농학석사 학위논문

Detection of horizontal gene transfer events using the tree-reconciliation method and the inter-taxon transfer trend of virulence factors and antibiotic resistance genes

계통수 조화법을 통한 유전자 이동 현상 규명과 병원성 인자 및 항생제 내성 유전자의 분류군 간 이동 경향

2023년 8월

서울대학교 대학원

농생명공학부 바이오모듈레이션 전공

최 영 석

A Thesis for the Degree of Master of Science

Detection of horizontal gene transfer events using the tree-reconciliation method and the inter-taxon transfer trend of virulence factors and antibiotic resistance genes

By

Youngseok Choi

Supervisor: Professor Hee-bal Kim

Aug, 2023

Major in Biomodulation

Department of Agricultural Biotechnology

Seoul National University

계통수 조화법을 통한 수평 유전자 이동 현상 규명과 병원성 인자 및 항생제 내성 유전자의 분류군 간 이동 경향

지도 교수 김 희 발

이 논문을 농학석사 학위논문으로 제출함

2023년 8월

서울대학교 대학원
농생명공학부 바이오모듈레이션 전공
최 영 석

최영석의 농학석사 학위논문을 인준함
2023년 6월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Abstract

Detection of horizontal gene transfer events using the tree-reconciliation method and the inter-taxon transfer trend of virulence factors and antibiotic resistance genes

Youngseok Choi

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Horizontal gene transfer (HGT) is a fundamental process in microbial evolution where genetic material is transferred between different organisms. It plays a significant role in shaping the genetic diversity and adaptive capabilities of microbial populations. With the advancement of bioinformatics techniques, studying HGT has become increasingly feasible and efficient. Bioinformatics tools enable the analysis of large-scale genomic data, identification of orthologous gene sets, construction of gene

and species trees, and the detection of HGT events. By integrating computational approaches with biological knowledge, researchers can uncover the intricate patterns and mechanisms of HGT, unraveling the genetic exchanges that contribute to the evolution and adaptation of microorganisms. The combination of bioinformatics methodologies and experimental validation has opened new avenues for studying HGT dynamics and its impact on the evolution of microbial genomes.

This paper summarizes two distinct chapters focusing on horizontal gene transfer (HGT) events and their implications in microbial evolution. Both studies aim to explore the occurrence and characteristics of HGT events, albeit with different emphases. The first paper introduces an improved method (the tree-reconciliation method) for detecting HGT events using gene trees and species trees, resulting in the development of an updated and enhanced database. The second paper examines the trends observed in inter-taxon HGT events related to virulence factors (VFs) and antibiotic resistance genes (ARGs), offering insights into the underlying mechanisms driving these events. Collectively, these studies contribute to our understanding of HGT dynamics and its impact on microbial genomes.

The first chapter of the research papers provides a comprehensive literature review on three key topics: horizontal gene transfer (HGT), the tree reconciliation method, and the significance of virulence factors (VFs) and antibiotic resistance genes (ARGs). The literature review explores the fundamental concept of HGT, a process critical to microbial evolution and the transfer of genetic material between organisms.

The second chapter focuses on detecting HGT events through the tree-

reconciliation method, utilizing gene trees and species trees. The study introduced the HGTree v2.0 database, an updated version of the previous HGTree database (constructed in 2015), in the purpose of coping up with the rapid rise of prokaryotic genome data since 2015. The update incorporates significantly larger datasets, with 5,405,040 putative HGT-related genes detected, improved visual interfaces, an enhanced user query tool, all of them with more reliable data curation procedures. By employing this method and leveraging the database improvements, researchers can more accurately identify and analyze HGT events, shedding light on the evolutionary dynamics of microbial genomes.

The third chapter investigates the trends observed in inter-taxon HGT events involving virulence factors (VFs) and antibiotic resistance genes (ARGs). The study revealed a notable disparity in the prevalence of ARGs compared to VFs in inter-genus to inter-phylum HGT events, with a higher occurrence of ARGs. Further analysis of the Cluster of Orthologous Groups (COG) categories demonstrated that essential COG categories related to bacterial survival are more frequently associated with inter-taxon HGT events, especially for ARGs. Additionally, the study associated various functional aspects and categories, providing comprehensive insights into the functional aspects of these inter-taxon HGT events-related genes. The investigation also extends to six multi-drug-resistant (MDR) candidate species, revealing an even greater involvement of ARGs in inter-taxon HGT events within each MDR species.

These research papers will enhance our understanding of HGT events and their impact on microbial evolution. The first paper introduces an improved method and the

HGTree v2.0 database for detecting HGT events and studying gene and species trees. The second paper highlights the prevalence of ARGs in inter-taxon HGT events and investigates the underlying factors, providing valuable insights into the genetic mechanisms shaping microbial genomes.

Key Words: Horizontal Gene Transfer, Tree Reconciliation, Virulence Factor, Antibiotic Resistance,

Student Number: 2021-21311

Contents

Abstract i
Contents v
List of Tables vi
List of Figures vii
Chapter 1. Literature Review 1
1.1 Horizontal Gene Transfer 2
1.2 Virulence Factors and Antibiotic Resistance Genes 4
Chapter 2. A comprehensive database update for horizontal gene transfer (HGT) events detected by the tree-reconciliation method 7
2.1 Abstract 8
2.2 Introduction 9
2.3 Data collection and processing 12
2.4 Results and Discussion 18
2.5 Future Works and Conclusion 31
Chapter 3. Exploring the trend of virulence factors and antibiotic resistance genes in inter-taxonomic horizontal gene transfer events 32
3.1 Abstract 33
3.2 Introduction 34
3.3 Method 37
3.4 Results and Discussion 39
3.5 Conclusion 66
General Discussion 68
References 70
국문 초록 82

List of Tables

Table 2.1 The summary of the comparison between HGTree v2.0 and the previous version **23**

Table 3.1 The number of inter-taxonomically transferred genes, virulence factors and antibiotic resistance genes detected by the tree-reconciliation method **41**

Table 3.2 A summary table of the total number of inter-taxonomically transferred VFs and ARGs of six MDR candidates and the ratio of them relative to VFs and ARGs in their whole genomes..... **61**

List of Figures

Figure 2.1 The flowchart of constructing the HGTree v2.0 database and the user query processor **16**

Figure 2.2 The overview of HGTree v2.0 ..... **19**

Figure 2.3 The ratio of total genes and HGT events per phylum and the ratio of HGT events occurred between and within taxa **25**

Figure 2.4 The summary of functional annotations for horizontally transferred genes **26**

Figure 2.5 The updated visualization of HGTree v2.0 interface **29**

Figure 3.1 The COG categories of VFs and ARGs in whole genome, HGT events and inter-taxon HGT events **43**

Figure 3.2 The COG categories of VFs transferred via inter-taxon HGT events **44**

Figure 3.3 The COG categories of ARGs transferred via inter-taxon HGT events **45**

Figure 3.4 The VFDB categories of VFs in HGT events and inter-taxon HGT events **49**

Figure 3.5 The ARO categories of ARGs in HGT events and inter-taxon HGT events **52**

Figure 3.6 The drug classes of ARGs transferred in HGT events and inter-taxon HGT events **58**

Figure 3.7 Graphs of MDR candidate species illustrating their inter-taxon (including inter-species) donors of VFs and ARGs, and the number and ratio of each inter-taxon events **62**

Chapter 1. Literature Review

1.1 Horizontal Gene Transfer

1.1.1 What are horizontal gene transfer events?

There are various ways that drive prokaryotic evolution, including mutation, genetic drift, selection, dispersal and horizontal gene transfer (Lynch et al. 2016; Arnold, Huang, and Hanage 2022). Among them, horizontally transferred genes are a major source of mechanism to shape the prokaryotic genomes, since prokaryotes (like bacteria and archaea) reproduce asexually, which often makes the offspring almost identical to their parents (Keeling and Palmer 2008). This shaping of genome by horizontal gene transfer favors the survival of the host bacteria, acting upon by natural selection. For example, organisms that received beneficial genes (such as antibiotic resistance genes) are more likely to adapt to the environment, whereas those received deleterious genes have less chances (Vogan and Higgs 2011). A significant fraction of prokaryotic genomes are known to be made up of HGT-related genes (Novichkov et al. 2004), thus the acquisition of genes does not merely mean a change in functions of genomes but is also critical in clarifying the lineage of bacteria, thus should be thoroughly studied.

The chance of a successful horizontal gene transfer decreases between organisms with distant phylogenetical relationships, since horizontal gene transfer events are likely to go through homologous gene recombination, where more likely genes are incorporated into the chromosome of the recipients with similar recombining sequences (Ochman, Lawrence, and Groisman 2000; Popa and Dagan 2011). However, studies say that HGT events do not have taxonomic boundaries, even

transferring genes to organisms of different domains (Koonin, Makarova, and Aravind 2001). Thus, prokaryotic species that get various types of genes through HGT can have much various characteristics, can even have a potential to be a multi-drug resistant (MDR) bacteria, which will be elucidated in the following sections.

1.1.2 The Tree-reconciliation Method

Various methods have been developed to identify HGT, each with its own strengths and limitations. Such methods include the HGT-DB and DarkHorse DB, which are typical implicit methods for HGT events detection that utilize the GC bias, nucleotide composition, and codon usage (Garcia-Vallve et al. 2003; Podell, Gaasterland, and Allen 2008), and the ShadowCaster (Sánchez-Soto et al. 2020), NearHGT (Adato et al. 2015) and RecentHGT (Li et al. 2018), which were built for HGT-related genes between closely related organisms. A common limitation they have is that the implicit methods are prone to false positive and negative results due to change in the features of horizontally acquired genes after long period of evolutionary history (Lawrence and Ochman 2002; Podell, Gaasterland, and Allen 2008), and for other methods are that they are not suitable for the detection of HGT events in organisms of long-distanced phylogenetic distance. One prominent and explicit phylogenetic method for detecting HGT is the tree-reconciliation method (Shikov et al. 2022).

The tree-reconciliation method has its long history back to when host-parasite cophylogeny was studied (Charleston and Perkins 2006). The method utilizes both species trees and corresponding gene trees to identify horizontally transferred genes.

The species tree represents the evolutionary relationships among different organisms, while the gene tree represents the evolutionary history of a specific gene or set of genes. By comparing these two types of trees, the method aims to identify incongruences or discrepancies providing a more accurate and reliable detection of HGT events (Ragan 2001; Sevillya, Adato, and Snir 2020). Upon calculation of the reconciliation, the DTL (Duplication, Transfer and Loss) events are considered and given a cost to each of them, so to find the lowest possible reconciliation cost (Bansal, Alm, and Kellis 2012). However, the detection of HGT events through the DTL-reconciliation model presents a complex challenge, in a way that it is computationally intense and conceptually difficult (Jeong et al. 2016).

Despite these challenges, the tree-reconciliation method remains a valuable tool in the study of HGT, especially when accurate inference trees are available, which can be derived by methods elucidated in the Chapter 2. It provides a robust framework for identifying and understanding the patterns and dynamics of HGT events, shedding light on the mechanisms that shape prokaryotic genomes and contribute to their evolutionary success.

1.2 Virulence Factors and Antibiotic Resistance Genes

The discovery of penicillin by Alexander Fleming in 1928 has significantly uplifted the clinical status of the mankind (Bud 2007). However, it was also a beginning of an emergence of the multi-drug resistance (MDR) bacterial strains, as bacterial strains disseminated and acquired various types of virulence factors and antibiotic resistance (Von Wintersdorff et al. 2016).

1.2.1 Virulence Factors

Virulence factors (VFs) are genetic elements or traits possessed by pathogens that contribute to their ability to cause disease in hosts (Weiss 2002). VFs play a significant role in the pathogenicity and virulence of bacterial strains, and their acquisition through HGT can have profound implications for the severity and outcome of infections. Understanding the transfer and expression of VFs is crucial for developing targeted strategies to combat pathogenic bacteria (Cross 2008). Virulence factors have various functions. For instance, genes with the 'Motility' and 'Adherence' functions are responsible for the movement of flagella and pili, and the 'Biofilm' function is responsible for giving protective niches to the cells (Stoodley et al. 2002). There are numerous other functions (all together, 14 categories according to the VFDB category (Liu et al. 2022), including the 'Others' function) and a collection of them will bring synergistic threat to mankind.

1.2.2 Antibiotic Resistance Genes

Antibiotic resistance genes (ARGs) confer resistance to antibiotics, allowing bacteria to survive exposure to antimicrobial agents (Martínez, Coque, and Baquero 2015). HGT plays a crucial role in the dissemination and accumulation of ARGs among bacterial populations, and the target is not constraint to pathogens, but can be as broad as commensal microorganisms too (Davison, Woolhouse, and Low 2000). The transfer of ARGs through HGT poses a serious threat to public health, as it can lead to the emergence of multi-drug resistant (MDR) bacterial strains. For instance, MRSA is a type of *Staphylococcus aureus* that has resistance to methicillin (Deurenberg

and Stobberingh 2008), and MRPA is Pseudomonas aeruginosa with comprehensive resistance to multi-drugs (Arruda et al. 1999). Other than the two examples, Enterococcus faecium, Klebsiella pneumoniae, Acinetobacter baumannii are also infamous pathogens with MDR, and collectively referred to as the ESKAPE groups (Pendleton, Gorman, and Gilmore 2013). Effective management of antibiotic resistance requires a comprehensive understanding of the mechanisms and dynamics of ARG transfer and evolution.

The mechanisms of antibiotics resistance are various. A mechanism like 'Antibiotic target alteration' can be induced by mutations or acquired functions through HGT events, and have a major impact on antibiotic resistance since targets of antibiotics are known to be specific (Kapoor, Saigal, and Elongavan 2017). Thus, a small change can bring a huge difference, like a change in D-alanyl-alanine (the target of glycopeptide antibiotics) to D-alanyl-lactase. Another example of antibiotic resistance mechanism is the 'Antibiotic efflux'. The mechanism is activated with proteins on the membrane (the efflux pumps) that pumps out the unwanted compounds entering the cell membrane, triggered by the difference in the concentration. (Lambert 2002; Giedraitienė et al. 2011). The gathering of all sorts of antibiotic mechanisms to a single bacteria species can ultimately create so called the 'Super Bacteria' that is resistance to many antibiotics' actions.

This chapter was published in *Nucleic Acids Research* as a partial fulfillment of Youngseok Choi's M.S program.

Chapter 2. A comprehensive database update for horizontal gene transfer (HGT) events detected by the tree-reconciliation method

2.1 Abstract

HGTree is a database that provides horizontal gene transfer (HGT) event information of 2,472 prokaryote genomes using the tree reconciliation method. HGTree was constructed in 2015, and a large number of prokaryotic genomes have been additionally published since then. To cope with the rapid rise of prokaryotic genome data, we present HGTree v2.0 (http://hgtree2.snu.ac.kr), a newly updated version of our HGT database with much more extensive data, including a total of 20,536 completely sequenced non-redundant prokaryotic genomes, and more reliable HGT information results curated with various steps. As a result, HGTree v2.0 has a set of expanded data results of 6,361,199 putative horizontally transferred genes integrated with additional functional information such as the KEGG pathway, virulence factor, and antimicrobial resistance. Furthermore, various visualization tools in the HGTree v2.0 database website provide intuitive biological insights, allowing the users to investigate their genomes of interest.

2.2 Introduction

Genetic materials are often inherited from parents to their offspring vertically, however, for prokaryotes like bacteria or archaea (that reproduce asexually), the inheritance can be demonstrated horizontally during species interactions, from adjacent organisms to another, for their evolutionary benefits. (Woese 1987) The asexual reproduction of prokaryotes gives nearly identical genetic information to the offspring, making horizontal gene transfer (HGT) an essential process in achieving genetic variations. (Keeling and Palmer 2008) HGT is one of the driving forces responsible for shaping the prokaryotic genomes and surviving natural selection. (Kunin and Ouzounis 2003) For example, taking up genes responsible for antimicrobial resistance could be critical in facilitating the organism's adaptation to a particular environment. Contrarily, organisms that have taken up unnecessary genes have more chances of excluding themselves from natural selection if the genes are either neutral or detrimental. (Vogan and Higgs 2011) As much as HGT dedicates to the evolution of prokaryotic organisms, it can always complicate the interpretation of the lineage or the history of species evolution, leading to an erroneous result regarding their classifications. Thus, it is fundamental in the research of prokaryotic evolution to deeply understand HGT. (Doolittle 1999)

In view of this, we constructed HGTree (https://hgtree.snu.ac.kr) (Jeong et al. 2016) in 2015. Other conventional databases like the HGT-DB (Garcia-Vallve et al. 2003) or DarkHorse HGT Candidate Resource (Podell, Gaasterland, and Allen 2008) utilize genome signatures (such as GC bias, nucleotide composition, and codon usage)

or implicit phylogenetic methods (such as comparison of evolutionary distance derived by sequence similarity). Unlike them, HGTree exploits the tree-reconciliation method, an explicit phylogenetic method that uses species trees and corresponding gene trees, to determine horizontally transferred genes. The tree-reconciliation method is well-known for its reliability and is a prevailing method for HGT event detection (Ragan 2001; Sevillya, Adato, and Snir 2020) and the HGTree database was able to implement the method while overcoming its challenges (Jeong et al. 2016). As a result, the database was built with 660,840 HGT event results of 2,472 prokaryotic genomes.

Since 2015, the number of prokaryotic genome data has increased dramatically, and so has the requirement for a more comprehensive HGT database with more reliable and efficient processes to detect HGT events. After the publication of HGTree, new methods to detect HGT events have been continuously developed. ShadowCaster hybridized the implicit method with the support-vector-machine (SVM) method (Sánchez-Soto et al. 2020), and tools like NearHGT (Adato et al. 2015) or RecentHGT (Li et al. 2018) were developed to detect HGT events between closely related taxa. However, these tools are not suitable for the accurate detection of the HGT events among various phyla, leaving the tree-reconciliation method the most reliable detection method as long as accurate inference trees are supported. (Shikov et al. 2022)

Therefore, we present a newly updated HGTree database in response to these demands, coping with the growing interest in HGT. The updated HGTree (referred to

as HGTree v2.0) includes approximately eight times larger genomes than the previous version, earning approximately over six times more putative HGT event results. The newly revised HGT events detection procedure ensured increased detection reliability, and the detected events are presented in the HGTree v2.0 website equipped with more user-friendly interfaces, including the modified user-query processor, which allows users to navigate their genomes.

2.3 Data collection and processing

2.3.1 Data processing

The genome data used in this research were retrieved from NCBI using two different search titles: ‘Bacteria’ and ‘Archaea’. Further options included ‘Assembly’, ‘Latest RefSeq’ (O'Leary et al. 2016), and ‘Complete genome’. (https://www.ncbi.nlm.nih.gov/assembly; May 5, 2021) (Sayers, Bolton, et al. 2021). A total of 20,179 bacterial and 357 archaeal genomes were retrieved at the strain level. Along with the nucleotide and amino acid sequence data, GenBank data were also retrieved, including each genome’s size, gene function, and the number of CDS. The taxonomy of each genome was further confirmed and retrieved from the GTDB database (https://gtdb.ecogenomic.org/) (Chaumeil et al. 2020).

To detect orthologous genes, 20,179 bacterial genomes and 357 archaeal genomes were processed separately using PorthoMCL (Tabari and Su 2017a). The all-versus-all blast search was performed to detect orthologous genes, and the genes were clustered and assigned to orthology groups in FASTA format. We set the alignment coverage to 80%, the E-value cutoff to 10^{-6}, and the minimum identity score was 98%. To retrieve 16S-rRNA, RNAmmer (ver. 1.2) (Lagesen et al. 2007) was used. The program uses an HMMER (ver. 3.1b2) (Potter et al. 2018) based scanning procedure to detect 16S-rRNA of both bacteria and archaea from input genome data file.

For each detected orthology group and its corresponding 16S-rRNA group, CLUSTAL Omega (ver. 1.2.3) (Sievers and Higgins 2018) was used for the sequence alignment, and FastTree2 (ver. 2.1.9) (Price, Dehal, and Arkin 2010) was used to

construct phylogenetic gene and species trees. As a result, an un-rooted gene tree (made of orthologous protein sequences) and a species tree (made of 16S-rRNA sequences of genomes from the same orthology group) were constructed for each orthology group. Newick Utility (newick_utils ver. 1.6) was used to re-root the species trees using the 18S rRNA sequence from '*Saccharomyces cerevisiae*' for the outgroup.

Ranger-DTL 2.0 (Bansal et al. 2018) was used to detect horizontally transferred genes among different orthologous group sets of phylogenic trees. The first round of Ranger-DTL 2.0 was done under default parameters (duplication cost: 2, transfer cost: 3, loss cost: 1) for a standard result. Next, we ran a second round with '3' and '4' for the transfer cost for a more rigorous analysis, leaving the duplication and loss cost as default. The two different transfer cost results were then aggregated as one result in the AggregateRanger step. The results of the first and second rounds were named as the 'Standard DB' and the 'Strict DB', respectively. This process will be further elucidated in the 'Results and Discussion' section.

The 20,536 genomes were scanned against PfamScan (ver. 1.6) (Mistry et al. 2021) under default parameters to conduct protein family level assignment. DIAMOND (ver. 0.9.14) (Buchfink, Xie, and Huson 2015) BLASTP (ver. 2.2.31+) (Camacho et al. 2009) search was conducted with all detected putative HGT-related genes against the VFDB database (http://www.mgc.ac.cn/VFs/main.htm) (Liu et al. 2022) to classify the virulence factors, and CARD-rgi (ver. 5.2.1) (Alcock et al. 2020) was conducted to identify the antimicrobial resistance genes. For the annotation of the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway (Kanehisa and Goto 2000), the emapper

command from the software eggNOG (eggNOG-mapper ver. 2.1.7) (Huerta-Cepas et al. 2019; Cantalapiedra et al. 2021) was used. The annotation of COG clusters was done by running BLASTP against the COG database (Galperin et al. 2021).

2.3.2 User query processor

We also modified the user query processor of the version, and the steps are as follows: First, Prokka (ver. 1.11) (Seemann 2014) and Barrnap (ver. 0.9) (Seemann 2013) are used to predict protein sequences and 16S-rRNA of the submitted genome. Second, both the protein sequences of the query genome and HGTree v2.0 data are used to generate two BLAST databases to find the best-hit results by performing the reciprocal BLAST search against each other. For the HGTree v2.0 BLAST database, users can choose between the 'Standard DB' and 'Strict DB'. The default parameters for the reciprocal BLAST are set as 80% for the alignment coverage, 80% for the identity score, and the *E*-value cutoff of 10^{-6}, but users can choose their parameters too. The best-hit results will be assigned to HGTree v2.0 database orthology groups to generate new orthology groups that include sequences of the query genome. Third, gene trees and corresponding species trees are generated through ClustalO and FastTree2. Finally, Ranger-DTL2 is performed to detect putative horizontally transferred genes from the input genome. Like how we built the HGTree v2.0 database, if users choose the 'Strict Mode' in the second step, Ranger-DTL will run twice, once with default transfer cost and second with transfer cost of '3' and '4'. Users will be given a final text file stating the ratio of HGT-related genes events and a list of them, separated into two sections: 'Donated Genes' and 'Received Genes'. Virulence factors (the VF ID), antimicrobial

resistance (the ARO ID), gene names, product names, and the genomes of the donors and recipients will also be included in the text file. The workflows of both database construction and the user query processor are illustrated in Figure 2.1.

2.3.3 Website visualization

The MariaDB (ver.10.5.8) (http://mariadb.org/) management system was used. The web-based user interface was generated with HTML5, django, CSS and JavaScript. DataTable (ver.1.11.5) (https://datatables.net/) and jQuery (ver.2.1.1) (http://jquery.com) were used to implement the user interface widgets. To generate circular phylogenetic trees, jsPhyloSVG-1.55 (Smits and Ouverney 2010) was used and graphics were illustrated with Google Chart SVG, JavaScript library, D3. (Bostock, Ogievetsky, and Heer 2011)

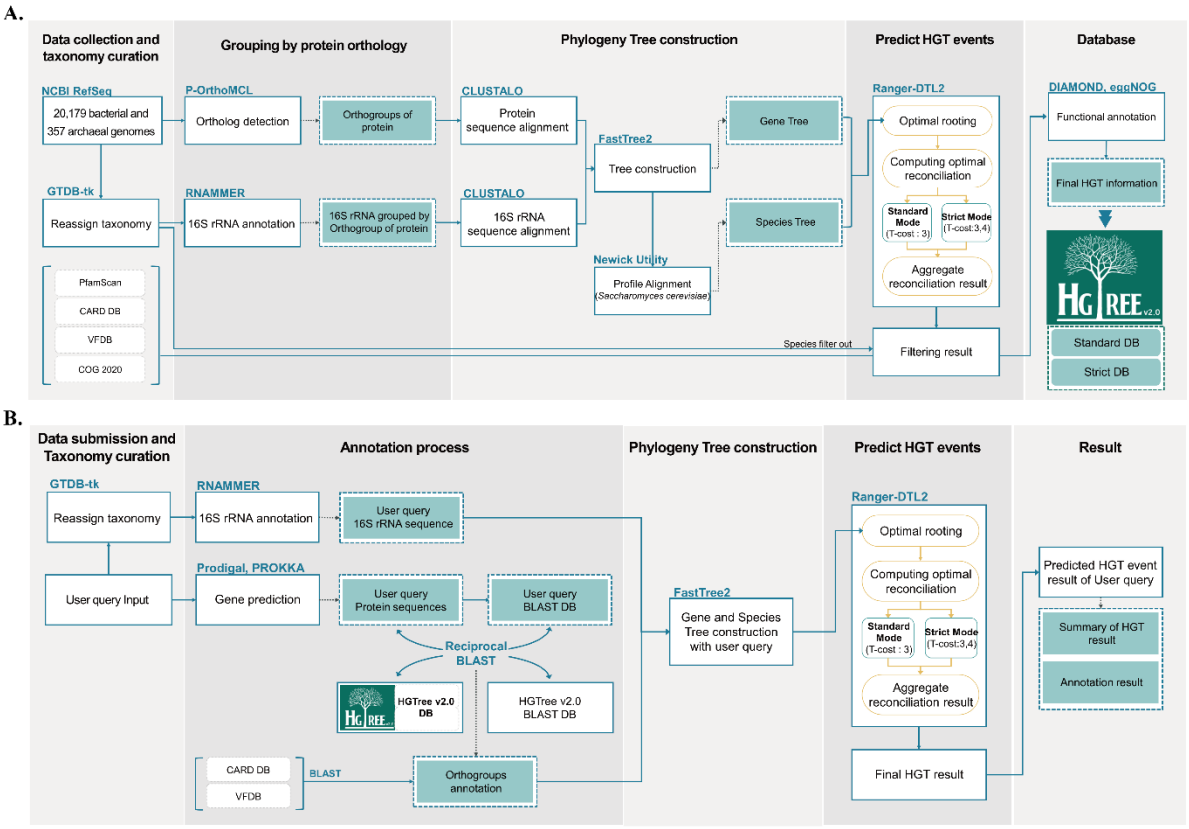

Figure 2.1 The flowchart of constructing the HGTree v2.0 database and the user query processor.

(A) The overall workflow of constructing the HGTree v2.0 Database. To construct the database, 20,179 bacterial and 357 archaeal complete genomes were downloaded from NCBI. The taxonomy of each genome was reclassified using GTDB-tk. The orthology groups were calculated with POrthoMCL and each group was aligned for phylogenetic gene trees construction. Corresponding species trees were made with 16S-rRNA sequences of genomes in the orthology group. Each group's gene tree and species tree are used to perform RANGER-DTL and putative HGT events and related genes are detected. HGT-related genes' Pfam, antimicrobial resistance, virulence factors, COG clusters and KEGG pathways annotations are also uploaded to the database.

(B) The overall algorithm of the user query processor. Users can upload their prokaryotic complete genomes to analyze HGT-related genes in them. In the user query processor, reciprocal blast exploits the HGTree v2.0 database results to calculate new orthology groups that include genes of the input genome. Other parts of the workflow follow the procedures of the previous version, except that users can choose the mode of the processor (the ‘Standard Mode’ and ‘Strict Mode’).

2.4 Results and Discussion

In HGTree v2.0, to keep pace with the flooding amount of new prokaryotic genome data, we acquired information on putative HGT-related genes through a more accurate tree-reconciliation method. Our major focus of the update was to ensure that the HGT results are more reliable and show their various functional data in more user-friendly ways. The overview of updates on HGTree v2.0 can be seen in Figure 2.2.

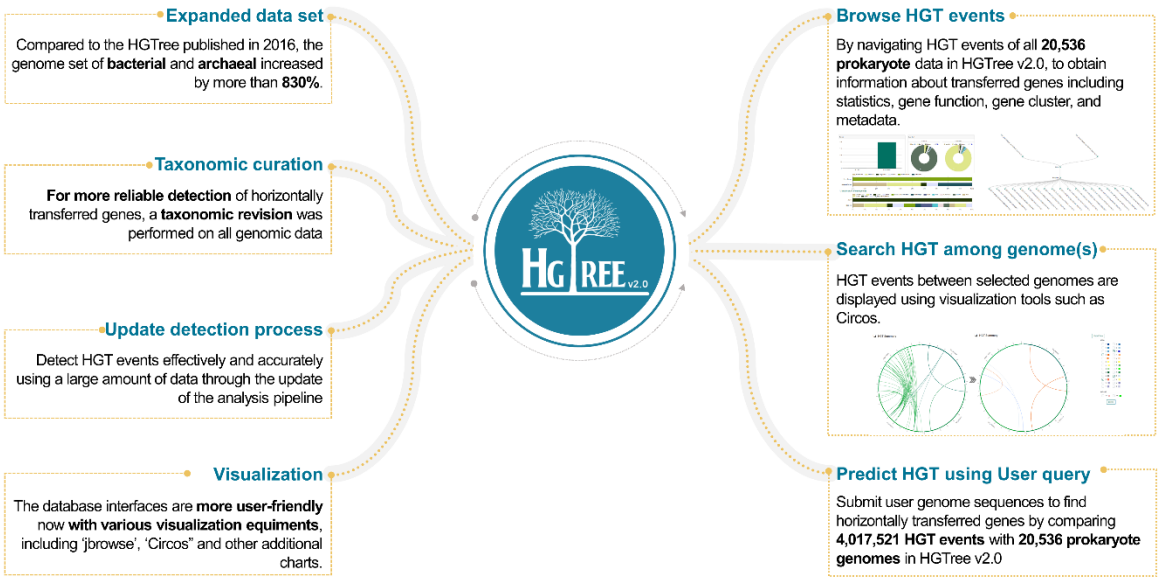

Fig 2.2 The overview of HGTree v2.0.

HGTree v2.0 has expanded data sets, changed the pipeline of HGT events detection, added more curations steps in detecting horizontally transferred genes, and is equipped with more visualization equipment for users. Users can 'Browse' and 'Search' HGT events and their related genes among different type of genomes in our vast database, and also can upload their genomes to find out what genes are engaged in HGT events.

2.4.1 Enhanced HGT events detection procedures

The orthology grouping step of the previous version was done using Mestortho. Mestortho is a powerful tool that uses the distance method to calculate orthology based on the phylogenetic criterion of minimum evolution (Kim et al. 2008). However, considering the vast amount of data input for constructing the database, Mestortho was replaced with PorthoMCL due to Mestortho's limitation on speed and time taken to calculate all minimum evolution of 20,536 genomes. PorthoMCL is designed to calculate orthology groups of a considerable number of genomes with its parallelized sparse file structure using a Markov Cluster algorithm (Tabari and Su 2017a), so we expected it to be more suitable for our dataset. PorthoMCL utilizes a similar algorithm to OrthoMCL (Li, Stoeckert, and Roos 2003), but according to Tabari et al., PorthoMCL was able to process at least five times larger size of the input genome than OrthoMCL, while the percentage of the process speedup was up to 455%.

As the detection of horizontally transferred genes relies on the reconciliation of gene and species phylogenetic trees, clarifying each genome's exact lineage and specie information is crucial. Therefore, we used a new approach by employing GTDB-tk to reclassify the taxonomy of all the genomes for more reliable detection of HGT-related genes. Furthermore, our method was not adequate for detecting HGT events between the strains of the same species since we used 16S-rRNA sequences to build phylogenetic trees. Using 16S-rRNA made the species trees less clear as it gets down to the strain level because 16s-rRNA is a universal species marker for prokaryotes. This could induce false-positive results about HGT events between the strains from the same

species. Constructing phylogenetic trees based on the whole-genome ANI method could have been an alternative, but the ANI method was considered inadequate to be used due to a decrease in discriminatory power among taxa higher than the family level. (Gosselin et al. 2022) Taking these into consideration, the GTDB-tk reclassification of each genome was important and ensured that no such false-positive events were included in the output result. GTDB-tk was also used in the user query processor so the taxonomy information of query genome could be reclassified.

One of the limitations the previous version had was that they missed other possibly calculable optimal reconciliations during the Ranger-DTL step. This problem was solved in the current update as Ranger-DTL 2.0 was used. In the OptRoot step, 3,028,789 pairs of trees (rooted species and un-rooted gene trees), constructed by aligning the same number of orthologous gene groups, were used to calculate all optimal gene tree rootings to additionally produce 1,500,857 trees, which summed up to 4,529,646 pairs of trees altogether. The first version of Ranger-DTL (Bansal, Alm, and Kellis 2012) did not allow this type of optimal calculation, so the users of HGTree had to investigate their genomes with only one optimal gene tree per orthology group. The subsequent Ranger-DTL step then calculated all optimal tree-reconciliations for the gene and species trees and returned one reconciliation result per each tree pair that was calculated to have the minimum reconciliation cost. For the ‘second round’ of the Ranger-DTL step, like prior research exemplified, the change in transfer costs from ‘3’ to ‘4’ resulted in more reliable HGT events detection, but the total quantity of the genes was much lesser, and the detected genes also had higher risks of false negatives. (Kundu and Bansal 2018; Kloub et al. 2021) However, by comparing the transfer cost ‘4’ results

with the cost ‘3’ results during the aggregation step, we were able to reduce the chances of false negatives. (Kloub et al. 2021) This was because the AggregateRanger step aggregated all the Ranger-DTL results into one result and provided the percentage of genes in each orthology group that showed 100% consistent ‘events’ and ‘mapping’ among all optimally calculated Ranger results. (Bansal et al. 2018) This percentage was additionally used as our filtering value and out of all aggregated reconciliation results, results that showed less than 90% were filtered.

2.4.2 Expanded database size

We used 20,536 non-redundant prokaryotic genomes (20,179 bacterial genomes and 357 archaeal genomes, 4,964 and 264 species, respectively) to detect 3,028,789 orthologous groups and by using them, we predicted 6,361,199 HGT-related genes (6,174,528 for bacteria genomes and 186,671 for archaeal genomes) (Table 2.1). As shown in Figure 2.3A, we could identify how many HGT-related genes were predicted per phylum, and it was clear that there was a positive correlation between the ratio of phyla and the ratio of predicted HGT-relate genes. Nevertheless, some phyla like *Actinobacteriota* and *Bacteriodota* had a higher ratio of predicted HGT events relative to their total gene ratios. Based on this result, we also calculated the frequency of the HGT events among each taxon, whether the events occurred within taxon or between different taxa (Figure 2.3B). As the figure shows, more genes were transferred within taxon, which also concurs with previous research that HGT occurs between closely related organisms more frequently (Ochman, Lawrence, and Groisman 2000; Soucy, Huang, and Gogarten 2015; Andam and Gogarten 2011).

Table 2.1 The summary of the comparison between HGTree v2.0 and the previous version.

Database		HGTree	HGTree V2.0
Total non-redundant microbial genome		2,472	20,536
Genome retrieval day		17 March, 2015	5 May, 2021
No. Phyla & Genera		41 & 700	36 & 1,542
Total protein sequence		7,748,306	74,339,979
Number of orthologous gene sets		154,805	3,028,789
Detected putative HGT events		660,840	Standard mode: 4,017,521, Strict mode: 1,314,170
Detected putative HGT-related genes		1,401,252	Standard mode: 6,361,199 Strict mode: 1,987,538
KEGG annotated genes		NA	1,827,784
Transferred virulence factors (for bacterial genomes)		NA	42,793
Transferred antimicrobial genes (for bacterial genomes)		NA	4,544
Programs Used			
Species re-classification		NA	GTDB-tk* (ver. 1.7.0)
Orthologous group detection		Mestortho	PorthoMCL
Sequence alignment		CLUSTAL Omega (ver. 1.2.1)	CLUSTAL Omega (ver. 1.2.3)
Phylogenetic tree construction		FastTree (ver. 2.0)	FastTree (ver. 2.1.9)
Tree reconciliation		RANGER-DTL-U (ver. 1.0)	Ranger-DTL 2.0*
Functional Annotation	Pfam	PfamScan (ver. 1.2)	PfamScan (ver. 1.6)
	COG	COG 2014 Database	COG 2020 Database
	KEGG	NA	eggNOG-mapper (ver. 2.1.7)
	Virulence genes	NA	VFDB 2019
	Antimicrobial genes	NA	CARD-rgi (ver. 5.2.1)

* The important changes in HGTree v2.0. By employing GTDB-tk program, all genomes' taxonomy was reclassified for more reliable results. Also, Ranger-DTL program allowed more reliable detection of HGT events by performing all possible optimal calculations.

In the previous version, only available Pfam IDs and COG (Clusters of Orthologous Genes) clusters were annotated to putative HGT-related genes. For a further analysis, the current update also focused on other functional annotations such as the KEGG pathway, virulence factors and antimicrobial resistance genes. As a result, 21,961 virulence factors and 7,140 antimicrobial resistance genes were identified from the HGT-related genes (Figure 2.4C), and 1,827,784 HGT-related genes were annotated with the KEGG pathways. The frequency of each cluster is illustrated in Figure 4A and B. Interestingly, a considerable number of genes was responsible for the 'Metabolism' of the organism, followed by 'Environmental Information Processing', 'Genetic Information Processing', 'Organismal Systems', 'Cellular Process' and 'Human Diseases'. Most genes responsible for 'Metabolism' was related to 'Energy Metabolism' and 'Carbohydrate Metabolism'.

In addition, the metadata of each genome was also retrieved from the NCBI database, which includes the 'Country', 'Isolation Source', and 'Collection Date'. The metadata is displayed on the website together with the world map (be further explained in the following 'Website Interface' section) and will provide further geographical insights to the users.

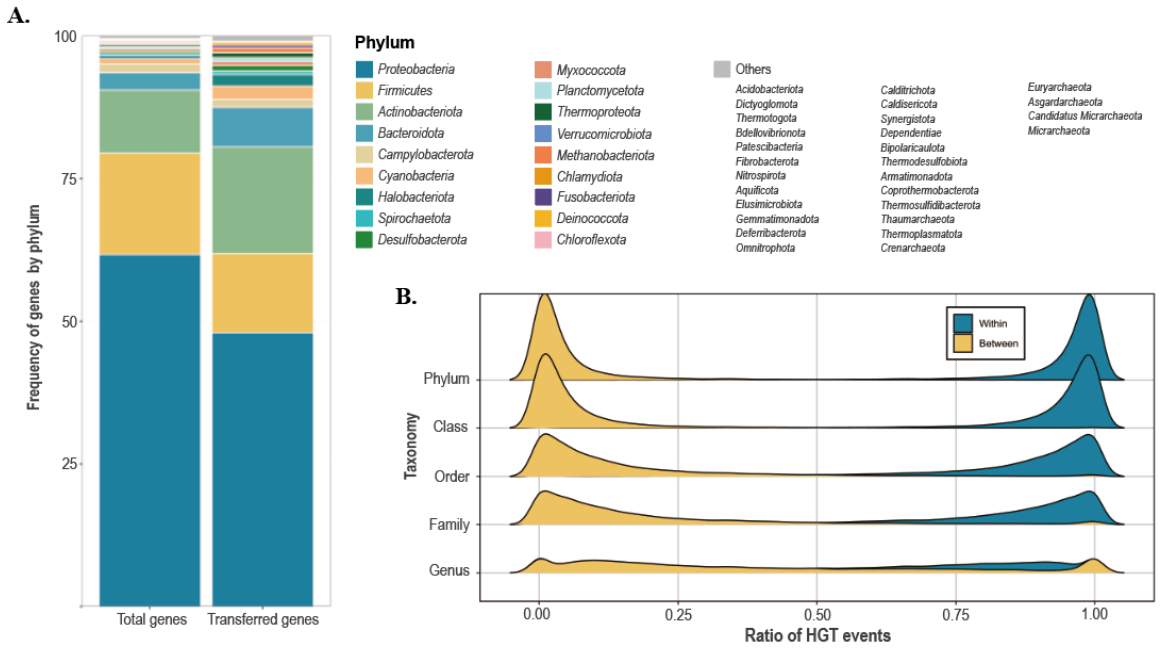

Figure 2.3 The ratio of total genes and HGT events per phylum and the ratio of HGT events occurred between and within taxa.

(A) The relative frequency of the total genes and HGT-related genes of genomes were visualized in the phylum level. More HGT-related genes were detected as the total number of genes in the phylum was larger.

(B) The x-axis shows the ratio of the HGT-related genes and the y-axis shows the taxonomy of the animal kingdom, from the genus level to the phylum level. The yellow peaks show the ratio of HGT events between different taxa and the blue peaks show the ratio of those that occurred within the same taxonomy. Both peaks were denser as the taxonomy levels were higher, meaning more genes were transferred between closely related organisms.

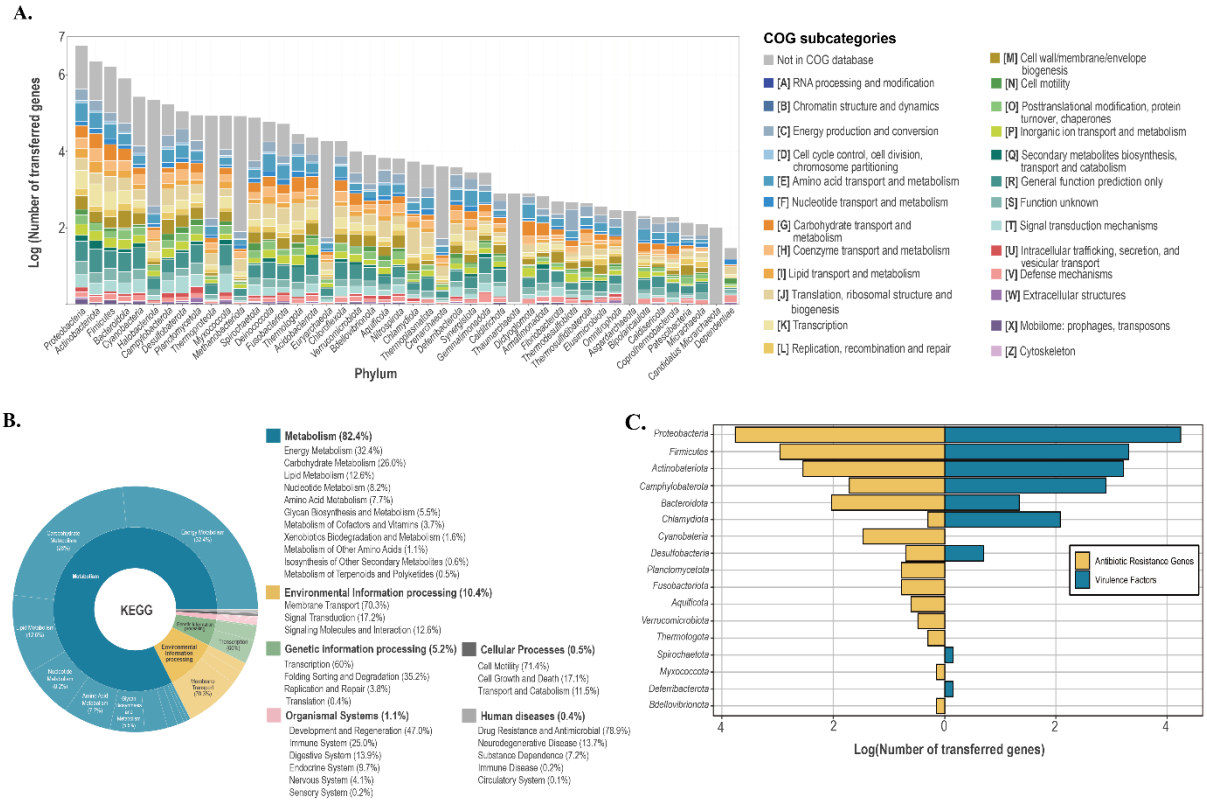

Figure 2.4 The summary of functional annotations for horizontally transferred genes.

(A) The ratio of COG clusters and the number of HGT-related genes for each phylum. Most of the clusters were responsible for the metabolism of the organism. A moderate portion of each genome was unclassified since no data was in the COG database.

(B) The ratio of KEGG pathway and subcategories of all bacterial genomes used. Most of the genes were classified in the 'Metabolism' category, followed by 'Environmental information processing', 'Genetic information processing', 'Organismal systems', 'Cellular Process' and 'Human diseases'.

(C) The number of virulence factors and antimicrobial resistance genes of bacterial genomes.

2.4.3 Upgraded data visualization

In the current update, we revised the website interface so that users can interpret results more thoroughly and quickly. The ‘Browse’ section provides all putative HGT events and the genes of 20,536 genomes. It also displays other information such as the genomes’ taxonomy, isolation source, geographical location, and the ratio of HGT-related genes in the genome (the HGT index). Users can choose a genome in the database to see its HGT-related genes (exhibited as ‘Standard DB’ and ‘Strict DB’, separately) and can also view charts illustrating the ratio of other genomes’ genera that share the same transferred genes. Furthermore, users can inspect the external links connecting to the corresponding website page of Pfam, COG, KEGG pathway, virulence factor, and antimicrobial resistance by clicking ‘Detail’. Like version 1, the HGT relationship plot related to each genome’s HGT events and their corresponding phylogenetic trees are concisely provided, along with the summary of HGT-related genomes in graphs and the world map. Lastly, with the Jbrowse program (Buels et al. 2016), gene clusters adjacent to the selected genes can be seen. The Jbrowse graph will give biological insights to the users about clusters of horizontally transferred genes by checking whether the surrounding genes are related to each other since HGT could often occur in the form of a gene cluster (or multiple genes) (Dilthey and Lercher 2015).

Unlike the ‘Browse’ section, the ‘Search’ section only shows HGT events among the selected genomes. Within this section, we utilized the program ‘Circos’ (Krzywinski et al. 2009) to visualize the predicted HGT events between organisms to make it easier for users to comprehend the events and their related genes. As shown in

Figure 2.5, the visualization can provide an intuitive understanding of the transferring genes by pointing out the start location (in the donor's chromosome) and the input location (in the recipient's chromosome) with different colors matched with selected donor genomes. Users can choose the COG clusters, KEGG pathway clusters, virulence factor, and antimicrobial resistance to only illustrate those horizontally transferred genes with corresponding functions to figure out what sort of genes could have affected the evolution of microorganisms.

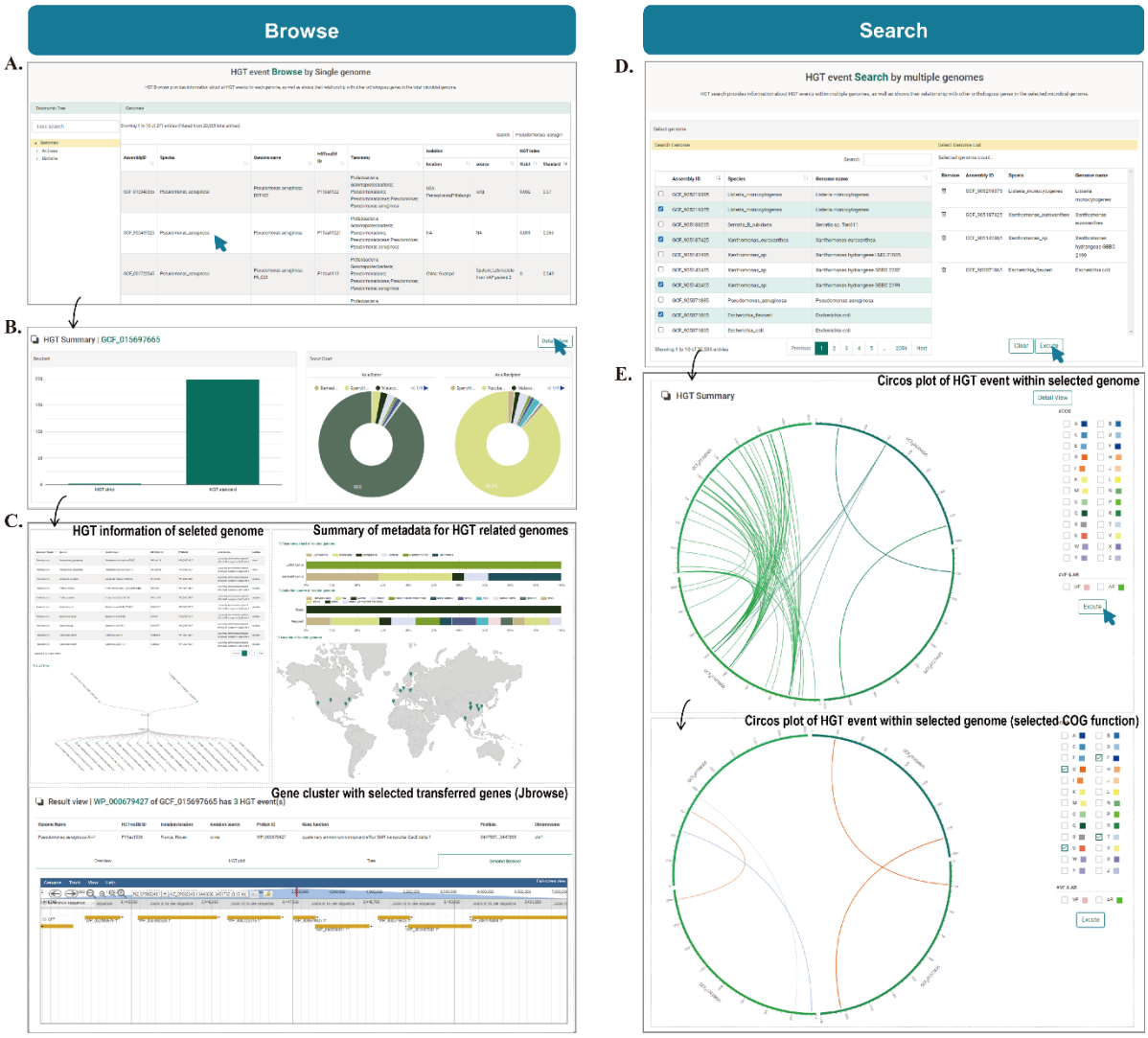

Figure 2.5 The updated visualization of HGTree v2.0 interface.

(A) In the 'Browse' section of the website, users can choose a GCF ID to find related HGT events. By clicking the genome, users can check the genome's taxonomy, isolation source, geographical location, and the ratio of HGT-related genes in the genome (the HGT index).

(B) By clicking 'Detail', users can check the 'HGT Summary' of the selected genome. The ratio of other genera engaged with the chosen genome by HGT events are shown in graph.

(C) Furthermore, the external links connecting to the Pfam, COG, and KEGG database are also included for the annotated HGT-related genes, along with the phylogenetic tree concisely showing the movement of the genes. The summary of HGT-related genomes is also shown in graph and the world-map. With the 'Jbrowse' program, users can also investigate gene clusters with selected transferred genes.

(D) The 'Search' section provides HGT events information only in the selected genomes.

(E) The 'Circos' visualization can provide an intuitive understanding of the transferring genes by lines matched with the colors of the donor chromosomes. The lines (HGT events) can be filtered by choosing the COG clusters, virulence factors or antimicrobial resistance genes to only illustrate HGT events that involve genes with such functions.

2.5 Future Works and Conclusion

In our future research, we intend to concentrate on constructing a database that can additionally accommodate data of other types of microorganisms that can demonstrate HGT, like viruses, fungi, and other eukaryotes. Future works will also involve finding the method to predict HGT-related genes within the strain level since the major drawback of our approach was that we implemented 16S-rRNA sequences when constructing the species phylogenetic trees. 16S-rRNA gives similar (or even identical) sequences as organisms are more closely related (Ward et al. 1998); thus, more research into finding an adequate method to specifically clarify species trees at the strain level is still necessary before obtaining more various information on horizontally transferred genes.

In summary, the HGTree database was updated with a more extensive number of prokaryotic data sets and predicted a much larger number of HGT-related genes than it could do before. Changes in the prediction procedures allowed more reliable detection of putative HGT-related genes. The website interfaces are now more user-friendly with various visualization equipment.

This chapter will be published elsewhere as a partial fulfillment of Youngseok Choi's M.S program.

Chapter 3. Exploring the trend of virulence factors and antibiotic resistance genes in inter-taxonomic horizontal gene transfer events

3.1 Abstract

Horizontal gene transfer is a major mechanism for prokaryotic evolution. The transfer of all sorts of genes makes up prokaryotic genomes and the process is spontaneous and massive. This subsequent research paper builds upon the finding of the database called HGTree v2.0. Among those HGT-related genes detected by the database, our interest lay in virulence factors and antibiotic resistance genes, because the dissemination and accumulation of them can be a huge threat to the wellbeing of the mankind, thus needed to be studied thoroughly. Therefore, we focused on the inter-taxonomic transfer of virulence factors and antibiotic resistance genes and found out that the HGT events of bot VFs and ARGs were more active as the inter-taxonomic transfer got higher from inter-genus to inter-phylum. In this paper, we present the overall trends of inter-taxonomic trends of VFs and ARGs, and possible explanations of them assessed by the essential COG categories in the survival of bacteria.

3.2 Introduction

Bacterial evolution often takes place by horizontal gene transfer (HGT), where each microorganism takes up genes from another. HGT is an important step in evolution of microorganisms for their genetic variation, along with gene mutation, since they reproduce asexually. (Woese 1987) As the name infers, the inheritance of genes is demonstrated horizontally, unlike other organisms that give their genes vertically through reproduction procedures from parents to their offspring. (Keeling and Palmer 2008; Woese 1987) Through this process, microorganisms that have taken up neutral or beneficial genes (such as antibiotic resistance genes (ARGs) or genes involved in metabolism) are more likely to adapt to certain environments than other organisms that have taken up deleterious genes (such as transposons or integrated prophages). (Hall et al. 2020) Therefore, HGT has a huge impact on the nature of the recipients' genetic traits and phenotypes since massive information can be acquired spontaneously and can shape up the genome to a whole new status (de Koning et al. 2000; Baugher, Durmaz, and Klaenhammer 2014).

The discovery of penicillin by Alexander Fleming in 1928 revolutionized medicine by effectively eliminating deadly pathogens (Bud 2007). However, on the other hand, it also led to the emergence of multi-drug resistant (MDR) bacterial strains that acquired and disseminated various antibiotic resistance genes (ARGs) (Von Wintersdorff et al. 2016). These MDR bacterial strains pose significant threats to the well-being of living organisms including the human (Nikaido 2009). In the absence of alternative therapies, once pathogens receive multiple AMR genes from other

organisms by HGT and become MDR, those pathogens can undermine our current clinical systems (Stanton 2013). In addition to antibiotic resistance, virulence factors (VFs) can equally contribute to the danger of MDR pathogenic strains, as infections caused by VFs are one of the major causes of mortality. Investigating the interactions between VFs and hosts is crucial for developing novel vaccination therapies in combination with antibiotic treatments (Cross 2008; Wu, Wang, and Jennings 2008).

Nowadays, very small is known about inter-taxon (inter-species, genus, family, order, class and phylum) HGT events, especially regarding the VFs and ARGs of bacteria. As previous research has verified, HGT is more likely to be demonstrated between closely related species than distant ones. The probability of HGT tends to decrease as the phylogeny gets farther, not only because the organisms could be physically distant, but since the donated genes are most likely to undergo homologous recombination during the incorporation of genes to chromosomes with similar recombining sequences. (Ochman, Lawrence, and Groisman 2000; Popa and Dagan 2011) Nonetheless, the processes of HGT are thought to have no taxonomic boundaries, even transferring genes between organisms of different domains. (Koonin, Makarova, and Aravind 2001) As for bacterial organisms, the inter-phylum HGT is known to take place among organisms living under extreme conditions (Caro-Quintero and Konstantinidis 2015), such as thermophilic and halophilic bacteria to increase their survival rates (Zhaxybayeva et al. 2009; Nelson-Sathi et al. 2012). Since VFs and ARGs are known to be crucial in the survival of bacteria (Sharma et al. 2017; Munita and Arias 2016), it is probable that VFs and ARGs both likely to undergo inter-taxon HGT events nonetheless of the phylogenetic distances between the donor

and recipient.

In light of these, we explored HGT events of 20,179 bacterial strains by employing the tree-reconciliation method. The tree-reconciliation method employs the phylogenetic gene trees and corresponding species trees (derived from sets of orthologous groups) to calculate HGT events, by using the program named RANGER-DTL 2.0. (Bansal et al. 2018) Such a method is known to be more relatively reliable than other methods that utilize GC bias or sequence similarity (so called the implicit methods) and is now one of the prevailing methods for HGT events detection (Sevillya, Adato, and Snir 2020). The results of putative HGT events and their related genes were used to construct HGTree v2.0 database (http://hgtree2.snu.ac.kr).

In this paper, by exploiting HGTree v2.0 database, we sought to investigate the VFs and ARGs of inter-taxonomic HGT events proposed to be occurred among 20,179 bacterial strains, to see what may have driven the extreme transfer between organisms with much farther evolutionary distance.

3.3 Method

The data used in this study was retrieved from the HGTree v2.0 database (Choi et al. 2023). In summary, the data was generated with the following methods:

A total of 20,179 bacterial genomes were downloaded from NCBI (Sayers, Beck, et al. 2021; O'Leary et al. 2016) The GTDB database (https://gtdb.ecogenomic.org/) (Chaumeil et al. 2020) was used to reclassify bacterial genomes' each lineage. An all-vs-all blast (Johnson et al. 2008) search was performed prior to the detection of orthologous genes. Orthologous gene sets were detected by using PorthoMCL (Tabari and Su 2017b), and RNAmmer (ver. 1.2) (Lagesen et al. 2007) was used to extract 16S-rRNA sequences of each genome. Multiple sequence alignment of orthologous genes and corresponding 16S-rRNA sequences was performed by using CLUSTAL Omega (ver. 1.2.3) (Sievers and Higgins 2014) The construction of the phylogenetic tree files was done by using FastTree2 (ver. 2.1.9) (Price, Dehal, and Arkin 2010) The species trees were re-rooted by using the Newick Utility (newick_utils ver. 1.6) program (Junier and Zdobnov 2010) with the 18S-rRNA sequence of '*Saccharomyces cerevisiae*' as the outgroup. To detect HGT events and the involved horizontally transferred genes, Ranger-DTL 2.0 (Bansal et al. 2018) was used by taking a set of gene tree and species trees.

The genes related to HGT events were blast searched against the data provided from the VFDB database (http://www.mgc.ac.cn/VFs/main.htm) (Liu et al. 2022) to search for virulence factors in them. Here, the BLAST-P (Johnson et al. 2008) was used with the alignment coverage to 80%, E-value cutoff to 10^-6, and the minimum identity

score to 95%. CARD-rgi (ver. 5.2.1) (Alcock et al. 2023) was conducted to clarify ARGs in them. With the lineages provided by the GTDB database, ARGs and VFs involved in each inter-taxon HGT event were detected with in-house python script and their categories were organized according to the ARO category table downloaded from the CARD website (https://card.mcmaster.ca/analyze/rgi) (Alcock et al. 2020) and the VF id table downloaded from the VFDB database. CARD-rgi was additionally used to classify each ARG's drug class. *Pseudomonas aeruginosa, Salmonella enterica, Klebsiella pneumoniae, Staphylococcus aureus, Enterococcus faecium and Enterococcus faecalis* were chosen as the MDR candidates according to the number of ARGs they have. Further studies on MDR bacteria were conducted with them with in-house python script, specifically investigating their inter-taxonomically transferred ARGs and VFs, to see how broad the range of inter-taxonomic donors and ARGs/ VFs involved could be.

3.4 Results and Discussion

3.4.1 The overall trend of inter-taxonomic transfer of VFs and ARGs

In our research, due to the limitation of our tree-reconciliation method that HGT events between strains could not be detected, all putative HGT events detected are basically inter-species HGT events. This was because the species trees were constructed with 16S-rRNA sequences, which is not a suitable method to distinguish bacteria in the strain levels, though it is universally used as a species marker for bacteria (Janda and Abbott 2007). Thus, the below results will take the 'Whole genome' or 'Inter-species' results as references and will look at events higher than inter-species (from inter-genus) as the inter-taxon HGT events. Furthermore, every gene number in this study was counted without duplicates.

The number of HGT-related genes, VFs, ARGs and their ratio per each inter-taxon and in whole genomes of the bacterial strains are tabulated in Table. 1. 578,924 genes were inter-phylum HGT related genes, 927,464 were inter-class, 1,940,342 were inter-order, 2,559,101 were inter-family, 4,349,718 were inter-genus, 5,405,040 were inter-species (the whole HGT events). Out of each inter-taxon HGT-related genes, 12,807 were VFs and 4,750 were ARGs in inter-species events, 1,400 were VFs 3,159 were ARGs in inter-genus events, 591 were VFs and 1,919 were ARGs in inter-family events, 425 were VFs and 1,469 were ARGs in inter-order events, 198 were VFs and 676 were ARGs in inter-class events, and 109 were VFs and 390 were ARGs in inter-phylum events.

The process of HGT is thought to be conducted in much less occasions between organisms that are distantly related (Ochman, Lawrence, and Groisman 2000; Popa and Dagan 2011; Rohner 2016; Kent et al. 2020). As can be seen from Table 1, the total count of horizontally transferred genes decrease from inter-species events to inter-phylum events. However, what was interesting about our finding was that the total number of ARGs overtook that of VFs as we investigated inter-taxonomic HGT events from species to phylum; the total number of ARGs exceeded since inter-genus events and the discrepancy in the ratio continued to rise until inter-phylum events (Table 3.1). The total number of inter-taxonomically transferred VFs sharply decreased after 'Species' and 'Genus' level, and the number of ARGs was as 3.6 times larger than VFs. The ratio of VFs also decreased, while ARGs relatively remained its constant ratio throughout whole inter-taxon events. This was an unexpected outcome since the total gene number of VFs was significantly greater than the gene number of ARs in the whole genomes and among the HGT events of 20,179 bacterial strains. Thus, although the inter-taxonomic transfer of ARGs are influenced by the phylogenetic distance between the donor and recipient, other reasons must be also underlying, making them more actively participate in inter-taxon HGT events than the VFs. Possible reasons will be further elucidated in the following sections.

Table 3.1 The number of inter-taxonomically transferred genes, virulence factors and antibiotic resistance genes detected by the tree-reconciliation method.

	Whole Genome	Inter-Species HGT Genes	Inter-Genus	Inter-Family	Inter-Order	Inter-Class	Inter-Phylum
Total Count	29,431,921	5,405,040	4,349,718	2,559,101	1,940,342	927,464	578,924
VF Genes	**151,162 (0.513%)**	**12,807 (0.237%)**	**1,400 (0.032%)**	591 (0.023%)	425 (0.022%)	198 (0.021%)	109 (0.019%)
AR Genes	46,369 (0.158%)	4,750 (0.088%)	3,159 (0.073%)	1,919 (0.075%)	1,469 (0.076%)	676 (0.073%)	390 (0.067%)

The percentages in brackets indicate the ratio of VFs and ARGs relative to the total count of genes. Note the relatively steady ratio of ARGs, contrary to the sharply decreasing ratio of VF genes as the inter-taxonomic events goes higher from species-level to phylum-level. The ratio of VF genes is also steady from inter-family to inter-phylum events and the ratio of ARGs exceeds that of VF's from inter-genus to inter-species events. The total number of transferred VF genes significantly decreases too, while the total number of transferred ARGs also decreases but not as much as VF genes do. Although it seems steady, the ratio also decreases as the inter-taxonomic event goes to the phylum-level, but at the phylum-level, the ARGs are as 3.6 times more abundant than VF genes. This was very notable because in the whole genome level, VF genes were as 3.3 times more abundant than AR genes.

3.4.2 Gene Function Approach

To get a possible answer for the trend of inter-taxonomic transfer of VFs and ARGs, we first approached it with the function of the VFs and ARGs in various ways, with the assumption that the inter-taxonomically transferred genes should be beneficial to the survival of bacteria, and that inter-taxonomically transferred ARGs should be more helpful than VFs, so that they are more actively involved in the inter-taxon HGT events. The results of COG categories suggested a possible reason for these, and we ought to suggest possible explanations, by relating to the VFDB category of the inter-taxonomically transferred VFs, and finally the ARO category and drug classes of the inter-taxonomically transferred ARGs.

3.4.2.1 The COG categories of VFs and ARGs in inter-taxonomic HGT events

Figure 3.1 is a collection of graphs illustrating the COG categories of VFs and ARGs in 1. Whole Genome (Graph A and D), 2. HGT events (inter-species events, Graph B and E) and 3. Inter-taxon HGT events (from inter-genus to inter-phylum events, Graph C and F). The COG categories for VFs are consisted of 22 COG categories, whereas for the ARGs, the categories consist of 18 COG categories. Figure 2 and 3 are collections of more comprehensive graphs of the VFs and ARGs COG categories of each inter-taxon HGT events. The COG categories in red are considered to be more critical in the survival of bacteria in terms of the conservation of essential genes in them, and in yellow are less probable to be critical in the survival due to the definition of the categories. This will be elucidated further below.

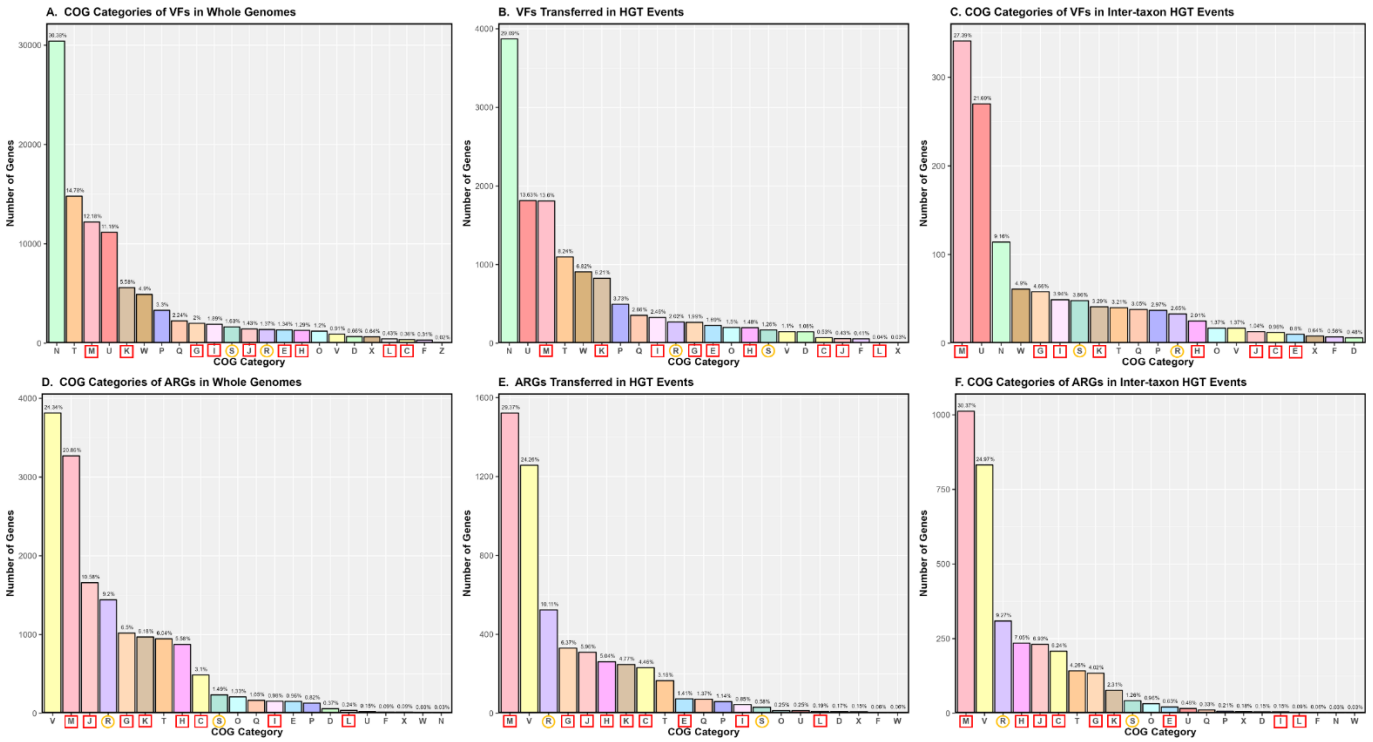

Figure 3.1 The COG categories of VFs and ARGs in whole genome, HGT events and inter-taxon HGT events.

The categories in red are the top COG categories that includes highly conserved essential genes vital for the survival of bacteria, whereas the categories in yellow (R and S) were excluded from the top list due to their definition of function, but still highly conserved. For the VFs, the trends seen in the whole genome are relatively similar with the trends seen in HGT events (inter-species events) and the trends shifts in inter-taxon HGT events, whereas for ARGs, the most frequent COG category changes to V to M. and the trends are similar in inter-taxon HGT events. Overall, higher ratio of top COG categories for bacterial survival can be seen in the ARGs.

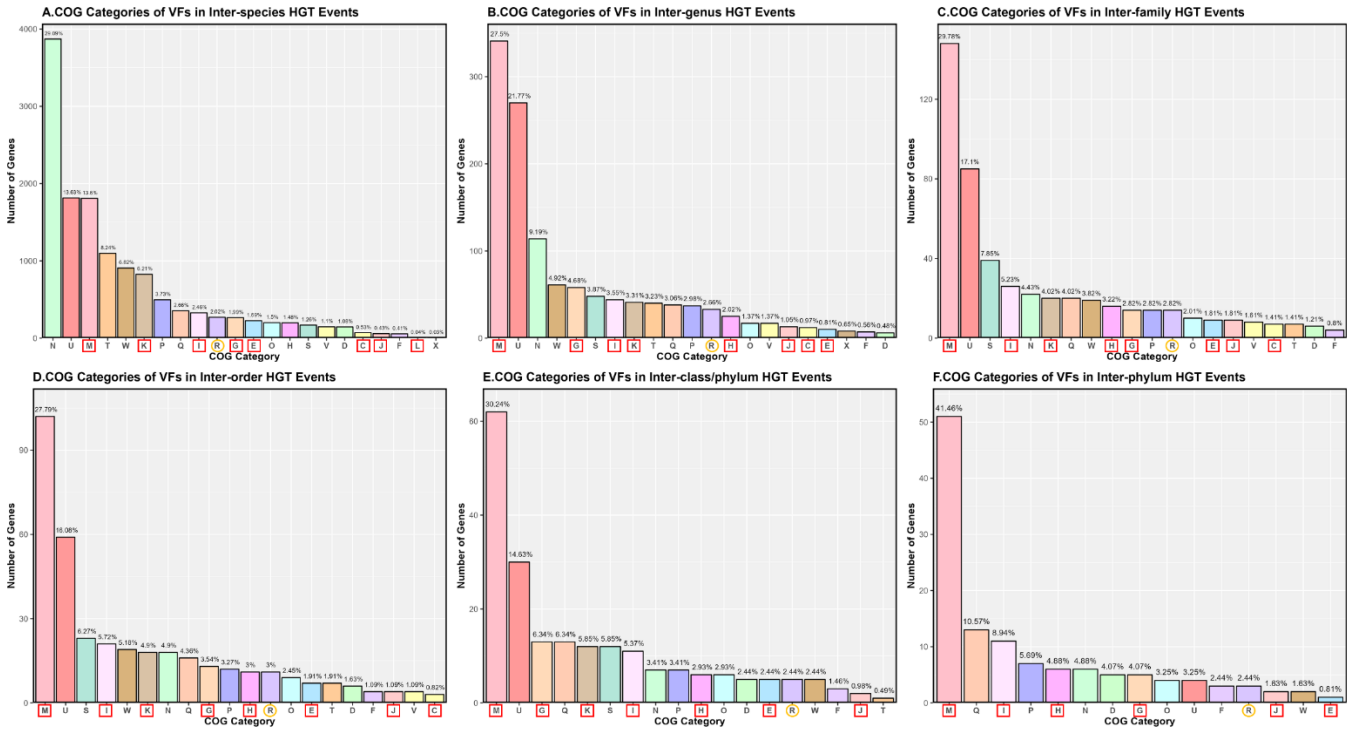

Figure 3.2 The COG categories of VFs transferred via inter-taxon HGT events.

The categories in red are the top COG categories that includes highly conserved essential genes vital for the survival of bacteria, whereas the categories in yellow (R and S) were excluded from the top list due to their definition of function, but still highly conserved. The COG categories of VFs involved in inter-taxon events were first consisted of 22 various categories in inter-species HGT events, and the number of categories gradually decreases as it goes to the inter-phylum HGT events. Note the high ratio of the category N (Cell Motility) in inter-species (where the transfers of VFs were most active) and the category M (Cell wall/membrane/envelope biogenesis) from inter-genus to inter-phylum HGT events; the category M was considered as one of the top COG categories for the survival of bacteria. Relative to the ARGs, the top COG categories are less in their ratio.

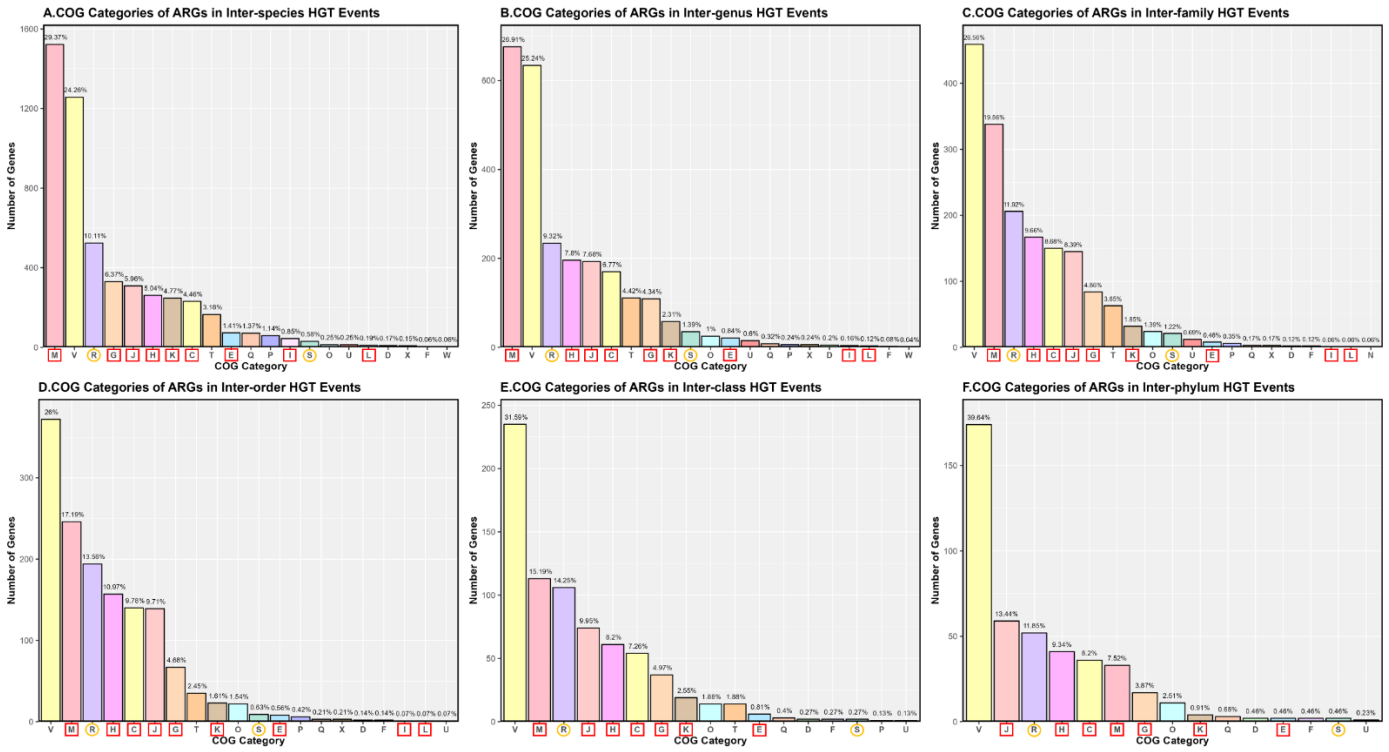

Figure 3.3 The COG categories of ARGs transferred via inter-taxon HGT events.

The categories in red are the top COG categories that includes highly conserved essential genes vital for the survival of bacteria, whereas the categories in yellow (R and S) were excluded from the top list due to their definition of function, but still highly conserved. The COG categories are consisted of 21 different categories, and the top COG categories are relatively highly distributed in ARGs than the VFs do. The category V (Defense mechanism) is prominent from inter-family to inter-phylum HGT events but other top COG categories can be readily seen in high percentage accumulatively.

We expected that the genes related in inter-taxonomic HGT events to be more beneficial in the adaptation and survival of the recipient, as other studies have mentioned that the inter-taxon HGT events are more often occurred in extreme conditions (Loy et al. 2017; Nelson-Sathi et al. 2012; Zhaxybayeva et al. 2009). To investigate the importance of each gene in the survival of the bacteria species, COG categories of each inter-taxon VFs and ARGs were clarified, as exhibited in Figure 1 and 2. The importance of each COG category in regard to the survival rate was determined by referring to the research done by Luo et al. (Luo, Gao, and Lin 2015), and the CEG (cluster of essential genes) database (Ye et al. 2013; Liu et al. 2020) . Luo et al. have taken 23 bacterial species to investigate the highly-conserved essential genes in them and figured out the COG categories of them, and the CEG database is also constructed in a similar way with essential genes and taken example of the COG category to elaborate its data. To conclude, top eight COG categories for the highly conserved essential genes were clarified as: C (Energy production and conversion), E (Amino acid transport and metabolism), G (Carbohydrate transport and metabolism), H (Coenzyme transport and metabolism), I (Lipid transport and metabolism), J (Translation, ribosomal structure and biogenesis), K (Transcription), L (Replication, recombination and repair), and M (Cell wall/membrane/envelope biogenesis) out of 26 categories. Many of them were related to the metabolism of organisms, and other profound functions like transcription and translation. The categories R and S were excluded from the top COG category list because they were defined as ‘General function prediction only’ and ‘Function Unknown’, respectively, but according to the CEG database, those categories also included various highly conserved essential genes

in them, thus illustrated in Figure 3.1 to 3.3 (Ye et al. 2013; Luo, Gao, and Lin 2015).

For the VFs involved in inter-taxon HGT events, the COG category N (Cell Motility) was prominent in whole genome and HGT events, but dramatically shifted to M (Cell wall/membrane/envelope biogenesis) from inter-genus HGT events to inter-phylum HGT events, which is considered one of the top COG categories in bacterial survival. As the inter-taxon HGT event proceeded, the category I (Lipid transport and metabolism) and Q (secondary metabolites biosynthesis, transport and catabolism) stood out too. Although the category I was one of the top COG categories that inter-taxonomically transferred ARGs do not have, the VFs included less top COG categories than ARGs. The category N (Cell motility)'s function will be cooperatively explained in the next section followed by the VFDB category explanation of 'Motility'

For the ARGs involved in inter-taxon HGT events, the COG category M (Cell wall/membrane/envelope biogenesis) stood out the most in inter-species and inter-genus events. It is after inter-family event where V (Defense mechanisms) was the most prominent category in inter-species HGT events (convincing in a way that we are dealing with ARGs, many genes should be included in the category V), but throughout inter-taxon HGT events the top COG category for the survival of bacteria, M (Cell wall/membrane/envelope biogenesis), G (Carbohydrate transport and metabolism), J (Translation, ribosomal structure and biogenesis), H (Coenzyme transport and metabolism), K (Transcription), C (Energy production and conversion) are all included. The category R (General function prediction only) and S (Function unknown) were also two of the prevalent COG categories in inter-taxonomically transferred ARGs,

collectively inferring that the ARGs include many of highly conserved essential genes crucial for the survival of bacteria.

As a conclusion, throughout the inter-taxon HGT event from inter-genus to inter-phylum, various essential genes were involved for both VFs and ARGs, according to the COG category of the genes, and both showed a greater number of the top COG categories as the inter-taxon event got higher from the genus to phylum level. There were more COG categories in ARGs that were critically involved in the survival of bacteria than the VFs had. This could suggest an explanation to the overall trend of inter-taxon HGT events that ARGs are more actively involved in higher inter-taxon events.

3.4.2.2 The VFDB categories of VFs in inter-taxon HGT events

The categories of both VFs are illustrated in Figure 4. The categories are consisted of 14 VFDB categories and the total numbers of genes in each specific category are illustrated from graph A to H; Graph A shows VFs in HGT events (inter-species events), and from Graph B to Graph F are the VFDB categories of VFs involved in inter-taxon (inter-genus to inter-phylum) HGT events. The number of VFDB category drops sharply when comparing the categories of Graph, A and F.

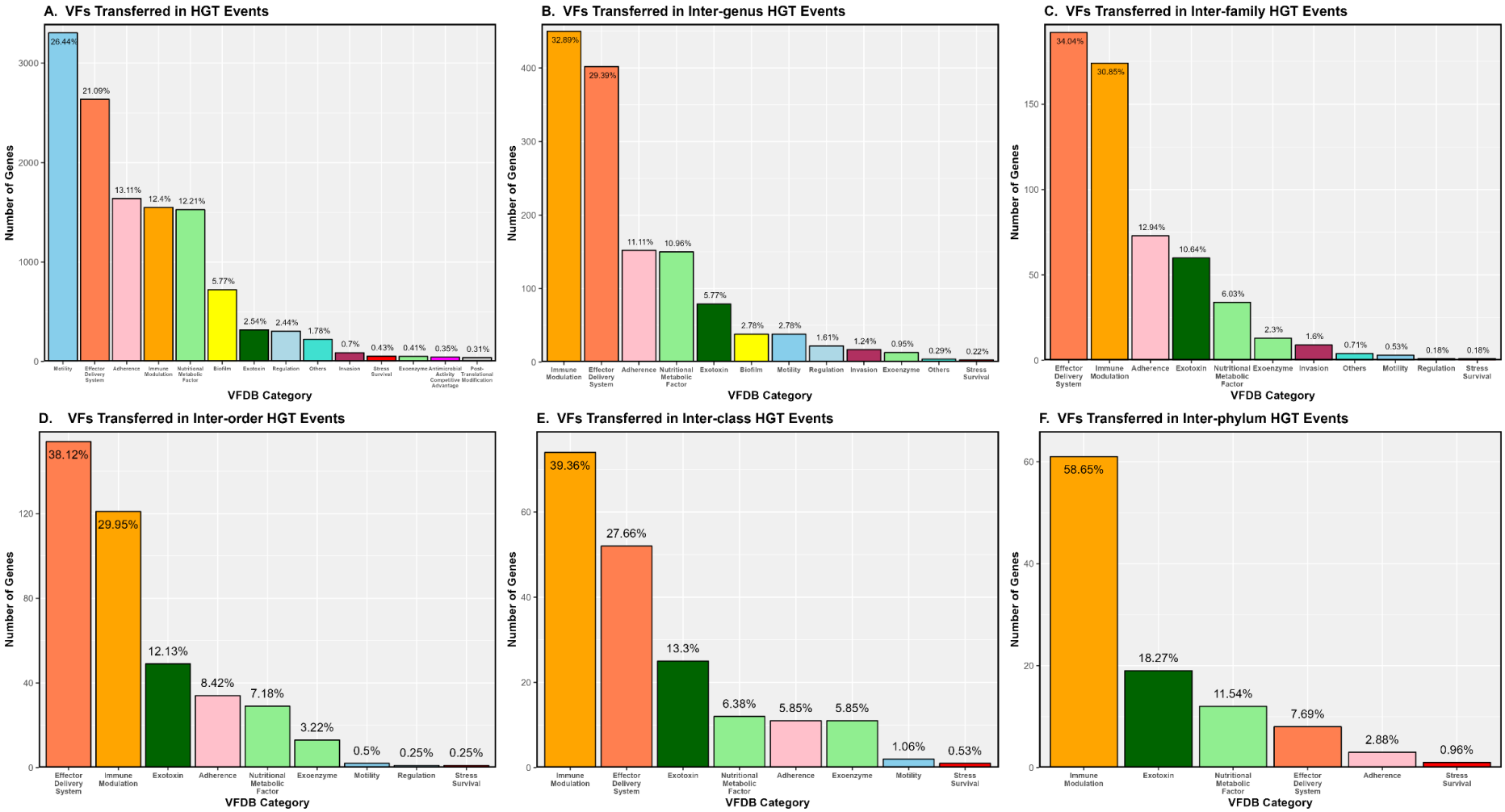

Figure 3.4 The VFDB categories of VFs in HGT events and inter-taxon HGT events.

Each graph shows categories of virulence factors (VFs) which are represented with designated colors. Graph A is illustrated with the VFs, verified by the VFDB, in HGT events, whereas from Graph B to Graph F are illustrated with VFs detected in inter-taxonomic (from inter-genus to inter-phylum) HGT events.

As reported in Figure 4, Graph A, the VFDB category ‘Motility’ was the one with the most transferred genes in HGT events. This was like what we could see from Figure 1, where the COG category N (Cell motility) was the top category in inter-species HGT events. The inter-taxon transfer of VFs was most active in inter-species HGT events (Table 1) and majority of them were related to the ‘Motility’ function, inferring the importance of ‘Motility’ in the survival of bacteria. ‘Motility’ is often responsible for flagellar movements such as flagellar motor stator and switch (Josenhans and Suerbaum 2002). In addition, the ‘Adherence’ category is responsible for the ‘Twitching motility’ of the bacteria (which is also often seen in inter-taxon HGT events), which generally is related to the movement of bacteria pili instead of flagellar (such as some species of gram-negative bacteria). Those proteins related to the movement of the bacteria are directly connected to the rate of the surface colonization; the faster the better (Harshey 2003). One category, the ‘Biofilm’ category, was not much seen in inter-taxon events (only in inter-genus HGT events) but was expected to be crucial in the survival of bacteria. Biofilm is also essential in the localization of bacteria. The biofilm produced by bacteria is not only responsible for providing protective niche to themselves, but can also induce a multispecies community, which ultimately can elevate their metabolic activities and survival (Stoodley et al. 2002).

However, from inter-genus to inter-phylum HGT events, ‘Immune modulation’ and ‘Effector Delivery System’ are the two prominent categories. Genes with ‘Immune modulation’ category can protect the bacteria by immune suppression or immune evasion (Lambertz et al. 2012) and is a direct mechanism that leads to the survival of the bacteria when infiltrating the immune system of the host organisms. The ‘Effector

Delivery System' was ranked high in almost every inter-taxon event (including the HGT events as a whole), except for the inter-phylum events. Effector proteins can contribute to the survival of bacteria by binding and inhibiting the enzymes of the host and suppress host's immune system (Zhao et al. 2017) and is one of the major survival strategy that many bacteria have, by means of manipulating host cell signaling pathways (Mattoo, Lee, and Dixon 2007).

3.4.2.3 The ARO categories of ARGS in inter-taxon HGT events

The ARO categories of the ARGs are illustrated in figure 5, consisting of seven different categories. The total numbers of genes in each specific category are collectively illustrated from graph A to H; Graph A shows ARGs of HGT events, and from Graph B to Graph F are ARGs involved in inter-taxon (inter-genus to inter-phylum) HGT events and can be seen that the number of categories remains relatively constant throughout the graphs. This could be just because the ARO category is not as specific as the VFDB category but considering the ratio of each category for both ARGs and VFs, and the shifts of places of the categories, it is probable that the ARO categories of ARGs infer the stability of ARGs in inter-taxon HGT events.

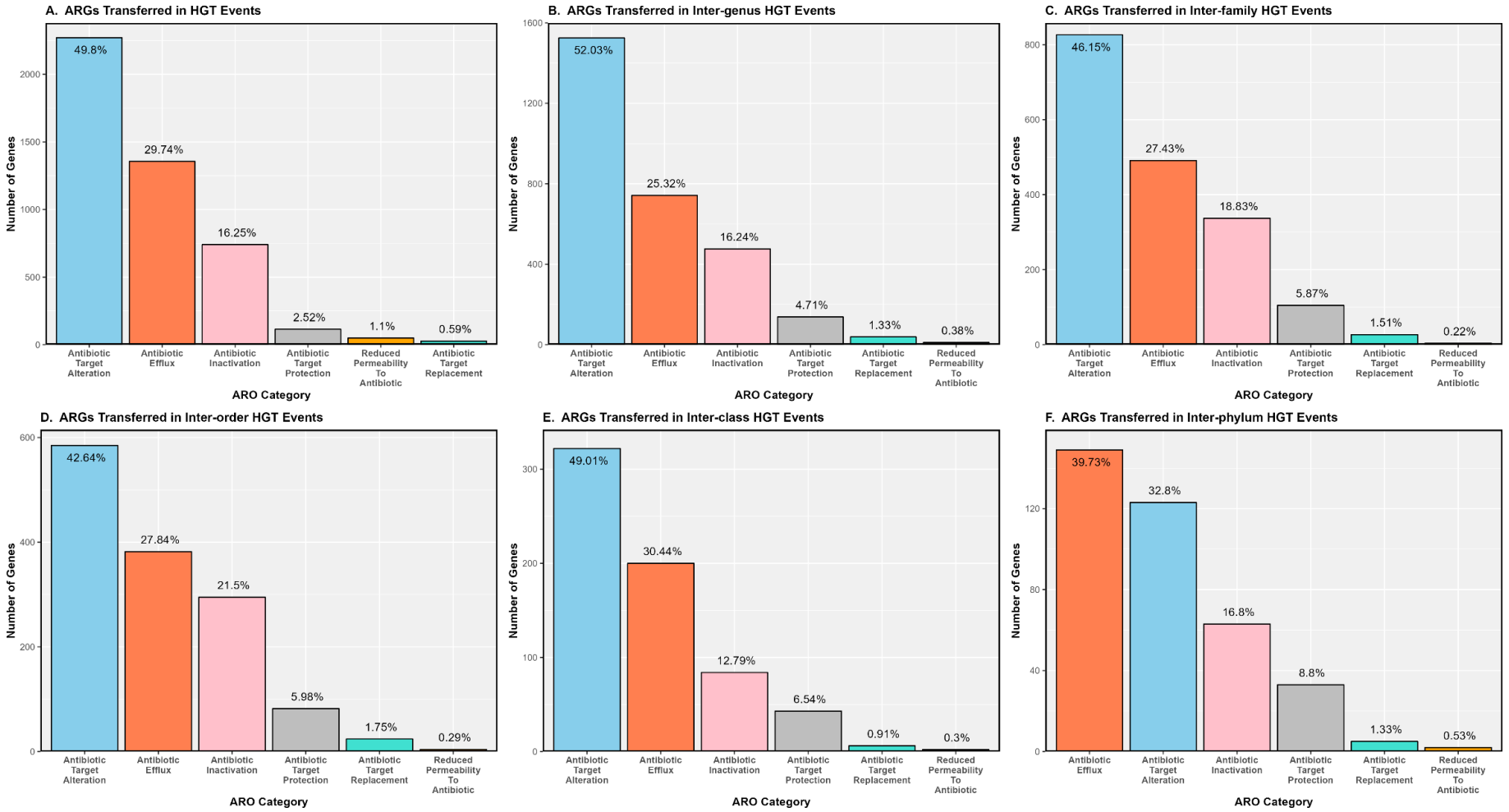

Figure 3.5 The ARO categories of ARGs in HGT events and inter-taxon HGT events.

Each graph shows ARO categories of antibiotic resistance genes (ARGs) which are represented with designated colors. Graph A is illustrated with the ARGs verified by the program CARD-rgi, in HGT events, whereas form Graph B to Graph F are illustrated with ARGs detected in inter-taxonomic (from inter-species to inter-phylum) HGT events.

For the ARO categories, ‘Antibiotic Target Alteration’ was the most frequent category throughout whole inter-taxon HGT events, followed by ‘Antibiotic Efflux’. The alteration can be in various format, either by natural variations like mutations or can be also acquired by processes like HGT, and since the target of antibiotics are quite specific, a minor change can have a major effect in resistance mechanism (Kapoor, Saigal, and Elongavan 2017). This could be the reason of high ratio of ‘Antibiotic Target Altering’ ARGs in the inter-taxon HGT events, due to its practicality. Here, the importance of cell wall comes again. One of the major types of resistance mechanism is by altering the precursors of the cell wall, typically, a change in D-alanyl-alanine to D-alanyl-lactase can bring resistance to the glycopeptides (which is known as the inhibitor of cell wall biosynthesis) (Giedraitienė et al. 2011). The efflux pumps are basically membrane proteins that are the underlying mechanism of resistance, triggered by change in concentration of intracellular area, and can be very specific to antibiotics. All antibiotics other than polymyxin can trigger the activation of efflux system of bacteria (Lambert 2002; Giedraitienė et al. 2011). The high percentage of ‘Antibiotic Efflux’ in inter-taxonomically transferred ARGs is probably due to its fundamental role in the resistance mechanism

3.4.2.4 The drug classes of ARGs in inter-taxonomic HGT events

In addition to the results above, since there were more essential COG categories in inter-taxonomically transferred ARGs, the drug classes searched by the CARD-rgi program was further studied to identify proliferating drug classes and get a possible explanation for the importance of ARGs in the survival of bacteria. Many ARO

categories found in inter-taxon HGT evens were related to ‘Antibiotic Target Alteration’. This could relate to the results seen in the types of drug classes (Figure 4), which many of the resistance mechanisms arose from the modification of the target of the antibiotics (like the resistance mechanism to glycopeptides, fluoroquinolone and cephalosporin (Yushchuk, Binda, and Marinelli 2020; Hooper 1999; Livermore 1987)). In total, 40 out of 49 drug classes from CARD-rgi was illustrated in Figure 4, Graph A (whole genome), 34 out of 49 drug classes in Graph B (HGT events) and 32 out of 49 in Graph B (inter-taxon HGT events). Many of the resistances were against drug classes that are highly important in human therapeutics (Van et al. 2020), and as illustrated, two classes were prominently seen in inter-taxon HGT events: ‘Glycopeptide Antibiotic’, ‘Disinfecting agents and Antiseptics’.

Glycopeptide antibiotic was the most frequently seen drug class in every inter-taxon event and is thought to be one of most effective antibiotics when treating infections from multidrug-resistant gram-positive bacteria (Yushchuk, Binda, and Marinelli 2020). The glycopeptide antibiotics are inhibitors of cell wall biosynthesis, and was introduced in 1958. (Acharya et al. 2021) This can also be related to the fact that majority of inter-taxonomically transferred VFs and ARGs were in the M category, which is defined to be ‘Cell wall/membrane/envelope biogenesis’, suggesting that the most crucial factor in the survival of bacteria can be associated with their cell wall.

The ‘Disinfecting agents and Antiseptic’ class was also readily seen in every inter-taxon HGT events just after the two antibiotics above. Both disinfectants and antiseptics are widely used in hospital as wide variety of active biocides, effective in

broad spectrum against gram-positive and negative bacteria. Antiseptics are biocides that can be more widely used as antimicrobial agents than antibiotics, destroying or inhibiting microorganism growth, while disinfectants are products of biocides (McDonnell and Russell 1999). The mechanisms of disinfectants and antiseptics are very various, can have numerous actions on bacteria including DNA and protein activity destruction, Membrane permeability increase (Chlorine), Cell wall denaturation (Alcohol) and disorganization (Quats) and can even attack essential cell components (Hydrogen peroxide) (Chapman 2003). Since many of the mechanism are driven by influx of those biocides (products) into the target cell, it is probable that this could also be related to the high M COG category of inter-taxon HGT-related genes. Further, various actions of disinfectants and antiseptics could account for the high ratio of 'Disinfecting agents and Antiseptic' category.

Other than the two classes, 'Aminoglycoside Antibiotic', 'Tetracycline antibiotic', 'Fluoroquinolone antibiotic' and 'Cephalosporin' were also four of the drug classes that were ranked high in Figure 4. The aminoglycoside antibiotic was first discovered in 1943 as streptomycin and is active on both gram-positive and negative bacteria. Aminoglycosides act on the 30S subunit of the ribosome to inhibit protein synthesis by inducing mistranslation (Schatz et al. 2005). Many of aminoglycoside antibiotics (such as kanamycin, neomycin, including streptomycin) are biosynthetic, can be produced by different species of *Streptomyces* and *Micromonospora* (Luan et al. 2020). This can again be related to the high number of J category (Translation, ribosomal structure and biogenesis) in the inter-taxonomically transferred ARGs.

Tetracycline antibiotics are broad spectrum antibiotics discovered in 1940, effective to both gram-positive and negative bacteria (Daghrir and Drogui 2013). The mechanism of tetracyclines are initiated by penetration of bacteria and induce interruption on protein synthesis or destruction of membrane. Here, we can see that the COG category M comes in action as defense mechanism. Cephalosporin is also a very widely used antibiotics in the world, and can be broadly used for gram-positive and negative bacteria (Thompson and Wright 1983). Antibiotics of cephalosporin typically bind to penicillin-binding proteins and the resistance is gained when the penicillin-binding proteins are altered by the 'Antibiotic Target Alteration' resistance mechanism (Livermore 1987).

Intriguingly, fluoroquinolone antibiotic resistance arise typically by mutation (Lehtinen et al. 2020), and was the fifth most drug class found in HGT events (inter-species events). The mechanism of fluoroquinolone antibiotic resistance can be categorized in two ways: 1. Alterations of drug target, and 2. Limiting the drug inflow (Hooper 1999). This could be related to the ARO category (described above), which 'Antibiotic Target Alteration' and 'Antibiotic Efflux' were the two ARO category primarily seen across all the inter-taxon HGT events. It was also a point to review that a typical mutation-derived resistance can be one of the highly horizontally transferred resistance mechanism.

What many of drug classes have in common was that they mainly interact with cell wall/membrane of bacteria for the penetration, by influx mechanism or binding to target proteins to induce antibiotic actions. This is probably related to the high ratio of

the M COG category (which is related to the cell wall/membrane biogenesis), and high ratio of ‘Antibiotic Target Alteration’ and ‘Antibiotic Efflux’ defense mechanism shown in the ARO categories.

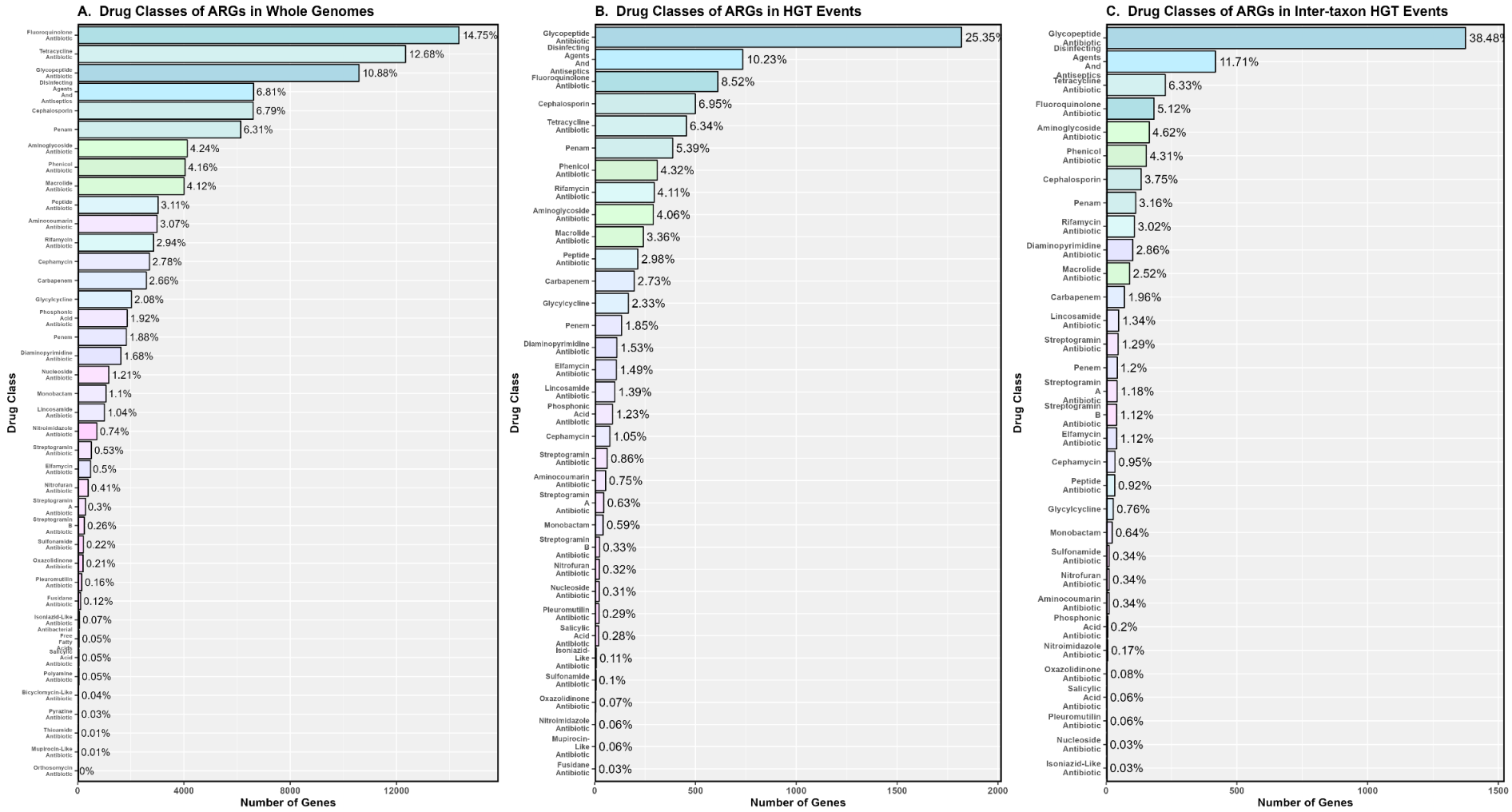

Figure 3.6 The drug classes of ARGs transferred in HGT events and inter-taxon HGT events.

The drug classes searched by the CARD-rgi program was studied to identify proliferating drug classes and get a possible explanation for the importance of ARGs in the survival of bacteria. Resistance to 'Glycopeptide antibiotic' and 'Disinfecting agents and Antiseptics' stood out the most in HGT events and inter-taxon HGT events.

3.4.3 MDR bacteria and their inter-taxonomic HGT events of VFs and ARGs

In addition to the results above, we also investigated famous MDR bacteria species to see the inter-taxon transfer events with different cases. Since MDR bacteria get their VFs and ARGs from wide range of different donors, we assumed that the trends in inter-taxon transfer VFs and ARGs of MDR bacteria species would have contributed to the overall trend of inter-taxonomical transfer of VFs and ARGs. Furthermore, many of drug classes involved in inter-taxon HGT events were related to MDR bacteria. Typically, a bacteria species with an antibiotic resistance become MDR upon acquiring another antibiotic resistance. For instance, MRSA is *S. aureus* with additional resistance to methicillin (Deurenberg and Stobberingh 2008) and MRPA is *P. aeruginosa* with collective resistance including aminoglycoside resistance (Arruda et al. 1999).

The trend of inter-taxonomically transferred VFs and ARGs in six MDR candidates (in order of *K. pneumoniae, S. enterica, S. aureus, P. aeruginosa, E. faecium* and *E. faecalis*) are tabulated in Table 2, listed with VFs and ARGs in whole genomes of each species, followed by the number of inter-taxon VFs and ARGs they received from different donor species (sum of all VFs and ARGs of all MDR strain per each species), and the ratio of them relative to the number of VFs and ARGs in whole genomes. For each MDR candidate, the ratio of inter-taxon ARGs were much higher than the ratio of inter-taxon VFs, average of 1.670% for VFs and 13.214% for ARGs, following the trend of the overall VFs and ARGs transfer in inter-taxon HGT events. The average percentage of inter-taxon ARGs to ARGs in whole genome was as 7.9 times greater than the percentage of inter-taxon VFs, even though the number of VFs in the

whole genomes and VFs transferred via inter-species HGT events were much greater than that of ARGs (Table 3.2). A deeper insight into the inter-taxon HGT events per each MDR bacteria candidates is illustrated in Graphs of Figure 3.7.

Table 3.2 A summary table of the total number of inter-taxonomically transferred VFs and ARGs of six MDR candidates and the ratio of them relative to VFs and ARGs in their whole genomes.

MDR Species	VFs in Whole Genomes	Inter-taxon VFs	ARGs in Whole Genomes	Inter-taxon ARGs
Klebsiella pneumoniae	6,175	191 (3.093%)	1,293	186 (14.385%)
Salmonella enterica	11,138	84 (0.754%)	1,570	135 (8.600%)
Staphylococcus auereus	4,848	13 (0.268%)	886	58 (6.546%)
Pseudomonas aeruginosa	10,603	23 (0.217%)	1,986	73 (3.680%)
Enterococcus faecium	232	6 (2.586%)	224	60 (26.786%)
Enterococcus faecalis	419	13 (3.103%)	140	27 (19.286%)
		Average: 1.670%		**Average: 13.214%**

The VFs and ARGs of each MDR candidate are listed as in whole genome and in inter-taxon HGT events. As seen, the average percentage of ARGs transferred in inter-taxonomic HGT events are much higher than that of VFs. Even the total number of ARGs are much higher than VFs (except for *K. pneumoniae* by small number), regardless of the total number of VFs in whole genome.

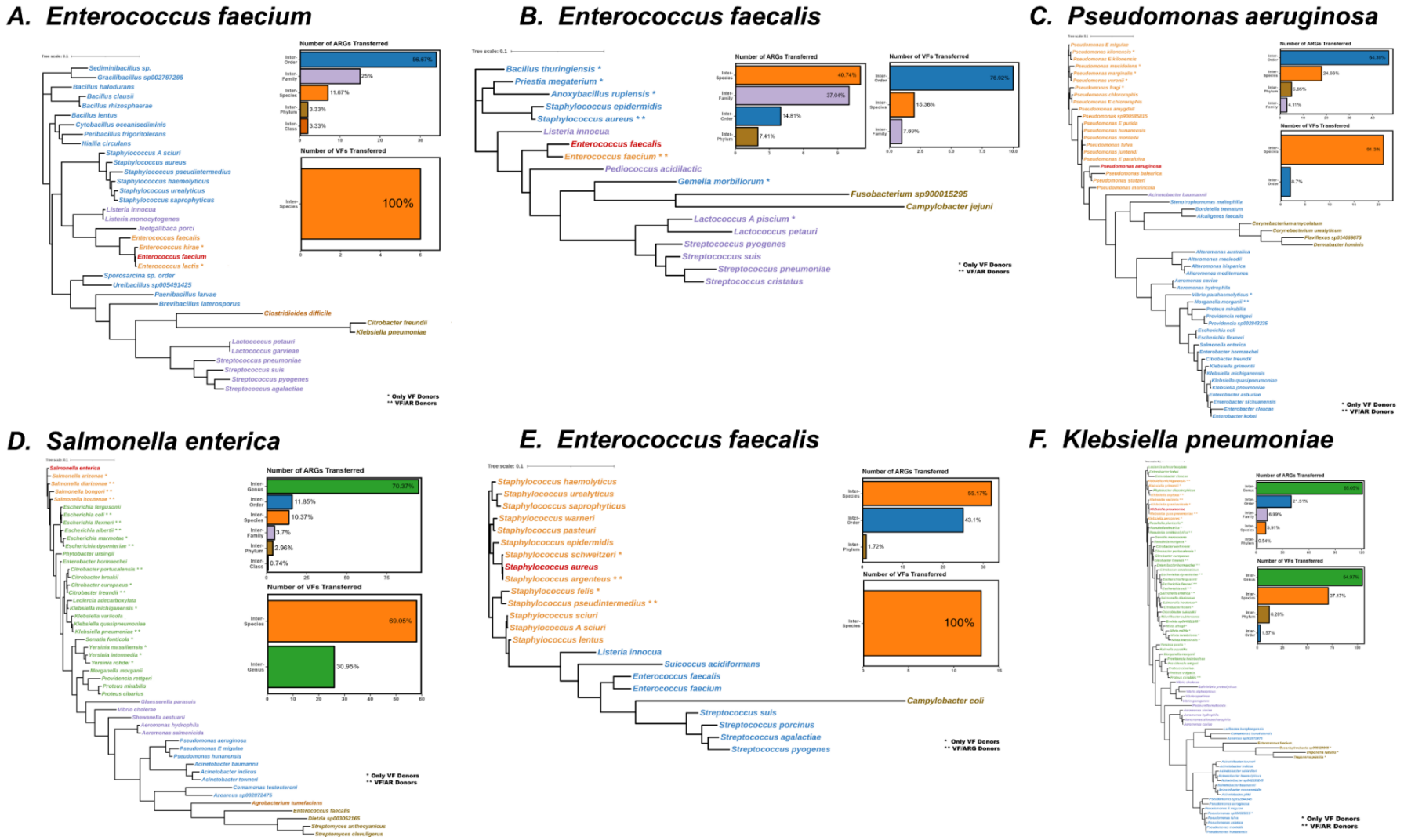

Figure 3.7 Graphs of MDR candidate species illustrating their inter-taxon (including inter-species) donors of VFs and ARGs, and the number and ratio of each inter-taxon events.

The candidates were chosen according to the number of inter-taxon HGT-related genes they have.

Graph A to F in Figure 3.7 are phylogenetic trees of MDR candidates showing their broad range of inter-taxonomic donors of VFs and ARGs and relative phylogenetic distances. On the right upper corner of each graph are two additional graphs showing the number of ARGs and VFs of each MDR candidate and the ratio of each inter-taxon HGT event involved. The donor species in the phylogenetic tree are colored with the same color as the variables in the two graphs, to illustrate which event the donors are associated. In the phylogenetic trees, species involved in inter-taxon HGT events of VFs only are indicated with a single asterisk, for those species involved in events of both VFs and ARGs are indicated with a double asterisk. Other species with no asterisks is involved in events of ARGs only. The trends and diversity of drug classes of inter-taxonomically transferred ARGs are further visualized in Supplementary Figure 1. The spectrum of donors for each candidate was very broad, all of them including species from other orders and phylum. As a whole (586 ARGs and 330 VFs in total), for inter-taxonomically transferred ARGs, most of them were transferred via inter-genus and inter-order HGT events (216 and 213 genes, respectively), while for VFs, the genes were mostly transferred through inter-species and inter-genus HGT events (171 and 131 genes, respectively). This result showed a concordance with the result of Table. 1 to some extent, that ARGs are more prone to be transferred via inter-taxonomic events, but also showed an unexpected result that ARGs are not always mostly transferred via inter-species events.

K. pneumoniae and *S. enterica* were the two MDR species that received most of their ARGs from inter-genus HGT events, and *K. pneumoniae* was also exceptional in a way that the total number of inter-taxonomically transferred VFs outnumbered that of ARGs (191 VFs to 186 ARGs). The inter-taxonomic VFs of *K. pneumoniae* are predicted to be mostly come from inter-genus HGT events, while for *S. enterica*, most of them came from inter-species HGT events. Except for *K. pneumoniae*, the trend was as expected for *S. enterica* that ARGs transfer more in inter-taxon HGT events than VFs do.

P. aeruginosa and *E. faecium* were the two MDR species that received most of their ARGs from inter-order HGT events. *P. aeruginosa* got most of its VFs from inter-species HGT events, and for *E. faecium*, 100% of inter-taxon VFs were from inter-species HGT events. The two candidates showed that ARGs can be transferred in inter-taxon HGT events as high as the inter-order level, even with a greater number than VFs transferred through inter-species HGT events.

S. aureus and *E. faecalis* were the two MDR species that received most of their ARGs from inter-species HGT events. For *S. aureus*, 100% of its inter-taxon VFs were retrieved from inter-species HGT events, while for *E. faecalis*, the VFs came from mostly inter-order HGT events. *E. faecalis* was exceptional too, that its VFs came through higher inter-taxon HGT events than its ARGs did. But for S. aureus, the total number of ARGs outnumbered VFs, with the higher inter-taxon HGT events involved in their transfer.

Each species in the MDR candidate showed its distinct pattern in the transfer of

VFs and ARGs in inter-taxon HGT events, but as a whole, the trend was as expected that ARGs are much actively involved in inter-taxon HGT events, on average. Based on the trend observed in this study, clarifying each pattern will require more observation and research to be done in the future, in order to conquer MDR bacteria for better clinical conditions.

3.5 Conclusion

By performing the tree-reconciliation method, which takes sets of pairs of gene trees and species trees derived from orthologous gene sets of 20,179 bacterial strains, genes that were involved in HGT events. In this research, we focused on inter-taxonomic transfer (inter-genus, family, order, class and phylum) of VFs and ARGs and found out that the ARGs were more involved in inter-taxon events than the VFs do. To explain this trend, we investigated the functional profiling of each gene, and could relate it to the COG categories of the inter-taxonomically transferred genes. According to the COG categories, more ARGs were classified as essential genes in the survival of bacteria and we supported this by giving explanations regarding the ARO category and drug classes of them.

Subsequently, we sought to find the same trend in six MDR bacteria species (*Pseudomonas aeruginosa, Salmonella enterica, Klebsiella pneumoniae, Staphylococcus aureus, Enterococcus faecium and Enterococcus faecalis*), and the trend of inter-taxonomically transferred VFs and ARGs was in accordance with the overall trend, even much higher ratio of ARGs being transferred in inter-taxon HGT events than the overall trend. However, each candidate showed distinct patterns, which some of them did not follow the trend. More works are needed to completely comprehend the specific trend of inter-taxonomic HGT events in MDR species.

Our limitation in this review was that we only could suggest a possible reason for the fact that there are more ARGs in inter-taxon HGT events than VFs but did not investigate for the reason why VFs are so less in inter-HGT events. Furthermore, HGT

evens inferred in this research is of inter-species events, no intra-species events were included to represent the whole HGT events. This was due to the limitation we also had before when constructing the HGTree v2.0 database, since the species trees were made with 16S-rRNA sequences. The trend of VFs and ARGs in inter-taxon HGT events will be much clearly seen with a proper method to clarify HGT events between species (inter-strain HGT events).

General Discussion

Understanding horizontal gene transfer events were very hard due to the extensive computational effort needed and the challenges in the conceptual approaches. However, with the rapid development in the field of bioinformatics, we are now getting more and more close to knowing HGT events in depth.

In chapter 2, I have conducted a comprehensive analysis of HGT events using the tree-reconciliation method. This method utilizes the tree-reconciliation method which takes pairs of species trees and gene trees to identify incongruences that indicate potential HGT events. By comparing these trees, I was able to detect and characterize HGT events in a dataset of 20,179 bacterial strains. I have developed the HGTree v2.0 database, which houses approximately eight times more genomes and HGT event results compared to the previous version. The database is a user-friendly platform that enables researchers to explore and analyze HGT events, providing valuable insights into the evolutionary dynamics and genetic variations driven by HGT.

In chapter 3, my focus was on understanding the inter-taxonomic HGT events involving virulence factors (VFs) and antibiotic resistance genes (ARGs) among a diverse range of bacterial strains. I explored the HGTree v2.0 database to investigate the transfer of these crucial genetic elements across different taxonomic groups. By analyzing the putative HGT events, I gained insights into the potential driving factors behind the extreme transfer of VFs and ARGs between organisms of different taxonomy. The findings emphasize the need for a comprehensive understanding of HGT in combating the challenges of infectious diseases and antimicrobial resistance.

In conclusion, these two papers contribute to our understanding of horizontal gene transfer (HGT) and its implications in prokaryotic evolution. Together, these studies enhance our knowledge of HGT and its impact on the genetic diversity and adaptation of prokaryotes, ultimately contributing to advancements in various fields such as medicine, microbiology, and evolutionary biology.

References

Acharya, Yash, Shaown Bhattacharyya, Geetika Dhanda, and Jayanta Haldar. 2021. 'Emerging roles of glycopeptide antibiotics: Moving beyond gram-positive bacteria', *ACS Infectious Diseases*, 8: 1-28.

Adato, Orit, Noga Ninyo, Uri Gophna, and Sagi Snir. 2015. 'Detecting horizontal gene transfer between closely related taxa', *PLoS computational biology*, 11: e1004408.

Alcock, Brian P, William Huynh, Romeo Chalil, Keaton W Smith, Amogelang R Raphenya, Mateusz A Wlodarski, Arman Edalatmand, Aaron Petkau, Sohaib A Syed, and Kara K Tsang. 2023. 'CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database', *Nucleic Acids Research*, 51: D690-D99.

Alcock, Brian P, Amogelang R Raphenya, Tammy TY Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, and Sihan Liu. 2020. 'CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database', *Nucleic Acids Research*, 48: D517-D25.

Andam, Cheryl P, and J Peter Gogarten. 2011. 'Biased gene transfer in microbial evolution', *Nature Reviews Microbiology*, 9: 543-55.

Arnold, Brian J, I-Ting Huang, and William P Hanage. 2022. 'Horizontal gene transfer and adaptive evolution in bacteria', *Nature Reviews Microbiology*, 20: 206-18.

Arruda, Erico AG, Ivan S Marinho, Marcos Boulos, Sumiko I Sinto, Caio M Mendes, Carmen P Oplustil, Helio Sader, Carlos E Levy, and Anna S Levin. 1999. 'Nosocomial infections caused by multiresistant Pseudomonas aeruginosa', *Infection Control & Hospital Epidemiology*, 20: 620-23.

Bansal, Mukul S, Eric J Alm, and Manolis Kellis. 2012. 'Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss', *Bioinformatics*, 28: i283-i91.

Bansal, Mukul S, Manolis Kellis, Misagh Kordi, and Soumya Kundu. 2018.

'RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss', *Bioinformatics*, 34: 3214-16.

Baugher, JL, E Durmaz, and TR Klaenhammer. 2014. 'Spontaneously induced prophages in Lactobacillus gasseri contribute to horizontal gene transfer', *Applied and environmental microbiology*, 80: 3508-17.

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. 2011. 'D³ data-driven documents', *IEEE transactions on visualization and computer graphics*, 17: 2301-09.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. 'Fast and sensitive protein alignment using DIAMOND', *Nature methods*, 12: 59-60.

Bud, Robert. 2007. *Penicillin: triumph and tragedy* (Oxford University Press on Demand).

Buels, Robert, Eric Yao, Colin M Diesh, Richard D Hayes, Monica Munoz-Torres, Gregg Helt, David M Goodstein, Christine G Elsik, Suzanna E Lewis, and Lincoln Stein. 2016. 'JBrowse: a dynamic web platform for genome visualization and analysis', *Genome biology*, 17: 1-12.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. 'BLAST+: architecture and applications', *BMC bioinformatics*, 10: 1-9.

Cantalapiedra, Carlos P, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. 'eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale', *Molecular biology and evolution*, 38: 5825-29.

Caro-Quintero, Alejandro, and Konstantinos T Konstantinidis. 2015. 'Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria', *The ISME journal*, 9: 958-67.

Chapman, John S. 2003. 'Disinfectant resistance mechanisms, cross-resistance, and co-resistance', *International biodeterioration & biodegradation*, 51: 271-76.

Charleston, Michael A, and Susan L Perkins. 2006. 'Traversing the tangle: algorithms and applications for cophylogenetic studies', *Journal of biomedical informatics*,

39: 62-71.

Chaumeil, Pierre-Alain, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. 2020. "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." In.: Oxford University Press.

Choi, Youngseok, Sojin Ahn, Myeongkyu Park, SaetByeol Lee, Seoae Cho, and Heebal Kim. 2023. 'HGTree v2. 0: a comprehensive database update for horizontal gene transfer (HGT) events detected by the tree-reconciliation method', *Nucleic Acids Research*, 51: D1010-D18.

Cross, Alan S. 2008. 'What is a virulence factor?', *Critical Care*, 12: 1-2.

Daghrir, Rimeh, and Patrick Drogui. 2013. 'Tetracycline antibiotics in the environment: a review', *Environmental chemistry letters*, 11: 209-27.

Davison, Helen C, Mark EJ Woolhouse, and J Chris Low. 2000. 'What is antibiotic resistance and how can we measure it?', *TRENDS in Microbiology*, 8: 554-59.

de Koning, Audrey P, Fiona SL Brinkman, Steven JM Jones, and Patrick J Keeling. 2000. 'Lateral gene transfer and metabolic adaptation in the human parasite Trichomonas vaginalis', *Molecular biology and evolution*, 17: 1769-73.

Deurenberg, Ruud H, and Ellen E Stobberingh. 2008. 'The evolution of Staphylococcus aureus', *Infection, genetics and evolution*, 8: 747-63.

Dilthey, Alexander, and Martin J Lercher. 2015. 'Horizontally transferred genes cluster spatially and metabolically', *Biology direct*, 10: 1-8.

Doolittle, W Ford. 1999. 'Phylogenetic classification and the universal tree', *Science*, 284: 2124-28.

Galperin, Michael Y, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. 2021. 'COG database update: focus on microbial diversity, model organisms, and widespread pathogens', *Nucleic acids research*, 49: D274-D81.

Garcia-Vallve, Santiago, Eduard Guzmán, MA Montero, and Antoni Romeu. 2003. 'HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes', *Nucleic acids research*, 31: 187-89.

Giedraitienė, Agnė, Astra Vitkauskienė, Rima Naginienė, and Alvydas Pavilonis.

2011. 'Antibiotic resistance mechanisms of clinically important bacteria', *Medicina*, 47: 19.

Gosselin, Sean, Matthew S Fullmer, Yutian Feng, and Johann Peter Gogarten. 2022. 'Improving Phylogenies Based on Average Nucleotide Identity, Incorporating Saturation Correction and Nonparametric Bootstrap Support', *Systematic Biology*, 71: 396-409.

Hall, Rebecca J, Fiona J Whelan, James O McInerney, Yaqing Ou, and Maria Rosa Domingo-Sananes. 2020. 'Horizontal gene transfer as a source of conflict and cooperation in prokaryotes', *Frontiers in Microbiology*, 11: 1569.

Harshey, Rasika M. 2003. 'Bacterial motility on a surface: many ways to a common goal', *Annual Reviews in Microbiology*, 57: 249-73.

Hooper, David C. 1999. 'Mechanisms of fluoroquinolone resistance', *Drug resistance updates*, 2: 38-55.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, and Lars J Jensen. 2019. 'eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses', *Nucleic acids research*, 47: D309-D14.

Janda, J Michael, and Sharon L Abbott. 2007. '16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls', *Journal of clinical microbiology*, 45: 2761-64.

Jeong, Hyeonsoo, Samsun Sung, Taehyung Kwon, Minseok Seo, Kelsey Caetano-Anollés, Sang Ho Choi, Seoae Cho, Arshan Nasir, and Heebal Kim. 2016. 'HGTree: database of horizontally transferred genes determined by tree reconciliation', *Nucleic acids research*, 44: D610-D19.

Johnson, Mark, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. 2008. 'NCBI BLAST: a better web interface', *Nucleic Acids Research*, 36: W5-W9.

Josenhans, Christine, and Sebastian Suerbaum. 2002. 'The role of motility as a virulence factor in bacteria', *International Journal of Medical Microbiology*,

291: 605-14.

Junier, Thomas, and Evgeny M Zdobnov. 2010. 'The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell', *Bioinformatics*, 26: 1669-70.

Kanehisa, Minoru, and Susumu Goto. 2000. 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic acids research*, 28: 27-30.

Kapoor, Garima, Saurabh Saigal, and Ashok Elongavan. 2017. 'Action and resistance mechanisms of antibiotics: A guide for clinicians', *Journal of anaesthesiology, clinical pharmacology*, 33: 300.

Keeling, Patrick J, and Jeffrey D Palmer. 2008. 'Horizontal gene transfer in eukaryotic evolution', *Nature Reviews Genetics*, 9: 605-18.

Kent, Alyssa G, Albert C Vill, Qiaojuan Shi, Michael J Satlin, and Ilana Lauren Brito. 2020. 'Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C', *Nature Communications*, 11: 4379.

Kim, Kyung Mo, Samsun Sung, Gustavo Caetano-Anolles, Jae Yong Han, and Heebal Kim. 2008. 'An approach of orthology detection from homologous sequences under minimum evolution', *Nucleic acids research*, 36: e110-e10.

Kloub, Lina, Sean Gosselin, Matthew Fullmer, Joerg Graf, Johann Peter Gogarten, and Mukul S Bansal. 2021. 'Systematic detection of large-scale multigene horizontal transfer in prokaryotes', *Molecular biology and evolution*, 38: 2639-59.

Koonin, Eugene V, Kira S Makarova, and L Aravind. 2001. 'Horizontal gene transfer in prokaryotes: quantification and classification', *Annual Reviews in Microbiology*, 55: 709-42.

Krzywinski, Martin, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. 2009. 'Circos: an information aesthetic for comparative genomics', *Genome research*, 19: 1639-45.

Kundu, Soumya, and Mukul S Bansal. 2018. 'On the impact of uncertain gene tree rooting on duplication-transfer-loss reconciliation', *BMC bioinformatics*, 19:

21-31.

Kunin, Victor, and Christos A Ouzounis. 2003. 'The balance of driving forces during genome evolution in prokaryotes', *Genome research*, 13: 1589-94.

Lagesen, Karin, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W Ussery. 2007. 'RNAmmer: consistent and rapid annotation of ribosomal RNA genes', *Nucleic Acids Research*, 35: 3100-08.

Lambert, PA1308633. 2002. 'Mechanisms of antibiotic resistance in Pseudomonas aeruginosa', *Journal of the royal society of medicine*, 95: 22.

Lambertz, Ulrike, Judith Maxwell Silverman, Devki Nandan, W Robert McMaster, Joachim Clos, Leonard J Foster, and Neil E Reiner. 2012. 'Secreted virulence factors and immune evasion in visceral leishmaniasis', *Journal of leukocyte biology*, 91: 887-99.

Lawrence, Jeffrey G, and Howard Ochman. 2002. 'Reconciling the many faces of lateral gene transfer', *TRENDS in Microbiology*, 10: 1-4.

Lehtinen, Sonja, Claire Chewapreecha, John Lees, William P Hanage, Marc Lipsitch, Nicholas J Croucher, Stephen D Bentley, Paul Turner, Christophe Fraser, and Rafał J Mostowy. 2020. 'Horizontal gene transfer rate is not the primary determinant of observed antibiotic resistance frequencies in Streptococcus pneumoniae', *Science advances*, 6: eaaz6137.

Li, Li, Christian J Stoeckert, and David S Roos. 2003. 'OrthoMCL: identification of ortholog groups for eukaryotic genomes', *Genome research*, 13: 2178-89.

Li, Xiangchen, Wenjun Tong, Lina Wang, Siddiq Ur Rahman, Gehong Wei, and Shiheng Tao. 2018. 'A novel strategy for detecting recent horizontal gene transfer and its application to Rhizobium strains', *Frontiers in Microbiology*, 9: 973.

Liu, Bo, Dandan Zheng, Siyu Zhou, Lihong Chen, and Jian Yang. 2022. 'VFDB 2022: a general classification scheme for bacterial virulence factors', *Nucleic acids research*, 50: D912-D17.

Liu, Shuo, Shu-Xuan Wang, Wei Liu, Chen Wang, Fa-Zhan Zhang, Yuan-Nong Ye, Candy-S Wu, Wen-Xin Zheng, Nini Rao, and Feng-Biao Guo. 2020. 'CEG 2.0:

an updated database of clusters of essential genes including eukaryotic organisms', *Database*, 2020.

Livermore, David M. 1987. 'Mechanisms of resistance to cephalosporin antibiotics', *Drugs*, 34: 64-88.

Loy, Alexander, Carina Pfann, Michaela Steinberger, Buck Hanson, Simone Herp, Sandrine Brugiroux, Joao Carlos Gomes Neto, Mark V Boekschoten, Clarissa Schwab, and Tim Urich. 2017. 'Lifestyle and horizontal gene transfer-mediated evolution of Mucispirillum schaedleri, a core member of the murine gut microbiota', *Msystems*, 2: e00171-16.

Luan, Yunxia, Nan Wang, Cheng Li, Xiaojun Guo, and Anxiang Lu. 2020. 'Advances in the application of aptamer biosensors to the detection of aminoglycoside antibiotics', *Antibiotics*, 9: 787.

Luo, Hao, Feng Gao, and Yan Lin. 2015. 'Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes', *Scientific reports*, 5: 1-8.

Lynch, Michael, Matthew S Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W Kelley Thomas, and Patricia L Foster. 2016. 'Genetic drift, selection and the evolution of the mutation rate', *Nature Reviews Genetics*, 17: 704-14.

Martínez, José L, Teresa M Coque, and Fernando Baquero. 2015. 'What is a resistance gene? Ranking risk in resistomes', *Nature Reviews Microbiology*, 13: 116-23.

Mattoo, Seema, Yvonne M Lee, and Jack E Dixon. 2007. 'Interactions of bacterial effector proteins with host proteins', *Current opinion in immunology*, 19: 392-401.

McDonnell, Gerald, and A Denver Russell. 1999. 'Antiseptics and disinfectants: activity, action, and resistance', *Clinical microbiology reviews*, 12: 147-79.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, and Lorna J Richardson. 2021. 'Pfam: The protein families database in 2021', *Nucleic acids research*, 49: D412-D19.

Munita, Jose M, and Cesar A Arias. 2016. 'Mechanisms of antibiotic resistance',

Virulence mechanisms of bacterial pathogens: 481-511.

Nelson-Sathi, Shijulal, Tal Dagan, Giddy Landan, Arnold Janssen, Mike Steel, James O McInerney, Uwe Deppenmeier, and William F Martin. 2012. 'Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea', *Proceedings of the National Academy of Sciences*, 109: 20537-42.

Nikaido, Hiroshi. 2009. 'Multidrug resistance in bacteria', *Annual review of biochemistry*, 78: 119-46.

Novichkov, Pavel S, Marina V Omelchenko, Mikhail S Gelfand, Andrei A Mironov, Yuri I Wolf, and Eugene V Koonin. 2004. 'Genome-wide molecular clock and horizontal gene transfer in bacterial evolution', *Journal of bacteriology*, 186: 6575-85.

O'Leary, Nuala A, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, and Danso Ako-Adjei. 2016. 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, 44: D733-D45.

Ochman, Howard, Jeffrey G Lawrence, and Eduardo A Groisman. 2000. 'Lateral gene transfer and the nature of bacterial innovation', *nature*, 405: 299-304.

Pendleton, Jack N, Sean P Gorman, and Brendan F Gilmore. 2013. 'Clinical relevance of the ESKAPE pathogens', *Expert review of anti-infective therapy*, 11: 297-308.

Podell, Sheila, Terry Gaasterland, and Eric E Allen. 2008. 'A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm', *BMC bioinformatics*, 9: 1-12.

Popa, Ovidiu, and Tal Dagan. 2011. 'Trends and barriers to lateral gene transfer in prokaryotes', *Current opinion in microbiology*, 14: 615-23.

Potter, Simon C, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, and Robert D Finn. 2018. 'HMMER web server: 2018 update', *Nucleic acids research*, 46: W200-W04.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2010. 'FastTree 2–approximately maximum-likelihood trees for large alignments', *PLoS One*, 5: e9490.

Ragan, Mark A. 2001. 'On surrogate methods for detecting lateral gene transfer', *FEMS Microbiology letters*, 201: 187-91.

Rohner, N. 2016. 'Genome Evolution's Role in Developmental Evolution'.

Sánchez-Soto, Daniela, Guillermin Agüero-Chapin, Vinicio Armijos-Jaramillo, Yunierkis Perez-Castillo, Eduardo Tejera, Agostinho Antunes, and Aminael Sánchez-Rodríguez. 2020. 'ShadowCaster: compositional methods under the shadow of phylogenetic models to detect horizontal gene transfers in prokaryotes', *Genes*, 11: 756.

Sayers, Eric W, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, and William Klimke. 2021. 'Database resources of the national center for biotechnology information', *Nucleic Acids Research*, 49: D10.

Sayers, Eric W, Evan E Bolton, J Rodney Brister, Kathi Canese, Jessica Chan, Donald C Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, Tom Madej, Aron Marchler-Bauer, Christopher Lanczycki, Stacy Lathrop, Zhiyong Lu, Francoise Thibaud-Nissen, Terence Murphy, Lon Phan, Yuri Skripchenko, Tony Tse, Jiyao Wang, Rebecca Williams, Barton W Trawick, Kim D Pruitt, and Stephen T Sherry. 2021. 'Database resources of the national center for biotechnology information', *Nucleic acids research*, 50: D20-D26.

Schatz, Albert, Elizabeth Bugie, Selman A Waksman, Arlen D Hanssen, Robin Patel, and Douglas R Osmon. 2005. 'The classic: streptomycin, a substance exhibiting antibiotic activity against Gram-positive and Gram-negative bacteria', *Clinical Orthopaedics and Related Research®*, 437: 3-6.

Seemann, Torsten. 2013. 'barrnap 0.9: rapid ribosomal RNA prediction', *Google Scholar*.

———. 2014. 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics*, 30: 2068-69.

Sevillya, Gur, Orit Adato, and Sagi Snir. 2020. 'Detecting horizontal gene transfer: a probabilistic approach', *BMC genomics*, 21: 1-11.

Sharma, Aditya Kumar, Neha Dhasmana, Neha Dubey, Nishant Kumar, Aakriti Gangwal, Meetu Gupta, and Yogendra Singh. 2017. 'Bacterial virulence factors: secreted for survival', *Indian journal of microbiology*, 57: 1-10.

Shikov, Anton E, Yury V Malovichko, Anton A Nizhnikov, and Kirill S Antonets. 2022. 'Current Methods for Recombination Detection in Bacteria', *International Journal of Molecular Sciences*, 23: 6257.

Sievers, Fabian, and Desmond G Higgins. 2014. 'Clustal Omega, accurate alignment of very large numbers of sequences', *Multiple sequence alignment methods*: 105-16.

———. 2018. 'Clustal Omega for making accurate alignments of many protein sequences', *Protein Science*, 27: 135-45.

Smits, Samuel A, and Cleber C Ouverney. 2010. 'jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web', *PloS one*, 5: e12267.

Soucy, Shannon M, Jinling Huang, and Johann Peter Gogarten. 2015. 'Horizontal gene transfer: building the web of life', *Nature Reviews Genetics*, 16: 472-82.

Stanton, Thaddeus B. 2013. 'A call for antibiotic alternatives research', *Trends in microbiology*, 21: 111-13.

Stoodley, Paul, Karin Sauer, David Gwilym Davies, and J William Costerton. 2002. 'Biofilms as complex differentiated communities', *Annual Reviews in Microbiology*, 56: 187-209.

Tabari, Ehsan, and Zhengchang Su. 2017a. 'PorthoMCL: parallel orthology prediction using MCL for the realm of massive genome availability', *Big Data Analytics*, 2: 1-5.

———. 2017b. 'PorthoMCL: parallel orthology prediction using MCL for the realm of massive genome availability', *Big Data Analytics*, 2: 1-5.

Thompson, RL, and AJ Wright. 1983. "Cephalosporin antibiotics." In *Mayo Clinic Proceedings*, 79-87.

Van, Thi Thu Hao, Zuwera Yidana, Peter M Smooker, and Peter J Coloe. 2020. 'Antibiotic use in food animals worldwide, with a focus on Africa: Pluses and minuses', *Journal of global antimicrobial resistance*, 20: 170-77.

Vogan, Aaron A, and Paul G Higgs. 2011. 'The advantages and disadvantages of horizontal gene transfer and the emergence of the first species', *Biology direct*, 6: 1-14.

Von Wintersdorff, Christian JH, John Penders, Julius M Van Niekerk, Nathan D Mills, Snehali Majumder, Lieke B Van Alphen, Paul HM Savelkoul, and Petra FG Wolffs. 2016. 'Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer', *Frontiers in Microbiology*, 7: 173.

Ward, David M, Michael J Ferris, Stephen C Nold, and Mary M Bateson. 1998. 'A natural view of microbial biodiversity within hot spring cyanobacterial mat communities', *Microbiology and Molecular Biology Reviews*, 62: 1353-70.

Weiss, Robin A. 2002. 'Virulence and pathogenesis', *TRENDS in Microbiology*, 10: 314-17.

Woese, Carl R. 1987. 'Bacterial evolution', *Microbiological reviews*, 51: 221-71.

Wu, Hsing-Ju, Andrew HJ Wang, and Michael P Jennings. 2008. 'Discovery of virulence factors of pathogenic bacteria', *Current opinion in chemical biology*, 12: 93-101.

Ye, Yuan-Nong, Zhi-Gang Hua, Jian Huang, Nini Rao, and Feng-Biao Guo. 2013. 'CEG: a database of essential gene clusters', *BMC genomics*, 14: 1-10.

Yushchuk, Oleksandr, Elisa Binda, and Flavia Marinelli. 2020. 'Glycopeptide antibiotic resistance genes: distribution and function in the producer actinomycetes', *Frontiers in Microbiology*, 11: 1173.

Zhao, Jianhua, Ksenia Beyrakhova, Yao Liu, Claudia P Alvarez, Stephanie A Bueler, Li Xu, Caishuang Xu, Michal T Boniecki, Voula Kanelis, and Zhao-Qing Luo. 2017. 'Molecular basis for the binding and modulation of V-ATPase by a bacterial effector protein', *PLoS pathogens*, 13: e1006394.

Zhaxybayeva, Olga, Kristen S Swithers, Pascal Lapierre, Gregory P Fournier, Derek

M Bickhart, Robert T DeBoy, Karen E Nelson, Camilla L Nesbø, W Ford Doolittle, and J Peter Gogarten. 2009. 'On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales', *Proceedings of the National Academy of Sciences*, 106: 5865-70.

국문 초록

계통수 조화법을 통한 수평 유전자 이동 현상 규명과 병원성 인자 및 항생제 내성 유전자의 분류군 간 이동 경향

최 영 석
농생명공학부 바이오모듈레이션 전공
서울대학교 대학원 농업생명과학대학

수평 유전자 이동 (Horizontal gene transfer, HGT)은 유전자가 서로 다른 생물체에게 전달되는 현상으로 미생물 진화의 근본적인 과정이다. 이는 미생물의 유전적 다양성과 적응 능력을 형성하는 데 중요한 역할을 한다. 생물정보학 기술의 발전으로 HGT의 연구가 점점 더 효율적으로 가능 해졌고 이를 통해 미생물 유전자 교환에 기여하는 복잡한 여러 패턴과 기작을 밝혀낼 수 있다.

이 논문은 HGT 현상과 그것의 의미를 파헤치는 데 초점을 맞춘 두

가지 다른 논문으로 구성되어 있다. 두 논문 모두 HGT 현상의 발생과 특성을 탐구하려는 목표를 가지고 있지만, 각각 강조하는 것은 다르다. 첫 번째 논문은 유전자 계통수 (gene tree)와 종 계통수 (species tree) 사용하여 HGT 현상을 규명하는 기존보다 개선된 계통수 조화 방법 (tree-reconciliation method)과, 이를 통해 개발된 새로운 버전의 데이터베이스를 소개한다. 두 번째 논문은 병원성 인자 (virulence factor, VFs)와 항생제 내성 인자 (antibiotics resistance genes, ARGs)와 관련된 분류군 간 (inter-taxon) HGT 현상에서 관찰된 동향을 조사하여 그것을 주도하는 기작에 대한 통찰 및 설명을 그 내용으로 한다.

이 논문의 첫 번째 챕터는 HGT, 계통수 조화 방법, 병원성 인자와 항생제 내성 인자라는 네 가지 주제에 대한 포괄적인 문헌 조사로 이루어져 있다. 이러한 문헌 조사로 본 논문의 내용을 간결하게 전달하고 미생물 진화와 유전 물질의 이동에 중요한 역할을 하는 HGT의 근본적인 개념을 탐구했다.

두 번째 챕터는 유전자 계통수와 종 계통수를 활용하여 HGT 현상을 탐지하는 계통수 조화 방법에 그 초점을 맞추고 있다. 본 연구에서는 2015년에 구축된 이전 버전의 HGTree 데이터베이스의 개선된 버전인 HGTree v2.0 데이터베이스를 소개한다. 본 업데이트는 2015년 이후 급증한 원핵 생물의 유전체 데이터에 대응하기 위함이 그 목적이었으며, 5,405,040개의 HGT관련 유전자를 포함한 더욱 많은 양의 데이터셋을 포함한다. 또한, HGTree v2.0은 본 업데이트를 통해 사용자 인터페이스와 데이터 검증 절차도 크게 개선되었으며 사용자가 자신의 유전체 속 HGT 관련 유전자를 직접 조사할 수 있는 User-query Processor 또한 포함하고 있다. 개선된 현 데이터베이스를 활용하여 연구자들은 보다 정확하게 HGT 현상을 식별하고 분석할 수 있으며, 미생물 유전체의 진화에 대한 HGT의 역학에 대한 통찰 또한 가능할 것이다.

세 번째 챕터에서는 VFs와 ARGs를 포함한 분류군 간 HGT 현상에서 관찰된 동향을 조사했다. 본 연구는 분류군 간 HGT 현상에서 ARGs의 이동이 VFs 보다 높은 것을 확인했으며, Cluster of Orthologous Genes (COG) 카테고리 분석을 통해 박테리아 생존과 관련된 필수적 COG 카테고리가 분류군 간 HGT 현상에서 더 자주 연관되어 있음 또한 발견했다. 또한, 본 연구에서는 VFs와 ARGs들의 여러가지 다른 기능적 측면을 조사해 분류군 간 HGT 현상에 관련된 VFs와 ARGs들의 포괄적인 통찰을 제공한다. 본 조사는 추가로 6가지의 다제내성 (MDR) 후보 박테리아 종들의 분류군 간 HGT 현상에서의 VFs와 ARGs의 세부적 동향을 분석해 MDR 후보군들의 ARGs가 전체적인 동향 보다 더 활발히 분류군 간 HGT 현상에 관여함을 알아냈다.

위의 연구들을 통해 HGT 현상이 가지는 여러가지 측면에 대하여 이해할 수 있었다. 본 연구들을 통해 미생물의 유전적 기능, 종 분류, 나아가 미생물의 진화에 대한 전반적인 연구에 활용될 수 있다.

주요어: 수평 유전자 이동, 계통수 조화 방법, 병원성/항생제 내성인자

학번: 2021-21311