COMPLETE GENOME SEQUENCING AND ANALYZING THE GENES OF *PSEUDOMONAS* SP. BIOMIG1^{BAC}

by

Recep Can Altınbağ

B.Eng. in Computer Engineering, Istanbul Technical University, 2019

Submitted to the Institute of Environmental Sciences in partial fulfillment of

the requirements for the degree of

Master of Science

in

Environmental Sciences

Boğaziçi University

2021

COMPLETE GENOME SEQUENCING AND ANALYZING THE GENES OF *PSEUDOMONAS* SP. BIOMIG1^{BAC}

APPROVED BY:

Assoc. Prof. Dr. Ulaş Tezel Thesis Advisor	. . . . .
Prof. Dr. İ. Raşit Bilgin	. . . . .
Assist. Prof. Dr. Emine Gözde Özbayram	. . . . .

DATE OF APPROVAL: 23/06/2021

ACKNOWLEDGEMENTS

I would like to express my genuine appreciation to my supervisor and the project leader Assoc. Prof. Dr. Ulaş Tezel. He always motivates me to push my limits and gives indescribable motivation with his knowledge, vision, and altitude. His approach to science has always impressed me. Every time I was stuck, he helped me with amazing ideas and support. He is also like a close friend who understands and guides me to overcome difficulties. I would like to thank to the members of my thesis committee, Prof. Dr. İ. Raşit Bilgin, and Assist. Prof. Dr. Emine Gözde Özbayram for evaluating my thesis and giving suggestions.

I would like to offer my special thanks to my undergraduate thesis advisor Assist. Prof. Dr. Mehmet Tahir Sandıkkaya. He showed me the possibilities I can't think of and he has a big impact on me to focus on environmental science.

I would like to thank my family members since they are very supportive and indulgent. Having an amazing family that I can always rely on has always given me great strength.

This research was supported by Bogazici University Research Fund under grant 19Y00P4.

ABSTRACT

COMPLETE GENOME SEQUENCING AND ANALYZING THE GENES OF *PSEUDOMONAS* SP. BIOMIG1^{BAC}

Disinfectant consumption has increased with the pandemic of SARS-CoV-2. Benzalkonium chlorides (BACs), which are a group of quaternary ammonium compounds (QACs), are widely used as active ingredients in many of those disinfectants in the market. Previously, a strain of *Pseudomonas,* i.e. BIOMIG1^{BAC}, that can completely mineralize BACs was isolated, and its draft genome was sequenced. Moreover, the *oxyBAC*, which is the key gene in the biotransformation of BACs, was identified and verified through genetic experiments. The main objective of this research was to understand the ecological significance of the strain BIOMIG1 and its *oxyBAC* gene at genomic level. The complete genome of the *Pseudomonas sp.* BIOMIG1^{BAC} was obtained with a hybrid method that combines Illumina short-read sequencing and Oxford Nanopore long-read sequencing technologies. *Pseudomonas sp.* BIOMIG1^{BAC} has a genome composed of a single circular chromosome with a 7,675,262 bp length. Phylogenomic analysis showed that the strain BIOMIG1^{BAC} is a new species classified under *Pseudomonas protegens* subgroup, which is tentatively named as *Pseudomonas alexanderii* sp. nov. The key genes of the BAC biodegradation pathway, including *oxyBAC*, are associated with transposons therefore they can be horizontally transferred among bacteria. Two transposon motifs that carry the *oxyBAC* and its accompanying genes, were identified in the *oxyBAC* containing genomes of other (QAC degrading) microorganisms. The outcomes of this study suggest that *oxyBAC* is present in phylogenetically diverse group of bacteria and has a potential to horizontally transfer within bacteria via transposition to plasmids and phage genomes in communities under the stress of disinfectants.

ÖZET

PSEUDOMONAS SP. BIOMIG1^{BAC}'ın TÜM GENOMUNUN ÇIKARTILMASI VE GENLERİNİN ANALİZİ

SARS-CoV-2 ile birlikte dezenfektan tüketimi artmıştır. Piyasadaki çoğu dezenfektan formülasyonunun aktif bileşenininde yaygın olarak dördüncül amonyum bileşikleri (DAB'ler) grubuna dahil olan benzalkonyum klorürler (BAK'ler) kullanılmaktadır. Laboratuvarımızdaki önceden yapılmış bir çalışmada, BAK'leri tamamen mineralize edebilen bir *Pseudomonas* suşu olan BIOMIG1^{BAC} izole edilmiş ve genomu parçalı olarak dizilenmiştir. Ayrıca, BAC'lerin biyotransformasyonunda anahtar gen olan *oxyBAC*, genetik deneylerle tanımlanmış ve doğrulanmıştır. Bu araştırmanın temel amacı, BIOMIG1^{BAC} suşunun ve *oxyBAC* geninin genomik düzeyde ekolojik öneminin belirlenmesidir. *Pseudomonas sp.* BIOMIG1^{BAC}'in tüm genomu Illumina kısa dizileme ve Oxford Nanopore uzun dizileme teknolojilerini birleştiren hibrit bir yöntemle oluşturuldu. *Pseudomonas sp.* BIOMIG1^{BAC} 7,675,262 bp uzunluğunda tek bir kromozomdan oluşan bir genoma sahiptir. Filogenetik analizler sonucunda BIOMIG1^{BAC} suşunun *Pseudomonas protegens* alt grubunda yer alan yeni bir tür olduğu ve *Pseudomonas alexanderii* olarak adlandırılabileceği gösterildi. Yapılan analizler sonucunda *oxyBAC* da dahil olmak üzere BAK parçalamadaki anahtar genlerin transpozonlarla ilişkili olduğu görüldü. Ek olarak, *oxyBAC* ve beraberindeki genleri taşıyan iki transpozon motifi, diğer (DAB parçalayıcı) mikroorganizmaların *oxyBAC* içeren genomlarında da yer almaktadır. Bu çalışma *oxyBAC*'ın filogenetik olarak çeşitli bakteri gruplarında bulunduğunu ve dezenfektanların stresi altında plazmitler ve faj genomları aracılığıyla yatay olarak diğer bakterilere transfer olma potansiyeline sahip olduğunu göstermektedir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZET v
TABLE OF CONTENTS vi
LIST OF FIGURES viii
LIST OF TABLES xii
LIST OF SYMBOLS/ABBREVIATIONS xiii
1. BACKGROUND 1
2. COMPLETE GENOME SEQUENCE of *PSEUDOMONAS* SP. BIOMIG1^{BAC} 3
2.1. Introduction 3
2.2. Materials and Methods 8
2.2.1. DNA Sequencing and Completing Genome with Hybrid Assembly 8
2.2.2. Comparative Analysis of Assemblers and Different Inputs 10
2.3. Result and Discussion 11
2.3.1. Completing the Circular Chromosome with Hybrid Assembly 11
2.3.2. Validation and Refinement of Genome Assembly 11
2.3.3. Comparison of Different Assembly Inputs 19
3. TAXONOMY of *PSEUDOMONAS* SP. BIOMIG1^{BAC} 23
3.1. Introduction 23
3.2. Materials and Methods 26
3.2.1. 16S rRNA Comparative Analysis 26
3.2.2. Whole Genome Comparative Analysis 27
3.2.3. Sub-Classification Based on Marker Genes of *Pseudomonas fluorescens* Group 28
3.2.4. Pan and Core Genome Analysis 28
3.3. Result and Discussion 29
4. STRAIN SPECIFIC METABOLIC FUNCTIONS OF *PSEUDOMONAS* SP. BIOMIG1^{BAC} 38
4.1. Introduction 38
4.2. Materials and Methods 40
4.2.1. Proteome Comparison 40
4.2.2. Gene Functions 40
4.3. Result and Discussion 40
5. POTENTIAL GENE MOBILITY MECHANISMS OF *PSEUDOMONAS* SP. BIOMIG1^{BAC} 52

5.1. Introduction 52
5.2. Materials and Methods 56
5.2.1. *oxyBAC* Gene Cluster 56
5.2.2. Metagenomic Analysis 56
5.2.3. Analysis of BIOMIG1 Phenotypes 57
5.3. Result and Discussion 57
6. CONCLUSIONS AND RECOMMENDATIONS 67
REFERENCES 69
APPENDIX A: GENOMES USED IN 16S rRNA ANALYSIS 80
APPENDIX B: *P. CHLORORAPHIS* AND *P. PROTEGENS* SUBGROUP GENOMES AND THEIR TAXONOMIC COMPARISON WITH GTDB AND NCBI TAXONOMY 85

LIST OF FIGURES

Figure 2.1. Four main challenges of the genome assembly…….....…….....……..…....6

Figure 2.2. Hybrid assembly pipeline to generate the complete genome…….……...9

Figure 2.3. The assembly graph produced by *SPAdes* with just short-reads before the hybrid assembly. The repeating sections are one of the major causes for topological complexity…….……12

Figure 2.4. Optimum point of contigs and dead-ends for the hybrid assembly…….……13

Figure 2.5. The comparison of the complete genome with hybrid assembly and the draft genome with short read assembly. The insertion sequences and the border of the contigs from the draft genome have a good correlation…….…….……14

Figure 2.6. (1) Hybrid assembly results of the different long read inputs show that Component-1 is circularized with long reads. (2) Short read assembly shows the exact path (The path is an IS) that connects the Component-1 to the chromosome, but it is not reliable because of insufficiency of short reads about repeating regions. (3) The long reads are successfully circularizing the Component-1 which is identified as TU. (4) IRL, IRR, DR, left flanks, and right flanks are showed. (5) The TU probably was in the genome previously, or it can be in the future…….……..…16

Figure 2.7. Annotated genes (green) of the TU with the insertion sequences (black) are drawn. IS91c2 (orange) was found in the chromosome; it can be effective for the insertion and excision…….……..….17

Figure 2.8. (1) Component-2 is unconnected in the hybrid assembly graph. (2) Component-2 is connected in the short-read assembly graph. (3) Blast result of the Component-2 shows that an IS element was divided it, so actually, there is no Component-2 in the genome. It was changed. (4) Left and right flanks of the mapped short reads. (5) The explanation of the situation. (6) The IS element (1213 bp) with IRL, IRR, DR, and flank sequences. It has one transposase with a 1006 bp length. …….…….……18

Figure 2.9. The comparison of different Nanopore (long read) inputs and their effects to the assembly quality. The optimum range is around 250 Mb and 500 Mb……...……..21

Figure 2.10. The comparison of different assembly inputs and their effects to the assembly quality. Hybrid assemblies are the best among all of them. ……..…….21

Figure 3.1. Flowchart of the genome based classification to define species. Overall genome relatedness index (OGRI) represent the similarity of two genomes. There are different examples of OGRI as ANI, dDDH, GGDC and Mash…….…....24

Figure 3.2. 16SrRNA tree of the representative genomes in *Pseudomonas* genus shows that *P.* sp. BIOMIG1^{BAC}, *P.* sp. CMR12a and *P.* sp. CMR5c are grouping with *P. putida* group members. In ANI values, it can be seen that 16S rRNA genes of *P.* sp. BIOMIG1^{BAC} and *P.* sp. CMR12a are identical. *Escherichia coli* K-12 MG1655 is the outgroup……..….30

Figure 3.3. Heatmap and dendrogram based on the whole-genome shared protein family statistics between *P. chlororaphis* and *P. protegens* subgroups. As an outgroup, *P. chlororaphis* subgroup was used. Clade 1, Clade 2, and Clade 3 are the three primary phylogroups of *P. protegens* subgroup. Subgroup names, species names, ANI values of the whole genome and 16S rRNA respect to *Pseudomonas alexanderii sp nov.* BIOMIG1^{BAC} can be found on top rows. When looking at the 16S rRNA ANI values, BIOMIG1^{BAC} is 99% similar to the strains in the Clade 1. When looking at the whole-genome ANI values, it is around 88-91%, similar to *P. piscis*, Clade 2, and Clade 3 strains. The maximum and minimum values in the heatmap are set to 6000 and 4000, respectively. As a result, the boundary color values even reflect the values that are higher or lower than the thresholds. The dark blue color tones represent subgroup differentiation, while the white color tones represent species differentiation. The same species can be identified by clusters of red color tones……..…….32

Figure 3.4. Phylogenetic tree of the strains in *P. protegens* and *P. chlororaphis* subgroups with species. General time reversible (GTR) model was used. The bootstrap consensus tree was built after 130 iterations with IQ-Tree (v. 2.1.1), the values and distant species were removed for clear visualization. *P.* sp. CF150 and *P. veronii* VI4T1 are outgroups……34

Figure 3.5. Core and pan-genome Venn diagrams of Clade 1, 2, and 3……..35

Figure 3.6. Strain specific protein families in Clade 1. BIOMIG1^{BAC} has the highest number as 938. Strains named MSSRFD41, MC042(T), FW507-12TSA, B6(2017), R26(2017), and CMR5c are strains of the *P. piscis*, and strains named as BIOMIG1^{BAC}(T), and CMR12a are strains of the *P. alexanderii sp nov*...……..…36

Figure 4.1. Proteome comparison of *Pseudomonas* sp. BIOMIG1^{BAC} with close relatives: *P.* sp. CMR12a, *P. piscis* MC042, *P. protegens* Pf-5, and *P. protegens* CHA0. Orange-colored regions in track 1 represent some of the xenobiotic degradation clusters such as BACs, BDMA, and Benzoate. Possible regions resulting from horizontal gene transfers can be seen…….41

Figure 4.2. SSGRs of *Pseudomonas* sp. BIOMIG1^{BAC} with the distribution of metabolic pathways. Outside to inside, SSGRs(blue), GC Content (heat map with plot), phages (purple), special genes (orange), core genome of Clade 1 (red), strain specific genes (blue), all genes (genes) and metabolic pathway links. The links of metabolic pathways are bolded if they are in the SSGRs. Percentages in the metabolic classes represent the proportion of SSGRs to the total. The ratio of xenobiotics biodegradation and metabolism is highest at 18%, and there are probably more undiscovered and unassigned genes related with xenobiotics…….…..43

Figure 4.3. BAC and aromatic hydrocarbon biodegradation pathways constructed based on the genes present in SSGR-24 and -7. Genes responsible for each reaction were annotated with gene names or numbers and given in Figure 4.4……...44

Figure 4.4. Annotation of the enzymes which are responsible for the xenobiotic degradation reactions in Figure 4.3. Transposases are also abundant in these regions. Thus, they can be indicators of horizontal gene transfers…….45

Figure 4.5. Annotation of prophages in *P.* sp. BIOMIG1^{BAC}. Hypothetical proteins and phage-like proteins are abundant. Also, phage-like proteins are not certain. Small prophages have head protein (capsid). On the other hand, large prophages also have tail proteins. Attachment sites and integrases are related, and most of them are found close to each other……48

Figure 5.1. Mostly used variations of QACs are BACs. BACs resistance mechanisms……...…53

Figure 5.2. Flanking ISs of the *oxyBAC* gene cluster are parts of the composite transposon. Left and Right IS Elements are 100% identical, and they have terminal inverted repeats (TIRs). 5 bp reverse repeat section can be a sign of the transposition. So, predicted TU could be driven by the right IS element, and the left IS element stays in the genome. The *oxyBAC* gene cluster contains 8 genes which are site-specific integrase, hypothetical protein, HNH endonuclease, SCP2 sterol-binding, *oxyBAC*, TetR, and two MFS transporters. IS elements are in the Tn3 family, and they have 3 genes which are transposase, recombinase, and toxin-antitoxin system…….……58

Figure 5.3. 16S rRNA tree of *oxyBAC* having (blue colored) strains and their close strains. The tree clearly shows the horizontal transfer of the *oxyBAC* gene cluster. *Methanothermobacter sp.* KEPCO1 is the outgroup……..….60

Figure 5.4. The *oxyBAC* gene cluster was found in two different composite transposon structures...62

Figure 5.5. (A) Assembly graph of the AS metagenome. (B) *MetaWRAP* Binning (C) Bin refinement with *GraphBin*…….……63

Figure 5.6. The insertion of a Tn3-Derived Inverted-Repeat Miniature Element (TIME) in Tn3 IS was created direct repeats (DRs). A predicted secondary structure of the TIME at the RNA level with MFE was built……...……....64

Figure 5.7. TCS network analysis for *oxyBAC* gene cluster, *oxyBAC* gene, and integrase gene in the cluster. The hatch marks in the edges represents the number of mutations…….66

LIST OF TABLES

Table 2.1. Comparison of the different technologies, assemblies and different inputs. The reference was created with the combination of 500 Mb and 700 Mb hybrid assemblies with the improvements mentioned above. 500 Mb (Hybrid)* is the submitted assembly to the NCBI database previously. In all hybrid assemblies, the used short reads are identical and 1 Gb…….20

Table 3.1. MinHash distances of close strains to the *Pseudomonas* sp. BIOMIG1^{BAC}. If it is higher than 0.05, they are probably different species……..…….31

Table 3.2. The combination of marker genes for *P. protegens* subgroup indication. *P. alexanderii sp nov.* BIOMIG1^{BAC}(T) and *P. alexanderii sp nov.* CMR12a are in the *P. protegens* subgroup. DGPf_0 is not significative…….…….33

Table 3.3. Core and pan genome size of two subgroups and clades of *P. protegens* subgroup..……35

Table 4.1. Comparison of insertion sequence numbers of different strains in Clade 1. There is a correlation between genome size and the number of IS…….46

Table 4.2. Clustering and proportioning based on specialty and specificity……..46

Table 4.3. Features of integrated prophages in *P.* sp. BIOMIG1^{BAC}……..47

Table 4.4. Antibiotic resistance genes which are found in the strain BIOMIG1^{BAC} and their mechanisms……..……....50

Table 5.1. The ratio differences of the mapped short reads to the Tn3 IS indicate that a Tn3 IS Element is separated with *oxyBAC* gene cluster……...…...61

Table 5.2. The control of the mapped short reads with the random segments……61

LIST OF SYMBOLS/ABBREVIATIONS

Abbreviation	Explanation
A	Adenine
ANI	Average Nucleotide Identity
BACs	Benzylalkyldimethylammonium Compounds
bp	Base Pair
C	Cytosine
CDS	Coding Sequence
COVID-19	Coronavirus Disease of 2019
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Deoxyribonucleic Acid
DR	Direct Repeat
EC	Enzyme Commission
EPA	Environmental Protection Agency
FTP	File Transfer Protocol
G	Guanine
Gb	Giga Base Pairs
GB	Gigabyte
GGDC	Genome to Genome Distance Calculation
GO	Gene Ontology
GTR	General Time Reversible
HGT	Horizontal Gene Transfer
IRL	Inverted Repeat Left
IRR	Inverted Repeat Right
K	Average Number of Nucleotide Differences
Kb	Kilo Base Pairs
KEGG	Kyoto Encyclopedia of Genes and Genomes
Mb	Mega Base Pairs
MFE	Minimum Free Energy
MGE	Mobile Genetic Element
MIC	Minimum Inhibitory Concentration
MLSA	Multilocus Sequence Typing
NCBI	National Center for Biotechnology Information

NGS	Next Generation Sequencing
NP	Nondeterministic Polynomial Time
OGRI	Overall Genome Relatedness Index
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
PGAP	Prokaryotic Genome Annotation Pipeline
proP	Prophage
QACs	Quaternary Ammonium Compounds
RefSeq	Reference Sequence
RNA	Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
Sd	Standard Deviation
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
SSGR	Strain Specific Genomic Region
T	Thymine
TIME	Tn3-Derived Inverted-Repeat Miniature Element
tRNA	Transfer Ribonucleic Acid
WGS	Whole Genome Sequencing

1. BACKGROUND

For many years, quaternary ammonium compounds (QACs) have been one of the most frequently used antibacterial agents in disinfection formulations. These compounds have become indispensable for surface cleaning and disinfection in households, industries, and medical settings due to their organic structure's durability and bioactivity against bacteria, fungi, and viruses at low concentrations (Pereira & Tagkopoulos, 2019). Especially benzalkonium chlorides (BACs), which are a group of QACs, have been some of the most preferred disinfectant chemicals during the SARS-CoV-2 epidemic (Hora et al., 2020). BACs consumption has increased twice during the pandemic (Zheng et al., 2020). Even though they are beneficial as a disinfectant, the effects that they create after consumption should not be ignored.

One of the impacts of QACs is their inhibitory effects on wastewater treatment systems. Approximately 75% of the QACs consumed reaches to wastewater treatment plants (Tezel and Pavlostathis, 2011). Because of their antimicrobial effects, QACs can decrease the performance of biological units of wastewater treatment plants. Given that many wastewater treatment plants are not designed for the removal of QACs, overall 25% of the QACs consumed are discharged into the environment. QAC residues in wastewater treatment and natural systems promote development and dissemination of QAC resistance mechanisms which also confer resistance to clinical antibiotics (Hora et al., 2020; Tezel and Pavlostathis, 2015). QACs in the environment also pose toxicity to aquatic organisms like fish, algae, daphnids, and microorganisms (Zhang et al., 2015). If the consumption of QACs continues to increase without taking precautions, it will cause malfunctioning of the wastewater fallowed by deterioration of the environment and dissemination of antibiotic resistance. Therefore, eliminating QACs from the environment is significant.

Biodegradation is the main mechanism that removes QACs in the wastewater and the environment (Hora et al., 2020). Biodegradation of all of the QAC homologs is initiated by dealkylation reaction which removes the long alkyl chain containing 8 to 18 carbons from the central quaternary nitrogen atom of the QACs (Ertekin et al., 2017; Lang et al., 2010; Tezel et al., 2012; van Ginkel and Kolvenbach, 1991). The reaction results a tertiary amine and an alkanoic acid. After this reaction, the biocidal activity of QACs decreases substantially (Tezel et al., 2012). The enzyme that catalyzes the key dealkylation reaction is a Rieske oxygenase called *oxyBAC*. This enzyme was first identified in *Pseudomonas* sp. BIOMIG1^{BAC} and its activity was genetically confirmed (Ertekin et al., 2017). *oxyBAC* gene has also detected in the genomes of other strains such as *Pseudomonas*

nitroreducens B and *Pseudomonas saponiphila* DSM9751 that can degrade various QACs including cetyltrimethylammonium chloride and benzalkonium chlorides (Lang et al., 2010; Oh et al., 2014). As a result, *oxyBAC* and the bacteria carrying this gene have an important role in cleaning up highly toxic and antibiotic resistance promoting QACs in the environment. Therefore, understanding the taxonomy of *oxyBAC* gene hosting bacteria and mobility of this gene are crucial to evaluate the current fate and project the future risks related to QAC pollution in the environment.

Therefore, the main objective of this research is to understand the taxonomy of *oxyBAC* gene hosting bacteria and potential niche and behavior of *oxyBAC* gene in the bacterial population using genomics. Previously, the draft genome of strain BIOMIG1^{BAC} was obtained using Illumina platform (Ertekin, 2017). In this study, the complete genome of the strain BIOMIG1^{BAC} which mineralizes BACs and other QACs was generated using a hybrid method that combines Illumina short-read sequencing and Oxford Nanopore long-read sequencing technologies. With the hybrid methodology, the high-quality circular chromosome of the strain BIOMIG1^{BAC} was obtained. Moreover, various sequencing inputs and their effects on the genome assembly process were compared for validation and improvement of the results. Following, taxonomy of the strain BIOMIG1^{BAC} was determined using its complete genome. *Pseudomonas protegens* subgroup, which contains QAC degrading strains BIOMIG1^{BAC} and DSM9751, was critically analyzed using comparative genomics. In addition, the complete genome of the strain BIOMIG1^{BAC} was evaluated in detail to identify strain specific genomic regions and their role in the QAC biotransformation. Transposons, prophages and insertion sequences were identified in order to understand the role of mobile genetic elements in the potential mobility of the *oxyBAC* gene and other strain specific functional genomic regions in bacterial populations. This study is the first study that evaluates the ecological importance the *oxyBAC* gene and bacteria hosting this gene using comparative genomics and mobilomics at complete genome level.

2. COMPLETE GENOME SEQUENCE of *PSEUDOMONAS* SP. BIOMIG1^{BAC}

This chapter was partially published in "Microbiology Resource Announcements" entitled "Complete Genome Sequence of Pseudomonas sp. Strain BIOMIG1^{BAC}, Which Mineralizes Benzalkonium Chloride Disinfectants (Altinbag et al., 2020)" Disclosure of the article is permitted by Ertekin, E. and Tezel, U.

2.1. Introduction

Bacteria are very diverse group of organisms that play a significant role in biogeochemical cycles on earth, synthesis of valuable chemicals and treatment of wastes. With the advancements in the sequencing technologies, genomes of bacteria that are important for earth and society are obtained to understand the mechanisms of novel bacterial functions.

There are three main generations of sequencing methods. Each method has different advantages and disadvantages. First-generation sequencing relies on the separation based on the size of DNA fragments, its accuracy is high, the read length is moderate (between 800-1000 bp) and the throughput is low (Mestan et al., 2011). Second-generation sequencing, which is also known as next-generation sequencing (NGS), relies on sequencing by synthesis technologies. Its accuracy is high, the read length is short (between 100-500 bp) and the throughput is high (Mestan et al., 2011). Third-generation sequencing relies on single molecule real time sequencing. Its accuracy is low, the read length is high (greater than 1000 bp) and the throughput is high (Mestan et al., 2011). If the genome size is high or the genome has complex repeating sections, the whole genome assembly can be challenging with using just one sequencing method (Wick et al., 2017). Hybrid methods like the combination of different sequencing methods can be used to overcome bottlenecks of one method; as an example, second-generation sequencing has higher accuracy but lower read length and third-generation sequencing has lower accuracy but higher read length, so the combination of these two generations of methods can create whole-genome assemblies (Wick et al., 2017).

Third-generation sequencing technology, also known as single molecule real time sequencing (SMRT), is relatively new. The first third-generation commercial sequencer was released by Pacific Biosciences (PacBio) in 2011 and another one was released by Oxford Nanopore Technologies (ONT) in 2015 (Brown and Clarke, 2016). Lack of imaging equipment for detection of the nucleotides brings down the size to the small enough to be portable, one of the examples is the MinION Mk1B

sequencer with 90 g weight (Kono and Arakawa, 2019). There is no need for PCR and DNA synthesis, the protein nanopore placed in an electrically resistant membrane enables the single-stranded DNA to pass through inside of the nanopore and the sensor records the changes of the current in real-time (Kono and Arakawa, 2019). The universal serial bus (USB) connection is enough to power the MinION devices, therefore sequencing can be done anywhere if there is a computer to connect. There is no limitation about the read length, the user can pick the desired read length. Unlike the NGS which gives the nucleotides and their qualities (*fastq* format) as output, nanopore sequencing gives the raw signal file (*FAST5* format) as output. To convert the raw signals to the nucleotides, base-calling softwares which are using machine learning algorithms are needed (Kono and Arakawa, 2019). Although third-generation sequencing has lower accuracy, a complete bacterial genome can be assembled *de novo* by using extra computational work such as multiple alignments to correct reads of overlapping segments, assembly algorithms, and polishing the assembly with probabilistic models (Loman et al., 2015). One of the results of this pipeline which has just third-generation long reads for a strain of *Escherichia coli* has 99.5 nucleotide identity (Loman et al., 2015). However, testing the genome identity and sequencing assessment is not easy for novel species.

Constructing whole-genome sequences without any reference genome or sequence is described as *de novo* assembly and it is a challenging process (Sohn and Nam, 2016). The main idea behind the genome assembly is joining the overlapping reads to obtain longer read lengths. However, it is not that easy and there are still too many challenges to overcome. First of all, although the sequencing technologies have higher throughputs, their read lengths are very small comparing with the organism's genome (NGS average read length bp /average bacteria genome size bp = 0.00005), so higher coverage rates and higher number of reads are important for better assemblies (Mestan et al., 2011). One of the popular algorithmic approaches for genome assemblies is using the de Bruijn graph which is partitioning the short reads to the different *k*-mers and building a directed graph to simplify the optimal path search (Sohn and Nam, 2016). There are two different approaches for creating the de Bruijn graph; Eulerian approach where k-mers are the nodes and the Hamiltonian approach where k-mers are the edges of the graph. The problem of the Hamiltonian approach is that if the topology of the graph increases, the computational time dramatically increases almost to infinity because the finding Hamiltonian path is nondeterministic polynomial time (NP)-complete problem (Sohn and Nam, 2016). So, the simplification of the de Bruijn graph with the Hamiltonian approach is important to reduce complexity, but it gives shorter contigs. On the other hand, the advantage of the Eulerian approach is a polynomial-time problem that it does not need any simplification and performs better for big genomes (Sohn and Nam, 2016). The most popular assembler using the Hamiltonian approach is *Velvet* whereas the *SPAdes* is a popular assembler using the Eulerian approach (Bankevich et al.,

2012; Zerbino and Birney, 2008). A comparison study of different assemblers for Standard (ECOLI-MC) Dataset shows that *SPAdes* is the best for completing genes and creating the largest contig, the second in N50 and close to the first one in the number of contigs among 8 assemblers which includes *Velvet* (Bankevich et al., 2012).

There are four main challenges for the *de novo* assemblies such as correction of sequencing errors, uneven read depth, topological complexity because of the repetitive elements, and algorithmic complexity which brings computational cost (Figure 2.1). The GC or AT rich sections of the DNA can cause sequencing errors more frequently for short-read sequencing platforms and a high number of consecutive nucleotide regions can cause sequencing errors for long-read sequencing platforms (Sohn and Nam, 2016; Wick et al., 2017). Substitution errors of high throughput platforms per 100 sequenced base are 0.26 for Illumina HiSeq, 0.24 for Illumina MiSeq, 0.05 for 454 GS Junior and 1.10 for Pacific Biosciences RS and indel error rates are 0.02 for Illumina HiSeq, 0.009 for Illumina MiSeq, 0.39 for 454 GS Junior and 15.56 for Pacific Biosciences RS (Laehnemann et al., 2016). It can clearly show that long-read sequencing methods have almost 5 to 10-fold higher substitution errors and 50 to 100-fold higher indel errors. So, error correction of long-read sequencing is still a big challenge (Figure 2.1B). It can be thought as NGS is very accurate. However, the assembly graph can be affected by these errors also. Moreover, NGS errors are more probably agglomerating to some section of the genome, on the other hand long read sequencing errors are occurring more randomly (Laehnemann et al., 2016). Most assemblers have error correction implementation in their pipeline and *k*-mer counting method is one of the popular ones. The main idea behind it that the erroneous *k*-mers probably have low *k*-mer depth and high frequency, because an ideal amplification process more probably generates normal distribution (Sohn and Nam, 2016). If the assumption of the normal distribution does not apply to the genomic data, different approaches are needed. Multiple sequence alignments for correction to the consensus especially for the long read sequencing methods and using suffix arrays for correction are other approaches (Laehnemann et al., 2016).

Genomes have repeating regions that make the assembly problem harder. If a repeating region size is bigger than the read length of the sequencing technology, it is almost impossible to resolve this problem with *de novo* assembly (Figure 2.1C). If the read length is higher than the repeating region, the the repeating region can be successfully represented (Sohn and Nam, 2016). There are two approaches in Unicycler hybrid assembler to solve fragmented genomes and repeating regions: using short-read assembly to find out the anchor points of the contigs of long read assembly and bridging short-read assembly contigs with semi-global alignments of the long-reads (Wick et al., 2017). NGS and third-generation sequencing reads can be used together for those approaches.

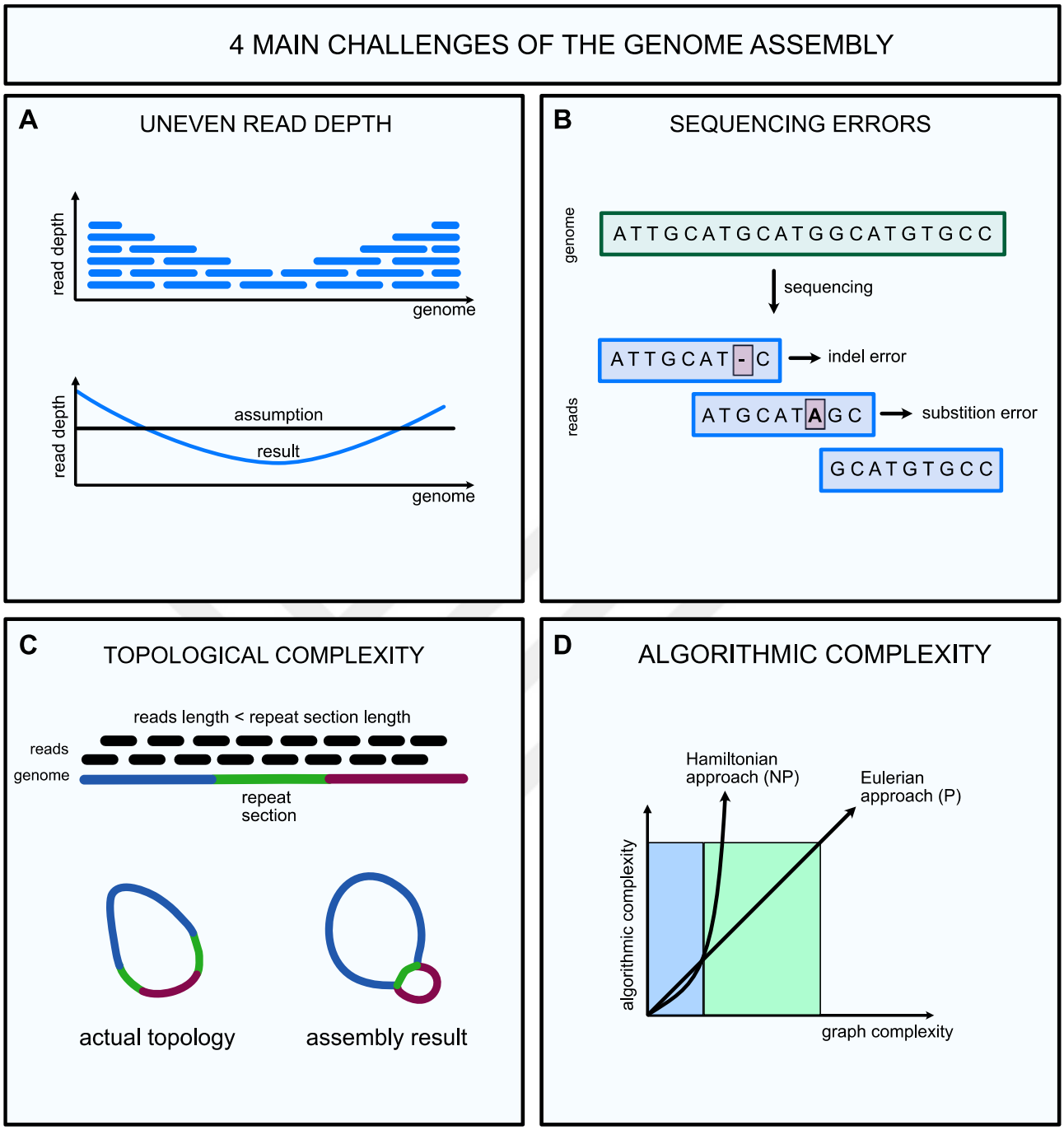


Figure 2.1. Four main challenges of the genome assembly (Sohn and Nam, 2016; Wick et al., 2017).

There are manual completion approaches to maximize assembly success. Assembler's main output type is the "contig file" that has sequences with the contig names, but the file is lacking information on the possible connections of the contigs. The assembly graphs can give probabilistic connections of the contigs, so the possible paths of the repeating sections can be identified. Also, sequencing errors can lead to false-positive connections especially for the polymorphic repeats (Miller et al., 2010). Another challenge is uneven sequencing depth which can cause gaps in the genome sequences, so the final assembly can be more fragmented because of the missing data (Figure 2.1A) (Sohn and Nam, 2016). The fourth challenge is the memory and time complexity of the

assembly approaches (Figure 2.1D). For the relatively small genomes like bacteria, the computational needs for the assembly can be provided with a regular computer which has at least 12 GB Ram and multi-core processor. However, it becomes challenging for bigger genomes like the human that some assembly programs need 512 GB memory for the human whole-genome assembly (Sohn and Nam, 2016). Interestingly, the assemblers with Hamiltonian approaches have lower computational cost than the Eulerian ones such as *ALLPATHS-LG*, which uses Eulerian approach for assembly, needs 3 weeks, 48 processors, and 500 GB of memory to complete the human genome, but *SOAPdenovo2*, which uses Hamiltonian approach, needs 3 days, 4 core processor and 35 GB memory (Sohn and Nam, 2016). The reason for that Hamiltonian one reduces the complexity of the graph by removing branches. Although this process gives better computational complexity, low sensitivity and higher fragmentation with lower N50 statistics are unavoidable (Sohn and Nam, 2016).

The sequence of the whole genome is like some randomly ordered letters without biological information about their meaning. The genes are the functional regions of DNA that are responsible for the synthesis of different products such as tRNA, RNA, and proteins (Waterman, 2018). Identification of gene regions and their products is crucial to understand their biological functions. There is a huge number of genes even in prokaryotic genomes, so there is a need for advanced computational tools for the detection and identification of those genes. One of the pipelines for gene annotation is NCBI Prokaryotic Genome Annotation Pipeline (PGAP). NCBI *PGAP* pipeline is using *GeneMarkS* algorithm for more precise gene prediction in prokaryotic genomes (Lomsadze et al., 2018). Also, *RefSeq* is a project which provides annotations of nearly 95000 prokaryotes that are used as a part of the pipeline (Haft et al., 2018). It is free software to annotate all the genes on the genome. The pipeline is containing also *RFAMs* for RNA family assignment, *PILER-CR* for identification of CRISPR repeats, *tRNAscan-SE* for finding tRNA genes in the genome, and *TIGRFAMs* for detecting protein families (Chan and Lowe, 2019; Edgar, 2007; Haft, 2003; Nawrocki et al., 2015; Selengut et al., 2007). Moreover, there are different genome annotation pipelines, one of them is *DFAST* (Tanizawa et al., 2018). Also, there is another tool called *Prokka* (Seemann, 2014). The main advantage of *Prokka* and *DFAST* versus *PGAP* is the speed. However, the results are less accurate concerning *PGAP* pipeline. For example, *PGAP* found 6243 CDS, *Prokka* found 5759 CDS, MiGAP found 5721 CDS, and *DFAST* found 5740 CDS in the genome of *E. coli* O26: H11 str. 11368 (Tanizawa et al., 2018). The speed difference between *MiGAP* is nearly 100 times slower than *DFAST* for the same strain of *E. coli*, and the values are respectively 4h 43m and 3m 27s (Tanizawa et al., 2018).

Topological complexities in the genome assembly graphs are related to repeating sections. It is significant to understand the reasons for the complexities to resolve them. One of the most significant factors in the formation of repeating sections is insertion sequences (ISs) because insertion sequences make copies of themselves in the genome (Vandecraen et al., 2017). ISs are small (about 3 Kb) transposons, and they are agents of horizontal gene transfer (HGT). Horizontal gene transfer (HGT) is the transfer of the genetic material without reproduction. Plasmids, bacteriophages, and transposons have a substantial role in the mobility of genes in a bacterial population, and they are the primary agents for horizontal gene transfer (Frost et al., 2005). ISs can move within the genome or between genomes (Vandecraen et al., 2017). They have their enzymes for independent transposition, and they are generally flanked by inverted repeats, which are binding sites of the transposase enzyme. Not just sequence, but the functionality of the sequences can be helpful to resolve complexities in assembly graphs for obtaining complete genomes.

The objective was to generate the complete genome of strain BIOMIG1^{BAC} by considering the impacts of the genome assembly challenges. Validation and further improvements of the genome assembly were performed with a comparison of different methodologies, algorithmic approaches, and inputs. The main idea was to use a hybrid methodology to combine long-read (Oxford Nanopore) and short-read (Illumina) sequencing for obtaining high-quality genome assembly. Most of the assemblers provide automatic pipelines; however, it is crucial to consider exceptional cases. Thus, computational and biological artifacts of the different approaches were evaluated, and assembly results for different inputs compared to show optimum parameters.

2.2. Materials and Methods

2.2.1. DNA Sequencing and Completing Genome with Hybrid Assembly

In this study, 10,757,826 paired-end short reads making a total of 1.1 Gb sequence data generated previously by the Illumina HiSeq 2000 platform were used. The draft genome of the *Pseudomonas* sp. BIOMIG1^{BAC} using that data set was 92.8% of the complete genome based on the estimation using the coverage on universal single-copy genes (Ertekin et al., 2017). Therefore, in order to get the whole genome, additional sequencing was performed on the SMRT technology platform called MinION from Oxford Nanopore Technologies (ONT; Oxford, England).

After two days cultivation of *Pseudomonas* sp. BIOMIG1^{BAC} cells with 200 mg/L BAC, EZ-DNA reagent (Bio Basic, Ontario, Canada) with ethanol precipitation and affinity column purification

(Macherey-Nagel) was used for isolation of the DNA (Altinbag et al., 2020). A sequencing kit (SQK-RAD004; ONT) was used to process approximately 400 ng DNA with the FLO-MIN106 flow cell (FAL49576). 2,681,861 long reads, which include a total of 8.5 Gb nucleotides with 30 Kb average read length, were generated by the MinION platform (Lu et al., 2016). The raw signal output of MinION platform (*FAST5* format) was converted to the nucleotides and their qualities file (*fastq* format) with *Guppy* (v. 3.4.4) base caller software (ONT) (Wick et al., 2019). As a configuration file, 'dna_r9.4.1_450bps_fast.cfg' was used because it was compatible with the flow cell. After this process, *Filtlong* (v. 0.2.0) was used to filter out the reads which have lower quality (<Q10), eliminate the reads lower than the threshold (<1000 bp), and shrink the total read length to the more suitable level for the assembly process (500 Mb).

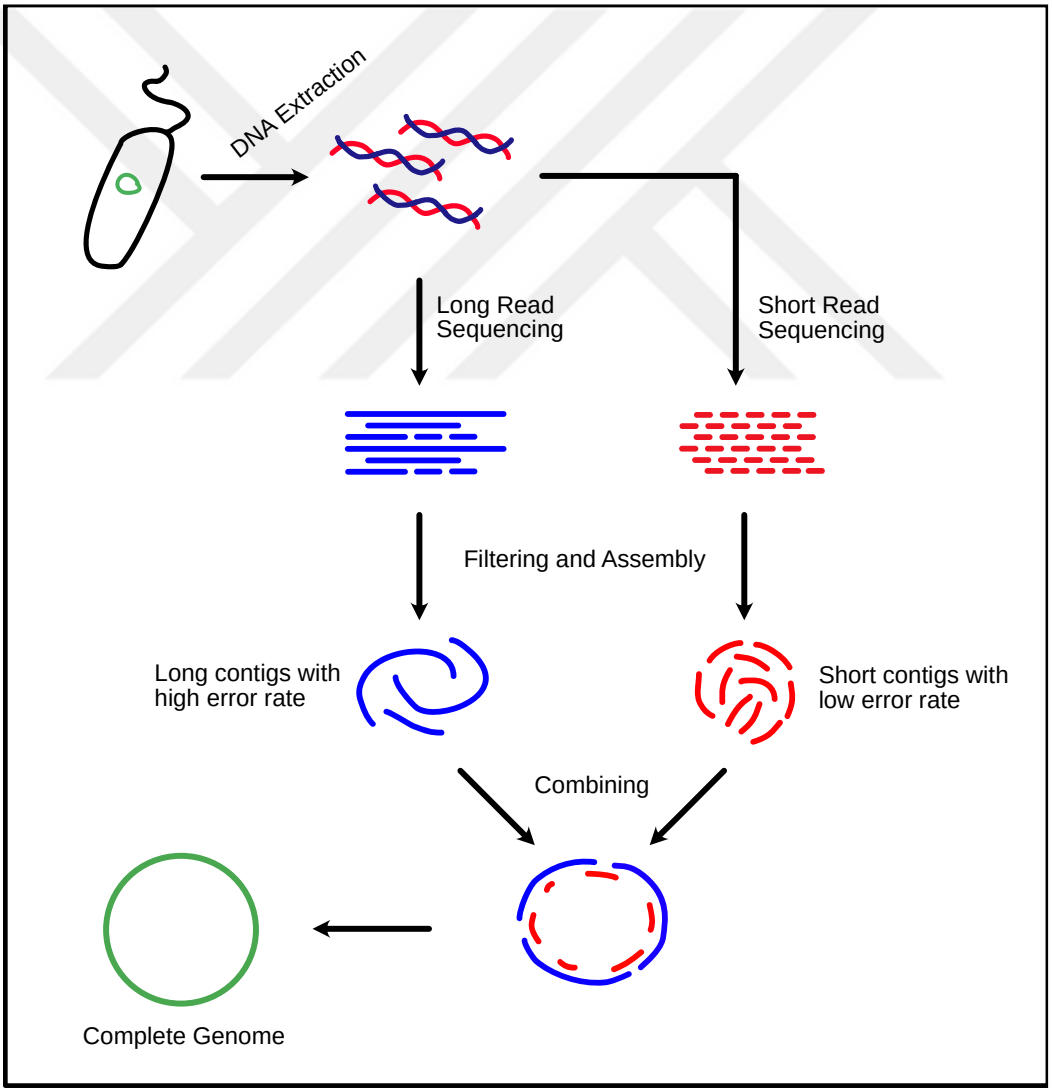


Figure 2.2. Hybrid assembly pipeline to generate the complete genome.

A hybrid assembly pipeline that uses long and short sequencing approaches was used to generate the complete genome of Pseudomonas sp. BIOMIG1^{BAC} (Figure 2.2). The two processes of the hybrid

assembler *Unicycler* (v. 0.4.8) are error correction and cleaning short reads before the assembly process (Wick et al., 2017). *SPAdes* (v. 3.11.1) was used as assembler of short reads with different *k*-mer options from 21 to 95 to find out the best *k*-mer which is maximizing the score function 1/(c*(d+2)) (where c is contig number and d is dead-ends in the assembly graph) (Bankevich et al., 2012; Wick et al., 2017). The best *k*-mer was 67 with 618 contigs and 3 dead-ends. Even the higher *k*-mer number creates a lower number of contigs, a small dead-end number is also important. *Unicycler* (v. 0.4.8) was used *Miniasm* (v. 2020-01-13) to assemble the filtered nanopore sequencing reads (28,128 long reads, 500,002,727 bp). Long read assembly was ended with 553 segments and 552 links that created 3 linear unitigs with the total 7,722,557 bp length. Polishing processes were done with *Pilon* (v. 1.23) and *Racon* (v. 1.4.3) to enhance the assembly. In the end of the hybrid assembly pipeline, *Unicycler* (v. 0.4.8) created a circular chromosome with a length of 7,675,262 bp and G+C content with 62.1%.

2.2.2. Comparative Analysis of Assemblers and Different Inputs

The short-read data generated by Illumina platform was used to compare different assembly approaches. There are two main algorithmic approaches (Eulerian and Hamiltonian) for the genome assembly. The effects of these approaches were compared with using two different assemblers: *Velvet* which is using Hamiltonian and *SPAdes* which is using Eulerian. The statistics of the *Velvet* (v. 1.2.10) short-read assembly were taken from the previous work (Ertekin et al., 2017). Also, *SPAdes* (v. 3.11.1) short-read assembly was used with default settings. The results of *SPAdes* and *Velvet* were compared with a table.

Furthermore, the sequencing methods can affect the genome assembly process. It is a fact that long-read technologies like MinION platform create noisy genomic data. Thus, the filtering according to the quality and quantity is very crucial. Long-read data generated by MinION platform was analyzed. From the long-read data, different inputs (50, 100, 250, 500, 750, 1000, and 1250 Mb long-read inputs filtered by *Filtlong* (v. 0.2.0)) were prepared with selecting the quality threshold Q10. *Miniasm* (v. 2020-01-13) long-read assembler in the *Unicycler* (v. 0.4.8) was generated genome assemblies from different long-read inputs (Li, 2016). *QUAST* was used for quality control of the different assembly results. The effects of the different inputs to the assembly process was shown. Bandage (v. 0.8.1) was used for visualization of the assembly graphs (Wick et al., 2015). Circos (v. 0.69-8) was used for the visualization of the comparisons (Krzywinski et al., 2009).

2.3. Result and Discussion

2.3.1. Completing the Circular Chromosome with Hybrid Assembly

Five hundred Mb long reads from the MinION platform and 1.1Gb short reads from Illumina HiSeq 2000 platform are combined with *Unicycler* to complete the genome. The total coverage of the reads was 135x which was enough to obtain complete genome of the *Pseudomonas* sp. BIOMIG1^{BAC} (Figure 2.3). The assembly yielded a single circular chromosome with 7,675,262 bp length. Linear chromosomes or plasmids were not detected. However, at the end of the assembly, there are two components that are excluded from the genome. One of them is a very tangled 31,111 bp component aligned in 6 contigs, and another one is 2,383 bp contig. The explanation to them will be made later. The circular chromosome was uploaded to the NCBI and it was announced previously (Altinbag et al., 2020). The GC content of the chromosome was 62.1%. There are 7,236 genes, 7,146 protein coding genes, 5 16S rRNA, and 72 tRNA in the genome. Also, one CRISPR locus is present.

2.3.2. Validation and Refinement of Genome Assembly

Although the hybrid assembly successfully created a circular chromosome, validation and further improvement of the assembly were performed manually using assembly graphs, read extractions with mappings, analysis of the assembly steps, and comparison of the different inputs. In addition, the contigs excluded from the genome were evaluated to understand the role of both computational and biological artifacts in bacterial genome assembly.

A graph is constructed during the assembly process to find the possible paths for combining the reads. The assembly graph produced with just short-reads can be seen in Figure 2.3. Dead-ends are the nodes which have only one connected edge in the graph. Bacterial chromosomes and plasmids are generally circular, so the graph without dead-ends is crucial for complete circularization. However, the number of dead-end may increase because of sequencing errors, different algorithmic approaches, and limited factors of the sequencing technologies. One of the factors that affect dead-end number is *k*-mer length. The assembly graph is created according to the substring length called *k*-mer and *k*-mer length affects both dead-end and contig number.

There is a trade-off between contig number and dead-end. If there are only short-reads, lower contig number is favorable because it will create longer contigs even the increase of the dead-ends. However, in the hybrid assembly process, the lower dead-end number is also very important because

it is not wanted to decrease the connectivity of the graph. It will increase success when long reads are involved. *Unicycler* is trying different *k*-mer inputs to find the optimum point between dead-ends and contig numbers. The optimum point for the short assembly process was *k*-mer as 67 that created 618 contigs and 3 dead-ends (Figure 2.4).

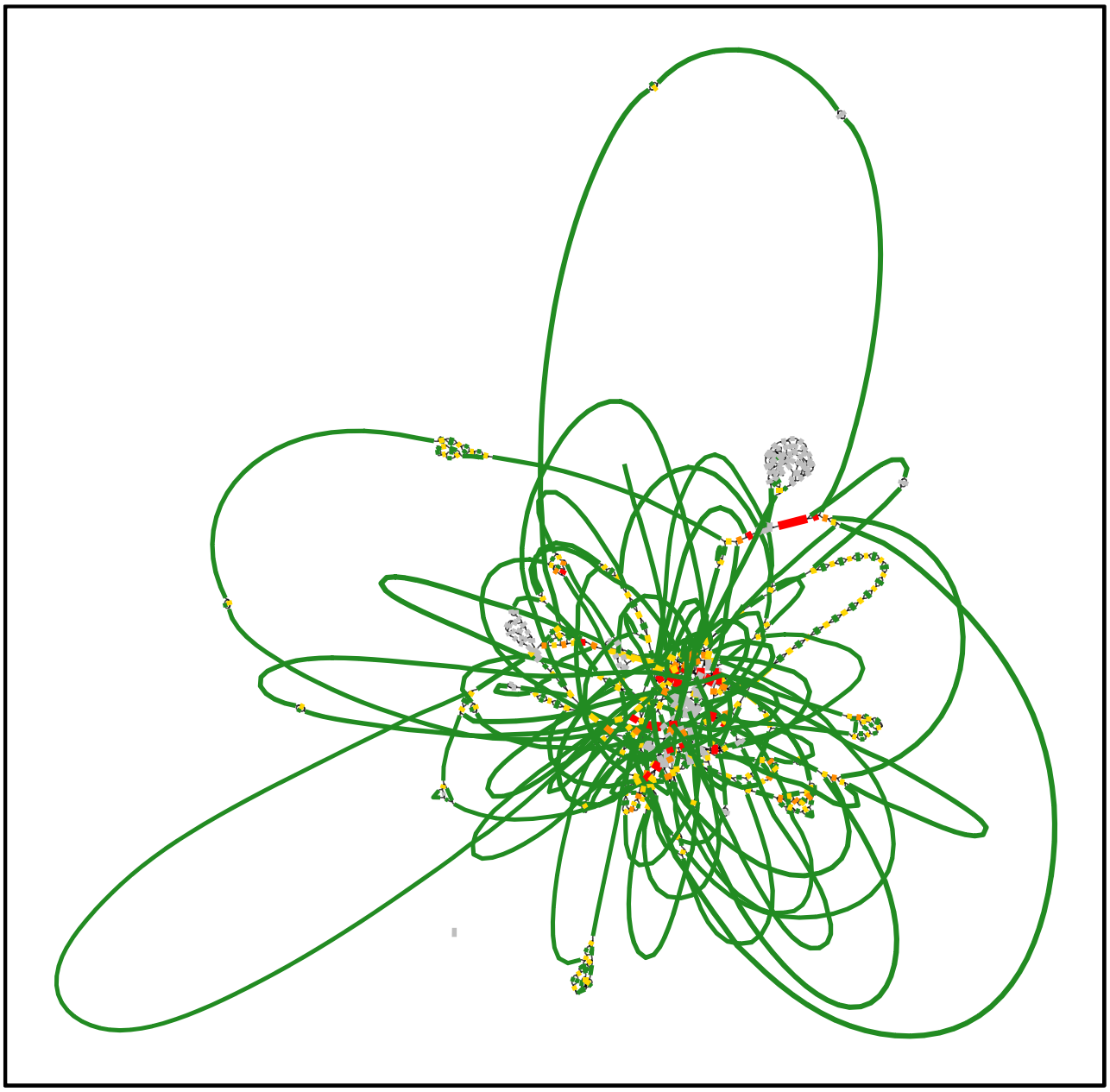


Figure 2.3. The assembly graph produced by *SPAdes* with short-reads before the hybrid assembly. The existence of the repeating sections is one of the major reasons for topological complexity.

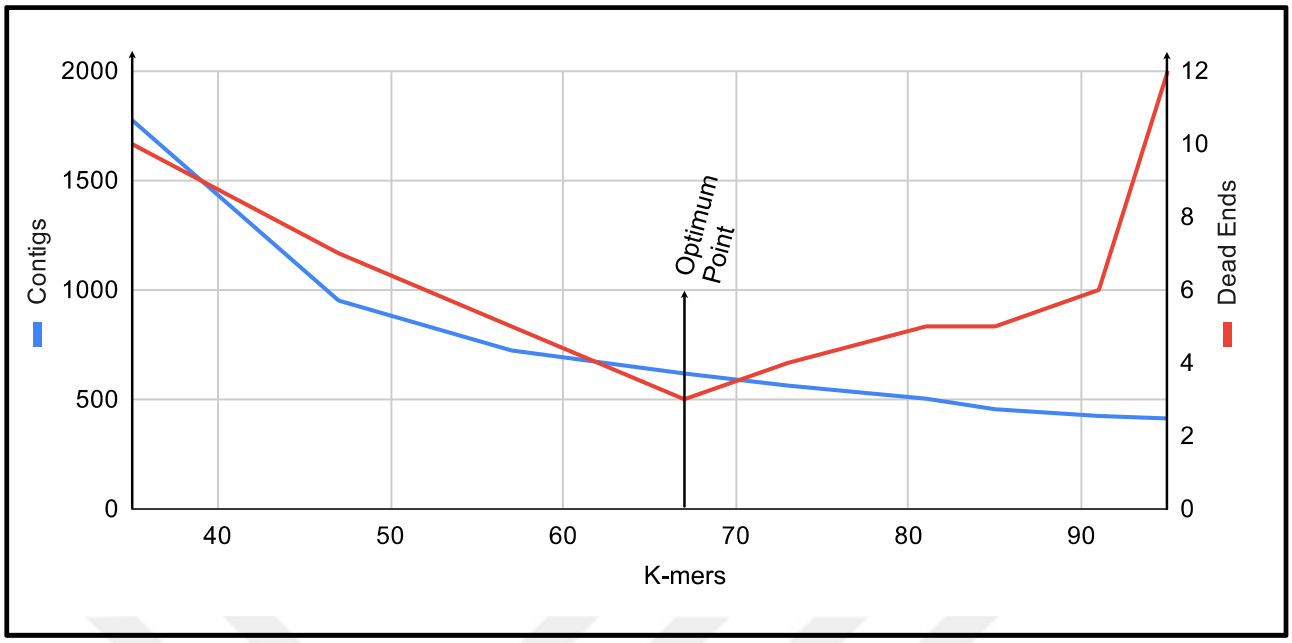


Figure 2.4. Optimum point of contigs and dead-ends for the hybrid assembly.

Ideally, the number of dead-ends must be zero, but the reads are not always equally distributed to the genome, or the quality values are not well enough to represent some regions. Furthermore, the contig number is decreasing with increasing *k*-mers, because the repeating sections are unraveling. Therefore, repeating sections affect the topological complexity and create a higher contig numbers. One of the reasons for the repeating sections is the insertion sequences because insertion sequences make multiple copies of themselves in the genome. A circular graph is made to compare the result of the hybrid assembly, insertion sequences, and the draft assembly to control the effect of the insertion sequences on the genome assembly (Figure 2.5). According to the graph result, 92 out of 96 insertion sequences are correlating with the borders of the contigs in the draft genome. There are very few insertion sequences that are not correlating with the border of the contigs because they have single-copy. Consequently, insertion sequences have a huge effect on the assembly of this genome, and they are one of the reasons for the high level of fragmentation in the draft genome assembly.

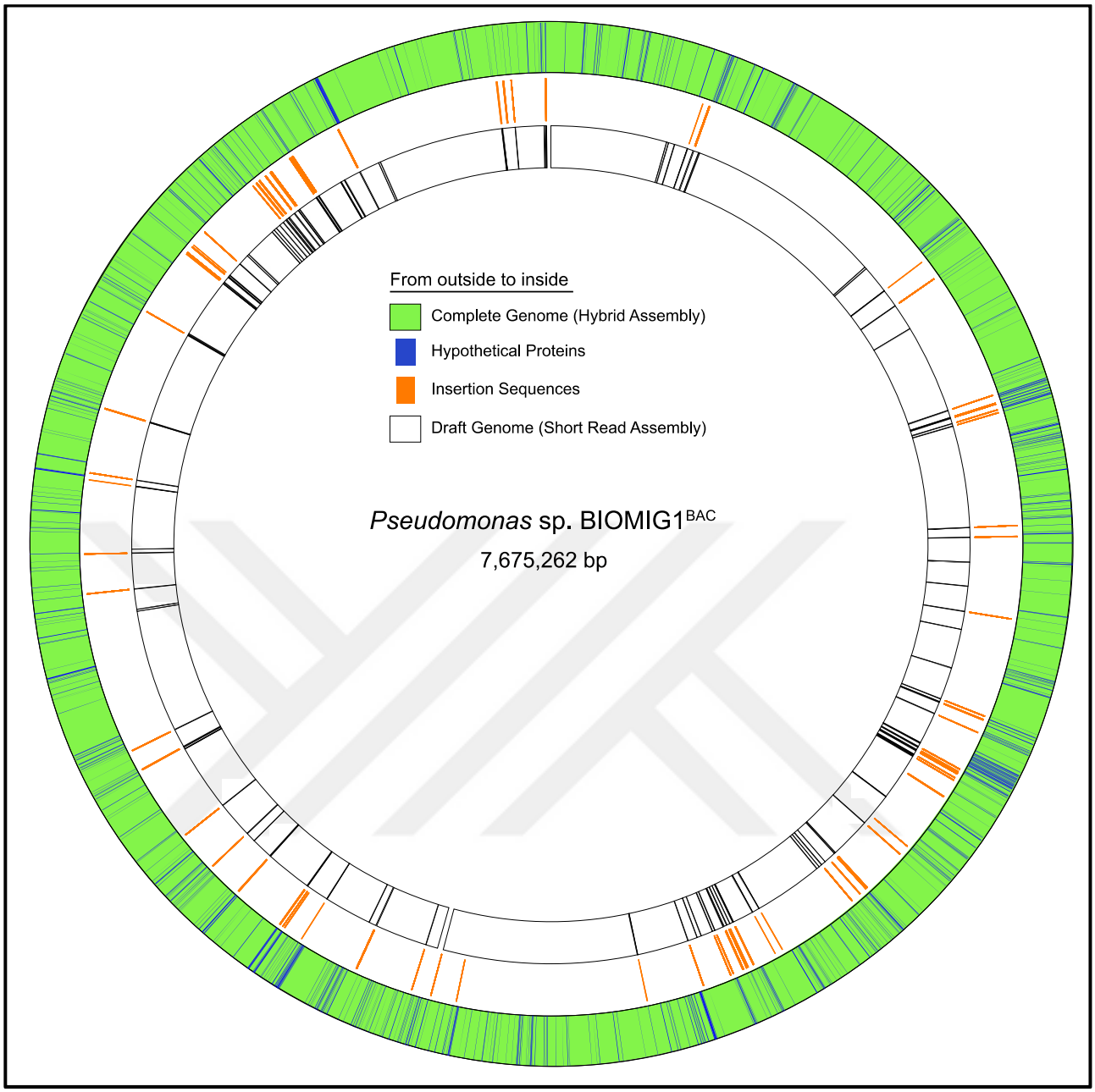


Figure 2.5. The comparison of the complete genome with hybrid assembly and the draft genome with short read assembly. The insertion sequences and the border of the contigs from the draft genome have a correlation.

Hybrid assembly output was good enough for completing the chromosome, but two components were excluded from the chromosomal alignment. The first one (Component-1) was composed of 6 contigs with a total length of 31,111 bp. The second one (Component-2) was composed of 1 contig with a total length of 2,832 bp. To find out why there were such components in the assembly outputs, assembly graphs were investigated with functional analyses.

Firstly, it is identified that Component-1 has 7 insertion sequences in it with functional analysis. Insertion sequences are one of the reasons for repeating sections which create topological complexity. Long-reads can resolve topological complexities, if there are enough number of reads. So, long-reads which covered the Component-1 were extracted with *minimap* (v. 2.17-r941) and *samtools* (v. 1.10), and the extracted reads were used to resolve the topological complexity. The mapped long-reads successfully circularized the Component-1. A circular sequence of 34,217 bp with 1.42x depth was created. The Component-1 was identified as a translocatible unit (TU) (Figure 2.6). It was not a plasmid because it did not have any origin of replication sequence and antibiotic resistance genes. TU's sequence has 7 insertion sequences which make up the 30% of its size. 18 hypothetical protein-coding genes and other annotated genes are scattered among insertion sequences. According to the *ISEScan*, there are two clusters of the IS91 family insertion sequences and the naming was made to distinguish them as IS91c1 and IS91c2 (Figure 2.7). IS91c1 (3204 bp), which is colored orange, is also found in the chromosome with 100% identity. It has 50 bp inverted repeats (IRL and IRR) and 2916 bp open reading frame, which is coding the transposase enzyme. Most of the Tn-3 type transposons have the resolvase gene, but IS91 lacks this gene. It is forward strand in the TU, but it is reverse strand in the chromosome. In the chromosome, it is flanked by 5 bp direct repeats that are clues of the insertion of IS. It probably was in the genome as a composite transposon structure; then, something may have triggered its separation. It can be good to research the reasons for the excision or insertion by finding the metabolic importance of this TU.

Secondly, Component-2 does not have an insertion or repeating sequence on it. It has a partial gene. Component-2 is composed of one contig. In the hybrid assembly graph, it is alone with no connections (Figure 2.8). However, in the short-read assembly graph, there is a path between C156, Component-2, and C199. So, according to the short-read assembly graph, it must be in the genome. It can be thought of as an artifact caused by the long reads or hybrid assembly, but it is not. Searching the sequence shows that Component-2 is divided into two parts in the hybrid assembly with an IS. The only explanation for this situation is that the genome changed during the long and short sequencing processes with an IS insertion. It is another indicator that the genome is very elastic. The IS is also found in other regions of the genome. The short reads are mapped to the Component-2 are investigated to confirm the insertion process. The 4 bp insertion site was found at the intersections of the left and right aligned reads to the region. Direct repeats confirm the occurrence of the insertion process. Insertion sequence size is 1213 bp, it has inverted repeats, and it has 1006 bp transposase coding gene. As a result, Component-2 can be disregarded, and it is not an artifact of the assembly process.

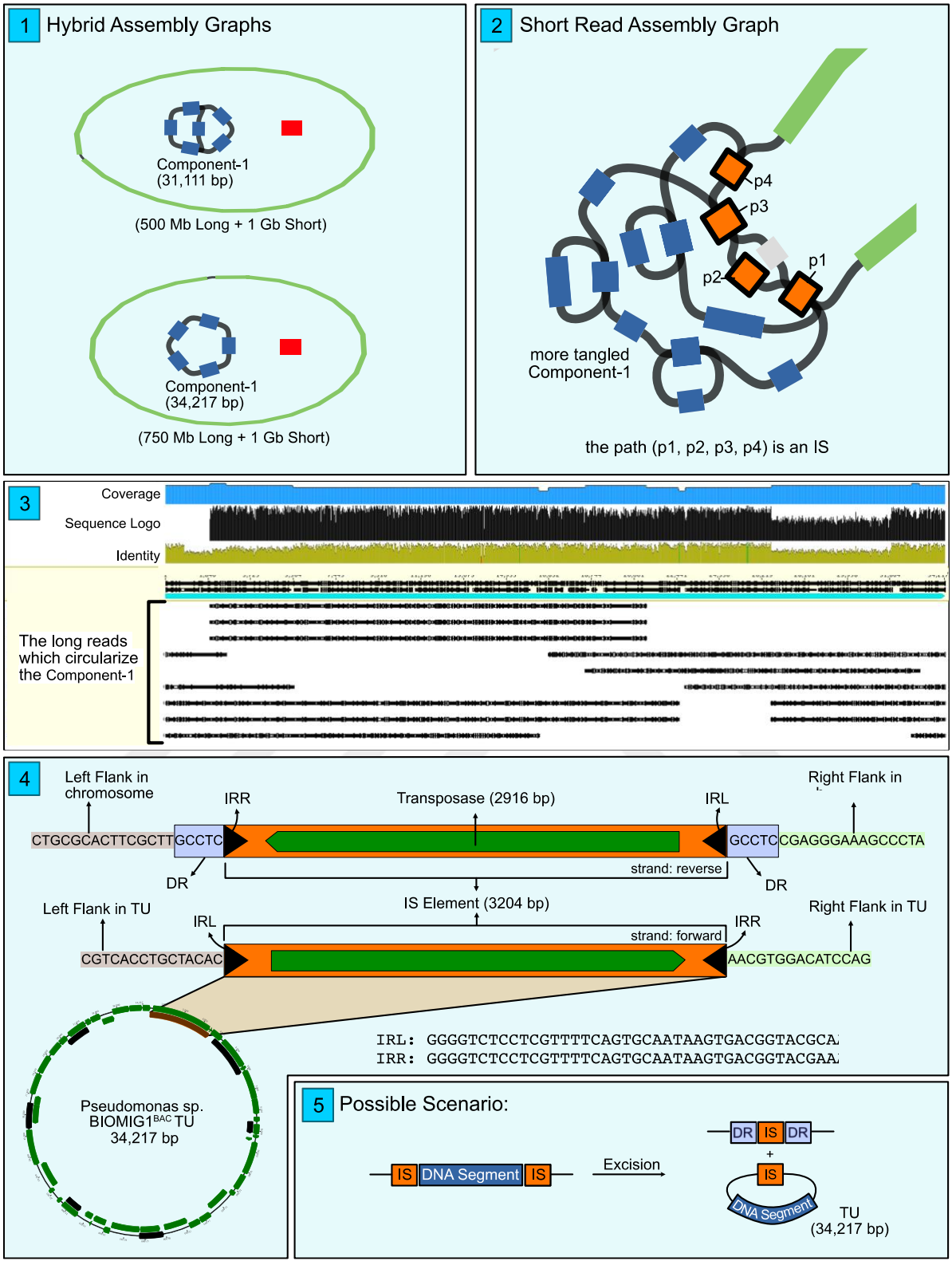


Figure 2.6. (1) Hybrid assembly results of the different long read inputs show that Component-1 is circularized with long reads. (2) Short read assembly shows the exact path (The path is an IS) that connects the Component-1 to the chromosome, but it is not reliable because of insufficiency of short reads about repeating regions. (3) The long reads are successfully circularizing the Component-1 which is identified as TU. (4) IRL, IRR, DR, left flanks, and right flanks are showed. (5) The TU probably was in the genome previously, or it can be in the future.

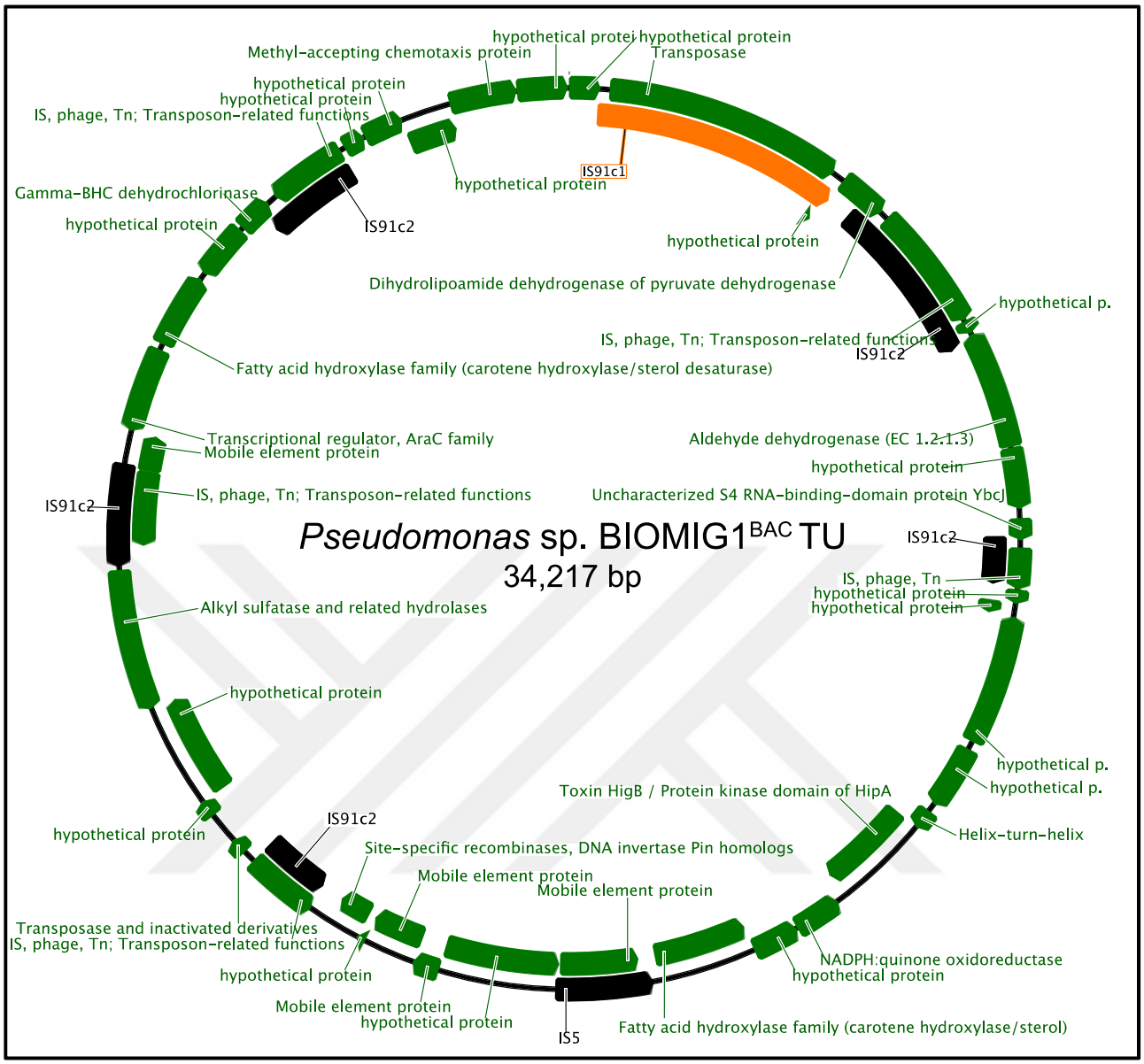


Figure 2.7. Annotated genes (green) of the TU with the insertion sequences (black) are drawn. IS91c2 (orange) was found in the chromosome; it can be effective for the insertion and excision.

Polishing is an essential step for improving the assembly. However, the probability of occurrence of false positive and false negatives changes must be considered. *Unicycler* is making polishing as a final process in its pipeline. In this process, it is using short reads as a reference because short reads have very good accuracy. The process was repeated until the time when there are no changes. It finished after 9 runs with total 261 SNP changes, 35 insertions, and 14 deletions. It is very hard to be sure that the polishing results are true positives and true negatives. Anyway, these changes do not have a high effect on the result. There are mainly SNPs, tiny insertions, and small deletions.

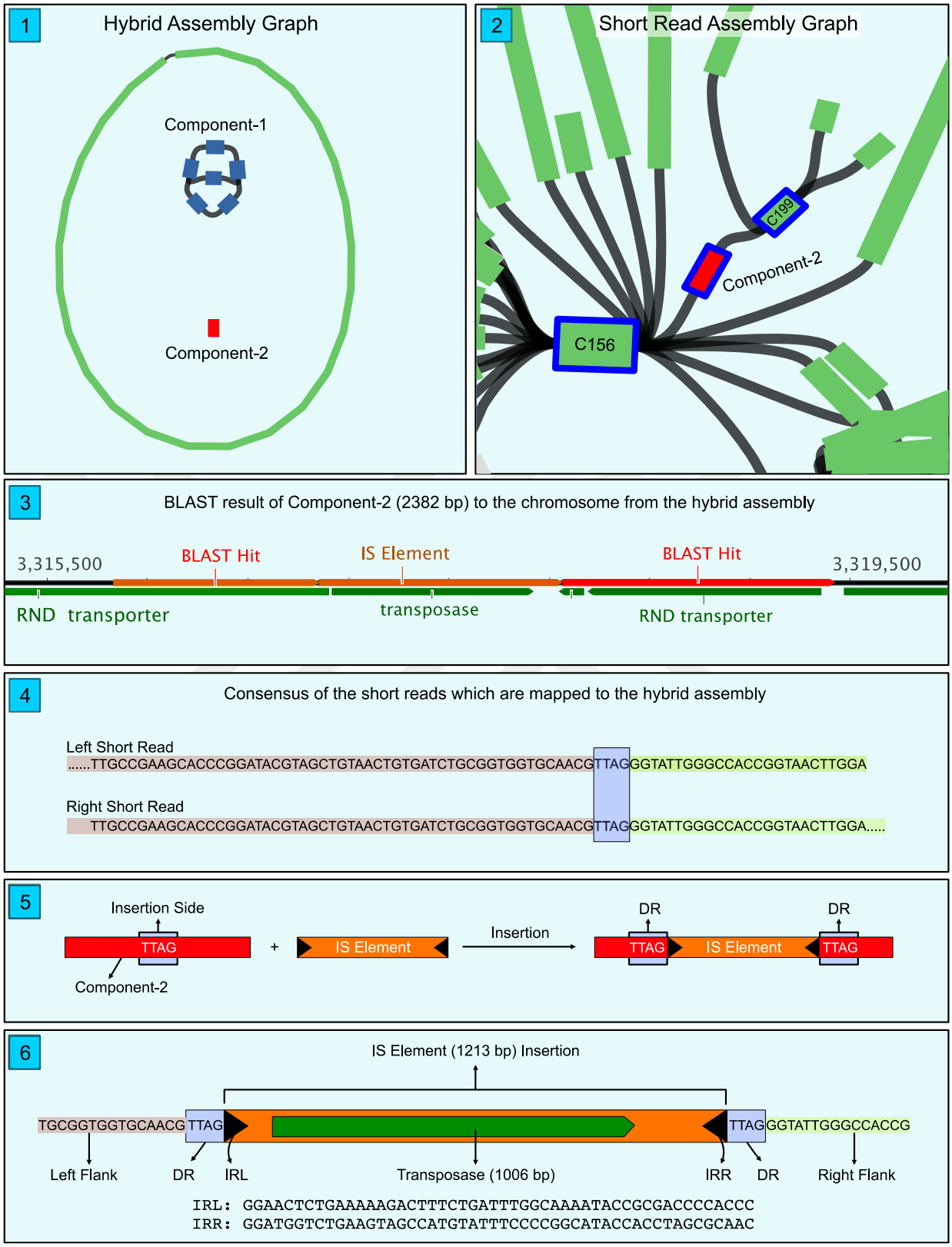


Figure 2.8. (1) Component-2 is unconnected in the hybrid assembly graph. (2) Component-2 is connected in the short-read assembly graph. (3) Blast result of the Component-2 shows that an IS element was divided it, so actually, there is no Component-2 in the genome. It was changed. (4) Left and right flanks of the mapped short reads. (5) The explanation of the situation. (6) The IS element (1213 bp) with IRL, IRR, DR, and flank sequences. It has one transposase with a 1006 bp length.

2.3.3. Comparison of Different Assembly Inputs

The polished and manually improved hybrid assembly with a combination of two different assemblies (500Mb long reads with 1 Gb short reads to build the chromosome and 750 Mb long reads with 1 Gb short reads to resolve uncompleted parts) was used as a reference to analyze the effects of different inputs and methods to the quality of the assembly. Because *Pseudomonas* sp. BIOMIG1^{BAC} is a novel species, and there is no reference genome to assess the exact quality of the assembly process, the combination of different inputs is used to create the reference assembly for the comparison. Nanopore represents the long reads, and Illumina represents the short reads. In all hybrid assemblies, the size of total short reads is 1 Gb. Using nanopore sequencing gives very good contig numbers even in the 50 Mb input; however, the genome fraction is just 57.038% in the 50 Mb with very high mismatches and indels (Table 2.1). With increasing input, the genome fraction, the number of misassemblies, the duplication ratio, and total length increase. On the other hand, the number of indels and the number of mismatches per 100 kbp decrease until the optimum point; then, they start to increase because of the accumulation of the errors in the reads.

The assembly results obtained using only Illumina reads, are very good for the number of mismatches and indels per 100 kbp. The number of N's per 100 kbp and number of contigs is not good in Velvet compare to SPAdes. Using the Hamiltonian approach and decreasing the graph complexity leads to the loss of many connections. Their duplication ratio and the number of misassemblies are better than the Nanopore. Although they give around 5 to100 fold lower error rates, they are not enough to represent the complete genome.

LA50, LA75, and the number of indels per 100 kbp are best for 500 Mb, and the number of mismatches per 100 kbp is best for 250 Mb among Nanopore sequencing. The expected output is that the average quality of long reads is not good and higher input means more errors. So, there is a trade-off relation and picking the optimum input is crucial for better assemblies (Figure 2.9).

Hybrid assemblies give excellent results for almost all parameters. There are tiny differences among hybrid assemblies. The largest contig from the 500 Mb* (Hybrid) and 750 Mb (Hybrid) is very close to each other; the size difference is just 3.5E-6%. The line graph for only long, only short and hybrid read assemblies can be seen in Figure 2.10.

Table 2.1. Comparison of the different technologies, assemblies and different inputs. The reference was created with the combination of 500 Mb and 700 Mb hybrid assemblies with the improvements mentioned above. 500 Mb (Hybrid)* is the submitted assembly to the NCBI database previously. In all hybrid assemblies, the used short reads are identical and 1 Gb.

Assembly methods	**Nanopore**					
	Unicycler (miniasm)					
Assembly Input	50 Mb	100 Mb	250 Mb	500 Mb	750 Mb	1000 Mb
# contigs	37	21	6	7	10	14
Largest contig	385455	944374	2463558	3870771	3871741	3571572
Total length	4420565	7166011	7535490	7870374	8055165	8233817
GC (%)	62.56	62.42	62.14	62.03	61.99	61.96
# misassemblies	12	22	27	29	34	39
Genome fraction (%)	57.038	92.57	96.684	98.962	99.513	99.548
Duplication ratio	0.999	1.004	1.011	1.032	1.049	1.07
# N's per 100 kbp	0	0	0	0	0	0
# mismatches per 100 kbp	172.11	34.92	8.37	15.09	25.56	29.43
# indels per 100 kbp	405.19	244.48	124.58	121.35	125.45	135.62
LA50	11	6	4	3	4	4
LA75	23	12	8	7	8	10
Assembly methods	**Nanopore**	**Illumina**		**Nanopore + Illumina**		
	Unicycler (miniasm)	**Velvet**	**SPAdes**	**Unicycler (miniasm + SPAdes)**		
Assembly Input	1250 Mb	1 Gb	1 Gb	500 Mb (Hybrid)*	750 Mb (Hybrid)	1000 Mb (Hybrid)
# contigs	13	316	113	5	5	5
Largest contig	3573185	275178	591200	7675262	7675279	7644056
Total length	8205582	7509453	7548207	7708330	7711403	7680180
GC (%)	61.99	62.2	62.19	62.1	62.1	62.09
# misassemblies	39	4	1	1	3	4
Genome fraction (%)	99.45	97.301	97.812	99.954	99.969	99.564
Duplication ratio	1.07	1.001	1	1	1.001	1.001
# N's per 100 kbp	0	24.6	1.32	0	0	0
# mismatches per 100 kbp	29.96	2.6	14.4	3.98	0.67	1.16
# indels per 100 kbp	133.35	0.67	0.61	1.13	0.86	0.59
LA50	5	40	12	1	1	1
LA75	10	86	28	1	2	2

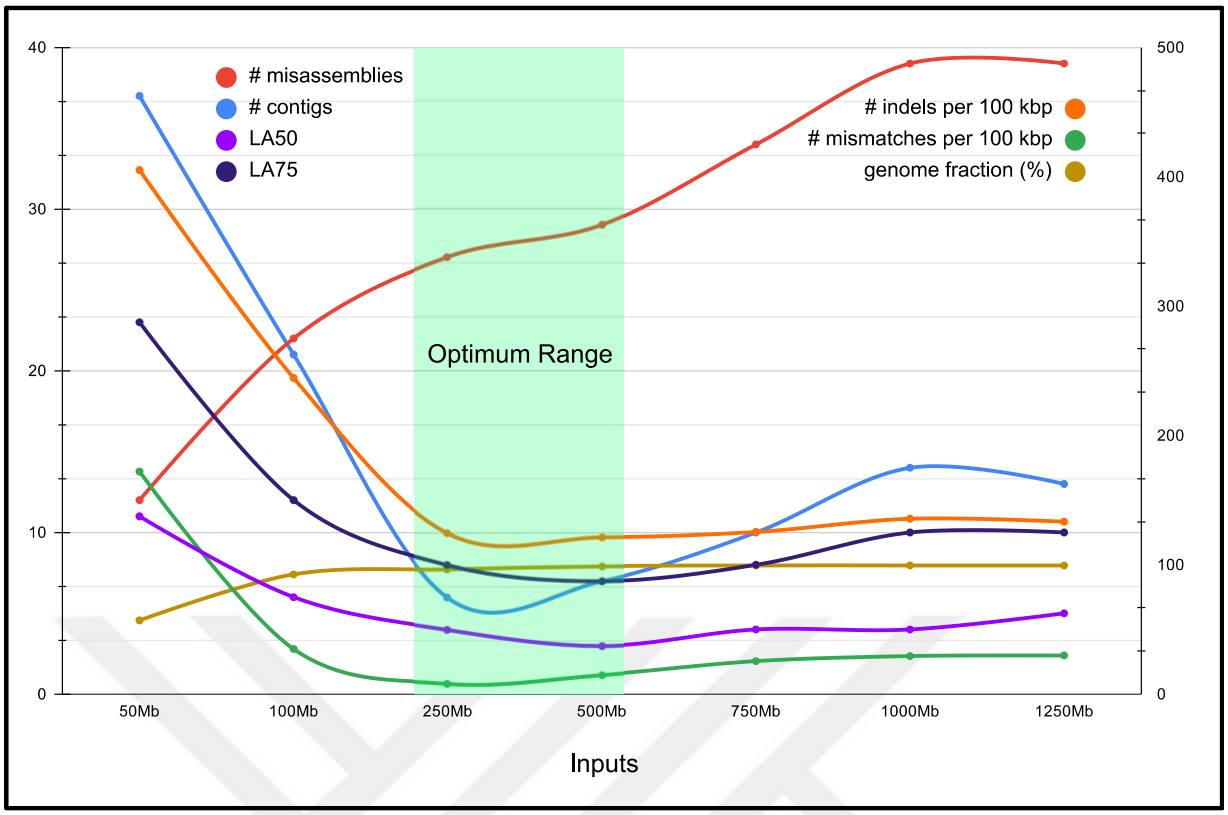


Figure 2.9. The comparison of different Nanopore (long read) inputs and their effects to the assembly quality. The optimum range is around 250 Mb and 500 Mb.

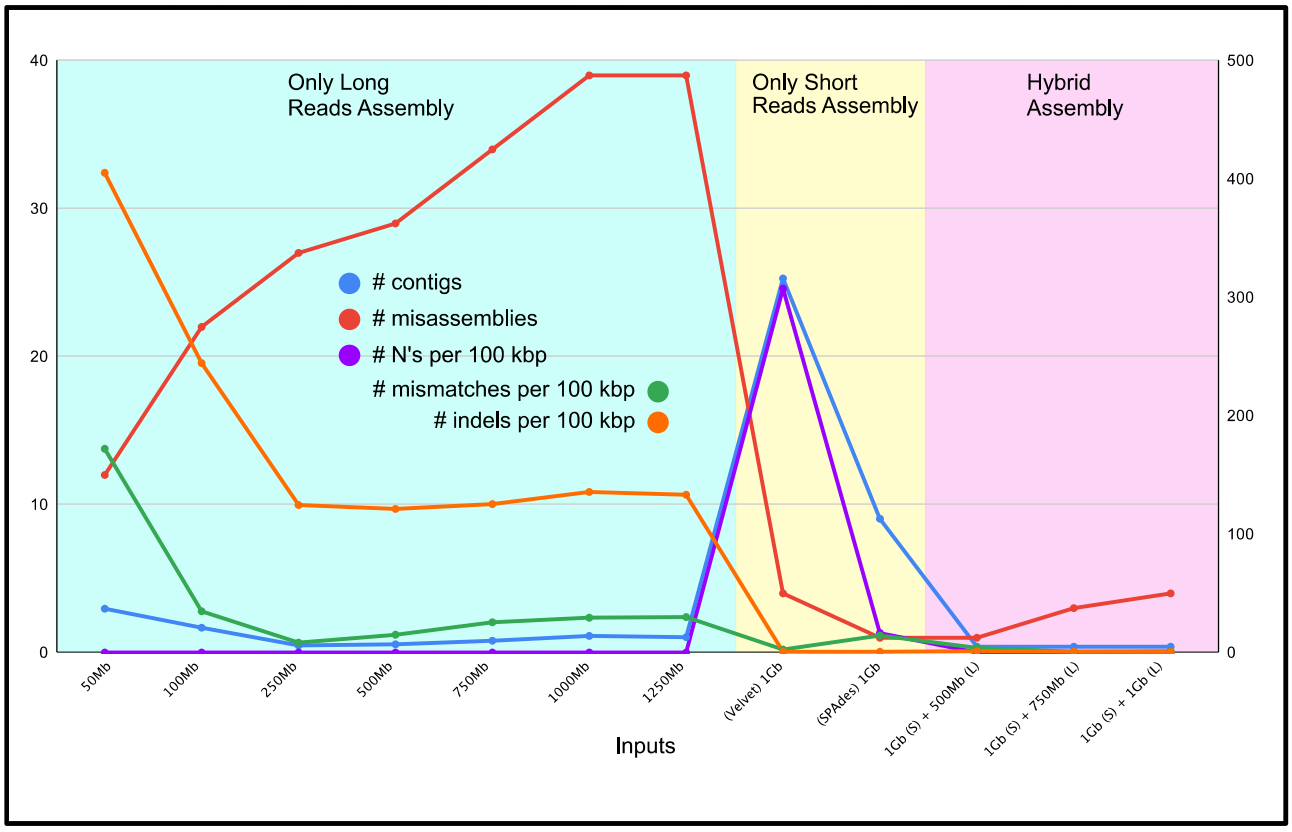


Figure 2.10. The comparison of different assembly inputs and their effects to the assembly quality. Hybrid assemblies are the best among them.

As a result, the genome of *Pseudomonas* sp. BIOMIG1^{BAC} was completed using a hybrid method that combines long (Oxford Nanopore) and short-read (Illumina) sequencing. A single circular chromosome with a length of 7,675,262 bp was generated with the hybrid assembly. There were no plasmids or linear chromosomes. As a comparison, only long-reads and only short-reads were used separately to present bottlenecks of each one. The hybrid method performed the best in almost all parameters. Manual curation based on assembly graphs increased the quality of the genome assembly. At the end of the assembly, there were two components excluded from the chromosome. After manual curation, one of the excluded component (Component-1) was identified as a translocatable unit and successfully circularized. Besides, second excluded component (Component-2) was eliminated after the investigation based on assembly graphs. To conclude, parameter selections, data filtering, functional analysis, and manual curations were key points for generating and validating the assembled genome of the strain BIOMIG1^{BAC}.

3. TAXONOMY of *PSEUDOMONAS* SP. BIOMIG1^{BAC}

This chapter was partially prepared to publish under the title "*Pseudomonas alexanderii* sp nov., an emerging *Pseudomonas protegens* subgroup of bacteria that degrades quaternary ammonium disinfectants, revealed by genome taxonomy".

3.1. Introduction

Recent bacterial taxonomy to classify species is based on the polyphasic approach, which is the combination of phylogenetic, genotypic, and phenotypic analyses (Hayashi Sant'Anna et al., 2019). To determine the phylogenetic classification of bacterial strains, traditional methods like DNA-DNA hybridization (DDH) and 16S rRNA similarity are used (Chun and Rainey, 2014). DDH is technically difficult, needs too much time, expensive, and only compare to genomes. Therefore, it is only used while defining a new species. On the other hand,16S rRNA molecule has long been used for phylogenetic classification of bacteria at species level if complete 16S rRNA sequence is obtained. (Hayashi Sant'Anna et al., 2019; Vinatzer et al., 2016). 16S rRNA can be used as a first step for taxonomical analysis, but it is not enough to reveal exact taxonomy (Chun and Rainey, 2014). It is recommended to use DDH if the 16S rRNA similarity between two genomes is higher than 97% (more recently 98.7%) to define a new species (Figure 3.1) (Oren and Garrity, 2013). A new species is affiliated only if DDH of its genome to the closest phylogenetic neighbor species is lower than 70%. Other methods that are considered fundamental like biochemical and fatty acid profiling can bias the taxonomy of bacteria, because there are not only lack of quantitative results but also inadequate for obtaining phylogenetic relations (Hayashi Sant'Anna et al., 2019). Another traditional method for taxonomy is the multilocus sequence analysis (MLSA) which is helpful to find out the differences between closely related strains and prevents the limitation of the single gene phylogeny (Hayashi Sant'Anna et al., 2019).

Recently, improvements in the sequencing technologies, besides the computational support, provided new methods to reveal the taxonomic classification of bacteria (Chun and Rainey, 2014). Next-generation sequencing (NGS) and third-generation sequencing have provided high throughput data to obtain the whole genomes of bacteria. Now, genome-based comparative analyses to determine more detailed phylogeny are feasible. Whole-genome comparative methods give powerful taxonomic classification which is called phylogenomics (Lalucat et al., 2020).

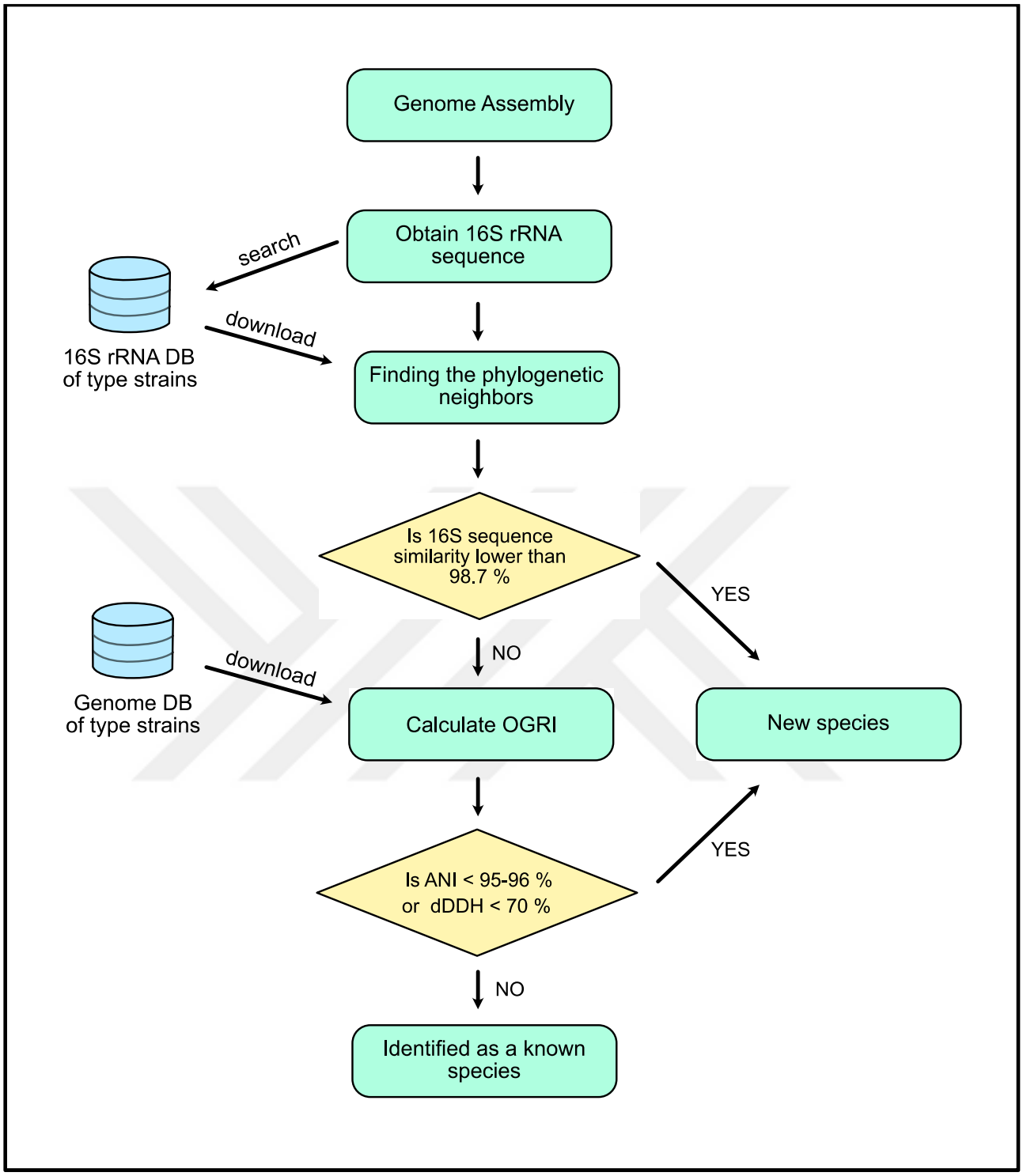


Figure 3.1. Flowchart of the genome based classification to define species (Chun et al., 2018). Overall genome relatedness index (OGRI) represent the similarity of two genomes. There are different examples of OGRI as ANI, dDDH, GGDC and Mash.

Average nucleotide identity (ANI) and genome to genome distance calculation (GGDC) are whole-genome comparisons based on sequence alignments that can be used instead of DDH (Konstantinidis and Tiedje, 2005; Lalucat et al., 2020; Meier-Kolthoff et al., 2013). ANI and GGDC analyses enables high resolution genome comparison with publicly available databases containing

genomes of thousands of bacteria which is faster than DDH (Lalucat et al., 2020). Seventy percent cutoff of DDH for species definition corresponds to the 94% cutoff for ANI (Konstantinidis and Tiedje, 2005). More recent work, which was used all prokaryotic genomes available in NCBI to compare ANI values with taxonomy, showed that 95% ANI is the determinative threshold for species with a higher rate of successful classification (Jain et al., 2018). In addition, considering the single-copy and housekeeping genes prevents the biases due to horizontal gene transfers (Hayashi Sant'Anna et al., 2019), maximum likelihood phylogenetic analysis based on single-copy proteins give high resolution accurate taxonomical classification of bacteria. For instance, phylogenetic classification and grouping of Pseudomonas genus of bacteria has recently been done using 100 single-copy genes (Hesse et al., 2018).

Description of the genus *Pseudomonas* was introduced in 1894 (Peix et al., 2018). Recently (the search was done on 14.12.2020), the number of the validly published and named species is 239 in 'The List of Prokaryotic names with Standing in Nomenclature' (Parte et al., 2020). This genus belongs to the *Gammaproteobacteria* class within the phylum of Proteobacteria in the Bacteria domain. It has the highest number of described species (Lalucat et al., 2020). Various environments like plants, animals, water, and soil can be home to the species of the genus *Pseudomonas*, and the species can use a huge number of organic compounds as an energy source (Hesse et al., 2018). Species in this genus are very crucial for environmental microbiology because of their role in bioremediation, biological control, and other environmental processes (Hesse et al., 2018). Moreover, it is known that some species of the genus have resistance to the broad range of antimicrobial compounds which have been used for agricultural and medical purposes (Breidenstein et al., 2011). It was shown based on the 100 single-copy genes that *Pseudomonas* genus has 13 groups, and the biggest one is *Pseudomonas fluorescens* group (Hesse et al., 2018). There are 10 subgroups inside the *Pseudomonas fluorescens* group and the average genome length of the group is 6.36 Mb (Hesse et al., 2018). Comparing with the average bacterial genome length, it is quite high. *Pseudomonas protegens* which is also a subgroup of *Pseudomonas fluorescens* group was initially described as species (Ramette et al., 2011). However, with the help of sequencing new strains and making new phylogenetic analyzes, *Pseudomonas protegens* was also introduced as a new subgroup which has two species as *Pseudomonas protegens* and *Pseudomonas saponiphila* (Hesse et al., 2018; Lang et al., 2010; Vinatzer et al., 2016). More recent taxonomic work named as GTDB-Tk shows that this subgroup has more species which are including 5 undescribed species and described ones such as *Pseudomonas protegens*, *Pseudomonas saponiphila* and *Pseudomonas piscis* (Chaumeil et al., 2019). *Pseudomonas piscis* was recently described as a new species, so there was a placeholder name for this species in the GTDB-Tk, but actually, both studies support each other (Liu et al., 2020).

The objective was to reveal the taxonomy of *Pseudomonas* sp. BIOMIG1^{BAC} with complete genome. Also, its novelty was presented by identifying shared and unshared protein families of the phylogenetic neighbors of the strain BIOMIG1^{BAC}. A comprehensive phylogenetic analysis was made with all possible phylogenetic neighbors. Other studies related to *Pseudomonas* genus, and more specifically *Pseudomonas fluorescens* group were examined. In this study, not only type strains but also undescribed strains were used in terms of being more comprehensive. It was hard to detect close undescribed strains, but the *Mash* algorithm made it possible to compare thousands of genomes in a very short amount of time, unlike other OGRI analyses. 16S rRNA, whole-genome, marker genes, shared protein families, and phylogenetic analyses were done for identification of the strain BIOMIG1^{BAC}.

3.2. Materials and Methods

3.2.1. 16S rRNA Comparative Analysis

Representative genomes of the genus *Pseudomonas* were downloaded from NCBI with a Python script to automatize the process. Entrez and BioPython libraries were used for downloading the genomes from the NCBI FTP server (Agarwala et al., 2018; Cock et al., 2009). 224 representative genomes were found on 10.11.2020, and the used taxonomy id was 286. As an addition to the representative genomes, genomes of *Pseudomonas* sp. BIOMIG1^{BAC}, *Pseudomonas* sp. CMR12a and *Pseudomonas* sp. CMR5c were downloaded. The genome of the *Escherichia coli* K-12 MG1655 was used as an outgroup. The extraction of the 16S rRNA from the genomes was made using a Python script. *MUSCLE* (v. 3.8.31) was used to align 16S rRNA genes, and the phylogenetic tree was made with cluster neighbor-joining (Edgar, 2004).

The groups and subgroups of the *Pseudomonas* genus were taken from the literature, and the same color codes were used in the 16S rRNA phylogenetic tree (Gomila et al., 2015; Hesse et al., 2018; Vinatzer et al., 2016). *Figtree* (v. 1.4.4) was used for simplifying the tree by collapsing the clades because the leaf number was too high for clear visualization (Rambaut, 2007). The total number of clades was 20 after the simplification. As an extra, ANI values were calculated with a Python package named pyani (v. 0.2.10), and the used option was the ANIb which is using *BLAST+* for alignments (Pritchard et al., 2016).

3.2.2. Whole Genome Comparative Analysis

Previously downloaded representative genomes were used to find close species to the *Pseudomonas* sp. BIOMIG1^{BAC} using *Mash*. *Mash* analysis has a very similar purpose to the ANI. *Mash* algorithm works faster, so it is preferable for big datasets. *Mash* (v. 2.2) was run with options which were 21 *k*-mer size, 1000 sketch size, and m 2 (Ondov et al., 2016). According to the result of *Mash*, *Pseudomonas protegens*, *Pseudomonas saponiphila*, and *Pseudomonas chlororaphis* were close species to the strain BIOMIG1. Moreover, *Pseudomonas fluorescens* group (taxid: 136843), *Pseudomonas protegens* (taxid: 380021), *unclassified Pseudomonas* (taxid: 196821), *Pseudomonas chlororaphis* (taxid: 587753) and *Pseudomonas saponiphila* (taxid: 556534) were downloaded to make more broad analysis. Unclassified species were downloaded because there could be some strains that are taxonomically closer to the strain BIOMIG1. Also, there was one recently described species, i.e. *Pseudomonas piscis*, which was a previously *unclassified Pseudomonas* (Liu et al., 2020). *P.* sp. BIOMIG1^{BAC} was selected as reference genome, and *Mash* analysis was done with the same previous options. According to shared hashes and p-value, the first 150 close genomes to the reference genome were picked. Threshold values for maximum contig number and minimum L90 was set to 400 and 200 respectively to eliminate low-quality assemblies. Moreover, duplicate genomes of the same strains were eliminated. After the quality filtering and elimination, a total of 141 genomes were remained for taxonomic analysis. *Prokka* (v. 1.12) was used to annotate genomes, and both *gff* and *fasta* files were used as input (Seemann, 2014). The annotated genomes were converted to the *PanACoTA* (v. 1.0) input format to perform pan and core genome analysis (Perrin and Rocha, 2020). Protein families were determined based on 80% identity with *MMseqs* (Hauser et al., 2016). The output of the pipeline created a genome to protein family matrix. In addition, genome to genome matrix was created based on the shared protein families among genomes with a Python script. The genomes which are not inside the *Pseudomonas protegens* and *Pseudomonas chlororaphis* groups were disregarded to eliminate the outliers because the focus was on the *P. protegens* group. Also, distant genomes were creating noise. *Pseudomonas chlororaphis* group was used as an outgroup for control. After the elimination, a square matrix of order 135 was created. *SciPy* and *Seaborn* were used for clustering based on protein families and visualizing the heatmap (Hill, 2015; Waskom et al., 2017). Whole-genome ANI values of the selected genomes were calculated with pyani (v. 0.2.10). When determining the borders of the phylogroups, the protein families distance matrix was used. The function named pdist in the *SciPy* library was used to process the distance matrix. Then, the linkage function was used with the average option to create linkage. Fluster function with distance method was used to calculate cluster groups. 5 different cluster groups were found, and *Seaborn* was used to visualize the matrix with clusters as a heatmap. The cluster groups were compatible as desired with

ANI values of the genomes. In the heatmap output, the color bar borders of the shared protein families were determined as 4000 and 6000 to create a more distinctive output. Also, a phylogenetic tree with options as '-nt 24 –m GTR –st DNA –B 1000' was created from the alignment output of the *PanACoTA* pipeline with *IQ-Tree* (v. 2.1.1) (Lanfear et al., 2020). 141 genomes were used to create the tree without elimination, unlike in the clustering step. A consensus tree with bootstrap values was completed after 130 iterations (Hoang et al., 2018). The tree was unrooted, *Figtree* (v. 1.4.4) was used to visualize and to determine root based on *P. veroonii* as outgroup (Rambaut, 2007). Also, other distant species were deleted for clear visualization. In *P. protegens* subgroup, there are three main clades, and they named Clade 1, Clade 2, and Clade 3. *P.* sp. BIOMIG1^{BAC} was in the Clade 1.

3.2.3. Sub-Classification Based on Marker Genes of *Pseudomonas fluorescens* Group

Pseudomonas fluorescens group of bacteria has 8 subgroups. These are *P. corrugate, P. koreensis, P. jessenii, P. mandelii, P. gessardii, P. fluorescens, P. protegens,* and *P. chlororaphis*. Garrido-Sanz *et al.* (2017) determined 8 indicator genes for *Pseudomonas fluorescens* group of bacteria i.e. DGPf_1, DGPf_2, DGPf_3, DGPf_4, DGPf_5, DGPf_6, DGPf_7, and DGPf_8 (Garrido-Sanz et al., 2017). Every subgroup has different composition of these genes. In addition, pyoluteorin synthesis genes (pltABCDEFG and pltRM) are also used as additional markers of the *Pseudomonas protegens* subgroup. The genes were downloaded from the NCBI database. *P.* sp. BIOMIG1^{BAC} and CMR12a were used to build the Blast database, and blastn (v. 2.2.31+) was used to search the marker genes with '-outfmt 6' option (Camacho et al., 2009). The output was filtered with thresholds equal to 80.0 percent for identity and 0.1 for e-value.

3.2.4. Pan and Core Genome Analysis

The taxonomy outputs from the previous section provided the clusters of the species in *Pseudomonas protegens* subgroup. Assignment of protein families to the previously determined species clusters was made with a Python script, which was used the pan-genome results (Perrin and Rocha, 2020). A Python library named *Pandas* were used for analyzing the data (McKinney, 2010). The calculation of pan and core genomes of the species in *P. protegens* subgroup was made with a Python script. Matplotlib library was used with Python to create Venn diagrams of the pan and core-genome (Hunter, 2007). Shared protein families were used for the core-genome, and a total number of protein families were used for the pan-genome diagrams.

3.3. Result and Discussion

Protein-based whole-genome comparisons provide a lot of information about the genome, and they are effective in resolving the differentiation between close species. Species with a lot of protein similarity with each other can be taken as close species. The whole-genome protein content was compared in order to determine the difference of *Pseudomonas* sp. BIOMIG1^{BAC} from other Pseudomonas species and to define its specific functional differences.

The number of genomes used in the 16S rRNA analysis was 189 (see Appendix A for the used genomes). *Pseudomonas* sp. BIOMIG1^{BAC}, *Pseudomonas* sp. *CMR12a*, and *Pseudomonas* sp. *CMR5c* were clustered together and the closest group to them is *Pseudomonas putida* group (Figure 3.2). *Pseudomonas* sp. BIOMIG1^{BAC} is in the *P. putida* group according to the 16S rRNA tree. Nevertheless, 16S rRNA analysis is not reliable most of the time because there are false positive and false negative examples, and the resolution of the 16S rRNA is not good (Vinatzer et al., 2016). Further analyses show that 16S rRNA is misleading for *Pseudomonas* sp. BIOMIG1^{BAC}. The average genome number in the clades of the tree is 8.8. Maximum and minimum ANI values of 16S rRNA genes in the *P. putida* are 0.991 and 0.985, respectively. *Pseudomonas* sp. BIOMIG1^{BAC} and *Pseudomonas* sp. *CMR12a* have a 100% identical 16S rRNA gene.

Existing phylogenomic studies for the *P. protegens* subgroup were assessed, and a new comprehensive study that includes all classified and unclassified genomes which belong to *P. protegens* subgroup to date (10.11.20) is created. One of the existing studies (Vinatzer et al., 2016) was only used 6 genomes from the *P. protegens* subgroup in the scope of *P. fluorescens* complex. Taxonomic errors in the NCBI affected the study that 2 of them are misclassified, 2 of them are unclassified, and 2 of them are truly classified. Another study (Hesse et al., 2018) was only used type strains of species to analyze *Pseudomonas* genus. Because of the limited scope, unclassified genomes were disregarded. According to this study, there are 2 species as *P. saponiphila* and *P. protegens*, in the *P. protegens* subgroup. Another taxonomic study was GTDB-Tk, the scope was comprehensive, but there was no information about subgroups and groups of the species (Chaumeil et al., 2019). GTDB-Tk has very reliable outputs, and it is combined with literature to create better results. The closest strains, according to the Mash analysis, were searched in GTDB-Tk to find their taxonomic classification in GTDB-Tk.

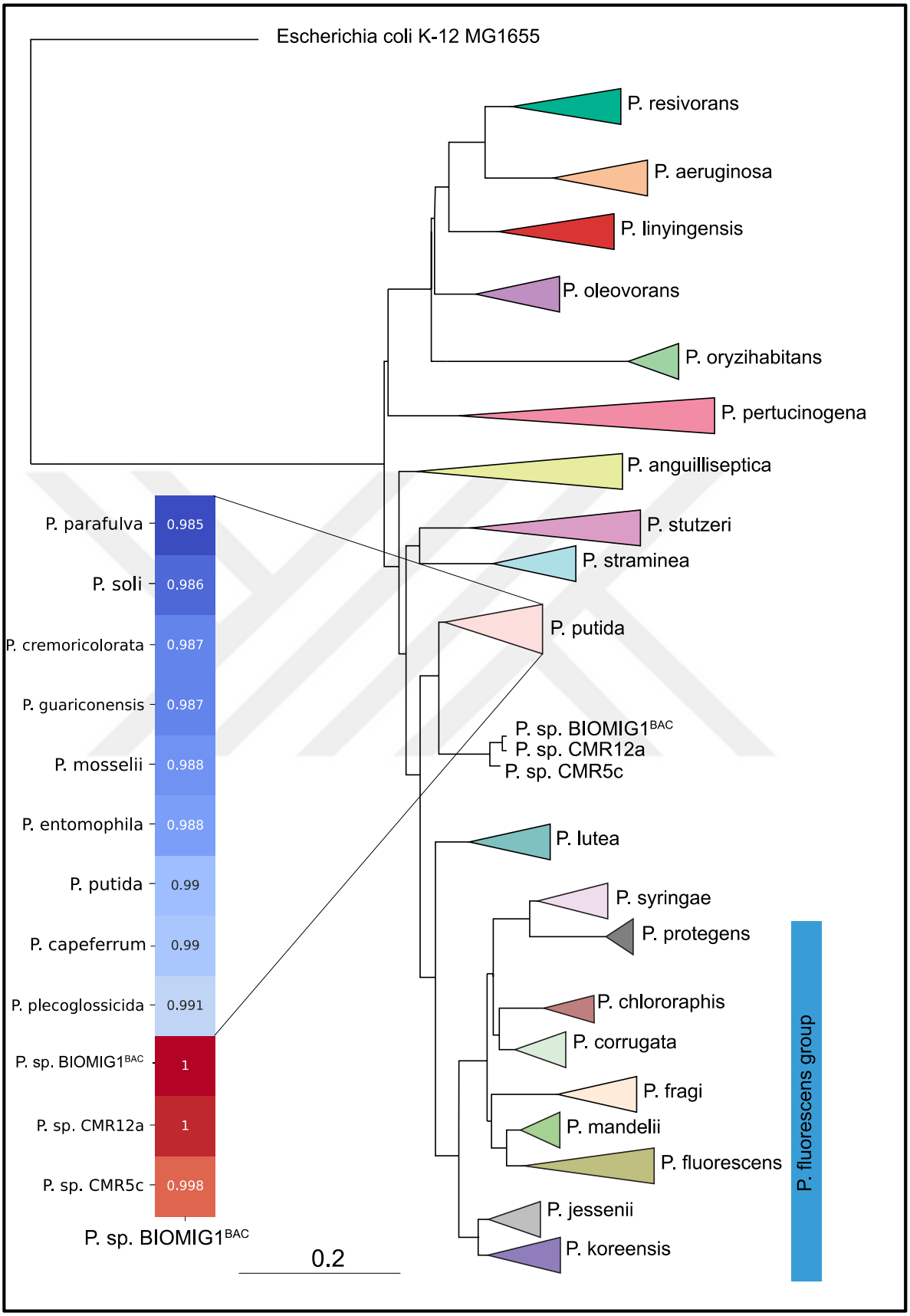


Figure 3.2. 16SrRNA tree of the representative genomes in *Pseudomonas* genus shows that *P.* sp. BIOMIG1^{BAC}, *P.* sp. CMR12a and *P.* sp. CMR5c are grouping with *P. putida* group members. In ANI values, it can be seen that 16S rRNA genes of *P.* sp. BIOMIG1^{BAC} and *P.* sp. CMR12a are identical. *Escherichia coli* K-12 MG1655 is the outgroup.

Possible genomes which can be a member of *P. protegens* subgroup were downloaded from NCBI. After quality filtering and clustering, whole-genome alignment based phylogenetic tree and protein family based clustering were created. According to the tree and protein based clustering, *P. protegens* subgroup has 3 main clades with a total of 93 strains (Figure 3.3). Comparison with the existing taxonomic studies were done to find out the optimum inference. Strains and species were controlled with GTDB-Tk (Chaumeil et al., 2019). GTDB-Tk uses placeholder names for the undescribed species. With the combination of the GTDB-Tk, *P. protegens* subgroup has 8 species, but only 3 of them (*P. protegens, P. saponiphila, P. piscis*) are described. Also, one of the undescribed species is eliminated because of the deficient quality. With the support of the clustering results, the tree, and GTDB-Tk, it can be said that NCBI taxonomy does not have an accurate taxonomic classification for *P. protegens* subgroup. There are strains which are 41.93% unclassified, 41.93% truly classified, and 16.12% misclassified in NCBI taxonomy (see Appendix B for more information about classification result and used strains).

P. sp. BIOMIG1^{BAC} is proposed to be a type strain of *P. alexanderii* sp *nov.* because it is different enough to be described as novel species according to this research results and GTDB-Tk. The proposed name is attributed to Martin Alexander, who discovered the QAC metabolism (Alexander, 1981). Also, the place where the species was discovered was once ruled by Alexander the Great. *P. protegens* subgroup has 3 main clades. *P. Piscis* and *P. alexanderii sp nov.* named as 's__Pseudomonas_E sp001269545' and 's__Pseudomonas_E sp001705835' respectively in the GTDB-Tk are in the Clade 1. Recently, *P. Piscis* was described. ANI values of the strains of *P. piscis* are not higher than 91%, so *P. alexanderii sp nov.* is very different, although they are clustered in the same clade. *P. protegens* is in the Clade 2 and *P. saponiphila* is in the Clade 3 (Figure 3.3). MinHash distances can be seen (Table 3.1).

Table 3.1. MinHash distances of close strains to the *Pseudomonas* sp. BIOMIG1^{BAC}. If it is higher than 0.05, they are probably different species.

***Pseudomonas* Strains**	**Genome size (bp)**	***MinHash* distance**
P. sp. CMR12a (*alexanderii sp nov.*)	6,896,611	0.012
P. sp. CMR5c (*piscis*)	6,796,817	0.067
P. sp. CMAA1215 (*piscis*)	6,658,235	0.068
P. protegens Pf-5	7,074,893	0.109
P. saponiphila DSM 9751	7,375,852	0.112
P. chlororaphis O6	6,980,251	0.129

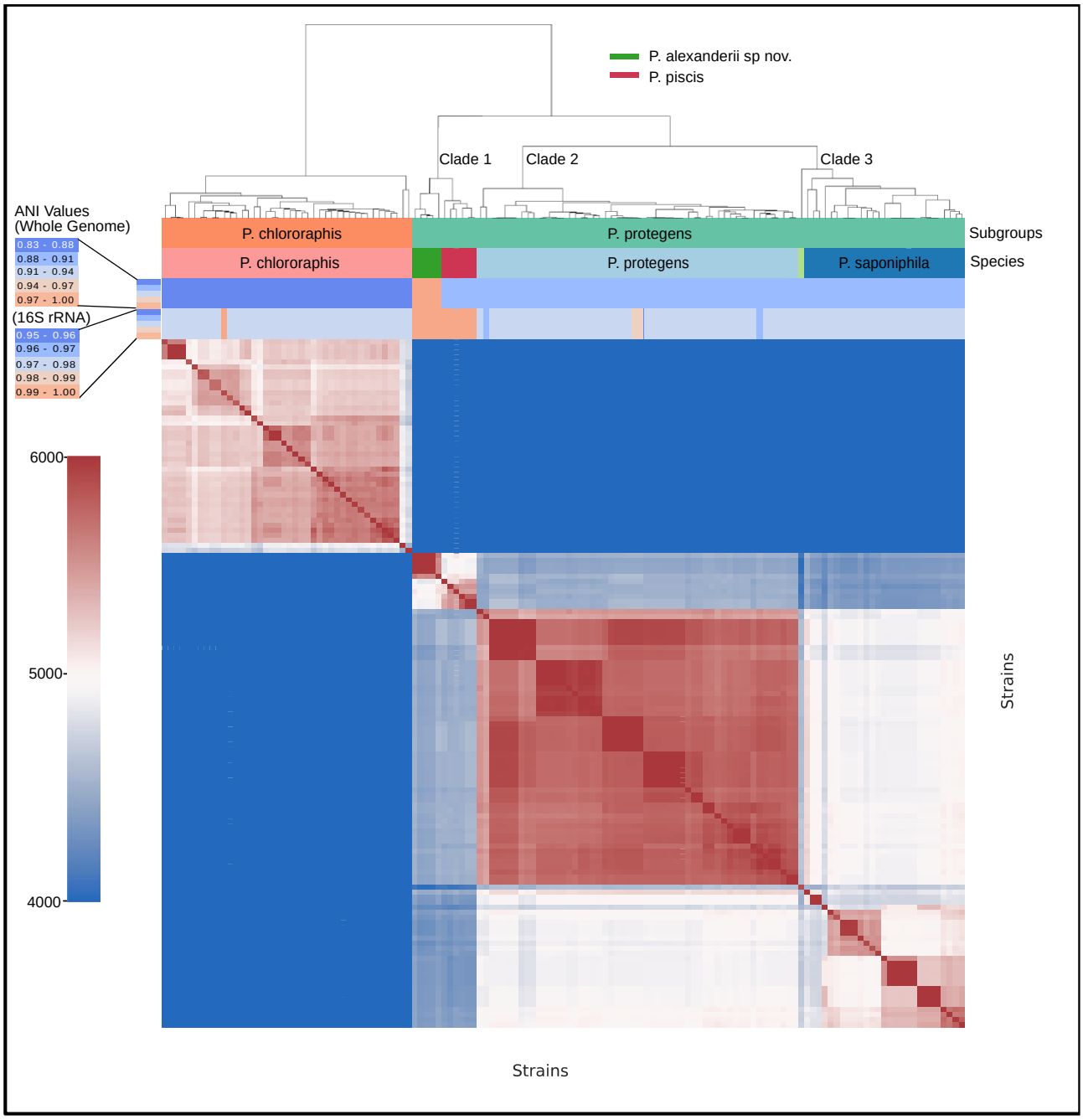


Figure 3.3. Heatmap and dendrogram based on the whole-genome shared protein family statistics between *P. chlororaphis* and *P. protegens* subgroups. As an outgroup, *P. chlororaphis* subgroup was used. Clade 1, Clade 2, and Clade 3 are the three primary phylogroups of *P. protegens* subgroup. Subgroup names, species names, ANI values of the whole genome and 16S rRNA respect to *Pseudomonas alexanderii sp nov.* BIOMIG1^{BAC} can be found on top rows. When looking at the 16S rRNA ANI values, BIOMIG1^{BAC} is 99% similar to the strains in the Clade 1. When looking at the whole-genome ANI values, it is around 88-91%, similar to *P. piscis*, Clade 2, and Clade 3 strains. The maximum and minimum values in the heatmap are set to 6000 and 4000, respectively. As a result, the boundary color values even reflect the values that are higher or lower than the thresholds. The dark blue color tones represent subgroup differentiation, while the white color tones represent species differentiation. The same species can be identified by clusters of red color tones.

Clade 3 is likely to have more reported species in the future due to the large distances around strains. While whole-genome and 16S rRNA ANI values were often overlapping, some strains were inconsistent. Among *P. alexanderii sp nov.* strains, 16S rRNA ANI values are 0.99 and higher, and whole-genome ANI values vary between 0.98 and 1.00.

Since distant species and strains were eliminated, the number of genomes used in protein-based clustering was reduced to 135. These genomes, however, were not removed from the tree because they had no effect on the output. The tree output matches the dendrogram, which was created based on shared protein families, as expected (Figure 3.4).

P. alexanderii sp nov. BIOMIG1^{BAC}(T) and *P. alexanderii sp nov.* CMR12a satisfy the requirements of P. *protegens* subgroup according to the blastn results for the *P. protegens* subgroup marker genes (Table 3.2). DGPf_2, DGPf_4 and DGPf_6 are present in both strains and DGPf_1, DGPf_3, DGPf_5, DGPf_7, and DGPf_8 are not present in both. E values are equal to zero and percent identities are higher than 80.0%.

Table 3.2. The combination of marker genes for *P. protegens* subgroup indication. *P. alexanderii sp nov.* BIOMIG1^{BAC}(T) and *P. alexanderii sp nov.* CMR12a are in the *P. protegens* subgroup. DGPf_0 is not significative.

Marker Gene Name	**P. protegens subgroup**	**P. alexanderii *sp nov. BIOMIG1^{BAC}***			**P. alexanderii *sp nov. CMR12a***		
		Existence	**Percent identity**	**E value**	**Existence**	**Percent identity**	**E value**
DGPf_0	±	NA			NA		
DGPf_1	-	-			-		
DGPf_2	+	+	84.00	0.0	+	84.32	0.0
DGPf_3	-	-			-		
DGPf_4	+	+	82.06	0.0	+	81.94	0.0
DGPf_5	-	-			-		
DGPf_6	+	+	87.67	0.0	+	87.33	0.0
DGPf_7	-	-			-		
DGPf_8	-	-			-		

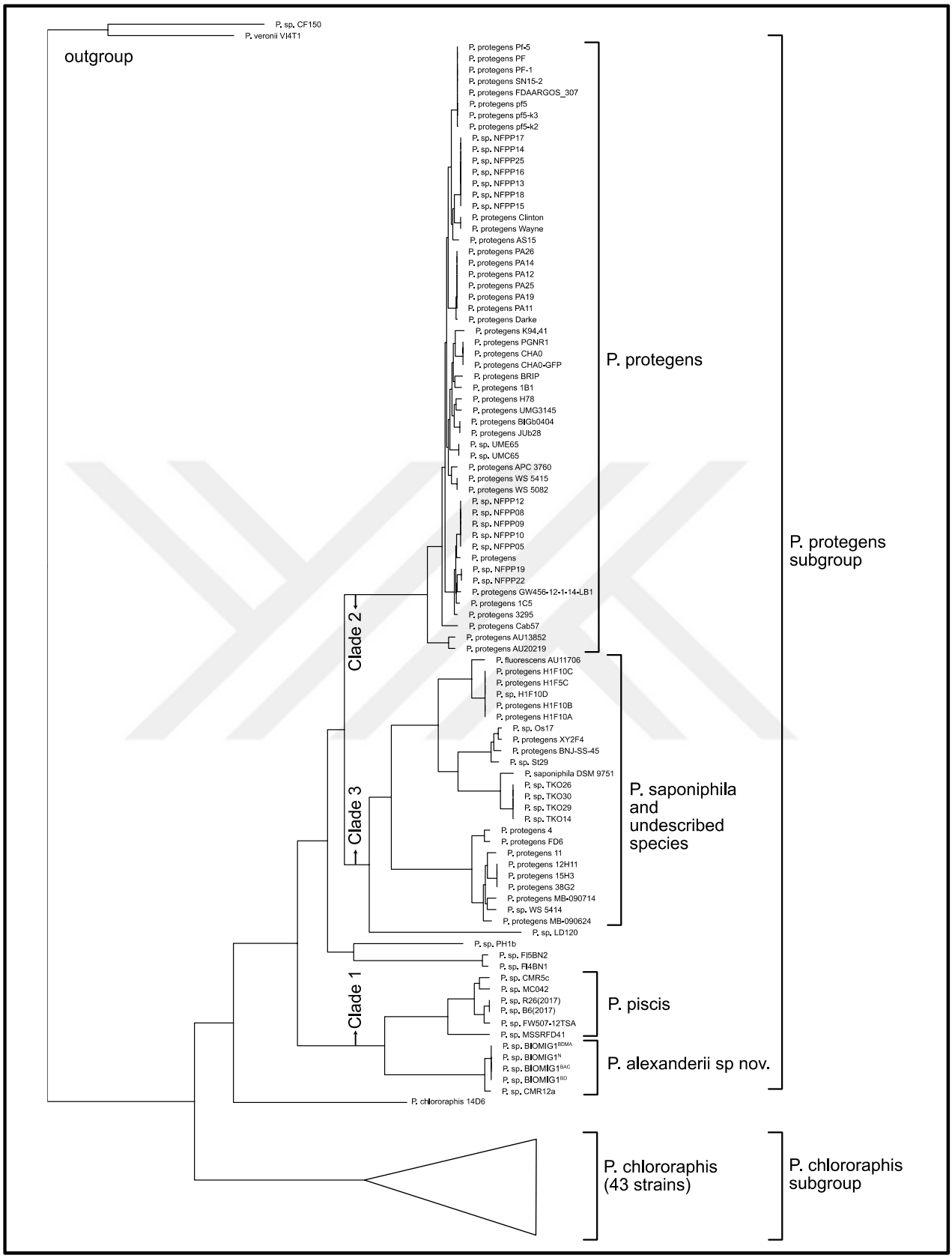


Figure 3.4. Phylogenetic tree of the strains in *P. protegens* and *P. chlororaphis* subgroups with species. General time reversible (GTR) model was used. The bootstrap consensus tree was built after 130 iterations with IQ-Tree (v. 2.1.1), the values and distant species were removed for clear visualization. *P.* sp. CF150 and *P. veronii* VI4T1 are outgroups.

Table 3.3 shows the outcomes of the core and pan-genome analyses. Clade 1's shared protein families number for the core genome is 4458, while the pan genome's total protein families number is 10210. Clade 1 contains 11 strains, Clade 2 contains 54 strains, and Clade 3 contains 27 strains. Clade 2 is the one with the most strains. Clade 3 has the largest pan-genome, and the lowest core-genome, contrary to its size since Clade 3's diversity is high. The core genome size of *P. chlororaphis* subgroup is higher than *P. protegens*.

Table 3.3. Core and pan genome size of two subgroups and clades of *P. protegens* subgroup.

Subgroup	**Core**	**Pan**
P. protegens	3066	23299
P. chlororaphis	3635	15806
Clades of *P. Protegens* subgroup	**Core**	**Pan**
Clade 1 (*P. alexanderii* sp nov. and *P. piscis*)	4458	10210
Clade 2 (*P. protegens*)	4780	11421
Clade 3 (*P. Saponiphila* and others)	4034	14151

Venn diagrams were created for core and pan-genome to show differences and intersections of the three clades (Figure 3.5). The core genome size of *P. protegens* subgroup in Table 3.3 is 3066; however, it is 3182 in Figure 3.5 because of the one unclustered strains in *P. protegens* subgroup.

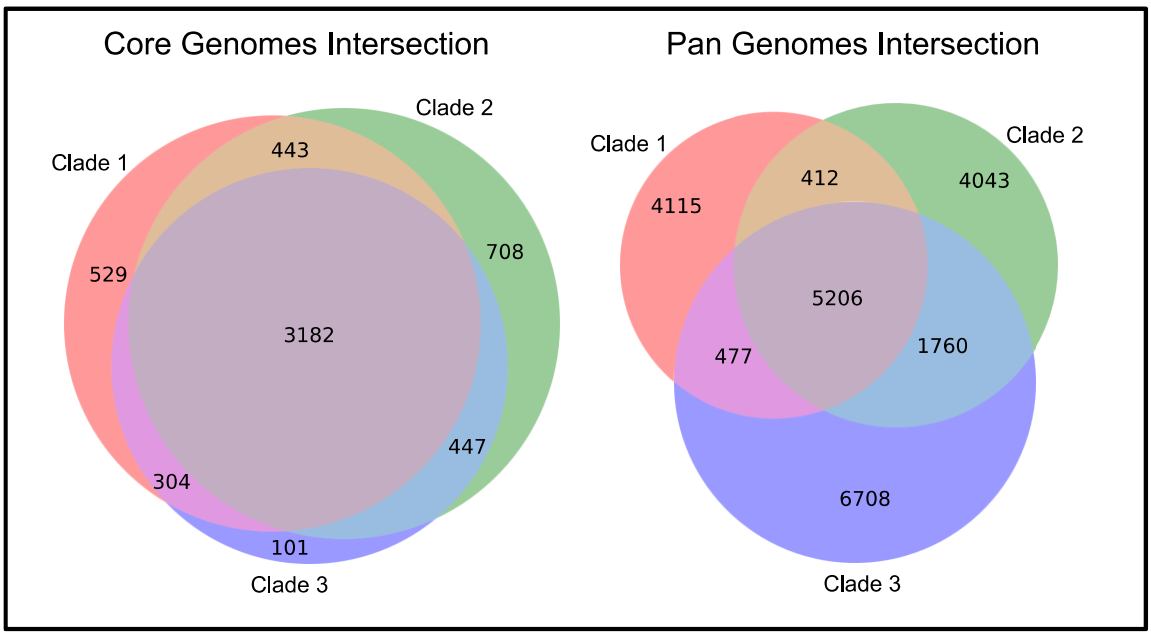


Figure 3.5. Core and pan-genome Venn diagrams of Clade 1, 2 and 3.

Clade 3's pan-genome difference is higher, and core genome difference is lower than others because of its diversity (Figure 3.5). It is another indicator for the existence of undescribed novel species in Clade 3. Clade 2's core genome difference is 708, and Clade 1's is 529. It is expected because Clade 1 has two species and Clade 2 has one species in it. The strain specific families of strains that are in the Clade 1 can be seen more detailed (Figure 3.6). The number of strain specific protein families is 938, and the number of strain specific genes is 1097 for BIOMIG1^{BAC}.

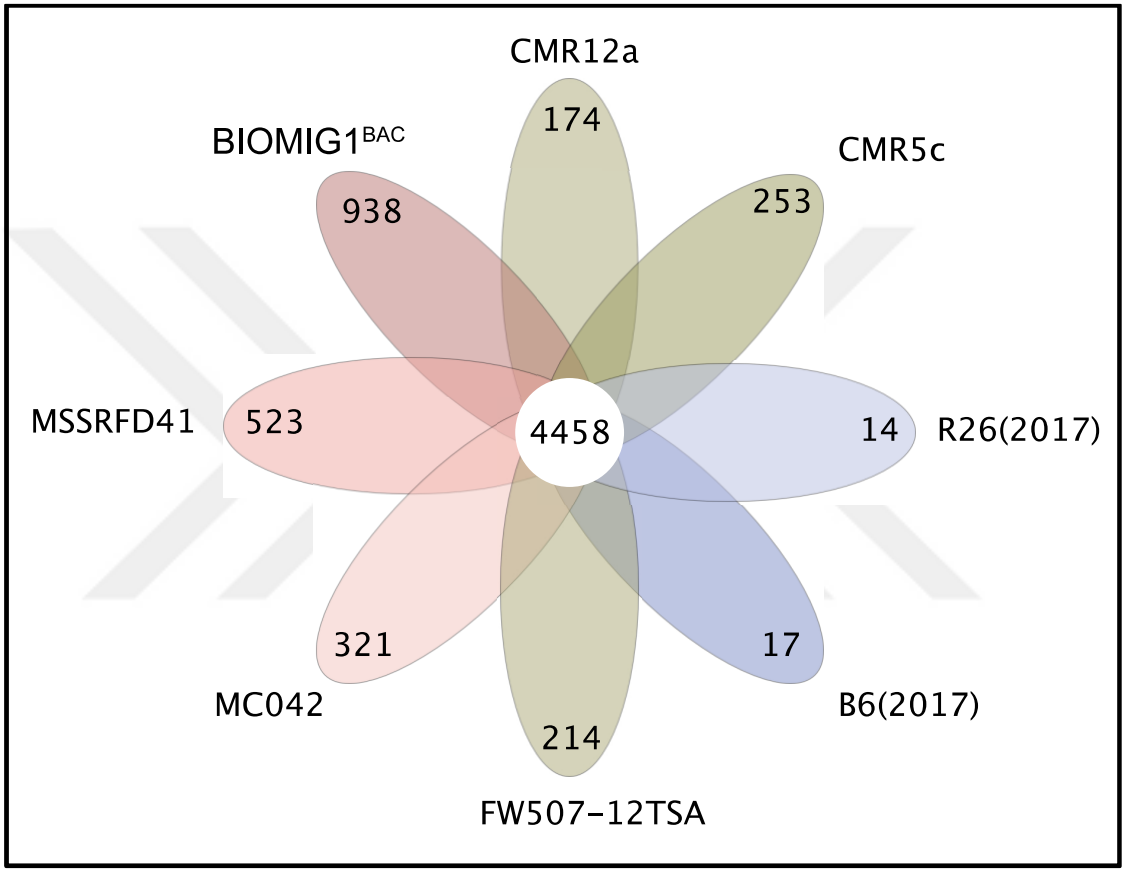


Figure 3.6. Strain specific protein families in Clade 1. BIOMIG1^{BAC} has the highest number as 938. Strains named MSSRFD41, MC042(T), FW507-12TSA, B6(2017), R26(2017), and CMR5c are strains of the *P. piscis*, and strains named as BIOMIG1^{BAC}(T), and CMR12a are strains of the *P. alexanderii sp nov.*.

One of the distinctive features of *P. sp* BIOMIG1^{BAC} is that it does not have a pyoluteorin synthesis gene cluster (pltABCDEFG and pltRM). Pyoluteorin synthesis is a characteristic difference for *P. protegens* species (Ramette et al., 2011). This is also another indicator for the novelty of *P. sp* BIOMIG1^{BAC}.

The total number of *P. protegens* subgroup strains that are used in this study is 91. Their average genome length is 6,962,584 bp. *P. sp* BIOMIG1^{BAC}, whose genome is 7,675,262 bp, has the largest

genome in the subgroup. *P. protegens* FD6 has a minimum genome length as 6,667,995 bp. The average GC content is 0.63. Maximum and minimum values for the GC content are 0.63 and 0.61, respectively.

Overall, the proposed minimal standards for the bacterial taxonomy were considered as a basis for the identification of the species (Chun et al., 2018). Furthermore, not only the type strains but also all strains that were likely to be a phylogenetic neighbor to the strain BIOMIG^{BAC} in the databases were compared in order to make a more detailed analysis. Firstly, 16S rRNA analysis was used as an initial step. There were type strains that have higher 16S rRNA similarity than 99%. Therefore, the next step was needed for identification because higher similarity values than the threshold (98.7%) need whole genome similarity comparisons. OGRI values were calculated as a next step using *Mash* and ANI. The closest type strain according to the ANI was *P. piscis* MC042 (91%), so it was quite far away from the threshold (95-96%). Therefore, the strain BIOMIG1^{BAC} identified as a novel species which was tentatively named as *Pseudomonas alexanderii* sp nov. Strain BIOMIG1^{BAC} was classified under *Pseudomonas protegens* subgroup. According to this research, *P. protegens* subgroup has 91 strains, and strain BIOMIG1^{BAC} has the highest genome length in the subgroup. *P. protegens* subgroup had 3 main clades according to the phylogenetic analysis. The errors and misclassifications in the NCBI taxonomy were presented for *Pseudomonas protegens* subgroup (see Appendix B). A more recent taxonomic study named GTDB-Tk had better results than NCBI taxonomy, and it supported the results of this study. One of the missing things in the GTDB-Tk was the group and subgroup information of the strains. In this study, the group and subgroup definitions was combined with comprehensive phylogenetic analysis to obtain detailed results.

4. STRAIN SPECIFIC METABOLIC FUNCTIONS OF *PSEUDOMONAS* SP. BIOMIG1^{BAC}

4.1. Introduction

The genome sequencing of an organism demonstrates the code for its present functional capacities as well as its evolutionary history (Madsen, 2015). There is a correlation between genome size with the number of genes, and giant genomes have more open reading frames (ORF). Roughly, one ORF has 300 amino acids, and the number of genes is equal to 10^{-3} x genome size. Gene annotation algorithms divide genomes according to the places of start and end codons based on the assumption of average ORF size. The divisions can be erroneous because they are theoretical. Furthermore, bacterial genome size is very diverse due to the ecological variety including diversity of nutrient sources and environmental conditions. After the ORF detection, alignment of the theoretical ORFs with known genes gives potential function of those ORFs. However, misannotations are common, and errors can spread quickly due to the automatic annotation processes (Madsen, 2015). Therefore, manual curation and genetic experiments are essential for the verification of the assigned function to an ORF.

The main categories of the ORFs based on their functions are cellular processes, metabolism, and information storage and processing. Also, more subcategories can be used to group genes with their possible functions. Moreover, there are still too much unknown or poorly categorized genes that approximately make up one-third of bacterial genomes. Although estimations of the gene functions are present, there is still a long way for comprehensively predictive inferences (Madsen, 2015). It can be inferred that increasing genome size increases the lack of knowledge about the genome's metabolism due to complex environmental pressures and metabolic cycles.

Bacterial genomes also contain foreign DNA that gives extra function to the bacteria. Those foreign DNA may be associated with bacterial viruses (phages) or plasmids (Frost et al., 2005; Sørensen et al., 2005; Wang et al., 2010). Some genes in the bacterial chromosomes or foreign DNA are linked to mobile genetic elements such as transposons, and integrons (Kichenaradja et al., 2010). Those genes favor bacterial survival and give advantage to bacteria to better compete with other bacteria in the population when conditions are not favorable. Among those genes, the ones conferring antibiotic resistance and biodegradation are particularly important since bacteria having those genes

can easily proliferate in the environment and facilitate the dissemination of those genes in the population through HGT (Hall et al., 2017).

The genus of *Pseudomonas* which are very dominant in soil and aquatic environments has a very diverse metabolism (Silby et al., 2011). As a result, they are very important group of bacteria for human health, plant protection and waste treatment. For instance, species of *P. fluorescens* synthesize novel secondary metabolites that promote plant growth. Some *P. putida* are main degraders of toxic chemicals such as pesticides and hydrocarbons in the environment. Moreover, *P. aeruginosa* strains are very known human pathogens with antibiotic resistance. Many novel functions of Pseudomonas strains are linked to specific genes that are associated with plasmids and mobile genetic elements in their genome (Silby et al., 2011).

Pseudomonas sp. BIOMIG1^{BAC} is (2-4 μm) a rod-shaped, aerobic bacterium that can mineralize BACs up to 1024 mg/L (Ertekin, 2017). It is a catalase and oxidase positive bacteria which uses a broad range of carbon sources. It has a distinct carbon utilization profile from *Pseudomonas putida*. 55% of *Pseudomonas* sp. BIOMIG1^{BAC} genes were not assigned to any functional category (Ertekin, 2017). Another study showed the strain BIOMIG1^{BAC} is very resistant to clinically used antibiotics (Gul, 2016). Moreover, various transporters in its genome can provide resistance to different substances. Also, there were HGT and phage-related genes (transposases, conjugative transfer proteins, and integrases) in *P.* sp. BIOMIG1^{BAC} (Ertekin, 2017). The objective of this part of the research was an analysis of the genome of the strain BIOMIG1^{BAC} to identify strain specific genes and functions and their role in BAC biodegradation.

4.2. Materials and Methods

4.2.1. Proteome Comparison

In the previous section, *Pseudomonas* sp. BIOMIG1^{BAC} was taxonomically classified. Then, the proteome of the strain BIOMIG1^{BAC} was compared simultaneously with the proteomes of *Pseudomonas* sp. CMR12a, *Pseudomonas piscis* MC042, *Pseudomonas protegens* Pf-5 and *Pseudomonas protegens* CHA0, which are phylogenomically closest strains to the BIOMIG1^{BAC}, using Pathosystems Resource Integration Center (PATRIC) (Davis et al., 2019; Overbeek et al., 2014). The genes which are specific to the *P.* sp. BIOMIG1^{BAC} were extracted. Then, a very basic 1d clustering algorithm was used to determine strain specific genomic regions (SSGR). For SSGR assignment, maximum distance between two consecutive genes (gene-distance) and the minimum number of genes in a region (gene-count) were used as end-points. In total, 27 SSGRs of *P.* sp. BIOMIG1^{BAC} were assigned using gene-distance as 5000 bp and gene-count equal to 5.

4.2.2. Gene Functions

The possible functions of the genes in the genome were predicted and assigned to a pathway using Kyoto Encyclopedia of Genes and Genomes (KEGG) and PATRIC (Davis et al., 2019; Kanehisa, 2000). Each gene received an Enzyme Commission (EC) and Gene Ontology (GO) number. KEGG and PATRIC were also used to detect antibiotic resistance genes in the genome (Davis et al., 2019; Kanehisa and Goto, 2000). The CRISPRFinder web tool was used to find CRISPRs (Couvin et al., 2018). Phage-related regions in the genome was identified and annotated by PHASTER web server (Arndt et al., 2016). Insertion sequences were annotated with ISEScan (v. 1.7.2), and ISFinder web server was used for manual curation to complete missing terminal inverted repeats (Xie and Tang, 2017). Visualization of the output was performed using Circos (v. 0.69-8) software (Krzywinski et al., 2009), an in-house program based on DNA Features Viewer (v. 3.0.3) and ChemSketch (v. 16.0).

4.3. Result and Discussion

Protein comparisons of closely related strains are very powerful in showing versatile results about differences. *Pseudomonas* sp. BIOMIG1^{BAC} was compared with *P.* sp. CMR12a, *P. piscis* MC042, *P. protegens* Pf-5 and *P. protegens* CHA0 (Figure 4.1).

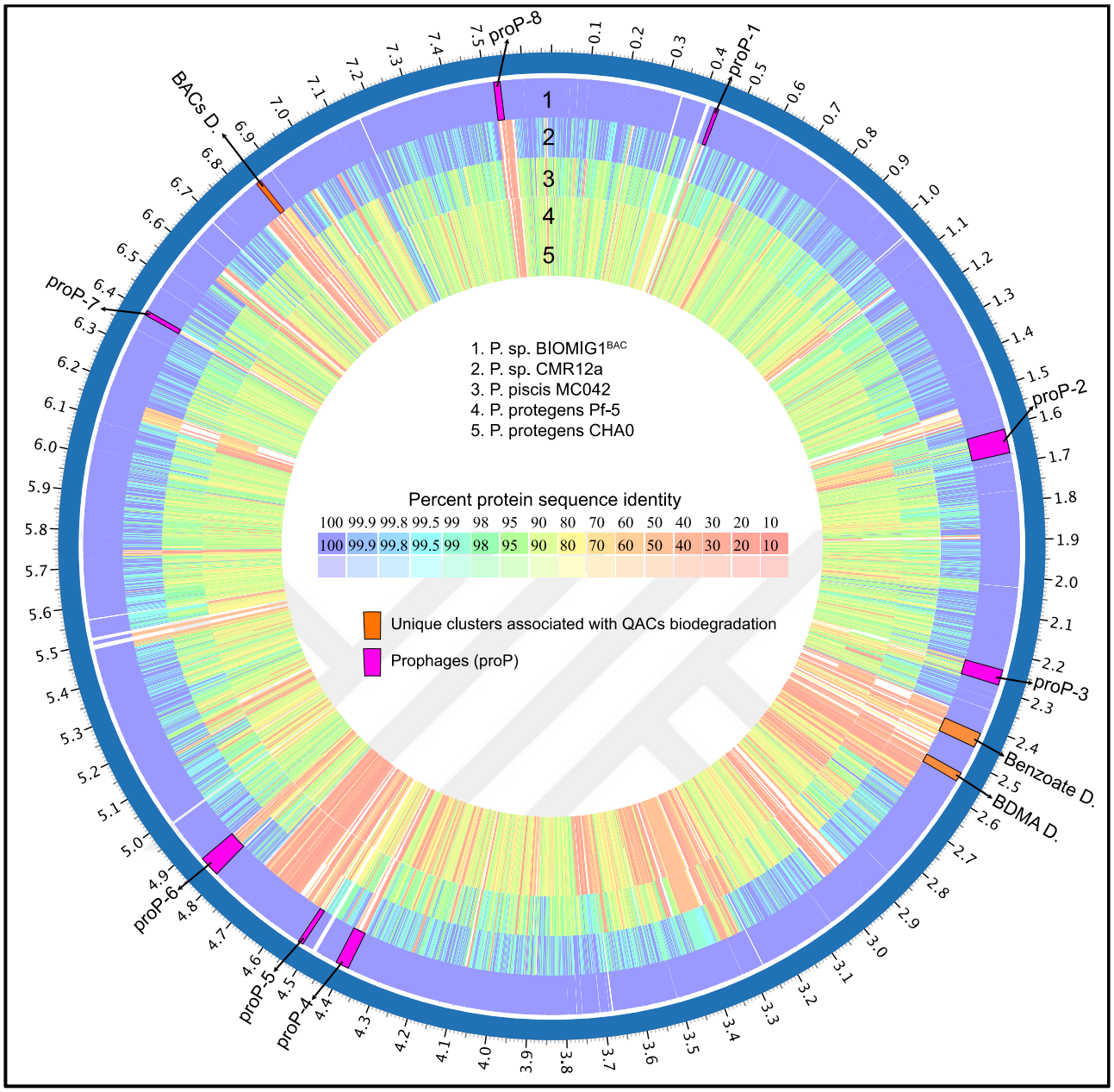


Figure 4.1. Proteome comparison of *Pseudomonas* sp. BIOMIG1^{BAC} with close relatives: *P.* sp. CMR12a, *P. piscis* MC042, *P. protegens* Pf-5, and *P. protegens* CHA0. Orange-colored regions in track 1 represent some of the xenobiotic degradation clusters such as BACs, BDMA, and Benzoate. Possible regions resulting from horizontal gene transfers can be seen.

P. sp. CMR12a and *P.* sp. BIOMIG1^{BAC} are the same species which belong to *P. alexanderii* sp. nov. according to the results of the previous section. *P. piscis* having the MC042 as the type strain is the closest species to the *P. alexanderii* sp. nov. *P. protegens* Pf-5 and *P. protegens* CHA0 are the type strains of *P. protegens*, and they are in the same *P. fluorescens* subgroup with *P.* sp. BIOMIG1^{BAC}. Overall proteome similarity between the strain BIOMIG1 and *P.* sp. CMR12a, *P. piscis* MC042, *P. protegens* Pf-5 and *P. protegens* CHA0 were 95.28%, 89.20%, 84.95%, and 83.63%, respectively. As expected, the difference is increasing from the outer tracks to the inner

tracks. Although *P.* sp. CMR12a and *P.* sp. BIOMIG1^{BAC}are the same species, the regions which are below 80% protein identity can be noticed in Figure 4.1. Those regions are indicators of possible horizontal gene transfer since these two strains share 100% 16S rRNA and 99.12% genome identity. There are 8 prophages in the *P.* sp. BIOMIG1^{BAC} genome. ProP-2 and ProP-3 are partially present in *P*. sp. CMR12a. ProP-2 is also partially found in *P. piscis* MC042. *P. protegens* strains do not have those phages. Therefore, it can be said that proP-2 is a phage that can infect the species in the Clade 1. Moreover, there are various gene clusters that are specific to *P*. sp. BIOMIG1^{BAC}.

In total, there are 27 SSGRs that contain 1103 genes in the genome of *P.* sp. BIOMIG1^{BAC}. Metabolic functions of the genes in the SSGRs are predicted using KEGG and PATRIC (Figure 4.2). Overall, 18% of xenobiotic biodegradation and metabolism-related genes are linked to SSGRs in the whole genome. The ratio is probably higher then calculated, since one-third of the genes in the SSGRs are hypothetical proteins which may have a role in biodegradation. There are also carbohydrate metabolism (15%), lipid metabolism (12%), polyketides with nonribosomal peptides (10%), and amino acid metabolism (10%) related genes in SSGRs. SSGR-2, SSGR-6, SSGR-14, SSGR-15, SSGR-16, SSGR-20 and SSGR-27 have prophages in their gene content. SSGR-7 and SSGR-15 are very long regions (~ 200 kb) that were subjected to many insertions.

The complete degradation of BACs by *P.* sp. BIOMIG1^{BAC} was experimentally proven previously (Ertekin et al., 2017). A novel Rieske oxygenase called *oxyBAC* is able to convert BACs to BDMA. BDMA was debezylated by a dehydrogenease to dimethyl amine and benzoate. Those genes responsible for mineralization of BACs (BAC→ CO_2 + NH_3, Figure 4.3) were all in the SSGRs. *oxyBAC* gene cluster is located in SSGR-24 between 6,839,160 and 6,852,283 bp, and it is flanked by identical insertion sequences (IS) (Figure 4.2 and 4.4.). BDMA degradation related genes are located between 2,545,754 and 2,565,402 bp in SSGR-7, and they are also flanked by ISs. Benzoate degradation related genes are also located in the SSGR-7 between 2,441,437 and 2,472,647 bp (Figure 4.2 and 4.4.). Besides benzoate, SSGR-7 contains genes responsible for biodegradation of other aromatic structures such as 3-fluorobenzoate, benzene, and 4-hydroxy-benzoate. Those genes are responsible for conversion of aromatic acids to succinyl-CoA which is processed to ATP and CO_2 via the citric acid cycle (Ogata et al., 1998). Moreover, there are transcriptional factors for regulating genes, receptors for mediating transportations, chemotaxis related genes, and transporters besides the enzyme coding genes in the SSGR-7.

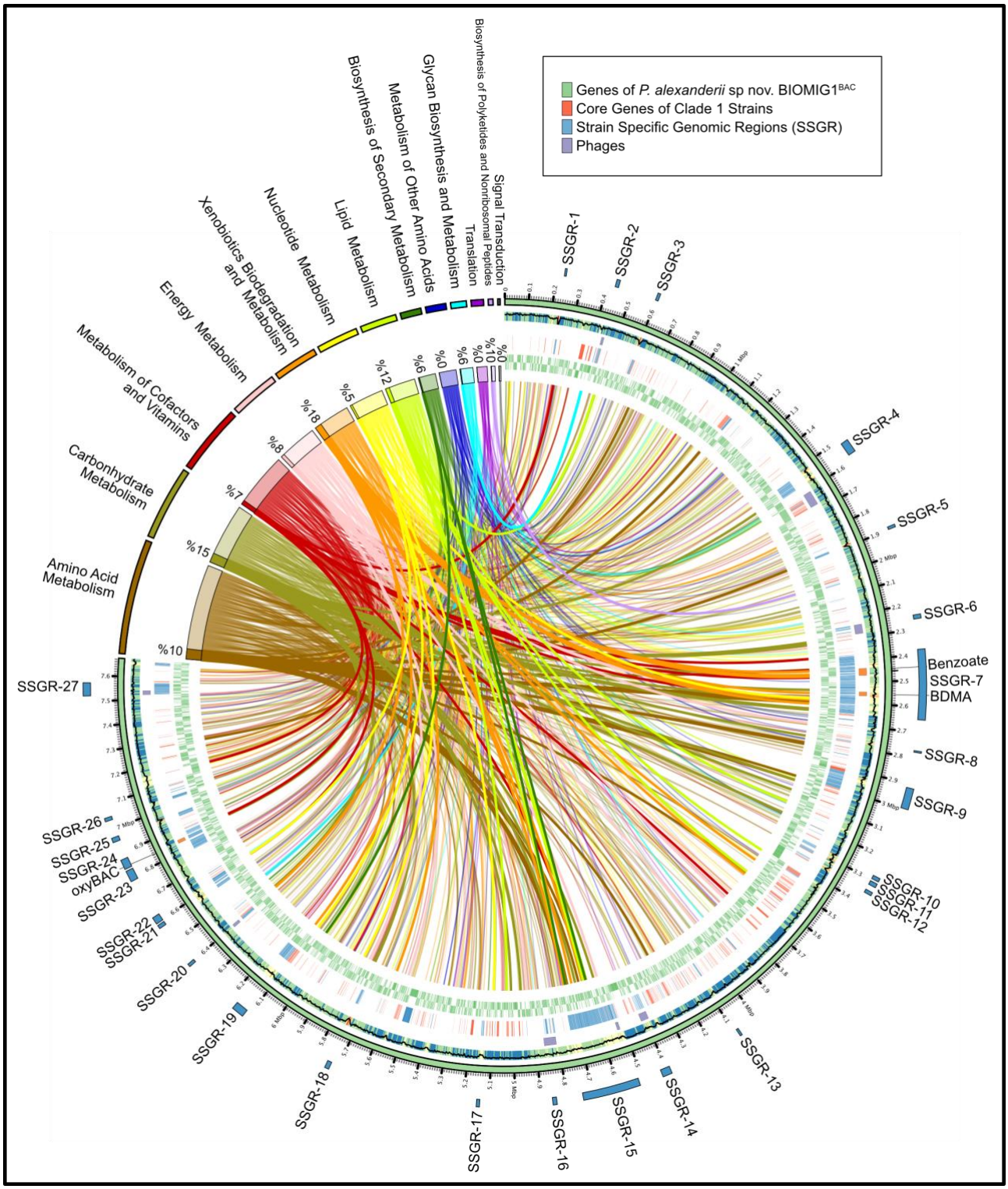


Figure 4.2. SSGRs of Pseudomonas sp. BIOMIG1^{BAC} with the distribution of metabolic pathways. Outside to inside, SSGRs(blue), GC Content (heat map with plot), phages (purple), special genes (orange), core genome of Clade 1 (red), strain specific genes (blue), all genes (genes) and metabolic pathway links. The links of metabolic pathways are bolded if they are in the SSGRs. Percentages in the metabolic classes represent the proportion of SSGRs to the total. The ratio of xenobiotics biodegradation and metabolism is highest at 18%, and there are probably more undiscovered and unassigned genes related with xenobiotics.

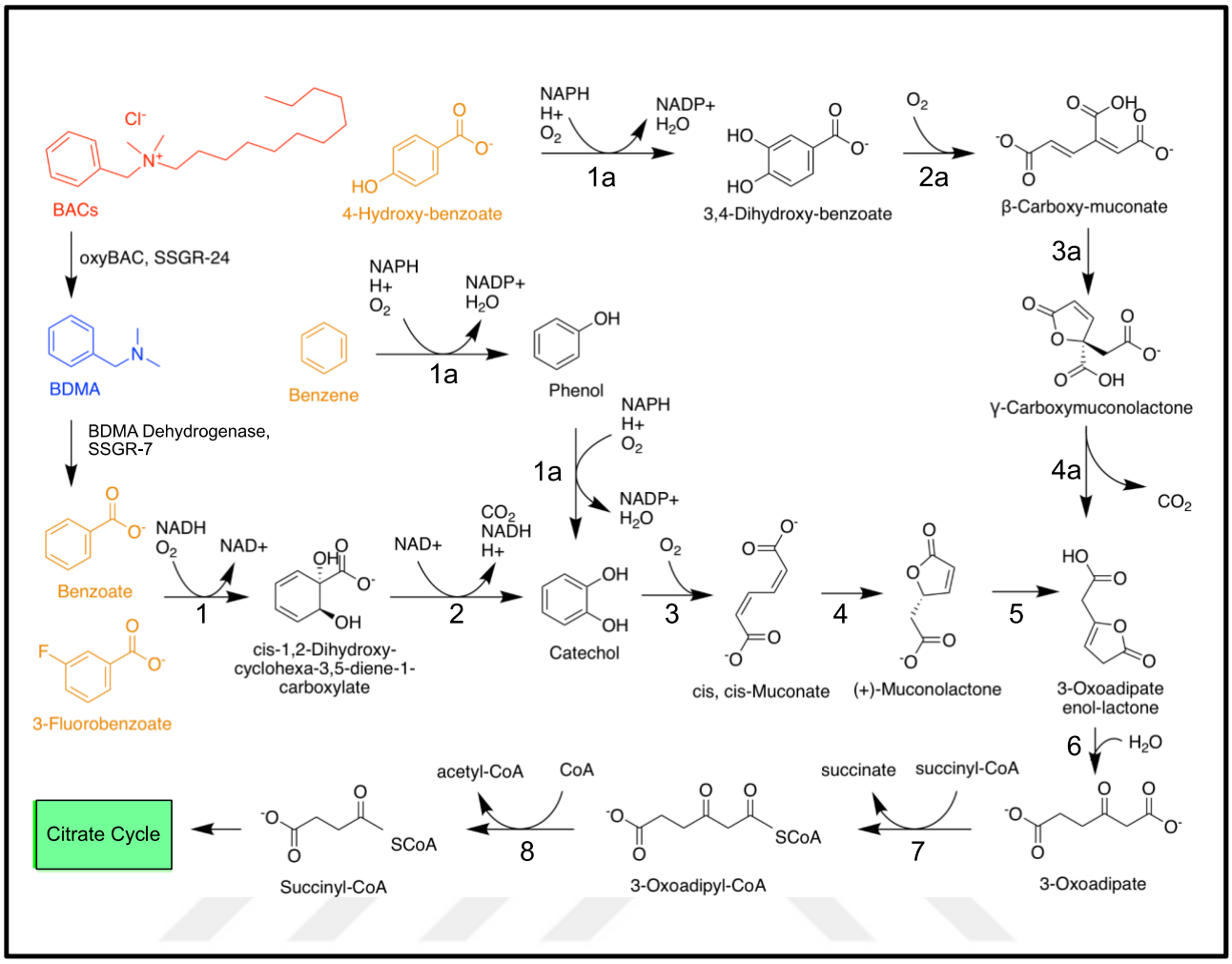


Figure 4.3. BAC and aromatic hydrocarbon biodegradation pathways constructed based on the genes present in SSGR-24 and -7. Genes responsible for each reaction were annotated with gene names or numbers and given in Figure 4.4.

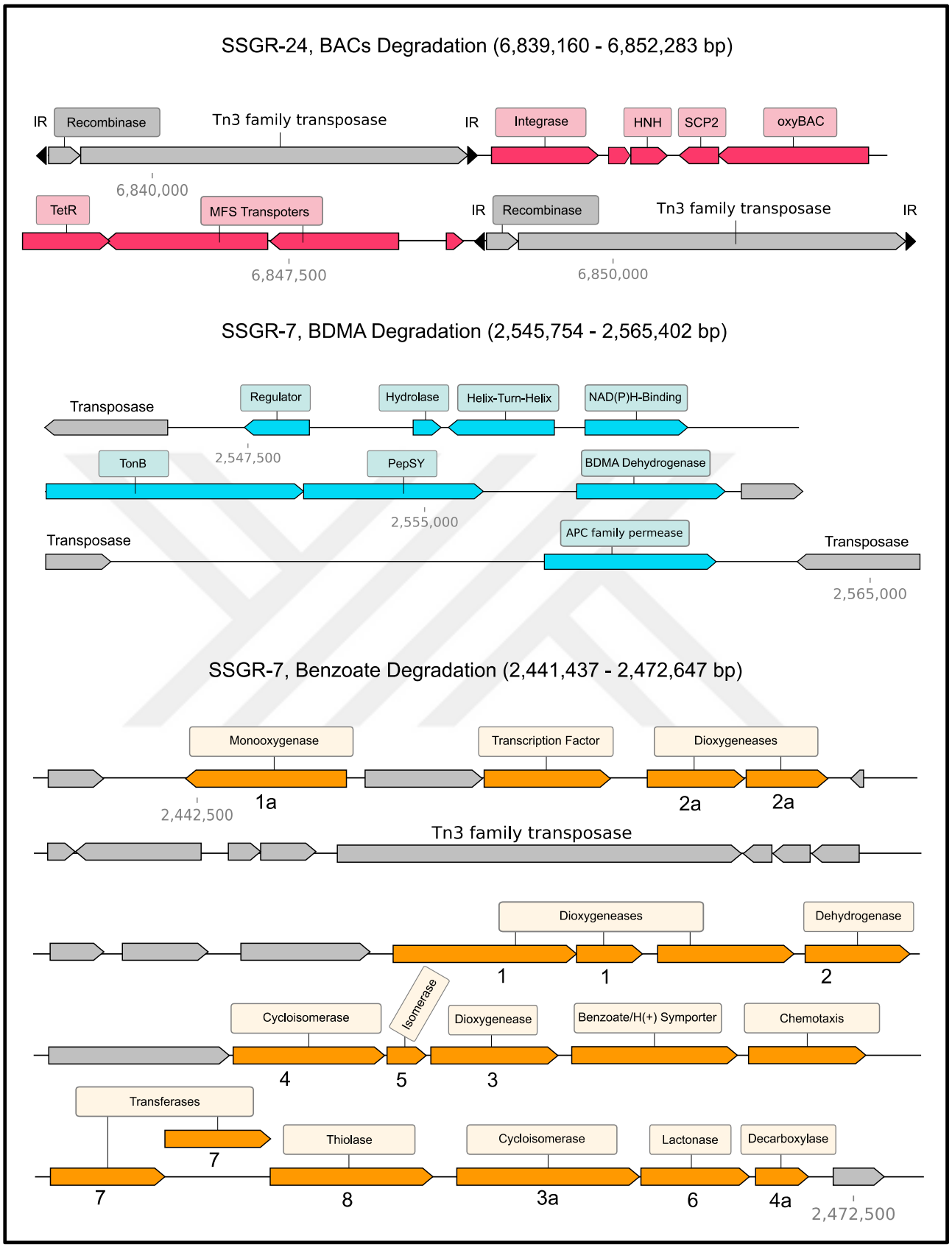


Figure 4.4. Annotation of the enzymes which are responsible for the xenobiotic degradation reactions in Figure 4.3. Transposases are also abundant in these regions. Thus, they can be indicators of horizontal gene transfers.

Interestingly, genes related to biotransformation of BACs have been assembled randomly in the genome and flanked with insertion sequences. This suggest that those genes have been horizontally transferred in to the genome of the strain BIOMIG1. Transposons are agents of horizontal gene transfers and organisms can acquire new metabolic functions, and antibiotic resistance via transpositions (Vandecraen et al., 2017). Insertion sequences (IS) are small transposable elements, and they have the enzyme which is coding transposable. Moreover, different genes can be found beside it. Insertion sequences of *Pseudomonas* sp. BIOMIG1^{BAC} were extracted, and a comparative analysis with close strains was done (Table 4.1). *P.* sp. BIOMIG1^{BAC} has 96 ISs and *P.* sp. CMR12a has 7 ISs, although they are the same species. Moreover, the maximum number of IS in other strains in Clade 1 is 20. Therefore, high IS number of *P.* sp. BIOMIG1^{BAC} is one of the reasons for its high genome size. 11 different IS families, and novel 5 ISs are found in the genome. The predominant family is IS5 (total number is 38) in the *P.* sp. BIOMIG1^{BAC}.

Table 4.1. Comparison of insertion sequence numbers of different strains in Clade 1. There is a correlation between genome size and the number of IS.

GID	Name	Subgroup	Species	# IS	%	Genome Size (bp)	Clade
7909361	P. sp. MSSRFD41	P. protegens	P. Piscis	8	0.15	6924660	CLADE 1
5061511	P. sp. MC042(T)	P. protegens	P. Piscis	8	0.13	6927622	CLADE 1
473781	P. sp. CMR5c	P. protegens	P. Piscis	13	0.17	6755578	CLADE 1
1541031	P. sp. FW507-12TSA	P. protegens	P. Piscis	20	0.25	6895883	CLADE 1
1083071	P. sp. B6(2017)	P. protegens	P. Piscis	12	0.17	7086907	CLADE 1
1082941	P. sp. R26(2017)	P. protegens	P. Piscis	11	0.16	7087578	CLADE 1
2124921	P. sp. CMR12a	P. protegens	P. alexanderii	7	0.10	6896611	CLADE 1
6095431	P. sp. BIOMIG1^{BAC}(T)	P. protegens	P. alexanderii	96	1.82	7675262	CLADE 1

Table 4.2. Clustering and proportioning based on specialty and specificity.

P. sp. BIOMIG1^{BAC}					
Hypothetical Protein		Integrase		CDS	
Specific	Whole	Specific	Whole	Specific	Whole
364 (41%)	880	28 (73%)	38	1097 (15%)	7053
Transcriptional Regulator		Insertion Sequence (IS)		TetR Regulator	
Specific	Whole	Specific	Whole	Specific	Whole
61 (12%)	470	67 (69%)	96	9 (17%)	52

Strain specific genes of *P.* sp. BIOMIG1^{BAC} were filtered according to their features. They were proportioned to the whole genome (Table 4.2). It is expected that the ratio of CDS (15%) and others should be close to each other. However, some of them have higher differences. Most of the insertion sequences (69%) and integrases (73%) are in the strain specific genomic regions of *P.* sp. BIOMIG1^{BAC}. Hypothetical proteins (44%) are also abundant.

Another reason for the increase in genome size is the insertions of various phages into the genome. There are 8 integrated prophages in *P.* sp. BIOMIG1^{BAC}. Although the analysis of the partial genome was revealed only 4 prophage (proP) regions, 8 proPs were found with the complete genome (Table 4.3). The attachment sites of the 5 of them were also extracted (Figure 4.5). The total length of the proPs is 244 kb, and it corresponds to 3.1% compared to the genome size. Although the genes of these 8 phages showed partial similarities with the phage genomes in the databases, no phages with 100% similarity were found in the databases. ProPs in the genome of BIOMIG1^{BAC} are most likely phages specific to this bacterium. The structures of proPs can be broadly divided into two groups. Phages with a genome size of around 10 kb generally consist of only the capsid (head) part (Figure 4.5, proP-1), while phages with large genomes also have a tail (tail) part (Figure 4.5, proP-3). Furthermore, it can be seen that the attachment sites and integrases are consecutive (proP-1, proP-3, and proP-7). The other 2 attachment sites in proP-6 and proP-8 have consecutive hypothetical proteins. Therefore, the hypothetical proteins close to attachment sites can be unknown integrases. The identity similarity of proP-1 and proP-7 is 95%. Differently, proP-7 has inserted transposase which is in an IS. There are also transposases in proP-6 and proP-8 that are also probably belonging to ISs. Transposases and more generally ISs are very interesting because they can cause the genes of the bacterial chromosome to be transported by phages or vice versa. A genomic region can be transported to the phages, or a phage can transport the genomic region to the bacteria with ISs.

Table 4.3. Features of integrated prophages in *P.* sp. BIOMIG1^{BAC}.

# ProP	Start	End	Length (kb)	Annotated Proteins	Hypothetical Proteins
1	435347	445941	10.5	13	4
2	1609497	1673714	64.2	78	10
3	2244050	2284719	40.6	54	18
4	4385300	4419717	34.4	49	7
5	4520015	4531960	11.9	17	7
6	4810446	4863786	53.3	54	17
7	6391620	6404087	12.4	16	2
8	7521245	7538449	17.2	18	8

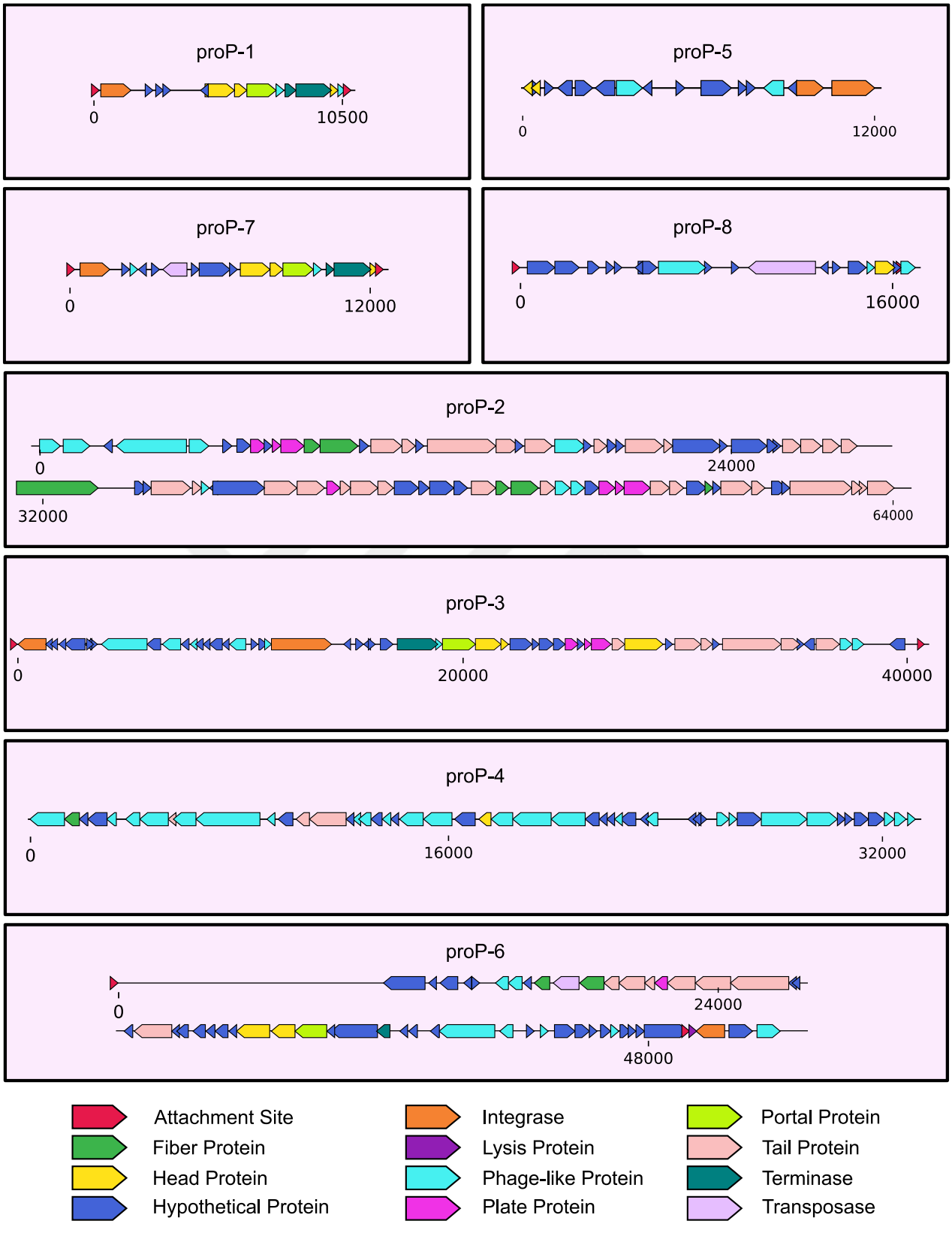


Figure 4.5. Annotation of prophages in *P.* sp. BIOMIG1^{BAC}. Hypothetical proteins and phage-like proteins are abundant. Also, phage-like proteins are not certain. Small prophages have head protein (capsid). On the other hand, large prophages also have tail proteins. Attachment sites and integrases are related, and most of them are found close to each other.

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are sequentially sequenced palindromic repeats and phage related short sequences called spacers that are observed in the genome of bacteria and archaea (Lino et al., 2018). There are sequences of phage DNA in the sections following the direct repeats (DR). As a result of the analysis for *Pseudomonas* sp. BIOMIG1^{BAC}, one possible CRISPR locus, was found in the genome. The total length of the loci is 268 bp, and there are 4 spacers inside. The nucleotide sequence of the DR is "GTTCACTGCCGCATAGGCAGCTAAGAAA" and it is 28 bp.

DRs are repeats which are specific to bacteria, and spacers are short sequences that are part of the phage genomes. The insertion of spacers into the chromosome is important to prevent future invasions of the phage. DRs of *P.* sp. *(alexanderii)* BIOMIG1^{BAC} were searched. It was observed that *P.* sp. *(alexanderii)* CMR12a, *P.* sp. *(piscis)* CMR5c, *P.* sp. *(piscis)* CMAA1215 species had the same DRs, but their spacers were different. On the other hand, the DR sequence of BIOMIG1^{BAC} did not coincide with the DR sequences in *P. protegens* Pf-5 and *P. saponiphila* DSM 9751 strains which are in *P. protegens* subgroup. Therefore, different DR sequences can also be supporters of taxonomic identifications. Also, they show the historical wounds of previous phage invasions. The DR sequences of the Clade 1 strains are the same, so that the DR sequence can be thought of as a common feature for Clade 1 strains. Their different spacers show that they are affected by various phages.

In addition, any of the cas and csy protein groups associated with CRISPR are not found in BIOMIG1^{BAC}. On the other hand, CMR12a contains CRISPR Cas1, helicase Cas3, Csy1, Csy2, Csy3, Csy4 genes. CMR5c and CMA1215 contain CRISPR helicase Cas3, Csy1, Csy2, Csy3 genes. Also, CMR12a has 38 spacers. CMR5c has 18 whereas BIOMIG1^{BAC} and CMAA1215 have only 4 spacers. This shows that although bacteria are phylogenetically close, the environments in which they are found and the phages they are affected by are quite different (Perneel et al., 2007). In particular, CMR12a is ahead in terms of CRISPR content. The phage DNAs found in spacers could not be found in databases, so it can be concluded that they belong to undiscovered phages.

Previously, tolerance of BIOMIG1^{BAC} to various antibiotics were experimentally shown (Gul, 2016). After completing the genome, antibiotic resistance mechanisms, related genes, and antibiotics are identified with bioinformatics (Table 4.4). According to the results, BIOMIG1^{BAC} has various resistance mechanisms for a wide range of antibiotics. ISs are also found alongside antibiotic resistance related genes, and ISs can be effective agents for their mobility. Biochemical experiments are needed for further analysis and their mobility.

Table 4.4. Antibiotic resistance genes which are found in the strain BIOMIG1^{BAC} and their mechanisms.

Mechanism	Genes	Antibiotic Resistances
Antibiotic inactivation enzyme	AAC(6')-Ic,f,g,h,j,k,l,r-z, CatB family	arbekacin, amikacin, azidamfenicol, chloramphenicol, dibekacin, sisomicin, gentamicin B, isepamicin, kanamycin A, netilmicin, neomycin, thiamphenicol, tobramycin
Antibiotic target in susceptible species	Alr, Ddl, dxr, EF-G, EF-Tu, folA, Dfr, folP, gyrA, gyrB, Iso-tRNA, kasA, MurA, rho, rpoB, rpoC, S10p, S12p	bicyclomycin, brodimoprim, clofazimine, ciprofloxacin, coumermycin A1, clorobiocin, coumermycin, D-cycloserine, dapsone, daptomycin, enacyloxin IIa, fusidic acid, fosfomycin, fosmidomycin, gatifloxacin, iclaprim, isoniazid, kirromycin, levofloxacin, moxifloxacin, mupirocin, nalidixic acid, novobiocin, ofloxacin, pulvomycin, rifamycin, rifabutin, rifampin, sparfloxacin, streptomycin, sulfadiazine, sulfadimidine, sulfadoxine, sulfamethoxazole, sulfisoxazole, sulfacetamide, mafenide, sulfasalazine, sulfamethizole, tetracycline, tetroxoprim, tigecycline, triclosan, trimethoprim, trovafloxacin
Antibiotic target protection protein	BcrC	bacitracin
Antibiotic target replacement protein	FabG, fabV, HtdX	triclosan
Efflux pump conferring antibiotic resistance	EmrAB-OMF, EmrAB-TolC, MacA, MacB, MdtABC-OMF, MdtABC-TolC, MexAB-OprM, MexCD-OprJ, MexCD-OprJ system, MexEF-OprN, MexEF-OprN system, MexJK-OprM/OpmH, TolC/OpmH, TriABC-OpmH	ampicillin, amoxicillin/clavulanic acid, azithromycin, aztreonam, ceftazidime, ceftriaxone, chloramphenicol, ciprofloxacin, colistin, erythromycin, gentamicin C, meropenem, nalidixic acid, novobiocin, norfloxacin, panipenem, tetracycline, ticarcillin, triclosan, trimethoprim, ofloxacin, sulfamethoxazole,
Gene conferring resistance via absence	gidB	streptomycin, ofloxacin
Protein altering cell wall charge conferring antibiotic resistance	GdpD, PgsA	daptomycin
Protein modulating permeability to antibiotic	OccD1/OprD, OccD2/OpdC, OccD3/OpdP, OccD4/OpdT, OccD6/OprQ, OccK10/OpdN, OccK5/OpdH, OccK8/OprE, OprB, OprB family, OprD family, OprF	imipenem
Regulator modulating expression of antibiotic resistance genes	OxyR	isoniazid

A gene cluster related to the copper homeostasis mechanism can be located between 4,533,727 and 4,552,212 bp in SSGR-5. There is no pathway assignment for the cluster. On the other hand, it can be possible to deduce the mechanism with the help of gene annotations. There are two annotated RND-Type transporter coding genes in the cluster. They can have similar functions like the RND-driven tripartite proteins, which perform transportation for copper homeostasis in E. coli (Kim et al., 2011). Moreover, TolC family and copper-binding proteins can have complementary functions. Other genes in the cluster can be effective, according to the more recent research (Solioz, 2018). ATPase gene in the cluster is probably used for exporting copper from the cytoplasm to the periplasmic space. Histidine kinase and transcription factors are very likely sensing copper to regulate efflux pumps.

Cu(I) to Cu(II) oxidation can be carried out by the oxidase gene. There are also hypothetical proteins. Experimental research is needed to demonstrate a complete mechanism. Besides, cobalt, nickel, and arsenate related gene clusters are located between 4,555,271 and 4,751,592 bp. Moreover, mercury resistance, regulator and transporter genes are located between 2,453,284 and 2,454,120 bp (SSGR-7), also between 7,548,356 and 7,551,632 bp (SSGR-27). They are flanked by ISs. Therefore, strain BIOMIG1^{BAC} has resistance to various numbers of metals, and it has the ability to detoxify some of them. They are located near ISs, and some of them are flanked with ISs.

There are proteins involved in the biodegradation of aromatic hazardous pollutants in the extracted metabolic functions of the strain BIOMIG1^{BAC}. Considering the genes present in its genome, other pollutants that strain BIOMIG1^{BAC} can degrade apart from QACs include tetrachloroethene, 2,4-dichlorobenzoate, benzoate, atrazine, 1- and 2-methyl naphthalene, styrene, naphthalene, anthracene, bisphenol A, caprolactam, g-hexachlorocyclohexane, fluorobenzoate, 1,4-dichlorobenzene, ethylbenzene, toluene, xylene, geraniol, 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT), biphenyl and trinitrotoluene.

Overall, strain specific genes of the strain BIOMIG1^{BAC} and their functions were identified with protein comparisons of the phylogenetic neighbors of the strain BIOMIG1^{BAC}. According to the proteome comparison, the genome of the strain BIOMIG1^{BAC} has strain specific prophages and genomic regions. Their exact locations in the genome were determined. The gene clusters (BAC, BDMA, and Benzoate degradation) associated with QAC biodegradation were identified as strain specific gene clusters. HGT mechanisms were effective for the acquisition of these clusters. Moreover, there were other strain specific genes that can encode proteins related to the biodegradation of other aromatic hazardous pollutants. Most of the pathways probably contain novel enzymes because there were missing links according to the searches in the pathway databases like KEGG and GO. As a result, strain BIOMIG1^{BAC} has an expanded flexible genome with unique ISs, prophages, metal resistance, antibiotic resistance, and biodegradation mechanisms of various pollutants.

5. POTENTIAL GENE MOBILITY MECHANISMS OF *PSEUDOMONAS* SP. BIOMIG1^{BAC}

5.1. Introduction

Quaternary ammonium compounds (QACs), which are also called quats are widely used as disinfectants in personal care products, biocides, and hygiene products (Zhang et al., 2015; Zheng et al., 2020). In the center of their molecular structure, there is a positively charged nitrogen atom that binds four alkyl groups which are consist of at least one hydrophobic hydrocarbon chain with other short-chain groups like methyl or benzyl (Zhang et al., 2015). Mostly used variations of QACs are benzylalkyldimethylammonium compounds (BACs, that the carbon atom number in the long hydrocarbon chain changes from 8 to 18), dialkyldimethylammonium compounds (DADMACs), and alkyltrimethylammonium compounds (ATMACs) (Zheng et al., 2020). QACs can easily stick to the cell wall surface because of their positively charged structure that causes penetration, insertion, and disruption (Alkhalifa et al., 2020). The cell death because of the leakage of crucial intracellular components and the loss of envelope of the viruses because of the damage can be the results of the QACs, so QACs are very effective to the bacteria and enveloped viruses like SARS-CoV-2 (Dev Kumar et al., 2020; Hora et al., 2020; McDonnell, 2009; Tezel and Pavlostathis, 2015). Compounds with the carbon number of 10-12 are more effective (Morrison et al., 2019). One of the lists of disinfectants against SARS-CoV-2 was published by the U.S. Environmental Protection Agency (EPA). There are 216 QACs containing products among 430 products in the list (Hora et al., 2020). EPA demonstrated BACs, DADMACs, and ATMACs as high production volume chemicals that their production or importation is over 1 million pounds per year. According to the report of the one manufacturer, the production of disinfectants in just one month in 2020 is the same as the production of the previous year (Hora et al., 2020). The changes between the usage of the C10-BAC, C12-BAC, and C14-BAC before and during the SARS-CoV-2 are 154%, 114%, and 146%, respectively (Zheng et al., 2020). Until the pandemic, usage of QACs was limited in European countries, unlike the U.S.; however, the usage of QACs has increased recently also in Europe.

Industrial, domestic, and hospital use of QACs cause increased concentrations of the QACs in the wastewater, approximately 25% of the QACs are released into the environment, and the rest is discharged to the wastewater treatment systems. However, QACs are not effectively removed in the treatment systems. One of the features of the QACs is the biodegradation possibility under aerobic conditions, so their concentrations are excessively changing (Tezel and Pavlostathis, 2015). Other

possibilities for the concentration decrease of QACs in the aquatic environments are photolysis and adsorption to the different materials because of their high affinity to organic and inorganic substances (Hora et al., 2020).

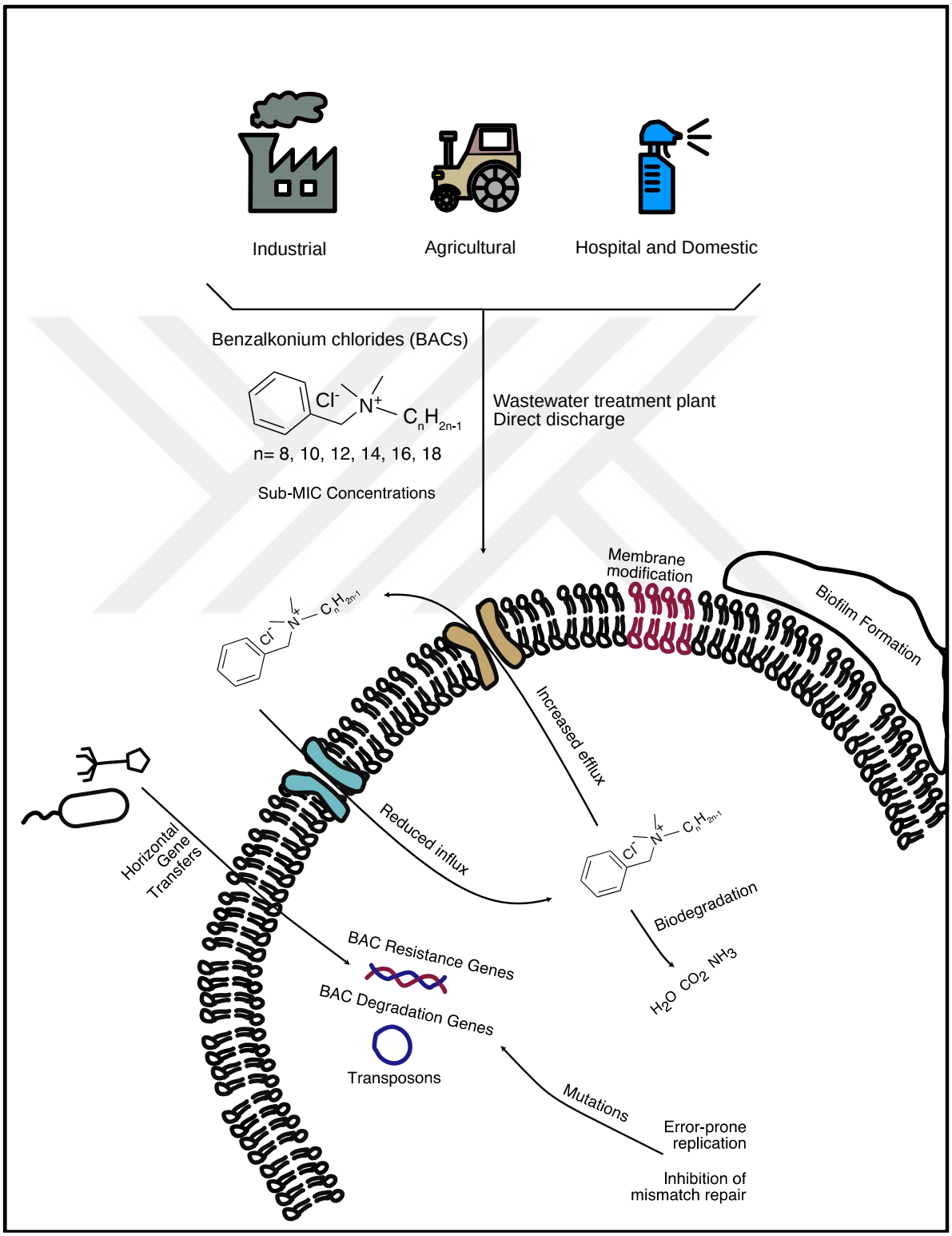


Figure 5.1. Mostly used variations of QACs are BACs. BACs resistance mechanisms (Pereira and Tagkopoulos, 2019; Tezel and Pavlostathis, 2015).

Consequently, the organisms are facing different concentrations of the QACs in different environments. The concentrations which are lower than the minimum inhibitory values for organisms can result in the evolution of resistance mechanisms (Figure 5.1); also, the QACs can be selective and has effects on the microbial communities through natural selection (Tezel and Pavlostathis, 2015). Researches show that QACs are present in different environments like soil, sediments, rivers, residential dust, also in different products like fruits, milk, and others (Zheng et al., 2020). Regarding the extensive discharge to the environment, damaging the aquatic life, causing several health issues, disordering the treatment plants, and promoting antibiotic resistance are concerns of the QACs usage (Hora et al., 2020; Zheng et al., 2020).

QACs can be eliminated from the environment by biodegradation. Biodegradation of QACs without benzyl group are subjected to several studies; however, there are limited studies about biodegradation of BACs (Zhang et al., 2015). The first isolated strain that can completely mineralize BACs is *Pseudomonas* sp. BIOMIG1^{BAC} (Ertekin, 2017). Before the isolation of strain BIOMIG1^{BAC}, the microbial communities from four different sources (Activated sludge (AS) and sewage (SEW) from Pasakoy Advanced Wastewater Treatment Plant, soil sample from Ataturk Arboretum (SOIL), sea sediment sample from Silivri (SEA)) were enriched with dilution to extinction method for obtaining BAC degraders (Ertekin, 2017). *Pseudomonas* sp. BIOMIG1^{BAC} was isolated from SEW sample. These four communities were sequenced with the short-read sequencing method. Furthermore, different phenotypes of the strain BIOMIG1^{BAC} were investigated to find out the responsible genes for BACs biodegradation. The strain BIOMIG1 phenotypes as BAC, BDMA, BD, and N were partially sequenced (Emine, 2017). The strain BIOMIG1^{BAC} can completely degrade BACs. BIOMIG1^{BDMA} is the BDMA accumulator which is converting BACs to BDMA. BIOMIG1^{BD} can't degrade BACs, but it can degrade BDMA. BIOMIG1^{N} can't degrade BACs and BDMA. BAC and BDMA have *oxyBAC* gene clusters. BAC and BD have BDMA degradation gene clusters. Therefore, *oxyBAC* and BDMA gene clusters can be lost according to the environmental factors, and their mobility mechanisms are active. Horizontal transfers of these gene clusters to other bacteria can be possible according to the results of the previous section.

Horizontal gene transfer is one of the driving forces of genetic adaptation for emerging new catabolic capacities. Horizontal gene transfer mechanisms are various. One study showed that horizontal transfer of the tfdA gene from plasmid pRO103 to the phenol degrading recipient bacteria dramatically enhanced phenoxyacetic acid degradation rates (Lipthay et al., 2001). Another study showed similar results that phenol degradation gene can be self-transmissible with the plasmid between the bacteria living in a plant and its rhizosphere (Wang et al., 2007). Another study showed

that the chlorocatechol, which is the aromatic pollutant degradation genes are self-transmissible horizontally between the strains of two different *Pseudomonas* species by bacterial conjugation (Ravatn et al., 1998). Plasmids are known to be highly effective agents of horizontal gene transfers; however, the strain BIOMIG1^{BAC} does not have any plasmid in its genome. Therefore, other mechanisms can be responsible for the *oxyBAC* gene cluster mobilization. Recent studies showed that antibiotic-resistance genes could be transferred via phage-related mobile elements (Billard-Pomares et al., 2014; Brown-Jaque et al., 2015; Miller and Balcazar, 2014). Furthermore, transposons can be effective for genetic adaptation. The existence of a transposase that detects direct repeats is required for integration; however, they can't horizontally transfer themselves. There are also examples of horizontal transfers of transposons via phages. Moreover, gene transfer agents (GTAs) have phage-like structures that can pack and transmit random segments of the host DNA. GTAs have mostly been found in bacterial species generally in alpha-proteobacteria isolated from seas. (Brown-Jaque et al., 2015). A recent study showed that phage-inducible chromosomal islands (PICIs) are novel mobile genetic elements that are blocking phage packaging and promoting PICI packaging (Fillol-Salom et al., 2019). Bacteria can use phage packing machinery with the help of the protein inside the PICIs to transport their genetic materials.

One of the interesting mobility mechanisms of genetic materials was identified recently. Harmen and Hall showed that a kanamycin resistance gene in between two identical ISs forming composite transposon could form translocatable unit (TU) with the help of the transposase. Kanamycin selection in the environment can trigger the insertion or excision of the translocatable unit. One of the flanking ISs is the active element for transposition. Adjacent DNA segments can be mobilized by the IS26, and it can be transported to the section which has an existing copy of the IS26. Therefore, composite transposons can occur. According to the study, frameshift mutations on the inverted repeat sections can inhibit the transposition process (Harmer and Hall, 2015). Therefore, the environmental stresses and different genetic transfer mechanisms are very effective for gaining or losing the genetic materials. The objective of this section was to determine and identify the structure of the *oxyBAC* gene cluster. Different phenotypes of the strain BIOMIG1 and other strains contain *oxyBAC* gene were compared to reveal the possible mobility mechanisms.

5.2. Materials and Methods

5.2.1. *oxyBAC* Gene Cluster

Previously, Ertekin and colleagues showed that a gene cluster consists of 8 genes was related to BACs biotransformation in *Pseudomonas* sp. BIOMIG1^{BAC}. BACs degradation ability of a novel Rieske-type oxygenase enzyme named *oxyBAC* was shown experimentally (Ertekin et al., 2017). However, the surrounding genes of the cluster were not determined because of the limitations of short-read sequencing. With the complete genome, the surrounding genes of the cluster were determined. 100% identical flanking regions with transposase coding gene were noticed with BLAST searches (v. 2.2.31+), and terminal inverted repeats (TIR) were extracted manually inside the regions. There was no similar insertion sequence in the databases (Camacho et al., 2009). ISEScan (v. 1.7.2) was also used for confirmation, and the flanking region was clustered as a novel insertion sequence (Xie and Tang, 2017).

Protein *BLAST* (blastp) was used for the search of *oxyBAC* to find other bacteria which have the *oxyBAC* gene. *Pseudomonas saponiphila* DSM 9751, *Pseudomonas nitroreducens* B, *Pseudomonas stutzeri* 95A6, *Novosphingobium* sp. B-7 and *Sinobacteraceae bacterium* Bin_59_2 have *oxyBAC* gene and *oxyBAC* gene cluster. Their genomes were downloaded from NCBI. Their genomes were draft. Only *Pseudomonas saponiphila* DSM 9751 has a complete cluster with flanking ISs. So, partial ones were assembled with reference mapping in *Geneious* (v. 7.1.8). Their *oxyBAC* gene clusters were aligned for comparative analysis. Also, their 16S rRNA genes were used to create a phylogenetic tree with *ClustalW* and Tamura-Nei Neighbor-Joining method. The closest two strains for each were added to the tree to show horizontal gene transfer probability of *oxyBAC* gene cluster. *Methanothermobacter* sp. KEPCO1 was used as an outgroup. *RNAfold* (v. 2.4.18) web server was used for prediction of TIME's RNA structure with minimum free energy (MFE) and partition function (Gruber et al., 2008; Mathews et al., 2004). *PopART* was used for haplotype network analysis of *oxyBAC* gene cluster with CS method (Clement and Bryant, 2017; Leigh et al., 2015).

5.2.2. Metagenomic Analysis

The metagenomic reads (AS, SEW, SOIL, and SEA) were investigated to obtain more detailed information about mobilization mechanism. Metagenomic reads were short, and the short-reads were limiting to obtain structure of *oxyBAC* gene cluster because the *oxyBAC* gene cluster is flanked by identical ISs. So, assembly graphs were used to show *oxyBAC* gene cluster owners in the community.

MetaWRAP pipeline was used for metagenomic analysis (Uritskiy et al., 2018). Three different binning algorithms (*metaBAT2*, *CONCOCT*, and *MaxBin2*) were used. *GraphBin* (v. 1.4) was also used for refining the binning process (Mallawaarachchi et al., 2020). To visualize the refinement, a Python program was written to visualize the output with python-igraph (v. 0.8) module.

5.2.3. Analysis of BIOMIG1 Phenotypes

Genome sequences of four phenotypes of BIOMIG1 as BAC, BDMA, BD, and N were analyzed to find out the mobility mechanism of *oxyBAC* and BDMA gene clusters. The raw reads were mapped to the flanked ISs of the cluster with *Geneious* (v. 7.1.8). As a control, the raw reads were mapped to the random regions of the genome. Finally, statistical analyzes were made with the mapping results.

5.3. Result and Discussion

Insertion sequences (ISs) and transposons are crucial agents for bacterial evolution. ISs and transposons are carrier agents that change the location of genes in the genome. When they carry these genes to migratory agents such as plasmids and phages, the genes can also be transferred to different bacteria horizontally (Vandecraen et al., 2017). So, their structures are important to find out their mobility mechanisms. With the hybrid assembly, partial ISs of *P.* sp. BIOMIG1^{BAC} were completed. Flanked IS structures of *oxyBAC* and BDMA gene clusters were revealed with the help of longer reads. Flanking ISs of the *oxyBAC* gene cluster are identical, and their lengths are 3853 bp (Figure 5.2). They have 49 bp terminal inverted repeats. They both have 2967 bp transposase (+), 550 bp recombinase (+), and 177 bp type II toxin-antitoxin system (-) coding genes. The novelty of the IS was shown with database searches and *ISEScan* results. It is in the Tn3 family, but it is very different from defined existing ones. Also, *ISEScan* has clustered it as a novel sequence (Xie and Tang, 2017). The total length of the composite transposon is 13,764 bp. There are no direct repeats. The right flank of the left IS and right flank of the right IS have 5 bp reverse repeats, so the transposition mechanism can be driven by the right IS. These reverse repeats can be the combination section of the mobile element when it circularizes itself in the excision. The predicted translocatable unit (TU) size is 9911 bp, and the possible structure is built (Figure 5.2). The *oxyBAC* gene cluster contains 8 genes which are site-specific integrase (+), hypothetical protein (+), HNH endonuclease (-), SCP2 sterol-binding (-), *oxyBAC* (-), TetR transcriptional regulator (+) and two MFS transporters (-). The function of the *oxyBAC* was previously represented (Ertekin et al., 2017). The length of the *oxyBAC* gene cluster is 6058 bp.

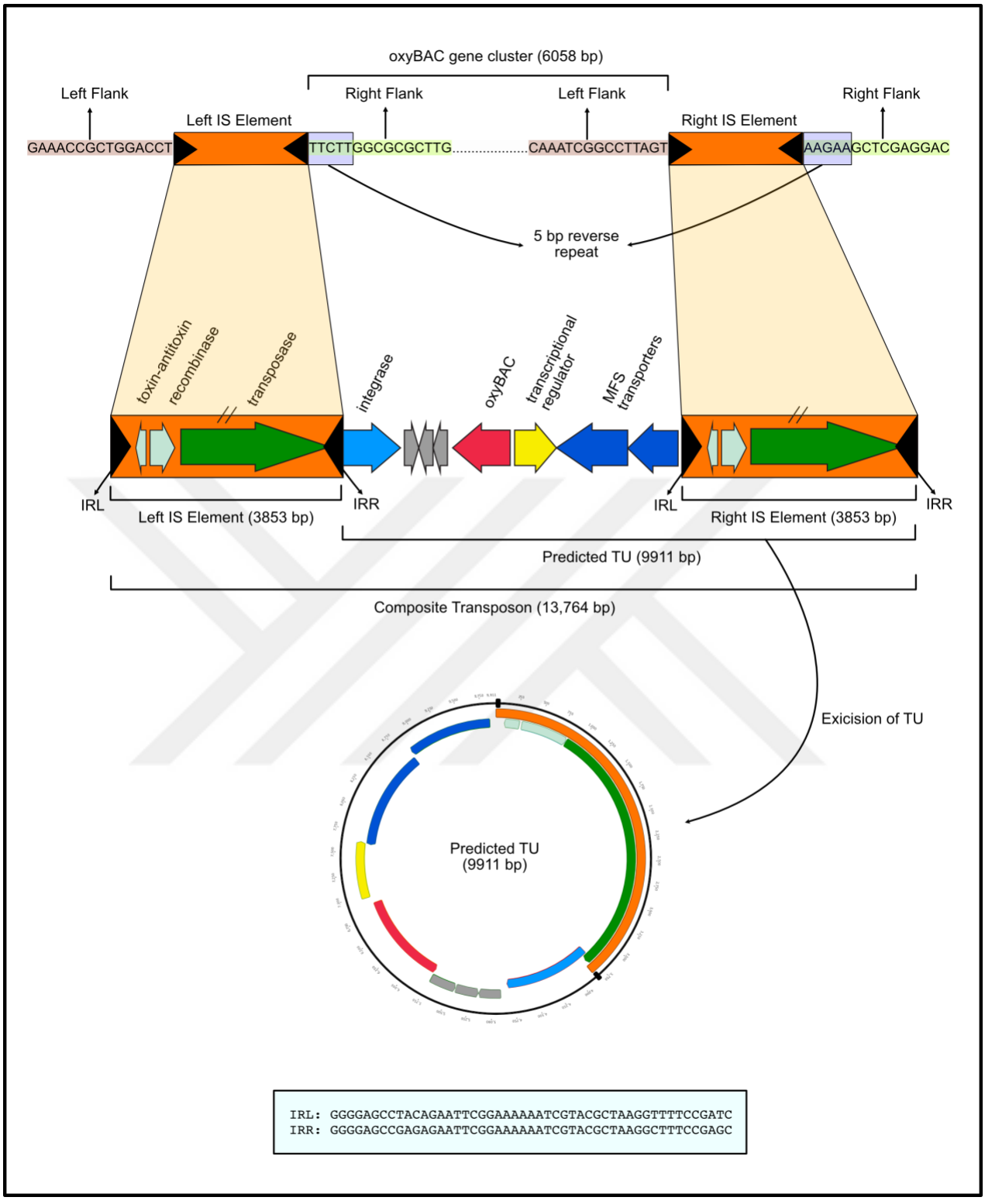


Figure 5.2. Flanking ISs of the *oxyBAC* gene cluster are parts of the composite transposon. Left and Right IS Elements are 100% identical, and they have terminal inverted repeats (TIRs). 5 bp reverse repeat section can be a sign of the transposition. So, predicted TU could be driven by the right IS element, and the left IS element stays in the genome. The *oxyBAC* gene cluster contains 8 genes which are site-specific integrase, hypothetical protein, HNH endonuclease, SCP2 sterol-binding, *oxyBAC*, TetR, and two MFS transporters. IS elements are in the Tn3 family, and they have 3 genes which are transposase, recombinase, and toxin-antitoxin system.

In addition, this composite transposon structure is not found in the genome of *Pseudomonas* sp. CMR12a, the strain closest to *Pseudomonas* sp. BIOMIG1^{BAC}. Although they are so close phylogenetically (ANI value is 99.1%), the absence of the *oxyBAC* gene cluster and the lack of carrier structures in CMR12a supports the horizontal acquisition of these genes. The gene cluster containing *oxyBAC* was also detected in five different bacterial species. These bacterial strains are *Pseudomonas saponiphila* DSM9751, *Pseudomonas nitroreducens* B, *Novosphigobium* sp. B-7, *Sinobacteraceae bacterium* Bin_59_2 and *Pseudomonas stutzeri* 95A6 (Figure 5.3). These species are phylogenetically quite distant from each other. However, each contains the *oxyBAC* gene cluster in its genomes. 16S rRNA tree was built to represent their close strains do not have *oxyBAC* gene cluster (Figure 5.3). According to the tree, it can be said that these genes were transferred horizontally. In these bacterial strains, 8 genes in the *oxyBAC* gene cluster are 90% or more similar. This gene cluster is flanked by ISs in other bacteria as well. Although the composite transposon structures are different because different ISs are found, the *oxyBAC* gene clusters in all strains are flanked by ISs. So, their mobility is related to ISs. The *oxyBAC* gene cluster, whose gene sequence is almost completely conserved, is found in different bacteria and different parts of their genomes. As a result, this gene cluster circulates between different types of bacteria by horizontal gene transfer.

The short reads of mutant strains of BIOMIG1^{BAC} were used to support the predicted TU of *oxyBAC* gene cluster in BIOMIG1^{BAC}. If the excision of the predicted TU is true, one of the flanking Tn3 IS must leave the genome after excision. Another assumption for testing the TU is that the short reads are evenly distributed. BIOMIG1^{BDMA} and BIOMIG1^{BAC} have *oxyBAC* gene cluster. On the other hand, BIOMIG1^{N} and BIOMIG1^{BD} do not have *oxyBAC* gene cluster. So, BIOMIG1^{N} and BIOMIG1^{BD} must have one less Tn3 IS. The short reads are mapped to the Tn3 IS. The numbers of the mapped short reads are compared (Table 5.1). Total reads were not equal for them, so all of them were proportioned to 10^7. The ratios were not perfect, but there were significant differences. As expected, BIOMIG1^{N} and BIOMIG1^{BD} have 3 Tn3 IS. BIOMIG1^{BDMA} and BIOMIG1^{BAC} have 4 Tn3 IS according to the mapped short read numbers. There are actual numbers for IS number and mapped reads for BIOMIG1^{BAC} because of the hybrid assembly. Others were predicted and compared with BIOMIG1^{BAC}. For the control of even distribution of the reads, short reads were mapped to the 5 different random segments of the genome (Table 5.2). They do not have any significant differences (the maximum difference with the average is 5%), unlike the mapping of the Tn3 IS. Therefore, it is an indicator of the excision of the predicted TU, and the one, the flanking ISs, is the active transposable element in the excision of the *oxyBAC* gene cluster.

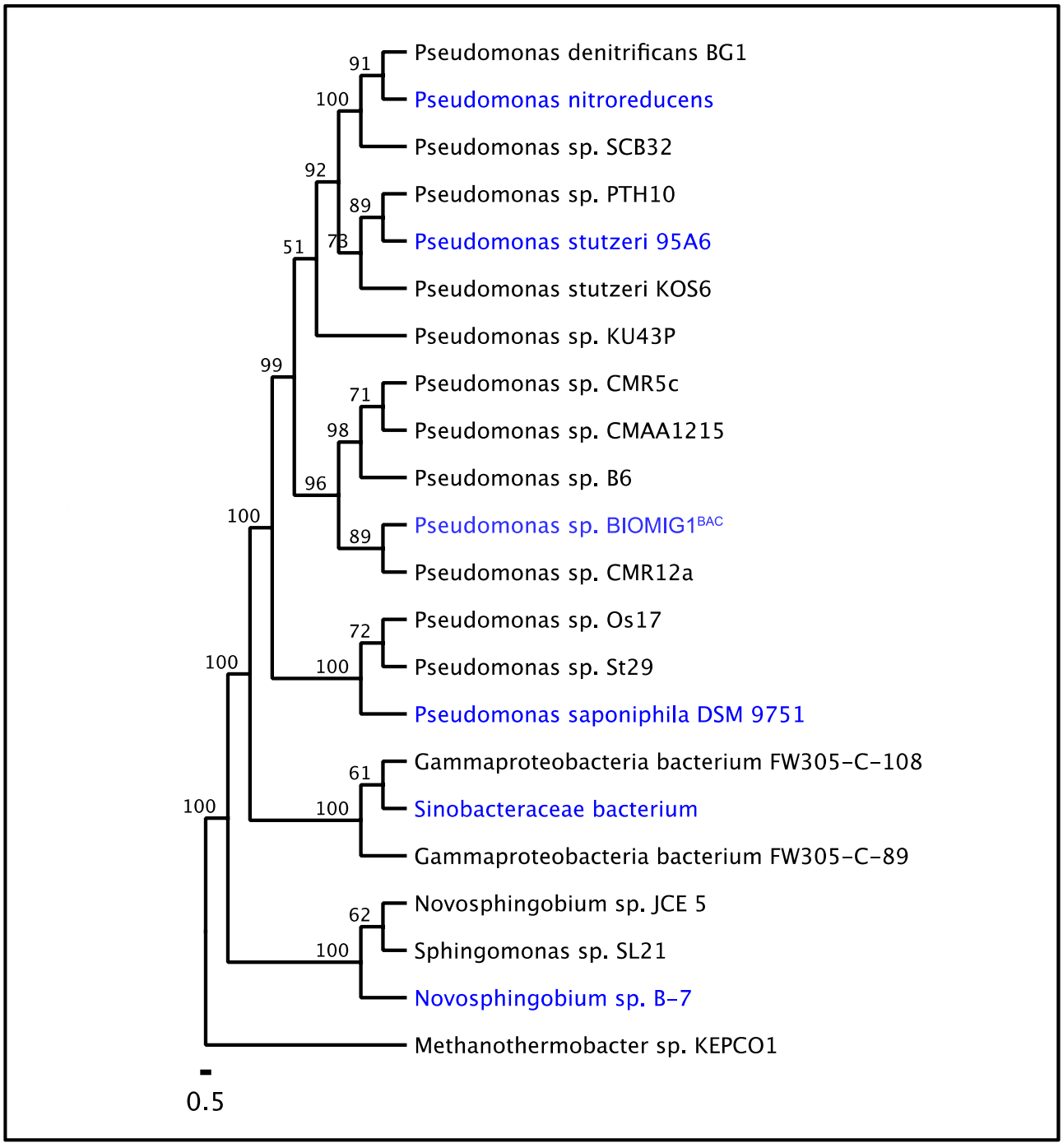


Figure 5.3. 16S rRNA tree of *oxyBAC* having (blue colored) strains and their close strains. The tree clearly shows the horizontal transfer of the *oxyBAC* gene cluster. *Methanothermobacter* sp. KEPCO1 is the outgroup.

The structures were aligned to reveal the differences of the composite transposon structures of the strains that have *oxyBAC* gene cluster. However, only the genome of *Pseudomonas* sp. BIOMIG1^{BAC} was complete; other genomes were draft. Although the genome of *Pseudomonas saponiphila* DSM9751 was draft, its composite transposon was complete. ISs in other genomes were not revealed because of the limitation of the draft genomes. The composite transposons of *Pseudomonas nitroreducens* B and *Pseudomonas stutzeri* 95A6 were completed with the help of the

reference mappings and sequence searches. However, these processes were not enough to complete the composite transposons of *Novosphigobium* sp. B-7 and *Sinobacteraceae bacterium* Bin_59_2. Moreover, the metagenomic reads of the activated sludge (AS) were used for comparison. Its transposon structure was completed with the help of *Pseudomonas saponiphila* DSM9751, because its composite transposon structure was different than *Pseudomonas* sp. BIOMIG1^{BAC}. There are mainly two different composite transposon structures (Figure 5.4). *P. saponiphila* DSM9751, *P. nitroreducens* B, *P. stutzeri* 95A6, and AS have flanked IS1380 (1667 bp) transposon with terminal inverted repeats. Also, there is a single copy IS21 family (2492 bp) IS in the composite transposon with some additional genes. On the other hand, *P.* sp. BIOMIG1^{BAC} has another composite transposon with Tn3 family ISs. The eight genes of the *oxyBAC* gene cluster are common for all of them. *Novosphigobium* sp. B-7 does not have all genes, but it is probably because of the deficient quality of its genome.

Table 5.1. The ratio differences of the mapped short reads to the Tn3 IS indicate that a Tn3 IS Element is separated with *oxyBAC* gene cluster.

	Pseudomonas sp. BIOMIG1 Phenotypes	# Actual Numbers	# Mapped Short Reads	# Mapped Short Reads (Proportioned to 10^7)
Total bp of Tn3 IS	BAC	14578	16872	15768
Number of Tn3 IS	BAC	4	4,37	4,327113063
Total bp of Tn3 IS	BD	-	12088	10697
Total bp of Tn3 IS (Prediction)	BD	-	3,13	2,992167832
Total bp of Tn3 IS	BDMA	-	18027	16095
Total bp of Tn3 IS (Prediction)	BDMA	-	4,67	4,416849616
Total bp of Tn3 IS	N	-	11332	10209
Total bp of Tn3 IS (Prediction)	N	-	2,94	2,85566433

Table 5.2. The control of the mapped short reads with the random segments.

		# Mapped Short Reads to Pseudomonas sp. BIOMIG1 Phenotypes (Control)				
	Random segments (bp)	N	BDMA	BD	BAC	Average
1	4500	3735	4476	4655	4130	4249
2	4362	8255	7730	8075	7576	7909
3	4401	5693	5936	5989	5429	5761,75
4	4364	5974	6811	5290	6271	6086,5
5	4449	8855	9602	8178	8866	8875,25
Total	22076	32512	34555	32187	32272	32881,5
Total	Difference with average (%)	1,12373219	-5,08948801	2,112129921	1,853625899	0

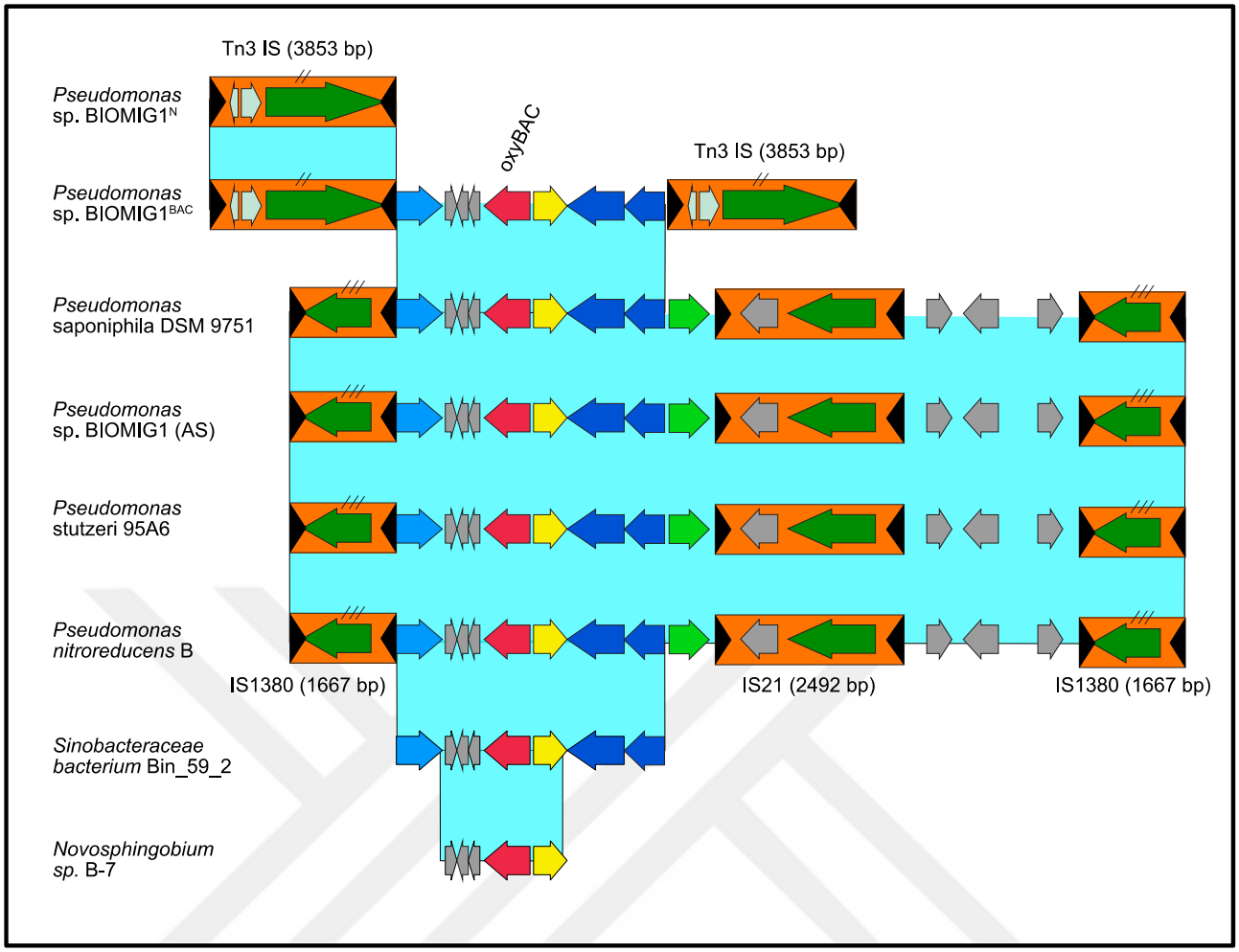


Figure 5.4. The *oxyBAC* gene cluster was found in two different composite transposon structures.

The *oxyBAC* gene cluster was found in metagenome reads of Activated Sludge (AS), but it was not clear which strains had the gene cluster because of the limitation of the short reads. Moreover, the binning algorithms are not good at clustering short contigs. Different approaches were compared to show which strains had the *oxyBAC* cluster. The assembly graph of the AS metagenome was drawn with the clusters. However, there were too many unbinned (black-colored) nodes in *MetaWRAP* binning (B) (Figure 5.5). It was not enough to determine the owner of the *oxyBAC* cluster. A more recent binning refinement method (*GraphBin*) which is using connectivity information from the assembly graph for unbinned nodes, was used (C) (Mallawaarachchi et al., 2020). With the help of the refinement, it can be said that *Pseudomonas* sp. BIOMIG1 (AS) has the *oxyBAC* gene cluster because the *oxyBAC* cluster is surrounded by *Pseudomonadales* assigned nodes. There is only one strain of *Pseudomonadales* in the AS metagenomes that is BIOMIG1. Two different composite transposon mechanisms can be effective in transferring *oxyBAC* gene cluster even for the same species. Therefore, the same species can have different structured composite transposons.

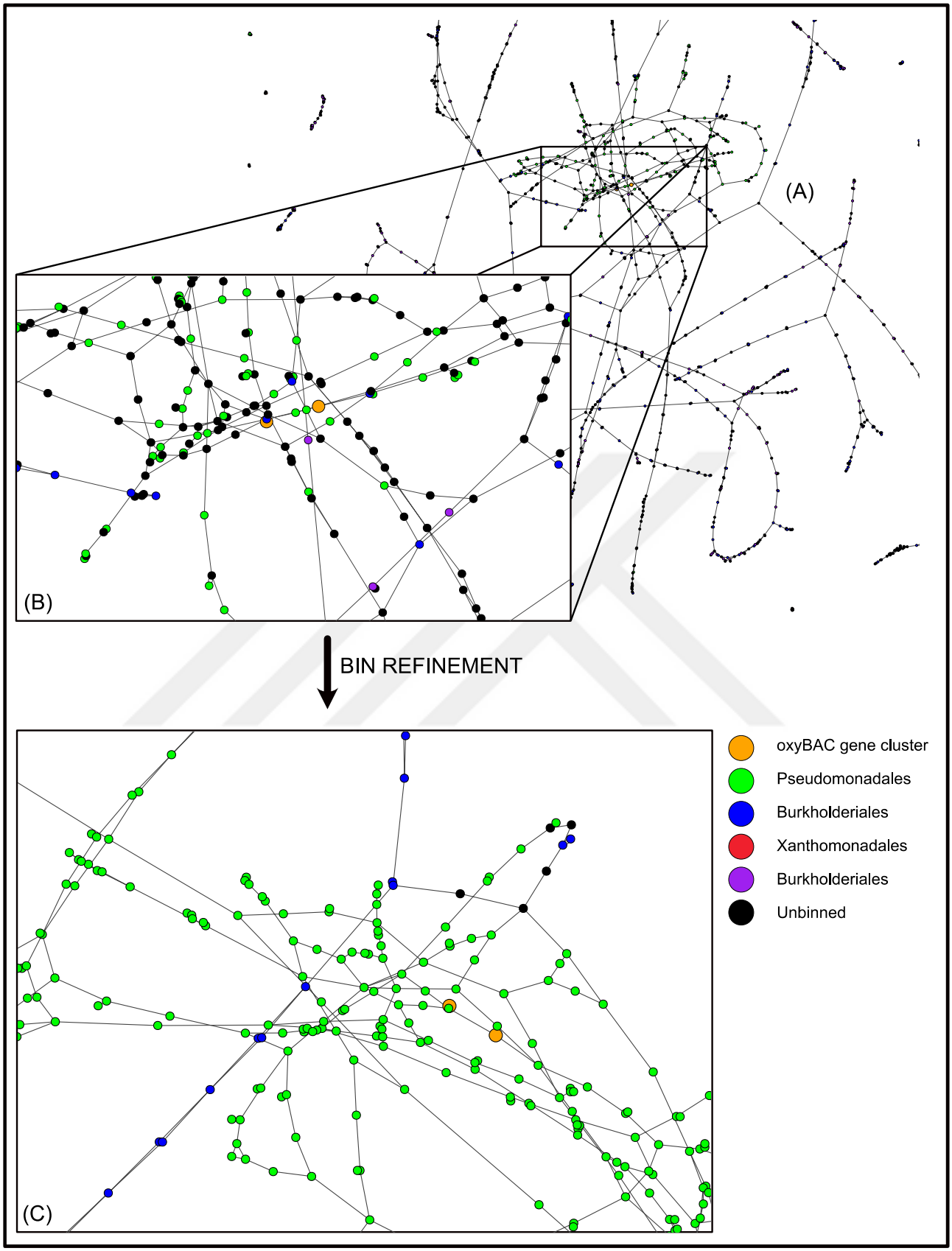


Figure 5.5. (A) Assembly graph of the AS metagenome. (B) *MetaWRAP* Binning (C) Bin refinement with *GraphBin*.

Flanking Tn3 family ISs of the *oxyBAC* gene cluster were also found in another region in the genome. It has almost 100% identity, but it is in a different orientation as reversed and divided into two parts. The 5 bp repeats were found between the parts. The insertion of a Tn3-Derived Inverted-Repeat Miniature Element (TIME) in Tn3 IS was revealed with the help of direct repeats (DRs) (Figure 5.6). 262 bp TIME has inverted repeats. It does not have any gene coding region.

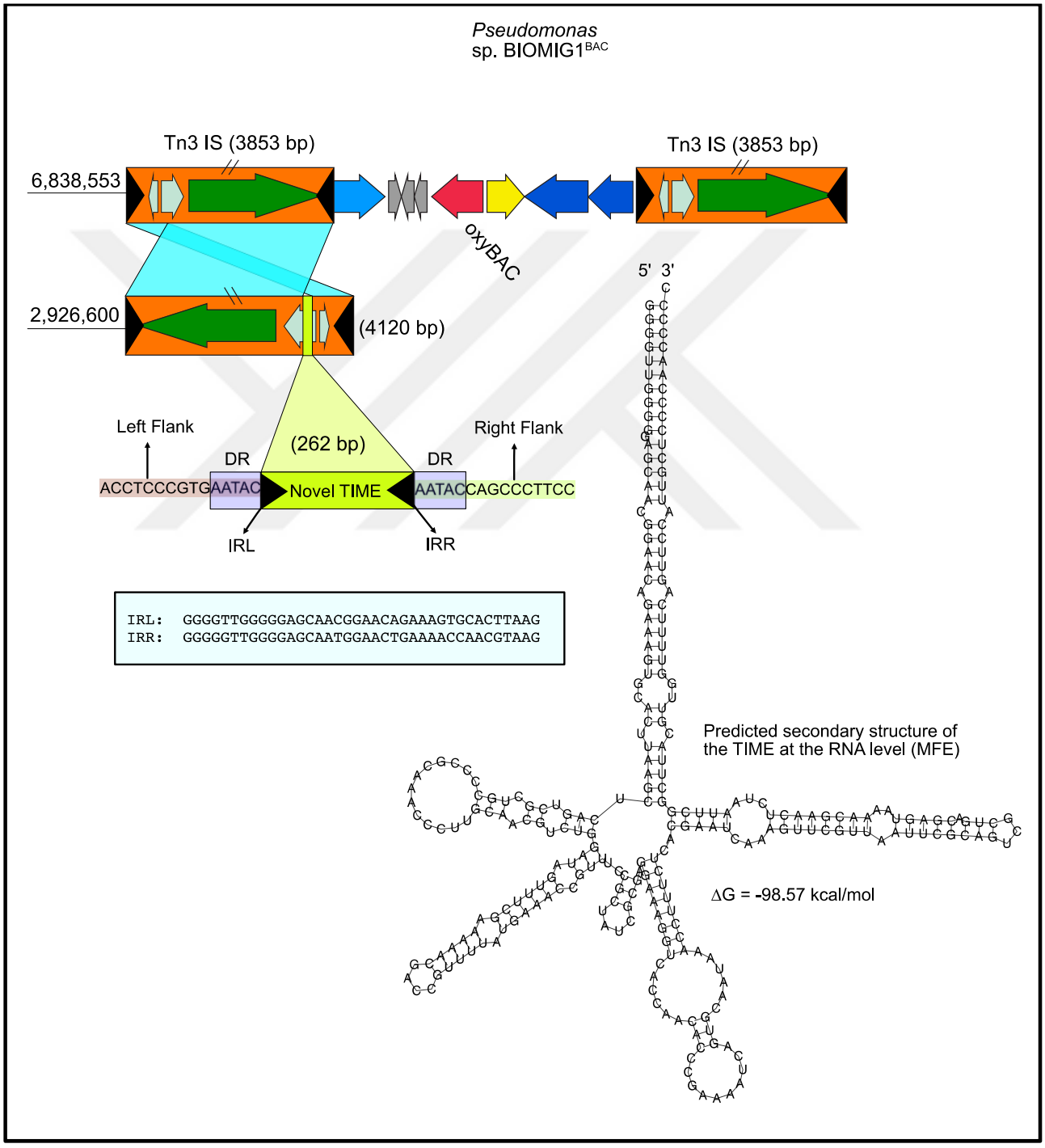


Figure 5.6. The insertion of a Tn3-Derived Inverted-Repeat Miniature Element (TIME) in Tn3 IS was created direct repeats (DRs). A predicted secondary structure of the TIME at the RNA level with MFE was built.

TIMEs are 262 bp non-autonomous elements (Hayes et al., 2014). They could be mobilized by transposons if the inverted repeats were compatible. The TIME in the BIOMIG1^{BAC} has divided the recombinase coding gene in the Tn3 transposon. It is not known how TIME affects the function of the recombinase, and there is only a single copy of this TIME in the genome. Another important thing is that the TIME can be moved to the flanking ISs of the *oxyBAC* gene cluster because the insertion sites are the same. It is unknown how this possibility can affect the *oxyBAC* cluster or transposition of the *oxyBAC* gene cluster.

Haplotype network analysis of *oxyBAC* gene cluster was made to find out the evolutionary history of the cluster. According to the result of TCS network analysis (Figure 5.7), there is no different haplotypes coming from an ancestral one, so it is impossible to find which is more ancestral. For integrase gene, *P. nitroreducens* B and *P. saponiphila* DSM 9751 are identical. Also, *Sinobacteraceae Bacterium* and *Pseudomonas stutzeri* 95A6 are identical. However, the difference is high (80 mutations) between identical groups. Although *oxyBAC* gene is more conserved than integrase gene, there is no identical grouping like integrase. Transcriptional regulator gene is identical in all strains. Consequently, the genes in the cluster are evolved independently from each other.

Overall, the structure of the *oxyBAC* gene cluster was identified with the help of the complete genome sequence and manual annotations according to the literature. The *oxyBAC* gene was found in two different composite transposon structures with a comparison of other species that have the *oxyBAC* gene. The *oxyBAC* gene mobility was associated with flanked ISs according to the comparison of other phenotypes of the strains BIOMIG1^{BAC}. According to the other phenotypes do not have *oxyBAC* gene, the lack of one IS can be sign of the translocatible unit formation. A translocatable unit could be formed in the mobilization of the *oxyBAC* gene. Flanked ISs formed the composite transposon structure for all strains that have *oxyBAC*. Future experimental studies are needed for the demonstration of the exact mobility mechanism.

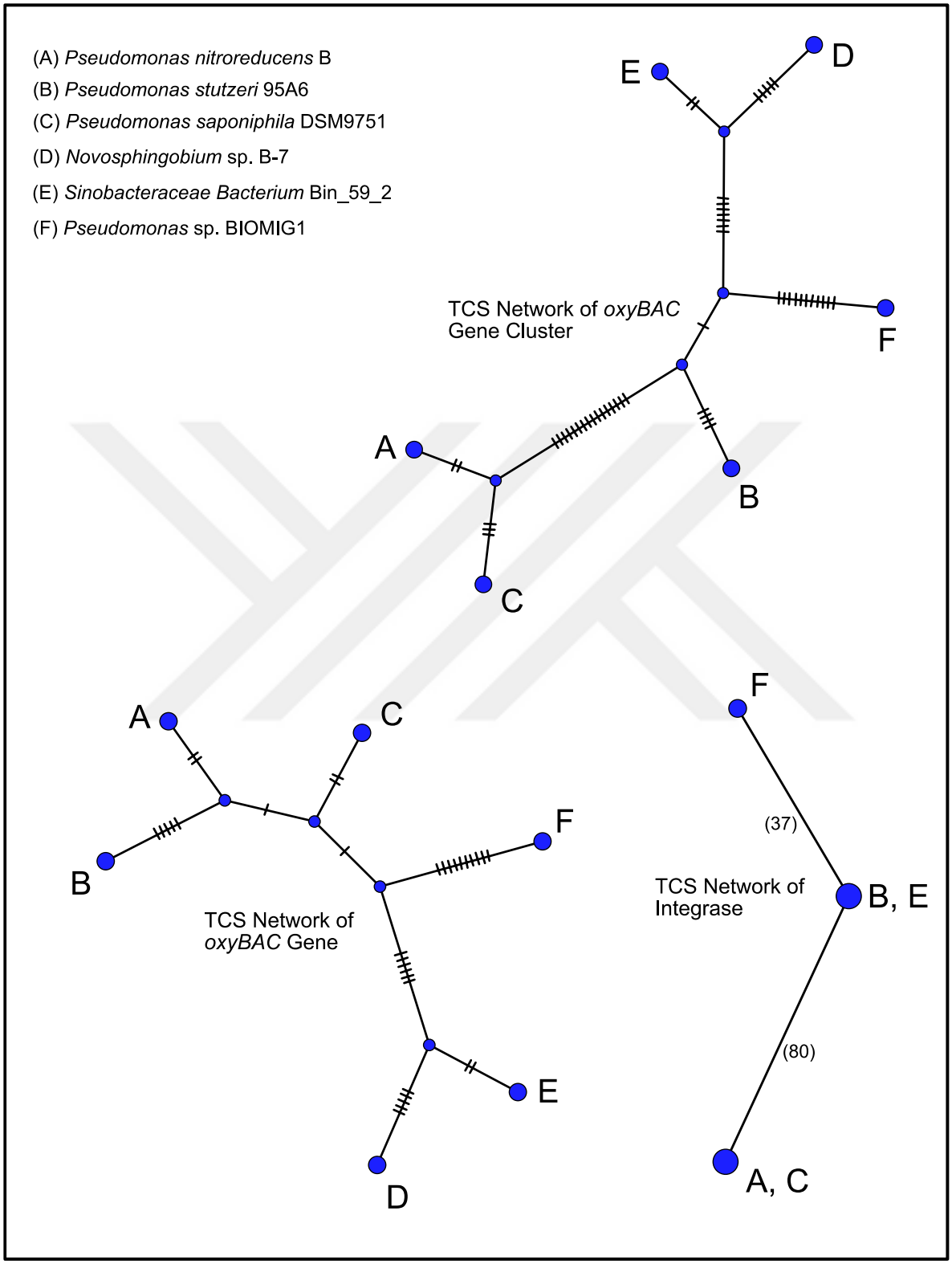


Figure 5.7. TCS network analysis for *oxyBAC* gene cluster, *oxyBAC* gene, and integrase gene in the cluster. The hatch marks in the edges represents the number of mutations.

6. CONCLUSIONS AND RECOMMENDATIONS

With the SARS-CoV-2 pandemic, the usage areas of disinfectants containing BACs, which are a group of QACs, have expanded, and their consumption has increased. On the other hand, the mechanisms that microorganisms develop or gain may reduce the antimicrobial effects of these compounds and can make it challenging to control pandemics in the future. In addition, it has been shown in previous studies that such compounds affect the spread of antibiotic resistance in nature, and they cause many environmental and health problems. Uncovering the complete genome of the *Pseudomonas* sp. BIOMIG1^{BAC}, the first to mineralize BACs completely, was crucial in this sense. The complete genome can be a vital resource for future environmental applications of the strain BIOMIG1^{BAC}. High-quality genome assembly was generated with a hybrid method. The strain BIOMIG1^{BAC} has a 7,675,262 bp circular chromosome and one 34 Kbp circular translocatible unit (TU). The strain BIOMIG1^{BAC} has an expanded flexible genome with some unique ISs, prophages, transposons, and abilities to degrade various organic pollutants.

According to the results of the comparison of different genome assembly inputs, the higher number of long-reads is not directly proportional to high quality assembly; so, picking the optimum intervals for filtering and processing the genomic data is important. However, if the average accuracy of long-read technologies like Oxford Nanopore will be increased in the future, it can be possible to obtain complete genomes with just long-reads. The hybrid methodology is very powerful; however, there can be some shortcomings. To overcome these shortcomings, using assembly graphs to resolve fragmentation can give successful results. Also, the assembly graphs can be used to extract composite transposons because composite transposons are very difficult to find in assemblies with just short reads. Taking into account that most of the genomes in the databases are still draft, extra computational works are needed.

Comparative genomics is very robust for revealing the taxonomy. The strain BIOMIG1^{BAC} is proposed as a type strain of the novel species tentatively named *alexanderii* that belongs to the *Pseudomonas protegens* subgroup in *Pseudomonas fluorescens* group. The results in this study sufficiently provided the minimal standards required for the new species identification. Furthermore, the strain BIOMIG1^{BAC} has a considerable number of differences even with the close strains in the subgroup. These differences made help to determine horizontal gene transfers and strain specific regions. Systematic and autonomous ways of the taxonomic assessment can be possible, but the literature is very dynamic. Although there are automatic tools that are helpful, manual curations are

crucial. The genes that are related to xenobiotics biodegradation and metabolism have the highest proportion when compared to other genes in strain specific regions. There are still too many unknown coding regions, especially in the strain specific regions; so, transcriptomic analyses are needed in future studies.

The genome of BIOMIG1^{BAC} has eight prophages in it. The ISs found in the prophages in the genome can be a sign of phage-related mobility of the genomic content. The excision of the prophages in the future can transport genomic content to other bacteria. Also, if the IS creates a translocatable unit structure on the prophage, it may be possible to transfer the DNA segment in the unit. Moreover, the phage integrase coding gene in the *oxyBAC* gene cluster can be related to phage-related mobility. Future experimental studies are essential to determine the exact mechanism.

The mobilization mechanisms such as transduction and transformation can be possible for the *oxyBAC* gene cluster. In that case, this gene cluster can soon be passed on many bacteria, causing the effect of disinfectants containing BACs to decrease. Therefore, infection control will be difficult in critical environments associated with human health. Future experimental studies will be crucial to show the mechanisms to help preventing the spread of microorganisms that degrade BACs in critical environments and develop strategies to maintain the disinfection effectiveness of these compounds. Also, the strain BIOMIG1^{BAC} can be used for various environmental applications to detect or eliminate the BACs in the environment.

REFERENCES

Agarwala, R., Barrett, T., Beck, J., Benson, D.A., Bollin, C., Bolton, E., … Zbicz, K. 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46(D1), D8-D13.

Alexander, M. 1981. Biodegradation of chemicals of environmental concern. Science 211(4478), 132-138.

Alkhalifa, S., Jennings, M.C., Granata, D., Klein, M., Wuest, W.M., Minbiole, K.P.C. and Carnevale, V. 2020. Analysis of the Destabilization of Bacterial Membranes by Quaternary Ammonium Compounds: A Combined Experimental and Computational Study. Chembiochem 21(10), 1510-1516.

Altinbag, R.C., Ertekin, E. and Tezel, U. 2020. Complete Genome Sequence of Pseudomonas sp. Strain BIOMIG1BAC, Which Mineralizes Benzalkonium Chloride Disinfectants. Microbiology Resource Announcements 9(20).

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y. and Wishart, D.S. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44(W1), W16-21.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., … Pevzner, P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19(5), 455-477.

Billard-Pomares, T., Fouteau, S., Jacquet, M.E., Roche, D., Barbe, V., Castellanos, M., … and Branger, C. 2014. Characterization of a P1-Like Bacteriophage Carrying an SHV-2 Extended-Spectrum β-Lactamase from an Escherichia coli Strain. Antimicrobial Agents and Chemotherapy 58(11), 6550-6557.

Breidenstein, E.B.M., de la Fuente-Núñez, C. and Hancock, R.E.W. 2011. Pseudomonas aeruginosa: all roads lead to resistance. Trends Microbiol. 19(8), 419-426.

Brown, C.G. and Clarke, J. 2016. Nanopore development at Oxford Nanopore. Nat. Biotechnol. 34(8), 810-811.

Brown-Jaque, M., Calero-Cáceres, W. and Muniesa, M. 2015. Transfer of antibiotic-resistance genes via phage-related mobile elements. Plasmid 79, 1-7.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1).

Chan, P.P. and Lowe, T.M., 2019. Gene Prediction, Humana, New York, NY, 1-14.

Chaumeil, P.A., Mussig, A.J., Hugenholtz, P. and Parks, D.H. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D.R., da Costa, M.S., Rooney, A.P., Yi, H., … Trujillo, M.E. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int. J. Syst. Evol. Microbiol. 68(1), 461-466.

Chun, J. and Rainey, F.A. 2014. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int. J. Syst. Evol. Microbiol. 64(Pt 2), 316-324.

Clement M, Snell Q, Walke P, Posada D, Crandall, K (2002). TCS: estimating gene genealogies. Proc 16th Int Parallel Distrib Process Symp 2:184.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., … de Hoon, M.J. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11), 1422-1423.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Neron, B., … Pourcel, C. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 46(W1), W246-W251.

Davis, J.J., Wattam, A.R., Aziz, R.K., Brettin, T., Butler, R., Butler, R.M., … Stevens, R. 2019. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res.

Dev Kumar, G., Mishra, A., Dunn, L., Townsend, A., Oguadinma, I.C., Bright, K.R. and Gerba, C.P. 2020. Biocides and Novel Antimicrobial Agents for the Mitigation of Coronaviruses. Front. Microbiol. 11.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5), 1792-1797.

Edgar, R.C. 2007. PILER-CR: Fast and accurate identification of CRISPR repeats. BMC Bioinformatics 8(1).

Ertekin, E., 2017. Microbial Ecology and Genetics of Benzalkonium Chloride Biotransformation in the Environment, Ph.D. Thesis, Boğaziçi University

Ertekin, E., Konstantinidis, K.T. and Tezel, U. 2017. A Rieske-Type Oxygenase of Pseudomonas sp. BIOMIG1 Converts Benzalkonium Chlorides to Benzyldimethyl Amine. Environ Sci Technol 51(1), 175-181.

Fillol-Salom, A., Bacarizo, J., Alqasmi, M., Ciges-Tomas, J.R., Martínez-Rubio, R., Roszak, A.W., … and Penadés, J.R. 2019. Hijacking the Hijackers: Escherichia coli Pathogenicity Islands Redirect Helper Phage Packaging for Their Own Benefit. Mol. Cell 75(5), 1020-1030.e1024.

Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. 2005. Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology 3(9), 722-732.

Garrido-Sanz, D., Arrebola, E., Martínez-Granero, F., García-Méndez, S., Muriel, C., Blanco-Romero, E., Martín, M., … Redondo-Nieto, M. 2017. Classification of Isolates from the Pseudomonas fluorescens Complex into Phylogenomic Groups Based in Group-Specific Markers. Front. Microbiol. 8.

Gul, G., 2016. Antibiotic Resistant Pseudomonas sp. BIOMIG1 Protects Susceptible Bacteria From Disinfectants, M.Sc. Thesis, Boğaziçi University

Gomila, M., Busquets, A., Garcia-Valdes, E., Michael, E., Cahan, R., Nitzan, Y. and Lalucat, J. 2015. Draft Genome Sequence of the Toluene-Degrading Pseudomonas stutzeri Strain ST-9. Genome Announc 3(3).

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R. and Hofacker, I.L. 2008. The Vienna RNA Websuite. Nucleic Acids Res. 36(Web Server), W70-W74.

Haft, D.H. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31(1), 371-373.

Haft, D.H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., … and Pruitt, K.D. 2018. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 46(D1), D851-D860.

Hall, J.P.J., Brockhurst, M.A. and Harrison, E. 2017. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. Philosophical Transactions of the Royal Society B: Biological Sciences 372(1735).

Harmer, C.J. and Hall, R.M. 2015. IS26-Mediated Precise Excision of the IS26-aphA1a Translocatable Unit. mBio 6(6), e01866-01815.

Hauser, M., Steinegger, M. and Söding, J. 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics 32(9), 1323-1330.

Hayashi Sant'Anna, F., Bach, E., Porto, R.Z., Guella, F., Hayashi Sant'Anna, E. and Passaglia, L.M.P. 2019. Genomic metrics made easy: what to do and where to go in the new era of bacterial taxonomy. Crit. Rev. Microbiol. 45(2), 182-200.

Hayes, F., Szuplewska, M., Ludwiczak, M., Lyzwa, K., Czarnecki, J. and Bartosik, D. 2014. Mobility and Generation of Mosaic Non-Autonomous Transposons by Tn3-Derived Inverted-Repeat Miniature Elements (TIMEs). PLoS One 9(8).

Hesse, C., Schulz, F., Bull, C.T., Shaffer, B.T., Yan, Q., Shapiro, N., … Loper, J.E. 2018. Genome-based evolutionary history of Pseudomonas spp. Environ. Microbiol. 20(6), 2142-2159.

Hill, C., 2015. Learning Scientific Programming with Python, Cambridge University Press, 333-401.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q. and Vinh, L.S. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35(2), 518-522.

Hora, P.I., Pati, S.G., McNamara, P.J. and Arnold, W.A. 2020. Increased Use of Quaternary Ammonium Compounds during the SARS-CoV-2 Pandemic and Beyond: Consideration of Environmental Implications. Environmental Science & Technology Letters 7(9), 622-631.

Hunter, J.D. 2007. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9(3), 90-95.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T. and Aluru, S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications 9(1).

Kanehisa, M. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28(1), 27-30.

Kanehisa, M. and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1), 27-30.

Kichenaradja, P., Siguier, P., Perochon, J. and Chandler, M. 2010. ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. Nucleic Acids Res. 38(Database issue), D62-68.

Kim, E.H., Nies, D.H., McEvoy, M.M. and Rensing, C. 2011. Switch or Funnel: How RND-Type Transport Systems Control Periplasmic Metal Homeostasis. J. Bacteriol. 193(10), 2381-2387.

Kono, N. and Arakawa, K. 2019. Nanopore sequencing: Review of potential applications in functional genomics. Development, Growth & Differentiation 61(5), 316-326.

Konstantinidis, K.T. and Tiedje, J.M. 2005. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences 102(7), 2567-2572.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19(9), 1639-1645.

Laehnemann, D., Borkhardt, A. and McHardy, A.C. 2016. Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. Brief. Bioinform. 17(1), 154-179.

Lalucat, J., Mulet, M., Gomila, M. and García-Valdés, E. 2020. Genomics in Bacterial Taxonomy: Impact on the Genus Pseudomonas. Genes 11(2).

Lanfear, R., von Haeseler, A., Woodhams, M.D., Schrempf, D., Chernomor, O., Schmidt, H.A., … Teeling, E. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37(5), 1530-1534.

Lang, E., Burghartz, M., Spring, S., Swiderski, J. and Sproer, C. 2010. Pseudomonas benzenivorans sp. nov. and Pseudomonas saponiphila sp. nov., represented by xenobiotics degrading type strains. Curr. Microbiol. 60(2), 85-91.

Leigh, JW, Bryant D (2015). PopART: Full-feature software for haplotype network construction. Methods Ecol Evol 6(9):1110–1116.

Li, H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32(14), 2103-2110.

Lino, C.A., Harper, J.C., Carney, J.P. and Timlin, J.A. 2018. Delivering CRISPR: a review of the challenges and approaches. Drug Deliv 25(1), 1234-1257.

Lipthay, J.R., Barkay, T. and Sørensen, S.r.J. 2001. Enhanced degradation of phenoxyacetic acid in soil by horizontal transfer of the tfdA gene encoding a 2,4-dichlorophenoxyacetic acid dioxygenase. FEMS Microbiol. Ecol. 35(1), 75-84.

Liu, Y., Rao, Q., Blom, J., Lin, Q. and Luo, T. 2020. Pseudomonas piscis sp. nov., isolated from the profound head ulcers of farmed Murray cod (Maccullochella peelii peelii). Int. J. Syst. Evol. Microbiol. 70(4), 2732-2739.

Loman, N.J., Quick, J. and Simpson, J.T. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nature Methods 12(8), 733-735.

Lomsadze, A., Gemayel, K., Tang, S. and Borodovsky, M. 2018. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res. 28(7), 1079-1089.

Lu, H., Giordano, F. and Ning, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 14(5), 265-279.

Madsen, E.L., 2015. Environmental microbiology : from genomes to biogeochemistry, John Wiley and Sons, Inc., Hoboken, New Jersey.

Mallawaarachchi, V., Wickramarachchi, A. and Lin, Y. 2020. GraphBin: refined binning of metagenomic contigs using assembly graphs. Bioinformatics 36(11), 3307-3313.

Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences 101(19), 7287-7292.

McDonnell, G., 2009. Encyclopedia of Microbiology, Academic Press, 529-548.

McKinney, W. 2010 Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference. Vol. 445, 56-61.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P. and Göker, M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14(1).

Mestan, K.K., Ilkhanoff, L., Mouli, S. and Lin, S. 2011. Genomic sequencing in clinical trials. J. Transl. Med. 9(1).

Miller, J.R., Koren, S. and Sutton, G. 2010. Assembly algorithms for next-generation sequencing data. Genomics 95(6), 315-327.

Miller, V. and Balcazar, J.L. 2014. Bacteriophages as Vehicles for Antibiotic Resistance Genes in the Environment. PLoS Pathog. 10(7).

Morrison, K.R., Allen, R.A., Minbiole, K.P.C. and Wuest, W.M. 2019. More QACs, more questions: Recent advances in structure activity relationships and hurdles in understanding resistance mechanisms. Tetrahedron Lett. 60(37).

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., … Finn, R.D. 2015. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 43(D1), D130-D137.

Ogata, H., Goto, S., Fujibuchi, W. and Kanehisa, M. 1998. Computation with the KEGG pathway database. Biosystems 47(1-2), 119-128.

Oh, S., Kurt, Z., Tsementzi, D., Weigand, M.R., Kim, M., Hatt, J.K., … Liu, S.J. 2014. Microbial Community Degradation of Widely Used Quaternary Ammonium Disinfectants. Appl. Environ. Microbiol. 80(19), 5892-5900.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S. and Phillippy, A.M. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17(1).

Oren, A. and Garrity, G.M. 2013. Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. Antonie Van Leeuwenhoek 106(1), 43-56.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., … Stevens, R. 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 42(Database issue), D206-214.

Parte, A.C., Sardà Carbasse, J., Meier-Kolthoff, J.P., Reimer, L.C. and Göker, M. 2020. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. Int. J. Syst. Evol. Microbiol. 70(11), 5607-5612.

Peix, A., Ramirez-Bahena, M.H. and Velazquez, E. 2018. The current status on the taxonomy of Pseudomonas revisited: An update. Infect. Genet. Evol. 57, 106-116.

Perneel, M., Heyrman, J., Adiobo, A., De Maeyer, K., Raaijmakers, J.M., De Vos, P. and Höfte, M. 2007. Characterization of CMR5c and CMR12a, novel fluorescentPseudomonasstrains from the cocoyam rhizosphere with biocontrol activity. J. Appl. Microbiol. 103(4), 1007-1020.

Perrin, A. and Rocha, E.P.C. 2020. PanACoTA: A modular tool for massive microbial comparative genomics, NAR genomics and bioinformatics, 3(1), lqaa106.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., and Toth, I. K. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Analytical Methods, 8(1), 12-24.

Rambaut, A. 2007. FigTree: Graphical viewer of phylogenetic trees. Github Repository https://github.com/rambaut/figtree/ (accessed June 2021).

Ramette, A., Frapolli, M., Fischer-Le Saux, M., Gruffaz, C., Meyer, J.M., Defago, G., Sutra, L. and Moenne-Loccoz, Y. 2011. Pseudomonas protegens sp. nov., widespread plant-protecting bacteria producing the biocontrol compounds 2,4-diacetylphloroglucinol and pyoluteorin. Syst. Appl. Microbiol. 34(3), 180-188.

Ravatn, R., Zehnder, A.J.B. and van der Meer, J.R. 1998. Low-Frequency Horizontal Transfer of an Element Containing the Chlorocatechol Degradation Genes from Pseudomonas sp. Strain B13 to Pseudomonas putida F1 and to Indigenous Bacteria in Laboratory-Scale Activated-Sludge Microcosms. Appl. Environ. Microbiol. 64(6), 2126-2132. Seemann, T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14), 2068-2069.

Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R. and White, O. 2007. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. 35(Database), D260-D264.

Silby, M.W., Winstanley, C., Godfrey, S.A.C., Levy, S.B. and Jackson, R.W. 2011. Pseudomonasgenomes: diverse and adaptable. FEMS Microbiol. Rev. 35(4), 652-680.

Sohn, J.-i. and Nam, J.-W. 2016. The present and future of de novo whole-genome assembly. Brief. Bioinform, 19(1), 23-40.

Solioz, M., 2018. Copper and Bacteria, Cham, Switzerland: Springer International Publishing, 49-80.

Sørensen, S.J., Bailey, M., Hansen, L.H., Kroer, N. and Wuertz, S. 2005. Studying plasmid horizontal transfer in situ: a critical review. Nature Reviews Microbiology 3(9), 700-710.

Tanizawa, Y., Fujisawa, T., Nakamura, Y. and Hancock, J. 2018. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 34(6), 1037-1039.

Tezel, U. and Pavlostathis, S.G. (2011) Antimicrobial Resistance in the Environment, pp. 349-387.

Tezel, U. and Pavlostathis, S.G. 2015. Quaternary ammonium disinfectants: microbial adaptation, degradation and ecology. Curr. Opin. Biotechnol. 33, 296-304.

Tezel, U., Tandukar, M., Martinez, R.J., Sobecky, P.A. and Pavlostathis, S.G. 2012. Aerobic Biotransformation ofn-Tetradecylbenzyldimethylammonium Chloride by an Enriched Pseudomonas spp. Community. Environ. Sci. Technol. 46(16), 8714-8722.

Uritskiy, G.V., DiRuggiero, J. and Taylor, J. 2018. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6(1), 158.

van Ginkel, C.G. and Kolvenbach, M. 1991. Relations between the structure of quaternary alkyl ammonium salts and their biodegradability. Chemosphere 23(3), 281-289.

Vandecraen, J., Chandler, M., Aertsen, A. and Van Houdt, R. 2017. The impact of insertion sequences on bacterial genome plasticity and adaptability. Crit. Rev. Microbiol. 43(6), 709-730.

Vinatzer, B.A., Garrido-Sanz, D., Meier-Kolthoff, J.P., Göker, M., Martín, M., Rivilla, R. and Redondo-Nieto, M. 2016. Genomic and Genetic Diversity within the Pseudomonas fluorescens Complex. PLoS One 11(2).

Wang, X., Kim, Y., Ma, Q., Hong, S.H., Pokusaeva, K., Sturino, J.M. and Wood, T.K. 2010. Cryptic prophages help bacteria cope with adverse environments. Nature Communications 1(1).

Wang, Y., Xiao, M., Geng, X., Liu, J. and Chen, J. 2007. Horizontal transfer of genetic determinants for degradation of phenol between the bacteria living in plant and its rhizosphere. Applied Microbiology and Biotechnology 77(3), 733-739.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S. and others, a. 2017. mwaskom/seaborn: v0.8.1 (September 2017). Zenodo.

Waterman, M.S., 2018. Introduction to Computational Biology, Chapman and Hall/CRC, 20-30.

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput. Biol. 13(6).

Wick, R.R., Judd, L.M. and Holt, K.E. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20.

Wick, R.R., Schultz, M.B., Zobel, J. and Holt, K.E. 2015. Bandage: interactive visualization ofde novogenome assemblies: Fig. 1. Bioinformatics 31(20), 3350-3352.

Xie, Z. and Tang, H. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics 33(21), 3340-3347.

Zerbino, D.R. and Birney, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5), 821-829.

Zhang, C., Cui, F., Zeng, G.-m., Jiang, M., Yang, Z.-z., Yu, Z.-g., Zhu, M.-y. and Shen, L.-q. 2015. Quaternary ammonium compounds (QACs): A review on occurrence, fate and toxicity in the environment. Sci. Total Environ. 518-519, 352-362.

Zheng, G., Filippelli, G.M. and Salamova, A. 2020. Increased Indoor Exposure to Commonly Used Disinfectants during the COVID-19 Pandemic. Environmental Science & Technology Letters 7(10), 760-765.

APPENDIX A: GENOMES USED IN 16S rRNA ANALYSIS

Assembly ID	Organism	Strain	GenBank Assembly Accession	RefSeq Assembly Accession
5497491	Pseudomonas juntendi	BML3	GCA_009932375.1	GCF_009932375.1
1200311	Pseudomonas abyssi	MT5	GCA_002307495.1	GCF_002307495.1
210741	Pseudomonas soli	SJ10	GCA_000498975.2	GCF_000498975.2
867521	Pseudomonas costantinii	BS2773	GCA_900105935.1	GCF_900105935.1
7059561	Pseudomonas multiresinivorans	populi	GCA_012971725.1	GCF_012971725.1
880551	Pseudomonas yangmingensis	DSM 24213	GCA_900114825.1	GCF_900114825.1
867151	Pseudomonas sabulinigri	JCM 14963	GCA_900105255.1	GCF_900105255.1
304091	Pseudomonas saudimassiliensis	12M76_air	GCA_000939975.1	GCF_000939975.1
357901	Pseudomonas weihenstephanensis	DSM 29166	GCA_001043055.1	GCF_001043055.1
867341	Pseudomonas mucidolens	LMG 2223	GCA_900106045.1	GCF_900106045.1
5497471	Pseudomonas asiatica	RYU5	GCA_009932335.1	GCF_009932335.1
869261	Pseudomonas deceptionensis	LMG 25555	GCA_900106095.1	GCF_900106095.1
60391	Pseudomonas vranovensis	DSM 16006	GCA_000425805.1	GCF_000425805.1
1460601	Pseudomonas baetica	LMG 25716	GCA_002813455.1	GCF_002813455.1
486591	Pseudomonas versuta	L10.10	GCA_001294575.1	GCF_001294575.1
882941	Pseudomonas argentinensis	LMG 22563	GCA_900113905.1	GCF_900113905.1
79911	Pseudomonas chloritidismutans	AW-1	GCA_000495915.1	GCF_000495915.1
534608	Pseudomonas poae	RE*1-1-14	GCA_000336465.1	GCF_000336465.1
7319111	Pseudomonas reactans	C5002	GCA_013385825.1	GCF_013385825.1
67571	Pseudomonas aestus	CMAA1215	GCA_000474765.1	GCF_000474765.1
1135121	Pseudomonas aestusnigri	VGXO14	GCA_002197985.1	GCF_002197985.1
2039781	Pseudomonas ficuserectae	ICMP 7849	GCA_003701615.1	GCF_003701615.1
45068	Pseudomonas entomophila	L48	GCA_000026105.1	GCF_000026105.1
863961	Pseudomonas extremaustralis	DSM 17835	GCA_900102035.1	GCF_900102035.1
862961	Pseudomonas arsenicoxydans	CECT 7543	GCA_900103875.1	GCF_900103875.1
1980081	Pseudomonas reidholzensis	CCOS 865	GCA_900536025.1	GCF_900536025.1
865521	Pseudomonas cannabina	ICMP 2823	GCA_900100365.1	GCF_900100365.1
866871	Pseudomonas pohangensis	DSM 17875	GCA_900105995.1	GCF_900105995.1
2513621	Pseudomonas kairouanensis	KC12	GCA_004682055.1	GCF_004682055.1
866631	Pseudomonas prosekii	LMG 26867	GCA_900105155.1	GCF_900105155.1
742261	Pseudomonas antarctica	PAMC 27494	GCA_001647715.1	GCF_001647715.1
5487321	Pseudomonas knackmussii	N1-2	GCA_009911755.1	GCF_009911755.1
553161	Pseudomonas meliae	ICMP6289	GCA_001400515.1	GCF_001400515.1
784271	Pseudomonas corrugata	RM1-1-4	GCA_001708425.1	GCF_001708425.1
1226141	Pseudomonas furukawaii	KF707	GCA_002355475.1	GCF_002355475.1
867041	Pseudomonas asplenii	ATCC 23835	GCA_900105475.1	GCF_900105475.1
357891	Pseudomonas helleri	DSM 28141	GCA_001043065.1	GCF_001043065.1
7319851	Pseudomonas salomonii	IPO3765	GCA_013386885.1	GCF_013386885.1
2125381	Pseudomonas synxantha	Feb-79	GCA_003851495.1	GCF_003851495.1

7319321	Pseudomonas gingeri	A6001	GCA_013386115.1	GCF_013386115.1
2639971	Pseudomonas leptonychotis	CCM 8849	GCA_004920405.1	GCF_004920405.1
1154581	Pseudomonas delhiensis	RLD-1	GCA_900187975.1	GCF_900187975.1
1683501	Pseudomonas plecoglossicida	KCJK7865	GCA_003062165.1	GCF_003062165.1
7343201	Pseudomonas kunmingensis	ZKA55	GCA_013409135.1	GCF_013409135.1
82401	Pseudomonas tremae	CC1513	GCA_000452765.2	GCF_000452765.1
867001	Pseudomonas mediterranea	DSM 16733	GCA_900106005.1	GCF_900106005.1
362071	Pseudomonas fildesensis	KG01	GCA_001050345.1	GCF_001050345.1
867471	Pseudomonas kilonensis	BS3780	GCA_900105635.1	GCF_900105635.1
864801	Pseudomonas koreensis	BS3658	GCA_900101415.1	GCF_900101415.1
7342391	Pseudomonas composti	ZKA28	GCA_013407905.1	GCF_013407905.1
879261	Pseudomonas bauzanensis	DSM 22558	GCA_900111225.1	GCF_900111225.1
864281	Pseudomonas azotoformans	LMG 21611	GCA_900103345.1	GCF_900103345.1
864181	Pseudomonas jinjuensis	JCM 21621	GCA_900103845.1	GCF_900103845.1
2224651	Pseudomonas hydrolytica	DSWY01	GCA_004123735.1	GCF_004123735.1
1180241	Pseudomonas fragi	F1786	GCA_002269585.1	GCF_002269585.1
568761	Pseudomonas taeanensis	MS-3	GCA_000498575.2	GCF_000498575.2
7068181	Pseudomonas lactis	WS 5000	GCA_012986545.1	GCF_012986545.1
1678021	Pseudomonas yamanorum	LBUM636	GCA_001612705.2	GCF_001612705.2
7331811	Pseudomonas hunanensis	xwS6	GCA_013393325.1	GCF_013393325.1
2335641	Pseudomonas dryadis	P27B	GCA_004327225.1	GCF_004327225.1
2037171	Pseudomonas songnenensis	NEAU-ST5-5	GCA_003696315.1	GCF_003696315.1
1876751	Pseudomonas parafulva	JBCS1880	GCA_003410295.1	GCF_003410295.1
869451	Pseudomonas xanthomarina	LMG 23572	GCA_900108535.1	GCF_900108535.1
866851	Pseudomonas litoralis	2SM5	GCA_900105005.1	GCF_900105005.1
1987781	Pseudomonas jilinensis	JS15-10A1	GCA_003586265.1	GCF_003586265.1
868331	Pseudomonas palleroniana	BS3265	GCA_900105975.1	GCF_900105975.1
862591	Pseudomonas panipatensis	CCM 7469	GCA_900099785.1	GCF_900099785.1
7067661	Pseudomonas proteolytica	WS 5126	GCA_012986025.1	GCF_012986025.1
5131551	Pseudomonas pelagia	Kongs-67	GCA_009497895.1	GCF_009497895.1
1545861	Pseudomonas oceani	DSM 100277	GCA_002903165.1	GCF_002903165.1
4560411	Pseudomonas brassicacearum	3Re2-7	GCA_008370715.1	GCF_008370715.1
910991	Pseudomonas punonensis	CECT 8089	GCA_900142655.1	GCF_900142655.1
1190381	Pseudomonas moraviensis	TYU6	GCA_002287825.1	GCF_002287825.1
84021	Pseudomonas canadensis	Feb-92	GCA_000503215.1	GCF_000503215.1
322041	Pseudomonas syringae pv. coryli	NCPPB 4273	GCA_000972175.1	GCF_000972175.1
2037191	Pseudomonas zhaodongensis	NEAU-ST5-21	GCA_003696365.1	GCF_003696365.1
1421181	Pseudomonas mosselii	PtA1	GCA_002736065.1	GCF_002736065.1
865761	Pseudomonas extremorientalis	BS2774	GCA_900104365.1	GCF_900104365.1
876751	Pseudomonas cuatrocienegasensis	CIP 109853	GCA_900110925.1	GCF_900110925.1
244011	Pseudomonas batumici	UCM B-321	GCA_000820515.1	GCF_000820515.1
40661	Pseudomonas caeni	DSM 24390	GCA_000421765.1	GCF_000421765.1
199151	Pseudomonas capeferrum	WCS358	GCA_000731675.1	GCF_000731675.1
1061021	Pseudomonas savastanoi	NCPPB 3335	GCA_000164015.3	GCF_000164015.3
2039371	Pseudomonas syringae pv. tagetis	ICMP 4092	GCA_003700835.1	GCF_003700835.1

803341	Pseudomonas humi	CCA1	GCA_001748265.1	GCF_001748265.1
1612001	Pseudomonas lurida	MYb11	GCA_002966835.1	GCF_002966835.1
867011	Pseudomonas granadensis	LMG 27940	GCA_900105485.1	GCF_900105485.1
884121	Pseudomonas straminea	JCM 2783	GCA_900112645.1	GCF_900112645.1
868081	Pseudomonas mohnii	DSM 18327	GCA_900105115.1	GCF_900105115.1
3522301	Pseudomonas psychrotolerans	CS51	GCA_006384975.1	GCF_006384975.1
867071	Pseudomonas xinjiangensis	NRRL B-51270	GCA_900104945.1	GCF_900104945.1
2335661	Pseudomonas daroniae	P18A	GCA_004327275.1	GCF_004327275.1
3171551	Pseudomonas nitroreducens	DSM 9128	GCA_005796065.1	GCF_005796065.1
884281	Pseudomonas borbori	DSM 17834	GCA_900115555.1	GCF_900115555.1
865251	Pseudomonas reinekei	BS3776	GCA_900104125.1	GCF_900104125.1
4692161	Pseudomonas profundi	M5	GCA_008638305.1	GCF_008638305.1
880441	Pseudomonas guineae	LMG 24016	GCA_900113745.1	GCF_900113745.1
2804701	Pseudomonas rhizoryzae	ZYY160	GCA_005250605.1	GCF_005250605.1
5849741	Pseudomonas laurentiana	JCM 32154	GCA_010671685.1	GCF_010671685.1
1154421	Pseudomonas segetis	CIP 108523	GCA_900188155.1	GCF_900188155.1
47571	Pseudomonas avellanae	BPIC 631	GCA_000444135.1	GCF_000444135.1
430471	Pseudomonas trivialis	IHBB745	GCA_001186335.1	GCF_001186335.1
867141	Pseudomonas guangdongensis	CCTCC 2012022	GCA_900105885.1	GCF_900105885.1
2180941	Pseudomonas stutzeri	NCTC10475	GCA_900638035.1	GCF_900638035.1
6279061	Pseudomonas otitidis	MrB4	GCA_011397855.1	GCF_011397855.1
87711	Pseudomonas oleovorans	MOIL14HWK12	GCA_000510765.1	GCF_000510765.1
7241581	Pseudomonas rhodesiae	NL2019	GCA_013285305.1	GCF_013285305.1
659521	Pseudomonas citronellolis	P3B5	GCA_001586155.1	GCF_001586155.1
5026041	Pseudomonas kitaguniensis	MAFF 212408	GCA_009296165.1	GCF_009296165.1
864621	Pseudomonas thivervalensis	BS3779	GCA_900102295.1	GCF_900102295.1
7067731	Pseudomonas peli	DSM 17833	GCA_012986145.1	GCF_012986145.1
219021	Pseudomonas saudiphocaensis	20_BN	GCA_000756775.2	GCF_000756775.1
1564761	Pseudomonas bohemica	IA19	GCA_002934685.1	GCF_002934685.1
868781	Pseudomonas anguilliseptica	DSM 12111	GCA_900105355.1	GCF_900105355.1
865621	Pseudomonas seleniipraecipitans	LMG 25475	GCA_900102335.1	GCF_900102335.1
6095431	Pseudomonas sp.	BIOMIG1BAC	GCA_001705995.2	GCF_001705995.2
7415041	Pseudomonas carnis	96A1	GCA_013522805.1	GCF_013522805.1
862791	Pseudomonas grimontii	BS2976	GCA_900101085.1	GCF_900101085.1
581321	Pseudomonas paralactis	DSM 29164	GCA_001439735.1	GCF_001439735.1
867301	Pseudomonas cedrina	BS2981	GCA_900104915.1	GCF_900104915.1
2124881	Pseudomonas chlororaphis subsp. piscium	ChPhzS135	GCA_003850485.1	GCF_003850485.1
1460581	Pseudomonas tolaasii	NCPPB 2192	GCA_002813445.1	GCF_002813445.1
884181	Pseudomonas sagittaria	JCM 18195	GCA_900115715.1	GCF_900115715.1
866571	Pseudomonas oryzae	KCTC 32247	GCA_900104805.1	GCF_900104805.1
868021	Pseudomonas saponiphila	DSM 9751	GCA_900105185.1	GCF_900105185.1
869661	Pseudomonas fuscovaginae	LMG 2158	GCA_900108595.1	GCF_900108595.1
1051841	Pseudomonas amygdali pv. lachrymans	NM002	GCA_002068135.1	GCF_002068135.1
242631	Pseudomonas balearica	DSM6083 (=SP1402)	GCA_000818015.1	GCF_000818015.1
301411	Pseudomonas simiae	PCL1751	GCA_000934565.1	GCF_000934565.1

862841	Pseudomonas gessardii	BS2982	GCA_900101185.1	GCF_900101185.1
558971	Pseudomonas endophytica	BSTT44	GCA_001411475.1	GCF_001411475.1
709011	Pseudomonas alcaligenes	NEB 585	GCA_001597285.1	GCF_001597285.1
7342741	Pseudomonas flavescens	BIGb0408	GCA_013408425.1	GCF_013408425.1
284948	Pseudomonas fulva	12-X	GCA_000213805.1	GCF_000213805.1
7113331	Pseudomonas nitrititolerans	AGROB37	GCA_013072755.1	GCF_013072755.1
63091	Pseudomonas chengduensis	EGD-AQ5	GCA_000465575.1	GCF_000465575.1
581291	Pseudomonas libanensis	DSM 17149	GCA_001439685.1	GCF_001439685.1
319611	Pseudomonas marginalis	H21	GCA_000967935.1	GCF_000967935.1
6086451	Pseudomonas psychrophila	KM02	GCA_011040435.1	GCF_011040435.1
4268931	Pseudomonas saxonica	DSM 108989	GCA_007858365.1	GCF_007858365.1
2323491	Pseudomonas bubulae	parafragi	GCA_900618535.1	GCF_900618535.1
633061	Pseudomonas agarici	NCPPB 2472	GCA_001543125.1	GCF_001543125.1
707708	Pseudomonas resinovorans	NBRC 106553	GCA_000412695.1	GCF_000412695.1
7319211	Pseudomonas edaphica	B7002	GCA_013385965.1	GCF_013385965.1
868441	Pseudomonas migulae	BS3662	GCA_900106025.1	GCF_900106025.1
2358871	Pseudomonas syringae pv. tomato	BRIP38746	GCA_004376045.1	GCF_004376045.1
865001	Pseudomonas abietaniphila	ATCC 700689	GCA_900100795.1	GCF_900100795.1
2513611	Pseudomonas nabeulensis	E10B	GCA_004682045.1	GCF_004682045.1
867121	Pseudomonas salegens	CECT 8338	GCA_900105655.1	GCF_900105655.1
867231	Pseudomonas vancouverensis	BS3656	GCA_900105825.1	GCF_900105825.1
863571	Pseudomonas congelans	DSM 14939	GCA_900103225.1	GCF_900103225.1
1839661	Pseudomonas kribbensis	46-2	GCA_003352185.1	GCF_003352185.1
2352741	Pseudomonas brenneri	BIGb0273	GCA_004363635.1	GCF_004363635.1
1912741	Pseudomonas gallaeciensis	V113	GCA_003444685.1	GCF_003444685.1
865201	Pseudomonas guariconensis	LMG 27394	GCA_900102675.1	GCF_900102675.1
1705051	Pseudomonas ovata	F51	GCA_003131185.1	GCF_003131185.1
863541	Pseudomonas moorei	BS3775	GCA_900102045.1	GCF_900102045.1
2355711	Pseudomonas inefficax		GCA_900277125.1	GCF_900277125.1
2353201	Pseudomonas mandelii	YF10-2(1) 51	GCA_004364415.1	GCF_004364415.1
7331821	Pseudomonas taiwanensis	xwS3	GCA_013393345.1	GCF_013393345.1
5732731	Pseudomonas fluorescens	DR397	GCA_010448615.1	GCF_010448615.1
213461	Pseudomonas cremoricolorata	ND07	GCA_000759535.1	GCF_000759535.1
1113821	Pseudomonas caspiana	FBF102	GCA_002158995.1	GCF_002158995.1
2353831	Pseudomonas helmanticensis	BIGb0525	GCA_004365765.1	GCF_004365765.1
5477191	Pseudomonas monteilii	ODNR6CL	GCA_009905655.1	GCF_009905655.1
1755521	Pseudomonas sichuanensis	WCHPs060039	GCA_003231305.1	GCF_003231305.1
880131	Pseudomonas linyingensis	LMG 25967	GCA_900109175.1	GCF_900109175.1
4167471	Pseudomonas oryzihabitans	DE0585	GCA_007665635.1	GCF_007665635.1
707668	Pseudomonas putida	NBRC 14164	GCA_000412675.1	GCF_000412675.1
7195081	Pseudomonas graminis	PgKB30	GCA_013201545.1	GCF_013201545.1
882871	Pseudomonas pachastrellae	JCM 12285	GCA_900114765.1	GCF_900114765.1
4729921	Pseudomonas marincola	YSy11	GCA_900682675.1	GCF_900682675.1
883381	Pseudomonas toyotomiensis	JCM 15604	GCA_900115695.1	GCF_900115695.1
4692541	Pseudomonas salina	XCD-X85	GCA_008641105.1	GCF_008641105.1

882131	Pseudomonas formosensis	JCM 18415	GCA_900115905.1	GCF_900115905.1
867361	Pseudomonas taetrolens	BS3652	GCA_900104825.1	GCF_900104825.1
2381931	Pseudomonas protegens	CHA0	GCA_900560965.1	GCF_900560965.1
6050361	Pseudomonas nitritireducens	WZBFD3-5A2	GCA_010994165.1	GCF_010994165.1
883901	Pseudomonas syringae	BS3827	GCA_900113625.1	GCF_900113625.1
1024641	Pseudomonas veronii	R02	GCA_002028325.1	GCF_002028325.1
867831	Pseudomonas jessenii	BS3660	GCA_900104905.1	GCF_900104905.1
5322091	Pseudomonas haemolytica	DSM 108988	GCA_009659615.1	GCF_009659615.1
1564441	Pseudomonas orientalis	F9	GCA_002934065.1	GCF_002934065.1
553271	Pseudomonas caricapapayae	ICMP2855	GCA_001400735.1	GCF_001400735.1
766821	Pseudomonas cerasi		GCA_900074915.1	GCF_900074915.1
213411	Pseudomonas lutea	DSM 17257	GCA_000759445.1	GCF_000759445.1
79781	Escherichia coli	K-12 MG1655	GCA_000005845.2	GCF_000005845.2
2124911	Pseudomonas sp.	CMR5c	GCA_003850545.1	GCF_003850545.1
2124921	Pseudomonas sp.	CMR12a	GCA_003850565.1	GCF_003850565.1

APPENDIX B: *P. CHLORORAPHIS* AND *P. PROTEGENS* SUBGROUP GENOMES AND THEIR TAXONOMIC COMPARISON WITH GTDB AND NCBI TAXONOMY

GID	Name	Subgroup	The taxonomy in this research	GTDB Species Taxonomy	NCBI Taxonomy*	Size
879581	P. sp. NFACC41-3	P. chlororaphis	P. chlororaphis			6492581
1081901	P. sp. NFIX51	P. chlororaphis	P. chlororaphis			6495783
1838551	P. chlororaphis HAMBI_1977	P. chlororaphis	P. chlororaphis			6621816
613588	P. chlororaphis subsp. chlororaphis GP72	P. chlororaphis	P. chlororaphis			6629881
2077721	P. chlororaphis 48B8	P. chlororaphis	P. chlororaphis			6652158
880931	P. sp. NFPP07	P. chlororaphis	P. chlororaphis			6652270
501768	P. chlororaphis subsp. aureofaciens 30-84	P. chlororaphis	P. chlororaphis			6666321
2125211	P. chlororaphis ATCC 17415	P. chlororaphis	P. chlororaphis			6670884
214361	P. chlororaphis subsp. aurantiaca JD37	P. chlororaphis	P. chlororaphis			6702062
2125221	P. chlororaphis subsp. aurantiaca M12	P. chlororaphis	P. chlororaphis			6730980
1541051	P. sp. GW531-T4	P. chlororaphis	P. chlororaphis			6731149
2125331	P. chlororaphis subsp. aureofaciens C50	P. chlororaphis	P. chlororaphis			6741275
2125351	P. chlororaphis subsp. aureofaciens ChPhzS23	P. chlororaphis	P. chlororaphis			6756093
4688861	P. chlororaphis JV395B	P. chlororaphis	P. chlororaphis			6771006
7068701	P. chlororaphis WS 5014	P. chlororaphis	P. chlororaphis			6787621
2125341	P. chlororaphis subsp. aureofaciens 66	P. chlororaphis	P. chlororaphis			6797278
88901	P. chlororaphis YL-1	P. chlororaphis	P. chlororaphis			6800980
473791	P. chlororaphis subsp. aureofaciens CD	P. chlororaphis	P. chlororaphis			6802143
2125271	P. chlororaphis subsp. aurantiaca M71	P. chlororaphis	P. chlororaphis			6807002
711711	P. chlororaphis	P. chlororaphis	P. chlororaphis			6837781
2125291	P. chlororaphis subsp. aurantiaca PCM 2210	P. chlororaphis	P. chlororaphis			6854483
5554551	P. chlororaphis subsp. aurantiaca zm-1	P. chlororaphis	P. chlororaphis			6861331
2125611	P. chlororaphis subsp. aureofaciens ChPhzTR18	P. chlororaphis	P. chlororaphis			6873200
4688851	P. chlororaphis JV497	P. chlororaphis	P. chlororaphis			6875649
1492231	P. sp. 09C 129	P. chlororaphis	P. chlororaphis			6897667
2125251	P. chlororaphis subsp. aurantiaca CW2	P. chlororaphis	P. chlororaphis			6925198
2125571	P. chlororaphis subsp. aureofaciens ChPhzTR36	P. chlororaphis	P. chlororaphis			6934077
2125641	P. chlororaphis subsp. aureofaciens ChPhzTR38	P. chlororaphis	P. chlororaphis			6947716
2125241	P. chlororaphis subsp. aurantiaca 449	P. chlororaphis	P. chlororaphis			6962068
2125541	P. chlororaphis subsp. aurantiaca 464	P. chlororaphis	P. chlororaphis			6964452
240261	P. chlororaphis subsp. aureofaciens NBRC 3521	P. chlororaphis	P. chlororaphis			6973338
392928	P. chlororaphis O6	P. chlororaphis	P. chlororaphis			6980251
299741	P. sp. MRSN12121	P. chlororaphis	P. chlororaphis			6986763
1226341	P. chlororaphis subsp. aurantiaca StFRB508	P. chlororaphis	P. chlororaphis			6997933

2125631	P. chlororaphis B25	P. chlororaphis	P. chlororaphis			7016593
2125601	P. chlororaphis subsp. aureofaciens ChPhzTR39	P. chlororaphis	P. chlororaphis			7046099
182551	P. chlororaphis PA23	P. chlororaphis	P. chlororaphis			7122173
1733261	P. sp. RW409	P. chlororaphis	P. chlororaphis			7154805
5435471	P. chlororaphis R47	P. chlororaphis	P. chlororaphis			7197297
2180071	P. chlororaphis NCTC7357	P. chlororaphis	P. chlororaphis			7209930
2124831	P. chlororaphis subsp. piscium ATCC 17411	P. chlororaphis	P. chlororaphis			7212397
2124821	P. chlororaphis subsp. piscium ATCC 17809	P. chlororaphis	P. chlororaphis			7218893
7909361	P. sp. MSSRFD41	P. protegens	P. Piscis	-	**sp.**	6924660
5061511	P. sp. MC042	P. protegens	P. Piscis	-	sp.	6927622
473781	P. sp. CMR5c	P. protegens	P. Piscis	s__Pseudomonas_E sp001269545	sp.	6755578
1541031	P. sp. FW507-12TSA	P. protegens	P. Piscis	s__Pseudomonas_E sp001269545	sp.	6895883
1083071	P. sp. B6(2017)	P. protegens	P. Piscis	s__Pseudomonas_E sp001269545	sp.	7086907
1082941	P. sp. R26(2017)	P. protegens	P. Piscis	s__Pseudomonas_E sp001269545	sp.	7087578
2124921	P. sp. CMR12a	P. protegens	P. alexanderii sp nov.	s__Pseudomonas_E sp001705835	sp.	6896611
781861	P. sp. BIOMIG1^{N}	P. protegens	P. alexanderii sp nov.	s__Pseudomonas_E sp001705835	sp.	7463297
781851	P. sp. BIOMIG1^{BDMA}	P. protegens	P. alexanderii sp nov.	s__Pseudomonas_E sp001705835	sp.	7463327
781871	P. sp. BIOMIG1^{BD}	P. protegens	P. alexanderii sp nov.	s__Pseudomonas_E sp001705835	sp.	7477097
6095431	P. sp. BIOMIG1^{BAC}	P. protegens	P. alexanderii sp nov.	s__Pseudomonas_E sp001705835	sp.	7675262
5454921	P. sp. LD120	P. protegens	-	-	sp.	6672566
4254931	P. protegens 3295	P. protegens	P. protegens	-	protegens	6907494
7067711	P. protegens WS 5082	P. protegens	P. protegens	-	protegens	6926615
7067681	P. protegens WS 5415	P. protegens	P. protegens	-	protegens	6930863
7277611	P. protegens UMG3145	P. protegens	P. protegens	-	protegens	6964104
5470871	P. protegens APC 3760	P. protegens	P. protegens	-	protegens	6995890
1760221	P. protegens	P. protegens	P. protegens	-	protegens	7037291
3217261	P. protegens PF-1	P. protegens	P. protegens	-	protegens	7052842
5385521	P. protegens SN15-2	P. protegens	P. protegens	-	protegens	7075587
4779711	P. protegens pf5-k2	P. protegens	P. protegens	-	protegens	7080416
4779721	P. protegens pf5-k3	P. protegens	P. protegens	-	protegens	7085242
7868941	P. sp. UME65	P. protegens	P. protegens	-	sp.	7120312
7869111	P. sp. UMC65	P. protegens	P. protegens	-	sp.	7193633
346601	P. protegens AU13852	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6678920
1534431	P. protegens AS15	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6756833
248021	P. protegens Cab57	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6827892
863591	P. sp. NFPP15	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6831633
2023001	P. protegens CHA0-GFP	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6834659
882011	P. sp. NFPP13	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6841688
895601	P. sp. NFPP16	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6842543
877751	P. sp. NFPP18	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6843250
895791	P. sp. NFPP14	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6843884
884831	P. sp. NFPP25	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6844013
862251	P. sp. NFPP17	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6844519
473741	P. protegens PGNR1	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6854017

687338	P. protegens CHA0	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6867980
877061	P. sp. NFPP19	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6883168
473761	P. protegens BRIP	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6886626
2077811	P. protegens Darke	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6896989
1633301	P. sp. NFPP22	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6900659
2077801	P. protegens Clinton	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6910241
2087661	P. protegens BIGb0404	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6926262
2077881	P. protegens Wayne	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6929928
2088021	P. protegens JUb28	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6930756
1541091	P. protegens GW456-12-1-14-LB1	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6945648
2077791	P. protegens 1C5	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6948607
879391	P. sp. NFPP10	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6957254
883961	P. sp. NFPP08	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6959206
473731	P. protegens K94.41	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6959659
895681	P. sp. NFPP09	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6960462
921051	P. protegens PA11	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6973524
767981	P. protegens PA26	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6980545
921021	P. protegens PA19	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6980620
921041	P. protegens PA14	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6981118
920221	P. protegens PA25	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6981405
768001	P. protegens PA12	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	6984535
865331	P. sp. NFPP12	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6995158
884321	P. sp. NFPP05	P. protegens	P. protegens	s__Pseudomonas_E protegens	sp.	6997636
2077781	P. protegens 1B1	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	7006377
1012841	P. protegens H78	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	7032394
473751	P. protegens PF	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	7068782
33368	P. protegens Pf-5	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	7074893
1620811	P. protegens FDAARGOS_307	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	7075815
4779691	P. protegens pf5	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	7087572
347011	P. protegens AU20219	P. protegens	P. protegens	s__Pseudomonas_E protegens	protegens	7275643
7067541	P. sp. WS 5414	P. protegens	P. saponiphila	-	sp.	6670970
7885221	P. protegens H1F10B	P. protegens	P. saponiphila	-	protegens	6811127
7946611	P. protegens H1F10A	P. protegens	P. saponiphila	-	protegens	6817972
7342441	P. protegens H1F10C	P. protegens	P. saponiphila	-	protegens	6817974
7343661	P. sp. H1F10D	P. protegens	P. saponiphila	-	sp.	6818519
7919481	P. protegens H1F5C	P. protegens	P. saponiphila	-	protegens	6818519
5478481	P. sp. Fl5BN2	P. protegens	P. saponiphila	-	sp.	7204762
346541	P. fluorescens AU11706	P. protegens	P. saponiphila	-	fluorescens	7256896
5478531	P. sp. Fl4BN1	P. protegens	P. saponiphila	-	sp.	7330768
1855741	P. protegens FD6	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	6667995
1741471	P. protegens MB-090714	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	6678986
2077741	P. protegens 12H11	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	6715230

2077761	P. protegens 15H3	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	6716467
2077771	P. protegens 38G2	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	6719051
1741561	P. protegens MB-090624	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	6775121
1528751	P. protegens 4	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	6832152
1528761	P. protegens 11	P. protegens	P. saponiphila	s__Pseudomonas_E protegens_A	protegens	7053517
1751771	P. sp. TKO26	P. protegens	P. saponiphila	s__Pseudomonas_E saponiphila	sp.	6865389
1751791	P. sp. TKO30	P. protegens	P. saponiphila	s__Pseudomonas_E saponiphila	sp.	6865389
1751841	P. sp. TKO29	P. protegens	P. saponiphila	s__Pseudomonas_E saponiphila	sp.	6865389
1751801	P. sp. TKO14	P. protegens	P. saponiphila	s__Pseudomonas_E saponiphila	sp.	6865808
868021	P. saponiphila DSM 9751	P. protegens	P. saponiphila	s__Pseudomonas_E saponiphila	saponiphila	7375852
149191	P. sp. PH1b	P. protegens	P. saponiphila	s__Pseudomonas_E sp000633395	sp.	7430140
2225491	P. protegens XY2F4	P. protegens	P. saponiphila	s__Pseudomonas_E sp0017547895	protegens	6811381
635431	P. sp. St29	P. protegens	P. saponiphila	s__Pseudomonas_E sp0017547895	sp.	6833117
635421	P. sp. Os17	P. protegens	P. saponiphila	s__Pseudomonas_E sp0017547895	sp.	6885464
1681341	P. protegens BNJ-SS-45	P. protegens	P. saponiphila	s__Pseudomonas_E sp0017547895	protegens	7116415

* Colors based on taxonomic accuracy: Green is true, red is wrong and grey is unclassified.