A Thesis for the Degree of Master

Evaluation of hybrid sequencing by comparing DNA digestion fragment pattern of *Xba*I, S1 nuclease in pulsed field gel electrophoresis (PFGE) and in *in silico* analysis of whole genome sequencing (WGS) data.

Sunwoo Lee

Department of Biotechnology

GRADUATE SCHOOL

JEJU NATIONAL UNIVERSITY

Nov, 2022

Evaluation of hybrid sequencing by comparing DNA digestion fragment pattern of *Xba*I, S1 nuclease in pulsed field gel electrophoresis (PFGE) and in *in silico* analysis of whole genome sequencing (WGS) data.

Sunwoo Lee
(Supervised by professor Tatsuya Unno)

A thesis submitted in partial fulfillment of the requirement
for the degree of Master of Science.
2022. 11.
This thesis has been examined and approved by

Sooje Park, Ph.D., College of Natural Sciences, Jeju National University

____

Tatsuya Unno, Ph.D., College of Applied Life Sciences, Jeju National University

____

Jongeun Park, Ph.D., College of Applied Life Sciences, Jeju National University

____

Department of Biotechnology
GRADUATE SCHOOL
JEJU NATIONAL UNIVERSITY

CONTENT

LIST OF FIGURES

LIST OF TABLES

국문 초록

최근 많은 유전체 분석 연구에 차세대 시퀀싱(next generation sequencing, NGS)이 적용되고 있으며 그 중요성에 대해서 보고되고 있다. 미생물의 유전자에 관련된 생명과학 연구에 전장 유전체 시퀀싱(whole genome sequencing, WGS)의 적용이 활발해지고 있으며, 그에 따라 sequencing 기술의 개발과 정확도에도 관심이 커지고 있다. 또한, sequencing 정확도를 높이기 위해서 hybrid sequencing을 하거나 추가 실험을 진행하여 비교, 분석하는 시도도 이루어지고 있다. 본 실험에서는 hybrid sequencing으로 WGS 데이터를 생산했으며, PFGE를 진행하여 그 결과를 비교함으로써 유사도를 확인했다. 펄스 필드 겔 전기영동(Pulse field gel electrophoresis, PFGE)은 *Xba*I 제한효소와 S1 nuclease를 적용하여 molecular typing과 plasmid profiling을 진행했다. *Xba*I 제한효소를 적용한 WGS 데이터와 PFGE 데이터의 molecular typing결과에서 샘플의 수가 적거나 여러 그룹으로 재분류되는 두 가지 그룹을 제외하고 모든 그룹에서 cluster가 유사하게 나타났다. 하지만, plasmid profiling의 경우에는 유사하게 나타나지 않았다. 이는 패턴으로 cluster를 비교할 때에는 만들어질 수 있는 패턴의 다양성, 그룹에 속한 샘플의 수, 그룹 내의 샘플정보의 일치가 결과 값에 영향을 끼친다는 것을 나타낸다. 또한 ANI와 PFGE 패턴의 비교 결과가 유사하게 나타난 샘플 간에 유사한 크기의 contig와 band가 72%로 확인되었다. 즉, 이는 WGS data와 PFGE 데이터 간에 28% 정도의 차이가 나타남을 암시한다. 본 연구를 통한 결과들을 바탕으로 WGS 데이터가 PFGE 패턴과 완벽히 일치하지 않다는 것을 확인하였으며 보다 정확한 비교와 분석을 위해서는 추가적인 연구가 필요하다고 사료된다.

ABSTRACT

Recently, next generation sequencing (NGS) has been applied to many genome analysis studies, and its importance has been reported. The application of whole genome sequencing (WGS) to life science research related to microbial genes is becoming more active, and interest in the development and accuracy of sequencing technology is also growing. Accordingly, to improve sequencing accuracy, attempts are being made to conduct and analyze hybrid sequencing or compare with additional experiments. In this experiment, WGS data was produced by hybrid sequencing, and similarity was confirmed by additionally performing PFGE. Pulse field gel electrophoresis (PFGE) was performed for molecular typing and plasmid profiling by applying *Xba*I restriction enzyme and S1 nuclease. In the molecular typing results of WGS data and PFGE data to which *Xba*I restriction enzyme was applied, clusters were similar in all groups except for two groups where the number of samples was small or reclassified into several groups. However, result of plasmid profiling did not appear similarly. This indicates that when comparing clustering by pattern, the diversity of patterns that can be made, the number of samples belonging to a group, and the matching of sample information in a group affect the result value. In addition, contigs and bands of similar size were identified at 78% among samples with similar ANI and PFGE pattern comparison results. That is, this implies that a difference of about 22% appears between the WGS data and the PFGE data. Based on the results of this study, it was confirmed that the WGS data did not completely match the PFGE results, and additional research is needed for more accurate comparison and analysis.

INTRODUCTION

Due to the development of NGS technology and the increase in demand, the application of WGS to biotechnology is becoming more active and its cost is also becoming cheaper [1]. Accordingly, the applied materials are also diversely applied to plants, fungi, and microorganisms, and the fields of application are also various [2-4]. In the field of precision medicine, WGS is being applied to human genome database construction and individual genome diagnosis analysis to provide personalized medicine [5]. As interest in WGS increases, the types of platforms that can perform WGS (Illumina, Pacific Biosciences, Thermo scientific, Oxford Nanopore Technologies, etc.) are also diversifying, and data production methods (ion torrent, DNA template sequencing, protein membrane sequencing) is also becoming more abundant [1, 6, 7]. Debate on sequencing accuracy is also issued by the increasing number of available platforms and methods that can produce WGS data. Recently, a hybrid sequencing method that assembles short-read sequences and long-read sequences together has been used in the WGS field, enabling more accurate identification of genetic mutations and gene analysis [8]. In the field of microbial research, WGS enables analysis of the entire genome including bacteria typing by analyzing DNA sequences using developed tools [9, 10]. As a method of bacterial typing, there are methods such as WGS data, PFGE, and MLST, and identification of different species of microorganisms, pathogenicity detection through phenotypic comparison, and classification through genome analysis can be implemented [11-13]. In the case of bacterial typing, conventional electrophoresis was used for the first and second generations, and pulsed field gel electrophoresis (PFGE) and DNA sequence analysis were used for the third and fourth generations, respectively [14]. PFGE is a third-generation bacteria typing method that was used as the gold standard for bacteria typing through DNA fingerprinting before sequencing was commercialized [15]. Plasmid profiling through conventional electrophoresis of the first generation allows us to know the pattern and size of plasmid, but DNA molecules of about 40 kb or more move at the same speed regardless of size, so it was difficult to separate and estimate the size of the bands [16, 17]. The disadvantages of conventional electrophoresis are also the same in the second-generation using restriction enzymes and probes, and PFGE, a third-generation technique that changes the direction of current to increase the separation ability of DNA molecules, has been developed [15]. The analysis method through WGS data corresponds to the 4th generation bacteria typing method. Recently, studies on the accuracy of WGS data have been conducted, along with studies on the accuracy of each platform [18]. The differences in accuracy by platform mostly comes from the sequencing methods and processes used by each platform.

Pacific Biosciences has the possibility of detecting unsynthesized bases to the zero-mode waveguide (ZMW), illumina has the possibility that DNA polymerase cannot be applied to all samples in the formation of a bridge-type template, and Oxford nanopore technology has 'computer signal' errors in the process of converting A, T, G, C' into base signals, Thermo scientific are known to cause sequencing errors in the process of reading homopolymers [19, 20]. To correct sequence errors of these WGS data, tools for error correction have been developed and related studies are being conducted [21, 22]. In addition, attempts are being made to improve sequencing accuracy through detecting mutation sequences, applying hybrid assembly that assembles short-read and long-read sequences together and comparing with other experimental results [22-24]. However, even if hybrid sequencing techniques are used, it is not easy to assemble into a perfect genome due to problems such as sequencing accuracy and assembly quality [25]. In this study, we tried to find out correlation and similarity between the two data by comparing the PFGE results and the WGS data. PFGE was performed by referring to previous studies on bacteria typing and plasmid typing using *Xba*I restriction enzyme and S1 nuclease [26, 27]. *Xba*I *Xb* digestion was performed by applying the *Xba*I restriction enzyme sequence ('TCTAGA') to WGS data (*Xba*I-WGS). Contigs not derived from the chromosome (S1-WGS) were extracted by excluding contigs containing the part of 16S rRNA sequence using blast. By comparing average nucleotide identity (ANI) and S1-PFGE results, the relationship between PFGE patterns and ANI results was identified. The results of this study provide insight into the similarity between the PFGE results and the results of the WGS data.

MATERIAL AND METHOD

Sample information

166 strains of E. coli used in this study were provided from a Specialized Bank for Multidrug Resistant Pathogens (Korea Disease Control and Prevention Agency, Drug Resistance Division). Samples (n = 166) were collected from various sites of Korea between 2017 and 2019 (Gangwon-do (n = 26), Gyeonggi-do (n = 13), Gyeongsang-do (n = 27), Seoul (n = 41), and Jeolla-do (n = 52), Chungcheong-do (n = 7)), and originated from various environments (human, animal, barn, hospital, etc.).

Molecular typing and plasmid profiling using PFGE

PFGE of all samples was performed according to the method of CDC PulseNet protocol [28]. The samples were incubated on Muller Hinton agar plates at 37°C for 14-18 hours, and the cultured bacteria were diluted in 1% suspension buffer and adjusted the O.D. value to be 0.8-1. Then, each sample was mixed with 1% SeaKem Gold Agarose (Lonza, 50150) at a ratio of 1:1 to make a plug. After proceeding the lysis and washing process of the plug, plugs were treated with enzyme.

In this experiment, *Xba*I restriction enzyme (Takara, 1093A) and S1 nuclease (Thermo Scientific, EN0321) were used for molecular typing and plasmid profiling. After each plug was treated with the enzyme, it was reacted at 37°C and room temperature for 1 hour. [26, 27]. Enzyme treated plugs were loaded onto 1% SeaKem Gold Agarose and tested with the CHEF-DRII system (Bio-rad, 1703615).

The PFGE conditions were adjusted and proceeded in two methods. The first PFGE method was conducted using a 48.5–1,000 kb ladder (Bio-rad, 1703635) at a voltage of 4.5 V/cm with an initial pulse of 6 seconds and a final pulse of 36 seconds. The second PFGE method was performed using an 8.3-48.5 kb ladder (Bio-rad, 1703707) at a voltage of 6 V/cm with an initial pulse of 1 second and a final pulse of 3 seconds. Additionally, to identify bands smaller than 8.3 kb (500 bp-10 kb), conventional electrophoresis was performed using a plasmid extraction method using a 1 kb ladder (BIONEER, D-1040) [29]. Gel photographs were visualized using a gel documentation system (CANNON, UNOK-8000HS) and a UV Transilluminator (Major Science, MUV21-312). Gel pictures were summarized in Figure S1-A, B, C.

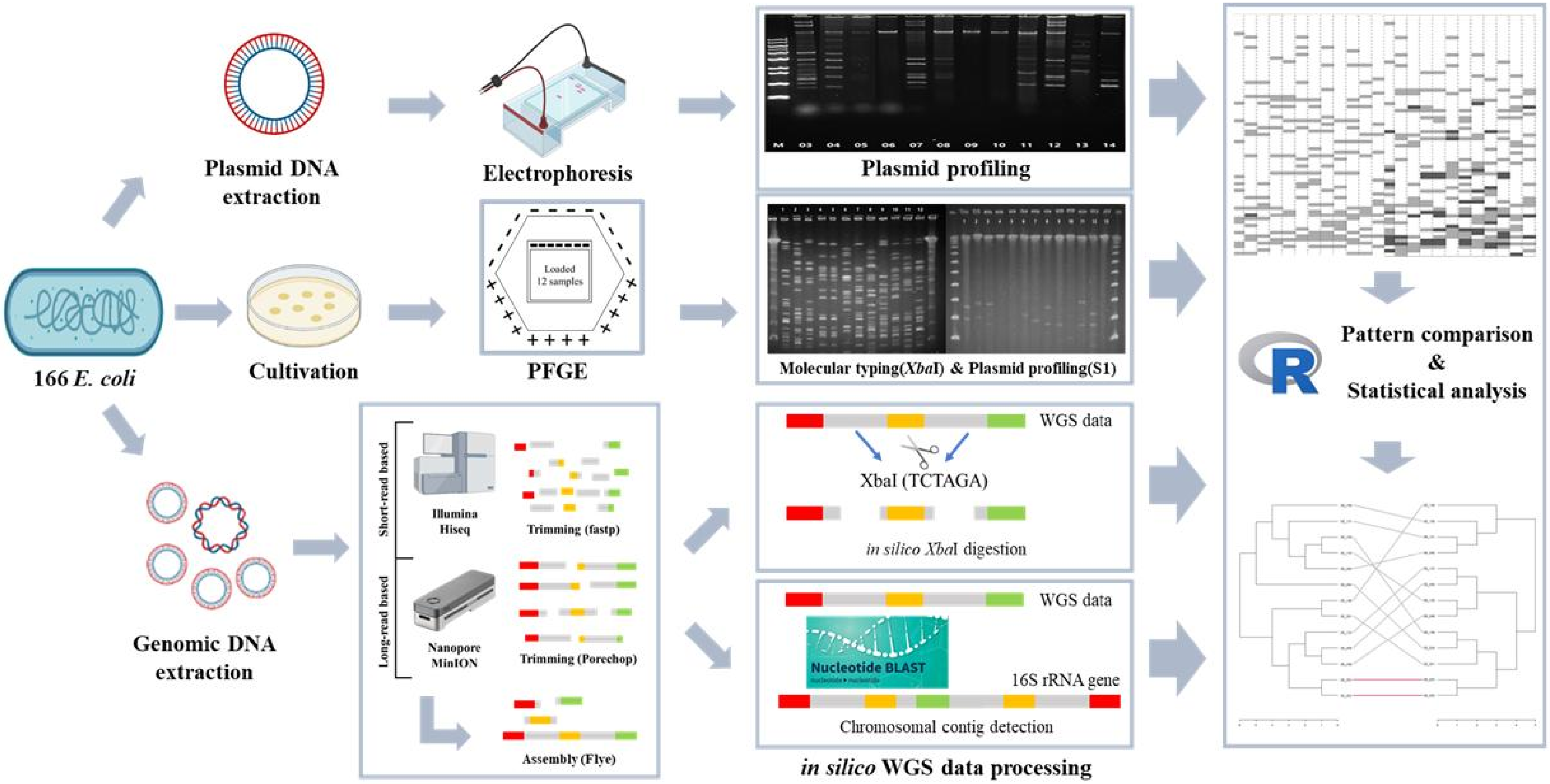


Figure 1. Summary of experimental work flow.

DNA extraction and sequencing

All samples were cultured on MacConkey (Difco™, NJ, USA) agar plates containing ampicillin (32 μg/ml) (Sigma-Aldrich, St Louis, USA) overnight at 37°C for 18-24 h. Single colonies were cultured on Mueller Hinton (Difco™, NJ, USA) agar plates at 37°C overnight for 18-24 h. DNA was extracted from the cultured bacteria using a Genomic DNA extraction kit (BIONEER, K-3032). The concentration and quality of the extracted DNA was measured using a Qubit fluorometer (Invitrogen, CA, USA). As the method of WGS, illumina Hiseq and Nanopore MinION were used. Briefly, the illumina Hiseq library was prepared using the xGen DNA Library Prep EZ Kit (IDTDNA, 10009821) and sent to Macrogen Inc. (Seoul, Korea) for sequencing with HiSeqXten (350bp × 2).

Nanopore MinION library was prepared using Nanopore Native Barcoding Kit (NANOPORE Tech, SQK-NBD112.24), NEBNext FFPE DNA Repair Mix (Biolabs, M6630L), NEBNext® Ultra™ II End Repair/dA-Tailing Module (Bio labs, E7546L), Blunt/TA Ligase Master Mix (Biolabs, M0367L), NEBNext® Quick Ligation Module (Biolabs, E6056L) and NEBNext® Quick Ligation Module (Biolabs, E6056L). Both methods were purified using the HiAccuBead (AccuGene, Incheon, Korea) purification kit during preparation of libraries. For Nanopore MinION sequencing, flow cell R9.4.1 (NANOPORE Tech, FLO-MIN106D) was used, and the prepared library was loaded into the device after checking the number of available pores (>900).

WGS data analysis

Hybrid sequence assembly was performed by a micropipe tool using sequence of both Illumina Hiseq and Oxford Nanopore MinION sequencing [1]. A bash shell script was used for *in silico* restriction enzyme digestion of assembled data (Table S2). Among the assembled contigs, chromosomal contigs were detected by comparing 16S rRNA sequences with blast [31]. The average nucleotide identity (ANI) value was measured using EZBioCloud's OAT (www.ezbiocloud.net/tools/orthoani) [32]. Phylogenetic trees based on homology were generated through the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method and visualized using R.

Patterning of WGS data and Gel photos

Gel photos of *Xba*I-PFGE and S1-PFGE were measured and organized using imageJ [33] and GelAnalyzer 19.1 (www.gelanalyzer.com, Istvan Lazar Jr., PhD and Istvan Lazar Sr., PhD, CSc). In addition, the fragment and contig lengths of *Xba*I-WGS and S1-WGS data were summarized, respectively. Patterns were measured and recorded based on the data of each length measurement.

For the pattern, 0 to 448,625 bp was set as the minimum and maximum length of range. The maximum length (448,625 bp) was divided into 76 sections with units of 6025.5 bp (1.315% of the maximum length). In the section including bands, fragments, and contigs, each number was counted and the corresponding number was entered. Otherwise, ‘0’ was entered. Microsoft Excel (Microsoft Corporation, 2018) was used to divide and organize the range.

Statistical analysis

Statistical analysis of this study was performed using the R programing. The ‘ecodist’ package was used to conducting mental-test and generating correlogram to compare the correlation of PFGE patterns. ‘dendextend’ package was used to analyze the similarity between dendrograms. In all analyses, a p-value lower than 0.05 was considered significant.

RESULT AND DISCUSSION

Comparison of *Xba*I-PFGE results and WGS data

166 E. coli strains were divided into 6 groups (Gangwon-do (GW), Gyeonggi-do (GG), Gyeongsang-do (GS), Seoul (SU), Jeolla-do (JR), Chungcheong-do (CC)) based on site. First, the PFGE result with *Xba*I restriction enzyme applied (*Xba*I-PFGE) and the data obtained by applying *Xba*I digestion to WGS data (*Xba*I-WGS) quantified the contig and fragment length data were patterned. Clustering between samples was confirmed by drawing a dendrogram after clustering with the Euclidean distance of each result.

Since there were limitations in comparing PFGE and WGS data with correlogram, the Mantel-r value and p-value that can confirm the significance of pattens were confirmed (Table 1.). The mantel-r value was 0.25 or more in the four groups of GW, GG, GS, and JR, and lowest value was 0.274 in the GG group and highest value was 0.404 in the GW group. p-value also showed significant data with a value of 0.05 or less in the four groups. However, in the SU group and the CC group, the absolute values of the Mantel-r were low at 0.009 and 0.175, respectively. p-value of each group was 0.561 and 0.237 respectively, higher than 0.05.

The graph was generated to check the correlogram through the mantel-test. As a result of the mantel-test, the change in mantel r value according to the distance between samples was confirmed (Fig. 2-A, B). In the GW, GG, GS, and JR group, cluster of *Xba*I-WGS samples showed more significant than correlation with the cluster of *Xba*I-PFGE. In the case of p-value, all four groups showed a value of 0.05 or less, and the SU group and CC group showed a value of 0.05 or more. In the case of the CC group, a significant relationship between *Xba*I-PFGE and *Xba*I-WGS patterns was found, but the p-value was 0.210.

The results of comparison between S1-PFGE and S1-WGS results were analyzed. In case of comparing the PFGE result of plasmid profiling and result of *in silico* digestion, the mantel-r value of the two results was 0.36 or more in the four groups of GW, GG, GS, and CC. The p-value was shown less than 0.05 except for the

GG group. Through the results of the correlogram, the correlation between the samples was showed significant only in the CC group (Fig. 2-B).

The patterns of PFGE and WGS data were compared and analyzed using enzyme. *Xba*I restriction enzyme and S1 nuclease were applied to check molecular type and plasmid profiling, respectively. The number of samples in the CC group was 7 regardless of the data analysis results, and it was considered that the number of samples to support the analysis results was insufficient. In addition, the SU group had 41 samples, but the reorganization of samples according to the isolation period was 23 groups. In the case of this study, because the comparison of clustering of all samples in each group was performed, the reorganized group could not be applied, that affected the results of this study. In addition, the samples of the remaining four groups seemed to have significant results in PFGE and WSG in molecular typing, but no significant results in plasmid profiling patterns. This indicated the possibility of dissemination of plasmids between clonal isolates.

In analysis of the results of band patterns, the maximum values of the number of bands and fragments were 22 and 63 in molecular typing. The maximum values of band length and fragments were 438,656 bp and 448,625 bp, respectively. In case of plasmid profiling, the maximum values of the number of bands and contigs were 17 and 13. The maximum values of band length and fragments were 272,450 bp and 257,002 bp, respectively. Therefore, molecular typing patterns can be entered in 76 sections (0 ~ 488,625 bp), but patterns in plasmid profiling can be entered in 43 sections (0 ~ 259096.5 bp). According to these results, the diversity of possible patterns can also affect the result values in comparison of clustering patterns.

Table 1. Comaprison *Xba*I-PFGE with *Xba*I-WGS patterns.

Site	*Xba*I-PFGE vs *Xba*I-WGS		S1-PFGE vs S1-WGS	
	Mantel-r	p-value	Mantel-r	p-value
GW	0.404	0.001	0.746	0.001
GG	0.274	0.023	0.241	0.068
GS	0.357	0.001	0.363	0.002
SU	-0.009	0.561	0.468	0.001
JR	0.263	0.001	0.259	0.001
CC	0.175	0.237	0.861	0.003

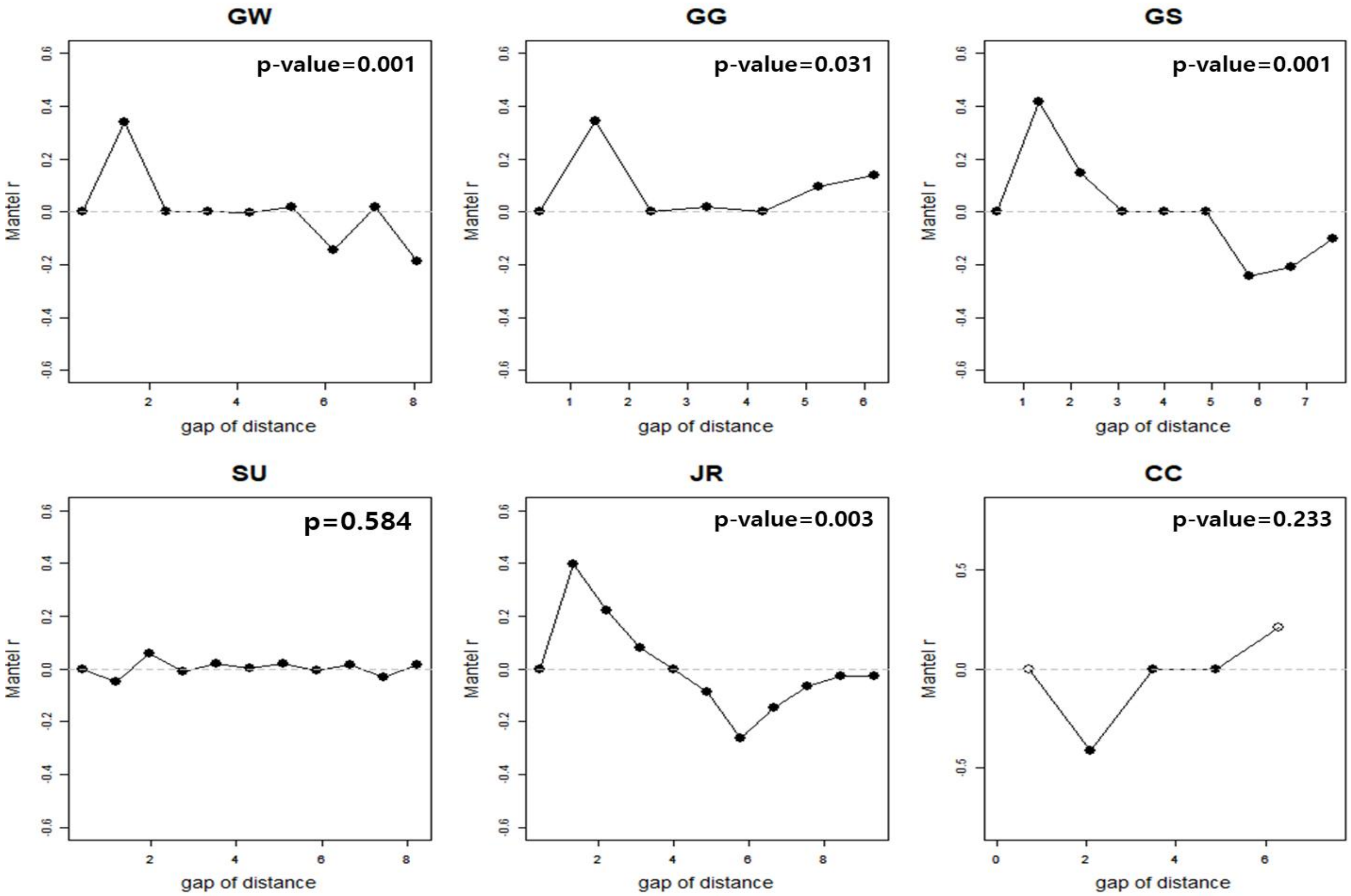


Figure 2-A. Correlogram of comparison *Xba*I-PFGE with *Xba*I-WGS.

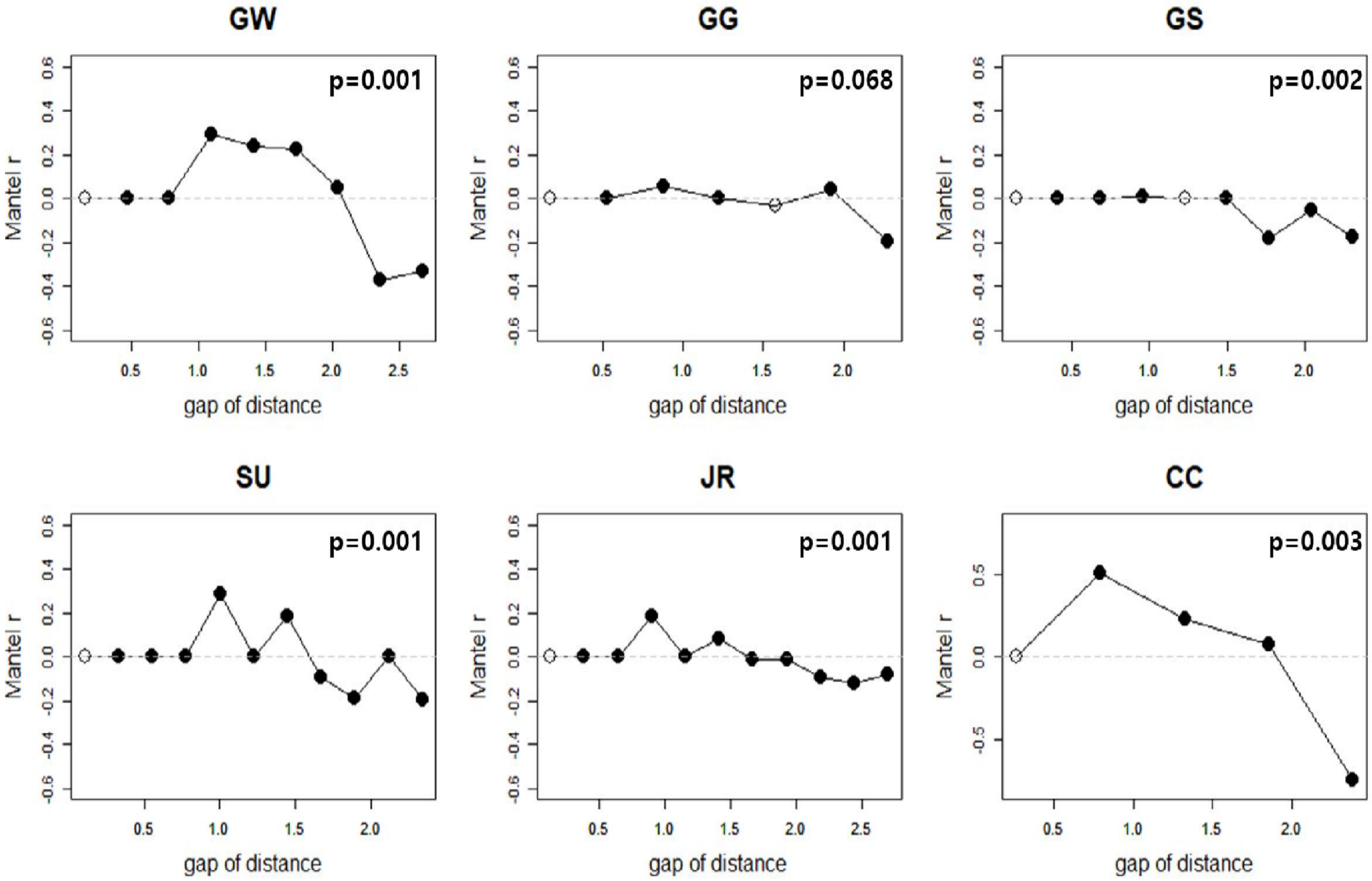


Figure 2-B. Correlogram of comparison S1-PFGE with S1-WGS.

Comparison of PFGE and ANI results

The results of ANI analysis were summarized by creating hierarchical clusters and heatmaps using the OrthoANI program (Figure 3-A, B.). In addition, tanglegrams were drawn comparing the PFGE results and the hierarchical clusters in Figure 2 (Figure 3-A,B.). The ANI results of all contigs (WGS-ANI) and plasmid contigs (Plasmid-ANI) of the WGS data were grouped as related animals to 94% (156/166) and 96% (161/166), respectively. In addition, similarity of molecular typing and plasmid profiling between contigs more than 99.9% ANI value rated 11% and 18%, respectively. The ratio of each site was as follows. WGS ANI and Plasmid ANI were 11.4% and 36.6% in the GW group, 10.3% and 38.9% in the GG group, 19.4% and 14.8% in the GS group, 4.3% and 4.3% in the SU group, 16.2% and 12.9% in the JR group, and 4.8% and 0% in the CC group.

ANI analysis was conducted to determine the match of genomes between WGS data of samples. Research on the possibility of judging the similarity of samples based on how well the ANI values match has already been conducted. The previous study established the criterion that samples can be judged similar when they are 99.9% or higher [34]. In addition, the other previous study was conducted showing that plasmids can move by horizontal gene transfer [35]. In the WGS-ANI value and Plasmid-ANI value of this study, the contigs with 99.9% or higher were 11.1% and 17.9%, respectively. In the GS, JR, and CC groups where the proportion of contigs with ANI values greater than 99.9% was WGS-ANI > Plasmid-ANI, plasmid transfer between clonal isolates was predicted. In addition, in the GW and GG groups with Plasmid-ANI > WGS-ANI, the fact that plasmid transfer mainly occurred between non-clonal isolates was predicted.

Each dendrogram was generated using the *Xba*I-PFGE value, WGS-ANI value, S1-PFGE value and Plasmid-ANI value. In addition, a tanglegram was generated to compare the two dendrograms (Figure 4-A, B). There were 6 samples each in which the clusters of the two dendrograms perfectly matched. In case of comparison of *Xba*I-PFGE and ANI values of all contigs, 6 couple of samples were OE_055, OE_057 of the GW group, OE_052, OE_053 and OE_051, OE_145 of the GG group, OE_031, OE_032 of the SU group, OE_161, OE_162 and OE_085, OE_086 of the JR group. When the ANI values of S1-PFGE and plasmid contigs were compared, OE_093, OE_094, OE_091, OE_109 in the GW group, OE_052, OE_053 and OE_050, OE_131 in the GG group, OE_161, OE_162 in the JR group, and OE_088 and OE_160 in the CC group.

The tanglegrams of Fig. 4-A, B. compare the clusters of PFGE patterns and ANI results. However, in process of generating each cluster, ANI compares all contigs one-to-one, but in the case of PFGE, simply measure the lengths of bands and generate the patterns. There was a high probability that different clusters will be appeared because a pattern is created and compared using the size of the contigs. Nevertheless, there were samples in which the Euclidean distance of the PFGE pattern and the ANI value cluster perfectly matched in the tanglegram. Among them, OE_052, OE_053 samples from the GG group and OE_161, OE_162 samples from the JR group were identical in both molecular typing and plasmid typing clusters. Thus, perfect matching of both cluster means that exact samples are clonal isolates and have the same plasmid. In addition, in previous study, clonal isolates are identified by clustering in *Xba*I-PFGE data [36]. Therefore, when only *Xba*I-PFGE and WGS-ANI values are the same, that samples are clonal isolates but have different plasmids. This results implies the possibility of horizontal gene transfer (HGT) of the plasmid between samples.

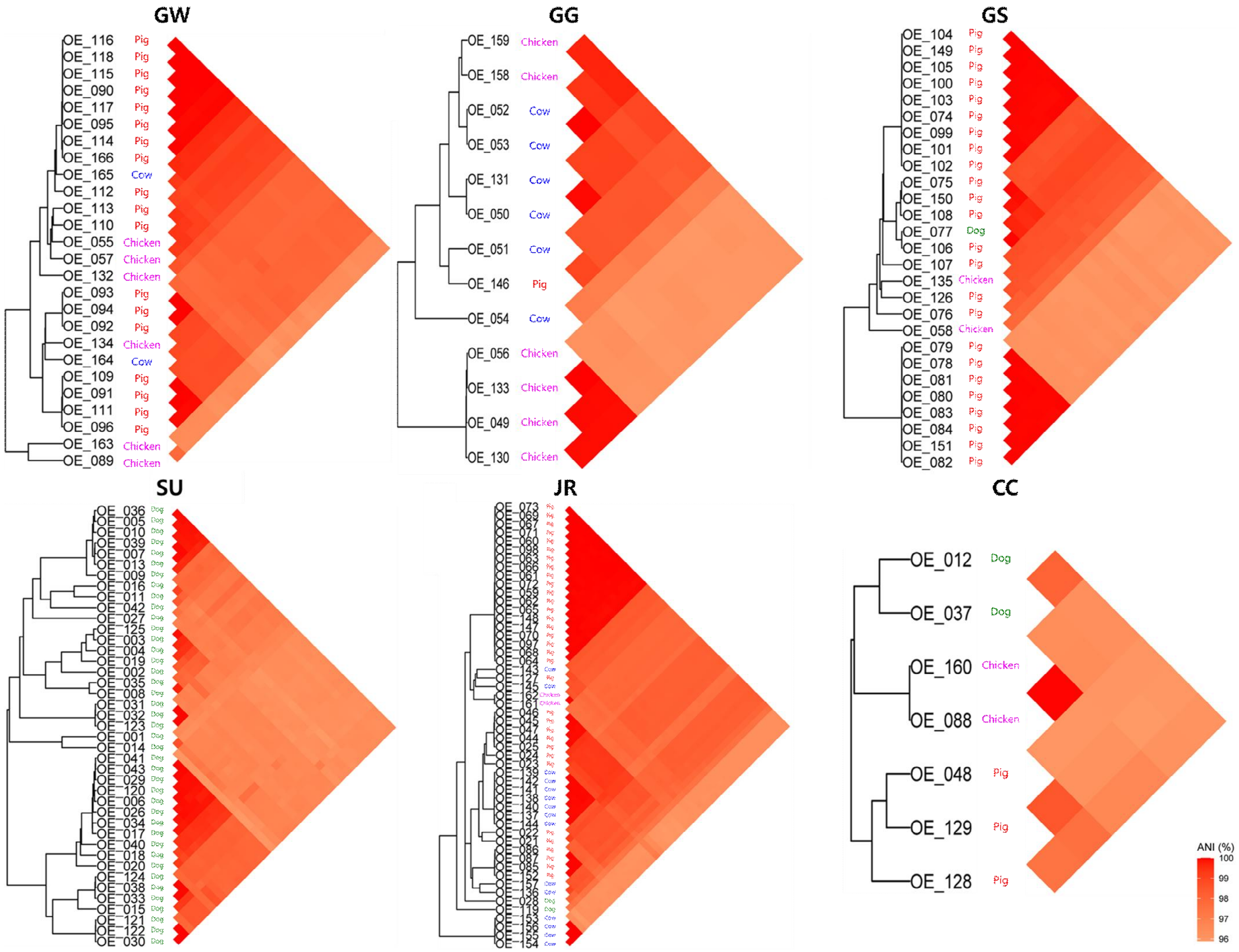


Figure 3-A. Heatmap and cluster of WGS-ANI result.

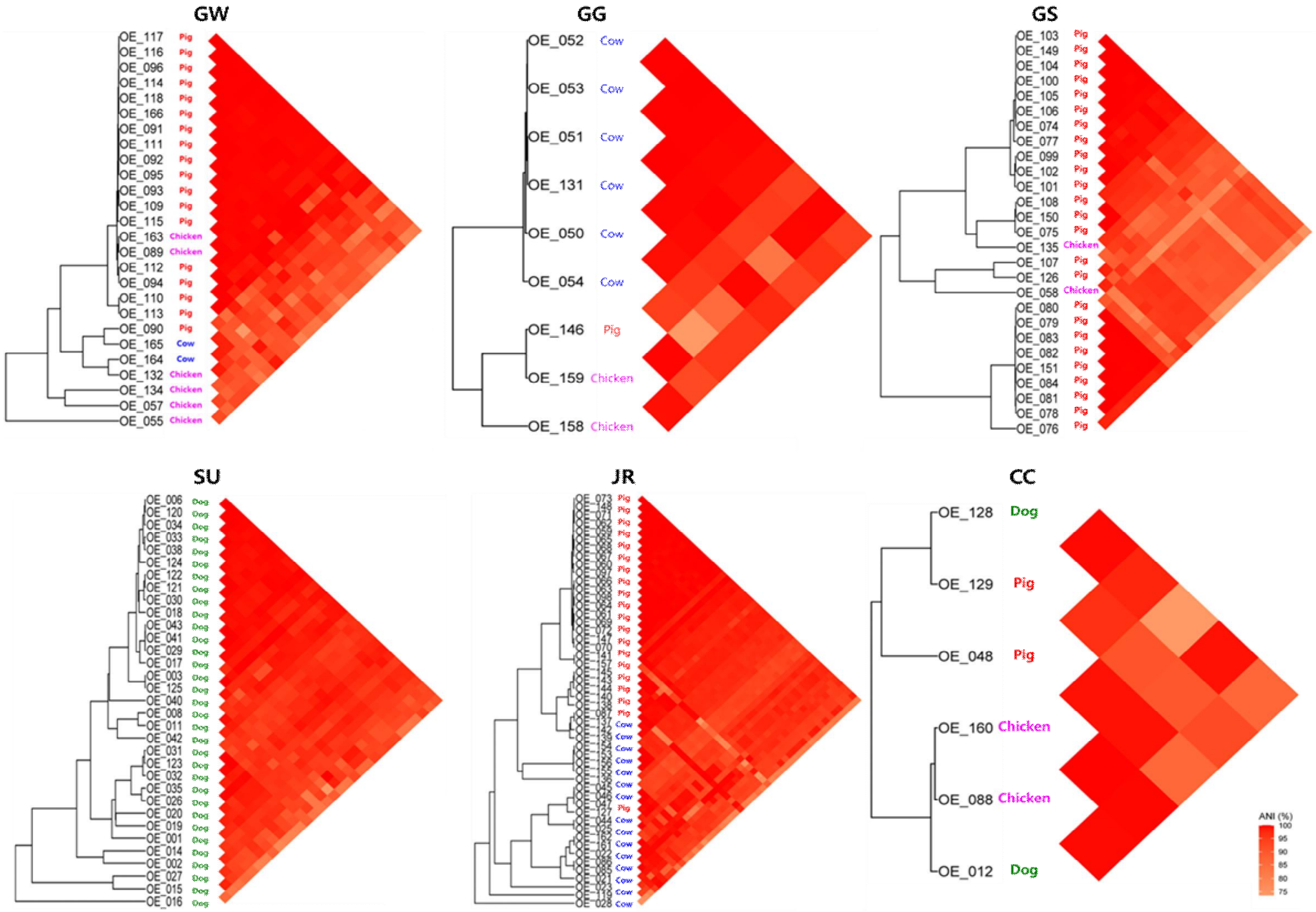


Figure 3-B. Heatmap and cluster of Plasmid-ANI result.

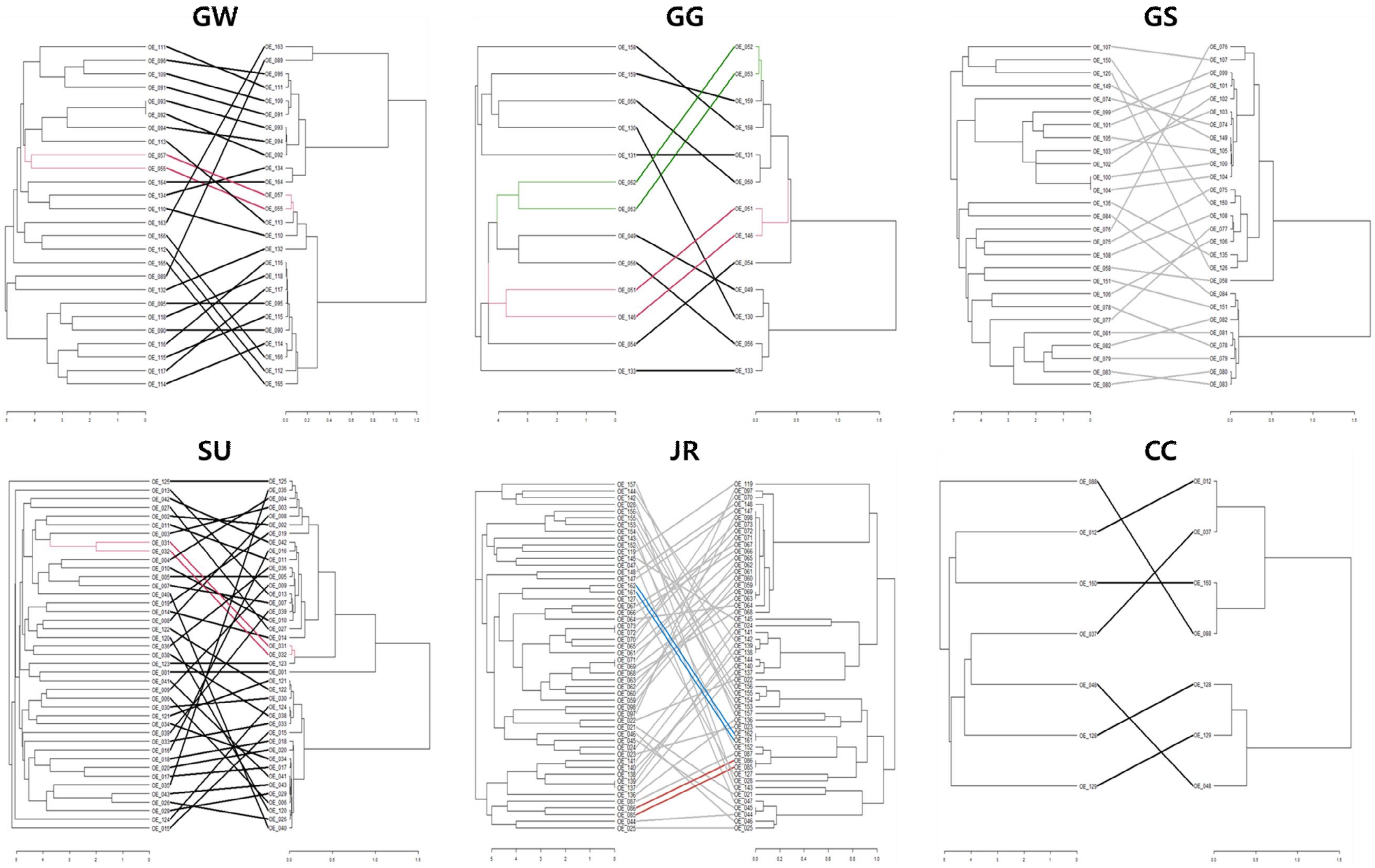


Figure 4-A. Tanglegram of comparison *Xba*I-PFGE cluster with WGS-ANI cluster.

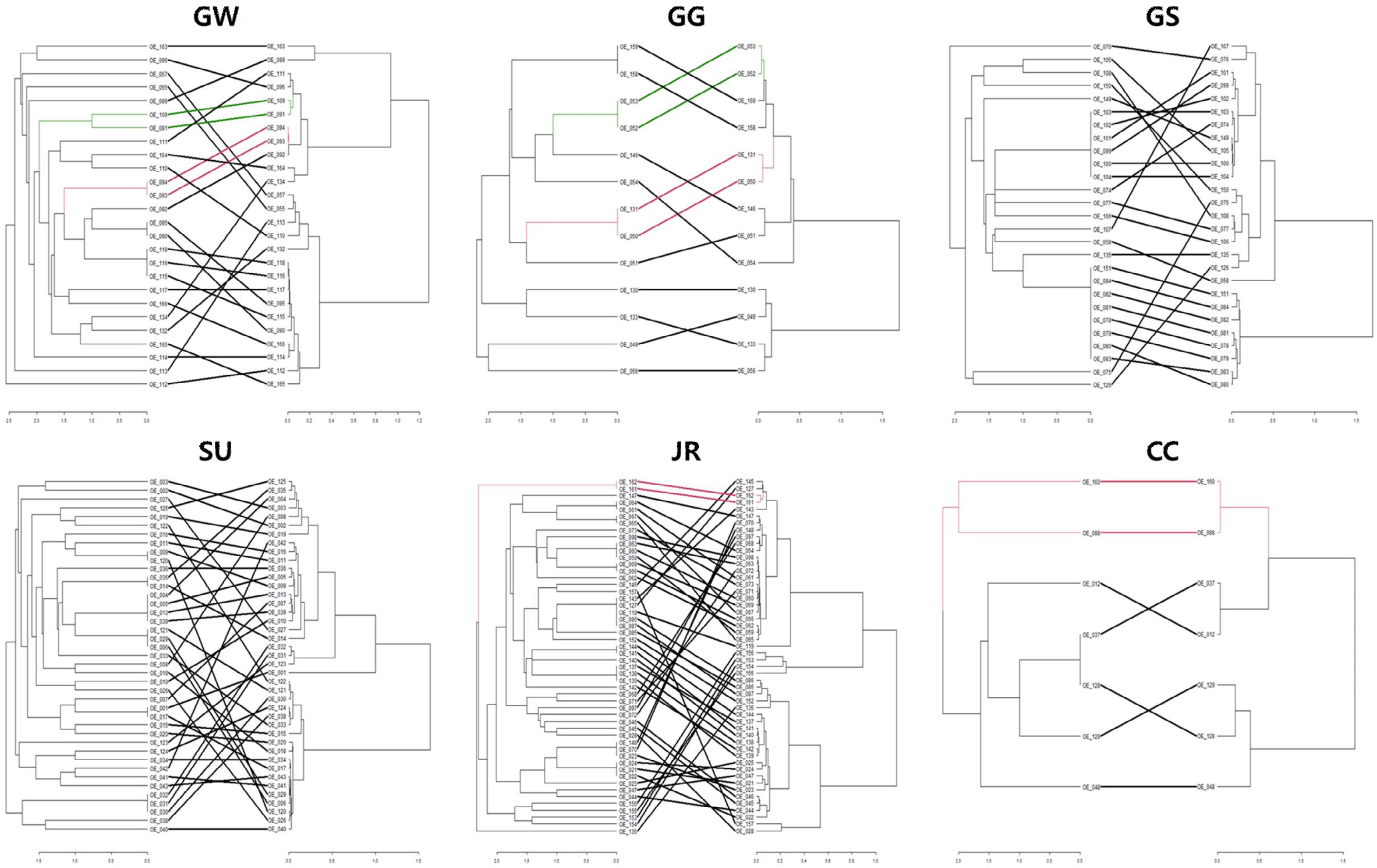


Figure 4-B. Tanglegram of comparison S1-PFGE cluster with Plasmid-ANI cluster.

Comparison of PFGE band and WGS data contig

In comparison S1-PFGE pattern with ANI value of samples, the clusters were perfectly matched among the six pairs of samples (Fig. 3-B.). The contig length of the WGS data and the band length of the PFGE result were compared. Since there was a previous study that the contig length and PFGE band length did not match perfectly [25], the PFGE bands that were ±10% of the contig length were regarded as similar bands and summarized (Table 2).

The correlation between samples was identified by analyzing the similarity of contigs judged to be plasmids between 6 couple of samples whose contigs and band lengths matched. The Euclidean distance between the patterns of the samples was 0, which perfectly matched the samples, and the ANI value was over 99.9%. Before comparing bands and contigs, PFGE results revealed that bands of 8.3 kb or less were difficult to distinguished by analyzing gel photo, so bands of 8.3 kb or less were excluded and compared with contigs. However, in case of the plasmid profiling, we performed conventional electrophoresis to check all plasmid of samples.

In the comparison of S1-PFGE pattern and plasmid-ANI results, OE_093, OE_094 and OE_091, OE_109 of the GW group, OE_052, OE_053 and OE_050, OE_131 of the GG group, and OE_161, OE_162, and CC of the JR group. It was OE_088 and OE_160 of the group were matched perfectly. Among the 33 contigs, 26 contigs had similar band lengths (78%). The 7 contigs were considered discrepant bands because they did not have bands of similar size.

In case of the unmatched PFGE band and the WGS data contig, it means an error in the WGS data. In this study, it was determined that 22% of WGS data had an error, which is a ratio based on length. However, it is difficult to conclude that the error rate is 22% because the subjects of the compared data were the Plasmid-ANI value and the S1-PFGE pattern. In addition, the form and analysis method of the values for comparison were different. Thus, additional research is required to verify the accurate correction of WGS data.

Table 2. Comparison S1-PFGE bands with WGS data contigs and statistical analysis.

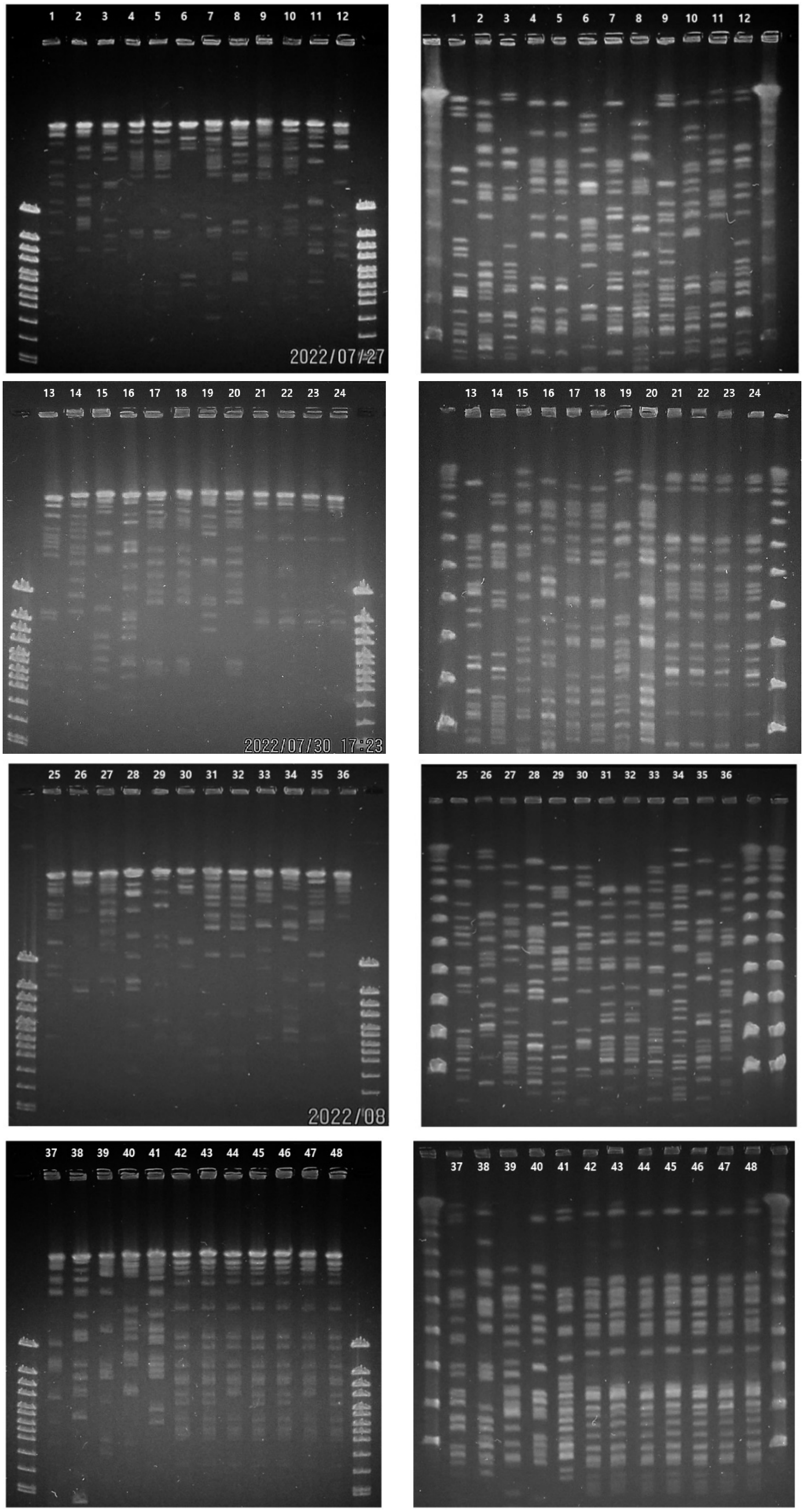
Site	Sample (contig#)	Euclidean distance	ANI(%)	Sample_contig#	WGS(bp)	PFGE(bp)
GW	OE 093 (3)	0	99.9913	OE_093_2	63528	-
				OE_093_3	251416	260947
				OE_093_4	43435	41770
	OE 094 (4)			OE_094_2	31766	29796
				OE_094_3	43439	41914
				OE_094_5	251417	261921
	OE 091 (4)	0	99.9918	OE_091_3	251416	259974
				OE_091_5	103257	98498
	OE 109 (3)			OE_109_4	103257	97570
				OE_109_6	251416	254948
GG	OE 052 (2)	0	99.9824	OE_052_2	124224	118941
				OE_052_3	104991	98884
	OE 053 (2)			OE_053_2	124222	120414
				OE_053_3	105617	98884
	OE 050 (2)	0	99.9489	OE_050_4	116218	112382
	OE 131 (8)			OE_131_9	57652	-
				OE_131_16	117663	113192
JR	OE 161 (7)	0	99.9719	OE_161_4	224711	218147
				OE_161_6	121852	-
				OE_161_7	154790	-
	OE 162 (6)			OE_162_3	224704	218147
				OE_162_4	55073	57518
				OE_162_8	49496	-
CC	OE 088 (8)	0	99.9862	OE_088_2	110897	-
				OE_088_3	125929	122014
				OE_088_4	228987	226300
				OE_088_5	100939	-
				OE_088_7	74321	70899
	OE 160 (10)			OE_160_2	125919	115961
				OE_160_3	74322	69331
				OE_160_4	89983	80996
				OE_160_5	229893	233241
				OE_160_8	99264	92461

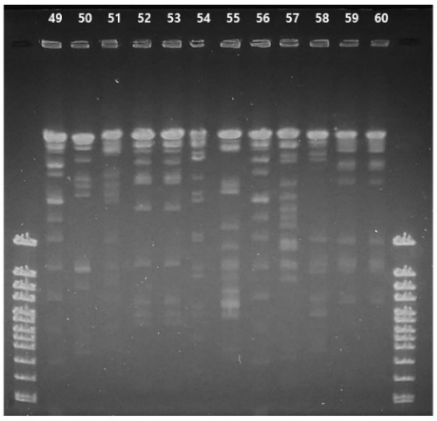
SUPPLEMENTARY INFORMATION

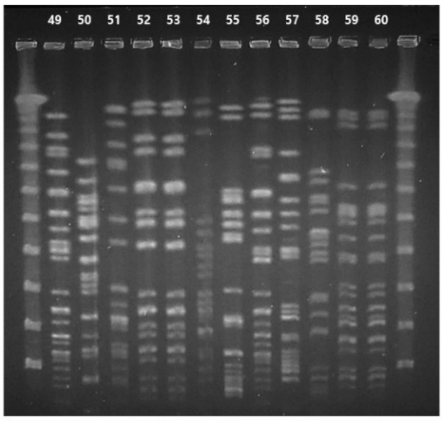
Table S1. *in silico* digestion shell script.

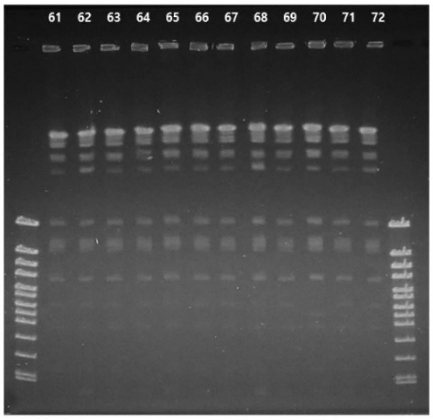
```
#!/bin/sh
#usage : ./restriction.sh    hybrid squencing fasta($1) sample_name($2)
XbaI="TCTAGA" #restriction enzyme    sequence

for i in `cat $2`; do
        sample=`grep    "$i" $1 | sort -t '_' -k2,3n`
        for m in    $sample; do
                echo $m >>    $[27]
                grep -A 1    "$m" $1 | sed -n 2p | sed "s/"$XbaI"/\n/g" >    temp
                line=`cat temp |    sed '$d' | wc -l`

                for l in `seq 2    1 $line`; do
                        sed -n $[1-11, 14-34, 37-43]p temp | wc -m    >>$[1-11, 14-34, 37-44]
                done
        done
done
```

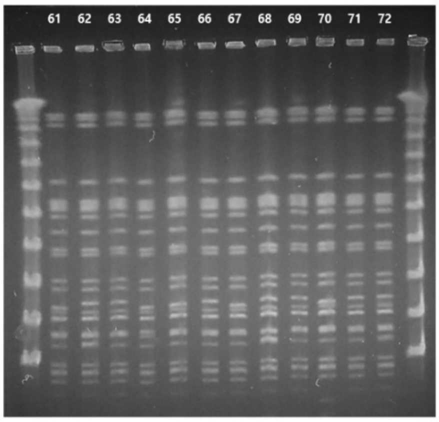
Figure S1-A. *Xba*I-PFGE gel photo.

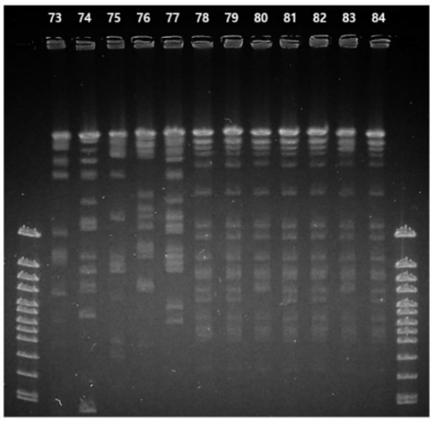


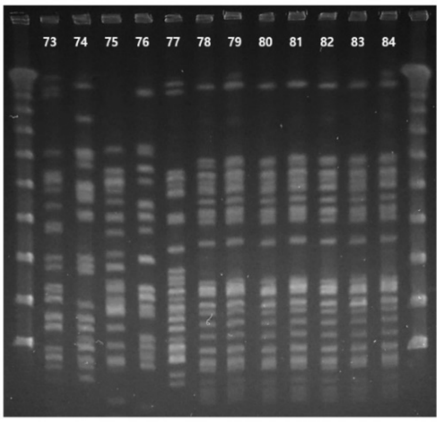

49 50 51 52 53 54 55 56 57 58 59 60

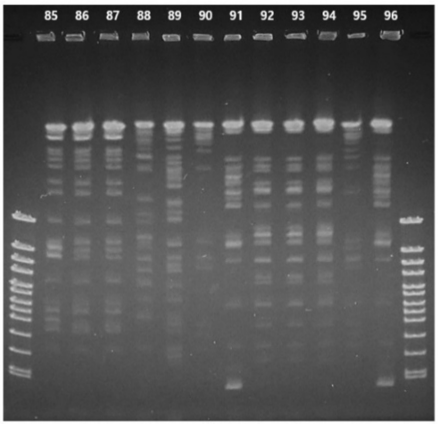

49 50 51 52 53 54 55 56 57 58 59 60

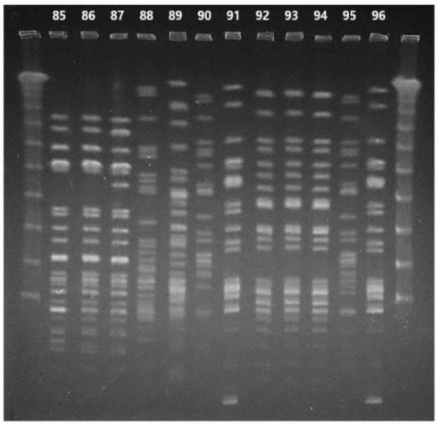

61 62 63 64 65 66 67 68 69 70 71 72

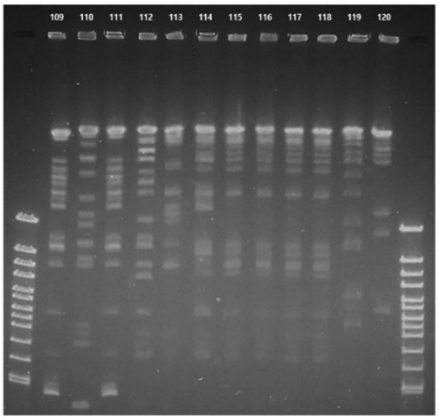

61 62 63 64 65 66 67 68 69 70 71 72

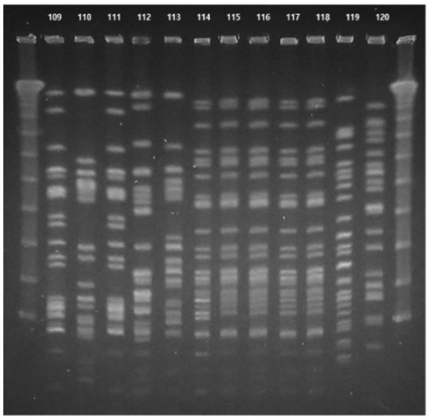

73 74 75 76 77 78 79 80 81 82 83 84

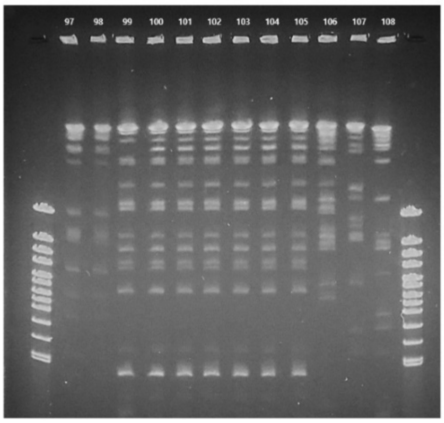

73 74 75 76 77 78 79 80 81 82 83 84

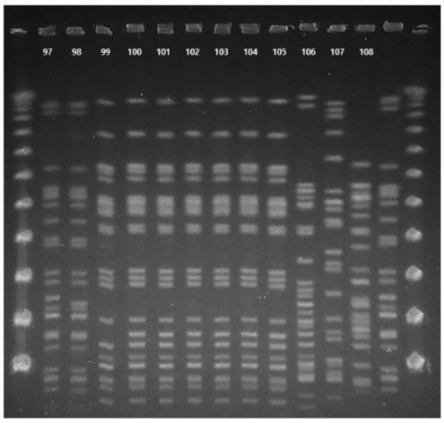

85 86 87 88 89 90 91 92 93 94 95 96

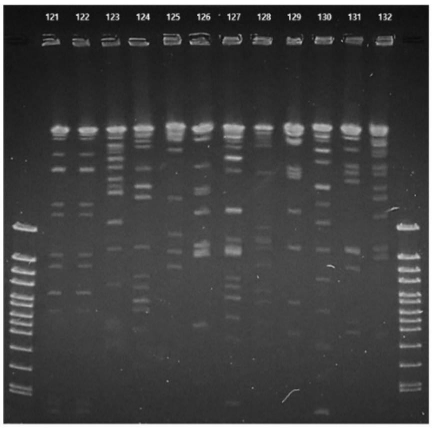

85 86 87 88 89 90 91 92 93 94 95 96

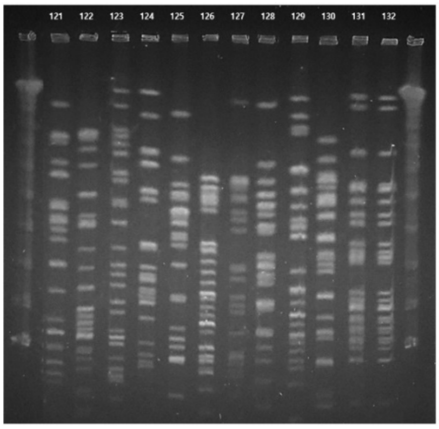

109 110 111 112 113 114 115 116 117 118 119 120

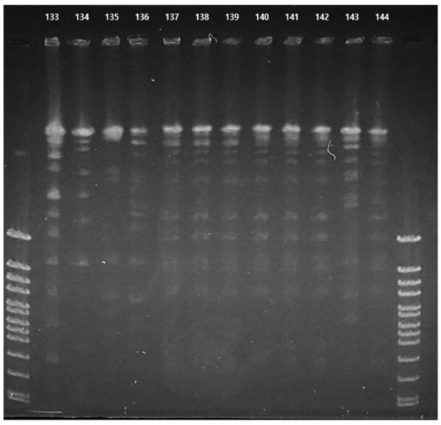

109 110 111 112 113 114 115 116 117 118 119 120

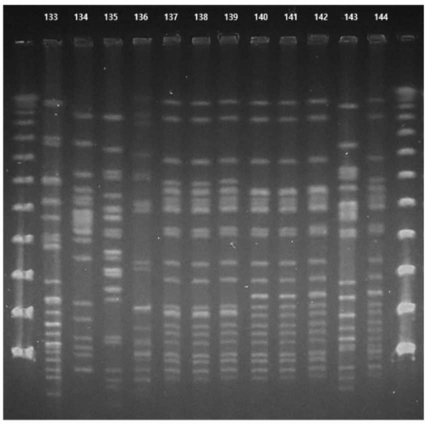

97 98 99 100 101 102 103 104 105 106 107 108


97 98 99 100 101 102 103 104 105 106 107 108


121 122 123 124 125 126 127 128 129 130 131 132


121 122 123 124 125 126 127 128 129 130 131 132


133 134 135 136 137 138 139 140 141 142 143 144


133 134 135 136 137 138 139 140 141 142 143 144

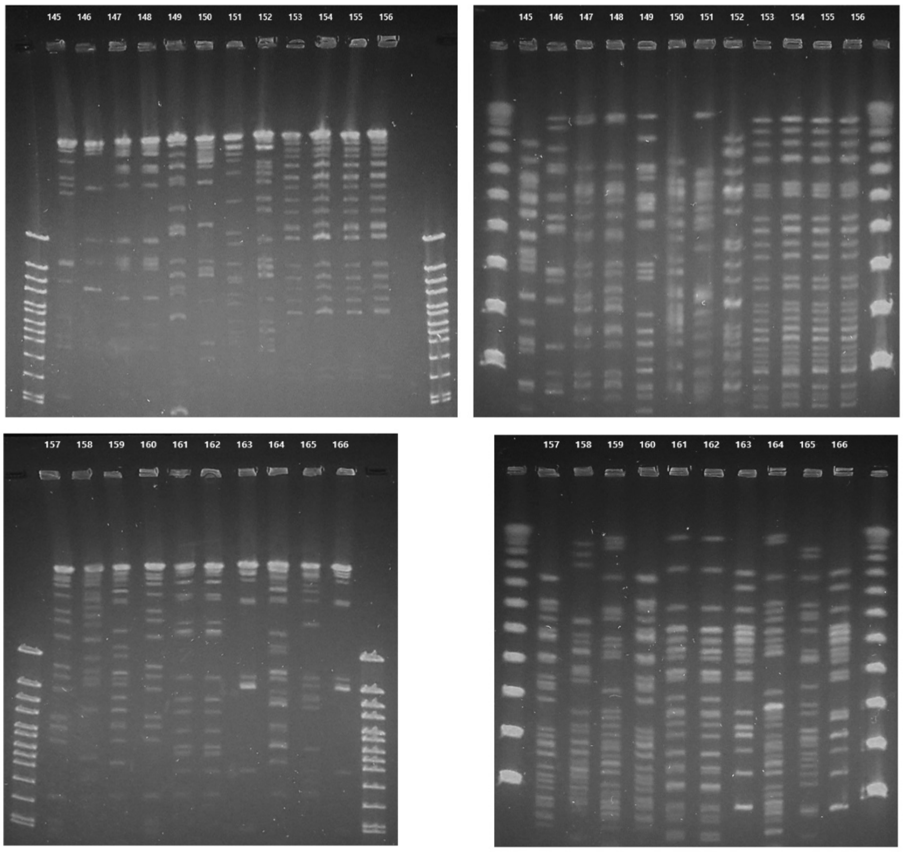
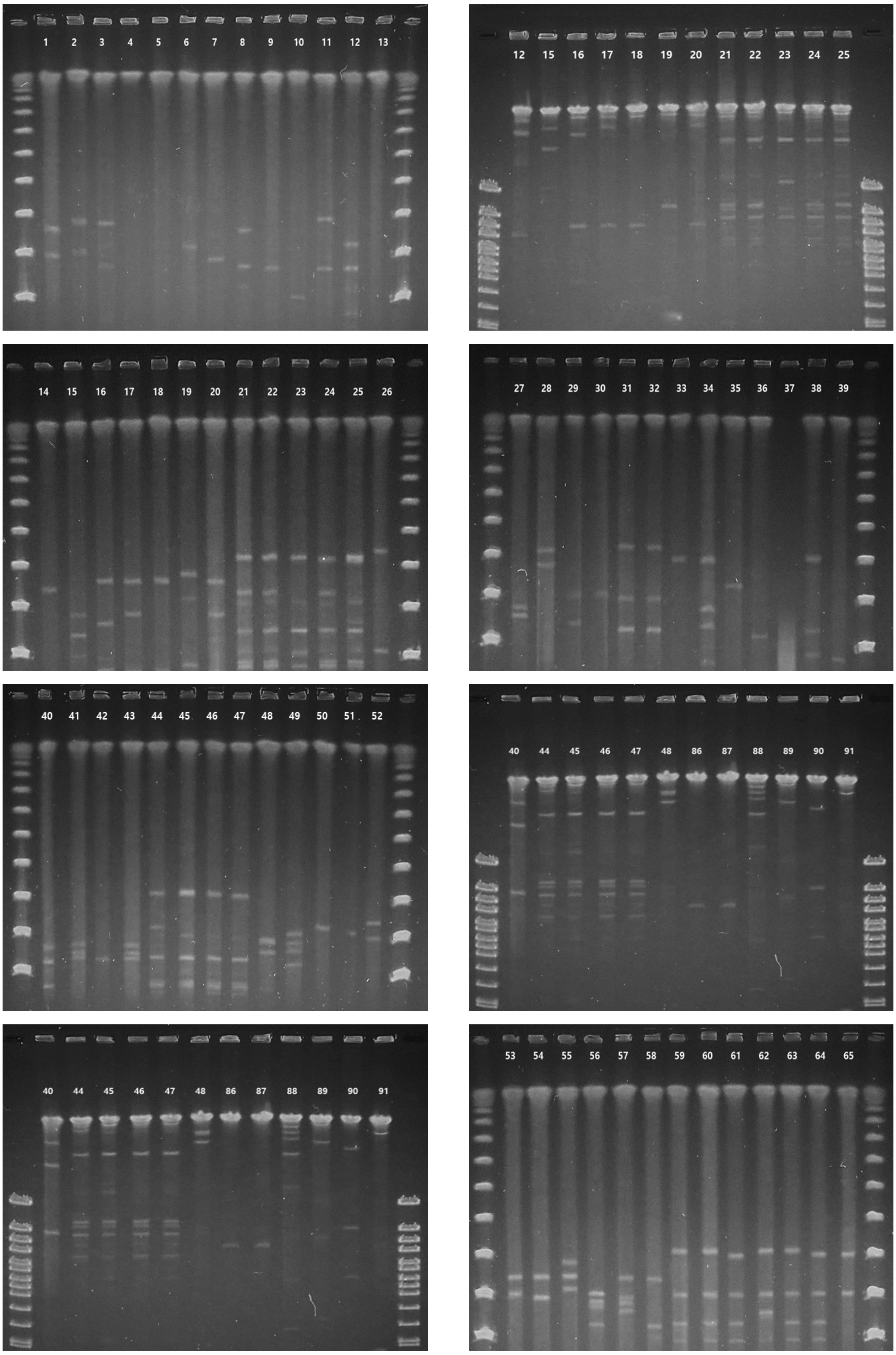
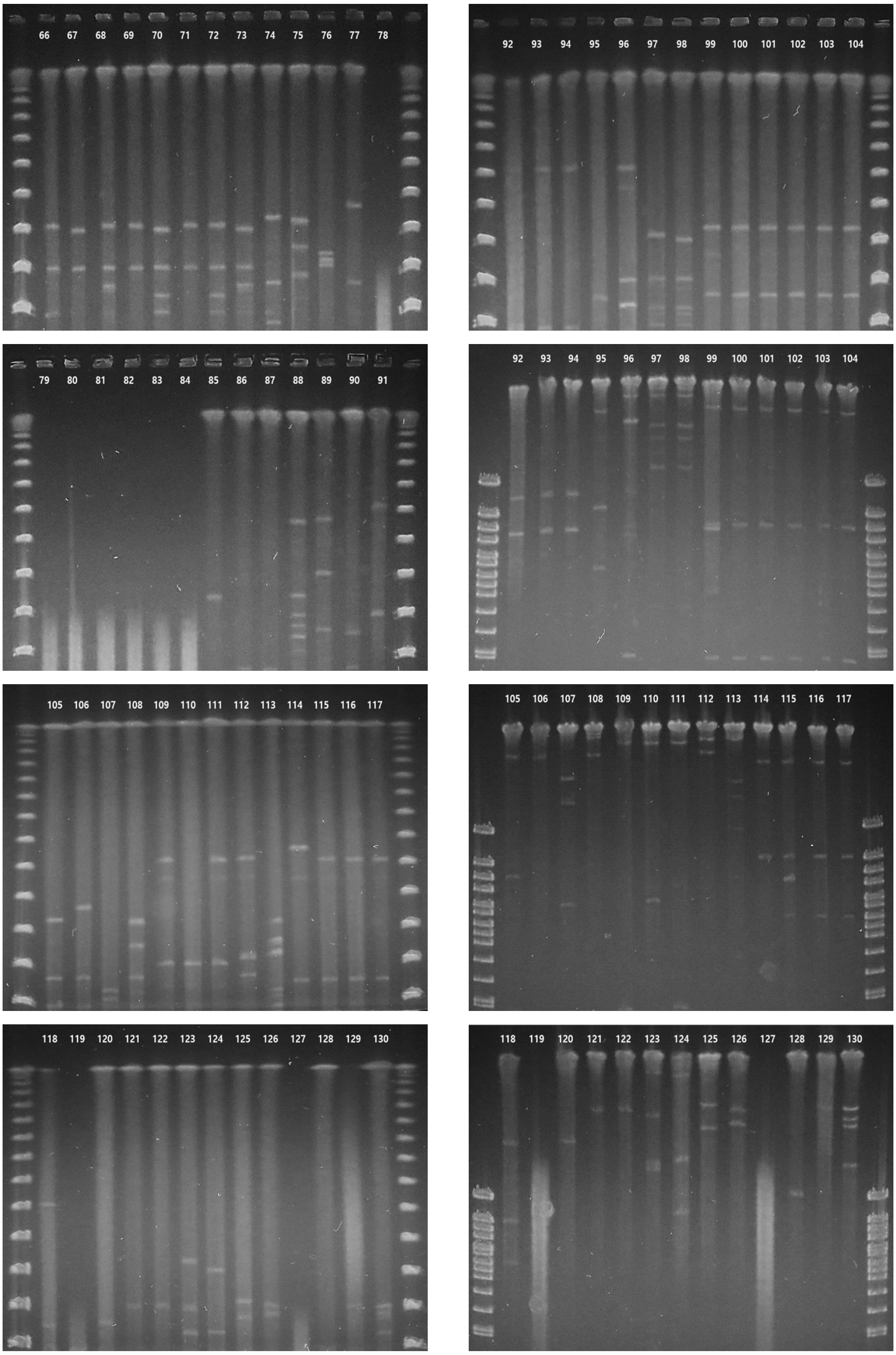
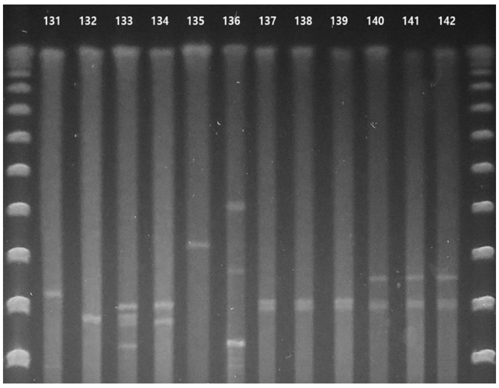
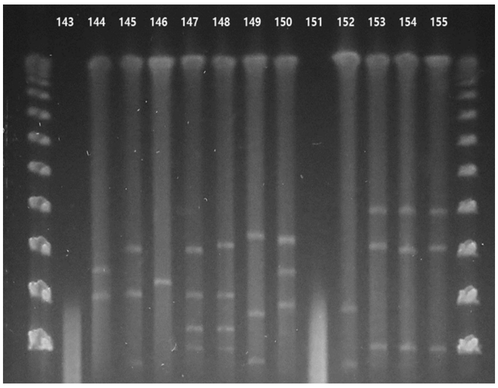

145 146 147 148 149 150 151 152 153 154 155 156
145 146 147 148 149 150 151 152 153 154 155 156
157 158 159 160 161 162 163 164 165 166
157 158 159 160 161 162 163 164 165 166

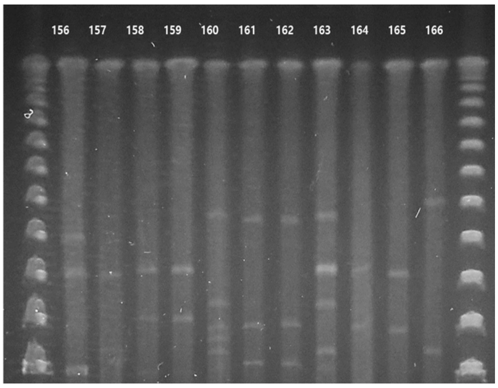
Figure S1-B. S1-PFGE gel photo.




66 67 68 69 70 71 72 73 74 75 76 77 78
92 93 94 95 96 97 98 99 100 101 102 103 104
79 80 81 82 83 84 85 86 87 88 89 90 91
92 93 94 95 96 97 98 99 100 101 102 103 104
105 106 107 108 109 110 111 112 113 114 115 116 117
105 106 107 108 109 110 111 112 113 114 115 116 117
118 119 120 121 122 123 124 125 126 127 128 129 130
118 119 120 121 122 123 124 125 126 127 128 129 130


131 132 133 134 135 136 137 138 139 140 141 142


143 144 145 146 147 148 149 150 151 152 153 154 155


156 157 158 159 160 161 162 163 164 165 166

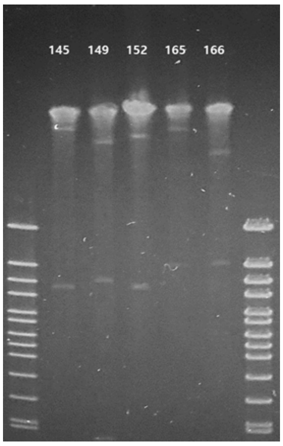
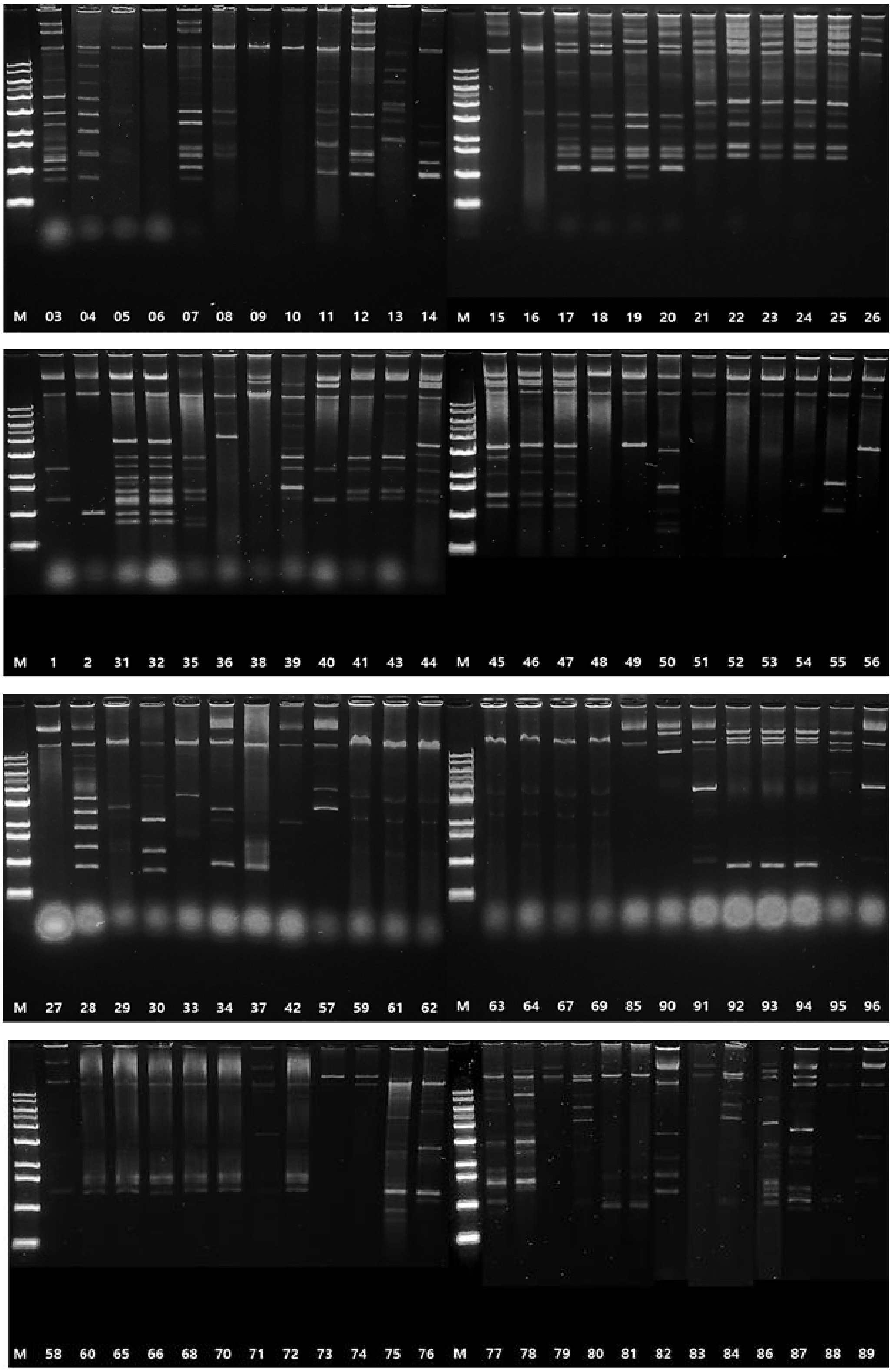
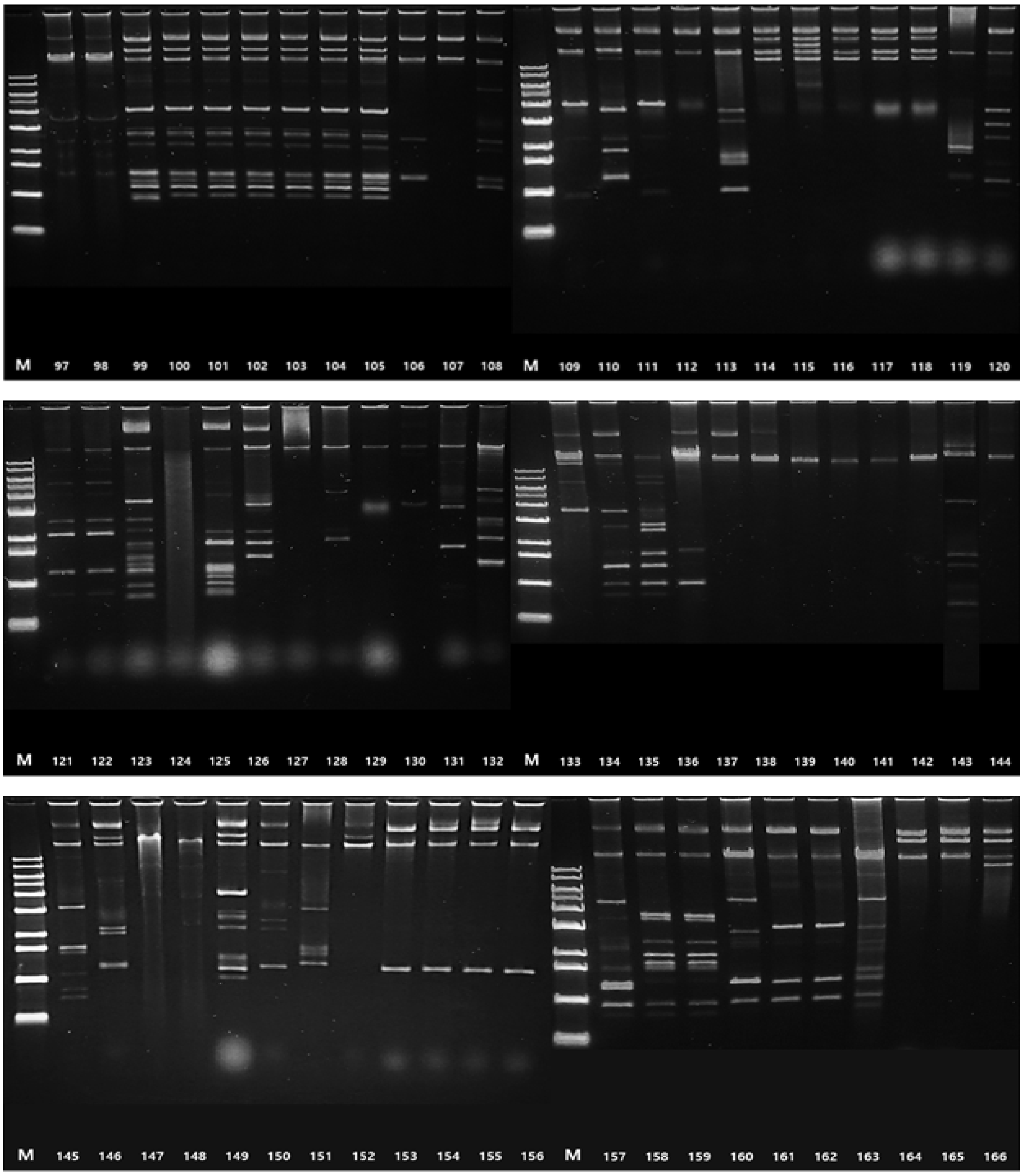

145 149 152 165 166

Figure S1-C. Plasmid profiling using conventional electrophoresis gel photo.


M 03 04 05 06 07 08 09 10 11 12 13 14 M 15 16 17 18 19 20 21 22 23 24 25 26
M 1 2 31 32 35 36 38 39 40 41 43 44 M 45 46 47 48 49 50 51 52 53 54 55 56
M 27 28 29 30 33 34 37 42 57 59 61 62 M 63 64 67 69 85 90 91 92 93 94 95 96
M 58 60 65 66 68 70 71 72 73 74 75 76 M 77 78 79 80 81 82 83 84 86 87 88 89


M 97 98 99 100 101 102 103 104 105 106 107 108 M 109 110 111 112 113 114 115 116 117 118 119 120
M 121 122 123 124 125 126 127 128 129 130 131 132 M 133 134 135 136 137 138 139 140 141 142 143 144
M 145 146 147 148 149 150 151 152 153 154 155 156 M 157 158 159 160 161 162 163 164 165 166

CONCLUSION

Molecular typing and plasmid profiling were performed by conducting PFGE and WGS of 166 E. coli, and the patterned results were compared and analyzed. In case of comparison the pattern of WGS data with PFGE results, it appeared similar in 4 out of 6 groups. However, we figured out difficulty determining the significance of the pattern when the number of samples belonging to the group is small or when the samples are again classified into several groups. Thus, the composition of the sample may also affect the experimental results in pattern analysis. Additionally, 6 pairs of samples showed perfect matching between samples with high significance (ANI > 99.9%) of plasmid sequences and clusters in S1-PFGE results. As a result of analyzing 6 pairs of samples, the ratio of similarity between PFGE band and WGS contig was 78%. Paradoxically, the ratio of mismatch between PFGE band and WGS data contigs was 22%. However, additional research is required to conclude that the contigs that do not match in PFGE and WGS data are caused by sequencing errors.

REFERENCE

1. Park, S.T. and J.J.I.n.j. Kim, Trends in next-generation sequencing and a new era for whole genome sequencing. 2016. 20(Suppl 2): p. S76.

2. Witney, A.A., et al., Clinical use of whole genome sequencing for Mycobacterium tuberculosis. 2016. 14(1): p. 1-7.

3. McNally, K.L., et al., Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. 2006. 141(1): p. 26-31.

4. Ranjan, R., et al., Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. 2016. 469(4): p. 967-977.

5. Van El, C.G., et al., Whole-genome sequencing in health care. 2013. 21(6): p. 580-584.

6. Lam, H.Y., et al., Performance comparison of whole-genome sequencing platforms. 2012. 30(1): p. 78-82.

7. Jain, M., et al., The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. 2016. 17(1): p. 1-11.

8. Ng, P.C. and E.F.J.G.v. Kirkness, Whole genome sequencing. 2010: p. 215-226.

9. Larsen, M.V., et al., Multilocus sequence typing of total-genome-sequenced bacteria. 2012. 50(4): p. 1355-1361.

10. Gupta, S.K., J.-M.J.A.a. Rolain, and chemotherapy, Reply to “Comparison of the web tools ARG-ANNOT and ResFinder for detection of resistance genes in bacteria”. 2014. 58(8): p. 4987-4987.

11. Struelens, M.J.J.M.d.I.O.C., Molecular epidemiologic typing systems of bacterial pathogens: current issues and perpectives. 1998. 93: p. 581-586.

12. Salipante, S.J., et al., Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. 2015. 53(4): p. 1072-1079.

13. Cooper, J.E. and E.J.J.T.i.m. Feil, Multilocus sequence typing–what is resolved? 2004. 12(8): p. 373-377.

14. Goering, R.V.J.I., Genetics and Evolution, Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. 2010. 10(7): p. 866-875.

15. Neoh, H.-m., et al., Pulsed-field gel electrophoresis (PFGE): A review of the “gold standard” for bacteria typing and current alternatives. 2019. 74: p. 103935.

16. Struelens, M.J., R. De Ryck, and A. Deplano, Analysis of microbial genomic macrorestriction patterns by pulsed-field gel electrophoresis (PFGE) typing, in New approaches for the generation and analysis of microbial typing data. 2001, Elsevier. p. 159-176.

17. Schwartz, D.C. and M.J.N. Koval, Conformational dynamics of individual DNA molecules during gel electrophoresis. 1989. 338(6215): p. 520-522.

18. Fox, E.J., et al., Accuracy of next generation sequencing platforms. 2014. 1.

19. Bragg, L.M., et al., Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. 2013. 9(4): p. e1003031.

20. Quainoo, S., et al., Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. 2017. 30(4): p. 1015-1063.

21. Lassmann, T., Y. Hayashizaki, and C.O.J.B. Daub, TagDust—a program to eliminate artifacts from next generation sequencing data. 2009. 25(21): p. 2839-2840.

22. MacLeod, I.M., et al., Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. 2013. 30(9): p. 2209-2223.

23. Salk, J.J., M.W. Schmitt, and L.A.J.N.R.G. Loeb, Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. 2018. 19(5): p. 269-285.

24. Faino, L. and B.P.J.T.i.P.S. Thomma, Get your high-quality low-cost genome sequence. 2014. 19(5): p. 288-291.

25. Juraschek, K., et al., Outcome of different sequencing and assembly approaches on the detection of plasmids and localization of antimicrobial resistance genes in commensal Escherichia coli. 2021. 9(3): p. 598.

26. Gaul, S.B., et al., Use of pulsed-field gel electrophoresis of conserved *Xba*I fragments for identification of swine Salmonella serotypes. 2007. 45(2): p. 472-476.

27. Barton, B.M., G.P. Harding, and A.J.J.A.b. Zuccarelli, A general method for detecting and sizing large plasmids. 1995. 226(2): p. 235-240.

28. PulseNet, C.J.C.D.C.P.A., Standard Operating Procedure for PulseNet PFGE of Escherichia coli O157: H7, Escherichia coli non-O157 (STEC), Salmonella serotypes, Shigella sonnei and Shigella flexneri. 2017. 157: p. 1-16.

29. Birnboim, H., [17] A rapid alkaline extraction method for the isolation of plasmid DNA, in Methods in enzymology. 1983, Elsevier. p. 243-255.

30. Murigneux, V., et al., MicroPIPE: Validating an end-to-end workflow for high-quality complete bacterial genome construction. 2021. 22(1): p. 1-15.

31. Johnson, M., et al., NCBI BLAST: a better web interface. 2008. 36(suppl_2): p. W5-W9.

32. Lee, I., et al., OrthoANI: an improved algorithm and software for calculating average nucleotide identity. 2016. 66(2): p. 1100-1103.

33. Schneider, C.A., W.S. Rasband, and K.W.J.N.m. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. 2012. 9(7): p. 671-675.

34. Olm, M.R., et al., Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. 2017. 27(4): p. 601-612.

35. Sørensen, S.J., et al., Studying plasmid horizontal transfer in situ: a critical review. 2005. 3(9): p. 700-710.

36. Seemann, T.J.B., Prokka: rapid prokaryotic genome annotation. 2014. 30(14): p. 2068-2069.

ACKNOWLEDGEMENT

This work was supported by the research program 'Integrated analysis of AMR genomic data derived from multiple sectors to build up One Health AMR system' funded by the Korea Disease Control and Prevention Agency from 2020 to 2022 (2020-ER5408-00).

감사의 글

이 논문을 쓰기까지 많은 도움들이 있었습니다. 우선, 저의 지도 교수님이신 운노 타쯔야 교수님께 깊은 감사의 말씀을 전하고 싶습니다. WGS data와 PFGE 결과를 비교할 아이디어를 주시지 않으셨다면 졸업 논문에 대한 연구를 시작하지 못했을 것입니다. 또한, 교수님의 실험실에서 석사 과정을 시작할 수 없었다면 생명 정보학을 접하기도 힘들었을 것입니다. 석사 과정동안 영어 논문을 함께 읽으면서 논문 읽기를 통해서 과학자로써 사고할 수 있는 방법에 대해서 알려주셨고, 소홀히 할 수 있었던 영어 공부를 미국 드라마 '프렌즈'를 롤플레잉 함으로써 재밌고 효과적으로 공부하는 쉐도잉 영어 공부 방법도 알려주셨습니다. 때론 교수님의 역할로써 실험실 학생들을 지도해 주시고, 때로는 생명공학도 대선배님으로써 학생들에게 과학 외에 중요한 영어, 사회생활 등을 조언해주셨음에 감사드립니다.

그리고 저의 부족한 졸업논문을 심사해 주시고 피드백까지 해주신 박수제 교수님, 박종은 교수님께 감사드립니다. 심사위원 교수님들의 조언과 피드백이 아니었다면 논문을 매끄럽고 간결하게 쓸 수 없었을 것입니다.

연구실에 있는 동안 참 많은 것들을 배웠습니다. 실험실에서 이루어지는 실험들, 연구실에서 이루어지는 분석들 모두 저에게는 처음이었습니다. 제가 실험실에 처음 들어왔을 때 옆에서 모르는 부분들을 알려주시고 적응할 수 있도록 도와주시고 막막했던 본 논문의 Result & discussion을 도와주신 김정만 박사님, 질병관리청 연구 과제를 함께 진행하고 본 논문을 쓸 때 사용했던 통계 분석을 자세히 알려주신 송호경 박사님 두 분께 감사드립니다. 또한, 지금은 졸업했지만 hybrid sequencing이 무엇인지 알려주고 조언해준 Raza Shahbaz와 힘들 때 옆에서 위로해주고 과학적 사고를 위해서 논문읽기의 중요성을 알려준 Singh Vineet에게 감사드립니다.

또한, 실질적으로 저와 가까운 위치에서 실험실 생활을 함께 해준 동료들에게 감사드립니다. 실험실에서 정식으로 석사 과정을 시작하기 전 실험실에서 이루어지는 실험과 기본적인 분석들과 흐름에 대해서 자세히 알려주고 옆에서 아낌없이 실질적인 조언해주고 때로는 실험실 생활 외적으로 신경을 많이 써준 선배이자 동생인 전다빈, 실험실에서 밤새 함께 Library를 제작 및 실험을 하고 기본적인 분석과 실험

들에 대해서 구체적으로 설명해주고 석사생활의 처음부터 끝까지 동고동락했던 정유진에게 특별히 감사를 표합니다.

저의 후배인 정지원, Bhandari Aprajita, 현수진, 김민우 학생들은 제가 도움을 요청할 때마다 자기 일처럼 도와주었으며 부족한 저의 조언도 흔쾌히 들어주었습니다. 특히 현수진 석사 과정 학생은 저의 졸업논문을 위한 실험을 옆에서 많이 도와주었습니다. 귀찮은 일들이 많았을텐데 볼멘소리 없이 따라주었습니다. 박사과정을 시작한지 얼마 되지 않아 저와 함께 일을 해보지는 못했지만 실험이나 연구 외적으로 동료가 되어준 고광표, 조혜준 박사과정 학생에게도 감사드립니다.

정말 감사의 글을 작성하다보니 저의 석사과정의 처음과 끝까지 저희 실험실 동료들이 곁에 없었던 적이 없었던 것 같습니다. 저의 석사 생활동안 제가 다른 분들의 마음을 불편하게 하거나 피해를 끼친 일들이 있었다면 너그러운 마음으로 용서해주시길 바랍니다. 2년 동안 매일매일 마주치면서도 실험실이 불편하다는 느낌을 받은 적 없었던 것 같습니다. 이러한 분위기를 만들고 유지해주신 실험실 구성원 분들께 건강과 안녕을 기원하며 다시 한 번 감사를 표합니다.