DEVELOPING A COMPUTATIONAL PIPELINE FOR DETECTING MULTI-FUNCTIONAL ANTIBIOTIC RESISTANCE GENES IN METAGENOMICS DATA

Ngoc Khoi Dang

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science and Applications

Liqing Zhang, Chair

Anuj Karpatne

Ismini Lourentzou

May 6, 2022

Blacksburg, Virginia

Keywords: multi-functional, antibiotic resistance genes, metagenomics

DEVELOPING A COMPUTATIONAL PIPELINE FOR DETECTING MULTI-FUNCTIONAL ANTIBIOTIC RESISTANCE GENES IN METAGENOMICS DATA

Ngoc Khoi Dang

ABSTRACT

Antibiotic resistance is currently a global threat spanning clinical, environmental, and geopolitical research domains. The environment is increasingly recognized as a key node in the spread of antibiotic resistance genes (ARGs), which confer antibiotic resistance to bacteria. Detecting ARGs in the environment is the first step in monitoring and controlling antibiotic resistance. In recent years, next-generation sequencing of environmental samples (metagenomic sequencing data) has become a prolific tool for the field of surveillance. Metagenomic data are nucleic acid sequences, or nucleotides, of environmental samples. Metagenomic sequencing data has been used over the years to detect and analyze ARGs. An intriguing instance of ARGs is the multi-functional ARG, where one ARG encodes two or more different antibiotic resistance functions. Multi-functional ARGs provide resistance to two or more antibiotics, thus should have evolutionary advantage over ARGs with resistance to single antibiotic. However, there is no tool readily available to detect these multi-functional ARGs in metagenomic data. In this study, we develop a computational pipeline to detect multi-functional ARGs in metagenomic data. The pipeline takes raw metagenomic data as the input and generates a list of potential multi-functional ARGs. A plot for each potential multi-functional ARG is also created, showing the location of the multi-functionalities in the sequence and the sequencing coverage level. We collected samples from three different

sources: influent samples of a wastewater treatment plant, hospital wastewater samples, and reclaimed water samples, ran the pipeline, and identified 19, 57, and 8 potentially bi-functional ARGs in each source, respectively. Manual inspection of the results identified three most likely bi-functional ARGs. Interestingly, one bi-functional ARG, encoding both aminoglycoside and tetracycline resistance, appeared in all three data sets, indicating its prevalence in different environments. As the amount of antibiotics keeps increasing in the environment, multi-functional ARGs might become more and more common. The pipeline will be a useful computational tool for initial screening and identification of multi-functional ARGs in metagenomic data.

DEVELOPING A COMPUTATIONAL PIPELINE FOR DETECTING MULTI-FUNCTIONAL ANTIBIOTIC RESISTANCE GENES IN METAGENOMICS DATA

Ngoc Khoi Dang

GENERAL AUDIENCE ABSTRACT

Antibiotics are the drug to fight against the infection of bacteria. Since the first antibiotic was discovered in 1928, many antibiotic drugs have been developed. At the same time, scientists discovered many genes responsible for the resistance of antibiotic drugs. Nowadays, antibiotic resistance is a global threat. Detecting antibiotic resistance genes in the environment is the first step towards monitoring and controlling antibiotic resistance. In recent years, next-generation sequencing has been widely used to get the DNA sequence from environment. Metagenomics analysis has been used over the years to detect and analyze ARGs. In the literature, it has been reported that a single gene could carry two parts of sequences corresponding to two different ARGs, thus conferring resistance to two different antibiotics. This fusion might have some evolutionary advantage. In this study, we developed a novel computational tool to detect multi-functional ARGs. We collected data from three sources: the treatment plant water, the hospital wastewater, and the reclaimed water, and identified 19, 57, and 8 potential bi-functional ARGs in each source, respectively. After we manually inspected the result, we found three most likely bi-functional ARGs. We also found one bi-functional ARG that appears in all three datasets. The gene is responsible for aminoglycoside and tetracycline resistance. The tool will serve as the initial screening step to detect multi-functional ARGs.

Dedicated to Virginia Tech.

Acknowledgments

I would like to express my deepest appreciation to my advisor, Dr. Liqing Zhang for her guidance and constant support during my study at Virginia Tech. Her encouragement and insightful suggestions greatly guided my research. I would also like to extend my deepest gratitude to Connor Brown for his valuable time and helpful comments. I also wish to thank Dr. Ismini Lourentzou and Dr. Anuj Karpatne for being my professors and serving as my committee members.

I am thankful to all of my lab members Justin Sein, Badhan Das, Nazifa Ahmed Moumi, Monjura Afrin Rumi, Muhit Islam Emon, and Joung Min Choi for their assistance during my research.

I also want to thank the Department of Computer Science at Virginia Tech for allowing me to work in the bioinformatics area.

As always, I would like to thank my family Đặng Ngọc Sơn, Trần Thị Ánh Trăng, Đặng Trần Khôi Nguyên, Đặng Hồng Ân, Võ Ngọc Bảo Trân, Võ Ngọc Bảo Vi and Lưu Phước Hóa for their unconditional love and support over the years.

Contents

List of Figures

List of Tables

List of Abbreviations

ARC Advanced Research Computing

ARGs Antibiotic resistance genes

CARD Comprehensive Antibiotic Resistance Database

CIWARS Cyberinfrastructure for Waterborne Antibiotic Resistance Risk Surveillance

DNA Deoxyribonucleic acid

NCBI The National Center for Biotechnology Information

ORF Open reading frame

SRA Sequence Read Archive

Chapter 1

Introduction

Since the discovery of the first antibiotic penicillin in 1928 [**?**], antibiotics have been first-line defenses against bacterial infection. Many antibiotic drugs have been discovered and classified into separate groups based on their antimicrobial mechanisms such as β-lactam (such as penicillin and carbapenems), aminoglycosides (e.g., gentamicin and streptomycin), macrolides (e.g., erythromycin and azithromycin), fluoroquinolones (e.g., Ciprofloxacin, levofloxacin), and glycopeptides (e.g., Vancomycin) [**?**]. Resistance to these antibiotics has been substantiated by overuse and misuse of these compounds. Thus, antibiotic resistance is now a global threat that requires exigent attention. [**?**]. In the presence of these antibiotics, the bacteria developed "smart" mechanisms to protect themselves. The antibiotic resistance genes (ARGs) in the organism are activated when the medium exists certain antibiotic compounds. These genes can inactivate the drug, limit the uptake of drugs, enhance the efflux of the drug, or modify the target proteins which neutralize the activity of antibiotic drugs. These genes can be transferred vertically or horizontally. Detecting ARGs in the environment can help researchers understand the resistance profile of micro-organisms in the environment. It has been found that antimicrobial resistance is a natural self-defensive mechanism of bacteria and ARGs exist in the bacteria living in areas that have never had contact with antibiotic drugs [**?**].

Since 2010, studies have shown that some ARGs have been found to have regions of homology to multiple drug resistance protein families and demonstrate multiple resistance functions

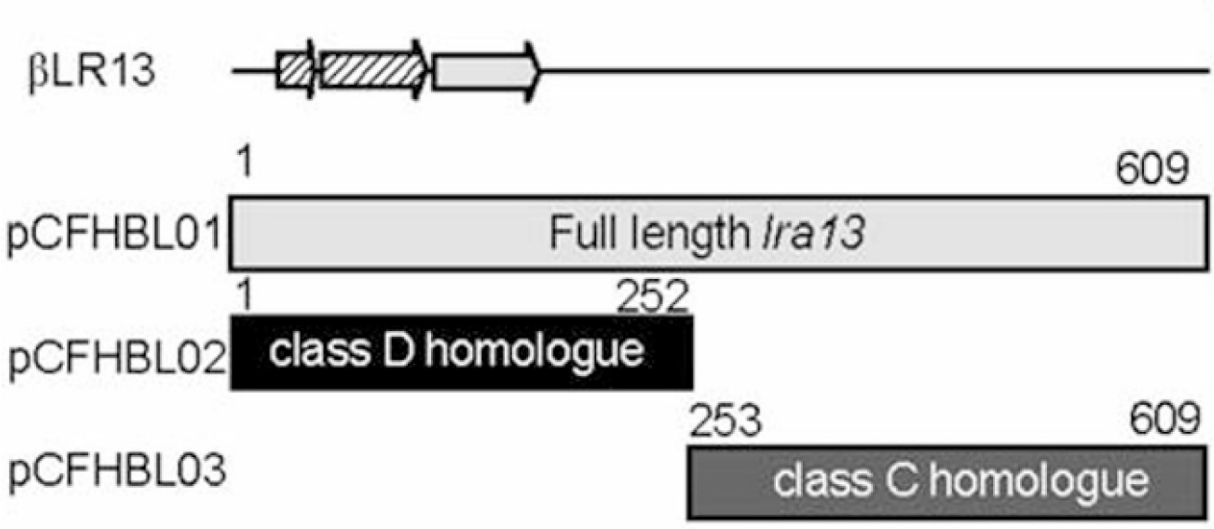

Figure 1.1: Example of a bi-functional ARG βLR13 (figure taken from the paper [**?**]). The protein has 609 amino acids. The C-terminus (356 amino acids) encodes a class C β-lactamase and the N-terminus (253 amino acids) a class D β-lactamase. pCFHBL01, pCFHBL02, and pCFHBL03 are clone IDs.

during experimental function validation (e.g., bi-functional ARGs [**?**]). As an example shown in Figure 1.1, the gene has a unusual long open reading frame (ORF). Each end of the ORF encodes a distinct resistance type. In the literature, we identified seven experimentally confirmed bi-functional ARGs [**?** **?** **?** **?** **?** **?** **?**], shown in Table 2.2. It is noted that even though these ARGs have two different resistance functions, the resistances are still in the same major antibiotic resistance group such as β-lactamase or transferase. No multi-functional ARGs with different major resistance classes have been discovered. As gene fusion events are known to contribute to the evolution of multi-functional bacterial proteins [**?**], it is unclear whether or not there are other types of multi-functional ARGs and how prevalent multi-functional ARGs are in the environment. Data mining of metagenomic sequencing data can be used to detect putative multi-functional ARGs and provide insights into these important yet unexplored questions. This study addresses a key gap in the breadth of bioinformatic tools for analyzing ARGs in metagenomics as there is no tool for detecting multi-functional ARGs at present.

Name	Year	Species	Antibiotic class	Resistance enzyme	Functions
Tp47	2004	T.palladium	β-lactam	β-lactamase	Penicillin binding protein and β-lactamase
βla_{LRA13}	2009	E.coli	β-lactam	β-lactamase	class C and class D β-lactamase
AAC(6')/ APH(2")	1986	E.faecalis, S.aureus	aminoglycoside	transferase	6'-N-aminoglycoside acetyltransferase and 2"-O-amino-glycoside phosphotransferase
AAC(3)-Ib/AAC(6')-Ib'	2006	P.aeruginosa	aminoglycoside	transferase	6'-N-aminoglycoside acetyltransferase and 2"-O-amino-glycoside phosphotransferase
ANT(3")-Ii/AAC(6')-IId	2002	P.aeruginosa	aminoglycoside	transferase	6'-N-aminoglycoside acetyltransferase and 2"-O-amino-glycoside phosphotransferase
AAC(6')-30/AAC(6')-Ib'	2004	P.aeruginosa	aminoglycoside	transferase	6'-N-aminoglycoside acetyltransferase and 2"-O-amino-glycoside phosphotransferase
AAC(6')-Ib-cr	2008	T.palladium	fluoroquinolone	N/A	norfloxacin resistance and ciprofloxacin resistance

Table 1.1: Current known bi-functional ARGs in literature

Thanks to the next generation sequencing techniques, nowadays researchers can generate a huge amount of DNA sequences in a short period of time at a low cost. Metagenomics is a technique to capture all the DNA sequences existing in a sample at a specific time. Metagenomics data could be collected independently from cultures and retrieve all DNAs in the sample without bias. Currently, shotgun sequencing is used widely to detect and quantify ARGs in the environment [**?**].

The first step in the metagenomic analysis is cleaning the data. It includes removing the adapters which are attached during sequencing and removing contaminated sequences from

other uninteresting species. After that, the clean short-read (200 base pairs) is assembled together into longer sequences (contigs). To verify the process, the coverage is calculated by mapping the short reads back to contigs. In the ideal case, the coverage plot should have a uniform distribution. Fig. 1.2.

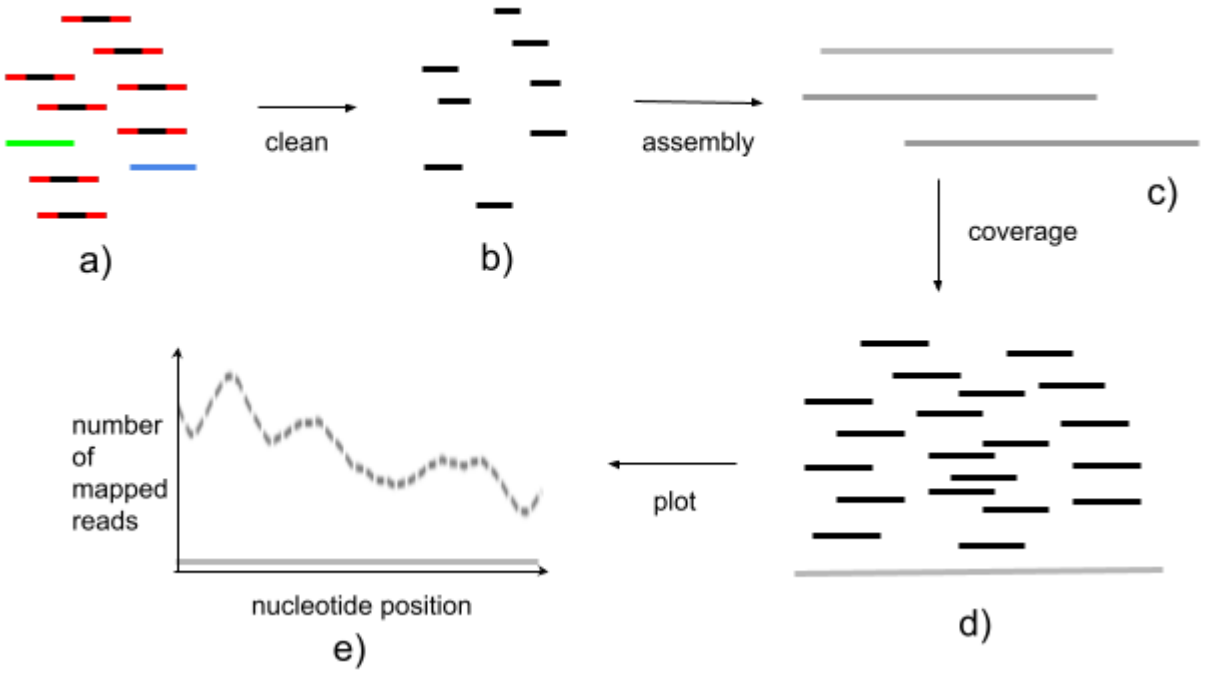

Figure 1.2: Basic steps in metagenomics analysis. a) Each line represents one read in the data. Red segments are adaptors. Green and blue lines are contaminated reads. b) All adaptors segments and contaminated reads are removed. c) Short reads are combined into longer reads (contigs) by assembler. d) The short reads are mapped back to contigs to calculate the coverage. e) Coverage visualization

In this study, we introduce a novel pipeline to detect the potential multi-functional ARGs from short-read sequences.

Chapter 2

Materials and Methods

2.1 Datasets

Data is collected from three different sources: wastewater treatment plant water, reclaimed water, and hospital wastewater. The wastewater treatment plan data is derived from a local wastewater treatment plant (hereafter referred to as the "CIWARS" samples). This project is the collaboration of Zhang Lab, Computer Science department, and Pruden Laboratory, Department of Civil and Environmental Engineering, at Virginia Tech which sampled influent and effluent wastewater in the Christiansburg, Virginia wastewater treatment plan from 2020 to 2021.

We get reclaimed water data from the ongoing project "Critical Barriers to Antibiotic Resistance during Water Reclamation and Reuse" in 2020 by Pruden Laboratory at Virginia Tech.

Lastly, the hospital sewage water data is obtained from a published research at Yeungnam University, Seoul in 2021 [**?**].

The accession numbers or file names, size of uncompressed files, and number of reads of these data files are shown in Table 2.1. The data from CIWARS project is currently stored at ARC server at Virginia Tech. The last two datasets are downloaded from the SRA. The bioproject accession in SRA of from Hospital wastewater data is PRJNA784332. The

bioproject accession in SRA of Reclaimed wastewater data is PRJNA669820.

Source	Accession number	Size (Gb)	Number of reads
CIWARS	Y20_M10_D12_EFF_S36_L001	19	26,951,055
CIWARS	Y20_M10_D16_EFF_S44_L001	18	25,726,121
CIWARS	Y20_M10_D19_EFF_S52_L001	17.2	24,547,589
CIWARS	Y21_M5_D17_EFF_S70_L001	16.4	23,287,756
CIWARS	Y21_M5_D21_EFF_S78_L002	12.8	18,072,355
CIWARS	Y21_M5_D28_EFF_S104_L001	18.6	26,382,296
Hospital wastewater	SRR17068071	37	52,732,927
Hospital wastewater	SRR17068072	37	50,963,293
Reclaimed water	SRR12900975	19.0	26,381,531
Reclaimed water	SRR12900976	11.6	16,137,090
Reclaimed water	SRR12900977	11.0	15,150,993
Reclaimed water	SRR12900978	13.6	18,848,916

Table 2.1: CIWARS dataset, Hospital wastewater dataset and Reclaimed water dataset

2.2 Methods

The pipeline includes pre-processing the raw data, assembling short-reads into contigs, finding ORFs in the contigs, aligning ORFs against the CARD database to annotate the ORFs, clustering aligned sequences within the pre-defined distance to detect multi-functional clusters and visualizing the clusters. A cluster represents a putative multi-functional ARG.

First, low-quality read and adaptors are removed from the raw short-read sequences using FASTP [**?**]. All the reads having a quality score lower than 10 or poly-G or poly-X are removed. We also enable the low-complexity-filter to remove low-complexity reads. After that, we apply BBDUK [**?**] to remove contaminated reads. All the reads from human, rat, and dog reference genomes are filtered out. The clean reads are assembled into contigs using MEGAHIT with meta-large option that uses multiple k-mer values to configure MEGAHIT [**?**]. From the contigs results, we get the coverage for the data using minimap2 [**?**] and

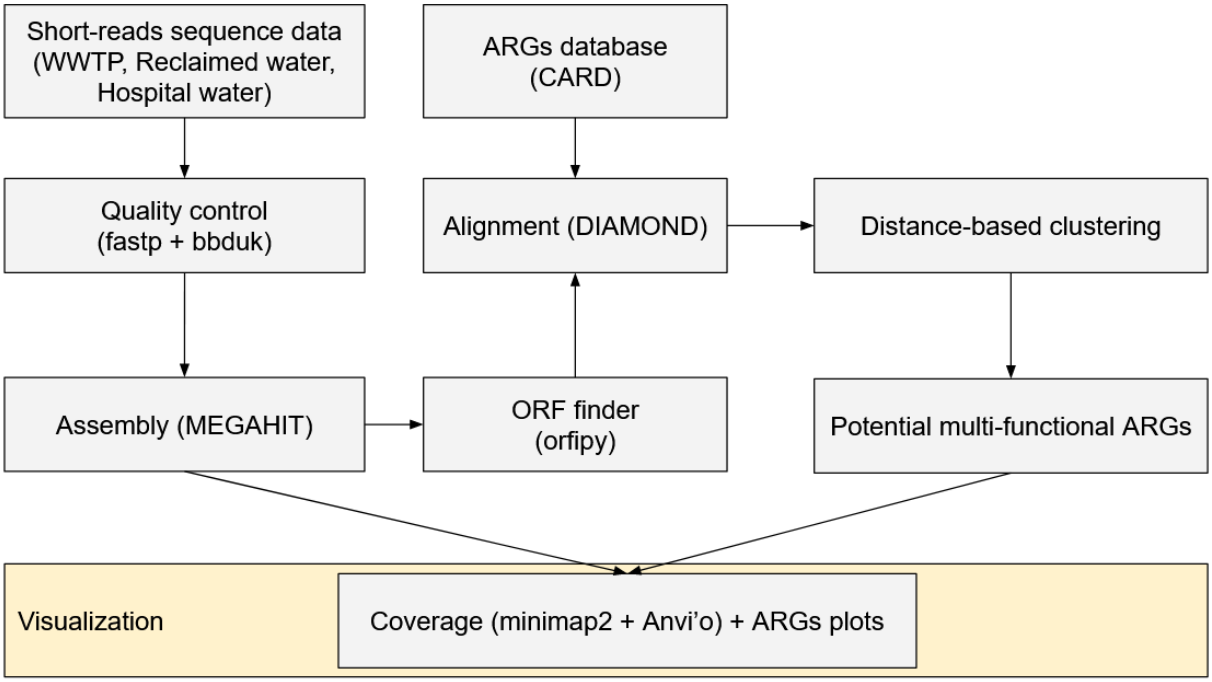

Figure 2.1: Pipeline: The raw short-read sequences are cleaned by FASTP and BBDUK. The assembly MEGAHIT assembles reads into contigs. The ORFs are defined from the contigs and aligned against the CARD database. All the aligned sequences are grouped if they have many different types of ARGs and are located very close to each other. The coverage of contigs is calculated. Then, all the results are plotted for visualization.

anvi'o [**?**].

For each contig, ORFIPY [**?**] is applied to find the ORFs present on each contig. After the ORFs are found, the ORFs are aligned against the CARD database [**?**] using DIAMOND [**?**] to detect ARGs on each ORF. DIAMOND is set in a sensitive mode which is designed to find all hits with the minimum of 40% identity. We also set the percentage of subject cover in DIAMOND to 70% to ensure the quality of the alignments. The aligned sequences are clustered by a distance-based clustering algorithm. Only the clusters having two or more different antibiotic resistance functions are kept. They are the potential multi-functional ARGs. Finally, all the selected clusters are plotted. The coverage plot and cluster plot are used as the initial screening clues for the biologist to validate the multi-functional ARGs.

The pipeline is demonstrated at Figure 2.1.

The distance-based clustering algorithm groups the closely aligned sequences together. First, it selects all the aligned sequences of a specific query sequence. All the aligned sequences are sorted by the start location of that aligned sequence in the query sequence. Then, the first aligned sequence is labeled as the first cluster. The end position of a cluster is defined as the maximum end position of all members in the cluster. If the distance of the start position of the next aligned sequence and the end position of a cluster is smaller than D distance. That next aligned sequence is added to the cluster and the end position of the cluster is updated. To avoid overlapping, the new aligned sequence must not overlap more than L percentage of the cluster. Whenever the algorithm fails to find a new member for the cluster, the cluster is recorded and the next available aligned sequence is assigned to a new cluster. The program runs for all aligned sequences. After the cluster is collected, we check its validity of the cluster. The size of the cluster must be larger than 1, and the cluster must contain at least 2 different ARGs from all the members. In this study, we set D to 20 and L to 10%.

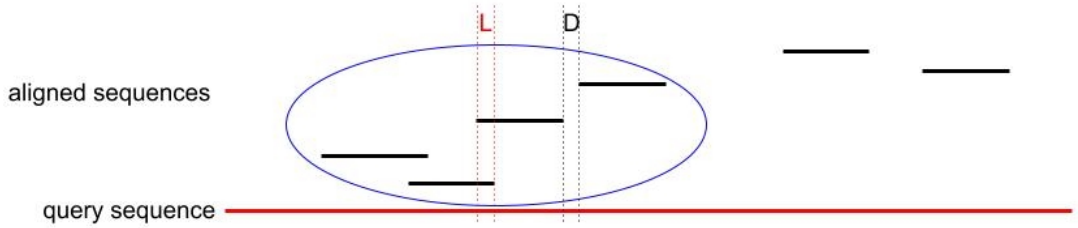

Figure 2.2: Distance-based clustering: The blue circle shows one cluster. The distance of 2 consecutive aligned sequences must less than distance D and the overlapping region must less than L percent. A cluster represents a potential multi-functional ARGs

The source code is implemented in Python and available at https://github.com/khoidnyds/bifunc.

Chapter 3

Results

From the CIWARS dataset, we found 5,334,703 contigs, 8,649,183 ORFs, 217,110 aligned sequences and 19 potential multi-functional ARGs from six different samples.

From the Hospital wastewater dataset, we found 2,374,309 contigs, 7,066,256 ORFs, 272,050 aligned sequences and 57 potential multi-functional ARGs. Finially, from Reclaimed water dataset, we found 2,365,096 contigs, 2,605,122 ORFs, 75,809 aligned sequences and eight potential multi-functional ARGs. The results are summarized in the table 3.1

	CIWARS	Hospital waste water	Reclaimed water
Number of reads	144,967,172	103,696,220	76,518,530
Number of contigs	5,334,703	2,374,309	2,365,096
Assembly size (bp)	3,341,433,922	2,213,679,336	1,438,226,446
N50 (bp)	653	1,400	608
ORFs	8,649,183	7,066,256	2,605,122
Number of aligned sequences	217,110	272,050	75,731
Number of clusters	19	57	8

Table 3.1: Results from the pipeline of 3 different datasets.

We manually investigate each multi-functional ARG in the results. All multi-functional ARGs have two different functions. Based on the description in CARD database, We remove transcription genes, regulation genes, activation genes, genes in a two-component system and sensor genes. The list of removed hits is attached in additional files. After the removal, we have six, 16 and three bi-functional ARGs left from CIWARS, hospital wastewater and reclaimed water, respectively. 3.1.

Finally, we run BLASTN for each bi-functional ARGs against the non-redundant protein sequence database to check for the presence of these genes in the public database. We found three unique bi-functional ARGs that match to existing genes in the BLAST database. The list of bi-functional ARGs is in the Table 3.2. All of these bi-functional ARGs have the function similar to RanA protein which is a part of ABC-type efflux system resistant to aminoglycoside antibiotic 3.2.

Bi-functional ARGs	Source	Gene 1	Gene 2
1	CIWARS	otrC - a tetracycline resistance efflux pump	RanA - a aminoglycoside resistance ABC-type efflux system
2	CIWARS, hospital wastewater, reclaimed water	tetA - a tetracycline efflux pump	RanA - a aminoglycoside resistance ABC-type efflux system
3	Hospital wastewater	oleC - an ABC transporter	RanA - a aminoglycoside resistance ABC-type efflux system

Table 3.2: 3 plausible potential bi-functional ARGs

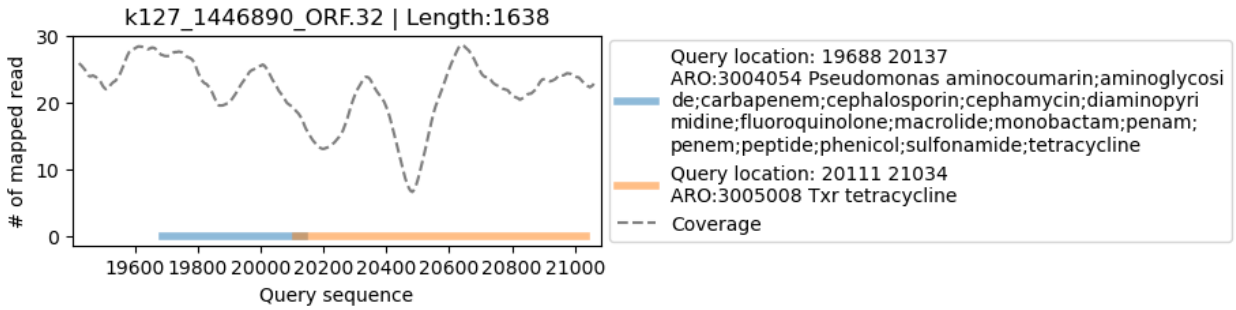

Figure 3.1: A potential bi-functional ARGs. The query sequence is an ORF in contig k127_1446890 having 1638 nucleotides from the hospital dataset. The orange line is an aligned sequence that is homologous alignment to ARO:3005008, a tetracycline antibiotic gene in the CARD database. The blue line is another aligned sequence that is homologous alignment to ARO:3004054, a aminoglycoside antibiotic gene in the CARD database. The location in query sequence and target sequence also is shown in the legend. The black dashed line shows the coverage in this area.

Figure 3.2: BLAST of k127_1446890_ORF.32. The red lines represent the hit with very high alignment scores (>200) from the non-redundant proteins BLAST database to our ORF sequence. This alignment makes our ARG a plausible potential bi-functional gene. All aligned sequences are the members of AtoC superfamily protein. AtoC superfamily is DNA-binding transcriptional response regulator, NtrC family, contains REC, AAA-type ATPase, and a Fis-type DNA-binding domains [Signal transduction mechanisms].

Interestingly, we find one bi-functional ARG that appears in all three datasets. This bi-functional ARG mapped to two different ARGs in the CARD database which are RanA and tetA. Gene ranA encodes a protein that is part of an ABC-type efflux system resistant to aminoglycoside antibiotic. Gene tetA encodes a tetracyline efflux pump resistant to tetracyline antibiotic. Fig. 3.3, Fig. 3.4.

We perform another validation check for the ARGs that appear in all three samples. We run the paired-ends read coverage to check the correctness of assembly on these gene Fig. 3.5 using IGView [**?**]. In the overlap region, one read from a paired-end read mapped to one ARG and another read from the same paired-end read mapped to the other ARG. It

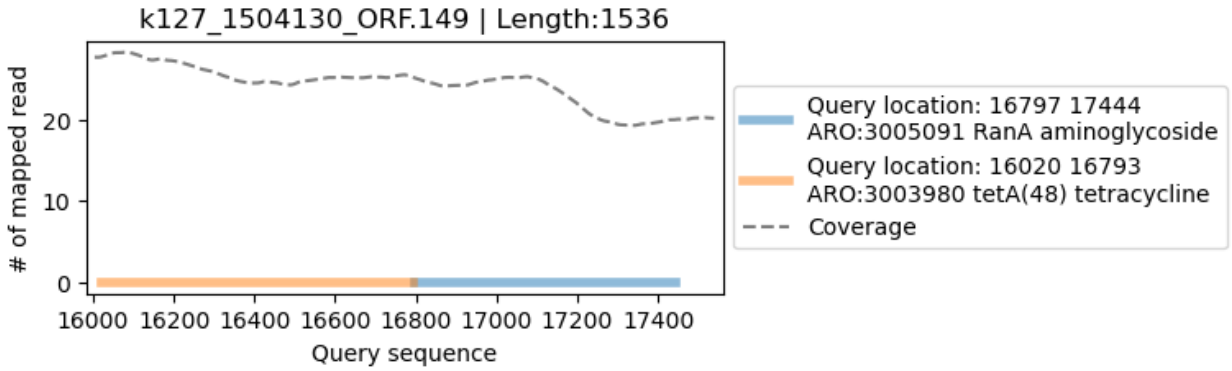

Figure 3.3: A potentially bi-functional ARG.This ARG appears in all three datasets. Similar to Fig:3.1. The query sequence is an ORF in contig k127_1504130. The orange line is gene ARO:3003980 which is a tetracycline antibiotic resistance gene. The blue line is gene ARO:3005091 which is a aminoglycoside antibiotic resistance gene.

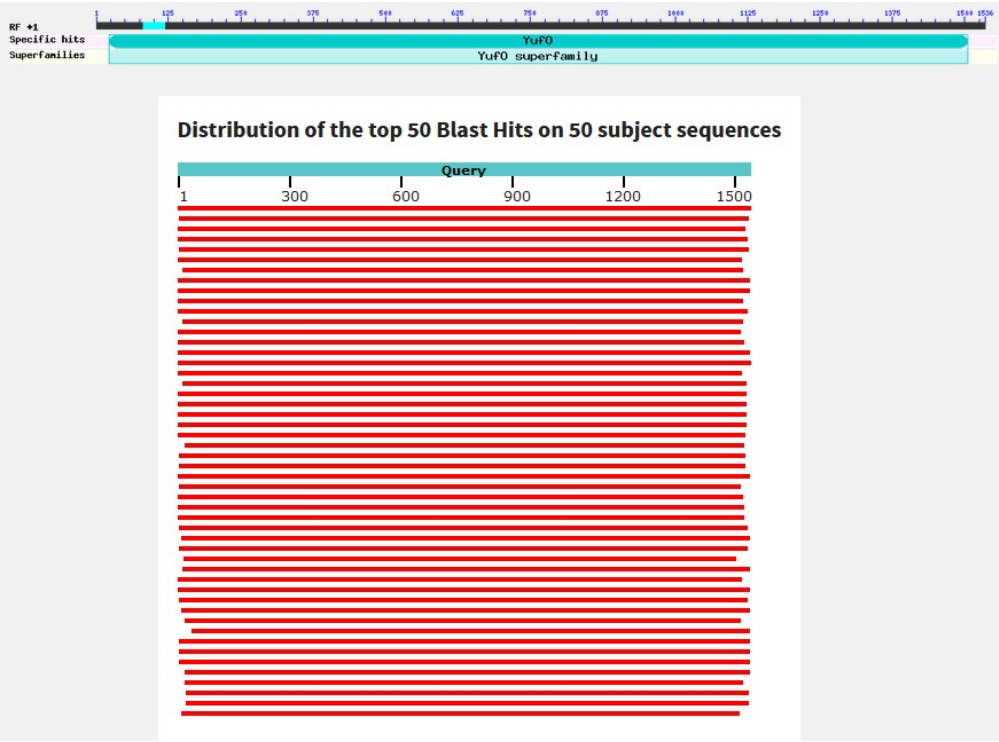

Figure 3.4: BLASTX of k127_1504130_ORF.149. All aligned sequences are the members of YufO superfamily protein. YufO is an uncharacterized ABC transporter ATP-binding protein.

strongly suggests that these two ARGs actually locate in a same region and thus are not the result of a misassembly.

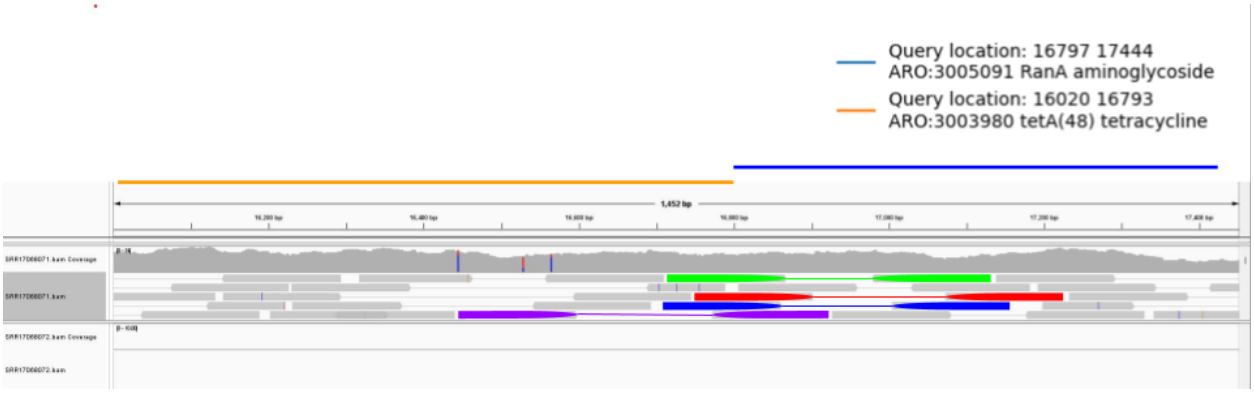

Figure 3.5: Paired-ends mapping of k127_1504130_ORF.149. The orange line is the aligned sequence which is homologous to ARO:3003980 encoded for tetracycline resistance. The blue line is the aligned sequence which is homologous to ARO:3005091 encoded for aminoglycoside resistance. The box below shows the coverage plot and paired-ends mapping of this location. Green, red, blue and purple colors represent 4 different paired-ends reads. Because of these pairs, we can confidently say that these 2 aligned sequences actually locate in a same region and therefor are not a result of misassembly.

Chapter 4

Discussions and Conclusions

From the table 3.1, The number of contigs found in the CIWARS dataset is much more than the other 2 datasets despite the size of the raw file of the CIWARS dataset being smaller. It suggests the CIWARS dataset is more complex than the other 2 datasets.

We notice that the number of clusters is found in the hospital wastewater sample is triple the number of clusters found in the wastewater treatment plant (CIWARS). We also note that the number of contigs in CIWARS is double the number of contigs in Hospital waste water. It suggests that the ratio of clusters found in Hospital water and CIWARS water should be much more than 3. On the contrary, reclaimed water contains very few clusters. The result is reasonable because the hospital wastewater probably contains more ARGs than the wastewater treatment plant water (CIWARS). And the Reclaimed water should contain the least ARGs.

Most of the components of our pipeline are the existing tools in the field. We connected these tools with proper configuration to build a completed metagenomics analysis pipeline. We defined the clustering algorithm step based on the need of clustering the neighbor-aligned sequences. This algorithm solely grouped nearby aligned sequences with 2 arguments to control the overlapping and allowing distance between members in the cluster.

We also build the visualization by gathering the alignment information as well as coverage information and putting them in a single plot. These plots are easy to understand and

quickly show the possible multi-functional ARGs with their positions in the query sequence.

After we manually inspect 84 bi-functional ARGs, we found 3 plausible potential bi-functional ARGs. They align well to single genes in public non-redundant protein database. Especially, one ARG appears in all 3 datasets. This ARG is composed of 2 genes: tetA and RanA. TetA(58) is a Tetracycline efflux pump described in Paenibacillus sp. LC231, a strain of Paenibacillus isolated from Lechuguilla Cave, NM, USA. This gene was first described by Pawlowski in 2016. RanA is a part of the RanARanB ABC-type efflux system. Alongside RanB, RanA confers resistance to aminoglycoside antibiotics [**?**]. Furthermore, the paired-ends mapping of this gene also shows that the gene is assembled correctly.

There're many existing tools in the field to detect and classify single ARG but there's no tool trying to detect multi-functional ARGs. From this study, we develop a new tool to run the complete metagenomic pipeline to detect the multi-functional ARGs from the metagenomics data. The tool generates the plots that show the location of the detected multi-functional ARGs and the coverage level of the region in the sample. This tool can assist researchers to search and explore their metagenomic sequencing datasets to detect multi-functional ARGs in the environment.

This study opens to many deeper studies about multi-functional ARGs in the future. Beside the short-read sequencing data, we can use long-read sequencing data which contain significant longer sequences. This type of data doesn't require the assembly step which could exclude many false positives from the results. Another improvement is the clustering step, we can include other sequence features to enhance the quality of the representation sequences. Therefore, we could have more accurate ARGs positions and labels for each cluster.