Pan-genome Analysis of *Alistipes* Genomes: Exploration of Diversity and Pathogenicity

By

Joy Deb Nath Tonmoy
17126033

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of
BS in Microbiology

Department of Mathematics and Natural Sciences
BRAC University
January 2023

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Joy Deb Nath Tonmoy

Student ID: 17126033

Approval

The thesis titled "Pan-genome Analysis of *Alistipes* Genomes: Exploration of Diversity and Pathogenicity" submitted by Joy Deb Nath Tonmoy (ID: 17126033) has been accepted as satisfactory in partial fulfillment of the requirement for the degree of BS in Microbiology on January 26, 2023.

Examining Committee:

Supervisor:

_____

Rafeed Rahman Turjya
Lecturer, Department of Mathematics and Natural Sciences
BRAC University

Program Coordinator:

_____

Dr. Nadia Sultana Deen
Associate Professor, Department of Mathematics and Natural Sciences
BRAC University

Departmental Head:

_____

Chairperson, Department of Mathematics and Natural Sciences
BRAC University

Acknowledgement

I am grateful to mother nature and my creator that I am still alive after the deadly corona virus pandemic and their blessing throughout my life. I am extremely grateful to my parents for their consistent unconditional love and support during this hectic journey.

Firstly, I would like to thank **Mr. Abdullah Al Kamran Khan**, Former Lecturer, Department of Mathematics and Natural Sciences, BRAC University, for encouraging me to do an undergrad thesis project on bioinformatics. If he did not support my enthusiasm at the first place, this project would not be even initiated. In this regard, I am thankful to **Dr. Iftekhar Bin Naser,** Associate Professor, for the motivation and **Dr. Mahbubul Hasan Siddiqee**, Associate Professor and former Program Coordinator of Microbiology program, for allowing me to do this project. Also, I am obliged to **Professor Dr. Mahboob Hossain** and the Chairperson of the Department of Mathematics and Natural Sciences, **Professor Dr. A. F. M. Yusuf Haider**, for making this journey smoother by providing effective administrative support when necessary.

Most importantly, I am indebted to my supervisor, **Mr. Rafeed Rahman Turjya**, Lecturer, Department of Mathematics and Natural Sciences, BRAC University. I am out of words expressing my gratitude for him. He has planned and designed this project from scratch and provided me step by step instructions, guided me to learn and use the relevant bioinformatic tools, and helped me to improve myself with constructive feedbacks. Without his consistent support, supervision, devotion to this project and patience, the completion of this project was impossible.

Lastly, I would like to appreciate my elder sister, Joya and her husband; younger sister, Joyita for their consistent love, encouragement and mental support. Moreover, I would like to thank my parents again for never giving up on me. I conclude by conveying my gratitude to my friends and everyone who was directly or indirectly involved during this journey at BRAC University and left a beautiful memory.

Abstract

Alistipes, a recently identified and comparatively unexplored bacterial genus of the *Bacteroidetes* phylum, has several species isolated from both healthy and diseased individuals. This gram-negative, anaerobic, rod-shaped genus of bacteria has multiple complete genomes of different species available; yet the pathogenicity and other clinically significant genomic characteristics have not been characterized. In this study, 16 genomes of different *Alistipes* species were used for pan-genome analysis to understand the diversity of the genes across the genus. Furthermore, the same genomes were used to identify the genes related to antimicrobial resistance, virulence, and other clinically significant genomic characteristics. In this comparative genomic analysis, the conserved characteristics as well as the variability of different species of *Alistipes* have been elucidated. Thirty-six antibiotic resistance genes that provide different strains and/or isolates of *Alistipes* resistance to antibiotics e.g., rifampicin, fluoroquinolone, tetracycline, etc. were identified. Additionally, the identification of bacterial type II secretion system and type II toxin-antitoxin systems as well as other virulence genes in the *Alistipes* genomes provides new insights on the impact of the bacteria on human health. Ultimately, the identification of pathogenicity-associated factors may lead to accurate therapeutic interventions when dealing with different *Alistipes* species.

Keywords: pan-genome analysis; *Alistipes;* antimicrobial resistance; virulence gene; toxin-antitoxin system; secretion system of *Alistipes* species.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 The genus *Alistipes*

The genus *Alistipes* belonging to the phylum *Bacteroidetes* is comparatively newer than other bacterial genus like *Escherichia, Pseudomonas*, et cetera. [1]. However, the significance and clinical importance of *Alistipes* species is increasing day by day. *Alistipes* falls in the group of anaerobic bacteria that are mostly found in the gastrointestinal tract (GI tract) of healthy human individuals [2]. As of July 2022, 16 species of the genus *Alistipes* have been recorded in the taxonomy database at The National Center for Biotechnology Information (NCBI, txid239759) and the species are: *Alistipes communis, Alistipes dispar, Alistipes finegoldii, Alistipes ihumii, Alistipes indistinctus, Alistipes inops, Alistipes massiliensis, Alistipes megaguti, Alistipes montrealensis, Alistipes okayasuensis, Alisitpes onderdonkii, Alistipes provencensis, Alistipes putredinis, Alistipes senegalensis, Alistipes shahii, Alistipes timonensis* [3].

Although species of *Alistipes* are mostly found in healthy human individuals, studies show different species have been isolated from blood, appendicular, abdominal, perirectal and brain abscesses, urine and intra-abdominal fluid of diseased individuals [2]. For instance, *A. ihumii* was isolated from the fecal sample of a 21 years old female individual who was suffering from anorexia nervosa [4]. Also, association of different *Alistipes* species have been noticed in diseases like liver diseases, cardiovascular diseases, hypertension, gut inflammation, inflammatory bowel disease, cancer [1]. Albeit the mode of infection and pathogenesis of *Alistipes* species is unclear; a study has found evidence of tumorigenesis via IL-6 / STAT3 pathway by *Alistipes finegoldii* [5]. Although most gut bacteria have type VI secretion system, no evidence of having T6SS was found in 9 *Alistipes* genomes that were analyzed for a study [6].

1.2 Antibiotic Resistance Gene Operons

Antibiotic resistance gene operon is essentially a set of genes that are controlled by a single promoter which gives a single messenger RNA (mRNA) resulting in encoding multiple antibiotic resistant proteins. Components of such operons are usually a single promoter, operator, structural genes, and regulatory genes [7]. These components of operon have been shown using the lac operon as an example (Figure 1) [35]. In general, there are two types of operons: inducible operon that is activated in presence of an activator molecule and repressible operon that usually remains activated but can be deactivated in presence of a repressor molecule. A microorganism may have multiple operons for successful gene expression. For example, *Escherichia coli* have around 700 operons that include antibiotic resistance gene operon as well [8].

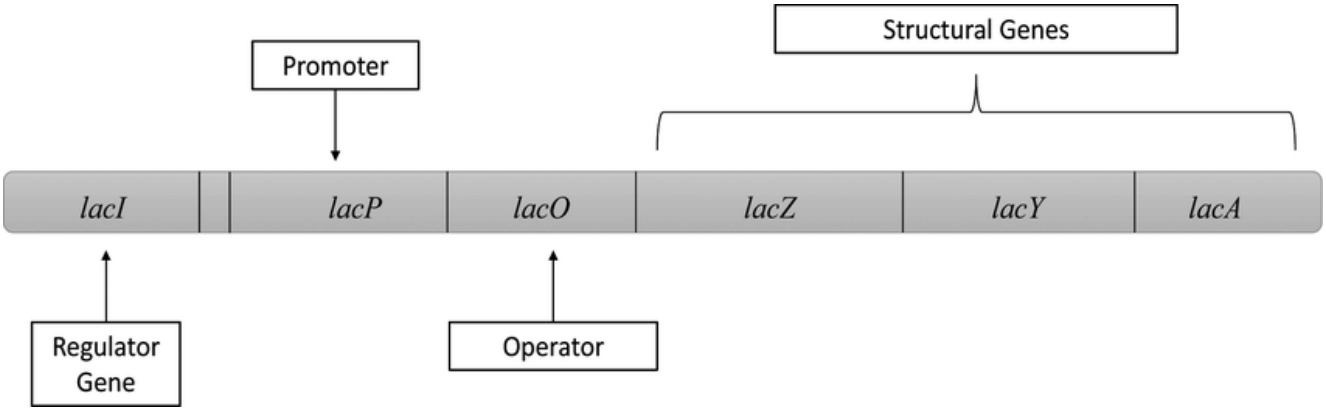

Figure 1: Components of operon has been described using the lac operon as an example. Adapted from [35].

Some antibiotic resistance gene operon inactivates the antibiotic enzymatically by changing pathways, some operons change the conformation or shape of the target with decreased affinity for the antibiotic where the antibiotic could bind via alteration, some operon contain efflux pumps, some operons trap the antibiotics. For instance, the vanA operon (Figure 2) in *Staphylococcus aureus* changes the target peptidoglycan of the antibiotic vancomycin with decreased affinity for vancomycin, thus acquiring resistance against vancomycin [9].

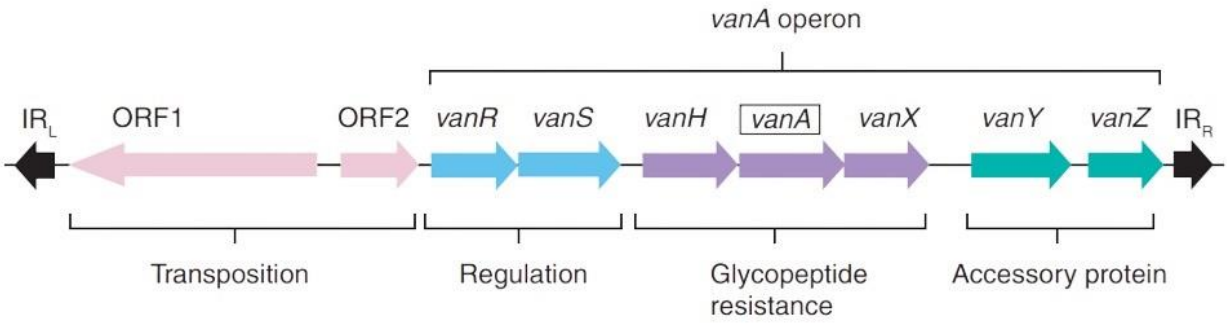

Figure 2: The vanA operon in *Staphylococcus aureus.* Adapted from [9].

Moreover, some antibiotic resistance gene are efflux protein-coding, another mechanism of antibiotic resistance and such efflux pump mediated antibiotic resistance is common in multi-drug (antibiotic) resistance (MDR) (Figure 3) [10].

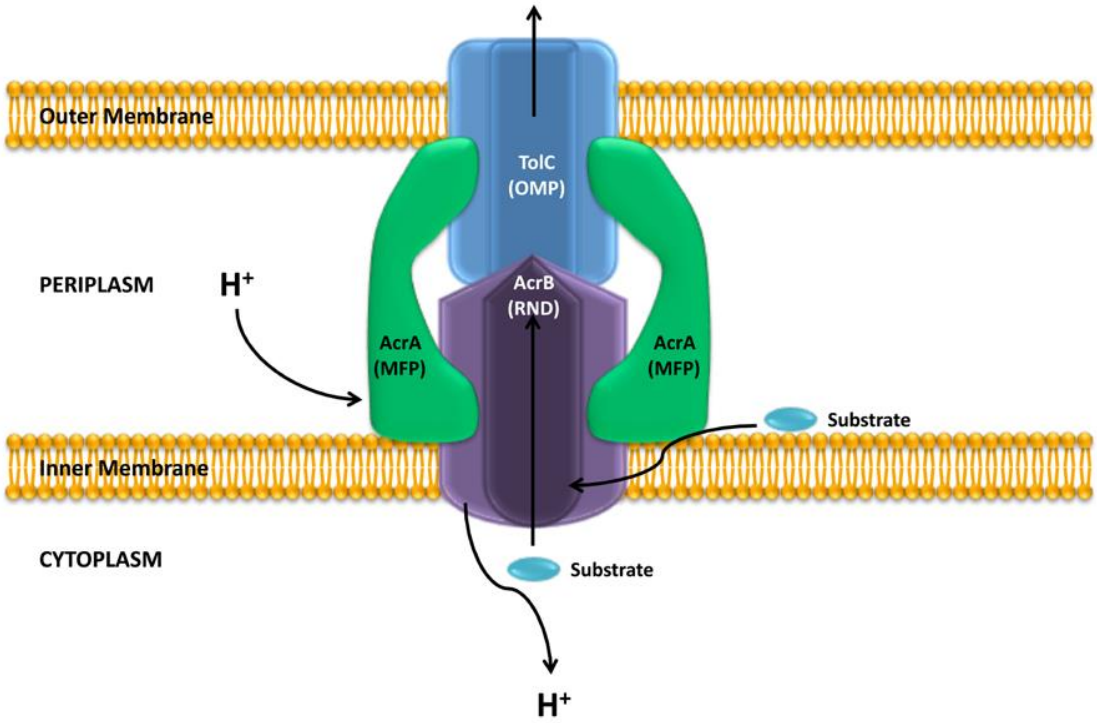

Figure 3: Structure of a Resistance-Nodulation-Division (RND) efflux pump that is found in AcrAB-TolC operon of *Escherichia coli.* Adapted from [10].

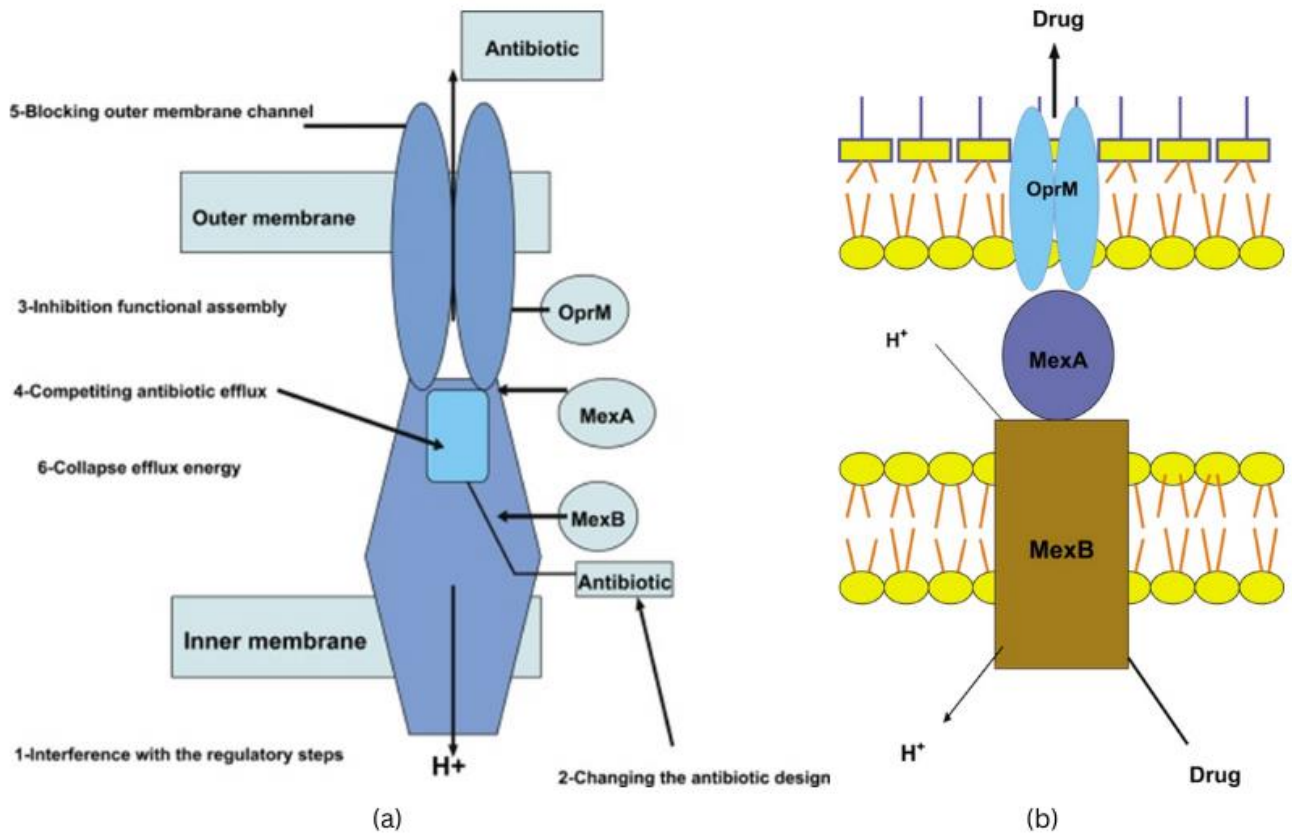

Figure 4: (a) Schematic diagram of the MexAB-OprM operon RND efflux pump in *Pseudomonas aeroginosa.* (b) Schematic diagram of the same operon shows its energy dependency on hydrogen protons for this process. Adapted from [36]

According to Piddock L. J. (2006), the MexAB-OprM operon in *Pseudomonas aeruginosa* (Figure 4, a) [36], AcrAB-TolC, multiple antibiotic resistance (mar) operon in *E. coli*, etc. are example of operons containing efflux proteins. Furthermore, one of the five structural families of bacterial MDR efflux pumps is the adenosine triphosphate (ATP)-binding cassette (ABC) superfamily. Nevertheless, various energy sources (Figure 4, b) [36] are needed to activate the efflux pump such as ATP hydrolysis for ABC transporters [11].

1.3 Bacterial Secretion Systems and Secreted Effectors

Bacterial secretion systems are one of the most important factors associated with virulence and pathogenicity. Some secretion systems are available only in gram negative bacteria while some are exclusive for gram positive bacteria only. The bacterial secretion system invades host cell defense mechanism and secrete the virulence associated proteins to cause disease and the secretion can be either one step secretion or two step secretion. Hence, multiple types of bacterial secretion systems have been identified in different bacteria [13].

1.3.1 Types of secretion system

The pathogenicity of any microorganism is surely related with the protein secretion system it possesses. These proteins encoded by the bacteria, mostly known as bacterial effector proteins and related bacterial pathogenesis are injected or transferred into the target eukaryotic cells by the secretion system that the bacteria have [12]. Till date, 11 such bacterial protein secretion systems or pathways (T1SS – T11SS) have been reported [13].

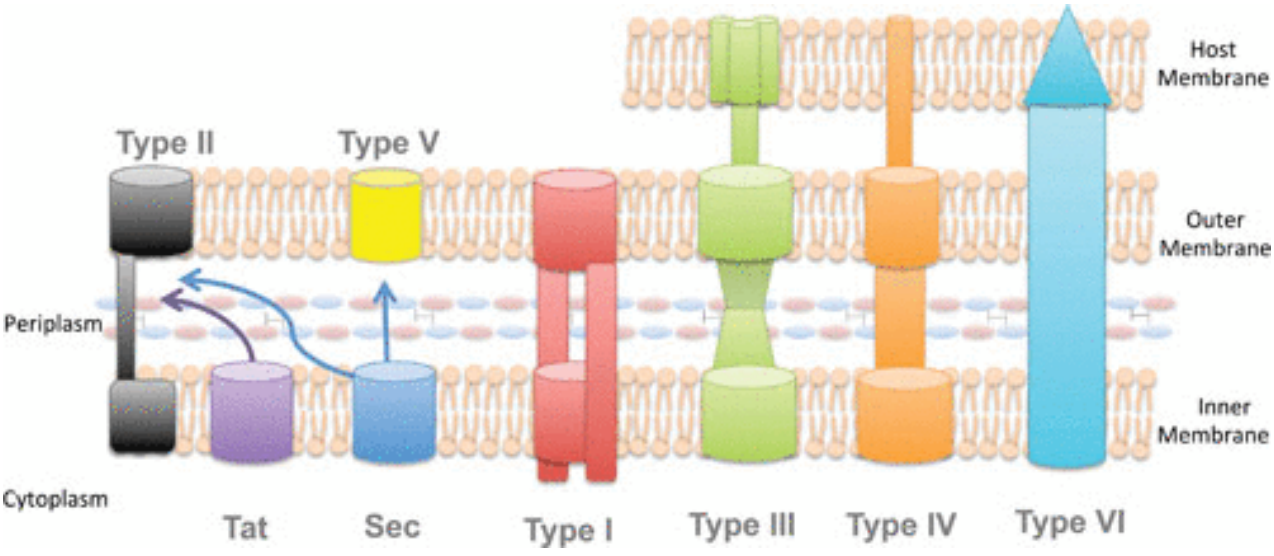

Figure 5: Different types of secretion system in gram negative bacteria. Adapted from [16].

Both gram-positive and gram-negative bacteria have some type of secretion systems in common, albeit type VII secretion system (T7SS) is unique for gram positive organisms. Nevertheless, some

secretion systems contain only housekeeping genes that are not involved in pathogenesis such as non-pathogenic bacteria containing T2SS while some secretion systems are full of pathogenic properties like T3SS and T6SS in most gram-negative bacteria [12,14,16]. Therefore, the components of each secretion system vary one to another. Multiple types of secretion system and associated pathways can be observed in gram negative bacteria (Figure 5) [16].

1.3.2 The Structures and Components of Different Secretion Systems

Starting with the components of type 1 secretion system (T1SS) are: an ABC transporter protein, membrane fusion protein (MFP), and an outer membrane factor (OMF) (Figure 6) [14] while the type 2 secretion system (T2SS) is much more complicated than T1SS consisting an apparatus made of 40-70 proteins from different families that has an inner membrane platform, an outer membrane complex, a secretion ATPase and a pseudopilus that are homologous to T4SS pili system [14,15].

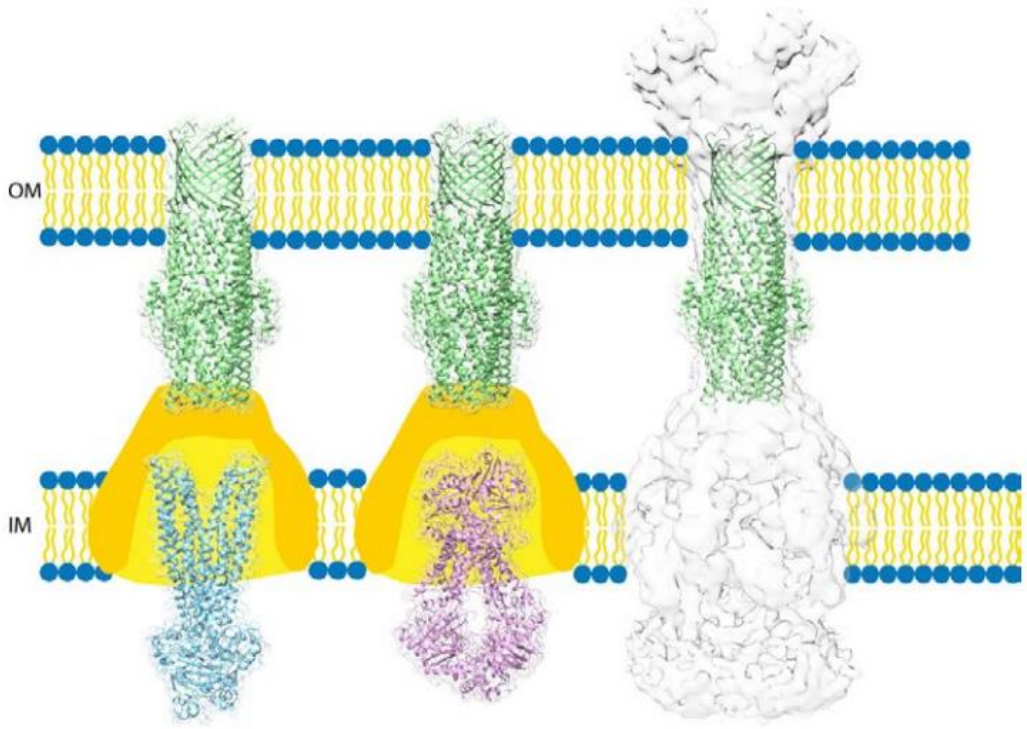

Figure 6: Proposed structure of Type I secretion system. Adapted from [14].

The schematic diagram (Figure 7(A)) is the expression of the operon composition of type II secretion system where the blue colored arrow indicates the secretin and the orange arrow is ATPase. This schematic diagram is a more detailed view of the secretin of the type II secretion system and its other components (Figure 7(B)) [14].

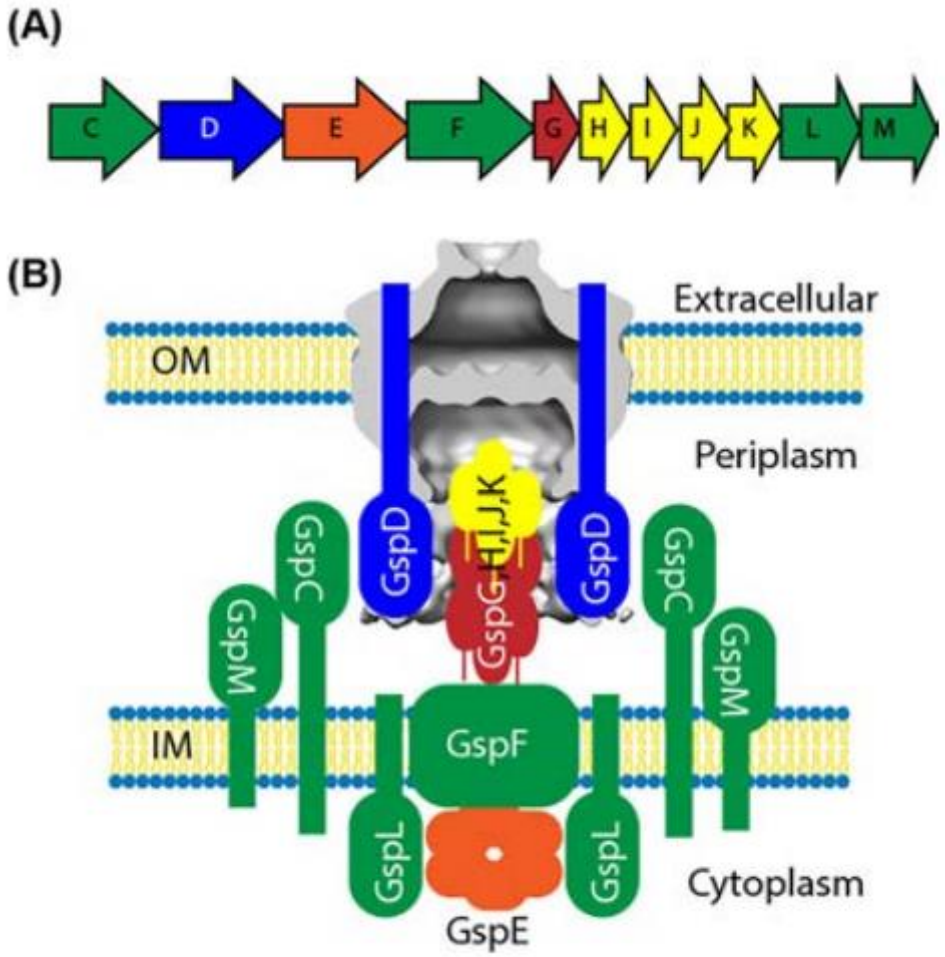

Figure 7: The structural organization of bacterial type II secretion system. Adapted from [14].

Similarly, the type 3 secretion system that is most common in gram negative organisms is comprised of about 30 different structural and accessory proteins including an ATPase, cytoplasmic ring, an inner membrane export complex, an inner membrane ring, an outer membrane ring, an inner rod, pili, and a translocon tip complex (Figure 8) [14,15,16].

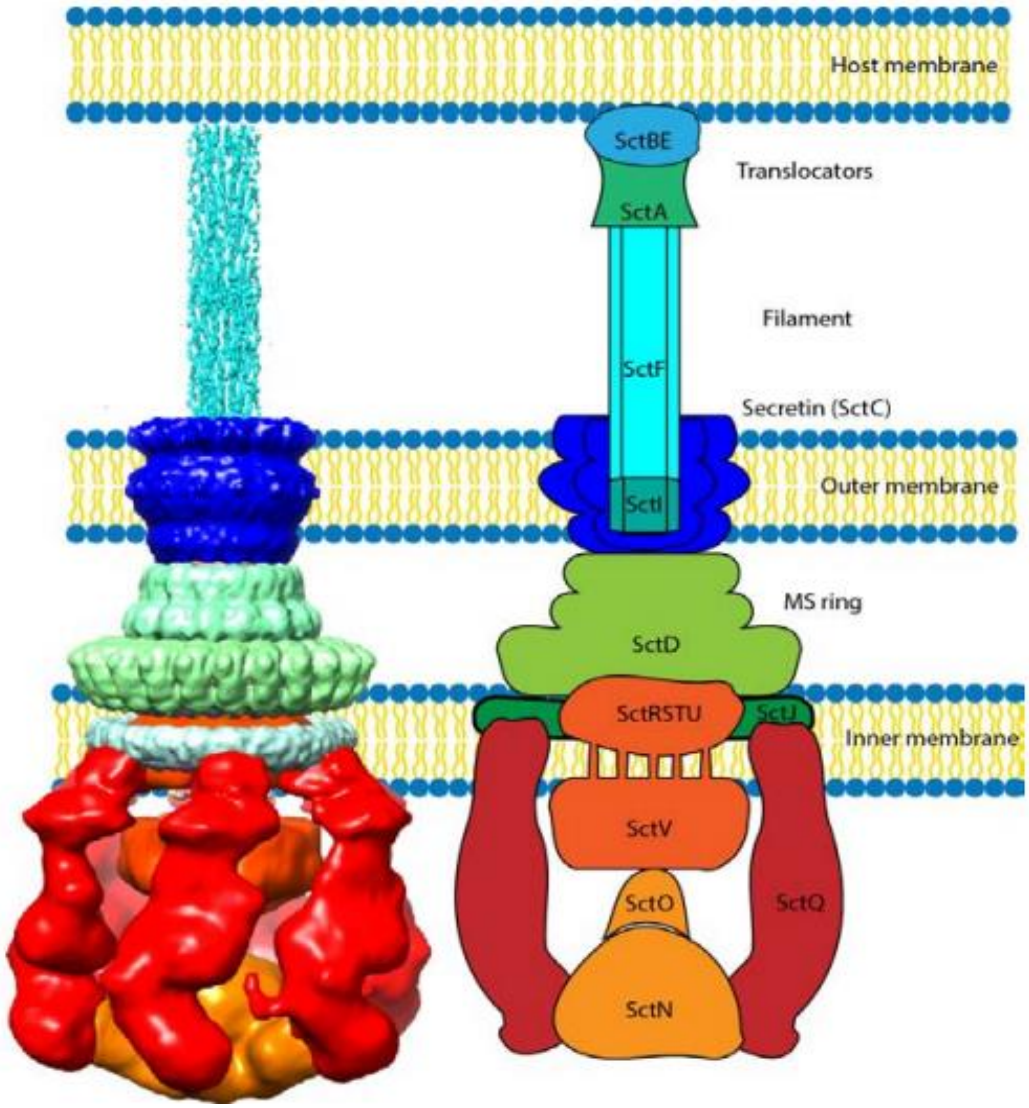

Figure 8: The structure of Type III secretion system. Adapted from [14].

According to Darbari and Waksman (2015), bacterial type iv secretion system (T4SS) has 12 components reported as VirB1-11, VirD4 and three ATPases from these 12 components VirB4, VirB11 and VirD4 fuels the system (Figure 9) [37]. Moving next, T5SS is composed of a unique protein, auto-transporter that contains a beta (β) barrel domain which enables this system to secrete the proteins to the outer membrane itself without needing any other secretion apparatuses like other secretion systems described so far [15,16].

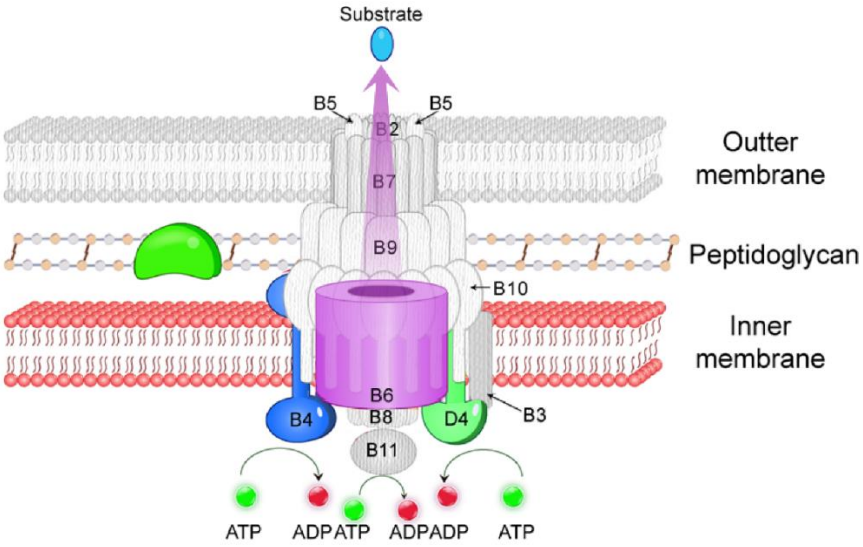

Figure 9: Schematic diagram of the structure of the type IV secretion system. Adapted from [37].

The type vi secretion system (T6SS) is mainly composed of 13 components that mirrors an inverted bacteriophage tail upon being assembled into a large complex (Figure 10, B) [14]. Along with the sheath structure T6SS has additional components known as PAAR proteins that are T6SS effectors as well as structural proteins [12].

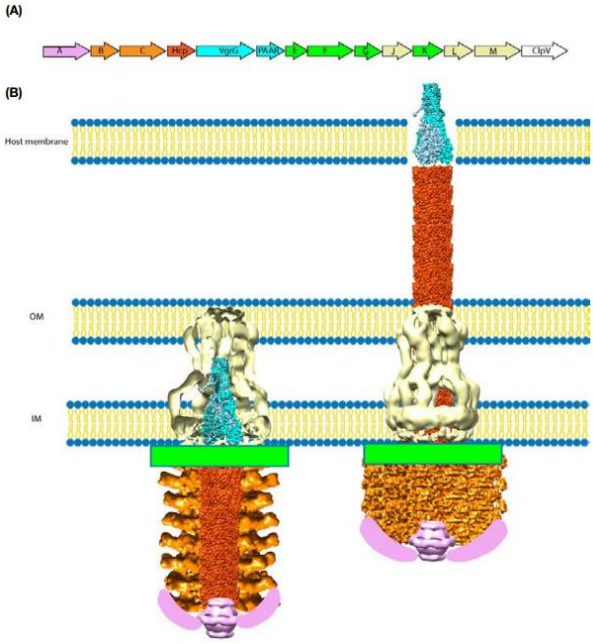

Figure 10: Schematic diagram of type VI secretion system. Adapted from [14]

Several core inner membrane proteins interacting with cytosolic chaperones and forming a channel to secrete effector proteins are the components of T7SS that is unique for gram positive organisms. (Figure 11) [14]. In gram negative organisms, T7SS is known as the Chaperone-Usher (CU) pathway [12].

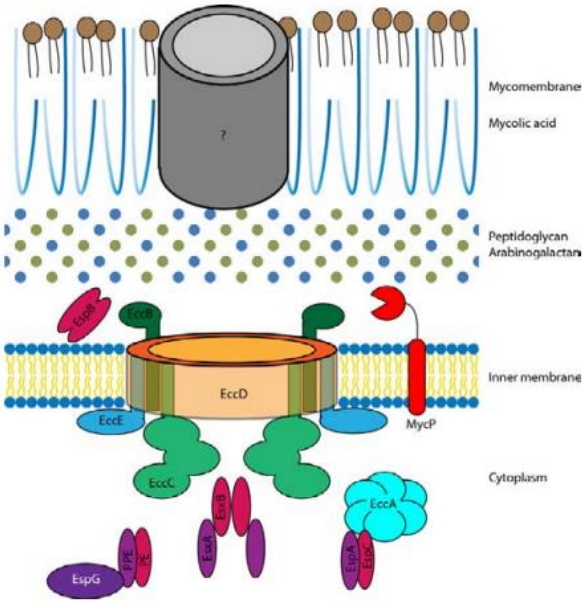

Figure 11: Schematic diagram of type VII secretion system. Adapted from [14]

Then, the T8SS or the curli secretion system is composed of non-covalent complex extracellular curli fibers that consist of CsgA, major subunit and CsgB, minor subunit along with other accessory proteins. Next, locating in bacterial outer membrane and that can form a transmembrane β-barrel the protein conducting translocon SprA is the major component of T9SS [15]. Lastly, an outer membrane β-barrel protein is the only known component of the newly described secretion system TXISS or T11SS [18].

1.3.3 The Effects of Different Secretion Systems

Now, the effects of the bacterial secretion system are not only associated with virulence but also with other attributes such as cell signaling, surviving in harsh conditions, surface attachment, biofilm formation, etc. For example, the curli secretion system or T8SS causes infection and host inflammation as well as involved in cell aggregation, biofilm formation and other sophisticated activities via their secreted effectors [15,19].

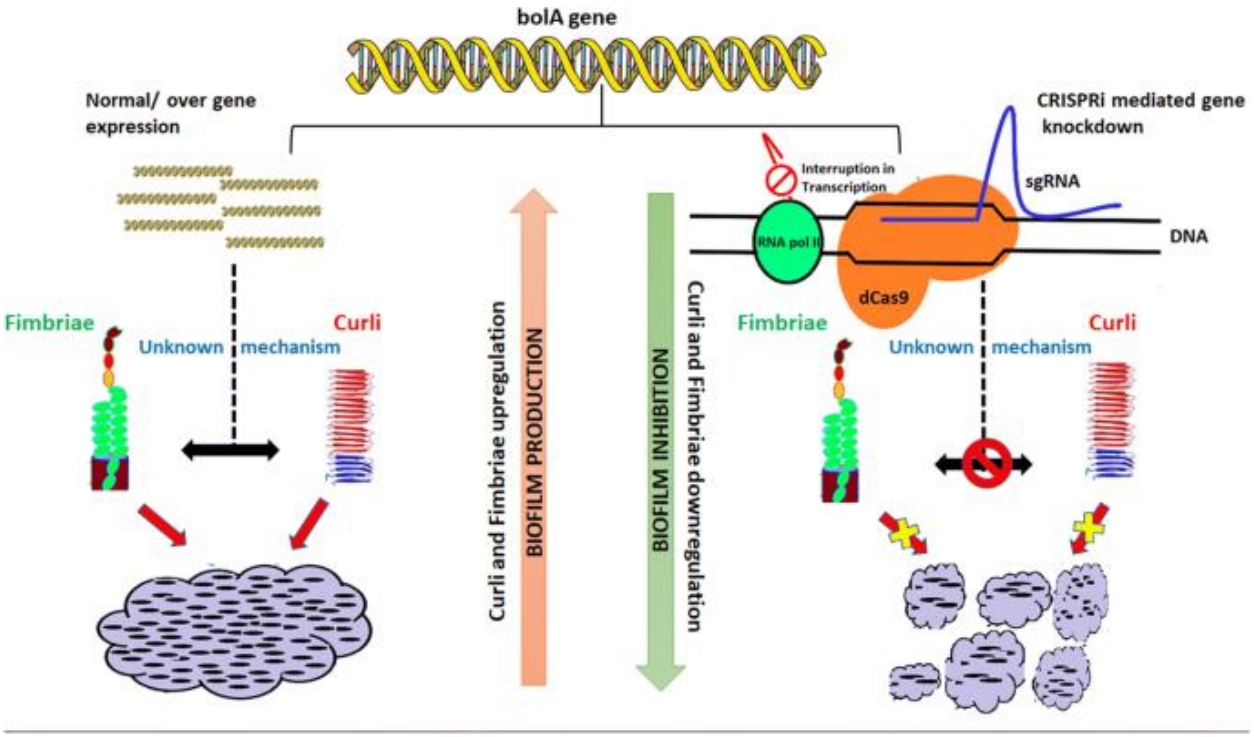

Figure 12: Type VIII secretion system affecting biofilm formation by regulating curli and fimbriae. Adapted from [38].

The type VIII secretion system (Figure 12) facilitates the acquiring process of the folding structure of curli fibers. These curli fibers are then responsible for biofilm formation and the upregulation of curli and fimbriae supports biofilm formation whereas downregulation works as an inhibitor of biofilm formation [38]. The T1SS secrete RTX proteins that are core virulence property of this secretion system. This secretion system is also useful in biological functions like cell attachment, biofilm formation, digestion systems, etc. in some bacteria.

Despite the T2SS in *Vibrio cholerae* is used to secrete cholera toxin by the bacteria and cause disease; T2SS is mostly required for surviving in the environment, facilitating bacterial colonization, providing nutrients, etc. [14,15]. Nevertheless, the type II secreted effectors in *vibrio cholerae* are necessary for the survival of the organism as it provides it nutrition at the first stage and in terms different effectors facilitated biofilm formation, release of toxin leading to pathogenicity and maintaining its life cycle (Figure 13) [39].

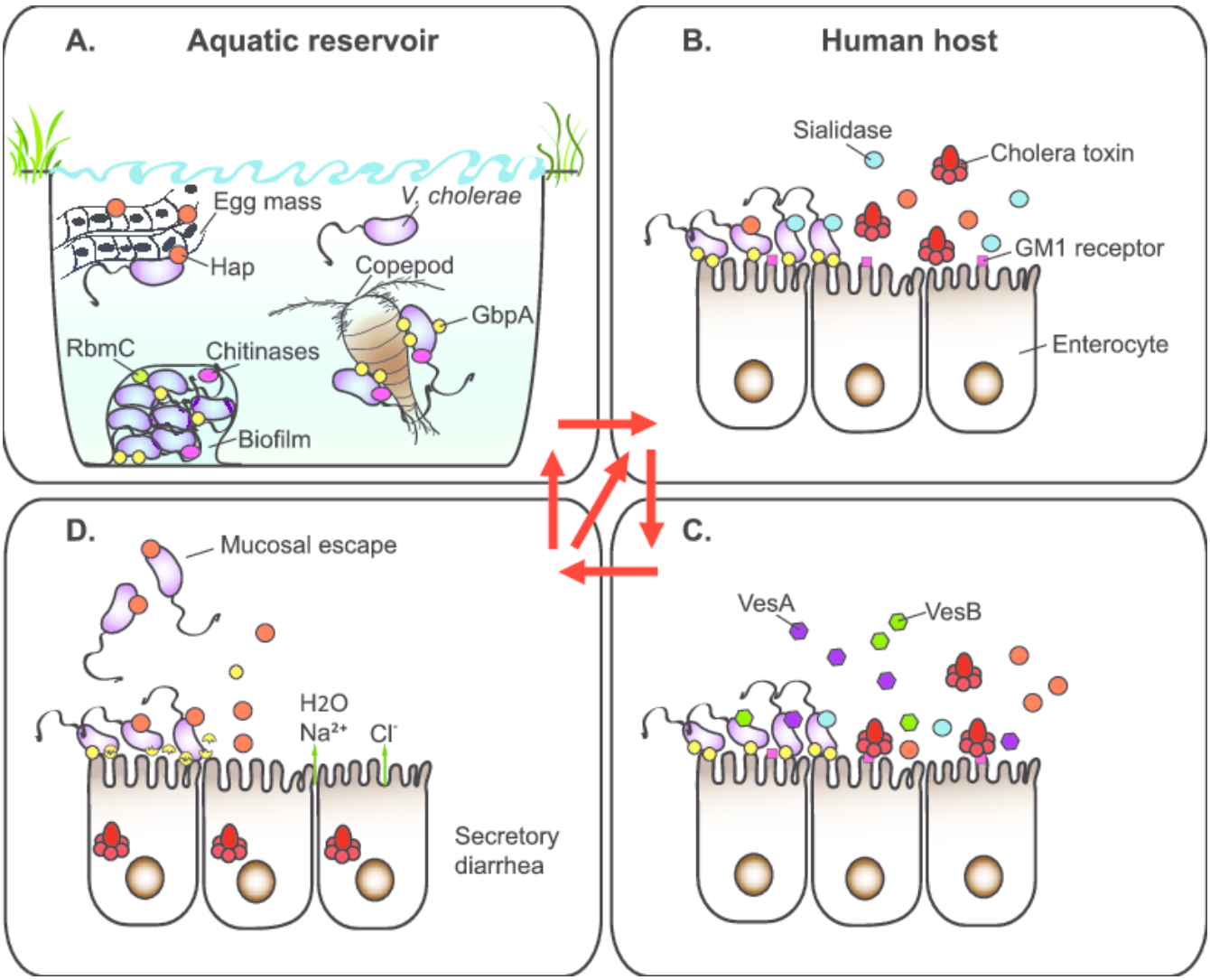

Figure 13: Type II secretion system in *vibrio cholerae* facilitating its dual life cycle. Adapted from [39].

The most common secretion systems T3SS and T6SS in gram negative microorganisms are responsible for mostly known bacterial diseases in humans, cattle, fish, etc. Furthermore, T4SS secreted effectors are effective in reproduction of the bacteria possessing T4SS, also in pathogenesis. Similarly, some T5SS secreted effectors play a role in virulence, but other secreted exoproteins are involved in attachment, escaping host immune system, etc. [14] All around, every function of all the bacterial secretion systems is mostly related to pathogenesis and leaves detrimental effects on the host.

1.3.4 Operons of Types of Secretion Systems

Nevertheless, bacterial protein secretion depends on the operons ultimately encoding the proteins that are secreted via these secretion systems. Different secretion systems are part of various operons and the operons control the protein secretion, pathogenesis, etc. in numerous microorganisms [7]. Besides, the secretion systems are basically encoded by the operons. For example, the virB operon in the genus *Brucella* encodes the type iv secretion system which is the sole virulence factor of this organism. The operon system has three functional groups and 11 proteins in total where short intergenic regions are found in between some proteins [20].

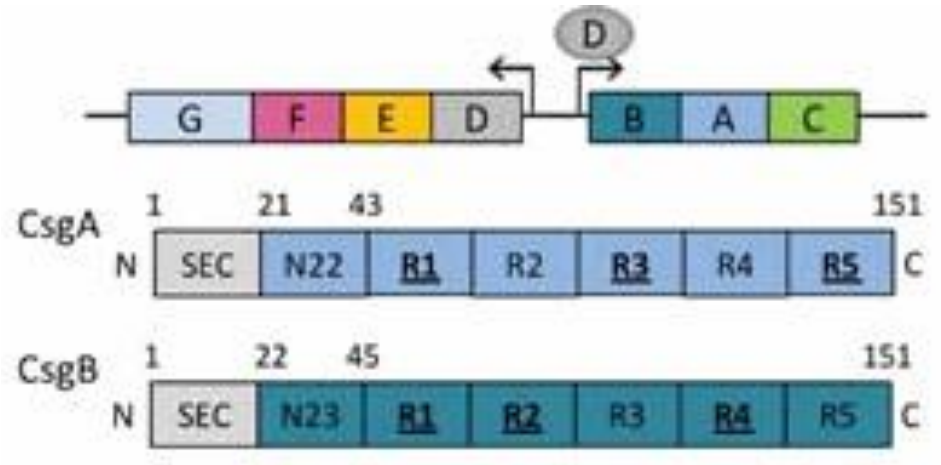

Figure 14: Schematic representation of curli secretion system gene operon. Adapted from [19].

Similarly, T8SS or curli secretion system have two operons csgBAC and csgDEFG, one for encoding structural proteins and another for encoding accessory proteins and the operons are dependent on sec pathway for protein secretion (Figure 14) [19].

From these two examples one thing is clear that the operons are constituted by the genes that are part of their aforementioned core components. Antibiotic resistance genes, translocator genes, various transporter genes, conserved genes, genes encoding efflux pumps, genes encoding different functional enzymes, operon regulatory genes. etc. are part of different operons. Each secretion system varies from one another in functions, structures, operons, etc.; however, all of them basically fulfill at least one similar function that is to escape the host immune system and cause disease.

1.4 Bacterial Toxin-Antitoxin Systems

Another system associated with bacterial pathogenesis is the Toxin-Antitoxin (TA) system that is commonly found in most pathogenic bacteria. Moreover, the name Toxin-Antitoxin system itself is quite self-explanatory that depicts this system is associated with two proteins: one is the toxin and another protein is the antitoxin for the toxin that holds the ability to counteract with its cognate toxin and mitigate the deleterious effects of the toxin. [21,22]

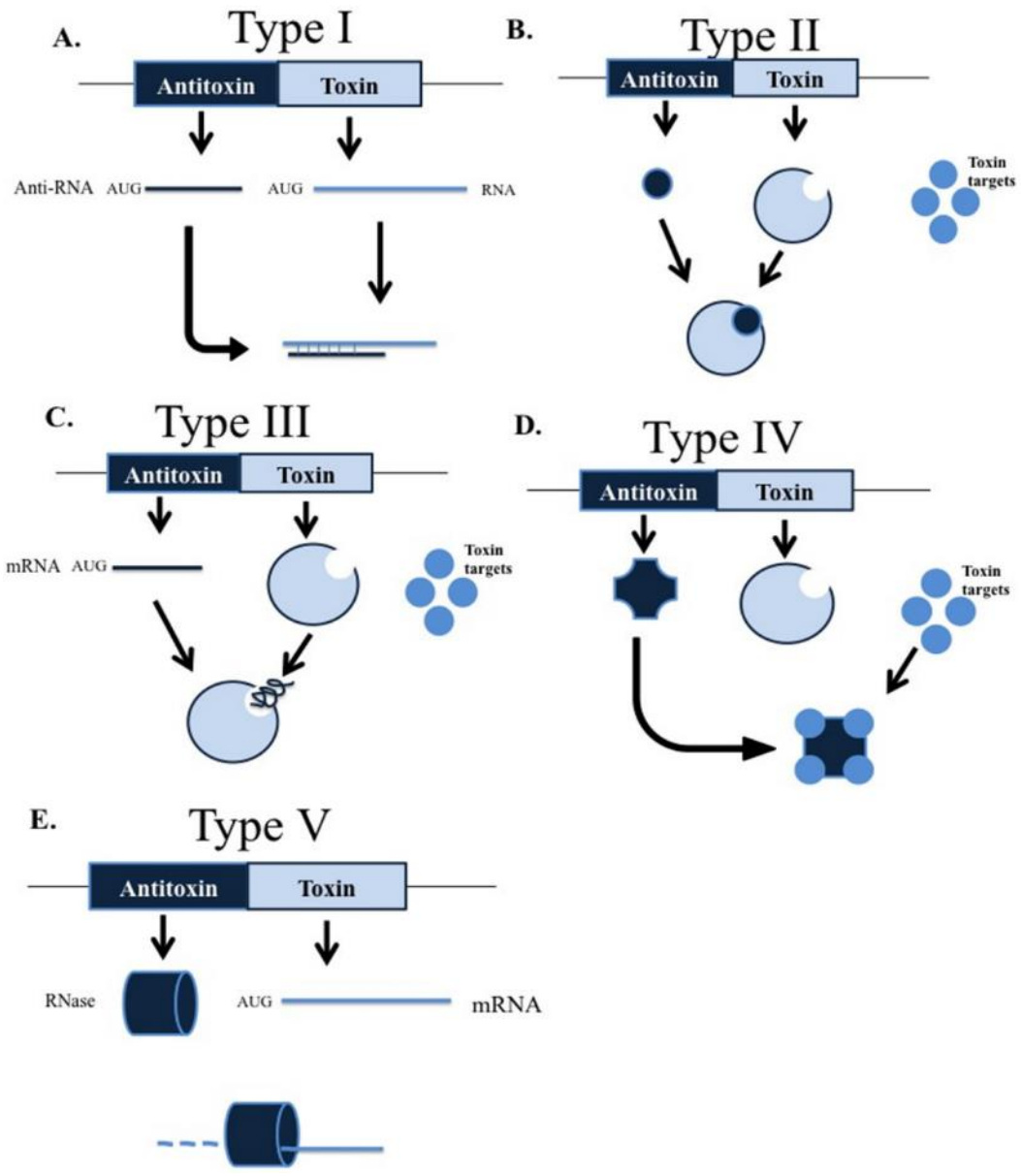

Figure 15: Types of Toxin-Antitoxin system. Adapted from [25].

Furthermore, pathogenic bacteria and archaea inherit the toxin-antitoxin system via horizontal gene transfer and some studies have mentioned that this system is associated with the maintenance of the plasmid consisting of the genes of antibiotic resistance, virulence, etc. Further, this system helps the organism to persist during host replication by releasing toxin to kill the host organism and releasing antitoxin within itself to protect itself against the toxin. Similarly, pathogenic organisms having TA systems affect other bacteria by releasing toxins when there is food and nutrient shortage in the environment and competition is high for these amongst other bacteria as well as other stress signals [21-25,27].

Now, both plasmid-encoded and chromosomally-encoded TA systems play role in pathogenesis via expression of virulence factors, persistence during stressed environment, biofilm formation, inactivating key metabolic functions, etc. while chromosomally encoded TA system in *Burkholderia pseudomallei* might be associated with human infections through bacterial persistence [22,25,26].

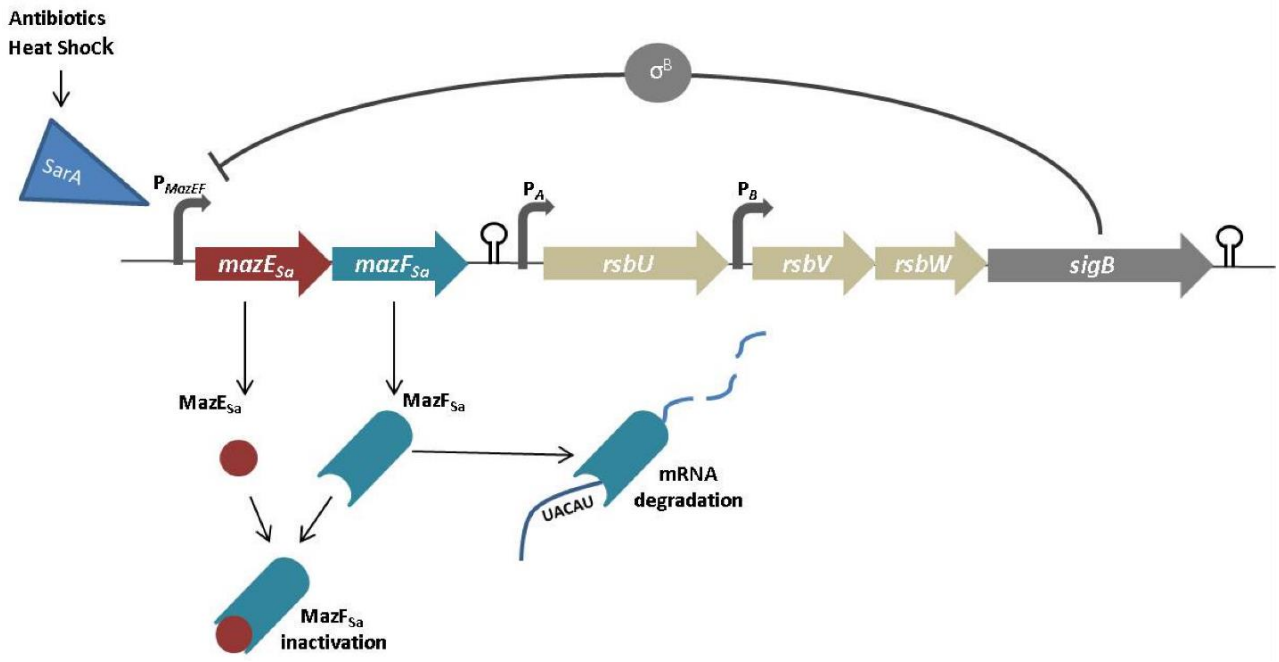

Figure 16: Toxin-Antitoxin system in MazEFSa operon. Adapted from [25].

As mentioned earlier, TA system operon only consist of two components the toxin and the antitoxin; albeit virulence related proteins are also encoded from this operon. Nevertheless, single stranded RNase activity of PIN domain that is mostly known as vapBC operon is also related to

TA system and considered as the toxin part of the TA system [21,27]. In addition to that, vapBC operon is common in pathogenic bacteria such as *Mycobacterium tuberculosis* causative agent of Tuberculosis in humans and *Pyrobaculum aerophilum*, *Neisseria gonorrhoeae,* etc. [27].

Similarly, MazEFSa operon in *Escherichia coli* a part of the host genetic network that is chromosomally encoded TA system is often related to programmed cell death. This operon is regulated by the promoter mazEF and the toxin MazF is neutralized by the antitoxin MazE. The antitoxin binds to the toxin site and inactivates it to infer a signal for programmed cell death in the host genetic network for persisting in harsh environment (Figure 16) [25].

1.5 Pan-genome Analysis

The knowledge of genomic characteristics from all the available genomes of a group of organisms is important to infer its diversity, complexity, pathogenesis, etc. and one way of knowing this is to perform pan genome analysis. The pan genome is the collection of all genomic features within a clade including the core genomes as well as the features that are not common in all of the species or strains [28,29,32]. Furthermore, a methodology for assessing genomic diversity in the available genome sequences and calculating the number of additional whole genome sequences required for defining the diversity adequately is provided by pan-genome analysis [30,31].

Now, the pan genome analysis returns the homologous genes from the provided analyzed dataset. Complete genomic data is used to perform the pan genome analysis while the use of scaffold or draft genomes in pan genome analysis is limited [32]. Since the use and importance of pan-genomics have increased, several online tools along with multiple stand-alone software have been developed to ease the process of pan genome analysis [30].

In the flow diagram (Figure 17), some of the essential steps for performing a pan-genome analysis have been presented with a flow diagram [32]. However, different methods and several other software online based tools have been introduced recently for performing this analysis in an efficient manner. Therefore, the steps for performing the analysis might vary depending on the method, tools and perspective.

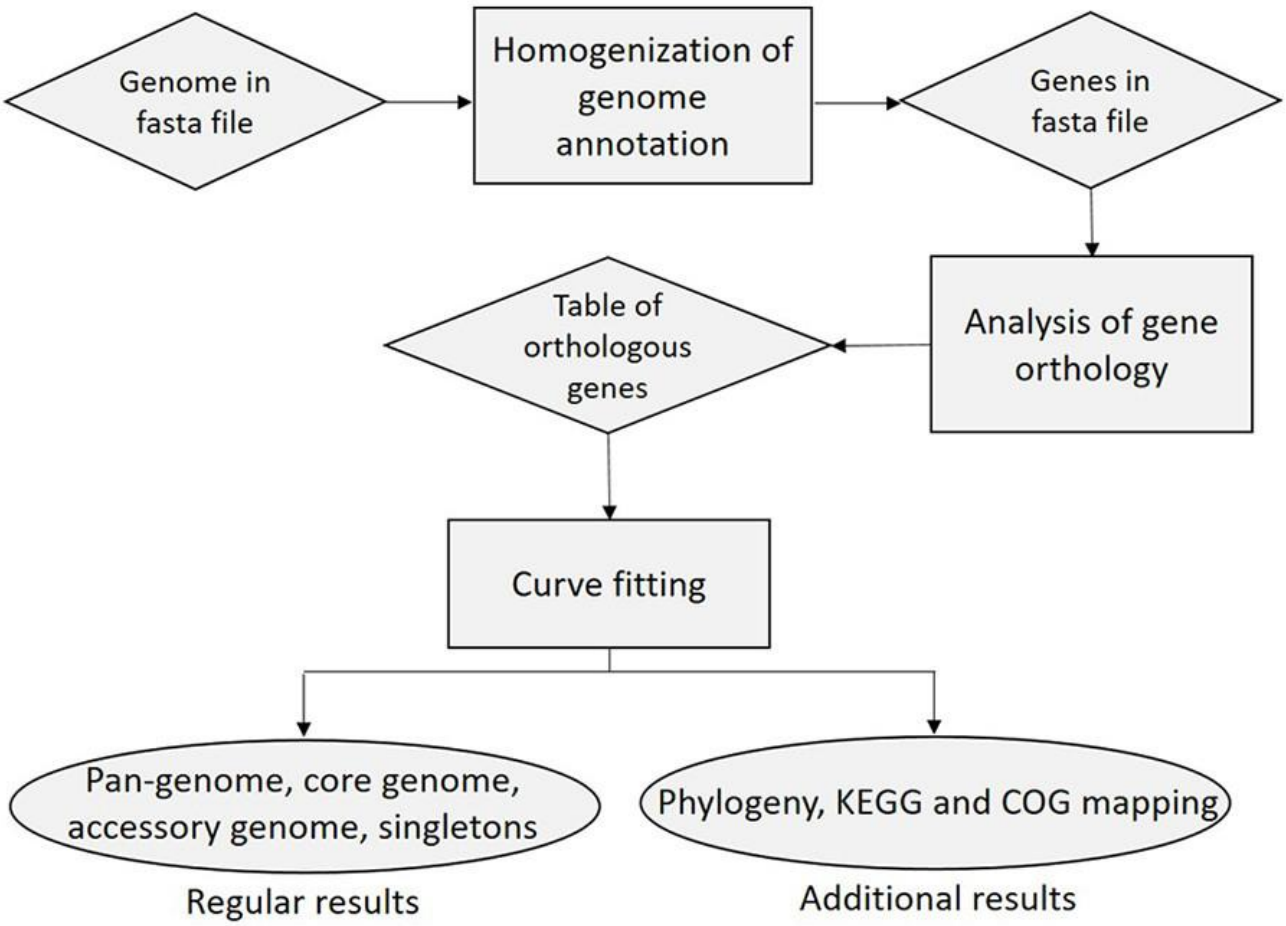

Figure 17: Flow diagram of main steps of pan genome analysis. Adapted from [32].

First process is to annotate the genomes with the same software or tool for homogenization of genome annotation. Softwares like GeneMark, RAST, PROKKA, etc. are usually used for this process. Next for clustering the genome tools like OrthoMCL, GET HOMOLOGUES, etc. and for the pan genome calculation tools like BPGA, GET HOMOLOGUES, PGAP, etc. are used [29, 30, 32].

Overall online tools like Panakeia, PGAweb can be used for online based pan genome analysis [33,34] whereas softwares such as Piggy, Roary, ClustAGE, DeNoGAP, EUPAN, micropan, Panaconda, PanCake, etc. can be used to successfully perform pan genome analysis [29,30,32]. The use of different tools and software are dependent on the respective purpose and objectives by the researchers.

1.6 The Rationale

To begin with, the genus *Alistipes* was selected based on a few criteria. It was ensured the genus that was to be chosen for this pan-genome analysis should have some clinical significance on animal kingdom, preferably human. Other criteria were to find a relatively less studied genus with at least more than five species having complete genome or at least chromosome level genomes available at the NCBI database. Despite searching for a less studied organism was the focus, the presence of significance on the other available researches on the organism was also checked. Thus, the background information and the knowledge gap were identified as well as gathering some key findings from other works to mention in this analysis.

Then, the genus *Alistipes* was chosen as it matched with all the aforementioned criteria. It has 16 species and some species with multiple genomes with complete as well as chromosome level genomes allowing to broaden the scope of analysis. Although the number of complete and chromosome level genomes are less than 20. Nonetheless, other works on this organism have shown that it is found in the gut microbiota of human, also having some other clinical significances.

Next, it was important to set the focus and what to look for in these genomes as well as setting the perspective of analysis for this research. From the previous studies, the virulence factors, pathogenicity, role to the current global issue that is the antibiotic resistance was not clear for these genomes. Also, it was not clear if this organism is good for the human microbiota or bad as it has been isolated from both healthy and diseased individuals.

Therefore, it was hypothesized that the organism might have some virulence properties that might be related to cause infection and further contribute to antibiotic resistance. Hence, the first thing to do was to find the secretion systems as well as toxin-antitoxin systems and virulence factors associated with these genomes. If known secretion system, toxin-antitoxin system, virulence factors, antibiotic resistance genes from other pathogenic and antibiotic resistant organism, can be found from this analysis that we aim for; then this knowledge might help to prevent diseases caused by this organism in future and might even help to limit use of wrong antibiotics against infections caused by this organism and prevent contribution to antibiotic resistance by this organism.

Hence, the objective of this project is to find the secretion systems, toxin-antitoxin systems, virulence factors, antibiotic resistance genes, also their operons and identify other significant proteins and genes from these genomes first. Second, the findings from these genomes will be matched with the known pathogenic properties from other organisms as mentioned earlier. Third, the diversity among the genomes and other genomic properties like complexity, pathogenicity, etc. will be identified by performing the pan-genome analysis on the genus *Alistipes*.

Chapter 2. Methodology

In the following flow diagram (Figure 18) the overall steps that were followed to perform this analysis have been represented.

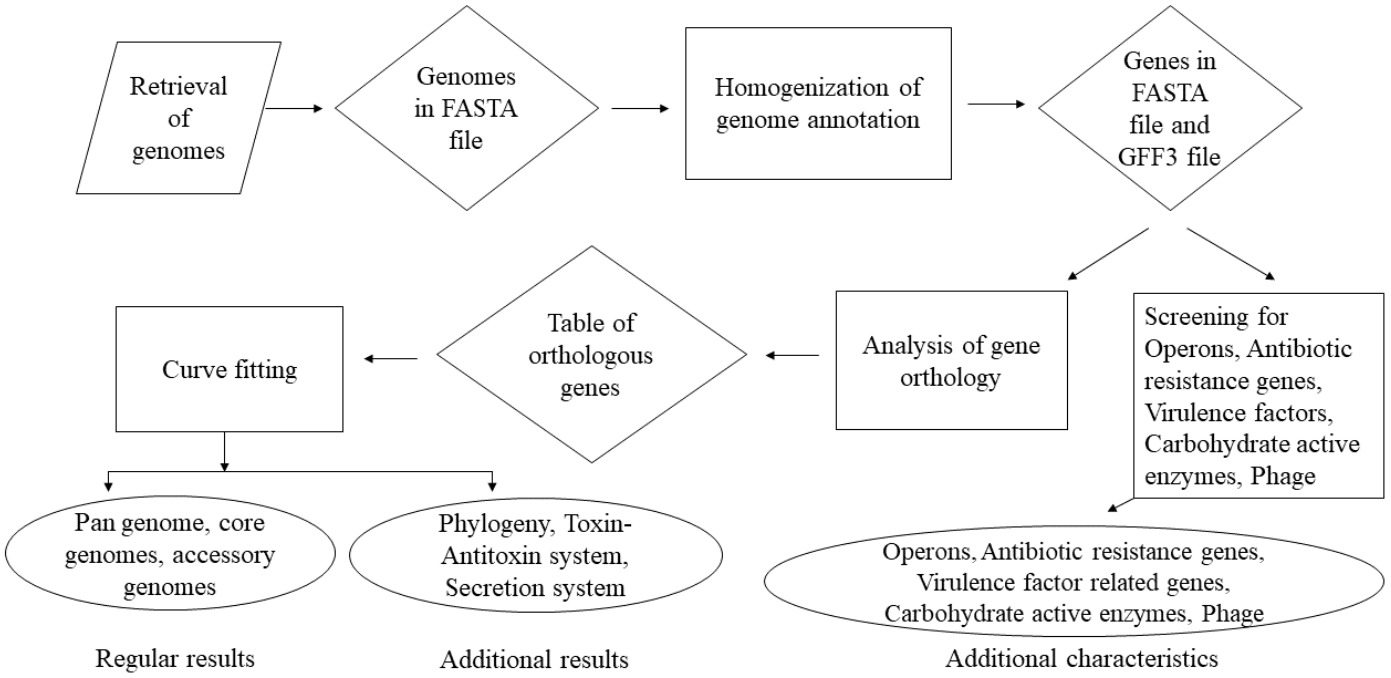

Figure 18: Methodology used for performing this pan-genome analysis

The retrieved genomes in FASTA format were used to perform the pan-genome analysis as well as to run in different online based bioinformatic tools to find out some additional characteristics as mentioned in the flow diagram (Figure 18).

2.1 Retrieval of Genomes

The genus *Alistipes* has over eleven hundred genomes available of the 16 identified species on the National Center for Biotechnology Information (NCBI) genome database [40]. However, the majority of them are at scaffold and contig level. Therefore, these scaffolds and contig leveled genomes were filtered out first and the complete level and chromosome level genomes were

considered for the analysis. Then, a note of the accession numbers of the complete and chromosome level genomes of the *Alistipes* species was taken.

Moving forward, the Galaxy web platform [41] was accessed online and was used for retrieval of the genomes. Since the accession numbers of the genomes were available, the genomes were retrieved from NCBI using the Get data option at Galaxy, then upon selecting NCBI dataset genomes the accession numbers were provided. The genomes were retrieved in FASTA format. These genomes were available in the galaxy environment, as well as in the system. Furthermore, the integration with galaxy had made this process much easier and more convenient as galaxy is an unrestricted, web-based platform that is free to use for anyone and can be used for all aspects of genomic analysis including data retrieval and integration from popular databases; for example, NCBI genome in this case.

Galaxy also enables researchers to do multi-step analysis, repeated analysis by providing workflows, collaboration and publication. Also, this environment has over 7500 bioinformatic tools that can be used as per interest and requirement without facing any trouble. Nevertheless, users can integrate their research within Galaxy and customize their own galaxy environment with the tools they may require for their analysis and they can perform their analysis and keep track of everything in the galaxy cloud system. Without the fear of losing any data from the retrieval of genomes to multi-step analysis with different tools and publication ready figures and visualizations, galaxy provides a suitable environment for the researchers who do not require computer programming experiences as well [41].

2.2 Annotation of Genomes

The 14 complete genomes and 2 chromosome level genomes of different *Alistipes* species that were retrieved using Galaxy from NCBI genome was next annotated using the Prokaryotic Genome Annotation System (Prokka), v1.14.5 [42] within Galaxy, as the retrieved genomic data in FASTA format of the 16 genomes were already stored in the Galaxy environment. Prokka was run within Galaxy as a command line program using the default parameters and it returned the annotated files in GFF3 format. Since Prodigal is integrated in Prokka and it is believed to be one of the efficient algorithms for predicting prokaryotic genes, the annotations generated using Prokka

can avoid false positive predictions compared to other available gene finding and gene annotation tools.

In Prokka, the genome annotation time is faster and accuracy of the annotation is better than other alternative tools. Prokka uses Prodigal, RNAmmer, Aragorn, SignalP, and Infernal for predicting features of the genomes; thus, providing an accurate annotation. User has to provide a complete DNA sequence in FASTA format as input. Next, the sequence will be compared to a primary database of known proteins using BLAST and blastp. Furthermore, the provided sequence is compared to multiple databases with known protein sequences and upon finding a significant match, Prokka provides an annotation for the sequence. Otherwise, the protein is annotated as a hypothetical protein. Prokka can be run in any Unix operating system as a stand-alone version; however, in this case it was run within Galaxy for convenience [42].

2.3 Pan-genome analysis

Roary is a pan genome analysis tool that can construct large scale pan genomes from given GFF files for example, GFF3 files that are generated via Prokka, faster with greater accuracy than other available pan genome analysis tools. It can identify the core genes and accessory genes with greater efficiency upon consuming less memory and less analysis time to calculate pan genome. Nevertheless, the only requirement for successful pan genome analysis via Roary is to use a single annotated assembly per sample from the same species [43].

Here, the GFF3 files generated by Prokka that is the annotated sequences of the 16 genomes of different *Alistipes* species was used for pan genome analysis via Roary v3.11.2 within Galaxy, as the galaxy environment already had the genomic data and annotated GFF3 files generated via Prokka; thus, reducing analysis time. Furthermore, there was less chance of occurrence of error due to mistakenly uploading the wrong files for analysis.

2.4 Operon finding

Operon Mapper is an online based bioinformatic tool that can directly predict operons from any given sequences or bacteria or archaea with great accuracy easily. The process of using this web-based tool is quite simple. The input has to be a FASTA sequence that can be uploaded or can be

pasted in the query box. Furthermore, Operon Mapper can predict ORFs from the same given sequence if the user provides GenBank coordinates or GFF files that also can be either pasted in the query box or uploaded. Finally, the user has to provide a valid job description for each run and provide a valid email address upon selecting the output options based on the requirement, and submit the job for testing in Operon Mapper. This artificial neural network (ANN) based tool can return results quite fast; however, it depends on the server load as well as the size of the provided sequence and output options [44].

Operon Mapper was accessed online and the Prokka generated GFF files and FASTA sequences of all 16 genomes of *Alistipes* species was used for running in Operon Mapper to find operon within these sequences. Since Operon Mapper deals with one sequence at a time, both FASTA sequence and GFF files were uploaded per genome per run. For output, the option "all possible outfiles and a compressed file with all of them" was selected as it contained all the data of predicted operonic gene pairs, predicted operons, Predicted ORFs coordinates, DNA sequences of the predicted ORFs, Protein sequences of the translated predicted ORFs, COGs assignations, and ORFs functional descriptions.

2.5 Identification of Antibiotic Resistance Gene

Antibiotic resistance has become a global concern and the number of antibiotic resistant bacteria is increasing day by day. Hence, the presence or absence of known antibiotic resistance genes was checked on the 16 selected genomes of different *Alistipes* species. Therefore, the Resistance Gene Identifier (RGI) from the Comprehensive Antibiotic Resistance Database (CARD) was used here [45].

RGI is an online based tool that is developed mainly for resistome analysis and prediction of resistant genes within a sequence. However, this tool is also available as a command line version, and can be also used via Galaxy wrapper. Although the command line version is advantageous than the web version as the command line version allows the user to analyze metagenomic reads as well as supports the prediction of k-mer of pathogen-of-origin for antimicrobial resistance genes. While the web version can be used for genome analysis, genome assemblies, metagenomic contigs, or proteomes except the additional features mentioned for the command line version [45].

The RGI web portal is pretty easy to use from any browser. CARD-RGI was accessed online and the DNA sequence in FASTA format was uploaded; albeit there was an option to provide GenBank accession. Next, the input Data type was selected as DNA sequence, the option of perfect, strict and loose hits criteria was also selected, nudge was included to identify loose hits ≥95% than to strict and the sequence quality was selected as high quality/ coverage and each job was submitted similarly. After completion of each job, the results were downloaded and saved in the system.

2.6 Identification of Virulence Factors

The virulence factor database (VFDB) consists of the virulence factors of different pathogenic bacteria. As well as, it consists of the structural features of these virulence factors, different functions, variety in their types, mechanism of pathogenesis, et cetera. This classified information helps researchers to identify virulence factors in new species of bacteria and help them to measure the severity of their pathogenesis and potential disease pattern. Furthermore, this also helps to take preventative steps well before these new organisms can cause diseases [46].

The VFanalyzer pipeline under the VFDB database does the job of identifying known or could be virulence factors in complete or draft level bacterial genomes. VFanalyzer searches the input sequence against the known datasets of several pathogenic bacterial genus. Upon mentioning the genus of the bacterial genome sequence to be inputted, strain name needs to be given as well as mentioning the type of the sequence file whether it is completed sequence or draft sequence et cetera before uploading the sequence in FASTA format. Then, a valid email address and name of the Institution needs to be addressed before submitting one run [46].

Since *Alistipes* is a comparatively new bacterial genus, the VFDB did not have it in its genus list. Hence, after analyzing the phylogenetic trees, the genus *Chlamydia* was found to be closest to *Alistipes*. Except for one draft genome, “The raw FASTA sequence(s) of a COMPLETE genome” was selected for the “type of upload file” and for the draft genome we the selected option was “The raw FASTA sequences of a DRAFT genome.” Then the sequence was submitted accordingly and upon receiving a job ID it was saved to retrieve results. Finally, the result table and associated files were downloaded and saved for analyzing later. The same step was followed for all 16 genomes.

Then, the tool Ngphylogeny was accessed online to generate phylogenetic trees for genes that had more than 3 sequences or at least 3 sequences [47]. Another online based tool, iTol was used for the visualization of the phylogenetic trees [48]. Next, InterProScan was accessed online and the single gene sequence was run here to crosscheck the protein domain family that was found in the Roary result [49]. For multiple sequences, in this case for less than three and more than one sequence was analyzed for multiple sequence alignment at T-coffee server using M-coffee online [50, 51].

2.7 Identification of Toxin-Antitoxin System Genes and related Operons

Toxin-Antitoxin system genes were identified alongside pan-genome analysis via Roary v3.11.2 within Galaxy. The annotated GFF files were used to screen out the toxin-antitoxin systems in the 16 selected genomes of aforementioned *Alistipes* species. The results were available within the same Roary output file of pan-genome analysis. Then, Microsoft Excel was used to sort and filter out the rows that contained the toxin-antitoxin system related genes. Next, this information was saved separately for identifying the corresponding operons for the toxin-antitoxin genes.

Moreover, the Prokka annotated GFF files were used to find the start and end position of the identified toxin-antitoxin genes. The operon mapper generated result file was used next to find the corresponding operons of the toxin-antitoxin genes by providing the positions. This step was repeated for rest of the genomes. Finally, the identified toxin-antitoxin genes from Roary and corresponding operons for the toxin-antitoxin genes from Operon Mapper was combined together for further analysis.

2.8 Identification of Secretion System related Genes and Operons

Similar to the Toxin-Antitoxin system genes, secretion system related genes were also identified alongside pan-genome analysis via Roary v3.11.2 within Galaxy. The annotated GFF files were used to screen out the toxin-antitoxin systems in the 16 selected genomes of aforementioned *Alistipes* species. The results were available within the same Roary output file of pan-genome analysis. Then, Microsoft Excel was used to sort and filter out the rows that contained the secretion system related genes. Next, this information was saved separately for identifying the corresponding operons for the secretion system genes.

Moreover, the Prokka annotated GFF files were used to find the start and end position of the identified secretion system genes. The operon mapper generated result file was used next to find the corresponding operons of the secretion system genes by providing the positions. This step was repeated for rest of the genomes. Finally, the identified secretion system genes from Roary and corresponding operons for the toxin-antitoxin genes from Operon Mapper was combined together for further analysis.

2.9 Annotation of Carbohydrate-Active enzyme

Carbohydrate active enzymes (CAZyme) are responsible for the metabolism of complex carbohydrates that are carbohydrates linked with other biopolymers such as protein or lipid; glycoprotein, glycolipid for example. The importance of CAZyme is observed in the biofuel industry, agriculture industry, human health, microbes et cetera. As microbes take complex carbohydrates as a food source, these are essential for their growth and other physiological importance. Henceforth, dbCAN was introduced for annotation of the carbohydrate active enzymes that are responsible for the synthesis, degradation and modification of these complex carbohydrates. The web server dbCAN2 is the updated version of dbCAN that is not only familiar with the updated database of CAZymes but also consist the hotpep search against the CAZyme peptide database along with previously introduced HMMER and DIAMOND search [52].

The updated dbCAN2 meta server is faster and more accurate than any other CAZyme annotation pipeline. Nevertheless, it can take both nucleotide and protein sequences as input and predict and return accurate results within a short time [52]. For the carbohydrate active enzyme annotation of the selected genomes of *Alistipes* species, the dbCAN2 meta-server was accessed online and one sequence in FASTA format was uploaded. Before submitting one job, the type of sequence was checked as complete nucleotide sequence and all 4 tools HMMER: dbCAN, DIAMOND: CAZy, HMMER: dbCAN-sub and CGCFinder was run with default parameters such as E-value and coverage cutoff. Here to mention, CGCFinder can be rerun with separate parameters after receiving results. However, this was not performed in this case. Similar process was followed for all the 16 genomes.

2.10 Prophage Screening

Presence of functional and non-functional bacteriophage genes in bacterial genomes as well as plasmids can be observed at least in >20% cases. This is related to emergence of antibiotic resistance as the presence of these bacteriophages urges bacteria to develop resistance to antibiotics. The tool PHAST was developed in order to screen bacteriophage gene clusters within bacterial genomes. PHASTER is the upgraded version of PHAST that is more accurate, fast and easier to maintain and perform query [53].

PHAge Search Tool- Enhanced Release, PHASTER is not only used for screening bacteriophage sequences within bacterial genomes and plasmids but also used for rapid identification of these phages and their annotation. PHASTER also provides genome visualization tools with an improved graphics interface that is helpful for better recognition of the prophage regions and later analysis. The webserver can run one sequence at a time; while the command line version can run multiple sequences at a time. Users can paste or upload the sequences in FASTA format or can provide GenBank accession and submit them for analysis in PHASTER with default parameters [53].

In this case, the PHASTER web server was accessed online and the selected genome sequences of *Alistipes* species in FASTA format was uploaded one by one and each job was submitted by unchecking “use pre-computed results” and checking “Remember my searches”. This provided an accurate result even though it took longer time than expected yet helped us to track the submission information and respective results easily. Later, the results were downloaded in available formats and screenshots of the visuals representations and figures were taken and saved for further analysis.

Chapter 3. Result and Interpretation

The results from different online based tools on the retrieved data from the database was analyzed both online and offline to find the significant results and opt out the insignificant outputs.

3.1 Retrieved Genomes

Among the 16 identified species of the genus *Alistipes*, only the ones that had complete or at least chromosome level genomes available on the National Center for Biotechnology Information (NCBI) genome database till March, 2022 were used for this analysis.

Assembly	GenBank	Scientific name	Modifier	Annotati...	Size (...	Level	Da	A...
ASM21057v1	GCA_000210575.1	Alistipes shahii WAL 8301	WAL 8301 (str...	NCBI RefS... Submitter	3.763	Chromosome	Ap	
ASM26536v1	GCA_000265365.1	Alistipes finegoldii DSM ...	DSM 17242 (st...	NCBI RefS... Submitter	3.734	Complete	Ju	
PRJEB28786	GCA_900604385.1	Alistipes megaguti	Marseille-P599...	NCBI RefS...	3.271	Complete	No	
Aond_1.0	GCA_006542645.1	Alistipes onderdonkii su...	3BBH6 (strain)	NCBI RefS... Submitter	3.507	Complete	Ju	
Acom_1.0	GCA_006542665.1	Alistipes communis	5CBH24 (strain)	NCBI RefS... Submitter	3.301	Complete	Ju	
Adis_1.0	GCA_006542685.1	Alistipes dispar	5CPEGH6 (stra...	NCBI RefS... Submitter	2.962	Complete	Ju	
Aond_2.0	GCA_006542705.1	Alistipes onderdonkii su...	5CPYCFAH4 (s...	NCBI RefS... Submitter	3.313	Complete	Ju	
Aond_3.0	GCA_006542725.1	Alistipes onderdonkii su...	5NYCFAH2 (str...	NCBI RefS... Submitter	3.313	Complete	Ju	
6CPBBH3	GCA_006542745.1	Alistipes communis	6CPBBH3 (stra...	NCBI RefS... Submitter	3.302	Complete	Ju	
ASM955745v1	GCA_009557455.1	Alistipes sp. dk3624	dk3624 (strain)	NCBI RefS... Submitter	3.020	Complete	De	
ASM1416349v1	GCA_014163495.1	Alistipes indistinctus	2BBH45 (strain)	NCBI RefS... Submitter	3.096	Complete	Se	
ASM2073572v1	GCA_020735725.1	Alistipes senegalensis	FDAARGOS_15...	NCBI RefS... Submitter	4.023	Complete	De	
ASM2284567v1	GCA_022845675.1	Alistipes onderdonkii	CE91-St18 (str...	NCBI RefS... Submitter	3.812	Complete	Ma	
ASM2284605v1	GCA_022846055.1	Alistipes finegoldii	CE91-St15 (str...	NCBI RefS... Submitter	4.117	Complete	Ma	
ASM2200991v1	GCA_022009915.1	Alistipes putredinis	nC33_bin.104.f...	Submitter	1.910	Complete	Ma	
min17_bin03	GCA_928852565.1	uncultured Alistipes sp.	min17_bin03 (i...		2.621	Chromosome	Ma	

Figure 19: Searching and filtering out the chromosome and complete level genomes of *Alistipes* species from the NCBI Genome database.

Next, the GenBank accession numbers (Figure 19) of these filtered genomes were collected and used at the Galaxy web platform for retrieving the genomic data in FASTA, CDS, GFF3, RNA, Protein file format. Then, the retrieved genomic data files were used for genome annotation. The list of genomes that were selected for this analysis have been presented below (Table 1).

Table 1: List of the selected genomes

SL No.	Genome name	Modifier	Level	GenBank	Accession
1.	*Alistipes shahii WAL 8301*	WAL 8301 (strain)	Chromosome	GCA_000210575.1	NC_021030.1
2.	*Alistipes finegoldii DSM 17242*	DSM 17242 (strain)	Complete	GCA_000265365.1	NC_018011.1
3.	*Alistipes megaguti*	Marseille-P5997 (strain)	Complete	GCA_900604385.1	NZ_LR027382.1
4.	*Alistipes onderdonkii subsp. vulgaris*	3BBH6 (strain)	Complete	GCA_006542645.1	NZ_AP019734.1
5.	*Alistipes communis*	5CBH24 (strain)	Complete	GCA_006542665.1	AP023049.1
6.	*Alistipes dispar*	5CPEGH6 (strain)	Complete	GCA_006542685.1	NZ_AP019736.1
7.	*Alistipes onderdonkii subsp. vulgaris*	5CPYCFAH4 (strain)	Complete	GCA_006542705.1	NZ_AP019737.1
8.	*Alistipes onderdonkii subsp. vulgaris*	5NYCFAH2 (strain)	Complete	GCA_006542725.1	NZ_AP019738.1
9.	*Alistipes communis*	6CPBBH3 (strain)	Complete	GCA_006542745.1	NZ_AP019739.1
10.	*Alistipes sp. dk3624*	dk3624 (strain)	Complete	GCA_009557455.1	NZ_CP045651.1

SL No.	Genome name	Modifier	Level	GenBank	Accession
11.	*Alistipes indistinctus*	2BBH45 (strain)	Complete	GCA_014163495.1	NZ_AP023049.1
12.	*Alistipes senegalensis*	FDAARGOS_1578 (strain)	Complete	GCA_020735725.1	NZ_CP085931.1
13.	*Alistipes onderdonkii*	CE91-St18 (strain)	Complete	GCA_022845675.1	NZ_AP025562.1
14.	*Alistipes finegoldii*	CE91-St15 (strain)	Complete	GCA_022846055.1	NZ_AP025581.1
15.	*Alistipes putredinis*	nC33_bin.104.fa (isolate)	Complete	GCA_022009915.1	CP091730.1
16.	*uncultured Alistipes sp.*	min17_bin03 (isolate)	Chromosome	GCA_928852565.1	OV789733.1

3.2 Annotated genomes

The retrieved genomic files, i.e., FASTA, CDS, GFF3, RNA, Protein of 14 complete genomes and 2 chromosome level genomes of different *Alistipes* species was next annotated using the Prokaryotic Genome Annotation System (Prokka), v1.14.5. Prokka provided 16 annotated result files in GFF3 format and annotated FASTA files of the genes and proteins for the 16 retrieved genomes.

Table 2: Prokka annotation summary for the selected 16 genomes

SL No.	Genome Name	GenBank	Prokka Annotation Summary						
			CDS	miscRNA	repeat RNA	rRNA	tmRNA	tRNA	Total genes
1.	*Alistipes shahii WAL 8301*	GCA_000210575.1	3049	23	1	3	1	47	3124
2.	*Alistipes finegoldii DSM 17242*	GCA_000265365.1	3187	23	N/A	6	1	50	3267

SL No.	Genome Name	GenBank	Prokka Annotation Summary						
			CDS	miscRNA	repeat RNA	rRNA	tmRNA	tRNA	Total genes
3.	*Alistipes megaguti*	GCA_900604385.1	2645	11	N/A	9	1	49	2715
4.	*Alistipes onderdonkii subsp. vulgaris*	GCA_006542645.1	2927	13	N/A	6	1	49	2996
5.	*Alistipes communis*	GCA_006542665.1	2675	17	N/A	6	1	52	2751
6.	*Alistipes dispar*	GCA_006542685.1	2436	13	N/A	6	1	48	2504
7.	*Alistipes onderdonkii subsp. vulgaris*	GCA_006542705.1	2781	19	N/A	6	1	49	2856
8.	*Alistipes onderdonkii subsp. vulgaris*	GCA_006542725.1	2785	19	N/A	6	1	49	2860
9.	*Alistipes communis*	GCA_006542745.1	2763	26	N/A	6	1	50	2846
10.	*Alistipes sp. dk3624*	GCA_009557455.1	2503	13	N/A	7	1	45	2569
11.	*Alistipes indistinctus*	GCA_014163495.1	2491	14	N/A	6	1	43	2555
12.	*Alistipes senegalensis*	GCA_020735725.1	3178	15	N/A	6	1	48	3248
13.	*Alistipes onderdonkii CE91-St18*	GCA_022845675.1	3088	18	N/A	6	1	45	3158
14.	*Alistipes finegoldii CE91-St15*	GCA_022846055.1	3540	22	N/A	6	1	49	3618
15.	*Alistipes putredinis*	GCA_022009915.1	2184	6	1	6	1	41	2239
16.	*uncultured Alistipes sp.*	GCA_928852565.1	2169	9	1	6	1	49	2235

From the summarized information (Table 2), it is evident that the number of transfer messenger RNA, tmRNA is same for all 16 genome that is 1 per genome whereas, repeat regions are only present in 3 genomes, *Alistipes shahii WAL 8301, Alistipes putredinis* and *uncultured Alistipes sp.* and among these 3 genomes only the genome of *Alistipes putredinis* is a complete genome while the other two are chromosome level genomes. Nevertheless, the presence of tRNA and mRNA, two types of non-coding RNA (ncRNA) in all the 16 genomes could be identified from this summarized information.

3.3 Pan-genome analysis

The GFF3 files of the 16 genomes of different *Alistipes* species that were generated via Prokka was used for pan-genome analysis via Roary v3.11.2 within Galaxy and upon changing some default values of the tool, for instance, changing the minimum percentage identity for blastp to 90% from the default value 95%. Next, results of the pan-genome analysis were collected in a "tsv" file, opened using Microsoft Excel and analyzed.

Table 3: Summarized Roary results.

SL No.	Presence of genes in the genomes	Number of genes	Type of gene
1.	>99%	14	Core
2.	95-99%	0	Soft core
3.	15-95%	3034	Shell

In the pan-genome analysis, 14 genes were found in all the 16 genomes, 9 genes were found in at least 15 of the 16 selected genomes. In this way, 122 genes were found in at least 12 or more genomes; that is 75% or more of the 16 genomes. Genes that are found in >99% genomes are considered to be the core genes; hence, the number of core genes are 14 in this case. The number

of soft-core genes is usually found within the range of 95-99% genomes. Since the number of genomes are low in this case, no soft-core genes could be identified within this range. Albeit, 122 genes were found in 12 or more genomes, 351 genes were present in at least 8 to 11 genomes, 2021 genes were found in at least 4 to 7 genomes. In total, 3034 shell genes were found within 15-95% match.

3.4 Identified Antibiotic Resistance Genes and Associated Operon

The online based tool, Resistance Gene Identifier (RGI) from the Comprehensive Antibiotic Resistance Database (CARD) was used to identify the possible antibiotic resistance genes from the 16 genomes of the genus *Alistipes.* Therefore, whole genome FASTA file of each genome was submitted by checking these options- Perfect, Strict, and Loose hits, Nudge excluded, High quality/Coverage and results were collected.

Table 4: Summary of CARD-RGI results

SL No.	Genome Name	GenBank	Number of Identified AMR genes
1.	*Alistipes shahii WAL 8301*	GCA_000210575.1	Strict hit: 2, Loose hit: 168; Total: 170
2.	*Alistipes finegoldii DSM 17242*	GCA_000265365.1	Strict hit: 2, Loose hit: 182; Total: 184
3.	*Alistipes finegoldii CE91-St15*	GCA_022846055.1	Strict hit: 3, Loose hit: 195; Total: 197
4.	*Alistipes onderdonkii subsp. vulgaris*	GCA_006542645.1	Strict hit: 1, Loose hit: 201; Total: 202
5.	*Alistipes onderdonkii subsp. vulgaris*	GCA_006542705.1	Strict hit: 1, Loose hit: 192; Total: 193
6.	*Alistipes onderdonkii subsp. vulgaris*	GCA_006542725.1	Strict hit: 1, Loose hit: 192; Total: 193
7.	*Alistipes onderdonkii CE91-St18*	GCA_022845675.1	Strict hit: 3, Loose hit: 213; Total: 216
8.	*Alistipes communis*	GCA_006542665.1	Strict hit: 4, Loose hit: 156; Total: 160
9.	*Alistipes communis*	GCA_006542745.1	Strict hit: 3, Loose hit: 157; Total: 160
10.	*Alistipes dispar*	GCA_006542685.1	Strict hit: 1. Loose hit: 163; Total: 164

SL No.	Genome Name	GenBank	Number of Identified AMR genes
11.	*Alistipes sp. dk3624*	GCA_009557455.1	Perfect: 1, Strict hit: 3, Loose hit: 211; Total: 215
12.	*Alistipes indistinctus*	GCA_014163495.1	Strict hit: 1, Loose hit: 188; Total: 189
13.	*Alistipes senegalensis*	GCA_020735725.1	Strict hit: 1, Loose hit: 201; Total: 202
14.	*Alistipes putredinis*	GCA_022009915.1	Loose hit: 121; Total: 121
15.	*Alistipes megaguti*	GCA_900604385.1	Strict hit: 1, Loose hit: 186; Total: 187
16.	*uncultured Alistipes sp.*	GCA_928852565.1	Strict hit: 1, Loose hit: 164; Total: 165

The result files were opened using Microsoft Excel and among the results, Perfect, Strict, and best of the Loose hits were chosen as probable AMR genes for each genome. In the summary (Table 4) of the predicted Antimicrobial Resistance (AMR) genes has been presented where it can be seen that the number of loose hit AMR genes dominate over the strict and perfect hit AMR genes.

Therefore, the best of the loose hit AMR genes was considered as significant AMR genes based on their percentage identity and best hit ARO score given by the tool. Total 105 AMR genes with perfect, strict and best of loose hits across the genomes were found after analyzing the result file. However, some genes could be found similar in multiple genomes, some genes could be found in all the genomes, some genes were identified as a variant of a same gene and so on. Hence, the total number of predicted AMR genes was 36, upon further analyzing the results.

The presence and absence of an AMR gene across different genomes has been depicted by the '+' sign and the '-' sign respectively in the summarized table (Table 5). From this table, it is evident that, three genes- 'Clostridioides difficile gyrB conferring resistance to fluoroquinolones', 'Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin' and 'Mycobacterium tuberculosis rpoB mutants conferring resistance to rifampicin' are present in all 16 genomes. Another AMR gene 'TaeA' that could be found in 15 genomes out of the 16 was absent only in the genome of *Alistipes putredinis* species.

Table 5: List of the predicted AMR genes across the genomes of *Alistipes* species.

Antibiotic resistance gene	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
adeF (Efflux pump membrane transporter BepE)	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+
Bifidobacterium adolescentis rpoB mutants conferring resistance to rifampicin	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Bifidobacterium bifidum ileS conferring resistance to mupirocin	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-
Capnocytophaga gingivalis gyrA conferring resistance to fluoroquinolones	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
ceoB	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
Clostridioides difficile gyrA conferring resistance to fluoroquinolones	+	+	-	+	+	+	+	+	+	+	+	+	+	-	+	+
Clostridioides difficile gyrB conferring resistance to fluoroquinolones	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
cmeB	+	-	-	+	+	+	+	-	-	-	-	+	+	-	+	-
ErmF	-	-	+	-	-	-	+	+	-	-	+	-	-	-	-	-

Antibiotic resistance gene	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Helicobacter pylori rpoB mutation conferring resistance to rifampicin	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	+
mdtC	-	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+
mefC	-	-	-	+	+	+	+	-	-	+	-	-	+	-	+	-
Mel	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
MexF	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-
MexK	+	+	+	+	+	+	+	-	-	+	+	+	+	-	+	-
MexW	+	+	+	+	+	+	+	-	-	+	+	-	+	+	+	+
msbA	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
MuxB	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mycobacterium tuberculosis rpoB mutants conferring resistance to rifampicin	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Mycobacterium tuberculosis thyA with mutation conferring resistance to para-aminosalicylic acid	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-
novA	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-

Antibiotic resistance gene	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
optrA	+	+	-	+	+	+	+	-	-	-	-	-	+	-	+	+
oqxB	-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-
qacG	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-
qacJ	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
rosA	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	-
rpoB2	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
Staphylococcus aureus fusA with mutation conferring resistance to fusidic acid	+	+	-	+	+	+	+	-	-	+	+	+	+	+	+	+
Staphylococcus aureus GlpT with mutation conferring resistance to fosfomycin	-	+	+	+	+	+	+	-	-	-	-	-	+	-	-	-
Staphylococcus aureus mupA conferring resistance to mupirocin	+	+	+	+	+	+	+	-	-	-	+	+	+	+	-	+
Staphylococcus aureus mupB conferring resistance to mupirocin	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
TaeA	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
tetQ	+	+	-	-	-	-	+	+	+	-	+	+	-	-	-	-
TriC	-	-	-	-	-	-	-	+	+	-	+	-	-	+	+	+

Antibiotic resistance gene	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
ugd	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	+

Furthermore, 'Bifidobacterium adolescentis rpoB mutants conferring resistance to rifampicin', 'Capnocytophaga gingivalis gyrA conferring resistance to fluoroquinolones', 'ceoB', 'Mel', 'msbA', 'MuxB', 'novA' and 'rpoB2'- these 8 genes could only be found in the corresponding single genome from the 16 genomes. Moreover, two of these eight genes belong to same family- rpoB (Table 5). Further, 'Bifidobacterium bifidum ileS conferring resistance to mupirocin', 'MexF', 'qacG', 'qacJ', and 'Staphylococcus aureus mupB conferring resistance to mupirocin'- these 5 genes could be found in only two respective genomes from the 16 genomes.

Therefore, there are total 13 genes that could be found in two or less than two genomes. Hence, the number of genes that could be found in more than two genomes is 23. Some of these genes could be found in multiple genomes of the same species and some genes could not be found in the other genomes of the same species. From the 16 genomes, there are only three species that have multiple genomes. *Alistipes finegoldii* and *Alistipes communis* have two genomes and *Alistipes onderdonkii* have four genomes.

AMR genes in multiple genomes of same species:

The two genomes of the species *Alistipes finegoldii* are *Alistipes finegoldii DSM 17342* and *Alistipes finegoldii CE91-St15*. Among these two genomes total 12 AMR genes could be found similar, that means these 12 genes could be found in both genomes of *A. finegoldii*. However, the number of AMR genes present in the *DSM 17342* genome is 16 and 17 AMR genes are present in the *CE91-St15* genome. The gene 'adeF' is the common strict hit in both of these genomes of *Alistipes finegoldii* whereas the rest of the genes are resultants of loose hits.

From the CARD-RGI results (Table 5), it can be observed that, 'adeF', 'Clostridioides difficile gyrB conferring resistance to fluoroquinolones', 'Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin', 'Helicobacter pylori rpoB mutation conferring resistance to rifampicin', 'mdtC', 'MexK', 'MexW', 'Mycobacterium tuberculosis rpoB mutants conferring resistance to rifampicin', 'oqxB', 'Staphylococcus aureus GlpT with mutation conferring resistance to fosfomycin', 'Staphylococcus aureus mupA conferring resistance to mupirocin' and 'TaeA' are the 12 AMR genes that are similar in both of the *A. finegoldii* species.

The predicted AMR genes that are present in *A. finegoldii DSM CE91-St15* genome but absent in the other *A. finegoldii* genome are- 'Capnocytophaga gingivalis gyrA conferring resistance to fluoroquinolones', 'ceoB', 'ErmF', 'qacG' and 'rosA' while 'Clostridioides difficile gyrA conferring resistance to fluoroquinolones', 'optrA', 'Staphylococcus aureus fusA with mutation conferring resistance to fusidic acid' and 'tetQ' are the predicted AMR genes that are present in the *DSM 17342* genome and absent in the other genome of *Alistipes finegoldii* (Table 5).

Similarly, 14 AMR genes could be found similar in the two *Alistipes communis* genomes- *Alistipes communis 5CBH24* and *Alistipes communis 6CPBBH3*. The tool predicted 15 possible AMR genes in the *5CBH24* genome and 14 AMR genes in case of the *6CPBBH3* genome. The only AMR gene that has been predicted in case of the genome *5CBH24* is also the gene that is absent in case of *6CPBBH3* genome and the predicted AMR gene is 'ErmF'. Except this gene the rest AMR genes are common in these two genomes (Table 5).

Next *Alistipes* species that have four genomes is *Alistipes onderdonkii* and the four genomes are *A. onderdonkii 3BBH6, A. onderdonkii 5CPYCFAH4, A. onderdonkii 5NYCFAH2* and *A. onderdonkii CE91-St18*. Individually, *A. onderdonkii 3BBH6* genome has 18 predicted AMR genes, both *A. onderdonkii 5CPYCFAH4* and *A. onderdonkii 5NYCFAH2* genome have 17 predicted AMR genes, and the genome *A. onderdonkii CE91-St18* has 19 predicted AMR genes. The 17 predicted AMR genes are similar in case of both *A. onderdonkii 5CPYCFAH4* and *A. onderdonkii 5NYCFAH2* genomes while these are the common AMR genes among these four genomes (Table 5).

A. onderdonkii 3BBH6 genome has the predicted AMR 'msbA' gene that is absent in the other three genomes of *A. onderdonkii* species and *A. onderdonkii CE91-St18* genome have 'ErmF' and

‘tetQ’ predicted AMR genes that are absent in the other three *A. onderdonkii* species; alongside the aforementioned 17 common AMR genes across different genomes of *Alistipes onderdonkii* species. Nevertheless, the AMR gene ‘adeF’ is the common strict hit resultant AMR gene for these genomes and only strict hit for *3BBH6, 5CPYCFAH4, 5NYCFAH2* genomes whereas the AMR genes ‘tetQ’ and ‘ErmF’ are also strict hit predicted AMR genes for the *CE91-St18* genome.

AMR genes in multiple genomes of different species:

So, after the same AMR gene in multiple genomes of the same species have been mentioned; however, there are some AMR genes that are similar in multiple genomes of different *Alistipes* species too. Also, some AMR genes that could be found similar in multiple genomes of the same species was found absent in other genomes of different *Alistipes* species. For example, ‘adeF’ and another AMR gene ‘Helicobacter pylori rpoB mutation conferring resistance to rifampicin’ is present in almost 14 genomes out of 16. These two AMR genes are present in multiple genomes of same species; however, absent in the genomes of *Alistipes indistinctus* and *Alistipes putredinis* (Table 5).

Similarly, another AMR gene ‘Clostridioides difficile gyrA conferring resistance to fluoroquinolones’ could be found in 14 genomes out of 16 and it was found present in one *Alistipes finegoldii* genome and absent in the other. This gene was found absent in the *Alistipes putredinis* genome as well. Moreover, the AMR gene ‘cmeB’ that could be found in all four genomes of *Alistipes onderdonkii* species was found absent in both genomes of *Alistipes finegoldii* and both genomes of *Alistipes communis* as well. However, this gene could be found in the single genomes of *Alistipes shahi, Alistipes indistinctus, Alistipes senegalensis* and *Alistipes megaguti* species too (Table 5).

Next, an AMR gene, ‘tetQ’ was not found in all the selected genomes of *Alistipes* species from the CARD-RGI results. It could be found in both genomes of *Alistipes communis,* absent in three of four genomes of *Alistipes onderdonkii* and one of the *Alistipes finegoldii* species and present in the *Alistipes finegoldii DSM 17342* and *Alistipes onderdonkii CE91-St18* genome as well as in the genome of *Alistipes shahi, Alistipes sp. dk3624* and *Alistipes indistinctus* species.

From CARD-RGI results (Table 5), it can be observed that the presence of AMR genes across difference genomes are quite notable and the lowest number of predicted AMR gene is 12 in case of the *Alistipes putredinis* genome and the highest number of predicted AMR gene is 19 in the *Alistipes onderdonkii CE91-St18* genome. 13 predicted AMR genes in the genome of *Alistipes indistinctus* species is the second lowest value of predicted AMR genes across the 16 genomes of *Alistipes* species and 14 AMR genes have been predicted by CARD-RGI in the *A. communis 6CPBBH3* and *uncultured Alistipes sp. isolate min17_bin03* genomes while 15 AMR genes were predicted each for the *A. communis 5CBH24* and *A. dispar 5CPEGH6* genomes. *Alistipes shahi WAL 8301* and *Alistipes finegoldii DSM 17342* genomes have 16 predicted AMR genes each whereas the *Alistipes finegoldii CE91-St15, A. onderdonkii 5CPYCFAH4* and *A. onderdonkii 5NYCFAH2* genomes have 17 predicted AMR genes each. Lastly, *A. onderdonkii 3BBH6* genome have 18 AMR genes, the second highest number of predicted AMR genes among the 16 genomes.

Therefore, the *Alistipes onderdonkii* species is the most antibiotic resistant species of among the other *Alistipes* species that have been mentioned in this project based on the number of predicted AMR genes. Similarly, it can be said that the *Alistipes putredinis* species is the least antibiotic resistant species among the others. The other species can be considered as moderate antibiotic resistant species of the genus *Alistipes* based on the same criteria.

Operon partners for the AMR genes:

Operons for these identified resistance genes were compared across genomes, and clusters were identified from the Operon Mapper results that were identified earlier. First, the AMR genes that could be found in all the 16 genomes and their operon partners were checked. For the gene 'Clostridioides difficile gyrB conferring resistance to fluoroquinolones' single gene operon has been predicted by Operon Mapper in the *Alistipes shahi WAL 8301* and *A. dispar 5CPEGH6* genomes and multiple gene operons in the other genomes. In the *Alistipes sp. dk3624* genome, the other gene in the operon along with this AMR gene is 'Putative protein YqeY' and in the other genomes have the gene 'Uridine kinase' in the operon instead of YqeY except the genome of *Alistipes putredinis* contain- 'DNA topoisomerase 4 subunit B', 'Cell division ATP-binding protein FtsE', 'Vitamin B12 import ATP-binding protein BtuD', two hypothetical proteins along with 'Uridine kinase' in the gyrB AMR gene operon.

Next, operon partners for the 'Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin' AMR gene was checked across the genomes and it was found the gene clusters were similar for all the genomes except for *Alistipes shahi* and *Alistipes putredinis* species. In case of *A. shahi* genome, it was a single gene operon. However, at least one gene from the gene cluster from the *Alistipes putredinis* genome was similar to the others. The common gene in all the genomes from this AMR gene operon is 'tRNA-Trp(cca)' and except *A. putredinis* all the multiple gene operons have Protein translocase subunit SecE, Transcription termination/ antitermination protein NusG, 50S ribosomal protein L11, 50S ribosomal protein L1, 50S ribosomal protein L10, 50S ribosomal protein L7/L12, DNA-directed RNA polymerase subunit beta, DNA-directed RNA polymerase subunit beta'- these genes in the gene cluster.

Interestingly, the operon for the AMR gene 'Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin' is similar for another AMR gene that is present in all the 16 genomes- 'Mycobacterium tuberculosis rpoB mutants conferring resistance to rifampicin'. Unlike the EF-Tu operon, *Alistipes shahi* genome have a multiple gene operon in this case. Moreover, a few genes from the operon of each genome for this AMR gene is common in all the genomes of *Alistipes* species. The operon for 'EF-Tu' and 'rpoB' gene could be found exactly similar in the following genomes- *Alistipes senegalensis, Alistipes dispar, Alistipes megaguti, Alistipes indistinctus, uncultured Alistipes sp. isolate min17_bin03,* both genomes of *Alistipes finegoldii* and *Alistipes communis,* all four genomes of *Alistipes onderdonkii* species. That means these genomes have similar type operon for these two AMR gene operons and the *Alistipes shahi, Alistipes putredinis* and *Alistipes sp. dk3624* genomes have another type operon.

Further, the AMR genes 'ErmF', 'mefC', 'mel' and 'qacJ' have single gene operons in the genomes these were predicted in. Another AMR gene 'Staphylococcus aureus fusA with mutation conferring resistance to fusidic acid' has single gene operon in most of the genomes it was identified; however, in the genomes of *Alistipes sp. dk3624, A. indistinctus 2BBH45,* and *Alistipes putredinis* this AMR gene was predicted in multiple gene cluster.

On the contrary, the AMR gene 'TaeA' have both single gene operon and multiple gene operon in multiple genomes it was identified. However, the pattern of gene clusters in multiple genomes of same species were found similar. For example, the four genomes of *A, onderdonkii* have both

single gene operon and multiple gene operon for this AMR gene and the multiple gene operon consisted with two hypothetical proteins, Histidine--tRNA ligase and the AMR gene itself within the operon for all the four genomes. This was found common in the genomes of *Alistipes finegoldii, Alistipes communis, Alistipes dispar, Alistipes senegalensis* species and the uncultured Alistipes sp. isolate min17_bin03 genome. The same AMR gene was predicted in two operons from a single genome of the *Alistipes indistinctus* species and one of the operons was single gene operon and the other operon consisted a few indifferent genes from the other genomes as mentioned earlier.

The CARD-RGI predicted AMR gene operons were mostly found similar across different genomes of *Alistipes* species. In most of the cases, AMR gene operons consisted similar genes in all respective genomes. The AMR genes that were present in the genomes of *Alistipes finegoldii, Alistipes communis* and *Alistipes onderdonkii*, they have the same set of genes in their operons. Operons and genes within the operons were similar for the multiple genomes of these species. For example, the set of genes that could be found in the AMR gene operon of *Alistipes finegoldii DSM 17342* genome, was found similar in the operon of the same gene of *Alistipes finegoldii CE91-St15* genome.

3.5 Identified Virulence Factors

The tool VFanalyzer compared the given FASTA sequences of the 16 genomes of *Alistipes* species with the virulence factor database (VFDB) and could identify 13 possible virulence factors from the genomes. Among the 13 identified virulence factors, 6 virulence factors genes were found within all the 16 genomes of *Alistipes* species and these genes are- clpB, rfbB, rfbC, tufA, kdsA, and gapA (Table 6).

The presence of virulence factor genes was found similar in both of the genomes of *Alistipes finegoldii- A. finegoldii DSM 17342* and *A. finegoldii CE91-St15*. Total 9 virulence factor genes could be found common between these two genomes- clpB, rffH, rfbB, rfbC, arnA, tufA, wbpA, kdsA, and gapA. Albeit, another virulence factor gene was identified in the *A. finegoldii CE91-St15* genome, wbgU and this gene was found to be absent in the other genome of *A. finegoldii* (Table 6).

Table 6: Virulence factors identified by VFanalyzer. Here, the '+' sign represents the presence of the virulence factor gene and the '-' sign represents the absence of the virulence factor gene in the respective genomes.

Virulence Factors and Related Genes	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
clpB: Chaperone protein ClpB	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rffH: Glucose-1-phosphate thymidylyltransferase 2	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
rfbB: dTDP-glucose 4,6-dehydratase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rfbC: dTDP-4-dehydrorhamnose 3,5-epimerase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
arnA: Bifunctional polymyxin resistance protein ArnA	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
tufa: Elongation factor Tu	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
wbpA: UDP-N-acetyl-D-glucosamine 6-dehydrogenase	+	+	+	-	-	-	-	+	+	-	-	+	-	-	-	-
rfbA: Glucose-1-phosphate thymidylyltransferase 1	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-

Virulence Factors and Related Genes	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
kdsA: 2-dehydro-3-deoxyphosphooctonate aldolase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rfbF: Glucose-1-phosphate cytidylyltransferase	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
wbgU: UDP-N-acetylglucosamine 4-epimerase	+	-	+	-	-	-	-	+	+	+	-	-	-	+	-	+
katA: Catalase	+	-	-	+	+	+	+	+	+	-	+	+	+	+	-	-
gapA: Glyceraldehyde-3-phosphate dehydrogenase A	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Similarly, the identified virulence factors in the four genomes of *Alistipes onderdonkii- A. onderdonkii 3BBH6, A. onderdonkii 5CPYCFAH4, A. onderdonkii 5CPYCFAH2* and *A. onderdonkii CE91-St18* were checked and all four of the *A. onderdonkii* genomes was found to have the similar virulence factor genes- clpB, rffH, rfbB, rfbC, arnA, tufA, kdsA, katA and gapA (Table 6). Moreover, the two genomes of *Alistipes communis- A. communis 5CBH24* and *A. communis 6CPBBH3* were also found to have the similar virulence factor genes among themselves. And the identified virulence factor genes are- clpB, rffH, rfbB, rfbC, tufA, wbpA, kdsA, wbgU, katA and gapA. The same virulence factors that were identified in the two genomes of *A. communis* could be found similar in the genome of *A. indistinctus* as well (Table 6).

In the same manner, the virulence factor genes could be found similar in the case of two different genomes of two different species, *A. finegoldii* and *A. putredinis*. The only genome of *A. putredinis* have the similar virulence factor gene pattern like the *A. finegoldii CE91-St15* genome (Table 6). Nevertheless, all of the identified virulence factor genes could be found in the genome of the *Alistipes shahi* species. The virulence factor gene 'rfbF' could be only found in the genome of *Alistipes shahi* and another virulence factor gene 'rfbA' could be found in the *Alistipes shahi* and *Alistipes sp. Dk3624* genomes.

Phylogenetic analysis:

Furthermore, phylogenetic trees were generated for the virulence factor genes that were found in more than 3 genomes and the genes are- 'clpB: Chaperone protein ClpB', 'rffH: Glucose-1-phosphate thymidylyltransferase 2', 'rfbB: dTDP-glucose 4,6-dehydratase', 'rfbC: dTDP-4-dehydrorhamnose 3,5-epimerase', 'arnA: Bifunctional polymyxin resistance protein ArnA', 'tufa: Elongation factor Tu', 'wbpA: UDP-N-acetyl-D-glucosamine 6-dehydrogenase', 'kdsA: 2-dehydro-3-deoxyphosphooctonate aldolase', 'wbgU: UDP-N-acetylglucosamine 4-epimerase', 'katA: Catalase' and 'gapA: Glyceraldehyde-3-phosphate dehydrogenase A'.

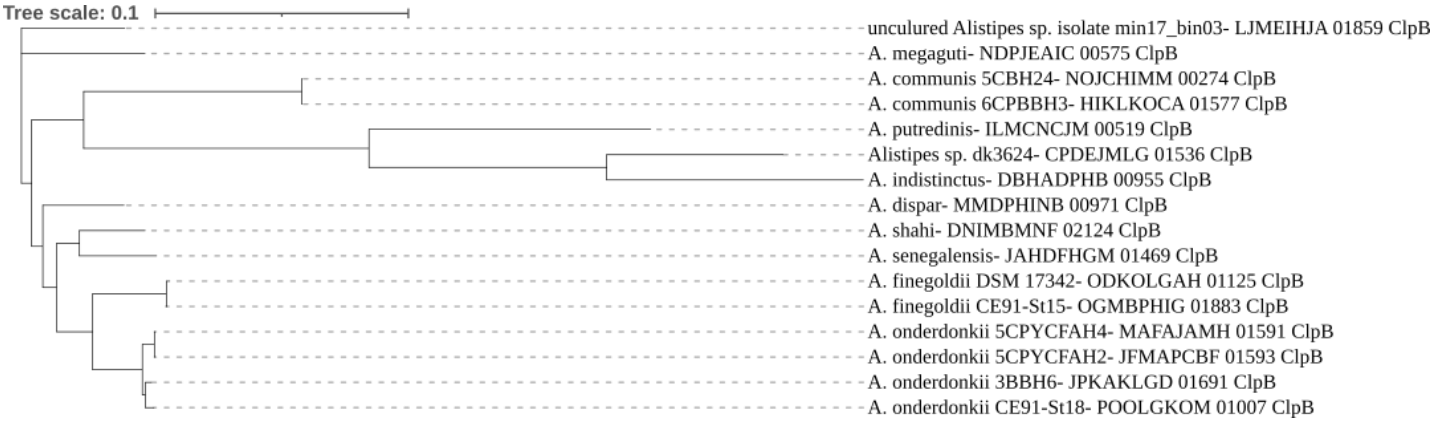

Figure 20: Phylogenetic tree of the virulence factor gene 'clpB'

The virulence factor gene, clpB could be found in all the 16 genomes and it appeared to have only one version. From the phylogenetic tree (Figure 20), it can be observed the clpB gene in the genomes of *Alistipes onderdonkii* species and *Alistipes finegoldii* species are closer to each other.

While the same gene in the multiple genomes of *Alistipes communis* are closer to *Alistipes megaguti* and *Alistipes putredinis*.

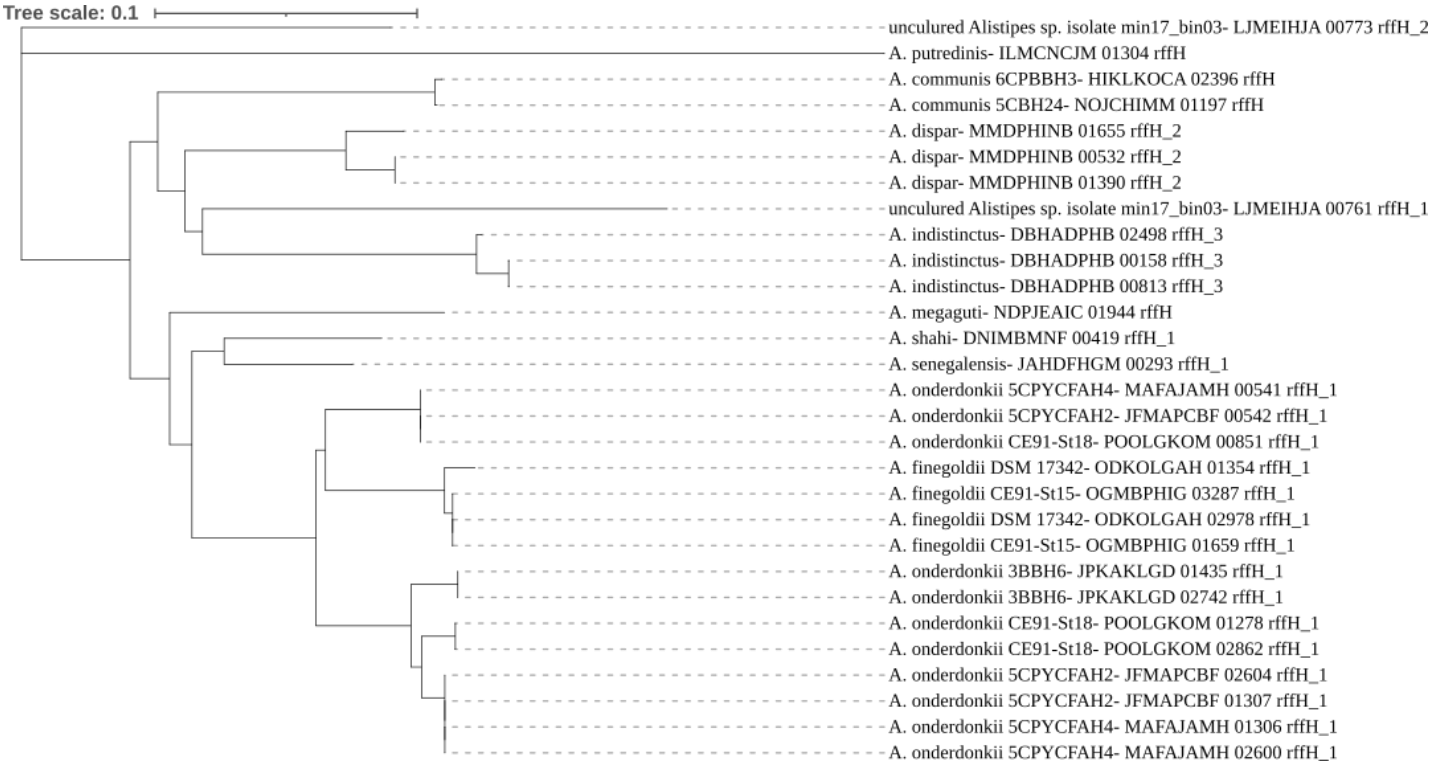

Figure 21: Phylogenetic tree of the virulence factor gene 'rffH'

Unlike the 'clpB' gene, four versions of another virulence factor gene, 'rffH' could be found- rffH, rffH_1, rffH_2, and rffH_3 where the rffH version could be found in the genomes of *Alistipes communis* and the single genome of *Alistipes putredinis* and *Alistipes megaguti*. This version of the rffH gene that was identified in *Alistipes megaguti* genome was found closer to the rffH_3 identified in *Alistipes indistinctus* and rffH_1 identified in *Alistipes shahi* genome. While the same rffH version that was identified in the other three genomes appeared closer in the tree (Figure 21).

The rffH_2 was identified in the *Alistipes dispar* genome and in the *uncultured Alistipes sp. isolate min17_bin03* genome and the rffH_1 version was identified in rest of the genomes appearing in the tree (Figure 21) except *Alistipes indistinctus* genome, rffH_3 version was identified in *Alistipes indistinctus*. However, both rffH_1 and rffH_2 versions of the rffH gene were identified in the *uncultured Alistipes sp. isolate min17_bin03* genome and the rffH_1 from this genome appeared closer to rffH_3 of *A. indistinctus* and rffH_2 identified in the *Alistipes dispar* genome (Figure 21).

While the rffH_2 from the *uncultured Alistipes sp.* genome was found closer to rffH identified in the *Alistipes putredinis* genome. In addition to that, the rffH_2 version identified in *A. dispar* was found closer to the rffH gene identified in the *Alistipes communis* genomes in the tree (Figure 21). The rffH_1 identified in the genomes of *Alistipes onderdonkii* and *Alistipes finegoldii* genomes appeared closer to each other in the phylogenetic tree. However, the two copies of the rffH_1 identified in *Alistipes onderdonkii CE91-St18* genome was found distant to each other yet closer to other genomes of *Alistipes onderdonkii* species.

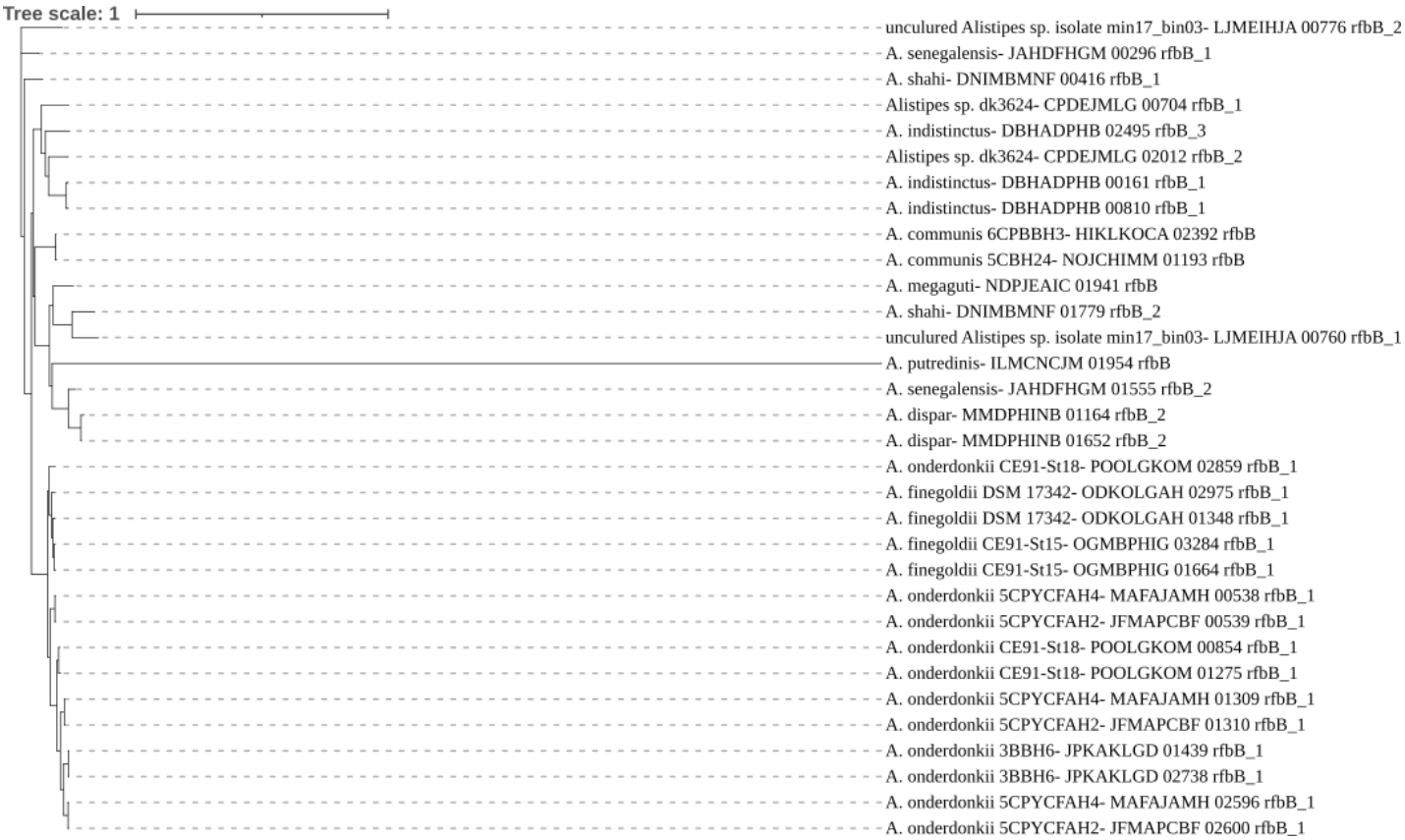

Figure 22: Phylogenetic tree of the virulence factor gene ‘rfbB’

Similar to the virulence factor gene ‘rffH’, ‘rfbB’ also have four versions of the rfbB gene- rfbB, rfbB_1, rfbB_2, and rfbB_3 (Figure 22). The rfbB version identified in the genomes of *Alistipes communis, Alistipes megaguti* were found closer to each other while another copy of the same version of this gene identified in *Alistipes putredinis* was found distant to these genomes and closer to the rfbB_2 version identified in the *Alistipes shahi* genome (Figure 22). Two different versions of the same gene- rfbB_1 and rfbB_2 was identified in the *uncultured Alistipes sp. isolate*

min17_bin03 genome and these two versions were found distant to each other. Moreover, same version of the same gene- rfbB_1 was identified twice in the *Alistipes onderdonkii CE91-St18* genome and one copy of rfbB_1 was found closer to the *A. finegoldii* genomes and the other was found closer to other genomes of *A. onderdonkii* species.

Furthermore, the virulence factor gene 'rfbC' also have four versions- rfbC, rfbC_1, rfbC_2, and rfbC_3 similar to 'rffH' and 'rfbB' genes. Here, the rfbC versions identified in *uncultured Alistipes sp. isolate min17_bin03, A. putredinis, A. megaguti, Alistipes sp. dk3624* and *A. communis* genomes and same version was found closer in the *A. megaguti* and *Alistipes sp. dk3624* genomes whereas the rfbC identified in the *uncultured Alistipes sp.* genome and *A. putredinis* genome was found distant to each other as well to these two genomes. The rfbC version identified in both genomes of *Alistipes communis* species was found quite distant to all these genomes identifying the same version of the virulence factor gene (Figure 23).

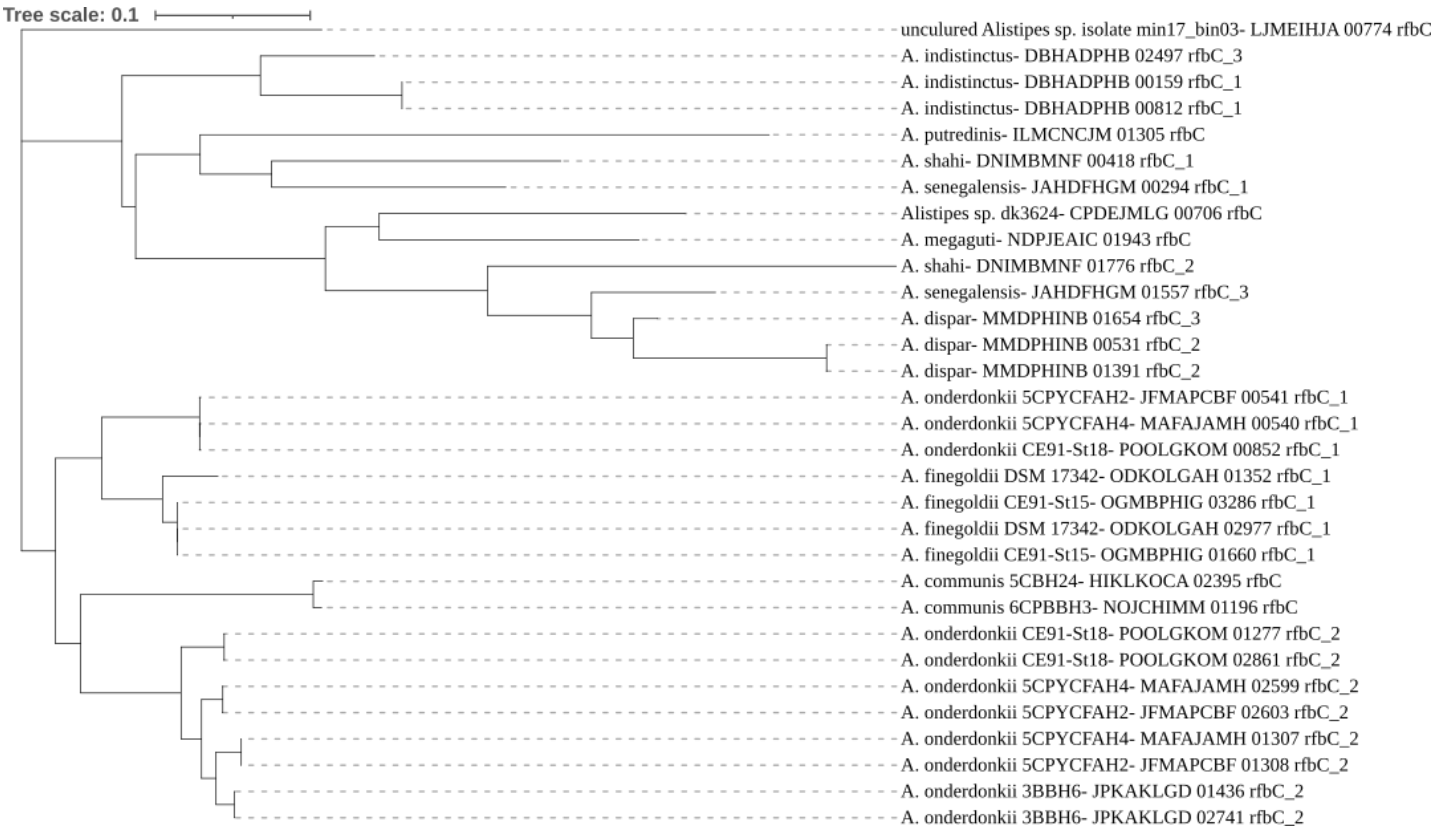

Figure 23: Phylogenetic tree of the virulence factor gene 'rfbC'

The other version rfbC_3 identified in *Alistipes indistinctus* genome and in the genome of *Alistipes senegalensis* was found distant to each other in the phylogenetic tree (Figure 23) but the same version identified in the *Alistipes dispar* genome was found closer to the genome of *Alistipes senegalensis*. Further, the rfbC_1 version identified in the multiple genomes of *A. finegoldii* and *A. onderdonkii* species were found closer to each other. The other version- rfbC_2 identified in these genomes was found closer to rfbC identified in the genomes of *Alistipes communis* species. However, the rfbC_1 version was not identified in the *A. onderdonkii 3BBH6* genome albeit two copies of the rfbC_2 versions of the same gene were identified in this genome and both copies were found closer to each other as well as the rest of the *A. onderdonkii* genomes (Figure 23).

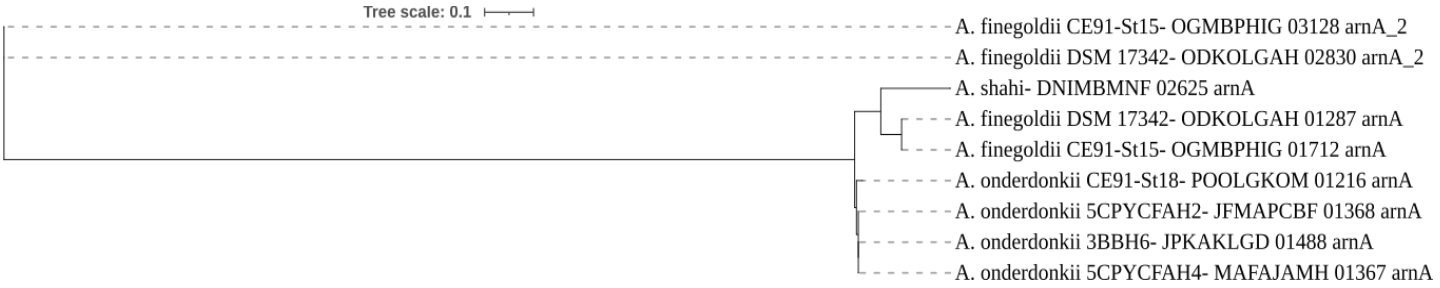

Figure 24: Phylogenetic tree of the virulence factor gene ‘arnA’

Another virulence factor gene ‘arnA’ was found in all of the genomes of *A. finegoldii* and *A. onderdonkii* species and in the genome of *Alistipes shahi*. Two copies of arnA_2 identified only in the genomes of *A. finegoldii* were found distant to the other version of the same gene, arnA in the same genomes (Figure 24). The arnA gene in the *Alistipes shahi* genome was found closer to the *Alistipes finegoldii DSM 17342* genome while the other genomes having the arnA gene were found close to each other.

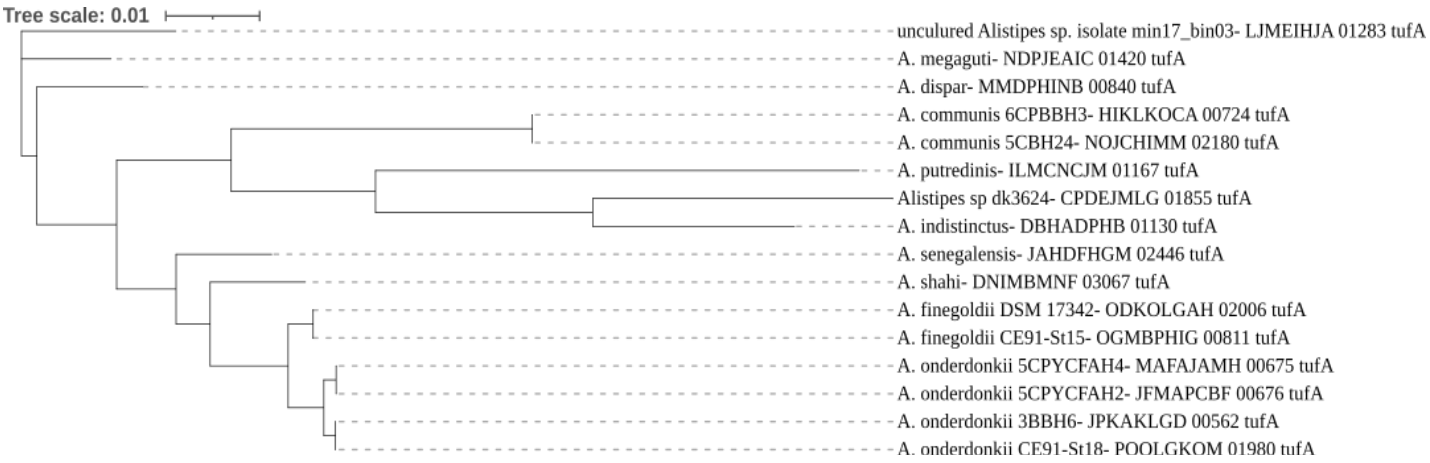

Figure 25: Phylogenetic tree of the virulence factor gene 'tufA'

In case of another virulence factor gene 'tufA' that was identified in all the 16 genomes, the genomes of *A. finegoldii* species and *A. onderdonkii* were closer to each other (Figure 25). Same scenario could be observed in the phylogenetic tree (Figure 26) of the virulence factor gene 'gapA' that was also identified in all the 16 genomes of the *Alistipes* species.

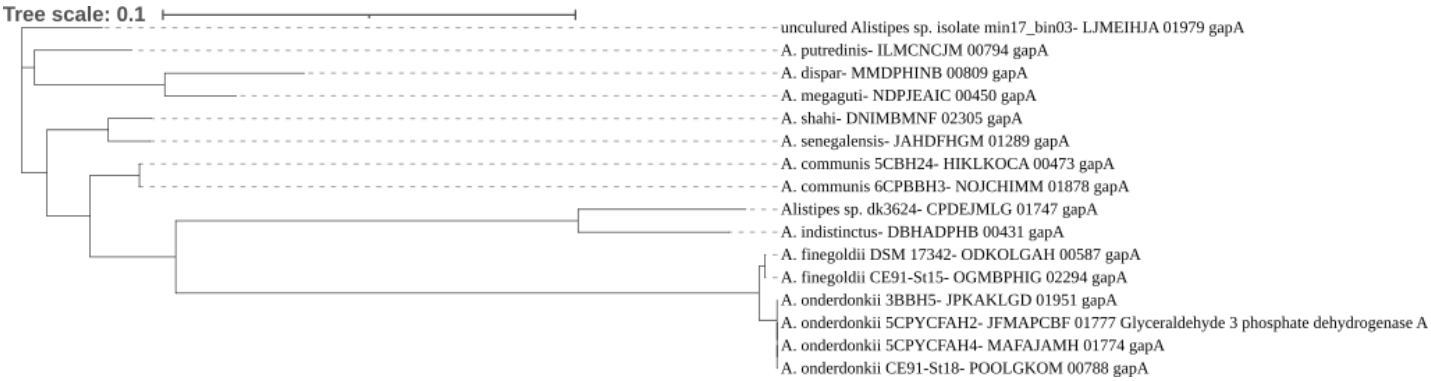

Figure 26: Phylogenetic tree of the virulence factor gene 'gapA'

The virulence factor genes 'wbpA' and 'wbgU' have been identified in considerably lower number of genomes and the wbpA gene identified in the *A. shahi* and *A. communis* species genomes were found closer in the phylogenetic tree (Figure 27). In addition to that. the wbgU gene in the *A. shahi* and *A. communis* genomes were found closer in the phylogenetic tree (Figure 28) similar to the 'wbpA' gene. That is the cluster was found similar for both of the genes in the aforementioned genomes of different species (Figure 27, 28).

Figure 27: Phylogenetic tree of the virulence factor gene ‘wbpA’

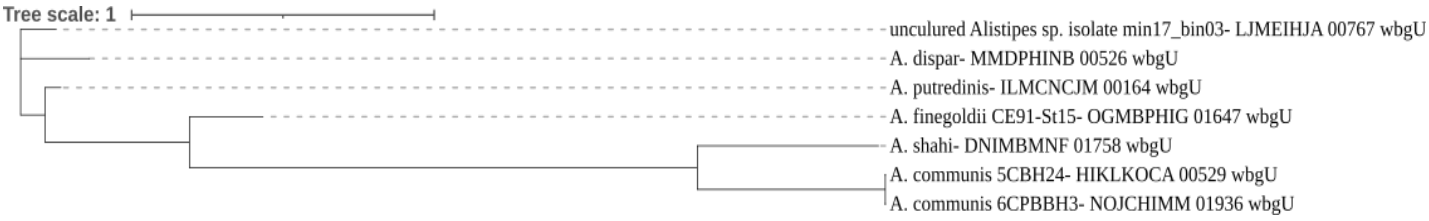

Figure 28: Phylogenetic tree of the virulence factor gene ‘wbgU’

Next, the virulence factor gene ‘kdsA’ have two versions kdsA_1 and kdsA_2 and these two versions were found in the 16 genomes of *Alistipes* species. In the phylogenetic tree (Figure 29), the gene kdsA_2 identified in the multiple genomes of *A. finegoldii* and *A. onderdonkii* species were found closer to each other and the same version identified in the *uncultured Alistipes sp. isolate* genome was found in the similar cluster. Another copy of the kdsA_2 gene identified in the *uncultured Alistipes sp. isolate* genome was found close to the kdsA_1 gene cluster of the genomes of the *A. finegoldii* and *A. onderdonkii* species.

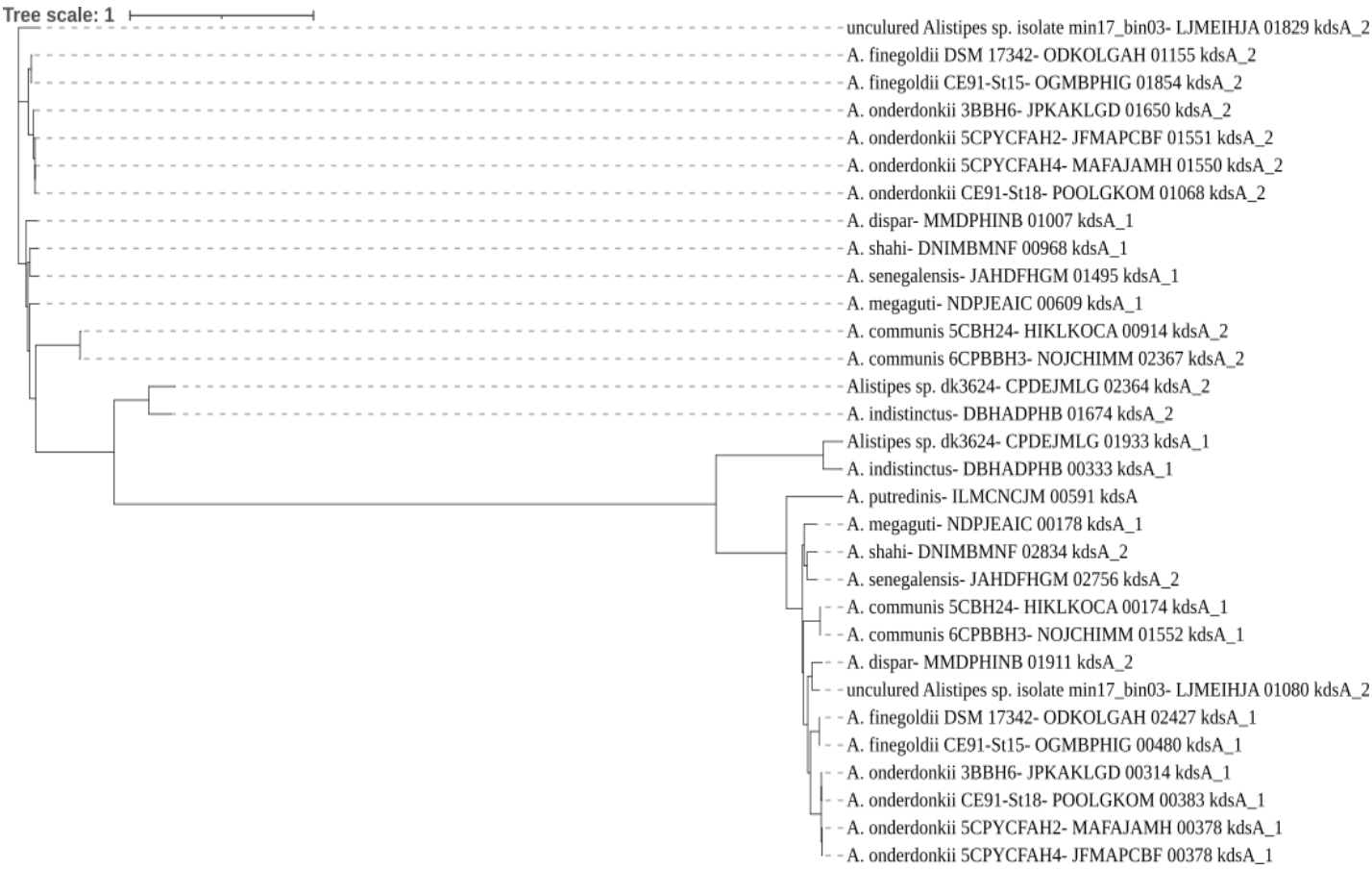

Figure 29: Phylogenetic tree of the virulence factor gene ‘kdsA’

Figure 30: Phylogenetic tree of the virulence factor gene ‘katA’

The virulence factor gene ‘katA’ was not identified in the *A. finegoldii* genomes. However, it was identified in the multiple genomes of the *A. onderdonkii* and *A. communis* species. In the phylogenetic tree (Figure 30), the katA gene identified in the multiple genomes of the same species appeared closer to each other.

3.6 Identified Toxin-Antitoxin System Genes and related Operons

Toxin-Antitoxin system genes were identified alongside pan-genome analysis via Roary v3.11.2 within Galaxy and 9 such genes could be identified from the 16 genomes of different *Alistipes* species. Toxin-Antitoxin genes were absent in the four genomes of *Alistipes onderdonkii- A. onderdonkii 3BBH6, A. onderdonkii 5CPYCFAH4, A. onderdonkii 5CPYCFAH2, A. onderdonkii CE91-St18* and the single genome of *Alistipes megaguti* and the 'uncultured *Alistipes* sp. Isolate min17_bin03' genome (Table 7).

Table 7: Identified Toxin-Antitoxin genes across the 16 genomes of *Alistipes* species. Here, the '+' sign represents the presence and the '-' sign represents the absence of the toxin-antitoxin gene in the respective genomes.

Toxin-Antitoxin genes	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
parD1: Antitoxin ParD1	+	+	+	-	-	-	-	-	+	-	-	+	-	-	-	-
socA: Antitoxin SocA	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
yefM: Antitoxin YefM	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
hipA: Serine/threonine-protein kinase toxin HipA	+	+	-	-	-	-	-	+	+	-	-	-	+	-	-	-

Toxin-Antitoxin genes	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
tdeA: Toxin and drug export protein A	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
parE1: Toxin ParE1	+	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-
apxIB: Toxin RTX-I translocation ATP-binding protein	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-
yoeB: Toxin YoeB	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
tabA: Toxin-antitoxin biofilm protein TabA	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-

Further, in the *Alistipes finegoldii DSM 17342* genome total 4 toxin-antitoxin genes and in the *Alistipes finegoldii CE91-St15* genome 2 such genes could be found. And only the gene 'parD1' could be found similar in both of the *Alistipes finegoldii* genomes. Similarly, the gene 'hipA' could be found in common in both of the *Alistipes communis* species where 2 toxin-antitoxin genes could be found in the *A. communis 5CBH24* genome and 3 toxin-antitoxin genes could be found in the *A. communis 6CPBBH3* genome respectively (Table 7).

Phylogenetic trees were generated for the 'parD1' (Figure 31) and 'hipA' (Figure 32) toxin-antitoxin genes. The antitoxin gene parD1 identified in the *A. communis 6CPBBH3* genome also in the multiple genomes of *Alistipes indistinctus* were found closer to each other (Figure 31). In

the tree of the hipA gene, the gene identified in the *A. communis 6CPBBH3* genome twice and both copies of the hipA gene was found distant to each other while the hipA gene identified in the other *A. communis 5CBH24* genome was found closer to the other genome of the same species (Figure 32). Moreover, two copies of the hipA gene identified in the *A. shahi* genome and one copy was found closer to the *A. communis 6CPBBH3* genome while the other one was found closer to the *A. communis 5CBH24* genome (Figure 32).

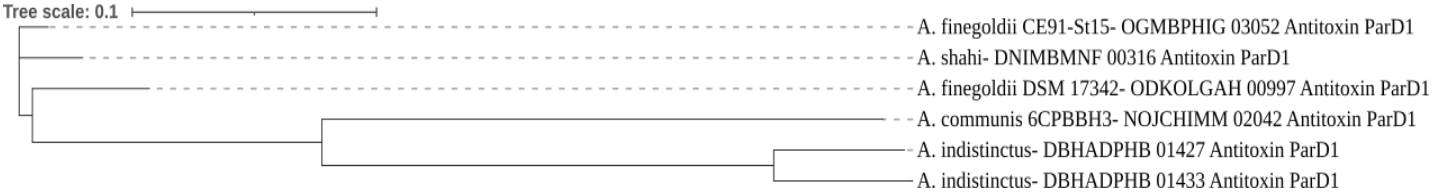

Figure 31: Phylogenetic tree for the Antitoxin gene parD1.

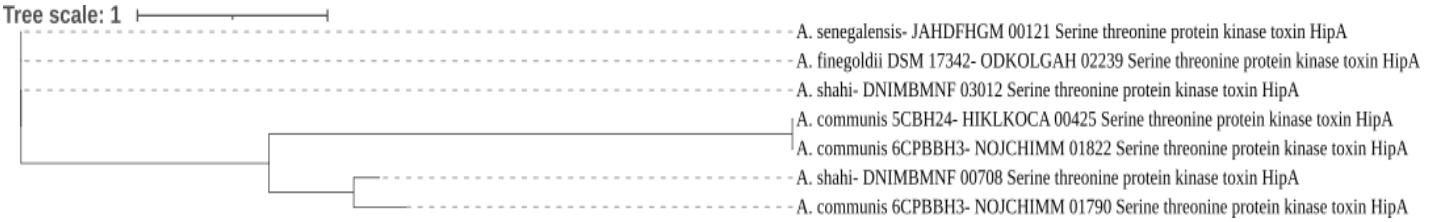

Figure 32: Phylogenetic tree for the toxin hipA gene.

Nevertheless, the operons of the toxin-antitoxin genes were checked from the operon mapper results and both single gene operon and multiple gene operon could be identified. The operon for the antitoxin gene parD1 and the toxin gene parE1 is the same. In addition to that, operons for the toxin apxIB gene in both *A. communis* and *A. dispar* genomes are actually the same. Moreover, the operon for the antitoxin yefM and toxin yoeB gene is the same and it could be only found in the *A. finegoldii DSM 17342* genome. This operon contained only two gene- yoeB and yefM. (Table 8).

Table 8: Summarized information on operons of the identified secretion system genes.

Genome name	Genes in the operon	Identified toxin-antitoxin genes from the operon
Alistipes shahi WAL 8301	Toxin ParE1, Integration host factor subunit beta, Antitoxin parD1	1. Antitoxin parD1 2. Toxin ParE1
Alistipes finegoldii DSM 17342	Antitoxin parD1	1. Antitoxin parD1
	Antitoxin yefM, Toxin YoeB	1. Antitoxin yefM 2. Toxin YoeB
A. finegoldii CE91-St15	Toxin ParE1, Integration host factor subunit beta, Antitoxin parD1	1. Antitoxin parD1 2. Toxin ParE1
A. communis 6CPBBH3	Toxin ParE1, Antitoxin parD1	1. Antitoxin parD1 2. Toxin ParE1
	Toxin RTX-I translocation ATP-binding protein, GTP 3'%2C8-cyclase	1. Toxin RTX-I translocation ATP-binding protein, apxIB
A. dispar 5CPEGH6	Toxin RTX-I translocation ATP-binding protein, GTP 3'%2C8-cyclase	1. Toxin RTX-I translocation ATP-binding protein, apxIB

In the *A. shahi* and the *A. finegoldii CE91-St15* genome, the toxin-antitoxin operon consist- Toxin ParE1, Integration host factor subunit beta and Antitoxin ParD1 genes. Other genomes that contained either of these two gene were part of a single gene operon. The other genes were part of single gene operon.

3.7 Identified of Secretion System related Genes and Operons

Secretion system related genes were identified across the 16 genomes of *Alistipes* species. The output has been represented below (Table 9). Only four secretion system related genes could be identified across the genomes and most of the genomes have no secretion system related genes. No secretion system related genes could be identified in the following genomes- *Alistipes shahi, A. finegoldii DSM 17342, A. onderdonkii 3BBH6, A. onderdonkii 5CPYCFAH4, A. onderdonkii 5CPYCFAH2, A. onderdonkii CE91-St18, A. dispar 5CPEGH6, Alistipes sp. Dk3624, A.*

senegalensis, A. putredinis, and *A. megaguti*. On the other hand, two secretion system genes could be identified in the genomes- *A. finegoldii CE91-St15, A. communis 5CBH24, A. communis 6CPBBH3* and *uncultured Alistipes sp. isolate min17_bin03* each. The number of identified secretion system related gene was found three, in the *A. indistinctus* genome.

Table 9: Summary of the identified secretion system genes across the 16 genomes of *Alistipes* species. Here, presence is denoted by '+' and absence is denoted by '-' sign respectively.

Secretion System related Genes	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
gspF: Putative type II secretion system protein F	-	-	+	-	-	-	-	+	+	-	-	+	-	-	-	-
sctC: Type 3 secretion system secretin	-	-	+	-	-	-	-	+	+	-	-	+	-	-	-	+
epsE: Type II secretion system protein E	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
epsF: Type II secretion system protein F	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+

Among the four identified secretion system genes, only one gene is related to Type III Secretion System (T3SS) and the rest three are related to Type II Secretion System (T2SS). The gene 'gspF: Putative type II secretion system protein F' was identified in the following genomes- *A. finegoldii*

CE91-St15, A. communis 5CBH24, A. communis 6CPBBH3 and *A. indistinctus* and a phylogenetic tree was generated (Figure 33). In the tree, the gspF gene identified in the multiple genomes of the *A. communis* species had appeared closer to each other (Figure 33).

Figure 33: Phylogenetic tree of the secretion system gene 'gspF' across the genomes.

The other two T2SS related genes 'espE' and 'espF' was taken along with 'gspF' to generate another phylogenetic tree where the relation between the T2SS genes could be identified (Figure 34) and it could be observed that the 'espF' gene identified in the *uncultured Alistipes sp.* genome and the 'espE' gene identified in *A. indistinctus* genome were closer to each other. Thus, the T2SS genes identified across different genomes were found closer to each other (Figure 34).

Another phylogenetic tree (Figure 35) was generated to identify the correlation between the 'gspF' and the 'espF' gene as they had similar annotation. It could be observed that the 'espF' gene identified in the *uncultured Alistipes sp.* genome and the gspF gene identified in other genomes appear closer in the phylogenetic tree (Figure 35).

Figure 34: Phylogenetic tree based on the identified T2SS genes across the genomes.

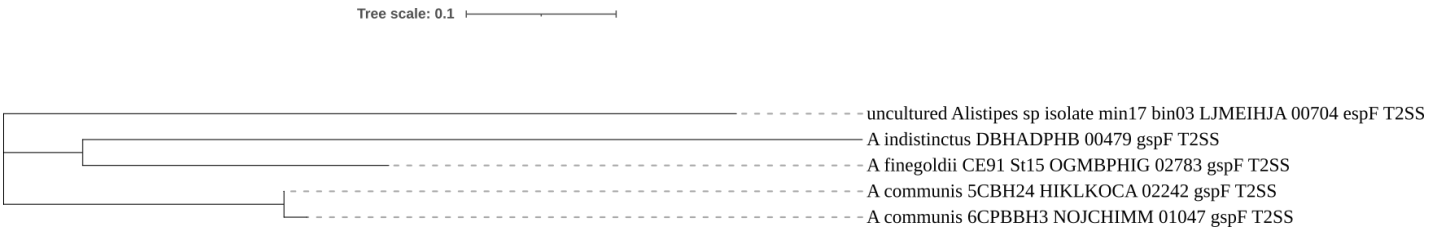

Figure 35: Phylogenetic tree based on the 'gspF' and 'espF' genes across the genomes.

Furthermore, the gene ‘sctC: Type 3 secretion system secretin’ that is related to Type III Secretion System (T3SS) was identified in the following genomes- *A. finegoldii CE91-St15, A. communis 5CBH24, A. communis 6CPBBH3, A. indistinctus* and *uncultured Alistipes sp. isolate min17_bin03*. Another phylogenetic tree (Figure 36) has been generated for this gene and the genomes consisting this gene appeared closer to each other in the tree where the sctC gene in both genomes from *A. communis* could be found side by side.

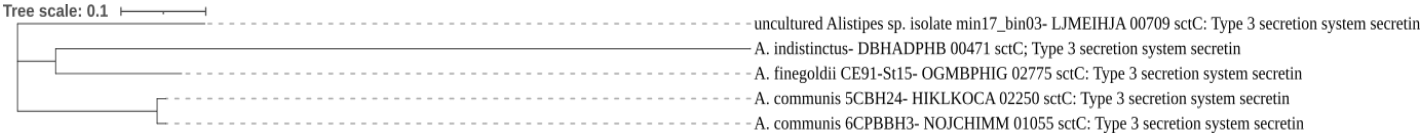

Figure 36: Phylogenetic tree of the secretion system gene ‘sctC’ across the genomes.

Next, the operons for the identified secretion system genes were identified as well and summarized (Table 10). First, a multiple gene operon could be identified for the genome *A. indistinctus* that encoded the aforementioned three secretion system genes- ‘gspF’, ‘sctC’ and ‘espE’. Same type of operon could be found for the *A. shahi, A. finegoldii CE91-St15, A. communis 5CBH24, A. communis 6CPBBH3* genomes where both the genes ‘gspF’ and ‘sctC’ were encoded from the same operon.

Table 10: Summarized information on operons of the identified secretion system genes.

Genome name	Genes in the operon	Identified secretion system genes from the operon
A. indistinctus 2BBH45	gspF, sctC, espE, and five hypothetical proteins	1. gspF 2. sctC 3. espE
A. communis 5CBH24	gspF, sctC, and seven hypothetical proteins	1. gspF 2. sctC
A. communis 6CPBBH3	gspF, sctC, and seven hypothetical proteins	1. gspF 2. sctC

Genome name	Genes in the operon	Identified secretion system genes from the operon
uncultured Alistipes sp. isolate min17_bin03	sctC and five hypothetical proteins	1. sctC
	espF	1. espF

On the contrary, the genome *uncultured Alistipes sp. isolate min17_bin03* was found to have a different type of operon as two different operons for the two identified secretion system genes was identified from the Operon Mapper results and the operon for the gene 'espF' had no other genes while the other operon of 'sctC' gene was a multiple gene operon.

3.8 Annotated Carbohydrate-Active enzymes

Carbohydrate active enzymes (CAZyme) were identified across the genomes and the tool dbCAN2 returned total 70 annotated carbohydrate active enzymes from different *Alistipes* genomes. The highest number of annotated CAZyme was found in two genomes- *A. senegalensis* and *A. finegoldii CE91-St15*, 32 CAZymes each. The other genome of *A. finegoldii* species- *A. finegoldii DSM 17342* was found to have 27 CAZymes. Next, 29 CAZymes were found in the following genomes- *A. shahi, A. dispar* and *A. megaguti*.

Then, the number of CAZymes were found in the *A. onderdonkii* genomes were- 28 in *A. onderdonkii 3BBH6* genome, 22 each in both *A. onderdonkii 5CPYCFAH6* and *A. onderdonkii 5CPYCFAH2* genomes and 26 CAZymes in the *A. onderdonkii CE91-St18* genome. Similarly, the number of CAZymes in the two genomes of *A. communis*- 23 in *A. communis 5CBH24* and 16 in *A. communis 6CPBBH3*. However, only those CAZymes were considered as significant who were found present in at least 75% of the total genome, that is equal to 12 genomes or more out of the 16 genomes and 13 such CAZymes could be identified.

Table 11: Summarized information of the significant Carbohydrate active enzymes (CAZymes). Here, presence is denoted by '+' and absence is denoted by '-' sign respectively.

Carbohydrate Active Enzyme		Name of species															
EC Number	Enzyme Name	*A. shahi WAL 8301*	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*Alistipes megaguti*	*uncultured Alistipes sp. isolate*
2.4.1.1	glycogen or starch phosphorylase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2.4.1.18	branching enzyme	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3.2.1.113	mannosyl-oligosaccharide α-1,2-mannosidase	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
3.2.1.114	mannosyl-oligosaccharide α-1,3-1,6-mannosidase	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
3.2.1.22	α-galactosidase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
3.2.1.52	β-hexosaminidase	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
3.2.1.24	α-mannosidase	+	+	+	+	+	+	+	+	-	+	+	+	+	-	+	+
3.2.1.21	β-glucosidase	+	+	+	+	-	-	+	+	+	-	+	+	+	+	+	+
3.2.1.23	β-galactosidase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-
3.2.1.40	α-L-rhamnosidase	+	-	+	+	+	+	+	+	+	+	+	+	+	-	+	-

Carbohydrate Active Enzyme		Name of species														
EC Number	Enzyme Name	*A. shahi WAL 8301*	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*Alistipes megaguti*	*uncultured Alistipes sp. isolate*
3.2.1.51	α-L-fucosidase	+	+	+	+	+	+	+	-	-	+	+	+	+	-	+	+
2.4.1.25	amylomaltase or 4-α-glucanotransferase	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+
3.2.1.63	α-1,2-L-fucosidase	+	+	+	+	+	+	+	-	-	+	+	+	-	-	+	+

Two CAZymes were found in all the 16 genomes of *Alistipes* species and these two CAZymes are- 'Glycogen or Starch phosphorylase' and 'Branching enzyme'. Another two CAZyme 'mannosyl-oligosaccharide α-1,2-mannosidase' and 'mannosyl-oligosaccharide α-1,3-1,6-mannosidase' could be found in 15 genomes out of the 16 genomes, absent in the *A. senegalensis* genome. Similarly, the CAZyme 'α-galactosidase' was found in 15 genomes of the 16 genomes as well; however, was found absent in the *uncultured Alistipes sp. isolate min17_bin03* genome and the CAZyme 'β-hexosaminidase' was only absent in the genome of *A. putredinis* and present in the other 15 genomes (Table 11).

The number of the significant annotated carbohydrate active enzyme is similar for the *A. shahi, A. finegoldii CE91-St15, A. onderdonkii 3BBH6* and *A. onderdonkii CE91-St18* genomes. Each of these genomes contain all the significant 13 CAZymes (Table 11). Therefore, the pattern of present significant CAZyme is similar in these genomes. Also, the pattern is similar for the *A. onderdonkii 5CPYCFAH4* and *A. onderdonkii 5CPYCFAH2* genomes as well since both of these genomes lack the same CAZyme.

On the contrary, the pattern is different in both *A. communis* and *A. finegoldii* genomes respectively, albeit these genomes have most of the CAZymes similar.The CAZyme 'α-L-rhamnosidase' could be found in the *Alistipes finegoldii CE91-St15* genome and was absent in the other genome. Further, the CAZyme 'α-mannosidase' was present in the *A. communis 5CBH24* genome and absent in the other genome of the same species. The CAZymes were checked across different genomes of different species as well and these happened to have different pattern of presence and absence of significant annotated carbohydrate active enzymes.

3.9 Identified Prophage Regions

The online based tool, PHASTER could identify 24 phage regions across the 16 genomes of *Alistipes* species. The phage regions were identified as Complete or Intact, Questionable, and Incomplete. The identified regions have been summarized in tabular format (Table 12) and the complete regions have been marked as 'C', and questionable have been marked as 'Q' while no such marking have been done for the incomplete regions.

Only one complete or intact phage region and two questionable phage regions could be identified across the 16 genomes of *Alistipes* species. The complete phage region 'PHAGE_Riemer_RAP44' was identified in the *A. shahi* genome and no further phages was identified from this genome. One of the questionable phage regions 'PHAGE_Burkho_phi1026b' was identified in the *A. onderdonkii CE91-St18* genome and the other questionable phage region 'PHAGE_Burkho_KS9' was identified in the *A. megaguti* genome.

No other phage regions could be identified from these two genomes but an incomplete phage region each. An incomplete phage region 'PHAGE_Cronob_ENT39118' in the *Alistipes megaguti* genome and another different incomplete phage region 'PHAGE_Synech_S_CBS3' was identified in the *A. onderdonkii CE91-St18* genome (Table 12).

Table 12: Summary of identified phage regions where '+' sign means presence and '-' means absence.

Name of the Phage region	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
PHAGE_Riemer_RAP44	+; C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PHAGE_Paenib_Lucielle	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PHAGE_Flavob_vB_FspM_pippi8	-	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-
PHAGE_Prochl_P_SSM2	-	-	+	+	+	+	-	-	-	+	-	-	+	-	-	+
PHAGE_Lactob_LP65	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
PHAGE_Flavob_vB_FspM_lotta8_1	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
PHAGE_Bacill_SPbeta	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
PHAGE_Flavob_vB_FspS _hattifnatt9_1	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
PHAGE_Clostr_c_st	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
PHAGE_Bacill_CP_51	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
PHAGE_Aeriba_AP45	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
PHAGE_Escher_phAPEC8	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
PHAGE_Flavob_vB_FspS _filifjonk9_1	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
PHAGE_Sulfit_NYA_2014a	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
PHAGE_Flavob_vB_FspS	-	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-

Name of the Phage region	Name of species															
	A. shahi WAL 8301	*A. finegoldii DSM 17342*	*A. finegoldii CE91-St15*	*A. onderdonkii 3BBH6*	*A. onderdonkii 5CPYCFAH4*	*A. onderdonkii 5NYCFAH2*	*A. onderdonkii CE91-St18*	*A. communis 5CBH24*	*A. communis 6CPBBH3*	*A. dispar 5CPEGH6*	*Alistipes sp. dk3624*	*A. indistinctus 2BBH45*	*A. senegalensis*	*A. putredinis*	*A. megaguti Marseille-P5997*	*uncultured Alistipes sp. isolate min17_bin03*
_laban6_1																
PHAGE_Escher_vB_EcoM _Schickermooser	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
PHAGE_Burkho_phi1026b	-	-	-	-	-	-	+; Q	-	-	-	-	-	-	-	-	-
PHAGE_Synech_S_CBS3	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-
PHAGE_Flavob_23T	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
PHAGE_Sinorh_phiM7	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
PHAGE_Cronob_ENT39118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
PHAGE_Burkho_KS9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+; Q	-
PHAGE_Entero_phiP27	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
PHAGE_Synech_S_CAM3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+

The identified 21 incomplete regions were found across the genomes and same phage region could be found in multiple genomes of same species and there were some regions too that could not be found in all the genomes of the same species. For example, the phage region 'PHAGE_Flavob_vB_FspM_pippi8_1' was found in both the genomes of *A. communis* species while the phage region 'PHAGE_Prochl_P_SSM2' was found in 3 of the 4 genomes of *A. onderdonkii* species but absent in the *A. onderdonkii CE91-St18* genome. Two incomplete phage regions were identified in the *A. finegoldii DSM 17342* genome- 'PHAGE_Paenib_Lucielle' and 'PHAGE_Flavob_vB_FspM_pippi8_1'. These two regions were absent in the other genome of *A.*

finegoldii while five different incomplete phage regions- 'PHAGE_Prochl_P_SSM2', 'PHAGE_Flavob_vB_FspS_laban6_1', 'PHAGE_Synech_S_CBS3', 'PHAGE_Flavob_23T' and 'PHAGE_Sinorh_phiM7' were identified in the *A. finegoldii CE91-St15* genome.

The incomplete region 'PHAGE_Prochl_P_SSM2' was found in the greatest number of genomes, 7 genomes- *A. finegoldii CE91-St15, A. onderdonkii 3BBH6, A. onderdonkii 5CPYCFAH4, A. onderdonkii 5NYCFAH2, A. dispar, A. senegalensis* and *uncultured Alistipes sp. isolate min17_bin03*. Moreover, the incomplete region 'PHAGE_Flavob_vB_FspM_pippi8' was found in 3 genomes- *A. finegoldii DSM 17342, A. communis 5CBH24* and *A. communis 6CPBBH3*. Apart from these phage regions, other identified regions were found in either one genome from the 16 genomes of *Alistipes* species. Only one incomplete sequence was identified each in *Alistipes sp. Dk3624, Alistipes indistinctus* and *Alistipes putredinis* genomes.

Chapter 4. Discussion

Here the complete and chromosome level genomes of the genus *Alistipes* was used to perform the pan-genome analysis. Till March 2022, genomic data was collected from NCBI genome database and 16 genomes could be found suitable for the analysis based on their completeness of genome sequencing and source of isolation. The chromosome level genome sequences were taken into consideration as these chromosome level genomes can also be analyzed to understand the genetic diversity of the organism similar to the complete level genomes. In this analysis, total 14 complete genomes from 8 different species of *Alistipes* and 2 chromosome level genomes two other species of *Alistipes* were used.

Further, the name of the species of one complete genome (*Alistipes sp. dk3624*) and another chromosome level genome (*uncultured Alistipes sp. isolate min17_bin03*) was not specified in the NCBI database. Albeit, other genomes had specified species name and some species had multiple complete genomes while only the *Alistipes shahi WAL 8301* species had chromosome level genome. Now, the genus *Alistipes* was chosen for performing this analysis despite having lower number of complete and chromosome level genomes available in the publicly available repositories as this genus is a comparatively new bacterial genus that is on the emergence of implications to different health conditions in human.

Beside pan-genome analysis, other important characteristics such as presence of virulence factors, antimicrobial resistant genes, toxin-antitoxin system related genes, secretion system genes, phage regions, carbohydrate-active enzymes and their operons were also analyzed in these 16 genomes of *Alistipes* species. Pan-genome analysis was performed to identify the core genes that is genes present in all genomes, soft-core genes, shell genes and cloud genes within these 16 genomes and to infer whether the genus contain open pan-genome or closed pan-genome while the analysis of other characteristics were done to identify the contribution to pathogenesis and other associated factors.

Genomic Data Retrieval, Genome Annotation, Pan-Genome analysis: The genomic data of the 16 genomes of *Alistipes* were retrieved using Galaxy upon collecting the respective accession numbers of the genomes. These data could be directly downloaded from the NCBI genome database; however, the get data option from the Galaxy environment was used for retrieving the genomes in FASTA file format via providing the accession numbers. The advantage was the

retrieved data was in the Galaxy cloud system and these data could be further used within the Galaxy environment to perform other analysis, such as pan-genome analysis, identification of toxin-antitoxin genes, identification of secretion system genes, etc. by running another bioinformatic tool- Roary.

Furthermore, one of the main prerequisites of pan-genome analysis is- annotation of the genomes. Prokka was used here for the annotation of the retrieved genomes and this tool was also run within the Galaxy environment. This was another advantage of using the Galaxy environment for retrieving the data. Genomic data retrieval led to annotation of the genomes and then pan-genome and further analysis via Roary within Galaxy and the whole process was easy and straight-forward, that makes Galaxy web platform suitable for data retrieval and cloud-based analysis platform for this project. Nevertheless, accidental loss of both of the retrieved and further analyzed data was merely impossible at this cloud-based Galaxy web platform that has made it more advantageous than the traditional approaches.

Although the genome annotations were provided at NCBI genome database for each genome in GFF3 file format, Prokka was used here for the homogenization of genome annotation that the NCBI provided annotation lacked. Also, the provided genome FASTA files were compared with the databases of known proteins to find a significant match during the Prokka annotation. In addition to that, the sequences with significant match were further analyzed via Prodigal, RNAmmer, Aragorn, SignalP, and Inferna within the Prokka tool itself to infer the features of the predicted protein, that provided an accurate as well as homogenized genome annotation for the retrieved genomes to perform pan-genome analysis.

Nevertheless, some sequences were annotated as ‘Hypothetical protein’ since these sequences were not subjected to any significant match from previously known proteins. Yet some features could be identified that had led the tool to annotate these sequences as hypothetical protein sequences. In this analysis, a good number of hypothetical proteins were found, yet considered for including in the analysis because by further analyzing their features and comparing them with their operon partners later on, their significant match could be inferred. This could be done easily because in the annotated GFF3 file, Prokka provided the Prodigal ID number within the annotation description for each sequence.

Also, the Prokka annotation summary provided the total number of genes, CDS, tRNA, mRNA, tmRNA, miscRNA and repeat RNA. The highest number of total genes could be identified from

the *Alistipes finegoldii CE91-St15* genome. Among the 3618 genes, 3540 were found to be protein coding, CDS gene. The total number of annotated sequences and sequences annotated as coding sequence in this *Alistipes finegoldii CE91-St15* genome varies greatly in the other genome- *Alistipes finegoldii DSM 17342* of the same species. In the other genome the total number of annotated sequences is 3267 and the number of sequences were annotated as protein coding was 3187. No repeat regions could be identified from both genomes and the non-coding RNAs could be found similar in both genomes of *Alistipes finegoldii.*

Unlikely to *A. finegoldii* species, the total number of annotated sequences as well as protein coding sequence were found quite similar in the multiple genomes of different species- *Alistipes onderdonkii* and *Alistipes communis* that have four genomes and two genomes respectively. The significance of the annotated repeat regions in 3 genomes- *Alistipes shahii WAL 8301, Alistipes putredinis* and *uncultured Alistipes sp.* are this repeat region might be related to the functionalities of the synthesized proteins. In addition to that, the repeat regions having variable length might introduce potential point mutation via mobile genetic elements or any other means that might lead to evolutionary distances among different species belonging to the same genus or might lead to diseases. This annotated information of repeat region as well as ncRNAs might be useful in the epigenetics analysis of the *Alistipes* species in future.

The core genes, soft-core genes, shell genes and other genes of the *Alistipes* species from the 16 genomes were identified via pan-genome analysis. The pan-genome analysis tool, Roary was used for the analysis within the Galaxy environment. Some default parameters of Roary had to be changed before initiating the analysis and the 'Minimum percentage identity for blastp' was changed from 95% to 90%. The number of genomes taken to perform the analysis was quite lower than usual pan-genome analysis; hence, the percentage identity had to be updated to cover as much genes as possible. However, InterProScan and multiple sequence alignments were performed to ensure the absence of false positive results from the analysis.

Roary analysis of the *Alistipes* species pan-genome identified 14 core gene, no soft-core genes, 3034 shell genes and over 20,000 cloud genes from the analyzed 16 genomes. The core genes were identified based on their presence on greater than 99% genomes. Likely, the soft-core genes could be identified from their presence on 95-99% genomes and the shell genes could be identified from their presence on 15-95% genomes. Further, the presence of genes on 0-15% genomes were identified as the cloud genes from the analyzed genomes of *Alistipes* species.

Here, the number of analyzed genomes were comparatively lower than other pan-genome analysis and initially no soft-core genes could be identified based on the traditional theory of presence of soft-core genes on 95-99% genomes of the total genomes. However, 108 genes could be identified in at least 12 or more genomes from the analyzed 16 genomes; that is 75-99% of the analyzed genomes of *Alistipes* species. Hence, these genes could be considered as soft-core genes of the *Alistipes* species as the total number of analyzed genome is lower in this analysis. If the number of complete genome increase, some genes that are now present on the 75-99% genomes might fall in the 95-99% range and could be identified as soft-core genes accordingly.

Now, the classification of a pan-genome whether it is open or closed is dependent on the identification of new genes upon an increase in the number of analyzed genome. In an open-pan genome, the number of gene families grows indefinitely upon addition of new genomes while the opposite can be observed in case of a closed pan-genome. Moreover, accessory genes are the genes that are not present in all the genomes of an analysis yet can be found in at least two or more than two genomes from the total number of analyzed genomes. Necessarily, the number of accessory genes in an open pan-genome is higher and the relatedness of the accessory genes with the core genes are not as much as the accessory genes of a closed pan-genome where these accessory genes can be found lower in number. In this analysis, the number of accessory genes as well as cloud gene are much higher than the identified core genes making the analyzed genomes of different *Alistipes* species an open pan-genome.

Antimicrobial resistance genes and operons: The web-based Resistance Gene Identifier (RGI) tool from the Comprehensive Antibiotic Resistance Database (CARD) was used to identify antimicrobial resistance genes (AMR genes) from the 16 genomes of different *Alistipes* species. The cut off options- perfect hit, strict hit and loose hit were selected and nudge was excluded. Here, nudge could have been included to annotate the loose hit results having percentage identity ≥95% to strict. However, this could introduce false positive predicted AMR genes since the number of analyzed genomes was low. Nevertheless, the best of the loose hit that is top 10% from the best hit bitscore for each genome was considered as significant AMR genes and the rest loose hit results were not considered for further analysis.

Along with the identified AMR genes based on their cut offs, the tool also provided useful information, such as, the best hit AMR gene annotation, their drug class, resistance mechanism, AMR gene family, the protein sequence from the AMR gene was identified, predicted DNA,

predicted protein, etc. In this analysis, the annotation that was provided for the identified AMR genes by CARD-RGI was mentioned along with their Prokka annotation as for the homogenization of gene annotation Prokka was already used. Nonetheless, it was convenient to use the Prokka annotations for the identified AMR genes, as this allowed to identify the same AMR gene that was predicted multiple times.

Across the genomes, the predicted AMR genes with loose hit cut off was found common. Although, at least 1 AMR gene could be identified with strict hit cut off in most of the genomes except for the genome of *Alistipes putredinis* species. On the other hand, the number of predicted AMR gene with perfect hit cut off was only one and it could be found in the *Alistipes sp. dk3624* genome. In this genome, the predicted AMR gene with perfect hit cut off was 'tetQ' that is resistant to the antibiotic tetracycline and belong to the antimicrobial resistance gene family of 'tetracycline-resistant ribosomal protection protein'.

Initially, total 2918 AMR genes were identified by the tool in the 16 genomes of *Alistipes* species where majority of the predicted AMR genes were loose hits. The highest number of identified AMR gene in a single genome was 216 and it was identified in *Alistipes onderdonkii CE91-St18* genome. Another genome where 215 AMR genes could be identified was the *Alistipes sp. dk3624* genome. On the contrary, the lowest number of identified AMR gene was 121 in the genome of *Alistipes putredinis* species. The number of identified AMR genes varied in the genomes of different species and in the multiple genomes of same species.

Although, the two genomes of the *Alistipes communis* species had similar number of identified AMR genes, 160 in each. However, most of the identified AMR genes were found common across different genomes of *Alistipes* species. In addition to that, the predicted AMR genes with aforementioned insignificant loose hit cut off were excluded from the list. Furthermore, the AMR genes that were identified in multiple genomes of both same and different species, some AMR genes were identified multiple times in different genomes, etc. these were excluded too. As a result, 36 AMR genes had been identified with perfect, strict and best of loose hit cut off across the 16 genomes of *Alistipes* species were included.

AMR genes identified in all the genomes: Three AMR genes could be found in all the genomes and these AMR genes are- '*Clostridioides difficile* gyrB conferring resistance to fluoroquinolones', '*Escherichia coli* EF-Tu mutants conferring resistance to Pulvomycin' and '*Mycobacterium tuberculosis* rpoB mutants conferring resistance to rifampicin'. The gyrB gene-

‘DNA gyrase subunit B’ from the Prokka annotation has been identified by CARD-RGI as an AMR gene that confers resistance to the antibiotic fluoroquinolone via antibiotic target alteration mechanism. Next, the tufA gene- ‘Elongation factor’ has been identified as EF-Tu mutants conferring resistance to the antibiotic elfamycin via antibiotic target alteration mechanism.

Some variants and/or mutants of the gene- DNA-directed RNA polymerase subunit beta (rpoB) have also been identified as AMR genes in the genomes of *Alistipes* species. The identified rpoB mutants are- *Mycobacterium tuberculosis* rpoB mutants, *Bifidobacterium adolescentis* rpoB mutants, and *Helicobacter pylori* rpoB mutants. Furthermore, other copies of the same gene were identified as rpoB_2 and rpoC. These mutants and different versions of the same gene were identified from different *Alistipes* species and they conferred resistance to the antibiotic rifamycin via two possible mechanisms- antibiotic target alteration and/ or antibiotic target replacement. These mutants and different versions of the same gene belongs to the same protein family- rifamycin-resistant beta-subunit of RNA polymerase (rpoB).

Although the *Mycobacterium tuberculosis* rpoB mutants could be identified from all the analyzed genomes of different *Alistipes* species, the *Bifidobacterium adolescentis* rpoB mutant and rpoB2 AMR genes could be identified only from the genome of *Alistipes putredinis* genome. Also, no other rpoB variants nor mutants could be identified from this genome. On the other hand, the *Helicobacter pylori* rpoB mutant AMR gene could be identified from 14 genomes out of 16, except the two genomes of *Alistipes indistinctus* and *Alistipes putredinis* species. Additionally, the presence of this AMR gene in multiple genomes of same species were found similar.

The operons for the AMR genes- ‘*Escherichia coli* EF-Tu mutants conferring resistance to Pulvomycin’ and ‘*Mycobacterium tuberculosis* rpoB mutants conferring resistance to rifampicin’ were found similar. Moreover, the rpoB_2, another variant of the AMR gene rpoB that was identified only from the genome of *Alistipes putredinis* was found to in the same operon. The genes in this operon- Elongation factor Tu, tRNA-Trp(cca), Protein translocase subunit SecE, Transcription termination/antitermination protein NusG, 50S ribosomal protein L11, 50S ribosomal protein L1, 50S ribosomal protein L10, 50S ribosomal protein L7/L12, DNA-directed RNA polymerase subunit beta (rpoB), and DNA-directed RNA polymerase subunit beta’ (rpoC).

The operon for the other rpoB mutant- *Helicobacter pylori* rpoB mutants contained most of the genes similar. On the other hand, *Bifidobacterium adolescentis* rpoB mutant was found in another operon where the other genes were found multiple copies of the same rpoB gene. It was found that

different versions of the same AMR gene could be part of different operons but in most of the cases it could belong to the same operon. The operons for the rpoB AMR gene were found similar for most of the *Alistipes* genomes except the genomes- *Alistipes shahi WAL 8301* and *Alistipes sp. dk3624*. Operons for these two genomes had some additional genes; although most of the genes were similar to the aforementioned operons of different genomes of *Alistipes* species.

Efflux pump membrane transporter, BepE: The tool CARD-RGI identified the different versions of the same AMR gene, BepE as- 'adeF', 'ceoB', 'cmeB', 'MexF', 'MexW', 'oqxB'. This could be identified upon revising the annotations from the Prokka generated GFF files. This gene BepE have different versions and different versions were annotated differently by CARD-RGI. All these versions belong to the protein family- 'Resistance-Nodulation-Cell Division (RND) antibiotic efflux pump'. By their antibiotic efflux mechanism, different versions of the BepE gene confers resistance to fluoroquinolone and tetracycline antibiotics.

In most of the genomes, different versions of the single AMR gene BepE, was identified as 'adeF' except in the genome of *Alistipes putredinis*. However, another version of the BepE gene- MexW was identified in the *A. putredinis* genome. Therefore, it can be considered that, another AMR gene, BepE was identified in all the genomes of different *Alistipes* species. Now, this version of BepE gene- MexW was also identified other genomes of *Alistipes* except the two complete genomes of *Alistipes communis* and the single complete genome of *Alistipes indistinctus* species. The operon that this version belongs to was found similar for both *A. finegoldii* genomes while the operon for the four genomes of *A. onderdonkii* species, *A. dispar* genome, *A. senegalensis* genome was identified similar. Albeit the operon for the *A. putredinis, A. megaguti, Alistipes sp. dk3624* and *uncultured Alistipes sp.* was different for this gene, most of the genes of their operons were found similar to the other operons that was identified in different genomes of *Alistipes* species.

In the multiple genomes of *Alistipes finegoldii* and *Alistipes onderdonkii* two different BepE versions- 'bepE_1' and 'bepE_4' were identified as the AMR gene 'adeF'. While bepE_1 and bepE_3 versions of the same resistance gene were identified in the multiple genomes of *Alistipes communis* species. Other bepE versions that were identified as adeF- bepE_2, bepE_4, bepE_5 and these different versions were found across different genomes of *Alistipes* species while the *A. senegalensis* genome had four copies of different versions of the bepE gene- bepE_1, bepE_2, bepE_3 and bepE_4. The operon partners for the different versions of the bepE gene were found mostly similar across multiple genomes of both same species and different species. The other genes

from the same operon were- Multidrug resistance protein MdtA, Outer membrane protein OprM, Efflux pump periplasmic linker BepF, Toluene efflux pump outer membrane protein TtgI.

Other AMR genes and their operons: Another AMR gene, gyrA- DNA gyrase subunit A was identified as '*Clostridioides difficile* gyrA conferring resistance to fluoroquinolones' in most of the genomes of *Alistipes* species except in the *Alistipes putredinis* and *Alistipes finegoldii CE91-St15* genome. However, another version of the same resistance gene was identified as '*Capnocytophaga gingivalis* gyrA conferring resistance to fluoroquinolones' and was found in the *Alistipes finegoldii CE91-St15* genome. Although the gyrA gene that was identified as an AMR gene because of mutation; later it was found that the *Clostridioides difficile* gyrA resistance gene further got mutated to *Capnocytophaga gingivalis* gyrA. Antibiotic target alteration is the mechanism of action of this AMR gene. Operon partners for this resistance gene were hypothetical proteins.

Energy-dependent translational throttle protein EttA, identified as an AMR gene TaeA by CARD-RGI is another AMR gene that was identified in all the different genomes of *Alistipes* species except in the genome of *Alistipes putredinis*. The AMR gene belongs to the 'ATP-binding cassette (ABC) antibiotic efflux pump' protein family and confers resistance to the pleuromutilin antibiotic through antibiotic efflux mechanism. Other operon partners of this gene were found mostly hypothetical proteins, Histidine--tRNA ligase, etc. Moreover, across different genomes of *Alistipes* species, the operon for this gene was found similar to all of them.

CARD-RGI has identified three versions of the 'Isoleucine--tRNA ligase (ileS)' gene as antimicrobial resistance genes and annotated them as- *Staphylococcus aureus* mupA conferring resistance to mupirocin, *Staphylococcus aureus* mupB conferring resistance to mupirocin, and *Bifidobacterium bifidum* ileS conferring resistance to mupirocin. This AMR gene belongs to the 'antibiotic-resistant isoleucyl-tRNA synthetase (ileS)' protein family and confers resistance to the antibiotic mupirocin via antibiotic target alteration. Further, the operon for this gene had no other gene except this AMR gene and 'RNA polymerase-binding transcription factor DksA' and the operons were similar for the different genomes of *Alistipes* species.

Antibiotic efflux, antibiotic target alteration, antibiotic target protection, antibiotic target replacement- these are the antibiotic resistance mechanisms that were observed in the identified AMR genes across different genomes of *Alistipes* species. Different *Alistipes* genomes were found having AMR genes that might lead those species resistant to the antibiotic- fluoroquinolone,

tetracycline, mupirocin, macrolide, rifamycin, peptide group antibiotics, etc. One AMR gene- OprM that confers resistance to the antibiotic carbapenem, was not identified by CARD-RGI in any of the genomes of *Alistipes* initially; yet this gene was identified as the operon partner of another AMR gene- BepE that was found in all the genomes of *Alistipes* species. The mutated AMR gene- OprM along with another BepE that belong to the same operon can confer resistance to macrolide antibiotic, fluoroquinolone antibiotic, monobactam, aminoglycoside antibiotic, carbapenem, cephalosporin, cephamycin, tetracycline antibiotic, peptide antibiotic, aminocoumarin antibiotic, diaminopyrimidine antibiotic, sulfonamide antibiotic, phenicol antibiotic, disinfecting agents and antiseptics at the same time.

The highest number of antimicrobial resistance genes was identified in the *Alistipes onderdonkii CE91-St18* genome and 19 AMR genes were identified in this genome alone. Nevertheless, 14 AMR genes were identified together in the four genomes of *Alistipes onderdonkii* species. Similarly, the same number of AMR genes were also identified in the two genomes of *Alistipes communis* at the same time and together 12 AMR genes were identified from the two genomes of *Alistipes finegoldii* species. Although, the presence or absence of the AMR genes in multiple genomes of same species and other species are more important than the number of AMR genes, these numbers help to make an educated guess that the *A. onderdonkii* species might be associated with antibiotic resistance among the other *Alistipes* species.

Virulence Factor, Toxin-Antitoxin system, Secretion system genes and operon: The identified antimicrobial resistance genes hints that different species of *Alistipes* might be associated with pathogenesis. Although, presence and absence of antimicrobial resistance genes cannot confirm the pathogenicity of any organism as these genes can be present in the organism for its survival in the environment. Therefore, the genomes of different *Alistipes* species were analyzed to identify presence or absence of- virulence factor genes (VF genes), toxin-antitoxin system genes and secretion system related genes. The screening for secretion system related genes and toxin-antitoxin system genes were performed using Roary alongside the pan-genome analysis and another online-based bioinformatic tool, VFanalyzer was used to identify the presence of the virulence factor genes.

The virulence factor database (VFDB) was used by the tool VFanalyzer for comparing the 16 genomes of different *Alistipes* species to identify the potential virulence genes based on the features, characteristics, functions of the genes, etc. To initiate the process, the genus of the

organism had to be mentioned in the VFanalyzer web-server. As the genus *Alistipes* is a comparatively new bacterial genus, it was not available in the list of the tool, VFanalyzer. Therefore, the tool was run for all 31 genera available in the genus list with one representative genome first. The *Alistipes finegoldii DSM 17342* was the reference genome and it was found closer to the genus *Chlamydia*. Later, the 16 genomes of different *Alistipes* species were run by selecting *Chlamydia* from the genus list and genes related to virulence factor were identified.

The number of identified virulence factor related genes was 13 and 6 of these genes were found present in all the genomes of different *Alistipes* species. The presence and absence of the virulence factor genes were found similar in the multiple genomes of *Alistipes finegoldii, Alistipes communis* and *Alistipes onderdonkii* species. The virulence factor genes identified in the *Alistipes indistinctus* genome was found similar to the genomes of *Alistipes communis*. The highest number of virulence factor genes were identified from the *Alistipes shahi WAL8301* genome and all 13 VF genes were found present in the *A. shahi* genome.

The pattern of virulence factor related genes in the multiple genomes of same species is expected to be similar and this was observed in the four genomes of *Alistipes onderdonkii* species. The two genomes of *Alistipes finegoldii* had most of the virulence genes similar; yet the *A. finegoldii CE91-St15* genome had an additional virulence gene that was absent in the other genome of the same species that was not expected. On the other hand, the virulence related genes were identified similar for the two genomes of *A. communis* species and the presence, absence pattern of the same VF genes were found in the *A. indistinctus* genome as well. That refers that the virulence pattern might be similar for the *A. communis* and *A. indistinctus* species.

Some of the identified VF genes were found to be associated with antibiotic resistance as well. For example, the VF gene arnA that have been identified from the multiple genomes of *Alistipes finegoldii* and *Alistipes onderdonkii* species as well as in the *Alistipes shahi WAL8301* genome, have antibiotic target alteration mechanism to acquire resistance against peptide group antibiotic. Therefore, the gene arnA has been annotated as Bifunctional polymyxin resistance gene, arnA.

Similarly, tufA that is the Elongation factor, has already been identified as an antibiotic resistantace gene in all the genomes of *Alistipes* species; the same gene was identified as a VF gene as well. Usually, this was supposed to be a positive/negative control result; yet this was considered to be a potential virulence factor related gene since multiple evidence of mutation of this same gene has been observed previously. Also, the genes within the same operon of this gene

were found virulent as well. Therefore, the classification and characterization of these VF genes were important to identify the pathogenicity pattern of different *Alistipes* species.

Toxin-antitoxin genes and their operons were screened for the similar purpose of understanding the pathogenicity of this organism. The pattern of the presence and absence of AMR genes and VF genes have been found similar for *Alistipes finegoldii* and *Alistipes onderdonkii* species. However, the pattern for toxin-antitoxin genes were found different for these two species. No toxin-antitoxin system related genes were identified from the four genomes of *A. onderdonkii* species. Additionally, the presence and absence of toxin-antitoxin genes within the two genomes of *Alistipes finegoldii* was found dissimilar. The similar scenario could be observed for the multiple genomes of *Alistipes communis* as well.

The identified toxin-antitoxin genes across different genomes of *Alistipes* species were not high in number but two significant toxin-antitoxin system was identified in this analysis. Toxin-antitoxin genes are usually transcribed together; hence, they are part of a single operon where the antitoxin holds the capability to mitigate the effects of the toxin. Here, two such system was identified and one of them was- toxin parE1 and antitoxin parD1, together these two are a single toxin-antitoxin system and this system was identified in the *Alistipes shahi WAL* 8301 genome, *Alistipes finegoldii CE91-St15* genome and *Alistipes communis 6CPBBH3* genome.

Although, the antitoxin parD1 gene was also identified in the *Alistipes finegoldii DSM 17342* genome; yet this was not considered to be the part of the same toxin-antitoxin system. Furthermore, another toxin-antitoxin system- toxin YoeB and antitoxin YefM, was identified only in the *Alistipes finegoldii CE91-St15* genome and none of these two genes were identified in any other genomes of different *Alistipes* species. Both of the identified toxin-antitoxin system (TA system) was classified as type II TA system that was found similar to other pathogenic bacteria.

Now, the significance of the identification of these two toxin-antitoxin systems are- these systems will surely contribute in the pathogenesis of the organism and this system will play a vital role for the development of antibiotic resistance and other crucial virulence factors of the organism. Additionally, the presence and absence of the toxin-antitoxin system depict the history of the horizontal gene transfer of the bacteria. Nevertheless, these genes also help to maintain the mobile genetic element of the bacteria (if present, such as plasmid) and help to develop the bacteria some additional characteristics and functionalities such as stress tolerance, develop protection against bacteriophage, etc.

Another marker for bacterial pathogenesis- the secretion system related genes were identified from the different genomes of *Alistipes* species and only four such genes could be identified across the genomes of *Alistipes* and three genes were identified from the Type II secretion system (T2SS) and the other gene was identified from the Type III secretion system (T3SS). Two T2SS related genes were identified in the *Alistipes indistinctus* genome while one T2SS gene and another T3SS gene was identified in the following genomes- *A. finegoldii CE91-St15, A. communis 5CBH24, A. communis 6CPBBH3* and *uncultured Alistipes sp. isolate*. Similar to the TA system related genes, no secretion system related genes were identified from any of the four genomes of *Alistipes onderdonkii* species.

The operons for the identified secretion system genes were checked for the other genes present in the same operon and it was found that the identified T2SS and T3SS genes were transcribed from the same operon in the multiple genomes of *A. communis* and single genome of *A. indistinctus,* while the operon was different for the *uncultured Alistipes sp. isolate* species. The phylogenetic tree analysis of the three identified T2SS genes confirmed that these genes belong to same system and the genomes where these genes were identified are closely related to each other.

After analyzing the AMR genes, VF genes, toxin-antitoxin genes and secretion system genes- *Alistipes finegoldii, Alistipes communis, Alistipes indistinctus* and *uncultured Alistipes sp. isolate* species were found to be the most pathogenic strains/ isolate from the genus *Alistipes;* although, the presence of AMR genes and VF genes were found considerably higher as well as significant in the *Alistipes onderdonkii* species. The pattern of pathogenic traits and other characteristics were found mostly similar for the *Alistipes finegoldii, Alistipes onderdonkii, Alistipes communis, Alistipes shahi, Alistipes megaguti, Alistipes dispar, uncultured Alistipes sp. isolate* species.

Carbohydrate Active enzyme and Phage region: The identification of phage regions and carbohydrate active enzyme (CAZyme) profiling was done in the different genomes of *Alistipes* species to infer additional characteristics of the organism. Two different web-based bioinformatic tool- dbCAN2 for carbohydrate active enzyme profiling and PHASTER for the identification of phage regions were used to classify these additional characteristics. Both dbCAN2 and PHASTER, could successfully identify some carbohydrate active enzymes and phage regions.

Initially, the number of the annotated carbohydrate active enzymes was 70. Some carbohydrate active enzymes were found repeated while some of the different versions of the same CAZyme was found and some CAZyme were found as the control for the dbCAN2 run. Therefore, these

results were further analyzed to reduce the number to 13 based on the presence of the CAZyme in 12 or more genomes and excluding the repeated results. If a CAZyme was not identified in at least 75% of the analyzed genomes, the annotation was not considered as significance because the web-based tool had compared the given genomic data in FASTA file with the vast Carbohydrate-Active Enzyme Database and annotated the sequences based on similarities that could often fail to separate the false positive and control annotations.

Now, the significance of the identified carbohydrate active enzymes in different *Alistipes* species was to understand the utilization of carbohydrates by the bacteria for surviving both in the environment and host. Moreover, the non-pathogenic strains and pathogenic strains could be classified based on the carbohydrate utilization since the pathogenic strains might have difference as it may utilize some specific carbohydrates to survive within the host, source of carbohydrate utilization, etc.

All the genomes of *Alistipes* species have carbohydrate-active enzymes from the Glycosyl Transferase family and Glycoside Hydrolase Family. The identified CAZymes were found mostly from these two carbohydrate families and only the enzyme 'glycogen or starch phosphorylase' belonged to the Glycosyl Transferase family while rest of the identified enzymes belonged to different glycoside hydrolase family. For example, the identified CAZyme, 'mannosyl-oligosaccharide α-1,2-mannosidase' belongs to the GH92- Glycoside Hydrolase family while another CAZyme, 'α-galactosidase' belongs to the GH31- Glycoside Hydrolase family. Hence, the presence of carbohydrate-active enzymes across different genomes of *Alistipes* were found similar.

Lastly, the presence or absence of phage regions across the genomes of *Alistipes* were identified and the tool, PHASTER, predicted 24 phage regions where 21 phage regions were incomplete regions, two of the predicted regions were questionable regions and only 1 phage region was predicted complete region. The complete phage region- 'PHAGE_Riemer_RAP44' was identified in the chromosome level genome of *Alistipes shahi WAL 8301*. The identification of the phage regions is significant to further understand the history of horizontal gene transfer and infer the evolution of the organism because these regions allow an organism to acquire new characteristics such as virulence, antibiotic resistance, et cetera.

Nevertheless, the identification of phage region can play a crucial role on the future genetic engineering applications on different species of *Alistipes*. Since, a complete phage region could be identified in the *A. shahi* genome, this region could be used to insert a marker gene for further

analysis on the species. Besides, the incomplete and questionable phage regions can also be used to check if these regions also allow the insertion of new genetic component from external environment. The phage regions might allow the organism to uptake foreign genes from the environment and develop pathogenic or other different characteristics.

Chapter 5. Conclusion

This study was conducted on the complete and chromosome level genomes of the recently classified and mostly unexplored bacterial genus, *Alistipes*, that is an anaerobic, gram-negative bacteria, and a resident of the Gastrointestinal tract (GI tract) in human. Although, different species of *Alistipes*, were identified from both healthy and diseased individuals; no comparative studies of their characteristics and pathogenesis, on the available genomes was done previously. Therefore, the study was focused on identifying the diversity of all the genes via pan-genome analysis as well as characterizing the antimicrobial resistance genes, virulence factor genes, and other modes of pathogenesis from the different genomes of *Alistipes* species.

The pan-genome analysis had classified all the genes from the 16 genomes of *Alistipes* species into four categories- core gene, soft-core gene, shell gene and cloud gene. Besides, the identification of the genes related to pathogenicity- antimicrobial resistance genes, virulence factor genes, toxin-antitoxin system genes, secretion system genes enabled to understand the host-pathogen interaction pattern of different *Alistipes* species. The combination of these two, helped to classify the virulent and non-virulent *Alistipes* species, as well as defining the genes that were shared by the multiple genomes of same species and different genomes of different species, along with identifying and classifying the genes that are unique to each genome.

Additional characteristic analysis, such as the identification of carbohydrate-active enzyme genes and identification of the presence of any phage region in the different *Alistipes* genomes, helped to understand the host-pathogen interaction from the perspective of carbohydrate utilization as well as the adaptability of different *Alistipes* strains and/or isolates to environment. Also, the identification of phage regions, further clarified the pathogenicity of different *Alistipes* species as these regions might allow insertion of foreign genetic elements like antibiotic resistance gene operon, virulence factor related genomic island, et cetera.

Based on the performed analysis, the pattern of pathogenesis and presence and/or absence of the pathogenesis associated genes, etc. were found mostly similar in the multiple genomes of the same species. The *Alistipes finegoldii CE91-St15* genome belonging to the *Alistipes finegoldii* species was found to be the most virulent genome across all the analyzed genomes of *Alistipes* species since antimicrobial resistance genes, virulence factor related genes, toxin-antitoxin related genes, secretion system genes, carbohydrate-active enzymes similar to other pathogens, and multiple

incomplete phage regions were identified in this genome alone. On the contrary, the species-*Alistipes putredinis* could be considered as non-virulence since no major virulence associated properties could be identified in the genome of this species.

To sum up, the outcome of this comparative pan-genome analysis is not only limited to understand the diversity of the genes present across different genomes of the *Alistipes* species and understand the pathogenic properties of this organism but also create potential opportunities of future applications such as preventative therapeutic interventions, phage therapy, genetic engineering and so on. The data generated from this analysis, that is the set of genes that are shared by different *Alistipes* species might be used to analyze other biological areas where this bioinformatic analysis will reduce the time of laboratory-based analysis as well as will increase the accuracy of results. Also, reverse vaccinology approaches can be taken from the comparison-based analyzed data of different strains and/or isolates of *Alistipes* species. Additionally, the same information might help to choose the right drug and/or antibiotic to treat diseases associated with different *Alistipes* species. Nonetheless, the success and accuracy of this type analysis is totally dependent on the knowledge and right choice of software, relevant bioinformatic tools and algorithms, and last but not the least, knowledge about setting different parameters to find out the best optimum results from each tool and/or software respectively.

References

[1] Parker, B. J., Wearsch, P. A., Veloo, A., & Rodriguez-Palacios, A. (2020). The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health. Frontiers in immunology, 11, 906. https://doi.org/10.3389/fimmu.2020.00906

[2] Shkoporov, A. N., Chaplin, A. V., Khokhlova, E. V., Shcherbakova, V. A., Motuzova, O. V., Bozhenko, V. K., Kafarskaia, L. I., & Efimov, B. A. (2015). Alistipes inops sp. nov. and Coprobacter secundus sp. nov., isolated from human faeces. International journal of systematic and evolutionary microbiology, 65(12), 4580–4588. https://doi.org/10.1099/ijsem.0.000617

[3] Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database : the journal of biological databases and curation, 2020, baaa062. https://doi.org/10.1093/database/baaa062

[4] Pfleiderer, A., Mishra, A. K., Lagier, J. C., Robert, C., Caputo, A., Raoult, D., & Fournier, P. E. (2014). Non-contiguous finished genome sequence and description of Alistipes ihumii sp. nov. Standards in genomic sciences, 9(3), 1221–1235. https://doi.org/10.4056/sigs.4698398

[5] Moschen, A. R., Gerner, R. R., Wang, J., Klepsch, V., Adolph, T. E., Reider, S. J., Hackl, H., Pfister, A., Schilling, J., Moser, P. L., Kempster, S. L., Swidsinski, A., Orth Höller, D., Weiss, G., Baines, J. F., Kaser, A., & Tilg, H. (2016). Lipocalin 2 Protects from Inflammation and Tumorigenesis Associated with Gut Microbiota Alterations. Cell host & microbe, 19(4), 455–469. https://doi.org/10.1016/j.chom.2016.03.007

[6] Age-6 Coyne, M. J., & Comstock, L. E. (2019). Type VI Secretion Systems and the Gut Microbiota. Microbiology spectrum, 7(2), 10.1128/microbiolspec.PSIB-0009-2018. https://doi.org/10.1128/microbiolspec.PSIB-0009-2018

[7] Osbourn, A. E., & Field, B. (2009). Operons. Cellular and molecular life sciences : CMLS, 66(23), 3755–3775. https://doi.org/10.1007/s00018-009-0114-3

[8] Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., & Collado-Vides, J. (2000). Operons in Escherichia coli: genomic analyses and predictions. Proceedings of the National Academy of Sciences of the United States of America, 97(12), 6652–6657. https://doi.org/10.1073/pnas.110147297

[9] Pantosti, A., Sanchini, A., & Monaco, M. (2007). Mechanisms of antibiotic resistance in Staphylococcus aureus. Future microbiology, 2(3), 323–334. https://doi.org/10.2217/17460913.2.3.323

[10] Piddock L. J. (2006). Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. Clinical microbiology reviews, 19(2), 382–402. https://doi.org/10.1128/CMR.19.2.382-402.2006

[11] Alvarez-Ortega, C., Olivares, J., & Martínez, J. L. (2013). RND multidrug efflux pumps: what are they good for?. Frontiers in microbiology, 4, 7. https://doi.org/10.3389/fmicb.2013.00007

[12] Galán, J. E., & Waksman, G. (2018). Protein-Injection Machines in Bacteria. Cell, 172(6), 1306–1318. https://doi.org/10.1016/j.cell.2018.01.034

[13] Arechaga, I., & Cascales, E. (2022). Editorial: Bacterial Secretion Systems, Volume II. Frontiers in microbiology, 13, 917591. https://doi.org/10.3389/fmicb.2022.917591

[14] Rapisarda, C., & Fronzes, R. (2018). Secretion Systems Used by Bacteria to Subvert Host Functions. Current issues in molecular biology, 25, 1–42. https://doi.org/10.21775/cimb.025.001

[15] Hui, X., Chen, Z., Zhang, J., Lu, M., Cai, X., Deng, Y., Hu, Y., & Wang, Y. (2021). Computational prediction of secreted proteins in gram-negative bacteria. Computational and structural biotechnology journal, 19, 1806–1828. https://doi.org/10.1016/j.csbj.2021.03.019

[16] Green, E. R., & Mecsas, J. (2016). Bacterial Secretion Systems: An Overview. Microbiology spectrum, 4(1), 10.1128/microbiolspec.VMBF-0012-2015. https://doi.org/10.1128/microbiolspec.VMBF-0012-2015

[17] Chandran Darbari, V., & Waksman, G. (2015). Structural Biology of Bacterial Type IV Secretion Systems. Annual review of biochemistry, 84, 603–629. https://doi.org/10.1146/annurev-biochem-062911-102821

[18] Grossman, A. S., Escobar, C. A., Mans, E. J., Mucci, N. C., Mauer, T. J., Jones, K. A., Moore, C. C., Abraham, P. E., Hettich, R. L., Schneider, L., Campagna, S. R., Forest, K. T., & Goodrich-Blair, H. (2022). A Surface Exposed, Two-Domain Lipoprotein Cargo of a Type XI Secretion System Promotes Colonization of Host Intestinal Epithelia Expressing Glycans. Frontiers in microbiology, 13, 800366. https://doi.org/10.3389/fmicb.2022.800366

[19] Bhoite, S., van Gerven, N., Chapman, M. R., & Remaut, H. (2019). Curli Biogenesis: Bacterial Amyloid Assembly by the Type VIII Secretion Pathway. EcoSal Plus, 8(2), 10.1128/ecosalplus.ESP-0037-2018. https://doi.org/10.1128/ecosalplus.ESP-0037-2018

[20] Boschiroli, M. L., Ouahrani-Bettache, S., Foulongne, V., Michaux-Charachon, S., Bourg, G., Allardet-Servent, A., Cazevieille, C., Lavigne, J. P., Liautard, J. P., Ramuz, M., & O'Callaghan, D. (2002). Type IV secretion and Brucella virulence. Veterinary microbiology, 90(1-4), 341–348. https://doi.org/10.1016/s0378-1135(02)00219-5

[21] Jurėnas, D., Fraikin, N., Goormaghtigh, F., & Van Melderen, L. (2022). Biology and evolution of bacterial toxin-antitoxin systems. Nature reviews. Microbiology, 20(6), 335–350. https://doi.org/10.1038/s41579-021-00661-1

[22] Van Melderen, L., & Saavedra De Bast, M. (2009). Bacterial toxin-antitoxin systems: more than selfish entities?. PLoS genetics, 5(3), e1000437. https://doi.org/10.1371/journal.pgen.1000437

[23] Hayes F. (2003). Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. Science (New York, N.Y.), 301(5639), 1496–1499. https://doi.org/10.1126/science.1088157

[24] Yamaguchi, Y., Park, J. H., & Inouye, M. (2011). Toxin-antitoxin systems in bacteria and archaea. Annual review of genetics, 45, 61–79. https://doi.org/10.1146/annurev-genet-110410-132412

[25] Fernández-García, L., Blasco, L., Lopez, M., Bou, G., García-Contreras, R., Wood, T., & Tomas, M. (2016). Toxin-Antitoxin Systems in Clinical Pathogens. Toxins, 8(7), 227. https://doi.org/10.3390/toxins8070227

[26] Alonso J. C. (2021). Toxin-Antitoxin Systems in Pathogenic Bacteria. Toxins, 13(2), 74. https://doi.org/10.3390/toxins13020074

[27] Bunker, R. D., McKenzie, J. L., Baker, E. N., & Arcus, V. L. (2008). Crystal structure of PAE0151 from Pyrobaculum aerophilum, a PIN-domain (VapC) protein from a toxin-antitoxin operon. Proteins, 72(1), 510–518. https://doi.org/10.1002/prot.22048

[28] Gaba, S., Kumari, A., Medema, M., & Kaushik, R. (2020). Pan-genome analysis and ancestral state reconstruction of class halobacteria: probability of a new super-order. Scientific reports, 10(1), 21205. https://doi.org/10.1038/s41598-020-77723-6

[29] Wu, Y., Zaiden, N., & Cao, B. (2018). The Core- and Pan-Genomic Analyses of the Genus Comamonas: From Environmental Adaptation to Potential Virulence. Frontiers in microbiology, 9, 3096. https://doi.org/10.3389/fmicb.2018.03096

[30] Vernikos, G. S. (2020). A Review of Pangenome Tools and Recent Studies. In H. Tettelin (Eds.) et. al., The Pangenome: Diversity, Dynamics and Evolution of Genomes. (pp. 89–112). Springer.

[31] Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. Current opinion in microbiology, 23, 148–154. https://doi.org/10.1016/j.mib.2014.11.016

[32] Costa, S. S., Guimarães, L. C., Silva, A., Soares, S. C., & Baraúna, R. A. (2020). First Steps in the Analysis of Prokaryotic Pan-Genomes. Bioinformatics and biology insights, 14, 1177932220938064. https://doi.org/10.1177/1177932220938064

[33] Beier, S., Thomson, N.R. Panakeia - a universal tool for bacterial pangenome analysis. BMC Genomics 23, 265 (2022). https://doi.org/10.1186/s12864-022-08303-3

[34] Chen, X., Zhang, Y., Zhang, Z., Zhao, Y., Sun, C., Yang, M., Wang, J., Liu, Q., Zhang, B., Chen, M., Yu, J., Wu, J., Jin, Z., & Xiao, J. (2018). PGAweb: A Web Server for Bacterial Pan-Genome Analysis. Frontiers in microbiology, 9, 1910. https://doi.org/10.3389/fmicb.2018.01910

[35] Stefanski, Katherine & Gardner, Grant & Seipelt-Thiemann, Rebecca. (2016). Development of a Lac Operon Concept Inventory (LOCI). Cell Biology Education. 15. ar24-ar24. 10.1187/cbe.15-07-0162.

[36] Askoura, Momen & Mottawea, Walid & Abujamel, Turki & Taher, Ibrahim. (2011). Efflux pump inhibitors (EPIs) as new antimicrobial agents against Pseudomonas aeruginosa. The Libyan journal of medicine. 6. 10.3402/ljm.v6i0.5870.

[37] Zhang W, Rong C, Chen C, Gao GF (2012) Type-IVC Secretion System: A Novel Subclass of Type IV Secretion System (T4SS) Common Existing in Gram-Positive Genus Streptococcus. PLOS ONE 7(10): e46390. https://doi.org/10.1371/journal.pone.0046390

[38] Azam, M.W., Zuberi, A. & Khan, A.U. *bolA* gene involved in curli amyloids and fimbriae production in E. coli: exploring pathways to inhibit biofilm and amyloid formation. J of Biol Res-Thessaloniki 27, 10 (2020). https://doi.org/10.1186/s40709-020-00120-7

[39] Sikora, Aleksandra. (2013). Proteins Secreted via the Type II Secretion System: Smart Strategies of Vibrio cholerae to Maintain Fitness in Different Ecological Niches. PLoS pathogens. 9. e1003126. 10.1371/journal.ppat.1003126

[40] Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2022 Dec 12]. Available from: https://www.ncbi.nlm.nih.gov/data-hub/genome/?taxon=239759

[41] The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, doi:10.1093/nar/gkac247

[42] Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England), 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

[43] Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics (Oxford, England), 31(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

[44] Blanca Taboada, Karel Estrada, Ricardo Ciria, Enrique Merino, Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes, Bioinformatics, Volume 34, Issue 23, 01 December 2018, Pages 4118–4120, https://doi.org/10.1093/bioinformatics/bty496

[45] Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., … McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research, 48(D1), D517–D525. https://doi.org/10.1093/nar/gkz935

[46] Liu, B., Zheng, D., Jin, Q., Chen, L., & Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic acids research, 47(D1), D687–D692. https://doi.org/10.1093/nar/gky1080

[47] Frédéric Lemoine, Damien Correia, Vincent Lefort, Olivia Doppelt-Azeroual, Fabien Mareuil, Sarah Cohen-Boulakia, Olivier Gascuel, NGPhylogeny.fr: new generation phylogenetic services for non-specialists, Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W260–W265, https://doi.org/10.1093/nar/gkz303

[48] Ivica Letunic, Peer Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W293–W296, https://doi.org/10.1093/nar/gkab301

[49] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, R. Lopez, InterProScan: protein domains identifier, Nucleic Acids Research, Volume 33, Issue suppl_2, 1 July 2005, Pages W116–W120, https://doi.org/10.1093/nar/gki442

[50] Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology, 302(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

[51] Wallace, I. M., O'Sullivan, O., Higgins, D. G., & Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic acids research, 34(6), 1692–1699. https://doi.org/10.1093/nar/gkl091

[52] Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic acids research, 46(W1), W95–W101. https://doi.org/10.1093/nar/gky418

[53] Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research, 44(W1), W16–W21. https://doi.org/10.1093/nar/gkw387