Identifying and characterizing bacterial pathogen-associated genes for antivirulence drug development

by

Wing Yin Venus Lau

BSc, Simon Fraser University, 2015

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2022

Declaration of Committee

Name:	**Wing Yin Venus Lau**
Degree:	**Doctor of Philosophy**
Title:	**Identifying and characterizing bacterial pathogen-associated genes for antivirulence drug development**
Committee:	**Chair: Mani Larijani** Associate Professor, Molecular Biology and Biochemistry
	Fiona Brinkman Supervisor Professor, Molecular Biology and Biochemistry
	Amy Lee Committee Member Assistant Professor, Molecular Biology and Biochemistry
	Margo Moore Committee Member Professor Emerita, Biological Sciences
	Steven Jones Committee Member Adjunct Professor, Molecular Biology and Biochemistry
	Jack Nansheng Chen Examiner Professor, Molecular Biology and Biochemistry
	Lori Burrows External Examiner Professor, Biochemistry and Biomedical Sciences McMaster University

Abstract

Development of novel therapeutic strategies is urgently required to counter the growing public health threat of antimicrobial resistance (AMR). While antibiotics/antimicrobials target essential processes, antivirulence is an emerging concept aiming to disrupt virulence, potentially reducing the selective pressure for AMR development. Pathogen-associated genes (PAGs), whose conservation only in bacterial pathogens indicates potential role in virulence, may be suitable candidates for antivirulence drug targets. In this thesis, I developed a computational workflow, with select experimental validation, for 1) characterizing and prioritizing PAGs as potential antivirulence drug targets, 2) analyzing the structure-activity relationship of the approved drug raloxifene as an antivirulence agent, and 3) assessing the application of PAGs towards metagenomics-based pathogen surveillance. First, I refined a previously developed PAG prediction algorithm to generate an updated PAG dataset. In addition, a genus-specific PAG analysis was performed, focusing on the antibiotic-resistant, priority pathogen *Pseudomonas aeruginosa* and the related *Pseudomonas* species. Seventeen PAGs identified from *P. aeruginosa* PA14 were subsequently analyzed *in silico* (predicting gene mobility, evolutionary selection, and protein subcellular localization using a newly expanded database), and *in vivo* (virulence assay), identifying novel targets, including two within a genomic island found in several multi-drug resistant *P. aeruginosa* isolates. Expanding a previous drug-repurposing study of raloxifene as an anti-pseudomonal antivirulence agent, I compared the properties of raloxifene analogs in a quantitative pyocyanin assay, a growth curve assay and a standardized *C. elegans* infection model, which revealed that at least one hydroxyl group of raloxifene likely contributed to its antivirulence activity. Lastly, to evaluate the potential application of PAGs in pathogen surveillance, I examined the prevalence of PAGs in lung microbiomes and watershed microbiomes. Cystic fibrosis lungs, relative to healthy lungs, were disproportionately enriched in PAGs, but interpretation of watershed analyses was affected by the abundance of uncharacterized bacterial species in freshwater. The analyses suggested that the application of PAGs in pathogen surveillance warrants further study but may be limited to well characterized microbial environments. Collectively, this work expands knowledge of PAGs and an associated proposed antivirulence drug, revealing new avenues for PAG research in surveillance and expanding potential drug targets for antivirulence therapies.

Keywords: pathogen-associated genes; bacterial virulence factors; antivirulence therapeutics; infectious disease; bioinformatics; microbial genomics

To my mother

Who has supported me through the whole journey

With unconditional love and patience

To my maternal grandfather 公公

Who has been my biggest cheerleader and role model

Who has always celebrated every step of my growth

Acknowledgements

The completion of this PhD thesis is done with tremendous support from many people whom I met and have stayed along my side during this journey. First of all, I would like to express my gratitude to my exceptional senior supervisor and mentor, Dr. Fiona Brinkman. She has provided me with a supportive environment and mentorship to transition from a wet-lab biologist to a bioinformatics scientist. I'd also like to thank my supervisory committee members, Dr. Margo Moore, Dr. Steven Jones and Dr. Amy Lee for their continuous support and constructive feedback on my thesis work as I navigated through my research.

I am grateful to have spent a few years with the wonderful Brinkman Lab members that have helped me on my projects in various ways. A big thank you to Geoff Winsor who guided me on my first large bioinformatics analysis; Gemma Hoad and Vivian Jin who I worked on the POSRT tools with; lab managers Raymond Lo and Dr. Hanadi Ibrahim who helped me with the raloxifene drug testing project; Dr. Patrick Taylor who I've worked closely on the pathogen-associated genes characterization; the Brinkman grads, Kristen Gray, Justin Jia and Emma Garlock who I have shared this grad school journey with since day one; and the rest of the lab members, Nolan Woods, Dr. Erin Gill and Justin Cook who have always given me insightful feedback during lab meetings and presentations.

Throughout the past few years, I've also had the opportunity to meet, work and build friendship with numerous graduate students within and beyond the department of Molecular Biology and Biochemistry. It was truly an unforgettable experience working with so many talented individuals in the MBB Graduate Caucus and SFUOmics Group, from organizing departmental socials, annual MBB Colloquiums to SFUOmics Research Days.

Of course, this achievement would not be possible without the endless support and love from my family. Thank you to the strongest and bravest woman in my life, my mom Jovy, for raising not just one, but three (sometimes difficult) daughters and always making sure we are well taken care of and supported in everything we do. To my maternal grandparents, 公公 and 婆婆, thank you for always being with me, celebrating every little achievements I made and teaching me how to be humble and kind. To my

sisters, Asa Pui Yin and Rica Ching Yin, you two have done an incredible job at keeping me going on this journey. Thank you for helping me out with anything I need and being my 24/7 emotional supporters. Last but not least, to my little husky, Luna Bear, you have given me so much joy and adrenaline rush to keep me on my toes all the time. Thank you for always reminding me to take breaks during a busy day and to enjoy the little moments of happiness together.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

aBSREL	adaptive Branch-Site Random Effects Likelihood
AMR	Antimicrobial resistance
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BUSTED	Branch-Site Unrestricted Statistical Test for Episodic Diversification
CF	Cystic fibrosis
COG	Cluster of Orthologous Groups
COPD	Chronic obstructive pulmonary disease
DIAMOND	Double Index Alignment of Next-generation sequencing Data
DUF	Domain of unknown function
E-value	Expected value
FEL	Fixed Effects Likelihood
GARD	Genetic Algorithm for Recombination Detection
GI	Genomic island
HyPhy	Hypothesis Testing using Phylogenies
MEME	Mixed Effects Model of Evolution
NCBI	National Center for Biotechnology Information
NGM	Nematode growth medium
PAG	Pathogen-associated gene
RefSeq	NCI Reference Sequence
rRNA	Ribosomal ribonucleic acid
SCL	Subcellular localization
T3SS	Type III secretion system
Tn	Transposon
VF	Virulence factor
WHO	World Health Organization

Glossary

Common gene	Gene predicted to be conserved in both pathogenic and non-pathogenic bacteria based on the PAGs analysis developed by the Brinkman Lab
Deduced protein	A protein whose amino acid sequence is predicted from the corresponding nucleotide sequence
E-value	Expected value; a parameter that describes the number of BLAST hits expected to see by chance in a database of a particular size
Episodic positive selection	Positive selection detected across all branches within a gene phylogeny
Non-pathogen-associated gene	Gene predicted to be conserved in only non-pathogenic bacteria based on the PAGs analysis developed by the Brinkman Lab
Orthogroup	A group of homologous genes, predicted by OrthoFinder
PA14	P. aeruginosa PA14 strain
Pathogen-associated gene	Gene predicted to be conserved in only pathogenic bacteria based on the PAGs analysis developed by the Brinkman Lab
Pervasive positive selection	Positive selection that occurs on some but not all branches within a gene phylogeny

Chapter 1.

Introduction

1.1. The global antimicrobial resistance (AMR) crisis

The discovery of penicillin by Alexander Fleming in 1928 (Fleming, 2001) gave rise to the golden era of antibiotic discovery between 1940s and 1960s, during which most modern antibiotic classes were discovered in natural compounds produced by microorganisms, mainly soil actinomycetes (Valiquette & Laupland, 2015). Although this breakthrough has revolutionized modern medicine in the treatment and prevention of bacterial infections, the surge of antibiotic use was immediately followed by the clinical emergence of resistance to nearly all antibiotics within one or two decades after development (Ventola, 2015). AMR exists in nature before the antibiotic era. It is a protective mechanism deployed by bacteria to endure the selective pressure for survival imposed by their diverse environments including intra-population competition with other microorganisms or exposure to antibiotics in clinical medicine. For example, metallo-β-lactamase resistance dates back to over 2 billion years ago based on phylogenetic evidence, illustrating the ancient history of AMR (Aminov, 2010). In recent decades, the the emergence and the growing spread of multidrug resistant bacterial pathogens at an accelerated rate is mainly attributed to the frequent and indiscriminate antibiotic use in clinical, agricultural and veterinary settings.

Effective antibiotics are required not only for controlling infectious disease transmission but also to prevent healthcare associated infections pertinent to common surgical procedures or high-risk patient groups, such as those with a compromised immune system. Failure to treat bacterial infections leads to prolonged illness and hospital stays, imposing a significant healthcare and economic burden to societies. The World Health Organization (WHO) published a list of highly antibiotic resistant, priority pathogens in urgent need of novel therapeutics in 2017 and also declared AMR as one of the top 10 global public health threats in 2019 (World Health Organization, 2017b). While the WHO previously estimated 700,000 AMR-related deaths annually, a more recent and comprehensive assessment of the global burden of AMR predicted 4.95 million AMR-associated deaths in 2019 (Antimicrobial Resistance, 2022; World Health

Organization, 2019a). The WHO also projected that AMR-related diseases will cause 10 million annual deaths by 2050 if the current AMR trajectory is not curbed with effective antimicrobial stewardship programs and novel therapeutics for bacterial infections (World Health Organization, 2019b).

The global AMR concern is further challenged by the global shortage of innovative antibacterial treatments that can minimize the risk of AMR development in bacterial. The continual rise in AMR pathogens is outpacing the clinical development of novel antibiotics, of which 82% are derivatives of known drug classes that are already associated with AMR (World Health Organization, 2021). Only 27 non-traditional antibacterial agents are under clinical development as of 2020 (Theuretzbacher et al., 2020). The public health urgency for novel and effective bacterial infection treatment strategies engendered the WHO's innovation criteria for novel antibacterial drugs, which must 1) be absent of cross-resistance to existing antibiotics, 2) have a new bacterial target, 3) confer a new mode of action and 4) belong to a novel drug class. These criteria help to identify truly innovative drugs with potentially more sustained effectiveness and reduced risk for rapid AMR development.

1.2. Alternative therapeutics for bacterial infections

1.2.1. Non-antibiotic strategies

The AMR challenge faced by current available antibiotics prompted the development of innovative strategies for bacterial infection management, such as restoring antibiotic sensitivity with potentiators, preventing antibiotic-induced dysbiosis and minimizing the risk of drug resistance development. There is a growing interest in applying the concept of precision medicine, commonly used in oncology, to infectious diseases. Narrowing the spectrum of antibacterial agent to precise targets associated with the disease-causing bacteria may reduce the disruption to the gut microbiota and the global selective pressure that drives AMR in all susceptible, pathogen and non-pathogenic, bacteria (Paharik et al., 2017; Spaulding et al., 2018).

With more attention being brought to the importance of microbiome in human health and diseases, microbiome-modulating agents are a holistic treatment approach that restore the healthy balance of a disrupted gut microbial community caused by

diseases or prolonged antibiotic use. Antibiotic exposure creates a competitive microbial environment for survival and has been shown to increase the prevalence of AMR genes within host-associated microbial communities. (Pilmis et al., 2020). A classic example of microbiome-modulating treatment is fecal microbiota transplantation, a highly effective *Clostridioides difficile* treatment in which the gut microbiota of a healthy donor is transferred to the colon of a patient. Other gut microbiome modulators include prebiotics and probiotics that can be consumed to enhance the health of gut microbiome. Prebiotics are food components (e.g. human milk oligosaccharide in breast milk) that promote the proliferation of commensal gut bacteria, while probiotics are living commensal bacteria (e.g. *Lactobacillus spp*.) derived from fermented food or cultured milk.

Another therapeutic strategy against bacterial infections is to improve pathogen clearance by infected host through immunomodulation. Immunomodulators can stimulate host immune response to potentiate pathogen killing by bactericidal drugs or immune-mediated pathogen clearance by bacteriostatic drugs. Instead of direct pathogen targeting, immunomodulators increase or rescue the efficacy of existing antibiotics to which AMR mechanisms may have already been developed. For example, the antioxidant N-acetylcysteine augments the clearance of *Mycobacterium tuberculosis (Mtb)* by reducing host production of reactive oxidative species, as a result of host inflammatory response, that drive the formation of drug-tolerant *Mtb* persister cells (Beam et al., 2021).

Phage therapy is also an emerging approach to target bacterial pathogens by employing the natural and highly specific interaction between bacteriophages (viruses that infect bacteria) and their bacterial hosts. The specificity of bacterial host receptors enables the use of lytic (virulent) phages to target, infect and lyse the bacterial pathogen of interest. Common phages associated with clinically important human pathogens belong to the viral orders *Caudovirales* (double-stranded DNA tailed phages) and *Microviridae* (single-stranded DNA tailless phages) (Lin et al., 2017). Use of phage-derived lytic proteins is a similar strategy to eliminate bacterial pathogens without the use of intact bacteriophages (Sao-Jose, 2018). Trade-off between phage resistance and antibiotic sensitivity in bacteria has been experimentally observed, warranting phage or phage-derived therapeutics as promising solution to restoring antibiotic sensitivity in resistant bacteria. This phenomenon is hypothesized to be an antagonistic pleiotropic

effect in which the bacterial component to which phage interacts is also responsible for AMR. Bacterial outer membrane proteins, often recognized by phages, are known to confer drug resistance so any mutations selecting for phage resistance may conversely restore drug sensitivity. For example, phage U136B infects *Escherichia coli* through the recognition of the antibiotic efflux protein TolC, so any mutation of this protein would simultaneously enable phage resistance and antimicrobial susceptibility due the loss of efflux function (Burmeister et al., 2020; Gurney et al., 2020).

Antivirulence, which targets bacterial virulence instead of essential biological processes, is the antibiotic alternative upon which this thesis work is based. More details are in the next section (Chapter 1.2.2).

1.2.2. Antivirulence

Antivirulence drugs control bacterial infections by attenuating bacterial virulence responsible for host damage without significant impact on bacterial survival. Contrary to antibiotics which target essential bacterial functions and impose strong selective pressure for drug resistance, antivirulence drugs aim to disarm pathogens and allow the natural pathogen clearance in immunocompetent hosts. As virulence factor production is often associated with a metabolic cost, targeting virulence factors with minimal evolutionary benefit to the pathogen at the site of infection likely reduces the risk of drug resistance development (Figure 1.1). By targeting virulence factors (VFs) unique to pathogens, antivirulence drugs may narrow the spectrum of activity to only organisms of concern while minimizing the drug perturbation of commensal bacteria at the site of drug action. Antivirulence drug targets are often involved in disease-causing processes such as toxin production and secretion, host adhesion/colonization and virulence regulation (e.g., quorum sensing). The monoclonal antibody Raxibacumab is one of the earliest Food and Drug Administration (FDA)- approved antivirulence drugs targeting the protective antigen component of the *Bacillus anthracis* anthrax toxin (Corey et al., 2013; Skoura et al., 2020). Many antivirulence agents have since been under clinical development. These include Shigamab against shiga toxin in *Escherichia coli* (Dickey et al., 2017), the repurposing antifungal drugs clotrimazole and miconazole against the quorum-sensing regulator PqsR in *Pseudomonas aeruginosa* (D'Angelo et al., 2018) and the natural product Baicalin (a Chinese herbal medicine) against the cell-surface sortase B in *Staphylococcus aureus* (Wang et al., 2018)

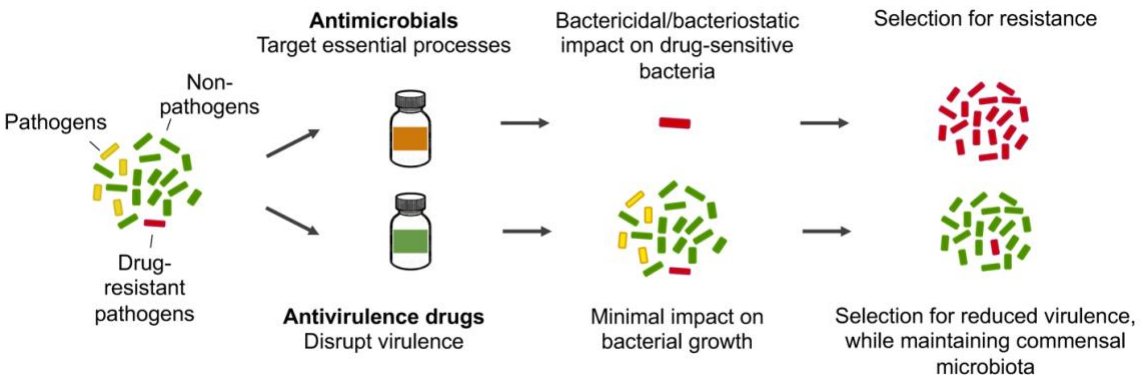


Figure 1.1 Comparison of antibiotics and antivirulence drugs.
The site of infection within a host is often populated by a combination of drug-sensitive pathogens, drug-sensitive non-pathogens (commensals) and drug-resistant pathogens. While conventional antimicrobials/antibiotics target essential bacterial processes and kill any susceptible pathogens and non-pathogens, antivirulence therapeutics targets only the bacterial pathogen-specific components (VFs) involved in host pathogenesis, while minimizing disruption to the host microbiota and potentially reducing the evolutionary selection for drug resistance.

While antivirulence drugs can potentially minimize the risk of AMR development, resistance towards antivirulence drugs, such as the anti-*P. aeruginosa* quorum-sensing inhibitor C-30 (resistance shown in *C. elegans* infection model), can still develop depending on factors such as drug target choice, bacterial population structure and site of action (Allen et al., 2014; Maeda et al., 2012). The cellular impact of the targeted VF must be evaluated with caution to ensure that an antivirulence drug truly satisfies the "disarm-don't kill" criterion without imposing any indirect consequence on the survival of the pathogen population. The development of resistance to antivirulence drugs has been speculated to be more complex than to antibiotics due to the range of impact of VFs on pathogen fitness in different ecological niches (Totsika, 2016). A review paper regarding the evolution of antivirulence drug resistance suggested that antivirulence drugs will likely select against resistance (i.e., favour drug-sensitive over drug-resistant bacteria) if the antivirulence drugs target VFs conferring no benefits but metabolic cost to the pathogen at the site of colonization or damage (Allen et al., 2014). Likely produced by opportunistic pathogens, these non-beneficial VFs likely confer alternative benefits in distinct, free-living environments but coincidentally have damaging effect on host cells while not necessarily required for host adaptation. For example, the adhesin P pili and the iron acquisition factor yersiniabactin, implicated in commensal niche colonization, are also associated with virulence in extraintestinal pathogenic *E. coli* during extraintestinal infections in the brain or urinary tract (Allen et al., 2014).

The development of effective antivirulence drugs that are less prone to resistance by bacteria therefore requires in-depth research on the evolutionary cost and benefit of the targeted VFs in different host environments. While data on the effectiveness of antivirulence drugs in immunocompetent versus immunocompromised hosts are currently unavailable, antivirulence drugs which rely on active host clearance of the pathogenic bacteria are speculated to be problematic for individuals with a compromised immune system (Hotinger et al., 2021). However, combination therapies of traditional antibiotics and antivirulence drugs may be useful in clearing bacterial infections in different hosts, as they compliment each other and can potentially overcome the limitations of an individual therapeutic agent.

Vaccines are a type of prophylactic antivirulence strategy, since many vaccine components are virulence factors (Bliven & Maurelli, 2012). However, this prophylactic approach necessitates the drug being given weeks before any exposure to the pathogen, to generate an immune response (Barie, 2000; Fleitas Martinez et al., 2019). The focus of this thesis is on antivirulence therapeutics, not prophylactics, to be used after exposure to the pathogen. This treatment approach may be most beneficial for targeting a specific bacterial pathogen(s) responsible for a diagnosed infection. Therefore, the high need for improved diagnostics is still important for this strategy, just as noted for antimicrobial resistance.

1.3. Bacterial virulence

1.3.1. Definition of pathogen, pathogenicity and virulence

Within the diverse realm of bacteria, many have the pathogenic potential to cause diseases in any host organisms. Published in 1890, Koch's postulates were a set of four criteria for establishing a causative relationship between a microbe (pathogen) and a disease (Figure 1.2): 1) the microorganism must be found in all diseased individuals and absent in healthy individuals, 2) the microorganism must be isolated from the diseased organism and grown in culture, 3) inoculation of the isolated microorganism into a healthy individual must inflict the same disease and 4) the microorganism must be reisolated from the inoculated host and confirmed to be the same as the original causative agent (Segre, 2013). Although historically used as a gold standard for establishing microbial etiology of infectious diseases, Koch's postulates were soon found

to be inadequately applicable to the diverse range of bacterial pathogens, including those that are unculturable or cause asymptomatic infections. In recent years, there has been a shift from a pathogen-centric view of pathogenesis a more encompassing definition which emphasizes on the interplay between a pathogenic bacterium, its host and the environment. A damage-response framework was more recently introduced to describe microbial pathogenesis as an integrated outcome of the interaction between a host and a microbe (Casadevall & Pirofski, 2003). Under this framework, pathogenicity is defined as an organism's ability to inflict physiological damage to a host while a pathogen is defined as a microorganism with such functional potential. A pathogen is thus a function of not only its inherent virulent capability but also host susceptibility, as evident in opportunistic pathogens which are normally harmless but can cause infections under specific host conditions such as weakened immune system or dysbiosis (Shapiro-Ilan et al., 2005).

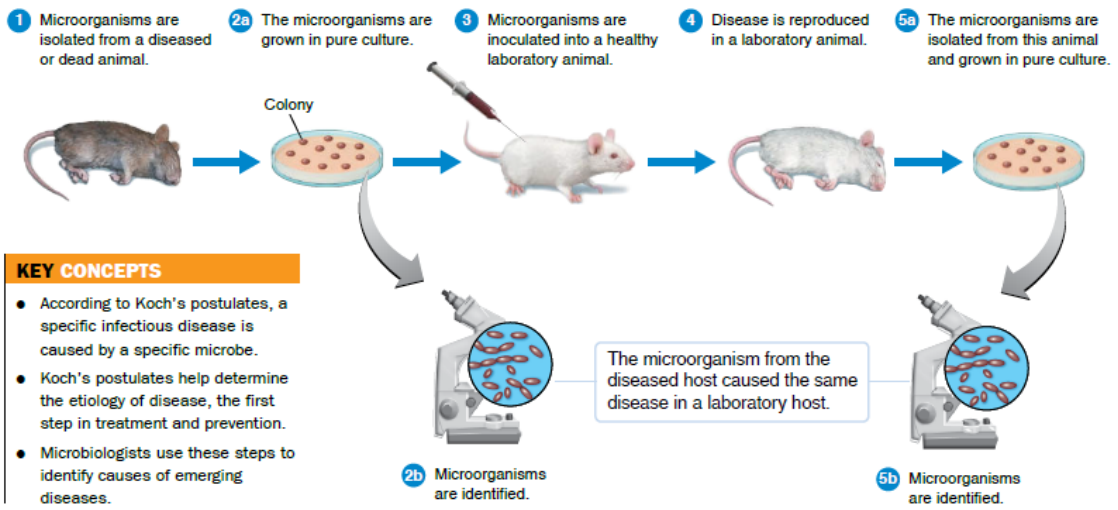


Figure 1.2 Koch's postulates for establishing a causative relationship between a microbe and a disease.
A set of four criteria to determine if a microbe causes an observed disease: 1) the microorganism must be found in all diseased individuals and absent in healthy individuals, 2) the microorganism must be isolated from the diseased organism and grown in culture, 3) inoculation of the isolated microorganism into a healthy individual must inflict the same disease and 4) the microorganism must be reisolated from the inoculated host and confirmed to be the same as the original causative agent. Figure from *Microbiology: An Introduction* (13th ed.) (Tortora, 2018).

While pathogenicity is the quality of being pathogenic, virulence is a measure of the disease-causing power (i.e., pathogenicity) of an organism and is variable based on the specific host-pathogen interaction, genetic property of a bacterial strain and the environmental condition at the infection site (Shapiro-Ilan et al., 2005). Virulence can be

assessed by the disease severity towards host organisms with measures like the lethal dose required to kill 50% of infected host (LD50) (Casadevall, 2017). Bacterial virulence is attributed to VFs, molecules produced by bacterial pathogens to directly cause diseases or trigger host-mediated pathogenesis in which the host is damaged by its own inflammatory responses. While some VFs may be widely conserved or species-specific, VFs are crucial for the process of bacterial infection and host adaptation (Peterson, 1996).

The molecular Koch's postulates were subsequently introduced in 1988 by Stanley Falkow to determine if a gene in a pathogenic organism encodes a disease-causing protein (i.e., VF) using the following criteria: 1) the gene of interest is found in all pathogenic strains and absent in all non-pathogenic strains of a genus/species, 2) the inactivation of the putative virulence gene should result in a loss of pathogenicity of the microorganism, and 3) the reintroduction of the gene should restore virulence (Falkow, 1988). More recently, Fredricks and Relman proposed an updated set of postulates for nucleic acid sequence-based detection of pathogens/VFs to incorporate the increasing use of microbial genomic data for infectious disease research (Fredricks & Relman, 1996). Regardless of the existing variations of the Koch's postulates, establishing disease causality is only the initial step in the discovery of novel VFs, whose functions are to be subsequently validated with extensive computational and experimental phenotypic studies.

Common VFs include microbial adherence and invasion factors (pili, fimbriae, adhesins), capsules (protection against phagocytosis and desiccation), exotoxins (secreted toxins such as cytotoxins, neurotoxins and enterotoxins) and endotoxins (cell-wall components like lipopolysaccharide that can induce bacteremia). Toxin-trafficking systems and membrane transporters also play a key role in host interaction by delivering effector molecules to the extracellular space or directly into host cells. The iron-binding siderophores are another class of VFs important for nutritional uptake by scavenging and chelating iron in competition with host cells. Recent attention has been brought to the complexity of bacterial pathogenesis that often involves other bacterial processes such as cell-to-cell communication, regulatory and metabolic pathways that are directly or indirectly implicated in virulence. Quorum sensing is a bacterial cell-cell communication strategy that regulates the expression of genes that are collectively beneficial to the bacterial population through responding to density-activated signalling

molecules (autoinducers). The contribution of quorum sensing to virulence has been documented in the bacterial pathogens *S. aureus, Bacillus cereus, Vibrio cholerae* and especially *P. aeruginosa* whose quorum sensing-induced VF production and biofilm formation have been well studied (Rutherford & Bassler, 2012). Similarly, metabolic genes are recently being explored for their functional roles in virulence and potential use as drug targets. For example, more than 16.5% of the *P. aeruginosa* PA14 core metabolic genes are involved in virulence regulation, particularly those related to beta-oxidation and biosynthesis of amino acids, succinate, citramalate and chorismate (Panayidou et al., 2020; Richardson, 2019).

Ultimately, host susceptibility and disease progression upon infection depend on both the virulence of the infecting pathogen and the immunologic response by the host. Within human hosts, neutrophils and macrophages act as the first line of non-specific, innate immune response in an attempt of pathogen clearance. The normal bacterial flora residing on or within the human host is another layer of host protection against the colonization and the uncontrolled proliferation of foreign or opportunistic pathogens within the host. Host factors such as age and health status may impact the effective host response to bacterial infections, with commonly infants, elderly or immunocompromised individuals being more susceptible to bacterial infections (Simon et al., 2015).

1.3.2. Evolution of bacterial pathogens

Bacterial pathogens often face a harsh and competitive host environment upon infection. They must overcome a multitude of physical challenges including nutrient acquisition, immune evasion, antibiotic killing and niche competition with other within-host organisms. There is a constant evolutionary arms race between the bacterial pathogen and its host, in the presence of multifactorial forces like mutations, genetic drift, transmission bottleneck and natural selection (Gatt & Margalit, 2021). Mutations beneficial for host adaptation and pathogen survival are often selected for and subsequently increase in frequency within the pathogen population. Whole genome sequencing of bacterial isolates can now reveal the underlying genetic factors responsible for host adaptive strategies, including resistance mechanisms against antibiotic exposure and virulence mechanisms for invasion and proliferation in host cells.

Hypermutations and rapid diversifying selection of genes in pathogens, especially those that are normally free-living, are imperative for their adaptation to a non-native niche (i.e., infection site within a host organism). The ability to disseminate beneficial genes among bacterial organisms through horizontal gene transfer further enables and accelerates the within-host diversification of pathogens. For example, *P. aeruginosa,* whose genome encodes many regulatory genes for environmental sensing, bacterial cell communication, virulence and resistance, readily adapts to different environments such as the human cystic fibrosis (CF) lungs during chronic infection (Gatt & Margalit, 2021; Hart & Winstanley, 2002; Moradali et al., 2017; Winstanley et al., 2016). Patterns of convergent evolution for AMR and niche specialization have also been observed among human pathogens (Fajardo-Lubian et al., 2019). Two aspects concerning AMR are the emergence of AMR genes which are ubiquitous by nature (in the natural environment or within host) and the dissemination of AMR genes to high frequency within the pathogen population under positive selection. Predicting the effectiveness of antibacterial drug candidates thus requires a deep understanding of not only the bacterial virulence but also the host-pathogen interaction, gene mobility and the within-host evolutionary dynamics that may increase the selective pressure for genes advantageous for host adaptation.

1.4. Identification of novel VFs

1.4.1. Existing tools and resources

Identifying novel VFs is necessary for the discovery of novel bacterial targets for antivirulence therapeutics. Genomics have revolutionized the drug discovery landscape by leveraging the wealth of publicly available genomic sequences and gene/protein functional data to enable rapid, high-throughput computational screening of appropriate drug targets. Initial development of antivirulence drugs focused on targeting VFs with direct involvement in host interaction. The Virulence Factor Database (VFDB) contains a curated set of VFs from clinically important pathogens (B. Liu et al., 2019). Similarly, Pathosystems Resource Integration Centre (PATRIC) is a comprehensive database of VFs as well as their functional annotation, protein-protein interaction, and transcriptomic data (Davis et al., 2020). MvirDB is also a centralized VF database that integrates VF and AMR sequences collected from multiple sources such as the VFDB, TVFac (toxin

and virulence database), Islander (genomic island predictor) and ARGO (AMR database) (C. E. Zhou et al., 2007). In addition to VF databases, novel VFs can now be predicted from genomic sequences using tools like VirulentPred (Garg & Gupta, 2008), VFanalyzer for bacterial draft/complete genomes (part of VFDB) (B. Liu et al., 2019) and PathoFact for metagenomics data (de Nies et al., 2021). Although existing VF resources have accelerated the progression of antivirulence drug research, up-to-date literature curation and experimental validation of novel VFs are essential for the continual expansion of current VF databases. Likewise, machine-learning based VF predictors rely heavily on curated VFs as training data; therefore, novel VFs that are highly divergent likely will be missed in VF predictions if they are not routinely incorporated into the updates of these bioinformatics tools.

1.4.2. Pathogen-associated genes (PAGs) as a novel avenue for VFs

Any bacterial gene involved in pathogenicity is defined as a VF regardless of its role in the virulence process (Wassenaar & Gaastra, 2001). Expanding the definition of VFs to not only virulence lifestyle genes (genes involved in host entry, colonization, immune evasion and proliferation upon infection) but also virulence-associated genes (genes involved in cellular metabolism, or the expression and the regulation of VFs) can help identify additional genes involved in virulence processes and can thus be used as therapeutic targets. Based on the PAGs analysis algorithm developed by the Brinkman Lab, PAGs are defined as genes predicted to be conserved only in bacterial pathogens through a comparative genome analysis of all curated bacterial pathogens and non-pathogens from the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database (S. J. Ho Sui, Fedynak, A., Hsiao, W.W.L, Langille, M.G.I & Brinkman, F.S.L., 2009). PAGs likely play a role in specialized functions important for the pathogenic lifestyle of bacteria under appropriate environmental conditions within a host. Development of pathogen-specific therapeutics to target specific proteins related to bacterial pathogens may reduce the likelihood of AMR development as often seen with broad-spectrum antibiotics that target all susceptible bacteria regardless of their pathogenicity (Simonson et al., 2021). For example, narrow-spectrum, pathogen-specific monoclonal antibodies have been shown to preserve intestinal microbiome while targeting specific disease-causing microorganisms in a mouse metagenomics study (Jones-Nelson et al., 2020)

The initial PAGs analysis in 2009 showed that PAGs are disproportionately enriched in “offensive functions” related to host invasion, Type III/IV secretion systems and toxins (S. J. Ho Sui et al., 2009). The potential application of PAGs towards antivirulence drug development was later demonstrated for an FDA-approved osteoporosis drug, raloxifene, with promising repurposing potential as an antivirulence agent. Raloxifene was computationally predicted, with preliminary experimental validation, to attenuate the virulence of *P. aeruginosa* by potentially interacting and interfering with the pathogen-associated PhzB2 protein involved in the biosynthesis of the major *P. aeruginosa* virulence factor, pyocyanin (i.e., exotoxin) (S. J. Ho Sui et al., 2012).

1.5. Functional characterization of PAGs

As more bacterial genomes are sequenced, many novel genes, including those that are pathogen-associated, have been predicted *in silico* by the presence of an open reading frame but are annotated as “hypothetical proteins” as their functions have yet to be determined. An estimated 30% of genes in bacteria do not have a well-defined function (Shahbaaz et al., 2016). To further highlight the lack of functional information for bacterial genes, a study by Antczak et al. showed that less than 1% of proteins in the minimal genome, the smallest set of genes required for survival, of *Mycoplasma mycoides* has been characterized (Antczak et al., 2019). Functional annotation of novel genes is often challenged by the high sequence divergence of these genes to genes with a defined function. A diverse combination of computational and laboratory efforts is thus necessary for the accurate functional annotation of uncharacterized genes, some of which may confer novel functions.

The PAGs analysis is a high-throughput and scalable genomics-based method for screening for putative pathogen-specific VFs from large bacterial genomic dataset. Different computational methods may help to infer the functional role of PAGs, especially those that are currently uncharacterized. A direct approach for functional annotation is the sequence similarity search against existing databases such as VFDB (VFs) (B. Liu et al., 2019), Pfam (protein families) (Mistry et al., 2021) and the NCBI Conserved Domain Database (protein domains) (M. Yang et al., 2020). Ortholog detection methods such as Ortholuge (Fulton et al., 2006), OrthoMCL (Fischer et al., 2011) and OrthoFinder (Emms & Kelly, 2019) can infer function of the genes of interest based on the annotation of their

orthologs which likely confer similar functions. PAGs whose sequences are too divergent to genes functionally annotated to date require alternative characterization strategies not dependent on sequence similarity.

Evolutionary selection inference of functionally unknown genes, in the context of pathogen evolution, may suggest important role in virulence-related processes. A common approach for detecting evolutionary selection in microbial genomes is the estimation of the dN/dS ratio (ω), which reflects the ratio of the number of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) (Kryazhimskiy & Plotkin, 2008). Synonymous substitutions are silent mutations that do not alter the corresponding amino acids in the protein. On the contrary, non-synonymous substitutions are nucleotide changes that alter the resultant amino acids. Synonymous site and non-synonymous site then refer to position of the nucleotide at which a mutation would result in a preservation or a change in the corresponding amino acid, respectively. A dN/dS ratio of one ($\omega=1$) implies a neutral selection under which the variations in a gene of interest are likely due to random mutations or genetic drift. A dN/dS ratio greater than one ($\omega>1$) suggests positive selection under which beneficial mutations are selected for. Lastly, a dN/dS ratio less than one ($\omega<1$) suggests purifying selection under which deleterious mutations are being purged from a population. dN and dS values form the basis of many widely used codon substitution models such as the Jukes and Cantor model (the first and simplest model assuming a constant rate of change among all nucleotides), K80 model (assumes rate of change differs between transitions and transversions) and the general time-reversible model (incorporates different rate for all nucleotide changes and frequencies) (Arenas, 2015) Other models also consider codon bias and physiochemical properties of the encoded amino acids.

Gene mobility may also provide insight into the functional nature of a gene. Horizontal gene transfer is a major contributor to bacterial pathogen evolution and versatility by enabling bacterial organisms to acquire genetic components important for niche adaptation and survival. Genomic islands (GIs) are clusters of bacterial or archaeal genes of probable horizontal origin (Bertelli et al., 2019; Langille et al., 2008). They can be transferred among bacterial cells to maintain gene flow and genome diversity within the bacterial population. There are many types of GIs with designated functions, including pathogenicity islands (containing virulence-related genes), metabolic

islands (containing metabolic genes), symbiotic islands (containing genes that facilitate the interaction with eukaryotic hosts) and resistance islands (containing AMR genes) (Dobrindt et al., 2004). However, some GIs may have a mosaic structure due to the physical association of gene clusters from different sources through independent horizontal transfer events (Jani & Azad, 2021). A high proportion of functionally uncharacterized genes resides in GIs, where VFs are also disproportionately enriched (S. J. Ho Sui et al., 2009; Hsiao et al., 2005). Assessing the mobility and co-localization of genes on GIs may provide some insight into the functions of unknown genes, given that genes conserved on the same non-mosaic GIs may be involved in similar cellular processes.

Likewise, other genomic contexts of a gene such as its genomic location and expression pattern in relation to other genes may also be useful in deducing its functional role within the organism. Bacterial genes are often arranged into operons, coregulated clusters of genes whose expression is upregulated or downregulated together (Assaf et al., 2021). Similarly, synteny, the physical co-localization of genes in the same order across taxa, may suggest functional context of an unknown gene. Colinear syntenic blocks that can span multiple operons and over long evolutionary distance across taxa may imply evolutionary selection against genomic rearrangement to maintain the complete functional unit of the syntenic region (Yelton et al., 2011). Physical interactions likely occur among proteins encoded by genes within a syntenic block, such as direct regulation of one another or interaction within a protein complex or cellular pathway. Functional association between an uncharacterized gene and a functionally annotated gene can thus be inferred by the co-evolution profile of their gene clusters based on conserved chromosomal proximity or co-occurrence in different taxa. (Sivashankari & Shanmughavel, 2006). In addition to genomic localization, the subcellular location where proteins (gene products) optimally operate within a bacterial cell may therefore provide insight into their functional roles. Typical gram-negative bacteria have 5 major subcellular localization (SCL) sites - cytoplasm, inner membrane, periplasm, outer membrane and extracellular space (secreted proteins), whereas typical gram-positive bacteria have 4 SCLs- cytoplasm, cytoplasmic membrane, cell wall and extracellular space (Yu et al., 2010).

Computational prediction of the properties of a gene requires validation by laboratory methods. To determine the virulence role of a bacterial gene, knockout

experiments are usually performed to study its loss-of-function impact in animal infection model organisms such as *Caenorhabditis elegans* (worm)*, Arabidopsis thaliana* (plant) and mouse (mammal). Transposon insertion mutagenesis is a biological process in which genes are transferred and integrated into specific or random sites within an organism's chromosome, therefore interrupting the function of the gene at the site of insertion. In the context of VF discovery and antivirulence therapeutic development, only non-essential genes, whose deletion does not impact bacterial viability, are considered for downstream virulence screening as ideal, druggable targets. (Cain et al., 2020). While transposon insertions are often random, more precise in-frame knockouts can be achieved using a 2-step allelic exchange by homologous recombination, a method has is optimized for generating clean knockouts of *P. aeruginosa* genes. The gene deletion is generated by cloning upstream and downstream sequence of a target sequence into a suicide vector that is subsequently inserted into bacterial chromosome via homologous recombination. Bacterial cells with the precise gene deletion of interest are then selected for (Hmelo et al., 2015).

1.6. *P. aeruginosa* - a WHO priority pathogen

P. aeruginosa is a gram-negative, aerobic, motile and non-spore-forming bacteria that commonly reside in soil and water but is also a major opportunistic pathogen affecting immunocompromised patients, and the leading cause of mortality in CF patients. Due to its intrinsic antimicrobial resistance, and the rise of multi drug resistance isolates, *P. aeruginosa* has been designated as one of top 3 WHO high priority pathogens that is in urgent need for novel therapeutics (Tacconelli et al., 2018; World Health Organization, 2017b). A 2017 WHO report estimated that multi-drug or pan-drug resistant *P. aeruginosa* infections attributed to 32,600 nosocomial infections and 2,700 deaths in United States alone (CDC, 2019). Ubiquitous in the environment, *P. aeruginosa* spreads not only through contaminated soil and water, but also by person-to-person transmission through contaminated hands, equipment (catheters and ventilators) and surfaces. This pathogen typically infects the human respiratory tract, urinary tract, burn wounds and blood (Bassetti et al., 2018). *P. aeruginosa* is equipped with an arsenal of intrinsic AMR mechanisms including the lack of non-specific porins for drug uptake (low outer membrane permeability), overexpressed efflux systems (Housseini et al., 2018) and the production of antibiotic-inactivating enzymes like the β-lactamase AmpC

(Moradali et al., 2017). The limited sensitivity to antibiotics is complicated by the acquisition of extrinsic AMR mechanisms via horizontal gene transfer. Particularly in chronic infection of *P. aeruginosa*, antibiotic failure is often caused by biofilm formation characterized by an overproduction of extracellular polysaccharide (alginate, Pel and Psl), leading to the formation of persister cells that are dormant and insensitive to antibiotics (Colvin et al., 2012).

1.6.1. Current drug therapies for *P. aeruginosa* infections

When possible, a culture is taken to identify the *P. aeruginosa* resistance profile and determine the best antimicrobial treatment that is also appropriate for the site of infection (i.e., skin, lungs, urinary tract, bloodstream). The empirical treatment of nosocomial P. aeruginosa infection that may have multi drug resistance is a combination of a β-lactam (e.g., penicillin, cephalosporin or carbapenem) and a fluoroquinolone (e.g., ciprofloxacin or levofloxacin) or an aminoglycoside (e.g., amikacin, gentamicin, tobramycin) (C. S. Curran et al., 2018). β-lactams mimic the terminal D-Ala-D-Ala moiety of the peptidoglycan pentapeptide, interfere with the cross-linking activity of penicillin-binding proteins and thus inhibit peptidoglycan synthesis (Pandey & Cascella, 2022). A common resistance mechanism against β-lactam is the expression of β-lactamases, such as the chromosomally encoded AmpC in *P. aeruginosa*, to inactivate β-lactams (Pandey & Cascella, 2022). Fluoroquinolones target DNA gyrase that is important for the relaxation of the supercoiled DNA strands and introduce DNA breaks during DNA replication (Fabrega et al., 2009). The predominant fluoroquinolone resistance mechanism in *P. aeruginosa* strains from cystic fibrosis is the overexpression of the MexCD-OprJ efflux pump (Fabrega et al., 2009). Aminoglycosides inhibit protein synthesis by binding to the 16S ribosomal RNA of the 30S ribosome (Krause et al., 2016). Resistance to aminoglycosides is commonly due to the methylation of the 16S rRNA which prevents effective target binding and the overexpression of efflux pumps (Krause et al., 2016).

Last-resort antibiotics include polymyxins (e.g., colistin and polymyxin B) that are used for extensively drug resistant *P. aeruginosa* infections (McEwen & Collignon, 2018; Vidaillac et al., 2012; Zavascki et al., 2007). Polymyxins interact with lipopolysaccharide (LPS) of the outer membrane of gram-negative bacteria and disrupt the bacterial membrane in a detergent-like manner (Zavascki et al., 2007). *P. aeruginosa* may

develop resistance to polymyxins by modifying the LPS to reduce its net negative charge for proper electrostatic interaction with the drugs. Other clinical strategies for treating multidrug *P. aeruginosa* infections include the combination therapy of a β-lactam and a β-lactamase such as ceftolozane-tazobactam and ceftazidime-avibactam treatments (Tummler, 2019). The difficulty in treating infections is a major reason why the World Health Organization prioritized the development of new therapies for this pathogen.

1.6.2. Genomic structure of *P. aeruginosa*

P. aeruginosa is noted for its high genomic diversity, with a large repertoire of genes that enables the species to colonize a wide variety of environments. With a genome size of 5.5 to 7Mbp, up to 20% of the total genome is comprised of accessory genomic elements such as GIs that are often associated with virulence and AMR (Subedi et al., 2018). *P. aeruginosa* also possesses a broad metabolic capacity and high proportion of regulatory genes for adapting to diverse environments and persisting in harsh clinical settings (Stover et al., 2000). *P. aeruginosa* is also equipped with a plethora of VFs (Figure 1.3). The signature blue-green phenazine pigment, pyocyanin, is secreted extracellularly by the type II secretion system to induce the formation reactive oxygen species which leads to cellular damage and cell death in hosts (Hall et al., 2016). Non-mucoid *P. aeruginosa* produces the proteolytic elastase capable of degrading plasma proteins like immunoglobulins, complement factors and cytokines and inflicting tissue damage in host (Kamath et al., 1998). The Type III secretion system (T3SS) is responsible for the direct delivery of the major exotoxins, exoenzyme S (ExoS), exoenzyme T (ExoT), exoenzyme U (ExoU) and exoenzyme Y (ExoY) into the host cytoplasm (Morrow et al., 2016; Newman et al., 2017; Wood et al., 2015). ExoS is a bifunctional enzyme with Rho GTPase-activating and ADP-ribosylating activities that are associated with increased invasiveness. ExoU is a cytotoxin with phospholipase activity that is associated with increased cytotoxicity, more severe infections and higher mortality in acute infections. Last but not least, ExoT induces host cell apoptosis while ExoY induces cytoskeletal disruption.

Most *P. aeruginosa* strains can be phylogenetically classified into one of the two major groups represented by the PAO1 clade (characterized by *exoS* expression) and the PA14 clade (characterized by *exoU* expression). The remaining strains often cluster into a third and more phylogenetically distant clade consisting of the taxonomic outlier

PA7 strain (Ozer et al., 2019). The *exoU* and *exoS* genes are typically mutually exclusive in *P. aeruginosa* isolates, which rarely carries both or none of these genes. PA14 is a hypervirulent strain characterized by the presence of two pathogenicity islands absent in PAO1 and the acquired mutation of *ladS* leading to a deleterious impact on biofilm formation, elevated T3SS activity and increased cytotoxicity towards mammalian cells in humans, mice or *C. elegans* (Mikkelsen et al., 2011).

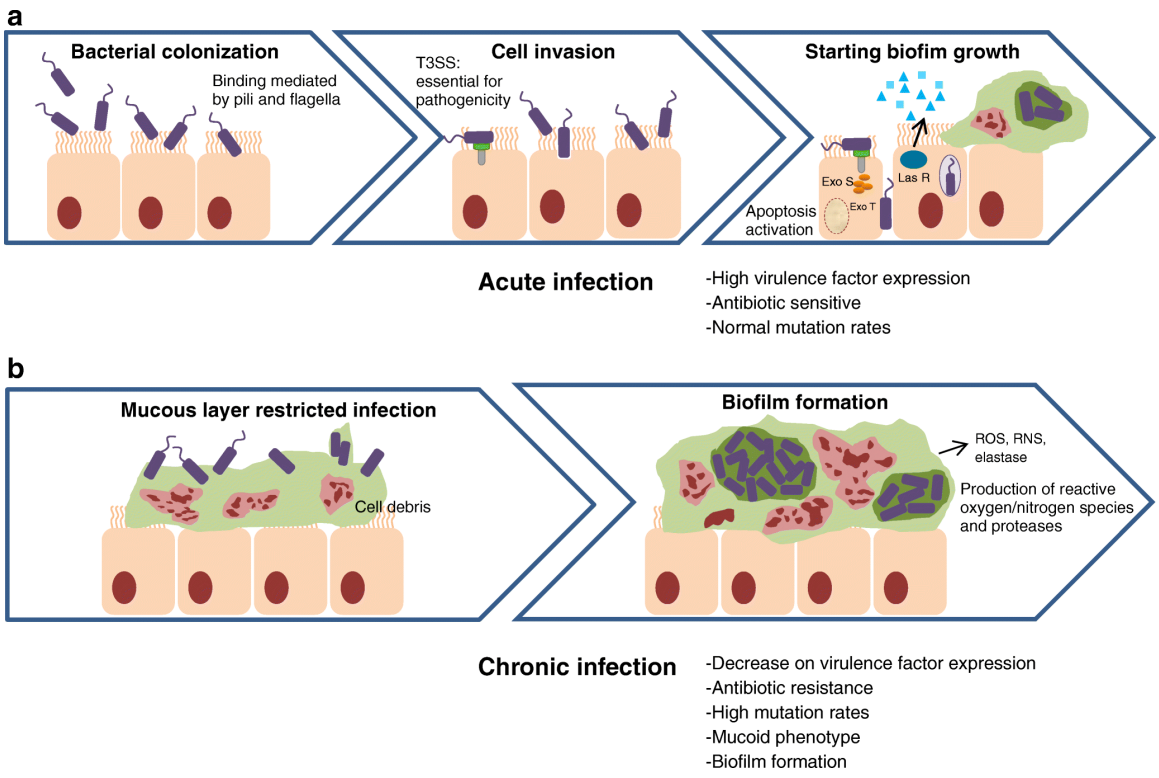


Figure 1.3 Pathogenesis of *P. aeruginosa* during acute and chronic infections. (A) Acute infection by P. aeruginosa is characterized by enhanced VF production and antibiotic sensitivity. (B) Chronic infection corresponds to a decrease in VF production and sensitivity to antibiotics due to biofilm formation. Figure from (Vilaplana & Marco, 2020).

1.6.3. *Caenorhabditis elegans* – an invertebrate infection model for *P. aeruginosa* pathogenesis

While mammalian model organisms such as mice or rats are useful for studying the complex pathogenesis of *P. aeruginosa* in mammalian chronic lung or acute burn wound infections (Bayes et al., 2016; Brandenburg et al., 2019), the nematode *C. elegans* is a simple yet powerful and cost-effective invertebrate model extensively used for the preliminary study of virulence in *P. aeruginosa.* Although lacking adaptive immunity, *C. elegans* employs an innate immune response, similar to plants and

mammals, to initial microbial infections through a conserved NF-kB signal transduction pathway, mediated by IL-1 and TOLL receptors to activate a defensive response with membrane-active small cationic peptides, defensins and other effector molecules (Ewbank & Zugasti, 2011). In addition, *C. elegans* lacks pattern recognition receptors (PRR) in its innate immune system to recognize pathogen-derived molecules. Notably, recent studies have also highlighted the importance of the *C. elegans* nervous system in pathogen sensing and immune regulation (Y. Liu & Sun, 2021). PA14 is the most virulent strain of *P. aeruginosa* (relative to PAO1, PA27, PO37, PAK) that can induce killing of *C. elegans* by a toxin-mediated fast killing or a slow infection-like killing process (Tan, Mahajan-Miklos, et al., 1999). During fast killing on high osmolarity peptone-glucose-sorbitol (PGS) agar, the presence of phenazine toxins, namely 1-hydroxyphenazine, phenazine-1-carboxylic acid, and pyocyanin, released by *P. aeruginosa* are sufficiently lethal to *C. elegans*, regardless of the viability of the pathogen (Cezairliyan et al., 2013). Slow killing over the course of a few days requires the infection and proliferation of live bacteria in the intestinal tract of worms grown on low osmolarity nematode growth medium (NGM) agar.

1.7. Metagenomics and One Health for pathogen detection in public health

One Health is a multisectoral and holistic approach in studying the intricate relationship among and maintaining the wellbeing of the biotic and abiotic components of an ecosystem. Life cycle and population dynamics of bacterial pathogens are often linked to the emergence, transmission and re-emergence of infectious diseases (Destoumieux-Garzon et al., 2018). In a constantly transforming ecosystem, habitat changes and physiological stress on microbial populations can drive evolutionary processes in pathogen such as the emergence of AMR genes and VFs as an adaptive response to the environmental changes. Habitat destruction, environmental pollution and climate change have profound impacts on the evolutionary selection and the geographic distribution of well-adapted organisms and their clinically important genes. Metagenomics, the study of the collective microbial genomic content from environmental samples, enables the assessment of functional potential and the risk of horizontal transmission of AMR and virulence-related genes within a microbial community. Metagenomic studies have shown that livestock is a major environmental reservoir for

AMR genes due to the frequent use of antimicrobials in animal feeding (Pilmis et al., 2020). AMR dissemination from animals to humans is enabled through biowaste, meat, milk, or direct contact. Surveillance of microbial communities across natural, agricultural, and urban settings is thus crucial for the rapid detection of pathogen emergence that could impose a public health threat and effective control of infectious disease transmission.

1.8. Goals of present research

1.8.1. Overall objective and hypothesis

While previous PAG analyses mainly focused on the functional assessment of known pathogen-associated VFs (Dhillon et al., 2015; S. J. Ho Sui et al., 2009; S. J. Ho Sui et al., 2012), I scaled up the analysis with a massively expanded bacterial genome dataset, and refined the predicted PAGs with additional analyses of taxonomic distribution and orthology inference. This thesis work also provided a more in-depth *in silico* functional characterization, with experimental lab-based validation, of a prioritized set of PAGs, some with undefined function at the start of the study, in *P. aeruginosa* PA14. Furthermore, I gained new structure-function insight into a candidate antivirulence drug initially identified through a PAG-based analysis. Lastly, I initiated the first study of PAGs in metagenomics datasets.

The overall objective of my research is to identify and prioritize potential pathogen-associated drug targets in priority pathogens, and characterize PAGs and initial antivirulence drug candidates, as a part of the antivirulence drug discovery efforts.

My main hypothesis is that coupling large scale bioinformatics analyses with in-depth functional analyses of select PAGs and antivirulence drug candidates will provide novel insight in themes in bacterial virulence and help identify and prioritize currently uncharacterized genes that may be suitable antivirulence drug targets. The discovery of novel virulence factors from pathogen-associated genes will also better our understanding of the bacterial mechanisms important for host-pathogen interactions.

1.8.2. Research aims

1. **Identify widely conserved PAGs, incorporating large-scale taxonomic and orthology analysis.**

   The first aim of the thesis, addressed in Chapter 2, was to identify widely conserved PAGs from an expanded set of over 8000 bacterial genomes. The hypothesis was that bacterial genes with important virulence-related functions are likely distributed in pathogens across multiple divergent taxa and are likely to be under positive selection. Taxonomic assessment of homologous gene groups revealed that PAGs are often conserved in only one or two closely related bacterial genera. For this reason, subsequent prioritization and characterization analyses were focused on PAGs that are conserved in pathogenic strains and absent in non-pathogenic strains within the chosen genus, *Pseudomonas*, under defined criteria.

2. **Characterize PAGs under positive selection.**

   The second thesis aim, also addressed in Chapter 2, was to characterize PAGs under positive selection with the hypothesis that PAGs under such selection are likely associated with adaptive functions in bacterial pathogens. This aim helps to identify and potentially prioritize PAGs that are likely functionally important for virulence for downstream computational and laboratory characterization in the third thesis aim.

3. **Further prioritize, including with laboratory-based validation, pathogen-associated genes as candidate antivirulence drug targets.**

   The third aim was to prioritize, including laboratory-based functional characterization, select PAGs as candidate antivirulence drug targets. Chapter 3 begins with a global protein SCL analysis of all bacterial genes analyzed in Chapter 2 to gain insight into the trends of SCL sites of PAGs in comparison to non-PAGs and common genes (found in both pathogens and non-pathogens). This analysis helped to prioritize PAGs as antivirulence drug targets based on drug accessibility, which is hypothesized to be higher in cell-surfaced or secreted pathogen-associated proteins. Chapter 3 then covers the first part of the third aim

by screening for virulence activity in the transposon insertion mutants of select PAGs in a worm infection model. This was followed by a gene/presence absence analysis of an interesting *P. aeruginosa* GI containing two prioritized PAGs, one with a virulence repressive activity and the other with evidence of positive selection. Chapter 4 addresses the second part of this third aim which involves the investigation of the structure-activity relationship of the FDA-approved osteoporosis drug, raloxifene, as a potential antivirulence agent against *P. aeruginosa*.

4. **Develop computational methods for metagenomics-based pathogen-associated gene characterization.**

The fourth thesis aim in Chapter 5 was to detect and characterize PAGs from metagenomics data, which has never been done prior to this study. The hypothesis was that read-based analysis using metagenomic data will complement existing genome-based methods in the characterization of PAGs and provide initial insight into the prevalence of such genes for surveillance purposes. Specifically, the prevalence of PAGs identified in Chapter 2 was assessed across lung microbiomes from individuals of different health status as well as freshwater microbiomes from upstream, downstream and at agriculturally affected (fecally polluted) watersheds.

Chapter 2.

Identification of PAGs, including those under positive selection

This chapter presents an updated version of the PAGs analysis, developed by the Brinkman Lab in 2009, with improvements to the algorithm to accommodate the massively expanded NCBI RefSeq bacterial genome dataset. To further explore the functional importance of PAGs in bacterial pathogens, I incorporated an orthology inference analysis to cluster PAGs into orthologous groups (i.e. groups of orthologs conserved only in bacterial pathogens) whose conservation across bacterial genera was then assessed. Using a subset of PAGs prioritized from the P. aeruginosa PA14 genome, I performed an evolutionary selection inference analysis to identify those with evidence of positive selection.

I completed all work presented in this chapter except for the initial development of the PAGs analysis algorithm (Dhillon et al., 2015; Fedynak, 2007; S. J. Ho Sui et al., 2009; S. J. Ho Sui et al., 2012), which I refined and built upon.

2.1. Abstract

As bacterial genomes become more abundantly available, large-scale comparative genome analyses are now possible to discover and better understand novel biological processes that help bacteria thrive in diverse environments. I improved the PAGs analysis algorithm, developed by the Brinkman Lab in 2009, with an updated dataset of over 8000 bacterial genomes and refined the results with an additional orthology analysis to identify genes with orthologs only in bacterial pathogens. Relative to genes present in both pathogens and non-pathogens, PAGs are disproportionately uncharacterized and taxonomically restricted to one or two bacterial genera. Those associated with known VFs were also more enriched in genes related to secretion systems, secreted proteins and toxins. In more focused analysis of the *Pseudomonas* genus, 17 PAGs from *P. aeruginosa* PA14 were prioritized for downstream analyses, including three genes evident of positive selection. These analyses altogether provide a preliminary prioritization strategy for bacterial genes that may be involved in important functions pertaining to the pathogenic lifestyle, as suggested from their unique conservation in pathogens and the evidence of positive selection in some of these genes.

2.2. Introduction

Bacterial pathogens deploy a plethora of virulence mechanisms to cause diseases in their host organisms (i.e., humans, animals and plants). Many bacterial VFs identified to date are directly involved in host colonization, infection, and immune evasion. In recent years, more research attention has been brought to other virulence-related genes such as regulatory and metabolic genes that control the expression and the production of VFs (Rutherford & Bassler, 2012; Suresh et al., 2021). Improved sequencing and bioinformatic tools enabled large-scale comparative genomic analyses to understand the underlying virulence determinants of disease-causing bacteria. Based on the first Molecular Koch's postulate, genes exclusively conserved in pathogenic bacteria and absent in non-pathogenic bacteria, referred to as PAGs in this thesis work, are likely contribute to bacterial pathogenicity (Falkow, 1988). A comparative genome analysis algorithm for identifying PAGs from comprehensive bacterial genome datasets has previously been developed by the Brinkman Lab in 2009 to analyze 613 genomes

and later updated in 2014 with 2794 genomes (Dhillon et al., 2015; S. J. Ho Sui, Fedynak, A., Hsiao, W.W.L, Langille, M.G.I & Brinkman, F.S.L., 2009). As newly sequenced bacterial genomes continue to increase, periodic updates of this PAGs analysis are needed to include and to refine the PAGs prediction using the ever-expanding bacterial genome datasets that are publicly available.

Understanding the functional role of PAGs in bacterial pathogens is crucial for identifying novel VFs for the development of novel antimicrobial alternatives, particularly antivirulence drugs, against bacterial infections. Homology detection of PAGs to known VFs, protein families and conserved protein domains may provide functional information, given that a significant sequence similarity exists between the PAGs of interest and their curated orthologs in the chosen databases. However, many genes remain uncharacterized owing to their sequence divergence from well annotated genes. The lack of functional characterization of some PAGs may be attributed to the rapid evolution and the selection of important and potentially virulence-related functions. As such, some PAGs may require alternative approaches such as non-sequence similarity-based computational methods to predict and infer their biological functions and features. Taxonomic distribution of PAGs among bacterial genomes may provide insight into the prevalence and nature of gene function as virulence genes with more specialized functions have been shown to be more taxon-specific than genes with more general functions that are often conserved among more bacterial taxa (Brambila-Tapia et al., 2014).

Positive selection plays an influential role in driving the evolution of virulence-related genes that are beneficial to a pathogen's adaptation and survival in the host environment. Particularly, novel genes that increase an organism's fitness in its pathogenic lifestyle are likely selected for and increased in prevalence within the bacterial population over time. For instance, VFs that are pathogen-associated have previously been shown to be disproportionately related to more offensive functions such as host invasion and T3SS (S. J. Ho Sui et al., 2009). Inferring positive selection in PAGs may reveal those with functional importance in pathogens and may facilitate the prioritization of pathogen-associated and positively selected genes for a more comprehensive characterization as candidate antivirulence drug targets.

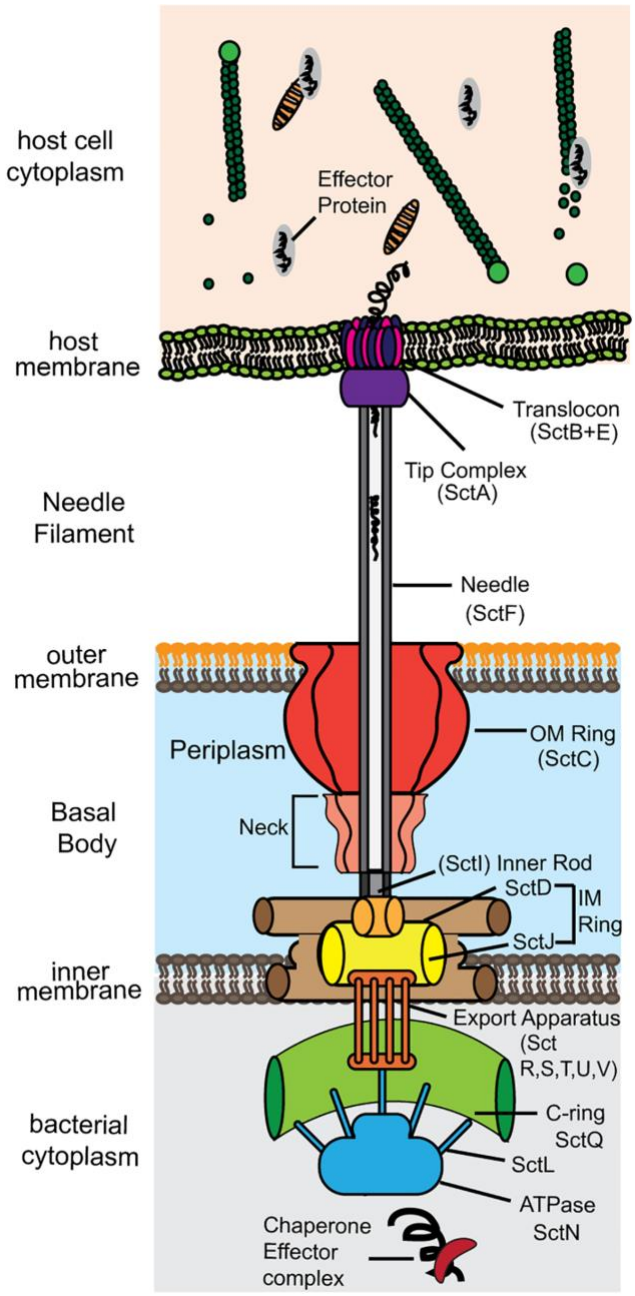


Figure 2.1 Components of the bacterial T3SS in the universal Sct nomenclature.

The T3SS apparatus delivers effector molecules from the bacterial cytoplasm directly into host cell. The basal body spanning the inner and outer membrane includes SctC (PscC), SctD (PscD), SctJ (PscJ), SctF (PscF), SctI (PscI) and is connected to the cytoplasmic C-ring SctQ (PscQ). The export apparatus consists of SctR (PscR), SctS (PscS), SctT (PscT), SctU (PscU) and SctV (PscV). The ATPase complex which provides energy for effector transport includes SctK (PscK), SctL (PscL), SctN (PscN) and SctO (PscO). T3SS is regulated by the needle length regulator SctP (PscP) and the export regulator SctW (PopN). Figure from (Dey et al., 2019). The translocator complex is formed by SctA (PcrV), SctB (PopD) and SctE (PopB). Text within the parentheses represents the species-specific nomenclature in *P. aeruginosa.*

Bacterial T3SSs are needle-like, protein transport machineries, found in many notorious gram-negative pathogens, that deliver VFs and effectors directly across the bacterial envelope and into the host cell (Deng et al., 2017). The structural components of the T3SS consist of approximately 20 highly conserved proteins (Figure 2.1). These T3SS proteins conserved across multiple bacterial species follow the unified secretion and translocation (Sct) nomenclature, initially proposed in 1998 and later expanded in 2016 (Hueck, 1998; Portaliou et al., 2016). The T3SS injectisome includes the basal body (SctC, SctD, SctJ, SctF and SctI), export apparatus (SctR, SctS, SctT, SctU and SctV), cytoplasmic ring (SctQ), ATPase complex (SctK, SctL, SctN and SctO), regulators (SctP and SctW) and translocators (SctA, SctB and SctE) (Deng et al., 2017). T3SS components that are found in bacterial pathogens and are more important for host interactions are hypothesized to be more likely to be pathogen-associated and under evolutionary pressure for positive selection.

Here, I updated and improved the scalability of the PAGs analysis in 2018 with 8646 bacterial genomes, triple in size relative to the genome dataset from the 2014 PAGs analysis. While the original algorithm identifies PAGs based solely on a single-way sequence similarity search, this 2018 update features an additional orthology analysis using OrthoFinder (Emms & Kelly, 2019) to infer orthologs of all genes from reciprocal best hits whose BLAST scores are normalized by gene length. This normalization step in the OrthoFinder algorithm improves both the recall of protein-coding genes with short sequences that are sometimes missed due to their inability to produce high bit scores or low e-values to pass the statistically significant threshold, as well as the precision of genes with long sequences that are sometimes falsely predicted as homologs due to the increased probability of producing high bit scores and low e-values. As the sequencing of novel bacterial genomes outpaces the characterization of their genomic and taxonomic features, I also assessed the potential implementation of a modified criterion in the PAGs detection to accommodate for any potential inaccurate taxonomic classification and pathogen status assignment in future analyses. Furthermore, I included the functional characterization and taxonomic conservation assessment of not only PAGs, but also non-PAGs (genes conserved only in non-pathogenic bacteria) that were not included in the 2009 and 2014 PAGs analyses. Moreover, with a focus on the WHO priority pathogen, *P. aeruginosa*, positive selection

inference was performed on a subset of PAGs in the hypervirulent *P. aeruginosa* PA14 strain as well as a set of T3SS structural and accessory genes.

The main goals of my research in this chapter are to 1) identify PAGs from the updated and expanded bacterial genome dataset, 2) compare trends in taxonomic conservation and biological functions among PAGs, non-PAGs and common genes (found in both pathogens and non-pathogens), and 3) infer positive selection in select PAGs as well as a set of T3SS genes to prioritize those with putative functional importance in pathogens for further downstream characterization.

2.3. Methods

2.3.1. Identification and preliminary characterization of PAGs from the complete NCBI RefSeq bacterial genome dataset

An updated set of 8646 Reference Sequence (RefSeq) bacterial genomes from the National Center for Biotechnology Information (NCBI) was retrieved from the MicrobeDB database v97 (Langille et al., 2012) on April 26th, 2018. Since The Institute For Genomic Research's Microbial Genome Properties Table, which provided the pathogen status of bacterial genomes (whether they are pathogens to humans, animals or plants) to the 2009 PAGs analysis, was no longer accessible at the time of my 2018 PAGs analysis update, I manually assigned each novel genome a pathogen status of "pathogen" or "non-pathogen" based on a manual curation of documented pathogenicity at the species level (Fedynak, 2007; Haft et al., 2005). Specifically, genomes were classified as a "pathogen" if their corresponding species had published evidence of infection in any host organism (most commonly in humans, animals or plants). On the contrary, genomes were classified as "non-pathogens" if their corresponding species have been documented in publications as environmental (non-host-associated) organisms or as host-associated organisms with no current evidence of pathogenicity. Genomes that were not reported or characterized in any publications at the time of the curation were assigned an "unknown" pathogen status and were excluded from the analysis. Due to the massive expansion of sequenced genomes since the previous analysis update in 2014 (Dhillon et al., 2015), the pathogen status of genomes was no longer curated individually at the strain level. All genomes/strains within a species were collectively assigned as a "pathogen" if at least one strain within the species shows

evidence of pathogenicity in at least one host organism, or as a “non-pathogen” if none of the strains within the species is reportedly pathogenic towards any host organisms. This species level curation of pathogen status mainly affects species like *E. coli* and *Listeria monocytogenes* that encompass both pathogenic and non-pathogenic strains (Chen et al., 2011; Lorenz et al., 2020).

The protein sequences of genes from a total of 5,196 pathogenic and 3,523 non-pathogenic RefSeq genomes were searched for sequence similarity against all genes from all other genomes, excluding the genome of origin, using DIAMOND (Buchfink et al., 2021; Buchfink et al., 2015). An e-value cut-off of 10^{-7} was used to exclude potentially distant homologs. A gene was considered pathogen-associated, non-pathogen-associated, or common if its protein sequence had significant sequence similarity (below the 10^{-7} e-value threshold) to genes found in the genomes of only bacterial pathogens, only non-pathogens, or both pathogens and non-pathogens, respectively. With the original intent of finding functional genes widely conserved among the diverse bacterial pathogens or non-pathogens, the unique genes were subdivided into “high quality” and “low quality” PAGs or non-PAGs based on three criteria: 1) broad phyletic distribution (conserved in three or more bacterial genera), 2) non-genus specificity (at least one genus in which the PAGs or non-PAGs were detected must containing both pathogenic and non-pathogenic species), and 3) proteins with probable function (more than 100 amino acids). PAGs and non-PAGs that satisfied all three criteria are considered “high quality” while the others are considered “low quality.”

The prevalence of functionally unknown and defined PAGs, non-PAGs and common genes was estimated from their NCBI gene annotations. Genes whose annotation containing “hypothetical protein” or “DUF” (domain of unknown function) were considered functionally unknown while the remaining genes were considered functionally defined. VF and protein family associations of these genes were also determined through a DIAMOND-based sequence similarity search against all known VFs in the Virulence Factor Database (B. Liu et al., 2019) as of January 10th, 2019 and all protein families within the Pfam database v33.0 (Mistry et al., 2021). An e-value cut-off of 10^{-7} and a sequence identity threshold of 90% were used. Genes associated with known VFs were further classified by the categories of bacterial VFs defined in VFDB. Statistical analysis of the over-presentation or under-representation of PAGs or non-PAGs, relative to common genes was performed by tabulating the number of genes, within each

pathogen association, that are or are not associated with functionally defined genes, known VFs or protein families, running a Chi-square test of association, followed by pairwise post-hoc tests with multiple testing correction using the Benjamini-Hochberg Procedure. Statistics were computed in R with the R package “rcompanion” v2.3.26.

To refine the results from the PAGs analysis, OrthoFinder v2.3.12 (Emms & Kelly, 2019) was used to infer orthologous relationship among all bacterial genes and cluster orthologs into “orthogroups.” In contrast to the original PAGs analysis algorithm which uses a single sequence similarity search of a gene against genes in other genomes, OrthoFinder implements a reciprocal best hit search, with bit scores normalized by gene length and phylogenetic distance, to identify orthogroups. A species tree and gene trees of each orthogroup are then constructed for gene tree-species tree reconciliation to identify gene duplication/loss and to infer probable paralogs within each orthogroup. Due to filesystem and runtime limitations on the Compute Canada’s Cedar high-performance computing cluster, the NCBI RefSeq bacterial genome dataset was reduced to 501 reference and representative genomes prior to the orthology inference by OrthoFinder. Pathogen-associated and non-pathogen-associated orthogroups were identified if all orthologs within the orthogroups belonged to genomes of only pathogenic or non-pathogenic bacteria, respectively. All orthologs within the pathogen-associated or non-pathogen-associated orthogroups are therefore identified as PAGs or non-PAGs, respectively. Taxonomic conservation of pathogen-associated and non-pathogen-associated orthogroups were assessed by counting the number of unique bacterial genera in which their orthologs were detected. Pairwise association of bacteria genera among all pathogen-associated or non-pathogen-associated orthogroups were visualized on circus plots created by the R package “circlize” v0.4.10. Specifically, an input file with the following fields was prepared: 1) the primary bacterial genus, 2) the paired bacterial genus and 3) number of pathogen-associated or non-pathogen-associated orthogroups shared among the genus pairs. The chordDiagram() function in “circlize” then mapped the data in a circos plot with the primary bacterial genera organized by alphabetical order in the clock wise direction, followed by the paired bacterial genera. The thickness of the links of the genus pairs is proportional to the number of orthogroups shared among each genus pair.

2.3.2. Genus-specific analysis of the NCBI RefSeq *Pseudomonas* genome dataset

A complete set of *Pseudomonas* genomes from the NCBI RefSeq Database was retrieved through MicrobeDB v102 on October 12, 2020. Each genome was assigned a "pathogen" or "non-pathogen" status based on manual literature search for evidence of pathogenicity in its respective strain. From a total of 605 *Pseudomonas* RefSeq genomes, a total of 306 pathogenic and 243 non-pathogenic *Pseudomonas* genomes were identified. With a focus on the hypervirulent *P. aeruginosa* UCBPP-PA14 strain (NCBI RefSeq genome assembly accession GCF_000014625.1), pathogen associated genes were identified by DIAMOND-based sequence similarity search (e-value cut-off of 10^{-7}) of the protein sequences of all genes in the *P. aeruginosa* PA14 genome against genes from other *Pseudomonas* genomes. Within this genus-specific PAGs analysis, a gene is considered pathogen-associated if found (i.e., had significant sequence similarity with genes) only in pathogenic *Pseudomonas* strains.

Orthology inference using OrthoFinder v2.3.12 required genome reduction to a smaller dataset that was manageable by Compute Canada's server. Phylogenetic distances among the 605 *Pseudomonas* genomes were estimated by MASH v2.3 using 10000 sketches of the default 21-mers, as employed and specifically tested on *Pseudomonas* genomes in IslandViewer4 for the selection of representation genomes (Bertelli et al., 2017). To address the over-representation of certain species like *P. aeruginosa,* with many sequenced clinical isolates, within the original genome dataset, pairwise MASH distances were used to determine a genomic distance cut-off for clustering highly similar genomes together and to select a single representative genome from each cluster. Specifically, the MASH distance matrix was used for hierarchical clustering and dendrogram visualization of all *Pseudomonas* genomes, using the R "stats" package v4.0.2. Upon the examination of a few tested MASH distance cut-offs of 0.21, 0.1, 0.08, 0.06 and 0.04 which generated 133, 206, 237, 295, and 380 genome clusters, respectively, the 0.1 cut-off was chosen as it was the highest MASH distance that resolved the three main *P. aeruginosa* lineages, PAO1, PA14 and PA7 into separate clusters while effectively reducing the original dataset to approximately a third in size. Within each cluster under the 0.1 MASH distance cut-off, an NCBI reference or representative genome was prioritized for selection. If no reference or representative genome was present in the cluster, one genome was randomly selected. The PAGs

analysis followed by orthology inference was performed on the reduced set of 206 *Pseudomonas* genomes. The PAGs analysis was done on both the complete and the reduced *Pseudomonas* genome dataset for comparison of results using the different genome dataset sizes. Results between the PAGs analysis and the OrthoFinder-based orthology inference were also compared using the reduced genome dataset. *P. aeruginosa* PA14 genes, whose pathogen association was supported by the PAGs analyses using both the complete and reduced *Pseudomonas* dataset, as well as the orthology analysis using the reduced dataset, were selected for more in-depth downstream characterization.

2.3.3. Evaluation of a more flexible criterion for PAGs detection

Re-assignment of pathogen association to genes

The pathogen association assignment step in the *Pseudomonas*-specific PAGs analysis, using the complete *Pseudomonas* genome dataset, in Chapter 2.3.2 was re-run with a new criterion allowing 1 to 10% of a pathogen-associated gene's significant sequence similarity search hits (probable homologs) to be of non-pathogen origin. A presence-absence matrix of PAGs and their probable homologs defined under each percentage in this criterion was mapped to a species tree constructed from the full *Pseudomonas* dataset (see section below). The identity of the non-pathogen genomes containing the probable homologs of the PAGs was compared among the different level of stringency (the percentage of non-pathogen genomes included) under this novel criterion.

Phylogenetic analysis of the Pseudomonas genus

The full RefSeq dataset of *Pseudomonas* genomes obtained from MicrobeDB v102 was used to construct a species tree using concatenated nucleotide sequences of four housekeeping genes, 16S ribosomal RNA (16s rRNA), DNA gyrase B (*gyrB*), RNA polymerase beta subunit (*rpoB*) and RNA polymerase sigma factor (*rpoD*), that have been shown to be effective in resolving the *Pseudomonas* phylogeny and identifying novel strains (Siehnel et al., 2010). Between the time at which the *Pseudomonas*-specific PAGs analysis (Chapter 2.3.2) was performed and the time at which this analysis (Chapter 2.3.3) was performed, a few RefSeq genomes were suppressed in the NCBI RefSeq Database due to reasons such as "missing strain identifier" or "containing

many frameshifted proteins", thus were excluded from this *Pseudomonas* species tree construction (Table 2.1). Multiple sequence alignment was done using MUSCLE v3.8.31 (Edgar, 2004), followed by phylogenetic tree construction using IQtree v2.0.3, which is reportedly ideal for concatenation-based species tree inference (Minh et al., 2020; X. Zhou et al., 2018). *E. coli* str. K-12 substr. MG1655 (GCF_000005845.2) was chosen as the outgroup.

Table 2.1 Supressed NCBI RefSeq *Pseudomonas* genome assemblies as of May 2021.

RefSeq genome assembly accession	Species	Strain	Pathogen status	Reason for suppression
GCF_000953455	*P. oleovorans*	Ppseudo_Pac	Non-pathogen	Missing strain identifier
GCF_002688705	*P. fulva*	SB1	Pathogen	Many frameshifted proteins
GCF_002843285	*P. sp. AK6U*	AK6U	Non-pathogen	Many frameshifted proteins
GCF_901472545	*P. aeruginosa*	NCTC13359	Pathogen	Many frameshifted proteins
GCF_901472565	*P. aeruginosa*	NCTC13620	Pathogen	Many frameshifted proteins
GCF_901472595	*P. aeruginosa*	NCTC13618	Pathogen	Many frameshifted proteins

2.3.4. Evolutionary selection analysis of PAGs and T3SS genes

Dataset

The 17 PAGs from the *P. aeruginosa* PA14 genome identified from the genus-specific PAGs analysis in Chapter 2.3.2 were selected for positive selection inference. T3SS structural genes and a variety of T3SS regulators, chaperones and effectors were retrieved through manual curation of published papers (Portaliou et al., 2016). NCBI protein accessions, locus tags, and nucleotide sequences were curated through cross-referencing NCBI gene/protein annotations. Orthologs of all PAGs and T3SS genes selected for this analysis were retrieved from the OrthoFinder-predicted orthogroups in Chapter 2.3.3.

Evolutionary selection inference

A more extensive evolutionary analysis of PAGs and T3SS-related genes in 14 *P. aeruginosa* representative genomes (retrieved from the reduced 206 Pseudomonas genome dataset) was performed using the HYpothesis testing using the PHYlogenies (HyPhy) v2.5 package (Kosakovsky Pond et al., 2020). HyPhy was the preferred tool for evolutionary analysis since it allows for synonymous rate (dS) variation to minimize

misidentification of positively selected sites in genes (Wisotsky et al., 2020). For each gene set, a codon-based multiple nucleotide sequence alignment, with masked internal and external stop codons, was generated from coding sequences for all *P. aeruginosa* genes within the orthogroup using MACSE v2.04 (Ranwez et al., 2018).

Multiple sequence alignments of the gene sets of interest were preprocessed and screened for recombination, which may affect phylogenetic tree construction, by HyPhy's Genetic Algorithm for Recombination Detection (GARD) method (Kosakovsky Pond et al., 2006). Alignments with detected recombination are partitioned into fragments based on the predicted location of recombination breakpoint. Selection inference by downstream HyPhy methods was done separately on the phylogeny of each alignment partition. GARD output is directly compatible with other HyPhy tools.

Genes were analyzed for positive selection by multiple HyPhy methods, with each method testing a slightly different hypothesis for codon evolution. Positive selection can be episodic (specific to some but not all lineages within a gene phylogeny) or pervasive evident across all phylogenetic lineages). Mixed Effects Model of Evolution (MEME) was chosen as the primary method for detecting site-specific episodic positive selection as it is better able to identify individual sites with evidence of positive selection under a proportion of the lineages in a phylogeny (Murrell et al., 2012). However, the predicted sites under positive selection were not analyzed in detail due to the limited data supporting the accuracy of such predictions. This is particularly important for capturing signatures of positive selection in genes relevant for pathogenesis and virulence as adaptive evolution of genes is usually attributed to a few amino acids, rather than the entire protein-coding sequence, undergoing episodic positive selection. (Guindon et al., 2004; Murrell et al., 2012).

The gene datasets were also analyzed by other complementary HyPhy methods that may provide additional information. Adaptive Branch Site Random Effects Likelihood (aBSREL) is a complementary test to MEME for identifying specific lineages on a phylogeny on which positive selection is evident (M. D. Smith et al., 2015). Contrary to the MEME and aBSREL which infer episodic selection, the Fixed Effects Likelihood (FEL) method tests for pervasive positive selection in small (less than 100 sequences) gene sets such as the PAGs and T3SS-related genes dataset used in this chapter (Kosakovsky Pond & Frost, 2005; Murrell et al., 2013). FEL assumes a constant

selective pressure at each site within a gene across the entire phylogeny (i.e., no phylogenetic lineages are under higher selection than others). The Branch-Site Unrestricted Statistical Test for Episodic Diversification (BUSTED) method, on the other hand, identifies gene-wide instead of site-specific episodic positive selection of genes (Murrell et al., 2015).

Controls

Bacterial genes known to be under purifying or positive selection are chosen as negative and positive controls, respectively, to ensure the interpretation of the HyPhy selection inference results was appropriate. Seven multilocus sequence typing genes (*acsA, aroE, guaA, mutL, nuoD, ppsA* and *trpE*) in *P. aeruginosa* were initially selected as negative controls (data not shown) (B. Curran et al., 2004). However, due to the lack of published dN/dS ratios or selection inference analyses on these genes, the RNA polymerase sigma factor (*rpoD*), DNA gyrase subunit B (*gyrB*), and DNA-directed RNA polymerase beta chain (*rpoB*) were used as negative controls for this analysis as purifying selection for these three genes have been supported by published data (Mulet et al., 2010). Likewise, the pyoverdine outer membrane receptor (*fpvA*) and the pyoverdine inner membrane protein (*fpvG*), with published evidence of positive selection (Tummler & Cornelis, 2005), were used as positive controls in this analysis.

2.4. Results

2.4.1. PAGs are disproportionately enriched in T3SS and toxin-related VFs, located on GIs, and functionally undefined

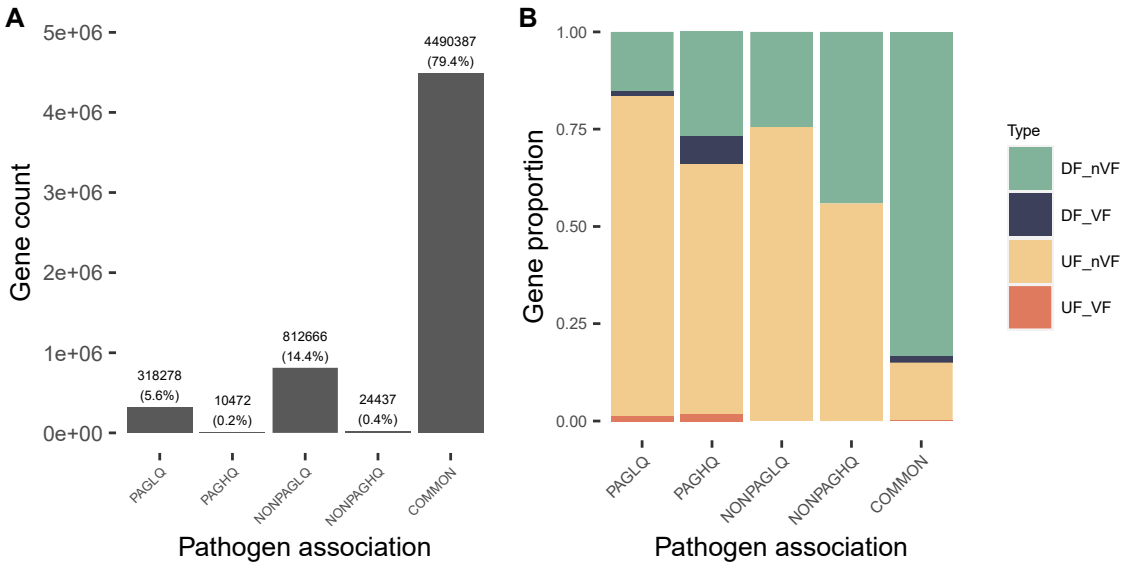
The third update of the PAGs analysis was done in 2018 to incorporate the expanded bacterial genome dataset since the previous analysis from 2014 (Table 2.2). From 8449 bacterial RefSeq genomes, 328,750 (5.8%) PAGs, 837,103 (14.8%) non-PAGs and 4,490,387 (79.4%) common genes (found in both pathogens and non-pathogens) were identified (Figure 2.2A). “High quality” PAGs and non-PAGs that satisfied all three criteria - broad taxonomic conservation, non-genus-specificity, and gene length of over 300 base pairs - only represented 0.2% and 0.4% of the total bacterial genes, respectively, suggesting at least one of the three criteria (broad taxonomic conservation – see Chapter 2.4.2) may be under-represented. A preliminary functional trend analysis of bacterial genes by pathogen association revealed that PAGs

relative to common genes are disproportionately enriched in genes currently with an unknown function but are associated with known VFs (Figure 2.2). On the contrary, no known VFs were detected in the non-PAGs, in support of the original hypothesis that PAGs are more likely to confer virulence-related functions and to be unique to pathogens while non-PAGs are not. A closer examination of all genes with VF association showed that, relative to common genes, PAGs are more enriched in VF-associated genes encoding T3SS components, secreted proteins and toxins (p-value = 5.0^{-4}) (Figure 2.3). All bacterial genes were also searched against the Pfam database for significant sequence similarity to known protein families curated to date. In agreement with previous assessment of functional annotation of genes within each pathogen association, PAGs as well as non-PAGs, relative to common genes, are significantly enriched in genes with uncharacterized protein families, suggesting that some genes unique and perhaps important for the pathogenic or non-pathogenic lifestyle may have been overlooked in the past (Figure 2.4). This set of preliminary functional analyses highlighted the lack of characterization of genes that are uniquely conserved in non-pathogens or pathogens, especially drawing attention to PAGs that may represent novel VF classes.

Table 2.2 Summary of the PAGs analysis performed in 2009, 2014 and 2018.

	2009	2014	2018
Total genomes	631	2,794	8,449
Total pathogen genomes	298	1,277	5,196
Total non-pathogen genomes	333	1,517	3,253
Pathogen status curation	Strain level	Strain level	Species level
Protein sequence similarity search	BLASTp	BLASTp	DIAMOND*

*DIAMOND is faster and more scalable than BLASTp (Buchfink et al., 2015)



C

Comparison	p.Chisq	p.adj.Chisq	Pathogen_association
DF_nVF : DF_VF	0.0005	0.0005	COMMON−PAGLQ
DF_nVF : UF_nVF	0.0005	0.0005	COMMON−PAGLQ
DF_nVF : UF_VF	0.0005	0.0005	COMMON−PAGLQ
DF_VF : UF_nVF	0.0005	0.0005	COMMON−PAGLQ
DF_VF : UF_VF	0.0005	0.0005	COMMON−PAGLQ
UF_nVF : UF_VF	0.0005	0.0005	COMMON−PAGLQ
DF_nVF : DF_VF	0.0005	0.0005	COMMON−NONPAGHQ
DF_nVF : UF_nVF	0.0005	0.0005	COMMON−NONPAGHQ
DF_nVF : UF_VF	0.0005	0.0005	COMMON−NONPAGHQ
DF_VF : UF_nVF	0.0005	0.0005	COMMON−NONPAGHQ
UF_nVF : UF_VF	0.0005	0.0005	COMMON−NONPAGHQ
DF_nVF : DF_VF	0.0005	0.0005	COMMON−PAGLQ
DF_nVF : UF_nVF	0.0005	0.0005	COMMON−PAGLQ
DF_nVF : UF_VF	0.0005	0.0005	COMMON−PAGLQ
DF_VF : UF_nVF	0.0005	0.0005	COMMON−PAGLQ
DF_VF : UF_VF	0.0005	0.0005	COMMON−PAGLQ
UF_nVF : UF_VF	0.0005	0.0005	COMMON−PAGLQ
DF_nVF : DF_VF	0.0005	0.0006	COMMON−PAGHQ
DF_nVF : UF_nVF	0.0005	0.0006	COMMON−PAGHQ
DF_nVF : UF_VF	0.0005	0.0006	COMMON−PAGHQ
DF_VF : UF_VF	0.0005	0.0006	COMMON−PAGHQ
UF_nVF : UF_VF	0.0005	0.0006	COMMON−PAGHQ
DF_VF : UF_nVF	0.2080	0.2080	COMMON−PAGHQ
DF_VF : UF_VF	NA	NA	COMMON−NONPAGHQ

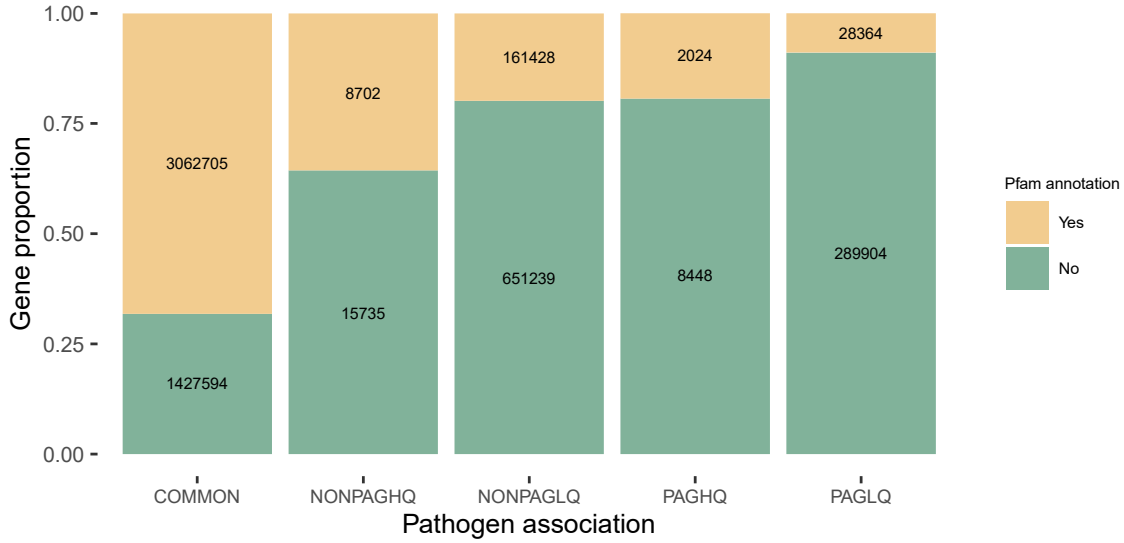
Figure 2.2 Classification of bacterial genes by pathogen-association revealed that relative to common genes, PAGs are disproportionately associated genes without a characterized function but are associated with known VFs.

(A) Distribution of bacterial genes in each of the five pathogen-association categories: “low-quality” PAGs (PAGLQ), “high-quality” PAGs (PAGHQ), “low-quality” non-PAGs (NONPAGLQ), “high-quality” non-PAGs (NONPAGHQ) and common genes (COMMON). Numbers with or without parentheses above each bar denote the total number or percent, respectively, of bacterial genes in each pathogen association. **(B)** Proportions of genes within each pathogen association with a defined function (DF) or unknown function (UF) and that are VF- or non-VF (nVF)-associated. Chi-squared test was first performed to show a significant, general association between the number of genes by pathogen association and their functional annotation (VF-association and functional annotation status) (p= 0.0005). **(C)** Post-hoc pairwise Chi-squared test were subsequently performed to compare the gene count in common genes and the 4 other pathogen association groups across the 4 functional annotation statuses (DF_VF, DF_nVF, UF_VF and UF_nVF). Relative to common genes, PAGs and non-PAGs showed a significant difference in gene count in majority of the functional annotation groups. “NA” indicates that the Chi-squared test could not be performed due to the absence of VF-associated genes in the non-PAGs. Associations are statistically significant if p-value is less than 0.05. In the post-hoc tests, p-values were adjusted for multiple testing by the Benjamini–Hochberg false discovery rate method.

Figure 2.3 Classification of VF-associated bacterial genes by VF classes showed that relative to common genes, PAGs are more enriched in genes encoding bacterial secretion components, secreted proteins and toxins.

VF classes were defined by the Virulence Factor Database used in this analysis. Pathogen-association types include “low-quality” PAGs (PAGLQ), “high-quality” PAGs (PAHGQ), “low-quality” non-PAGs (NONPAGLQ), “high-quality” non-PAGs (NONPAGHQ) and common genes (COMMON). Number above each bar represents the number of genes categorized by pathogen association.

A



B

Comparison	p.Chisq	p.adj.Chisq
COMMON : NONPAGHQ	0.0005	0.000556
COMMON : NONPAGLQ	0.0005	0.000556
COMMON : PAGHQ	0.0005	0.000556
COMMON : PAGLQ	0.0005	0.000556
NONPAGHQ : NONPAGLQ	0.0005	0.000556
NONPAGHQ : PAGHQ	0.0005	0.000556
NONPAGHQ : PAGLQ	0.0005	0.000556
NONPAGLQ : PAGLQ	0.0005	0.000556
PAGHQ : PAGLQ	0.0005	0.000556
NONPAGLQ : PAGHQ	0.1790	0.179000

Figure 2.4 Classification of bacterial genes by Pfam annotations showed that PAGs disproportionately lack association to known protein families.

(A) Proportion of bacterial genes with or without Pfam protein family annotation in each pathogen association category. Pathogen association types: “low-quality” PAGs (PAGLQ), “high-quality” PAGs (PAHGQ), “low-quality” non-PAGs (NONPAGLQ), “high-quality” non-PAGs (NONPAGHQ) and common genes (COMMON). Numbers within the bar represent gene count. Chi-squared test was first performed to show a significant, general association between the number of genes by pathogen association and their Pfam protein family association (p= 0.0005). **(B)** Post-hoc pairwise Chi-squared test were subsequently performed to compare the gene counts in each pair of pathogen association groups and their protein family association. Relative to common genes, PAGs and non-PAGs showed significantly higher prevalence of genes without a detected association to known protein families within the Pfam database v33. Associations are statistically significant if p-value is less than 0.05. In the post-hoc tests, p-values were adjusted for multiple testing by the Benjamini–Hochberg false discovery rate method.

2.4.2. PAGs are mostly conserved within one or two bacterial genera

PAGs and non-PAGs identified in the 2018 PAGs analysis (Chapter 2.4.1) were further refined and clustered into orthologous groups called orthogroups using OrthoFinder (Emms & Kelly, 2019). Contrary to the PAGs analysis algorithm which relies on a single sequence similarity search to identify sequence similarity hits (putative homologs) of a query gene, OrthoFinder is a more sophisticated ortholog inference approach that includes a reciprocal best BLAST, ortholog clustering and gene tree-species tree reconciliation to distinguish orthologs from paralogs. From a subset of 501 representative or reference RefSeq bacterial genomes, 106 pathogen-associated orthogroups (Figure 2.5) and 169 non-pathogen-associated orthogroups (Figure 2.6) were identified. Only 11 pathogen-associated orthogroups and 4 non-pathogen-associated orthogroups were conserved across 3 or more bacterial genera. Moreover, only 2 pathogen-associated orthogroups and 1 non-pathogen-associated orthogroups are detected across 4 bacterial genera. The majority of genes that are only found in either pathogens or non-pathogens are conserved among two phylogenetically or ecologically related genera, suggesting that PAGs and non-PAGs may confer niche-specific functions shared among bacteria inhabiting a similar environment. For example, pathogen-associated orthogroups are commonly shared between *Borrelia* and *Borreliella* (many vector-borne pathogens), *Escherichia* and *Shigella* (many foodborne pathogens) and *Vibrio* and *Alivirio* (aquatic bacteria with some pathogenic species). Similarly, non-pathogen-associated orthogroups are found in genus pairs such as *Clostridium* and *Bacillus* (endospore-forming organisms), *Bacteroides* and *Prevotella* (gut microbes), and *Rhodococcus* and *Gordonia* (environmental organisms). The limited taxonomic conservation of PAGs and non-PAGs showed that selecting “high-quality” PAGs found in 3 or more genera in the PAGs analysis (Chapter 2.3.1) is not a suitable criterion for prioritizing PAGs for downstream functional analyses. Instead, focusing on PAGs conserved in pathogenic species within a genus maybe be useful in identifying novel VFs that may be used as precise, pathogen-specific antivirulence drug targets.

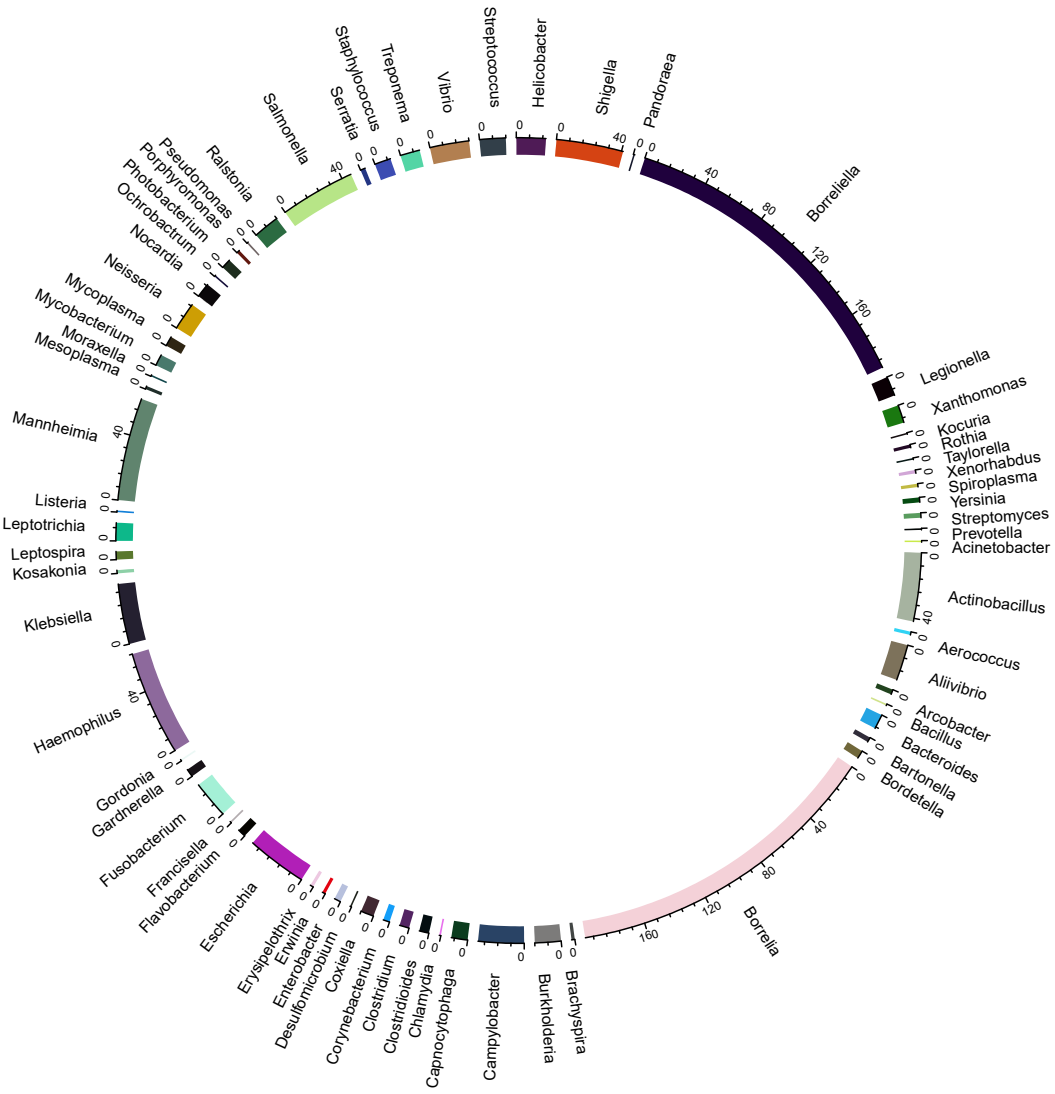


Figure 2.5 Pairwise genus association of pathogen-associated orthologs among the 501 RefSeq reference/ representative bacterial genomes showed that majority of the 106 detected pathogen-associated orthogroups are conserved in two ecologically similar genera.

Number on scale represents the number of PAGs, identified from 501 RefSeq reference/ representative bacterial genomes, shared between each pair of bacterial genera. For orthogroups conserved in more than two genera, all genus pairs are illustrated individually (e.g., an orthogroup conserved in 3 genera, A, B and C, will show the pairwise association between A-B, B-C and A-C). Orthogroups conserved only in one genus were excluded.

Figure 2.6 Pairwise genus association of non-pathogen-associated orthologs among the 501 RefSeq reference/ representative bacterial genomes showed that majority of the 169 detected non-pathogen-associated orthogroups are also conserved in two ecologically similar genera.

Number on scale represents the number of non-PAGs, shared between each pair of bacterial genera. For orthogroups conserved in more than two genera, all genus pairs are illustrated individually (e.g., an orthogroup conserved in 3 genera, A, B and C, will show the pairwise association between A-B, B-C and A-C). Orthogroups conserved only in one genus were excluded.

2.4.3. 17 *Pseudomonas*-specific PAGs identified in *P. aeruginosa* PA14 were prioritized for downstream functional analyses

Since PAGs are most likely found among one or two bacterial genera based on the results from the previous section (Chapter 2.4.2), a genus-specific approach to PAGs identification was adopted, with a focus on the *Pseudomonas* genus. PAGs in the

P. aeruginosa PA14 genome were identified using the full *Pseudomonas* genome dataset. In this genus-specific analysis where the pathogen status of genomes was curated at the strain level, a pathogen-associated gene is defined as a gene predicted to be found only in the pathogenic *Pseudomonas* strains within the genome dataset used. The genus-specific PAGs analysis was repeated on a reduced *Pseudomonas* genome dataset (Table 2.3) created for the purpose of performing the downstream OrthoFinder-based orthology analysis within the filesystem limits of the Compute Canada's Cedar server. However, it is important to note that reducing the genome dataset may also reduce the precision of PAGs prediction as genes may be falsely predicted as pathogen associated if some non-pathogen genomes containing those genes are missing from the analysis. Differences between PAGs analysis with or without the subsequent orthology inference as well as between the usage of the full and the reduced *Pseudomonas* genome dataset were assessed (Figure 2.7). Within the *P. aeruginosa* PA14 genome, 330 (77.6%) of 425 PAGs identified by the PAGs analysis were also supported by the orthology analysis using the reduced genome dataset. The higher number of PAGs identified from the reduced dataset may contain genes that are falsely identified as pathogen-associated due to the potential absence of ortholog-encoding genomes from non-pathogenic bacteria that were originally present in the complete genome dataset but excluded in the reduced dataset. For this reason, the 55 PAGs detected by both the PAGs with the complete and reduced genome datasets, as well as the orthology analysis using the reduced genome dataset, were of interest. Seventeen of these genes (Table 2.4) were found in more than 3 genomes within the reduced *Pseudomonas* genome dataset and were selected for more in-depth characterization.

Table 2.3 Genome count of the complete and reduced *Pseudomonas* RefSeq genome dataset.

	Complete dataset	Reduced dataset
Total genomes	605	206
Pathogen genomes	306	61
Non-pathogen genomes	243	145
Species		
P. aeruginosa	220	14
P. chlororaphis	45	9
P. fluorescens	23	18
P. putida	35	21
P. syringae	27	9
P. stutzeri	18	14

Only the top *Pseudomonas* species are shown.

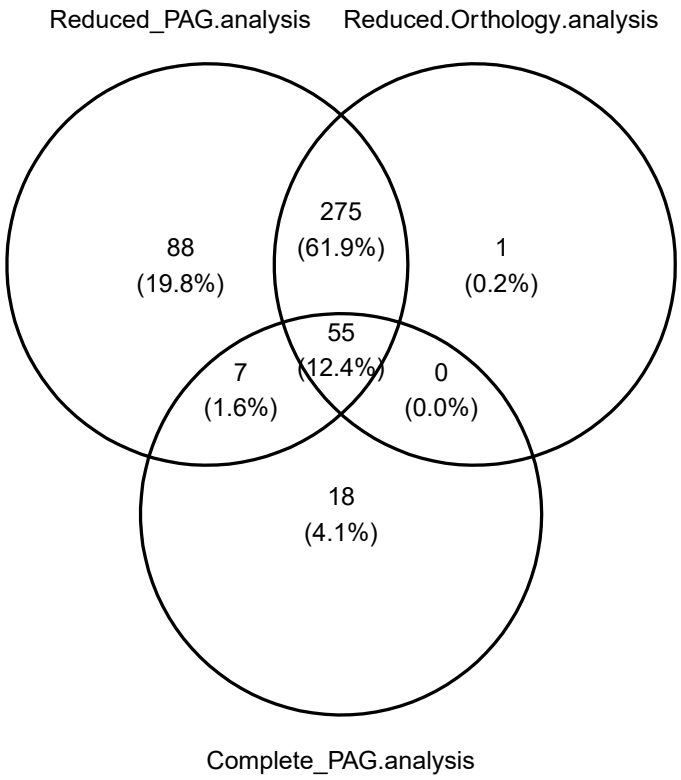


Figure 2.7 Venn diagram of PAGs identified 55 PAGs by the PAGs analysis using the full and reduced *Pseudomonas* genome dataset, as well as the follow-up orthology inference using the reduced dataset.

Numbers without and with parentheses represent the count and percentage of PAGs that were detected by one or more methods and datasets.

Table 2.4 17 PAGs identified in *P. aeruginosa* PA14 and at least two other genomes within the reduced *Pseudomonas* genome dataset.

RefSeq protein accession	Protein name	Protein length (aa)	Ortholog-containing species	Number of genomes in orthogroup	PSORTb SCL
WP_003139757.1[a]	Pentapeptide repeat-containing protein	215	*P. aeruginosa*	14	Extracellular
WP_003141470.1[a]	DUF3218 domain-containing protein	221	*P. aeruginosa*	10	Unknown
WP_003095466.1	Hypothetical protein	248	*P. aeruginosa*	8	Unknown
WP_003116317.1	Class I SAM-dependent methyltransferase	253	*P. aeruginosa; P. mendocina; P. oryzihabitans*	8	Cytoplasmic
WP_003111331.1[b]	Hypothetical protein	128	*P. aeruginosa*	7	Unknown
WP_003140882.1[a]	Hypothetical protein	212	*P. aeruginosa*	7	Cytoplasmic
WP_003118883.1[a]	Hypothetical protein	384	*P. aeruginosa*	6	Cytoplasmic Membrane
WP_025297936.1	Penicillinase repressor	132	*P. aeruginosa*	6	Unknown
WP_016254216.1	M56 family metallopeptidase	331	*P. aeruginosa*	6	Cytoplasmic Membrane
WP_003134054.1[a]	Hypothetical protein	137	*P. aeruginosa*	6	Cytoplasmic
WP_003089506.1[a]	Hypothetical protein	177	*P. aeruginosa*	5	Unknown
WP_003139692.1	Hypothetical protein	486	*P. aeruginosa*	5	Unknown
WP_071535563.1[a]	Amino acid permease	55	*P. aeruginosa*	5	Unknown
WP_003096896.1	Hypothetical protein	78	*P. aeruginosa*	4	Unknown
WP_071534232.1[a]	Response regulator GacA	105	*P. aeruginosa*	4	Extracellular
WP_003140763.1[a]	Hypothetical protein	94	*P. aeruginosa*	4	Unknown
WP_011347386.1	Type I toxin-antitoxin system ptaRNA1 family toxin	76	*P. aeruginosa*	4	Unknown

Pathogen-association was supported by the single-way BLAST analysis and OrthoFinder analysis using the full or the reduced Pseudomonas RefSeq genome datasets.
PSORTb SCL: protein subcellular localization (SCL) site predicted by the PSORTb 3.0 SCL predictor.
[a] PAGs (PAGs) with BLAST hits in only *Pseudomonas* based on the original, all bacterial PAGs analysis
[b] PAG with BLAST hits in *P. aeruginosa* and *Burkholderia multivorans* based on the original, all bacterial PAG analysis.

2.4.4. Detection of positive selection in the PAGs identified in *P. aeruginosa* PA14

Based on the HyPhy selection inference results of the control genes (Table 2.5), a pathogen-associated gene is considered under positive selection if the positive selection is inferred by MEME and at least one other HyPhy selection inference methods. Combining evidence of positive selection from two or more methods minimizes false positive prediction by a single selection inference test. Of the 17 PAGs that are conserved in *P. aeruginosa* PA14 and at least two other *Pseudomonas* genomes, three genes show evidence of positive selection in at least one site in a proportion of lineages in their orthogroup phylogeny, supported by at least two HyPhy selection methods, namely MEME and FEL or aBSREL or BUSTED (Table 2.6).

Table 2.5 Selection inference of negative and positive control genes.

Gene	Gene name	MEME[1]	BUSTED[2]	FEL[3]	aBSREL[4]
Negative controls					
rpoD	RNA polymerase sigma factor RpoD	No	No	No	No
gyrB	DNA gyrase subunit B	No	No	No	Yes
rpoB	DNA-directed RNA polymerase beta chain	No	No	No	No
Positive controls					
fpvA	pyoverdine outer membrane receptor	Yes (15)	No	Yes (2)	Yes
fpvG	iron inner-membrane reductase	Yes (1)	No	No	Yes

[1] MEME (Mixed Effects Model of Evolution): inference method for site-specific episodic positive selection [2] BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification): inference method for gene-wide episodic positive selection
[3] FEL (Fixed Effects Likelihood): inference method for site-specific pervasive positive selection
[4] aBSREL (adaptive Branch-Site Random Effects Likelihood): inference method for branch-specific episodic positive selection
Number in parentheses indicates the number of sites (codons) within a gene with evidence of positive selection, detected by MEME or FEL.

The first *P. aeruginosa* PA14 PAG with evidence of positive selection is the class I S-adenosyl methionine (SAM)-dependent methyltransferase (PA14_RS20430; WP_003116317.1). Significant results from MEME and FEL indicate that this gene has at least one site under pervasive positive selection (i.e., under positive selection across all lineages of the gene phylogeny). With a broad spectrum of substrate specificity, class I SAM-dependent methyltransferases are important for SAM-dependent biochemical processes across all domains of life (Kozbial & Mushegian, 2005), including bacterial pathogenesis. Examples of virulence-related SAM-dependent methyltransferases are

the enteropathogenic *E. coli* T3SS effector Nle which blocks host host NF-κB signaling important for immune defense (Yao et al., 2014) and the *Burkholderia glumae* ToxA responsible for the biosynthesis of toxoflavin, a major *B. glumae* toxin (Kang et al., 2019). Evidence of positive selection and pathogen-association suggests PA14_RS20430 may also play a functional role in the pathogenic lifestyle or host adaptation of *P. aeruginosa*.

Another PAG in *P. aeruginosa* PA14 with evidence of positive selection is a DUF3218 family protein (PA14_RS24305; WP_003141470.1), whose positive selection was also supported by both MEME and FEL. As the gene annotation implies, this gene is yet to be functionally characterized and show minimal homology to any known protein families; however, the NCBI Gene Expression Omnibus suggests that this gene is upregulated in host airways, which may be related to respiratory infection by *P. aeruginosa*. Its pathogen-associated nature as well as the evidence of positive selection indicate potential functional importance in the pathogenesis of *P. aeruginosa* that warrants further investigation.

The M56 family metallopeptidase (PA14_RS12695; PA14_31050; WP_016254216.1) was also predicted to be under episodic positive selection (detected in some but not all lineages of the gene tree) both at the site and gene-level based on combined results from MEME and BUSTED, respectively. The positive selection inference of this gene is further supported by aBSREL in which a proportion of branches on the gene tree were under positive selection. This metallopeptidase is predicted by PSORTb 3.0 (Yu et al., 2010) to localize in the cytoplasmic membrane. Metallopeptidases are a large family of proteolytic enzymes distributed across all kingdoms of life. Pertaining to microbial pathogenesis, some metallopeptidases are VFs associated with host defence depletion and cellular damage, implicated in infection-associated symptoms such as inflammation and pneumonia (Cerda-Costa & Gomis-Ruth, 2014). PA14_RS12695 in *P. aeruginosa* PA14 may be functionally similar to other M56 proteins, including BlaR1 and MecR1 in *S. aureus* that are involved in β-lactam resistance (Dalbey et al., 2012). Based on a multiple sequence alignment with other M56 family proteases using Clustal Omega (Figure 2.8) (Sievers & Higgins, 2014) and a transmembrane helix prediction using TMHMM Server 2.0 (Krogh et al., 2001) respectively, PA14_RS12695 contains the M56 family signature HEXXH zinc-binding motif, between the third and fourth transmembrane domains, followed by an aspartate

five residues downstream of the second histidine in the motif (Figure 2.9) (Rawlings et al., 2018). The function of PA14_RS12695 is likely related to the downstream gene PA14_RS1270 (WP_025297936.1), annotated as a BlaI/MecI/CopY family transcriptional regulator. Both the positively selected PA14_RS12695 and its neighbouring transcriptional regulator PA14_RS1270 are located on a GI predicted by Islander and IslandPick within IslandViewer 4 (Bertelli et al., 2017).

Additionally, at least one site in the hypothetical protein (PA14_33980; WP_003089506.1) and the type I toxin-antitoxin system ptaRNA1 family toxin (PA14_RS06215; WP_011347386.1) was detected by MEME to be under episodic positive selection, though the results are not confirmed by other HyPhy analyses.

Table 2.6 Selection inference of PAGs in *P. aeruginosa* PA14.

Locus tag	NCBI RefSeq protein accession	Gene name	MEME[1]	BUSTED[2]	FEL[3]	aBSREL[4]
PA14_RS13860	WP_003089506.1	hypothetical protein	N/A[6]			
PA14_RS26325	WP_003095466.1	hypothetical protein	Yes (5)	No	No	No
PA14_RS31315	WP_003096896.1	hypothetical protein	N/A[6]			
PA14_RS18450	WP_003111331.1	hypothetical protein	Yes (1)	No	No	No
PA14_RS20430	WP_003116317.1	class I SAM-dependent methyltransferase	Yes (1)	No	Yes (2)	No
PA14_RS02925	WP_003118883.1	hypothetical protein	No	No	No	No
PA14_RS20955	WP_003134054.1	hypothetical protein	N/A[6]			
PA14_RS15515	WP_003139692.1	hypothetical protein	No	No	No	No
PA14_RS15795	WP_003139757.1	pentapeptide repeat-containing protein	No	No	No	No
Record removed	WP_003140763.1	hypothetical protein	No	No	No	No
PA14_RS21825	WP_003140882.1	hypothetical protein	No	No	No	No
PA14_RS24305	WP_003141470.1	DUF3218 family protein	Yes (5)	No	Yes (1)	No
PA14_RS06215	WP_011347386.1	type I toxin-antitoxin system ptaRNA1 family toxin	Yes (1)	No	No	No
PA14_RS12695	WP_016254216.1	M56 family metallopeptidase	Yes (3)	Yes	No	Yes
PA14_RS12700	WP_025297936.1	BlaI/MecI/CopY family transcriptional regulator	No	No	No	No
Record removed[5]	WP_071534232.1	response regulator GacA	No	No	No	No

Locus tag	NCBI RefSeq protein accession	Gene name	MEME[1]	BUSTED[2]	FEL[3]	aBSREL[4]
Record removed[5]	WP_071535563.1	amino acid permease	No	No	No	No

[1] MEME (Mixed Effects Model of Evolution): inference method for site-specific episodic positive selection [2] BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification): inference method for gene-wide episodic positive selection

[3] FEL (Fixed Effects Likelihood): inference method for site-specific pervasive positive selection

[4] aBSREL (adaptive Branch-Site Random Effects Likelihood): inference method for branch-specific episodic positive selection

[5] The NCBI RefSeq protein accession, previously in *P. aeruginosa* UCBPP-PA14 genome assembly (GCF_000014625.1), is no longer annotated on any genomes

[6] No selection inference performed due to the presence of less than three unique sequences in the gene orthogroup. Number in parentheses indicates the number of sites (codons) within a gene with evidence of positive selection, detected by MEME or FEL.

Figure 2.8 Multiple sequence alignment of the conserved HEXXH zinc-binding motif followed by an aspartate (D) five residues downstream of the second histitine (H), found in M56 family proteins, in the the positively selected and pathogen-associated M56 family metallopeptidase.

The metallopeptidase of interest (PA14_RS12695; WP_016254216.1) shown on top was aligned to six M56 family proteins. Multiple sequence alignment was performed by Clustal Omega v1.2.4

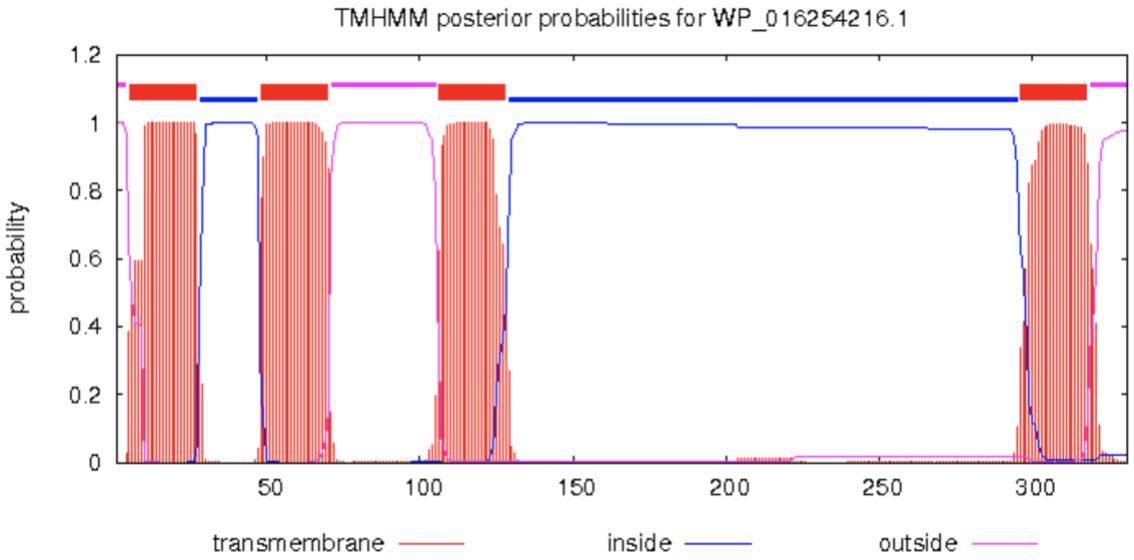


Figure 2.9 TMHMM transmembrane helices predictor further validated that the metallopeptidase of interest (PA14_RS12695; WP_016254216.1) contains the M56 family HEXXH zinc-binding motif between the third and the fourth predicted trasmembrane domains.

TMHMM v2.0 was used for transmembrane helices prediction.

2.4.5. Detection of positive selection in T3SS genes of clinically important human pathogens

Since T3SS has been widely reported to be used for the delivery of bacterial VFs, twenty structural T3SS genes and a small set of T3SS transcriptional regulators, chaperones and effectors were analyzed for positive selection under the same combination of HyPhy methods described above (Table 2.7). In the set of structural T3SS genes, PcrD (SctV) in the T3SS inner membrane machinery showed strong evidence of positive selection in four HyPhy methods, MEME, BUSTED, FEL and aBSREL. PcrD is thought to form the entrance of the translocation channel in the inner membrane machinery, located above the ATPase complex and below the secretion pore (Deng et al., 2017; Wagner et al., 2018). PcrD/SctV is not an essential component in some T3SS, but may function to lift the helical SctRST complex to support the interaction with the inner membrane ring-forming protein PscJ/SctJ of the basal body that also showed evidence of gene-wide and site-specific episodic positive selection, according to BUSTED and MEME results (Wagner et al., 2018). Though PcrD/SctV is not an essential component of the T3SS, the inferred positive selection suggests that PcrD may be important for effector secretion.

Another positively selected T3SS component is the translocator PopD (SctB) which, together with PopB, forms the translocation pore of the T3SS injectosome and contains transmembrane segments that insert into the host membrane (Wagner et al., 2018). Positive selection in PopD was supported by BUSTED, MEME as well as FEL. Since PopD directly interacts with host cell membrane, mutations may confer selective advantage in the adhesion and invasion of *P. aeruginosa* into host cells.

The only non-structural T3SS gene detected for positive selection is the cytotoxin, exoenzyme T (ExoT) in which at least one site showed evidence of pervasive positive selection and a few others for episodic positive selection. ExoT, which is also pathogen associated, confers GTPase activating and ADP ribosyltransferase activities that alter host cytoskeleton, therefore blocking host phagocytosis during pathogenesis (Hauser, 2009). Of the four most studied *P. aeruginosa* effectors (ExoS, ExoT, ExoU and ExoY), HyPhy selection inference was only done on ExoT, ExoU and ExoY as ExoS is not present in the PA14 strain (Wareham et al., 2005). Notably, ExoT is the only known *P. aeruginosa* effector that is present in nearly all clinical and environmental isolates (Feltman et al., 2001).

Table 2.7 Selection inference of Type III secretion system genes.

Gene	Gene Description	NCBI RefSeq protein accession	MEME[1]	BUSTED[2]	FEL[3]	aBSREL[4]
Basal body						
PscC (SctC)	Secretin	WP_003140040.1	No	No	No	No
PscD (SctD)	Major IM ring component	WP_003132859.1	No	No	No	Yes
PscJ (SctJ)	Lipoprotein ring component	WP_003120329.1	Yes (1)	Yes	No	No
Inner rod						
PscI (SctI)	Inner rod	WP_003120330.1	No	No	No	No
Needle filament						
PscF (SctF)	Needle component	WP_003087729.1	No	No	No	No
Inner membrane machinery						
PcrD (SctV)	Major component, external to channel	WP_003087696.1	Yes (5)	Yes	Yes (7)	Yes
PscR (SctR)	Translocase channel	WP_003087674.1	No	Yes	No	No
PscS (SctS)	Translocase channel	WP_003087672.1	No	No	No	Yes
PscT (SctT)	Translocase channel	WP_003132884.1	No	Yes	No	No
PscU (SctU)	Minor component, external to channel	WP_003087668.1	Yes (3)	No	No	No
Needle tip and translocon						
PopB (SctE)	Translocon pore protein (secreted)	WP_003087710.1	No	No	Yes (1)	No
PopD (SctB)	Translocon pore protein (secreted)	WP_003109504.1	Yes (1)	Yes	Yes (6)	No
PcrV (SctA)	Tip/filament protein (secreted)	WP_003109502.1	Yes (1)	No	No	No
ATPase						
PscN (SctN)	Hexameric ring-structure ATPase	WP_003100796.1	No	No	No	Yes
Coiled coil linker						
PscO (SctO)	Central stalk, inserting in ATPase ring	WP_003140046.1	No	No	No	No
Sorting platform						
PscQ (SctQ)	Component of 6 Pod assembly	WP_003140061.1	No	No	Yes (1)	No
PscK (SctK)	Connector of Pods with SctD	WP_003087734.1	No	No	No	No

Gene	Gene Description	NCBI RefSeq protein accession	MEME[1]	BUSTED[2]	FEL[3]	aBSREL[4]
PscL (SctL)	External stator, connecting ATPase and Cytoplasmic ring	WP_003087735.1	No	No	No	No
Needle length regulator						
PscP (SctP)	Molecular ruler regulating needle/filament length	WP_011666708.1	No	No	Yes (1)	No
Export regulator						
PopN (SctW)	Gatekeeper/Affinity switch (secreted)	WP_003087692.1	No	No	No	No
Effectors/toxins						
ExoU	Acute cytotoxin; phospholipase	WP_003134060.1	No	No	No	No
ExoT	Toxin	WP_003136948.1	Yes (6)	No	Yes (1)	No
ExoY	adenylate cyclase	WP_011666674.1	No	No	No	Yes
Transcriptional regulators						
ExsA	activator of type III gene transcription	WP_003120334.1	No	No	No	No
ExsD	binds ExsA to inhibit transcription	WP_003109509.1	No	No	No	No
ExsC	inhibits ExsD activity	WP_003109505.1	No	No	No	No
ExsE	inhibits ExsC activity	WP_003109506.1	No	No	No	No
Chaperones						
PscB	Chaperone for PopN[5]	WP_003109510.1	No	No	No	No
PscG	Chaperone for PscF	WP_003140037.1	No	No	No	No
PscE	Chaperone for PscF	WP_003100751.1	No	No	No	No
PscH	YopR family T3SS polymerization control protein	WP_003100725.1	No	No	No	No
Pilotin						
ExsB	Regulator of T3SS apparatus assembly	WP_003117721.1	No	No	No	No

[1] MEME (Mixed Effects Model of Evolution): inference method for site-specific episodic positive selection [2] BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification): inference method for gene-wide episodic positive selection

[3] FEL (Fixed Effects Likelihood): inference method for site-specific pervasive positive selection

[4] aBSREL (adaptive Branch-Site Random Effects Likelihood): inference method for branch-specific episodic positive selection

[5] Gene description is based on homolog YscB's interaction with YopB in Yersinia (H. Yang et al., 2007)

"SctX" in parentheses in the "Gene" field indicates the Sct nomenclature (Notti & Stebbins, 2016), if available. Number in parentheses in the "MEME" and "FEL" fields indicates the number of sites (codons) within a gene with evidence of positive selection, detected by MEME or FEL, respectively.

2.4.6. Modifying the PAGs analysis parameter to allow PAGs be found in a few non-pathogen genomes may improve the identification of PAGs, while accommodating for the potential errors in taxonomic classification of novel genomes

Within the *P. aeruginosa* PA14 genome (GCF_000014625.1), common genes, found in both pathogen and non-pathogen genomes within the *Pseudomonas* genus, were reanalyzed for their pathogen-association. Specifically, for each common gene, the genomes of origin were categorized as pathogen or non-pathogen and the total genomes within each category. Fifty seven genomes with no available information on pathogen status, from the literature, were ignored. Common genes with at least 90% of the putative homologs in pathogen genomes were selected for visualization in a presence/absence matrix in combination with a species tree of all *Pseudomonas* genomes (not shown due to the large figure size).

During this re-analysis, the NCBI RefSeq assembly of the non-pathogen *Pseudomonas* sp. AK6U (GCF_002843285.1) was suppressed due to sequence quality issue (many frameshifted proteins). This genome was removed from this analysis, which updated the pathogen association of 14 genes from being a common gene to a pathogen-associated gene as none of their predicted homologs are now found in a non-pathogen genome (Table 2.8). While the majority encodes hypothetical proteins, the few genes with functional annotations include the quorum sensing gene expression regulator QteE (Siehnel et al., 2010), an exo-alpha-sialidase important for sialic acid-dependent host invasion (Kiyohara et al., 2011), and a GNAT family N-acetyltransferase whose protein family is involved in a wide range of bacterial functions including antibiotic resistance (aminoglycoside acetyltransferases) and peptidoglycan synthesis (FemABX aminoacyl transferase) (Siehnel et al., 2010).

Table 2.8 Common genes that became pathogen-associated after genome assembly supression in the RefSeq database.

RefSeq protein accession	Name
WP_003088334.1	hypothetical protein
WP_003090369.1	quorum threshold expression protein QteE
WP_003111044.1	GNAT family N-acetyltransferase
WP_003111314.1	hypothetical protein
WP_003119164.1	hypothetical protein
WP_003120198.1	methyltransferase domain-containing protein

WP_003136949.1	hypothetical protein
WP_003138781.1	exo-alpha-sialidase
WP_003139695.1	DUF1127 domain-containing protein
WP_003139696.1	hypothetical protein
WP_010791699.1	hypothetical protein
WP_031632737.1	hypothetical protein
WP_123794454.1	hypothetical protein
WP_123903201.1	hypothetical protein

Allowing up to 1% of the predicted homologs of gene to be found in non-pathogen genomes led 272 common genes to become pathogen-associated. At most two non-pathogen genomes were detected under this 1% criterion. Many genes that fit within this criterion are known bacterial VFs, including many T3SS components and effectors, other transport proteins, and metallophores (for metal homeostasis in host environment). Fifty three of these genes are found in all *P. aeruginosa* isolates within the reduced *Pseudomonas* genome dataset (Table 2.9). Many of these functionally annotated genes are virulence related. Examples include the mucoid induction factor MucE (biofilm formation), phenazine-1-carboxylate N-methyltransferase PhzM (biosynthesis of phenazine VFs), and LasA elastase (tissue damage). Three notable genes, now considered as pathogen-associated, appeared in multiple *P. aeruginosa* and *P. stutzeri* genomes. One of them is a copper resistance (CopD) family protein (WP_003141586.1) which is crucial for copper tolerance in bacteria, as a defence mechanism against the host inflammatory response, in which serum copper is elevated and supplied to the phagolysosomal compartments of macrophages to eliminate bacterial pathogens (Ladomersky & Petris, 2015). Another bacterial gene now considered as pathogen-associated and conserved in multiple *P. aeruginosa* and *P. stutzeri* is the MotA/TolQ/ExbB proton channel family protein (WP_003088234.1) which belongs to a family of outer membrane receptor energizer including the TonB-dependent transporters for siderophore uptake (Kuehl & Crosa, 2010). A DUF2149 domain-containing protein (WP_003088238.1) also had a similar pathogen-associated conservation in *P. aeruginosa* and *P. stutzeri*. As the criterion was further relaxed to allow for up to 2% of the predicted homologs to be found in non-pathogen genomes. An additional 180 genes changed from common to pathogen-associated; however, it was unclear whether these genes may be involved in virulence as the majority of them did not have a known function.

Due to the ongoing addition of novel bacterial genomes to the RefSeq database, some previously identified PAGs such as the PopB translocator gene of the *Pseudomonas* T3SS were no longer detected as pathogen-associated, due to their significant sequence similarity to genes in novel non-pathogen genomes in the more recent PAG analyses. In the case of PopB, it was detected in the novel *P. fluorescens* NCTC10783 strain, whose genome is more similar to the pathogenic *P. aeruginosa* group than to the non-pathogenic *P. fluorescens* group. PopB was also found in two novel environmental species, *P. sp.* CCOS 191 and *P. soli* SJ10, which shares high sequence similarity to the pathogenic *P. putida* group (Figure 2.10). Since there are no available data on the characteristics, particularly virulence, of these novel isolates, there is a possibility that these genomes were inaccurately classified into a species. In future PAGs analysis updates, relaxing the criteria for pathogen-associated gene detection to include up to 1% of BLAST hits to be found in non-pathogens may improve the prediction of bacterial genes that are uniquely associated with pathogens. Since the PAGs analysis is considered as the initial step in the prioritization strategy of potentially novel VFs, the virulence-related role of any identified PAGs still need to be validated with subsequent functional analyses.

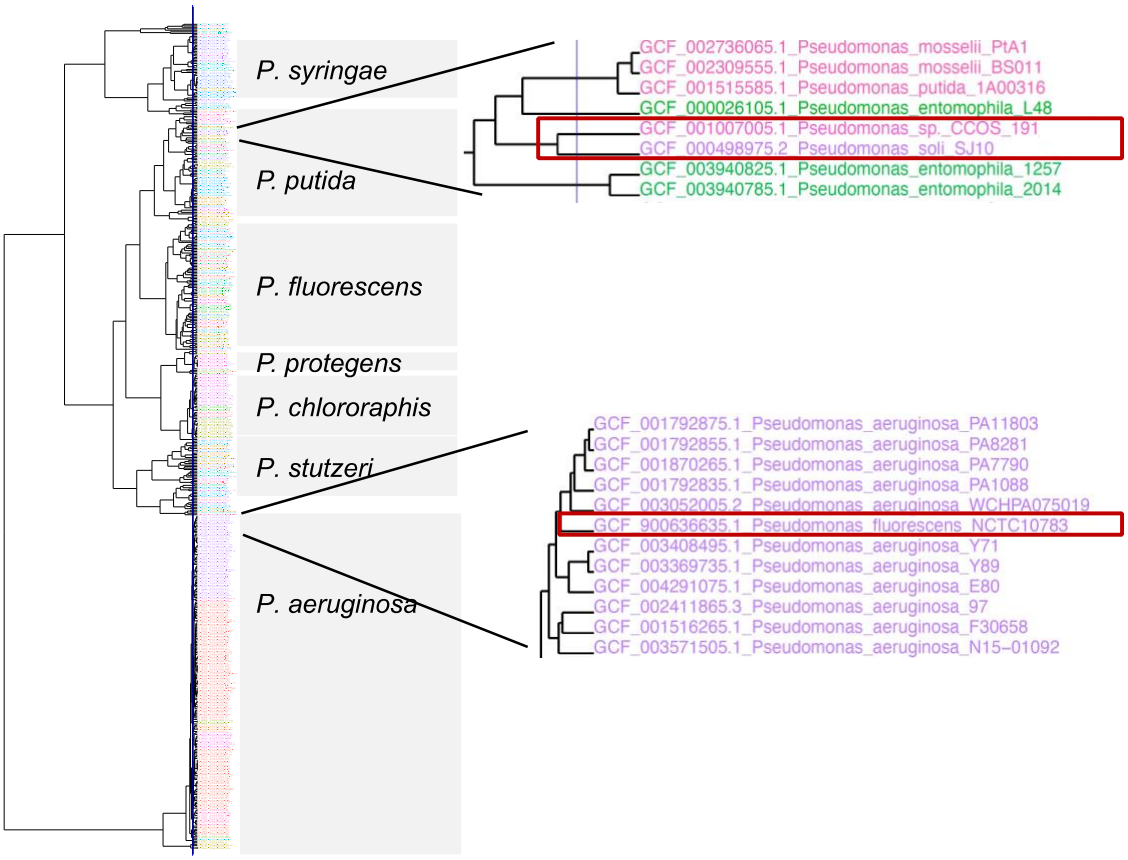


Figure 2.10 A simplified dendrogram of *Pseudomonas* genomes, constructed from MASH distance matrix shows that a novel *P. fluorescens* (commonly known as a non-pathogen) strain NCTC10783 and two novel environmental strains *P. sp*. CCOS 191 and *P. soli* SJ10 are clustered, by sequence similarity, to the pathogenic *P. aeruginosa* and *P. putida* genomes.

Colours on dendrogram represents the clusters of *Pseudomonas* defined by the 0.1 MASH distance cut-off. The red boxes capture the novel non-pathogen genomes that are highly similar to the genomes within the pathogenic *P. putida* and *P. aeruginosa* groups.

Table 2.9 PAGs with up to 1% of their DIAMOND BLAST hits in non-pathogen genomes.

NCBI RefSeq Protein Accession	Gene Name
Novel PAGs detected due to NCBI's recent removal of low-quality non-pathogen genomes	
WP_003088334.1	hypothetical protein
WP_003090369.1	quorum threshold expression protein QteE
WP_003111044.1	GNAT family N-acetyltransferase
WP_003111314.1	hypothetical protein
WP_003119164.1	hypothetical protein
WP_003120198.1	methyltransferase domain-containing protein
WP_003136949.1	hypothetical protein
WP_003138781.1	exo-alpha-sialidase
WP_003139695.1	DUF1127 domain-containing protein
WP_003139696.1	hypothetical protein
WP_010791699.1	hypothetical protein
WP_031632737.1	hypothetical protein
WP_123794454.1	hypothetical protein
WP_123903201.1	hypothetical protein
Novel PAGs with up to 1% of DIAMOND blast hits in non-pathogen genomes	
WP_003082488.1	DUF1857 family protein
WP_003082734.1	hypothetical protein
WP_003083879.1	hypothetical protein
WP_003084489.1	GNAT family N-acetyltransferase
WP_003085474.1	hypothetical protein
WP_003085997.1	hypothetical protein
WP_003086013.1	hypothetical protein
WP_003087555.1	DUF5086 domain-containing protein
WP_003088027.1	type VI secretion system effector peptidoglycanhydrolase Tse1
WP_003088087.1	hypothetical protein
WP_003088145.1	hypothetical protein
WP_003089862.1	hypothetical protein
WP_003090022.1	hypothetical protein
WP_003090024.1	hypothetical protein
WP_003090566.1	hypothetical protein
WP_003090681.1	hypothetical protein
WP_003090686.1	hypothetical protein
WP_003090820.1	hypothetical protein
WP_003090821.1	hypothetical protein
WP_003091038.1	hypothetical protein
WP_003091768.1	hypothetical protein
WP_003092748.1	DUF4345 domain-containing protein
WP_003093256.1	mucoid induction factor MucE

NCBI RefSeq Protein Accession	Gene Name
WP_003093617.1	phenazine-1-carboxylate N-methyltransferase PhzM
WP_003095356.1	pseudopaline biosynthesis dehydrogenase CntM
WP_003095359.1	pseudopaline biosynthesis protein CntL
WP_003095507.1	hypothetical protein
WP_003097848.1	K(+)-transporting ATPase subunit F
WP_003102243.1	hypothetical protein
WP_003102692.1	sterol desaturase family protein
WP_003106166.1	suppressor of fused domain protein
WP_003109295.1	hypothetical protein
WP_003110435.1	Vicinal oxygen chelate (VOC) family protein
WP_003118870.1	hypothetical protein
WP_003130258.1	hypothetical protein
WP_003137358.1	GatB/YqeY domain-containing protein
WP_003138008.1	hypothetical protein
WP_003138107.1	hypothetical protein
WP_003138171.1	type VI secretion system effector muramidase Tse3
WP_003138256.1	fucose-binding lectin PA-IIL
WP_003138700.1	hypothetical protein
WP_003139897.1	protease LasA
WP_003139919.1	hypothetical protein
WP_003140206.1	hypothetical protein
WP_003140628.1	hypothetical protein
WP_003142074.1	hypothetical protein
WP_003142075.1	hypothetical protein
WP_004365177.1	YfiM family lipoprotein
WP_004365414.1	GtrA family protein

2.5. Discussion

This chapter presents an update of the PAGs analysis with over 8600 bacterial genomes. A few changes have been implemented to improve the scalability of the analysis as the number of genomes grows exponentially. In this update, the pathogenic status of all genomes was manually curated at the species level rather than at the strain level as done in previous analyses. Since this curation was historically done through manual literature search of the strain of each genome individually (Dhillon et al., 2015; Fedynak, 2007; S. J. Ho Sui et al., 2009; S. J. Ho Sui et al., 2012), this was not practical with the substantial expansion of the genome dataset. The distinction between pathogenic and non-pathogenic strains within a few species was traded off for

completing the pathogen status curation in a manageable amount of time. While most species contain strains that are universally pathogenic or non-pathogenic, few species such as *E. coli* and *Salmonella enterica* include non-pathogenic strains (e.g., *E. coli* str. K-12 and *S. enterica* subsp. *enterica* serovar Typhi str. Ty21a) despite being a predominantly pathogenic species. This change must be taken into consideration when interpreting the predicted pathogen association of genes belonging to species with mixed pathogenicity, as some genes may be identified as pathogen-associated if they are present in non-pathogenic strains that have been generalized as a "pathogen" in the analysis. It is important to note that the pathogen status of the genomes may change in future PAGs updates when more published data on the lifestyle and pathogenicity become available for the novel bacterial species. Determining whether a bacterium is pathogenic may sometimes be challenging due to context dependency of a pathogen. Some bacteria are pathogenic only under specific conditions that may or may not be well studied. While disease-causing capability exists as a spectrum among bacteria, such as those that are opportunistic pathogens under specific environmental/host conditions, this analysis considered a species to be pathogenic if any evidence of virulence towards any eukaryotic organisms has been documented. Another change in the PAGs analysis is the replacement of the protein sequence alignment search tool from BLAST to DIAMOND for better handling of the large genome dataset and reducing the analysis runtime (Altschul et al., 1990; Buchfink et al., 2021; Buchfink et al., 2015). DIAMOND reduces the runtime while maintaining a similar accuracy as BLAST with the following features: double indexing of both query and reference sequences, spaced seeds (the use of a subset of positions within a subsequence for comparison between the query and reference sequences) and the use of a reduced alphabet (grouping of similar amino acids). As more genomes are added to the analysis, the time and the computational resources to perform the all-against-all sequence similarity search will also rise exponentially.

Moving forward, scalable modifications to the PAGs analysis algorithm are needed to keep up with the growing number of bacterial genomes. The pathogen status curation was a major bottleneck in the analysis update and should be ideally automated in the future to ensure consistency in literature interpretation as well as to potentially re-introduce the more refined strain-level pathogen curation from previous analyses. From a technical standpoint, better data handling and storage are also needed to utilize the

available computational resources more efficiently for down analyses. For reference, 23 terabytes of data were generated from the 2018 PAGs analysis. From the *Pseudomonas* genome dataset, the allowance of up to 1% of a gene's sequence similarity hits to be found in non-pathogen genomes was shown to be effective in minimizing the false negatives in the identification of bacterial genes that are supposedly pathogen-associated but missed due to their presence in bacterial genomes that have been misclassified in non-pathogenic taxa. As bacterial genomic research advances, taxonomic classification transitioned from the laboratory-based gold standard of DNA-DNA hybridization for species delineation to the use of higher-resolution, genome-based approaches such as average nucleotide identity, phylogenetic distances and taxon-specific genes (Barco et al., 2020; Gupta & Sharma, 2015; Hugenholtz et al., 2021). These genome-based methods show that while most bacterial taxa are monophyletic, several taxa at the species, genus and even up to the phylum level required revision of their taxonomic assignment due to incongruence in their phenotypic and phylogenetic features (Nouioui et al., 2018). As such, some novel genomes may belong to bacterial pathogens but have insufficient data on pathogenicity, have been incorrectly assigned to a taxon or mis-classified as non-pathogens. Implementation of more lenient criteria for detecting PAGs may improve the recall of the algorithm. Other aspects of improvement may include the use of a phylogenetic-based based method for assessing the extent of gene conservation across the large diversity of bacteria. Bacterial genera exhibit a wide spectrum of diversity. Some genera like *Pseudomonas* and *Escherichia* may have more intra-genus genomic variations than others.

In comparison to genes conserved in both bacterial pathogens and non-pathogens, PAGs and non-PAGs are disproportionately uncharacterized, evident from the high prevalence of genes annotated as "hypothetical protein" or "domain of unknown function," and the lack of sequence similarity to protein families curated to date. These genes are often homologous to other functionally undefined genes found across multiple bacterial strains or species, making sequence similarity-based functional prediction difficult without the incorporation of other gene information, such as their genomic context, gene expression or phenotypic data. Based on the unique conservation of these genes in either pathogens or non-pathogens, some of these genes may be involved in biological processes important for the adaptation to diverse bacterial lifestyles but yet to be characterized. The limited taxonomic distribution of most PAGs and non-PAGs in one

to two bacterial genera further suggest that these genes may confer more specialized functions for fine tuning the organism's adaptation to a particular environment. Consistent with a previously reported trend that pathogen-associated VFs from the Virulence Factor Database are enriched in "offensive" functions such as host invasion and toxin secretion (S. J. Ho Sui et al., 2009), PAGs homologous to known VFs are also disproportionately related to bacterial secretion systems, secreted genes and toxins. Novel VFs and mechanisms may therefore lie within the set of PAGs whose biological role, potentially in pathogenesis, has not been explored. Increasing research efforts have been focused on the discovery and characterization of novel bacterial VFs from hypothetical proteins in clinically important pathogens such as *Shigella flexneri*, *Mycobacteirum tuberculosis, Klebsiella pneumoniae* as well as *P. aeruginosa* (Galperin & Koonin, 2004; Pranavathiyani et al., 2020; Reem et al., 2021; Sen & Verma, 2020; Shahbaaz et al., 2016; Z. Yang et al., 2019). This PAGs analysis provides a comprehensive platform for identifying novel VFs not only from previously uncharacterized genes, but also prioritizing those that are pathogen-specific to be used in precise antivirulence drug development.

Evolutionary selection inference was used to computationally prioritizing putative virulence related PAGs for laboratory functional analyses. Shifting from a commensal to a pathogenic lifestyle requires bacterial adaptation to a new niche. Novel environments drive evolution of the organisms by imposing selective pressure on traits beneficial for survival such as protection against host immune defence, nutrient uptake, resource competition with other microbial colonizers. Genes capable of increasing the evolutionary fitness of pathogenic bacteria including host-to-host transmission are thus more likely involved in processes related to pathogenesis and disease progression in host (Wickham et al., 2007). In genetically diverse pathogens like *P. aeruginosa* which can colonize a wide range of ecological niches (e.g., infects, fungi, worms plants and mammals), identifying genes, including their functions, that are positively selected for in host adaptation as well as AMR is crucial for the development of effective therapeutics for treating bacterial infections. Several studies have identified genes related to metabolism, energy production and stress-response from the *P. aeruginosa* core as well as accessory genome to be under positive selection during CF lung colonization (Dettman & Kassen, 2021; E. E. Smith et al., 2006; Winstanley et al., 2016). Notable from the evolutionary selection analysis presented in this chapter, specific components

of the *P. aeruginosa* T3SS, namely the inner membrane component PcrD and translocator gene PopD, that come into direct contact with effectors that modulate host interactions were also detected to be under positive selection. It was also not surprising that one of the main *P. aeruginosa* VFs, exotoxin ExoT, also showed evidence of positive selection, suggesting that these genes may play an important role in the host-pathogen interaction of *P. aeruginosa.* Likewise, the inferred positive selection in the few PAGs detected in *P. aeruginosa* PA14 warrants further study into the potential involvement in virulence of these genes, which may be used as pathogen-specific targets for novel antivirulence drug development.

In summary, the update and refined PAGs analysis with orthology and positive selection inference methods presented in this chapter may serve as the preliminary, large-scale, comparative genomic screening strategy to identify potentially novel virulence-related genes from the growing collection of publicly available bacterial genomes. A global analysis of the functional annotation and virulence association shows that PAGs are disproportionately annotated as “hypothetical proteins” with no defined functions and those associated with known VFs are enriched in functions related to the bacterial secretion systems, secreted proteins and toxins. Despite the initial goal to functionally explore PAGs that widely conserved across different bacterial genera, I observed that most genes that are unique to either pathogens or non-pathogens (PAGs or non-PAGs) are only found in one or two genera. The limited taxonomic conservation of PAGs may be beneficial for the development of precise, pathogen-specific antivirulence therapeutics, which may address the non-selective killing of existing antibiotics and may therefore minimize the selective pressure for AMR development in the human microbiota. This set of *in silico* analyses helped to prioritize 17 PAGs found in the hypervirulent *P. aeruginosa* PA14 strain for more in-depth functional characterization through both computational prediction and laboratory validation. Although the work in this thesis focused on *P. aeruginosa*, the preliminary PAGs screening and characterization methods described in this chapter are also applicable to other bacterial pathogens. PAGs may thus represent an untapped repertoire of pathogen-specific and potentially virulence-related genes for uncovering novel bacterial virulence determinants and identifying candidate targets for antivirulence drug development.

Chapter 3.

Further prioritization of PAGs as candidate antivirulence drug targets

This chapter begins with a protein subcellular localization (SCL) analysis of all bacterial genes by pathogen association as predicted in Chapter 2, for the purpose of identifying SCL trends in PAGs and prioritizing more drug accessible PAGs as potential antivirulence drug targets. As a part of this SCL analysis, I also led the update of the PSORTdb 4.0 protein SCL database and the optimization of the PSORTm protein SCL predictor for metagenomic sequences. I tested for software bugs and helped troubleshoot issues related to PSORTdb and PSORTm. I wrote and published a manuscript for both PSORTdb 4.0 (Lau et al., 2021) and PSORTm (Peabody et al., 2020) as first and first co-author, respectively. The second part of this chapter presents an in vivo virulence screening assay of the subset of PAGs prioritized from P. aeruginosa PA14 in Chapter 2. The laboratory work was performed by Dr. Patrick Taylor. The subsequent Kaplan-Meier survival analysis and visualization were done by me. The final part of this chapter focuses on the gene mobility analysis of a pair of PAGs of interest and their associated GI in P. aeruginosa isolates.

I completed all work presented in this chapter with the following exceptions: Dr. Mike Peabody and Gemma Hoad led the initial development and set-up of the PSORTm Docker image; Justin Jia assisted in the benchmarking work of PSORTm. Gemma Hoad and Vivian Jin implemented the PSORTdb 4.0 web update. Dr. Patrick Taylor performed the in vivo screening of the P. aeruginosa PA14 transposon mutants of the select PAGs.

3.1. Abstract

The application of PAGs as potential antivirulence drug targets requires an in-depth characterization of their encoding proteins. Features such as protein SCL and molecular functions of the PAGs may be useful in the prioritization of drug targets and the identification of antivirulence drug candidates. The first part of this study compares the global trend in protein SCL encoded by PAGs, non-PAGs as well common genes (found in both pathogens and non-pathogens). Genes predicted to be conserved in only pathogens or non-pathogens disproportionately encode proteins in the more drug-accessible localizations, cell wall and extracellular space. For antivirulence drug target prioritization, a subset of PAGs in *P. aeruginosa* PA14 were screened for virulence-modulating activity using the corresponding *P. aeruginosa* PA14 transposon insertion mutants in a *C. elegans* infection model. Of 11 screened PAGs, 6 PAGs showed direct virulence-enhancing activity (mutants improved worm survival) while 1 PAG, a BlaI/MecI/CopY family transcriptional regulator (PA14_RS12700), showed virulence-repressing activity (mutant enhanced worm killing). A subsequent gene presence-absence analysis of a 43kb *P. aeruginosa* GI carrying PA14_RS12700 as well as the adjacent PA14_RS12695, a pathogen-associated M56 family metallopeptidase with evidence of positive selection, revealed that these two PAGs as well as the entire gene set of this GI is associated with multiple multidrug resistant *P. aeruginosa* strains of global concern. While the *C. elegans*-based virulence screening of transposon mutants represents a quick and effective method for identifying PAGs with putative virulence functions, additional laboratory analyses are required to characterize the exact biological function of these PAGs.

3.2. Introduction

The development of a novel antivirulence drug requires a collection of information on traits such as a putative bacterial target's localization within or external to the bacterial cell, functional role in pathogenesis and prevalence in clinically important isolates. Despite constant advances in genome sequencing and analysis approaches, computational predictions must still be complemented with laboratory assessments to validate the function and features of novel genes. Determining the SCL site at which protein targets reside is an essential step to designing drugs that can effectively reach

their cellular targets and elicit its therapeutic effects. Membrane trafficking and subcellular targeting govern the pharmacokinetics (how the organism affects the drug) and the pharmacodynamics (how the drug affects the organism) that are important for proper drug-target interactions. Translocation of drugs across the bacterial cell envelope is often a challenge in the delivery of antimicrobials to their intracellular targets. AMR is notoriously common in gram-negative bacteria like *P. aeruginosa* owing to poor permeability and abundant efflux transporters associated with the additional outer membrane that is generally absent in gram-positive bacteria (Breijyeh et al., 2020).

Subcellular fractionation has been a widely used method for isolating and identifying proteins residing in specific cellular compartments; however, this laboratory-based method is time-consuming and prone to cross-contamination of proteins from neighbouring compartments (Gardy & Brinkman, 2006). Computational methods for protein SCL prediction have since been developed to complement the previous approach and to enable a more high-throughput SCL analysis of novel protein sequences as the number of available bacterial genomes increases (Rey, Gardy, et al., 2005). The PSORTb bacterial and archaeal protein SCL predictor, developed by the Brinkman Lab, is the most precise tool of its kind, featuring a multi-component approach to predict where a protein is localized based on its protein sequence (Gardy & Brinkman, 2006; Gardy et al., 2005; Gardy et al., 2003; Yu et al., 2010). PSORTb version 3.0 and onwards includes protein SCL prediction for the typical gram-positive and gram-negative organisms, and organisms with a non-conventional cell envelope structure such as gram-negative bacteria without an outer membrane (e.g., *Mycoplasma spp.*) or gram-positive bacteria with an outer membrane (e.g., *Deinococcus spp*.). Protein SCL predictors like PSORTb can help researchers gain insight into the functional role of novel genes based on the localization sites of the proteins they encode.

In addition to protein SCL, genomic localization of the corresponding gene may also suggest its functional significance based on gene mobility. Pathogenicity islands (PAIs), usually carrying one or more virulence-related genes of similar function, enhance virulence in multiple human pathogens. Well known PAIs include VP-1 PAI in *Vibrio cholera*e for enabling the drastic change from avirulent to virulent phenotype via viral transduction, SP-1 and SP-2 in *Salmonella enterica* for adaptation to intracellular environment, and the locus of enterocyte effacement (LEE) PAI in enteropathogenic *E. coli* for establishing adherence to intestinal epithelial cells (Gal-Mor & Finlay, 2006). VFs,

especially those that play a more offensive role during host interaction, are disproportionately more prevalent in GIs than in chromosomes (S. J. Ho Sui et al., 2009). Using bioinformatics tools like IslandViewer 4 (Bertelli et al., 2017), GIs with their gene content can be predicted from their bacterial genomes of origin, typically based on their atypical sequence composition bias and sporadic phylogenetic distribution compared to the rest of the genomes (Bertelli et al., 2019)

While bioinformatics analyses may predict the biological function of a gene from its sequence, examining the functional role of a gene under laboratory setting is still the gold standard for gene function validation. In this chapter, virulence screening in a *C. elegans* nematode infection model was the chosen *in vivo* method for the preliminary characterization of PAGs in *P. aeruginosa* PA14. The biological role of PAGs was investigated by examining the impact on the nematode survival upon infection by a mutant strain of *P. aeruginosa*, in which the wild-type function of the gene is disrupted by transposon insertion mutagenesis. The Ausubel Lab has constructed a non-redundant transposon mutant library of *P. aeruginosa* PA14 mutant strains with each carrying a unique transposon insertion. Consisting of approximately 80% of the non-essential open reading frames in *P.* aeruginosa PA14 (Liberati et al., 2006), this transposon mutant library is useful for screening virulence phenotype in the yet-to-be characterized PAGs, given that the corresponding mutant is readily available. *C. elegans* is a well-studied nematode model for studying the pathogenesis of *P. aeruginosa* (Tan, Mahajan-Miklos, et al., 1999; Tan, Rahme, et al., 1999). In comparison to other model hosts such as mouse, thale cress (*Arabidopsis thaliana*) and fruit fly (*Drosophila melanogaster*), *C. elegans* offers many advantages such as having a short life cycle of 3 days (from embryo to adult), ease of propagation (300 progeny per hermaphrodite adult) and small size. It has an ancestral immune system capable of eliciting antimicrobial responses against intestinal infection by microbial pathogens (Balla & Troemel, 2013). Though the intestinal infection of *C. elegans* does not reflect the natural site of infection by *P. aeruginosa* (i.e., predominantly in human lung, skin and blood) the slow killing of *C. elegans* under low osmolarity minimal media resembles the infection process and been widely used with transposon mutants to identify novel virulence-related genes in *P. aeruginosa*, particularly in the PA14 strain.

This chapter extends the functional characterization of PAGs detected in Chapter 2. First, I led the update of the PSORTdb bacterial and archaeal protein SCL database

and conducted an analysis on the global SCL trends of proteins encoded by all bacterial genes, categorized by their pathogen association predicted from the 2018 PAGs update (Chapter 2.3.1). As the preliminary virulence screening strategy, the *P. aeruginosa* PA14 transposon insertion mutant strains of a subset of the 17 *Pseudomonas*-specific PAGs, identified in Chapter 2.4.3, were used to infect *C. elegans* in a slow killing assay and were examined for any virulence attenuation or exacerbation based on changes in the survival of infected worms. A pathogen-associated transcriptional regulator and ametallopeptidase were detected on a *P. aeruginosa* GI, whose complete gene content and prevalence in *P. aeruginosa* strains were subsequently explored in a gene presence-absence matrix across the complete NCBI RefSeq *Pseudomonas* genome dataset available at the time of study.

3.3. Methods

3.3.1. Global SCL analysis of bacterial genes, by pathogen association, from the NCBI RefSeq genomes

The precomputed protein SCL data in PSORTdb 4.0 (Lau et al., 2021) were used to analyze the distribution of bacterial proteins at different SCL sites. SCL sites were assigned to all proteins encoded by genes in the 4807 pathogenic and 3019 non-pathogenic bacterial RefSeq genomes used in Chapter 2.3.1. In each pathogen genome, genes were first classified as either pathogen-associated or common, based on the 2018 PAGs analysis (Chapter 2.3.1). Likewise, in each non-pathogen genome, genes were classified as non-pathogen-associated or common in the same manner. Subsequently, within the pathogen and non-pathogen genome datasets, genes were categorized by their protein SCL sites and the proportion of genes associated with each SCL site was calculated. The proportion of total genes, regardless of pathogen association, in each SCL site was also calculated as a reference for comparison. Finally, these proportions were averaged across all pathogen or non-pathogen genomes within each cell envelope type. The pathogen genome dataset included 3397 gram-negative isolates, 892 gram-positive isolates, 165 gram-negative isolates without an outer membrane and 353 gram-positive isolates with an outer membrane. The non-pathogen genome dataset included 1675 gram-negative isolates, 1208 gram-positive isolates, 25 gram-negative isolates without an outer membrane and 111 gram-positive isolates with an outer membrane.

Due to the compositional nature of the data, the differences in protein SCL distribution of PAGs, non-PAGs and common genes were statistically analyzed by an initial log-ratio transformation of the gene proportions, followed by two-way ANOVA and Tukey's Honestly Significant Difference post-hoc test for pairwise comparison between the different pathogen association groups. The R packages "compositions" v2.0-1 and "stats" v4.0.2 were used for statistical analyses.

3.3.2. *C. elegans*-based virulence screening of PAGs in *P. aeruginosa* PA14

Laboratory work was done by Dr. Patrick Taylor from the Brinkman and Lee Labs.

Preliminary virulence screening was done on 11 of the 17 PAGs from *P. aeruginosa* PA14, that were identified and prioritized from the *Pseudomonas*-specific PAGs analysis (Chapter 2.4.3). PA14 MrT7 insertion mutant strains for these 11 PAGs from the Transposon Insertion Mutant Library (Liberati et al., 2006) were available and thus were screened for an change in virulence in a *C. elegans* infection model. *C. elegans* (wild-type Bristol N2) were maintained on NGM plates coated with *E. coli* OP50 as food source. To prepare for the slow-killing assay, low-osmolarity NGM plates were seeded with 50 µL of overnight culture of the wild-type or transposon mutant strains of *P. aeruginosa* PA14, or *E. coli* OP50 as positive control, per 5.5 cm diameter plates for bacterial lawn formation at 37°C overnight. The plates were then equilibrated at 24°C overnight prior to infection. To prevent progeny development during the assay, 100 ug/mL of 5'-fluoro-2'-deoxyuridine (FUdR) added to the surface of the NGM plates 1 hour prior to the transfer of worms synchronized to the fourth larval (L4) stage. Thirty worms were transferred onto each plate and monitored for viability for up to 8 days (196 hours). Worms were considered dead if they no longer moved or responded to touch (i.e., touching the NGM agar near the worm using the worm picker). The slow killing assay was performed for a minimum of three biological replicates, each with three technical replicates.

Kaplan-Meier survival curves were plotted for each set of worms grown in the presence of 1) transposon insertion mutant of *P. aeruginosa* PA14, 2) wild-type *P. aeruginosa* PA14 (positive control) and 3) *E. coli* OP50 (negative control). Differences in worm survival between infection with the wild-type and mutant strains of *P. aeruginosa*

were assessed using a log-rank test with a statistically significant p-value threshold of 0.05. Survival curve analysis and visualization were done with the R packages "survminer" v0.4.8.

3.3.3. Gene presence-absence analysis of a GI of interest in *P. aeruginosa* PA14

The complete gene set of a 43kb PAGs-containing GI (GI; 2,678,167 to 2,721,433bp) within the *P. aeruginosa* UCBPP-PA14 genome (GCF_000014625.1) was retrieved from IslandViewer 4 (Bertelli et al., 2017) on May 18th, 2021. Using the computed data from *Pseudomonas*-specific PAGs analysis using the full *Pseudomonas* genome dataset (Chapter 2.3.2), a presence-absence matrix of all 45 genes on this GI was constructed and mapped to the species tree of the 605 *Pseudomonas* genome generated in Chapter 2.3.3. Taxonomic trends in the conservation of the full GI gene set were assessed. *Pseudomonas* genomes containing the complete gene set within this GI were characterized by manual search in publications for clinically important features of the corresponding isolates such as hypervirulence or AMR.

3.4. Results

3.4.1. Update, optimization, and maintenance of the PSORT family of archaeal and bacterial SCL tools

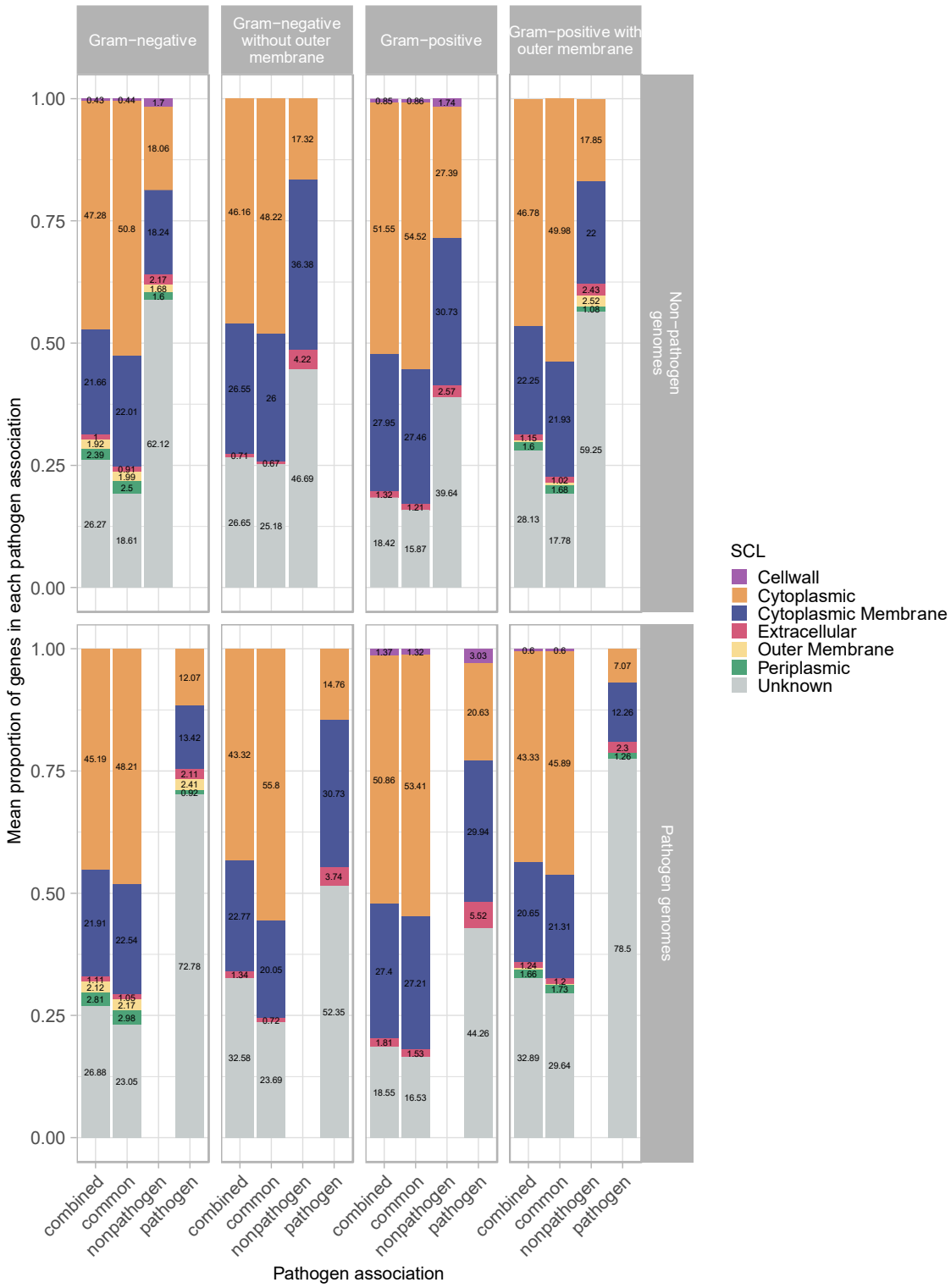
As a part of the efforts in characterizing PAGs by function and protein localization within the bacterial cell, I was actively involved in the routine update, optimization and troubleshooting of the PSORT family protein SCL tools, including the PSORTb SCL predictor for genomic data (Gardy et al., 2005; Yu et al., 2010), the PSORTm SCL predictor for metagenomic data (Peabody et al., 2020) and the PSORTdb SCL database (Lau et al., 2021; Peabody et al., 2016; Rey, Acab, et al., 2005; Yu et al., 2011). I led the update of the PSORTdb 4.0 database, which included an enhancement of the web interface for improved queries and data display, the implementation of a novel unique protein sequence identifier system and the incorporation of the bacterial outer membrane vesicle (OMV) as a novel secondary protein localization site. The secondary localization sites provide a higher resolution of a protein's localization site in addition to the primary localization site such as cytoplasm, cytoplasmic membrane, periplasm, cell

wall, outer membrane and extracellular space. Like previous updates, the PSORTdb 4.0 also features an expanded cPSORTdb database of bacterial proteins with computationally predicted SCL sites as well as an expanded ePSORTdb database of bacterial proteins with experimentally verified SCLs.

The optimized protein SCL predictor for metagenomics data, PSORTm, and the updated protein SCL database, PSORTdb 4.0 have been published separately in two manuscripts, written as the first co-author and first author, respectively (Lau et al., 2021; Peabody et al., 2020).

3.4.2. PAGs are disproportionately localized in the cell wall or extracellular space and have more unknown localization predictions in many bacterial genomes

PAGs had statistically significant differences in the SCL distribution compared to common genes or total genes regardless of pathogen association in gram-negative bacteria and gram-negative bacteria without outer membrane (Figure 3.1). Non-PAGs also differed significantly from common genes or the total genes in gram-negative bacteria, gram-negative bacteria without an outer membrane, as well as gram-positive bacteria with an outer membrane. Specifically, pathogen and non-PAGs are disproportionately enriched extracellular and cell wall proteins. They also have a large proportion of genes with unknown SCL, potentially due to the large number of hypothetical proteins with low sequence similarity to well characterized genes for proper SCL prediction. These SCL results supported the initial functional characterization data and the hypothesis that PAGs are likely involved in host interactions. As such, their protein products are likely found on bacterial cell surface or in the extracellular space where they are more accessible to host cells.


Gram−negative
Gram−negative without outer membrane
Gram−positive
Gram−positive with outer membrane
Non−pathogen genomes
Pathogen genomes
Mean proportion of genes in each pathogen association
1.00
0.75
0.50
0.25
0.00
combined
common
nonpathogen
pathogen
Pathogen association
SCL
Cellwall
Cytoplasmic
Cytoplasmic Membrane
Extracellular
Outer Membrane
Periplasmic
Unknown

B

Pathogen_association_of_genes	p adj	gram_stain	dataset
common−combined	0.9980794128	Gram−positive	Pathogen genomes
pathogen−combined	0.4565084939	Gram−positive	Pathogen genomes
pathogen−common	0.4861196562	Gram−positive	Pathogen genomes
common−combined	0.9639047325	Gram−negative	Pathogen genomes
pathogen−combined	0.0080627974	Gram−negative	Pathogen genomes
pathogen−common	0.0057135326	Gram−negative	Pathogen genomes
common−combined	0.8315726343	Gram−positive with outer membrane	Pathogen genomes
common−combined	0.1565981921	Gram−negative without outer membrane	Pathogen genomes
pathogen−combined	0.0063533387	Gram−negative without outer membrane	Pathogen genomes
pathogen−common	0.0019850654	Gram−negative without outer membrane	Pathogen genomes
common−combined	0.9744928042	Gram−positive	Non−pathogen genomes
nonpathogen−combined	0.2638917854	Gram−positive	Non−pathogen genomes
nonpathogen−common	0.3424827801	Gram−positive	Non−pathogen genomes
common−combined	0.9633512808	Gram−negative	Non−pathogen genomes
nonpathogen−combined	0.0009340469	Gram−negative	Non−pathogen genomes
nonpathogen−common	0.0013579558	Gram−negative	Non−pathogen genomes
common−combined	0.8274066007	Gram−positive with outer membrane	Non−pathogen genomes
nonpathogen−combined	0.0034471924	Gram−positive with outer membrane	Non−pathogen genomes
nonpathogen−common	0.0016703766	Gram−positive with outer membrane	Non−pathogen genomes
common−combined	0.9695525021	Gram−negative without outer membrane	Non−pathogen genomes
nonpathogen−combined	0.0164094089	Gram−negative without outer membrane	Non−pathogen genomes
nonpathogen−common	0.0139408832	Gram−negative without outer membrane	Non−pathogen genomes

Figure 3.1 SCL profile comparison of bacterial genes revealed a significantly different SCL distribution of PAGs and non-PAGs, than common genes, across bacteria of certain cell envelope type.

(A) The proportions (labeled in the plot) within each SCL site and gene pathogen association were averaged across 4807 pathogenic and 3019 non-pathogenic bacteria in each gram-stain/cell envelope type. Each stacked bar represents the average proportion of genes, by pathogen association, in each possible SCL site across all pathogen or non-pathogen genomes of a particular cell envelope type. Note that certain SCL sites are absent in some cell envelope types (e.g., gram-positive bacteria do not have an outer membrane). **(B)** Compositional data-based statistical analysis of the distribution of genes by pathogen association in the different SCL sites. Gene proportions in each SCL site were log-ratio transformed followed by analysis by a two-way ANOVA and Tukey's Honestly Significant Difference post-hoc test for pairwise comparison between the different pathogen association groups. Adjusted p-value less than 0.05 indicates statistical significance between the compared groups. Relative to common genes, the distribution of PAGs and non-PAGs across the multiple SCL sites was significantly different, with more PAGs encoding proteins localized to the extracellular space or cell-wall in gram-negative bacteria (adjusted p=0.00571) and gram-negative bacteria without an outer membrane (adjusted p=0.00198).

3.4.3. 7 of the 11 screened PAGs significantly affected the survival of *C. elegans* infected with the gene-specific mutants

As a follow-up analysis of the 11 PAGs prioritized in Chapter 2.4.3, the *P. aeruginosa* PA14 transposon insertion mutants of 6 genes significantly increased and 1 gene significantly decreased survival of the infected *C. elegans* over 8 days (Figure 3.2). Transposon insertion mutants of an amino acid permease (PA14_RS15305; WP_071535563.1), a GI-localized type I toxin-antitoxin system ptaRNA1 family toxin (PA14_RS06215; WP_011747386.1) and a class I SAM-dependent methyltransferase (PA14_RS20430; WP_003116317.1) improved worm survival, suggesting that the wild-type function of these 3 PAGs likely contributes to virulence. The class I SAM-dependent methyltransferase was previously inferred to be under positive selection in Chapter 2.4.4. Virulence activity was also inferred for 3 uncharacterized genes, PA14_RS13860 (WP_003089506.1), PA14_RS26325 (WP_003095466.1) and PA14_RS02925 (WP_003118883.1), that are currently annotated as “hypothetical proteins” in the NCBI RefSeq database. Functional data for these genes were unavailable and sequence similarity-based analyses were not informative due to high sequence divergence from any known genes. PSORTb v3.0 (Yu et al., 2010) SCL prediction gave a cytoplasmic membrane localization for PA14_RS02925. Other genes had an unknown localization based on the current PSORTb 3.0 prediction.

The BlaI/MecI/CopY family transcriptional regulator (PA14_RS12700, WP_025297936.1) accelerated worm killing, suggesting that the wild-type gene may be a virulence repressor. This transcriptional regulator is of particular interest due its immediate proximity to the gene encoding the M56 family metallopeptidase (PA14_RS12695; WP_016254216.1), that is both pathogen-associated and evident of positive selection. This pair of genes is predicted to be conserved within the same operon and localized to the same GI based on data from the Pseudomonas Genome Database (Winsor et al., 2016) and IslandViewer 4 (Bertelli et al., 2017). This pair of pathogen-associated transcriptional regulator and metallopeptidase, as well as the GI on which they reside were further explored in the next section (Chapter 3.4.4).

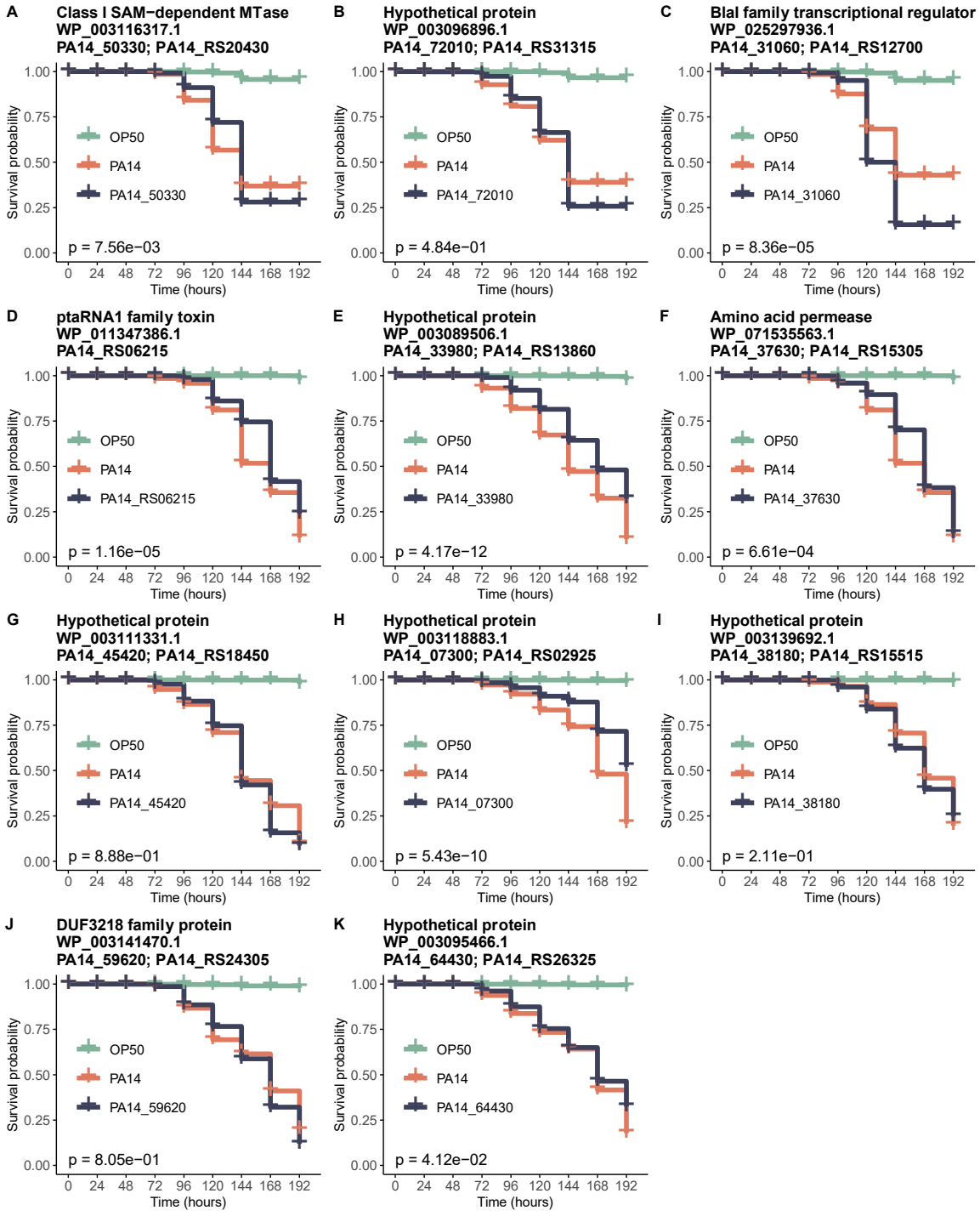


Figure 3.2 Kaplan-Meier curves of *C. elegans* infected with *P. aeruginosa* PA14 transposon mutants of select PAGs indicated virulence activity in six PAGs and virulence-repressing activity in one PAG.

Each *P. aeruginosa* PA14 transposon insertion mutant strain (blue) was compared to the wild-type *P. aeruginosa* PA14 strain (green). *E. coli* OP50 was included as a negative control (orange). Worm viability was assessed over 8 days. All Kaplan-Meier curves are averaged across three biological and three technical replicates. The PAGs conferring transposon insertion are

identified by their deduced protein names, NCBI RefSeq protein accessions (WP_), old PA14 locus tags (PA14_) and updated PA14 locus tags (PA14_RS) within the *P. aeruginosa* UCBPP-PA14 genome (GCF_000014625.1). P-values (p) of the log-rank test between survival of worms infected with the PA14 wild-type and transposon mutant strains are indicated within each figure. P-value below 0.05 represents a statistically significant difference in worm survival. Based on this data, contribution to virulence was inferred for the following genes: an amino acid permease (PA14_RS15305; WP_071535563.1), a type I toxin-antitoxin system ptaRNA1 family toxin (PA14_RS06215; WP_011747386.1), a class I SAM-dependent methyltransferase (PA14_RS20430; WP_003116317.1) and 3 uncharacterized genes, PA14_RS13860 (WP_003089506.1), PA14_RS26325 (WP_003095466.1), and PA14_RS02925 (WP_003118883.1). Virulence-repressing activity was inferred in the BlaI/MecI/CopY family transcriptional regulator (PA14_RS12700, WP_025297936.1).

3.4.4. The 43kb GI containing PA14_RS12695 and PA14_RS12700 is found in PA14 as well as several multidrug-resistant strains of *P. aeruginosa*

M56 family metallopeptidase (PA14_RS12695; WP_016254216.1) and the adjacent BlaI/MecI/CopY family transcriptional regulator (PA14_RS12700, WP_025297936.1) are located on the same 43kb GI (2678167 to 2721432bp) in the *P. aeruginosa* UCBPP-PA14 genome. The gene presence-absence analysis revealed that the complete gene content of the GI is conserved only in *P. aeruginosa* isolates. The presence of the full GI in isolates in the PA14 and PAO1 phylogenetic groups suggests that it is not lineage-specific. This GI contains multiple metal transporter genes and efflux genes encoding products such as a HupE/UreJ family protein, CusA/CzcA family heavy metal efflux RND transporter, TolC family protein and the major facilitator superfamily transporter (Table 3.1; Figure 3.3).

The complete gene set of the 43kb GI containing the M56 metallopeptidase and the BlaI/MecI/CopY family transcriptional regulator, in the *P. aeruginosa* PA14 genome were detected in multiple multidrug and extensively drug resistant *P. aeruginosa* strains from both the PA14 and the PAO1 lineages (Figure 3.4). Within the PA14 lineage, the full GI was found in the copper resistant M1608 and M37351 strains (Petitjean et al., 2017), the multiphage- and multidrug-resistant Pcyll-40 strain (Pourcel et al., 2020), the extensively drug-resistant Pa58, Pa124 and Pa127 strains (Espinosa-Camacho et al., 2017) and the global epidemic AR_0353 strain (Chew et al., 2019). Within the PAO1 lineage, strains carrying the full GI include the CF-associated RIVM-EMC2982 (Santi et al., 2021), the extensively drug-resistant AG1 (Molina-Mora et al., 2020) and the Carbo01 63 (van der Zee et al., 2018).

Two strains within the PA7 group, AR441 (GCF_003073695.1) and AR0356 (GCF_002968755.1), carry all except one efflux pump gene, the efflux resistance-nodulation-division transporter periplasmic adaptor subunit, in this GI. Additionally, a *pBT2436*-like megaplasmid carrying antimicrobial resistant genes against aminoglycoside, phenicol, sulfonamide and tetracycline has also been documented in these strains (Cazares et al., 2020). The association of this GI with antimicrobial resistant (AMR) strains suggests the genes on this island, potentially including the pathogen-associated M56 metallopeptidase and the BlaI/MecI/CopY family transcriptional regulator, may play an important role not only in virulence, but also in AMR in clinically relevant isolates.

Table 3.1 45 genes found on the 43kb GI of interest in the *P. aeruginosa* PA14 genome.

NCBI RefSeq protein accession	Locus tag	Name	Start	End
WP_016254220.1	PA14_RS12590	Hypothetical	2678195	2678438
WP_003138971.1	PA14_RS12595	conjugative transfer protein TrbI	2678434	2679706
WP_003138972.1	PA14_RS12600	P-type conjugative transfer protein TrbG	2679708	2680698
WP_003138973.1	PA14_RS12605	Conjugal transfer protein TrbF	2680694	2681399
WP_003138975.1	PA14_RS12610	Conjugal transfer protein TrbL	2681411	2682782
WP_016254219.1	PA14_RS12615	Hypothetical protein	2682778	2683099
WP_003138979.1	PA14_RS12620	Conjugal transfer protein TrbJ	2683110	2683836
WP_003138981.1	PA14_RS12625	Conjugal transfer protein TrbE	2683832	2686286
WP_003138984.1	PA14_RS12630	Conjugal transfer protein TrbD	2686298	2686571
WP_003138985.1	PA14_RS12635	Conjugal transfer protein TrbC	2686567	2686960
WP_003138986.1	PA14_RS12640	P-type conjugative transfer ATPase TrbB	2686956	2688027
WP_003138987.1	PA14_RS12645	CopG family transcriptional regulator	2688023	2688488
WP_003138988.1	PA14_RS12650	Conjugal transfer protein TraG	2688484	2690485
WP_025297938.1	PA14_RS12655	EexN family lipoprotein	2690721	2691000
WP_022580529.1	PA14_RS12660	D-alanyl-D-alanine endopeptidase	2691004	2691940
WP_003138990.1	PA14_RS12665	Hypothetical protein	2692126	2692432
WP_003138992.1	PA14_RS12670	Hypothetical protein	2692482	2693151
WP_022580528.1	PA14_RS12675	membrane protein	2693163	2693856
WP_003138997.1	PA14_RS12680	CusA/CzcA family heavy metal efflux RND transporter	2693903	2697041
WP_003139001.1	PA14_RS12685	Efflux RND transporter periplasmic adaptor subunit	2697051	2698281
WP_003139003.1	PA14_RS12690	TolC family protein	2698277	2699525
WP_016254216.1	PA14_RS12695	M56 family metallopeptidase	2699665	2700661
WP_025297936.1	PA14_RS12700	BlaI/MecI/CopY family transcriptional regulator	2700660	2701059
WP_016254215.1	PA14_RS12705	relaxase/mobilization nuclease	2701169	2703155
WP_003139016.1	PA14_RS12710	S26 family signal peptidase	2703603	2704179
WP_016254214.1	PA14_RS12715	DUF2840 domain-containing protein	2704175	2704733

NCBI RefSeq protein accession	Locus tag	Name	Start	End
WP_016254213.1	PA14_RS12720	Hypothetical protein	2704729	2705014
WP_003139018.1	PA14_RS12725	AAA family ATPase	2705010	2705649
WP_016254212.1	PA14_RS12730	Replication initiator protein A	2705902	2706760
WP_003107109.1	PA14_RS12735	Helix-turn-helix domain-containing protein	2706786	2707068
WP_011666628.1	PA14_RS12740	DUF2285 domain-containing protein	2707179	2707950
WP_016254211.1	PA14_RS12745	DUF2958 domain-containing protein	2708293	2708644
WP_016254210.1	PA14_RS12750	Helix-turn-helix transcriptional regulator	2708886	2709183
WP_016254209.1	PA14_RS12755	DUF736 domain-containing protein	2709576	2709891
WP_016254208.1	PA14_RS12760	Hypothetical protein	2710671	2710881
WP_003139031.1	PA14_RS12765	ParB N-terminal domain-containing protein	2710944	2712999
WP_003139033.1	PA14_RS12770	DUF945 domain-containing protein	2713080	2713905
WP_003139034.1	PA14_RS12775	Hypothetical protein	2714624	2714903
WP_016254207.1	PA14_RS12780	DUF2958 domain-containing protein	2714968	2715319
WP_003139161.1	PA14_RS12785	MFS transporter	2715632	2716943
WP_025297819.1	PA14_RS12790	Metal-sensing transcriptional repressor	2716952	2717210
WP_003139163.1	PA14_RS12795	Hypothetical protein	2717345	2717738
WP_003139036.1	PA14_RS12800	DNA repair protein RadC	2718014	2718521
WP_003139037.1	PA14_RS12805	PDDEXK nuclease domain-containing protein	2718847	2720011
WP_003139038.1	PA14_RS12810	Site-specific integrase	2720007	2721213

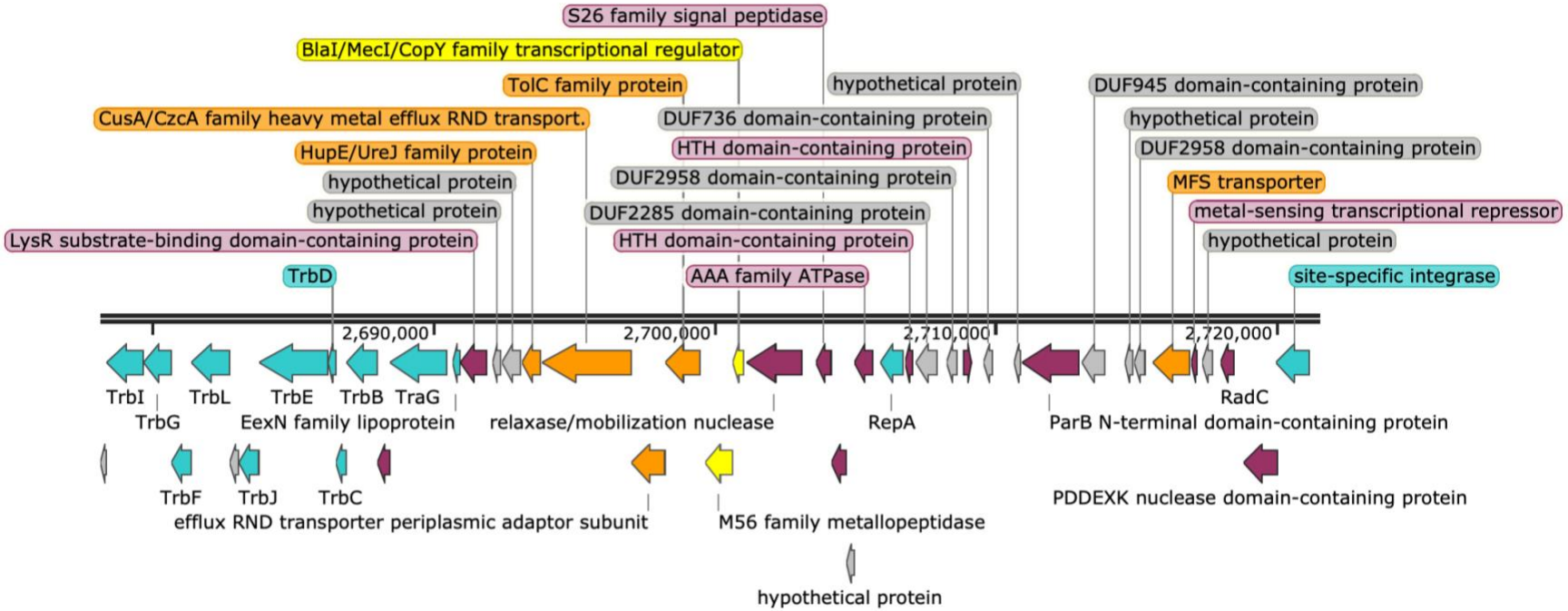
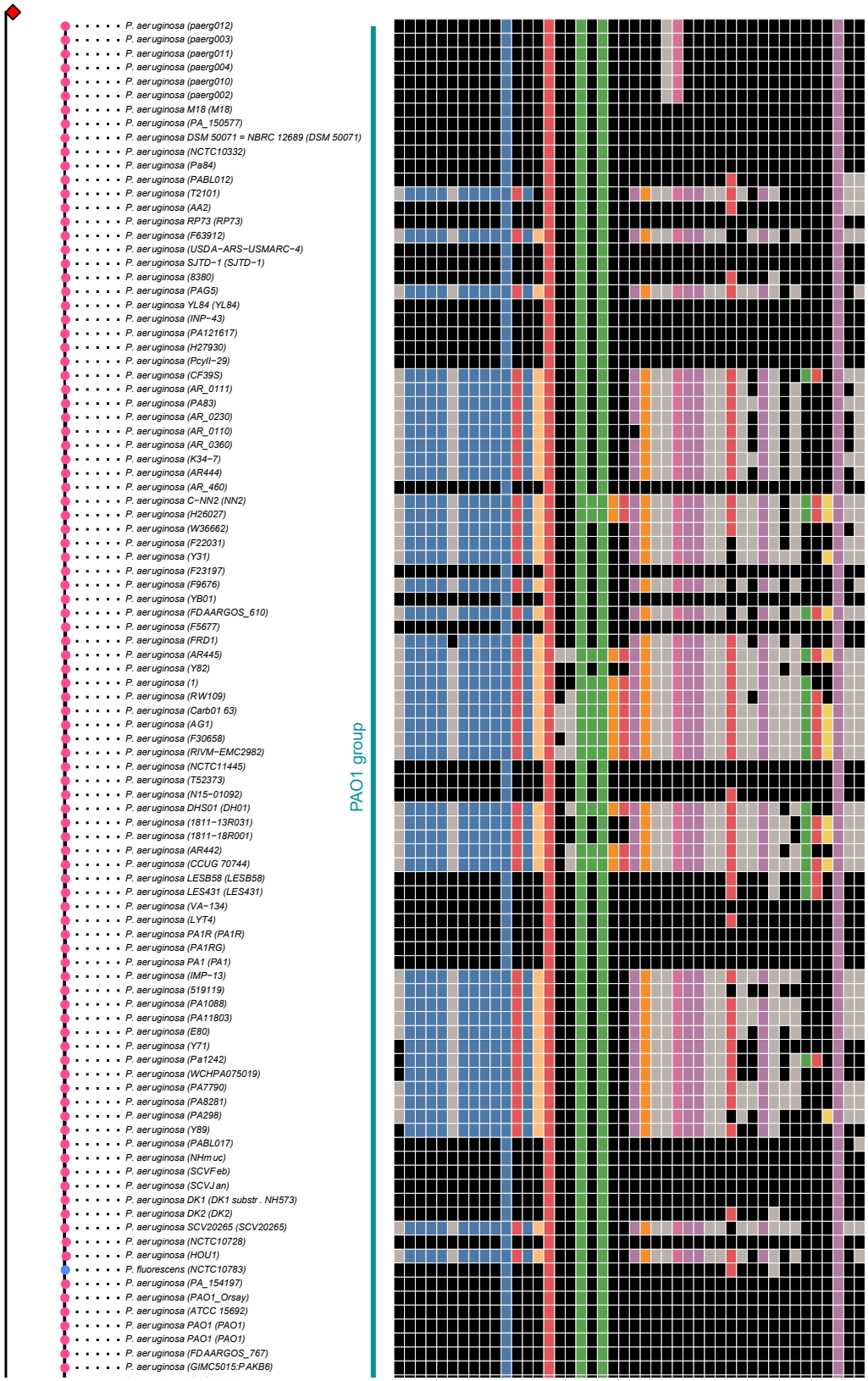
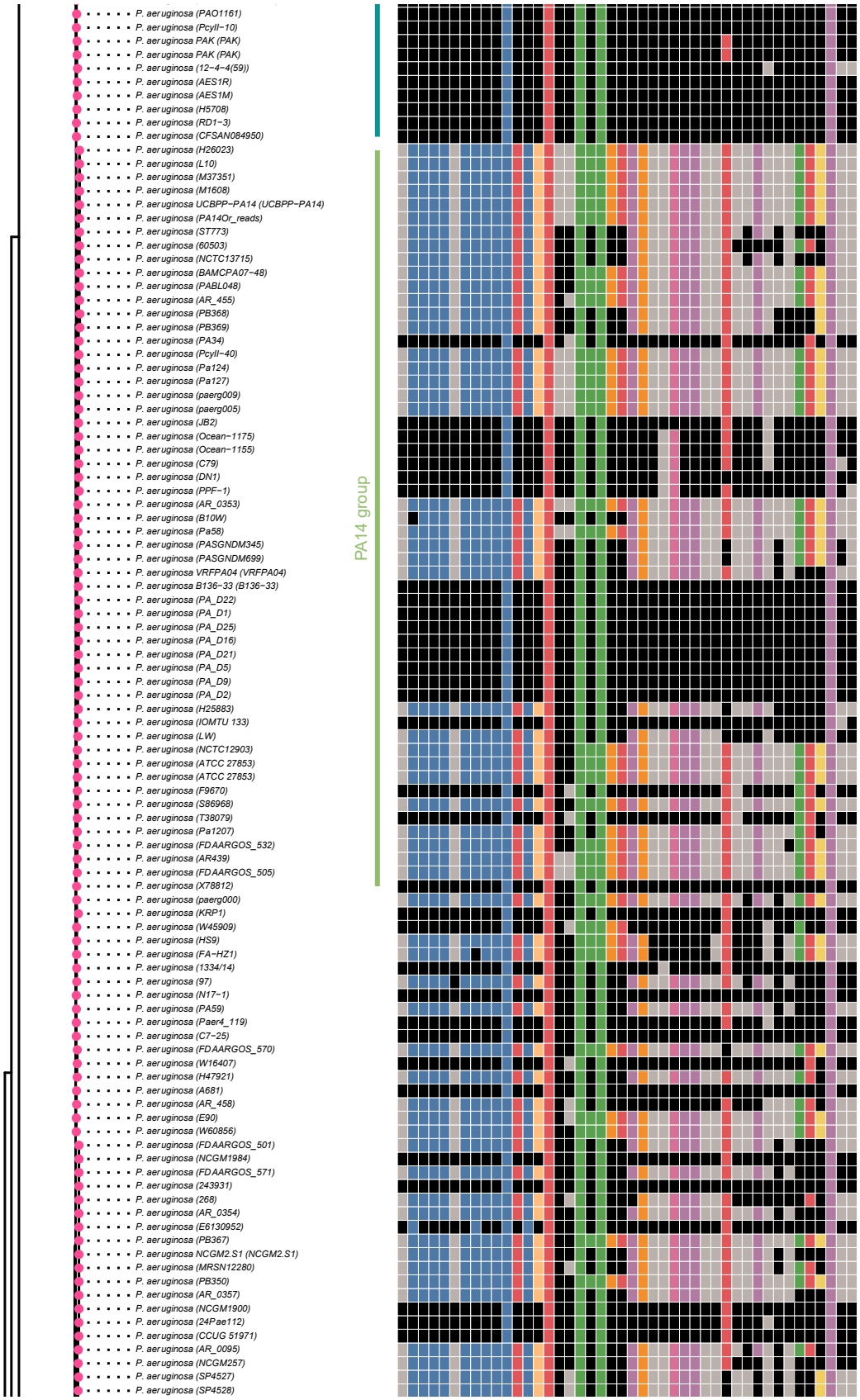


Figure 3.3 The 43kb GI (2,678,167 – 2,721,432bp) in the *P. aeruginosa* PA14 genome (GCF_000014625.1) contains the pathogen-associated M56 metallopeptidase and BlaI/MecI/CopY family transcriptional regulator, as well as several genes related to metal resistance and transporters.

The 45 genes on this GI are categorized into plasmid conjugation system (teal), transporters (orange), the two PAGs of interest (yellow), and uncharacterized genes (gray). Figure was generated with SnapGene v6.0.2.


P. aeruginosa (paerg012)
P. aeruginosa (paerg003)
P. aeruginosa (paerg011)
P. aeruginosa (paerg004)
P. aeruginosa (paerg010)
P. aeruginosa (paerg002)
P. aeruginosa M18 (M18)
P. aeruginosa (PA_150577)
P. aeruginosa DSM 50071 = NBRC 12689 (DSM 50071)
P. aeruginosa (NCTC10332)
P. aeruginosa (Pa84)
P. aeruginosa (PABL012)
P. aeruginosa (T2101)
P. aeruginosa (AA2)
P. aeruginosa RP73 (RP73)
P. aeruginosa (F63912)
P. aeruginosa (USDA−ARS−USMARC−4)
P. aeruginosa SJTD−1 (SJTD−1)
P. aeruginosa (8380)
P. aeruginosa (PAG5)
P. aeruginosa YL84 (YL84)
P. aeruginosa (INP−43)
P. aeruginosa (PA121617)
P. aeruginosa (H27930)
P. aeruginosa (PcyII−29)
P. aeruginosa (CF39S)
P. aeruginosa (AR_0111)
P. aeruginosa (PA83)
P. aeruginosa (AR_0230)
P. aeruginosa (AR_0110)
P. aeruginosa (AR_0360)
P. aeruginosa (K34−7)
P. aeruginosa (AR444)
P. aeruginosa (AR_460)
P. aeruginosa C−NN2 (NN2)
P. aeruginosa (H26027)
P. aeruginosa (W36662)
P. aeruginosa (F22031)
P. aeruginosa (Y31)
P. aeruginosa (F23197)
P. aeruginosa (F9676)
P. aeruginosa (YB01)
P. aeruginosa (FDAARGOS_610)
P. aeruginosa (F5677)
P. aeruginosa (FRD1)
P. aeruginosa (AR445)
P. aeruginosa (Y82)
P. aeruginosa (1)
P. aeruginosa (RW109)
P. aeruginosa (Carb01 63)
P. aeruginosa (AG1)
P. aeruginosa (F30658)
P. aeruginosa (RIVM−EMC2982)
P. aeruginosa (NCTC11445)
P. aeruginosa (T52373)
P. aeruginosa (N15−01092)
P. aeruginosa DHS01 (DH01)
P. aeruginosa (1811−13R031)
P. aeruginosa (1811−18R001)
P. aeruginosa (AR442)
P. aeruginosa (CCUG 70744)
P. aeruginosa LESB58 (LESB58)
P. aeruginosa LES431 (LES431)
P. aeruginosa (VA−134)
P. aeruginosa (LYT4)
P. aeruginosa PA1R (PA1R)
P. aeruginosa (PA1RG)
P. aeruginosa PA1 (PA1)
P. aeruginosa (IMP−13)
P. aeruginosa (519119)
P. aeruginosa (PA1088)
P. aeruginosa (PA11803)
P. aeruginosa (E80)
P. aeruginosa (Y71)
P. aeruginosa (Pa1242)
P. aeruginosa (WCHPA075019)
P. aeruginosa (PA7790)
P. aeruginosa (PA8281)
P. aeruginosa (PA298)
P. aeruginosa (Y89)
P. aeruginosa (PABL017)
P. aeruginosa (NHmuc)
P. aeruginosa (SCVFeb)
P. aeruginosa (SCVJan)
P. aeruginosa DK1 (DK1 substr. NH573)
P. aeruginosa DK2 (DK2)
P. aeruginosa SCV20265 (SCV20265)
P. aeruginosa (NCTC10728)
P. aeruginosa (HOU1)
P. fluorescens (NCTC10783)
P. aeruginosa (PA_154197)
P. aeruginosa (PAO1_Orsay)
P. aeruginosa (ATCC 15692)
P. aeruginosa PAO1 (PAO1)
P. aeruginosa PAO1 (PAO1)
P. aeruginosa (FDAARGOS_767)
P. aeruginosa (GIMC5015:PAKB6)
PAO1 group


P. aeruginosa (PAO1161)
P. aeruginosa (PcyII−10)
P. aeruginosa PAK (PAK)
P. aeruginosa PAK (PAK)
P. aeruginosa (12−4−4(59))
P. aeruginosa (AES1R)
P. aeruginosa (AES1M)
P. aeruginosa (H5708)
P. aeruginosa (RD1−3)
P. aeruginosa (CFSAN084950)
P. aeruginosa (H26023)
P. aeruginosa (L10)
P. aeruginosa (M37351)
P. aeruginosa (M1608)
P. aeruginosa UCBPP−PA14 (UCBPP−PA14)
P. aeruginosa (PA14Or_reads)
P. aeruginosa (ST773)
P. aeruginosa (60503)
P. aeruginosa (NCTC13715)
P. aeruginosa (BAMCPA07−48)
P. aeruginosa (PABL048)
P. aeruginosa (AR_455)
P. aeruginosa (PB368)
P. aeruginosa (PB369)
P. aeruginosa (PA34)
P. aeruginosa (PcyII−40)
P. aeruginosa (Pa124)
P. aeruginosa (Pa127)
P. aeruginosa (paerg009)
P. aeruginosa (paerg005)
P. aeruginosa (JB2)
P. aeruginosa (Ocean−1175)
P. aeruginosa (Ocean−1155)
P. aeruginosa (C79)
P. aeruginosa (DN1)
P. aeruginosa (PPF−1)
P. aeruginosa (AR_0353)
P. aeruginosa (B10W)
P. aeruginosa (Pa58)
P. aeruginosa (PASGNDM345)
P. aeruginosa (PASGNDM699)
P. aeruginosa VRFPA04 (VRFPA04)
P. aeruginosa B136−33 (B136−33)
P. aeruginosa (PA_D22)
P. aeruginosa (PA_D1)
P. aeruginosa (PA_D25)
P. aeruginosa (PA_D16)
P. aeruginosa (PA_D21)
P. aeruginosa (PA_D5)
P. aeruginosa (PA_D9)
P. aeruginosa (PA_D2)
P. aeruginosa (H25883)
P. aeruginosa (IOMTU 133)
P. aeruginosa (LW)
P. aeruginosa (NCTC12903)
P. aeruginosa (ATCC 27853)
P. aeruginosa (ATCC 27853)
P. aeruginosa (F9670)
P. aeruginosa (S86968)
P. aeruginosa (T38079)
P. aeruginosa (Pa1207)
P. aeruginosa (FDAARGOS_532)
P. aeruginosa (AR439)
P. aeruginosa (FDAARGOS_505)
P. aeruginosa (X78812)
P. aeruginosa (paerg000)
P. aeruginosa (KRP1)
P. aeruginosa (W45909)
P. aeruginosa (HS9)
P. aeruginosa (FA−HZ1)
P. aeruginosa (1334/14)
P. aeruginosa (97)
P. aeruginosa (N17−1)
P. aeruginosa (PA59)
P. aeruginosa (Paer4_119)
P. aeruginosa (C7−25)
P. aeruginosa (FDAARGOS_570)
P. aeruginosa (W16407)
P. aeruginosa (H47921)
P. aeruginosa (A681)
P. aeruginosa (AR_458)
P. aeruginosa (E90)
P. aeruginosa (W60856)
P. aeruginosa (FDAARGOS_501)
P. aeruginosa (NCGM1984)
P. aeruginosa (FDAARGOS_571)
P. aeruginosa (243931)
P. aeruginosa (268)
P. aeruginosa (AR_0354)
P. aeruginosa (E6130952)
P. aeruginosa (PB367)
P. aeruginosa NCGM2.S1 (NCGM2.S1)
P. aeruginosa (MRSN12280)
P. aeruginosa (PB350)
P. aeruginosa (AR_0357)
P. aeruginosa (NCGM1900)
P. aeruginosa (24Pae112)
P. aeruginosa (CCUG 51971)
P. aeruginosa (AR_0095)
P. aeruginosa (NCGM257)
P. aeruginosa (SP4527)
P. aeruginosa (SP4528)
PA14 group

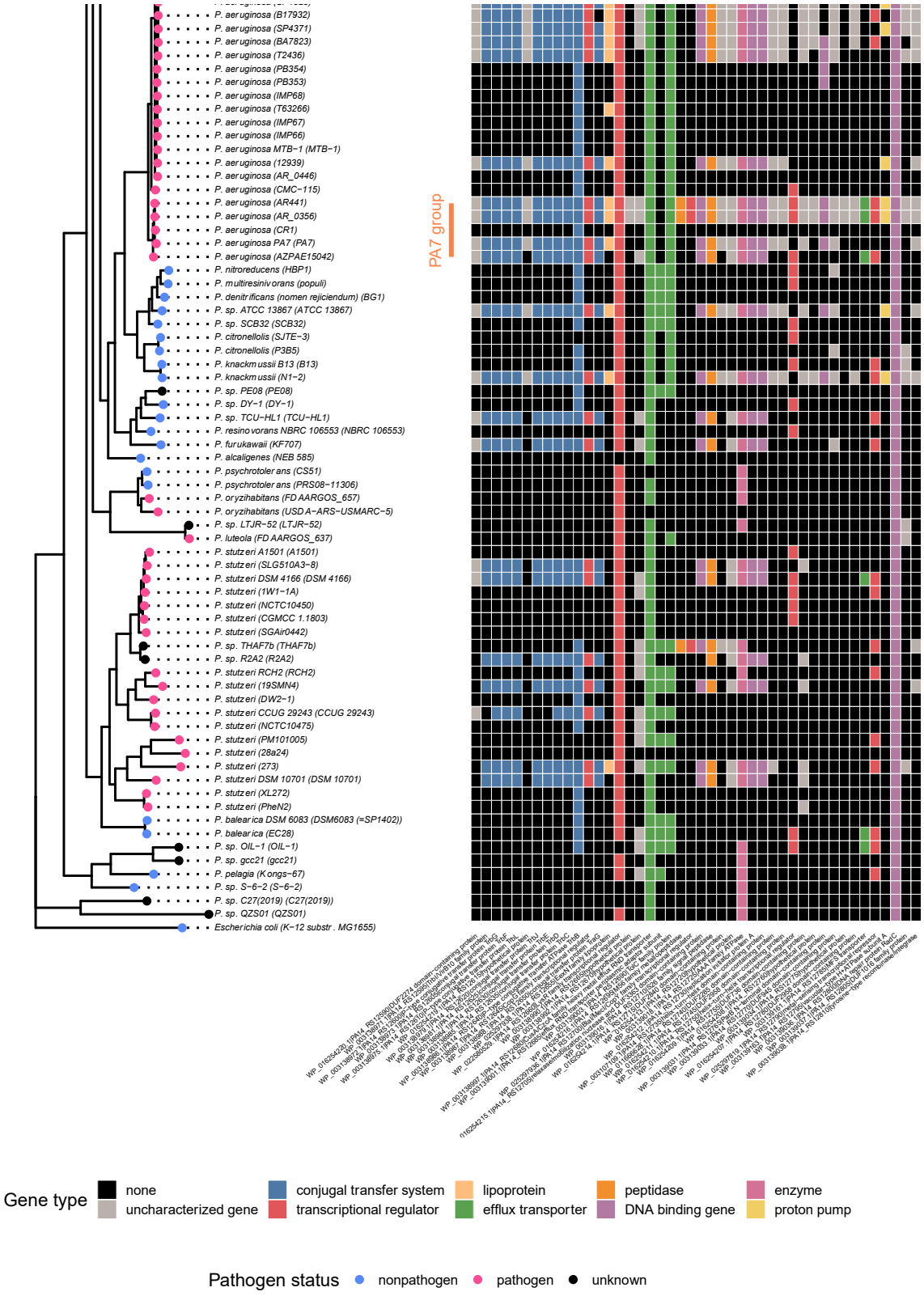

P. aeruginosa (B17932)
P. aeruginosa (SP4371)
P. aeruginosa (BA7823)
P. aeruginosa (T2436)
P. aeruginosa (PB354)
P. aeruginosa (PB353)
P. aeruginosa (IMP68)
P. aeruginosa (T63266)
P. aeruginosa (IMP67)
P. aeruginosa (IMP66)
P. aeruginosa MTB−1 (MTB−1)
P. aeruginosa (12939)
P. aeruginosa (AR_0446)
P. aeruginosa (CMC−115)
P. aeruginosa (AR441)
P. aeruginosa (AR_0356)
P. aeruginosa (CR1)
P. aeruginosa PA7 (PA7)
P. aeruginosa (AZPAE15042)
P. nitroreducens (HBP1)
P. multiresiniv orans (populi)
P. denitrificans (nomen rejiciendum) (BG1)
P. sp. ATCC 13867 (ATCC 13867)
P. sp. SCB32 (SCB32)
P. citronellolis (SJTE−3)
P. citronellolis (P3B5)
P. knackm ussii B13 (B13)
P. knackm ussii (N1−2)
P. sp. PE08 (PE08)
P. sp. DY−1 (DY−1)
P. sp. TCU−HL1 (TCU−HL1)
P. resino vorans NBRC 106553 (NBRC 106553)
P. fur ukawaii (KF707)
P. alcaligenes (NEB 585)
P. psychrotoler ans (CS51)
P. psychrotoler ans (PRS08−11306)
P. oryzihabitans (FD AARGOS_657)
P. oryzihabitans (USD A−ARS−USMARC−5)
P. sp. LTJR−52 (LTJR−52)
P. luteola (FD AARGOS_637)
P. stutz eri A1501 (A1501)
P. stutz eri (SLG510A3−8)
P. stutz eri DSM 4166 (DSM 4166)
P. stutz eri (1W1−1A)
P. stutz eri (NCTC10450)
P. stutz eri (CGMCC 1.1803)
P. stutz eri (SGAir0442)
P. sp. THAF7b (THAF7b)
P. sp. R2A2 (R2A2)
P. stutz eri RCH2 (RCH2)
P. stutz eri (19SMN4)
P. stutz eri (DW2−1)
P. stutz eri CCUG 29243 (CCUG 29243)
P. stutz eri (NCTC10475)
P. stutz eri (PM101005)
P. stutz eri (28a24)
P. stutz eri (273)
P. stutz eri DSM 10701 (DSM 10701)
P. stutz eri (XL272)
P. stutz eri (PheN2)
P. balear ica DSM 6083 (DSM6083 (=SP1402))
P. balear ica (EC28)
P. sp. OIL−1 (OIL−1)
P. sp. gcc21 (gcc21)
P. pelagia (K ongs−67)
P. sp. S−6−2 (S−6−2)
P. sp. C27(2019) (C27(2019))
P. sp. QZS01 (QZS01)
Escherichia coli (K−12 substr . MG1655)
PA7 group
Gene type
none
uncharacterized gene
conjugal transfer system
transcriptional regulator
lipoprotein
efflux transporter
peptidase
DNA binding gene
enzyme
proton pump
Pathogen status
nonpathogen
pathogen
unknown

Figure 3.4 Presence-absence matrix of the 45 genes on the 43kb GI (2678167 to 2721432) on *P. aeruginosa* PA14, mapped to a *Pseudomonas* species tree, revealed that the complete gene set of this GI is found in multiple multidrug and extensively drug resistant strains of *P. aeruginosa*.

Genes are displayed in the order of locus tags from left to right and colored by function. Species tree was constructed from the 16s rRNA, *gyrB*, *rpoB*, *rpoD* genes from the complete 605 *Pseudomonas* RefSeq genomes, using IQtree. Pathogen status of the genomes are denoted by colored nodes on the tree. Red diamond on the species tree represents a collapsed clades of the remaining *Pseudomonas* genomes that did not carry the full gene set of the GI.

3.5. Discussion

This chapter began with an analysis of the global trend in bacterial protein SCL among PAGs, non-PAGs and common genes. Relative to common genes, PAGs and non-PAGs disproportionately encoded proteins with a predicted cell wall and extracellular localization, suggesting that genes that are conserved in only pathogens or non-pathogens may be more involved in sensing and interacting with their external environments. Under the assumption that PAGs and non-PAGs are likely non-essential genes specialized in niche-adaptation, this observation is in agreement with a smaller-scale PSORTb-based SCL study of 24 bacteria, which showed that while essential bacterial proteins are enriched in the cytoplasm, non-essential proteins are predominantly localized in the cytoplasmic membrane, periplasm, outer membrane, cell wall and extracellular space while (Peng & Gao, 2014). In addition, PAGs and non-PAGs also encode more proteins with unknown SCL. This is likely due to the challenge of PSORTb prediction of the abundant uncharacterized proteins that are conserved only in pathogens or non-pathogens. These uncharacterized proteins may have high sequence divergence from proteins with known SCL that were used to develop and train the PSORTb training modules. As proteins need to be transported to the appropriate cellular site to perform their inherent function, their localization site may shed light on their molecular function and protein interaction if no other functional data are available.

Using results from this SCL analysis, researchers in antivirulence drug development may prioritize PAGs as antivirulence drug targets based on the drug accessibility of the cellular compartment in which their encoded proteins reside. Outer membrane or extracellular proteins are often implicated in virulence and are involved in pathogen-host interactions. For example, bacterial toxins are secreted proteins associated with virulence processes, including physical damage to host cells,

biochemical degradation and interrupted cellular signalling in hosts, which altogether dampens the host ability to elicit an effective immune response and pathogen clearance. Some well-known secreted toxins include the hemolysin in *S. aureus*, ExoS and ExoT in *P. aeruginosa*, toxin A and toxin B in *Clostridioides difficile*, shiga toxin Stx1 and Stx2 in *E. coli* and *Shigella dysenteriae*. OMVs have garnered increased attention over recent years as membrane-associated VFs. OMVs are spherical bilayer vesicles released from the outer membrane of gram-negative bacterium such as *P. aeruginosa. E. coli, Shigella spp. Vibro sp.* And *Neisseria sp.* (Cecil et al., 2019; Jan, 2017; Schwechheimer & Kuehn, 2015) Carrying a diverse range of cargoes including lipopolysaccharides, peptidoglycan, outer membrane proteins, signalling molecules, toxins and many VFs, OMVs play an important role in the pathogen lifestyle including host adaptation, stress response, and AMR. Due to their implication in bacterial pathogenesis, OMVs have been added as a secondary localization site to version 4.0 of PSORTdb.

Of the 11 PAGs screened *in vivo*, virulence activity was inferred from the transposon mutants of 6 genes: an amino acid permease (PA14_RS15305; WP_071535563.1), a type I toxin-antitoxin system ptaRNA1 family toxin (PA14_RS06215; WP_011747386.1), a class I SAM-dependent methyltransferase (PA14_RS20430; WP_003116317.1), and 3 uncharacterized genes, PA14_RS13860 (WP_003089506.1), PA14_RS26325 (WP_003095466.1) and PA14_RS02925 (WP_003118883.1). Virulence-repressing activity was inferred only in the BlaI/MecI/CopY family transcriptional regulator (PA14_RS12700, WP_025297936.1). Although functional data for many of these genes were either lacking or only available at the general protein family level at the time of this study, the laboratory evidence of virulence activity from this study warrants further investigation of the role and impact on the host of these genes in *P. aeruginosa*. While the intestinal accumulation of bacteria and the lack of an adaptive immune system in *C. elegans* do not fully reflect many extraintestinal infections and the antimicrobial immune response in mammalian organisms, *C. elegans* has been a simple yet powerful model organism for identifying large set of universally conserved VFs responsible for the pathogenesis of *P. aeruginosa* in plant and animal hosts (Mahajan-Miklos et al., 2000). This *C. elegans* study served as a preliminary and cost-effective method for selecting putative virulence related PAGs for more refined functional assays. Future *in vivo* functional analyses of the mutant of these PAGs will transition to more complex mammalian models with an immune system more

related to the human's against *Pseudomonas* infections. Numerous murine models have been designed for studying chronic and acute lung infections, as well as burn wound infections by *P. aeruginosa* (Bayes et al., 2016; Turner et al., 2014). While the use of transposon insert mutants is widely used for functional screening, a few caveats must be considered and addressed with follow-up analyses. Transposon insertion mutagenesis inserts a transposable element into a random site in a gene to disrupt its proper expression of a functional protein. Dependent on the site of insertion, transposon inserting near the 3' end is more likely to not disrupt the proper function of a gene than an insertion at the 5' end (start of the gene) (Liberati et al., 2006). For this reason, *P. aeruginosa* PA14 transposon mutant strains with insertion only in the beginning half of the PAGs of interest were chosen. A subsequent functional study will validate the virulence of the select PAGs using clean in-frame knock outs.

Two adjacent PAGs, PA14_RS12695 and PA14_RS12700, located on a GI in *P. aeruginosa* PA14 have garnered attention from the virulence screening in this chapter as well as from the evolutionary selection inference from Chapter 2. PA14_RS12700 is a BlaI/MecI/CopY family transcriptional regulator with a probable virulence repressor function as its transposon mutant reduced the survival of *P. aeruginosa*-infected *C. elegans*. While PA14_RS12700, a M56 family metallopeptidase, was not screened for virulence in the worm infection model due to the lack of readily available transposon mutant at the time of study, it was inferred to be under positive selection in Chapter 2.4.4. This pair of PAGs were predicted to reside on a GI with a few metal and AMR genes. Cross-resistance to heavy metals and antimicrobials due to the sharing of the same efflux mechanisms, has also been reported among human pathogens in the environmental reservoirs (Dickinson et al., 2019). Since GIs can drive genetic diversification and pathogen adaptation to different ecological environments, the presence of this GI in many global, multi-drug resistant *P. aeruginosa* strains may suggest an important role in enhanced virulence and AMR.

In conclusion, PAGs may be practical drug targets due to their disproportionate enrichment in drug-accessible cellular compartments such as the cell wall and extracellular space, particularly in gram-negative bacteria that are notoriously associated with AMR. With Dr. Patrick Taylor, we identified 6 PAGs with virulence activity and 1 PAG with virulence-repressing activity. This assay helped us further prioritize the virulence repressing PAG, a BlaI/MecI/CopY family transcriptional regulator

(PA14_RS12700, WP_025297936.1) and its neighbouring positively selected PAG, a M56 family metallopeptidase (PA14_RS12695; WP_016254216.1) for a more comprehensive *in vivo* functional characterization study now being conducted by the Brinkman and Lee Labs. Our computational pathogen-associated gene characterization method, complemented with laboratory virulence screening, is a promising workflow for detecting putative virulence from bacterial genes whose function has yet to be examined. Information on the protein SCL and genomic island localization of these virulence-related PAGs may help to prioritize candidate antivirulence drug targets based on potential drug accessibility and to design effective antivirulence therapeutics for treating bacterial infections.

Chapter 4.

Structure-activity relationship analysis of raloxifene as a potential antivirulence agent against *P. aeruginosa*

This chapter presents a follow-up analysis to the peer-reviewed article "Raloxifene attenuates P. aeruginosa pyocyanin production and virulence" published by the Brinkman Lab in collaboration with the Lei Xie Research Group at Hunter College, New York (S. J. Ho Sui et al., 2012). I compared various analogs of raloxifene to investigate the structural component(s) responsible for raloxifene's antivirulence properties against P. aeruginosa PA14.

I performed all work presented in this chapter with the following exception: Members of Dr. Peter Wilson's laboratory in the SFU Chemistry Department synthesized all raloxifene analogs. In particular, Jacob Duerichen-Parfitt, the undergraduate student in the Wilson Lab, had work extensively on the synthesis. Former Brinkman Lab managers, Raymond Lo and Dr. Hanadi Ibrahim, assisted with the bacterial and worm cultures.

4.1. Abstract

Antivirulence drugs are a promising alternative for treating bacterial infections and combatting the global concern of AMR. Previously, an *in silico* docking analysis coupled with laboratory validation, identified the selective estrogen receptor modulator, raloxifene, as a potential antivirulence agent against *P. aeruginosa*. Here, the structure-activity relationship of raloxifene in the context of antivirulence was investigated by comparing the performance of raloxifene analogs in pyocyanin production inhibition, improved survival of *Pseudomonas*-infected *C. elegans*, and bacterial growth. The removal of the piperidinylethoxy group, essential for mammalian estrogen receptor binding, from all analogs had no effect on virulence attenuation, indicating that this side chain is not responsible for off-target effects in *P. aeruginosa*. The absence of the 6' hydroxy and the 4' hydroxy in raloxifene made the analog Compound 1 an inactive antivirulence agent. On the contrary, the presence of only the 6' hydroxy in Compound 3 or both hydroxy groups in Compound 2 resulted in comparable antivirulence activities to raloxifene across all assays. Compound 2S, also known as the raloxifene core in Drug Bank, seemed to be toxic to *C. elegans* despite being effective at pyocyanin reduction. These results provided structure-activity information for raloxifene against *P. aeruginosa* and suggested that raloxifene, Compound 2 and Compound 3 may be potent antivirulence agents to alleviate the current challenges of treating antimicrobial resistant *P. aeruginosa* infections.

4.2. Introduction

Emergence of multi-drug resistant strains of many threatening human pathogens is surpassing the rate at which novel antibiotics are being deployed in clinical settings. *P. aeruginosa*, the opportunistic gram-negative bacterium accountable for many nosocomial infections in people with CF or compromised immune system, is one of the top priority pathogens requiring urgent public health interventions and novel therapeutic development (World Health Organization, 2017b). The extensive resistance of this organism due to intrinsic and acquired mechanisms, such as drug efflux, target alternation, challenges the provision of appropriate drug regime for infection treatment (Nguyen et al., 2018). AMR is associated with an alarming increase in morbidity and mortality from healthcare-associated infections, impacting the global burden of

healthcare and economics (World Health Organization, 2017a). Novel therapeutics are urgently needed to treat *Pseudomonas* infections and to prevent the spread of drug resistance among individuals.

Antimicrobial development is further impeded by the shift in pharmaceutical investment from infectious diseases to chronic diseases (Morel & Mossialos, 2010). The collapsing efforts in antibiotics research and development are attributed to the high cost of clinical testing and the low profit generation from infrequent use and short treatment regime of antibiotics compared to the regular administration of cancer and other chronic disease drugs. Additionally, the rapid evolution of resistance shortens the clinical lifespan and thus jeopardizes the return of investment of antimicrobials (Bettiol et al., 2015). To revive the antibiotic development pipeline, more cost-effective and reduced-risk strategies such as drug repositioning, finding new indications for known drugs, are required to encourage pharmaceutical companies to redirect their focus to antimicrobial products (Murteira et al., 2013). An alternative therapeutic approach is antivirulence which attenuates virulence without affecting the viability or the growth of bacterial pathogens. Antivirulence therapeutics inhibit actions of VFs rather targeting bacterial essential functions and theoretically should reduce the evolutionary pressure to select for drug resistance if pathogens no longer require to compete for survival (Zambelloni et al., 2015).

Previously, bioinformatics analyses identified a potentially novel indication for raloxifene, an approved drug for osteoporosis treatment and breast cancer prevention in post-menopausal women, as an antivirulence drug against *P. aeruginosa* (S. J. Ho Sui et al., 2012). From an *in silico* drug-target analysis workflow, raloxifene was predicted to bind to the pathogen-associated pseudomonal PhzB2 protein. Encoded by *phzb2* gene within the *phz2* phenazine biosynthetic operon, PhzB2 is involved in the synthesis of phenazine-1-carboxylic acid which can be subsequently converted to pyocyanin, 1-hydroxyphenazine, and phenazine-1-carboxamide (Mavrodi et al., 2001). Phenazine-1-carboxylic acid and its downstream products in *P. aeruginosa* are the main redox-active VFs important for *P. aeruginosa* growth and survival under iron-deficient environments in a wide range of host organisms. However, these compounds are implicated in host toxicity by inducing oxidative stress that mediates the killing of host cells as well as other microbial competitors in the same niche (Briard et al., 2015). Antivirulence activity of raloxifene was validated in a quantitative pyocyanin assay and a *C. elegans* infection

model. In particular, raloxifene reduced the production of the redox-active pyocyanin, a major *P. aeruginosa* exotoxin, that is well-reported for its destructive impact in CF lungs (Mavrodi et al., 2001).

Raloxifene (DrugBank accession: DB00481) is an approved selective estrogen receptor modulator optimized to bind to human estrogen receptors to elicit tissue-specific and gene-specific responses. Its structure consists of a hydroxyl benzothiophene core with a direct linkage to a phenol and a flexible carbonyl hinge to a phenyl 4-piperidinoethoxy (amine-containing) side chain (Figure 4.1a). Primarily as a selective estrogen receptor modulator, the antiestrogenic property of raloxifene lies within the piperidine ring (nitrogen-containing cyclic side chain) in which the piperidine nitrogen interacts with Asp351 of the human estrogen receptor to modulate downstream estrogen signalling pathways (I. K. Jordan et al., 2002; V. C. Jordan et al., 2015; Levenson & Jordan, 1998).

To better understand the alternative drug indication of raloxifene, as predicted in the Ho Sui et al. paper (S. J. Ho Sui et al., 2012), this study investigates the structure-activity relationship of raloxifene and its antivirulence activity against *P. aeruginosa*, through comparison antivirulence activity among various drug analogs.

4.3. Methods

4.3.1. Bacterial strains and compounds

Wild-type *P. aeruginosa* strain PA14 was used to study the antivirulence properties of raloxifene and its analogs. *E. coli* strain OP50 served as the nematode food source and the non-pathogenic control in the *C. elegans* infection model. Raloxifene was purchased from Cayman Chemical (Ann Arbor, MI). Four analogs of raloxifene were synthesized by Dr. Peter Wilson's synthetic organic chemistry lab (Simon Fraser University, Burnaby, BC). Stock solutions of raloxifene and its analogs were prepared at 20 millimolar (mM) with dimethyl sulphoxide.

4.3.2. Bacterial growth curve assay

Frozen stock culture of *P. aeruginosa* PA14 was streaked onto Luria-Bertani (LB) agar plate for overnight incubation at 37°C. A single bacterial colony on the plate was chosen to inoculate 2 mL of overnight liquid LB culture. To set up the growth assay, the PA14 culture was diluted 1 in 100 into 24 mL of fresh LB. Raloxifene or one of the analogs was added to the culture at a final concentration of 0.2mM. Bacterial growth at 37°C was monitored by optical density at 600nm. Absorbance of the cultures were taken hourly until the 2nd hour then every 30 minutes until the 9th hour. Measurements were blanked with LB broth at each time point. 3 replicates of the growth assay were conducted.

4.3.3. Pyocyanin extraction and quantitative assay

Concentration of pyocyanin pigment was determined spectrophotometrically at the 520nm absorbance as described by Essar et al. (Essar et al., 1990). Bacteria were harvested from 1 mL of a single-colony *P. aeruginosa* PA14 overnight culture. Pyocyanin was extracted from the supernatant with 1mL of chloroform, followed by re-extraction with 1 mL of 0.2M Hydrochloric acid (HCl). Absorbance of the resultant pink solution containing pyocyanin was measured at 520 nm, blanked with equivalent volume of HCl. Concentration was expressed as micrograms of pyocyanin in millilitres of supernatant (ug/mL) by multiplying the absorbance by 17.072. Pyocyanin quantitative assay was done in triplicates.

4.3.4. *C. elegans* infection model

C. elegans (wild-type Bristol N2) were maintained on NGM plates with *E. coli* OP50 as food source. To prepare for the slow-killing assay, low-osmolarity NGM plates were seeded with 50 µL of overnight culture of *P. aeruginosa* PA14, or *E. coli* OP50 as control, per 3.5cm diameter plates for bacterial lawn formation at 37°C overnight. The plates were then equilibrated at 24°C overnight prior to infection. To prevent progeny development during the assay, 100 µg/mL of f-fluoro-2'-deoxyuridine (FUdR) added to the surface of the NGM plates 1 hour prior to the transfer of worms synchronized to the fourth larval (L4) stage. 30 worms were transferred onto each plate and were monitored for viability for up to 144 hours. Worms are considered dead if they no longer moved or

responded to touch. For the drug plates, 0.2 mM of raloxifene or its analogs was incorporated into the NGM prior to bacterial seeding. The experiment ended when the viability of untreated PA14-infected worms dropped below 20%. The killing assay was performed for a minimum of three replicates.

Kaplan-Meier curves of drug-treated and untreated worms infected with *P. aeruginosa* PA14 were estimated and visualized with the R packages “survminer” and “ggsurvplot.” The mean survival curves of PA14-infected worms under the different drug/analog treatments were compared using a log-rank test with a statistically significant p-value threshold of 0.05. Survival curve analysis and visualization were done with the R packages “survminer” and “ggsurvplot.”

4.3.5. Synthesis of raloxifene analogs

Several analogs were synthesized and assessed for their biological effects on *P. aeruginosa* to uncover the structural components of raloxifene that are responsible for virulence attenuation (Figure 4.1). The removal of the antiestrogenic piperidinylethoxy group and subsequent structural modifications in raloxifene (Figure 4.1a) generated the drug analogs used in this study. The 6’ hydroxyl attached to the benzothiophene and the 4’ hydroxy attached to the benzene group of raloxifene were removed in Compound 1 while preserved in Compound 2 (Figure 4.1b and c). The hydroxy benzothiophene remained unchanged in while the 4’ hydroxyl group was removed in Compound 3 (Figure 4.1d). Compound 2 was further simplified into another analog named Compound 2S, also known as the raloxifene Core in DrugBank (Accession: DB08773), in which only the benzothiophene and the phenol groups in raloxifene were retained (Figure 4.1e). Another analog named Compound 4, which would have only the 4’hydroxyl group removed from the benzothiophene, was still being synthesized at the end of my thesis work, so could not be included in this study.

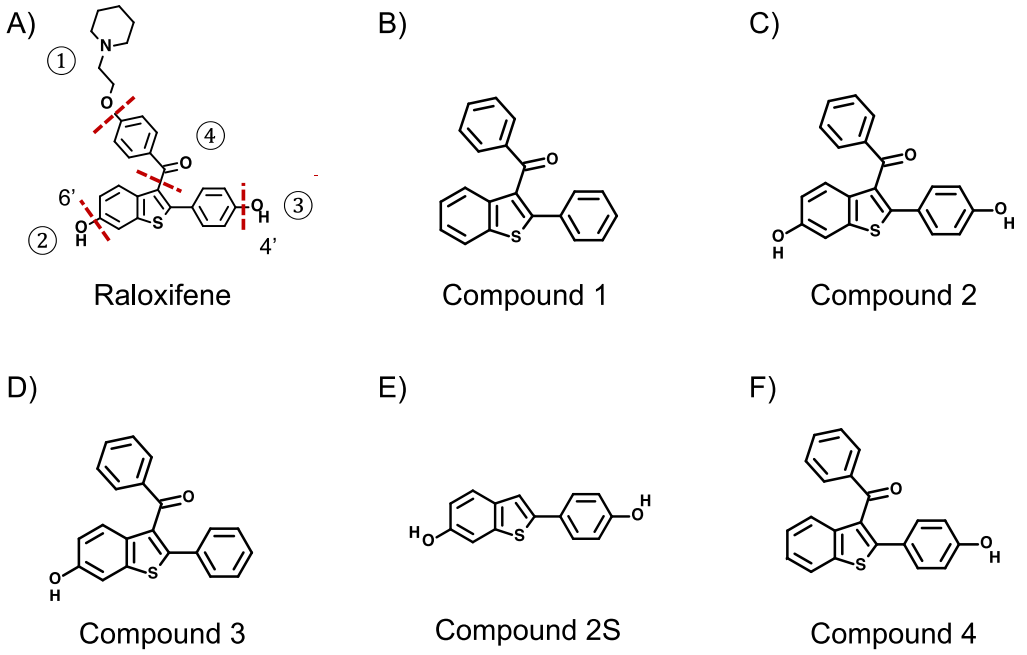


Figure 4.1 Chemical structure of raloxifene and the analogs used in this study. **A)** raloxifene - the FDA-approved selective estrogen receptor modulator for the prevention and treatment of osteoporosis in postmenopausal women. Numbers in circle represents position at which modifications were made in the analogs. Piperidinylethoxy group at 1 was removed from all analogs. Additional modifications were made as follows: **(B)** Compound 1 – removal of hydroxy (-OH) at 2 and 3; **(C)** Compound 2 – no additional modifications made; **(D)** Compound 3 – removal of the 4' hydroxyl only at 3; **(E)** Compound 2S – Removal of 1 and 4; **(F)** Compound 4 (synthesis still in process thus not included in the analysis) – removal of the 6' hydroxyl only at 2. Red dotted line represents cleavage site.

4.4. Results

4.4.1. Compound 1, compound 2 and compound 3 showed minimal impact on the growth of *P. aeruginosa*

Raloxifene analogs, except Compound 2S which showed toxicity in *C. elegans*, were screened for effects on bacterial growth. *P. aeruginosa* PA14 was grown in LB with no drug, 0.2 mM of raloxifene, Compound1, Compound 2 or Compound 3. The growth curves were similar across all culture conditions, suggesting that these raloxifene analogs had minimal effect on the growth of *P. aeruginosa* PA14 during the exponential growth phase at 37°C (Figure 4.2). This assay suggested that these analogs do not target essential functions of *P. aeruginosa* under our experimental conditions.

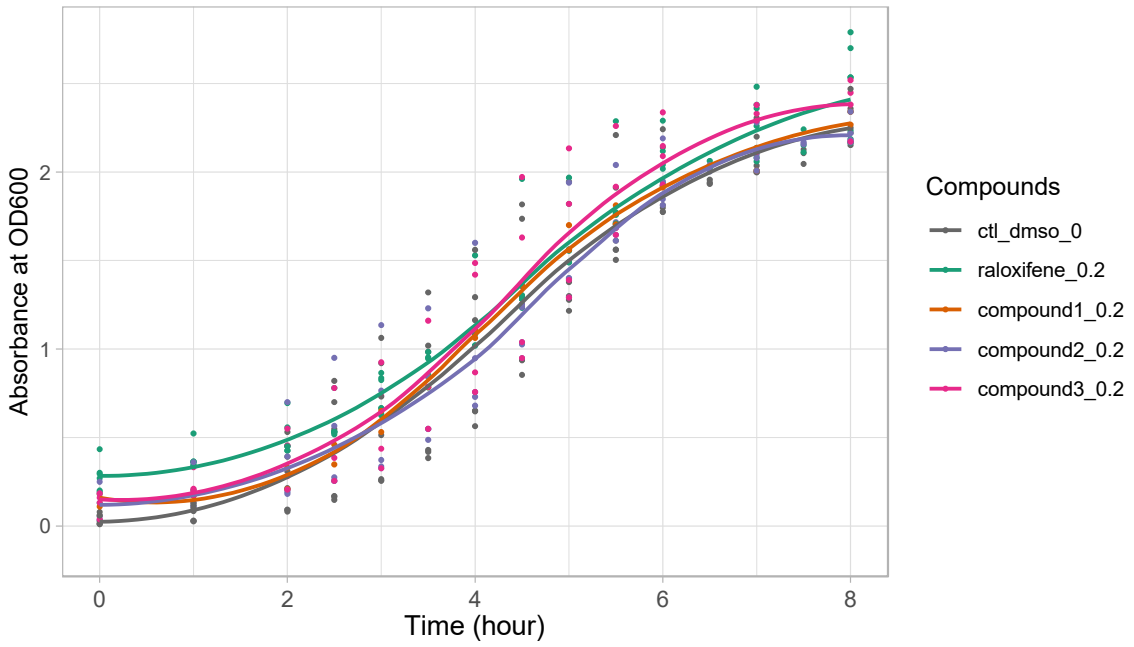


Figure 4.2 Growth curve assay of *P. aeruginosa* PA14 treated with raloxifene or its analogs showed that Compound 1, Compound 2 and Compound 3 had minimal impact on bacterial growth.

DMSO was used as negative control. The remaining samples were treated with 0.2 mM of raloxifene, Compound 1, Compound 2 or Compound 3. Compound 2S was excluded from this analysis due to the observed toxicity in the worm infection model (Figure 4.4).

4.4.2. Analogs preserving at least one of the two hydroxyl groups in raloxifene (compound 2 and compound 3) significantly reduced pyocyanin production in *P. aeruginosa*

In a quantitative assay, the level of pyocyanin production in overnight culture supernatant of *P. aeruginosa* was compared after treatment with raloxifene and its analogs (Figure 4.3). Administration of 0.2 mM raloxifene as well as 0.2mM and 0.1mM Compound 2 reduced the pyocyanin concentration by half, compared to the untreated *P. aeruginosa* sample. There was no statistically significant difference among these compounds which contain both the hydroxyl groups. Titration of Compound 2 from 0.025mM up to 0.2mM reduced pyocyanin concentration in a dose-dependent manner. PA14 culture treated with Compound 1, with all hydroxyl groups removed, showed only a slight reduction in pyocyanin production relative to the negative control. Compound 2S, the simplified version of Compound 2 with only the hydroxyl-containing benzothiophene moiety and the phenol, was more effective than Compound 1 but less potent than raloxifene at pyocyanin reduction. Contrary to the observation that the presence of both hydroxyl groups is necessary for raloxifene's antivirulence activity as seen in Compound

2 versus Compound 1, Compound 3 with only the 6' hydroxyl group, also drastically lowered pyocyanin production in *P. aeruginosa*, with no statistically significant difference to raloxifene and Compound 2.

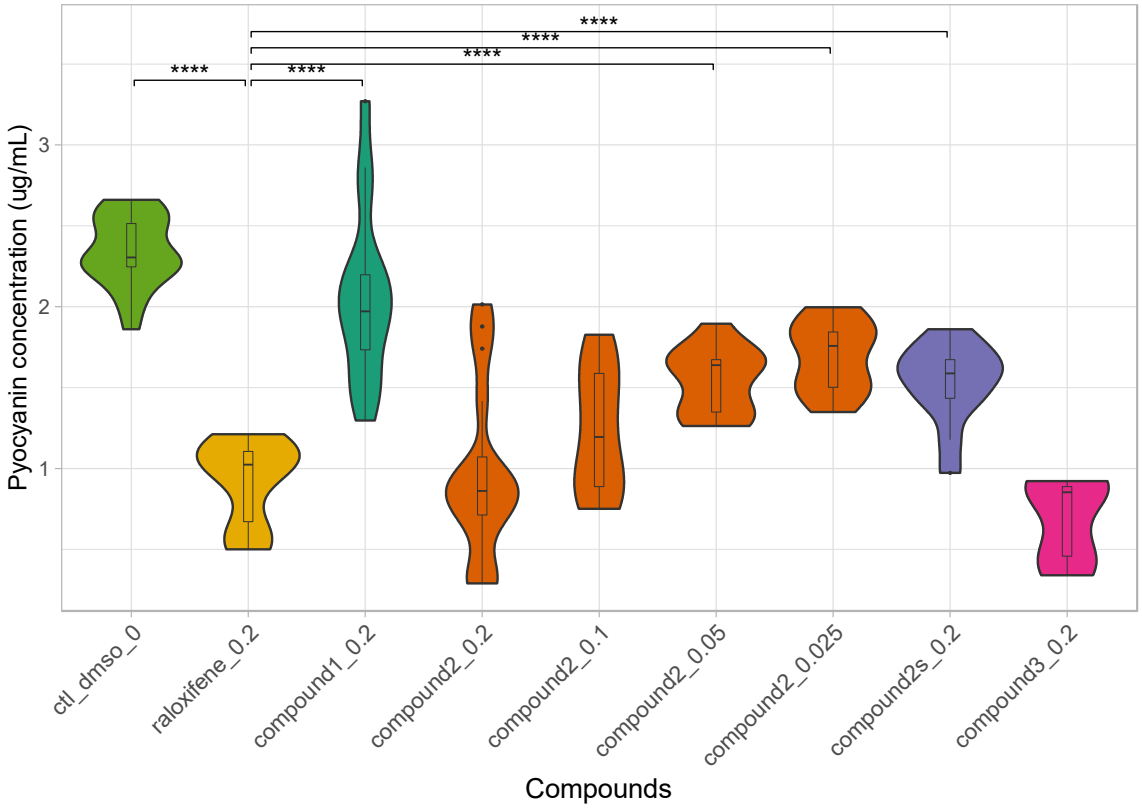


Figure 4.3 Quantitative chemical assay of pyocyanin production in *P. aeruginosa* PA14 cultured with raloxifene and its analogs showed that Compound 2 and Compound 2 reduced pyocyanin production to a level comparable to raloxifene.

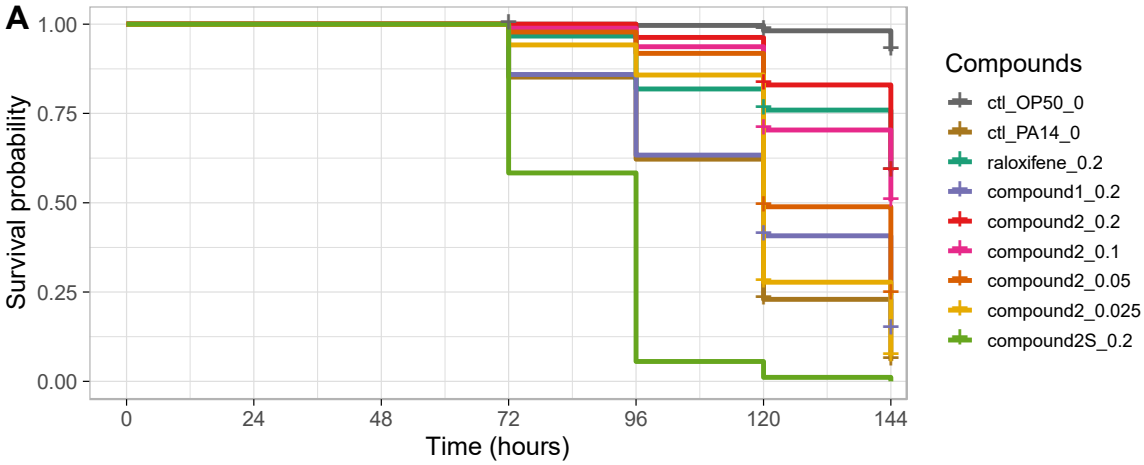
Left to right: negative control with dimethyl sulfoxide, raloxifene (0.2 mM), Compound 1 (0.2 mM), Compound 2 (0.2 mM, 0.1 mM, 0.025 mM and 0.025 mM), Compound 2S (0.2 mM) and Compound 3 (0.2 mM). Statistically significant P-values (<0.05) of the unpaired two-samples tests involving raloxifene are indicated by asterisks (*) as follows: P-value ≤ 0.0001 (****) and 0.0001 < P ≤ 0.001 (***).

4.4.3. Compound 2 and compound 3 showed comparable improvement in survival of *Pseudomonas*-infected worms, relative to raloxifene

Virulence attenuation by raloxifene analogs was assessed in a *C. elegans* slow-killing assay, mimicking the natural infection process in which the worms feeding on minimal medium were killed over several days with a gradual intestinal accumulation of *P. aeruginosa* PA14. Corresponding to the notable reduction in pyocyanin production, no

significant pairwise difference in the survival of PA-14 infected worms was observed between treatment with 0.2mM of raloxifene and Compound 2 or Compound 3 (Figure 4.4). These three compounds provided the most protection of worms from *P. aeruginosa*. Worms were also treated with Compound 2 at 0.2 mM, 0.1 mM, 0.05 mM and 0.025 mM and showed a dose-dependent improvement in survival with higher drug concentrations. Compound 1 provided a low level of protection of worms against *P. aeruginosa* infection, as evident by the slight improvement in the survival relative to the infected yet untreated worm with a survival probability of less than 20% at 144 hours. Compound 2S strikingly showed signs of toxicity in *C. elegans* as the worm survival dropped more quickly than in infected worms not treated with any compounds. The survival drastically dropped to below 10% when treated with Compound 2S while the survival of untreated worms or worms treated with other compounds stayed above 50% at the 96-hour time point.

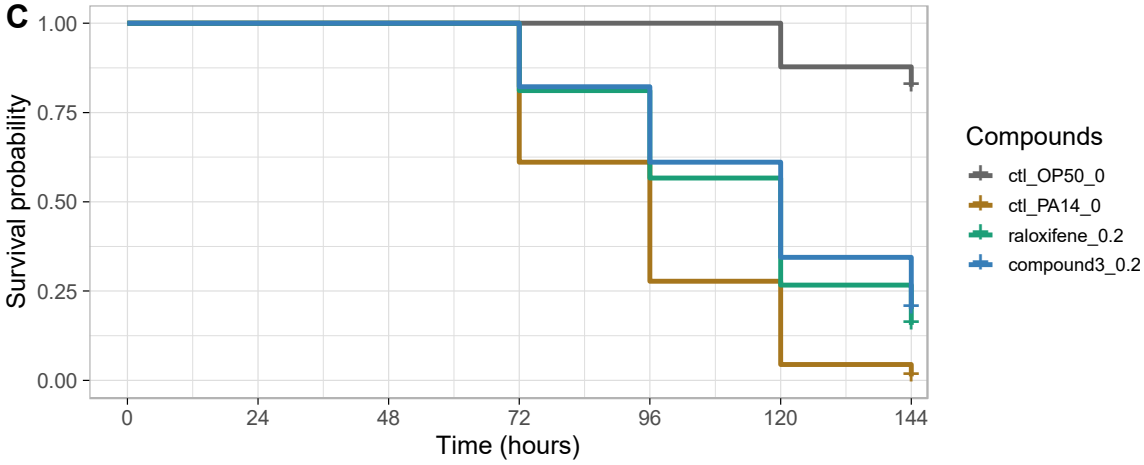
A



B

	ctl_OP50 0	ctl_PA14 0	raloxifene 0.2	compound1 0.2	compound2 0.2	compound2 0.1	compound2 0.05	compound2 0.025
ctl_PA14_0	1.54e−86	NA	NA	NA	NA	NA	NA	NA
raloxifene_0.2	2.29e−16	3.47e−38	NA	NA	NA	NA	NA	NA
compound1_0.2	8.01e−61	5.96e−04	1.58e−21	NA	NA	NA	NA	NA
compound2_0.2	7.15e−13	6.54e−53	1.12e−01	7.78e−31	NA	NA	NA	NA
compound2_0.1	1.33e−21	2.61e−38	2.78e−01	1.06e−19	3.49e−03	NA	NA	NA
compound2_0.05	1.58e−46	5.24e−18	5.18e−10	7.93e−06	1.87e−17	2.09e−08	NA	NA
compound2_0.025	1.69e−78	1.20e−04	4.97e−28	7.11e−01	6.75e−43	5.30e−28	5.24e−09	NA
compound2S_0.2	4.55e−109	9.69e−33	6.31e−71	2.64e−36	8.67e−96	1.54e−86	3.13e−77	6.11e−63

C



D

	ctl_OP50_0	ctl_PA14_0	raloxifene_0.2
ctl_PA14_0	1.12e−36	NA	NA
raloxifene_0.2	9.54e−22	9.74e−07	NA
compound3_0.2	8.12e−19	1.66e−08	3.54e−01

Figure 4.4 Kaplan-Meier curve with log-rank tests of *Pseudomonas*-infected *C. elegans* under treatment with raloxifene or analogs showed that Compound 2 and Compound 3 are as effective as raloxifene in improving worm survival

Positive control worms were fed on *E. coli* OP50. Worms infected with *P. aeruginosa* PA14 but not treated any compounds were used as a negative control. Numbers following the compound

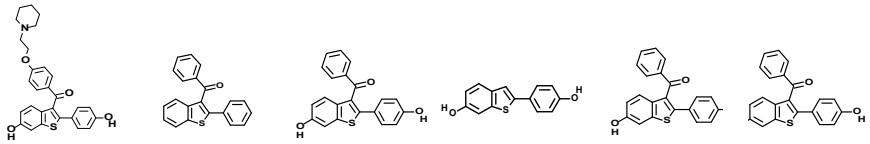
name and underscore represent concentration in mM (e.g., “raloxifene_0.2” represents raloxifene at 0.2 mM). **(A)** Survival curve of raloxifene (0.2 mM), Compound 1 (0.2 mM), Compound 2 (0.2-0.0025 mM) and Compound 2S over 6 days. **(B)** P-values table generated from multiple pairwise comparisons of the survival curves in A) using the log-rank test. P-values were corrected for multiple testing with the Benjamin–Hochberg (BH) procedure in the R package “survminer” v0.4.8. **(C)** Kaplan-Meier survival curve of raloxifene (0.2 mM) and Compound 3 (0.2 mM). **(D)** P-value table generated from multiple pairwise comparison of survival curves in C using the same method as described in B).

4.5. Discussion

This study investigated the structure-activity relationship of raloxifene in the context of antivirulence against *P. aeruginosa* infection. Multiple analogs of raloxifene have been synthesized and assessed in biological assays including the pyocyanin quantitative analysis, *C. elegans* slow-killing assay and growth curve assay (Figure 4.5). To test the hypothesis that the side chain(s) responsible for raloxifene’s antivirulence properties are different from that conferring antiestrogenic property important for osteoporosis treatment, the piperidinylethoxy group thought to be associated with the latter was excluded in the analogs. Compound 1, Compound 2, Compound 2S and Compound 3 each represent a unique modification to the remaining structure of raloxifene. Compound 1, with both the 6’ and 4’ hydroxyl groups removed, overall showed the weakest inhibition of pyocyanin production as well as the least improvement of worm survival upon *P. aeruginosa* infection. Compound 2S, although reduced pyocyanin production at a level comparable raloxifene and Compound 2, accelerated the death of *Pseudomonas*-infected worms relative to infected worms with no drug treatment.

Compound 2 (with both hydroxy groups in raloxifene) and Compound 3 (with only the 6’ hydroxyl group in raloxifene) showed significantly similar antivirulence properties as raloxifene. These observations suggest that at least one of the non-piperidinylethoxy side chains, is responsible for the virulence attenuation in *P. aeruginosa*. The drug repurposing potential of raloxifene as an antivirulence drug is likely related to its ability to interfere with the biosynthesis of phenazines, particularly the major *P. aeruginosa* VF, pyocyanin which was measured in the quantitative pyocyanin assay. Pyocyanin and phenazine-1-carboxylic acid are two major phenazines produced by *P. aeruginosa* PA14. Due to their redox potentials, they act as redox mediators (i.e., electron shuttling) to allow the oxidation of major intracellular reductants (such as NADH and glutathione) (Glasser et al., 2017). The subsequent reduction of extracellular oxidants triggers the

formation of reactive oxygen species that may cause cell damage (Blankenfeldt & Parsons, 2014; Hall et al., 2016). Since Compound 4 (with only the 4' hydroxyl group in raloxifene) was still in the synthesis step at the end of this thesis work, its antivirulence property could not be compared to the other analogs to determine whether specifically the 6' hydroxyl group or either one hydroxyl group (4' or 6') was required to elicit an antivirulence activity. While the initial study of raloxifene as an antivirulence agent against *P. aeruginosa* computationally predicted that raloxifene binds to the PhzB2 protein involved in the phenazine biosynthetic pathway, the exact bacterial target of raloxifene has not been verified *in vivo*. Moving forward, a ligand binding assay of raloxifene, Compound 2 and Compound 3 against *P. aeruginosa* proteins would be useful in elucidating the drug-target interaction important for the antivirulence activity of these candidate antipseudomonal agents.



	Raloxifene	**Compound 1**	**Compound 2**	**Compound 2s**	**Compound 3**	**Compound 4**
Compound synthesis	Completed	Completed	Completed	Completed	Completed	In progress
Impact on bacterial growth	No	No	No	ND	No	ND
Reduced pyocyanin production	Yes	No	Yes	Yes	Yes	ND
Improved survival of infected worms	Yes	No	Yes	No	Yes	ND

Figure 4.5 Results summary of the structure-activity relationship analysis of raloxifene as an antivirulence agent against *P. aeruginosa*.
Similar to raloxifene, Compound 2 and Compound 3 showed the best antivirulence properties, including minimal impact on bacterial growth, reduction of pyocyanin production in overnight *P. aeruginosa* cultures and improved survival of *P. aeruginosa*-infected *C. elegans*. Although Compound 1 did not impact bacterial growth, it also did not reduced virulence in *P. aeruginosa* in the pyocyanin assay and *C. elegans* infection model. Compound 2S, though reduced pyocyanin production, did not improved the survival of infected *C. elegans*. Compound 4 was still being synthesized at the time of this study so was not included in the laboratory assays. ND: not determined.

The triphenylethylene class of tamoxifen-related selective estrogen receptor modulators, including raloxifene, has been characterized as a group of broad-spectrum drugs effective not only against mammalian targets, but also microbial targets in fungal, viral, parasitic and bacterial pathogens (Montoya & Krysan, 2018). Tamoxifen is bacteriolytic to *Enterococcus faecium* and *Acinetobacter baumannii* (Jacobs et al., 2013)

and synergizes with some antibiotics to restore susceptibility of antimicrobial resistant *P. aeruginosa* (M. Hussein et al., 2018; M. H. Hussein et al., 2017). While previous studies showed antibacterial (bacterial killing) effect of some SERMs, this structure-activity relationship study of raloxifene, a tamoxifene-related SERM, instead showed a virulence attenuation in *P. aeruginosa* without affecting the survival of the pathogen, which represents a potentially novel antivirulence agent against *P. aeruginosa*.

In conclusion, the work outlined this thesis chapter illustrates the possibility of combining the comprehensive bacterial PAGs analysis described in Chapter 2 and a computational drug screening platform to effectively identify novel antivirulence indication in FDA-approved drugs that may be able to target pathogens-specific bacterial proteins (encoded by PAGs). Though a combination of bacterial growth assay, quantitative pyocyanin assay and *C. elegans* infection model, I uncovered the antivirulence-associated structural components of the FDA-approved raloxifene that was identified as a candidate drug for repurposing as antivirulence agent against *P. aeruginosa* in a previous study. The wealth of bacterial genomic data and pharmaceutical data, coupled with improved *in silico* PAG analysis and drug-target interaction analysis, now enables quicker discovery of candidate drug and drug targets. Effective treatment of bacterial infections while minimizing the risk of AMR emergence requires the development of therapeutics alternative to conventional antimicrobials. Exploring novel indications of existing drugs (i.e., drug repurposing) like raloxifene may be a promising approach to address the unmet clinical need of novel therapeutics against bacterial infections. Repurposing existing drugs as precise, antivirulence agents that disarm rather than kill bacterial pathogens may help to curb the continual rise in AMR and to accelerate the development of more sustainable therapeutic solutions for infectious disease control.

Chapter 5.

Metagenomics-based PAG detection

This chapter presents the first metagenomics-based analysis of PAGs. This work used the publicly available human lung and freshwater microbiome datasets from the peer-viewed articles "Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality" published by the Brinkman Lab and its collaborators (Van Rossum et al., 2015) and "Sputum DNA sequencing in CF: non-invasive access to the lung microbiome and to pathogen details" published by Feigelman et al. (Feigelman et al., 2017), respectively. I performed all analyses of these datasets, from sequence read processing to PAGs detection and data visualization.

I performed all work presented in this chapter with the following exception: sample collection and sequencing of both metagenomic datasets.

5.1. Abstract

Metagenomics enables the assessment of both taxonomic composition and gene content within microbial communities. While PAGs have been identified from high-quality NCBI RefSeq genomes, evaluating their prevalence in clinical and agricultural environments would provide insight into the potential, real world application of PAGs as biomarkers for monitoring pathogen transmission and drug targets for antivirulence therapeutics. This chapter consists of metagenomic analyses of PAGs in several lung microbiomes as well as freshwater microbiomes. Comparison of lung microbiomes of CF patients, chronic obstructive pulmonary disease (COPD) patients, smokers and healthy individuals revealed that the CF lungs are disproportionately enriched in PAGs. On the contrary, no significant difference in the prevalence of PAGs was detected between pristine watersheds and watersheds at or downstream of agricultural sites with fecal contamination, likely due to the abundance of uncharacterized bacterial species in freshwater. Both analyses revealed challenges in understanding the functional role of PAGs at the enriched sites (CF lungs, agriculturally contaminated and downstream sites) owing to the lack of characterization of these genes. Based on this analysis, pathogen surveillance using PAGs may currently be limited to more well-characterized microbiomes, such as those sampled from clinical settings. Ongoing advancement in metagenomic sequencing technologies and gene characterization methods would be beneficial for identifying important virulence determinants from the detected PAGs, especially those found in novel or uncharacterized environmental bacteria, that may be important for infectious disease surveillance.

5.2. Introduction

Metagenomics, also known as “community genomics,” is the study of genetic materials of microbial populations in a culture-independent manner (Handelsman, 2004). This approach provides the taxonomic profile and functional potential of the microbiome of interest through sequencing the total DNA within a sample. While traditional public health pathogen detection relies on culture-based assays, shotgun metagenomic sequencing now enables the detection of the complete bacterial population and gene set within an environmental sample. Without *a priori* knowledge of the microbial constituents, metagenomics is capable of detecting rare and emerging pathogens,

including those that are not culturable under laboratory conditions (Miller et al., 2013). Metagenomics-based functional profiling may also provide insight into the risk of the acquisition and dissemination of AMR and virulence-related genes within microbial communities (de Abreu et al., 2020; M. Zhou et al., 2021). Metagenomics has a broad range of application from clinical diagnoses of bacterial infections, environmental pathogen surveillance to agricultural assessment of soil health (Bengtsson-Palme et al., 2017; Moragues-Solanas et al., 2021; Orellana et al., 2018).

CF is a common autosomal genetic disease most prevalent in North America, Europe and Australia (Elborn, 2016). It is characterized by thick mucus build-up in the human airway, due to a mutation in the cystic fibrosis transmembrane regulator (CFTR) responsible for chloride ion transport across the epithelial cell membrane. The impaired CFTR predisposes CF lungs to bacterial infections which often negatively impact the prognosis of CF patients (Coutinho et al., 2008). Metagenomics is a useful tool for understanding microbial dynamics in the CF respiratory tract to inform proper management of CF-related infections. The CF lung is a hostile environment for bacterial colonizers owing to mucus-induced osmotic stress, host immune response and antibiotic exposure (Winstanley et al., 2016). This creates a competitive environment in which the initial colonization by *S. aureus* and *Hemophilus influenzae* is often replaced by *P. aeruginosa* and *Burkholderia cepacia* after the first decade of life of CF patients (Coutinho et al., 2008). Subsequent chronic CF lung infections are often associated with biofilm formation as well as other genetic adaptations to the CF lungs, such as loss of VFs and development of AMR in *P. aeruginosa*. Uncovering the microbiome composition, virulence determinants and bacterial adaptive mechanisms in CF lungs is thus critical to the improvement of the prognosis of individuals with CF.

Multiple CF respiratory tract metagenomic studies have investigated the microbial communities and functional potential the CF lungs of adults and children using throat swabs, bronchoalveolar lavage or sputum (Feigelman et al., 2017; Kirst et al., 2019; Pust et al., 2020). In one study comparing the lung microbiome of CF patients, chronic pulmonary obstructive disease (COPD) patients, smokers and healthy individuals (Feigelman et al., 2017), healthy subjects had the highest lung microbiome diversity with many species from known oral flora. In contrast, the CF lung microbiome was the least diverse, with colonization by one or a few bacterial colonizers such *Pseudomonas*, *Staphyloccocus*, *Strenophomonas* and *Achromobacter* species. Lung

microbiome diversity was slightly reduced in smokers and varied drastically among COPD patients relative to healthy participants. Pathogen identification and AMR characterization of the CF lung microbiome were congruent between clinical culture diagnosis and sputum metagenomics analysis, suggesting that the latter may represent a more informative, non-invasive procedure in the diagnosis and surveillance of CF-related infections. As an extension to this published study, I compared the prevalence of PAGs, non-PAGs and common genes in the lung microbiome among the four participant groups. The hypothesis of this follow-up study is that CF lungs are enriched in PAGs, of which some may represent novel virulence-related genes important for the bacterial pathogenesis in CF-related respiratory infections.

Healthcare-associated infections by multidrug resistant pathogens remains a threat to morbidity and mortality of patients, such as those with CF (Haque et al., 2018; Ogunsola & Mehtar, 2020). Common nosocomial pathogens include Vancomycin-resistant enterococci, methicillin-resistant *S. aureus*, *Clostridium difficile* and *P. aeruginosa* (Suleyman et al., 2018). Those capable of forming endospores are even more persistent in hospital environments They are commonly acquired through cross contamination between patients and healthcare workers or environmental sources such as medical devices, faucets and tap water. Biofilm-forming *P. aeruginosa* is one of the common risks of nosocomial infections due to its persistence in hospital water systems or on surfaces for an extended period (Lalancette et al., 2017; Moore et al., 2021). Identification of the environmental reservoirs of nosocomial pathogens is also key to the design and implementation of effective infection control strategies. This is now possible with the use of metagenomics under the One Health approach.

Defined by the US Centers for Disease Control and Prevention as “a collaborative, multisectoral, and transdisciplinary approach …with the goal of achieving optimal health outcomes recognizing the interconnection between people, animals, plants, and their shared environment,” One Health has been gaining attention in recent years in the management of emerging and multidrug resistant infectious diseases (Evans & Leighton, 2014). It focuses on the animal-human-environmental interface for the surveillance and control of pathogen transmission. There is increasing evidence of AMR transmission from livestock to clinical setting as a consequence of the indiscriminate use of clinical and veterinary antibiotics (Essack, 2018; Mackenzie & Jeggo, 2019). Soil and water are considered reservoirs for AMR genes acquired through

livestock and human waste. Environmental exposure to antibiotics may also impact the natural soil and water microflora, thus exerting selective pressure for resistance development (Kim & Cha, 2021). Though most studies have focused mainly on AMR transmission at the environmental-clinical interface, parallel phenomenon can be hypothesized for the dissemination of virulence-related genes among bacterial pathogens.

Watersheds, also known as drainage basins, are land areas to which all freshwater streams and rivers drain into a common outlet and are a major source of drinking and irrigation water. Monitoring the wellbeing of the watershed ecosystem ensures the safety in water consumption and thus helps to maintain human and animal health. Bacterial ecosystem and pathogen transmission are governed by environmental fluctuations in temperature, weather, and run-offs. In Genome Canada's Applied Metagenomics of the Water Microbiome project, rainfall was shown to have a large impact on bacterial diversity and composition at agricultural watersheds due to its effect on microbial and nutritional transport overland or within water stream (Van Rossum et al., 2015; Van Rossum et al., 2018). A temporal and spatial metagenomic analysis within this Watershed project revealed that the transition from the dry to rainy season in British Columbia, Canada, was accompanied by an increase in alpha diversity (the mean species diversity at a given site) in watersheds at the site as well as downstream of agricultural activity/contamination, suggesting potential run-offs from upstream of the agricultural contamination site induced by the increased rainfall (Peabody, 2017). In the same study, an increase in *Pseudomonadaceae*, the bacterial family containing a diverse range of environmental and pathogenic species like *P. aeruginosa*, was also observed in the agricultural contaminated and downstream sites in the rainy season.

Incorporating the concept of One Health in monitoring and controlling the transmission of infectious disease, this chapter presents the first metagenomics-based analyses to examine the prevalence of PAGs in both the lung microbiome dataset (Feigelman et al., 2017) and the Genome Canada's watershed microbiome dataset (Peabody, 2017; Van Rossum et al., 2015) . These analyses explored whether enrichment of PAGs is associated with specific lung conditions or impacted by rainwater run-offs in agricultural watersheds. Specifically, I tested the hypotheses that PAGs are more prevalent 1) in CF lungs that are more commonly prone to infections and 2) at agriculturally affected watersheds where fecal contamination from agricultural activity is

most prominent, compared to the upstream and downstream watershed sites. In the watershed microbiome dataset, I also hypothesized that the transition from drier to rainy season would cause an increase in PAGs in the downstream watershed site, given these genes are enriched at the agriculturally affected site. The goal of this chapter was to evaluate the applicability of PAGs as biomarker for poor CF prognosis or environmental pathogen transmission.

5.2.1. Shotgun metagenomics datasets

Lung microbiome from sputum DNA

Metagenomic sequences were retrieved from a published lung microbiome study (Feigelman et al., 2017) through the NCBI Sequence Read Archive (BioProject ID: 316588). In brief, sputum DNA samples were collected from 17 participants: 6 CF patients, 4 patients with COPD, 3 smokers and 4 healthy non-smokers. Sputum was produced spontaneously by CF and COPD patients and was induced by hypertonic saline nebulization in the smokers and healthy controls.

Freshwater microbiome from Canadian watersheds

Freshwater samples were collected from agricultural watersheds across southwestern British Columbia, Canada, as described in a previously published analysis from the Genome Canada Watershed Project (Van Rossum et al., 2015). Raw sequences were retrieved from the NCBI Sequence Read Achieve (BioProject ID: 287840). Agricultural watershed samples were collected at a highly irrigated and farmed floodplain (APL), a site upstream (AUP) and a site downstream (ADS) of APL between April 2012 and April 2013 (Figure 5.1, Table 5.1). Based on the collection date, samples were categorized into “drier season” (May to October) or “rainy season” (November to April).

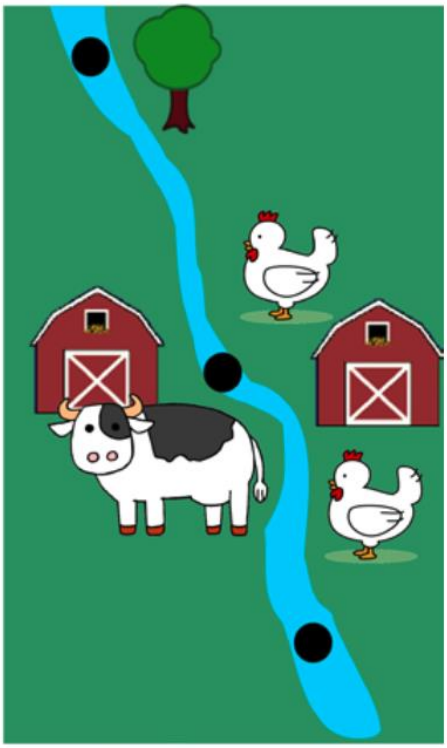


Figure 5.1 Sampling sites within the agricultural watershed.
The agricultural watershed contains three sampling sites (black circles): a site upstream of agricultural pollution, a polluted site surrounding by intense agricultural activity, and a downstream site. Figure adapted from Dr. Mike Peabody's PhD thesis (Peabody, 2017).

Table 5.1 Description of sampling sites across the agricultural watershed.

Watershed site name	Catchment land use	Description
Agri-Upstream (AUP)	Forest and minimal housing	Upstream of agricultural "pollution". Not affected by agricultural activity. Collected from a small rocky stream near the base of a forested hill with minimal housing nearby.
Agri-Pollution (APL)	Agriculture	At site of agricultural "pollution". Collected from a slough in an intensely farmed and irrigated floodplain with minimal tree cover. AUP is upstream of floodplain, separated by 9 km.
Agri-Downstream (ADS)	Agriculture and some urban	Downstream of agricultural "pollution". Collected from a river fed by an agricultural floodplain (site of APL) as well as a separate tributary from a more distant agricultural and urban area. Minimal tree cover throughout catchment. ADS is 2.5 km from APL.

Adapted from (Van Rossum et al., 2015) with permission.

5.2.2. Sequence processing and assembly

Sequence reads from both datasets were processed in the same manner. After read quality assessment using FASTQC v0.11.9 (Trivedi et al., 2014), low quality bases and adapter sequences were trimmed from the paired-end raw reads with Trimmomatic v0.39 under the default parameters (Bolger et al., 2014). Human contaminants were removed by discarding reads that aligned to the human reference genome GrCh38 (hg38) using Bowtie2 v2.4.2 (Langmead & Salzberg, 2012). The remaining paired-end reads were merged by PEAR v0.9.6 and assembled into contigs by metaSPAdes (SPAdes v3.15.1 with the "--meta" option) (Nurk et al., 2017).

5.2.3. Pathogen-association and functional analyses

Gene prediction was done using MetaProdigal (Prodigal v2.6.3 under the metagenomic setting) (Hyatt et al., 2010), followed by the categorization of bacterial genes as PAGs, non-PAGs or common genes (found in both pathogens and non-pathogens) based on a protein sequence similarity search, using DIAMOND v2.0.9 (minimum 90% identity; e-value of 10^{-7}) (Buchfink et al., 2021), against the genes with pathogen-association computed in the PAGs analysis update on the complete NCBI RefSeq bacterial genome dataset in Chapter 2.3.1. Using the same DIAMOND-based sequence similarity search (same parameters as described earlier), these genes were subsequently assessed for VF association based on the Virulence Factor Database as of January 2019 (B. Liu et al., 2019) and were classified into 26 functional categories based on the COG database updated in 2020 (Galperin et al., 2021).

5.2.4. Statistical analysis

Due to the varying taxonomic diversity across some samples particularly in the lung microbiome dataset, gene counts in each pathogen association and functional category were normalized by the number of unique lowest common taxonomic nodes, defined by the NCBI RefSeq non-redundant protein record linked to each gene (NCBI, 2017). The normalized gene counts were then averaged across the replicates within each sample groups. Over-representation of PAGs and non-PAGs, relative to common genes was assessed by performing the Fisher exact test on a contingency table constructed from the averaged and normalized gene counts in each pathogen

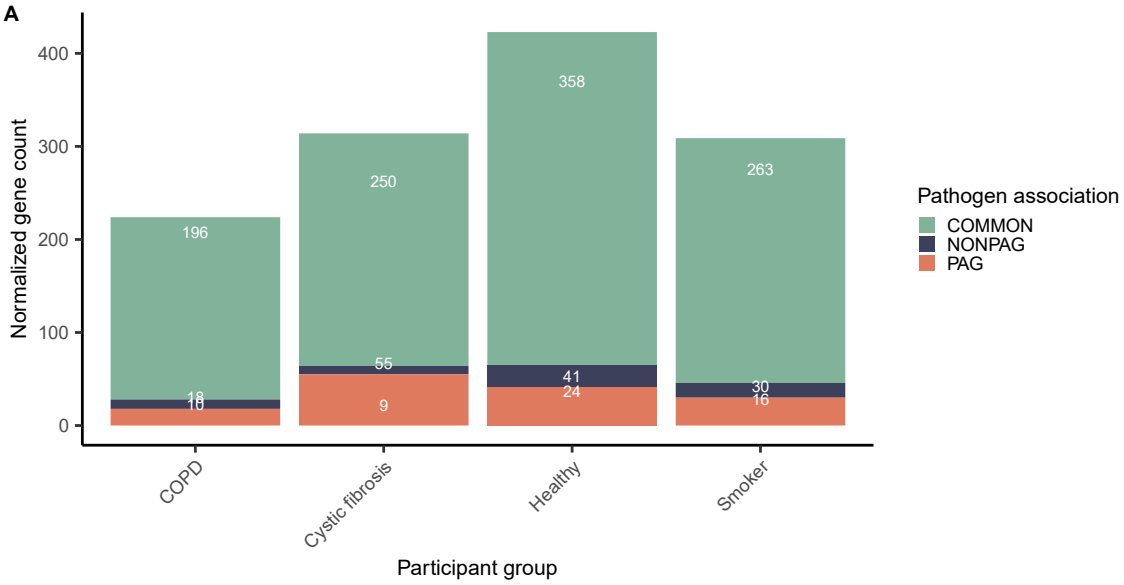
association and in each sample group (participant groups in the lung microbiome dataset and collection sites in the watershed dataset). A post-hoc pairwise Fisher's exact test, with p-values adjusted by the Benjamini–Hochberg false discovery rate method, was performed to identify whether PAGs or non-PAGs were disproportionately represented in certain sample types. Only the adjusted p-values in the post-hoc pairwise Fisher's exact tests, for any statistically significant association, were reported in the Results Section. Statistical analyses were done using the R "stats" v4.0.2 and the "rcompanion" v2.3.26 package.

5.3. Results

5.3.1. Lung microbiome of CF patients is disproportionately enriched in PAGs

Based on clinical culture diagnoses and microbiome analyses, the original study of this lung microbiome dataset reported that CF lungs harbour a less diverse microbial community and are commonly dominated by a single or a few bacterial colonizers such as *Pseudomonas*, *Staphylococcus* and *Streptococcus* species (Feigelman et al., 2017). To avoid any potential confounding effects due to the varying taxonomic diversity across CF patients, chronic pulmonary obstructive disease (COPD) patients, smokers and healthy participants, gene counts were normalized by the number of unique taxa within each pathogen association in each sample. Using Fisher's exact test on a 4x3 contingency table constructed with the pathogen association of genes (pathogen-associated, non-pathogen-associated or common) on one axis and the participant groups on another, a statistically significant association was detected between pathogen association of genes and the health status of the study participants (P=0.008) (Figure 5.2A). The post-hoc pairwise Fisher's exact tests with multiple testing correction, comparing differences in gene count between common genes and PAGs or non-PAGs indicated that CF lungs are associated with higher proportion of pathogen-association genes relative to common genes than COPD patients (P=0.011), smokers (adjusted p-value =0.014), and healthy individuals (adjusted p-value =0.011) (Figure 5.2B).

A



B

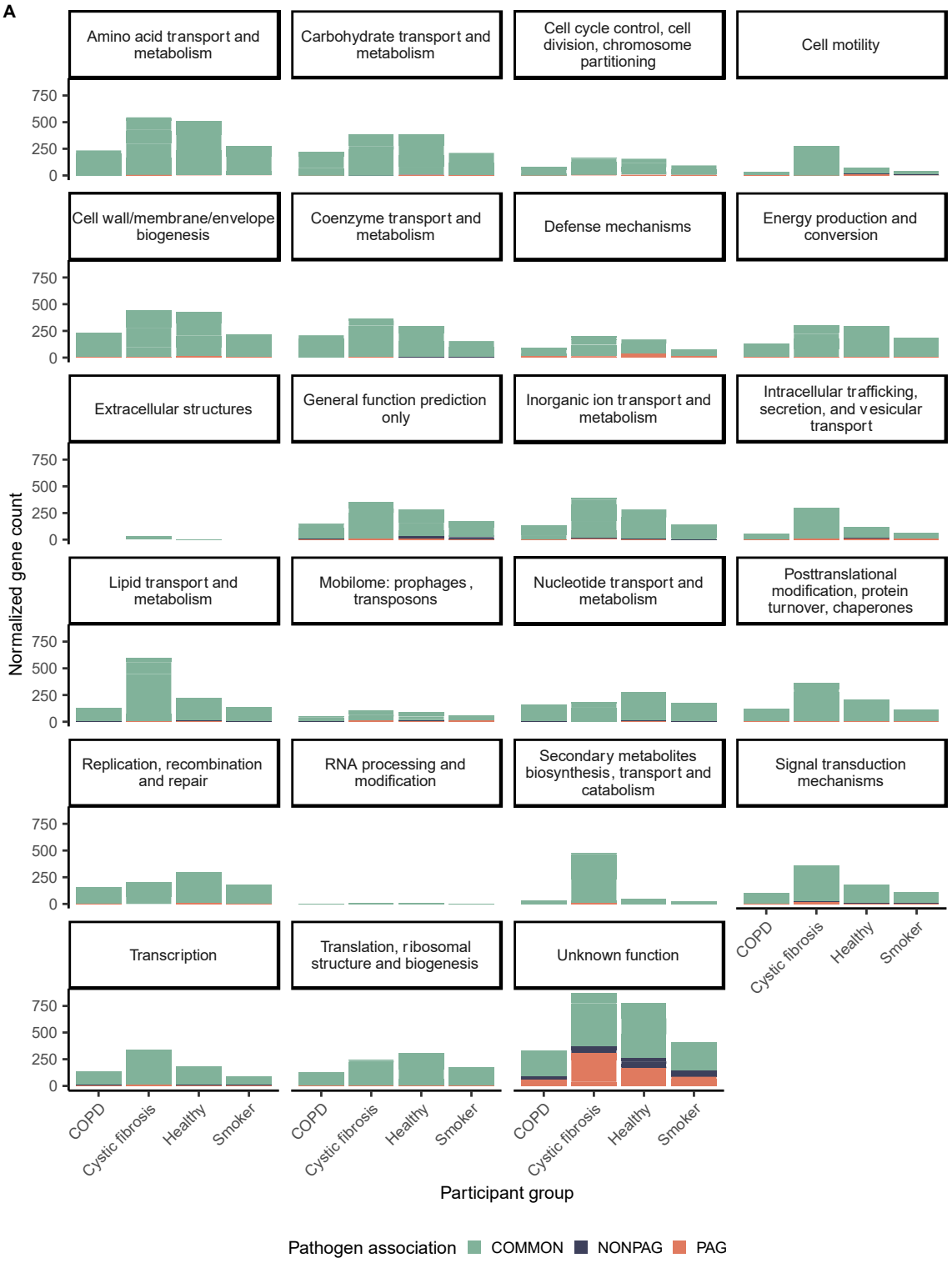
Comparison	p.Fisher	p.adj.Fisher	Pathogen_association
COPD : Cystic fibrosis	0.00197	0.0114	COMMON−PAG
Cystic fibrosis : Healthy	0.00381	0.0114	COMMON−PAG
Cystic fibrosis : Smoker	0.00698	0.0140	COMMON−PAG
COPD : Healthy	0.56600	0.6790	COMMON−PAG
COPD : Smoker	0.54100	0.6790	COMMON−PAG
Cystic fibrosis : Healthy	0.14500	0.6810	COMMON−NONPAG
Cystic fibrosis : Smoker	0.22700	0.6810	COMMON−NONPAG
COPD : Cystic fibrosis	0.48700	0.8690	COMMON−NONPAG
COPD : Healthy	0.58000	0.8690	COMMON−NONPAG
COPD : Smoker	0.83900	0.8690	COMMON−NONPAG
Healthy : Smoker	0.86900	0.8690	COMMON−NONPAG
Healthy : Smoker	1.00000	1.0000	COMMON−PAG

Figure 5.2 Distribution of bacterial genes by pathogen association showed that PAGs, relative to common genes are more prevalent in CF lungs.

(A) Gene counts (annotated in text) were normalized by number of NCBI lowest common taxonomic nodes and averaged within each sample type. Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). Fisher's exact test was first performed to detect a significant association of the normalized, mean gene counts between participant groups (COPD, CF, healthy individuals, smokers) and the pathogen association of genes (P= 0.008). **(B)** Post-hoc pairwise Fisher's exact tests were subsequently performed to detect any significant association between COMMON and PAG or NONPAG in each pair of participant groups. Relative to common genes, PAGs are more prevalent in the lungs of CF patients than COPD (adjusted P=0.0114), healthy individuals (adjusted P=0.0114) and smokers (adjusted P=0.0140). No statistically significant association was detected between COMMON and NONPAG in any pairs of participant groups. Associations are statistically significant if p-value is less than 0.05. In the post-hoc tests, p-values were adjusted for multiple testing by the Benjamini–Hochberg false discovery rate method.

COG functional analysis was performed to explore the functional trends of genes within each lung microbiome environment. Within each COG category, the prevalence of bacterial genes by pathogen association was compared across the lungs of the four participant groups. Based on the Fisher's exact tests, performed on a contingency table constructed from genes of the three pathogen association groups and the four participant groups and subsequent post-hoc pairwise Fisher's exact tests with multiple testing correction, CF lungs are disproportionately enriched in PAGs, relative to common genes, in the "Unknown function" COG category (Figure 5.3A-C). This observation is consistent with results in Chapter 2 highlighting that PAGs are less well characterized than common genes. In addition, common genes were significantly more prevalent than PAGs in CF lungs when compared to COPD lungs (P=0.0288) (Figure 5.3A,B,D).

To explore the prevalence of VFs in the different lung conditions, the predicted genes were classified by known or unknown association with VFs annotated in the Virulence Factor Database (Figure 5.4). Less than a quarter of the genes within each pathogen association across all samples had VF association. Genes with no VF association may either have no virulence function or have no significant sequence similarity to known VFs, in the case of many uncharacterized genes (often annotated as "hypothetical proteins"). Within each participant group, Fisher's exact tests were performed on a 2x3 contingency table constructed from the two types of VF association (VF-associated or Unknown) and the three pathogen association groups of bacterial genes. No significant difference was detected between the participant groups and pathogen association of genes related to known VFs, (Figure 5.5). Although further classification of VF-associated genes by VF classes and overall Fisher's exact test revealed significant association between the pathogen association of genes and the participant groups in the "Toxin" and "Enzyme" VF classes, the post-hoc pairwise tests with multiple testing did not support these association (Figure 5.6).


A
Amino acid transport and metabolism
Carbohydrate transport and metabolism
Cell cycle control, cell division, chromosome partitioning
Cell motility
Cell wall/membrane/envelope biogenesis
Coenzyme transport and metabolism
Defense mechanisms
Energy production and conversion
Extracellular structures
General function prediction only
Inorganic ion transport and metabolism
Intracellular trafficking, secretion, and vesicular transport
Lipid transport and metabolism
Mobilome: prophages, transposons
Nucleotide transport and metabolism
Posttranslational modification, protein turnover, chaperones
Replication, recombination and repair
RNA processing and modification
Secondary metabolites biosynthesis, transport and catabolism
Signal transduction mechanisms
Transcription
Translation, ribosomal structure and biogenesis
Unknown function
Normalized gene count
750
500
250
0
COPD
Cystic fibrosis
Healthy
Smoker
Participant group
Pathogen association
COMMON
NONPAG
PAG

B

COG_category	P_value
Cell motility	0
Unknown function	0.033
Mobilome: prophages , transposons	0.077
Intracellular trafficking, secretion, and v esicular transport	0.298
Nucleotide transport and metabolism	0.415
Lipid transport and metabolism	0.486
General function prediction only	0.568
Defense mechanisms	0.734
Carbohydrate transport and metabolism	0.735
Posttranslational modification, protein tur nover, chaperones	0.751
Coenzyme transport and metabolism	0.772
Amino acid transport and metabolism	0.777
Cell wall/membrane/envelope biogenesis	0.81
Inorganic ion transport and metabolism	0.839
Transcription	0.875
Signal transduction mechanisms	0.89
Energy production and con version	0.914
Translation, ribosomal structure and biogenesis	0.988
Cell cycle control, cell division, chromosome par titioning	1
Replication, recombination and repair	1
Secondary metabolites biosynthesis , transport and catabolism	1
RNA processing and modification	N/A
Extracellular structures	N/A

C

Comparison	p.Fisher	p.adj.Fisher	Pathogen_association	COG_function
COPD : Cystic fibrosis	0.00146	0.00876	COMMON−PAG	Unknown function
Cystic fibrosis : Healthy	0.01750	0.04940	COMMON−PAG	Unknown function
Cystic fibrosis : Smoker	0.02470	0.04940	COMMON−PAG	Unknown function
COPD : Healthy	0.22800	0.34200	COMMON−PAG	Unknown function
COPD : Smoker	0.32200	0.38600	COMMON−PAG	Unknown function
COPD : Cystic fibrosis	0.65900	0.86600	COMMON−NONPAG	Unknown function
COPD : Healthy	0.45400	0.86600	COMMON−NONPAG	Unknown function
COPD : Smoker	0.42200	0.86600	COMMON−NONPAG	Unknown function
Cystic fibrosis : Healthy	0.85100	0.86600	COMMON−NONPAG	Unknown function
Cystic fibrosis : Smoker	0.69500	0.86600	COMMON−NONPAG	Unknown function
Healthy : Smoker	0.86600	0.86600	COMMON−NONPAG	Unknown function
Healthy : Smoker	1.00000	1.00000	COMMON−PAG	Unknown function

D

Comparison	p.Fisher	p.adj.Fisher	Pathogen_association	COG_function
COPD : Cystic fibrosis	0.0048	0.0288	COMMON−PAG	Cell motility
Cystic fibrosis : Healthy	0.0323	0.0969	COMMON−PAG	Cell motility
Cystic fibrosis : Healthy	0.0435	0.1300	COMMON−NONPAG	Cell motility
Cystic fibrosis : Smoker	0.0299	0.1300	COMMON−NONPAG	Cell motility
Cystic fibrosis : Smoker	0.0781	0.1560	COMMON−PAG	Cell motility
COPD : Healthy	0.5000	0.7500	COMMON−NONPAG	Cell motility
COPD : Smoker	0.4810	0.7500	COMMON−NONPAG	Cell motility
COPD : Healthy	0.6750	0.8100	COMMON−PAG	Cell motility
COPD : Smoker	0.6520	0.8100	COMMON−PAG	Cell motility
Healthy : Smoker	1.0000	1.0000	COMMON−PAG	Cell motility
COPD : Cystic fibrosis	1.0000	1.0000	COMMON−NONPAG	Cell motility
Healthy : Smoker	1.0000	1.0000	COMMON−NONPAG	Cell motility

Figure 5.3 COG functional analysis of predicted genes in the lung microbiome of the four participant groups showed that CF lungs are disproportionately enriched in PAGs, relative to common genes, with unknown function.

(A) Normalized counts of genes classified by COG functional categories and pathogen association in the lung microbiome of CF patients, COPD patients, smokers and healthy controls. Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). Due to the low gene count in most COG categories, **(B)** Fisher's exact test, performed in each COG category, detected a significant association between pathogen association of genes and the participant groups only in the "Unknown function" (p=0.033) and "Cell motility" (p=0.001) COG categories. **(C)** Post-hoc pairwise Fisher's exact tests were subsequently performed to detect any significant association between COMMON and PAG or NONPAG in each pair of participant groups. Relative to common genes, PAGs are more enriched in the "Unknown Function" COG category in the lungs of CF patients than COPD (adjusted p-value = 0.009), healthy individuals (adjusted p-value = 0.049) and smokers (adjusted p-value = 0.049). **(D)** Post-hoc pairwise Fisher's exact tests within the "Cell motility" category detected an enrichment of PAG, relative to COMMON, in CF lungs when compared to COPD lungs. Associations are statistically significant if p-value is less than 0.05. In the post-hoc tests, p-values were adjusted for multiple testing by the Benjamini–Hochberg false discovery rate method. "N/A" refers to COG category in which Fisher's exact test could not be performed due to the lack of PAG or NONPAG in that category.

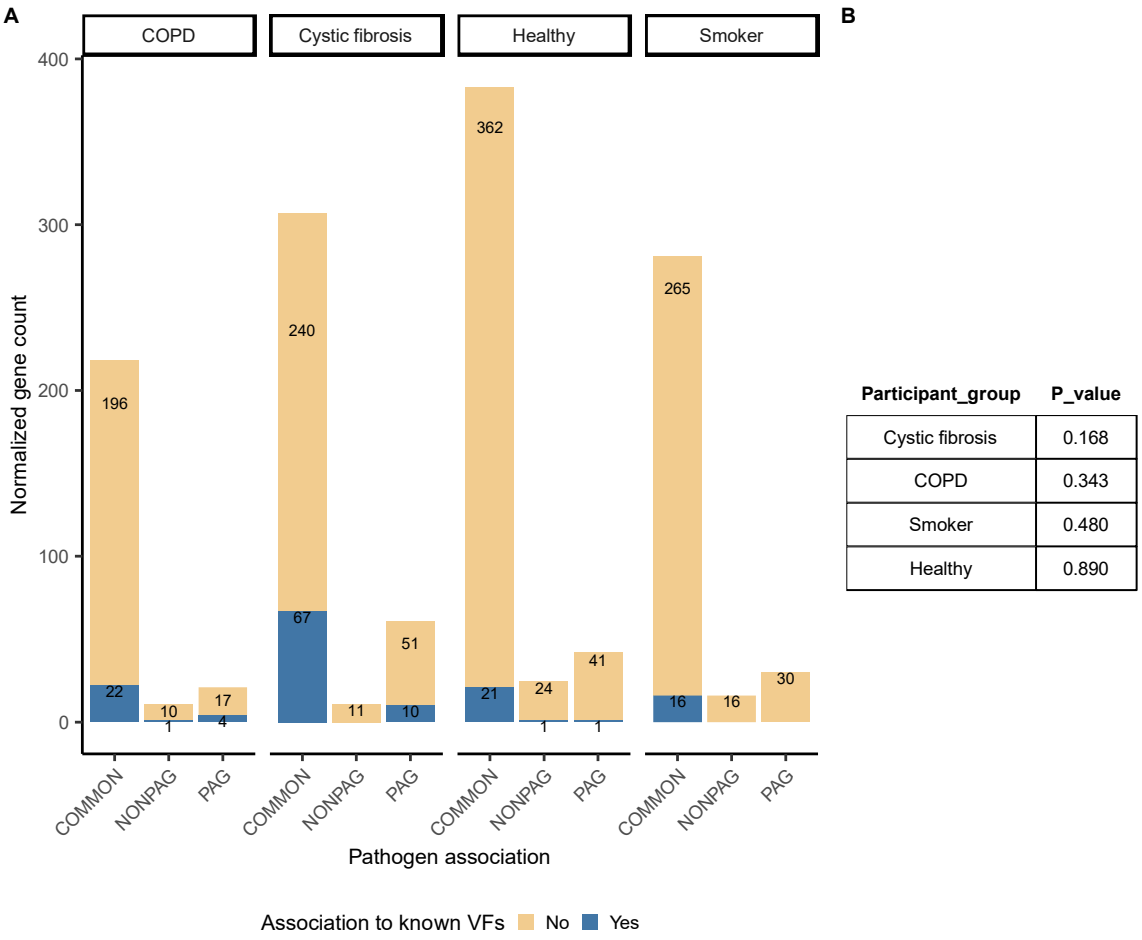


Figure 5.4 No significant association was detected between pathogen association of genes and their association to known VFs.

(A) Normalized gene counts (annotated in text) averaged among the lung microbiomes of CF patients, COPD patients, smokers and healthy controls. Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). In each participant group, genes in each of the three pathogen association groups were classified by association to known VFs based on whether a significant sequence similarity to VFs from the Virulence Factor Database was detected (>90% sequence identity and e-value cut-off of 10^{-7}). **(B)** Fisher's exact test was conducted to test the hypothesis that more PAGs than common genes or non-PAGs are associated with known VF, however, no statistically significant association was detected between the pathogen association and VF association of genes within any of the sample (participant) type.

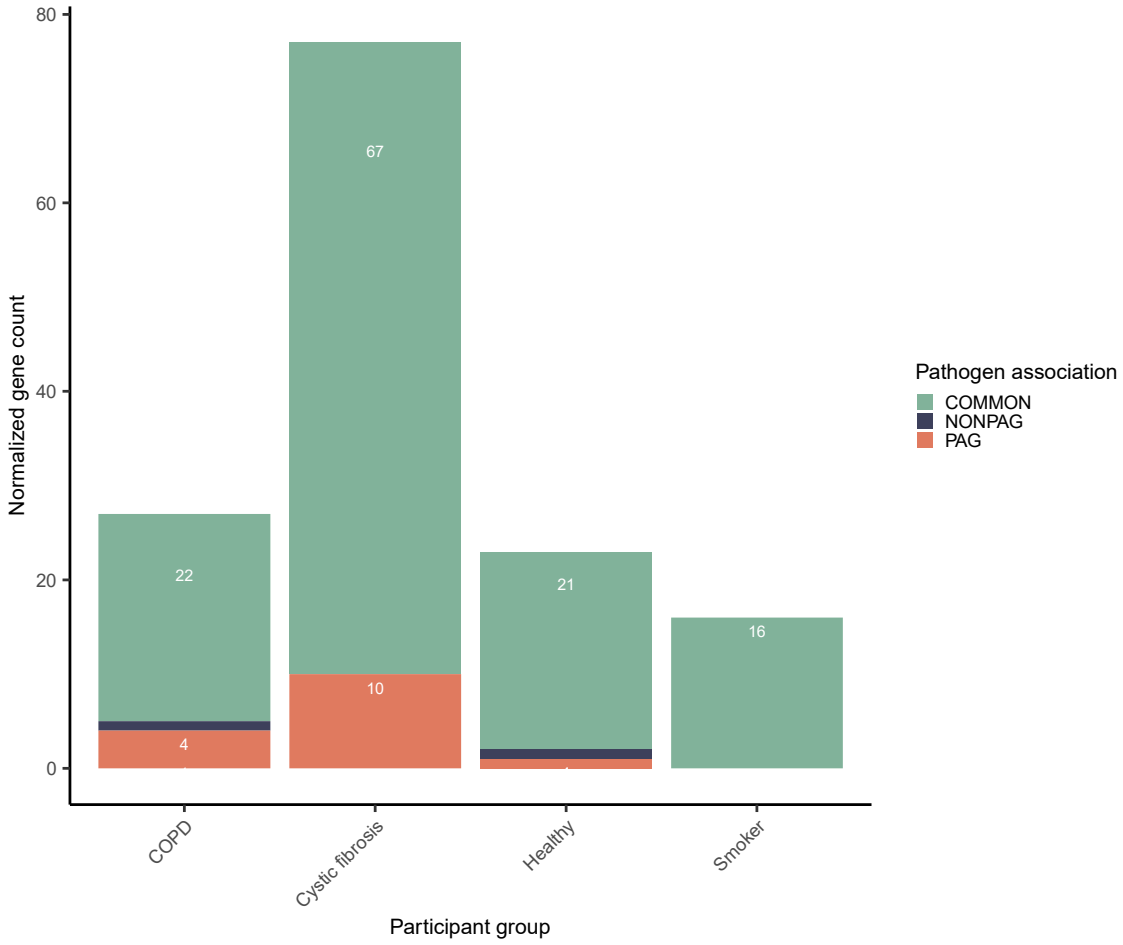
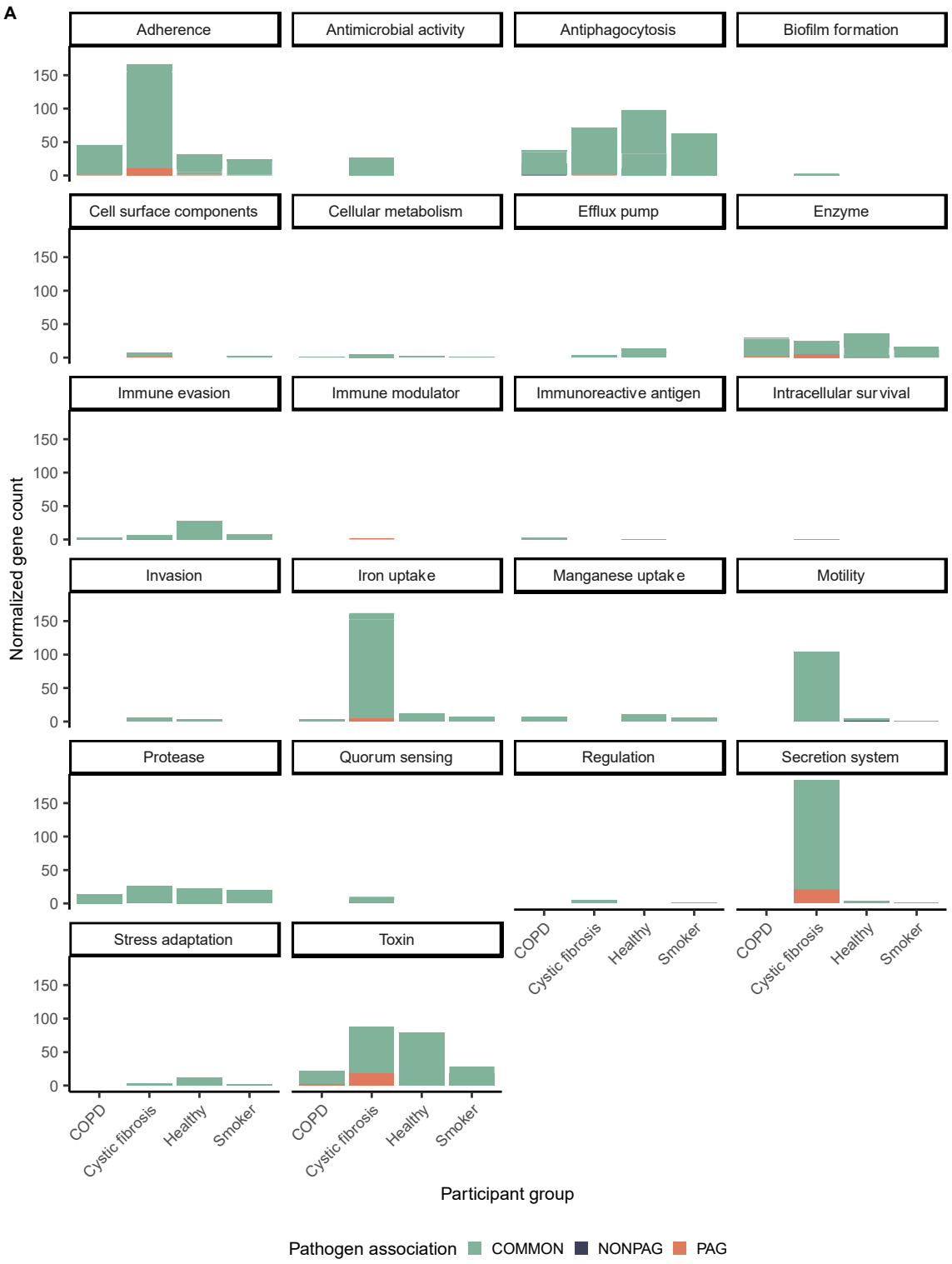


Figure 5.5 Bacterial genes with VF association revealed no significant association between pathogen association of VF-associated genes and lung conditions.

Normalized gene count, averaged within each sample (participant) type, with significant sequence similarity to known VFs in the Virulence Factor Database were categorized by pathogen association in each participant group: COPD patients, CF patients, healthy controls, and smokers). Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). Based on a Fisher's exact test to examine the hypothesis that genes associated with bacterial pathogens and known VFs are more enriched in CF lungs, no significant association between the pathogen association of VF-associated genes and sample type (p=0.197).


A
Adherence
Antimicrobial activity
Antiphagocytosis
Biofilm formation
Cell surface components
Cellular metabolism
Efflux pump
Enzyme
Immune evasion
Immune modulator
Immunoreactive antigen
Intracellular survival
Invasion
Iron uptake
Manganese uptake
Motility
Protease
Quorum sensing
Regulation
Secretion system
Stress adaptation
Toxin
Normalized gene count
0
50
100
150
COPD
Cystic fibrosis
Healthy
Smoker
Participant group
Pathogen association
COMMON
NONPAG
PAG

B

VF_class	P_value
Enzyme	0.007
Toxin	0.01
Antiphagocytosis	0.122
Motility	0.144
Immune evasion	0.353
Adherence	0.554
Iron uptake	1
Protease	1
Cell surface components	1
Secretion system	1
Cellular metabolism	N/A
Immunoreactive antigen	N/A
Manganese uptake	N/A
Antimicrobial activity	N/A
Biofilm formation	N/A
Efflux pump	N/A
Intracellular survival	N/A
Invasion	N/A
Quorum sensing	N/A
Regulation	N/A
Stress adaptation	N/A
Immune modulator	N/A

C

Comparison	p.Fisher	p.adj.Fisher	Pathogen_association	VF_class
Cystic fibrosis : Healthy	0.0309	0.185	COMMON−PAG	Toxin
COPD : Healthy	0.0887	0.206	COMMON−PAG	Toxin
Cystic fibrosis : Smoker	0.1030	0.206	COMMON−PAG	Toxin
COPD : Smoker	0.1420	0.213	COMMON−PAG	Toxin
COPD : Cystic fibrosis	1.0000	1.000	COMMON−PAG	Toxin
Healthy : Smoker	1.0000	1.000	COMMON−PAG	Toxin

D

Comparison	p.Fisher	p.adj.Fisher	Pathogen_association	VF_class
Cystic fibrosis : Healthy	0.0325	0.108	COMMON−PAG	Enzyme
Cystic fibrosis : Smoker	0.0359	0.108	COMMON−PAG	Enzyme
COPD : Cystic fibrosis	0.0635	0.127	COMMON−PAG	Enzyme
COPD : Healthy	1.0000	1.000	COMMON−PAG	Enzyme
COPD : Smoker	1.0000	1.000	COMMON−PAG	Enzyme
Healthy : Smoker	1.0000	1.000	COMMON−PAG	Enzyme

Figure 5.6 Classification of VF-associated genes by VF did not reveal a significant association between the prevalence of genes by pathogen association and the participant groups in any of the VF classes .

(A) Normalized counts of genes, averaged within each sample type, with known VF association, further classified by VF classes according to the Virulence Factor Database. Sample types include COPD patients, CF patients, smokers and healthy controls. Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). **(B)** Fisher's exact test, performed in each VF class, detected a significant association between pathogen association of genes and the sample (participant) types in "Toxin" (p=0.007) and "Cell motility" (p=0.01). **(C)** Post-hoc pairwise Fisher's exact tests within the "Toxin" VF class were subsequently performed to detect any significant association between COMMON and PAG in each pair of sample types. However, this post-hoc test did not detect any association of PAGs enrichment with sample (participant) types upon adjusting p-values for multiple testing. **(D)** Post-hoc pairwise Fisher's exact tests within the "Enzyme" VF class, however, did not detect any association of PAGs enrichment with sample (participant) types upon adjusting p-values for multiple testing. Associations are statistically significant if p-value is less than 0.05. In the post-hoc tests, p-values were adjusted for multiple testing by the Benjamini–Hochberg false discovery rate method. "N/A" refers to COG category in which Fisher's exact test could not be performed due to the lack of PAG or NONPAG in that category.

5.3.2. The prevalence of virulence associated PAGs increased at the agricultural affected and downstream sites during the rainy season.

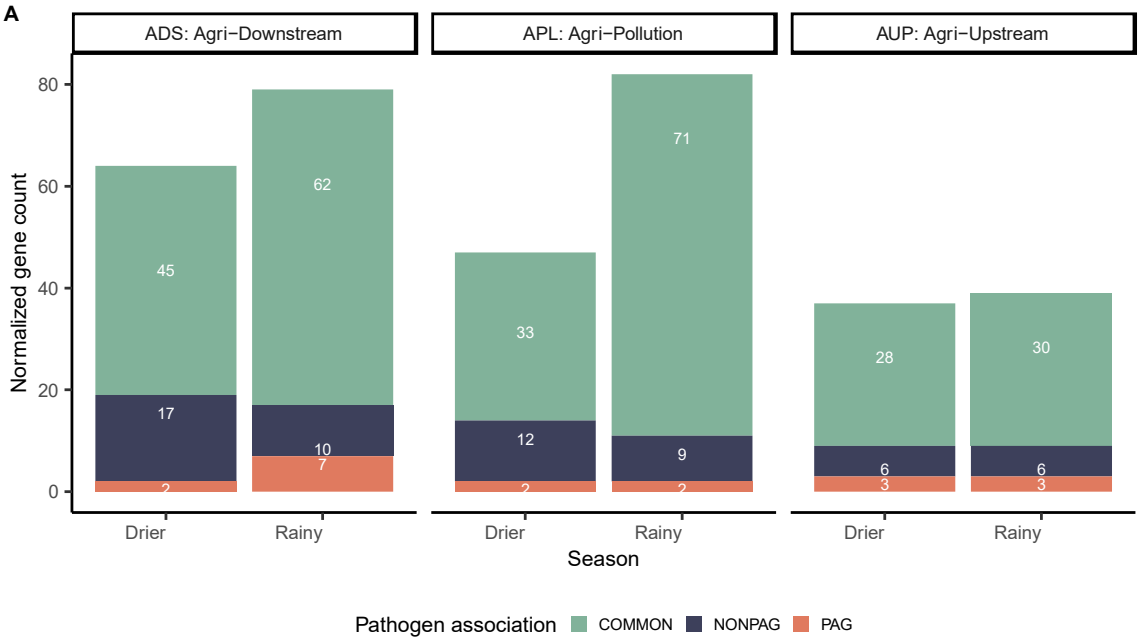
The same analysis was applied to the watershed dataset to test the hypothesis that PAGs are more enriched in agricultural affected (fecally contaminated) sites. Although Fisher's exact test detected a borderline association between the pathogen association of genes and the collection season at the Agri-Downstream site (P=0.05), subsequent post-hoc tests with multiple testing correction did not support such association (Figure 5.7). In a previous analysis of this watershed dataset (Peabody, 2017), the mean percentage of reads that could be assigned to the bacterial family level ranged from 11.1% to 42.4% across the Agri-Downstream, Agri-Pollution and Agri-Upstream sites depending on whether 16S rRNA or shotgun metagenomic sequences were used in the analysis (Table 5.2.). The low percentage of assigned reads may indicate the presence of novel bacterial genomes which may reduce the sensitivity of this metagenomics-based detection of PAGs, since the genes from novel genomes with unknown pathogen status, excluded from the list of PAGs identified in Chapter 2, would also be excluded from this analysis. Using the same statistical analysis method on the COG functional categories and VFDB VF association described in Chapter 5.3.1, no significant increase in prevalence of any pathogen association gene groups was detected in any watershed sites during the drier season (Figure 5.8). In addition, none of the pathogen association gene groups was significantly enriched in VF-associated genes in any of the 6 combinations of collection seasons (drier vs rainy) and collection sites (APL, AUP, ADP) (Figure 5.9). Bacterial genes collected from each watershed site and season were not classified into VF classes due to the extremely low gene counts and lack of significant association of the pathogen association gene groups and collection sites/seasons.

Table 5.2 Percent of reads that could be assigned to the family level in the 16S rRNA amplicon sequencing and shotgun metagenomic sequencing datasets across watershed sites.

Sequence method	Site name	Min	Max	Mean	SD
16S	ADS	34.2	56.8	42.2	6.9
16S	APL	16.9	58.5	40.4	10.3
16S	AUP	11.9	38.4	23.6	7.9
Shotgun	ADS	11.2	37.5	21.2	6.7
Shotgun	APL	14.7	31.5	20.9	5.8
Shotgun	AUP	8.1	19.8	11.1	3.4

ADS: Agri-Downstream. APL: Agri-Pollution. AUP: Agri-Upstream.
Table adapted from Dr. Mike Peabody's PhD thesis (Peabody, 2017).

A



B

Season	P_value
Rainy	0.379
Drier	0.627

C

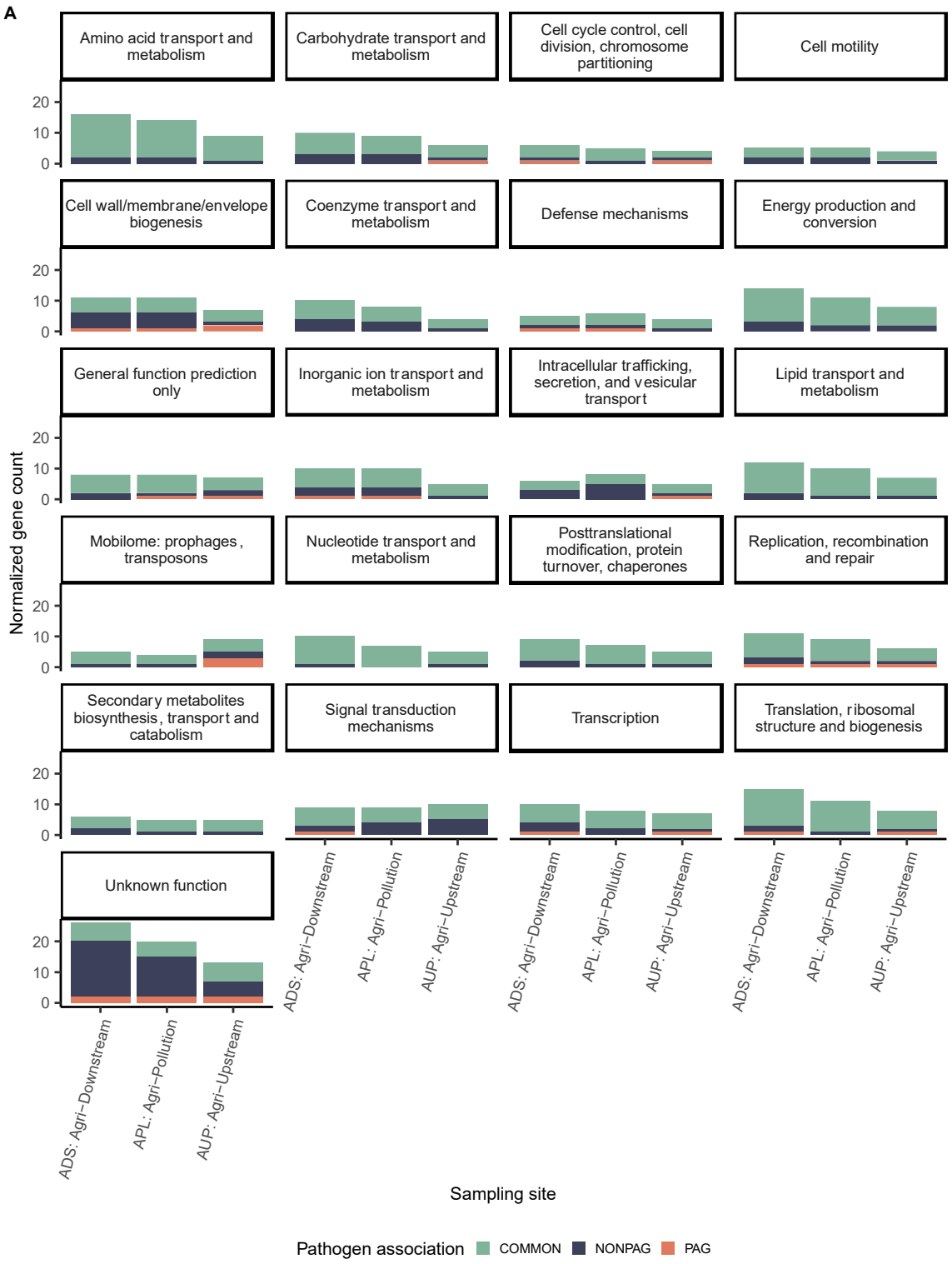
Site	P_value
ADS: Agri−Downstream	0.050
APL: Agri−Pollution	0.067
AUP: Agri−Upstream	1.000

D

Comparison	p.Fisher	p.adj.Fisher	Pathogen_association	Site
Drier : Rainy	0.0551	0.0551	COMMON−NONPAG	ADS: Agri−Downstream
Drier : Rainy	0.3080	0.3080	COMMON−PAG	ADS: Agri−Downstream

Figure 5.7 Distribution of genes by pathogen association collected at different agricultural watersheds and seasons did not reveal an enrichment of genes by pathogen association in any collection season or site.

(A) Gene counts (annotated in text) were normalized by number of NCBI lowest common taxonomic nodes. Samples are categorized by the season at which they were collected: drier season (May-Oct) and rainy season (Nov-Apr). Sampling sites include agricultural downstream site (ADS), agricultural pollution site (APL) and agricultural upstream site (AUP). Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). **(B)** Fisher's exact test detected no significant association between pathogen association of genes and site of collection in either season. **(C)** Fisher's exact test detected a borderline significant association between pathogen association of genes and collection season in the agricultural downstream site. **(D)** However, post-hoc pairwise Fisher's exact test performed within samples collected at the agricultural downstream site did not detect a significant difference in gene count between COMMON and PAG or NONPAG. Associations are statistically significant if p-value is less than 0.05. In the post-hoc tests, p-values were adjusted for multiple testing by the Benjamini–Hochberg false discovery rate method.


A
Amino acid transport and metabolism
Carbohydrate transport and metabolism
Cell cycle control, cell division, chromosome partitioning
Cell motility
Cell wall/membrane/envelope biogenesis
Coenzyme transport and metabolism
Defense mechanisms
Energy production and conversion
General function prediction only
Inorganic ion transport and metabolism
Intracellular trafficking, secretion, and vesicular transport
Lipid transport and metabolism
Mobilome: prophages, transposons
Nucleotide transport and metabolism
Posttranslational modification, protein turnover, chaperones
Replication, recombination and repair
Secondary metabolites biosynthesis, transport and catabolism
Signal transduction mechanisms
Transcription
Translation, ribosomal structure and biogenesis
Unknown function
Normalized gene count
0
10
20
ADS: Agri−Downstream
APL: Agri−Pollution
AUP: Agri−Upstream
Sampling site
Pathogen association
COMMON
NONPAG
PAG

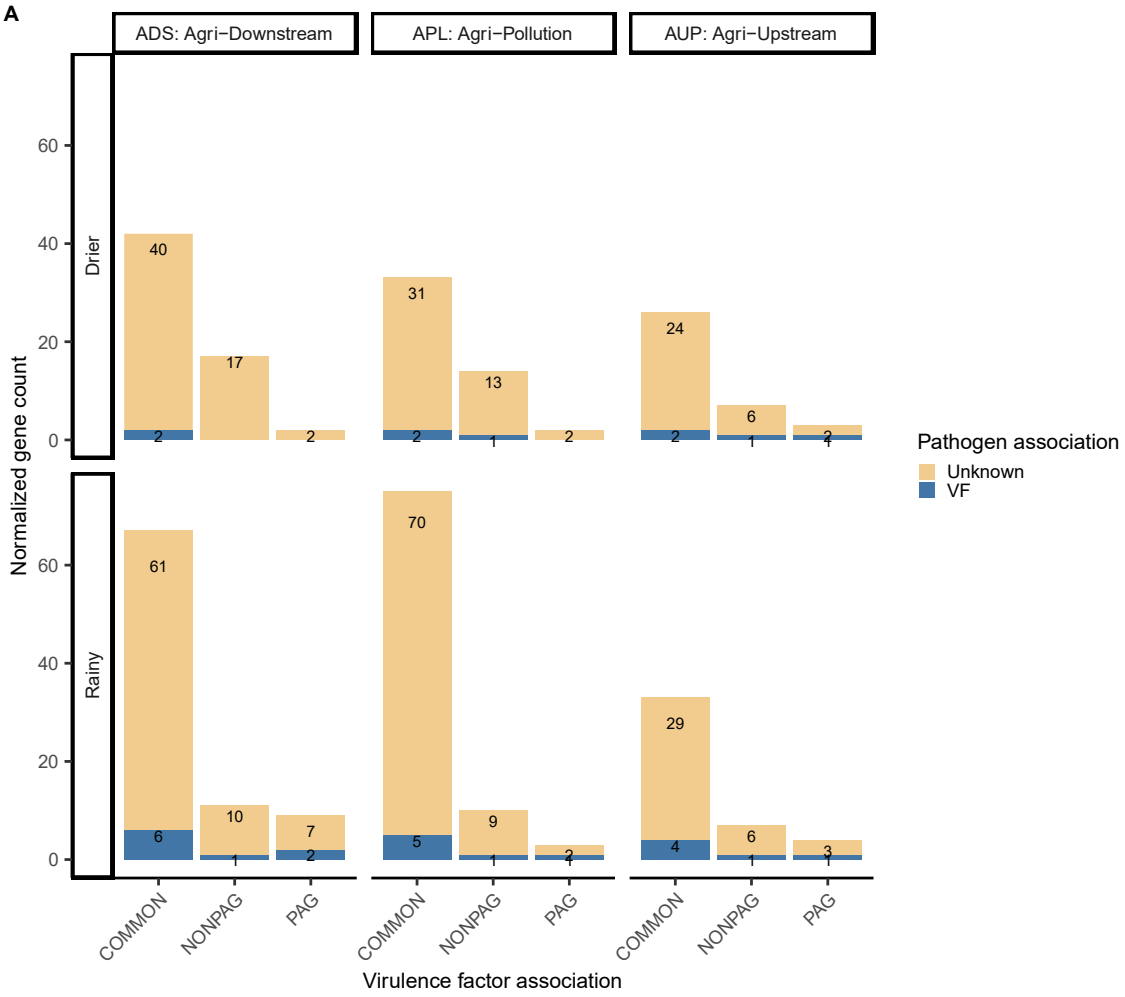
B

COG_category	P_value
Unknown function	0.432
Intracellular trafficking, secretion, and v esicular transport	0.508
Signal transduction mechanisms	0.563
Mobilome: prophages , transposons	0.587
Cell wall/membrane/envelope biogenesis	0.613
Nucleotide transport and metabolism	0.678
Carbohydrate transport and metabolism	0.679
General function prediction only	0.799
Transcription	0.926
Cell cycle control, cell division, chromosome par titioning	0.930
Translation, ribosomal structure and biogenesis	0.943
Amino acid transport and metabolism	1.000
Cell motility	1.000
Coenzyme transport and metabolism	1.000
Defense mechanisms	1.000
Energy production and con version	1.000
Inorganic ion transport and metabolism	1.000
Lipid transport and metabolism	1.000
Posttranslational modification, protein tur nover, chaperones	1.000
Replication, recombination and repair	1.000
Secondary metabolites biosynthesis , transport and catabolism	1.000

Figure 5.8 COG functional analysis of bacterial genes in the watershed samples collected in the drier months (May-Oct) showed no significant association between pathogen association of genes and collection site in each of the COG category.

(A) Normalized counts of genes classified by COG functional categories and pathogen association in the samples collected, in the drier season (May-Oct), from the agricultural downstream site (ADS), agricultural pollution site (APL) and agricultural upstream site (AUP). Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). **(B)** Within each COG category, Fisher's exact test was performed to test the hypothesis that there is an association between the prevalence of genes by pathogen association and their collection site, however, no significant association was detected in any COG category. Associations are statistically significant if p-value is less than 0.05.

A



B

Season	Site	P_value
Rainy	APL: Agri−Pollution	0.153
Rainy	ADS: Agri−Downstream	0.302
Drier	AUP: Agri−Upstream	0.303
Rainy	AUP: Agri−Upstream	0.746
Drier	ADS: Agri−Downstream	1.000
Drier	APL: Agri−Pollution	1.000

Figure 5.9 Classification of genes by VF association at each watershed sampling site showed no detected association between known VFs and pathogen association of genes in each of the 6 combinations of collection season and collection site.

(A) Normalized gene counts (annotated in text) are shown for agricultural downstream site (ADS), agricultural pollution site (APL) and agricultural upstream site (AUP). Types of pathogen association includes common genes (COMMON), non-PAGs (NONPAG) and PAGs (PAG). In each participant group, genes in each of the three pathogen association groups were classified as VF-associated or unknown association (Unknown) based on whether a significant sequence similarity to VFs from the Virulence Factor Database was detected (>90% sequence identity and e-value cut-off of 10-7). **(B)** Within each of the six combinations of collection site and collection

season, Fisher's exact test did not detect any significant association between the pathogen association and VF association of bacterial genes. Associations are statistically significant if p-value is less than 0.05.

5.4. Discussion

Since the start of The Human Microbiome Project, in 2007, which aimed to better understand how microbial communities influence human health and predisposition to diseases, there has been an increasing interest in clinical metagenomics as a promising and non-invasive approach to the surveillance and diagnosis of human diseases (Chiu & Miller, 2019; Turnbaugh et al., 2007). With the ease of fecal sampling, numerous metagenomic studies identified distinct microbiome signatures of various intestinal diseases, ranging from inflammatory bowel disease, colorectal cancer to Crohn's disease (Gigliucci et al., 2018; Jiang et al., 2021; Ma et al., 2021). On the contrary, metagenomic surveillance of respiratory diseases is more challenging owing to the invasive sampling of the respiratory tract (Carr & Chaguza, 2021). However, a few recent respiratory metagenomic studies, including the one used in my analysis, have effectively profiled the microbiome within the CF lungs using sputum samples (Bacci et al., 2020; Feigelman et al., 2017; Huang et al., 2021; Zheng et al., 2021). Shotgun metagenomics in particular enables not only the taxonomic profiling of and pathogen detection but also the assessment of the AMR and virulence potential within disease-associated microbiomes (De et al., 2020; Sanabria et al., 2021).

As PAGs likely represent an under-characterized reservoir of novel bacterial VFs, I performed the first PAGs-based metagenomic analysis of the CF lung microbiome using a published sputum DNA dataset (Feigelman et al., 2017). This analysis revealed that relative to healthy lungs, CF lungs were disproportionately enriched in PAGs, indicating potential role of some of these bacterial genes in CF airway adaptation. Decline in CF lung function is often associated with persistent colonization by *P. aeruginosa* and *Burkholderia cepacia* with hypermutations, AMR and other phenotypes that aid in pathogen survival (Pust et al., 2020; E. E. Smith et al., 2006; Winstanley et al., 2016). Lung infections in CF patients begins with bacterial attachment to mucosal surfaces in the respiratory tract followed by an escape or resistance of host immune responses and inflammation. Prolonged infections involve a successive change in lung microbiota with new predominant species and are often characterized by the establishment of a mucoid phenotype with alginate production in biofilms to prevent host

neutrophil-mediated killing of bacteria. This process results in bacterial persistence and drug resistance which make complete bacterial eradication in the lungs a challenge (Hart & Winstanley, 2002). Although CF therapies such as the latest Trikafta (triple combination therapy of elexacaftor, tezacaftor and ivacaftor) can mitigate respiratory system complications such as mucus obstruction and infections, complete eradication of persistent *P. aeruginosa* airway infection remains an unsolved challenge due to the likelihood of recurrence after the first year of CF treatment (Hisert et al., 2017). Therefore, uncovering the virulence determinants in CF airway infections and implementing proper infection management are crucial for improving the prognosis of CF patients. For future directions, longitudinal metagenomic studies of CF lungs would be useful in examining the population dynamics and microbial profiles of bacterial pathogens over the course of CF disease progression under the perturbation by antibiotic treatments. PAGs, enriched in CF airways, may provide a novel avenue for discovering key genetic players in the establishment and persistence of chronic respiratory infections.

While pathogen transmission among healthcare and community settings poses a health threat to the susceptible individuals, researchers are now turning to a more holistic, One Health approach to understand pathogen transmission in the interactions between humans, animals (e.g., livestock) and the natural environment. For example, one metagenomic study of a North American commercial farm revealed that swine harbour a diversity of AMR genes (e.g., those against tetracycline, macrolides, lincosamides, and streptogramin) as well as the major foodborne pathogen *L. monocytogenes* that may increase the risk of transmission into the human population at the human-animal interface (Ramesh et al., 2021). In addition, anthropogenic activities in urban and agricultural areas drive the co-enrichment and dissemination of AMR and VFs genes in the nearby aquatic ecosystem through the increased use of antibiotics in livestock and fish farming (Liang et al., 2020). The co-occurrence and the possible co-evolutionary selection of AMR and VF genes in the human gut microbiome (Escudeiro et al., 2019) prompted my PAGs analysis of the Genome Canada's watershed microbiome dataset to investigate if agricultural affected watersheds are more enriched in PAGs (which are likely involved in virulence) than the pristine watershed upstream of the agricultural site. However, no significant enrichment of PAGs was detected among the watershed sites or (dry vs rainy) seasons, likely owing to the low number of genes with a

predicted pathogen association and the lack of characterization of the freshwater microbiome at these collection sites (compared to the more well studied human-associated microbiomes) as previously reported (Peabody, 2017). Likewise, majority of genes at each sample collection site was not associated with the known VFs curated in the Virulence Factor Database, indicating that they either have no true virulence function or represent novel classes of VF that are yet to be characterized. Since a limited number of genes had significant association with known VFs, it was a challenge to compare the trend in virulence genes across the sample collection sites and seasons. It is important to note that this type of VF analysis was highly dependent on the breadth of VFDB and the number of bacterial genes with significant sequence similarity to the existing VFs curated in VFDB; if novel genes such as those annotated as “hypothetical proteins” represent novel VFs and are not yet curated in VFDB, their VF association would not be detected.

The lung microbiome and watershed microbiome analyses both highlighted the need for better functional characterization method of PAGs, especially those currently annotated as “hypothetical proteins” whose gene expression and function have yet to be validated in the laboratory. Additional functional characterization method of the largely uncharacterized PAGs to better understand the impact of environmental, agricultural and human factors on virulence and pathogen transmission through a One Health approach. PAGs that are important for virulence during host infections as well as are prevalent in natural environments closely linked to human activities can potentially be used as biomarkers for monitoring the transmission of pathogens and virulence genes among the different interconnected clinical and non-clinical ecosystems. However, based on the results presented in this chapter, the application of PAGs in pathogen surveillance may currently be limited to well-studied microbiomes such as those associated with known clinical infections, rather than environmental microbiomes in which the microbial organisms are less characterized. Nonetheless, this data chapter illustrates the potential application of this thesis work, including the previous chapters 2-4, in the discovery of novel virulence factors from PAGs that would help progress the development of novel antivirulence therapeutics and potentially One Health strategies in infectious disease management.

In conclusion, this chapter represents the first set of metagenomic analyses of PAGs in microbial communities using both clinical and environmental datasets. I showed

that in comparison to healthy lungs, CF lungs, commonly associated with chronic infections, are more enriched in PAGs. Further investigation into the role of PAGs and their associated bacterial pathogens in CF lung pathogenesis would be beneficial to improve the disease management strategies and prognosis of people with CFs. I could not identify any statistically significant PAGs enrichment in any watershed sites (upstream, downstream or at agriculturally affected sites), which I hypothesized to be due to the lack of characterization of freshwater microbial communities at these sites as observed in a previous study of this dataset (Peabody, 2017). Since the identification of PAGs fundamentally relies on proper curation of NCBI RefSeq bacterial genomes as pathogens or non-pathogens based on published literature (see Chapter 2.3.1), they are likely under-represented in environmental bacteria whose genomes are not currently in the RefSeq Database. Better sampling and characterization of environmental bacteria, such as those found in freshwater, would help to improve the identification of PAGs in novel bacterial species and the detection of these genes in less characterized microbiomes. Under the One Health perspective, the interconnected ecosystems of human, animal and the natural environments enables the flow of microbial communities as well as genes via horizontal gene transfer that may help to disseminate pathogens to and from the human population. Metagenomics thus serves as an informative tool in understanding the environmental and microbial dynamics of infectious disease and aiding in public health decisions in controlling the spread of pathogens, especially those of clinical importance.

Chapter 6.

Concluding remarks

6.1. Summary

This thesis presents a computational approach for identifying PAGs from a large-scale comparative bacterial genome analysis and characterizing their functional significance through analyses of taxonomic conservation, evolutionary selection, and protein SCL. To prioritize PAGs as candidate targets for antivirulence therapeutics, virulence of a subset of these genes was tested in a worm infection model. A few PAGs from *P. aeruginosa* PA14, including the pair of GI-localized BlaI/MecI/CopY family transcriptional regulator (PA14_RS12700, WP_025297936.1) and M56 family metallopeptidase (PA14_RS12695; WP_016254216.1), are now under in-depth functional characterization as a part of a follow-up study to this work. In addition, I performed a laboratory-based structure-activity relationship analysis of an FDA-approved osteoporosis drug, raloxifene, as a potential antivirulence agent against *P. aeruginosa*. Lastly, the prevalence of PAGs was investigated in lung microbiomes among individuals with different health status, as well as in freshwater microbiomes across different sites near an agriculturally affected watershed.

Chapter 2 lays the foundation of the collection of this thesis work with an update of the PAGs analysis on over 8600 NCBI RefSeq bacterial genomes. Preliminary, sequence similarity-based analyses showed that PAGs disproportionately lack functional characterization as many are still annotated as “hypothetical proteins” or are too divergent from protein families currently curated in the Pfam database. Sequence similarity search against the Virulence Factor Database revealed that PAGs associated with known virulence factors were enriched in functions related to bacterial secretion system, toxins or secreted proteins, suggesting potential role in pathogen-host interaction in some PAGs that are yet to be characterized. An orthology analysis was also introduced to cluster PAGs into orthologous groups, which was subsequently assessed for taxonomic conservation. While genes common to both pathogens and non-pathogens are often found across multiple bacterial genera, majority of PAGs as well as non-PAGs was conserved in up to two genera that seemingly shared similar ecological

niches. Based on this finding, the original PAGs analysis was modified to run on genomes within a single bacterial genus, *Pseudomonas*. Seventeen PAGs in *P. aeruginosa* PA14 were selected for downstream analyses, including an evolutionary selection inference which identified 3 of these genes to be evident of positive selection: a class I SAM-dependent methyltransferase (PA14_RS20430; WP_003116317.1), a DUF3218 family protein (PA14_RS24305; WP_003141470.1) and a M56 family metallopeptidase (PA14_RS12695; WP_016254216.1). Positive selection was also detected in three structural components and one effector gene of the bacterial T3SS in *P. aeruginosa* PA14. These includes the inner membrane component PcrD, the basal body component PscJ, the translocator PopD and the exotoxin exoenzyme T.

Chapter 3 focuses on a more in-depth characterization of the PAGs identified in Chapter 2, with the goal to prioritize those with putative virulence function as candidate antivirulence drug targets. The chapter begins with a global subcellular localization analysis of all bacterial genes. Relative to common genes that are found in both pathogens and non-pathogens, PAGs predominantly reside in the more drug-accessible cell wall or extracellular space in many bacteria. As a continuation study of the 17 PAGs selected from the *P. aeruginosa* PA14 genome, the contribution to virulence of 11 of these genes, whose transposon insertion mutants were readily available, was tested in a *C. elegans* infection model. Of these PA14 PAGs, virulence promoting activity was detected in 6 genes while virulence repressing activity was detected in one gene, the BlaI/MecI/CopY family transcriptional regulator (PA14_RS12700, WP_025297936.1). This transcriptional regulator was particularly interesting because of its genomic proximity to the M56 family metallopeptidase (PA14_RS12695; WP_016254216.1) which was detected to be under positive selection in the previous chapter. This pair of genes were located on a 43kb GI that carries metal- and AMR-associated genes and is found in multiple multi-drug and extensively drug resistant strains of *P. aeruginosa*.

Chapter 4 describes a follow-up analysis of the FDA-approved osteoporosis drug, raloxifene, as a potential antivirulence agent. Raloxifene was previously predicted to interact with the phenazine biosynthesis protein, PhzB2, whose pathogen association was predicted by the 2009 PAGs analysis. To explore the structural components of raloxifene potentially responsible for its antivirulence activity against *P. aeruginosa*, multiple drug analogs, with slight structural modifications, were synthesized and experimentally assessed for their impact on pyocyanin production growth of *P.*

aeruginosa, as well as the survival of *P. aeruginosa*-infected *C. elegans*. Analogs with at least one (the 6') of the two hydroxyl groups present on the raloxifene core showed significant reduced pyocyanin production, survival improvement of infected worms as well as minimal impact on bacterial growth, suggesting that these structural components may be important for raloxifene's antivirulence properties.

Chapter 5 presents the first PAGs analysis of a clinical and an environmental microbiome dataset. In the lung microbiome comparison between CF patients, chronic obstructive pulmonary disease patients, smokers and healthy individuals, PAGs were significantly more prevalent in CF lungs. Functional analysis of the PAGs detected in the lung microbiomes was challenging owing to the abundance of genes without an annotated function. Re-analysis of this data in the future when more functional data become available could potentially reveal functional differences of genes across the different participant groups. This metagenomics-based PAGs analysis was also applied to the Genome Canada's Applied Metagenomics of the Watershed Microbiome dataset, assessing the prevalence of PAGs between freshwater collection sites at, upstream and downstream of an agriculturally affected (fecally contaminated) site. No significant association was detected among the prevalence of genes by pathogen association and the collection sites. Since environmental isolates are often not as well characterized or sequenced as clinical isolates, their gene sequences may be too divergent from the gene set used in the RefSeq-based PAGs analysis. Nonetheless, agriculturally affected watershed and its downstream site seemingly had an increase in PAGs during the transition from dry to rainy season in western Canada. This observation indicates a potential role of rainwater run-offs in the dissemination of bacteria and bacterial genes across environmental habitats. This chapter highlights the potential application of PAGs as biomarkers for pathogen surveillance in well characterized microbiomes (i.e., those associated with clinical settings) in which the functional and virulence potentials of their bacterial constituents are better profiled.

6.2. Future directions

This thesis work illustrates a combined, computational and laboratory approach for discovering, characterizing and validating potentially novel virulence-related genes from bacterial PAGs. This methodology is a work in progress. With the growing wealth of bacterial genomic data, it will be necessary to streamline the process of identifying and

prioritizing candidate genes for the discovery of novel virulence factors that may potentially be selected as novel antivirulence drug targets.

The PAGs analysis requires regular updates to incorporate the ever-growing number of bacterial genomes in the NCBI RefSeq Database. Automating genome retrieval, pathogen status assignment for each novel genome (currently done manually) and running the all-vs-all protein sequence similarity search would help to streamline future updates of this analysis. More efficient data parsing and storage at the sequence similarity search stage will be necessary to effectively analyse thousands of genomes within the filesystem upper limit of file size and number of high-performance computing resources such as Compute Canada's Cedar cluster. Defining a pathogen vs non-pathogen is non-trivial and developing a more robust algorithm, that incorporates phyletic distance of pathogens and non-pathogens, is needed to better prioritize those PAGs of highest interest. For example, we may want to prioritize genes conserved in multiple pathogens, that have non-pathogens phyletically interspersed among them, versus identifying a gene found in pathogens that may simply be a unique gene for that lineage not necessarily involved in pathogenicity. While sequence-similarity based approaches for characterizing functionally unknown PAGs may be challenging, owing to the sequence divergence in some of these genes, there are numerous alternative methods to computationally infer the functional importance or virulence role of a gene. For example, if a PAG is predicted to reside in an operon, it is likely to share a related function with its adjacent genes since genes within an operon are often expressed together to affect the same biological process. Incorporating transcriptomic data of bacterial pathogens, such as those available in the NCBI Gene Expression Omnibus, would also be useful in screening for PAGs that are differentially expressed in host infections, even if their functions are still unknown (Clough & Barrett, 2016). It's important to note that while computational functional prediction is suitable for large-scale screening of PAGs with potential virulence function, laboratory assays are required to confirm their function.

This thesis work identified an interesting pair of transcriptional repressors (PA14_RS12700, WP_025297936.1) and metallopeptidase (PA14_RS12695; WP_016254216.1) that warrants further investigation into their virulence role in *P. aeruginosa*. The metallopeptidase is hypothesized to also confer virulence activity since it is pathogen-associated as well as under positive selection. Assessing the virulence of

a knockout of this pathogen-associated gene with or without the wild-type transcriptional repressor PA14_RS12700 would provide more insight into the virulence function of these genes, as well as the molecular pathways in which these genes may be co-involved. Should the virulence function of these genes be more thoroughly confirmed with additional laboratory assays, these genes may be selected as candidate antivirulence drug targets in the future. The drug repurposing methodology that identified raloxifene as a potential anti-Pseudomonal agent may be applied to the discovery of existing drugs that may serve as novel antivirulence agent against either the transcriptional repressor or the metallopeptidase.

6.3. Relevance and impact

This thesis work built and improved upon previous PAGs analyses to provide further insight into functional importance of PAGs using additional methodologies such as orthology and positive selection analyses. This is also the first analysis of non-PAGs, which were shown to have limited taxonomic conservation similar to PAGs, suggesting that these genes may have specialized function for bacterial adaptation to the non-pathogenic lifestyle. This work contributes to the ongoing antivirulence drug discovery effort by prioritizing pathogen-specific genes with probable virulence function and accessible subcellular localization as drug targets. To gain a better understanding of the drug accessibility of PAGs, I was involved in the update of the PSORTdb protein SCL database which I subsequently used to perform the global protein SCL analysis of all bacterial proteins by pathogen association. My work with PSORT family of tools extends to the optimization and maintenance of the PSORTb and PSORTm protein SCL predictors for bacterial genomic data and metagenomic data, respectively. In addition, the structure-activity relationship analysis in Chapter 4 revealed the structural components of raloxifene responsible for its antivirulence activity. These results help to advance the repurposing of raloxifene as an antivirulence therapeutic against the high priority pathogen *P. aeruginosa* noted for its intrinsic AMR and the need to develop new therapies against it. I also performed the first metagenomics-based analysis of PAGs to explore the prevalence of these genes in both clinical and environmental conditions with varying microbial diversity.

Overall, I have developed a computational screening approach to identify putative virulence-related candidates from PAGs that may help uncover novel bacterial

virulence mechanisms. This approach can serve as a preliminary genomic-based platform for identifying and prioritizing candidate antivirulence drug targets, that is applicable to a variety of bacterial pathogens.

References

Allen, R. C., Popat, R., Diggle, S. P., & Brown, S. P. (2014). Targeting virulence: can we make evolution-proof drugs? *Nat Rev Microbiol, 12*(4), 300-308. doi:10.1038/nrmicro3232

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Aminov, R. I. (2010). A brief history of the antibiotic era: lessons learned and challenges for the future. *Front Microbiol, 1*, 134. doi:10.3389/fmicb.2010.00134

Antczak, M., Michaelis, M., & Wass, M. N. (2019). Environmental conditions shape the nature of a minimal bacterial genome. *Nat Commun, 10*(1), 3100. doi:10.1038/s41467-019-10837-2

Antimicrobial Resistance, C. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*. doi:10.1016/S0140-6736(21)02724-0

Arenas, M. (2015). Trends in substitution models of molecular evolution. *Front Genet, 6*, 319. doi:10.3389/fgene.2015.00319

Assaf, R., Xia, F., & Stevens, R. (2021). Detecting operons in bacterial genomes via visual representation learning. *Sci Rep, 11*(1), 2124. doi:10.1038/s41598-021-81169-9

Bacci, G., Taccetti, G., Dolce, D., Armanini, F., Segata, N., Di Cesare, F., . . . Bevivino, A. (2020). Untargeted Metagenomic Investigation of the Airway Microbiome of Cystic Fibrosis Patients with Moderate-Severe Lung Disease. *Microorganisms, 8*(7). doi:10.3390/microorganisms8071003

Balla, K. M., & Troemel, E. R. (2013). Caenorhabditis elegans as a model for intracellular pathogen infection. *Cell Microbiol, 15*(8), 1313-1322. doi:10.1111/cmi.12152

Barco, R. A., Garrity, G. M., Scott, J. J., Amend, J. P., Nealson, K. H., & Emerson, D. (2020). A Genus Definition for Bacteria and Archaea Based on a Standard Genome Relatedness Index. *mBio, 11*(1). doi:10.1128/mBio.02475-19

Barie, P. S. (2000). Modern surgical antibiotic prophylaxis and therapy--less is more. *Surg Infect (Larchmt), 1*(1), 23-29. doi:10.1089/109629600321263

Bassetti, M., Vena, A., Croxatto, A., Righi, E., & Guery, B. (2018). How to manage Pseudomonas aeruginosa infections. *Drugs Context, 7*, 212527. doi:10.7573/dic.212527

Bayes, H. K., Ritchie, N., Irvine, S., & Evans, T. J. (2016). A murine model of early Pseudomonas aeruginosa lung disease with transition to chronic infection. *Sci Rep, 6*, 35838. doi:10.1038/srep35838

Beam, J. E., Rowe, S. E., & Conlon, B. P. (2021). Shooting yourself in the foot: How immune cells induce antibiotic tolerance in microbial pathogens. *PLoS Pathog, 17*(7), e1009660. doi:10.1371/journal.ppat.1009660

Bengtsson-Palme, J., Larsson, D. G. J., & Kristiansson, E. (2017). Using metagenomics to investigate human and environmental resistomes. *J Antimicrob Chemother, 72*(10), 2690-2703. doi:10.1093/jac/dkx199

Bertelli, C., Laird, M. R., Williams, K. P., Simon Fraser University Research Computing, G., Lau, B. Y., Hoad, G., . . . Brinkman, F. S. L. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res, 45*(W1), W30-W35. doi:10.1093/nar/gkx343

Bertelli, C., Tilley, K. E., & Brinkman, F. S. L. (2019). Microbial genomic island discovery, visualization and analysis. *Brief Bioinform, 20*(5), 1685-1698. doi:10.1093/bib/bby042

Bettiol, E., Wetherington, J. D., Schmitt, N., Harbarth, S., & Consortium, C. (2015). Challenges and solutions for clinical development of new antibacterial agents: results of a survey among pharmaceutical industry professionals. *Antimicrob Agents Chemother, 59*(7), 3695-3699. doi:10.1128/AAC.00638-15

Blankenfeldt, W., & Parsons, J. F. (2014). The structural biology of phenazine biosynthesis. *Curr Opin Struct Biol, 29*, 26-33. doi:10.1016/j.sbi.2014.08.013

Bliven, K. A., & Maurelli, A. T. (2012). Antivirulence genes: insights into pathogen evolution through gene loss. *Infect Immun, 80*(12), 4061-4070. doi:10.1128/IAI.00740-12

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Brambila-Tapia, A. J., Armenta-Medina, D., Rivera-Gomez, N., & Perez-Rueda, E. (2014). Main functions and taxonomic distribution of virulence genes in Brucella melitensis 16 M. *PLoS One, 9*(6), e100349. doi:10.1371/journal.pone.0100349

Brandenburg, K. S., Weaver, A. J., Jr., Karna, S. L. R., You, T., Chen, P., Stryk, S. V., . . . Leung, K. P. (2019). Formation of Pseudomonas aeruginosa Biofilms in Full-thickness Scald Burn Wounds in Rats. *Sci Rep, 9*(1), 13627. doi:10.1038/s41598-019-50003-8

Breijyeh, Z., Jubeh, B., & Karaman, R. (2020). Resistance of Gram-Negative Bacteria to Current Antibacterial Agents and Approaches to Resolve It. *Molecules, 25*(6). doi:10.3390/molecules25061340

Briard, B., Bomme, P., Lechner, B. E., Mislin, G. L., Lair, V., Prevost, M. C., . . . Beauvais, A. (2015). Pseudomonas aeruginosa manipulates redox and iron homeostasis of its microbiota partner Aspergillus fumigatus via phenazines. *Sci Rep, 5*, 8220. doi:10.1038/srep08220

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods, 18*(4), 366-368. doi:10.1038/s41592-021-01101-x

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods, 12*(1), 59-60. doi:10.1038/nmeth.3176

Burmeister, A. R., Fortier, A., Roush, C., Lessing, A. J., Bender, R. G., Barahman, R., . . . Turner, P. E. (2020). Pleiotropy complicates a trade-off between phage resistance and antibiotic resistance. *Proc Natl Acad Sci U S A, 117*(21), 11207-11216. doi:10.1073/pnas.1919888117

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. *Nat Rev Genet, 21*(9), 526-540. doi:10.1038/s41576-020-0244-x

Carr, V. R., & Chaguza, C. (2021). Metagenomics for surveillance of respiratory pathogens. *Nat Rev Microbiol, 19*(5), 285. doi:10.1038/s41579-021-00541-8

Casadevall, A. (2017). The Pathogenic Potential of a Microbe. *mSphere, 2*(1). doi:10.1128/mSphere.00015-17

Casadevall, A., & Pirofski, L. A. (2003). The damage-response framework of microbial pathogenesis. *Nat Rev Microbiol, 1*(1), 17-24. doi:10.1038/nrmicro732

Cazares, A., Moore, M. P., Hall, J. P. J., Wright, L. L., Grimes, M., Emond-Rheault, J. G., . . . Winstanley, C. (2020). A megaplasmid family driving dissemination of multidrug resistance in Pseudomonas. *Nat Commun, 11*(1), 1370. doi:10.1038/s41467-020-15081-7

CDC. (2019). *Antibiotic Resistance Threats in the United States*. Retrieved from Atlanta, GA:

Cecil, J. D., Sirisaengtaksin, N., O'Brien-Simpson, N. M., & Krachler, A. M. (2019). Outer Membrane Vesicle-Host Cell Interactions. *Microbiol Spectr, 7*(1). doi:10.1128/microbiolspec.PSIB-0001-2018

Cerda-Costa, N., & Gomis-Ruth, F. X. (2014). Architecture and function of metallopeptidase catalytic domains. *Protein Sci, 23*(2), 123-144. doi:10.1002/pro.2400

Cezairliyan, B., Vinayavekhin, N., Grenfell-Lee, D., Yuen, G. J., Saghatelian, A., & Ausubel, F. M. (2013). Identification of Pseudomonas aeruginosa phenazines that kill Caenorhabditis elegans. *PLoS Pathog, 9*(1), e1003101. doi:10.1371/journal.ppat.1003101

Chen, J., Xia, Y., Cheng, C., Fang, C., Shan, Y., Jin, G., & Fang, W. (2011). Genome sequence of the nonpathogenic Listeria monocytogenes serovar 4a strain M7. *J Bacteriol, 193*(18), 5019-5020. doi:10.1128/JB.05501-11

Chew, K. L., Octavia, S., Ng, O. T., Marimuthu, K., Venkatachalam, I., Cheng, B., . . . Teo, J. W. P. (2019). Challenge of drug resistance in Pseudomonas aeruginosa: clonal spread of NDM-1-positive ST-308 within a tertiary hospital. *J Antimicrob Chemother, 74*(8), 2220-2224. doi:10.1093/jac/dkz169

Chiu, C. Y., & Miller, S. A. (2019). Clinical metagenomics. *Nat Rev Genet, 20*(6), 341-355. doi:10.1038/s41576-019-0113-7

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods Mol Biol, 1418*, 93-110. doi:10.1007/978-1-4939-3578-9_5

Colvin, K. M., Irie, Y., Tart, C. S., Urbano, R., Whitney, J. C., Ryder, C., . . . Parsek, M. R. (2012). The Pel and Psl polysaccharides provide Pseudomonas aeruginosa structural redundancy within the biofilm matrix. *Environ Microbiol, 14*(8), 1913-1928. doi:10.1111/j.1462-2920.2011.02657.x

Corey, A., Migone, T. S., Bolmer, S., Fiscella, M., Ward, C., Chen, C., & Meister, G. (2013). Bacillus anthracis protective antigen kinetics in inhalation spore-challenged untreated or levofloxacin/ raxibacumab-treated New Zealand white rabbits. *Toxins (Basel), 5*(1), 120-138. doi:10.3390/toxins5010120

Coutinho, H. D., Falcao-Silva, V. S., & Goncalves, G. F. (2008). Pulmonary bacterial pathogens in cystic fibrosis patients and antibiotic therapy: a tool for the health workers. *Int Arch Med, 1*(1), 24. doi:10.1186/1755-7682-1-24

Curran, B., Jonas, D., Grundmann, H., Pitt, T., & Dowson, C. G. (2004). Development of a multilocus sequence typing scheme for the opportunistic pathogen Pseudomonas aeruginosa. *J Clin Microbiol, 42*(12), 5644-5649. doi:10.1128/JCM.42.12.5644-5649.2004

Curran, C. S., Bolig, T., & Torabi-Parizi, P. (2018). Mechanisms and Targeted Therapies for Pseudomonas aeruginosa Lung Infection. *Am J Respir Crit Care Med, 197*(6), 708-727. doi:10.1164/rccm.201705-1043SO

D'Angelo, F., Baldelli, V., Halliday, N., Pantalone, P., Polticelli, F., Fiscarelli, E., . . . Rampioni, G. (2018). Identification of FDA-Approved Drugs as Antivirulence Agents Targeting the pqs Quorum-Sensing System of Pseudomonas aeruginosa. *Antimicrob Agents Chemother, 62*(11). doi:10.1128/AAC.01296-18

Dalbey, R. E., Wang, P., & van Dijl, J. M. (2012). Membrane proteases in the bacterial protein secretion and quality control pathway. *Microbiol Mol Biol Rev, 76*(2), 311-330. doi:10.1128/MMBR.05019-11

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., . . . Stevens, R. (2020). The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res, 48*(D1), D606-D612. doi:10.1093/nar/gkz943

de Abreu, V. A. C., Perdigao, J., & Almeida, S. (2020). Metagenomic Approaches to Analyze Antimicrobial Resistance: An Overview. *Front Genet, 11*, 575592. doi:10.3389/fgene.2020.575592

de Nies, L., Lopes, S., Busi, S. B., Galata, V., Heintz-Buschart, A., Laczny, C. C., . . . Wilmes, P. (2021). PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome, 9*(1), 49. doi:10.1186/s40168-020-00993-9

De, R., Mukhopadhyay, A. K., & Dutta, S. (2020). Metagenomic analysis of gut microbiome and resistome of diarrheal fecal samples from Kolkata, India, reveals the core and variable microbiota including signatures of microbial dark matter. *Gut Pathog, 12*, 32. doi:10.1186/s13099-020-00371-8

Deng, W., Marshall, N. C., Rowland, J. L., McCoy, J. M., Worrall, L. J., Santos, A. S., . . . Finlay, B. B. (2017). Assembly, structure, function and regulation of type III secretion systems. *Nat Rev Microbiol, 15*(6), 323-337. doi:10.1038/nrmicro.2017.20

Destoumieux-Garzon, D., Mavingui, P., Boetsch, G., Boissier, J., Darriet, F., Duboz, P., . . . Voituron, Y. (2018). The One Health Concept: 10 Years Old and a Long Road Ahead. *Front Vet Sci, 5*, 14. doi:10.3389/fvets.2018.00014

Dettman, J. R., & Kassen, R. (2021). Evolutionary Genomics of Niche-Specific Adaptation to the Cystic Fibrosis Lung in Pseudomonas aeruginosa. *Mol Biol Evol, 38*(2), 663-675. doi:10.1093/molbev/msaa226

Dey, S., Chakravarty, A., Guha Biswas, P., & De Guzman, R. N. (2019). The type III secretion system needle, tip, and translocon. *Protein Sci, 28*(9), 1582-1593. doi:10.1002/pro.3682

Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., . . . Brinkman, F. S. (2015). IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res, 43*(W1), W104-108. doi:10.1093/nar/gkv401

Dickey, S. W., Cheung, G. Y. C., & Otto, M. (2017). Different drugs for bad bugs: antivirulence strategies in the age of antibiotic resistance. *Nat Rev Drug Discov, 16*(7), 457-471. doi:10.1038/nrd.2017.23

Dickinson, A. W., Power, A., Hansen, M. G., Brandt, K. K., Piliposian, G., Appleby, P., . . . Vos, M. (2019). Heavy metal pollution and co-selection for antibiotic resistance: A microbial palaeontology approach. *Environ Int, 132*, 105117. doi:10.1016/j.envint.2019.105117

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol, 2*(5), 414-424. doi:10.1038/nrmicro884

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi:10.1093/nar/gkh340

Elborn, J. S. (2016). Cystic fibrosis. *Lancet, 388*(10059), 2519-2531. doi:10.1016/S0140-6736(16)00576-6

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol, 20*(1), 238. doi:10.1186/s13059-019-1832-y

Escudeiro, P., Pothier, J., Dionisio, F., & Nogueira, T. (2019). Antibiotic Resistance Gene Diversity and Virulence Gene Diversity Are Correlated in Human Gut and Environmental Microbiomes. *mSphere, 4*(3). doi:10.1128/mSphere.00135-19

Espinosa-Camacho, L. F., Delgado, G., Soberon-Chavez, G., Alcaraz, L. D., Castanon, J., & Morales-Espinosa, R. (2017). Complete Genome Sequences of Four Extensively Drug-Resistant Pseudomonas aeruginosa Strains, Isolated from Adults with Ventilator-Associated Pneumonia at a Tertiary Referral Hospital in Mexico City. *Genome Announc, 5*(36). doi:10.1128/genomeA.00925-17

Essack, S. Y. (2018). Environment: the neglected component of the One Health triad. *Lancet Planet Health, 2*(6), e238-e239. doi:10.1016/S2542-5196(18)30124-4

Essar, D. W., Eberly, L., Hadero, A., & Crawford, I. P. (1990). Identification and characterization of genes for a second anthranilate synthase in Pseudomonas aeruginosa: interchangeability of the two anthranilate synthases and evolutionary implications. *J Bacteriol, 172*(2), 884-900. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2153661

Evans, B. R., & Leighton, F. A. (2014). A history of One Health. *Rev Sci Tech, 33*(2), 413-420. doi:10.20506/rst.33.2.2298

Ewbank, J. J., & Zugasti, O. (2011). C. elegans: model host and tool for antimicrobial drug discovery. *Dis Model Mech, 4*(3), 300-304. doi:10.1242/dmm.006684

Fabrega, A., Madurga, S., Giralt, E., & Vila, J. (2009). Mechanism of action of and resistance to quinolones. *Microb Biotechnol, 2*(1), 40-61. doi:10.1111/j.1751-7915.2008.00063.x

Fajardo-Lubian, A., Ben Zakour, N. L., Agyekum, A., Qi, Q., & Iredell, J. R. (2019). Host adaptation and convergent evolution increases antibiotic resistance without loss of virulence in a major human pathogen. *PLoS Pathog, 15*(3), e1007218. doi:10.1371/journal.ppat.1007218

Falkow, S. (1988). Molecular Koch's postulates applied to microbial pathogenicity. *Rev Infect Dis, 10 Suppl 2*, S274-276. doi:10.1093/cid/10.supplement_2.s274

Fedynak, A. (2007). *Quantifying trends in bacterial virulence and pathogen-associated genes through large scale bioinformatics analysis*

. (PhD). Simon Fraser Univesity,

Feigelman, R., Kahlert, C. R., Baty, F., Rassouli, F., Kleiner, R. L., Kohler, P., . . . von Mering, C. (2017). Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. *Microbiome, 5*(1), 20. doi:10.1186/s40168-017-0234-1

Feltman, H., Schulert, G., Khan, S., Jain, M., Peterson, L., & Hauser, A. R. (2001). Prevalence of type III secretion genes in clinical and environmental isolates of Pseudomonas aeruginosa. *Microbiology (Reading), 147*(Pt 10), 2659-2669. doi:10.1099/00221287-147-10-2659

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., . . . Stoeckert, C. J., Jr. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr Protoc Bioinformatics, Chapter 6*, Unit 6 12 11-19. doi:10.1002/0471250953.bi0612s35

Fleitas Martinez, O., Cardoso, M. H., Ribeiro, S. M., & Franco, O. L. (2019). Recent Advances in Anti-virulence Therapeutic Strategies With a Focus on Dismantling Bacterial Membrane Microdomains, Toxin Neutralization, Quorum-Sensing Interference and Biofilm Inhibition. *Front Cell Infect Microbiol, 9*, 74. doi:10.3389/fcimb.2019.00074

Fleming, A. (2001). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929. *Bull World Health Organ, 79*(8), 780-790. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11545337

Fredricks, D. N., & Relman, D. A. (1996). Sequence-based identification of microbial pathogens: a reconsideration of Koch's postulates. *Clin Microbiol Rev, 9*(1), 18-33. doi:10.1128/CMR.9.1.18

Fulton, D. L., Li, Y. Y., Laird, M. R., Horsman, B. G., Roche, F. M., & Brinkman, F. S. (2006). Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics, 7*, 270. doi:10.1186/1471-2105-7-270

Gal-Mor, O., & Finlay, B. B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol, 8*(11), 1707-1719. doi:10.1111/j.1462-5822.2006.00794.x

Galperin, M. Y., & Koonin, E. V. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res, 32*(18), 5452-5463. doi:10.1093/nar/gkh885

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res, 49*(D1), D274-D281. doi:10.1093/nar/gkaa1018

Gardy, J. L., & Brinkman, F. S. (2006). Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol, 4*(10), 741-751. doi:10.1038/nrmicro1494

Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., & Brinkman, F. S. (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics, 21*(5), 617-623. doi:10.1093/bioinformatics/bti057

Gardy, J. L., Spencer, C., Wang, K., Ester, M., Tusnady, G. E., Simon, I., . . . Brinkman, F. S. (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res, 31*(13), 3613-3617. doi:10.1093/nar/gkg602

Garg, A., & Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics, 9*, 62. doi:10.1186/1471-2105-9-62

Gatt, Y. E., & Margalit, H. (2021). Common Adaptive Strategies Underlie Within-Host Evolution of Bacterial Pathogens. *Mol Biol Evol, 38*(3), 1101-1121. doi:10.1093/molbev/msaa278

Gigliucci, F., von Meijenfeldt, F. A. B., Knijn, A., Michelacci, V., Scavia, G., Minelli, F., . . . Morabito, S. (2018). Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients. *Front Cell Infect Microbiol, 8*, 25. doi:10.3389/fcimb.2018.00025

Glasser, N. R., Wang, B. X., Hoy, J. A., & Newman, D. K. (2017). The Pyruvate and alpha-Ketoglutarate Dehydrogenase Complexes of Pseudomonas aeruginosa Catalyze Pyocyanin and Phenazine-1-carboxylic Acid Reduction via the Subunit Dihydrolipoamide Dehydrogenase. *J Biol Chem, 292*(13), 5593-5607. doi:10.1074/jbc.M116.772848

Guindon, S., Rodrigo, A. G., Dyer, K. A., & Huelsenbeck, J. P. (2004). Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A, 101*(35), 12957-12962. doi:10.1073/pnas.0402177101

Gupta, A., & Sharma, V. K. (2015). Using the taxon-specific genes for the taxonomic classification of bacterial genomes. *BMC Genomics, 16*, 396. doi:10.1186/s12864-015-1542-0

Gurney, J., Pradier, L., Griffin, J. S., Gougat-Barbera, C., Chan, B. K., Turner, P. E., . . . Hochberg, M. E. (2020). Phage steering of antibiotic-resistance evolution in the bacterial pathogen, Pseudomonas aeruginosa. *Evol Med Public Health, 2020*(1), 148-157. doi:10.1093/emph/eoaa026

Haft, D. H., Selengut, J. D., Brinkac, L. M., Zafar, N., & White, O. (2005). Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics, 21*(3), 293-306. doi:10.1093/bioinformatics/bti015

Hall, S., McDermott, C., Anoopkumar-Dukie, S., McFarland, A. J., Forbes, A., Perkins, A. V., . . . Grant, G. D. (2016). Cellular Effects of Pyocyanin, a Secreted Virulence Factor of Pseudomonas aeruginosa. *Toxins (Basel), 8*(8). doi:10.3390/toxins8080236

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev, 68*(4), 669-685. doi:10.1128/MMBR.68.4.669-685.2004

Haque, M., Sartelli, M., McKimm, J., & Abu Bakar, M. (2018). Health care-associated infections - an overview. *Infect Drug Resist, 11*, 2321-2333. doi:10.2147/IDR.S177247

Hart, C. A., & Winstanley, C. (2002). Persistent and aggressive bacteria in the lungs of cystic fibrosis children. *Br Med Bull, 61*, 81-96. doi:10.1093/bmb/61.1.81

Hauser, A. R. (2009). The type III secretion system of Pseudomonas aeruginosa: infection by injection. *Nat Rev Microbiol, 7*(9), 654-665. doi:10.1038/nrmicro2199

Hisert, K. B., Heltshe, S. L., Pope, C., Jorth, P., Wu, X., Edwards, R. M., . . . Singh, P. K. (2017). Restoring Cystic Fibrosis Transmembrane Conductance Regulator Function Reduces Airway Bacteria and Inflammation in People with Cystic Fibrosis and Chronic Lung Infections. *Am J Respir Crit Care Med, 195*(12), 1617-1628. doi:10.1164/rccm.201609-1954OC

Hmelo, L. R., Borlee, B. R., Almblad, H., Love, M. E., Randall, T. E., Tseng, B. S., . . . Harrison, J. J. (2015). Precision-engineering the Pseudomonas aeruginosa genome with two-step allelic exchange. *Nat Protoc, 10*(11), 1820-1841. doi:10.1038/nprot.2015.115

Ho Sui, S. J., Fedynak, A., Hsiao, W. W., Langille, M. G., & Brinkman, F. S. (2009). The association of virulence factors with genomic islands. *PLoS One, 4*(12), e8094. doi:10.1371/journal.pone.0008094

Ho Sui, S. J., Fedynak, A., Hsiao, W.W.L, Langille, M.G.I & Brinkman, F.S.L. (2009). The Association of Virulence Factors with Genomics Islands. *PLoS One, 4*(12).

Ho Sui, S. J., Lo, R., Fernandes, A. R., Caulfield, M. D., Lerman, J. A., Xie, L., . . . Brinkman, F. S. (2012). Raloxifene attenuates Pseudomonas aeruginosa pyocyanin production and virulence. *Int J Antimicrob Agents, 40*(3), 246-251. doi:10.1016/j.ijantimicag.2012.05.009

Hotinger, J. A., Morris, S. T., & May, A. E. (2021). The Case against Antibiotics and for Anti-Virulence Therapeutics. *Microorganisms, 9*(10). doi:10.3390/microorganisms9102049

Housseini, B. I. K., Phan, G., & Broutin, I. (2018). Functional Mechanism of the Efflux Pumps Transcription Regulators From Pseudomonas aeruginosa Based on 3D Structures. *Front Mol Biosci, 5*, 57. doi:10.3389/fmolb.2018.00057

Hsiao, W. W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B. B., & Brinkman, F. S. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet, 1*(5), e62. doi:10.1371/journal.pgen.0010062

Huang, C., Chen, H., Ding, Y., Ma, X., Zhu, H., Zhang, S., . . . Feng, Y. (2021). A Microbial World: Could Metagenomic Next-Generation Sequencing Be Involved in Acute Respiratory Failure? *Front Cell Infect Microbiol, 11*, 738074. doi:10.3389/fcimb.2021.738074

Hueck, C. J. (1998). Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev, 62*(2), 379-433. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9618447

Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D. H., & Soo, R. M. (2021). Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J, 15*(7), 1879-1892. doi:10.1038/s41396-021-00941-x

Hussein, M., Han, M. L., Zhu, Y., Schneider-Futschik, E. K., Hu, X., Zhou, Q. T., . . . Velkov, T. (2018). Mechanistic Insights From Global Metabolomics Studies into Synergistic Bactericidal Effect of a Polymyxin B Combination With Tamoxifen Against Cystic Fibrosis MDR Pseudomonas aeruginosa. *Comput Struct Biotechnol J, 16*, 587-599. doi:10.1016/j.csbj.2018.11.001

Hussein, M. H., Schneider, E. K., Elliott, A. G., Han, M., Reyes-Ortega, F., Morris, F., . . . Velkov, T. (2017). From Breast Cancer to Antimicrobial: Combating Extremely Resistant Gram-Negative "Superbugs" Using Novel Combinations of Polymyxin B with Selective Estrogen Receptor Modulators. *Microb Drug Resist, 23*(5), 640-650. doi:10.1089/mdr.2016.0196

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11*, 119. doi:10.1186/1471-2105-11-119

Jacobs, A. C., Didone, L., Jobson, J., Sofia, M. K., Krysan, D., & Dunman, P. M. (2013). Adenylate kinase release as a high-throughput-screening-compatible reporter of bacterial lysis for identification of antibacterial agents. *Antimicrob Agents Chemother, 57*(1), 26-36. doi:10.1128/AAC.01640-12

Jan, A. T. (2017). Outer Membrane Vesicles (OMVs) of Gram-negative Bacteria: A Perspective Update. *Front Microbiol, 8*, 1053. doi:10.3389/fmicb.2017.01053

Jani, M., & Azad, R. K. (2021). Discovery of mosaic genomic islands in Pseudomonas spp. *Arch Microbiol, 203*(5), 2735-2742. doi:10.1007/s00203-021-02253-2

Jiang, P., Wu, S., Luo, Q., Zhao, X. M., & Chen, W. H. (2021). Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models. *mSystems, 6*(3). doi:10.1128/mSystems.00112-21

Jones-Nelson, O., Tovchigrechko, A., Glover, M. S., Fernandes, F., Rangaswamy, U., Liu, H., . . . Sellman, B. R. (2020). Antibacterial Monoclonal Antibodies Do Not Disrupt the Intestinal Microbiome or Its Function. *Antimicrob Agents Chemother, 64*(5). doi:10.1128/AAC.02347-19

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res, 12*(6), 962-968. doi:10.1101/gr.87702

Jordan, V. C., Curpan, R., & Maximov, P. Y. (2015). Estrogen receptor mutations found in breast cancer metastases integrated with the molecular pharmacology of selective ER modulators. *J Natl Cancer Inst, 107*(6), djv075. doi:10.1093/jnci/djv075

Kamath, S., Kapatral, V., & Chakrabarty, A. M. (1998). Cellular function of elastase in Pseudomonas aeruginosa: role in the cleavage of nucleoside diphosphate kinase and in alginate synthesis. *Mol Microbiol, 30*(5), 933-941. doi:10.1046/j.1365-2958.1998.01121.x

Kang, Y., Kim, H., Goo, E., Jeong, H., An, J. H., & Hwang, I. (2019). Unraveling the role of quorum sensing-dependent metabolic homeostasis of the activated methyl cycle in a cooperative population of Burkholderia glumae. *Sci Rep, 9*(1), 11038. doi:10.1038/s41598-019-47460-6

Kim, D. W., & Cha, C. J. (2021). Antibiotic resistome from the One-Health perspective: understanding and controlling antimicrobial resistance transmission. *Exp Mol Med, 53*(3), 301-309. doi:10.1038/s12276-021-00569-z

Kirst, M. E., Baker, D., Li, E., Abu-Hasan, M., & Wang, G. P. (2019). Upper versus lower airway microbiome and metagenome in children with cystic fibrosis and their correlation with lung inflammation. *PLoS One, 14*(9), e0222323. doi:10.1371/journal.pone.0222323

Kiyohara, M., Tanigawa, K., Chaiwangsri, T., Katayama, T., Ashida, H., & Yamamoto, K. (2011). An exo-alpha-sialidase from bifidobacteria involved in the degradation of sialyloligosaccharides in human milk and intestinal glycoconjugates. *Glycobiology, 21*(4), 437-447. doi:10.1093/glycob/cwq175

Kosakovsky Pond, S. L., & Frost, S. D. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol, 22*(5), 1208-1222. doi:10.1093/molbev/msi105

Kosakovsky Pond, S. L., Poon, A. F. Y., Velazquez, R., Weaver, S., Hepler, N. L., Murrell, B., . . . Muse, S. V. (2020). HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol, 37*(1), 295-299. doi:10.1093/molbev/msz197

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., & Frost, S. D. (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatics, 22*(24), 3096-3098. doi:10.1093/bioinformatics/btl474

Kozbial, P. Z., & Mushegian, A. R. (2005). Natural history of S-adenosylmethionine-binding proteins. *BMC Struct Biol, 5*, 19. doi:10.1186/1472-6807-5-19

Krause, K. M., Serio, A. W., Kane, T. R., & Connolly, L. E. (2016). Aminoglycosides: An Overview. *Cold Spring Harb Perspect Med, 6*(6). doi:10.1101/cshperspect.a027029

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol, 305*(3), 567-580. doi:10.1006/jmbi.2000.4315

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genet, 4*(12), e1000304. doi:10.1371/journal.pgen.1000304

Kuehl, C. J., & Crosa, J. H. (2010). The TonB energy transduction systems in Vibrio species. *Future Microbiol, 5*(9), 1403-1412. doi:10.2217/fmb.10.90

Ladomersky, E., & Petris, M. J. (2015). Copper tolerance and virulence in bacteria. *Metallomics, 7*(6), 957-964. doi:10.1039/c4mt00327f

Lalancette, C., Charron, D., Laferriere, C., Dolce, P., Deziel, E., Prevost, M., & Bedard, E. (2017). Hospital Drains as Reservoirs of Pseudomonas aeruginosa: Multiple-Locus Variable-Number of Tandem Repeats Analysis Genotypes Recovered from Faucets, Sink Surfaces and Patients. *Pathogens, 6*(3). doi:10.3390/pathogens6030036

Langille, M. G., Hsiao, W. W., & Brinkman, F. S. (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics, 9*, 329. doi:10.1186/1471-2105-9-329

Langille, M. G., Laird, M. R., Hsiao, W. W., Chiu, T. A., Eisen, J. A., & Brinkman, F. S. (2012). MicrobeDB: a locally maintainable database of microbial genomic sequences. *Bioinformatics, 28*(14), 1947-1948. doi:10.1093/bioinformatics/bts273

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Lau, W. Y. V., Hoad, G. R., Jin, V., Winsor, G. L., Madyan, A., Gray, K. L., . . . Brinkman, F. S. L. (2021). PSORTdb 4.0: expanded and redesigned bacterial and archaeal protein subcellular localization database incorporating new secondary localizations. *Nucleic Acids Res, 49*(D1), D803-D808. doi:10.1093/nar/gkaa1095

Levenson, A. S., & Jordan, V. C. (1998). The key to the antiestrogenic mechanism of raloxifene is amino acid 351 (aspartate) in the estrogen receptor. *Cancer Res, 58*(9), 1872-1875. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9581827

Liang, J., Mao, G., Yin, X., Ma, L., Liu, L., Bai, Y., . . . Qu, J. (2020). Identification and quantification of bacterial genomes carrying antibiotic resistance genes and virulence factor genes for aquatic microbiological risk assessment. *Water Res, 168*, 115160. doi:10.1016/j.watres.2019.115160

Liberati, N. T., Urbach, J. M., Miyata, S., Lee, D. G., Drenkard, E., Wu, G., . . . Ausubel, F. M. (2006). An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proc Natl Acad Sci U S A, 103*(8), 2833-2838. doi:10.1073/pnas.0511100103

Lin, D. M., Koskella, B., & Lin, H. C. (2017). Phage therapy: An alternative to antibiotics in the age of multi-drug resistance. *World J Gastrointest Pharmacol Ther, 8*(3), 162-173. doi:10.4292/wjgpt.v8.i3.162

Liu, B., Zheng, D., Jin, Q., Chen, L., & Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res, 47*(D1), D687-D692. doi:10.1093/nar/gky1080

Liu, Y., & Sun, J. (2021). Detection of Pathogens and Regulation of Immunity by the Caenorhabditis elegans Nervous System. *mBio, 12*(2). doi:10.1128/mBio.02301-20

Lorenz, B., Ali, N., Bocklitz, T., Rosch, P., & Popp, J. (2020). Discrimination between pathogenic and non-pathogenic E. coli strains by means of Raman microspectroscopy. *Anal Bioanal Chem, 412*(30), 8241-8247. doi:10.1007/s00216-020-02957-2

Ma, Y., Zhang, Y., Xiang, J., Xiang, S., Zhao, Y., Xiao, M., . . . Xiao, Z. (2021). Metagenome Analysis of Intestinal Bacteria in Healthy People, Patients With Inflammatory Bowel Disease and Colorectal Cancer. *Front Cell Infect Microbiol, 11*, 599734. doi:10.3389/fcimb.2021.599734

Mackenzie, J. S., & Jeggo, M. (2019). The One Health Approach-Why Is It So Important? *Trop Med Infect Dis, 4*(2). doi:10.3390/tropicalmed4020088

Maeda, T., Garcia-Contreras, R., Pu, M., Sheng, L., Garcia, L. R., Tomas, M., & Wood, T. K. (2012). Quorum quenching quandary: resistance to antivirulence compounds. *ISME J, 6*(3), 493-501. doi:10.1038/ismej.2011.122

Mahajan-Miklos, S., Rahme, L. G., & Ausubel, F. M. (2000). Elucidating the molecular mechanisms of bacterial virulence using non-mammalian hosts. *Mol Microbiol, 37*(5), 981-988. doi:10.1046/j.1365-2958.2000.02056.x

Mavrodi, D. V., Bonsall, R. F., Delaney, S. M., Soule, M. J., Phillips, G., & Thomashow, L. S. (2001). Functional analysis of genes for biosynthesis of pyocyanin and phenazine-1-carboxamide from Pseudomonas aeruginosa PAO1. *J Bacteriol, 183*(21), 6454-6465. doi:10.1128/JB.183.21.6454-6465.2001

McEwen, S. A., & Collignon, P. J. (2018). Antimicrobial Resistance: a One Health Perspective. *Microbiol Spectr, 6*(2). doi:10.1128/microbiolspec.ARBA-0009-2017

Mikkelsen, H., McMullan, R., & Filloux, A. (2011). The Pseudomonas aeruginosa reference strain PA14 displays increased virulence due to a mutation in ladS. *PLoS One, 6*(12), e29113. doi:10.1371/journal.pone.0029113

Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M., & Tang, P. (2013). Metagenomics for pathogen detection in public health. *Genome Med, 5*(9), 81. doi:10.1186/gm485

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol, 37*(5), 1530-1534. doi:10.1093/molbev/msaa015

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., . . . Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res, 49*(D1), D412-D419. doi:10.1093/nar/gkaa913

Molina-Mora, J. A., Chinchilla-Montero, D., Chavarria-Azofeifa, M., Ulloa-Morales, A. J., Campos-Sanchez, R., Mora-Rodriguez, R., . . . Garcia, F. (2020). Transcriptomic determinants of the response of ST-111 Pseudomonas aeruginosa AG1 to ciprofloxacin identified by a top-down systems biology approach. *Sci Rep, 10*(1), 13717. doi:10.1038/s41598-020-70581-2

Montoya, M. C., & Krysan, D. J. (2018). Repurposing Estrogen Receptor Antagonists for the Treatment of Infectious Disease. *mBio, 9*(6). doi:10.1128/mBio.02272-18

Moore, M. P., Lamont, I. L., Williams, D., Paterson, S., Kukavica-Ibrulj, I., Tucker, N. P., . . . Winstanley, C. (2021). Transmission, adaptation and geographical spread of the Pseudomonas aeruginosa Liverpool epidemic strain. *Microb Genom, 7*(3). doi:10.1099/mgen.0.000511

Moradali, M. F., Ghods, S., & Rehm, B. H. (2017). Pseudomonas aeruginosa Lifestyle: A Paradigm for Adaptation, Survival, and Persistence. *Front Cell Infect Microbiol, 7*, 39. doi:10.3389/fcimb.2017.00039

Moragues-Solanas, L., Scotti, R., & O'Grady, J. (2021). Rapid metagenomics for diagnosis of bloodstream and respiratory tract nosocomial infections: current status and future prospects. *Expert Rev Mol Diagn, 21*(4), 371-380. doi:10.1080/14737159.2021.1906652

Morel, C. M., & Mossialos, E. (2010). Stoking the antibiotic pipeline. *BMJ, 340*, c2115. doi:10.1136/bmj.c2115

Morrow, K. A., Ochoa, C. D., Balczon, R., Zhou, C., Cauthen, L., Alexeyev, M., . . . Stevens, T. (2016). Pseudomonas aeruginosa exoenzymes U and Y induce a transmissible endothelial proteinopathy. *Am J Physiol Lung Cell Mol Physiol, 310*(4), L337-353. doi:10.1152/ajplung.00103.2015

Mulet, M., Lalucat, J., & Garcia-Valdes, E. (2010). DNA sequence-based analysis of the Pseudomonas species. *Environ Microbiol, 12*(6), 1513-1530. doi:10.1111/j.1462-2920.2010.02181.x

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol, 30*(5), 1196-1205. doi:10.1093/molbev/mst030

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., . . . Kosakovsky Pond, S. L. (2015). Gene-wide identification of episodic selection. *Mol Biol Evol, 32*(5), 1365-1371. doi:10.1093/molbev/msv035

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet, 8*(7), e1002764. doi:10.1371/journal.pgen.1002764

Murteira, S., Ghezaiel, Z., Karray, S., & Lamure, M. (2013). Drug reformulations and repositioning in pharmaceutical industry and its impact on market access: reassessment of nomenclature. *J Mark Access Health Policy, 1*. doi:10.3402/jmahp.v1i0.21131

NCBI. (2017). RefSeq non-redundant proteins. Retrieved from https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/

Newman, J. W., Floyd, R. V., & Fothergill, J. L. (2017). The contribution of Pseudomonas aeruginosa virulence factors and host factors in the establishment of urinary tract infections. *FEMS Microbiol Lett, 364*(15). doi:10.1093/femsle/fnx124

Nguyen, L., Garcia, J., Gruenberg, K., & MacDougall, C. (2018). Multidrug-Resistant Pseudomonas Infections: Hard to Treat, But Hope on the Horizon? *Curr Infect Dis Rep, 20*(8), 23. doi:10.1007/s11908-018-0629-6

Notti, R. Q., & Stebbins, C. E. (2016). The Structure and Function of Type III Secretion Systems. *Microbiol Spectr, 4*(1). doi:10.1128/microbiolspec.VMBF-0004-2015

Nouioui, I., Carro, L., Garcia-Lopez, M., Meier-Kolthoff, J. P., Woyke, T., Kyrpides, N. C., . . . Goker, M. (2018). Genome-Based Taxonomic Classification of the Phylum Actinobacteria. *Front Microbiol, 9*, 2007. doi:10.3389/fmicb.2018.02007

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res, 27*(5), 824-834. doi:10.1101/gr.213959.116

Ogunsola, F. T., & Mehtar, S. (2020). Challenges regarding the control of environmental sources of contamination in healthcare settings in low-and middle-income countries - a narrative review. *Antimicrob Resist Infect Control, 9*(1), 81. doi:10.1186/s13756-020-00747-0

Orellana, L. H., Chee-Sanford, J. C., Sanford, R. A., Loffler, F. E., & Konstantinidis, K. T. (2018). Year-Round Shotgun Metagenomes Reveal Stable Microbial Communities in Agricultural Soils and Novel Ammonia Oxidizers Responding to Fertilization. *Appl Environ Microbiol, 84*(2). doi:10.1128/AEM.01646-17

Ozer, E. A., Nnah, E., Didelot, X., Whitaker, R. J., & Hauser, A. R. (2019). The Population Structure of Pseudomonas aeruginosa Is Characterized by Genetic Isolation of exoU+ and exoS+ Lineages. *Genome Biol Evol, 11*(1), 1780-1796. doi:10.1093/gbe/evz119

Paharik, A. E., Schreiber, H. L. t., Spaulding, C. N., Dodson, K. W., & Hultgren, S. J. (2017). Narrowing the spectrum: the new frontier of precision antimicrobials. *Genome Med, 9*(1), 110. doi:10.1186/s13073-017-0504-3

Panayidou, S., Georgiades, K., Christofi, T., Tamana, S., Promponas, V. J., & Apidianakis, Y. (2020). Pseudomonas aeruginosa core metabolism exerts a widespread growth-independent control on virulence. *Sci Rep, 10*(1), 9505. doi:10.1038/s41598-020-66194-4

Pandey, N., & Cascella, M. (2022). Beta Lactam Antibiotics. In *StatPearls*. Treasure Island (FL).

Peabody, M. A. (2017). *Applying metagenomics analysis towards a better understanding of freshwater microbial communities.* Simon Fraser University,

Peabody, M. A., Laird, M. R., Vlasschaert, C., Lo, R., & Brinkman, F. S. (2016). PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res, 44*(D1), D663-668. doi:10.1093/nar/gkv1271

Peabody, M. A., Lau, W. Y. V., Hoad, G. R., Jia, B., Maguire, F., Gray, K. L., . . . Brinkman, F. S. L. (2020). PSORTm: a bacterial and archaeal protein subcellular localization prediction tool for metagenomics data. *Bioinformatics, 36*(10), 3043-3048. doi:10.1093/bioinformatics/btaa136

Peng, C., & Gao, F. (2014). Protein localization analysis of essential genes in prokaryotes. *Sci Rep, 4*, 6001. doi:10.1038/srep06001

Peterson, J. W. (1996). Bacterial Pathogenesis. In *Medical Microbiology* (4 ed.): University of Texas Medical Branch at Galveston.

Petitjean, M., Martak, D., Silvant, A., Bertrand, X., Valot, B., & Hocquet, D. (2017). Genomic characterization of a local epidemic Pseudomonas aeruginosa reveals specific features of the widespread clone ST395. *Microb Genom, 3*(10), e000129. doi:10.1099/mgen.0.000129

Pilmis, B., Le Monnier, A., & Zahar, J. R. (2020). Gut Microbiota, Antibiotic Therapy and Antimicrobial Resistance: A Narrative Review. *Microorganisms, 8*(2). doi:10.3390/microorganisms8020269

Portaliou, A. G., Tsolis, K. C., Loos, M. S., Zorzini, V., & Economou, A. (2016). Type III Secretion: Building and Operating a Remarkable Nanomachine. *Trends Biochem Sci, 41*(2), 175-189. doi:10.1016/j.tibs.2015.09.005

Pourcel, C., Midoux, C., Vergnaud, G., & Latino, L. (2020). The Basis for Natural Multiresistance to Phage in Pseudomonas aeruginosa. *Antibiotics (Basel), 9*(6). doi:10.3390/antibiotics9060339

Pranavathiyani, G., Prava, J., Rajeev, A. C., & Pan, A. (2020). Novel Target Exploration from Hypothetical Proteins of Klebsiella pneumoniae MGH 78578 Reveals a Protein Involved in Host-Pathogen Interaction. *Front Cell Infect Microbiol, 10*, 109. doi:10.3389/fcimb.2020.00109

Pust, M. M., Wiehlmann, L., Davenport, C., Rudolf, I., Dittrich, A. M., & Tummler, B. (2020). The human respiratory tract microbial community structures in healthy and cystic fibrosis infants. *NPJ Biofilms Microbiomes, 6*(1), 61. doi:10.1038/s41522-020-00171-7

Ramesh, A., Bailey, E. S., Ahyong, V., Langelier, C., Phelps, M., Neff, N., . . . Gray, G. C. (2021). Metagenomic characterization of swine slurry in a North American swine farm operation. *Sci Rep, 11*(1), 16994. doi:10.1038/s41598-021-95804-y

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., & Delsuc, F. (2018). MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol Biol Evol, 35*(10), 2582-2584. doi:10.1093/molbev/msy159

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res, 46*(D1), D624-D632. doi:10.1093/nar/gkx1134

Reem, A., Zhong, Z. H., Al-Shehari, W. A., Al-Shaebi, F., Amran, G. A., Moeed, Y. A. G., . . . Askary, A. E. (2021). Functional Annotation of Hypothetical Proteins related to Antibiotic Resistance in Pseudomonas Aeruginosa PA01. *Clin Lab, 67*(8). doi:10.7754/Clin.Lab.2021.210536

Rey, S., Acab, M., Gardy, J. L., Laird, M. R., deFays, K., Lambert, C., & Brinkman, F. S. (2005). PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res, 33*(Database issue), D164-168. doi:10.1093/nar/gki027

Rey, S., Gardy, J. L., & Brinkman, F. S. (2005). Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *BMC Genomics, 6*, 162. doi:10.1186/1471-2164-6-162

Richardson, A. R. (2019). Virulence and Metabolism. *Microbiol Spectr, 7*(2). doi:10.1128/microbiolspec.GPP3-0011-2018

Rutherford, S. T., & Bassler, B. L. (2012). Bacterial quorum sensing: its role in virulence and possibilities for its control. *Cold Spring Harb Perspect Med, 2*(11). doi:10.1101/cshperspect.a012427

Sanabria, A. M., Janice, J., Hjerde, E., Simonsen, G. S., & Hanssen, A. M. (2021). Shotgun-metagenomics based prediction of antibiotic resistance and virulence determinants in Staphylococcus aureus from periprosthetic tissue on blood culture bottles. *Sci Rep, 11*(1), 20848. doi:10.1038/s41598-021-00383-7

Santi, I., Manfredi, P., Maffei, E., Egli, A., & Jenal, U. (2021). Evolution of Antibiotic Tolerance Shapes Resistance Development in Chronic Pseudomonas aeruginosa Infections. *mBio, 12*(1). doi:10.1128/mBio.03482-20

Sao-Jose, C. (2018). Engineering of Phage-Derived Lytic Enzymes: Improving Their Potential as Antimicrobials. *Antibiotics (Basel), 7*(2). doi:10.3390/antibiotics7020029

Schwechheimer, C., & Kuehn, M. J. (2015). Outer-membrane vesicles from Gram-negative bacteria: biogenesis and functions. *Nat Rev Microbiol, 13*(10), 605-619. doi:10.1038/nrmicro3525

Segre, J. A. (2013). What does it take to satisfy Koch's postulates two centuries later? Microbial genomics and Propionibacteria acnes. *J Invest Dermatol, 133*(9), 2141-2142. doi:10.1038/jid.2013.260

Sen, T., & Verma, N. K. (2020). Functional Annotation and Curation of Hypothetical Proteins Present in A Newly Emerged Serotype 1c of Shigella flexneri: Emphasis on Selecting Targets for Virulence and Vaccine Design Studies. *Genes (Basel), 11*(3). doi:10.3390/genes11030340

Shahbaaz, M., Bisetty, K., Ahmad, F., & Hassan, M. I. (2016). Current Advances in the Identification and Characterization of Putative Drug and Vaccine Targets in the Bacterial Genomes. *Curr Top Med Chem, 16*(9), 1040-1069. doi:10.2174/1568026615666150825143307

Shapiro-Ilan, D. I., Fuxa, J. R., Lacey, L. A., Onstad, D. W., & Kaya, H. K. (2005). Definitions of pathogenicity and virulence in invertebrate pathology. *J Invertebr Pathol, 88*(1), 1-7. doi:10.1016/j.jip.2004.10.003

Siehnel, R., Traxler, B., An, D. D., Parsek, M. R., Schaefer, A. L., & Singh, P. K. (2010). A unique regulator controls the activation threshold of quorum-regulated genes in Pseudomonas aeruginosa. *Proc Natl Acad Sci U S A, 107*(17), 7916-7921. doi:10.1073/pnas.0908511107

Sievers, F., & Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol, 1079*, 105-116. doi:10.1007/978-1-62703-646-7_6

Simon, A. K., Hollander, G. A., & McMichael, A. (2015). Evolution of the immune system in humans from infancy to old age. *Proc Biol Sci, 282*(1821), 20143085. doi:10.1098/rspb.2014.3085

Simonson, A. W., Mongia, A. S., Aronson, M. R., Alumasa, J. N., Chan, D. C., Lawanprasert, A., . . . Medina, S. H. (2021). Pathogen-specific antimicrobials engineered de novo through membrane-protein biomimicry. *Nat Biomed Eng*. doi:10.1038/s41551-020-00665-x

Sivashankari, S., & Shanmughavel, P. (2006). Functional annotation of hypothetical proteins - A review. *Bioinformation, 1*(8), 335-338. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17597916

Skoura, N., Wang-Jairaj, J., Della Pasqua, O., Chandrasekaran, V., Billiard, J., Yeakey, A., . . . Tan, L. K. (2020). Effect of raxibacumab on immunogenicity of Anthrax Vaccine Adsorbed: a phase 4, open-label, parallel-group, randomised non-inferiority study. *Lancet Infect Dis, 20*(8), 983-991. doi:10.1016/S1473-3099(20)30069-4

Smith, E. E., Buckley, D. G., Wu, Z., Saenphimmachak, C., Hoffman, L. R., D'Argenio, D. A., . . . Olson, M. V. (2006). Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients. *Proc Natl Acad Sci U S A, 103*(22), 8487-8492. doi:10.1073/pnas.0602138103

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol, 32*(5), 1342-1353. doi:10.1093/molbev/msv022

Spaulding, C. N., Klein, R. D., Schreiber, H. L. t., Janetka, J. W., & Hultgren, S. J. (2018). Precision antimicrobial therapeutics: the path of least resistance? *NPJ Biofilms Microbiomes, 4*, 4. doi:10.1038/s41522-018-0048-3

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., . . . Olson, M. V. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature, 406*(6799), 959-964. doi:10.1038/35023079

Subedi, D., Vijay, A. K., Kohli, G. S., Rice, S. A., & Willcox, M. (2018). Comparative genomics of clinical strains of Pseudomonas aeruginosa strains isolated from different geographic sites. *Sci Rep, 8*(1), 15668. doi:10.1038/s41598-018-34020-7

Suleyman, G., Alangaden, G., & Bardossy, A. C. (2018). The Role of Environmental Contamination in the Transmission of Nosocomial Pathogens and Healthcare-Associated Infections. *Curr Infect Dis Rep, 20*(6), 12. doi:10.1007/s11908-018-0620-2

Suresh, S., Alva, P. P., & Premanath, R. (2021). Modulation of quorum sensing-associated virulence in bacteria: carbohydrate as a key factor. *Arch Microbiol, 203*(5), 1881-1890. doi:10.1007/s00203-021-02235-4

Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., . . . Group, W. H. O. P. P. L. W. (2018). Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis, 18*(3), 318-327. doi:10.1016/S1473-3099(17)30753-3

Tan, M. W., Mahajan-Miklos, S., & Ausubel, F. M. (1999). Killing of Caenorhabditis elegans by Pseudomonas aeruginosa used to model mammalian bacterial pathogenesis. *Proc Natl Acad Sci U S A, 96*(2), 715-720. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9892699

Tan, M. W., Rahme, L. G., Sternberg, J. A., Tompkins, R. G., & Ausubel, F. M. (1999). Pseudomonas aeruginosa killing of Caenorhabditis elegans used to identify P. aeruginosa virulence factors. *Proc Natl Acad Sci U S A, 96*(5), 2408-2413. doi:10.1073/pnas.96.5.2408

Theuretzbacher, U., Outterson, K., Engel, A., & Karlen, A. (2020). The global preclinical antibacterial pipeline. *Nat Rev Microbiol, 18*(5), 275-285. doi:10.1038/s41579-019-0288-0

Tortora, G. J., Funke, B. R., & Case, C. L. . (2018). *Microbiology: An introduction* (13 ed.): Pearson.

Totsika, M. (2016). Benefits and Challenges of Antivirulence Antimicrobials at the Dawn of the Post-Antibiotic Era. *Drug Delivery Letters, 6*(1). doi:10.2174/2210303106666160506120057

Trivedi, U. H., Cezard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., & Gharbi, K. (2014). Quality control of next-generation sequencing data without a reference. *Front Genet, 5*, 111. doi:10.3389/fgene.2014.00111

Tummler, B. (2019). Emerging therapies against infections with Pseudomonas aeruginosa. *F1000Res, 8*. doi:10.12688/f1000research.19509.1

Tummler, B., & Cornelis, P. (2005). Pyoverdine receptor: a case of positive Darwinian selection in Pseudomonas aeruginosa. *J Bacteriol, 187*(10), 3289-3292. doi:10.1128/JB.187.10.3289-3292.2005

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature, 449*(7164), 804-810. doi:10.1038/nature06244

Turner, K. H., Everett, J., Trivedi, U., Rumbaugh, K. P., & Whiteley, M. (2014). Requirements for Pseudomonas aeruginosa acute burn and chronic surgical wound infection. *PLoS Genet, 10*(7), e1004518. doi:10.1371/journal.pgen.1004518

Valiquette, L., & Laupland, K. B. (2015). Digging for new solutions. *Can J Infect Dis Med Microbiol, 26*(6), 289-290. doi:10.1155/2015/971858

van der Zee, A., Kraak, W. B., Burggraaf, A., Goessens, W. H. F., Pirovano, W., Ossewaarde, J. M., & Tommassen, J. (2018). Spread of Carbapenem Resistance by Transposition and Conjugation Among Pseudomonas aeruginosa. *Front Microbiol, 9*, 2057. doi:10.3389/fmicb.2018.02057

Van Rossum, T., Peabody, M. A., Uyaguari-Diaz, M. I., Cronin, K. I., Chan, M., Slobodan, J. R., . . . Brinkman, F. S. (2015). Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. *Front Microbiol, 6*, 1405. doi:10.3389/fmicb.2015.01405

Van Rossum, T., Uyaguari-Diaz, M. I., Vlok, M., Peabody, M. A., Tian, A., Cronin, K. I., . . . Brinkman, F. S. L. (2018). Spatiotemporal dynamics of river viruses, bacteria and microeukaryotes. *bioRxiv*.

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *P T, 40*(4), 277-283. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25859123

Vidaillac, C., Benichou, L., & Duval, R. E. (2012). In vitro synergy of colistin combinations against colistin-resistant Acinetobacter baumannii, Pseudomonas aeruginosa, and Klebsiella pneumoniae isolates. *Antimicrob Agents Chemother, 56*(9), 4856-4861. doi:10.1128/AAC.05996-11

Vilaplana, L., & Marco, M. P. (2020). Phenazines as potential biomarkers of Pseudomonas aeruginosa infections: synthesis regulation, pathogenesis and analytical methods for their detection. *Anal Bioanal Chem, 412*(24), 5897-5912. doi:10.1007/s00216-020-02696-4

Wagner, S., Grin, I., Malmsheimer, S., Singh, N., Torres-Vargas, C. E., & Westerhausen, S. (2018). Bacterial type III secretion systems: a complex device for the delivery of bacterial effector proteins into eukaryotic host cells. *FEMS Microbiol Lett, 365*(19). doi:10.1093/femsle/fny201

Wang, G., Gao, Y., Wang, H., Niu, X., & Wang, J. (2018). Baicalin Weakens Staphylococcus aureus Pathogenicity by Targeting Sortase B. *Front Cell Infect Microbiol, 8*, 418. doi:10.3389/fcimb.2018.00418

Wareham, D. W., Papakonstantinopoulou, A., & Curtis, M. A. (2005). The Pseudomonas aeruginosa PA14 type III secretion system is expressed but not essential to virulence in the Caenorhabditis elegans-P. aeruginosa pathogenicity model. *FEMS Microbiol Lett, 242*(2), 209-216. doi:10.1016/j.femsle.2004.11.018

Wassenaar, T. M., & Gaastra, W. (2001). Bacterial virulence: can we draw the line? *FEMS Microbiol Lett, 201*(1), 1-7. doi:10.1111/j.1574-6968.2001.tb10724.x

Wickham, M. E., Brown, N. F., Boyle, E. C., Coombes, B. K., & Finlay, B. B. (2007). Virulence is positively selected by transmission success between mammalian hosts. *Curr Biol, 17*(9), 783-788. doi:10.1016/j.cub.2007.03.067

Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., & Brinkman, F. S. (2016). Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids Res, 44*(D1), D646-653. doi:10.1093/nar/gkv1227

Winstanley, C., O'Brien, S., & Brockhurst, M. A. (2016). Pseudomonas aeruginosa Evolutionary Adaptation and Diversification in Cystic Fibrosis Chronic Lung Infections. *Trends Microbiol, 24*(5), 327-337. doi:10.1016/j.tim.2016.01.008

Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D., & Muse, S. V. (2020). Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril. *Mol Biol Evol, 37*(8), 2430-2439. doi:10.1093/molbev/msaa037

Wood, S. J., Goldufsky, J. W., Bello, D., Masood, S., & Shafikhani, S. H. (2015). Pseudomonas aeruginosa ExoT Induces Mitochondrial Apoptosis in Target Host Cells in a Manner That Depends on Its GTPase-activating Protein (GAP) Domain Activity. *J Biol Chem, 290*(48), 29063-29073. doi:10.1074/jbc.M115.689950

World Health Organization. (2017a). Guidelines for the Prevention and Control of Carbapenem-Resistant Enterobacteriaceae, Acinetobacter baumannii and Pseudomonas aeruginosa in Health Care Facilities. In. Geneva.

World Health Organization. (2017b). *WHO publishes list of bacteria for which new antibiotics are urgently needed*. Retrieved from http://www.who.int/mediacentre/news/releases/2017/bacteria-antibiotics-needed/en/

World Health Organization. (2019a). *2019 Antibacterial agents in clinical development: an analysis of the antibacterial clinical development pipeline*. Retrieved from https://apps.who.int/iris/handle/10665/330420

World Health Organization. (2019b). New report calls for urgent action to avert antimicrobial resistance crisis [Press release]

World Health Organization. (2021). Global shortage of innovative antibiotics fuels emergence and spread of drug-resistance.

Yang, H., Shan, Z., Kim, J., Wu, W., Lian, W., Zeng, L., . . . Jin, S. (2007). Regulatory role of PopN and its interacting partners in type III secretion of Pseudomonas aeruginosa. *J Bacteriol, 189*(7), 2599-2609. doi:10.1128/JB.01680-06

Yang, M., Derbyshire, M. K., Yamashita, R. A., & Marchler-Bauer, A. (2020). NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. *Curr Protoc Bioinformatics, 69*(1), e90. doi:10.1002/cpbi.90

Yang, Z., Zeng, X., & Tsui, S. K. (2019). Investigating function roles of hypothetical proteins encoded by the Mycobacterium tuberculosis H37Rv genome. *BMC Genomics, 20*(1), 394. doi:10.1186/s12864-019-5746-6

Yao, Q., Zhang, L., Wan, X., Chen, J., Hu, L., Ding, X., . . . Shao, F. (2014). Structure and specificity of the bacterial cysteine methyltransferase effector NleE suggests a novel substrate in human DNA repair pathway. *PLoS Pathog, 10*(11), e1004522. doi:10.1371/journal.ppat.1004522

Yelton, A. P., Thomas, B. C., Simmons, S. L., Wilmes, P., Zemla, A., Thelen, M. P., . . . Banfield, J. F. (2011). A semi-quantitative, synteny-based method to improve functional predictions for hypothetical and poorly annotated bacterial and archaeal genes. *PLoS Comput Biol, 7*(10), e1002230. doi:10.1371/journal.pcbi.1002230

Yu, N. Y., Laird, M. R., Spencer, C., & Brinkman, F. S. (2011). PSORTdb--an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res, 39*(Database issue), D241-244. doi:10.1093/nar/gkq1093

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., . . . Brinkman, F. S. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics, 26*(13), 1608-1615. doi:10.1093/bioinformatics/btq249

Zambelloni, R., Marquez, R., & Roe, A. J. (2015). Development of antivirulence compounds: a biochemical review. *Chem Biol Drug Des, 85*(1), 43-55. doi:10.1111/cbdd.12430

Zavascki, A. P., Goldani, L. Z., Li, J., & Nation, R. L. (2007). Polymyxin B for the treatment of multidrug-resistant pathogens: a critical review. *J Antimicrob Chemother, 60*(6), 1206-1215. doi:10.1093/jac/dkm357

Zheng, Y., Qiu, X., Wang, T., & Zhang, J. (2021). The Diagnostic Value of Metagenomic Next-Generation Sequencing in Lower Respiratory Tract Infection. *Front Cell Infect Microbiol, 11*, 694756. doi:10.3389/fcimb.2021.694756

Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., & Slezak, T. (2007). MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res, 35*(Database issue), D391-394. doi:10.1093/nar/gkl791

Zhou, M., Wu, Y., Kudinha, T., Jia, P., Wang, L., Xu, Y., & Yang, Q. (2021). Comprehensive Pathogen Identification, Antibiotic Resistance, and Virulence Genes Prediction Directly From Simulated Blood Samples and Positive Blood Cultures by Nanopore Metagenomic Sequencing. *Front Genet, 12*, 620009. doi:10.3389/fgene.2021.620009

Zhou, X., Shen, X. X., Hittinger, C. T., & Rokas, A. (2018). Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Mol Biol Evol, 35*(2), 486-503. doi:10.1093/molbev/msx302