing large and growing microbial pangenomes. bioRxiv 2023:2023.07.03.547471. https://doi.org/10.1101/2023.07.03.547471.

[5] 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.

[6] 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.

[7] 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.

[8] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 2020;37:1530-4. https://doi.org/10.1093/molbev/msaa015.

[9] Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 2018;6:e4958. https://doi.org/10.7717/peerj.4958.

[10] 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.

[11] 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.

Table 1. Indicator genome sequences of the clinical isolates.

Clinical Isolate	Number of ORFs detected in the clinical isolate	Number of Co-detected ORFs	Ratio of Co-detected ORFs (%)	Indicator genome sequece deisgnated by ORF composition		Indicator genome sequece deisgnated by Shall-seq procedure		Concordance rate (%)
				Acccession No.	Phylogenetic group	Acccession No.	Phylogenetic group	
22-847[(A)]	4951	4630	93.5	NZ_CP116480.1	B1	NZ_CP010235.1	B1	92.0
23-165[(B)]	5212	4837	92.8	NZ_AP022161.1	F	NZ_CP103468.1	F	96.7
23-389[(C)]	4979	4816	96.7	NZ_CP013835.1	B2	NZ_CP082831.1	B2	97.6
	4979	4816	96.7	NZ_HG941718.1	B2	NZ_CP082831.1	B2	98.0
23-395	4985	4755	95.4	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
23-396	4969	4697	94.5	NZ_CP097360.1	B2	NZ_CP097360.1	B2	-
23-412[(D)]	5105	4898	95.9	NZ_CP023826.1	B2	NZ_AP022362.1	B2	97.9
23-418	5088	4948	97.2	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
23-435[(E)]	5161	4933	95.6	NZ_AP022362.1	B2	NZ_CP082831.1	B2	96.1
23-439	5045	4764	94.4	NZ_CP054343.1	D	NZ_CP054343.1	D	-
23-449[(F)]	4909	4597	93.6	NZ_UFZT01000001.1	B2	NZ_CP048920.1/ NZ_CP051849.1	B2/ B2	99.0/ 99.0
23-475	4659	4301	92.3	NZ_CP014111.1	B2	NZ_CP014111.1	B2	-
23-480	5133	4824	94.0	NZ_CP097360.1	B2	NZ_CP097360.1	B2	-
23-496[(G)]	5321	5083	95.5	NZ_CP103562.1	B2	NZ_CP082831.1	B2	97.6
23-501[(H)]	5124	4920	96.0	NZ_JAOSCG010000001.1	B2	NZ_AP022362.1	B2	98.7
N4	5047	4532	89.8	N.D.	-	NZ_CP051719.1	B2	-
N21	5004	4772	95.4	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
N55	4457	4324	97.0	NZ_OW849064.1	B2	NZ_OW849064.1	B2	-
N57[(I)]	4844	4674	96.5	NZ_JAOSFD010000001.1	B2	NZ_LR130545.1	B2	99.3
N58	4743	4602	97.0	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
N71	4838	4711	97.4	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
N75[(J)]	4450	4060	91.2	NZ_CP083534.1	B2	NZ_CP051719.1	B2	93.8
N137	3867	3741	96.7	NZ_OW849064.1	B2	NZ_OW849064.1	B2	-
N165[(K)]	4982	4788	96.1	NZ_CP080119.1	B2	NZ_OW849064.1	B2	99.1
N211	5059	4941	97.7	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
N214	4971	4878	98.1	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
N222	5065	4901	96.8	NZ_AP022362.1	B2	NZ_AP022362.1	B2	-
N223[(L)]	4944	4761	96.3	NZ_CP080119.2	B2	NZ_CP023853.1	B2	97.5
N247[(D)]	5165	4983	96.5	NZ_CP023826.1	B2	NZ_AP022362.1	B2	97.9
N254[(D)]	5134	4970	96.8	NZ_CP023826.1	B2	NZ_AP022362.1	B2	97.9

superscripts, comparison of two indicator genome sequences are shown in Figure 1. the letter of superscripts links to the figure labels.

N.D., Not Dtermined.

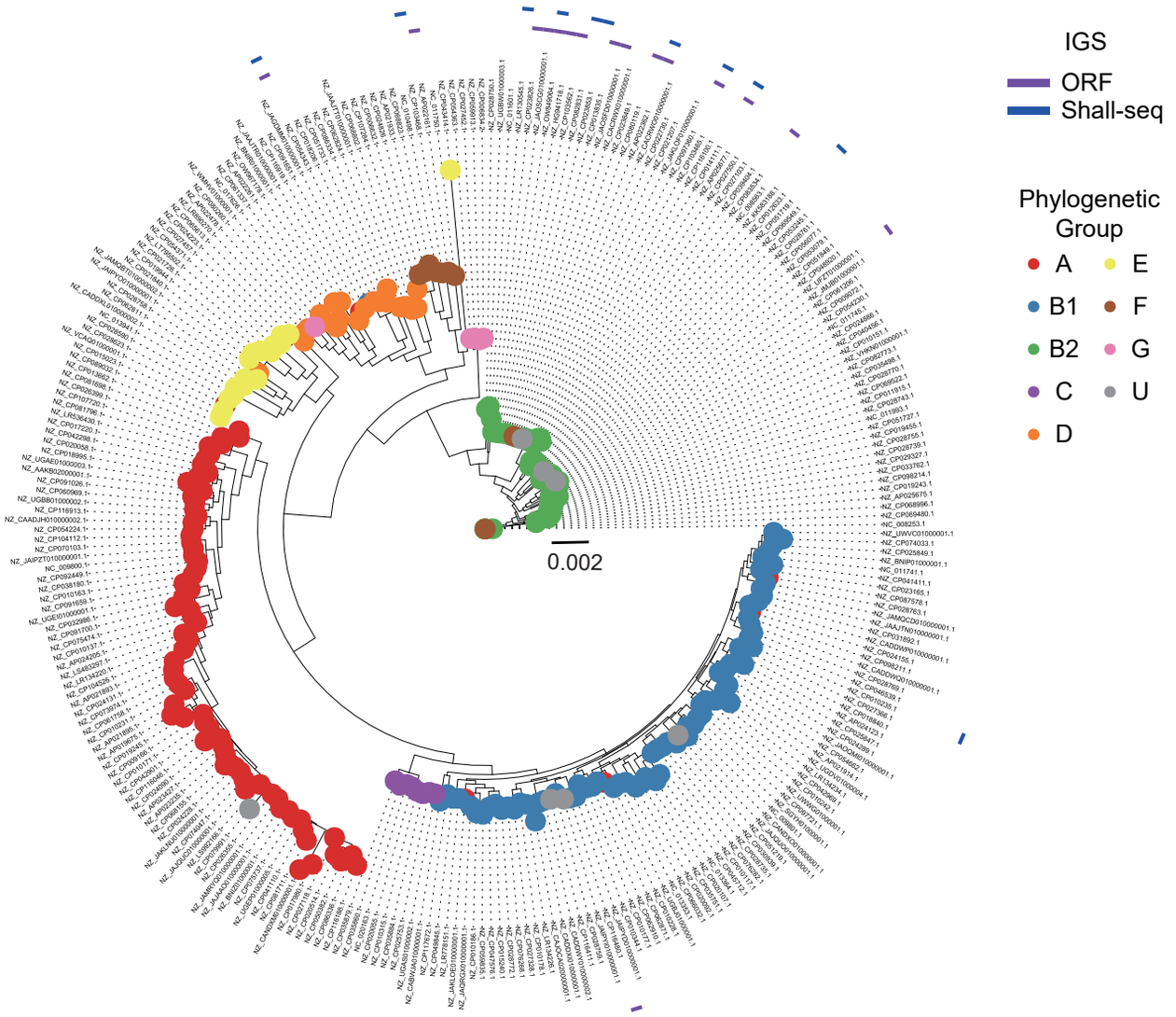
IGS
ORF
Shall-seq
Phylogenetic Group
A
B1
B2
C
D
E
F
G
U
0.002

Fig. 1 The genetic relationships between IGS_{SS} and IGS_{ORF} by core-genome analysis. The phylogenetic tree for 281 *E. coli* genome sequencies which containing a total of 23 *E. coli* genome sequences determined as IGS_{SS} and IGS_{ORF} was constructed by the Roary. The phylogenetic tree was constructed by 1474 genes detected as core gene with the Roary and IQtree2. The circle plot at the end of the branch means the phylogenetic group of each *E. coli* reference genomes. The heat map at the outside of phylogenetic tree means the *E. coli* genomes detected as IGS by ORF composition (purple) and/or Shall-seq procedure (blue).

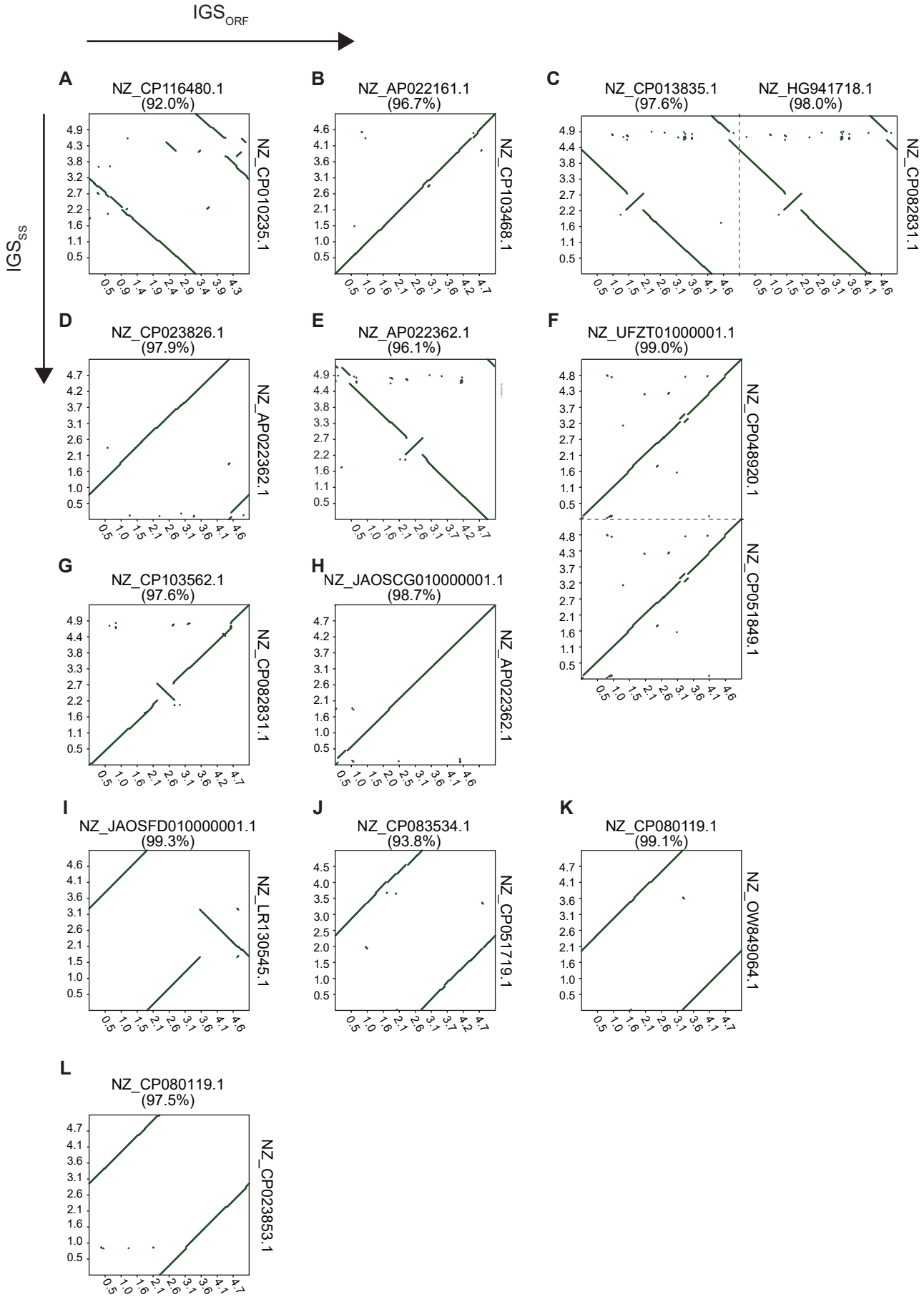
IGS_ORF
IGS_SS
A
NZ_CP116480.1
(92.0%)
NZ_CP010235.1
B
NZ_AP022161.1
(96.7%)
NZ_CP103468.1
C
NZ_CP013835.1
(97.6%)
NZ_HG941718.1
(98.0%)
NZ_CP082831.1
D
NZ_CP023826.1
(97.9%)
NZ_AP022362.1
E
NZ_AP022362.1
(96.1%)
F
NZ_UFZT01000001.1
(99.0%)
NZ_CP048920.1
NZ_CP051849.1
G
NZ_CP103562.1
(97.6%)
NZ_CP082831.1
H
NZ_JAOSCG010000001.1
(98.7%)
NZ_AP022362.1
I
NZ_JAOSFD010000001.1
(99.3%)
NZ_LR130545.1
J
NZ_CP083534.1
(93.8%)
NZ_CP051719.1
K
NZ_CP080119.1
(99.1%)
NZ_OW849064.1
L
NZ_CP080119.1
(97.5%)
NZ_CP023853.1

Fig. 2 Comparison between IGS_{ORF} and IGS_{SS}. In 15 clinical isolates, different reference genome sequences were selected between IGS_{ORF} and IGS_{SS}, and IGS_{ORF}s (horizontal axis) and compared with IGS_{SS}s (vertical axis) by drawing dot plots using D-Genies. Each dot represents > 75% nucleotide identity. Within the plot, diagonal lines indicate homologous regions between the two indicator genomes; regions lacking such diagonals comprise each indicator's genome-specific regions. Figure labels are linked to the superscripts in **Table 1**.