

Investigation of the atypical type III secretion system in *Pseudomonas syringae* strains using long read sequencing technology

by

Janet Sui Hoa Lorv

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Biology

Waterloo, Ontario, Canada, 2024

© Janet Sui Hoa Lorv 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Boris Vinatzer Professor, School of Plant and Environmental Sciences, Virginia Polytechnic Institute and State University
Supervisor(s)	Dr. Brendan J. McConkey Professor, Department of Biology, University of Waterloo
Internal Member	Dr. Trevor Charles Professor, Department of Biology, University of Waterloo
Internal Member	Dr. Matthew Scott Professor, Department of Applied Mathematics, University of Waterloo
Internal-external Member	Dr. Dan Brown Professor, Cheriton School of Computer Science, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Ubiquitous across many environments, *Pseudomonas syringae* is a phytopathogen that is spread through the water cycle. With a broad host range, *Pseudomonas syringae* has been reported to cause outbreaks across many agricultural crops including potato, mango, and kiwifruits. Although this species is highly studied, interest in agricultural impacts has meant that sequencing efforts have focused primarily on virulent strains isolated from diseased plants.

In this work, we explored the genetic differences of a recently diverged, closely related and putatively non-virulent (i.e., not visibly detected by host plant) subphylogroup of *P. syringae* strains, subphylogroup 2c. Strains in this monophyletic group are a part of the larger virulent phylogroup 2 and more broadly grouped with other agriculturally virulent strains in phylogroups 1 and 3. Virulence of this bacterial species complex has been attributed to the genetic mobility of genes encoding type III effector proteins that are secreted by the type III secretion system. Gain or loss of these genes can cause *P. syringae* strains to switch between virulent and avirulent (i.e., detected by host plant and rendered not virulent) phenotypes. Many other niche adaption genes such as phytotoxin biosynthesis genes can also contribute to a virulence phenotype. We propose that the mechanisms contributing to the putative non-virulence phenotype of subphylogroup 2c may be elucidated via identification and examination of the divergent gene clusters within *P. syringae* strains.

To capture the full complement of genes, we sequenced and assembled complete genomes of two subphylogroup 2c strains, a biocontrol strain *P. syringae* pv. *syringae* 508, and a strain isolated from a healthy leaf, *P. syringae* TLP2. These strains were sequenced using Oxford Nanopore Technology's MinION long read sequencer to generate complete genomes without additional sequencing. Using one strain per R9.4 flowcell, we were able to generate >300x coverage for each strain with maximum read lengths exceeding 500kb and 250kb respectively allowing easy assembly of the full genome lengths. However, the ease-of-use of nanopore sequencing is hampered somewhat by the relatively high sequencing error rate of the reads, requiring extensive testing and optimization of recently developed software to maximize assembly quality. Despite optimization, many genes in the initial assembly remained fragmented. To address this issue, we developed the software tool Kastor that identifies and corrects draft assembly errors through comparisons to reference genomes and/or protein sequence information. Users can reference related genomes for more precise error identification, or use reference proteins for a more generalized identification method that doesn't rely

on data from closely related species. For both the Pss508 and PsyTLP2 sequenced genomes, fewer than 3000 errors (comprising <0.05% of the assemblies) were corrected, however assembly completeness was increased from <85% to >99.8% as benchmarked using single copy orthologs. The completed assemblies were then annotated using NCBI's PGAP.

The annotated gene sets of the two *P. syringae* strains were then compared to 428 *P. syringae* strains across the species complex. We identified several divergent gene clusters in both sequenced and subphylogroup 2c genomes, with the most prominent diverging gene cluster set encoding the atypical variant of the type III secretion system (T3SS). Using PIRATE, we identified a related gene cluster in the distantly related phylogroup 13. With the analysis of both gene sequence and gene content, we propose that the atypical variant is a result of horizontal transfer and not divergence from the canonical variant in closely related strains. The horizontal acquisition of this T3SS variant in the most recent common ancestor of subgroup 2c likely triggered the divergence of the subgroup from group 2, the loss of common type III effector genes, and a likely niche/virulence change, resulting in putatively weakly-virulent strains.

Acknowledgements

This thesis would not be made possible with the guidance and support of many individuals that have assisted me during this work. I would like to express my deepest gratitude to my supervisor, Dr. Brendan McConkey, for foremost giving me the opportunity to pursue my work. His unending patience, kindness, and insight were a great benefit during stressful times. I would like to also give special thanks to members Dr. Trevor Charles, Dr. Dan Brown and Dr. Matthew Scott for their feedback, support and understanding throughout the work. Also, I want to thank Dr. Boris Vinatzer for taking the time to read my work with a critical eye.

I am very grateful to have received funding from the Ontario government (OGS) and the government of Canada (NSERC) to pursue this project.

I am forever grateful to both past and present members of the McConkey Lab for their helpful advice and insights as well as their joyful enthusiasm during our daily routines.

Lastly, I would like to give all my love and thanks to my family and friends for their unwavering support of me throughout this work and my life.

Table of Contents

Examining Committee Membership.....	ii
Author’s Declaration	iii
Abstract.....	iv
Acknowledgements	vi
List of Figures.....	xi
List of Tables	xiv
List of Abbreviations	xvi
Chapter 1 : Background.....	1
1.1 Phytopathogens and host plant interactions	1
1.2 Variability in pathogenicity	2
1.3 <i>Pseudomonas syringae</i> , classification and phytopathology	3
1.4 Typical model of pathogenic host-bacterial interactions.....	6
1.4.1 Pre-pathogenicity: The epiphytic microbial lifecycle and plant colonization.....	6
1.4.2 Barriers to successful host plant colonization	10
1.4.3 Early stages of pathogenicity: MAMP triggered immunity	14
1.4.4 Plant immune system suppression.....	16
1.4.5 Mid-late stage of pathogenicity	16
1.5 Sequencing technology and bacterial genomics.....	26
1.5.1 History of development of sequencing and use in genomics of microorganisms	26
1.5.2 Nanopore sequencing technology: a new form of sequencing to fully resolve bacterial genomes	28
1.5.3 Data analysis of long reads in whole genome sequencing	32
1.6 Thesis Objectives: Analysis of pathogens using long reads.....	34

Chapter 2 : Assessment of ONT’s Nanopore sequencing use in assembling complete bacterial genomes.....	36
2.1 Introduction	36
2.1.1 Chapter objectives and overview.....	36
2.1.2 Chapter background.....	36
2.2 Material and Methods.....	38
2.2.1 Bacterial strains, DNA extraction and library preparation.....	38
2.2.2 Nanopore sequencing, offline basecalling, and analysis of sequencing results	39
2.2.3 Draft genome assembly	39
2.3 Results	40
2.3.1 Assessment of two R9.4 flowcell MinION sequencing runs	40
2.3.2 Bacterial genome assembly using nanopore long reads	49
2.4 Discussion.....	53
Chapter 3 : Kastor: A reference-based assessment of fragmentation errors in long-read assemblies..	58
3.1 Introduction	58
3.1.1 Chapter objective and overview	58
3.1.2 Chapter background.....	58
3.2 Material and Methods.....	61
3.2.1 Software Implementation	63
3.3 Results	67
3.3.1 Draft assembly improvements.....	67
3.3.2 Profile of Kastor detected errors.....	72
3.3.3 Limitations and workarounds.....	77
3.4 Discussion.....	80

Chapter 4 : Kastor-P: Reference protein-based gene fragmenting error assessment module	83
4.1 Introduction	83
4.1.1 Chapter objective and overview	83
4.1.2 Chapter background.....	84
4.2 Material and Methods.....	85
4.2.1 Software Implementation	87
4.3 Results	92
4.3.1 Improvements to draft assembly	92
4.3.2 Assessment of detected errors using Kastor-P	94
4.4 Discussion.....	97
4.5 Future Directions	99
Chapter 5 : Divergence of subphylogroup 2c <i>Pseudomonas syringae</i> strains linked with horizontal gene transfer of atypical type III secretion system.....	101
5.1 Introduction	101
5.1.1 Chapter objective and overview	101
5.1.2 Chapter background.....	102
5.2 Material and Methods.....	105
5.3 Results	108
5.3.1 Pangenome Overview.....	108
5.3.2 Identified divergent gene cluster of subphylogroup 2c involves the T3SS.....	111
5.3.3 Horizontal gene transfer of T3SS variants	115
5.3.4 Analysis of AT-PAI encoding the third major T3SS variant in <i>P. syringae</i>	119
5.3.5 Analysis of genetic factors that impact pathogenicity in subphylogroup 2c strains	131
5.4 Discussion.....	152

Chapter 6 : Conclusions.....	160
References	162
Appendices	192
Appendix A : Confirmation of plant interaction and subphylogroup 2c <i>Pseudomonas syringae</i>	192
Appendix B : Chapter 3 Supplementary Materials.....	195
Appendix C : Chapter 5 Supplementary Materials.....	201
C.1 Incongruencies in core gene phylogeny with past research.....	201
C.2 Imbalance in accessory genome from strain selection.....	210
C.3 Incongruence strains grouping with gene set clustering at different protein level	211

List of Figures

Figure 1.1 Schematic overview of the bacterial plant pathogen and host plant interaction	13
Figure 1.2 Schematic diagram of the <i>Pseudomonas syringae</i> type III secretion system (T3SS) injection apparatus.....	18
Figure 1.3 Overview of the indirect detection of pathogenic T3 effector proteins with its respective Resistance proteins (R-protein) under the guard and decoy model.....	24
Figure 2.1 Quantity and quality of reads sequenced with DNA prepared using the RAD002 Rapid library preparation method on a R9.4 flowcell over the course of a 48 hr sequencing run.....	43
Figure 2.2 Expected coverage of Pss508, with an estimated genome size of 6.0Mb, across time based on total sequence yield	44
Figure 2.3 Heatmap of the number of reads sequenced per multiplex channel across the entire Pss508 sequencing run.....	45
Figure 2.4 Average quality of reads vs read length.....	49
Figure 2.5 Relative 6mer composition between a draft assembly and a reference assembly.....	51
Figure 3.1 Kastor error assessment and correction workflow	62
Figure 3.2 IGV visualized examples of draft assembly regions (Pss508) containing deletion error candidates	65
Figure 3.3 Reduction of fragmentation of identified genes before and after Kastor correction.....	73
Figure 3.4 Breakdown of the genomic sequence preceding detected errors in the Pss508 assembly ..	76
Figure 3.5 Post Kastor correction Average Nucleotide Identity (ANI) percent differences	80
Figure 4.1 Flowchart of the protein-based Kastor-P workflow.....	89
Figure 4.2 A simplified schematic of an overlap with a candidate error detection.....	90
Figure 5.1 Phylogenetic tree generated from 2087 single copy gene families using FastTree with SH-like support.	111
Figure 5.2 Common variants of the observed type III secretion system gene clusters	113

Figure 5.3 Representative diagram of hierarchical clustering of <i>P. syringae</i> strains by T3SS PAI genes	114
Figure 5.4 Multi-locus sequence analysis tree of conserved T3SS genes, grouped as single gene families, shared across all <i>P. syringae</i> strains.....	121
Figure 5.5 Maximum likelihood gene tree of <i>hrpR</i> and <i>hrpS</i> using a TIM1+G4 model	122
Figure 5.6 Maximum likelihood gene tree of some of the most conserved genes <i>hrcT</i> , <i>hrcS</i> , <i>hrcR</i> encoding the inner membrane of the T3SS injectisome	126
Figure 5.7 Gene trees of <i>hrpA</i>	128
Figure 5.8 Gene trees of <i>hrpF</i>	129
Figure 5.9 Icicle plot of the hierarchical set clustering of <i>hrp</i> box regulated gene families	132
Figure 5.10 Icicle plot of the hierarchical set clustering of the <i>hrp</i> box regulated gene families and known virulence gene	135
Figure 5.11 Hierarchical set clustering and binary heatmap of the presence and absence of gene families that are <i>hrp</i> box regulated genes, and/or blast-detected virulence genes reveal general virulence strategies of <i>P. syringae</i> strains	138
Figure A.1 Representative images of the HR assay injection sites, at 1×10^8 mg/ml, after 24 hours	193
Figure B.1 Reduction of fragmentation in identified genes before and after Kastor correction (PsyTLP2).....	197
Figure B.2 Breakdown of the genomic sequence preceding detected errors in the PsyTLP2 assembly	198
Figure B.3 Post Kastor correction Average Nucleotide Identity percent differences (PsyTLP2).....	199
Figure C.1a Enlarged view of Phylogroup 2, excluding 2c, in Figure 5.1 phylogenetic tree	203
Figure C.1b Enlarged view of phylogroups 2c, 3, and 6 plus parts of phylogroups 1b and 2, from figure 5.1 phylogenetic tree	205
Figure C.1c Enlarged view of the figure 5.1 phylogenetic tree.....	207

Figure C.1d Enlarged view of the figure 5.1 phylogenetic tree	208
Figure C.2 An upset plot representing shared gene families between phylogroups (i.e., intersection in a Venn diagram)	211
Figure C.3 Tanglegram comparison of the hierarchical clustering of shared gene families at 50% identity (left) compared to 90% identity (right)	213

List of Tables

Table 2.1 R9.4 flowcell sequencing statistics using standalone Albacore basecalling	42
Table 2.2 Assessment of draft assembly quality via detection of known 452 single universal orthologs found in Gammaproteobacteria using BUSCO.	52
Table 2.3 Draft assembly statistics generated using an optimized Minimap-Miniasm-Racon pipeline, and post-assembly correction with Nanopolish and Medaka.	52
Table 3.1 BUSCO assessment of assembly completeness	69
Table 3.2 Fragmentation status of each <i>Pseudomonas syringae</i> assembly post Kastor correction with the equidistant reference genome set.	70
Table 3.3 Number and types of errors detected by Kastor	74
Table 3.4 Influence of genome set on Kastor corrected assemblies using BUSCO to assess completeness of Pss508	78
Table 4.1 A comparison of universal single orthologs detected by BUSCO post Kastor corrected assemblies for using both approaches	93
Table 4.2 Number of annotated genes predicted for pre and post corrected assemblies using NCBI's PGAP	94
Table 4.3 A comparison of the number and location of errors detected using both Kastor error assessment modules.	96
Table 5.1 General virulence genes detected using blastp in two group 2c strains (Pss508 and PsyTLP2)	142
Table 5.2 A gene family cluster (cluster 784-787) in group 2c strains that contain genes that have diverged from both closely related phylogroups 2 and 3	147
Table 5.3 Accessory genome clusters that appear in select group 2c strains that contain genes that have diverged from homologs found in other Psy strains: Pss508 (top) and PsyTLP2 (bottom)	150
Table A.1 Variation in plant response in HR assay between three closely related <i>P. syringae</i> strains after 24 hour post-inoculation	194

Table B.1 BUSCO assessment of the completeness of long-read assembly PsyTLP2 at different pipeline stages	195
Table B.2 Number of corrections using different combinations of genome sets	196
Table B.3 Reference genome nucleotide distance chart.....	200

List of Abbreviations

Avr	avirulence
AT-PAI	atypical tripartite T3SS pathogenicity island
CCS	circular consensus sequencing
CEL	conserved T3 effector loci
CLR	consensus long read sequencing
DAMP	Damage associated molecular pattern
EEL	exchangeable effector loci
ETI	Effector triggered immunity
ETS	Effector triggered suppression
Hop	Hrp outer protein
Hrc	hypersensitive response and conserved
Hrp	hypersensitive response and pathogenicity
JA	Jasmonic acid
MAMP	Microbe associated molecular pattern
mmr	Minimap-Miniasm-Racon assembly pipeline
MTI	MAMP triggered immunity
ONT	Oxford Nanopore Technologies
Pco	<i>Pseudomonas congelans</i>
PG-2c	Subphylogroup 2c <i>Pseudomonas syringae</i> strains
PRR	Pattern recognition receptor
Psy	<i>Pseudomonas syringae</i>
Pss508	<i>Pseudomonas syringae</i> pv. <i>syringae</i> 508
PsyTLP2	<i>Pseudomonas syringae</i> TLP2
ROS	Reactive oxygen species
S-PAI	single partite T3SS pathogenicity island
SA	salicylic acid
SMRT	single molecule read-time
T-PAI	tripartite T3SS pathogenicity island
T3SS	type III secretion system
T3Es	type III effector proteins

Chapter 1:

Background

1.1 Phytopathogens and host plant interactions

The effects of climate change are becoming an increasing concern for food sustainability. Substantial reduction of arable land and overuse of monoculture crops have increased uncertainty of sufficient yearly agricultural yield. In response, focus has shifted to developing agricultural practices including maximizing crop yield via transgenic crops, and application of beneficial plant growth promoting microorganisms [1]. Despite these advances, transgenic plants typically are engineered to target specific issues. Under conditions favouring pathogens, a singular phytopathogen species can cause annual losses ranging from 5-50% in the Ontario agricultural industry [2]. These devastating outbreaks can occur from a variety of microorganisms including fungi [3], bacteria [4], oomycetes [5], and viruses [6].

To mitigate these losses, one approach to fostering healthy plants is augmenting the composition of the plant's microbial community, such as increasing plant stress resilience via plant growth promoting bacteria [1], and pathogen prevention through biocontrol strains of microorganisms [7]. In the latter approach, biocontrol strains can directly target pathogenic species via toxic secondary metabolite biosynthesis [8], direct competition (e.g., via T6SS cytotoxicity) [9], or microbial induced plant defence (i.e., endophytic mediated resistance) [10]. Consequently, the susceptibility of plants to pathogens is dependent on complex interactions between the host plant and its associated microbiota, and between other microbes [11]. The host and community, known as the holobiont, collectively evolve together with reciprocal impacts of host and community species. Imbalances in the microbial community can increase the susceptibility [12] or resistance to plant pathogens [7,10,13]. In turn, the host plant defence response shapes its own microbiome, as previously demonstrated through mutagenesis-based reduction of early broad immunity genes revealed a shift in microbial composition towards Proteobacteria in the microbiome [14]. For some phytopathogens, including *Pseudomonas syringae*, plant infection occurs at different stages of the host's lifecycle [15]. Furthermore, less virulent to potentially non-virulent variants of the same phytopathogen species have been found, indicating conditional virulence [16,17]. Despite being highly similar genetically, these variant strains can have different phenotypes and elicit significant variability in plant responses [18]. Understanding the differentiating genetic factors between these related strains with differing phenotypes can provide insight into the survivability of these low virulence strains within host plants.

1.2 Variability in pathogenicity

Taxonomic classification of this species has grown in complexity with the current trend coalescing traditionally individual *Pseudomonas* species into a larger species complex (See section 1.3). Putatively non-pathogenic strains of phytopathogenic species have been isolated from a number of sources, such as healthy plants and other environmental sources [7,16,19]. Until an host plant can be identified, these strains lack the known pathogenicity of virulent strains isolated from diseased crops [18].

Among these strains include more distantly related plant-associated strains, such as *P. syringae* G12-2, which was previously classified as *P. putida* [20]. Analyses of host interaction were focused on the genetic factors involved in an epiphytic lifestyle (see section 1.4.1), such as plant promoting secondary metabolite biosynthesis and expression of ice controlling proteins, while pathogenic host colonization and infection by these strains have not been observed. However, in the past 15 years, non-pathogenic strains closely related to known virulent *P. syringae* strains have been found to colonize host plants, a lifestyle typically involving host manipulation leading to pathology (see section 1.4.5) [21]. Non-pathogenicity in many of these *P. syringae* strains was initially attributed to missing a type III secretion system (T3SS), a key pathogenic factor involved in direct host interaction and immune system suppression via injection of T3 effectors into the host. In past studies, knockout mutants of conserved core T3SS genes have rendered pathogenic variants non-virulent [21–23]. In one study of isolated strains lacking pathogenicity, the T3SS expression was disrupted by a mobile genetic element, and the researchers speculate removal of the genetic element may result in phenotype reversion in T3SS function [17].

Putatively non-pathogenic strains that have an active T3SS may also be potentially weakly virulent. Comparative studies of these non-pathogenic strains and their virulent counterparts have highlighted that non-pathogenic variants often have a reduced effectome [24,25]. Since large and diverse effectome allows for flexibility in MTI suppression, a reduced effectome can suggest a highly specialized host target [26,27]. However, over reliance of short read draft assemblies in bacterial sequencing has made confirmation of the lack of effectors difficult. The exchangeable effector loci containing the diversity of genes encoding T3 effector proteins are frequently fragmented due to high discontinuity in short read assemblies [28]. Full resolution of effector loci at the genomic level can be

instrumental to determining the link between expressible effectome and non-pathogenicity (see section 1.5).

Isolated single strains are not always reflective of the plant's microbial community and population, and mixed populations of microbe can complement poor virulence [29]. Using confocal microscopy, non-pathogenic strains (lacking T3SS) were able to proliferate within a mixed population that includes virulent strains. Here, secreted effectors complemented the lack of plant defence suppression. A similar effect was seen in mixtures of weakly and strongly virulent populations [30]. When different mixtures of variably virulent populations were observed, weakly virulent strain populations were more successful and abundant during early infection stages whereas strongly virulent strains dominated during late infection. Thus, community effects can propagate less virulent strains. These heteromixtures likely occur frequently *in planta*, as was found with isolates from kiwifruits that revealed both virulent and non-virulent strains (i.e. strains with disrupted T3SS gene clusters) [17]. In the same manner, the inverse can occur, where inoculation of mixed populations with avirulent strains conferred a protective effect, limiting the host susceptibility and pathogen spread with increased activation of immune defence response [29].

Therefore, in order to elucidate the genetic factors involved in the variability in pathogenicity between *P. syringae* strains, an in-depth background of the host-bacterial lifestyle and the genes involved is needed to investigate their genomes.

1.3 *Pseudomonas syringae*, classification and phytopathology

This work focuses on the ubiquitous bacterial plant pathogen *Pseudomonas syringae* [2,4]. *P. syringae* is commonly prevalent in more tropical environments, where warm temperate and rainfall facilitate the dispersal and spread leading to infectious outbreaks contributing to crop losses [4,31,32]. Canada's cold weather has largely prevented bacterial phytopathogen induced crop loss, however with milder winters pathogens, such as *P. syringae*, are becoming a concern in temperate climates to northern climates [33]. The bacterium *Pseudomonas syringae* was initially isolated from *Syringa vulgaris* (lilac trees)[34] as a singular species. Since then, researchers have expanded the classification of this phytopathogen into a species complex to include strains isolated from a large diversity of host plants in North America [35], Europe [36,37], Asia [22], and Oceania [38]. This bacterial species has a very broad host range and is capable of surviving on a variety of agricultural

crops (e.g., kiwifruit [22], mangos [39], bean [35], cherry [37], etc.), flowers [40,41], and wild plants [42]. Furthermore, non-plant environments such as snowbanks have also been discovered to harbour the species [38]. The extensive host range is not a distinct characteristic of representative strains but instead reflects the diversity of virulence of individual strains within the species complex, with each individual strain adapting to a specific host using a curated set of virulence genes [43,44].

Initially, the diversity of this species was underestimated and *Pseudomonas* species classifications were largely assigned via host-strain interactions. However, with advancements in DNA sequencing and the resultant evolution of taxonomical classification, these distinct species have been collapsed into a singular *P. syringae* complex [44]. Traditionally, *P. syringae* strains were identified using a phenotype-based biochemical classification, for example through the LOPAT test for pseudomonads [45]. After isolation of numerous virulent strains from agriculturally related samples, the *Pseudomonas syringae* complex was split into 64 pathovars based on host range and pathogenic characteristics [46,47]. Subsequently, with the development and popularization of DNA-DNA hybridization technology, high genetic similarity between phenotypically distinct pseudomonads was reported leading to partitioning of the species into nine “genomespecies” groups based on capability of DNA hybridization of 16S rRNA and other genes between strains [48]. As genetic-based techniques matured, the species is now delineated into 13 phylogroups based on 16S rRNAs, multilocus sequence analyses, and/or core genome analyses [44,47,49]. However, there is still incongruence in strain placement, dependent on the selected genes for multi-locus sequence alignments (MLSAs) and genes within the core genome can vary depending on the strains being examined [50].

As the classification methods were refined, many previously separate *Pseudomonas* spp. have been combined into a singular species. Several well-studied and previously classified as individual pathogenic *Pseudomonas* species, such as *P. avellanae* and *P. savastanoi*, have been placed within phylogroups 1 and 3 within *P. syringae* [44,51]. Many of these virulent species have very similar genomes, commonly grouping within the primary phylogroups 1, 2, and 3 [43]. Other more ambiguously virulent pseudomonads are included within the species complex as well: *Pseudomonas congelans*, isolated from the phyllosphere (above-ground portion of the plant) of grasses, was previously classified as part of the *Pseudomonas fluorescens* group, but are now proposed to be more closely related to *P. syringae* or as a sub-species of *P. syringae* instead [19,49]. The primary phylogroups (late branching phylogroups in the species complex) comprise the largest number of

sequenced strains, likely due to high relevance to crop disease. Collectively, *P. syringae* strains from these phylogroups have a diverse host plant range known to cause outbreaks in lilac trees, mangoes, kiwifruits, and many other host plants. Comparatively, the secondary phylogroups (early branching phylogroups in the species complex) have fewer representative genomes, including *P. syringae* strains isolated from non-agricultural sources including clouds, rivers, snow banks, and wild plants [16,52,53]. However, these groups also include known virulent *P. syringae* strains. Virulent *P. viridiflava* have a broad range of hosts including *A. thaliana*, *Prunus cerasus* (cherry), *Primula officinalis* (Cowslip), *Rhododendron sp.* and epilithon [37,38,40,41,54,55], and are typically classified as belonging as part of phylogroup 7. Similarly, *P. cichorii* has a very broad range of hosts and has been placed within *P. syringae* phylogroup 11 [44,56].

Although current classification is helpful in defining the species and phylogroup boundaries genetically, reconciliation of species genotype with phenotype still proves to be difficult, especially for defining the ecological boundaries for these strains [43]. For these bacteria, host virulence is closely related to host-specific effector expression [24]. Strains with a similar host range have been found in different phylogroups [57], along with putatively non-pathogenic variants as well. Since *P. syringae* strains with undetermined plant hosts have also been found in water samples, ecological groupings are further compounded by the potential dissemination of the *P. syringae* strains through the water cycle [58]. This method of bacterial distribution likely has contributed to its high prevalence throughout the world and its broad host range due to high level of contact with diverse sets of plant species. In fact, Walker and Parker (1964) and Taylor et al. (1979) have demonstrated favourable rainfall and wind were major contributing factors to *P. syringae* dispersal and potential outbreaks of halo blight in potatoes in Wisconsin, USA [31,32]. Consequently, *P. syringae* are typically found in the plant phyllosphere and have primarily been studied in terms of its impact within host plant leaves [15]. Although *P. syringae* is typically reported as a leaf colonizing bacteria, other forms of pathogenicity have also been reported, such as more external saprophytic lifestyles. At least one strain has been associated with fruit rot, a common factor contributing to economic losses during fruit storage post-harvest [59].

1.4 Typical model of pathogenic host-bacterial interactions

1.4.1 Pre-pathogenicity: The epiphytic microbial lifecycle and plant colonization

The phyllosphere is one of the most abundant habitats in the world with continuous exposure to a wide variety of environmental and organic-based stressors, at both macro and micro level. In the field, plants commonly come into contact with a variety of microorganisms, either through passive (i.e., rain, irrigation, soil, etc.) or active (i.e., agricultural treatment, herbivore contact, etc.) means [58,60,61]. Consequently, the phyllosphere can harbour a diverse community of microorganisms including fungi, bacteria, protists, and viruses [62]. However, continuous exposure to the environmental conditions creates a harsh environment with limited sources of nutrients, high levels of abiotic stresses, including UV radiation, oxidative stress, and desiccation [63–65]. For some phytopathogens such as *P. syringae*, the microbe spends a large portion of its lifecycle in a non-pathogenic epiphytic state on the surface of plant leaves [15,66]. To prevent cell desiccation while on the leaf surface, *P. syringae* pv. *syringae* B728a synthesizes a biosurfactant, syringafactin, that has been seen to improve diffusion of water across the waxy leaf surface of *Phaseolus vulgaris* (bean), helping maintain proper moisture levels and nutrient acquisition for cell survival [67]. Other abiotic stresses can be mitigated through forming bacterial aggregates via biofilm production, creating stable communities on leaf surfaces [68,69]. Biofilm overproducing mutants of *P. syringae* pv. *syringae* UMAF0158 were found to have lower rates of virulence on leaves, and was suggested that the increased polysaccharide production is tied to adhesion to mango tree leaves, reducing leaf colonization [70]. Conversely, with poor nutrient availability, expression of motility genes and resultant mobility of epiphytic microbes is also a key factor in successful nutrient acquisition and leaf colonization.

In contrast, the host plant's internal environment is a more nutrient and moisture abundant habitat without many of the external abiotic stresses, providing a favourable environment for microbes. However, successful colonization of a host plant relies on three major factors: environmental conditions, host-microbe interactions, and microbe-microbe interactions. Although nutrient availability is a major determinant in host colonization, the specific conditions to trigger a phenotypic switch from an epiphytic to parasitic lifestyle in *P. syringae* has not yet been determined [15]. Changes to ambient conditions, such as increase in temperature [33], decrease in light duration [71], and decrease in nutrient availability [63] can favour one phenotype over the other.

A cuticle layer primarily composed of cutin and wax on the leaf surface separates the two habitats, forming a protective layer that prevents water, solutes and organic material (such as microbes) from penetrating the epidermal layer of the leaf [72]. This cuticle layer both prevents leaf desiccation through water seepage, while also regulating its microbial community through prevention of colonization by external microbes. With the correct combination of abiotic stresses triggering a phenotypic transition from an epiphytic to endophytic lifestyle, microbial chemotaxis and motility genes will be upregulated in the colonizing microbe while downregulating sessile-related genes, such as biofilm production [73,74]. Chemosensing genes have been implicated in locating alternative openings for colonization. For leaf colonizers, common routes include natural openings such as stomatal openings needed for gas exchange, hydathodes openings, and wounds [60,61,75]. Wounds can be induced either through external means (e.g. insect bites), or directly via expression of lytic enzymes or physically destructive proteins such as ice nucleation proteins. Although stomatal openings can be restricted via closure of two guard cells surrounding the entry point, the primary function of closure is to regulate gas exchange within the leaf rather than active restriction of access [76]. For root colonizers in the rhizosphere (i.e. all structures below ground), alternative openings include epidermal cracks and root hairs. To reach these entry points, motility gene expression is paramount for colonization, downregulation of these genes can reduce and thereby hinder the virulence of a phytopathogen [77]. Furthermore, the inability to downregulate sessile-related genes such as polysaccharide production may also hinder motility, indirectly affecting colonization ability [73].

Upon entry to the intracellular spaces of the plant, the colonizing microbes typically reside initially in the apoplastic space between plant cells and where gas is exchanged during photosynthesis [78]. It is within this space that the invading microbe encounters two main barriers affecting successful host colonization: interaction with the host plant, and its existing microbiome. Although water availability is comparatively more abundant than on the leaf surface, the colonizer needs to compete with both host and microbiota for water [79]. In fact, genes encoding proteins which control osmotic pressure, such as *avrE1*, are among the most conserved genes among phytopathogenic bacteria such as *P. syringae* [80–82]. Humidity control via these genes is a key determinant of early colonization where redundancies of these genes have been found for some phytopathogens. For *P. syringae*, highly conserved genes *avrE1* and *hopM1* both control osmolarity through different mechanisms, with the latter protein increasing water availability through the plant hormone abscisic

acid signalling pathway [82]. Increased localization of water in the apoplastic space is dominant contributing factor for successful colonization and production of the characteristic water-soaked lesions symptom common in infections [83].

In addition to water, the invading microbe also competes for other resources through the expression of niche adaptation genes, along with synthesis and secretion of small molecules to generate a favourable habitat for itself [84]. Secreted proteins can include cell wall degrading enzymes, such as pectinase, endoglucanase, endomannase, arabinase, xylanase, and more [85–87]. Degradation of the host plant cell wall in conjunction with direct injection of effector proteins into the host cell via the type III (T3SS), type IV (T4SS), and type VI secretion systems allows the invading microbe to manipulate the host environment [88,89]. For many pathogens, effector protein secretion via the type III secretion system is the most common form of virulence for both mammalian and plant host variants [88]. Secreted effector proteins are a very broad protein class defined by its functional contribution to an effect on the host machinery and includes transcription factor mimics, metabolite sequesters, and proteases [90,91]. Depending on the bacterium, the effector protein repertoire (the effectome) can be highly variable and is specific to the species' pathovar strain [24,26]. Compared to *Ralstonia solanacearum*, a phytopathogen with an effectome of approximately 64 effector proteins, *P. syringae* can have a considerably smaller effectome [92]. The effectomes of *P. syringae* strains can typically range from 9 up to 50 proteins, containing proteins used to alter host RNA or protein stability, turnover, and activity [24,90,93]. Effectively, the invading pathogen can change the plant metabolome [84], encourage cell growth [91], interfere with cell signalling [82,94], and most importantly suppress the plant cell defence response [90,95,96].

The T4SS-based form of niche manipulation has primarily been documented in *Agrobacterium tumefaciens*, where a Ti plasmid is injected instead of effector proteins [89]. In this case, plant-like promoters encoded on the plasmid induces the host plant to express effector proteins needed to assume control over the plant cell machinery. Alternatively, the T4SS can also be used to interact with existing microbiota via conjugation [97]. Mobile genetic elements are a common source of virulence genes, and interaction with less virulent subpopulations may influence future virulence. Similar to the T4SS, the T6SS, found in gram-negative bacteria, is another pilus-based secretion system that has been tied to virulence in mammalian pathogens [98]. However, the secretion system has primarily been reported to be involved in microbe-microbe interactions via a toxin-antitoxin mechanism [9]. Rather than host manipulation, this system can benefit colonization via direct removal

of hostile microbes, such as the existing host microbiota, and maintaining subpopulations expressing the same antitoxin.

During host colonization, phytopathogens can also secrete phytotoxins, another key determinant of virulence [99]. In some cases, these toxins can function similarly to effector proteins in host manipulation. In some *Burkholderia* species, synthesis of the phytotoxin toxoflavin can stimulate production of hydrogen peroxide, indirectly affecting the existing microbiota, whereas tropolone strongly sequesters metal ions away from plant use [33]. These metabolites can also act as general immune suppressors, for example coronatine synthesis by *P. syringae* which is used to bypass stomatal closure from restricting further colonization [60,100].

Although phytopathogens, such as the xylem infecting bacterium *Ralstonia solanacearum*, can migrate to other sites of infection establishment in the local area surrounding entry point is critical in early host-colonization [101]. Expression of adhesion genes and exopolysaccharide synthesis has been correlated with higher virulence *in planta* [25]. Conversely, chemotaxis and motility genes are downregulated in favour of genes involved with biofilm production [80]. In fact, an exo-polysaccharide (EPS) deficient mutant strain of *Ralstonia solanacearum* displayed reduced virulence [102]. The EPS has been suggested to both form a general protective layer against dehydration and reduce early build up of plant immunity via calcium chelation [103].

Besides genes encoding secreted proteins, a colonizing microbe expresses a myriad of other niche specific genes unrelated to host/microbiota interaction. Although not directly involved in pathology, these genes are critical in niche adaptation where downregulation or gene deletion can indirectly reduce virulence via poor persistence in the host [104]. For some microbes, biosynthesis of trehalose can alleviate osmotic stress in host plant [105]. Expression of protective genes such as detoxification proteins and phenylpropanoid degrading genes can protect the microbe from cell damage for example from plant-produced reactive oxygen species and metabolites [63,106].

Among these niche adaptation genes, metabolic-related genes are among the most differentiated genes in response to the new habitat [105]. These genes include carbon and amino acid catabolic genes and associated transport genes to maximize nutrient acquisition from the plant metabolome [84]. For example, colonization of a sugar-deficient xylem has shown to induce upregulation of carbon metabolic genes where sucrose acquisition was paramount to xylem colonization [101]. In summary, many phytopathogens passively gain entry through openings in

cuticle walls of its host. Upon entry, host colonization typically occur within the apoplastic space where the microbe expresses a variety of niche adaptation genes needed to establish its niche outcompeting both host and the pre-existing microbial community.

1.4.2 Barriers to successful host plant colonization

There are two main barriers to successful host colonization: the holobiont microbial community and the plant immune system. These factors are not mutually exclusive, and each can directly influence the effectiveness of the other [10,78]. Genomic studies on the natural microbial community have revealed that plants are a home to large diversity of microbial types, including bacteria, fungi, and protists [62]. Typically, these microbes can have a direct neutral (commensal) or beneficial (symbiotic) relationship to their host, and can be beneficial overall due to their role in mediating colonization by incoming microorganisms [107]. Although microbial communities can vary depending on tissue, a study of *Arabidopsis thaliana* metagenome samples reveal that these communities have large taxonomical and functional overlap indicating that the microbiota is relatively uniform throughout the plant [108].

Leaf phytopathogens typically need to contend with species from a range of bacterial phyla including Proteobacteria, Firmicutes, Bacteroides, and Actinobacteria [78]. A diverse set of inter- and intraspecies interactions with these phyla can prevent proliferation of phytopathogens via both direct interactions and indirect mechanisms. Proliferation of invading phytopathogens can be directly inhibited using competition and synthesis of secondary metabolites. Endophytes with identifiable non-ribosomal peptide synthetases (NRPS), and polyketide synthases (PKS) genes were also suggested to have biocontrol potential via biosynthesis of antimicrobial peptides and polyketides using these genes [99,109]. For example, *P. syringae* pv. *syringae* 508 was found to produce the antimicrobial secondary metabolite syringopeptin capable of inhibiting fungal pathogen *Venturia inaequalis*, reducing apple scab formation [7]. Similarly, syringomycin and syringopeptins can also be antagonistic towards bacteria, inhibiting gram-positive bacterial populations [8,110]. Conversely, these same antimicrobial biosynthesis gene clusters have also been identified in phytopathogen *P. syringae* strains with implications in virulence suggesting that secondary metabolite synthesis is a common mechanism of antagonism from both the invading microbe and the plant microbiota during host colonization [71,110,111].

Between bacteria, the type VI secretion system also provides another form of direct competition between a colonizing bacteria and other bacterial species within the holobiont. Found across a diverse range of both commensal and pathogenic bacteria, this phage-like needle secretion system is used to directly inject toxins into competing species [9]. Toxification of itself and related strains is prevented through expression of a companion antitoxin to neutralize the injected toxins. For example, injected effectors such as DNase (Tde) from *Agrobacterium tumefaciens*, have been found to have antibacterial activity against *Pseudomonas aeruginosa*, allowing for former bacteria to outcompete the latter *in planta* [112]. However, effectiveness may be niche dependent as an inverse result was shown *in vitro*. Non-pathogenic plant-associated bacteria, such as commensal *Pseudomonas putida* KT2440, were also demonstrated to be T6SS active competitors, expressing T6 effector toxin Tke2 and its corresponding immunity protein Tki2 [113]. With the prevalence of the T6SS utilization as general form of interbacterial competition creates a second microbial-based barrier to niche adaptation during host colonization.

Members of the plant microbiota can also indirectly inhibit proliferation and mitigate the virulence that more deleterious microbes can impart on the host plant. Residing endophytes have been reported to express quorum-sensing inhibitors [114]. Rather than direct antagonism against the phytopathogens, these molecules can disrupt the exchange of quorum sensing intraspecies signalling needed for both *in planta* survival and higher level of pathogenicity [115,116]. Expression of AiiA, an AHL (alpha-homoserine lactone) lactonase was capable of attenuating AHL-based quorum sensing in *Erwinia carotovora* leading to a reduction of virulence in a variety of plants, including potatoes and carrots [114].

Other ways that the plant microbiota can modulate community members is indirectly through host plant interactions. This phenomenon has been observed as endophytic mediated resistance. Introduction of endophytic fungi, such as *Rhizopus*, were able to reduce virulence with subsequent applications of phytopathogens, such as *P. syringae*, on lima bean (*Phaseolus lunates*) [117]. Although the mechanism for determining activation is not fully understood, endophytes are speculated to pre-emptively activate the host plant's non-specific immune system to induce early resistance making a host plant less susceptible to colonization. Unlike mammalian cells, plants lack dedicated defence cells thus its immune system is primarily comprised of a reactive defence system that is typically localized to site of infection, but with increasing degree of intensity as infection persists [75]. Since the plant's innate immunity is not specific to invading organism and can be

triggered by any microorganism whether it is beneficial, commensal or deleterious [75,118]. In fact, host immune response is one of the main mechanisms used to shape the microbial community.

Arabidopsis thaliana ein2 mutants, with the core ethylene immune signal system disrupted, showed an altered microbial community compared to wild-type plants [119]. The reciprocal nature of immune system priming and microbial community changes in response to immune system activation may ultimately facilitate retention of healthy microbiota in healthy plants.

Although the methods of microbial clearance are consistent during the duration of activation, the plant immune system is typically split into two distinguishable phases depending on the mechanism of activation: the microbial/pathogen associated molecular pattern (MAMP/PAMP) triggered immunity (MTI), and the effector triggered immunity (ETI) (Figure 1.1). The MTI is typically characterized as the early immune response with the detection of microbe presence via generalized molecular patterns common to all microbes. Conversely, the ETI is a follow-up immune response to targeted attempts by the invading microbes to suppress early immune response [75,90]. This latter phase is typically designated as strain-specific due to an observed effector (microbe)-Resistance gene (plant) pair response [120]. However, unlike direct detection of specific antigens common to mammalian immune systems, the ETI is activated through a self-monitoring system [121]. Since both phases rely on indirect detection of microorganisms, the plant immune system is a catchall system incapable of differentiating between beneficial and deleterious microbes.

ETI has been reported to suppress invading microbes similar to MTI, but at an elevated intensity tying both systems as a singular temporal immune system [75]. Jones and Dangl (2006) proposed a model that depicted the plant-pathogen immune response as a race between plant defence activation and phytopathogenic defence suppression with four distinct phases, known as the zigzag model (Figure 1.1). In this model, a quantitative intensity level of plant defence response increases with every cycle of immune activation and suppression. The first three cycles have been attributed to MTI activation, MTI suppression (Effector triggered suppression, or ETS), and subsequent activation of the ETI. After activation of the ETI, the final phase of the model requires novel methods of plant defence suppression or evasion by the microbe by either acquiring novel effectors (gene gain), losing immunity triggering effectors (gene loss), or novel mechanisms through horizontal gene transfer with the surrounding community. Failure to do so allows for immune system signalling to accumulate above the plant specific threshold triggering localized apoptosis known as the hypersensitive response

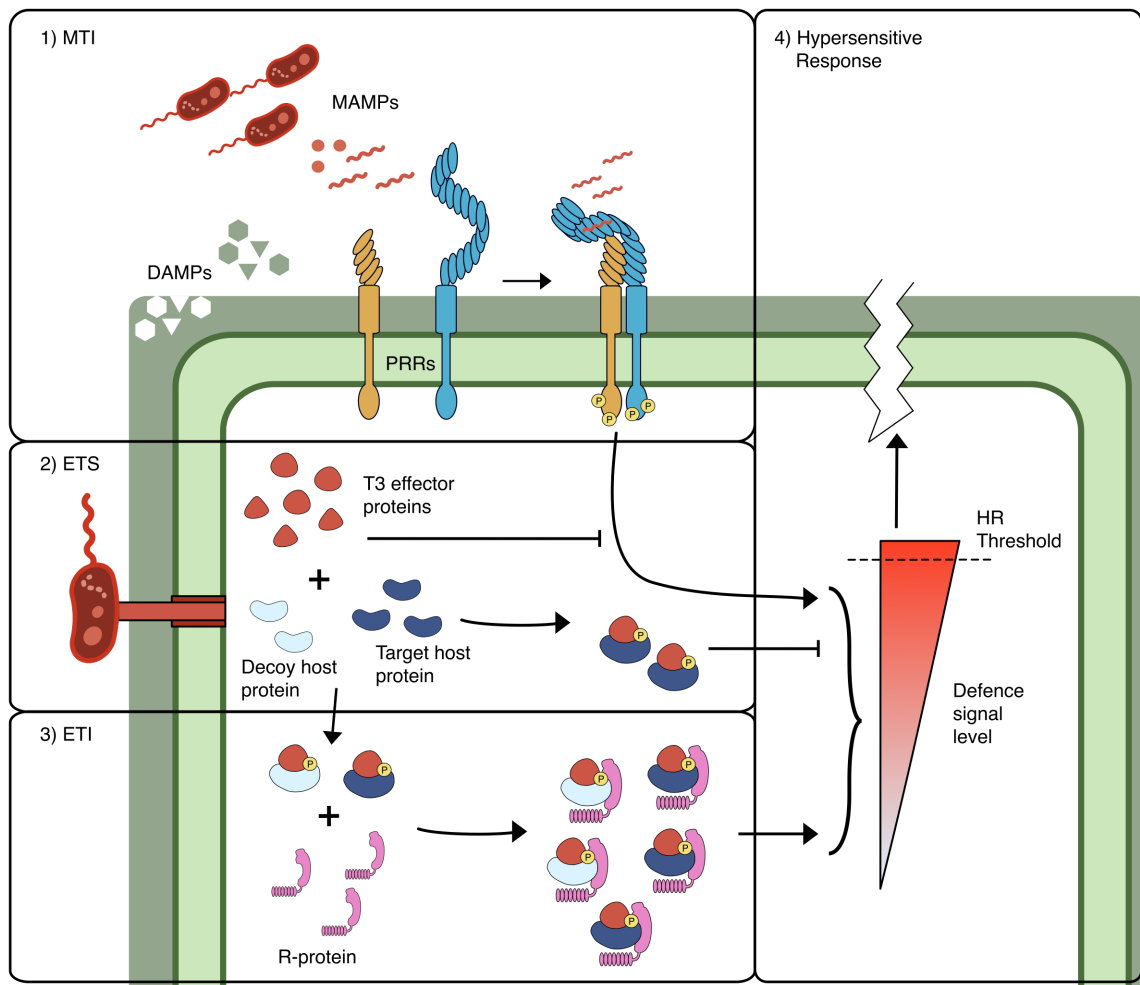


Figure 1.1 Schematic overview of the bacterial plant pathogen and host plant interaction modelled after the four phases of plant defence activation and suppression in Jones and Dangl's (2006) proposed zigzag model. 1) microbial associated molecular patterns (MAMPs) and MAMP triggered immunity (MTI): Plant defence activation is initiated with plant response receptors (PRRs) detecting MAMPs or detecting damage associated molecular patterns (DAMPs). 2) Effector triggered suppression (ETS): Injection of T3 effector proteins by the bacterial type III secretion system suppress plant defence signalling and response through interactions with host protein and DNA. 3) Effector triggered immunity (ETI): indirect detection of T3 effectors via host protein/decoy monitoring triggers increased defence response. 4) Hypersensitive response (HR): accumulation of defence signals beyond the HR threshold triggers apoptosis, reducing pathogen spread from local infection.

(HR), and when combined with priming the defence in distal cells the host limits pathogen proliferation and dispersal to surrounding tissues [122].

1.4.3 Early stages of pathogenicity: MAMP triggered immunity

During a host colonization event, the MAMP triggered immunity is the first phase of the plant defence response that is activated against an invading microbe. Upon entry into the internal organs of the plant, cell surface pattern recognition receptors (PRRs) bound to plant cells detect microbial signatures via molecular patterns found in highly conserved molecules across microorganisms (Figure 1.1a) [123]. These molecular patterns serve as PRR ligands, and are typically conserved domains belonging to common cell components that are secreted or released during cell death [124]. From early plant immunity studies on MTI activation, MAMPs were initially considered virulence and pathogen-associated (PAMPs) some of which include cell wall degrading enzymes, exopolysaccharides, glycolipids, cytokinins, and proteases, but not all MAMPs are virulence-related and can be considered as niche factors [104]. Furthermore, since PRR can also bind damage-associated molecular patterns (DAMPs) as well, whether some of these molecules are from the microbe or their degraded products is uncertain [123].

The archetypal example of a MAMP is the conserved stretch of 22 amino acids situated at the N-terminus of the bacterial flagellum [125]. It is a common point of microbial detection since general motility has been tied to increased survival of both epiphytes and endophytes and is of paramount importance to virulence [126]. Commonly designated as flg22, this conserved domain has been demonstrated to bind to PRR FLS2 in *A. thaliana*. Another highly studied MAMP is the conserved stretch of 18 amino acids at the N-terminus of elongation factor ef-Tu [127]. This growth factor is necessary for bacterial proliferation and is released upon cell death indicating that MAMPs are primarily indicators of microbial presence.

Despite targeting highly conserved ligands, PRR are not synonymous across all plant species [26]. Genetic composition of the plant can influence binding strength of the PRR with its respective ligand. FLS2 receptors from tomato (*Lycopersicon esculentum*) had different binding capabilities between FLS2 receptors found in *Nicotiana benthamiana*, and *A. thaliana* [128]. Although *L. esculentum* was capable of detecting both flg22 and a truncated flg15 variant from *Escherichia coli*, the latter two plants showed reduced sensitivity to the truncated MAMP. Differences in binding can likely influence the degree of defence response and can shape the host microbial community.

Upon microbial contact, when a PRR binds with its respective ligand a signalling cascade is triggered within the plant. For example, when the MAMP flg22 is bound to the FLS2 receptor, the PRR dimerizes with a neighbouring co-receptor BAK1, which is stabilized by ligand binding [129]. Proximity of the two cytosolic domain triggers transphosphorylation of both domains and a bound BIK1 protein [130]. Release of the now active BIK1 through monoubiquitination initiates phosphorylation and signalling of downstream immune responses [131]. Many downstream responses can include opening of calcium channels to trigger synthesis and accumulation of plant hormone salicylic acid (SA), production of reactive oxygen species (ROS) as a general bactericide and secondary messenger, strengthening of cell walls by increased callose deposition, production of nitric oxide triggering guard cells to close the stomatal openings, secretion of general antimicrobial compounds such as acetosyringone, and lastly initiating mitogen-activated protein signalling cascades [123,132–135].

Simultaneous to MTI activation, the plant begins to synthesize hormone signalling molecules salicylic acid, jasmonic acid (JA), ethylene among others [136,137]. Accumulation of these hormones facilitates cross talk and defence activation in both the current and neighbouring plant cells creating a localized defence response. For a specific example, SA triggers the monomerization of nonrepressor pathogen related protein 1 (NPR1) [138]. These monomers then form complexes with WRKY transcription factors to bind and upregulate the expression of SA-related defence genes while downregulating other genes [139]. The resultant complex defence expression profile defence is dependent on hormone signalling and consequently which WRKY transcription factors are active. For example, NPR1 binding to WRKY70 triggers upregulation of SA-mediated defence genes, such as expression of pathogen related (PR) genes including PR1, PR2, and PR5, while downregulating expression of JA-mediated genes such plant defensin 1.2 (PDF1.2) [140]. The inverse occurs when NPR1 binds to WRKY18, WRKY40, or WRKY60 resulting in increased expression of JA-mediated genes and downregulation of SA-mediated genes. In fact, some pathogens take advantage of this complex signal interplay through hormone mimicking to reduce specific defence responses [100]. As the immune reaction persists, accumulation of these hormones builds along with the intensity of defence upregulation.

1.4.4 Plant immune system suppression

Elicitation of the plant immune response can hinder host colonization. Although some phytopathogens have been suggested to evade early immune response, such as through synthesis of polysaccharides that can mask microbes from detection, full evasion of MTI activation has not been reported [70,102]. Instead, a key to phytopathogen persistence against plant defence is via immune suppression to bring MTI response to manageable levels. In fungi, expression of the Ec6 protein pre-emptively minimizes MTI response by binding its own chitin released by plant chitinase activity, minimizing the MAMPs available to act as PRR ligands [141]. Other MTI suppression strategies directly target the defence response. In phytopathogenic bacteria, biosynthesis of some phytotoxins can mediate defence response through hormone signal manipulation [142]. For *P. syringae*, the phytotoxin coronatine acts as a jasmonic acid mimic, taking advantage of the inverse relationship between jasmonic acid and salicylic acid mediated defence response [100]. The MTI activated stomatal closure in *A. thaliana* is a common early immune response to counter continual invasion [60]. Closure via the ingress by guard cells is signalled and mediated by both salicylic acid and abscisic acid. To prevent closure, detection of the toxin coronatine as jasmonic acid triggers the upregulation of JA-mediated genes *lox2*, *cor1*, and *coil* [143]. When combined with pathogen stimulated expression of plant transcription factor MYC2, via ABA signalling, the interaction between COI1 and MYC2 upregulates expression of NAC family transcription factors [100,144]. In turn, these transcription factors inhibit SA-signalling with downregulation of its mediated defence genes, stimulating the re-opening of the stoma [60]. However, these strategies are general methods that nonspecifically target conserved defence signalling pathways present in most plants. More commonly, many phytopathogens also utilize more direct methods of immune response suppression.

1.4.5 Mid-late stage of pathogenicity

1.4.5.1 The Type III secretion system and plant immune suppression

For many bacterial pathogens, the primary mode of virulence is effector protein injection into the target host via the type III secretion system (T3SS) [88,90]. With a close evolutionary relationship to flagella, this secretion system shares many structural and functional similarities, but only limited sequence similarity with flagellar proteins (Figure 1.2) [145]. Internally, the assembly apparatus

shares a similar basal body spanning the periplasmic space, a cytoplasmic c-ring, an inner membrane export apparatus, and ATPase complex to fuel protein export [88]. Externally, the T3SS differs from the flagella with the T3SS injectisome housing an outer membrane ring rather than L-ring and p-ring needed for flagellar motion. Although both external structures are built from short interlinked protein subunits, the T3SS system forms a rigid pilus, which is capped with a tip complex in mammalian pathogen variants [146,147]. Whereas in phytopathogens, a longer pilus has been reported to be required to span the thick cell walls of plants. Once T3SS translocators form a pore within the host cell membrane, attachment of the pilus with the pore creates a direct channel between the bacteria and plant cell for direct secretion of effector protein into the host cell [148]. To ensure proper folding within the host cell, secreted effector proteins are exported with a cognate chaperone that aids in folding [149]. Highly specific to their respective protein, these chaperones are commonly found neighbouring the effector genes, allowing simultaneous regulation and expression. Once in host cells, these proteins then manipulate the cellular processes within the host including disruptions to immune response and transcription factor mimicry [90].

The T3SS injectisome is structurally complex, requiring a specific expression order of its subunits and strict regulation of the expression hierarchy [88]. Although this system plays a major role in virulence, the T3SS is a core genome component and is highly conserved in both beneficial and parasitic microbes [150]. Unlike the vertical inheritance of flagella, the myriad of structural and regulatory genes encoding the T3SS forms mobile genomic clusters that are typically spread as a genomic island. For some *P. syringae*, this pathogenicity island (PAI) is situated within a large mobile plasmid, but in other strains the gene cluster is located within the bacterial chromosome [97,151].

Similar to flagella comparisons, T3SS across bacterial species share structural and functional similarities, but are often divergent in both sequence and gene arrangement [150,152]. Although a universal nomenclature has been proposed to unify the functional similar structural components, uptake of this descriptive system has been slow [88]. In the absence of a broadly accepted universal nomenclature, there are differences between bacterial T3SS with similar nomenclature but are not synonymous between species, i.e., *Ralstonia* and *Xanthomonas* spp. when compared to *P. syringae*. All references to T3SS genes and order within this thesis refers to the *P. syringae* T3SS, with conserved structural proteins such as *hrc* (hypersensitive response / conserved), and non-conserved proteins as *hrp* (hypersensitive response / pathogenicity), unless specified (Figure 1.2) [151].

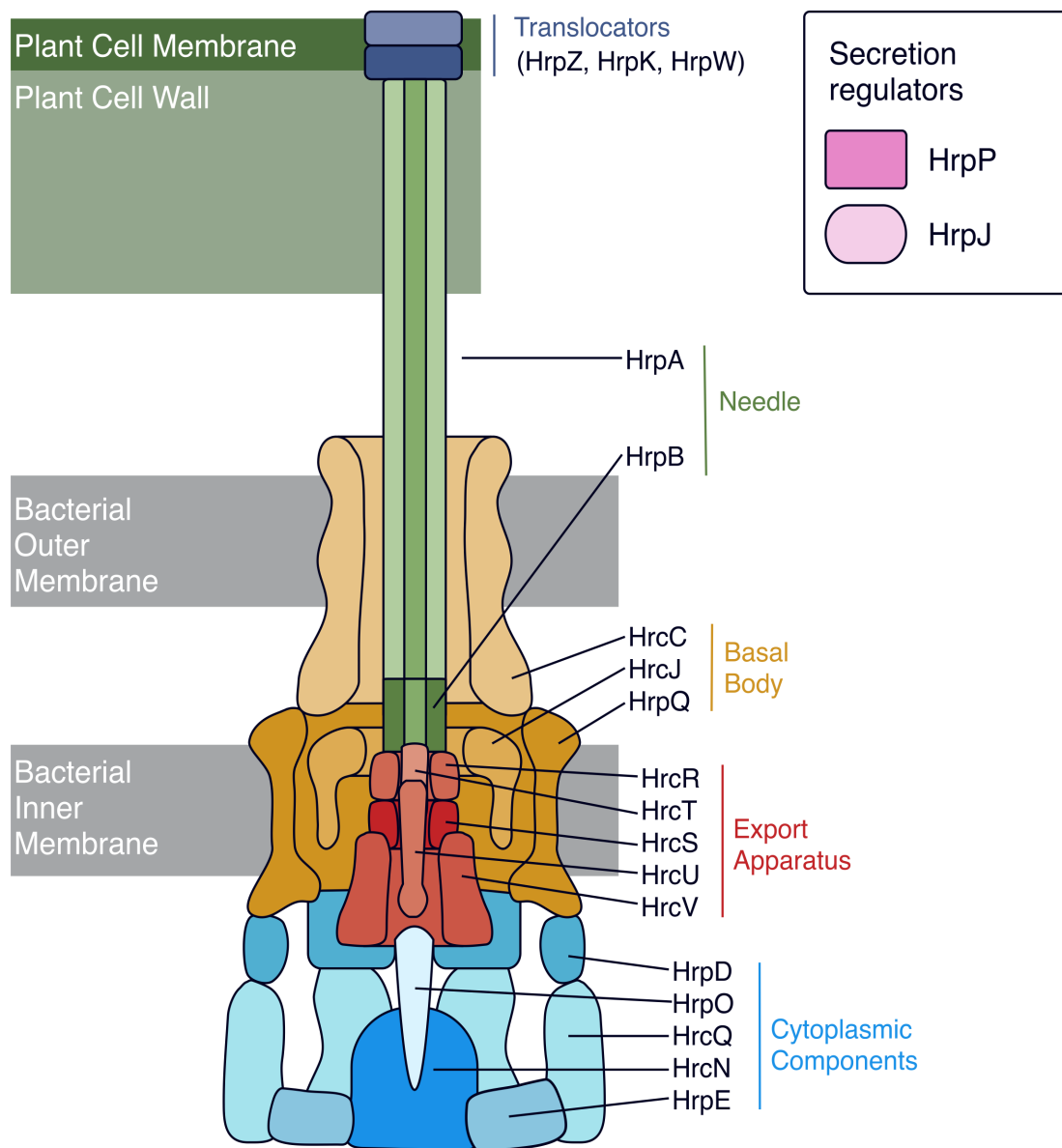


Figure 1.2 Schematic diagram of the *Pseudomonas syringae* type III secretion system (T3SS) injection apparatus based on functional homology to *Pseudomonas aeruginosa* and *Salmonella* spp. T3SS structure. Due to the complexity of the system, expression and assembly of the structure occurs in three major phases: assembly of the basal body and export apparatus, assembly of pilus, and secretion/assembly of the translocator and tip protein. Secretion regulators interact with the injection apparatus to regulate secretion of T3SS effector proteins. Other T3SS regulators not shown include HrpR/HrpS, HrpF, HrpG, HrpT, HrpV, HrpL.

Different T3SS PAIs have also been found in a single bacterial strain with different host targets suggesting host specializations depending on T3SS type [27]. A *P. syringae* strain was found to have a canonically virulent T3SS and a secondary type T3SS (Rhc-like) common to mutualistic rhizobia with no known host targets [153]. Comparatively, both of these T3SS gene clusters do not share any gene arrangement or sequence similarities. Host specialization of each T3SS type likely enable a bacterial species to expand to new niches [27] and gene cluster disparity may be an indicator of specialization. Conversely, PAI variants of the canonical virulent-associated T3SS (not Rhc-like) have also been found in *P. syringae* species complex, although there are some cluster rearrangements, these PAIs appear to be mutually exclusive with all strains harbouring only one of these variants [43].

There are three known T3SS variants of the virulent T3SS in the *P. syringae* species complex: a canonical tripartite T3SS PAI (T-PAI) common to many agriculturally virulent strains, a single partite T3SS (S-PAI) found in some virulent phylogroups, particularly among *P. viridiflava* and *P. cichorii* strains, and an atypical tripartite variant (AT-PAI) identified in a single late branching subphylogroup [18,151,154]. All three PAIs share a core gene cluster plus one to two effector loci. Both T-PAI and AT-PAI have a core gene cluster that is flanked by two effector loci, with the effector loci having different gene and nucleotide composition with the AT-PAI was initially undetectable via DNA-DNA hybridization due to high sequence divergence [18]. In the tripartite PAIs, the upstream effector loci typically contains many highly conserved core effectors including translocator genes and humidity control genes required for the T3SS [148,155]. Downstream of the core gene cluster is an exchangeable effector locus, forming the most variable portion of the pathogenicity island and housing many of the strain-specific effectome and virulence for the respective species [156]. Besides the high diversity in number and types of effector genes in this loci, presence of residual mobile genetic elements (e.g., insertion sequences, partial phages, transposon elements, etc.) indicate that novel effector gain and loss required for host adaptation frequently occurs within this loci [157]. In contrast to the T-PAI, the S-PAI T3SS has a single gene cluster containing both the core and exchange effector loci [154], with the effector locus flanked by the core genes.

All three *P. syringae* PAI variants have the same organization of the core gene cluster, with two exceptions: the copy number of the *hrpS/R* regulator gene and a gene fusion/fission of the conserved *hrcQ* gene [152,154]. Based on sequence data the distribution of the *P. syringae* T3SS is polyphyletic, and debate on the evolutionary path of the PAI remains ongoing [43]. Additionally, the presence of the atypical variant in a sub-phylogroup among a broader group of T-PAI containing

phylogroups supports some form of non-vertical inheritance [18]. Conversely, T3SS integration in the bacterial chromosome, and strong congruence in phylogroup distribution of conserved core genes may indicate low mobility of the core PAI clusters compared to the effector loci [158]. However, non-conserved *hrp* genes *hrpZ* and *hrpA* are divergent enough for some researchers to suggest they should be used in grouping of *P. syringae* strains, especially when *hrpZ* is involved in contact and pore formation in host cells along with other translocator proteins containing harpin domains [159]. For AT-PAI, sequence divergence across all T3SS genes likely indicates a different origin, and that acquisition may be indicative of a niche change [27,43]. Despite this, experimental confirmation of an active T3SS against plant hosts for all three variants has been confirmed [18,151,154]. Furthermore, mutual exclusivity of these clusters two suggest overlapping function when compared to the two distinct T3SS in mammalian pathogen *Salmonella* spp. [160]. Both SPI-1 and SPI-2 have differentiated PAI arrangement and specialize in different host targets. Therefore, the link between T3SS variation, function, and niche in *P. syringae*, remains to be determined.

1.4.5.2 Regulation of T3SS for precise plant response

Expression of the T3SS injectisome is an energy intensive process, and it has been debated whether it is transiently expressed or constitutively expressed, with some *in planta* transcriptome studies suggesting constitutive expression [63,161]. Although the exact signal for T3SS expression is uncertain, T3SS expression studies suggest that both minimal media, i.e., nutrient deficiency, and presence of plant signal from exudates are needed for significant upregulation of T3SS expression [162]. With a substrate secretion hierarchy, many of the non-conserved *hrp* genes in the T3SS encode a complicated network of regulator genes to ensure correct temporal expression [88]. When the mammalian pathogen T3SS variant is fully expressed, the secretion hierarchy is speculated to occur in three major phases (Figure 1.2). Prior to the first phase, internal structural components are (basal body, export apparatus, and cytoplasmic components) are assembled either with inside-out or outside-in model via the sec-dependent secretion pathway. In the first phase after host detection, the earliest secreted substrates are the subunits of needle and inner rod forming the pilus. This phase is followed with secretion of the second batch substrates, the translocators and tip proteins (in mammalian variants). These protein form the first contact with the host cell, forming an open pore for pilus docking [148,163,164]. With a direct channel into host cells, the final phase fully expresses the late

stage substrates, the effector proteins to be translocated into the host [88]. Strict regulation of secretion is required for efficient injection into host cells. When expression of effector protein AvrPto was coupled with a SA-inducing promoter, the effector protein was demonstrated to be prematurely secreted during pilus assembly, prior to plant cell contact [146].

To initiate T3SS expression in the plant pathogen *R. solanacearum*, the outer membrane receptor PrhA was identified to induce the expression of its T3SS regulatory gene HrpG on contact with plant signals [165]. If a similar mechanism is present in *P. syringae* the expression of the gene cluster is likely under repression until host colonization. In *P. syringae*, expression of the entire *hrp* regulon is initiated by global regulators *hrpR/S*, which encode the sigma factor RpoN and its respective enhancer-binding protein [166,167]. Unlike *R. solanacearum*, the environmental signal regulator has not been identified in *P. syringae* since an equivalent receptor domain was not identified HrpR/S indicating they are unlikely to directly receive the plant signal [168,169]. Upon *hrpR/S* expression, heterodimerization of the pair allows for full expression of the *hrpL*, an ECF sigma factor, the direct global regulator of the *hrp* regulon [168,170]. However, without dimerization, *hrpL* expression can still occur with a singular copy of the pair as seen in the single partite variant of T3SS [154,167]. Like other sigma factors, when HrpL is bound to RNA polymerase, the RNAP complex identifies *hrp*-box promoter motifs inducing expression of the PAI gene cluster, associated effector genes, and general virulence-associated genes [171]. However, tight regulation of the initial HrpR/S expression, particularly HrpS which is directly regulated by HrpR based on data from *P. syringae* pv. *phaseolicola* [172], controls when T3SS-based virulence becomes active. Besides enhancing *hrpL* transcription with HrpR, HrpS can also solely weakly induce HrpL and an entire suite of T3SS-independent genes making it a prime target for regulation [167,170,173].

Upon initial activation of the canonical T-PAI T3SS, accumulation of *hrpR* transcripts occurs within an hour of colonization, but HrpA1 appears to be required for expressed HrpR to activate expression of *hrpS* suggesting that basal body assembly is complete prior to full expression in T-PAI [169,174,175]. Presence of HrpA1 likely triggers a signalling cascade that promotes the HrpR/S heterodimerization and HrpL expression. When HrpA1 is present, negative repressor HrpF was identified to stabilize HrpA1 for T3SS biogenesis [176]. The protein complex sequesters HrpF from binding to the derepressor HrpG, leaving the free HrpG to bind to negative repressor HrpV [177]. The HrpGV complex thus prevents free HrpV from binding to HrpS, freeing the protein to dimerize with HrpR to induce HrpL expression [178]. Conversely, S-PAI lacks enhancer HrpS containing only a

singular global regulator HrpR [154]. Without HrpS, T3SS regulatory genes may target the HrpR directly for stricter regulation.

The requirement of HrpA1 for full expression indicates that T3SS expression occurs in stages because HrpL induces the PAI gene cluster including HrpA1 [179]. In fact, HrpL is an indirect auto-regulator of its own expression [180]. Binding of RNAP-HrpL complex to the *hrpJ* promoter blocks access to the *hrpL* promoter, repressing its own expression after the initial activation phase. Thus, auto-repression allows for the expression of HrpA1, for injectisome assembly, and HrpJ, a secretion regulator that physically plugs the export channel after complexing with HrpGV [181]. Detection of external HrpJ suggests secretion of the protein may be tied to freeing the T3SS channel for export [182,183]. A negative regulatory function has also been identified with *hrpT* expression, but its role has yet to be identified in the expression cascade. The complicated network of regulation indicates the need to fine-tune virulence regulation until the appropriate signals are detected.

1.4.5.3 Effector protein suppression and ETI activation determines success in infection

When in contact with the plant cell, secreted translocator proteins have been suggested to dimerize forming a pore on the host cell surface [148,164]. The tip complex on the T3SS then docks onto this pore forming a direct channel for injection of effector proteins into the host plant [88]. Although not comprising the entire effectome, some of these secreted effector proteins function as immunity suppressors at different cellular levels [75,90]. The T3 effector AvrPto disrupts MTI immunity signalling through antagonistic binding to the kinase domains of PRR receptors FLS2 and EFR, preventing activation of the signalling cascade [95]. Similarly, AvoPtoB secreted by *P. syringae* pv. *tomato* DC3000 disrupts more downstream immune signalling by recruiting the host's 26S proteasome, triggering ubiquitination of the plant Fen kinases [184]. At the RNA level, HopU1 from the same strain was found to antagonistically bind to the host protein GRP7 preventing binding to *fls2* and *efr* mRNA, reducing the expression of new PRRs [185]. As a collective, the suppression by secreted effector proteins brings the quantity of immune response to a manageable level [75]. Consequently, large and diverse effectomes have been identified in many phytopathogens [24,92].

Under Jones and Dangl's zigzag model (Figure 1.1), plant disease resistance and susceptibility is the result of the balance between immunity activation and pathogen suppression [75]. Furthermore, both MTI and ETI were found to activate similar transcriptome profiles indicating

similar modes of immune response for both defence response corroborating that plant resistance is quantitative, increasing in intensity as infection persists [122]. Effectively, ETI is an enhanced MTI response, upregulating many of the same genes to counteract pathogen suppression [75,186]. In some cases, without MTI suppression ETI is unnecessary, and MTI can trigger full resistance. When the purified cell wall glycoprotein CBEL (a MAMP) from oomycete *Phytophthora parasitica* was exogenously applied tobacco plants, it is capable of triggering HR-like symptoms [187]. Since MTI activation continues throughout infection, when combined with unsuppressed ETI, the plant immune signal can overwhelm the pathogen virulence signal, tipping the balance towards plant resistance.

Without the dedicated immune cells seen in animal immunity, to trigger ETI, the pathogen's effector based suppression is indirectly detected in host plants. Even with the diversity of effector genes, many of the effectors have similar targets [188]. Under the guard and decoy model (Figure 1.3), plants are able to detect effector protein function through self-monitoring of the integrity and activity of immune genes and proteins [121,189]. Early MTI activation primes the host cells for immune resistance by activation of plant Resistance (R) proteins containing nucleotide binding leucine-rich repeating domains (NLRs) [122]. Some R proteins act as guard, where they bind to a host protein or decoy proteins, where they act as the effector protein target. Typically, when host protein is compromised, the guard R protein is unbound and free to activate the respective defence response [189]. As an example, *P. syringae* pv. *tomato* T1 secretes the T3 effector AvrRpt2, a protease that directly cleaves RIN4 in *A. thaliana*, a regulator of plant immunity [190]. However, RIN4 is defended by two R proteins, the guard protein RPS2 [191], and the decoy protein RPM1 [188]. Degradation of RIN4 via AvrRpt2 releases the bound RPS2 triggering activation of ETI genes. Similarly, RIN4 can be inactivated through phosphorylation by effector proteins AvrRpm1. In this case, R protein RPM1 acts as a decoy protein with accessible sites for phosphorylation. Alternate phosphorylation of RIN4 and RPM1 was detected in samples with AvrRpm1, thus activating RPM1 for immune signalling. Decoys are not limited to proteins, with *upa20*-like promoters acting as decoy promoter upstream of the R gene Bs3 in peppers [192]. The effector protein AvrBs3, expressed in *Xanthomonas campestris* pv. *Vesicatoria*, targets these promoters and inadvertently upregulates the plant defence response.

Typically, both effector suppression and R gene detection have been classified as strain-specific/host-specific virulence factors with each recognizable T3E gene having a corresponding R gene [120,121,189]. When an R gene is identified as a target of an effector protein, the effector gene

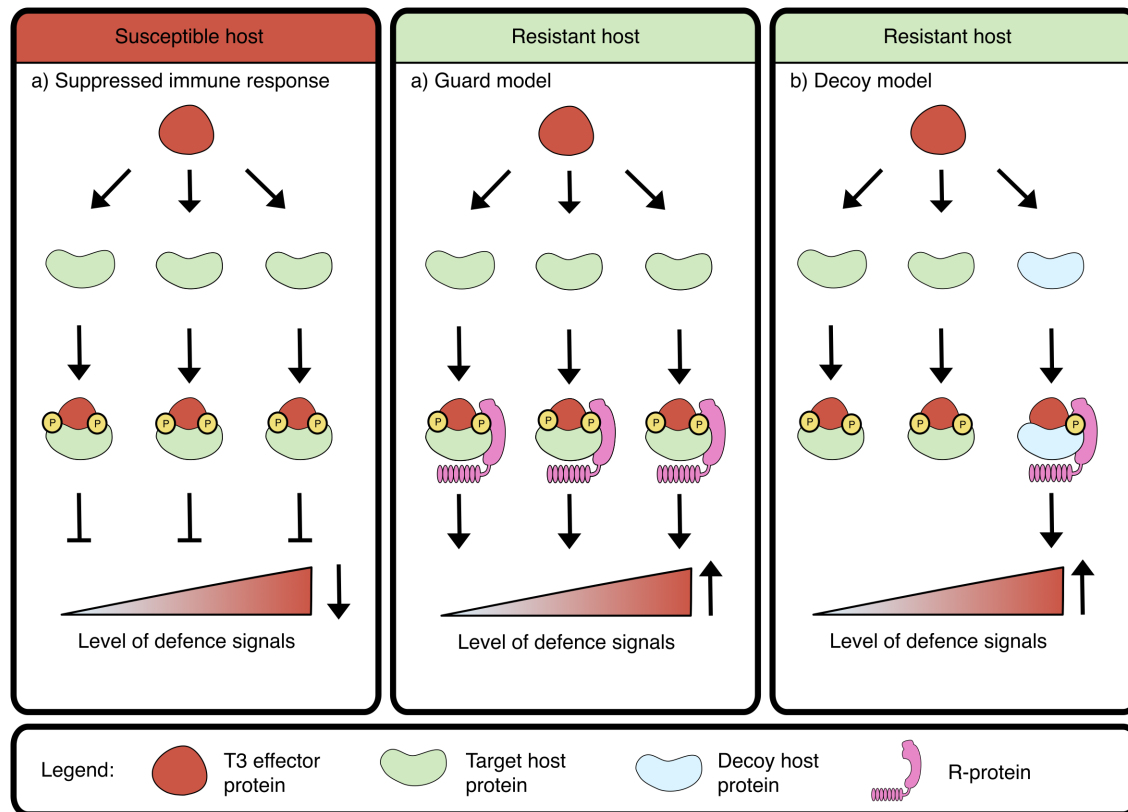


Figure 1.3 Overview of the indirect detection of pathogenic T3 effector proteins with its respective Resistance proteins (R-protein) under the guard and decoy model. The T3 effector in this example depicts only one of many different method of host protein manipulation (i.e., phosphorylation). a) T3 effector protein binding to unguarded target host protein leads to suppressed immune response b) Under the guard model, the R-protein detects unusual changes to the target host protein (guardee) triggering increased immune response. c) Under the decoy model, the R-protein detects changes to a separate sensor protein (decoy) triggering increased immune response.

is re-classified as an avirulence gene (*avr* and the respective R gene name, i.e., RPM1 and effector gene *avrRpm1*) as detection renders the phytopathogen avirulent. This indirect effector detection system allows the plant broader protection than direct recognition. For example, the R gene RPM1 can detect phosphorylation from both effector proteins AvrRpm1 and AvrB, indicating it is not a direct one-to-one pairing system [188].

Activation of the ETI also plays a major role in developing localized acquired defence to prevent systemic infection by limiting pathogen spread [122]. The effects of ETI include potentiating salicylic acid (SA) accumulation, triggering of the hypersensitive response, and activating the defence

system of more distal cells [132]. When localized resistance is activated, unless the pathogen acquires a novel gain-of-function gene or a gene from the surrounding community, the phytopathogen is considered avirulent and unable to systemically infect the host plant [193]. When immunity is triggered, a subset of defence expressed proteins contain a specific CC-NLR domain that can function as a calcium channel [122]. Cytoplasmic calcium is a major regulator of the hormone SA and increased cytoplasmic calcium can increase SA production [103,194]. When immune signals and SA levels exceed some threshold, it triggers localized resistance through the hypersensitive response (HR)[195]. In this case, plant cells in the local area of the infection trigger apoptosis, combined with the rapid influx of reactive oxygen species and ions. Induced cell death limits plant resources available for the pathogen to utilize. Simultaneously, SA also signals to distal cells to prime defence responses, such as increasing callose deposition in neighbouring cells to thicken cell walls against future microbial invasions and inducing SA-mediated systemic acquired resistance [196]. Recently ETI alone has been postulated to also induce a similar process in cells surrounding the local infection [122]. Jacob et al. (2023) suggests that ETI does not need to overcome effects of effector protein suppression. They found that endophytic communities outside of the local infection sites did not contain injected pathogens in all samples compared to high proliferation within local sites. NLR containing proteins released during HR interacts with neighbouring cells to activate the defence response prior to bacterial contact, thereby containing the pathogen to the local site.

Typically, a plant-pathogen pair eliciting both local and systemic acquired resistance is indicative of both a resistant host and an avirulent pathogen or incompatibility between the phytopathogen and a non-host plant, which can still be virulent in other plants [132,197]. In cases of compatibility, virulent pathogens can spread beyond local infection sites in susceptible plants. For example *P. syringae* pv. tomato DC3000 can cause bacterial speck disease in tomato while *P. syringae* pv. *syringae* B728a causes chlorosis and bacterial speck in bean plants [142,158]. Other diseases induced by *P. syringae* can include leaf blight, stem root and leaf collapse. Mitigation of the immune system prior to localized resistance is dependent on the pathogen's MAMPs, and effector profile, with even singular genes behaving as potential determining factors in host specificity [26]. Against pepper plants, xanthomonads lacking the avirulence gene *avrBsT* were capable of infecting the host, indicating detection of the corresponding protein by the plant is instrumental in pepper resistance against *Xanthomonas* spp. [198]. Similarly, phytopathogens with effectomes expressing AvrRpt2, AvrRpm1, and AvrB will trigger resistance in *A. thaliana* [188,190]. Consequently, some

studies attempted to determine host specificity of a phytopathogen from its effectome [43,57], however predictive capabilities of this approach have been limited to date.

1.5 Sequencing technology and bacterial genomics

1.5.1 History of development of sequencing and use in genomics of microorganisms

Since the development of Sanger sequencing and linear PCR cycling, the ability to automate sequencing has opened up the exploration of all organisms at the genetic level, especially exploration of microbial genomes [199]. Building upon the idea of individual base detection via synthesis, next generation high throughput technology relied on isolating individual nucleic acid strands, either on beads (Roche's 454 pyrosequencing) or wells (Illumina's synthesis based sequencing). With fluorescent dNTPs, individual nucleotide base can be identified during synthesis via detection of specific light emission corresponding to incorporated dNTPs. Due to a focus on throughput, the cost per base for Illumina short read sequencing dropped, facilitating the accessibility of sequencing for many genetic studies. With the newest iteration of this technology, the NextSeq 1000/2000 benchtop sequencer, can output yield equivalent of 600 *P. syringae* strains at 100x coverage in 48 hours [111,158,200]. As a consequence, the number of microbial genome sequences has exponentially increased with many microbial genome research groups dedicated to comprehensive sequencing of their respective microbe. The ability to mass sequence has provided invaluable research including tracking outbreaks, tracing micro-evolution, expression analyses, and analyses of microbial populations [14,50,179].

Illumina's technology is limited by physical space, and as a synthesis-based technology, the fidelity of DNA polymerase [199,201]. The need for strand anchoring and bridge amplification for clonal cluster formation limits the insert size that can be sequenced. Read quality steadily decreases during strand extension so high-quality short reads can range between 75-300 bp. Conversely, the technology leverages the clonal clusters to amplify sequence signal creating a consensus signal for light emission detection accuracy thereby increasing the quality of their reads. During sequencing runs with high enough coverage of the target assembly (>30x yield of the expected assembly size), the resultant sequence quality can have an error rate of <1% [202]. However, despite these advantages, the short read length has significant disadvantages in downstream analyses. Since these reads are shorter than an average bacterial gene, reads may not fully span the target DNA fragment [203]. This creates ambiguity in overlap between reads containing repeat sequences leading to poor

resolution of repetitive regions [204]. Without additional targeted sequencing of these regions, contigs cannot be ordered and joined into a singular contiguous contig. The resulting assemblies are often highly fragmented, hindering analyses on regions involving repetitive sequences, structural rearrangements, transpositions and insertion sequences [205]. Using short reads alone, fragmentation of draft assemblies of bacterial genomes into many contigs is common even at high (100x) base coverage [18]. In some cases, missing genes need to be considered during data analysis. Since draft assemblies are pervasive among bacterial genomes, analysis of gene presence/absence in genomes is often done in terms of a 'hard core genome' (genes present in 100% of genomes) vs. 'soft core genome' (genes present in ~95% of genomes) [206]. These issues become compounded in studies where a high degree of fragmentation can hinder species resolution and gene set analyses, preventing accurate identification of effectors in *P. syringae* effector loci [207].

Since 2011, PacBio has innovated on the synthesis-based sequencing method to develop its single molecule real-time (SMRT) sequencing technology capable of generating significantly longer reads, currently averaging around 30kb in length [208]. Although longer reads can be sequenced with their SMRT continuous long read methodology, the error rate of individual reads increases, with sequencing accuracy dropping to 86-89%. Unlike Illumina's sequencing by synthesis (SBS) technology, the SMRT system double stranded insert sequences are ligated with single stranded adapters forming an looping bell-like structure where one adapter binds to an embedded DNA polymerase in one of the 8 million wells on the SMRT chip. As with other synthesis methods, light emission during incorporation of fluorescent dNTP by the DNA polymerase is recorded. The use of a high fidelity polymerase allows for the generation of very long reads using two complementary approaches [209]. Earlier methods of continuous long reads were intended for very long insert sequence where the sequencing was allowed to run as long as quality allowed [208]. This method can create reads up to 100kb in size. Furthermore, with a single molecule isolated in each well, amplification into sequence cluster and consensus is unneeded reducing confounding light signals of neighbouring clusters allowing for more precise basecalling [210]. Conversely, any errors are amplified without a consensus leading to generation of long but error prone reads, where a single pass had an error rate of ~13% (RSII P6-C4 chemistry) [211]. However, even with poorer read accuracy, the large overlaps between reads are highly advantageous allowing the assembly of complete bacterial chromosomes with only long reads [212]. Additionally, an analysis of synthetic reads claims that long reads can be preferable to short reads for metagenomic studies [213]. To improve read quality of their

long reads, PacBio released their HiFi sequencing method in 2020, which takes advantage of the circular shape of the SMRTbell via circular consensus sequencing (CCS) [211]. This approach uses smaller insert sizes between 10-25kb, but the DNA polymerase makes multiple passes over the insert by looping around the circular DNA. Sequenced adapters are recognized allowing for several passes of the insert sequence to be collapsed into a consensus sequence during basecalling, effectively reducing the error rate to <1%.

Besides poorer read accuracy, another disadvantage of longer reads was that the throughput of long reads are considerably lower than short read technologies. A PacBio 8 million well chip and 25kb CCS reads can generate up to 200Gb, but with the newest release of the Revio SMRT cell containing 25 million wells, throughput can be theoretically increased up to 375Gb, comparable to Illumina sequencing [214]. For microbial genomics, there are a few other drawbacks to use of PacBio HiFi especially for smaller sequencing projects when compared to large mammalian genomes. Large operating costs (e.g., a dedicated facility, complex system upkeep, and technician costs) can make PacBio systems less accessible [215]. Furthermore since microbes have smaller genomes, lower throughput systems such as Illumina's NextSeq system can be more cost effective compared to SMRT sequencing minimum of 8 million wells.

1.5.2 Nanopore sequencing technology: a new form of sequencing to fully resolve bacterial genomes

Since 2015, nanopore sequencing technology has emerged as alternative long read sequencing approach that can fully resolve genomes in an efficient sequencing system [216]. The concept was initiated with the identification of pore-forming channels produced by pathogens, sparking researchers to look for applications of this ability to transport molecules across a gradient potential separated by a membrane [217]. Upon discovery that single molecules could be possibly identified during transport across the *Staphylococcus aureus* alpha-hemolysin channel, the protein channel drew interest in its potential for sequencing applications [218]. When Garaj et al. (2010) were able to create an ionic gradient across a thin graphene barrier, they discovered that a graphene sheet paired with a nanopore could be used as a trans-electrode for the detection of current changes [219]. Furthermore, when paired with a modified pore channel, disruptions to the electric current from molecules passing the channel could be identified at high resolution [220].

From this discovery, Oxford Nanopore Technology (ONT) established a new innovative form of non-synthesis based sequencing with the early access release of MinION in 2015: nanopore sequencing [216]. In this system, nanopores are affixed to a graphene matrix that creates a salt gradient between two chambers, similar to a cell membrane. An ASICs chip then records the ionic current that passes the nanopore. During sequencing, motor proteins pre-bound to DNA adapters feed one strand of the DNA insert through the nanopore. In order to control the rate of nucleotide feed through, a modified motor protein was developed to bind to a mutated CsgG (Curlin sigma S-dependent growth subunit G) nanopore to feed bound DNA at an approximate rate of 420bp/s in initial iterations of the technology (i.e., R7 flowcells) [220]. Feeding the DNA in a stepwise manner, DNA translocation was slow enough to allow a set 5-6 nucleotide stretches of DNA to become lodged within the nanopore long enough to record alterations to the ionic current through the nanopore [216]. For R9.4 flowcells, the current changes between 6 nt kmers were then used to identify the DNA sequence during basecalling. Similar to synthesis methods, the stepwise data acquisition from 5' to 3' allows for real time sequencing and concurrent basecalling.

Boasting a smaller form factor than both Illumina and PacBio platforms, the first release ONT (2015) has managed to fit the entire sequencing system into a 10.5cm (L) x 2.3cm (W) x 5.5cm (D) machine [221]. The MinION sequencer contains 2048 nanopores embedded in the ASICs board as a hexagonal array as groups (called channels) of 4 nanopores [220]. Without the need for DNA synthesis, nanopore sequencing does not require additional reagents allowing each flowcell to come ready-to-use, improving both portability and ease-of-use compared to the more traditional sequencing systems. Consequently, beyond initial setup of the overall system, the sequencer does not require upkeep when flowcells can be easily swapped between projects with little downtime. Impressively, the compact form does not sacrifice sequence yield. A MinION flowcell with 2048 nanopores has reported a theoretical maximum yield of 50Gb, nearly 5x greater than the ~13Gb maximum yield for a MiSeq with v3 reagent kit, a sequencing run at a similar price point [221,222]. However, ONT only guarantees 800 nanopores in these flowcells suggesting that sequencing output can be order of magnitudes lower than expected. One multiplexed MinION sequencing run, reporting 1432 active pores, generated 10.48Gb a comparable yield to MiSeq [223]. The read throughput is significantly lower at ~837k reads compared to Illumina's MiSeq and PacBio's SMRT flowcells that can generate millions of reads. However, Nanopore throughput can be increased via joining up to 5 flowcells in ONT's gridION system released in 2017 [221]. Recently, even high yield can be achieved using the

recent promethION flowcell, a 2675 channel (10700 nanopore) flowcell with a guaranteed 5000 active nanopores [221]. Although the higher throughput is beneficial for larger sequencing projects such as human genome assembly, nanopore sequencing use has been primarily supplemental to other sequencing methods [224,225]. For microbes, the smaller genomes of bacterial pathogens make the higher throughput systems less cost effective outside of larger sequencing projects commonly found in clinical setting [226–228]. Since PacBio does not offer SMRT chip with fewer than 8M wells, nanopore sequencing offers a more cost-effective approach to long read sequencing for microbial genetics projects with a smaller scope. With the release of the Flongle flowcells, offering 126 channels (guaranteeing 50 of 504 nanopores), long reads will become more accessible at a highly reduced cost (~\$100 compared to ~\$1500 for minION flowcells). These flowcells can be used in targeted microbial sequencing where a theoretical maximum of 2.8Gb can produce >560x coverage of the average 5.0Mb bacterial genome [203,221].

Unlike both Illumina and PacBio systems, nanopore sequencing does not rely on robust DNA polymerase fidelity for accurate base detection. Therefore, read length using nanopore sequencing is theoretically unconstrained without base accuracy decline towards the 3' end of reads [229]. Although insert size is limitless, the largest read size achieved using this platform to date is 2 Mb [230]. A larger read size has yet to be reported due to physical limitations on the system including the library preparation methods, and molecular dynamics for the delivery of the DNA to the pore [231]. During library preparation, both methods of adapter ligation reduce the size of the DNA insert. Using a standard ligation kit, sample DNA is recommended to be sheared to size (i.e., Covaris g-Tube shearing can yield 20kb DNA inserts) and magnetically isolated for standardization of read size. In the non-standard library preparation workflow, the rapid library kit, transposase-based random attachment of DNA adapters can yield a highly variable left-skewed distribution of read sizes ranging from a few nucleotides to ultra long reads (100kb+) [221]. However, sequencing with these ultra long DNA inserts is hampered by molecular dynamics [231]. A long DNA strand has high mass, but only two motor proteins for nanopore binding, whereas, a solution of 10x smaller DNA strands at same molarity are more likely to encounter and bind to the nanopores. Furthermore, long DNA strands produce more viscous solutions prone to manual shearing and poor diffusion across sequencing chamber, but these issues can be mitigated with dedicated protocols for ultra long reads (100 kb+).

Although read quality does not decline for longer reads, Nanopore read sequences have high error rates, similar to PacBio's long reads [216]. With sequencing of single molecules it is not

possible to average out individual errors in signal, in contrast to consensus calling in Illumina's clonal clusters. The high sensitivity of an electrical-based signal creates higher levels of noise which make it harder to distinguish kmers, compared to the four possible light signals with bases called using fluorescent dyes [204,220]. ONT and third party developers address the issue of low signal-to-noise with trained neural networks based basecallers. As the software and current detection methods mature, nanopore reads should continue to improve, even for previously collected sequence run data [232]. In one of the earliest public releases, R7 flowcells had a reported error rate of 36% (64% sequence accuracy) using official basecallers, but some experimental third party software have been reported to further reduced error rates to 22% (78% accuracy) [233,234]. ONT have also continual worked to reduce noise through chemistry improvements. Release of the R9.4 flowcell showed substantial improvement reducing the error rate to 15%, with official basecallers, which is comparable to PacBio's sequencing with its error rate 13% [211,216]. Similar to PacBio's CCS method, ONT have attempted to incorporate sequencing of complement DNA strands, with Nanopore's 2D and 1D² reads [235]. The former type (2D), released in 2016, incorporates a hairpin adapter on one end of the DNA insert to physically pull the complementary strand through the nanopore after sequencing of the primary strand. The latter type (1D²), released in 2018, was introduced in R9.5 flowcells and is intended to improve on the 2D read concept; these reads utilized a special adapter, that also carried a bound motor protein, on both ends [221]. After feed through of the first strand, the special adapter helps bring the complementary motor protein within proximity to the nanopore increasing the probability of sequencing the complement though this is not guaranteed. The error was reportedly reduced to 5%, but both R9.5 and 1D² read technology has since been discontinued.

Besides a higher error rate, the error profile for both long read technologies are different than that for Illumina short reads. High prevalence of short insertion and deletion were found in long read data, accounting for 92% of errors in PacBio CCS reads [211]. Unlike substitution errors, these indel errors have been attributed to miscalls in homopolymer lengths, and are especially prevalent in low complexity regions [236,237]. Since nanopore sequencing rely on changes in current, the low complexity from homopolymer signal around the detected kmer size can confound the true homopolymer length. These indel errors are systematic making them hard to resolve through consensus sequence, and can greatly hinder downstream analyses. Consequently, highly accurate genome assemblies using long reads were found to contain gene-fragmenting errors [238]. An

example human assembly with 99.8% identity, generated from both long and short reads had at least 2746 genes contain an indel error. As this is a known issue, many error correction algorithms have been continually developed targeting this issue, and can come pre-packaged in assemblers such as wtdbg2 [239] and Canu [240]. In the newer iterations of the flowcells (R10+), ONT addresses homopolymer errors by implementing two reader heads that independently detect read signal [241]. Combined with a flip-flop model that utilizes both signals, ONT's new GPU basecaller Guppy (replacing previous basecaller Albacore) shows improvement in individual read accuracy to ~89%, but with custom models targeted towards the study organisms accuracy can be marginally improved to ~91% [232]. Between ONT and third party researchers, software dedicated to improved homopolymer basecalling are in development and read accuracy of nanopore sequencing will continue to improve as the technology evolves.

1.5.3 Data analysis of long reads in whole genome sequencing

With the emergence of the nanopore sequencing, bioinformatic software targeting nanopore long read sequences are still developing. As the technology continues to mature, these newer software are continuously being released while older software have been discontinued or updated/improved. With the rapid expansion of this field, many of these tools have been documented at <https://long-read-tools.org/>.

1.5.3.1 Base calling of nanopore sequenced reads

To simplify use of the platform, the operating software for sequencing, MinKNOW, comes with web-based basecalling pre-implemented, allowing for live basecalling with ONT's official basecaller during sequencing [221]. Metadata, raw information, and basecalling information are all stored in a hierarchical data format, the Fast5 file format, but raw reads are provided as fastq files as well. During basecalling, changes in ion flow are identified in the raw current data, segmenting the data into “events” representing each k-mer. Progression of the DNA through the nanopore is reported at a fixed rate (420bp/s for R7 and 450bp/s for R9.4), however the rate of kmer events is influenced by preceding events [216,220,242]. Thus, early basecallers implemented hidden Markov models for kmer event identification before transitioning to trained neural networks (NN) in current basecallers

[202,232,243]. Official basecallers have shown improvements in event differentiation over time; error rates in later versions of ONT's original basecaller Albacore reduced error rates from 36% in R7 flowcells to ~15% in R9.4 flowcells [216,220,242]. ONT's newer GPU-based basecaller Guppy further improved read accuracy to 89%, which is sufficient for many applications [232]. With more developmental versions of their newest basecaller, Flappie, implementation of flip-flop model targeting homopolymer errors is expected to further reduce error rates with newer releases [244]. Since ONT basecallers are trained on human, yeast, and *Escherichia coli*, users can further improve accuracy in basecalling by applying custom trained NN models to their target organism [232,237]. Additionally, other third-party basecallers have been developed to improve efficiency and detection of common errors to varying success [233,241,243]. Due to its hierarchical format, the fast5 file format can store multiple basecalling runs as nested runs. Prior to genome assembly, read quality may also be improved with pre-corrector or consensus calling. Unless the user customizes their own assembly pipeline, both pre-correction and read filtering are common pre-processing steps implemented in many assembler pipelines [239,240,245].

1.5.3.2 Genome assembly using nanopore long reads

Two main approaches used to assemble high quality whole genomes from long reads are *de novo* assembly and hybrid assembly [237]. Both methods typically rely upon additional short read sequencing to supplement the error-prone long reads. *De novo* assembly is the simplest form. Since the release of PacBio's SMRT system, whole bacterial genomes could be assembled solely using long reads [212]. Similarly, Loman's research group in 2015 demonstrated that nanopore long reads were also capable of complete bacterial assembly, recreating the *Escherichia coli* K-12 MG1655 genome with 99.5% nucleotide identity [229]. Unlike with short reads, *de novo* assembly of long read can use a classic 2000's era overlap-layout-consensus approach to genome assembly [237]. In the first step, reads are typically mapped together and ordered to create an overlap graph. Large overlap spans between long reads and the relatively low number of reads compared to short reads highly simplifies the overlap graph, making it a feasible approach for long-read assembly [246]. At the layout stage, the overlap graph is resolved into contigs while the consensus step maps reads to this contig to determine the consensus sequence of the contig [247,248]. Since homopolymer errors are non-randomly located, appearing in similar aligned locations across reads, resolution of these errors in the consensus sequence is difficult and leads to error prone assemblies [238]. Post-assembly polishing software has been developed to address these errors. One popular software, Nanopolish, is a signal-based hidden

Markov model that tries to resolve mapped reads to the assembly by reconciling raw read signals at each mapped position [229]. ONT also released its own polishing software Medaka, a neural network-based polisher used to identify and correct homopolymers; other third party researchers have also released their own toolkits for *de novo* polishing [241,249,250]. Although these methods have improved sequence quality of *de novo* assemblies, the standardization of Illumina sequencing in sequencing workflows has led to many studies opting to supplement long reads with additional higher quality short reads with a hybrid assembly approach.

With hybrid assemblers, both long and short reads are incorporated during genome assembly. Unicycler, a hybrid assembler using SPAdes as the base assembler, builds its assembly graph using kmer data from short reads [251]. Long reads are then used to resolve nodes with multiple (3+) edges by creating a bridge between singular nodes. Taking account the node copy number, the long read helps path finding through the assembly graph resolving them into singular contigs. As another form of hybrid assembly, long reads can also be used to align and order the contigs generated from short reads acting as a scaffold in low complexity regions [28,252]. These methods are particularly useful for follow-up studies of draft genomes where access to the original short reads and assembly graph may not be available. Lastly, another common form of hybrid assembly is post-assembly polishing of long read assemblies with higher quality short reads [253,254]. Supplemental short reads mapped to *de novo* assemblies are used to call a consensus sequence for the assembly. All three approaches to hybrid assembly can significantly improve the quality of the final assembly. Although superior to solely *de novo* assemblies, poor alignment between higher quality short reads and the more error-prone long reads can lead to mismatches in mapping and subpar correction [237]. Despite these approaches in long read based methods, a high level of errors were found in all forms of long read assemblies, including hybrid assemblies of the human genome [238,255]. Even assemblies with a reported >99.8% identity to the target genome, presence of these gene-fragmenting errors suggests that improvements in long read assemblies are still needed for assembly of high quality complete genomes.

1.6 Thesis Objectives: Analysis of pathogens using long reads

T3 Effector protein secretion is one of the primary modes of virulence of phytopathogens [88,90,156]. For the phytopathogen *P. syringae*, these effector genes have been primary located

within the effector locus of the T3SS encoding genomic island, but they can also be found throughout the bacterial genome [158]. Many analyses of these effectomes use data annotated from draft genomes. Analysis of the size and diversity of effectome as well as presence of specific genes can then provide insight into respective virulent strains including identification of virulence genes of newly isolated strains and determination of recombination rates between virulent strains [24,43]. However, many of these studies do not provide a complete picture of gene order and arrangement, as many of these genomes are draft assemblies with a high degree of fragmentation. Location and interpretation of exchangeable effector loci are highly impacted by genome fragmentation, creating difficulties in identifying loci variations linked to T3SS regulation in different phytopathogenic strains [156,157]. Since the *hrp* regulon encompasses other genes beyond effector genes, a full effector locus can provide insight in gene retention and mobility of these areas [193].

The objective of this thesis is to explore the virulence capability of two putatively non-virulent *P. syringae* strains, *P. syringae* pv. *syringae* 508 and *P. syringae* TLP2, that were previously determined to have an active T3SS, but elicited variable plant (*P. vulgaris*) responses when compared to reference strain *P. syringae* B728a [18]. Furthermore, high genetic similarity between all three strains allows for insightful comparison on genetic differences that may be linked to phenotype differentiation. Therefore, I hypothesize that some of the genetic contributors to the phenotypic variation between virulent and putatively non-virulent (biocontrol) strains of phytopathogenic *P. syringae* can be identified through comparison of complete genomes of closely related strains. However, the variability and high mobility of effector loci is difficult to resolve in fragmented assemblies necessitating full assembly of the complete genome. Therefore, another objective of the work is to test the then newly released nanopore sequencing system to generate full assemblies as an alternative to the common workflow where disjointed bacterial draft assemblies are generated using short read sequencing. I further hypothesize that with complete assemblies, identification of divergent gene clusters can provide insight on mobile regions and phenotypic variance between these *P. syringae* strains. A reduced genome and effectome size from gene loss likely are the main contributors to poor virulence in these strains. Since the sequencing focus of this thesis is solely on two *P. syringae* strains, a secondary objective of this thesis is the exploration of the capabilities of the recently developed nanopore sequencing technology to fully assemble bacterial genome without additional sequencing. This objective will have a minor focus on practicality of nanopore technology for smaller sequencing projects.

Chapter 2:

Assessment of ONT's Nanopore sequencing use in assembling complete bacterial genomes

2.1 Introduction

2.1.1 Chapter objectives and overview

With the release of ONT's nanopore sequencing technology in 2015, early adoption of nanopore sequencing in genomic workflow was low. But with improvements to the technology and release of more community-developed software list in <https://long-read-tools.org/>, use of the technology in genomics is gaining popularity. In this work, our objective was to test the effectiveness of this system for smaller projects as an efficient alternative long read sequencing approach to PacBio. We also wanted to investigate whether this sequencing system is capable of generating high quality full, complete bacterial genomes allowing for in-depth analysis of the assemblies via full resolution of commonly disjoint regions in short read assemblies, such as the T3SS exchangeable effector loci [156]. In this chapter, we tested the nanopore sequencing system by sequencing two *Pseudomonas syringae* bacteria using the MinION sequencer [7,256]. Both experiments generated a considerable amount of data from each flowcell used, with >1.8Gbps (>300x coverage), for each bacterium in less than 3 days, including library preparation. These sequenced long reads were then sufficient to fully resolve both genomes into complete error-prone assemblies, with genes fragmented as a result of indel errors. Despite ease of use, early adoption of Nanopore sequencing for assembly of the two target genomes was hindered by the lack of standardized and tested software.

2.1.2 Chapter background

Although readily available, DNA sequencing to date has been primarily obtained from a few short read sequencing systems with Illumina's HiSeq and MiSeq becoming the standard for many sequencing projects. Illumina utilizes a synthesis based sequencing method where data is determined by fluorescence from dyed-dNTPs during DNA synthesis of the DNA fragments [257]. Anchoring and amplification of these DNA fragments is required to form DNA clusters in order to reach the minimum light sensitivity for accurate base recognition [210]. Constraints in DNA cluster size and

fidelity of DNA polymerase during both amplification and sequencing limits potential read length with base accuracy declining during later extension steps of DNA cycling. Consequently, trimming the tails of Illumina reads can further reduce read length [258]. The short length of these reads limits subsequent downstream analyses. For genome analyses, use of short reads can produce highly discontinuous draft assemblies; even with a high (100x) genomic coverage, a draft assembly with 20 or more reported contigs for a singular chromosome bacterium is not unusual [18].

Pacific BioSciences provided a solution to these issues with the 2011 release of its Single-molecule real-time (SMRT) third-generation sequencing platform, a synthesis-based system capable of producing longer reads [259]. Single molecules embedded into individual wells are sequenced using a robust DNA polymerase attached to adapters ligated to DNA fragments. In contrast to Illumina systems, PacBio sequencing is not restricted by a PCR amplification step, however read length is still constrained based on fidelity of DNA polymerase during incorporation of fluorescently labelled dNTPs [201,259]. SMRT sequel systems can produce more error-prone long reads (>20kb) with Consensus Long Read (CLR) sequencing, or shorter but more accurate reads with Circular Consensus Sequencing (CCS) [211]. Nevertheless, with a longer read length, SMRT reads can resolve full bacterial genomes without additional sequencing or scaffolding required [212].

However these sequencing systems have a high initial equipment cost and appropriate technical expertise is required to operate these systems. Combined with the large size of the hardware, access to these systems is frequently limited to sequencing centers and/or shared across entire departments limiting access to sequencing. Smaller sequencing projects may be delayed until a sufficient pool of samples is available for cost efficient sequencing. Targeting this niche, Oxford Nanopore Technologies (ONT) released an innovative nanopore-based sequencing platform marketed as an ease-of-use, portable, and time-efficient method of sequencing [216]. Nanopores are capable of detecting single molecule using changes in electrical current during translocation across a graphene surface, simulating the ionic gradient created by the cytoplasmic membrane [260,261]. ONT's R9.4 flowcell uses a modified motor protein, attached via ligated adapters, to feed DNA fragments through a mutant CsgG nanopore at an approximate rate of 450bp/s [220]. With this controlled rate of DNA feed-through, the electric current associated with each instance of 6bp nucleotide stretches passing through the nanopore can be recorded. A model can then generated for the current differential between all combinations of 6bp DNA. These models can then be used to translate changes in electrical current patterns to corresponding nucleotides for real-time base calling [216]. Unlike the

synthesis-based sequencing methods, nanopore sequencing does not rely on robust DNA polymerase fidelity for correct dNTP incorporation; therefore, base accuracy does not decline during sequencing so ultra long read lengths of > 100 kb are possible [229]. However, detection of minute changes in electrical signal makes this system sensitive to physical and chemical conditions. Clogging of the nanopore and variability in the speed of DNA passage can create spurious signals and noise that are difficult to differentiate [262]. Consequently, a high average error rate of up to 15% has been observed in R9.4 cells, comparable to ~13% error rate of PacBio's CLR sequencing. The observed error rate is expected to decrease with maturation of the sequencing technology [242].

Comparatively, ONT's MinION system is highly portable with the entire system packed in a small 10.5 cm (L) x 2.3 cm (W) x 5.5 cm (D) machine, excluding the computer required to interface with the sequencer [263]. With the required components packaged in a disposable flowcell and lacking moveable parts, technician upkeep is not required after the initial setup. For sequencing consistency a standard workflow library preparation kit requires third party consumables and controlled DNA shearing, physically limiting read length (i.e., Ligation Sequencing Kit). However, ONT offers a transposase library preparation method (i.e., Rapid Sequencing Kit) for low DNA input and is unbounded by DNA length, at the cost of a reduction in consistency.

2.2 Material and Methods

2.2.1 Bacterial strains, DNA extraction and library preparation

Two putatively non-pathogenic strains of the ubiquitous plant associated bacteria *Pseudomonas syringae* [21], *Pseudomonas syringae* pv. *syringae* 508 [7] and *Pseudomonas syringae* TLP2 [256], were selected for sequencing using Oxford Nanopore Technologies' MinION sequencing technology [216]. A single colony of each bacterial isolate was grown for 24 hours in King's B broth at 28 degrees Celsius. With the intent to maximize the likelihood of a full genome assembly, methods tailored to maximizing read length were used in all downstream steps. Genomic DNA was isolated using Quick's long read DNA extraction protocol [231]; extra care was taken during extraction to avoid shearing of the DNA. In brief, both gram-negative bacteria were lysed using a gentle SDS-lysis buffer. After multiple rounds of buffered phenol:chloroform:isoamyl-alcohol (25:24:1), the extracted genomic DNA was then precipitated using a 3M potassium acetate solution and 100% ethanol. The

DNA was isolated via DNA spooling, in a single attempt, with a glass rod and gently washed with multiple rounds of 10mM Tris (pH 8) and slowly resuspended in the same solution. Since no plasmids were detected in either bacteria using targeted plasmid isolation, DNA missed from the first attempt was discarded. DNA intactness was assessed for minimal shearing via agarose gel electrophoresis. Concentration and quality of the DNA was assessed using a Qubit fluorometer as a 1:10 diluted and sheared sample.

In order to maximize read length, DNA libraries of both genomes were prepared using the Rapid Sequencing kit; RAD002 for Pss508 and RAD004 for PsyTLP2. Each MinION R9.4 flowcell quality was checked using the provided lambda DNA and flushed at least 2 times prior to library loading. With this method, the prepared DNA is expected to be sheared with DNA transposase resulting in low fragment molarity but high DNA concentration from the higher mass of intact genomic DNA [231]. Total DNA loaded was increased to 16ug to ensure sufficient DNA strands for nanopore docking. The provided library loading beads were also excluded to avoid possible interference of the beads with the longer strands of DNA.

2.2.2 Nanopore sequencing, offline basecalling, and analysis of sequencing results

Real time sequencing in 2018 was initiated on a 6-core, 128 Gb RAM Ubuntu 14.04 workstation using MinKnow software (versions 1.5.12 and 1.10.23 for Pss508 and PsyTLP2 respectively) without live basecalling. The Pss508 sequencing run proceeded for a full 48 hours whereas the PsyTLP2 sequencing run was stopped at 22 hours after sufficient data was acquired. Using the available software at the time, Raw fast5 sequencing data were basecalled using the standalone version of ONT's original basecalling software Albacore (version 2.1.3) and converted to a fastq file [232]. Sequencing results were analysed using a custom Rscript. All tables and plots were generated using R version 3.4.2 and Rstudio version 1.1.383 using packages dplyr (version 0.7.8), tidyr (version 0.7.2), ggplot2 (version 2.2.1) and ggpubr (version 0.1.8).

2.2.3 Draft genome assembly

Reads that pass ONT's quality threshold (Phred score of ~16) were checked for Lambda DNA contamination by aligning all reads, via Bowtie2 [264], to a reference Lambda genome

(Accession: NC_001416). Lambda-aligning reads were then removed. The reads were then further processed by trimming potential adapters by removing 50 nucleotides from either end of each read. Any reads shorter than the trimmed length were removed (i.e., <100bp). To minimize the errors, the reads were further filtered with a high cut-off value set at a Phred score of 21 (3rd quartile in quality scores). The final assemblies for each of the bacterial strains were generated using an optimized Overlap-Layout-Consensus (OLC) pipeline. Overlaps were generated using Minimap2 (version 2.10-r761) [247], the overlap layout graph and its corresponding unitig was generated using Miniasm (version 0.2-r159-dirty)[265], and the consensus sequence for the final contig was called using Racon using ONT provided optimized parameters for the nanopore sequencing system, -M 8 -X -1 -G -4 -E -1 [248]. For each sequenced strain, different combinations of relevant parameters such as overlap stringency were tested to generate a single contig between 5.7 to 6.0Mb in size [18,158]. The quality of post-assembly genomes were evaluated for completeness with a targeted search for intact single, universal orthologs via BUSCO (version 3.0)[266].

For each bacterium, two assemblies (the most BUSCO complete, and the longest single contig) were then further polished using the signal-based HMM software Nanopolish (version 1.6) with methylation-aware (-ma) and homopolymer focus (-hp) parameters turned on [229]. For further improvement, a secondary polishing step was done using the developmental version of the neural network software Medaka (version 0.2.0) under the default models [249]. After inspection for circularization via a self-overlap check, read coverage and reference synteny (Mauve), each assembly was manually circularized with a custom script setting *dnaA* at position 1 [267,268]. OLC pipeline scripts were run on a 4-core, 64 Gb RAM Linux workstation running the Fedora OS. Both Nanopolish and Medaka polishing steps were run using Compute Canada's Graham cluster using general compute nodes (32-cores, 128Gb RAM) and GPU nodes (16-core, 64Gb RAM) respectively.

2.3 Results

2.3.1 Assessment of two R9.4 flowcell MinION sequencing runs

Using two separate nanopore sequencing runs, the R9.4 flowcell generated 2.53 and 1.83 Gb of data for *P. syringae* pv. *syringae* 508 and *P. syringae* TLP2 respectively (Table 2.1). Since the genomes of strains of this bacterial species can range in size from 5.7 Mb to >6.0Mb, a genome size

assumption of 6.0Mb indicates each flowcell yielded a theoretical coverage of 421x and 310x, more than ten times the previously recommended minimum 30x coverage of past human whole genome short read sequencing studies [269]. Despite the high coverage, the total yield significantly was considerably less than the reported theoretical maximum of 50Gb [263]. These yields are also comparatively lower than the expected yield of 11-13Gb (at 48 hours) from the short read sequencing equivalent at a similar price point (~\$1000 – 1500 CAD including first party consumables only), the MiSeq system (v3 reagent kit) [270].

In contrast to short read sequencing, this lower yield is spread over a considerably smaller number of reads. Only 121,883 and 231,484 reads were generated for each sequencing run respectively, with the majority of sequencing occurring within the first 24 hours (Figure 2.1). For Pss508, the first 24 hours accounts for 81.3% (99138) of total reads, 83.4% of the total yield, and provides ~351x coverage of expected genome. In fact, when discounting sequence quality, MinION sequencing generated 100x theoretical coverage of the Pss508 genome within the first 6 hours (Figure 2.2). Beyond 24 hours, the sequenced reads were largely redundant and slightly lower in quality, with a lower mean in both read length (average of 21374 bp reduced to 18357 bp) and read quality (Phred score of 24 decreased to 20). This uneven distribution in read generation suggests that more than 24 hours of sequencing is largely redundant for a single bacterial genome. With PsyTLP2, the sequencing run was halted at 22 hours. Similar to the Pss508 sequencing run, a sequencing yield of 2.11 Gb was generated, providing more than sufficient coverage (310x) for the targeted genome size (6.0Mb).

In both sequencing runs, the highest numbers of reads were generated at the beginning of the run (Figure 2.1). When segmented into 48 hours for the Pss508 sequencing, the greatest of yield is generated within ten hours and decreases in quantity over time at a relatively consistent rate. At each 8 hour interval, there is a temporary increase in reads sequenced that may correspond to programmed reversal of currents to induce rejection of attached molecules and reduce clogging [262]. The benefits are initially effective and boosts the number of sequenced reads by more than 1000 in the next hour. However, the sequencing rate typically returns to the pre-reversal level quickly and continues to decline. Later de-clogging attempts become less effective, with hours 33 and 41 showing no discernable increase in read output and yield (Figure 2.2).

Similar to read quantity, the quality of sequenced reads decreases over time albeit at slower rate, with initial reads having the highest median quality. Furthermore, there were slight

Table 2.1 R9.4 flowcell sequencing statistics using standalone Albacore basecalling

	<i>P. syringae</i> pv. <i>syringae</i> 508	<i>P. syringae</i> TLP2
Sequencing hours	48	22
Total Yield (Gb)	2.53	1.83
Number of Reads	121883	231484
Mean/Median read length (bp)	20379/9304	7990/4553
Longest read length (bp)	569,475	267,384
Mean/Median Phred Score quality	17.1/18.6	20.0/20.8
Max Quality	29.4	27.7
Estimated coverage ¹	~351x	~310x

¹ – Based on an estimated genome size of 6.0Mb

improvements in read quality following each 8 hour interval supporting that the systematic refresh can boost overall quality, but the impact was relatively minor (Figure 2.1). Compared to short read sequencing [202], the quality of nanopore long reads never exceeds a Phred score of 30 (1 error in 1000 bp), consistent with the higher error rate of the nanopore sequencing system previously reported [232]. Although there is a decrease in interquartile range across time, the full range of read quality remains relatively stable throughout the sequencing run, including when the quality significantly drops following the 43rd hour. The stability of this quality range suggests that the decrease in read quality may indicate that the influencing factor is localized to individual nanopore conditions rather than a systematic change in the flowcell environment. Even up to the final hour, some pores can produce both high quality reads and long reads.

I also investigated whether individual pore activity affects changes in the quantity and quality of reads sequenced. At the onset of sequencing, almost half (~1000) of the embedded 2048 nanopores were inactive, though this is not unexpected and is consistent with an ONT guarantee of 800 active pores [263]. Embedded as a hexagonal array, the 2048 nanopores are grouped into multiplex channels of 4 pores across the ASICS board. Although the MinKnow output does not record specific nanopores, we can extract channel positioning, via software configurations, and map each read to a

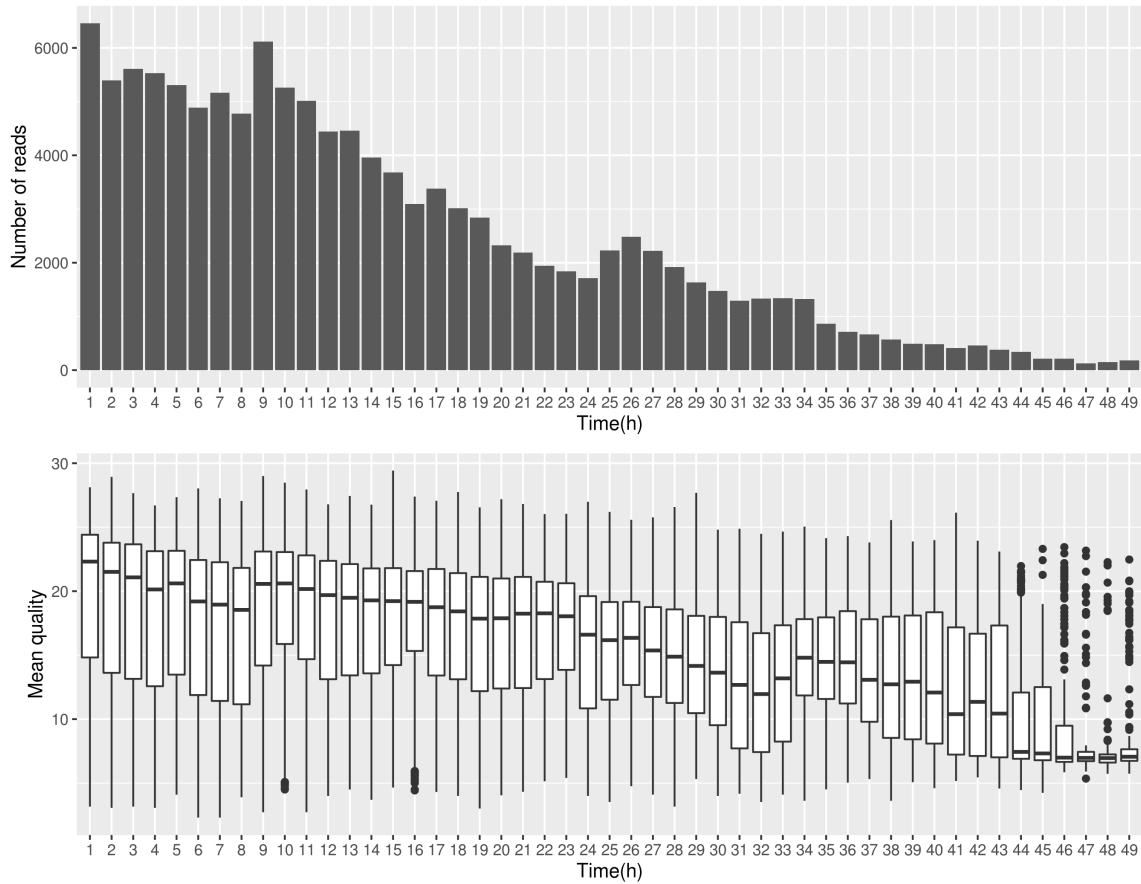


Figure 2.1 Quantity and quality of reads sequenced with DNA prepared using the RAD002 Rapid library preparation method on a R9.4 flowcell over the course of a 48 hr sequencing run. (Top) Number of reads sequenced across time. (Bottom) Average (ONT reported) quality of reads over time (Phred scores: an estimate of the probability of an incorrect basecall, at the specified nucleotide position, on a negative log scale; $-10\log_{10}(P)$).

particular channel position. From this data, 512 channels are arranged as two adjacent 16 x 16 arrays with DNA input, presumably, entering the well from rightmost position (Figure 2.3a left); where channel 1 is situated in the top rightmost corner. Across the entire run, sequencing activity is relatively uniformly scattered throughout the well with exception to the top left quadrant and a few channels in the bottom right. Among these regions, there are completely inactive channels (all four pores are inactive) across all time points indicating "dead" spots that can hinder total yield. Without proper reporting, we could not deduce whether pores were inactive prior to the run. However, under

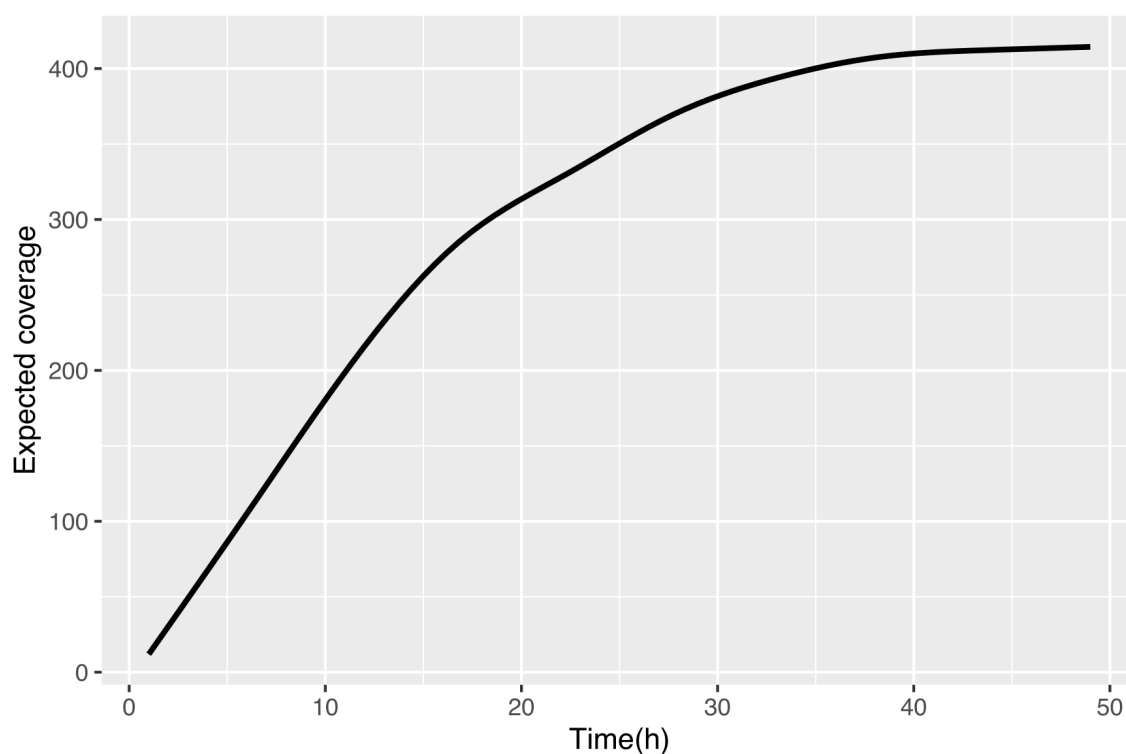
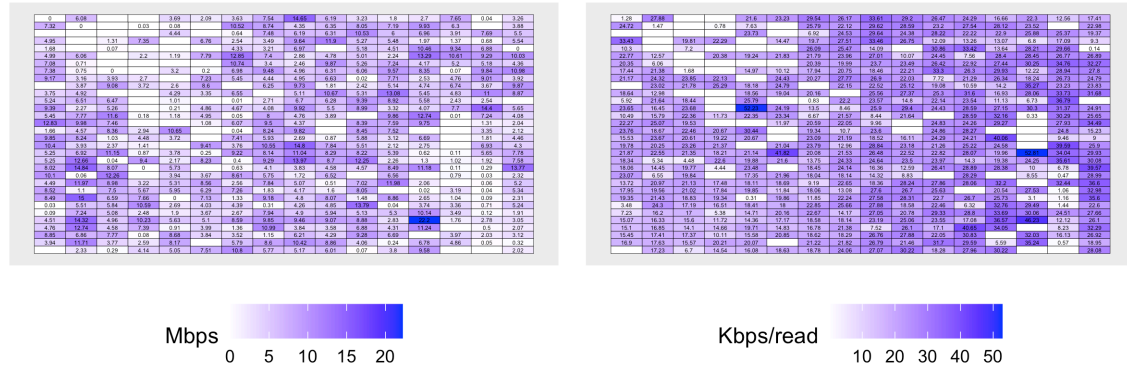


Figure 2.2 Expected coverage of Pss508, with an estimated genome size of 6.0Mb, across time based on total sequence yield.

the assumption of the right side DNA input, consistent channel inactivity on the left side suggests that initial inactivity can be localized into “dead spots”.

When the activity of these channels are normalized by read output, a light gradient indicates that nanopore channels producing on average the longest reads are concentrated in the bottom right, with a few exceptions (Figure 2.3a right). The poor diffusion of high mass DNA across the flowcell indicates that long reads are likely physically limited to a subsection of available nanopores. Furthermore, there is a slight inverse relationship between total sequence output and the longer read output. Bottom right channels producing the longest reads have typically lower read output (Figure 2.3a right). Across the flowcells, channels produced a median of 242 reads, yet the top 20 long read-sequencing channels produce a median of 140 reads. The specific relationship between producing longer reads is unclear as more than the contribution of multiple nanopores towards channel activity

a) Total channel activity



b) Channel activity over time

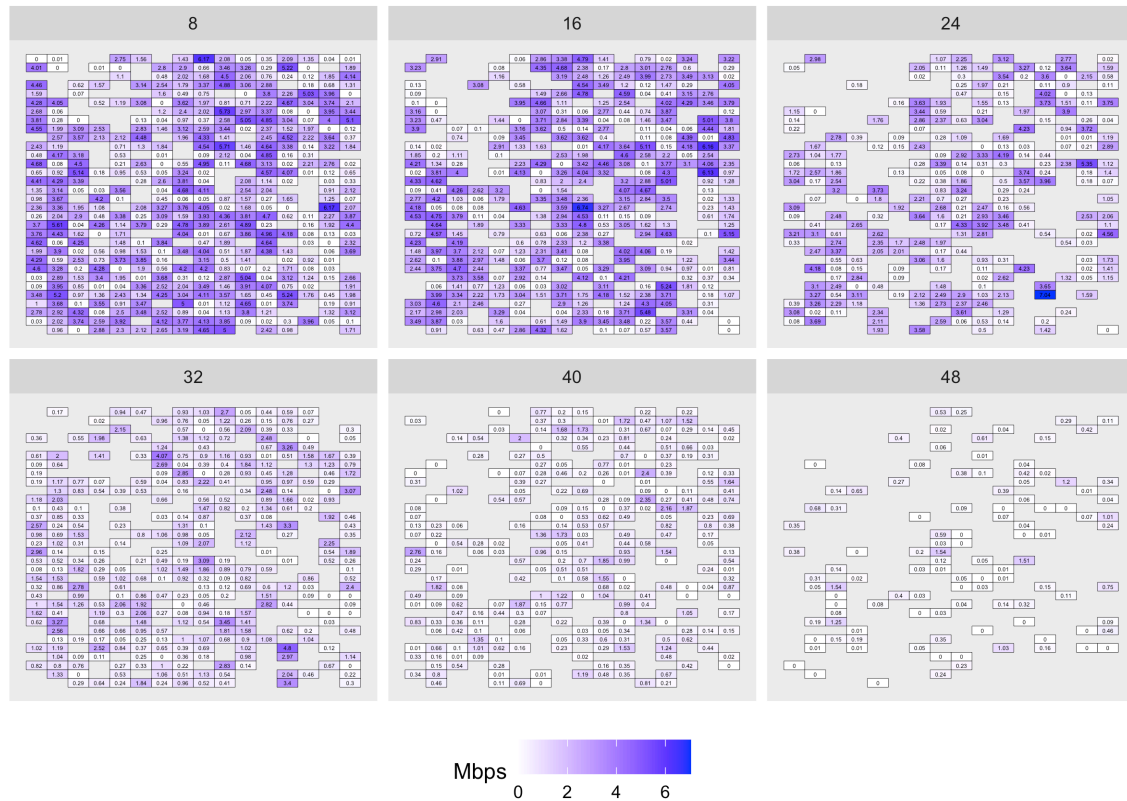


Figure 2.3 Heatmap of the number of reads sequenced per multiplex channel across the entire Pss508 sequencing run. Each of the 512 channels reflects four nanopores arranged in a hexagonal array, but individual nanopore position could not be tracked. Channel positioning within the heat map is indicative of the relative position on the 32 (horizontal) x 16 (vertical) asics chip. Sample input diffuses from right side to the left. (A) Total channel activity by total yield (left) and yield per read (right). (B) Yield of each channel in 8 hour intervals. Missing channels indicate all four nanopores are inactive for the time interval.

can confound this effect; the top 2 channels produce a collective 6 reads while a 3rd channel produced 38 reads. However, when we track channel activity across time, these channels typically are quicker to become inactive (Figure 2.3b). Among the top 20 channels producing longer reads, 12 of the 20 (60%) are inactive within 32 hours compared to 150 of all 512 (29.2%) channels. Four of the top 5 channels were inactive in the first 8 hours with fifth channel inactivating within 16 hours. Rather than read length, persistence in channel activity correlates with high initial activity rather than overall yield. Clusters of channels with a higher initial sequencing yield tended to persist to the final 16 hours (Figure 2.3b), but generated an average total yield (Figure 2.3a left). In fact, 12 of the top 20 read generating channels (60%) were active throughout the entire sequencing run and the other 8 channels were active up to 40 hours. Although these channels produce on average slightly shorter reads (median of 15.0 kb/read vs the overall of 22.0 kb/read), these channels were highly active and generated greater than 600 reads. In fact, the most active channel (156), sequenced 607 reads in for a combined total yield of 22.2 Mbps with an average read length of 36.6kbps per read, and remained active for 48 hours. In comparison, another channel (314), producing 2372 reads for a total yield of 12.7 Mbps with an average read length of 5.34 kbps per read, also remained active throughout the sequencing run. All together, pore activity persistence is affected by initial pore activity and consistency in activity. A drop in channel activity leads to earlier inactivation with pores generating initial longer reads slightly correlated to lower overall read output, but not sequence yield.

MinION nanopore reads can produce long, but error prone reads

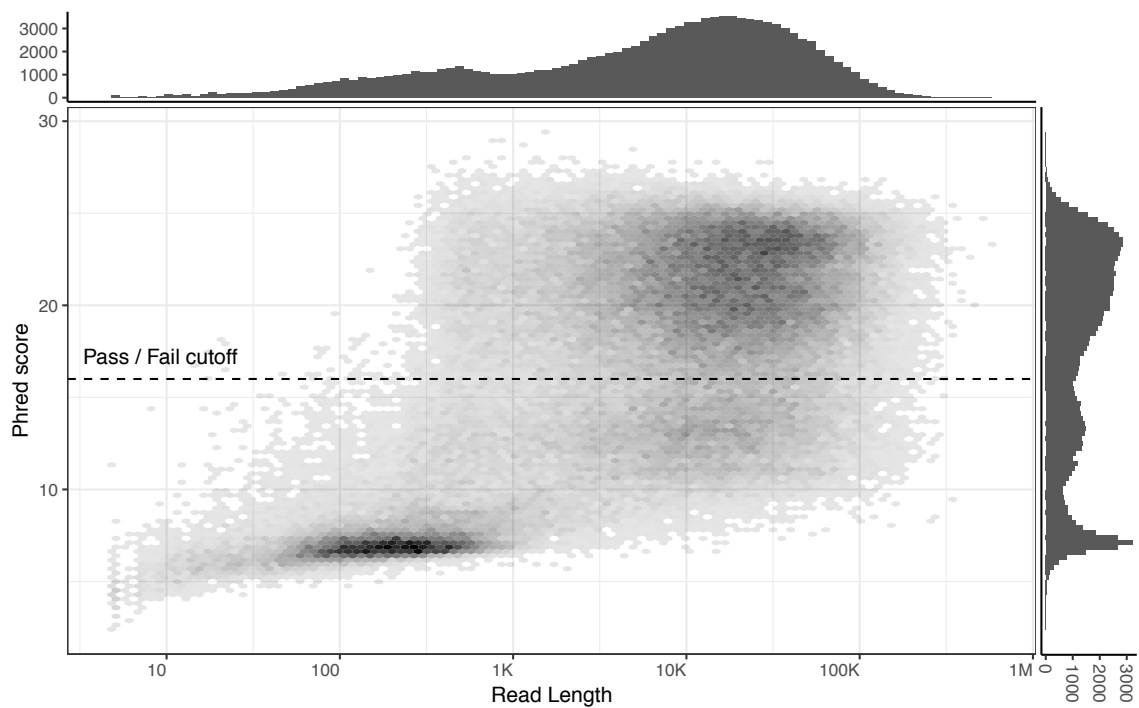
In standardized long read sequencing workflows, controlled shearing of DNA into large fragments is a common prior step to adapter ligation for selective and uniform sequencing of long reads [255]. With this study's interest in unrestricted read lengths, an alternate, specialized workflow using transposase-mediated adapter ligation was selected for both sequencing experiments (e.g. the Rapid sequencing kit). Without controlled DNA shearing, the sequencing yields for bacteria were highly variable in quantity, quality, and length of the reads sequenced (Figure 2.4). Although the Pss508 sequencing run generated 0.7Gb more total yield, the sequences were concentrated in ~1.90x fewer reads than the PsyTLP2 with a higher variability in the range of read length (Table 2.1). Although for both sequencing runs, read lengths ranged from 5bp, the smallest recordable length, to greater than 100,000bp, the Pss508 sequencing run on average generated reads with nearly twice the

length of PsyTLP2. Pss508 reads had a mean and median length of 20379bp and 9304bp whereas PsyTLP2 reads were 7990 and 4553. Furthermore, for Pss508, the longest read passing the average quality threshold reached a remarkable size of 569,475bp whereas the longest read for PsyTLP2 was still quite long but notably smaller at 267,384 bp (Table 2.1). Although reads exceeding 600 kbps were recorded for Pss508, these reads did not pass the pass/fail quality threshold (Phred score of ~16). Despite the higher range, Pss508 had a broader bimodal distribution with a heavy left skew for lower Phred scores (Figure 2.4). Comparatively, PsyTLP2 reads were more concentrated in a narrower, more uniform distribution with fewer outliers contributing to long tails at both extremes.

To further contrast the two datasets, the higher variation in the Pss508 reads is mirrored in the average read qualities with Pss508 read sequences forming a trimodal distribution compared to the left skewed distribution as described by PsyTLP2 reads. Pss508 reads are typically of a slightly lower quality with a mean Phred score of 17.1 (median = 18.6) compared to PsyTLP2's mean score of 20.0 (median = 20.8)(Table 2.1). This lower mean read quality in Pss508 is attributed in part to a large abundance of short reads, primarily less than 1kbp in size, with a quality score less than 10 accounting for 21.7% of all reads. The higher concentration of these smaller reads in Pss508, when compared to PsyTLP2, may be linked to poor diffusion of large DNA fragments where leftmost channels primarily encounter shorter DNA fragments (Figure 2.3). This concentration of reads does not appear in PsyTLP2 due to additional DNA shearing creating a less viscous solution, allowing for greater diffusion of all DNA fragments in the flowcell.

There is also a weak correlation between lower quality and short read length can be seen in reads <316bp ($\log_{10} = 2.5$) however this trend falls apart with longer reads (Figure 2.4); the majority of reads between 316 to 1000bp were below cutoff threshold in Pss508 compared to the higher quality reads of PsyTLP2 in the same size range. Reads greater than 316bp in length encompassed the full read quality range in both sequencing runs indicating little impact on sequence quality during prolonged read sequencing, unlike the reduced read quality during read extension of synthesis-based sequencing methods [211,257,259]. To verify this observation, the quality of the first and last 500 bp of each read (excluding reads smaller than 500bp in size) was analysed for any potential decline in sequence quality. The sequence quality was consistent, with Phred scores typically ranging between 14-16 throughout the reads except the initial and final 2 bp, which displayed only a slightly greater range in read quality (not shown). In summary, the Pss508 had higher variation in both sequence

A. Pss508



B. PsyTLP2

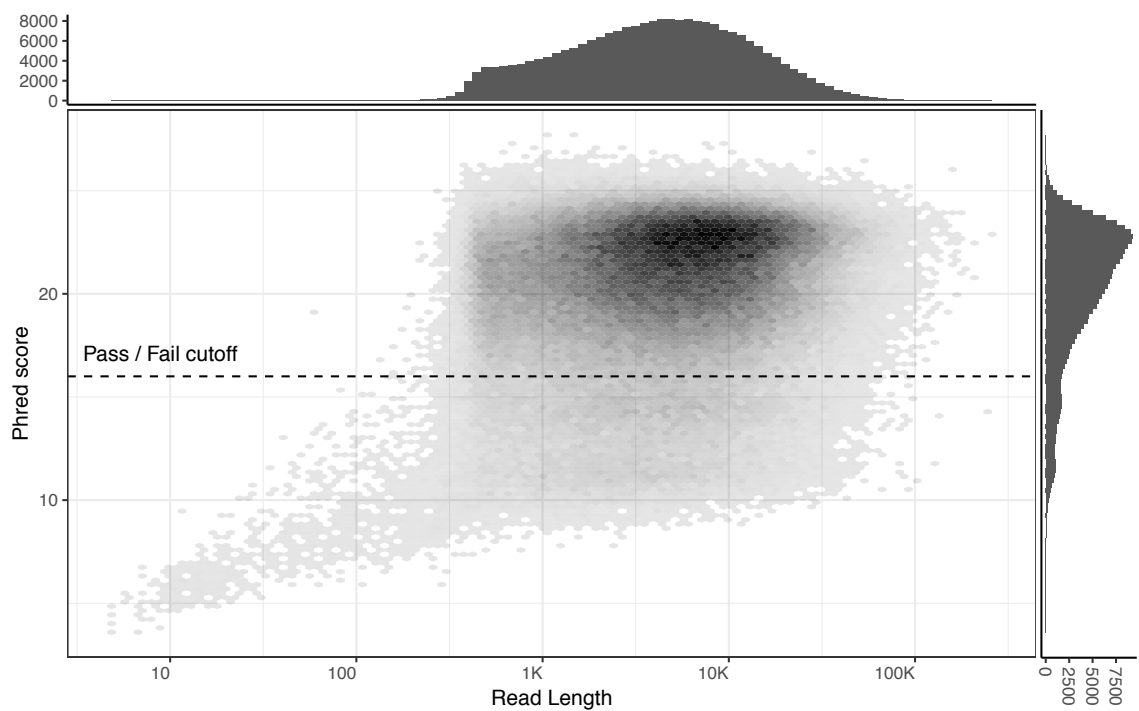


Figure 2.4 Average quality of reads vs read length. Each point ($\alpha=0.2$) represents an individual read with darker regions indicating higher concentrations of reads sequenced. Phred quality scores were called from the native quality scores provided by the Albacore Basecaller. The horizontal line marks the pass/fail threshold. Distributions for read length and read quality are at the top and right of each main plot respectively. (A) Pss508. (B) PsyTLP2.

length and quality whereas reads from PsyTLP2 were typically more uniform. With uncontrolled fragment size and unrestricted input DNA, sequencing runs can be uneven between two runs using similar workflows and protocols indicating heavy influence of experimental conditions [231]. In our case of high input DNA and total yield, we mitigated the lower quality of reads in Pss508 from run variability by increasing the quality threshold to isolate the top 25% of reads. Third quartile Phred score thresholds of 22.1 and 22.6 were set, yielding 30470 and 57853 reads for Pss508 and PsyTLP2 respectively. Due to high yield, even this restrictive threshold provides a high 147x and 97x theoretical coverage for each respective genome.

2.3.2 Bacterial genome assembly using nanopore long reads

Long read sequences assemble full, but functionally incomplete genomes

Genome assembly of each bacteria was attempted using three of available generalized long read assemblers: Canu [240], SMARTdenovo [271], and Minimap-Miniasm-Racon (mmr), a multi-software OLC pipeline [247,248,265]. Unlike more mature sequencing system, available software specific to nanopore sequences were limited [272]. Since the assembly of both Pss508 and PsyTLP2, improvements to nanopore sequencing has lead to the development of more robust, advanced assemblers, such as Flye2 and wtgdb, [239,273] with improved assembly capabilities. Comparison of assembler performance was carried out on the Pss508 strain. Out of the three assemblers, only Canu was capable of assembling a singular, full contig of 5.99Mb using default parameters. Despite a high abundance of sequence information, both SMARTdenovo and mmr both produced discontinuous assemblies comprised of 4 and 12 contigs respectively. These assemblies were improved upon increasing stringency and optimization of assembly parameters, particularly those pertaining to read overlap. For SMARTdenovo, pre-correction of the long reads with an untested read corrector, daccord [274], was necessary for complete single-contig assembly. Optimization of the Miniasm parameters, primarily of pre-selection parameters for read overlaps, and ONT-specific Racon

parameters for the mmr pipeline similarly yielded a singular contig. With optimization, both SMARTdenovo and the mmr pipeline were capable of assembling a complete genome within the expected genomic range (5.94Mb and 5.83Mb, respectively).

All three assemblies were complete using genome-level metrics. However, at the nucleotide level the sequence composition of k-mers differed from expectations when compared to a reference strain (*P. syringae* B728a)[158] and a similar strain (*P. syringae* 642)[18], both belonging to the same phylogroup and sub-phylogroup respectively [44]. Ideally, kmer composition of the assemblies should be highly correlated ($R=0.99+$) between closely related strains with only minor strain-specific differences (Figure 2.5a). However, long-read assemblies can significantly differ in kmer composition ($R<0.95$) as many outlying kmers involving homopolymers may be present, a known issue in long read sequencing (Figure 2.5b)[229]. Errors in these assemblies became apparent during gene annotation, with the number of genes predicted far exceeding expectations, and many genes appearing truncated [275]. Genes, fragmented by insertion/deletion errors, are likely detected as multiple genes [238].

Assembly completeness was assessed using BUSCO, a software that detects the completeness of known single universal orthologs [266]. Only one of the assemblies, the mmr assembly, contained more than 25% of the expected 452 orthologs commonly found in Gammaproteobacteria; the other two assemblies had fewer than 15% of the orthologs expected to be present (Table 2.2)[276]. The majority of single copy orthologs (>200) are fragmented with more than 100 of these universal orthologs not found. The best assembly (mmr) was further polished using Nanopolish, a signal-based HMM that utilizes the raw signal of each read mapped to the assembly to correct the sequence [229]. Post-Nanopolish, assembly completeness significantly improves from 115 orthologs (25.4%) to 322 orthologs (71.2%). Although a second Nanopolish run did minimally improve the assembly (not shown), the use of the Neural Network polishing software Medaka, developed by ONT [249], further improved the assembly from 322 detected orthologs to 385 orthologs (85.2% completeness). Even after two full rounds of polishing, both Pss508 and PsyTLP2's error-prone assemblies (Table 2.3) are incomplete at the sequence level, remaining in a gene-fragmented state with pseudogenes accounting for more than 10% of genes. In Chapter 3 I present a reference-based method for error correction of long-read assemblies to address the issue of gene fragmentation.

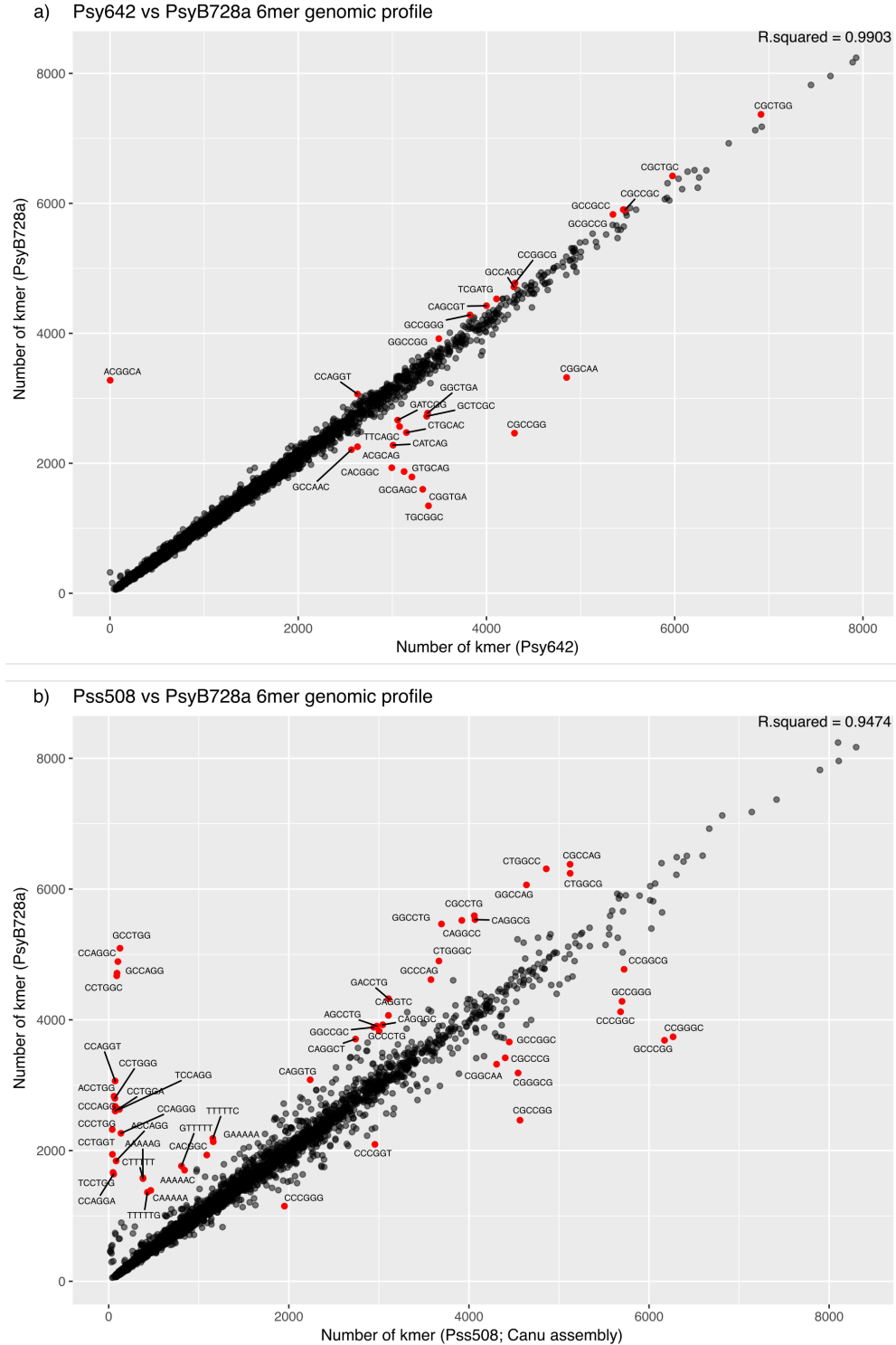


Figure 2.5 Relative 6mer composition between a draft assembly and a reference assembly (PsyB728a). (A) Short read assembly (Ps642) (B) Long read assembly (Pss508).

Table 2.2 Assessment of draft assembly quality via detection of known 452 single universal orthologs found in Gammaproteobacteria using BUSCO.

Assembler ¹	Polisher	Universal Orthologs Found			
		Complete	Duplicated	Frameshifted	Incomplete
Canu	N/A	58	0	220	174
SMARTdenovo ²	N/A	44	0	215	193
Mmr pipeline	N/A	115	0	234	103
Post-polishing					
Mmr pipeline	Nanopolish	322	0	102	28
Mmr pipeline	Nanopolish + Medaka	385	0	54	13

¹ – Long read assemblers capable of generating a single contig genome assembly for Pss508. mmr pipeline represents an OLC assembly pipeline of software Minimap, Miniasm, and Racon using a custom script.

² - SMARTdenovo required pre-correcting of reads using daccord. Reads for other assemblers were not corrected prior to assembly.

Table 2.3 Draft assembly statistics generated using an optimized Minimap-Miniasm-Racon pipeline, and post-assembly correction with Nanopolish and Medaka.

	<i>P. syringae</i> pv. <i>syringae</i> 508	<i>P. syringae</i> TLP2
Size (Mb)	5.84	5.71
Contigs	1	1
Plasmids	No	No
GC content	59.22	59.26
Reads Used	62622	199549
Coverage	328	293

2.4 Discussion

Nanopore sequencing removes barriers to sequencing for smaller research projects by providing accessible low cost, quick, and efficient system

Oxford Nanopore Technologies' developed a highly compact, portable, and efficient sequencing system that can be used for all forms of sequencing projects ranging from smaller projects, e.g. bacterial sequencing [223], to supplementary sequencing for larger projects [224]. In this work, we tested whether nanopore sequencing could provide the data required to fully assemble two small bacterial genomes without additional sequencing methods. The entire MinION nanopore system was highly compact and space efficient easily fitting within a small lab while providing ease of access compared to the bulkier Illumina and PacBio machines [222,259]. Quick access allows for reduced downtime for data acquisition without relying on coordination with sequencing centers/departments where turnaround time can range from 4-8 weeks [277].

Unlike sequencing-by-synthesis based technology, the lack of DNA cycling and amplification allows for nanopore sequencing components to be packaged as a disposable flowcell, eliminating much of the complex machinery of larger systems which require skilled maintenance [216]. The necessary sequencing reagents are combined with input DNA as a single process. Beyond the initial computer setup, MinION nanopore users require only basic lab skills for sequencing, improving the ease of use over other sequencing methods. Resource efficiency can be further increased if a non-standard workflow such as the rapid library preparation method is implemented. The low DNA input and unrestricted read length of this preparation method (RAD002/004) allowed us to both simplify the process and bypass the use of third-party single-use consumables [255]. When samples contain non-robust cells, many of the reagents, from DNA isolation to library preparation, are both more easily accessible and commonly found in wet labs using basic molecular techniques [231].

Each *P. syringae* genome was sequenced using a single MinION R9.4 flowcell and produced between 1.83 and 2.53 Gb of sequence data (Table 2.1). Despite a high yield per genome, the output was 6x lower than the potential output of a Illumina MiSeq run, at a comparable price point [270]. Similarly, 121k to 231k reads generated using MinION flowcells were at least 60x fewer than the theoretical 8 million reads (4-5 million read expected in practice) that can be generated with a similarly priced SMRT chip [278]. The lower yield from MinION is sufficient for smaller organism genomes such as those of bacteria and single cellular eukaryotes. however the data may not be

sufficient on its own for multi-cellular organisms [224]. A yield of 2.53 Gb can provide 30x or 60x coverage for an 84Mbps or 42Mbps genome, respectively. When poor quality reads are discarded (below standard quality cutoff), this is reduced to 65.8Mbps or 32.9Mbps genomes respectively. For fungal genomes, this total yield can provide 60x coverage from *Aspergillus nidulans* [279], and ~50x for the slightly large *Neurospora crassa* genome [280]. However, less than 20x coverage would be obtained for the larger eukaryote, *Caenorhabditis elegans* [281], and a single flowcell is not feasible for larger eukaryotes [255].

Read lengths however can compensate for low yield MinION read lengths average 23-33x longer than Illumina reads, allowing assembly of more contiguous assemblies (Figure 2.4). MinION read lengths are on average comparable to either PacBio's CCS reads or CLR reads, depending on the library preparation method used [211]. However, 26.5% and 5.9% of reads from the MinION runs were longer than the 25kb length limit of PacBio's CLR synthesis method [259]. In fact, the longest reads, reaching up to 267-569kb in length, can encompass ~4.5-9.8% of the full *P. syringae* genome length, reducing the yield required when a large portion of the assembly is in a single read. For viral genomes these reads can cover 22-47% or more of the genome [282]. However, at a lower genomic coverage, the quality of each read has a greater impact. As previously reported, the MinION reads are more error prone without any reads exceeding a Phred score of 30 (Table 2.1, Figure 2.4), which is the quality baseline for Illumina SBS sequencing. The average mean Phred score quality of 17 (Pss508) and 20 (PsyTLP2) for nanopore sequencing indicates approximately one to two erroneous bases per 100 bp. Since the nanopore sequencing error profile indicates systematic issues during basecalling of homopolymers, these reads may have localized error rates of up to 13%, comparable to PacBio's CLR read [202,242].

For both sequenced bacterial genomes, the systematic errors of these reads hinder the quality of the full assembly with fragmented genes below acceptable levels even with multiple rounds of correction (Table 2.2)[238]. Although currently sequencing groups can circumvent quality issues with hybrid assembly approaches via additional sequencing with high quality short reads, similar to the release of PacBio's CCS technology, quality of nanopore sequencing will continue to improve with maturation of nanopore technology and software [232,283]. Overall, at a comparable price to PacBio/Illumina, we were able to generate enough sequence data capable of assembling a full genome for each bacterium within 72 hours of receiving the MinION system (with 48 hours of sequencing). For organisms with small genomes, the MinION can produce an overabundance of sequence data,

research labs can improve yield efficiency with fragment barcoding or the more recently released smaller flowcells [284]. Even with a comparatively lower yield to more established sequencing methods, the MinION can be an attractive option for a low maintenance, time and resource efficient sequencing system for smaller sequencing projects.

Nanopore sequencing's capability of generating long reads is physically limited by the chosen library preparation workflow, and the properties of the system.

Both MinION nanopore sequencing runs generated sequence yield significantly lower than its expected output and in comparison to other sequencing systems. The reported maximum output for a flowcell is 50Gb of sequence data, yet both runs yielded less than < 2.6Gb accounting for less than 5.2% of total potential output. In comparison, PacBio outputs typically 4-5M reads out of 8M potential reads (50-63% of output)[278] whereas an expected MiSeq range from 7.5 to 8.5Gb out of a maximum 15Gb (50-57%)[270]. This yield difference may be due to the fundamental physicochemical differences between the methods. Unlike the other two technologies, nanopore sequencing was not consistent across different pores, with a few channels generating the highest sequence output (Figure 2.3a). Even at the start of the run, 76 channels (304 pores) are initially inactive, with only 800 pores per flowcell being guaranteed by ONT [263]. These inactive nanopores can severely limit the sequence yield when the highest amount of sequence data is generated at the start of run with nanopore activity decreasing as the run progresses (Figure 2.3b). In fact the 2.07Gb (81.8%) of the total yield was sequenced in the first 24 hours of the Pss508 run. Of the 436 channels initially active for the Pss508 sequencing run, only 210 (48.1%) of them provided more than 2 Mb of data in the first 8 hours. In addition to the number of active pores gradually inactivate with use, persistently active pores showed significant decrease in activity over time (Figure 2.3b). After 24 hours, only 259 (50.5%) channels were active with only 93 (36%) of these sequencing greater than 2Mb. At 48 hours, none of the 109 active channels could produce over 2 Mb. Every 8 hour intervals, output was temporarily boosted by flow reversal for all pores to likely remove clogged nanopores [262]. Although active pores were temporarily increased to 330 channels at 8 hours, only 34 (10%) channels had a 2Mb output suggesting had only a minor impact in recovering the nanopores.

Nanopore channels that remain active throughout the run typically have the high read activity but not necessarily highest output (Figure 2.3a). Conversely, other channels producing longer reads

but an equivalent total output become inactive earlier suggesting that the fluid dynamics of the interaction of DNA fragments with nanopores is linked to persistence of nanopore activity. Unlike synthesis-based methods that utilize DNA fragment anchoring for consistency but limit read length, ONT's nanopore is reliant on stochastic interaction of the adapter attached motor proteins to bind with nanopore, allowing feed-through of the DNA fragment [220]. With sample input localized to one side of Asics chip, sufficient diffusion for uniform distribution is paramount for sequencing. However, in this work, the choice of the length unrestricted rapid sequencing kit and a tailored workflow towards larger DNA fragments resulted in a viscous, high mass but low molarity sample. Poor diffusion of this sample is suggested with channels closer to input side (Figure 2.3a, right) sequencing on average longer reads than farther pores. Consequently, these closer pores lose activity earlier suggesting that larger DNA fragments can reduce motor protein interaction with a nanopore hindering sequencing output. Larger DNA fragments have a higher potential to form knots leading to clogged nanopores reducing available nanopores for sequencing [262]. Furthermore, the low molarity of motor proteins compared to DNA mass can severely limit the probability of interactions with the nanopore; a 500kb fragment has two points of interaction (i.e., motor proteins) whereas 25 20kb fragments, with a similar mass, would have 50 potential points of interaction. Once inactive, channels tend not to recover suggesting that without sequencing activity nanopores may fully inactivate.

Taken together, the choice of library preparation to maximize read length may have reduced the sequencing potential of both MinION flowcells due to the physiochemical dynamics of the sequencing system. For better yield output and lower read variability, the standardized workflow, ONT's ligation sequencing kit, with controlled shearing is advisable [285]. Even with shorter long reads, full assembly can be completed with high yield [212]. However, the rapid sequencing method is serviceable for small projects when lower yield is sustainable. Total output can be maximized by skipping the first Lambda control run if needed. Even with multiple flushes, the lambda DNA fragments were found to remain in the flowcell and contaminate the run needing post-processing removal (not shown). Unlike Illumina sequencing, use of additional DNA (e.g., PhiX) is not necessary and omitting it may reduce contamination [210]. For smaller projects the resource-efficient rapid sequencing kit can provide an overabundance of data, but this could be limiting for larger projects. If output is a concern, the standard workflow can provide an output similar to PacBio, but more time-efficient, at a higher cost.

Available Software limits the quality of long read assemblies generated from early nanopore data

Although software is available for the ONT system, much of it has not been thoroughly tested and is somewhat limited, which may be attributed to its recent development. Even with an official basecaller (Albacore), there were many other unofficial basecallers (including the developmental version of Guppy, the most recent official basecaller that was unreleased at the time of this study) [232]. Long read correction software not targeted to PacBio were relatively rare, and only three generalized assemblers could handle the incredibly long but error-prone reads [240,247,248,265,271]. Although Canu, an assembler for PacBio reads but also suitable for ONT reads, has default parameters, it can be restrictive as specific genome size is needed which may introduce biases. *P. syringae* bacteria can have genome sizes ranging from 5.7 to >6 Mb [18,158]; specifying a size produced an assembly closest to that value; for Pss508, Canu produced a 5.99Mb draft assembly just short of the expected 6Mb. Although this is unlikely to be an issue for well-known genomes, it may produce lower quality results for *de novo* assemblies if the target genome size is unknown. In this study, higher quality assemblies required chaining of multiple different programs using non-standard parameters, which can be a challenge for unfamiliar users.

Without additional sequencing, assembly correction was limited to Nanopolish which takes advantage of raw signal information to direct correction via mapped reads [229]. However, this software alone was unable to improve assembly accuracy to an acceptable complete state (Table 2.2). Even with additional polishing with Medaka, we were only able to assemble a full, but still highly gene-fragmented genome from available software (Table 2.3). However, with technology maturation, improved software will be continue to be released and push the boundaries of sequence accuracy, targeting the weaknesses of this sequencing technology [241]. With future developments, ONT's nanopore sequencing system can evolve to become an attractive and efficient alternative for all types of sequencing applications.

Chapter 3:

Kastor: A reference-based assessment of fragmentation errors in long-read assemblies

3.1 Introduction

3.1.1 Chapter objective and overview

Despite multiple rounds of polishing, genomic assemblies generated from long read sequences remain error-prone. Insertion and deletion errors are common causes of gene-fragmentations within the assembly. We hypothesize that errors can be detected and corrected using reference information. In this study we present Kastor, an assembly error assessment software tool focused on detecting and correcting indel errors. In addition to genome correction, this program may be used for comparative assessment of multiple long-read assembly pipelines. Kastor compares the final assembled genome with a set of locally aligned reference genomes to identify potential assembly errors. The frequency of alignment differences between the reference genomes and the draft assembly is used to identify consistent differences, which are then marked as candidate errors. These candidates are adjusted or removed, conditional on support from raw read data. Unless specified otherwise, Kastor uses the finalized detected errors as targets for correction and assembly defragmentation to produce a higher quality and more complete assembly. To assess and validate error identification with Kastor, two *Pseudomonas syringae* genomes were assembled and polished using minION sequence data only. Here, Kastor identified potential assembly errors and defragmented genes within the two MinION assemblies and a reference Illumina *Pseudomonas syringae* draft assembly. After correcting the detected errors, we were able to improve BUSCO-assessed assembly completeness from 85.2% to 99.8% and reduce the number of annotated pseudogenes from 23.3% to 5.6%, comparable to the number of pseudogenes in a corresponding Illumina assembly (4.4%). Thus with Kastor correction, we were able to complete two high quality nanopore-only assemblies of bacterial genomes.

3.1.2 Chapter background

The number of microbial genome sequences has increased exponentially in response to the decreasing cost and maturation of sequencing technology. The increased availability of sequencing

has been invaluable to the research community and has expanded to include uses such as tracking of bacterial outbreaks [286] and providing insight into micro-evolution processes [43]. However, sequencing technologies are not without challenges [238]. The most common sequencing technology, Illumina, provides high yield and high quality (<1% error rate) short paired-end reads ranging from 75-300 bp [202]. Although highly accurate, assemblies generated using solely these short fragments suffer from relatively high discontinuity. These assemblies often have poorly resolved repetitive regions leading to difficulties in ordering and joining of sequences flanking these regions [251]. This often results in fragmented draft assemblies that can hinder downstream studies, such as those involving repetitive sequences, structural rearrangements, transpositions and insertion sequences [43,204,205].

In part to alleviate these concerns, long read sequencing technologies, such as PacBio's single molecule read-time (SMRT) and Oxford Nanopore Technology's nanopore sequencers, have been developed and continue to be refined. Depending on the long-read sequencing platform, library preparation, and sequencing chemistry used, generated reads can on average be greater than 10 kilobases and in some cases up to 2M base pairs [230,259]. These longer reads can encompass full-length repeat regions allowing for resolution of both complete circular bacterial genomes [229,259] and larger eukaryotic chromosomes [287].

One disadvantage to long read sequencing is that the current methodology is more error prone than its short-read counterparts [205,242,288]. PacBio's SMRT sequencing initially had a single pass error rate of ~13% (RSII with P6-C4 chemistry), but this error rate has been reduced to <1% using circular consensus sequencing (CCS) [211]. This improvement utilizes higher sequencing of sub-reads and calls a sequence consensus per read, compensating for random errors in each individual sequencing pass of the read. ONT's MinION sequencer has a similar error rate, where 1D reads sequenced on a minION R9.4 flowcell had an average error rate of up to 15% [216]. Recent improvements to both chemistry and software has reduced the consensus error rate to below 5% [255] and a claimed consensus quality comparable to PacBio's CCS with its recent iteration of the sequencer (R10) [289]. Despite improvements in sequencing quality for both systems, effective sequence correction is still needed [237,287]. Errors in long read sequences are predominantly short insertions and deletions associated with homopolymers [242,288,289]. In fact, homopolymer indels account for 92% of errors in PacBio CCS reads [211]. For nanopore sequencing, sequence errors have been speculated to be inherent to the base-calling method, resulting from k-mer miscalls due to

software limitations during signal interpretation, particularly for low complexity regions [236,237,289]. These base-calling errors may be reduced as the technology continues to mature. In response to these errors in long-read sequencing, many error correction algorithms have been developed and can come pre-packaged with assemblers such as wtdbg2 [239] and Canu [240]. To remove final errors, it is common practice to polish the assembly with correction software such as Arrow [290] and Nanopolish [229].

Upon assembly, long reads are capable of resolving genomes into single, contiguous sequences. Despite multiple rounds of correction, these assemblies can still fall short in per-base quality compared to Illumina assemblies [223]. To minimize cost and maintain high quality complete genomes, many researchers are opting for a hybrid approach incorporating both long and short reads [205,291]. Generally, these hybrid assemblies are generated either using supplemental long reads to scaffold shorter, high quality contigs together [251,292], or a long-read assembly is assembled [212,240] then polished using short reads for improved sequence quality [253]. Either strategy is sufficient in assembling single, contiguous genomes with accuracy greater than 99.99% [287].

Unfortunately, an assembly with high sequence quality may still contain errors that impede accurate prediction of protein coding regions in long-read based assemblies [238]. Indel errors in long-read assemblies can pseudogenize full genes by introducing frameshifts and premature stop codons [287]. This fragmentation can lead to erroneous interpretations in downstream analyses [205]. As an example, an analysis by Watson and Warr [238] revealed that at least 2746 genes contained indel errors in a 99.8% identity hybrid human assembly [255], and even greater amount of indels were found in a PacBio-only (RSII) assembly [293]. Currently there are difficulties in assessing these inaccuracies without additional sequencing of higher quality short reads, since these errors account for a low percentage of the assembly sequence. Traditional assembly evaluation measures such as N50, GC content, L90, and percent identity to homologs do not directly assess gene fragmentation. The benchmarking software BUSCO is one of the few tools available to estimate gene completeness of an assembly by detecting the presence and status of universal single orthologs within a target assembly [266]. However, its focus on orthologs and lack of customizability may lead to potential overestimation of assembly completeness.

3.2 Material and Methods

Design Overview

Kastor is a reference-based correction software written in Python and is separated into three main modules: reference-based detection of candidate errors, error adjustment and confirmation using raw read information, and assembly correction (Figure 3.1). The first module is designed to compare the draft assembly with locally aligned reference genomes to identify candidate sites for assembly errors. Candidate sites are then compared with aligned raw read data to adjust or confirm previously detected error candidates with sufficient read support. Read support is used to minimize reference bias and retain strain variability, within range of the candidate errors. For indel errors, nearby read-supported error equivalents are favoured over the initial candidate and the detected error site is adjusted accordingly. The third module uses both the error information and suggested corrections to correct the corresponding draft assembly.

Bacterial strains, reference genome selection, and pre-processing

Two strains of plant-associated *Pseudomonas syringae* strains, *Pseudomonas syringae* 508 (Pss508) and *Pseudomonas syringae* TLP2 (PsyTLP2), were sequenced using ONT's MinION sequencing technology [18]. Both genomes were assembled to their pre-correction assembly using a overlap-layout-consensus pipeline comprised of Minimap2, Miniasm and Racon, and then polished with Nanopolish and Medaka [229,247–249,265]. The previously published genome for the closely related strain *Pseudomonas syringae* 642 (Psy642) was also used as an Illumina assembly for comparison [18] and is expected to have relatively few indel errors.

Three different sets of *P. syringae* (Psy) genomes with varying diversity and number of genomes were retrieved from the NCBI genome database and used for reference-based assembly error correction. Set one is a high-quality collection of 11 closely related Psy strains with an assembly quality classified as scaffold or higher, with the exception of Psy642. The strains for this primary set were selected to have approximately equal phylogenetic distances to the target strain and each other based on both gapped (94 - 98%) and symmetrical (78 - 82%) identity as provided by NCBI (Table S3). To minimize bias towards any particular taxonomic groups, groups with higher shared gapped identity (~98%) were limited to three strains. Genome sets two and three include a greater variety of

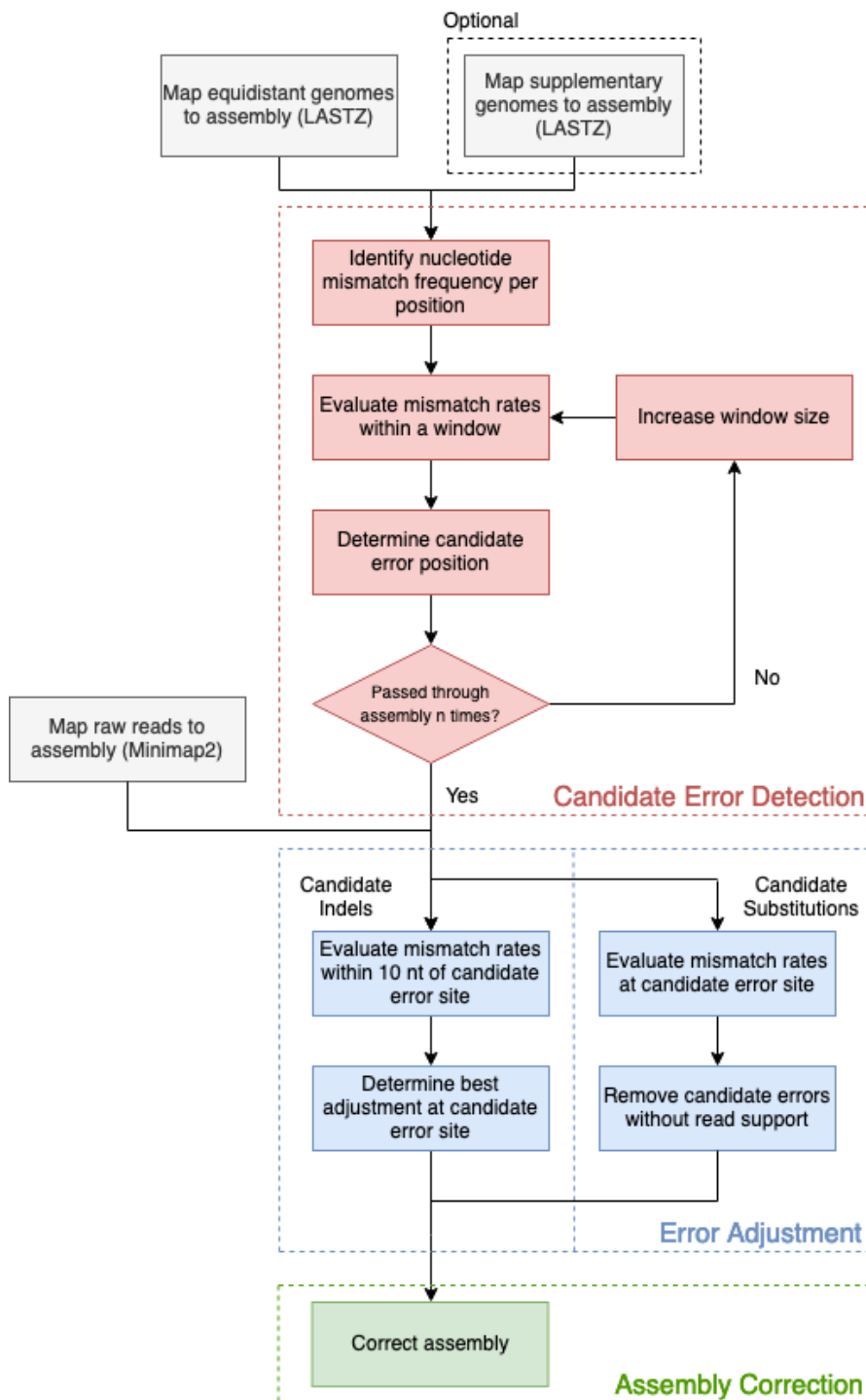


Figure 3.1 Kastor error assessment and correction workflow. Each module in Kastor is highlighted as a separate colour: Red – Error detection, Blue – Error adjustment, Green –Assembly Correction. External steps in the workflow are indicated in grey.

draft genomes. Set two includes 170 Psy strains (*sensu lato*) that belong to phylogroup 2, the closest related monophyletic group [44]; set three includes a selection of 441 Psy genomes predominately classified as genomespecies 1 in NCBI and select strains from most of the 13 broad monophyletic groups.

Each reference genome set and the target strain raw reads were aligned to the target assembly using LASTZ [294], with a minimum identity of 80%, no chaining and no interpolation. Poor alignments and short alignments (<1000 nt) were filtered to minimize spurious reference mapping. The full alignment was then converted to an mpileup file (Samtools version 1.2) using the `-aa` flag to ensure that all zero coverage positions are reported within the file to maintain consistency between the alignment files.

3.2.1 Software Implementation

Module One: Candidate error detection

For each contig, the software cycles through each position and assesses whether there is sufficient reference genome coverage (≥ 5 genomes) to infer a candidate error site. At sites with sufficient depth, Kastor calculates the frequencies of substitutions and indels in reference genomes as compared to the target genome, which are then used to calculate an alignment difference probability (ADP) for each error type. ADP is calculated as the proportion of reference genomes that have a difference with the draft assembly at the current site. The ADP for substitution errors is based on the frequency of nucleotide mismatches. For insertion errors, Kastor calculates the proportion of reference genomes missing sequence(s) at the position (i.e., a “deletion” in the reference genome). For deletion errors, the inverse is considered. The software uses this ADP for the initial identification of potential sites for error correction. When the probability of a putative error equals or exceeds the set error threshold, the position is identified as a candidate error site. For example, if a G at position X in the draft assembly aligns with a highly conserved A in all the reference genomes, the ADP is calculated as the proportion of reference nucleotides at position X that are not G. In the case where the ADP exceeds the error threshold, the substitution $G \rightarrow A$ at position X is classified as a candidate error.

The ADP is calculated and evaluated independently for substitutions, deletions and insertions. As longer indels are likely more reflective of a genome-specific variant of the sequence rather than a candidate assembly error, candidate insertion and deletion errors greater than two nucleotides are ignored. To avoid collisions with longer indels, the ADP is calculated only at the starting position of the alignment indel difference. For example, if some reference genomes are missing nucleotides AG at positions YZ in the draft assembly, an insertion ADP of < 1.0 is calculated for position Y only and considered as a single putative insertion error two nucleotides long.

The software then scans through saved alignment information with a minimum of two and a default of four elective iterations. These iterations are intended to disambiguate highly variable regions containing multiple candidate errors in close proximity. The software attempts to reduce the complexity of the region by first flagging simpler cases in earlier passes then disambiguating additional cases in subsequent iterations (Figure 3.2). This is achieved by varying the window size in each iteration. The first iteration has a window size of one and in the second iteration the window size varies with homopolymer size. The third pass (if used) has a starting window size of 11, where five nucleotides flank each side of the candidate site. This window increases by five nucleotides each side for subsequent iterations.

During each pass, Kastor evaluates the ADP for all error types at each position and retains the greatest ADP. In the infrequent case of a combined substitution error and deletion error, this would be parsed as two adjacent errors in the first and second mandatory pass. If the ADP exceeds a threshold of 0.25 (i.e., 25% of aligned reference genomes contained alignment differences), the position is considered as a candidate error location. The software then calculates a regional ADP for the set window size. During this stage, Kastor keeps track of which sequence contributed to the regional ADP rate to ensure that multiple differences on a single aligned sequence cannot contribute to the regional rate.

To maximize error finding, Kastor also employs two additional search strategies that are limited to elective passes at every second pass. The first strategy shifts the search window left and right while maintaining window size to facilitate detection in situations where the candidate site is not central to an erroneous region. The second strategy is the consideration of different error lengths for the same error type. The software defaults to considering different error lengths (i.e., indels of one and two) as independent errors to simplify early detection. With this search strategy, a putative two-

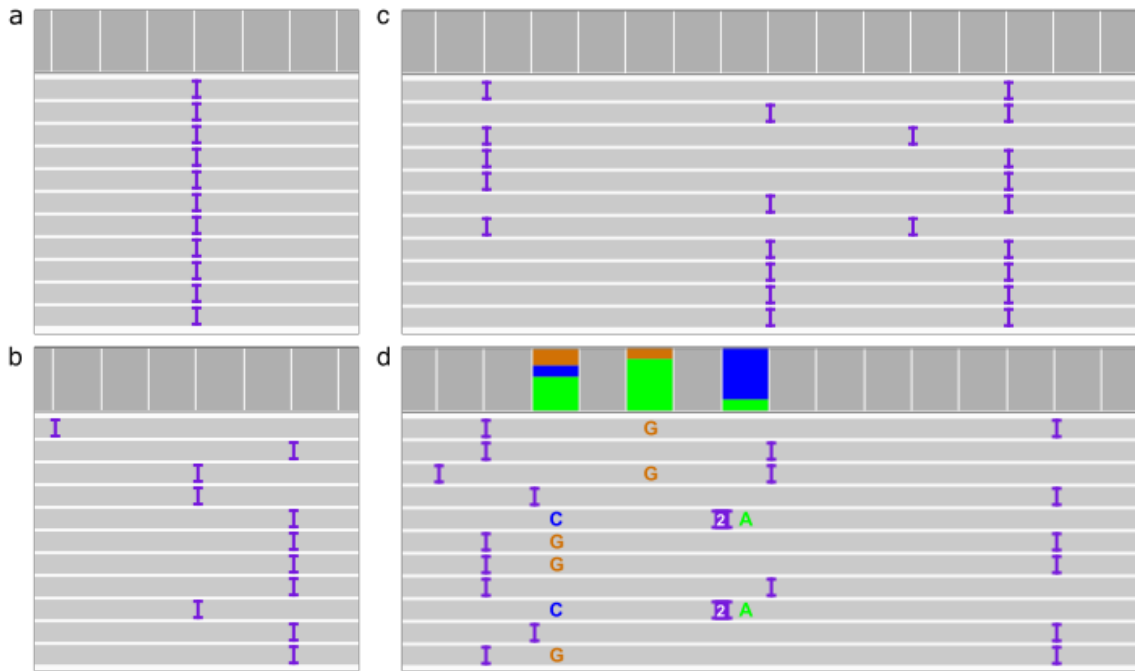


Figure 3.2 IGV visualized examples of draft assembly regions (Pss508) containing deletion error candidates. The topmost row represents single positions (grey/coloured blocks) in the draft assembly. Each subsequent grey row represents an aligned reference genome. Visible nucleotides are mismatches while purple indicators represent unaligned additional nucleotides found in the reference genome. a) Simple case. Single position with an adjacent deletion error candidate. b) Single deletion error with variable candidate sites. c) A region with two deletion error candidates in close proximity. Smaller windows in earlier passes disambiguate the region as two simpler error candidates side by side. d) Ambiguous region with either two single nucleotide or one double nucleotide deletion error candidate.

nucleotide deletion can be treated as a single deletion. This is intended for cases where there is a disagreement in error length rather than a presence of a candidate error.

If the ADP exceeds the error threshold, the software then evaluates the region to identify potential erroneous nucleotide(s). Kastor scores each candidate location containing a reference alignment mismatch as a putative true state, where the score = (mismatch frequency x homopolymer length). If multiple locations belong to the same homopolymer and are of the same correction type, the sites are merged before score calculation. Since Kastor focuses on errors found in long read assemblies where many indel errors are associated with homopolymers, the score is weighted slightly

towards mismatches involving longer homopolymers [216]. However, if the alignment difference suggests a correction that would split a homopolymer, the score is penalized by half. The candidate position with the highest score is considered the candidate error site for correction.

Module Two: Error Adjustments

The second module aims to use target genome read data to support candidate error locations and minimize bias towards reference genomes (Figure 3.1). An alignment of the raw sequence data to the draft assembly is required. In this work, raw reads were mapped using LASTZ due to longer read length. For every putative indel error, Kastor sets a 21-nucleotide window with the candidate error position at the center. Similar to the first module, a read ADP of the respective error type is calculated for every position in the search window. Adjacent repeat nucleotides are grouped into a single representative candidate site to prevent distributing potential errors over multiple equivalent positions within a homopolymer. An exception to this is an error that introduces a break within a homopolymer, which is considered as a unique and independent position. Any position with a read ADP exceeding the user-defined adjustment error threshold is considered a read-supported candidate error site(s) for adjustment.

Kastor assesses all read-based candidate sites for support of the initial reference-based candidate error prior to any evaluation of alternative candidate error sites. We reason that with sufficient support for the initial error candidate no adjustment is needed and alternate read-based candidates are not evaluated. In cases where the initial candidate has no read support, Kastor considers the previously identified read-supported candidate error sites as possible alternate candidate locations for the detected error. The alternative with the greatest read ADP is selected to replace the initial reference-detected error candidate site. In the event that no candidates of the same length are found, multiple smaller error candidates may be considered for adjustment. More specifically, an indel candidate of 2 nucleotides may be adjusted to become two single indels if there is read data support. However, a lack of any or insufficient amounts of shorter supported candidates will result in no adjustments.

In contrast, substitutions are evaluated differently. From module one, unlike indels, all possible substitutions are considered and evaluated for read-support. For example, a candidate site with both adenine and guanine mismatches in the reference alignments will consider both nucleotide

as a possible correct state. For the same position, Kastor calculates a read ADP for all possible nucleotides and finalizes the candidate substitution correction with greatest ADP that exceeds the adjustment threshold as the likely true state. Unlike indel errors, this adjustment module searches for read support only for substitutions. A lack of support will result in the removal of the substitution as a detected error. Once all candidate error adjustments are made, the finalized list of errors are passed to the third module for assembly correction, unless specified as an output.

Module Three: Assembly Correction

The finalized list of errors is then used to correct the assembly and output a fasta file. If desired, users can manually inspect the error correction list as a tab-delimited output file. This error list may be used as input to reproduce the assembly correction without re-running the program.

Correction and Analysis

Completeness of the corrected assemblies was analysed using BUSCO, a software that benchmarks the status and copy number of single copy universal orthologs [266]. Gene fragmentation of the pre and post-correction assembly was analysed using NCBI's PGAP (build 4456) [295]. The RefSeq protein set used in the gene annotation was used to assess whether improvements are primarily in gene de-fragmentation or whether percent identity to reference genomes is also altered. NCBI's tblastn was used for all comparison of RefSeq annotated proteins to pre- and post-corrected assemblies. The analysis of pre- and post-corrected assembly relatedness to reference genomes was assessed using orthoANI [296]. All tables and plots were generated using R version 3.4.2 and Rstudio version 1.1.383 using packages dplyr (version 0.7.8), tidyr (version 0.7.2), ggplot2 (version 2.2.1) and ggpubr (version 0.1.8).

3.3 Results

3.3.1 Draft assembly improvements

Kastor improves long read assembly BUSCO completeness to ~99.8%

To assess the impact of Kastor on genome assemblies, we generated two long-read only draft assemblies for *Pseudomonas syringae* pv. *syringae* 508 (Pss508) and *Pseudomonas syringae* TLP2 (PsyTLP2), and used a published single short-read draft assembly for *Pseudomonas syringae* Psy642, for evaluation (Genbank Accession: NZ_ADGB000000000). As results were highly similar for both long-read genome assemblies, we focus on the results for Pss508, with PsyTLP2 results presented in supplementary material (Appendix Tables B.1-B.2 and Figures B.1-B.3). Prior to Kastor correction, each long read assembly was resolved to a single contig with a calculated size, ANI, and GC content similar to closely related strains [18]. The benchmarking software BUSCO [266] was used to assess draft assemblies for completeness by measuring the presence/absence of 452 intact universal single copy orthologs belonging to organisms of the Gammaproteobacteria class (Table 3.1). Despite multiple rounds of polishing, both draft genomes initially exhibited poor completeness. Here, 54 (11%) of BUSCO orthologs were considered fragmented while 13 (3%) of the orthologs were missing. A manual inspection of the 54 fragmented orthologs using tblastn and BUSCO's Gammaproteobacteria ortholog database revealed that 41 (76%) of these orthologs were in fact present but contained indels resulting in out-of-frame errors.

To address the presence of gene-fragmenting indel errors, we applied the Kastor software as an error assessment and correction tool which uses reference genome data to identify and correct these types of assembly errors. We curated a primary reference genome set (set one) containing 11 closely related *Pseudomonas syringae* genomes with 94 to 98% pairwise gapped identity, striking a balance between genetic closeness and sequence variability between all genomes, including the target genome (Table B.3). With this primary genome set, we applied Kastor to all three *Pseudomonas* draft genomes with an error threshold of 0.99. The selected parameters require all eleven aligned reference genomes to report the same alignment difference within the search window. This consistent difference in alignment is indicative of an error candidate (See Section 3.2.1). Using this threshold, Kastor was able to improve the long-read assemblies to 99.8% completeness, comparable with a short-read draft assembly of the closely related strain Psy642, which had a completeness of 99.3% before and after Kastor correction (Table 3.1). Here, missing orthologs in the Psy642 short read assembly may be due to gaps in genome coverage.

Table 3.1 BUSCO assessment of assembly completeness. A total of 452 single universal orthologs found in *Gammaproteobacteria* were determined to be complete, duplicated, fragmented or missing. BUSCO was used to monitor improvements to assembly completeness as assembly pipeline progresses with multiple rounds of polishing. Assembly completeness with a final Kastor correction is comparable with a related Illumina genome assembly (Psy642).

Assembly	Status of universal single orthologs			
	<i>Complete</i>	<i>Duplicated</i>	<i>Fragmented</i>	<i>Missing</i>
Pss508				
Unitig	0	0	54	398
Consensus called	115	0	234	103
Signal-HMM polished	322	0	102	28
NN polished	385	0	54	13
Kastor corrected	451	0	0	1
Psy642				
Draft	449	0	0	3
Kastor corrected	449	0	0	3

Kastor correction de-fragments genes

One caveat to BUSCO assessment is that the program can overestimate genome completion by only considering conserved and core orthologs within the selected ortholog database, neglecting the harder to resolve accessory genome, i.e., non-conserved genes shared between smaller subsets of the species. For a more comprehensive view, we investigated the completeness of the whole genome by annotating each assembly using NCBI's prokaryotic genome annotation pipeline (PGAP) to locate both complete genes and pseudogenes (i.e., frameshifted and incomplete genes) [295]. PGAP detected higher levels of gene-fragmentation in the pre-correction draft Pss508 assembly, with 1234 observed pseudogenes (23.3% of total genes), compared to BUSCO (14.8%). Of these, 1101 (89.2%)

of the PGAP annotated pseudogenes were considered frameshifted only, suggesting high levels gene-fragmenting indels when compared to the short-read assembly Psy642, which contained 25 frameshifted pseudogenes (Table 3.2).

Following Kastor correction all assemblies were reassessed with PGAP. Both long read assemblies showed considerable improvements (Table 3.2); for Pss508, Kastor increased the number of completed genes from 3967 (75.0%) to 4890 (92.7%). Tracking the genes between the pre- and post-correction assemblies reveals that 949 completed genes were added after Kastor correction and 20 genes were removed. The majority (909) of added genes were previously annotated as frameshifted, while a minority (30) were previously incomplete genes. Unexpectedly, 10 new complete genes were also found. Of the 20 completed genes removed from the pool, 16 were no longer predicted in the assembly while three became frameshifted and one became incomplete. Overall, there is a large decrease in the number of frameshifted and incomplete genes. Eight pseudogenes were reclassified as a different type; five frameshifted pseudogenes became incomplete whereas three incomplete pseudogenes became frameshifted.

Table 3.2 Fragmentation status of each *Pseudomonas syringae* assembly post Kastor correction with the equidistant reference genome set.

Assembly	Number of genes ¹		
	Complete	Frameshifted ²	Incomplete ^{2,3}
Pss508			
Pre-correction	3967	1101	133
Post-correction	4890 (+949 / -20)	190 (+8 / -917)	106 (+8 / -36)
Psy642 (Illumina)			
Pre-correction	4981	25	208
Post-correction	4982 (+9 / -8)	21 (+1 / -5)	213 (+6 / -2)

¹Number of genes annotated using NCBI's PGAP. This number can fluctuate slightly between pre- and post-corrected assemblies depending on the sensitivity of the PGAP in gene prediction.

²General types of pseudogenes as annotated by the PGAP software.

³Genes were considered incomplete if the gene lacks a conserved start/stop codon, contains an internal stop codon, or is a partial sequence.

Compared to Pss508, Kastor identified relatively few potential error sites in Psy642, the Illumina draft assembly. Post correction there is a gain of nine complete genes along with a loss of eight complete genes for a net increase of one gene. There was also a net loss of five frameshifted genes and a net gain of six incomplete genes (Table 3.2). Of the eight complete genes that were lost, three became pseudogenized and five were no longer predicted. Similar to Pss508, Kastor can identify and defragment frameshifted pseudogenes in the Psy642 short-read assembly. Here, three genes were converted from pseudogenes to complete genes. There is an increase in incomplete pseudogenes, but similar to Pss508, two are new, two are a shift in pseudogene type, and only two represent a pseudogenization of previously complete genes. In both Pss508 and Psy642 assembly correction, there is a very low incidence of fragmentation of complete genes (four and three genes respectively) consistent with a low level of false positive corrections.

To investigate whether Kastor correction has additional impacts beyond de-fragmenting pseudogenes, we assessed the protein product similarity and coverage to the respective RefSeq ortholog used in PGAP's homology-based annotation [295]. The RefSeq protein orthologs were extracted for each gene with a known product; *de novo* predicted genes were excluded. These RefSeq protein sets were then mapped back to the respective assembly using tblastn and restricted to the best high scoring pair (HSP). We reason that the top HSP would yield the assembly gene equivalent, and fragmented genes should have either poor query coverage or low similarity. We plotted gene query coverage (qcov) to percent of positive-matching sequence similarity (ppos) for each HSP (Figure 3.3).

After error correction, most genes with a correction in both Pss508 and Psy642 showed an average increase in query coverage and similarity regardless of the annotation prior to correction; even previously complete genes showed an improvement (Figure 3.3). For Pss508, 1076 genes showed improvements in either query coverage or similarity, with 55.6% of these genes increasing in both scales, 37.4% increasing in query coverage only, and 7.0% increasing in only query similarity. Only 35 genes, representing 3.1% of edited genes, decreased in query coverage and similarity and only 12 of these decreased by greater than 5% for either coverage or similarity. Notably, 6 of these genes were de-fragmented pseudogenes. Despite being reclassified as complete, these genes had a decrease in similarity with their respective RefSeq orthologs.

Since the majority of genes remain uncorrected, we compared the general protein product profiles between Pss508 and its sister strain Psy642. Prior to correction, the Pss508 assembly is

highly fragmented compared Psy642, but post correction Pss508 and Psy642 become highly similar in terms of coverage and sequence similarity. This profile similarity suggests that the remaining genes with low query coverage may reflect strain differences rather than assembly errors. However, the post-corrected Pss508 assembly profile still has a greater number of fragmented genes. These non-corrected genes may be false negatives, possibly stemming from insufficient information for error correction.

3.3.2 Profile of Kastor detected errors

Long read assemblies are expected to be more error prone than short-read assemblies [238]. Many of these errors were previously attributed to poor disambiguation of homopolymer length during assembly [216,229]. We investigated whether Kastor detected errors in Pss508 follow the same pattern. Unexpectedly, Kastor initially identified a higher level of substitution error candidates than gene-fragmenting indel errors common to long reads with 24600 and 2000 candidates respectively [242,288]. However, many of these initial substitution candidates were subsequently dropped as they lacked read support. Only 900 (3.6%) substitutions were retained using a lenient threshold of 10% support, from raw unprocessed reads, of a singular replacement nucleotide. In contrast, 561 (28%) of indel error candidates had read support using the indel threshold of 25% of reads. By design, Kastor initially retains all detected indel error candidates. Rather than removing these candidates, the software adjusts the positioning of the error and expected correction. We reason that the systematic association of indel errors with homopolymer repeats in long reads could lead to low levels of read support for the detected indel candidates [216,238,289].

After error adjustment, Kastor found a total of 2900 errors in Pss508, accounting for less than 0.05% of the assembly (Table 3.3). Indel errors make up the majority of errors at 69% of all errors passing thresholds, with the majority of indels being false deletions. Comparatively, many fewer errors were found overall in Psy642 which is likely reflective of the overall increased accuracy of nucleotide calling in Illumina assemblies [202]. Although a higher number of deletion errors were detected than insertion errors at 124 and 71 respectively, the disparity between the two error types is less pronounced than in the Pss508 assembly.

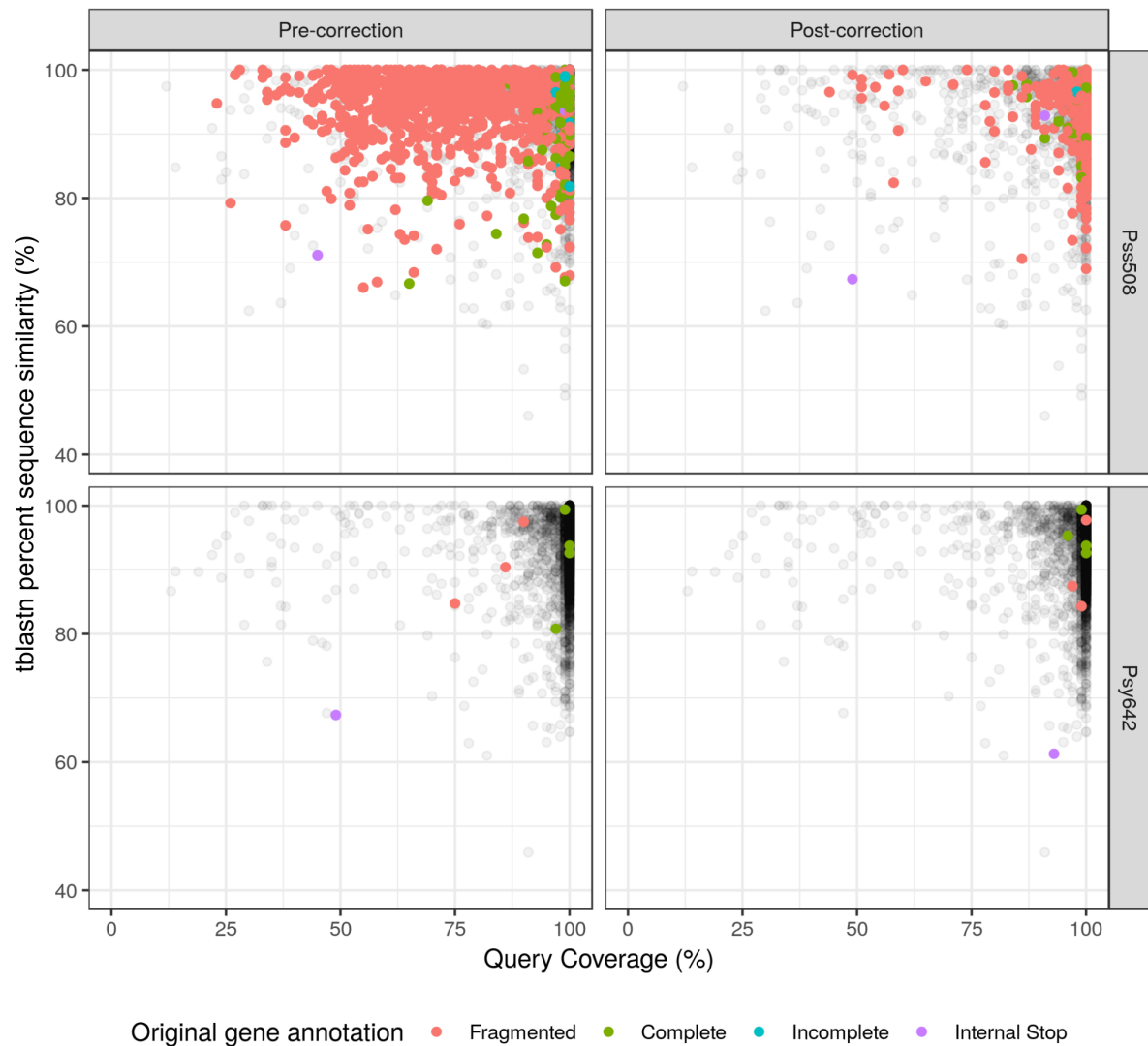


Figure 3.3 Reduction of fragmentation of identified genes before and after Kastor correction. RefSeq proteins used in PGAP's protein homology annotation of pre-correction assemblies were mapped back to each assembly using tblastn, and plotted as percent sequence similarity vs. percent sequence coverage. Corrected genes are coloured according to the original annotations while non-corrected genes are represented as grey. In detail, fragmented indicate genes where two or more frameshifted gene fragments were found; complete indicates complete genes; incomplete indicate truncated genes at the 5' or 3' end; and internal stop indicates genes containing a stop codon truncating the full gene. Gene fragmentation or incompleteness is inferred from lower query coverage when tblastn detects only a fragment of expected gene. Longer and more similar genes were found after correction. (A) Pss508 long-read assembly. (B) Psy642 short-read Illumina assembly.

Table 3.3 Number and types of errors detected by Kastor

	Assembly	
	Pss508	Psy642 (Illumina)
Detected Errors		
Total	2900	195
Substitutions ¹	900	-
Deletions ²	1708	124
Insertions ²	292	71
Indel Error Locations³		
Within gene	1317	17
Intergenic	683	178

¹ Number of substitution errors with read support. Reduced from 24600 candidate substitution errors in Pss508. No substitution errors were supported for Psy642 due to a lack of read data.

² For Pss508, the genome location of 561 (19%) detected indels were adjusted based on supplementary raw data during the error adjustment module.

³ Error locations were determined based on assembly annotations using NCBI's PGAP.

There is also a difference in the location of indel errors between the two assemblies. Although coding sequences account for 89.1% of the Pss508 genome, only 1317 (67%) of the indels found in Pss508 were within annotated genes while the remaining 683 (32.6%) were in intergenic regions. In contrast, only 17 (~9%) of the errors in Psy642 were associated with annotated genes. This may be due to the low rate of indel changes within genes when comparing between bacterial genomes, so detected indels are more likely to represent errors. Conversely, intergenic regions may have more indel associated differences between genomes, making it more difficult to distinguish indel errors from strain-to-strain variation.

When taken together with overall improvement and defragmentation of error containing genes (Figure 3.3), the number of indels found in pseudogenes and low number of gene-associated errors in Illumina indicates the corrected within-gene indel errors in the Pss508 assembly are unlikely to be spurious errors. Conversely, there is an elevated number of indel errors found in the intergenic region for both assemblies, despite less than 11% of the genome being intergenic. When manually

inspected, some of these errors were found in variable length homopolymers in otherwise conserved regulatory RNA regions. In non-conserved intergenic regions, high nucleotide variability may lead to poor mapping of the region and consequently high levels of indel differences. The increases of indels may make these regions prone to false detection that may contribute to higher level of intergenic corrections.

Since nanopore sequences are base-called as 5 or 6-mer events for R9.4 flowcells, we investigated whether detected errors are associated with specific k-mers in Pss508 by extracting the upstream 4-mer for each detected error [216]. When the top 20 k-mers were plotted for each error type (Figure 3.4), the majority of errors are associated with homopolymers which are expected to be more error-prone [229]. However, the associated homopolymers differ for each error type. In general, substitution errors appear to occur frequently following small homopolymers, length 3 or less, or single nucleotides. There is no dominant 4-mer associated with substitutions suggesting that they are more randomly distributed in the assembly. Furthermore, substitution errors are more likely to be either a guanine or cytosine regardless of the upstream 4-mer.

In contrast, deletion errors are highly associated with homopolymers from two to eight nucleotides long (Figure 3.4). Although there is a large range, the greatest number of identified deletion errors occur following four to five stretches of the same nucleotide [289]. Unexpectedly, despite a relatively high GC content of 59%, the most common errors were associated with long stretches of adenine and thymine followed by around half as many errors associated with guanine and cytosine. In all cases, the detected error tends to match the associated homopolymer, but there is a small incidence of errors where a non-repeated nucleotide is deleted. Taken together, the prominence of deletion errors occurring after long stretches of the same nucleotide may reflect difficulties in determining accurate homopolymer length during basecalling. The uncertainty and variability in homopolymer length between reads can lead to more conservative estimates of length and therefore introduce erroneous deletions [216].

Many of the detected insertion errors follow either adenine or thymine, but are not associated with any particular type of homopolymer (Figure 3.4). Although insertion errors are more frequently associated with homopolymers of length two and five, this trend differs for PsyTLP2 (Supplementary Tables S1-2 and Figures 1-3) indicating some variation between sequencing runs. Unexpectedly, two

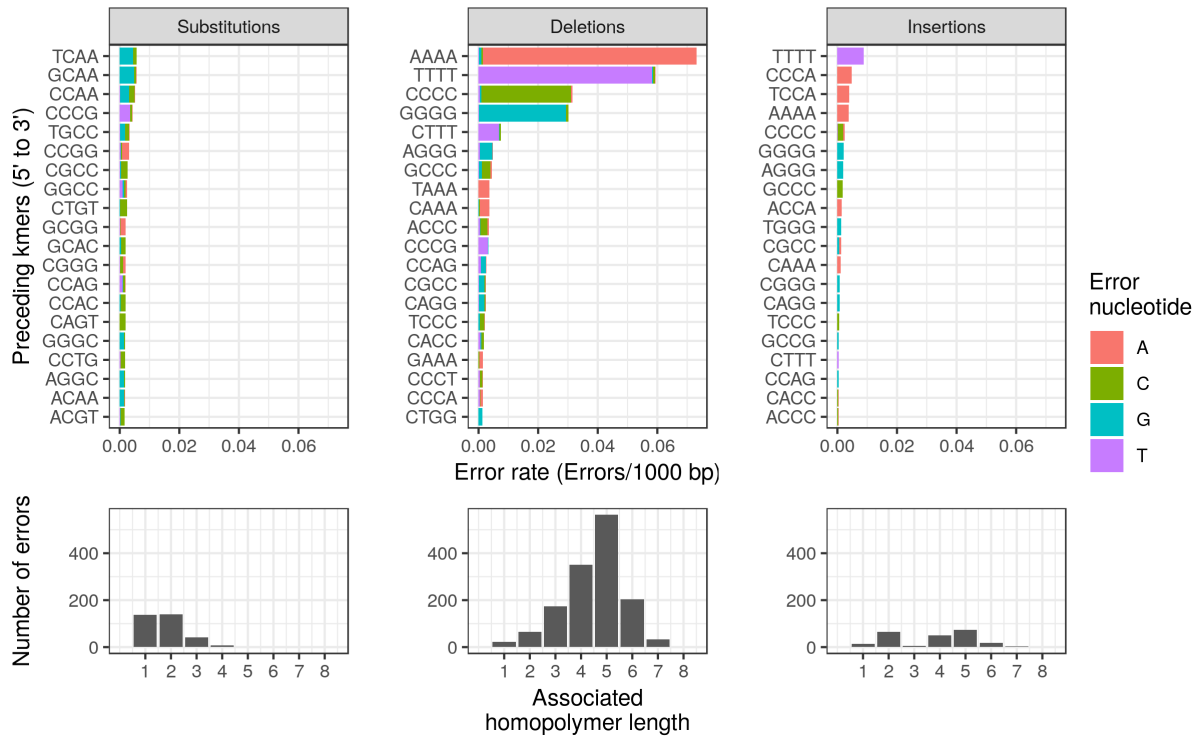


Figure 3.4 Breakdown of the genomic sequence preceding detected errors in the Pss508 assembly. Top plots: Top 20 k-mers ($k = 4$) upstream of detected errors separated by error type. Each bar is coloured with the associated erroneous nucleotide. Bottom plots: the distribution of the length of upstream homopolymers for the respective error type.

of the top four 4-mers for insertions are not homopolymer repeats. Both CCCA and TCCA, as well as ACCA (#9), are two or more cytosine followed by a single adenine and a second erroneous adenine. The combination of these 4-mers likely contributes the majority of errors associated with shorter homopolymers. This suggests that there is potential difficulty discerning the number of adenines following a nucleotide run of cytosines. Collectively, the high abundance of homopolymer containing k-mers is indicative of non-random errors in minION long read assemblies associated largely with homopolymer regions [229,289].

3.3.3 Limitations and workarounds

Non-corrected pseudogenes

Since Kastor candidate error correction is dependent on reference information, a lack of reference genome coverage can limit the error detection leading to possible false negatives and consequently non-corrected genes. For *Pseudomonas syringae* related strains, Dillon and colleagues (2019) report that the core genome (2410 genes) of greater than 400 *Psy* (*sensu lato*) genomes represent slightly less than 50% of a typical *Psy* genome assembly [43]. Reliable reference coverage of the remaining genome is contingent on the user's choice of reference genomes due to an expansive accessory genome reflective of the species high diversity [44]. Using the primary equidistant reference set (set one), approximately 10% of the Pss508 assembly had insufficient reference genome coverage (<5 genomes) to confidently search for error candidates.

Within the Pss508 assembly, 953 genes annotated as incomplete or frameshifted were corrected, and 296 frameshifted or incomplete genes were retained in the post-Kastor assembly (Table 3.2). For comparison, 233 frameshifted or incomplete genes are present in the post-correction *Psy*642 Illumina assembly, with a larger proportion of incomplete genes. Within both genomes a proportion of the frameshifted or incomplete genes are likely true pseudogenes, however the greater proportion of frameshifted genes in the Pss508 assembly suggests that many of the non-corrected frameshifts may represent retained indel errors. We investigated the role of reference genome coverage on these non-corrected pseudogenes. Excluding the four newly annotated pseudogenes and four previously complete genes, 283 of the remaining 288 pseudogenes in Pss508 did not contain detected errors (Table 3.2). We determined 148 (52.3%) of these pseudogenes did not have sufficient coverage for at least 50% of the gene. These genes were also found in specifically low coverage regions. Notably, the median gene length for many of these pseudogenes is well below an average bacterial gene [297]. Upon manual inspection, at least 3 of the corresponding reference genes have since been depreciated in NCBI's database suggesting that a subset of these short genes may be spurious annotations [298].

Maximizing error correction

To assess the impact of increasing reference genome coverage on error correction, we applied Kastor to Pss508 using two other reference genome sets (set two and three) with increasing number

and diversity of Psy strains to maximize accessory genome coverage by introducing more divergent sequences that may be passed through recombination [299]. The second set includes 170 strains constrained to the Group 2 Psy phylogroup, and the third set includes 441 strains chosen to also include select strains from all Psy phylogroups [44]. In both cases, the error threshold was increased to 0.995 and 0.999 to ensure a consensus across the higher number of genomes. These genome sets improved the target assembly coverage with ~97% and ~98% of Pss508 determined to have sufficient reference genome coverage respectively. Despite the higher coverage in both cases, when these corrected assemblies were assessed using BUSCO, they were found to have an additional 2-3 universal orthologs shown as fragmented indicating slightly poorer correction when compared to using the equidistant reference dataset (set 1) (Table 3.4). Despite sufficient coverage of core genes, there was less sequence consensus due to increased nucleotide variability between more distant genomes. This variation leads to lower error detection in previously covered regions using reference genome set one.

To offset the tradeoff between higher coverage and detection of very conserved assembly errors, we implemented an optional hybrid approach where Kastor can take advantage of multiple genome sets with a different diversity of sequences. A primary reference set can be supplemented with reference information from the secondary genome set only in regions of low coverage. This strategy allows for coverage in the distant accessory genome while maintaining the higher candidate detection rate of the primary set. We tested the effect of three different combinations of multiple reference genomes sets. In all cases, the primary set was the equidistant 11 genome set (set one). This

Table 3.4 Influence of genome set on Kastor corrected assemblies using BUSCO to assess completeness of Pss508. Three genome sets were used with Kastor. Each set contains a different number and variety of genomes.

Genome set	Status of universal single orthologs			
	<i>Complete</i>	<i>Duplicated</i>	<i>Fragmented</i>	<i>Missing</i>
None	385	0	54	13
Equidistant set	451	0	0	1
Psy Group 2 set	448	0	11	4
Psy 441 set	426	0	18	7

set was either supplemented with either no additional reference genomes, the group 2 psy set (set two), or the psy 441 set (set three). When using a combination of reference sets, an additional 150 errors were detected in the Pss508 long-read assembly. The majority of these errors appear within genes in previously low coverage regions. Correction of these errors was beneficial in defragmenting an additional ~50-60 pseudogenes. Overall, a single, well-curated reference genome set is sufficient for Kastor to correct the majority of assembly errors, but error correction can be maximized with an additional curated genome set with a slightly higher diversity of genomes.

A concern about using reference information is biasing the correction of the final assembly towards one or a small group of reference genomes. Even after implementing a strain-specific read-based adjustment, greater than half of detected candidate indel errors lacked any read support for target strain specific corrections. Here, Kastor relies solely on reference alignment differences to determine the error site and suggested correction. Since the error site for a region is determined by scoring all possible sites, it is possible for an error sites originating from a single or group reference genome to be consistently scored higher. Without any read-based adjustments, these corrections would move the target assembly closer towards those reference genomes. Successful de-fragmentation, with minimal bias, should improve assembly similarity to the closely related *P. syringae* genomes in the reference set, especially for conserved genes. To determine if a potential bias towards specific genomes existed, we compared each pre- and post-correction assemblies to the reference *P. syringae* genomes in the equidistant genome dataset.

Sequence similarity was assessed by comparing the average nucleotide identity (ANI) between each assembly and each reference genome using orthoANI [296]. As expected, comparing the post-corrected Pss508 to Psy642 and PsyUB0390 (group 2c strains) had ~98% ANI which is consistent with strains belonging to phylogroup 2c [43]. All other *P. syringae* strains had an ANI of ~94% to the two assemblies which is similar to distance of Psy642 to the other strains (Table B.3). Comparing ANI differences between pre- and post-correction assemblies, the shift in ANI was minimal, less than 0.3%, as expected from the low rate of corrections (Figure 3.5). In fact, ANI increased for almost all strains when compared to Pss508. Unexpectedly, one exception is that the Pss508 assembly decreased slightly in similarity to both group 2c strains. Notably, there is an increase in similarity to *avellanae* pathovar strains for Pss508. A correction bias towards *avellanae* pathovars is less likely when the second highest ANI increase was with a *P. syringae* pv. *Lapsa* strain which is more distant to the *avellanae* pathovars than other strains [43]. Overall, Kastor corrections do not

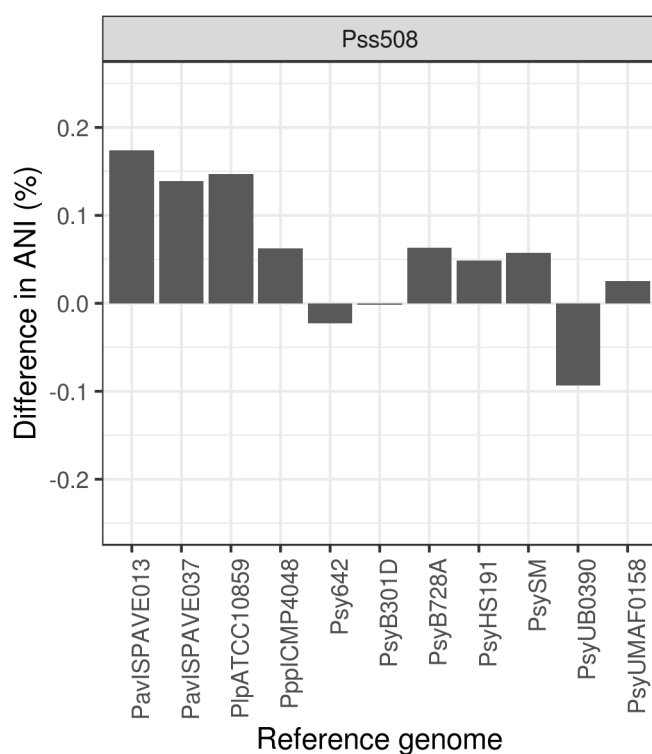


Figure 3.5 Post Kastor correction Average Nucleotide Identity (ANI) percent differences. ANI between Pss508 and reference genomes were compared before and after error correction.

appear biased towards any particular strain or group and have minimal impact on nucleotide similarity to reference strains.

3.4 Discussion

Detected assembly errors in Illumina draft assemblies

We designed Kastor as a tool to use reference genome information to correct draft assemblies to improve sequence quality. Ideally, a draft assembly with high sequence quality would have minimal corrections reflecting a low error rate. We assessed the performance of Kastor in error correction by applying the software to error-prone long-read assemblies, Pss508 and PsyTLP2, and an Illumina draft assembly, Psy642. As the current standard to low error sequencing, we expect the

Illumina sequence to have minimal nucleotide errors and serve as a baseline for background error correction [202]. Kastor detected many more errors in Pss508, with up to 15x more errors than the Illumina equivalent Psy642 (Table 3.3). However, Kastor also found 195 potential assembly errors in Psy642 suggesting some errors may be present despite a higher sequence quality. Alternately, if we assume the Illumina sequence assembly represents a reasonable ground truth, we can estimate an upper false positive limit of one erroneous correction per ~29,800 basepairs. We define this rate as the upper limit since Illumina sequences may still contain errors that require correction suggesting some errors may be true positives [300].

Unlike in the Pss508 assembly, the majority of indel errors (91%) were found in intergenic regions of Psy642 (Table 3.3). These errors may represent false positives due to lower sequence conservation in intergenic regions. This higher variability can lead to poor alignment and spurious error detection. A false positive rate is difficult to quantify even with detected intergenic assembly errors in an Illumina assembly since 18 - 32% of intergenic regions greater than 60 nt were found to be conserved in a variety of bacteria [301]. These conserved regions can contain sRNAs, ncRNA, and regulatory elements under evolutionary pressure [302,303]. The sequence conservation suggests that poor alignment is unlikely to introduce spurious error detection and that some of these intergenic errors may not be false positives. In contrast, variable length homopolymeric repeats in small RNAs have also been implicated in regulatory function [304], and it is challenging to separate true and false positives in these regions.

Software use and limitations

Although Kastor is effective in detecting putative assembly errors, it is reliant on accurate and precise mapping of the reference genomes. This reliance is affected by three main factors: reference genome selection, reference quality, and mapping quality. Therefore, a limitation to Kastor is the availability of high-quality reference genomes for selection. Inclusion of highly divergent genomes may introduce too many variations and can mask clade specific mutations even in universal orthologs (Table 3.4) [44]. However, the genome set used still needs to be sufficiently diverse to not bias correction towards any particular reference sequence. Therefore, we recommend a reference genome set of around 10 closely related but approximately equidistant reference genomes that belong to the same monophyletic groups as the target organism. Including at least one sister strain (~98% identity)

is also recommended to reduce potential false detection of single variants. The genome collection used here was sufficient to detect the majority of assembly errors (Table 3.2). An optional supplementary reference set with more divergent sequences may be provided to increase coverage and maximize error detection, but is not required as the gain in accuracy is not pronounced. When provided with a secondary set, there was slight improvement to the Pss508 assembly with an additional 68 complete genes (~7%).

The quality of chosen reference genomes also influences error detection. Lower quality (i.e., draft) genomes can still be used, but missing sequences in the reference genome can lead to lower coverage of the assembly and consequently missed errors. Furthermore, the discontinuity can introduce spurious regions of higher coverage from overlapping contigs or regions of low coverage [202,251]. A well-rounded and high quality genome set can be most effective in the assessment of errors present in long read assemblies, thus careful selection of reference genomes is recommended.

Conclusion

Kastor was capable of effectively detecting errors in Pss508 and PsyTLP2 minION-only assemblies using reference genome information. Upon correction of these errors, both assemblies were sufficiently accurate and complete for effective gene annotation. This form of guided error correction can be used as an alternative method of final polishing without the need for additional short read sequencing [238,287]. Additionally, our software can be used to estimate the prevalence of small indel errors in long-read assemblies [289]. Such estimates may provide researchers with a useful benchmark to compare assembly quality across different assembly pipelines.

Chapter 4:

Kastor-P: Reference protein-based gene fragmenting error assessment module

4.1 Introduction

4.1.1 Chapter objective and overview

One major limitation of genome-based error assessment with Kastor is requirement of a diverse range of closely related assemblies to the target assembly (Chapter 3). When high quality reference genomes are unavailable, genome-based correction can bias towards the limited data. As an alternative, we hypothesize that the high diversity of reference protein sequences available in the public space can be leveraged to directly assess gene-fragmenting errors in target assemblies. In this study, we present a preliminary protein-based error assessment module in Kastor (Kastor-P) to complement the genome-based assessment module. This module takes advantage of the availability of the large and diverse public protein database to identify gene-fragmenting insertion/deletion errors in draft assemblies [238]. The high variety of accessible protein sequences combined with sequence conservation of homologous proteins should improve input sequence availability for error detection in scarcely studied organisms. For example, mobile genetic elements, such as coding sequence for virulence proteins [305] can be mapped to proteins from distant relatives (Chapter 5)[306]. To assess error identification using Kastor-P, we ran the module on two *Pseudomonas syringae* draft assemblies, each generated solely from long reads or short reads, from a previous study (Chapter 2 and 3). Here, Kastor-P made more conservative, targeted error identifications compared to the genome module. After correcting for these errors, we were once again able to improve BUSCO-assessed assemblies from 85.2% to 99.8%, comparable to an Illumina assembly [266]. This conservative assessment leaves undetected errors within intergenic regions but is sufficient for annotation-targeted studies. In summary, with Kastor-P's protein-based assessment module, we were able to correct protein-coding regions within the long read draft assemblies to a near-complete state (Chapter 3).

4.1.2 Chapter background

Reduced costs and streamlined workflows in sequencing technologies have exponentially increased our ability to collect data and answer biological inquiries at the genomic level, from analysing evolutionary processes [307] to epidemiological tracking of global pandemics [308]. Improvements to genomic research have been further compounded with the continual maturation and innovation in long read sequencing approaches within the last decade [237]. Although long read sequencing systems have a greater error rate for individual positions (9% - 15% [211,232,242]) compared to the more mature short read systems (<1% [202]), the longer read length can compensate for assembly discontinuities that trouble short read assembly methods [251].

With short and long read systems compensating for the disadvantages of the other system, many researchers have opted to combine both technologies for comprehensive resolution of the target genomes, with large projects utilizing more than 5 sequencing methods [224,251,253]. A common hybrid approach sequences DNA using one short read (Illumina) and one long read sequencer (PacBio or Oxford Nanopore Technologies), followed by either using long reads to resolve assembly graphs [251], scaffolding of assembled short read contigs (e.g. during assembly graph resolution) [205,252,288], or polishing of long-read assemblies with more accurate short reads [253,292]. However, there are limitations to this approach. At the analysis level, sequence discrepancy arising from the error-prone long reads can lead to poor mapping between the two systems and ultimately inefficient correction in final results [238]. Research projects with a smaller scope may use only a single sequencing system, or can be restricted by sample (e.g. environmental DNA) and/or budget (e.g. smaller research labs) limitations. In these cases, gene fragmentation in error-prone long read assemblies making long read sequencing the unattractive option despite higher assembly continuity (Chapter 3). Although assembly correctors such as Nanopolish [229] and Medaka [249] have been developed, up to 14% of known single, universal orthologs remain fragmented or missing in final assemblies generated solely from long reads (Chapter 3) [266].

In response, we have previously developed Kastor, a genome-based error assessment software that aims to defragment long read assemblies (Chapter 3). This software takes advantage of the large availability of reference genomes to identify fragmenting insertion/deletion errors common to these systems (Chapter 3). To minimize reference specific biases, a diverse set of closely related genomes was selected for the test set, two *Pseudomonas syringae* genomes [7,256,309] from a subspecies of the ubiquitous and extensively researched environmental *Pseudomonas* bacteria [4].

Although many sequencing consortiums [310] have been expanding the diversity of genome assemblies available in the public space, genome availability may still be insufficient for many research projects. As an example, environmental metagenomic studies have continually unveil *Candidatus* species revealing that a large genetic diversity remains undiscovered [207]. For these studies, a diverse set of related genomes may not be available to meet Kastor's requirements. In contrast, reference protein sequence availability can exceed the genomic counterpart with protein database collecting sequences from DNA, RNA, and protein sequencing. Unrestricted by the need to assemble the full genome, new proteins from diverse species can be reported for metagenomic studies where full genome assembly may be unavailable [311]. Furthermore, public protein data are also available from other, more generalized sequencing studies [312]. Lastly, proteins undergo a slower rate of evolution compared to DNA, creating a wider pool of available input sequences improving the probability of sufficient mapping for error assessment in the target genome.

4.2 Material and Methods

Draft genome assembly, Reference protein curation, and pre-processing

The test data used is comprised of a draft assembly from *Pseudomonas syringae* pv. *syringae* strain 508 (Pss508) [7,309]. The draft genome assembly was sequenced solely using Oxford Nanopore Technology's R9.4 MinION flowcell [242] and assembled with an Minimap-Miniasm-Racon pipeline optimized specifically for the long read data used [247,265,265,313]. Prior to Kastor correction, these assemblies were corrected with read-based methods Nanopolish [229] and Medaka [249]. A control Illumina draft assembly for closely related bacterium *Pseudomonas syringae* 642 (Genbank Accession NZ_ADGB00000000) was retrieved from NCBI [18].

For comparisons of both Kastor error detection methods, the same draft assembly genome for Pss508 was also corrected with Kastor's genome-based module (as per Chapter 3). A first correction pass used a curated set of 11 approximately equidistant *P. syringae* genomes (94 to 98% identity) supplemented with 170 *P. syringae* spp. genomes for low coverage regions. A second correction pass utilized a sub-phylogroup specific set of 9 genomes containing *P. syringae* and *P. congelans* strains [44]. For details on the genome-based correction, refer to Chapter 3.

A curated set of reference proteins was obtained for each of the draft genomes from NCBI's RefSeq database as described below. In the current Kastor-P implementation, the software requires three input files: the target draft assembly, a set of single unique reference proteins per expected gene, and a list of the locations of mapped reference protein to the target assembly (via tblastn). The latter two input files are up to the user. In our test dataset, the reference protein id used in homology-based predictions (PGAP build 4456) of their respective candidate gene were extracted [295]. Remaining candidate genes that were identified using *ab initio* models were first extracted, based on gene position, as DNA gene sequences which were then translated into the corresponding protein sequence. Predicted protein sequences were then used with blastp to identify top-scoring equivalents in the RefSeq database. Due to limitations in the current implementation, protein duplicates and redundant protein variants were removed manually to avoid conflicts during error detection. Overall, the final protein set contains a single, unique reference protein for every candidate gene that was identified using PGAP. To utilize sequence conservation information, the reference proteins were then mapped back onto the draft assembly using tBlastn[314], grouped by query id and ordered by percent coverage then by bitscore. A custom output of tblastn containing reference protein id, contig, genomic location, reading frame, protein sequence positions, and both mapped genomic DNA and peptide sequence, stored as a tab-delimited file.

Correction and analysis

The Kastor-P protein module was written natively in Python v3.6.3 without third party dependencies. Post-correction, completeness of each assembly was estimated using BUSCO, a software that benchmarks the status and copy number of single copy universal orthologs [266] with the expectation that most organisms within a specified group will have close to 100% of the BUSCO genes present. Gene fragmentation in the pre- and post-correction assembly was analysed using NCBI's PGAP (build 4456) [295]. The curated reference protein set was used to assess whether improvements were primarily due to gene de-fragmentation. All tables and plots were generated using R version 3.4.2 and Rstudio version 1.1.383 using packages dplyr (version 0.7.8), tidyr (version 0.7.2), ggplot2 (version 2.2.1) and ggpubr (version 0.1.8).

4.2.1 Software Implementation

Kastor-P's protein-based method for error detection is implemented as a secondary, separate module from the genome-based Kastor correction tool. Users can select either detection method, or both can be executed in parallel. Since both modules were developed separately, use of modules from both methods are run as separate commands and users must specify if they want to execute "Kastor genome" or "Kastor protein".

Candidate error region detection

A flowchart of the full error assessment pipeline is summarized in Figure 4.1. The first phase of error correction identifies potential error locations by identifying gaps and overlaps between aligned reference protein segments. Kastor-P sequentially processes the mapping between reference proteins and the draft genome by first collecting aligned segments for each of the mapped reference protein. Then, overlaps and gaps between these aligned protein fragments are located (Figure 4.2) and used to identify possible error locations as described below. As the first step for each mapped reference protein, the aligned protein fragment best matching the putative gene (i.e., highest percent coverage, with ties decided by higher bitscore) is used as the initial seed for the most probable location and reading frame for the gene within the draft genome assembly. From this seed, candidate start and end positions of the respective putative gene is calculated based on the expected length of the reference protein. Before proceeding to error detection this seed is initially screened for full coverage. Reference protein queries with a fully mapped seed (100% coverage) correspond to intact genes and are removed from the error correction pipeline. For reference protein queries with seeds representing incomplete or fragmented genes, the remaining collected aligned protein segments are ordered based on percent coverage and bitscore. Assessment of these protein segments occurs sequentially during candidate error detection. To begin the candidate error detection, the genomic region spanning the mapped seed is designated as an “assessed region” (confirmed fragment of the putative gene) for downstream comparisons.

For each of the non-seed alignments, a preliminary position check is done, and any protein alignments mapped more than 50 bp outside the putative gene boundaries are discarded. If the alignment is found within these bounds, the protein alignment is then compared to all previously assessed regions sequentially. During each comparison, a secondary initial check is performed to determine if the

current aligned protein segment is completely within the compared assessed region, by genomic position, and if so they are dropped and Kastor-P begins evaluating the next aligned protein segment. After passing both checks, the aligned protein segment is considered to represent a fragment of the putative gene. The aligned protein segment and the assessed region being compared are rearranged relative to its mapped genomic positions. The end positions of each alignment/region closest to other is then used to identify whether the ends of these two neighbouring segments either overlap or form a gap. Since overlaps do not require additional information, the region of overlap is marked as a potential error location. Gaps are considered further downstream in the process. This process is then repeated for each of the previously assessed genomic regions. These primary comparison pass avoids conflicts from any early processing of sequences and designates aligned protein segment as one of four categories: no overlap (gaps on both sides); single overlap between the aligned protein segment and assessed region; double overlap, overlap of aligned protein segment with two other previously assessed region ($r + n + r$); and irregular (a catch-all category), indicating more than two overlaps were found, a case that was not observed in our test dataset. If there are no overlap, we designate the genomic region that the aligned protein segment covers as a another assessed region indicating that this genomic region contains a fragment of the putative gene, and can be used in future comparisons of subsequent aligned protein fragments from the same reference proteins.

In the event of a single or double overlap, the temporarily marked potential error locations are designated as candidate error locations to be evaluated. The aligned protein segment is then merged with the overlapping assessed region updating it into a single, larger assessed genomic region ($r1$) containing a fragment of the putative gene. In the case of double overlaps, the software handles merging sequentially in the same order as alignment comparison. After the first merge, the other overlapping assessed region ($r2$) is then merged to the newly updated assessed region ($r1$) to reupdate it into a single, even larger assessed region. The second merge effectively collapses the genomic region containing the evaluated aligned protein segment and the flanking compared assessed regions ($r1/2 + \text{alignment} + r1/2$) into a singular assessed genomic region with new start and stop positions.

After all overlapping segments have been merged into a set of non-overlapping assessed regions. The list of these assessed regions is then ordered by genomic position. Traversing the entire putative gene location from 5' to 3', these regions are merged sequentially with identified gap regions are flagged as candidate sites for correction. After merging, all assessed region are collapsed into a single genomic region spanning the entire known putative gene, including putative error locations.

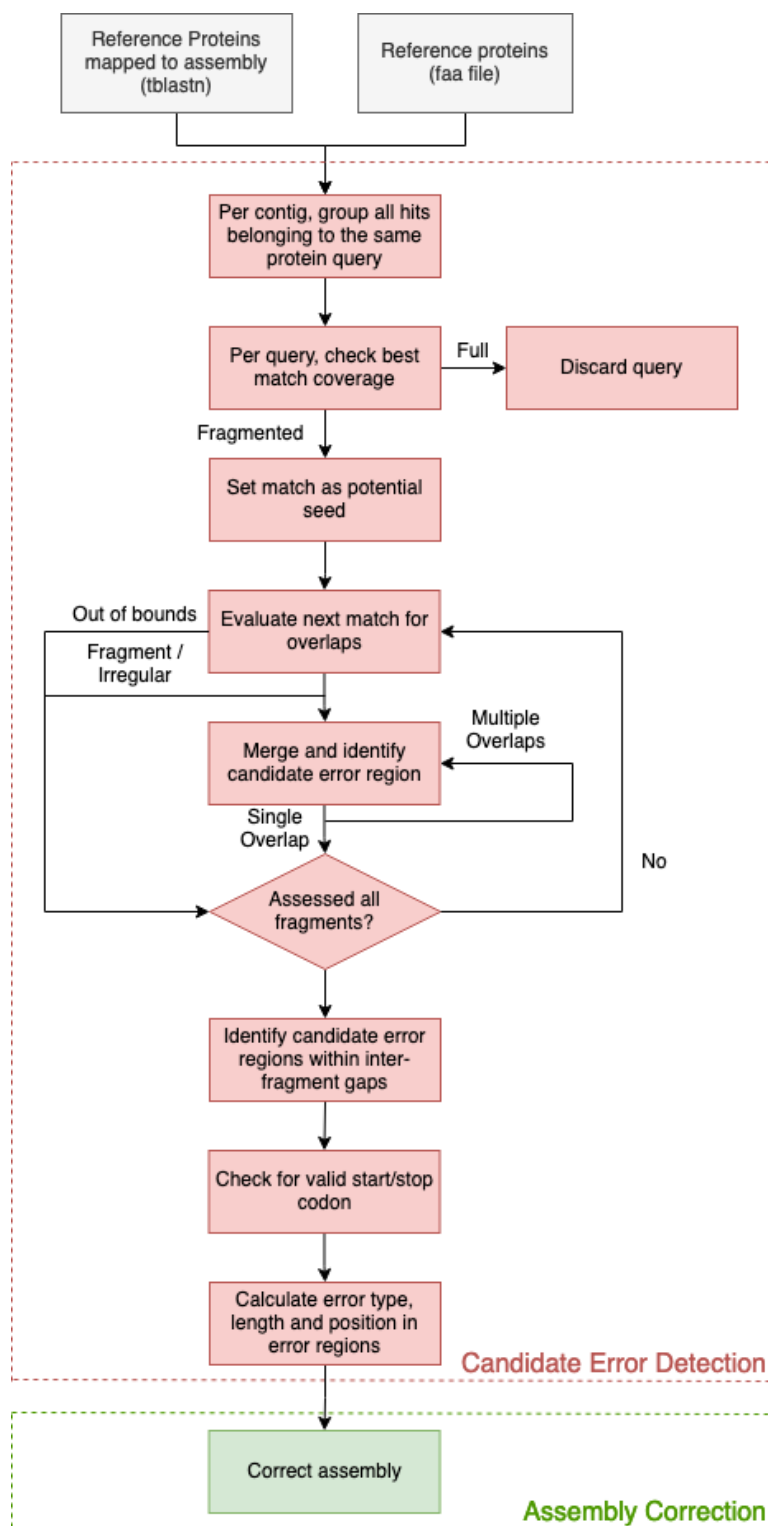


Figure 4.1 Flowchart of the protein-based Kastor-P workflow.

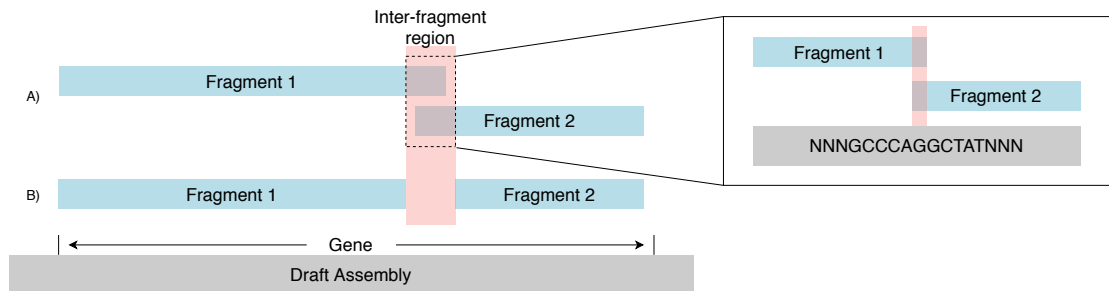


Figure 4.2 A simplified schematic of an overlap with a candidate error detection. The distance between closely mapped of protein fragments is assessed for candidate error regions. Negative distances (overlaps) were evaluated as candidate regions with an erroneous deletion (A) whereas positive distances (gaps) were considered as candidate regions with an erroneous insertion (B). Sequence(s) at the edge of these regions (breakpoints) are evaluated as candidates for regional error with respect to differences in reading frame between fragments.

Kastor-P finalizes the full genomic region as a putative gene by checking for a valid start and stop codons at the ends of the segment. If start and stop codons are present in the expected locations, the assessed genomic region is flagged as complete with a putative gene. For incomplete gene segments, 50 bp upstream and downstream of the expected start/stop position is searched for valid start and stop codons in all reading frames. In the case of multiple valid codons, the start/stop codon closest to the expected site is marked as the gene's start/stop. For codons found in an alternate reading frame, Kastor-P marks the gap between these found codons and the edge of the finalized assessed genomic region as a candidate error region. For assessed genomic regions without identified start/stop codons, the case is marked as incomplete indicating that the putative gene could not be fully resolved, but candidate error regions found through between aligned protein segments are still evaluated downstream.

Error determination

In the current implementation of Kastor-P, the software uses a heuristic-based, inverse logic model for error detection. This model indirectly detects the error position by determining the error correction that best reconstructs the fragmented gene into a complete in-frame coding sequence. Each candidate error region is evaluated with the same model regardless whether the candidate error region was initially assessed as an overlap or gap (Figure 4.2).

An expected size for the candidate error region (inter-fragment region in Figure 4.2) is calculated from an assessment of the mapped protein sequence relative to the encoding assembly sequence. Since frameshift sequencing errors are less than three nucleotides in nearly all cases, the current implementation limits the length of putative correction to two nucleotides, the maximum correction needed to bring gene fragments to the same reading frame.

In this assessment the amino acid sequence involved in the overlap is used to extract the relative assembly sequence. The software attempts to resolve the difference between the aligned protein segments and codons within the draft gene sequence, where indels in the putative error location are corrected so the gene has a consistent reading frame between the segments. For example, in one of the simplest cases (Figure 4.2b), a single nucleotide gap error between two aligned protein segments, but with no corresponding amino acid in this gap region indicates an additional nucleotide creating a frameshift in the putative gene.

Lastly, the putative insertion/deletion correction needed is determined by the incongruence between expected and actual region size. To simplify, a candidate deletion error is determined when an putative insertion of either one or two nucleotide(s) is required to adjust an error region to match the expected size (Figure 4.2a). For example in Figure 4.2a, the last amino acid of fragment 1 and first amino acid of fragment 2 shares the same guanine in their codon indicating an additional nucleotide is needed to shift both fragments to the same reading frame. As the initial strain specific nucleotide is unknown for candidate error deletions, the inserted nucleotide during correction is marked as 'N'. A candidate insertion error is determined in the inverse situation where a deletion is required for congruency between expected and actual region (Figure 4.2b). In the previous Figure 4.2b example, a gap region with size of one but an expected size of 0 is flagged as candidate insertion error where a deletion correction of the one nucleotide is required. Additional differences such as extra amino acids are considered species-specific variants and ignored.

Kastor-P establishes the error position and nucleotide marked for correction based on heuristics informed from previous data extracted from the genome-based detection test data. In all cases, the software considers the edge of each candidate error region as the probable target position. We reason that these positions mark the breakpoint where mapping cannot be extended suggesting the location may correspond to a shift in the reading frame. Previously, fragmenting errors have been attributed to homopolymer uncertainty [232] with a high prevalence of errors involving longer

homopolymers especially with a length of 4 to 5 nucleotides (Figure 3.4). In summary, Kastor-P estimates the most likely insertion error position as at the edge of longer homopolymer, where a homopolymer is present. In the case of a candidate deletion error, the corrective insertion is set to ‘N’ as the original nucleotide is unknown. Once the errors have been determined they are collected to be used in the final assembly correction. Similar to the genome-based Kastor error detection method, Kastor-P can adjust the detected errors found using raw read information as a separate process within the genome-based correction module.

4.3 Results

4.3.1 Improvements to draft assembly

Protein-based error correction improves assembly completeness similar to genome-based correction

The effect of Kastor-P error assessment was evaluated on both a long read (MinION only) assembly belonging to *Pseudomonas syringae* pv. *syringae* 508, Pss508 (Accession No. CP053030), and a short read Illumina draft assembly belonging to closely related *Pseudomonas syringae* 642, Psy642 (Genbank Accession NZ_ADGB000000000)[18]. We also assessed the effectiveness of this new methodology with our previous genome-based method. Therefore, for Kastor-P's protein correction, the same pre-corrected Pss508 draft assembly from our previous work was used in this assessment (Chapter 3). Prior to correction, when the assembly was evaluated for known single universal orthologs in Gammaproteobacteria, BUSCO classified the Pss508 draft assembly as incomplete with 54 (11%) and 13 (3%) of orthologs as fragmented and missing respectively (Table 4.1)[266]. After either protein or genome based correction, Kastor was able to improve the Pss508 assembly to 99.8% completeness with the same remaining single universal ortholog missing in both cases. Similar results were obtained with the Psy642 assembly, with 99.3% of universal orthologs found in the Kastor-P correction, indicating our corrections to detected errors do not noticeably fragment intact annotated genes.

To ensure that Kastor-P's error assessment is not limited to highly conserved genes and proteins, i.e. easily mapped genes due to sequence conservation, we indirectly evaluated our correction to genes in the accessory genome by annotating both pre- and post-corrected assemblies

using PGAP (build 4456)[295]. With both Kastor protein and genome correction, the annotation of complete genes greatly increased, from 3967 in the pre-corrected assembly to 4754 and 4890, respectively in the final assemblies (Table 4.2). Although genome correction is within the expected range when compared to Psy642's 4981 complete genes, the protein corrected assembly corrected 136 fewer genes or approximately 15% less. Around 108 of these genes remain fragmented with 298 annotated genes containing observed frameshifts in the coding region, compared to genome-corrected 190 genes with observed frameshifts, while there is an additional 35 incomplete genes annotated in the protein-corrected assembly. However, in both cases identified Pss508 frameshifted genes found still outnumber the genes with observed frameshifts found in the Psy642 Illumina assembly (less than 21), likely indicating that both the genome and protein based error detection is conservative and do not detect all cases. Conversely, the number of incomplete genes found in the Psy642 assembly is greater than those found in the long read assemblies by 50 or more genes. Many of the incomplete genes in the Psy642 assembly can be attributed to assembly discontinuities leading to gene truncation on contig edges (i.e., 297 contigs in the Psy642 vs. one contig in the Pss508 assembly) rather than sequence errors [251].

Table 4.1 A comparison of universal single orthologs detected by BUSCO post Kastor corrected assemblies for using both approaches. A total of 452 single universal orthologs found in Gammaproteobacteria were determined to be complete, duplicated, fragmented or missing.

Assembly		Status of universal single orthologs			
		<i>Complete</i>	<i>Duplicated</i>	<i>Fragmented</i>	<i>Missing</i>
Pss508					
	Draft	385	0	54	13
	Reference genomes	451	0	0	1
	Reference proteins	451	0	0	1
Psy642					
	Draft	449	0	0	3
	Reference genomes	449	0	0	3
	Reference proteins	449	0	0	3

Table 4.2 Number of annotated genes predicted for pre and post corrected assemblies using NCBI's PGAP (build 4456). Annotated genes are designated as complete, frameshifted, and incomplete. Incomplete genes encompass all gene pseudogenes, such as genes with an internal stop, missing start/stop codon, etc.

Assembly	Number of annotated genes		
	<i>Complete</i>	<i>Frameshifted</i>	<i>Incomplete</i>
Pss508			
Draft	3967	1101	133
Kastor correction (Genome)	4890	190	106
Kastor correction (Protein)	4754	298	141
Psy642			
Draft	4981	25	208
Kastor correction (Genome)	4982	21	213
Kastor correction (Protein)	4989	17	199

4.3.2 Assessment of detected errors using Kastor-P

Protein-based error detection is inherently more conservative than its genome-based equivalent

The magnitude of errors detected varies greatly between the two methodologies. Protein error assessment with Kastor-P detected one-third to one-fourth as many errors in both Pss508 and Psy642 (Table 4.3) and is algorithmically limited to genic region of the genome. For Pss508, approximately one third (900) of undetected errors were substitution errors, which are not resolved using protein based correction. For insertion and deletion errors, the protein-based assessment module detected half as many errors (1087 indels) compared to 2000 found with a genome-based approach. However, 683 of undetected errors with the latter method were found in intergenic regions of genome, which are not corrected by protein mapping. Among errors within genic regions, Kastor-P's method detected 1087 indel errors within mapped genes, 230 (18%) fewer than that obtained using the genome-based

method. The annotation difference of 136 with genes using PGAP is largely attributed to these putative false negatives (Table 4.2).

For the Illumina assembly of Psy642, Kastor-P detected higher than expected genic indel errors (44) compared to the detected 17 indel errors using a genome-based method (Table 4.3). If we consider all errors detected in the Illumina assembly as presumed false positives, the protein-based method has a higher maximum false positive rate at ~4% (44/1087) compared to the genome-based method's rate of ~1.3% (17/1317). Despite a low sequencing error rate of less than 1% [202], some of the errors identified in Psy642 may represent true finds. Among the 44 indel error corrections, 18 were within genes previously annotated as incomplete, and these corrections contribute an additional 8 complete genes in the annotated genome (Table 4.2). Conversely, 26 of the errors were found around previously complete genes, yet improved genome annotation of the post-corrected Psy642 assembly suggests the introduced true false positives may not be visibly detrimental due to external factors compensating for indels. For example, multiple start codons in all reading frames were found in the end of some genes. Genes with an introduced frameshifts by a false positive correction will still be in frame of a viable start codon, thus the gene will be falsely annotated as “complete”. All together, the more conservative detection of errors may be sufficient to complete a draft assembly, but placement of false positive suggest that precise error determination could be further refined in future implementations.

Comparison of errors detected between the two methodologies

Previously, with the Pss508 assembly corrected using the genome-based approach 1708 of the 2000 indel errors (85.4%) were deletions and 292 (14.6%) were insertions (Table 4.3). The higher frequency of deletions has been attributed to uncertainty in homopolymer length determination due to variation in the translocation speed through the nanopore [315]. Comparatively, the protein-based methodology may overestimate the amount of insertion errors with 240 (22.1%) insertion errors found in Pss508 while 24 (55%) of these errors were found in Psy642. However, both methods found that errors exceeding a single nucleotide were uncommon, where only 70 of the indel errors were two nucleotides long in Pss508 using mapped proteins. Similarly, 870 of 974 (89.3%) of mapped genes contained only a single identified error. Only 104 (10.7%) of the genes had 2 or greater errors, and

Table 4.3 A comparison of the number and location of errors detected using both Kastor error assessment modules.

	Assembly			
	Reference genome		Reference protein	
	Pss508	Psy642 ^a	Pss508	Psy642 ^a
Detected Errors				
Total	2900	195	1087	44
Substitutions	900 ¹	-	-	
Deletions	1708 ²	124	847	20
Insertions	292 ²	71	240	24
Indel Error Locations ³				
Within gene	1317	17	1087	44
Intergenic	683	178	-	-

^a Illumina draft assembly of *P. syringae* pv. *syringae* 642 (Genbank Accession: NZ_ADGB000000000) [18]

¹ Number of substitution errors with read support. Reduced from 24600 candidate substitution errors in Pss508. No substitution errors were supported for Psy642 due to a lack of read data.

² For reference-based correction of Pss508, the genome location of 561 (19%) detected indels were adjusted based on supplementary raw data during the error adjustment module.

³ Number of errors and genetic location based on gene annotation of pre-corrected assembly

nine (1%) genes had three errors. Therefore, the most common cases found with the protein-based approach are fragmented genes with a single deletion error one nucleotide long.

Lastly, another discrepancy between the two detection algorithms is the types of errors found. With the genome-based approach many of the errors in the Pss508 assembly were associated with homopolymers of A or T of four or more nucleotide, 2x greater than homopolymer runs of C or G (Figure 3.4). Somewhat surprisingly, errors detected using proteins were not particularly biased towards a given nucleotide. Out of the 405 errors associated with homopolymers of length four or more, 194 (47.9%) were adenine and thymine homopolymers. When broken down by nucleotide base, the homopolymers types are near equally prevalent at 106,104,107, and 88 for adenine, cytosine, guanine, and thymine respectively. This inconsistency suggests there may be differences in

positioning of the final error based on mapped reference proteins and genome based correction, but both methods were effective at defragmentation of genes in the draft assembly.

4.4 Discussion

Intended as a final error correction step in genome assembly, Kastor was designed to target insertion and deletion errors that commonly fragment long read draft assemblies using high quality external data (Chapter 3). However, one major concern of our previous genome-based approach was the limited availability of a diverse set of closely related genomes and can provide a challenge for sequencing projects with a broader scope, such as metagenomic studies. In response, the protein-based error assessment module was designed to utilize the expansive reference protein database to directly evaluate draft assemblies for gene fragmenting indel errors and to aid in downstream annotation analyses [238]. Utilizing this method, highly mobile genes, such as virulence genes, can be detected with homologs for more distant relatives. For example, a core effector gene, *avrE*, of the atypical type III secretion system in Pss508 was defragmented using a homolog belonging to a *Pseudomonas congelans* strain (classified as part of *Pseudomonas fluorescens* group) [19,316]. This gene was initially missed in the first pass of the genome-based approach from the lack of similarity to genomes within Pss508's subphylogroup (2c)[43,44], and this particular *avrE* gene variant is missing from the genomes belonging to its parent phylogroup (group 2) (Chapter 5). Therefore, poor sequence coverage prevented the assessment of errors in this gene.

Using an Illumina draft assembly as a control, the protein-based error assessment model was able to detect 1087 and 44 indel errors in Pss508 and Psy642 respectively. Correction of less than 2% of the genome can improve assembly completeness by an additional 14.4% on a gene level, to a final 99.8% completeness as assessed with BUSCO. However, a small number of putative false positive errors have been observed in the Illumina Psy642 draft assembly. Since fragmenting indel sequencing errors are relatively rare among Illumina sequences [202], we treated the 44 detected errors as false positives despite the slight improvement in genome annotation, where an additional 8 complete genes were identified (Table 4.2). In fact, an error was detected in complete gene, *prfB*, designated as "fragmented" due to programmed ribosomal slippage [317] therefore, instances where corrected genes use non-standard translations can lead to common cases of false positives. These natural intragenic insertions, commonly used for autoregulation, are absent in the final protein product and

which Kastor-P will detect as erroneous [318]. If it is assumed that all corrections to Psy642 are false positives, we can combine the 4982 draft genes and the average gene size of the mapped proteins, 1059 bp, to estimate an upper false positive rate. For the protein error assessment module, we determined that with a curated set of reference proteins, the software makes one erroneous correction per ~119,900 bp. Although the rate is lower than the genome assessment module's 1 per ~29,800 bp, the false positives of the protein module are limited to genic and more conserved regions of the genome and is likely an underestimate of the error rate.

Software use and limitations

Similar to the genome assessment module, the protein module is constrained by the same three factors: reference protein selection, reference quality and mapping quality. The current implementation requires a curated set of unique proteins for the genome. Errors found in genes with a high copy number [319] or non protein coding genes such as rRNAs are frequently missed and are a likely factor in inferior correction of the accessory genome. This poor correction of the accessory genome is missed by BUSCO (Table 4.1), a method that solely identifies single universal orthologs [266]. This discrepancy in the number of corrections between the core and accessory genome is further compounded by higher sequence conservation in tested single, universal orthologs of the core genome, which are the ideal cases for protein-based correction, where ease of high quality mapped protein enhances error detection by Kastor-P [276].

Distantly related homologs reduce the mapping quality between the reference and draft assembly, where sequence divergence can confound the location of candidate error regions for precise determination of error position, making it imperative that users provide high quality protein sequences that have been curated from closely related proteins. As an example, use of a more distant homolog of one reference protein can account for at least one false positive in Psy642. The sub clade encompassing both Pss508 and Psy642 has only been sparsely sequenced and thus close homologs are poorly represented in the RefSeq database [43,44]. A more distantly related TonB homolog was mapped as a single fragment to the Psy642 draft assembly at 86% coverage [320]. One end of the *tonB* gene remained unmapped due to sequence divergence. Without reference information, the unmapped edge was considered a candidate region with a likely error. Since the putative *tonB* gene length was determined from a more distant homolog, the correct in-frame start codon located further

upstream was ignored by Kastor-P in favour of a valid, but frameshifted start codon, which was identified nearest to the expected start site. The software then identifies a false error, and introduces an erroneous correction in this region to resolve the expected gene. Alignment of a close TonB homolog would likely have identified the full gene suggesting no errors in the region. Therefore, future implementation of conflict resolution via masking of regions can allow users to expand into larger reference protein database, such as NCBI's non-redundant database, for more homologous proteins and duplicate gene resolution.

An improved protein set can also improve mapping quality between protein and assembly. Poor protein mapping may be a significant contributor to the discrepancy in errors observed between the two modules. The current implementation of the protein module utilizes a strict rule set in determining the candidate error region. Depending on sequence similarity between the translated assembly and reference, the seed extension step of tblastn can extend beyond the most probable error position and mask it from detection [314]. Inaccurate detection of the candidate error region can impede proper positioning and ultimately introduce an erroneous correction. In these cases, the gene will be defragmented sufficiently for annotation and supported by positive BUSCO and PGAP annotation results, but a small section of the gene may have introduced errors. Therefore, draft assemblies with the current implementation of the protein-based can be sufficient for broad downstream analyses such as presence-absence based comparative genomics [321] rather than targeted studies such as multi-locus evolutionary analyses [307]. In summary, the current state of protein-based error assessment is successful as a proof of concept, but requires further improvements for use beyond as a benchmark for assessment of assembly fragmentation.

4.5 Future Directions

General improvements and streamlining the workflow

The current implementation of Kastor-P's error correction is general proof of concept on the capability of using widely available reference data to correct draft assembly without the additional sequencing that is common to modern hybrid sequencing projects [225]. Thus, this preliminary concept can be improved and expanded upon at both a general and detection level. One major general improvement with both the genome and protein detection methods is the implementation of multithreading and batching of data for faster processing and better scalability for larger datasets, e.g.

metagenomic sequencing. At the base level, the software treats each candidate error section (i.e., genome sequence windows and protein queries) as independent of other candidates and processed sequentially. Thus, input data can be done as batch analyses, reducing the required memory bandwidth with larger datasets.

A second major improvement to consider is the streamlining of the preprocessing methodology to reduce the onus on the user to manually curate their dataset prior to use and raise reproducibility. Currently, users are required to provide a curated single, unique collection of reference proteins mapped to their assembly genomes. This requirement would not be feasible for metagenomic projects with a wide range of contigs, e.g. *Candidatus* spp., without definitive references [207]. In these cases, the software can be bundled, or scripts provided, to implement a pipeline similar to the testing dataset (See section 4.2.1). Individual contigs are split and annotated with automated pipelines capable of detecting gene fragments such as PGAP [295] and DFAST [322]. Translated fragments can then be used to pull single semi-curated or closest matching reference proteins from NCBI for mapping with *tblastn* [314]. However, this approach requires heavy reliance on the stability of third-party availability of both software and API, which can be subject to change.

Incorporation of more generalized models or k-mer error profiles can expand and improve correction determination beyond initial heuristic methods

Kastor-P's current scope of error detection excessively relies on a priori knowledge reported on the Oxford Nanopore MinION sequencing technology [323]. The resultant heuristic error profile model is heavily biased towards errors associated with this sequencing system such as long homopolymer indel errors [232,315]. Datasets from projects where genomes were sequenced with a variety of sequencing technologies can be assessed against the finalized genome to extract error frequency distributions. Collectively, a diverse dataset on the error distributions can be used to generate a generalized weighted probability model to infer error nucleotide and position within a candidate error region. Alternatively, for single genome projects, a k-mer error profile approach, similar to Pollux [300], can be implemented to identify unique k-mers within an error candidate region that can signify the position of the error. In both cases, the implementation of methodology using empirical-based models will help avoid biases of errors detected and improve confidence in the results.

Chapter 5:

Divergence of subphylogroup 2c *Pseudomonas syringae* strains linked with horizontal gene transfer of atypical type III secretion system

5.1 Introduction

5.1.1 Chapter objective and overview

The rapid gene transfer, gain, and loss of virulence genes, resulting in hyper sequence variation, are key determining factors in *P. syringae*-host interaction and pathogenicity. In this chapter, we hypothesize that identification of divergent gene clusters, and comparison of virulence gene repertoire can provide insight into the pathogenic capabilities and ecological niche of biocontrol strain *P. syringae* pv. *syringae* 508 [7] and *P. syringae* TLP2 [256]. Geneset comparisons with 411 other Psy strains, 7 *P. congelans* strains, and 10 *Pseudomonas* spp. (outgroup), confirm that Pss508 and PsyTLP2 formed a unique monophyletic subphylogroup 2c with two other Psy strains and *P. congelans* strains. In this work, we expanded our analysis to divergent clusters unique to this ambiguously pathogenic subphylogroup. At least 10 of approximately 50 genes unique to the group were found to be divergent genes shared with phylogroup 13 strains at lower clustering thresholds. When mapped to our Psy assemblies, these genes were identified to encode the AT-PAI T3SS variant. With a combination of MLSA, and geneset clustering, our data suggests that subphylogroup 2c likely indirectly inherited this pathogenicity island (PAI).

Based on comparison of PAI characteristics including genomic location and gene arrangement, we propose that the atypical T3SS is a third major T3SS variant with hybrid characteristics shared with the two more common variants, T-PAI and S-PAI. This data suggests that subphylogroup 2c horizontally inherited the entire AT-PAI rather than simple recombination of distantly related T-PAI. From our gene level analysis, we determined that acquisition of this AT-PAI is correlated with a major loss of a large number effector genes in subphylogroup 2c, with remaining identified conserved effector genes attributed to be niche factors. Effector gene loss does not extend to more general modes of virulence.

On direct comparison between our sequenced Psy strains, we found that at least one putative T3 effector was found in one but not the other, and vice versa suggesting pathogenic capability. As putative targets of host detection, these genes alone do not explain the variation of pathogenicity within subphylogroup 2c (see appendix A). In summary, this work speculates that AT-PAI acquisition in the MRCA of subphylogroup 2c potentially triggered divergence from phylogroup 2, and loss of many effector genes. But retention of generalized virulence also suggests a possible niche change to similar to other weakly virulent *P. syringae* strains [29].

5.1.2 Chapter background

Pseudomonas syringae is one of the most ubiquitous phytopathogens. It has a broad range of hosts and causes a variety of crop diseases including halo blight, fruit speck, and fruit rot [4]. Although many *P. syringae* infections have been attributed to sowing of contaminated seeds, the presence of the bacterium has been reported in water samples outside of agricultural sites suggesting close ties to the water cycle [15,58]. With more than 28 known plant hosts, this bacterial species complex poses an ongoing threat to agricultural crops [4]. When given favourable conditions, the bacterium can quickly spread causing large disease outbreaks for example in fruit crops such as kiwifruits [324] and mangos [39].

The broad host range of the bacterium is not a distinct characteristic of particular strains but is a result of the diversity of virulence of individual strains within the species complex [44,325]. Constantly adapt to plant host immunity, individual *P. syringae* strains typically exert virulence to a specific host through an array of virulence genes injected into host cells via the type III secretion system (T3SS) [326,327]. Encoded in a pathogenicity island (PAI), the mobility genes associated with the T3SS secretion system makes it a common form of pathogenicity in bacterial pathogens [151].

Across different bacterial pathogens, these genomic islands can be highly diverse in both gene sequence and arrangement, but the diversity within a bacterial species is usually limited to a small number of variants. *P. syringae* broadly show two overall variations in arrangements of genes within the same gene clusters of the PAI [151,154]: the canonical tripartite PAI (T-PAI), , and its atypical subvariant AT-PAI both of which has a conserved core gene cluster flanked by two effector gene loci, whereas the single partite PAI (S-PAI) has an inverse gene arrangement where the effector loci is flanked by a split core gene cluster [152]. Other slight variations in S-PAI include a missing

duplicated regulator gene and a fused conserved gene in the S-PAI. Despite high genetic diversity in many T3SS PAI genes, the core gene cluster encodes a highly conserved injection-like apparatus constructed from the same corresponding gene in each PAI [88].

When the bacterium comes in contact with the host cell, T3SS expressed harpin proteins dimerize and bind to the host cell, forming a pore [328]. Pili-subunits HrpA proteins are then recruited and assembled into a needle-like pilus that bind to this pore creating a channel to inject bacterial proteins directly into host cells [329]. The HrpA protein forms the external portion of the secretion apparatus and has a higher genetic variability than other T3SS proteins, likely in order to evade detection from host plants [88]. The T3SS is initiated by expression and dimerization of global regulators HrpR and HrpS which subsequently induce expression of downstream genes [168]. Due to the complexity of the secretion apparatus, the precise timing of gene expression is vital resulting in a highly controlled T3SS regulon by more than 7 PAI genes that regulating expression and protein localization to ensure accurate gene expression during early, middle, and late stages of construction [88,330]. After assembly of the secretion apparatus, full expression of master regulator HrpL [166], activates gene expression of the full suite of *hrp* regulon comprising of genes encoding the secreted T3SS effector proteins (T3Es) and other virulence genes.

Injection of the expressed T3Es help pathogens modulate the target host environment by regulating host gene expression and production through mimicry, binding, and/or catabolism, consequently increasing a nutrient rich intracellular environment while hindering plant immunity [331]. Conversely, host plants can indirectly detect the phytopathogen via self-regulation and identification of anomalies via Resistance (R) proteins that can act as decoys or guard effector-bound host proteins [121]. As postulated by the zigzag model, detection via R-proteins triggers upregulation of plant immunity leading to an alternating exchange of activation and suppression of immunity between host R-proteins and pathogen effector proteins [75]. Due to necessity of protein binding between effector and R proteins, both proteins are typically highly specific, forming pairs with the effector protein [332]. Effector genes recognized by a known R-protein are considered avirulence genes (*avr*): recognition of T3 effector *avrRPM1* by R-protein *RPM1* renders *P. syringae* expressing the gene avirulent [188]. Thus to evade plant immunity, it is highly advantageous for the bacterial pathogen to carry an effector repertoire composed of T3Es that can avoid recognition by target host R-proteins [121]. However, a large effectome is also energetically unfavourable to maintain and regulate for the pathogen. An average effectome size of 30-50 effector genes is frequently observed in

well-characterized *P. syringae* strains [24,93] and is regulated through the singular sigma factor HrpL. Due to constant competition with plant immunity proteins, the *P. syringae* effectome is a continuously adapting to the evolving host R-protein repertoire, through gene exchange, gene loss, and gene gain of effector genes [193].

Loss of a detectable avirulence proteins in the bacteria can prolong evasion of the host immune system. Conversely, a previously avirulent *P. syringae* strain can gain virulence by acquisition of a gene through recombination or via a novel gain-of-function via SNPs [333], deletions [334], and transposable elements [335]. The gain of an advantageous mutation has been seen to cause a selective sweep in pathogen population [336]. Curation of the pathogen effectome to match a particular host leads to strong host-specificity, though in many cases the host can remain unknown for less well-characterized *P. syringae* strains [43,57].

Attempts to distinguish *P. syringae* groups based on correlation between the phenotype of host-specificity and genetic features have proven difficult. At the sequence level, previously distinct species, such as *Pseudomonas amygdali* and *Pseudomonas savastanoi*, were found to be closely related and collapsed into the *P. syringae* species complex based on core genome content [44,337]. Strains belonging to different pathovars, assigned based on the strain's host range, similarly form monophyletic groups based on both sequence identity and gene content. Analysis of the high recombination rates of a small set of core genome genes within some strains indicate a high level of intra-phylogroup homologous recombination compared to relatively rare inter-phylogroup transfer [338]. The robust division of major phylogroups supports the variation in recombination based on genetic distance where increased sequence mismatch can reduce the probability of inter-phylogroup transfer [339].

Distinguishing host-specificity using genetic content is further confounded by the presence of alternative modes of virulence with similar functions. Among later branching strains of the *P. syringae* complex, classified as phylogroup 2 based on core genome analysis [44,57], some strains are considered generalists without a strong host specificity. Reduced host specificity was correlated with a reduced effectome size [57,340,341]. The loss of effector genes can be compensated by addition of phytotoxin biosynthesis genes, expression of non-ribosomal peptides and polyketide synthetase [99]. These alternatives can serve similar virulence function by targeting generalized plant innate immunity (e.g. molecular-pattern mediated immunity), such as the prevention of stomata closure rather than

specific host targets [100,342]. However, generalist pathogens often still rely on an active albeit reduced effectome to target the later stages of plant immunity, the effector mediated immunity [75]. Not all effectors directly manipulate plant immunity however, with some highly conserved effectors improving survivability within the host plant environment. This may indicate T3 effectors can be a niche factor rather than a virulence factor [80,104,343]. Therefore, gene content can be informative of general niche adaptations, rather than as a specific host-plant identifier.

There is a relatively recent push to increase sequencing of the *Pseudomonas* species complex including less characterized phylogroups, including putatively non-virulent *P. syringae* strains and environmental strains of uncertain virulence [18,22,29]. Previously lack of virulence was attributed to missing T3SS components which prevented infection, but later studies refuted this and identified T3SS in these strains [18,21]. For some non-virulent strains, deleterious mutations and embedded transposable elements can prevent proper T3SS expression, however these strains can easily revert back to virulence [22]. Lack of detectable virulence in other strains has been attributed to differences in T3SS expression, where a strain is highly virulent in combined inoculations with more virulent strains but weakly virulent on its own [30]. Some *P. syringae* strains are only putatively non-virulent as they don't have an identified host, primarily for strains isolated from environmental samples and those within less characterized phylogroups [43,44,49]. Strains of atypical sub-phylogroup 2c (PG-2c) in the species complex were identified with active T3SS expression, but activity was only confirmed through introducing known avirulence T3 effectors via mutagenesis [18]. Unusually, the T3SS has been reported as a novel atypical variant with high sequence divergence in core conserved T3SS genes despite the common tripartite gene arrangement (AT-PAI) of the encoding pathogenicity island [151]. Inoculation at high levels with some of these native strains does not trigger hypersensitive host response despite proliferation within the sample host [18]. Currently, none of these strains have identifiable virulence and at least one strain was initially considered as a fungi-targeted biocontrol strain of interest [8].

5.2 Material and Methods

Genome selection and pangenome analysis

A collection of 428 *Pseudomonas* spp. genomes was collected from NCBI [344]. The majority of the genomes are predominately closely related *P. syringae* genomes (n = 411), based on

sequence relatedness of these genomes, to the target group of the work (subphylogroup 2c) [48]. Many of these genomes are classified as part of genomespecies I group in NCBI, an older sub-group classification of *P. syringae* based on DNA-DNA hybridization[48]. Supplementary reference strains, belonging to alternate genomespecies, from missing phylogroups were included to ensure representation of the broad monophyletic groups [44]. Two phylogroups, namely 8 and 12, were excluded due to poor sequence availability. *P. congelans* strains (n= 7), under the *P. fluorescens* species complex, were also included due to high sequence similarity of core T3SS genes [19,49]. Lastly, an outgroup of 10 strains from other major *Pseudomonas* species, including *P. aeruginosa*, *P. fluorescens*, *P. putida*, were used to root the core genome phylogeny.

Each genome was re-annotated using DFAST [322], tailored specifically towards *Pseudomonas syringae*, to standardize the files into a readable format for downstream software. Gene families and clustering were then identified using PIRATE [321]. Default percentage identity thresholds (50%, 60%, 70%, 80%, 90%, 95%, 98%) were used during de-clustering steps with higher stringency. InterPROSCAN was then used for functional annotation of the identified gene families on a representative protein product from each group [345]. Representative proteins were selected based on longest product to ensure identification of all possible domains. Collectively, this data, along with strain data, were consolidated into a SQLite database for easy parsing of gene families in relation to strains and gene loci [346]. Through PIRATE, the 'hard'-core genome (genes present in all strains) was aligned using MAFFT and then concatenated [347]. Due to alignment length, the core genome phylogeny was generated using fastTree2 [348] using the 10 outgroup strains to root the tree.

Shared gene sets between *P. syringae* strains were analysed using the R package Hierarchical Sets [349]. The ordered gene families and respective genome was pre-processed into binary presence and absence data format using R. To reduce biases towards larger or more divergent genomes, singletons (i.e., unique gene families from a single gene and genome) were removed before analysis. Gene presence threshold was reduced to 95% to account for missing genes in draft genomes due to contig fragmentation.

Phylogeny analysis of the Type III secretion system

Separately, T3SS protein sequences were aligned using MAFFT [347]. For HrpR and its paralog HrpS, only the former was aligned due to higher similarity to HrpR in T3SS pathogenicity

islands with a single homolog in multi-locus sequence analysis of T3SS genes [170]. In the single gene tree analysis of HrpR genes, HrpS was also aligned as a separate gene due to known homology with HrpR. Some other T3SS genes, i.e., HrpA and HrpF, had high divergence in protein product sequence (<50% identity) resulting in multiple gene families, and alignments for these sequences were generated according to gene family clusters. Each multi-sequence alignment was manually adjusted to remove large uninformative gaps. Minor gaps were retained to maintain higher diversity for highly conserved *hrc* genes. A multi-sequence alignment was generated for the T3SS phylogeny by concatenating *hrpR* and *hrc* genes present in all strains (i.e., *hrpR* with *hrcJUTSRQN* and *V*). All alignments were then tested for the best evolutionary model using ModelTest [350]. A maximum likelihood tree was then generated for each alignment using RaxML with the selected evolutionary model [351]. Support for the consensus tree was created with 1000-3000 bootstrap trees depending on the convergence of the consensus tree with bootstrap trees.

T3SS *hrp* regulon analysis

To identify the *hrp* regulon, a hidden Markov model of the canonical reference *hrp* promoter was developed from a collection of 216 known *hrpL* promoter boxes from the three *P. syringae* reference strains PsyB728a, PtoDC3000, and Pph1448 that was extracted from www.pseudomonas-syringae.org. The HMM model was generated using MEME [352] and incorporated spacer size to minimize the effect of specific spacer sequences. Prior to promoter prediction, the intergenic region including 100bp of downstream genes in each *P. syringae* genome was extracted to reduce the search space. Overlapping intergenic regions spanning very small genes (<100bp) were merged to avoid repetitive scanning of the same sequence. After extraction, the *hrp* HMM was used to scan these intergenic regions with FDR q-value cutoff of 0.1.

Visualization and Descriptive statistics

Visualization and data analysis were done in R using a collection of tidyverse packages (dplyr v1.0.2, tidyr v1.1.2, readr v1.4.0, stringr v1.4.0, and ggplot2 v3.3.2)[353]. Data was extracted from the SQLite database using R packages RSQLite (v 2.2.1), DBI (v1.1.0), and dbplyr (v1.4.4). All phylogeny trees were visualized and annotated using R packages ape (v5.4-1) [354], tidytree

(v0.3.3) [355], and ggtree (v2.3.7)[356]. Dendrograms and comparisons between them were visualized using phylogram (v2.1.0) [357], and dendextend(v1.14.0) [358]. Combined figures were generated using a combination of ggplot2, ggpubr(v0.4.0)[359], and egg (v0.4.5)[360]. T3SS gene cluster images were generated using draw.io and inkscape. Colours were chosen using Hues (v0.2.0) [361] to select distinguishable, colour-blind friendly colours using K-means clustering.

5.3 Results

5.3.1 Pangenome Overview

The pangenome of a collection of 428 *Pseudomonas* spp. strains was clustered and evaluated using the PIRATE software (Pangenome iteration Refinement and Threshold Evaluation) [321]. This collection includes 418 strains within *Pseudomonas syringae* (Psy) species complex, plus an outgroup of 10 other reference *Pseudomonas* spp. strains. The diversity of the outgroup helps ensure the accurate placement of Psy strains including those previously identified as other pseudomonads (i.e., *P. congelans*), and highly divergent strains belonging to the earliest branching monophyletic group [19,362]. With PIRATE, the proteome for each strain was compared in a pairwise blast search at multiple percent identity levels starting at 50% identity and proceeding upwards to 98% identity [321]. Here, clustered genes are represented as gene families, especially at lower identities, rather than orthologs.

Core genome phylogeny

The overall pangenome results for *P. syringae* identified 37282 gene family groups. 2968 of these contained a fission/fusion event while 2056 contained a duplication/loss of genes. 3085 gene families were identified as part of the ‘soft-core’ genome, appearing in more than 95% of the strains. PIRATE also identified that 998 gene families of the soft-core genome were multiple copy genes, which is expected in gene families such as rRNA genes [319]. From the identified gene families, a subset of 2087 single copy genes found in all strains, or the ‘hard-core’ genome was used by PIRATE to generate a core genome alignment and a FastTree phylogeny using SH-like support (Figure 5.1). The tree was then rooted using the 10 outgroup reference strains. From the tree, the 428 *Pseudomonas* strains were grouped into 11 out of the 13 main phylogroups and one large outgroup consistent with

previous work [44]. Phylogroups 8 and 12 were not populated due to missing genomes or poor genome sequence quality and were filtered out.

From this data, I was able to confirm that both sequenced strains were grouped the last branching phylogroup known as group 2 [44]. Furthermore, with greater than 98% similarity to *P. syringae* 642 (Psy642), I also confirm that these two strains cluster with Psy642 and one other *P. syringae* strain forming the earliest diverging monophyletic cluster from group 2, previously classified as the subphylogroup 2c [18].

Unexpectedly, a few *P. syringae* strains were clustered with outgroup species. These were designated as outliers and excluded along with the outgroup from further analysis. One possibility for the presence of strains outside the standard phylogroups is a result of changes to classification methodology over time with more precise techniques, e.g. biochemical (DNA-DNA hybridization) vs. genetic methods (core genome phylogeny) [48,363]. One recent and prominent example is the grouping of *P. congelans* strains within sub-phylogroup 2c as part of late branching *P. syringae* strains compared (Figure 5.1)[18,44,49]. Initially, *P. congelans* isolates were classified as part of the *P. fluorescens* group based biochemical identification (LOPAT system) of pyoverdine synthesis, a fluorescent pigment [19]. However, concordant with previous studies, the present genome analysis places *P. congelans* isolates consistently within *P. syringae* subphylogroup 2c with the three known strains of the subphylogroup (Figure 5.1), averaging approximately 98% similarity (data not shown)[7,18,256]. All 7 sequenced *P. congelans* strains available in NCBI in this work along with reference *P. fluorescens* strains as an outgroup were incorporated in this work confirming that *P. congelans* species are exclusively part of subphylogroup 2c within the *P. syringae* species complex.

From metadata available on NCBI, subphylogroup 2c strains were either isolated from health plant tissues, or environmental samples, indicating that this understudied subphylogroup of ambiguously phytopathogenic strains represents a possible transition in phenotype after divergence from its closely related virulent subphylogroups [7,19,21,66]. With the evolution of species classification, discontinuity in these genomes and classification suggests reclassification should be considered for *P. congelans*. Consequently, this data suggests that genetic factors, i.e., divergent gene clusters, unique to this subgroup may correlate with this phenotypic transition.

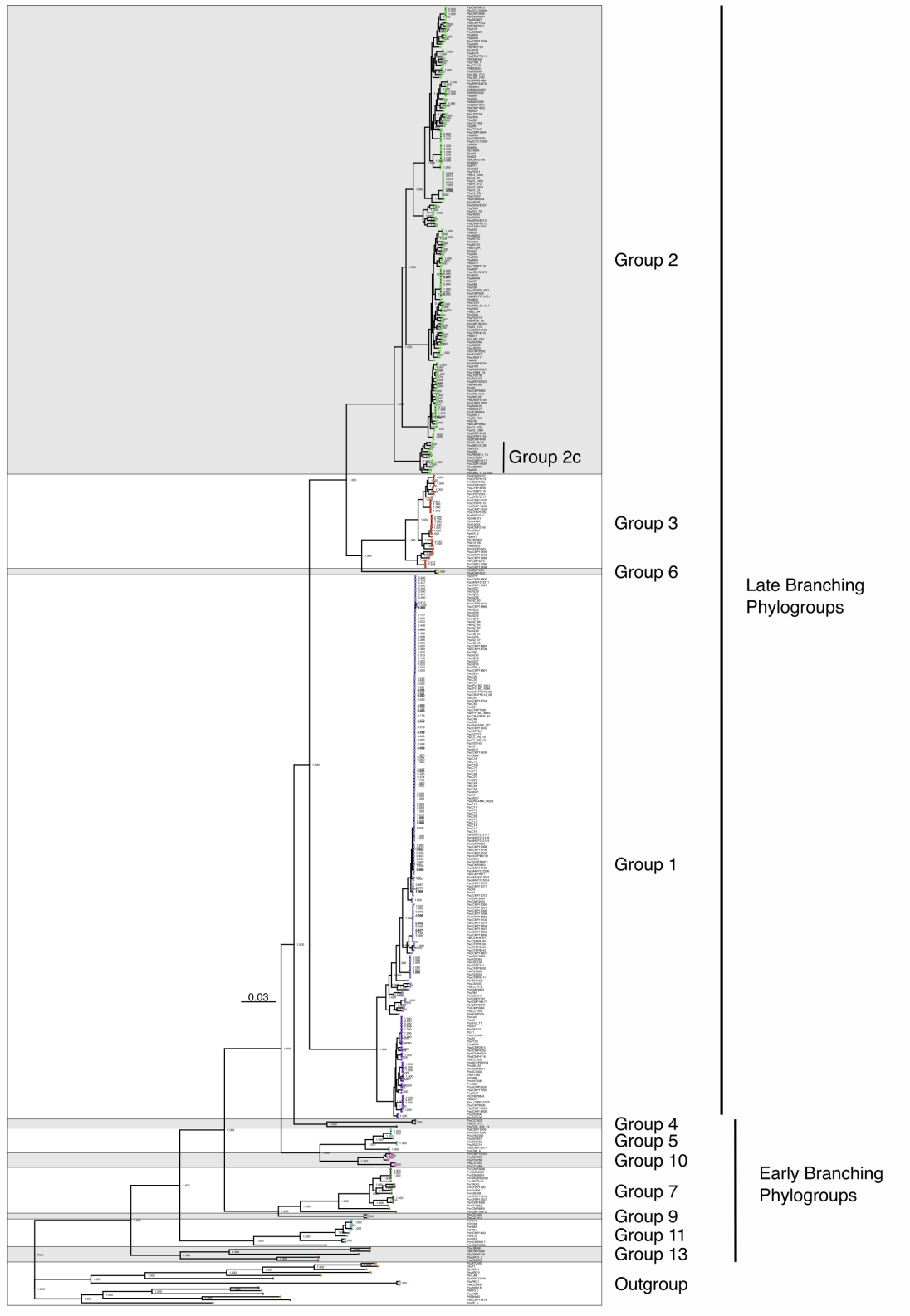


Figure 5.1 Phylogenetic tree generated from 2087 single copy gene families using FastTree with SH-like support. Each phylogroup is highlighted in alternating boxes with the corresponding group number on the right. These phylogroups can be described as two broader sets of strains representing the early branching strains (i.e., secondary phylogroups) and late branching strains (i.e., primary phylogroups). Placement of *P. syringae* strains Pss508 and PsyTLP2 along with *P. congelans* and *P. syringae* in subphylogroup 2c. For readability, an enlarged version can be seen in Figure C.1(a-d).

5.3.2 Identified divergent gene cluster of subphylogroup 2c involves the T3SS

Accessory genes unique to subphylogroup 2c

In comparison, 52 gene families were found to belong exclusively to sub-group 2c (PG-2c) indicating that these gene families accounting for > 50% of all gene families unique to phylogroup 2 as a whole. Although this sub-phylogroup shares the most recent common ancestor with the remainder of phylogroup 2, these PG-2c specific genes are one of the predominant factors that separate the subgroup from other group 2 strains. If these genes are removed from the analysis subgroup 2c strains cluster slightly more closely with other group 2 strains, as much of the remaining accessory genome is shared (data not shown). Using positional data from the sequenced complete assemblies (Chapter 2 and 3) to determine the location of these unique gene families in PG-2c, many of them were clustered within the same genomic location in both Pss508 and PsyTLP2 indicating this region contains a divergent cluster of interest. This suggests these genes are a key factor that separates PG-2c from other group 2 strains. Conversely, the other prominent distinction of this subgroup is the lack of accessory genes common to later branching strains, such as highly conserved effector genes [151]. These missing accessory genes may account for the shift in its putative phenotype, a reduced or absent virulence in phytopathology of PG-2c relative to the primary phylogroups that are populated with agriculturally damaging and archetypal *P. syringae* strains [44,158,324,338].

Divergent gene cluster in subphylogroup 2c is shared with some phylogroup 13 strains

Horizontal gene transfer can be highly influential on the niche of *P. syringae* strains, aiding in the gain and loss of key genes involved in host invasion and evasion [325]. Therefore, sequence divergence of conserved core genome may not correlate well to its virulent phenotype. To investigate this possibility we clustered *P. syringae* strains by gene content at different protein identity levels and

compared this to the core gene phylogeny. To match previous standards of pangenome practices for closely related genes [206], we first split the gene families into smaller gene clusters at 90 percent identity, a common threshold for qualifying orthologs in closely related strains. Genes were then clustered using a sequential agglomerative method, where strains sharing the greatest shared gene sets are clustered first, from Hierarchical sets R package [349]. This method was then repeated at different protein identity levels. Comparison of strain grouping between 90% and other identity level revealed that subphylogroup 2c and a subset of phylogroup 13 strains cluster below 90% together suggesting shared divergent genes between the two group (Figure C.3). When the shared gene families from both groups were extracted, at least 10 genes of these genes were previously mentioned as “unique” gene families to subphylogroup 2c at the ortholog level (pid = 90%). These gene families showed homology ranging between 50-70% identity and are predicted to encode hypothetical protein.

Divergent gene cluster identified as part of T3SS encoding pathogenicity island

Across the *P. syringae* species complex, each strain harbours at least one of the T3SS variants encoded in pathogenicity islands. Similar to previous studies, the subtypes found in our dataset include the canonical *P. syringae* tripartite pathogenicity island (T-PAI), the single partite pathogenicity island (S-PAI), and the most recently reported atypical tripartite variant (AT-PAI) (Figure 5.2). Each of these gene clusters is capable of translocation of effector proteins [88], indicating plant-bacterial communication even among putatively non-phytopathogenic strains [18,150]. We targeted this particular cluster to identify gene relationships between shared PAIs especially when gene mobility of exchanged effectors can influence interaction with host plants [156].

Among the T3SS variants, the canonical T-PAI is the major subtype found in *P. syringae*, appearing in 378 strains (90.4%) within our dataset. Most of these strains belong to strains that diverged after phylogroup 7 with one exception (Figure 5.1). Phylogroup 2c deviates from these late branching strains by harbouring a reported the atypical and genetically divergent T-PAI subtype (Figure 5.3)[18]. Originally, this variant was endemic to phylogroup 2c, however, when the identified divergent genes was paired with positional data, I was able to determine that these hypothetical genes were the AT-PAI equivalent of the non-conserved *hrp* genes (green genes in Figure 5.2). I was also

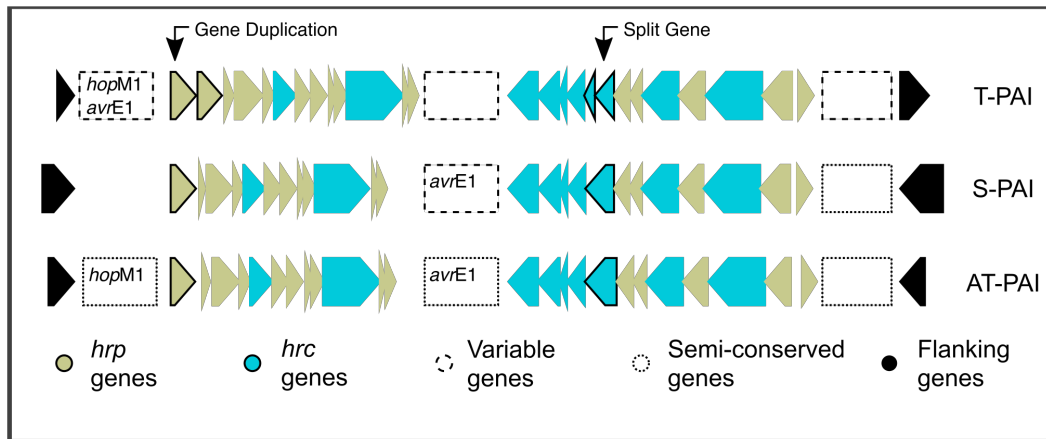


Figure 5.2 Common variants of the observed type III secretion system gene clusters: the tripartite T3SS pathogenicity island (T-PAI), the singular T3SS pathogenicity island (S-PAI), and the atypical T3SS pathogenicity island (AT-PAI). Full PAI of each subtype highlighted based on gene type; *hrc* genes encode conserved core genes; *hrp* genes encode non-conserved core T3SS genes. Effector regions are boxes denoted by dashed borders; gene variability in the region is denoted by the dash type in the legend. The conserved effector loci (upstream) of T-PAI were found to be variable in a small subset of *P. syringae* genomes but otherwise are conserved in the majority of strains. Gene differences between the PAI variants are denoted by a bold black border. For T-PAI, global regulator *hrpR* underwent gene duplication into *hrpS* (Green) and the singular gene *hrcQ* was split into two genes encoding subunits, *hrcQa* and *hrcQb* (blue). Differences in gene arrangement of highly conserved effector genes are labelled within the respective effector region. Gene *hopM1* is missing in S-PAIs.

able to confirm that all core T3SS genes in the PAI for both subphylogroup 2c and two strains of group 13 clustered at below 70% identity suggesting homology across the T3SS gene cluster. Above 70% identity, the conserved core T3SS (structural) genes were shared across all strains, but the less conserved core T3SS formed distinct gene families. However, the gene organization of these shared gene families was confirmed to be identical to AT-PAI including an *avrE* gene, a pseudogene in group 13 strains, residing within the hypervariable region situated at the center of the PAI (Figure 5.2) [364].

This gene cluster was further confirmed to match AT-PAI using the MLSA analysis of conserved T3SS *hrp/hrc* genes (Figure 5.4). Although discovered in two distinct phylogroups, this T3SS subtype only appears in a subset of group 13 strains within this dataset and is the least prevalent

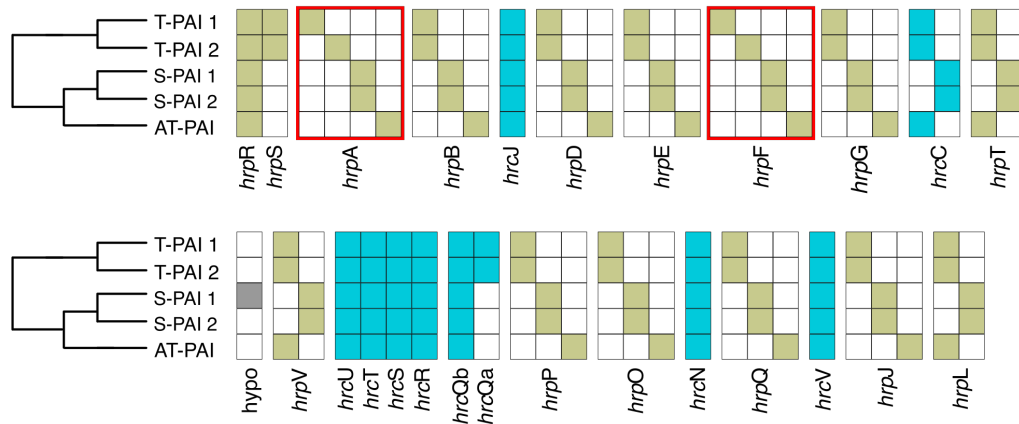


Figure 5.3 Representative diagram of hierarchical clustering of *P. syringae* strains by T3SS PAI genes with the respective presence (coloured)/absence (uncoloured) of shared gene families. Clustered phylogroups (PG): T-PAI1 (PG-1,4,6,9,10), T-PAI2 (PG-2,3,5,7), S-PAI1 (PG-7), S-PAI2 (PG-11), and AT-PAI (PG-2c,13). Functionally equivalent T3SS genes with less than <50% protein product identity were grouped as separate gene families with PIRATE. Strains are grouped into five clusters reflecting T3SS subvariants and sequence divergence of select T3SS genes. High sequence divergence of genes separates functionally equivalent T3SS genes into separate gene families. Clustered strains are incongruent with core genome phylogeny (Figure 5.1) with T-PAI subvariants grouping more distant phylogroups together. Among the genes, both *hrpA* and *hrpF* are the most divergent genes separating the PAIs into 4 distinct PAI groups, T-PAI1, T-PAI2, S-PAI, and AT-PAI (see Figure 5.7 and 5.8 for details on strain clusters). S-PAI can further be divided into S-PAI1 (phylogroup 7) and S-PAI2 (phylogroup 11) by the presence of an additional hypothetical gene present only in S-PAI1. AT-PAI clusters between the two major variants: it shares 4 more gene families with T-PAI (*hrcC*, *hrpT*, *hrpV*, and *hrpL*), but clusters closely with S-PAI due to a lack of the global regulator *hrpS* gene family and has higher similarity in the conserved *hrcQ(b)* gene family. S-PAI is the most genetically divergent with highly conserved T3SS *hrcC* gene clustering as a separate gene family.

of the T3SS variants. The remaining group 13 strains did not contain any other PAI variant indicating neither a loss of T3SS capabilities in the remaining group 13 nor a gain potential T3SS-based virulence in these select group 13 strains.

The S-PAI variant also appears in only two phylogroups, the closely related groups 7 and 11 (Figure 5.4) [154]. The S-PAI in both phylogroups are highly similar, except S-PAI in phylogroup 7 contains an additional hypothetical protein situated between *hrpT* and *hrpV* (Figure 5.3). Furthermore, within our dataset, group 7 strains do not exclusively contain this T3SS subtype and are interspersed in group 7 with these strains forming a polyphyletic group in the core gene phylogenetic tree (Figure

5.1 and Figure C.1d compared to Figure 5.4). The other strains belonging to group 7 are consistent with the more canonical T-PAI subtype [151]. In fact, each *P. syringae* strain in our dataset has only one of three aforementioned T3SS subtypes, consistent with the T3SS variations being mutually exclusive [150].

When gene organization of the PAIs was assessed, all three variants share very similar gene arrangements suggesting that all subtypes are derived from the same common ancestor (Figure 5.2). In contrast, a secondary T3SS with an entirely different gene structure has been observed in *P. syringae* strains primarily belonging to phylogroup 3 and a subset of group 13 [153]. This secondary gene cluster contains features similar to *Rhizobium* T3SS, and a gene organization distinct from the primary T3SS suggesting a separate evolutionary path filling functionally different niche similar to the two main T3SS variants of *Salmonella* [87]. When taken together, the incompatibility and prevalence of primary T3SS subtypes suggests that the three variants share the same functional role in niche adaptation, and is the primary form of effector protein translocation in *P. syringae* [88]. The specifics of these different T3SS may provide insight into differences in niche adaptations.

5.3.3 Horizontal gene transfer of T3SS variants

Acquisition of the canonical T3SS, the tripartite pathogenicity island (T-PAI) within *P. syringae*, is confounded by inconsistency that has arisen in part from different approaches to species delimitation [49]. Consequently, PAI variants across species form multiple patchworks of polyphyletic groups depending on study gene selection [43]. Similar to previous studies, there were incongruencies between core gene phylogeny and our MLSA of conserved core T3SS genes. This incongruency between our phylogenetic trees suggests multiple points of potential PAI acquisition indicating horizontal gene transfer between phylogroups as one of the confounding factors. For example, I found that of the five earliest branching strains (phylogroup 13), three of the earlier diverging group 13 genomes lack any discernable T3SS related to the three T3SS variants, whereas the two later branching strains forming a monophyletic group (PgIICMP2236 and PsyUB246) maintaining the non-canonical, atypical variant found only in late-branching subphylogroup 2c [18]. This divergence suggests a likely horizontal inheritance of PAI along later branches within group 13, rather than a loss.

I investigated the possibility of horizontal gene transfer of the homologous AT-PAI found in both subphylogroup 2c and group 13 strains by analysing the distribution of T3SS within the *P. syringae* complex. Multi-loci of conserved *hrp/hrc* genes were extracted, aligned in RAxML (model GTR+I+G4, bootstrap n=3000), and analysed. The visualized multi-locus sequence tree suggests that the distribution of T3SS variants is incongruent with the core genome phylogeny (Figure 5.1/Figure C.1d and Figure 5.4) indicating that horizontal acquisition of a T3SS PAI has potentially occurred at least three times and horizontal gene transfer for PAI variants has occurred more than once. With the prevalence of the canonical T3SS and congruency in phylogroup partitioning in both trees, gradual sequence divergence of conserved core T3SS after separation of phylogroup 9 suggests that acquisition the main variant (T-PAI) occurred in singular event after branching from phylogroup 13 (Figure 5.4). However, the lack of data for some of the early branching phylogroups (i.e., 8 and 12) makes elucidating the dissemination of T3SS across the species difficult. Whether the canonical T-PAI appears prior to phylogroup 11 could not be discerned due to the lack of a sufficient quality phylogroup 12 genomes. Without an identifiable ancestral variant, the earliest likely point of the canonical T-PAI gain would be post divergence all phylogroups from the MRCA at phylogroup 9. However, the presence of the strains containing PAI variants forming polyphyletic clades (on the core genome phylogenetic tree, Figure 5.1/C.1) obscures the evolution of T3SS PAI across the species complex.

Assuming acquisition of T-PAI in the MRCA of phylogroup 9, the acquisition of AT-PAI likely has happened at least twice among *P. syringae* strains, once for the two most divergent subphylogroup 13 strains and late branching subphylogroup 2c respectively. In addition, the atypical variant is concentrated within distinct subgroups of larger phylogroups suggesting that acquisition occurred within an ancestral sub-clade (Figure 5.4). Furthermore, 35 (67%) of the unique subphylogroup 2c gene families, previously identified in the phylogroup gene set analysis (Figure C.2), are these highly differentiated AT-PAI T3SS genes and are the only unique gene families found in all subphylogroup 2c strains (Figure C.2 and Figure 5.3); group 13 AT-PAI genes clustered as separate gene families at <70% identity. The remaining 17 (33%) unique gene families are found in smaller subset of subphylogroup 2c strains. Taken together, it is likely sub-specification of this group can possibly be defined by the acquisition of the AT-PAI as the unifying characteristic of this subphylogroup.

Even though both subphylogroups 2c and 13 share the same T3SS variant, clustering together in both conserved *hrc/hrp* consensus tree (Figure 5.4) and T3SS gene set clustering of the entire T3SS loci (data not shown), the AT-PAI in subphylogroup 13 lacks a discernable CEL, the highly truncated EEL and a short hypervariable region primarily comprising of the T3E *avrE1*. In fact, removing the core T3SS genes, hierarchical clustering of the known PAI genes, both subgroup form distinct clusters lacking any shared effector genes (data not shown). The only shared effector gene *avrE1* in subphylogroup 13 is truncated. Regardless, the protein product of this gene fragment share higher sequence similarity between both subphylogroups than the canonical T-PAI gene variant effectively differentiate this PAI in group 13 as a pseudo-single partite variant of the tripartite AT-PAI. However, without the external effector loci, this difference indicates that a direct transfer between the two subgroups is unlikely and that the MRSA for each group acquired the AT-PAI as either a single or tripartite PAI from an external source.

In both cases, a survey of the upstream and downstream genes supports this hypothesis as both PAI locations differ from the T-PAI defined as between genes *gstA* and *queA* [151]. We determined the flanking genes by searching upstream and downstream of genes *hrpR* and *hrpL* for the either canonical flanking genes or a ubiquitous core genome gene unassociated with T3SS, whichever gene is nearest to the respective T3SS core gene. Subgroup 13 AT-PAI is located in a different genomic location between a *lysR* regulator and a peroxide stress gene *yaaA* [365] while for subgroup 2c the AT-PAI is situated between a gene encoding a bile:sodium symporter and an *rdgC* homolog, the recombination dependent growth gene [366]. For the latter, all 2c genomes indicated that AT-PAI was directly inserted in the current position when compared to the uninterrupted region in closely related PsyB728a reference strains. Conversely, the canonical T3SS genomic location of PG-2Cc contains remnants of the T-PAI CEL (n = 8, 53%), specifically the non-T3 effector genes, suggesting almost direct excision of the genomic island. Loss of T-PAI following acquisition of the AT-PAI indicates that the variants are incompatible with each other, supporting the hypothesis that mutual exclusivity of these PAI indicate a shared function for niche survival [43].

Since interphylogroup gene transfer between distantly related bacterial strains is uncommon, we analysed the incongruencies between core genome phylogeny (Figure 5.1) and multi-loci sequence analysis (Figure 5.4) for other examples supporting horizontal acquisition of T3SS encoding genomic islands. From this data, two additional examples were identified. Horizontal gene transfer of the T-PAI and S-PAI variants possibly occurred at least twice each creating patches of differing PAI

within phylogroup 7 along the core gene phylogeny where monophyletic groups have both strains containing one or the other variant. For S-PAI, the *hrc/hrp* genes cluster together separating into smaller monophyletic groups according to species phylogroup. This polyphyletic distribution of S-PAI containing phylogroup 7 (dark green) strains in the phylogroup suggests T-PAI was vertically inherited with subsets of phylogroup 7 gained this S-PAI after divergence from MRCA with phylogroup 9. Furthermore, the clear sequence conservation of core gene within phylogroup 7 variant, but gene divergence from phylogroup 11's S-PAI (Figure 5.3), suggests the gain of S-PAI was a singular event with a singular ancestral sub-strain rather than multiple transfer events. In fact, the S-PAI is not isolated with any particular subclade of phylogroup 7 suggests that PAI mobility is occurring at the strain level.

Similar to AT-PAI. *P. syringae* strains carrying the S-PAI variant also lack the T-PAI in its canonical genomic location, with it located instead in between genes encoding a n-methyl-l-tryptophan oxidase and ketoglutarate semialdehyde dehydrogenase. In fact, this insertion occurs in same location for both phylogroup 7 and 11 strains opposing the possibility of a second T3SS acquisition event in phylogroup 7. The method of PAI variant mobility, whether recombination or HGT, remains unclear until further strains carrying this T3SS variant can be assessed. A phylogroup-focused examination would likely be needed to differentiate the variability in phylogroup 7.

Unexpectedly, a very small subset of strains also contain uncommon flanking genes. Around three strains in phylogroup 7 with the T-PAI, possibly have T3SS rearrangements. However, a closer examination reveals that the T3SS are in the same genomic position, suggesting partial recombination of the genomic island may occur when genes (including the CEL) are relocated downstream of the core T3SS gene cluster into the EEL region. A further 4 strains (from phylogroups 1b, 2, and 3) have both a canonical and atypical flanking gene support, suggesting changes may be due to genomic recombination. For the phylogroup 2 strains, instead of the canonical *gstA* gene, Psy547 has an upstream xanthine phosphoribosyl-transferase gene while Psy551 has an upstream *terE* gene.

In reference strain PsyB728a, both these uncommon genes are typically found 1109 and 372 genes upstream respectively. An assessment of the core genome topology (Figure 5.1) compared to the conserved *hrc/hrp* consensus gene tree reveals evidence of a second example of PAI mobility (Figure 5.4). A major incongruence is the clustering of earlier branching phylogroup 5 with phylogroup 2 indicating both groups likely inherited their T-PAI from the same MRCA rather than a

MRCA with groups 2 and 3 as suggested by the core genome phylogeny. A gene tree of the more divergent *hrpA* [159,329] and *hrpF* [176] genes, the most divergent T3SS genes, partially corroborates the relatedness between these PAI where these strains group together with select strains from each group, but in these cases all phylogroup 5 strains also group in the same subclade (Figure 5.7 and 5.8). Furthermore, a hierarchical clustering of shared T3SS gene sets, including more sequence variable *hrp* genes and flanking effector genes, clusters all phylogroup 5 strains closer to both these more divergent phylogroups rather than closely related phylogroup 1 strains (see Figure 5.9 in later sections). In contrast to the aforementioned transfer events, the genomic position of the T-PAI for all three phylogroups (2,3, and 5) remained in the canonical location, which suggests a direct recombination of the PAI between the most recent common ancestor (MRCA) of phylogroup 5 with the MRCA of phylogroups 2 and 3.

Taken together, the similarity in shared T3SS gene sets between divergent phylogroups provides some evidence that multiple events of interspecies recombination of the core T3SS cluster affecting the shared effector loci has occurred in the *P. syringae* complex. This data thus supports horizontal inheritance of AT-PAI in the MRCA subphylogroup 2c strains with group 13 strains as possible indirect source of the PAI.

5.3.4 Analysis of AT-PAI encoding the third major T3SS variant in *P. syringae*

The Atypical T3SS gene cluster is hybrid between T-PAI and S-PAI

MLSA results of the conserved *hrp/hrc* genes shared across strains suggests that sequence divergence of the AT-PAI were nearly equally divergent from the other two variants, only marginally closer in sequencing similarity to T-PAI (Figure 5.4). This data suggests that the AT-PAI may be a third major variant of T3SS in *P. syringae* strains. Since gene arrangement and order of the genomic island are key differentiating factors for T3SS variants [150], I investigated this hypothesis by assessing each of the primary T3SS for every *P. syringae* strain, comparing both the gene presence and organization of the pathogenicity islands. Across all three T3SS variations, the genes encoding the core T3SS gene cluster are largely homogenous in content, order, and organization (see Figure 5.2)[150]. These genes comprise of structural components of the injectisome apparatus largely

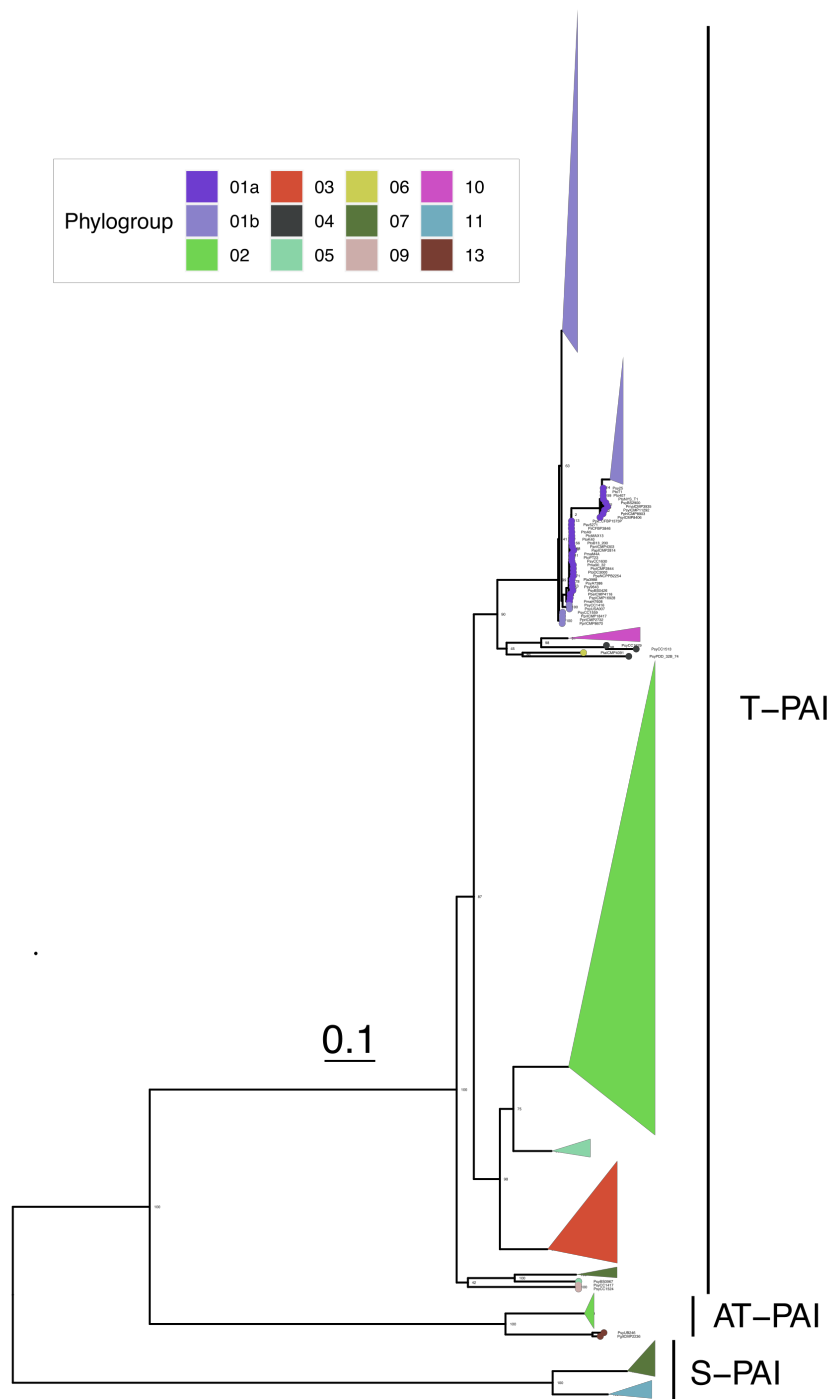


Figure 5.4 Multi-locus sequence analysis tree of conserved T3SS genes, grouped as single gene families, shared across all *P. syringae* strains (GTR+I+G4, Bootstrap n=3000). Conserved genes used in the analysis include *hrpR*, *hrcJ*, *hrcU*, *hrcT*, *hrcS*, *hrcR*, *hrcQ*, *hrcN*, *hrcV*. Divergence of conserved T3SS genes is not congruent with strain phylogeny (figure 5.1) indicating multiple inter-phylogroup transfer events of the T3SS PAI. The atypical PAI is more similar to the canonical PAI (T-PAI) than the single partite variant (S-PAI). Each leaf tip is coloured according to each strain's phylogroup with a triangle leaf tip representing a collapsed monophyletic group of the same phylogroup (with its representative colour). Tree branch scale represents the genetic difference between the leaf tips.

conserved across T3SS containing bacterium [88]. Many of the structural genes are highly conserved (Figure 5.2, blue genes), and can be clustered in the same gene family with greater than 70% identity for sequence divergent T3SS variants (Figure 5.3, blue *hrc* genes) whereas they have high sequence divergence for genes encoding other components of the T3SS, with less than 50% identity grouping into separate gene families (Figure 5.3, green *hrp* genes).

Within the core gene cluster there are two main distinctions between the canonical T-PAI and the other variants. Both AT-PAI and S-PAI, lack a second paralog of the global regulators *hrpR/S* (Figure 5.2 and 5.3). For the canonical T-PAI, the duplicate global regulator, *hrpS*, gene encodes a secondary T3SS enhancer. In fact, a gene tree of the global regulator homologs indicates that duplication of *hrpR* occurs after divergence from the MRCA shared between *hrpR* of AT-PAI and S-PAI (Figure 5.5). Although there is evidence HrpS can solely activate T3SS expression[56], dimerization of *hrpR* is crucial for high T3SS expression suggesting that absence of this paralog may influence T3SS expression in AT-PAI and S-PAI containing strains [168,170]. A second difference is that the conserved protein HrcQ, involved in forming the cytoplasmic ring, is split into two gene families, marked as a fusion/fission event, representing the two subunits in T-PAI (HrcQa and HrcQb) instead of a singular gene family representing the singular protein in AT-PAI and S-PAI (Figure 5.2 and 5.3)[152].

For phylogroup 7 and 11 genomes containing the single partite variant (S-PAI), the core gene cluster comprises the entire T3SS (Figure 5.2 and 5.3). For the tripartite PAI variants (T-PAI and AT-PAI), the core gene cluster is flanked by two effector protein containing regions: a upstream conserved effector loci (CEL) and a more variable downstream exchangeable effector loci (EEL)(Figure 5.2) [151]. For the majority of these tripartite containing strains (i.e., only canonical variants, T-PAI), gene content and order were observed to be highly conserved in the CEL. A

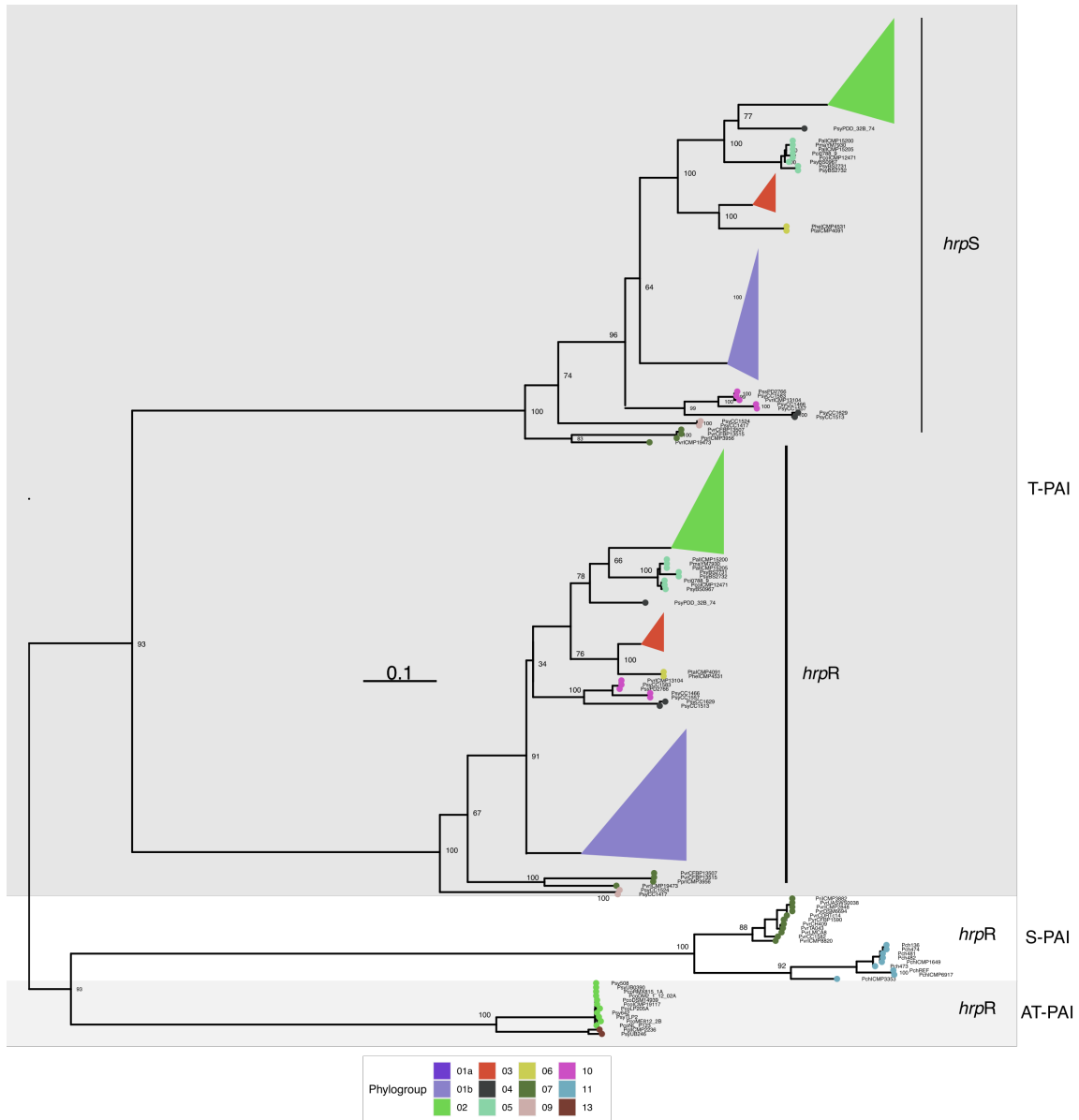


Figure 5.5 Maximum likelihood gene tree of *hrpR* and *hrpS* using a TIM1+G4 model (Bootstrap $n=1000$). Each leaf tip is coloured according to each strain's phylogroup. Gene duplication of *hrpR* into *hrpS* in the T-PAI occurs after divergence from the MRCA of *hrpR* shared by the non-canonical T3SS variants (S-PAI and AT-PAI). Unlike like the conserved *hrc* genes (Figure 5.4 and 5.6), *hrpR* of AT-PAI share closer homology with S-PAI's *hrpR* than T-PAI's *hrpR*.

tripartite arrangement is also apparent in the AT-PAI, indicating that this atypical PAI is more similar to T-PAI in its over-arching genetic structure.

However, the content of the flanking effector loci is different between AT-PAI and T-PAI. In T-PAI containing strains, the conserved genes upstream of T3SS regulator *hrpR* until *gstA* include, in the same order, genes encoding effector proteins *avrE1*, *hopM1*, and *hrpW1* with their respective chaperone genes; conserved hypothetical proteins; an oligotransporter; and an ISPsy9 transposase *orfA*. To contrast, the downstream EEL is a highly variable loci that typically extends from the gene encoding *hrpL* to *queA*, and range from three genes, as seen in the AT-PAI, to remnant prophages. Although full sized phages are rare, phage integrases were more commonly observed in many *P. syringae* genomes. In conjunction with the presence of other mobile gene elements, including transferases, general integrases, and insertion sequences, these genes likely facilitate horizontal gene transfer boosting gene diversity within the region [156,367].

The ubiquity of this tripartite PAI arrangement and its genomic position between a *gstA* and *queA* gene is prevalent amongst the strains but may reflect a sequencing bias towards agriculturally relevant strains [44]. It is not universal however. In a small subset of phylogroup 7 genomes (n=3) that contain the T-PAI rather than S-PAI, conserved CEL genes have been found in EEL and the inverse was also observed. In these cases, the gene directly upstream was not the canonical *gstA* gene but another conserved gene, encoding a substrate transporter, which were found 27 genes further upstream of the T3SS PAI in other *P. syringae* strains. Relocation of the CEL downstream in these strains indicates that the both effector loci should be considered as a collective effectome rather than distinct regions since these regions are one of the key differentiating factors between the variants. Four other strains from different phylogroups (01b, 2, and 3) also had different conserved genes directly flanking one side of the T3SS PAI. Furthermore, in 89 strains (~21%), genes flanking the T3SS could not be determined due to assembly discontinuity.

In a similar case, an analysis of the genes upstream of the CEL and downstream of the EEL indicates an altered genomic location in all strains containing AT-PAI, suggesting either prior recombination of the pathogenicity island or loss of the canonical T-PAI and simultaneous gain of the full AT-PAI. Many of the conserved CEL genes remain in the canonical location between genes *gstA* and *queA*, however, whether these genes are actively transcribed as part of the *hrpL* regulon remains unclear [93]. In PG-2C, the AT-PAI has been relocated > 270 kbp downstream becoming situated

between a gene encoding a sodium:bile acid symporter and *rdgC* encoding a recombination associated protein whereas, group 13 strains are unclear due to contig discontinuity.

Despite retaining a tripartite gene arrangement, in all strains with atypical T3SS, the AT-PAI was found to contain shorter CEL and EEL regions compared to the canonical T-PAI. The CEL is reduced to five hypothetical genes from the typical 15 genes, without any identifiable motifs, plus a singular encoding a T3E chaperone belonging to the TIR chaperone family [368]. The hypothetical gene directly downstream of the chaperone gene is likely a novel variant of highly conserved T3 effector gene *hopM1* with the encoded protein product sharing a reciprocal top match of 50% identity and 60% coverage. The other closest match to CEL effector genes is one hypothetical gene encoding a protein product with a top reciprocal match to HrpK, a harpin translocator. HrpK is typically found in the EEL but an equivalent to the HrpW1 translocator is commonly found in canonical CEL [155]. However, the match falls below the twilight zone and has poor coverage so whether it can act as a translocator (i.e., involved in the direct interaction to host plants) warrants further exploration [148]. Lastly, the largest conserved gene, *avrE1*, encodes a protein integral in maintaining an aqueous host environment [80], is missing from the CEL and instead appears in the hypervariable region between the two core gene clusters, sharing a similar cluster architecture to S-PAI variants (Figure 5.2)[152].

Similarly, the EEL in AT-PAI was found to be highly truncated in all PG-2C strains, containing only three genes, and lacking any discernable mobile genetic elements. The only non-hypothetical gene found in the EEL encodes a necessary SLT transglycosylase likely used to degrade the outer membrane during injectisome assembly [369]. In nearly all strains with a canonical T-PAI, this gene equivalent was typically found in CEL as one of the 15 conserved upstream genes. When taken together, it appears that the effectome of AT-PAI containing subphylogroup 2c strains have been further reduced when compared to closely related phylogroup 2 strains, a group with an already reduced T3 effectome [43].

To contrast the tripartite system, S-PAI contains a single effector loci situated in the hypervariable region between the core T3SS gene clusters, rather than flanking it [364]. This core gene cluster is further split into two gene clusters by an internal hypervariability region across all genomes. We were unable to confirm the region's gene hypervariability due to lack of available high-quality genomes containing S-PAI. Among the complete genome set (n=418), only 7 genomes, typically belonging to phylogroup 7, contained a full sequence spanning the entire hypervariable

region. Among the 7 genomes the hypervariable region was small, and relatively conserved whereas at least two strains (Group 11's PchREF and PchICMP6917) had incomplete but larger effector loci. Given the hypervariability of the region, erroneously missing a gene due to assembly discontinuity is possible. Similarly for T-PAI containing strains, 318 genomes (76.8%) also contained a fragmented hypervariable region preventing full evaluation of the region across the majority of strains. For genome complete reference strains (e.g., PsyB728A (group 2), and Pph1448 (group 3)), we observed that this hypervariable region is relatively small, containing few to no genes, whereas in S-PAI it is variable in length and encompasses a large portion of the effectome. Despite containing relatively few genes in this region, the AT-PAI shares another commonality with S-PAI: the largest conserved effector gene *avrEI* and its respective chaperone gene are both localized to this region rather than the CEL (Figure 5.2)[152].

Although AT-PAI core gene cluster contains many similarities to S-PAI in gene arrangement, protein coded in the AT-PAI have a higher sequence similarity to the canonical T-PAI than the S-PAI and cluster together at a lower stringency threshold supports the longer branch length between AT-PAI and S-PAI than between AT-PAI and T-PAI (Figure 5.3 and 5.4). A gradual increase in protein identity thresholds can splinter the broad gene family clusters into more specific groups based on protein similarity. As mentioned previously, T3SS, both S-PAI and AT-PAI lack the paralog *hrpS* found in T-PAI and required for polymerization and transcription regulation [152]. At the protein sequence level, the S-PAI *hrpR* forms a unique cluster at 90% id while the AT-PAI remains clustered with other T-PAI *hrpR* until reaching the strictest threshold of 98%, indicating higher sequence similarity. However, the high sequence conservation implies orthology of this gene across all PAI. Similarly, *hrcQ* is a singular gene in both S-PAI and AT-PAI yet split into two genes encoding subunits *hrcQa* and *hrcQb* in T-PAI (Figure 5.3). AT-PAI HrcQ also has a higher sequence similarity to the T-PAI oligomer, forming its own unique cluster at 95% identity threshold compared to S-PAI's threshold at 70% identity. This genetic feature supports that changes in the core gene cluster of T-PAI occurred following divergence from AT-PAI (Figure 5.5). The pattern of sequence similarity also occurs with other gene families include T3SS regulators, *hrpT* [370], *hrpV*[178], and *hrpL*[166], as well as secretin encoding gene *hrcC* sharing high sequence similarity between AT-PAI and T-PAI (Figure 5.6). High protein conservation suggests a similar selective pressure on these genes for AT-PAI and T-PAI, despite different PAI gene structures. Of the remaining genes forming the core

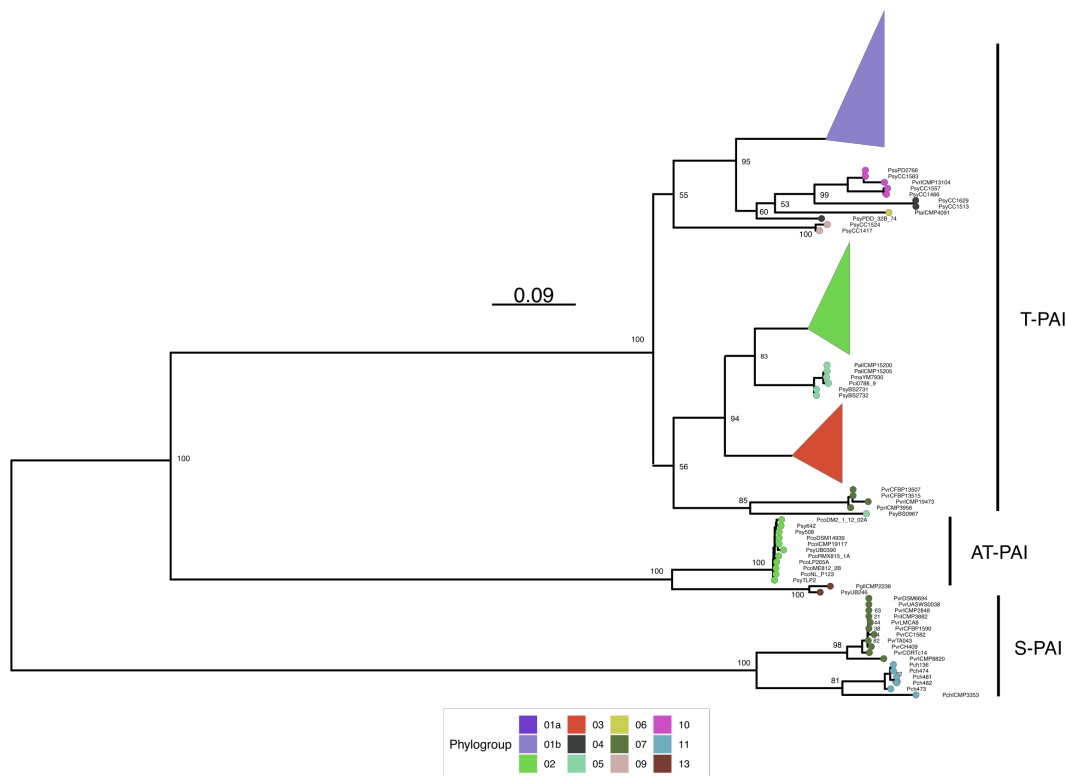


Figure 5.6 Maximum likelihood gene tree of some of the most conserved genes *hrcT*, *hrcS*, *hrcR* encoding the inner membrane of the T3SS injectisome (Bootstrap n=1000). Each leaf tip is coloured according to each strain's phylogroup. Conserved T3SS genes indicate that AT-PAI is more sequence similar to T-PAI than S-PAI sharing a more recent common ancestor.

cluster, 6 genes of the non-conserved core genes, encoding either regulators (*hrpA*, *hrpF*, *hrpG*[176], and *hrpP*[371]) or the ATPase complex (*hrpD*, *hrpE* and *hrpO*[88]) are highly divergent and form distinct clusters with less than <50% identity separate from the other variants (Figure 5.3).

The AT-PAI appears to have a mixture of features from both T-PAI and S-PAI. The core gene region of AT-PAI, and partially the hypervariable region are closer in arrangement to S-PAI. Conversely, it also retains an overall gene arrangement and sequence similarity, in conserved structural genes, to the canonical T-PAI. When taken together, the MLSA trees and *hrpR/S* gene tree suggests that not long after the MRCA of AT-PAI and T-PAI diverged from S-PAI, the MRCA of T-PAI underwent gene duplication (*hrpR/S*) and gene fusion (*hrcQ*) diverging from AT-PAI. Therefore,

the high sequence divergence from S-PAI and gene arrangement divergence from T-PAI indicates that AT-PAI should be considered separately as a third major T3SS variant.

Sequence variation in T3SS correlates to gene clustering of pilus proteins HrpA and HrpF

When we analysed each gene in the PAIs, we found that the three T3SS variants correspond to the gene cluster variations for the gene families of *hrpA* and *hrpF* (Figure 5.3). High sequence variability is speculated to be tied to HrpA function. Akin to long filament in flagella, the needle of the injectisome is comprised of a uniform polymerization of many copies of a singular protein, HrpA [329]. The plant host has been hypothesized to recognize the structure of injectisome via HrpA, similar to detection of flagella via the flg22 molecular pattern in the filament. Consequently, high selective pressure to evade host plant defence has been attributed to the high sequence variability of HrpA. As expected, *hrpA* genes in my results was one of the most divergent T3SS genes, splitting into four distinct clusters: two clusters for T-PAI, *hrpA1* and *hrpA2* [159], and one cluster each for the S-PAI and AT-PAI variants (Figure 5.3 and 5.7). This data supports the idea that AT-PAI is a third separate variant, but also that T-PAI has evolved into two sub-variants. Among the other core non-conserved *hrp* genes, *hrpF* follows a similar pattern, also grouping into four distinct clusters. In fact, the genomes represented in each cluster are the same for both genes, and is the defining gene pattern differentiating the PAIs into distinct clusters and variants (Figure 5.3 and Figure 5.8). The separation of *hrpA* and *hrpF* of the AT-PAI as distinct clusters from both T-PAI phylogenetic trees also supports a more recent acquisition for genes within the AT-PAI between the otherwise distant phylogroups 2c and 13.

As for T-PAI, the pattern in strain clustering is the same for both *hrpA* and *hrpF* genes (Figure 5.7 and 5.8) where phylogroup 5, exclusively forms a distinct cluster with homologs from phylogroup 2 and 3, and separate from phylogroup 1 supporting the MLSA of conserved T3SS genes in Figure 5.4. The lack of progressive divergence in these genes compared to the other T-PAI containing phylogroups suggests this T-PAI in groups 2,3, and 5, or at least this subset of genes, were acquired from an external source rather than the gradual divergence expected with vertical inheritance based on the core genome phylogeny (Figure 5.1).

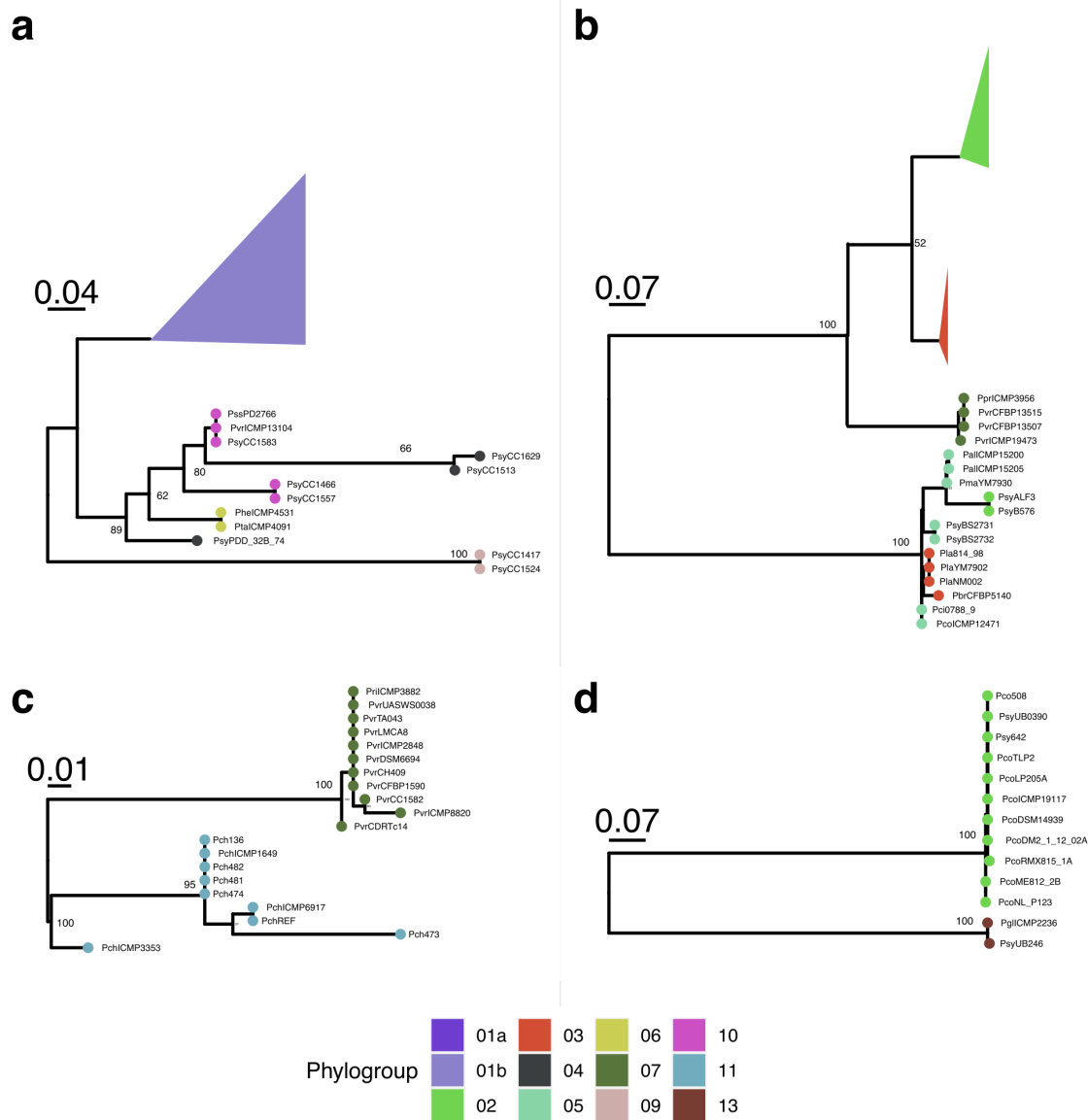


Figure 5.7 Gene trees of *hrpA* (Bootstrap n=2000). Although encoding functional pilus subunits, *hrpA* forms four separate gene families with high sequence divergence between gene families but low sequence divergence within these groups. Although functionally equivalent, separate gene trees were generated for *hrpA* since was grouped as separate gene families sharpening with less than <50% protein product identity. Two for canonical T-PAI, *hrpA1* (a) and *hrpA2* (b), and one for the other two variants S-PAI (c) and AT-PAI (d). Gene trees with likely horizontal inheritance (b, d) have slightly greater genetic divergence. Each leaf tip is coloured according to each strain's phylogroup except for phylogroup 1 represented by a singular purple. Large phylogroups 1, 2, and 3 have been collapsed into triangles with proportional branch length. Overall size of these triangles are not representative of number of genomes (Tree branch scale represents the genetic difference between the leaf tips).

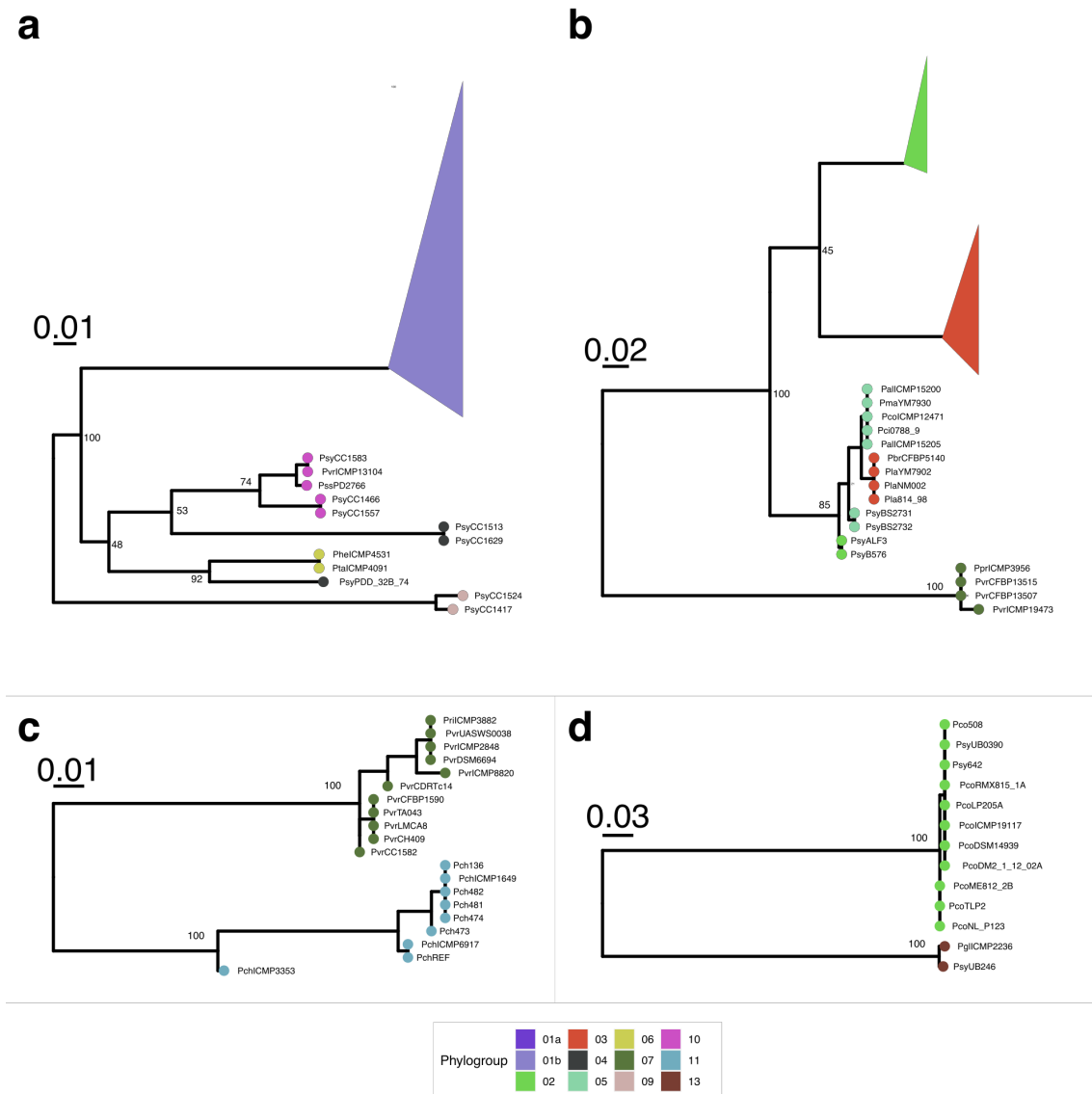


Figure 5.8 Gene trees of *hrpF* (Bootstrap n=2000). Sequence divergence splinters the core T3SS genome into the same four gene families equivalent to *hrpA* (Figure 5.6). Similarly, the four gene families have high divergence between gene families but low sequence divergence within these groups. Two for canonical T-PAI, (a/b), and one for the other two variants S-PAI (c) and AT-PAI (d). Although functionally equivalent, separate gene trees were generated for *hrpA* since was grouped as separate gene families sharpening with less than <50% protein product identity. Gene trees with likely horizontal inheritance (b, d) have slightly greater genetic divergence. Each leaf tip is coloured according to each strain's phylogroup except for phylogroup 1 represented by a singular purple. Large phylogroups 1, 2, and 3 have been collapsed into triangles with proportional branch length. Overall size of these triangles are not representative of number of genomes (Tree branch scale represents the genetic difference between the leaf tips).

When considering *hrpA* and *hrpF* gene trees, the T3SS tree topology suggests an initial acquisition of this T-PAI variant occurred in an ancestral strain to phylogroups 2 and 3 followed by a secondary HGT between the MRSA of phylogroups 2 and 5. However, gene tree phylogeny of both *hrp* genes indicate that a group 5 strains consistently group with a small subset of phylogroup 2 and 3 strain. Although the shortness of the *hrpA* gene prevents sufficient information for pinpoint grouping, the more conserved *hrpF* corroborates this grouping. The close grouping with only a subset rather than the entire phylogroups 2 and 3 may suggest possible recent recombination of individual genes after HGT of the full PAI. Unlike the other T3SS variants, the T-PAI of phylogroup 5 remains within the original genomic location suggests that T-PAI HGT may have occurred through direct homologous recombination.

The distinct clustering pattern for both *hrpA* and *hrpF* is likely due to protein-protein interaction and compatibility between their respective protein products [176]. HrpF has been speculated to prevent HrpA degradation through binding in the cytosol prior to pilus assembly suggesting that sequence diversity in HrpA will be reflected in HrpF. When we generated gene trees for both *hrpA* (Figure 5.7) and *hrpF* (Figure 5.8), the similarity in tree topology supports this hypothesis. However, the difference in the tree scale indicates that *hrpA*'s diversity is only partially reflected in *hrpF*, where the latter has shorter branch lengths and thus less sequence divergence. Higher HrpF conservation is expected since high variation is required only in the regions associated with protein interaction. Furthermore, HrpF is hypothesized to interact with the less variable T3SS regulator HrpG, supporting that HrpF sequence conservation needs be partially maintained for regulation.

When taken together with the multi-locus sequencing analysis, our data suggests that the AT-PAI is distinctly a separate T3SS variant that is shared via horizontal transfer between subphylogroup 2c, a subset of group 13 strains, and an yet unidentified intermediate strain. Furthermore, this same pattern is not uncommon with a T-PAI subvariant shown in phylogroups 2, 3, and 5 as well as S-PAI with phylogroup 7.

Lastly, we investigated whether this diversity can provide insight into host specificity by mapping known hosts to their species. This analysis yielded no direct pattern (data not shown) indicating the singular method of host plant recognition of HrpA does not strongly influence host specificity of the bacterium.

5.3.5 Analysis of genetic factors that impact pathogenicity in subphylogroup 2c strains

Grouping virulence reveal differences in general pathogenic strategies of *Psy* sub-phylogroups

To investigate potential links between subphylogroup 2c strains lack of pathogenicity compared to the other strains and T3SS regulated genes, we generated a HMM prediction model of known *hrp* promoter box using MEME [171,352]. Using this model we scanned the Intergenic regions and up to 100 bp of the respective genes with FIMO for *hrp* promoters indicating T3SS regulation. To reduce false positives from spurious hits, we also applied an FDR q-value cutoff of 0.1. The number of predicted *hrp* promoters range from zero to 95, forming a bimodal histogram with more than two sharp peaks (data not shown), with the primary and more virulent phylogroups forming the higher range of *hrp* promoters detected. More specifically, *P. syringae* strains in phylogroups 1,3,4,5, and 6 have the highest number of detected promoters with strains averaging ~55 *hrp* boxes found. Conversely, phylogroup 2 strains, typically targeting a broader host range, have a comparatively reduced effectome consistent with previous reports, averaging around ~30 *hrp* box promoters [93].

An even further reduced effectome is found for the subphylogroup 2c with these strains containing between 14 to 20 *hrp* boxes. Earlier branching strains excluding phylogroups 4 and 5, have fewer *hrp* promoters compared to later branching strains. Phylogroups 7,9,10,11, and 13, along with subphylogroup 2c, have a maximum of 25 promoters, with the phylogroup 13 containing the least amount of *hrp* boxes found; as expected, *hrp* boxes were undetected in group 13 strains lacking a T3SS. However, the lower level of *hrp* detection is potentially confounded by overall sequence divergence of HrpL, the RpoN sigma factor recognizing *hrp* boxes, with the *hrp* consensus sequence in earlier branching strains likely differing from later branching phylogroups.

For every predicted *hrp* box, we extracted downstream genes that are potentially transcribed in a single unit. In this work, we describe a transcriptional unit as all genes, found downstream of a promoter, that are in the same orientation separated by a maximum of 200 bp per gene [372]. Strains were then were grouped together using agglomerative clustering of the gene sets via Hierarchical Sets

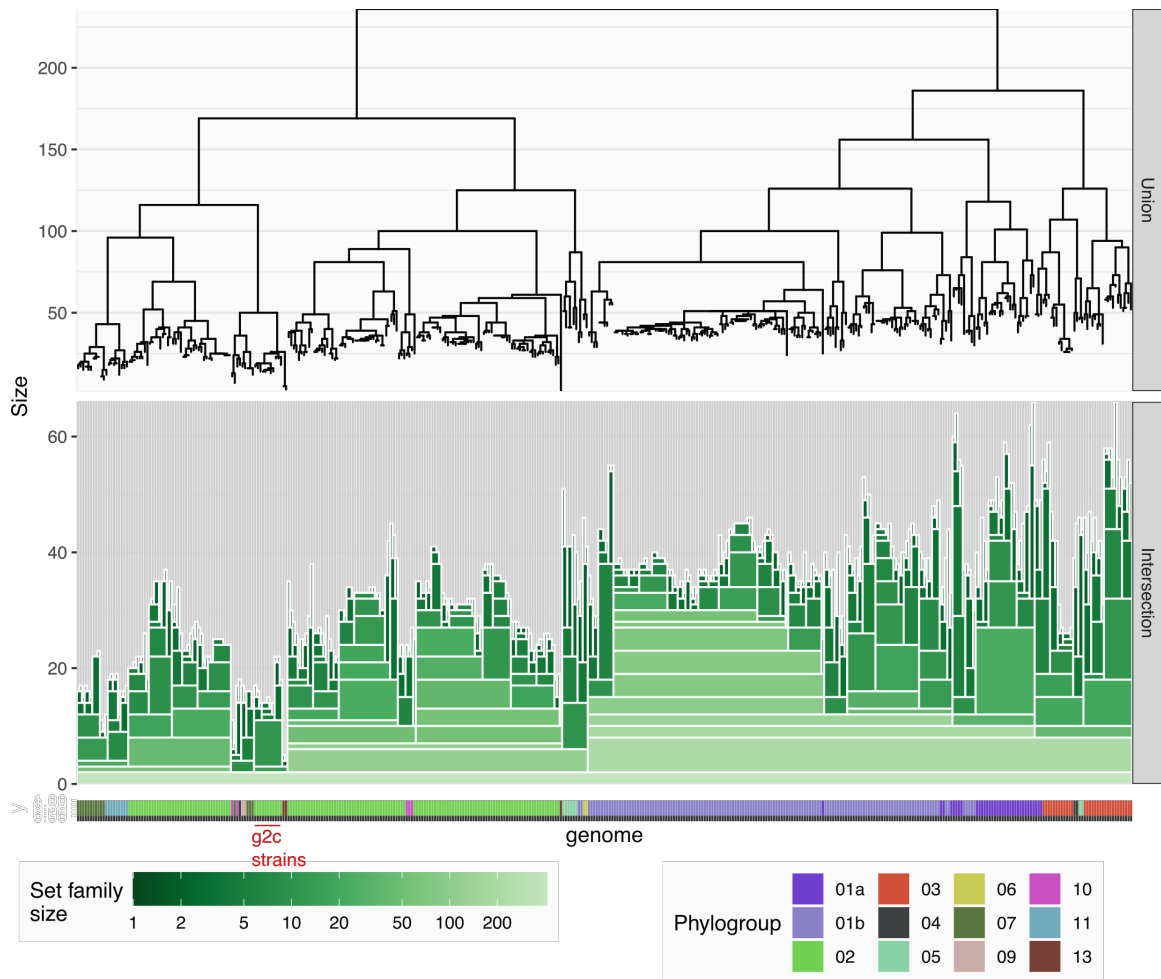


Figure 5.9 Icicle plot of the hierarchical set clustering of *hrp* box regulated gene families. A dendrogram of clustered strains (Top) are grouped based on the shared gene families sets with the edge length represented by the union size of the sets. Bottom “icicle” graph represents the shared sets between groups of clustered strains with size correlating to the number of shared genes in the set (intersect). Subphylogroup 2c (highlighted red) is closer grouped with other early branching strains due to the absence of shared gene families. Beyond *hrp* regulated PAI genes (light green bar at base), the subphylogroup 2c strains shares no *hrp* regulated gene families with other *P. syringae* strains.

[349]. This clustering method begins at the strain level where strains sharing greatest number of gene sets, largest number intersection of gene (shared) to union, are initially grouped before clustering more divergent groups together via shared gene sets. To avoid emphasizing uniqueness in strains gene singletons were removed.

The *P. syringae* strains were split into 2 major clusters according to their respective *hrp*-regulated with latest branching group 2 strains grouping with majority of early branching phylogroups, with a few exceptions (Figure 5.9)[43]; phylogroup 6 also clustered with early branching strains while a group 4 and a few group 5 strains clustered with late branching phylogroups 1, and 3. From the icicle plot, intersection size reveals that the *hrp*-regulated genes unifying the two groups is minimal comprising solely of T3SS PAI genes known to be part of *hrpL* regulon [373]. In this case, phylogroup 3 strains cluster with phylogroup 1 with overlap indicating many shared *hrp*-regulated gene families between the phylogroups possibly sharing the same function during host infection [374]. Unlike phylogroups 1 and 3, which have great overlap in *hrp* regulon genes, phylogroup 2 clusters together with early branching phylogroups due to minimal overlap and divergence from the other late branching phylogroups (1 and 3).

Consistent with previous studies, both group 2 and group 10 strains group together due to a reduced identifiable effectome [47]. This clustering is more due to a lack of T3E reliance rather than shared genes between the groups as highlighted by the lack of intersection (i.e., no box) between the groups. Furthermore, the differentiation of subphylogroup 2c from other group 2 genomes clustering together with early branching phylogroups 9 and 10, the *P. syringae* strains also with a smaller *hrp* regulon, primarily due to a lack of shared *hrp* regulon genes with group 2 strains (Figure 5.9). Subphylogroup 2c clusters together with phylogroup 13 strains solely sharing conserved T3 effector, *hrpA* and *hrpF*.

Similar to *hrpA* and *hrpF*, other known *hrp*-regulated genes were observed to be ubiquitous among the strains but were not considered shared due to great sequence divergence. This is attributed to high diversifying pressure [159,325] or origination from an alternative MRCA ortholog, placing them in separate gene families. This same gene diversity reduces the clustering efficacy and is evident with the genes encoding conserved T3Es HopM1 and AvrE1. These T3E genes cluster as separate gene families from closely related strains. As mentioned above, the alternative gene position of *avrE1* in the hypervariable region for non-canonical T3SS variants suggests an origin from horizontal gene

transfer rather vertical inheritance (Figure 5.2). For example, *avrE1* in AT-PAI of subphylogroup 2c is highly divergent from other *avrE1*, clustering as a separate gene family, with the respective protein product sharing less than 50% identity with other *avrE1* gene family homologs found in other strains. This variance indicates the genes do not share the same MRCA and are not direct orthologs with the T-PAI and S-PAI variants. The closest equivalent gene outside of PG-2c is in AT-PAI containing group 13 strains, which is similar enough to be classified as the same gene family with the respective AvrE1 sharing 50-60% identity with each group. However, this shared gene does not unify the two groups due to both sequence divergence and its putative pseudogenization in group 13 strains. With the importance of this gene in niche survivability in its plant host [80,343], the pseudogenization of this gene may be an artifact of draft gene assembly due its truncation at the edge of a contig.

Overall, the strains are clustered in two major groups that are largely consistent with the bimodal distribution of *hrp* boxes found (Figure 5.9). One primary group, comprised of groups 1, 3, and 5 strains, has more detected *hrp* promoter boxes and also a greater number of shared gene sets among the subgroups (Figure 5.10). Conversely, the second large group was comprised of the remaining phylogroups and has fewer *hrp* boxes. In general, strains within this group form more insular clusters. Taken together, the former large cluster suggests that these phylogroups heavily rely on the diversity of its effectome.

To address the diversity of *hrp* regulated genes, especially T3 effector genes, the gene families were collapsed into larger functional groups based on additional annotations of the encoded protein product as identified by InterPROSCAN. Smaller gene families were clustered either by known molecular function, or known protein (super)family allowing for grouping of similar functioning genes. Unexpectedly, only a small set of gene families were compressed, reducing 323 gene families to 310 larger family groups, e.g., the 4 *hrpA* gene families were collapsed into a single gene family. With these functional groups, there were minimal changes in clustering between all strains, highly similar to Figure 5.9, with the overarching group shared gene set including collapsed gene families of *hrpA* and *hrpF* (data not shown). Beyond the core T3SS genes (*hrp/hrc* genes in Figure 5.3), there were no universal T3E shared by all strains. Highly conserved T3E genes, such as *hopM1* and *avrE1*, are missing from select strains either due to possible gene loss due to high T3SS divergence like the former gene, or artificial pseudogenization of the latter gene.

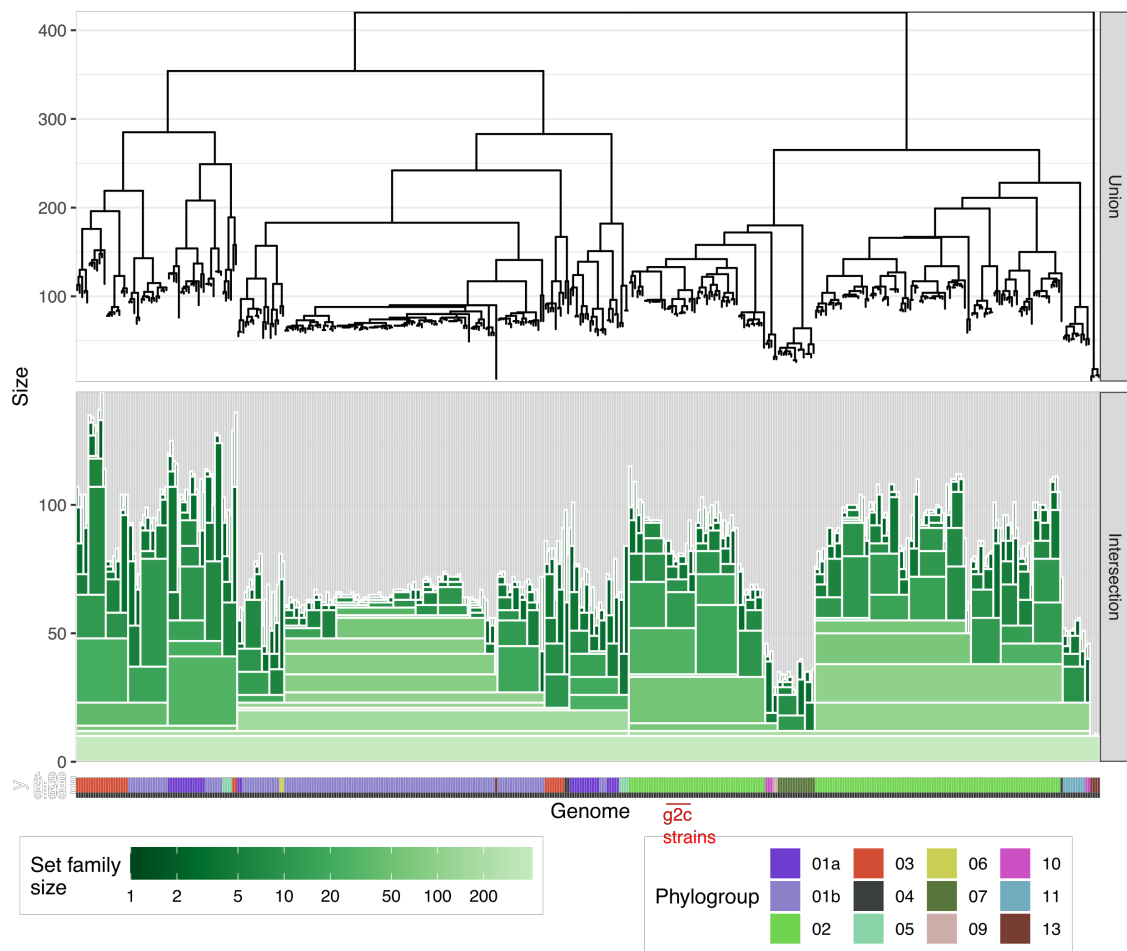


Figure 5.10 Icicle plot of the hierarchical set clustering of the *hrp* box regulated gene families and known virulence gene (collapsed into molecular function groups, protein superfamilies, and protein families). The icicle plot is a summary plot of the Figure 5.11 heatmap. A dendrogram of clustered strains (Top) are grouped based on the shared gene families sets with the edge length represented by the union size of the sets. Bottom “icicle” graph represents the shared sets between groups of clustered strains with size correlating to the number of shared genes in the set (intersect). When including broader gene groups compared to only *hrp* regulated genes in Figure 5.9, subphylogroup 2c strains cluster closely with other group 2 strains, sharing many genes known to be involved in virulence (e.g. phytotoxins). A presence/absence breakdown of shared gene families is shown in Figure 5.11.

A limitation in this method of inference comes from the limited annotations of many proteins providing uncertain evidence of their putative virulence. Even with current protein annotation software, many of the encoded protein remain hypothetical and require further functional characterization. One subset of the uncharacterized genes of interest under the *hrp* regulon in subphylogroup 2c encodes a protein annotated only as either disordered and/or contains a secretion signal. Although both characteristics are indicative of a role in potential virulence[375], common T3Es with disordered domains utilize chaperones for correct protein folding after translocation into the host plant preventing analysis of these effector genes with similar functions [149].

We further investigated whether other known virulence and protein groups can be informative of general pathogenicity strategies between subphylogroup 2c and other *P. syringae* strains. In this analysis, we combined the expanded *hrp*-regulated functional gene families with additional blastp-identified gene families of known virulence genes. For T3 effectors, we used a custom database created from three T3 effector databases. Other virulence gene families include genes involved in phytotoxin biosynthesis [325,376], plant cell wall degradation, and general virulence genes previously identified in reference phylogroup 2 strain PsyB728a [158]. Previously identified effector and phytotoxin biosynthesis gene families were also included.

We did not use a general virulence database, such as VFDB and Victor, to avoid identification of niche survivability genes, such as the flagella, and focus on proteins previously characterized to be involved in manipulation of the host environment (e.g., defence and evasion of host defence system). The collected virulence proteins were then compared against the representative protein for each gene family using Blastp. Due to the potential sequence diversity of T3 effectors, we considered gene family-query hits with >50% identity and >60% query coverage as putative virulence genes. An icicle plot summary of this clustering is shown in Figure 5.10. Under this methodology, the strains cluster more closely with their core genome phylogroup with greater intersection of shared gene families. In fact, subphylogroup 2c strains cluster strongly with other group 2 strains, and away from other early branching strains, such as phylogroup 10. The larger shared gene sets between group 2c strains and other group 2 strains suggests that many of its potential virulence are not under direct *hrp* regulation, and may involve other mechanisms.

A breakdown of the shared gene sets was visualized with five broad categories of genes, core T3SS genes (i.e., Figure 5.3 genes), general virulence (i.e., cell wall degrading proteins), *hrpL*

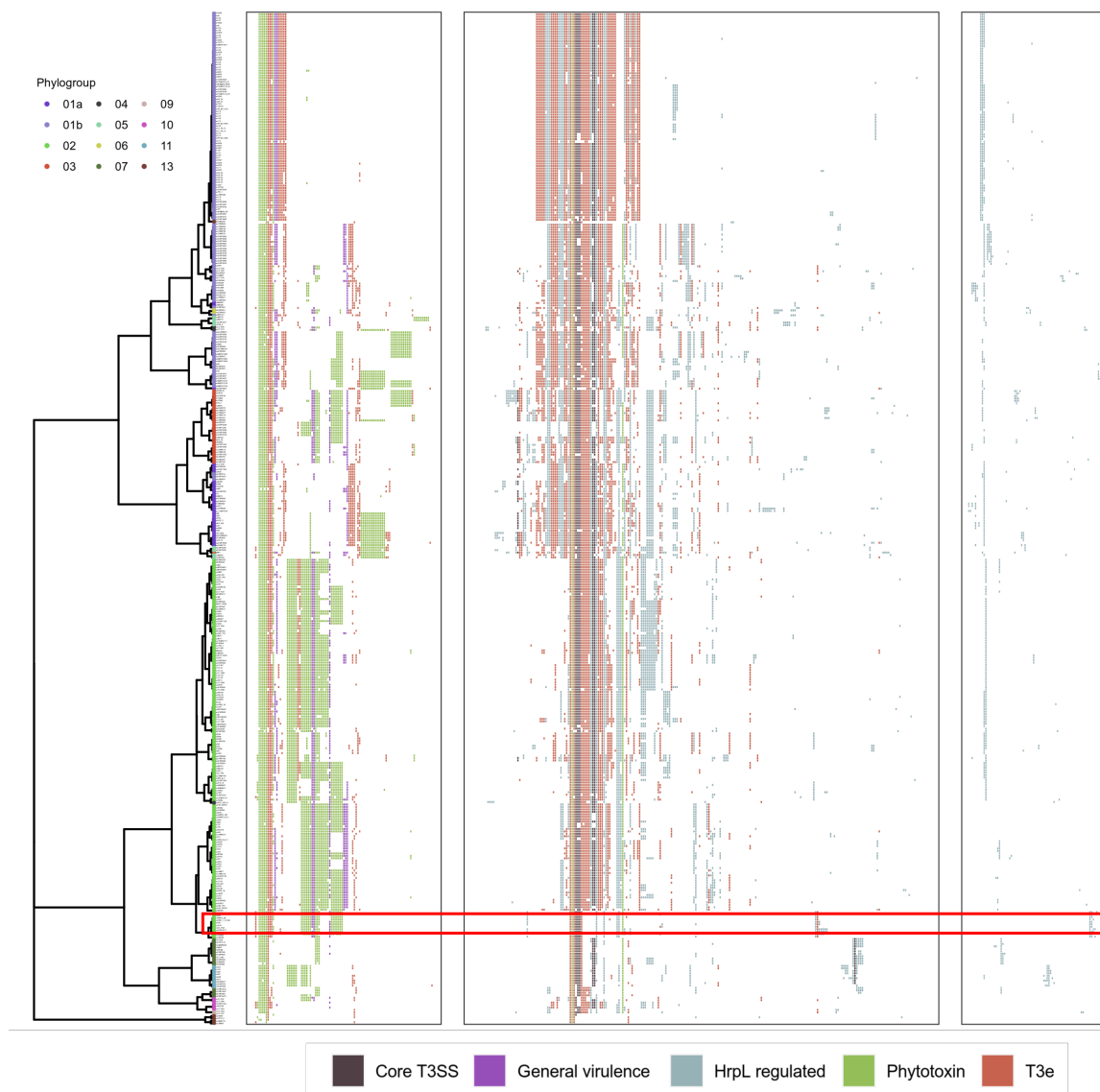


Figure 5.11 Hierarchical set clustering and binary heatmap of the presence and absence of gene families that are *hrp* box regulated genes, and/or blast-detected virulence genes reveal general virulence strategies of *P. syringae* strains. A icicle plot summary of strain clustering based on these gene families is shown in Figure 5.10. Divergence of subphylogroup 02c (highlighted in red) from MRCA of group 2 is correlated with reduction of virulence genes found in late branching *P. syringae* strains (phylogroups 01, 02, and 03) similar to more distant strains (below red box). However, presence phytotoxin genes and conserved T3E clusters it closely with phylogroup 02 strains indicating similar virulence strategies are maintained in subphylogroup 02c. In the plot, similar gene families are grouped together if the encoded protein products share a molecular function, protein superfamily, or family. These collapsed gene families are separated into three major heatmap plots: virulence genes detected by blast (left), known function/superfamilies/families detected by interproscan (middle), and unknown function/families (right). Tree tips are coloured based on each strain's phylogroup. Each heatmap block is coloured based on blast matches to known virulence genes at 50% sequence identity and 60% query coverage.

regulated, phytotoxins, and T3 effector genes in Figure 5.11. Clustering with this collection of genes reveals that nearly all strains share a very small subset of 7 virulence genes, appearing as nearly vertical lines in Figure 5.11. These genes include a CheY-like response regulator (purple), a non-ribosomal peptide synthase *syfA*, a gene involved in syringolin A biosynthesis (green), and genes partially encoding the mangotoxin biosynthetic pathway (green)[8,377]. In contrast, the remaining non-conserved genes typically appear in less than half of all strains.

Compared to other phylogroups, *P. syringae* strains belonging to phylogroup 1 (purple strains in top half of the plot) are the most reliant on T3Es, visibly containing larger concentration of red T3 effectors (Figure 5.11). The overabundance of T3Es is likely to compensate for the general inability to synthesize any known phytotoxin common to later branching phylogroups 2 and 3 [100,342,377,378]. However, there is a small subgroup of phylogroup 01a strains (dark purple strains in the middle of the plot) that appear to synthesize phaseolotoxin and coronatine (green), two phytotoxins important to repress stomata closure and ensure continuous invasion of the host plant. These same strains also contain arginine catabolism genes in the *hrp* regulon, involved with triggering chlorosis in host plants, via arginine deficiency, producing effects similar to other phytotoxins [100].

On the opposite end of the spectrum, the latest branching strains in phylogroup 2 (green strains in bottom half of the plot) have the least number of T3Es amongst the three main studied late branching strains carrying highly conserved T3Es (red) that are also found in phylogroup 1 and 3. Lacking in a diverse set of T3Es, this *P. syringae* group comparatively expresses the most phytotoxin synthesis genes (green). As a group, these strains are capable of expressing non-ribosomal polyketide

synthases capable of synthesizing toxic syringopeptin, syringomycin, and syringolin A [99]. Furthermore, other genes in this phylogroup encode peptidyl synthases have also been tied short chain peptides capable of phytotoxin-like capabilities. Situated between the two major phylogroups, strains in phylogroup 3 have a more variable set of genes; this group has a slightly smaller effectome, but is more capable of synthesizing phytotoxins including phaseolotoxin and coronatine similar to phylogroup 1.

Taken together with the core genome tree topology (Figure 5.1), there is a gradient in pathogenicity strategies in the commonly studied late branching strains. The earliest of the three late branching phylogroups (group 1) is highly dependent on a targeted approach with a high variety in its T3 effectome directly engaging with the plant defence response suggestive of higher level of host specificity when tailored to particular host plant [341]. As the species diverges and branches into later branching phylogroups (group 2, and 3) it adopts a more generalized approaching in evading host plants. In these strains, the lower virulence from a reduced effectome, likely via T3E gene loss, can be possibly compensated by the larger repertoire of phytotoxin biosynthesis genes in phylogroup 2, the last branching phylogroup [100].

In the remaining primary phylogroups, patterns in virulence genes are less clear, but all strains carry at least a few T3E genes (Figure 5.11). In fact, phylogroup 10 strains (pink strains at the bottom of the plot) share some *hrp* regulated genes that cluster the strains with phylogroup 2 [43]. Conversely, these two phylogroups primarily differ in phytotoxin biosynthesis genes with the former group (group 10) apparently lacking production of these common phytotoxins as a virulence strategy, with fewer genes (green bars) in group 10, but whether this group contains yet uncharacterized phytotoxin biosynthesis genes remain uncertain.

Besides T3 effectors and phytotoxins, other virulence genes are also ubiquitous across all the phylogroups. Many of these genes are likely involved in niche adaptation via general host defence evasion. These ubiquitous genes are an alginate lyase gene, involved in alginate catabolism[379] and regulation of the T3SS (e.g., core T3SS genes) [330,380], and an apde-like flavin transferase gene. Other virulence genes are more group specific with phylogroups 1, 2, and 6 expressing a xylanase gene for heteropolysaccharide degradation yet missing from phylogroup 3 [381,382]. We attempted to analyze whether the clustering of specific shared virulence gene sets can provide insight into host specificity; however, possible functional redundancy in virulence genes can mask any discernable

pattern of linkage to hosts. For example, two highly conserved *hrpW1* and *hopAK1* genes are putative pectate/pectin lysis genes that have similar functions in cell wall degradation and translocation of T3 effectors [383]. Loss of one gene is unlikely to influence host specificity, but can shift clustering patterns.

For virulence genes known to belong to the *hrp* regulon, a detectable *hrp* box was not consistently predicted, possibly due to sequence differentiation in *hrp* boxes from canonical sequence. More specifically, these genes may not be part of the *hrp* regulon for phylogroups 1,7,11, and partially 4, unlike other groups such as phylogroup 3, which has a clear recognition sequence. For strains containing the canonical T3SS, a lack or lower expression may signify a shift in the importance of these genes in their pathogenic strategy during the transition from a epiphytic lifestyle to an endophytic one [15]. Similarly, an *hrp* box upstream of the *apdb*-like gene was undetectable in all phylogroup 2 strains, in contrast to those in phylogroup 1. The spurious nature of *hrp* box detection indicates that gene presence may partially account for variation in *P. syringae* strain virulence.

Virulence genes are common to putatively non-pathogenic strains with an atypical T3SS

To compensate for a reduced effectome, phylogroup 2 strains utilize a more broad approach to virulence by expressing phytotoxin biosynthesis genes, cell wall degrading genes, and ice nucleation genes; many of these same genes also appear in group 2c strains (Figure 5.11 red box and Table 5.1)[15]. For the putatively non-pathogenic strains of phylogroup 2c, the majority of virulence genes suggest that these strains have retained virulence capability similar to their close relatives. For example, we were able to identify the genes involved in the secretion of syringopeptin and syringolin A [8]. These biosynthesis genes cluster with other phylogroup 2 strains equivalent at greater than 90% protein identity. These genes are located in the same genomic location as well, indicating vertical inheritance from the same MRCA as other group 2 strains. Similarly, ice nucleation genes implicated to trigger endophytic colonization of host plants in group 2 strains are present in phylogroup group 2c.

These strains also contain a unique set of minimally characterized *hrp*-regulated genes, possibly encoding unidentified T3Es (Figure 5.11, rightmost section, light blue genes/bars). The

encoded protein was annotated as disordered, a common trait in T3Es [375], however, corresponding chaperone genes were not found near these genes so it is uncertain whether the protein will be properly folded upon injection into the plant host [149]. Whether these genes are functionally unique to group 2c is currently unclear as they are not well characterized. Similarly, group 13 strains have a minimal effectome where only *hopAJI*, one of the most universal T3E genes, a lytic transglycosylase implicated in T3SS assembly, and an *exoU*-like gene are likely to be part of the T3SS [369]. However, the early branching *P. syringae* group 13 strains lack many of the other virulence genes, with the exception of a non-ribosomal peptide synthase gene and a single gene involved with syringolin A biosynthesis pathway [99]. Collectively, both subphylogroups have the fewest identifiable virulence genes.

Virulence genes in putative non-phytopathogenic *P. syringae* sub-phylogroup 2c

To assess the full gamut of potential virulence genes in the subphylogroup 2c, we analysed the previously identified genes using blastp [384]. Group 2c synthesizes many of the same phytotoxins including syringopeptin, syringomycin, syringolin A, and phaseolotoxin (green bars, Figure 5.11) [99,342], the latter of which is more commonly found in other closely related phylogroup 3. In contrast, this subgroup 2c has a smaller effectome compared to both of the larger phylogroups (group 2 and 3) with only 12 identifiable putative T3 effectors (Table 5.1). Of these 12 effectors, only seven of them appear in all subphylogroup 2c strains with the remaining effectors appearing in smaller subsets. Six of these T3Es are highly conserved in the *P. syringae* complex: *hopAH1*, *hopAJI*, *hopANI*, *hopB2*, *hopM1*, and *avrE1* [151]. The four former genes appear in the same genomic locations as reference group 2 strain PsyB728a, supporting vertical transmission. Conversely, the proximity of the highly divergent variants of the latter two genes, *hopM1* and *avrE1*, indicate concurrent transfer during horizontal acquisition of the AT-PAI (Figure 5.2). A high selective pressure to maintain these genes with the PAI suggests that the genes are possibly PAI-specific and integral in maintaining a suitable host environment for niche survivability [80].

The last conserved T3 effector in the subphylogroup 2c is *hopBHI* [340], an effector that does not appear in either the remainder of phylogroup 2 or 3, suggesting that it was acquired after the gain of the AT-PAI. This gene also appears in a small subset of other strains from a variety of phylogroups including 1, 4, 9, and 10. Similarly, those strains carry the T3E as a singular gene rather

Table 5.1 General virulence genes detected using blastp in two group 2c strains (Pss508 and PsyTLP2). An accession number for a reference protein equivalent is included.

Loci Number						
Gene Family	Pco508	PcoTLP2	Virulence type	Product	Hrp Box	In T3SS cluster
g000489	01619	01490	Misc	Ice nucleation protein	No	No
g001978	01786	01658	Misc	Non-ribosomal peptide synthetases	No	No
g000596	01787	01659	Misc	Non-ribosomal peptide synthetases	No	No
g004746	01788	01660	Misc	non-ribosomal peptide synthetases	No	No
g003681	00915	00814	Misc	Pectate lyase	No	No
g001155_2	01700	01574	Misc	Syringolin A	No	No
g000697	04437	04239	Misc	Xylanase	No	No
g003220	04915	04780	Phytotoxin	Mangotoxin	No	No
g002558	04916	04781	Phytotoxin	Mangotoxin	No	No
g000379	04917	04782	Phytotoxin	Mangotoxin	No	No
g001308	04918	04783	Phytotoxin	Mangotoxin	No	No
g006890	02523	02404	Phytotoxin	Phaseolotoxin	No	No
g006227	02524	02405	Phytotoxin	Phaseolotoxin	No	No
g006821	02525	02406	Phytotoxin	Phaseolotoxin	No	No
g009236	02526	02407	Phytotoxin	Phaseolotoxin	No	No
g005528	02527	02408	Phytotoxin	Phaseolotoxin	No	No
g005840	02528	02409	Phytotoxin	Phaseolotoxin	No	No
g004390	02535	02416	Phytotoxin	Phaseolotoxin	No	No
g000002	01936	01814	Phytotoxin	Syringopeptin	No	No
g002161	02557	02430	Phytotoxin	Syringopeptin	No	No
g003925	02604	02467	Phytotoxin	Syringopeptin	No	No
g005176	02606	02468	Phytotoxin	Syringopeptin	No	No

g005343	02613	02475	Phytotoxin	Syringopeptin	No	No	WP_060709030.1
g004199	02615	02477	Phytotoxin	Syringopeptin	No	No	WP_016568288.1
g006526	02616	02478	Phytotoxin	Syringopeptin	No	No	WP_024692058.1
g005850	02619	02481	Phytotoxin	Syringopeptin	No	No	WP_003408051.1
g000217	00721	-	Phytotoxin	Tabtoxin	No	No	WP_058396127.1
g013649	01583	01454	T3Es	AvrE1	Yes	Yes	WP_201419679.1
g007421	00446	00457	T3Es	ExoU	No	No	WP_024672918.1
g007625	-	-	T3Es	HopAB1e	Yes	No	WP_229687585.1
g003020	-	-	T3Es	HopAF1d	Yes	No	WP_103364865.1
g000262	02990	02874	T3Es	HopAH1as	No	No	WP_236425917.1
g002256	04294	04096	T3Es	HopAJ1	No	No	WP_248688721.1
g003250	00451	00462	T3Es	HopAN1	No	No	WP_096115189.1
g003873	-	-	T3Es	HopAZ1b	Yes	No	WP_004418454.1
g000225	00513	00524	T3Es	HopB2ad	No	No	WP_065831354.1
g006071	01567	01438	T3Es	HopBH1c	Yes	Yes	WP_244986618.1
g013602	01565	01436	T3Es	HopM1	No	Yes	WP_201419664.1
g003872	01258	-	T3Es	HopX1a	No	No	WP_003409095.1

than in the PAI. For a subset of 2c strains, there is also a neighbouring putative T3E gene encoding a disordered protein [375]. Whether this second gene is expressed in conjunction to *hopBHI* is unclear.

Among the remaining effector genes, four of them were not conserved across the subphylogroup: *hopABI*, *hopAFI*, *hopAZI*, and *hopXI*. The genes *hopAFI* and *hopXI* appear frequently among late branching strains but only appear in two strains in group 2c. Furthermore, both genes appear in the same genomic location as reference strain PsyB728a suggesting that poor gene conservation in the subphylogroup 2c is likely due to gene loss rather than gene gain/gene transfer [193]. More specifically, *hopXI* remains in the same genomic location in *P. syringae* pv. *syringae* 508, within the canonical T3SS conserved effector loci, but has since been lost in other strains such as *P. syringae* TLP2. In contrast, the genes *hopABI* and *hopAZI* also appear in only two strains of group 2c, but appear in a different genomic location compared to other strains. In fact, the close proximity of *hopAZI* to a recombinase in a *P. congelans* strains suggests possible gene recombination.

The last putative T3 effector gene in subphylogroup 2c is somewhat anomalous, encoding a *exoU*-like protein. It is found in some group 2c strains but only appears in a small subset of other *P. syringae* strains. In *P. aeruginosa*, ExoU is a T3SS mediated phospholipase cytotoxin capable of killing *Saccharomyces cerevisiae* cells via disruption of lipid metabolism [385]. With a high sequence similarity, at >88% identity with 100% coverage between the *P. syringae*/*P. congelans* and *P. aeruginosa* versions, the ExoU-like protein likely operates in a similar manner. Both proteins share a lipase domain in the C-terminus but differ in the N-terminus. The *P. congelans* ExoU-like protein lacks the disordered N-terminus common to T3Es suggesting that it may not require a chaperone for secretion and folding, which is further corroborated by the lack of a neighbouring chaperone gene. Furthermore, the group 2c strains also lack the remaining *exo* operon indicating that the gene possibly plays a different role [386]. However, an *hrp* box was not detected so the link to *hrp* regulon and T3SS secretion is uncertain.

Many of the aforementioned genes may not belong directly to the *hrp* regulon. Of the 12 effectors found in group 2c, only *hopBHI* and *avrEI* have a discernable *hrp* box upstream of their respective gene loci calling into question whether these strains express these genes via the T3SS [166]. We investigated whether an *hrp* box can be found in equivalent genes in other phylogroup strains. For conserved T3 effectors, only *avrEI* has a consistently detectable *hrp* box using MEME [352]. Conversely, *hrp* boxes have been observed upstream of accessory T3 effectors indicating these genes are expressed as part of the *hrp* regulon, unlike most effector genes in subphylogroup 2c. Notably, a *hrp* box was also not predicted for *hopXI* in group 2c strain *P. syringae* pv. *syringae* 508

compared to other phylogroup 2 strains. However, this gene is situated in the canonical T3SS genomic location rather than the PAI and may be part of the same regulation loop as other genes in that location [151]. The lack of a *hrp* box makes it unclear whether these conserved genes in the previous effector loci retain T3SS-mediated expression and activity during host colonization. However, as an energy intensive process, the retention of some of these genes suggests that they may still play a role in niche adaptation of subphylogroup 2c strains.

Among highly conserved T3 effectors, subphylogroup 2c strains are most notably missing three common T3SS genes: *hrpK*, *hrpW*, and *hopAKI*. All three genes encode harpin T3 translocator and have been implicated in pore-formation crucial to injectisome attachment along with HrpZ [148,155,328]. Group 2c strains appear to retain only the core translocator gene *hrpZ* in core T3SS cluster of the AT-PAI. In the canonical T-PAI, both *hrpK* and *hrpW* are located in the conserved effector loci but are absent in both the AT-PAI effector loci and the canonical genomic location of the T-PAI's effector loci in subphylogroup 2c strains. The encoded translocators contain either a pectate and pectin lyase fold that has been speculated to degrade the host's cell wall and facilitate docking of the pilus to the host cell [328]. For targeting mammalian hosts, these proteins require polymerization with each other for efficient pore formation and pilus attachment. However, compensatory genes for additional translocators were not identified in group 2c suggesting the AT-PAI may possibly have reduced efficiency or variability in T3 effector translocation.

Other missing highly conserved T3E genes include *hopAAI*, and *hopDI*. The gene *hopAAI* encodes a T3effector that triggers chlorosis through the catabolic disruption of the arginine pathway in the host plant [374]. The absence of this T3E gene and alternative chlorosis inducing genes, such as phaseolotoxin biosynthesis [342], are supported by the inability of *P. syringae* pv. *syringae* strains 508 and *P. syringae* TLP2 to trigger chlorosis in *P. vulgaris* cv. Tendergreen leaves (Appendix A). The lack of conserved *hopDI*, a gene involved in maintaining a suitable host environment by inhibiting NTL9 induced genes, indicates that this subgroup may be less able to manipulate host plant environment via traditional means [387]. However, subphylogroup 2c strains may express alternative genes capable of functioning similarly, or, have currently uncharacterized genes that may fill this functional role. Overall, the lack of T3 effectors in subphylogroup 2c with known interactions with host defence system suggests that this subphylogroup have a highly reduced virulence restricting it to in host survivability and specificity.

Although reduced, the subphylogroup 2c still retains an effectome and is capable of proliferating in some host plants [18], suggesting it may have as-of-yet unidentified effectors. Four out of the 10 genes identified as within the group 2c *hrp* regulon are hypothetical genes encoding proteins unique to these strains. Three contain intrinsically disordered domains, common to T3 effectors, while the fourth gene encodes a ParB-like nuclease, likely involved gene partitioning and may be remnant of plasmid integration [335]. Although not part of the *hrp* regulon, a subset of subphylogroup 2c strains also contains a gene encoding a Rap1a-like immunity protein with a signal peptide. The expression and secretion of this protein may confer some protection against Ssp1 endopeptidases during bacterial competition rather than host plants [388]. The uniqueness of this reduced effectome suggests it may have transitioned to a niche role, either in target host plant selectivity or ecological niche within the microbial population of its target host.

Divergence of group 2c genes in common gene clusters

We analysed the pangenome for other niche adaptation genes, similar to the atypical T3SS, which are specific to the subphylogroup 2c strains. In this analysis we targeted highly divergent gene families that are shared between subphylogroup 2c strains and the *P. syringae* complex. All Gene families were accessed for unique clusters comprised solely of subphylogroup 2c strains when stringency was increased to 90% identity. Similar to our *hrp* regulon, the divergent clusters were assigned two transcriptional unit ids based on proximity to each other according to genomic location on Pss508 and PsyTLP2, the only two complete genomes in the subphylogroup (Chapter 3). Furthermore, closely positioned transcriptional units (<2000 bp) were manually identified as a potentially divergent region.

As expected, the majority of divergent gene families were single genes shared by a small subset of *P. syringae* strains and subphylogroup 2c strains. The differences between the strains are scattered in the genome, perhaps due to differential selective pressure unique to each strain. Among the gene families conserved in subphylogroup 2c strains, the two largest divergent gene clusters were the two core clusters in the atypical PAI suggesting it is one of the key distinguishing genomic features separating this subphylogroup from other closely related strains. Although these gene clusters are shared with phylogroup 13, this variant of the AT-PAI now shares less than 80% identity with the most conserved gene family and is reduced to 50% identity for the most dissimilar gene families of the region.

Another divergent region containing clusters of interest encode type I fimbriae-like genes putatively involved in general cell adhesion during pathology (Table 5.2)[389]. This genomic region is a highly variable loci appearing in all strains. Evidence of phages, such as a P22-like autotransporter gene, suggests that gene mobility contributes to the diversity of this region. The most variable element is a pair of RHS-repeating genes, which support that this region is a rearrangement hotspot [390]. Although these genes appear as a pair in *P. syringae* pv. *syringae*_508 and TLP2, many strains contain a sole copy; whether this discrepancy in copy number arose from gene gain/loss or limitations in short read sequencing of repeat regions is unclear. In conjunction with the putative use of type I fimbriae-like proteins for cell adhesion, the presence of these RHS-repeating putative toxins suggests the region confer some niche adaptation for *P. syringae* strains. With the possibility of secretion by the T6SS, these toxins likely play a role in niche survival against other bacterial competition upon cell adhesion, especially during early stages of host colonization [29].

Table 5.2 A gene family cluster (cluster 784-787) in group 2c strains that contain genes that have diverged from both closely related phylogroups 2 and 3. Encoded proteins from these genes cluster away from orthologs at 70-80% sequence identity. Number of Psy genomes indicate all the strains that have the gene family. Gfam pid indicates the lowest percent identity needed to cluster the encoded protein products. Allele breakpoint indicates the percent identity threshold needed for encoded protein from group 2c variants to form its own unique cluster. A breakpoint below 90% is considered sufficient sequence divergence from closely related orthologs.

Gene family	No. of Group2c genomes	No. of Psy genomes	Gfam pid	Allele breakpoint	Product
g005056	11	25	60	80	hypothetical protein
g003698	11	194	50	80	hypothetical protein
g000003	11	417	50	80	RHS repeat-associated core domain-containing protein
g000509	11	356	50	80	hypothetical protein
g006390	11	20	60	70	hypothetical protein
g007610	11	20	60	70	lipoprotein
g000003	11	417	50	80	RHS repeat-associated core domain-containing protein
g000703	11	387	50	80	helix-turn-helix transcriptional regulator
g000016	5	231	60	70	hypothetical protein
g007584	10	12	60	60	hypothetical protein

Differences between group 2c strains *P. syringae* pv. *syringae* 508 and TLP2

Despite the close similarities of subphylogroup 2c strains, colonization and persistence of sister strains, such as *P. syringae* pv. *syringae* 508 and *P. syringae* pv. *syringae* 642, in a target plant can trigger varying host response [18]. We observed a similar result effect when *P. syringae* pv. *syringae* 508 and *P. syringae* TLP2 were directly inoculated in *Phaseolus vulgaris* cv. Tendergreen (Appendix A). As previously reported, *P. syringae* pv. *syringae* 508 had a consistent response only after direct inoculation of high concentration (1×10^8 cells/ml) of the bacterium triggering hypersensitive (HR)-like symptoms in the bean host similar to *P. syringae* pv. *syringae* 642. Conversely, *P. syringae* TLP2 was incapable of triggering a HR-like response in the same host regardless of inoculation concentration suggesting some variability in host defence response to similar strains. To investigate this variability at the genomic level, we directly compared the gene sets of both *P. syringae* pv. *syringae* 508 and *P. syringae* TLP2 to each other.

At the sequence level, the genome of Pss508 is larger than PsyTLP2 by approximately 100 kbp (Table 2.3). The majority of this difference is housed in two regions composed of 50kb and 10kb respectively. The remaining ~40kb is scattered throughout the genome primarily as single gene differences. Many of the genes within the largest two regions are strain specific indicating that potential horizontal inheritance. In fact, mobile gene elements such as integrases, transposases, and insertion sequences, appear throughout the additional regions in Pss508. In the 10kb region there is evidence of phage insertion with the presence of genes encoding phage tails, Cro/CI suppression, and SOS response genes, which are hallmarks of a prophage [391]. Notably, this region is situated in the canonical location of the T-PAI exchangeable effector loci where prophage insertion has been observed in other *P. syringae* strains within the EEL of the tripartite genomic island (T-PAI). Furthermore, the parts of the region associated with lateral transfer are marked by a change in GC content from the typical 58-59% common to *Pseudomonas* strains. PsyTLP2 in contrast hosts a separate, much smaller, collection of hypothetical genes that are specific to the strain, as expected with high gene mobility of the region. These genes were not detected to be a part *hrp* regulon so the role as putative effectors is unclear.

Among known virulence factors, the genome of Pss508 contains an additional effector gene *hopX1* [369] and also a gene encoding a TraG-like translocator putatively DNA transfer during bacterial conjugation [392]. Comparatively, additional effectors can be extra factors for host

recognition for response variation especially with T3SS-mediated virulence where gene loss can be imperative to host evasion. In contrast, PsyTLP2, like Psy642, contains an additional but unusual putative T3 effector gene in AT-PAI that encodes a Rap1a immune protein homolog. In other bacteria, Rap1a provides immunity in type VI secretion system mediated toxin-antitoxin competition between microbes through binding inactivation of SSP1 endopeptidases [388]. The homolog may provide a more general protective effect rather than specific host manipulation.

The *hrp* regulon also differs between the two strains. During T3SS expression, PsyTLP2 appears to regulate a gene encoding an oxidoreductase found only in subphylogroup 2c and subphylogroup 13 strains. The encoded protein has been annotated as a putative NAD(P)-dependent alcohol dehydrogenase with an enoylreductase domain. This domain is proposed to synthesize antibacterial type I polyketides and has been involved in fungal virulence of rice plants [393].

Conversely, one known general virulence factor in Pss508 and not PsyTLP2 is a singular gene that has been identified to be a part of the tabtoxin biosynthesis pathway (Table 5.3)[378]. This gene presence can indicate Pss508 is capable of synthesis of an additional putative secondary molecule similar to tabtoxin.

Unlike Pss508, the genome of PsyTLP2 also contains a putative virulence gene cluster, comprised of up to three genes, that is shared between 14 strains in the observed pangenome. The first gene encodes a LysE translocator, a lysine, histidine, and arginine transporter followed by a gene encoding a caspase-like C14 endopeptidase, which may be involved in plant programmed cell death [394]. The last gene encodes a hypothetical protein and appears in eight genomes including four subphylogroup 2c strains (*P. syringae* 642, *P. congelans* ME812-2B-00425, *P. congelans* NL-P123-02549, and *P. syringae* TLP2). Although this region appears in the four subphylogroup 2c strains, Psy642 clusters more closely with phylogroup 1 strains. Furthermore, Psy642 and the phylogroup 1 strains lack the last gene. In contrast, the other three strains cluster with phylogroup 2 and include the third gene that is missing in Psy642. Each of these gene family clusters are maintained at >90% identity indicating different origins for the Psy642 version and the other three phylogroup 2c strains, despite very similar functional roles.

Table 5.3 Accessory genome clusters that appear in select group 2c strains that contain genes that have diverged from homologs found in other Psy strains: Pss508 (top) and PsyTLP2 (bottom). Number of Psy genomes indicate all the strains that have the gene family. Gfam pid indicates the lowest percent identity needed to cluster the encoded protein products. Allele breakpoint indicates the percent identity threshold needed for encoded protein from group 2c variants to form its own unique cluster. A breakpoint below 90% is considered sufficient sequence divergence from closely related orthologs. An accession number for a reference protein equivalent is included.

Gene family	No. of Group2c genomes	No. of genomes	Gfam pid	Allele breakpoint	Product	Ref protein (Accession No.)
<i>Cluster 1530-1551</i>						
g008928	1	10	50	80	integrase	WP_116817155.1
g010562	1	6	60	80	hypothetical protein	WP_054091190.1
g016121	1	8	70	70	hypothetical protein	WP_214423019.1
g007534_2	1	19	80	80	DUF262 domain-containing protein	WP_214423024.1
g002325	1	61	60	80	hypothetical protein	WP_032673169.1
g009994	1	4	60	70	hypothetical protein	WP_103688886.1
g013484	1	12	50	60	hypothetical protein	WP_214423035.1
g010663	1	17	60	80	hypothetical protein	WP_005779491.1
g007231	1	19	50	70	DNA recombination protein RecF	WP_024694182.1
g002343	1	52	50	80	hypothetical protein	WP_235389248.1
g015565	1	2	60	60	hypothetical protein	WP_046482354.1
g009137	1	7	60	60	hypothetical protein	WP_214423048.1
g009237	1	7	80	80	Cro/CI family transcriptional regulator	WP_057418604.1
g012618	1	3	50	60	hypothetical protein	WP_044344620.1
g007442	1	17	60	70	serine/threonine protein phosphatase	WP_010954525.1

g005513	1	17	50	70	hypothetical protein	WP_046463730.1
g008874	1	18	70	80	hypothetical protein	WP_020308634.1
<i>Cluster 1036</i>						
g010508	1	5	70	70	LysE family translocator	WP_122398342.1
g010140	3	13	80	80	peptidase C14	WP_044308077.1
g009221	3	8	50	80	hypothetical protein	WP_141250276.1

5.4 Discussion

Acquisition of the atypical PAI in subphylogroup 2c strains is a key differentiating factor of the subgroup from other late-branching phylogroups

The subphylogroup 2c is a small but distinct sub-group that was recently identified within its parental phylogenetic clade, group 2, the latest branching monophyletic group of the *P. syringae* species complex [44]. However, among known *P. syringae* strains, this subgroup is small and underrepresented with only a few strains reported in previous studies [43,44,49]. Isolated from healthy plants, at least one subphylogroup 2c strain, *P. syringae* 642, has been found to be persist in *Phaseolus vulgaris* (bean), but is putatively non-pathogenic unless overexpressing known avirulence gene, *avrRpt2*, via mutagenesis triggering an HR response in the host [18]. This subphylogroup 2c also contains *P. congelans*, a *Pseudomonas* species first isolated in the phyllosphere of grasses, and was previously classified as belonging to the *P. fluorescens* species group in NCBI [19].

In this work, using core genome phylogeny, we were able to expand this subgroup to 11 strains, including our two previously sequenced complete genomes for strains *P. syringae* pv. *syringae* 508 (Pss508) and *P. syringae* TLP2 (PsyTLP2)(Chapter 3). Similar to *P. syringae* 642, Pss508 was previously reported to be non-virulent to *P. vulgaris* without mutagenesis, along with PsyTLP2, having an undetectable host reaction despite an active T3SS [18]. Avirulence in other strains has been typically attributed to T3SS disruption in the encoding PAI [22]. Despite an active T3SS and without a detectable host plant for any strains, the ambiguously non-virulent phenotype of these strains separates this subgroup from the well characterized, agriculturally virulent of closely related late branching *P. syringae* phylogroups (Figure 5.1, groups 1,2, and 3).

At the genomic level, these strains strictly adhere to this subphylogroup (PG-2c) clustering using both the core genome sequence and gene set presence/absence clustering. These strains are highly similar to each other sharing around 98%+ sequence similarity to each other compared to 94% or lower with other subgroup strains (Chapter 3), forming a distinguishable subclade consistent with earlier studies. In our analysis, other phylogroups show minor branching variations with a few group 2 strains clustering together somewhat differently than previously reported, suggesting that some subgroups within phylogroup 2 may not be as clearly defined [43,44,57].

Gene set clustering provides some insight in the divergence of subphylogroup 2c (Figure C.3). As a whole, subphylogroup 2c shares the largest portion of its accessory genome with the other group 2 strains. However, our data suggests two gene-set based factors are key differentiators. The

first factor is a shared group of genes that are exclusive to this subgroup (Figure C.2). Upon investigation, we found that this gene set is a part of the T3SS PAI. Despite the conserved genes of the T3SS PAI being found in nearly all strains, with two exceptions, the more variable genes of the pathogenicity island were highly divergent and shared less than 50% protein identity with the two more canonical counterparts (Figure 5.3). This differentiated them as distinct gene families and suggests an alternative source of PAI acquisition rather vertical inheritance. The secondary key distinction of this group is the reduced number of shared gene sets with other late branching groups when compared to general group 2 strains (Figure 5.9). The high mobility of T3Es has been postulated as a key contributing factor to the high host diversity of the species complex, and is supported by remarkably high intra-phylogroup transfer of genes between closely related strains [43]. However, unlike other *P. syringae* subgroups, gene set clustering indicates that only g2 strains collectively demonstrate the same loss of accessory genes that are otherwise common to late-branching strains (Figure 5.11). Taken together, both key subphylogroup differentiating factors suggestive of a divergence event at the MRCA of this subgroup and later branching group 2 strains following the acquisition of atypical PAI encoding the third, non-canonical T3SS.

Horizontal acquisition of the T3SS encoding AT-PAI in *P. syringae* subphylogroup 2c

A defining characteristic of all *P. syringae* is the presence of T3SS and a feature of the subphylogroup 2c strains is the presence of the atypical T3SS. The high sequence divergence of the AT-PAI separates it from the canonical variant (T-PAI) [151], making it difficult to detect using standard DNA probes. Initial characterization of the AT-PAI previously revealed the differences in the genes flanking the PAI indicate a change in the localization of the genomic island within the genome. Our analysis of the AT-PAI supports the notion of the MRCA of subphylogroup 2c strains acquiring AT-PAI through horizontal gene transfer. With the complete assembled genome of Pss508 and PsyTLP2, we were able to confirm that AT-PAI are no longer flanked by canonical T-PAI genes *gstA* and *queA*, and is situated between a gene encoding a sodium:bile acid symporter and *rdgC* gene with the latter gene indicating possible recombination [366]. The PAI location was confirmed in all subphylogroup 2c strains supporting subphylogroup 2c MRCA acquisition in a singular event (Figure 5.4). However, recombination of highly divergent sequences is a barrier to gene transfer, with most

inter-phylogroup gene transfer occurring between closely related strains allowing high levels of accessory sharing.

To address sequence divergence from PAI origin, our gene set clustering analysis of the PAI gene products focused on broader protein families formed by grouping homologous genes/proteins at a lower protein identity. For example, protein products of gene duplicates *hrpR* and *hrpS* are considered a singular broad protein family but separate paralogs at higher stringency [170,321]. All strains were clustered into groups falling within the three main T3SS variants, with the T3SS phylogenetic tree having long branches that were entirely incongruent with core genome phylogeny (Figure 5.4). In addition, a subset of group 13 *P. syringae* strains, distant from group 2 in the core phylogeny, shared nearly the entire PAI gene set with the only exceptions being genes found in the effector loci flanking the AT-PAI. Other than the highly conserved *avrE*, effector genes were undetectable in group 13. At higher identity thresholds the grouping between subphylogroup 2c and group 13 disappears, sharing only the conserved T3SS (*hrc*) genes in common. Less conserved T3SS genes typically had less than 70% identity between the two groups, placing them within broad protein families. Gene equivalency of these protein families was confirmed via gene order and orientation in their respective genomic islands, including the conserved avirulence gene *avrE* (Figure 5.2).

Our MLSA analysis of all shared PAI gene families across all *P. syringae* strains confirms that the core T3SS gene cluster in both subphylogroup 2c and group 13 are closely related (Figure 5.4). The PAI of these two groups are nearly equally distant from core genes of both S-PAI and the canonical T-PAI, supporting that horizontal inheritance of this divergent T-PAI from a shared origin. An alternate hypothesis for PAI relatedness is a possible recombination event between the MRCA of both subgroups. However, both the presence of an external effector loci flanking the core genes solely in PG-2c and non-canonical genomic localization invalidates this possibility. In order for this to happen two recombination events would need to occur: a re-localization event with a tripartite PAI and recombination of core genes, where sequence divergence would be a barrier for recombination [395].

Effector loci proteins for the subphylogroup 2c AT-PAI did not cluster with known effectors from other *P. syringae* strains, and yielded no hits within NCBI's nr database indicating no known homolog. These suggest that an intermediary remains to be discovered. Furthermore, for group 13, only 40% of strains carry the AT-PAI, with the remaining three strains missing any detectable form of

Pseudomonas T3SS encoding genes. Vertical inheritance of the group 13 PAI from the MRCA of *P. syringae* species complex is possible, along with loss in the other group 13 strains, however this is unlikely.

Inter-phylogroup transfer of the AT-PAI in subphylogroup 2c is further supported by some evidence of a similar event occurring between the MRCA of late branching phylogroups 2 and 3 with early branching phylogroups 5. Core T3SS gene families MLSA suggests that the T-PAI between phylogroups 2 and 5 share a MRCA that diverged from the MRCA of phylogroup 3's T-PAI (Figure 5.4). However, the core T3SS phylogeny tree contradicts previous phylogenetic analyses based on highly conserved T3SS core genes, such as *hrcV* [43]. The latter analyses are more concordant with species phylogeny indicating vertical inheritance of the T-PAI.

To further contrast with previous work, our analysis of non-conserved *hrpA* and associated *hrpF* appears to further support horizontal recombination of the T3SS cluster across phylogroups. Forming the external pilus and susceptible to plant detection, HrpA is one of the most variable protein products in the T3SS gene cluster [159]. Across all the strains, *hrpA* form four distinct gene families: AT-PAI's *hrpA*, S-PAI's *hrpA*, and T-PAI's variants, *hrpA1* and *hrpA2* [396]. With less than 50% protein identity, these homologs are unlikely to be direct orthologs obtained via direct vertical inheritance. This variance is corroborated with the grouping of HrpF, a more conserved protein that binds to HrpA during regulation of pilus construction [176]. Taken together, the PAI between these phylogroups may have exchanged the PAI either as a whole or in parts. Furthermore, the incongruence between different phylogenetic analyses, sparking debates on T3SS divergence in *P. syringae*, may be partly confounded by higher recombination of core gene cluster between phylogroups than expected [338].

The atypical PAI encodes a separate hybrid-like T3SS variant distinct from the two common PAIs

Variations in T3SS gene clusters have been observed in across other bacteria species [150] . For the well-characterized mammalian pathogen *Salmonella*, the two variants SPI-1, and SPI-2 serving differing functions and expressed under different environmental stresses [160]. The former is expressed during active invasion of non-phagocytic cells for short-term survival whereas the latter promote long term survival via host environment manipulation. The different functional niche of both

T3SS is reflected in the different structural composition of the respective PAI including rearrangement and missing genes between the two T3SS [150]. In contrast, *P. syringae* T3SS variants have very similar gene arrangement and order with the atypical variant displaying hybrid-like aspects from both of the more common variants, T-PAI and S-PAI (Figure 5.2 and Figure 5.4).

All three variants of the T3SS encoding PAI carry a core T3SS gene cluster split by a hypervariable region containing additional T3E genes [364]. However, unlike the S-PAI, both tripartite PAIs have effector loci flanking the core gene cluster. While the canonical T-PAI typically contains an upstream conserved effector loci (CEL) and a downstream highly variable exchangeable effector loci [156], the AT-PAI has a significantly reduced effector loci on either end of the gene cluster. These clusters likely contain one to two possible effector genes with the conserved *hopMI* having an identifiable accompanying chaperone gene [96]. In contrast, the hypervariable portion of S-PAI acts as the respective effector loci containing the conserved effector protein *avrEI*, and other effector genes [152]. Unusually, the AT-PAI shares this characteristic with S-PAI with its *avrEI* situated within the hypervariable region rather than CEL typical of T-PAI (Figure 5.2). In addition, AT-PAI shares the gene arrangement of S-PAI with a singular initial PAI regulator, *hrpR/S*, and a fused *hrcQ*, unlike the split subunits in T-PAI. Conversely, conserved core genes of AT-PAI are slightly closer in sequence to T-PAI core genes, forming higher identity gene family clusters compared to S-PAI (Figure 5.3, 5.4, and 5.6). The hybrid like characteristics of AT-PAI to both of the other two PAI may suggest a link between all three variants and its dissemination across the *P. syringae* species complex. With the incongruence of T3SS gene trees and core genome phylogeny of the species, recombination of the T3SS across phylogroups indicates a rapid change in its niche.

Correlation of putative non-pathogenicity of known subphylogroup 2c strains with AT-PAI acquisition suggests a potential transition in niche

For microorganisms, genomic adaptations typically reflect a transition where virulence genes for pathogens can be viewed synonymously as niche adaptation genes [104]. For subphylogroup 2c strains, use of an atypical T3SS has been speculated to indicate an adaptation for different niche than its closely related group 2 strains. The gain of a novel *hrpA* uncommon to *P. syringae* strains may aid in evading a common method of plant recognition. Furthermore, subphylogroup 2c strains lack the *hrpK* gene common to the CEL of canonical T-PAIs [148]. Other common T3 harpin translocators

hrpW1 and *hopAK1* are also missing from the subgroup genomes. These respective expressed proteins, along with core gene *HrpZ*, initialize interaction with target host cells via pore formation prior to pilus attachment. However, we were able to identify an *HrpK*-like protein in the truncated CEL supporting an active state of expressed T3SS. Since these harpins were reported to dimerize during pore-formation, the concomitant loss of these genes would not be unusual [328]. Although still capable of pore-forming with the novel *hrpK*-like gene, without dimerization efficacy the activity may be reduced, similar to the case where high T3SS expression by *HrpR* requires dimerization with *HrpS* for enhanced protein activity[168], or indicate a separate target entirely.

Additionally, the severely truncated effector loci on either end of AT-PAI suggest that gain of the AT-PAI is correlated with a substantial loss of effectome for subphylogroup 2c strains. For many *P. syringae* strains, the EEL constitutes a large section of a strain's effectome diversity [156]. Furthermore, our analysis of common known effector genes and *hrpL* regulated genes reveals that highly conserved genes across the genome, e.g. *hopAA1*, were also missing, and of the effector genes found, e.g. *hopX*, a canonical *hrpL* promoter could not be identified raising uncertainty in the expression and activity of these effector genes [397]. Conversely, a novel set of potential effector genes were identified using *hrpL* promoter box analysis, many of which are currently uncharacterized. Protein products of these genes were predicted to have disordered regions, a common characteristic of effector proteins. Others contained putative secretion signals suggesting some role in subphylogroup 2c strain niche adaptation and possible virulence.

Although the effectome is further reduced when compared to other *P. syringae* phylogroups, the expression of the phytotoxin phaseolotoxin similar to phylogroup 3 strains and non-ribosomal polyketide synthases, such as syringolin A, common to phylogroup 2 [158] may compensate for the lack of effectors. In contrast to effectors, these genes target initial host invasion. Phaseolotoxin is a tripeptide inhibitor of ornithine transcarbamoylase activity in *P. vulgaris* and prevents accumulation of innate immune response [342]. Syringolin A prevents stomal closure via protease inhibition [398]. None of the identified virulence genes target late-stage effector-based immunity [75]. Furthermore, identifiable subphylogroup 2c effector proteins AvrE1 and HopM1 primarily target bacterial survival with osmoregulation via promoting an aqueous environment [80,343]. Direct plant immunity targeting effector proteins, e.g., effector protein HopA1 inactivation of plant MAPK signalling proteins [374], were not identified raising uncertainty in the ability of subphylogroup 2c strains to directly combat the plant immune response. However, the sequence divergence of *hrpL* in AT-PAI

hinders full effectome identification, therefore subphylogroup 2c may contain novel T3Es that would facilitate virulence in these strains especially for a novel hosts. The subphylogroup 2c effectome can also be easily expanded via recombination or a novel gain-of-function gene triggering a transition from non-virulence to virulence [395].

Subgroup 2c's reduced effectome may fill a similar niche to recently reported low T3SS expressing subpopulations of *P. syringae* strains [30], especially when AT-PAI have only a singular copy of *hrpR/S*. In the canonical arrangement, both variants are required to dimerize to upregulate T3SS expression [168]. Without dimerization, canonical HrpS was only capable of low level expression suggesting AT-PAI may also have lower level of expression than T-PAI strains. In contrast, S-PAI contain a singular HrpR and these *P. syringae* strains are fully capable of virulence [154]. With an active T3SS, subphylogroup 2c strains may have higher fitness at initial phases of host invasion with many of the identified genes either promoting mobility into the target host and sustainable environment. Pss508 was identified with a second T6SS gene set in canonical EEL genomic location suggesting increased fitness in inter-microbial interactions [388]. During later stage of invasion, similar to weakly virulent low T3SS expressing *P. syringae* strains, subphylogroup 2c strains may rely on more virulent *P. syringae* strains for immunity suppression [29]. However, the effect of different *P. syringae* strain mixtures on plant colonization remains largely uncharacterized.

Minor differences in gene sets between individual strains, *P. syringae* pv. *syringae* 508 and *P. syringae* TLP2, can account for variation in host plant response

As with other *P. syringae* strains, individual gene sets can confer different responses in host plants. Slight changes in host response for Pss508 and PsyTLP2 can be attributed to differing gene sets. Although there are no indications of *hrpL* based expression, Pss508 contains an additional effector gene, HopX [369], and partial tabtoxin biosynthesis pathway [378]. Conversely, PsyTLP2 shares an ExoU-like toxin with *P. syringae* 642 [18,386]. With the variability in expressible virulence gene between the two strains, any of these genes can provide a point of contact for plant recognition where a single gene can trigger pathogen avirulence via the hypersensitive response [75]. However, even with current wealth of genomic data, identification of effector genes needs to be paired with expression data. Host invasion and virulence is highly temporally regulated with the expression, and assembly, of T3SS occurring in precise stages [330]. With T3Es targeting different

aspects of the plant immunity, future study on the T3Es expression of subphylogroup 2c will provide further insight into the niche of this putatively non-virulent subphylogroup.

Chapter 6:

Conclusions

In this dissertation, I aimed to identify genetic factors (i.e., divergent gene clusters, and pathogen-associated genes) that contribute to phenotypic and phylogenetic differentiation of subphylogroup 2c *Pseudomonas syringae* strains from closely related virulent strains [18]. For comprehensive coverage of highly mobile pathogen-associated genes, I sequenced and assembled the complete genomes of two subphylogroup 2c strains, *P. syringae* pv. *syringae* 508 and *P. syringae* TLP2, using Oxford Nanopore Technology's nanopore sequencing [7,18,66]. I also developed Kastor, an error assessment tools that utilizes publically available reference information to polish both assemblies to completion.

With these complete genomes, I was able to discern that the differentiation of *P. syringae* subphylogroup 2c strains likely arose from the horizontal acquisition of the atypical type III secretion system (T3SS) pathogenicity island (PAI), a divergent gene cluster previously only identified in this subphylogroup [18]. Two gene clusters with similar gene arrangement, but divergent in sequence, were also identified in a subset of group 13 strains. Through protein homology of core T3SS, these gene clusters were identified as the core T3SS gene cluster encoding the T3SS in group 13 strains. Using multilocus sequence analysis of homologous core genes, PAIs in phylogroups 2c and 13 were revealed to share the same recent ancestor. Although variation in T3SS has been linked to different specialization in other pathogens [160], not only does the Atypical T3SS PAI share similarity to the other *P. syringae* T3SS variants, the group 2c and 13 PAI are mutually exclusive with the other T3SS PAI variants with *P. syringae* strains carrying a single copy of the T3SS PAI [151,153,154]. Thus, I suggest this T3SS shares the same functional niche as the other two variants, and this is supported by activity assays from a previous study by Clarke et al (2010).

This comparative study supports the conclusion that the subphylogroup 2c strains likely fulfill a similar ecological niche as other weakly virulent strains [29,30]. Complete genome sequencing with long reads provided the resolution necessary to confirm a highly reduced effector loci in both the hypervariable region and flanking regions of the PAI, decreasing the subphylogroups' ability to enact common infection strategies [90]. Consequently, acquisition of the atypical T3SS coincided with a highly reduced effectome and missing conserved effector proteins. However, presence of likely replacement novel effector genes supports weak virulence for niche adaptation.

Therefore, I propose that the non-virulence phenotype of the *P. syringae* subphylogroup 2c strains arose from the acquisition of the atypical T3SS leading to loss of non-essential effector gene for plant colonization triggering adaptation to a new lifecycle niche apart from pathogenicity (e.g., epiphyte specialization, or early plant colonization specialization) [15,30].

Identification of the structural variation between the PAIs and former effector loci was possible with complete genome assembly of the two subphylogroup 2c strains, *P. syringae* pv. *syringae* 508 and *P. syringae* TLP2 using nanopore long reads. With the low upkeep, portable, quick turnaround time and ease of use, I found that the accessibility of this sequencing system reduces the barrier to complete genome sequencing for this work's smaller and more targeted sequencing project, especially for the need for good resolution of a highly mobile genetic cluster. A single sequencing run can provide ample yield to assembly more than one bacterial genome. Conversely, I also found that the high error rate and, at the time, lack of software, can hamper downstream analyses of the assemblies [238]. I was able to compensate for higher error rate by developing the error detection software, Kastor. Kastor utilizes either reference genome or protein mapped to the draft assembly to identify gene-fragmenting errors. Correction of the detected errors can polish the draft assembly to completion. When combined with maturation of the technology, I suggest that the use of this sequencing system would be highly beneficial to future comparative bacterial research.

References

1. Glick BR. Beneficial plant-bacterial interactions. Heidelberg: Springer International Publishing; 2015.
2. Brown C. Agronomy Guide for Field Crops - Publication 811. Toronto, Ontario: Queen's Printer for Ontario; 2022.
3. Dean R, Van Kan JL, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, et al. The top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*. 2012;13:414–30.
4. Mansfield John, Genin Stephane, Magori Shimpei, Citovsky Vitaly, Sriariyanum Malinee, Ronald Pamela, et al. Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology*. 2012;13:614–29.
5. Kamoun S, Furzer O, Jones JDG, Judelson HS, Ali GS, Dalio RJD, et al. The top 10 oomycete pathogens in molecular plant pathology. *Molecular Plant Pathology*. 2015;16:413–34.
6. Scholthof K-BG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, et al. Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology*. 2011;12:938–54.
7. Burr TJ, Matteson MC, Smith CA, Corral-Garcia MR, Huang T-C. Effectiveness of bacteria and yeasts from apple orchards as biological control agents of apple scab. *Biological Control*. 1996;6:151–7.
8. Grgurina I, Bensaci M, Pocsfalvi G, Mannina L, Cruciani O, Fiore A, et al. Novel cyclic lipodepsipeptide from *Pseudomonas syringae* pv. *lachrymans* strain 508 and syringopeptin antimicrobial activities. *Antimicrobial Agents and Chemotherapy*. 2005;49:5037–45.
9. Jani AJ, Cotter PA. Type VI secretion: not just for pathogenesis anymore. *Cell Host & Microbe*. 2010;8:2–6.
10. Conn VM, Walker AR, Franco CMM. Endophytic *Actinobacteria* induce defense pathways in *Arabidopsis thaliana*. *Molecular Plant-Microbe Interactions*. 2008;21:208–18.
11. de Oliveira Costa LE, de Queiroz MV, Borges AC, de Moraes CA, de Araújo EF. Isolation and characterization of endophytic bacteria isolated from the leaves of the common bean (*Phaseolus vulgaris*). *Brazilian Journal of Microbiology*. 2012;43:1562–75.
12. Haney CH, Wiesmann CL, Shapiro LR, Melnyk RA, O'Sullivan LR, Khorasani S, et al. Rhizosphere-associated *Pseudomonas* induce systemic resistance to herbivores at the cost of susceptibility to bacterial pathogens. *Molecular Ecology*. 2018;27:1833–47.
13. de Lamo FJ, Takken FLW. Biocontrol by *Fusarium oxysporum* using endophyte-mediated resistance. *Frontiers in Plant Science*. 2020;11:37.

14. Rastogi G, Coaker GL, Leveau JHJ. New insights into the structure and function of phyllosphere microbiota through high-throughput molecular approaches. *FEMS Microbiology Letters*. 2013;348:1–10.
15. Hirano SS, Upper CD. Bacteria in the leaf ecosystem with emphasis on *Pseudomonas syringae* - a pathogen, ice nucleus, and epiphyte. *Microbiology and Molecular Biology Reviews*. 2000;64:624–53.
16. Monteil CL, Cai R, Liu H, Llontop MEM, Leman S, Studholme DJ, et al. Nonagricultural reservoirs contribute to emergence and evolution of *Pseudomonas syringae* crop pathogens. *New Phytologist*. 2013;199:800–11.
17. Xie T, Wu X, Luo L, Qu Y, Fan R, Wu S, et al. Natural variation in the *hrpL* promoter renders the phytopathogen *Pseudomonas syringae* pv. *actinidiae* nonpathogenic. *Molecular Plant Pathology*. 2023;24:262–71.
18. Clarke CR, Cai R, Studholme DJ, Guttman DS, Vinatzer BA. *Pseudomonas syringae* strains naturally lacking the classical *P. syringae* *hrp/hrc* locus are common leaf colonizers equipped with an atypical type III secretion system. *Molecular Plant-Microbe Interactions*. 2010;23:198–210.
19. Behrendt U, Ulrich A, Schumann P. Fluorescent pseudomonads associated with the phyllosphere of grasses; *Pseudomonas trivialis* sp. nov., *Pseudomonas poae* sp. nov. and *Pseudomonas congelans* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*. 2003;53:1461–9.
20. Sun X, Griffith M, Pasternak JJ, Glick BR. Low temperature growth, freezing survival, and production of antifreeze protein by the plant growth promoting rhizobacterium *Pseudomonas putida* GR12-2. *Canadian Journal of Microbiology*. 1995;41:776–84.
21. Mohr TJ, Liu H, Yan S, Morris CE, Castillo JA, Jelenska J, et al. Naturally occurring nonpathogenic isolates of the plant pathogen *Pseudomonas syringae* lack a type III secretion system and effector gene orthologues. *Journal of Bacteriology*. 2008;190:2858–70.
22. Li Y, Zhu Q, Zhi T, Fan R, Xie T, Zhao Z, et al. Genetic Causes of Non-pathogenic *Pseudomonas syringae* pv. *actinidiae* Isolates in Kiwifruit Orchards. *Frontiers in Microbiology*. 2021;12.
23. Marchetti M, Capela D, Glew M, Cruveiller S, Chane-Woon-Ming B, Gris C, et al. Experimental evolution of a plant pathogen into a legume symbiont. *PLoS Biology*. 2010;8:e1000280.
24. Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, Cherkis K, et al. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. *PLoS Pathogens*. 2011;7:e1002132.
25. Qian W, Jia Y, Ren S-X, He Y-Q, Feng J-X, Lu L-F, et al. Comparative and functional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. *campestris*. *Genome Research*. 2005;15:757–67.

26. Roux B, Bolot S, Guy E, Denancé N, Lautier M, Jardinaud M-F, et al. Genomics and transcriptomics of *Xanthomonas campestris* species challenge the concept of core type III effectome. *BMC Genomics*. 2015;16:975.
27. Du J, Reeves AZ, Klein JA, Twedt DJ, Knodler LA, Lesser CF. The type III secretion system apparatus determines the intracellular niche of bacterial pathogens. *Proceedings of the National Academy of Sciences*. 2016;113:4794–9.
28. Coombe L, Li JX, Lo T, Wong J, Nikolic V, Warren RL, et al. LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics*. 2021;22:534.
29. Rufián JS, Macho AP, Corry DS, Mansfield JW, Ruiz-Albert J, Arnold DL, et al. Confocal microscopy reveals in planta dynamic interactions between pathogenic, avirulent and non-pathogenic *Pseudomonas syringae* strains. *Molecular Plant Pathology*. 2018;19:537–51.
30. Rufián JS, Sánchez-Romero M-A, López-Márquez D, Macho AP, Mansfield JW, Arnold DL, et al. *Pseudomonas syringae* differentiates into phenotypically distinct subpopulations during colonization of a plant host. *Environmental Microbiology*. 2016;18:3593–605.
31. Walker J, Patel P, others. Splash dispersal and wind as factors in epidemiology of halo blight of bean. *Phytopathology*. 1964;54:140–1.
32. Taylor JD, Phelps K, Dudley CL. Epidemiology and strategy for the control of halo-blight of beans. *Annals of Applied Biology*. 1979;93:167–72.
33. Naughton LM, An S, Hwang I, Chou S-H, He Y-Q, Tang J-L, et al. Functional and genomic insights into the pathogenesis of *Burkholderia* species to rice. *Environmental Microbiology*. 2016;18:780–90.
34. van Hall C. De seringenziegte, veroorzaakt door *Pseudomonas syringae* nov. sp. Bijdragen tot Kennis der bakterieele Plantenziekten. 1902.
35. Loper JE, Lindow SE. Lack of evidence for the *in situ* fluorescent pigment production by *Pseudomonas syringae* pv. *syringae* on bean leaf surfaces. *Phytopathology*. 1987;77:1449–54.
36. Murillo J, Bardaji L, Navarro de la Fuente L, Elena Führer M, Aguilera S, Álvarez-Morales A. Variation in conservation of the cluster for biosynthesis of the phytotoxin phaseolotoxin in *Pseudomonas syringae* suggests at least two events of horizontal acquisition. *Research in Microbiology*. 2011;162:253–61.
37. Ruinelli M, Blom J, Pothier JF. Complete genome sequence of *Pseudomonas viridiflava* CFBP 1590, isolated from diseased cherry in france. *Genome Announcements*. 2017;5:e00662-17.
38. Morris CE, Sands DC, Vanneste JL, Montarry J, Oakley B, Guilbaud C, et al. Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in north america, europe, and new zealand. *mBio*. 2010;1:e00107-10.

39. Gutiérrez-Barranquero JA, Cazorla FM, de Vicente A. *Pseudomonas syringae* pv. *syringae* associated with mango trees, a particular pathogen within the “hodgepodge” of the *Pseudomonas syringae* complex. *Frontiers in Plant Science*. 2019;10.
40. Baltrus DA, Yourstone S, Lind A, Guilbaud C, Sands DC, Jones CD, et al. Draft genome sequences of a phylogenetically diverse suite of *Pseudomonas syringae* strains from multiple source populations. *Genome Announcements*. 2014;2.
41. Lefort F, Calmin G, Crovadore J, Osteras M, Farinelli L. Whole-genome shotgun sequence of *Pseudomonas viridiflava*, a bacterium species pathogenic to *Arabidopsis thaliana*. *Genome Announcements*. 2013;1:e00116-12.
42. Kniskern JM, Barrett LG, Bergelson J. Maladaptation in wild populations of the generalist plant pathogen *Pseudomonas syringae*. *Evolution*. 2011;65:818–30.
43. Dillon MM, Thakur S, Almeida RND, Wang PW, Weir BS, Guttman DS. Recombination of ecologically and evolutionarily significant loci maintains genetic cohesion in the *Pseudomonas syringae* species complex. *Genome Biology*. 2019;20:3.
44. Berge O, Monteil CL, Bartoli C, Chandeysson C, Guilbaud C, Sands DC, et al. A user’s guide to a data base of the diversity of *Pseudomonas syringae* and its application to classifying strains in this phylogenetic complex. *PLoS ONE*. 2014;9:e105547.
45. Psallidas PG. *Pseudomonas syringae* pv. *avellanae* pathovar nov., the bacterium causing canker disease on *Corylus avellana*. *Plant Pathology*. 1993;42:358–63.
46. Young JM. Taxonomy of *Pseudomonas Syringae*. *Journal of Plant Pathology*. 2010;10.
47. Sarkar SF, Guttman DS. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology*. 2004;70:1999–2012.
48. Gardan L, Shafik H, Belouin S, Broch R, Grimont F, Grimont PAD. DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremiae* sp. nov. and *Pseudomonas cannabina* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*. 1999;49:469–78.
49. Gomila M, Busquets A, Mulet M, García-Valdés E, Lalucat J. Clarification of taxonomic status within the *Pseudomonas syringae* species group based on a phylogenomic analysis. *Frontiers in Microbiology*. 2017;8.
50. Baltrus DA, Dougherty K, Beckstrom-Sternberg SM, Beckstrom-Sternberg JS, Foster JT. Incongruence between multi-locus sequence analysis (MLSA) and whole-genome-based phylogenies: *Pseudomonas syringae* pathovar *pisi* as a cautionary tale. *Molecular Plant Pathology*. 2014;15:461–5.
51. Scortichini M, Marcelletti S, Ferrante P, Firrao G. A genomic redefinition of *Pseudomonas avellanae* species. *PLoS ONE*. 2013;8:e75794.

52. Monteil CL, Yahara K, Studholme DJ, Mageiros L, Méric G, Swingle B, et al. Population-genomic insights into emergence, crop adaptation and dissemination of *Pseudomonas syringae* pathogens. *Microbial Genomics*. 2016;2.
53. Morris CE, Kinkel LL, Xiao K, Prior P, Sands DC. Surprising niche for the plant pathogen *Pseudomonas syringae*. *Infection, Genetics and Evolution*. 2007;7:84–92.
54. Samad A, Antonielli L, Sessitsch A, Compant S, Trognitz F. Comparative genome analysis of the vineyard weed endophyte *Pseudomonas viridiflava* CDRTc14 showing selective herbicidal activity. *Scientific Reports*. 2017;7:1–15.
55. Samad A, Trognitz F, Antonielli L, Compant S, Sessitsch A. High-quality draft genome sequence of an endophytic *Pseudomonas viridiflava* strain with herbicidal properties against its host, the weed *Lepidium draba* L. *Genome Announcements*. 2016;4:e01170-16.
56. Bartoli C, Berge O, Monteil CL, Guilbaud C, Balestra GM, Varvaro L, et al. The *Pseudomonas viridiflava* phylogroups in the *P. syringae* species complex are characterized by genetic variability and phenotypic plasticity of pathogenicity-related traits. *Environmental Microbiology*. 2014;16:2301–15.
57. Sarkar SF, Gordon JS, Martin GB, Guttman DS. Comparative genomics of host-specific virulence in *Pseudomonas syringae*. *Genetics*. 2006;174:1041–56.
58. Morris CE, Sands DC, Vinatzer BA, Glaux C, Guilbaud C, Buffière A, et al. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. *The ISME Journal*. 2008;2:321–34.
59. Adetuyi FC, Isogai A, Di Giorgio D, Ballio A, Takemoto JY. Saprophytic *Pseudomonas syringae* strain M1 of wheat produces cyclic lipodepsipeptides. *FEMS Microbiology Letters*. 1995;131:63–7.
60. Melotto M, Underwood W, Koczan J, Nomura K, He SY. Plant stomata function in innate immunity against bacterial invasion. *Cell*. 2006;126:969–80.
61. Frank AC, Saldierna Guzmán JP, Shay JE. Transmission of bacterial endophytes. *Microorganisms*. 2017;5:70.
62. Hardoim PR, Overbeek LS van, Berg G, Pirttilä AM, Compant S, Campisano A, et al. The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. *Microbiology and Molecular Biology Reviews*. 2015;79:293–320.
63. Yu X, Lund SP, Scott RA, Greenwald JW, Records AH, Nettleton D, et al. Transcriptional responses of *Pseudomonas syringae* to growth in epiphytic versus apoplastic leaf sites. *Proceedings of the National Academy of Sciences*. 2013;110:E425–34.
64. Rosier A, Bishnoi U, Lakshmanan V, Sherrier DJ, Bais HP. A perspective on inter-kingdom signaling in plant-beneficial microbe interactions. *Plant Molecular Biology*. 2016;90:537–48.

65. Vorholt JA. Microbial life in the phyllosphere. *Nature Reviews Microbiology*. 2012;10:828–40.
66. Lindow SE. Competitive exclusion of epiphytic bacteria by ice-*Pseudomonas syringae* mutants. *Applied and Environmental Microbiology*. 1987;53:2520–7.
67. Burch AY, Zeisler V, Yokota K, Schreiber L, Lindow SE. The hygroscopic biosurfactant syringafactin produced by *Pseudomonas syringae* enhances fitness on leaf surfaces during fluctuating humidity. *Environmental Microbiology*. 2014;16:2086–98.
68. Keith LMW, Bender CL. AlgT (ζ 22) controls alginate production and tolerance to environmental stress in *Pseudomonas syringae*. *Journal of Bacteriology*. 1999;181:7176–84.
69. Monier J-M, Lindow SE. Differential survival of solitary and aggregated bacterial cells promotes aggregate formation on leaf surfaces. *Proceedings of the National Academy of Sciences*. 2003;100:15977–82.
70. Heredia-Ponce Z, Gutiérrez-Barranquero JA, Putschert-Montenegro G, Eberl L, Cazorla FM, de Vicente A. Biological role of EPS from *Pseudomonas syringae* pv. *syringae* UMAF0158 extracellular matrix, focusing on a Psl-like polysaccharide. *NPJ Biofilms Microbiomes*. 2020;6:37.
71. Río-Álvarez I, Rodríguez-Herva JJ, Martínez PM, González-Melendi P, García-Casado G, Rodríguez-Palenzuela P, et al. Light regulates motility, attachment and virulence in the plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000. *Environmental Microbiology*. 2014;16:2072–85.
72. Müller C, Riederer M. Plant surface properties in chemical ecology. *Journal of Chemical Ecology*. 2005;31:2621–51.
73. Pfeilmeier S, Saur IM-L, Rathjen JP, Zipfel C, Malone JG. High levels of cyclic-di-GMP in plant-associated *Pseudomonas* correlate with evasion of plant immunity. *Molecular Plant Pathology*. 2016;17:521–31.
74. Aragón IM, Pérez-Mendoza D, Gallegos M-T, Ramos C. The c-di-GMP phosphodiesterase BifA is involved in the virulence of bacteria from the *Pseudomonas syringae* complex. *Molecular Plant Pathology*. 2015;16:604–15.
75. Jones JDG, Dangl JL. The plant immune system. *Nature*. 2006;444:323–9.
76. Underwood W, Melotto M, He SY. Role of plant stomata in bacterial invasion. *Cellular Microbiology*. 2007;9:1621–9.
77. Tans-Kersten J, Huang H, Allen C. *Ralstonia solanacearum* needs motility for invasive virulence on tomato. *Journal of Bacteriology*. 2001;183:3597–605.
78. Chen T, Nomura K, Wang X, Sohrabi R, Xu J, Yao L, et al. A plant genetic network for preventing dysbiosis in the phyllosphere. *Nature*. 2020;580:653–7.

79. Aung K, Jiang Y, He SY. The role of water in plant-microbe interactions. *The Plant journal*. 2018;93:771–80.
80. Xin X-F, Nomura K, Aung K, Velásquez AC, Yao J, Boutrot F, et al. Bacteria establish an aqueous living space in plants crucial for virulence. *Nature*. 2016;539:524–9.
81. Badel JL, Shimizu R, Oh H-S, Collmer A. A *Pseudomonas syringae* pv. *tomato* *avrE1/hopM1* mutant is severely reduced in growth and lesion formation in tomato. *Molecular Plant-Microbe Interactions*. 2006;19:99–111.
82. Roussin-Léveillé C, Lajeunesse G, St-Amand M, Veerapen VP, Silva-Martins G, Nomura K, et al. Evolutionarily conserved bacterial effectors hijack abscisic acid signaling to induce an aqueous environment in the apoplast. *Cell Host & Microbe*. 2022;30:489-501.e4.
83. Xin X-F, Kvitko B, He SY. *Pseudomonas syringae*: what it takes to be a pathogen. *Nature Reviews Microbiology*. 2018;16:316–28.
84. Ward JL, Forcat S, Beckmann M, Bennett M, Miller SJ, Baker JM, et al. The metabolic transition during disease following infection of *Arabidopsis thaliana* by *Pseudomonas syringae* pv. *tomato*. *The Plant Journal*. 2010;63:443–57.
85. Kubicek CP, Starr TL, Glass NL. Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. *Annual Review of Phytopathology*. 2014;52:427–51.
86. An S-Q, Febrer M, McCarthy Y, Tang D-J, Clissold L, Kaithakottil G, et al. High-resolution transcriptional analysis of the regulatory influence of cell-to-cell signalling reveals novel genes that contribute to *Xanthomonas* phytopathogenesis. *Molecular Microbiology*. 2013;88:1058–69.
87. Norman-Setterblad C, Vidal S, Palva ET. Interacting signal pathways control defense gene expression in *Arabidopsis* in response to cell wall-degrading enzymes from *Erwinia carotovora*. *Molecular Plant-Microbe Interactions*. 2000;13:430–8.
88. Deng W, Marshall NC, Rowland JL, McCoy JM, Worrall LJ, Santos AS, et al. Assembly, structure, function and regulation of type III secretion systems. *Nature Reviews Microbiology*. 2017;15:323–37.
89. Păcurar DI, Thordal-Christensen H, Păcurar ML, Pamfil D, Botez C, Bellini C. *Agrobacterium tumefaciens*: from crown gall tumors to genetic transformation. *Physiological and Molecular Plant Pathology*. 2011;76:76–81.
90. Block A, Li G, Fu ZQ, Alfano JR. Phytopathogen type III effector weaponry and their plant targets. *Current Opinion in Plant Biology*. 2008;11:396–403.
91. Kay S, Hahn S, Marois E, Hause G, Bonas U. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science*. 2007;318:648–51.

92. Ailloud F, Lowe T, Cellier G, Roche D, Allen C, Prior P. Comparative genomic analysis of *Ralstonia solanacearum* reveals candidate genes for host specificity. *BMC Genomics*. 2015;16:270.
93. Lam HN, Chakravarthy S, Wei H-L, BuiNguyen H, Stodghill PV, Collmer A, et al. Global analysis of the HrpL regulon in the plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000 reveals new regulon members with diverse functions. *PLoS ONE*. 2014;9:e106115.
94. Zhang J, Shao F, Li Y, Cui H, Chen L, Li H, et al. A *Pseudomonas syringae* effector inactivates MAPKs to suppress PAMP-induced immunity in plants. *Cell Host & Microbe*. 2007;1:175–85.
95. Xiang T, Zong N, Zou Y, Wu Y, Zhang J, Xing W, et al. *Pseudomonas syringae* effector AvrPto blocks innate immunity by targeting receptor kinases. *Current Biology*. 2008;18:74–80.
96. Lozano-Durán R, Bourdais G, He SY, Robatzek S. The bacterial effector HopM1 suppresses PAMP-triggered oxidative burst and stomatal immunity. *The New phytologist*. 2014;202:259–69.
97. Carter MQ, Chen J, Lory S. The *Pseudomonas aeruginosa* pathogenicity island PAPI-1 is transferred via a novel type IV pilus. *Journal of Bacteriology*. 2010;192:3249–58.
98. Navarro-Garcia F, Ruiz-Perez F, Cataldi Á, Larzábal M. Type VI secretion system in pathogenic *Escherichia coli*: structure, role in virulence, and acquisition. *Frontiers in Microbiology*. 2019;10.
99. Bender CL, Alarcón-Chaidez F, Gross DC. *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. *Microbiology and Molecular Biology Reviews*. 1999;63:266–92.
100. Zheng X, Spivey NW, Zeng W, Liu P-P, Fu ZQ, Klessig DF, et al. Coronatine promotes *Pseudomonas syringae* virulence in plants by activating a signaling cascade that inhibits salicylic acid accumulation. *Cell Host & Microbe*. 2012;11:587–96.
101. Jacobs JM, Babujee L, Meng F, Milling A, Allen C. The *In Planta* transcriptome of *Ralstonia solanacearum*: conserved physiological and virulence strategies during bacterial wilt of tomato. *mBio*. 2012;3:10.1128/mbio.00114-12.
102. Milling A, Babujee L, Allen C. *Ralstonia solanacearum* extracellular polysaccharide is a specific elicitor of defense responses in wilt-resistant tomato plants. *PLoS ONE*. 2011;6:e15853.
103. Aslam SN, Newman M-A, Erbs G, Morrissey KL, Chinchilla D, Boller T, et al. Bacterial polysaccharides suppress induced innate immunity by calcium chelation. *Current Biology*. 2008;18:1078–83.
104. Hill C. Virulence or niche factors: what's in a name? *Journal of Bacteriology*. 2012;194:5725–7.
105. Liu W, Yu Y-H, Cao S-Y, Niu X-N, Jiang W, Liu G-F, et al. Transcriptome profiling of *Xanthomonas campestris* pv. *campestris* grown in minimal medium MMX and rich medium NYG. *Research in Microbiology*. 2013;164:466–79.

106. Mitter B, Petric A, Shin MW, Chain PSG, Hauberg-Lotte L, Reinhold-Hurek B, et al. Comparative genome analysis of *Burkholderia phytofirmans* PsJN reveals a wide spectrum of endophytic lifestyles based on interaction strategies with host plants. *Frontiers in Plant Science*. 2013;4:120.
107. Lebeis SL, Paredes SH, Lundberg DS, Breakfield N, Gehring J, McDonald M, et al. Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science*. 2015;349:860–4.
108. Bai Y, Müller DB, Srinivas G, Garrido-Oter R, Potthoff E, Rott M, et al. Functional overlap of the *Arabidopsis* leaf and root microbiota. *Nature*. 2015;528:364–9.
109. Strieker M, Tanović A, Marahiel MA. Nonribosomal peptide synthetases: structures and dynamics. *Current Opinion in Structural Biology*. 2010;20:234–40.
110. Scholz-Schroeder BK, Hutchison ML, Grgurina I, Gross DC. The contribution of syringopeptin and syringomycin to virulence of *Pseudomonas syringae* pv. *syringae* strain B301D on the basis of *sypA* and *syrB1* biosynthesis mutant analysis. *Molecular Plant-Microbe Interactions*. 2001;14:336–48.
111. Joardar V, Lindeberg M, Jackson RW, Selengut J, Dodson R, Brinkac LM, et al. Whole-genome sequence analysis of *Pseudomonas syringae* pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition. *Journal of Bacteriology*. 2005;187:6488–98.
112. Ma L-S, Hachani A, Lin J-S, Filloux A, Lai E-M. *Agrobacterium tumefaciens* deploys a superfamily of type VI secretion DNase effectors as weapons for interbacterial competition in planta. *Cell Host & Microbe*. 2014;16:94–104.
113. Bernal P, Allsopp LP, Filloux A, Llamas MA. The *Pseudomonas putida* T6SS is a plant warden against phytopathogens. *The ISME Journal*. 2017;11:972–87.
114. Dong YH, Xu JL, Li XZ, Zhang LH. AiiA, an enzyme that inactivates the acylhomoserine lactone quorum-sensing signal and attenuates the virulence of *Erwinia carotovora*. *Proceedings of the National Academy of Sciences*. 2000;97:3526–31.
115. Caballo-Ponce E, Meng X, Uzelac G, Halliday N, Cámara M, Licastro D, et al. Quorum sensing in *Pseudomonas savastanoi* pv. *savastanoi* and *Erwinia toletana*: role in virulence and interspecies interactions in the olive knot. *Applied and Environmental Microbiology*. 2018;84:e00950-18.
116. Quiñones B, Dulla G, Lindow SE. Quorum sensing regulates exopolysaccharide production, motility, and virulence in *Pseudomonas syringae*. *Molecular Plant-Microbe Interactions*. 2005;18:682–93.
117. Adame-Álvarez R-M, Mendiola-Soto J, Heil M. Order of arrival shifts endophyte–pathogen interactions in bean from resistance induction to disease facilitation. *FEMS Microbiology Letters*. 2014;355:100–7.

118. Arnaud D, Hwang I. A sophisticated network of signaling pathways regulates stomatal defenses to bacterial pathogens. *Molecular Plant*. 2015;8:566–81.
119. Bodenhausen N, Bortfeld-Miller M, Ackermann M, Vorholt JA. A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. *PLoS genetics*. 2014;10:e1004283.
120. Potnis N, Krasileva K, Chow V, Almeida NF, Patil PB, Ryan RP, et al. Comparative genomics reveals diversity among xanthomonads infecting tomato and pepper. *BMC Genomics*. 2011;12:146.
121. Block A, Alfano JR. Plant targets for *Pseudomonas syringae* type III effectors: virulence targets or guarded decoys? *Current Opinion in Microbiology*. 2011;14:39–46.
122. Jacob P, Hige J, Dangl JL. Is localized acquired resistance the mechanism for effector-triggered disease resistance in plants? *Nature Plants*. 2023;9:1184–90.
123. Boller T, Felix G. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review of Plant Biology*. 2009;60:379–406.
124. Furukawa T, Inagaki H, Takai R, Hirai H, Che F-S. Two distinct EF-Tu epitopes induce immune responses in rice and *Arabidopsis*. *Molecular Plant Microbe Interact*. 2014;27:113–24.
125. Felix G, Duran JD, Volko S, Boller T. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *The Plant Journal*. 1999;18:265–76.
126. Walitang DI, Kim K, Madhaiyan M, Kim YK, Kang Y, Sa T. Characterizing endophytic competence and plant growth promotion of bacterial endophytes inhabiting the seed endosphere of Rice. *BMC Microbiology*. 2017;17:209.
127. Kunze G, Zipfel C, Robatzek S, Niehaus K, Boller T, Felix G. The N terminus of bacterial elongation factor Tu elicits innate immunity in *Arabidopsis* plants. *The Plant Cell*. 2004;16:3496–507.
128. Robatzek S, Bittel P, Chinchilla D, Köchner P, Felix G, Shiu S-H, et al. Molecular identification and characterization of the tomato flagellin receptor LeFLS2, an orthologue of *Arabidopsis* FLS2 exhibiting characteristically different perception specificities. *Plant Molecular Biology*. 2007;64:539–47.
129. Chinchilla D, Zipfel C, Robatzek S, Kemmerling B, Nürnberger T, Jones JDG, et al. A flagellin-induced complex of the receptor FLS2 and BAK1 initiates plant defence. *Nature*. 2007;448:497–500.
130. Lu D, Wu S, Gao X, Zhang Y, Shan L, He P. A receptor-like cytoplasmic kinase, BIK1, associates with a flagellin receptor complex to initiate plant innate immunity. *Proceedings of the National Academy of Sciences*. 2010;107:496–501.

131. Li L, Li M, Yu L, Zhou Z, Liang X, Liu Z, et al. The FLS2-associated kinase BIK1 directly phosphorylates the NADPH oxidase RbohD to control plant immunity. *Cell Host & Microbe*. 2014;15:329–38.
132. Bordiec S, Paquis S, Lacroix H, Dhondt S, Ait Barka E, Kauffmann S, et al. Comparative analysis of defence responses induced by the endophytic plant growth-promoting rhizobacterium *Burkholderia phytofirmans* strain PsJN and the non-host bacterium *Pseudomonas syringae* pv. *pisi* in grapevine cell suspensions. *Journal of Experimental Botany*. 2011;62:595–603.
133. Varnier A-L, Sanchez L, Vatsa P, Boudesocque L, Garcia-Brugger A, Rabenoelina F, et al. Bacterial rhamnolipids are novel MAMPs conferring resistance to *Botrytis cinerea* in grapevine. *Plant, Cell & Environment*. 2009;32:178–93.
134. Newman M-A, Sundelin T, Nielsen J, Erbs G. MAMP (microbe-associated molecular pattern) triggered immunity in plants. *Frontiers in Plant Science*. 2013;4:139.
135. Baker C, Mock N, Smith J, Aver'Yanov A. The dynamics of apoplast phenolics in tobacco leaves following inoculation with bacteria. *Frontiers in Plant Science*. 2015;6.
136. de Torres-Zabala M, Truman W, Bennett MH, Lafforgue G, Mansfield JW, Rodriguez Egea P, et al. *Pseudomonas syringae* pv. *tomato* hijacks the *Arabidopsis* abscisic acid signalling pathway to cause disease. *The EMBO Journal*. 2007;26:1434–43.
137. Niu D-D, Liu H-X, Jiang C-H, Wang Y-P, Wang Q-Y, Jin H-L, et al. The plant growth-promoting rhizobacterium *Bacillus cereus* AR156 induces systemic resistance in *Arabidopsis thaliana* by simultaneously activating salicylate- and jasmonate/ethylene-dependent signaling pathways. *Molecular Plant Microbe Interact*. 2011;24:533–42.
138. Cao H, Glazebrook J, Clarke JD, Volko S, Dong X. The *Arabidopsis* NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell*. 1997;88:57–63.
139. Li J, Brader G, Palva ET. The WRKY70 transcription factor: a node of convergence for jasmonate-mediated and salicylate-mediated signals in plant defense. *The Plant Cell*. 2004;16:319–31.
140. Xu X, Chen C, Fan B, Chen Z. Physical and functional interactions between pathogen-induced *Arabidopsis* WRKY18, WRKY40, and WRKY60 transcription factors. *The Plant Cell*. 2006;18:1310–26.
141. de Jonge R, Peter van Esse H, Kombrink A, Shinya T, Desaki Y, Bours R, et al. Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. *Science*. 2010;329:953–5.
142. Zhao Y, Thilmony R, Bender CL, Schaller A, He SY, Howe GA. Virulence systems of *Pseudomonas syringae* pv. *tomato* promote bacterial speck disease in tomato by targeting the jasmonate signaling pathway. *The Plant Journal*. 2003;36:485–99.

143. Brooks DM, Bender CL, Kunkel BN. The *Pseudomonas syringae* phytotoxin coronatine promotes virulence by overcoming salicylic acid-dependent defences in *Arabidopsis thaliana*. *Molecular Plant Pathology*. 2005;6:629–39.
144. Bu Q, Jiang H, Li C-B, Zhai Q, Zhang J, Wu X, et al. Role of the *Arabidopsis thaliana* NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Research*. 2008;18:756–67.
145. Abby SS, Rocha EPC. The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. *PLoS Genetics*. 2012;8:e1002983.
146. Jin Q, He SY. Role of the Hrp pilus in type III protein secretion in *Pseudomonas syringae*. *Science*. 2001;294:2556–8.
147. Büttner D. Protein export according to schedule: architecture, assembly, and regulation of type III secretion systems from plant- and animal-pathogenic bacteria. *Microbiology and Molecular Biology Reviews*. 2012;76:262–310.
148. Kvitko BH, Ramos AR, Morello JE, Oh H-S, Collmer A. Identification of harpins in *Pseudomonas syringae* pv. *tomato* DC3000, which are functionally similar to HrpK1 in promoting translocation of type III secretion system effectors. *Journal of Bacteriology*. 2007;189:8059–72.
149. Akeda Y, Galán JE. Chaperone release and unfolding of substrates in type III secretion. *Nature*. 2005;437:911–5.
150. Tampakaki AP. Commonalities and differences of T3SSs in rhizobia and plant pathogenic bacteria. *Frontiers in Plant Science*. 2014;5.
151. Alfano JR, Charkowski AO, Deng W-L, Badel JL, Petnicki-Ocwieja T, van Dijk K, et al. The *Pseudomonas syringae* Hrp pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. *Proceedings of the National Academy of Sciences*. 2000;97:4856–61.
152. Araki H, Innan H, Kreitman M, Bergelson J. Molecular evolution of pathogenicity-island genes in *Pseudomonas viridiflava*. *Genetics*. 2007;177:1031–41.
153. Gazi AD, Sarris PF, Fadoulglou VE, Charova SN, Mathioudakis N, Panopoulos NJ, et al. Phylogenetic analysis of a gene cluster encoding an additional, rhizobial-like type III secretion system that is narrowly distributed among *Pseudomonas syringae* strains. *BMC Microbiology*. 2012;12:188.
154. Araki H, Tian D, Goss EM, Jakob K, Halldorsdottir SS, Kreitman M, et al. Presence/absence polymorphism for alternative pathogenicity islands in *Pseudomonas viridiflava*, a pathogen of *Arabidopsis*. *Proceedings of the National Academy of Sciences*. 2006;103:5887–92.
155. Kajihara S, Hojo H, Koyanagi M, Tanaka M, Mizumoto H, Ohnishi K, et al. Implication of *hrpW* in virulence of *Pseudomonas cichorii*. *Plant Pathology*. 2012;61:355–63.

156. Deng W-L, Rehm AH, Charkowski AO, Rojas CM, Collmer A. *Pseudomonas syringae* exchangeable effector loci: sequence diversity in representative pathovars and virulence function in *P. syringae* pv. *syringae* B728a. *Journal of Bacteriology*. 2003;185:2592–602.
157. Charity JC, Pak K, Delwiche CF, Hutcheson SW. Novel exchangeable effector loci associated with the *Pseudomonas syringae* *hrp* pathogenicity island: evidence for integron-like assembly from transposed gene cassettes. *Molecular Plant-Microbe Interactions*. 2003;16:495–507.
158. Feil H, Feil WS, Chain P, Larimer F, DiBartolo G, Copeland A, et al. Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. *Proceedings of the National Academy of Sciences*. 2005;102:11064–9.
159. Inoue Y, Takikawa Y. The *hrpZ* and *hrpA* genes are variable, and useful for grouping *Pseudomonas syringae* bacteria. *Journal of General Plant Pathology*. 2006;72:26–33.
160. van der Heijden J, Finlay BB. Type III effector-mediated processes in *Salmonella* infection. *Future Microbiol*. 2012;7:685–703.
161. Ailloud F, Lowe TM, Robène I, Cruveiller S, Allen C, Prior P. *In planta* comparative transcriptomics of host-adapted strains of *Ralstonia solanacearum*. *PeerJ*. 2016;4:e1549.
162. Haapalainen M, van Gestel K, Pirhonen M, Taira S. Soluble plant cell signals induce the expression of the type III secretion system of *Pseudomonas syringae* and upregulate the production of pilus protein HrpA. *Molecular Plant-Microbe Interactions*. 2009;22:282–90.
163. Armentrout EI, Rietsch A. The type III secretion translocation pore senses host cell contact. *PLoS Pathogens*. 2016;12:e1005530.
164. Haapalainen M, Engelhardt S, Küfner I, Li C-M, Nürnberger T, Lee J, et al. Functional mapping of harpin HrpZ of *Pseudomonas syringae* reveals the sites responsible for protein oligomerization, lipid interactions and plant defence induction. *Molecular Plant Pathology*. 2011;12:151–66.
165. Marena M, Brito B, Callard D, Genin S, Barberis P, Boucher C, et al. PrhA controls a novel regulatory pathway required for the specific induction of *Ralstonia solanacearum* *hrp* genes in the presence of plant cells. *Molecular Microbiology*. 1998;27:437–53.
166. Hendrickson EL, Guevera P, Ausubel FM. The alternative sigma factor RpoN is required for *hrp* activity in *Pseudomonas syringae* pv. *maculicola* and acts at the level of *hrpL* transcription. *Journal of Bacteriology*. 2000;182:3508–16.
167. Lee JH, Sundin GW, Zhao Y. Identification of the *HrpS* binding site in the *hrpL* promoter and effect of the RpoN binding site of HrpS on the regulation of the type III secretion system in *Erwinia amylovora*. *Molecular Plant Pathology*. 2016;17:691–702.
168. Hutcheson SW, Bretz J, Sussan T, Jin S, Pak K. Enhancer-binding proteins HrpR and HrpS interact to regulate *hrp*-encoded type III protein secretion in *Pseudomonas syringae* strains. *Journal of Bacteriology*. 2001;183:5589–98.

169. Grimm C, Panopoulos NJ. The predicted protein product of a pathogenicity locus from *Pseudomonas syringae* pv. *phaseolicola* is homologous to a highly conserved domain of several procaryotic regulatory proteins. *Journal of Bacteriology*. 1989;171:5031–8.
170. Jovanovic M, James EH, Burrows PC, Rego FGM, Buck M, Schumacher J. Regulation of the co-evolved HrpR and HrpS AAA+ proteins required for *Pseudomonas syringae* pathogenicity. *Nature Communications*. 2011;2:177.
171. Xiao Y, Hutcheson SW. A single promoter sequence recognized by a newly identified alternate sigma factor directs expression of pathogenicity and host range determinants in *Pseudomonas syringae*. *Journal of Bacteriology*. 1994;176:3089–91.
172. Grimm C, Aufsatz W, Panopoulos NJ. The *hrpRS* locus of *Pseudomonas syringae* pv. *phaseolicola* constitutes a complex regulatory unit. *Molecular Microbiology*. 1995;15:155–65.
173. Wang J, Shao X, Zhang Y, Zhu Y, Yang P, Yuan J, et al. HrpS is a global regulator on type III secretion system (T3SS) and non-T3SS genes in *Pseudomonas savastanoi* pv. *phaseolicola*. *Molecular Plant-Microbe Interactions*. 2018;31:1232–43.
174. Wei W, Plovianich-Jones A, Deng W-L, Jin Q-L, Collmer A, Huang H-C, et al. The gene coding for the Hrp pilus structural protein is required for type III secretion of Hrp and Avr proteins in *Pseudomonas syringae* pv. *tomato*. *Proceedings of the National Academy of Sciences*. 2000;97:2247–52.
175. Ortiz-Martín I, Thwaites R, Macho AP, Mansfield JW, Beuzón CR. Positive regulation of the Hrp type III secretion system in *Pseudomonas syringae* pv. *phaseolicola*. *Molecular Plant-Microbe Interactions*. 2010;23:665–81.
176. Huang Y-C, Lin Y-C, Wei C-F, Deng W-L, Huang H-C. The pathogenicity factor HrpF interacts with HrpA and HrpG to modulate type III secretion system (T3SS) function and T3SS expression in *Pseudomonas syringae* pv. *averrhoi*. *Molecular Plant Pathology*. 2016;17:1080–94.
177. Wei C-F, Deng W-L, Huang H-C. A chaperone-like HrpG protein acts as a suppressor of HrpV in regulation of the *Pseudomonas syringae* pv. *syringae* type III secretion system. *Molecular Microbiology*. 2005;57:520–36.
178. Jovanovic M, Lawton E, Schumacher J, Buck M. Interplay among *Pseudomonas syringae* HrpR, HrpS and HrpV proteins for regulation of the type III secretion system. *FEMS Microbiol Letters*. 2014;356:201–11.
179. Ferreira AO, Myers CR, Gordon JS, Martin GB, Vencato M, Collmer A, et al. Whole-genome expression profiling defines the HrpL regulon of *Pseudomonas syringae* pv. *tomato* DC3000, allows *de novo* reconstruction of the Hrp *cis* Element, and identifies novel coregulated genes. *Molecular Plant-Microbe Interactions*. 2006;19:1167–79.
180. Waite C, Schumacher J, Jovanovic M, Bennett M, Buck M. Negative autogenous control of the master type III secretion system regulator HrpL in *Pseudomonas syringae*. *mBio*. 2017;8:e02273-16.

181. Charova SN, Gazi AD, Mylonas E, Pozidis C, Sabarit B, Anagnostou D, et al. Migration of type III secretion system transcriptional regulators links gene expression to secretion. *mBio*. 2018;9.
182. Fu ZQ, Guo M, Alfano JR. *Pseudomonas syringae* HrpJ is a type III secreted protein that is required for plant pathogenesis, injection of effectors, and secretion of the HrpZ1 harpin. *Journal of Bacteriology*. 2006;188:6060–9.
183. Crabill E, Karpisek A, Alfano JR. The *Pseudomonas syringae* HrpJ protein controls the secretion of type III translocator proteins and has a virulence role inside plant cells: The *Pseudomonas syringae* HrpJ control protein. *Molecular Microbiology*. 2012;85:225–38.
184. Rosebrock TR, Zeng L, Brady JJ, Abramovitch RB, Xiao F, Martin GB. A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. *Nature*. 2007;448:370–4.
185. Nicaise V, Joe A, Jeong B, Korneli C, Boutrot F, Westedt I, et al. *Pseudomonas* HopU1 modulates plant immune receptor levels by blocking the interaction of their mRNAs with GRP7. *The EMBO Journal*. 2013;32:701–12.
186. Tena G. PTI and ETI are one. *Nature Plants*. 2021;7:1527–1527.
187. Séjalon-Delmas N, Mateos FV, Bottin A, Rickauer M, Dargent R, Esquerré-Tugayé MT. Purification, elicitor activity, and cell wall localization of a glycoprotein from *Phytophthora parasitica* var. *nicotianae*, a fungal pathogen of tobacco. *Phytopathology*. 1997;87:899–909.
188. Mackey D, Holt BF, Wiig A, Dangl JL. RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. *Cell*. 2002;108:743–54.
189. van der Hoorn RAL, Kamoun S. From guard to decoy: a new model for perception of plant pathogen effectors. *The Plant Cell*. 2008;20:2009–17.
190. Axtell MJ, Staskawicz BJ. Initiation of RPS2-specified disease resistance in *Arabidopsis* is coupled to the AvrRpt2-directed elimination of RIN4. *Cell*. 2003;112:369–77.
191. Kunkel BN, Bent AF, Dahlbeck D, Innes RW, Staskawicz BJ. RPS2, an *Arabidopsis* disease resistance locus specifying recognition of *Pseudomonas syringae* strains expressing the avirulence gene *avrRpt2*. *The Plant Cell*. 1993;5:865–75.
192. Römer P, Hahn S, Jordan T, Strauss T, Bonas U, Lahaye T. Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. *Science*. 2007;318:645–8.
193. Sokurenko EV, Hasty DL, Dykhuizen DE. Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends in Microbiology*. 1999;7:191–5.
194. Wang L, Tsuda K, Sato M, Cohen JD, Katagiri F, Glazebrook J. *Arabidopsis* CaM binding protein CBP60g contributes to MAMP-induced SA accumulation and is involved in disease resistance against *Pseudomonas syringae*. *PLoS Pathogens*. 2009;5:e1000301.

195. Hofius D, Schultz-Larsen T, Joensen J, Tsitsigiannis DI, Petersen NHT, Mattsson O, et al. Autophagic components contribute to hypersensitive cell death in *Arabidopsis*. *Cell*. 2009;137:773–83.
196. Conrath U. Systemic acquired resistance. *Plant signaling & behavior*. 2006;1:179–84.
197. Tao Y, Xie Z, Chen W, Glazebrook J, Chang H-S, Han B, et al. Quantitative nature of *Arabidopsis* responses during compatible and incompatible interactions with the bacterial pathogen *Pseudomonas syringae*. *The Plant Cell*. 2003;15:317–30.
198. Kim NH, Choi HW, Hwang BK. *Xanthomonas campestris* pv. *vesicatoria* effector AvrBsT induces cell death in pepper, but suppresses defense responses in tomato. *Molecular Plant Microbe Interact*. 2010;23:1069–82.
199. Schuster SC. Next-generation sequencing transforms today's biology. *Nature Methods*. 2008;5:16–8.
200. Buell CR, Joardar V, Lindeberg M, Selengut J, Paulsen IT, Gwinn ML, et al. The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. *Proceedings of the National Academy of Sciences*. 2003;100:10181–6.
201. Eckert KA, Kunkel TA. DNA polymerase fidelity and the polymerase chain reaction. *Genome Research*. 1991;1:17–24.
202. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*. 2012;30:434–9.
203. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*. 2015;15:141–61.
204. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nature Methods*. 2011;8:61–5.
205. Smits THM. The importance of genome sequence quality to microbial comparative genomics. *BMC Genomics*. 2019;20:662.
206. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31:3691–3.
207. Konstantinidis KT, Rosselló-Móra R. Classifying the uncultivated microbial majority: a place for metagenomic data in the *Candidatus* proposal. *Systematic and Applied Microbiology*. 2015;38:223–30.
208. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. *Genome Biology*. 2013;14:405.

209. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA Sequencing from single polymerase molecules. *Science*. 2009;323:133–8.
210. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in Genomic Sciences*. 2015;10:18.
211. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*. 2019;37:1155–62.
212. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*. 2013;10:563–9.
213. Pearman WS, Freed NE, Silander OK. Testing the advantages and disadvantages of short- and long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*. 2020;21:220.
214. PacBio Revio: Long-read sequencing at scale. Available from: <https://www.pacb.com/revio/>
215. Birla B. PacBio vs. Oxford nanopore sequencing. *Genohub Blog*. 2017. Available from: <https://blog.genohub.com/2017/06/16/pacbio-vs-oxford-nanopore-sequencing/>
216. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*. 2016;17:239.
217. Wang H, Branton D. Nanopores with a spark for single-molecule detection. *Nature Biotechnology*. 2001;19:622–3.
218. Howorka S, Cheley S, Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores. *Nature Biotechnology*. 2001;19:636–9.
219. Garaj S, Hubbard W, Reina A, Kong J, Branton D, Golovchenko JA. Graphene as a subnanometre trans-electrode membrane. *Nature*. 2010;467:190–3.
220. Carter J-M, Hussain S. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. *Wellcome Open Research*. 2018;2:23.
221. Oxford Nanopore Technologies: Products. Available from: <https://nanoporetech.com/products>
222. Illumina: Sequencing and array-based solutions for genetic research. Available from: <https://www.illumina.com/>
223. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*. 2017;3:e000132.
224. Nurk S, Koren S, Rhie A, Rautiainen M, Bizkadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022;376:44–53.

225. Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford nanopore sequencing. *BMC Genomics*. 2020;21:631.
226. Xu Y, Lewandowski K, Lumley S, Pullan S, Vipond R, Carroll M, et al. Detection of viral pathogens with multiplex nanopore MinION sequencing: be careful with cross-talk. *Frontiers in Microbiology*. 2018;9.
227. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology*. 2019;37:783–92.
228. Taxt AM, Avershina E, Frye SA, Naseer U, Ahmad R. Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Scientific Reports*. 2020;10:7622.
229. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods*. 2015;12:733–5.
230. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*. 2019;35:2193–8.
231. Quick J. Ultra-long read sequencing protocol for RAD004. 2018. Available from: <https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-mrxc57n>
232. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*. 2019;20:129.
233. Boža V, Brejová B, Vinař T. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS ONE*. 2017;12:e0178751.
234. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*. 2015;33:296–300.
235. Seki M, Katsumata E, Suzuki A, Sereewattanawoot S, Sakamoto Y, Mizushima-Sugano J, et al. Evaluation and application of RNA-Seq by MinION. *DNA Research*. 2019;26:55–65.
236. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*. 2018;19:90.
237. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*. 2020;21:30.
238. Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology*. 2019;37:124–6.

239. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*. 2020;17:155–8.
240. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Research*. 2017;27:722–36.
241. Huang Y-T, Liu P-Y, Shih P-W. Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing. *Genome Biology*. 2021;22:95.
242. Jain M, Tyson JR, Loose M, Ip CLC, Eccles DA, O’Grady J, et al. MinION analysis and reference consortium: phase 2 data release and analysis of R9.0 chemistry. *F1000Research*. 2017;6:760.
243. Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*. 2018;7:giy037.
244. Oxford Nanopore Technologies. GitHub. Available from: <https://github.com/nanoporetech>
245. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*. 2020;17:1103–10.
246. Cunnac S, Chakravarthy S, Kvitko BH, Russell AB, Martin GB, Collmer A. Genetic disassembly and combinatorial reassembly identify a minimal functional repertoire of type III effectors in *Pseudomonas syringae*. *Proceedings of the National Academy of Sciences*. 2011;108:2975–80.
247. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.
248. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Research*. 2017;27:737–46.
249. Oxford Nanopore Technologies. Medaka. Available from: <https://github.com/nanoporetech/medaka>
250. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. *Nature Biotechnology*. 2020;1–10.
251. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*. 2017;13:e1005595.
252. Luo J, Lyu M, Chen R, Zhang X, Luo H, Yan C. SLR: a scaffolding algorithm based on long reads and contig classification. *BMC Bioinformatics*. 2019;20:539.

253. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*. 2014;9:e112963.
254. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*. 2020;36:2253–5.
255. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*. 2018;36:338–45.
256. Lindow SE. Ecology of *Pseudomonas syringae* relevant to the field use of ice-deletion mutants constructed in vitro for plant frost control. *Engineered Organisms in the Environment: Scientific Issues*. Washington, D.C.: American Society for Microbiology; 1985. p. 23–35.
257. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456:53–9.
258. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*. 2016;9:88.
259. Rhoads A, Au KF. Pacbio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*. 2015;13:278–89.
260. Diabate M, Munro P, Garcia E, Jacquel A, Michel G, Obba S, et al. *Escherichia coli* α -hemolysin counteracts the anti-virulence innate immune response triggered by the Rho GTPase activating toxin CNF1 during bacteremia. *PLoS Pathogens*. 2015;11:e1004732.
261. Schneider GF, Xu Q, Hage S, Luik S, Spoor JNH, Malladi S, et al. Tailoring the hydrophobicity of graphene for its use as nanopores for DNA translocation. *Nature Communications*. 2013;4:2619.
262. Kubota T, Lloyd K, Sakashita N, Minato S, Ishida K, Mitsui T. Clog and release, and reverse motions of DNA in a nanopore. *Polymers*. 2019;11:84.
263. Oxford Nanopore Technologies: MinION. Available from: <https://nanoporetech.com/products/minion>
264. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9:357–9.
265. Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics*. 2016;32:2103–10.
266. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210–2.

267. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*. 2004;14:1394–403.
268. Sernova NV, Gelfand MS. Identification of replication origins in prokaryotic genomes. *Briefings in Bioinformatics*. 2008;9:376–91.
269. Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*. 2014;15:121–32.
270. Kastanis GJ, Santana-Quintero LV, Sanchez-Leon M, Lomonaco S, Brown EW, Allard MW. In-depth comparative analysis of Illumina MiSeq run metrics: development of a wet-lab quality assessment tool. *Molecular Ecology Resources*. 2019;19:377–87.
271. Liu H, Wu S, Li A, Ruan J, Wu S, Li A, et al. SMARTdenovo: a *de novo* assembler using long noisy reads. *Gigabyte*. 2021;2021:1–9.
272. PacBio. pbccoda: PacBio secondary analysis tools on bioconda. Available from: <https://github.com/PacificBiosciences/pbccoda>
273. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*. 2019;37:540–6.
274. Tischler G, Myers EW. Non hybrid long read consensus using local de bruijn graph assembly. *bioRxiv*; 2017;106252.
275. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068–9.
276. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*. 2013;41:D358–65.
277. TCAG DNA / Sequencing Facility. Available from: <https://www.tcag.ca/facilities/dnaSequencingSynthesis.html>
278. University of Minnesota Genomics Center. PacBio Sequel II Sequencing. Available from: <https://genomics.umn.edu/service/pacbio-sequel-sequencing>
279. Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*. 2005;438:1105–15.
280. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*. 2003;422:859–68.
281. The *C. elegans* Sequencing Consortium. Genome sequence of the *Nematode C. elegans*: a platform for investigating biology. *Science*. 1998;282:2012–8.

282. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1.2-megabase genome sequence of Mimivirus. *Science*. 2004;306:1344–50.
283. Zeng J, Cai H, Peng H, Wang H, Zhang Y, Akutsu T. Causalcall: nanopore basecalling using a temporal convolutional network. *Frontiers in Genetics*. 2020;10:1332.
284. Avershina E, Frye SA, Ali J, Taxt AM, Ahmad R. Ultrafast and cost-effective pathogen identification and resistance gene detection in a clinical setting using nanopore flongle sequencing. *Frontiers in Microbiology*. 2022;13:822402.
285. Wick RR, Judd LM, Wyres KL, Holt KE. Recovery of small plasmid sequences via Oxford nanopore sequencing. *Microbial genomics*. 2021;7:000631.
286. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 2016;530:228–32.
287. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. 2020;585:79–84.
288. Koren S, Schatz MC, Walenz BP, Martin J, Howard J, Ganapathy G, et al. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotechnology*. 2012;30:693–700.
289. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods*. 2021;1–5.
290. Pacific Bioscience. Genomic Consensus. Available from: <https://github.com/PacificBiosciences/GenomicConsensus>
291. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics*. 2019;5:e000294.
292. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*. 2016;32:1009–15.
293. Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*. 2015;12:780–6.
294. Harris RS. Improved Pairwise Alignment of Genomic DNA. Pennsylvania State University; 2007. Available from: http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf
295. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*. 2016;44:6614–24.

296. Lee I, Ouk Kim Y, Park S-C, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*. 2016;66:1100–3.
297. Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*. 2005;33:3390–400.
298. Bork P, Bairoch A. Go hunting in sequence databases but watch out for the traps. *Trends in Genetics*. 1996;12:425–7.
299. Lovell HC, Mansfield JW, Godfrey SAC, Jackson RW, Hancock JT, Arnold DL. Bacterial evolution by genomic island transfer occurs via DNA transformation *In Planta*. *Current Biology*. 2009;19:1586–90.
300. Marinier E, Brown DG, McConkey BJ. Pollux: platform independent error correction of single and mixed genomes. *BMC Bioinformatics*. 2015;16:10.
301. Tsai C-H, Liao R, Chou B, Palumbo M, Contreras LM. Genome-wide analyses in bacteria show small-rna enrichment for long and conserved intergenic regions. *Journal of Bacteriology*. 2015;197:40–50.
302. Khademi SMH, Sazinas P, Jelsbak L. Within-host adaptation mediated by intergenic evolution in *Pseudomonas aeruginosa*. *Genome Biology and Evolution*. 2019;11:1385–97.
303. Thorpe HA, Bayliss SC, Hurst LD, Feil EJ. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*. 2017;206:363–76.
304. Pernitzsch SR, Tirier SM, Beier D, Sharma CM. A variable homopolymeric G-repeat defines small RNA-mediated posttranscriptional regulation of a chemotaxis receptor in *Helicobacter pylori*. *Proceedings of the National Academy of Sciences*. 2014;111:E501–10.
305. Arnold DL, Jackson RW, Vivian A. Evidence for the mobility of an avirulence gene, *avrPpiA1*, between the chromosome and plasmids of races of *Pseudomonas syringae* pv. *pisi*. *Molecular Plant Pathology*. 2000;1:195–9.
306. Hulin MT, Armitage AD, Vicente JG, Holub EB, Baxter L, Bates HJ, et al. Comparative genomics of *Pseudomonas syringae* reveals convergent gene gain and loss associated with specialization onto cherry (*Prunus avium*). *New Phytologist*. 2018;219:672–96.
307. Marcelletti S, Scortichini M. Definition of plant-pathogenic *Pseudomonas* genomospecies of the *Pseudomonas syringae* complex through multiple comparative approaches. *Phytopathology*. 2014;104:1274–82.
308. Daughton CG. Wastewater surveillance for population-wide Covid-19: the present and future. *Science of The Total Environment*. 2020;736:139631.

309. Lorv JSH, McConkey BJ. Complete genome sequences for *Pseudomonas syringae* strains 508 and TLP2.
310. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, et al. The genomic standards consortium. PLoS Biology. 2011;9:e1001088.
311. Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microbial genomics. 2020;6:mgen000409.
312. Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. Proceedings of the National Academy of Sciences. 1986;83:6233–7.
313. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nature Communications. 2016;7:ncomms11307.
314. Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. BMC Biology. 2006;4:41.
315. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. PLoS One. 2021;16:e0257521.
316. Ham JH, Majerczak DR, Arroyo-Rodriguez AS, Mackey DM, Coplin DL. WtsE, an AvrE-family effector protein from *Pantoea stewartii* subsp. *stewartii*, causes disease-associated cell death in corn and requires a chaperone protein for stability. Molecular Plant-Microbe Interactions. 2006;19:1092–102.
317. Craigen WJ, Cook RG, Tate WP, Caskey CT. Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. Proceedings of the National Academy of Sciences. 1985;82:3616–20.
318. Farabaugh PJ. Programmed translational frameshifting. Annual Review of Genetics. 1996;30:507–28.
319. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. Applied and Environmental Microbiology. 2000;66:1328–33.
320. Noinaj N, Guillier M, Barnard TJ, Buchanan SK. TonB-dependent transporters: regulation, structure, and function. Annu Rev Microbiol. 2010;64:43–60.
321. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. Gigascience. 2019;8:1–9.
322. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics. 2018;34:1037–9.

323. Jain M. High-coverage long read DNA sequencing with the Oxford nanopore MinION. UC Santa Cruz; 2017.
324. Butler MI, Stockwell PA, Black MA, Day RC, Lamont IL, Poulter RTM. *Pseudomonas syringae* pv. *actinidiae* from recent outbreaks of kiwifruit bacterial canker belong to different clones that originated in china. PLoS ONE. 2013;8:e57464.
325. Dillon MM, Almeida RND, Laflamme B, Martel A, Weir BS, Desveaux D, et al. Molecular evolution of *Pseudomonas syringae* type III secreted effector proteins. Frontiers in Plant Science. 2019;10.
326. Alfano JR, Collmer A. Type III secretion system effector proteins: double agents in bacterial disease and plant defense. Annual Review of Phytopathology. 2004;42:385–414.
327. Hueck CJ. Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiology and Molecular Biology Reviews. 1998;62:379–433.
328. Haapalainen M, Dauphin A, Li C-M, Bailly G, Tran D, Briand J, et al. HrpZ harpins from different *Pseudomonas syringae* pathovars differ in molecular interactions and in induction of anion channel responses in *Arabidopsis thaliana* suspension cells. Plant Physiology and Biochemistry. 2012;51:168–74.
329. Taira S, Tuimala J, Roine E, Nurmiaho-Lassila E-L, Savilahti H, Romantschuk M. Mutational analysis of the *Pseudomonas syringae* pv. *tomato* *hrpA* gene encoding Hrp pilus subunit. Molecular Microbiology. 1999;34:737–44.
330. O'Malley MR, Anderson JC. Regulation of the *Pseudomonas syringae* type III secretion system by host environment signals. Microorganisms. 2021;9:1227.
331. Janjusevic R, Abramovitch RB, Martin GB, Stebbins CE. A bacterial inhibitor of host programmed cell death defenses is an E3 ubiquitin ligase. Science. 2006;311:222–6.
332. Tsiamis G, Mansfield JW, Hockenhull R, Jackson RW, Sesma A, Athanassopoulos E, et al. Cultivar-specific avirulence and virulence functions assigned to *avrPphF* in *Pseudomonas syringae* pv. *phaseolicola*, the cause of bean halo-blight disease. The EMBO Journal. 2000;19:3204–14.
333. Lu X, Kracher B, Saur IML, Bauer S, Ellwood SR, Wise R, et al. Allelic barley MLA immune receptors recognize sequence-unrelated avirulence effectors of the powdery mildew pathogen. Proceedings of the National Academy of Sciences. 2016;113:E6486–95.
334. Hartmann FE, Croll D. Distinct trajectories of massive recent gene gains and losses in populations of a microbial eukaryotic pathogen. Molecular Biology and Evolution. 2017;34:2808–22.
335. Landgraf A, Weingart H, Tsiamis G, Boch J. Different versions of *Pseudomonas syringae* pv. *tomato* DC3000 exist due to the activity of an effector transposon. Molecular Plant Pathology. 2006;7:355–64.

336. Baltrus DA, McCann HC, Guttman DS. Evolution, genomics and epidemiology of *Pseudomonas syringae*. *Molecular Plant Pathology*. 2017;18:152–68.
337. Gomila M, Peña A, Mulet M, Lalucat J, García-Valdés E. Phylogenomics and systematics in *Pseudomonas*. *Frontiers in Microbiology*. 2015;6:214.
338. Yan S, Liu H, Mohr TJ, Jenrette J, Chiodini R, Zaccardelli M, et al. Role of recombination in the evolution of the model plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000, a very atypical tomato strain. *Applied and Environmental Microbiology*. 2008;74:3171–81.
339. Majewski J, Cohan FM. The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics*. 1998;148:13–8.
340. Mucyn TS, Yourstone S, Lind AL, Biswas S, Nishimura MT, Baltrus DA, et al. Variable suites of non-effector genes are co-regulated in the type III secretion virulence regulon across the *Pseudomonas syringae* phylogeny. *PLoS Pathogens*. 2014;10:e1003807.
341. Lindeberg M, Cunnac S, Collmer A. *Pseudomonas syringae* type III effector repertoires: last words in endless arguments. *Trends in Microbiology*. 2012;20:199–208.
342. Aguilera S, López-López K, Nieto Y, Garcidueñas-Piña R, Hernández-Guzmán G, Hernández-Flores JL, et al. Functional characterization of the gene cluster from *Pseudomonas syringae* pv. *phaseolicola* NPS3121 involved in synthesis of phaseolotoxin. *Journal of Bacteriology*. 2007;189:2834–43.
343. Hu Y, Ding Y, Cai B, Qin X, Wu J, Yuan M, et al. Bacterial effectors manipulate plant abscisic acid signaling for creation of an aqueous apoplast. *Cell Host & Microbe*. 2022;30:518-529.e6.
344. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 2022;50:D20–6.
345. European Bioinformatics Institute. InterProScan 5: genome-scale protein function classification. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998142/>
346. SQLite. Available from: <https://www.sqlite.org/index.html>
347. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*. 2013;30:772–80.
348. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 2010;5:e9490.
349. Pedersen TL. Hierarchical sets: analyzing pangenome structure through scalable set visualizations. *Bioinformatics*. 2017;33:1604–12.

350. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution*. 2020;37:291–4.
351. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. 2019;35:4453–5.
352. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. *Nucleic Acids Research*. 2015;43:W39–49.
353. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. *Journal of Open Source Software*. 2019;4:1686.
354. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35:526–8.
355. Yu G. Data Integration, Manipulation and visualization of phylogenetic trees. New York: Chapman and Hall/CRC; 2022.
356. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. 2017;8:28–36.
357. Wilkinson SP, Davy SK. phylogram: an R package for phylogenetic analysis with nested lists. *Journal of Open Source Software*. 2018;3:790.
358. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*. 2015;31:3718–20.
359. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots. 2023. Available from: <https://cran.r-project.org/web/packages/ggpubr/index.html>
360. Auguie B. egg: Extensions for “ggplot2”. Available from: <https://cran.r-project.org/web/packages/egg/index.html>
361. Baums J. hues. 2023. Available from: <https://github.com/johnbaums/hues>
362. Blakney AJC, Patten CL. A plant growth-promoting pseudomonad is closely related to the *Pseudomonas syringae* complex of plant pathogens. *FEMS Microbiology Ecology*. 2011;77:546–57.
363. Lelliott RA, Billing E, Hayward AC. A determinative scheme for the fluorescent plant pathogenic pseudomonads. *Journal of Applied Bacteriology*. 1966;29:470–89.
364. Gropp SJ, Guttman DS. The PCR amplification and characterization of entire *Pseudomonas syringae* hrp/hrc clusters. *Molecular Plant Pathology*. 2004;5:137–40.

365. Liu Y, Bauer SC, Imlay JA. The YaaA protein of the *Escherichia coli* OxyR regulon lessens hydrogen peroxide toxicity by diminishing the amount of intracellular unincorporated iron. *Journal of Bacteriology*. 2011;193:2186–96.
366. Mehr IJ, Long CD, Serkin CD, Seifert HS. A Homologue of the Recombination-Dependent Growth Gene, *rdgC*, Is Involved in Gonococcal Pilin Antigenic Variation. *Genetics*. 2000;154:523–32.
367. Romantschuk M, Richter GY, Mukhopadhyay P, Mills D. IS801, an insertion sequence element isolated from *Pseudomonas syringae* pathovar *phaseolicola*. *Molecular Microbiology*. 1991;5:617–22.
368. Elliott SJ, Hutcheson SW, Dubois MS, Mellies JL, Wainwright LA, Batchelor M, et al. Identification of CesT, a chaperone for the type III secretion of Tir in enteropathogenic *Escherichia coli*. *Molecular Microbiology*. 1999;33:1176–89.
369. Oh H-S, Kvitko BH, Morello JE, Collmer A. *Pseudomonas syringae* lytic transglycosylases coregulated with the type III secretion system contribute to the translocation of effector proteins into plant cells. *Journal of Bacteriology*. 2007;189:8277–89.
370. Deng W-L, Preston G, Collmer A, Chang C-J, Huang H-C. Characterization of the *hrpC* and *hrpRS* operons of *Pseudomonas syringae* pathovars *syringae*, *tomato*, and *glycinea* and analysis of the ability of *hrpF*, *hrpG*, *hrcC*, *hrpT*, and *hrpV* mutants to elicit the hypersensitive response and disease in plants. *Journal of Bacteriology*. 1998;180:4523–31.
371. Morello JE, Collmer A. *Pseudomonas syringae* HrpP is a type III secretion substrate specificity switch domain protein that is translocated into plant cells but functions atypically for a substrate-switching protein. *Journal of Bacteriology*. 2009;191:3120–31.
372. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proceedings of the National Academy of Sciences*. 2000;97:6652–7.
373. Castillo-Lizardo MG, Aragón IM, Carvajal V, Matas IM, Pérez-Bueno ML, Gallegos M-T, et al. Contribution of the non-effector members of the HrpL regulon, *iaaL* and *matE*, to the virulence of *Pseudomonas syringae* pv. *tomato* DC3000 in tomato plants. *BMC Microbiology*. 2015;15:165.
374. Munkvold KR, Russell AB, Kvitko BH, Collmer A. *Pseudomonas syringae* pv. *tomato* DC3000 type III effector HopAA1-1 functions redundantly with chlorosis-promoting factor PSPTO4723 to produce bacterial speck lesions in host tomato. *Molecular Plant-Microbe Interactions*. 2009;22:1341–55.
375. Marín M, Uversky VN, Ott T. Intrinsic disorder in pathogen effectors: protein flexibility as an evolutionary hallmark in a molecular arms race. *The Plant Cell*. 2013;25:3153–7.
376. Dong X, Lu X, Zhang Z. BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database*. 2015;2015.

377. Carrión VJ, Gutiérrez-Barranquero JA, Arrebola E, Bardaji L, Codina JC, Vicente A de, et al. The mangotoxin biosynthetic operon (*mbo*) is specifically distributed within *Pseudomonas syringae* Genomospecies 1 and was acquired only once during evolution. *Applied and Environmental Microbiology*. 2013;79:756–67.
378. Kinscherf TG, Willis DK. The biosynthetic gene cluster for the beta-lactam antibiotic tabtoxin in *Pseudomonas syringae*. *The Journal of Antibiotics*. 2005;58:817–21.
379. Keith RC, Keith LMW, Hernández-Guzmán G, Uppalapati SR, Bender CL. Alginate gene expression by *Pseudomonas syringae* pv. *tomato* DC3000 in host and non-host plants. *Microbiology*. 2003;149:1127–38.
380. Wang H, Yang Z, Swingle B, Kvitko BH. AlgU, a conserved sigma factor regulating abiotic stress tolerance and promoting virulence in *Pseudomonas syringae*. *Molecular Plant-Microbe Interactions*. 2021;34:326–36.
381. Lotan T, Fluhr R. Xylanase, a novel elicitor of pathogenesis-related proteins in tobacco, uses a non-ethylene pathway for induction 1. *Plant Physiology*. 1990;93:811–7.
382. Bhardwaj N, Kumar B, Verma P. A detailed overview of xylanases: an emerging biomolecule for current and future prospective. *Bioresources and Bioprocessing*. 2019;6:40.
383. Jakob K, Kniskern JM, Bergelson J. The role of pectate lyase and the jasmonic acid defense response in *Pseudomonas viridiflava* virulence. *Molecular Plant-Microbe Interactions*. 2007;20:146–58.
384. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
385. Rabin SDP, Hauser AR. *Pseudomonas aeruginosa* ExoU, a toxin transported by the type III secretion system, kills *Saccharomyces cerevisiae*. *Infection and Immunity*. 2003;71:4144–50.
386. Gendrin C, Contreras-Martel C, Bouillot S, Elsen S, Lemaire D, Skoufias DA, et al. Structural basis of cytotoxicity mediated by the type III secretion toxin ExoU from *Pseudomonas aeruginosa*. *PLoS Pathogens*. 2012;8:e1002637.
387. Block A, Toruño TY, Elowsky CG, Zhang C, Steinbrenner J, Beynon J, et al. The *Pseudomonas syringae* type III effector HopD1 suppresses effector-triggered immunity, localizes to the endoplasmic reticulum, and targets the *Arabidopsis transcription* factor NTL9. *New Phytologist*. 2014;201:1358–70.
388. English G, Trunk K, Rao VA, Srikannathasan V, Hunter WN, Coulthurst SJ. New secreted toxins and immunity proteins encoded within the Type VI secretion system gene cluster of *Serratia marcescens*. *Molecular Microbiology*. 2012;86:921–36.
389. Feil H, Feil WS, Lindow SE. Contribution of fimbrial and afimbrial adhesins of *Xylella fastidiosa* to attachment to surfaces and virulence to grape. *Phytopathology*. 2007;97:318–24.

390. Koskiniemi S, Lamoureux JG, Nikolakakis KC, Roodenbeke C t’Kint de, Kaplan MD, Low DA, et al. Rhs proteins from diverse bacteria mediate intercellular competition. *Proceedings of the National Academy of Sciences*. 2013;110:7032–7.
391. Hernandez-Doria JD, Sperandio V. Bacteriophage transcription factor Cro regulates virulence gene expression in enterohemorrhagic *Escherichia coli*. *Cell Host & Microbe*. 2018;23:607-617.e6.
392. Schröder G, Krause S, Zechner EL, Traxler B, Yeo H-J, Lurz R, et al. TraG-like proteins of dna transfer systems and of the *Helicobacter pylori* type IV secretion system: inner membrane gate for exported substrates? *Journal of Bacteriology*. 2002;184:2767–79.
393. Kwan DH, Leadlay PF. Mutagenesis of a modular polyketide synthase enoylreductase domain reveals insights into catalysis and stereospecificity. *ACS Chemical Biology*. 2010;5:829–38.
394. Minina EA, Coll NS, Tuominen H, Bozhkov PV. Metacaspases versus caspases in development and cell fate regulation. *Cell Death & Differentiation*. 2017;24:1314–25.
395. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. *Trends Microbiol*. 2010;18:315–22.
396. Ruinelli M, Blom J, Smits THM, Pothier JF. Comparative genomics and pathogenicity potential of members of the *Pseudomonas syringae* species complex on *Prunus* spp. *BMC Genomics*. 2019;20:172.
397. Gimenez-Ibanez S, Boter M, Fernández-Barbero G, Chini A, Rathjen JP, Solano R. the bacterial effector HopX1 targets JAZ transcriptional repressors to activate jasmonate signaling and promote infection in *Arabidopsis*. *PLoS Biology*. 2014;12:e1001792.
398. Schellenberg B, Ramel C, Dudler R. *Pseudomonas syringae* virulence factor syringolin A counteracts stomatal immunity by proteasome inhibition. *Molecular Plant-Microbe Interactions*. 2010;23:1287–93.
399. McInerney JO, McNally A, O’Connell MJ. Why prokaryotes have pangenomes. *Nature Microbiology*. 2017;2:17040.

Appendices

Appendix A: Confirmation of plant interaction and subphylogroup 2c *Pseudomonas syringae*

Material and Methods

Hypersensitive (HR) Assay of Bean plants

Prior to each assay, *Phaseolus vulgaris* cv. Tendergreen plants were grown in 4 inch rounded pots for 2 weeks. These plants were grown in a communal growth chamber with a 16h:8h day-night cycle with temperatures of 22 degrees Celsius during the day and 18 degrees at night. A high light intensity of 200-250 $\mu\text{mol}/\text{m}^2$ was used for the day cycle. Each plant was grown to ensure full expansion of primary leaves prior to exposure to the *P. syringae* strains.

A streak plate culture on King's B agar was grown from stock cultures of each *Pseudomonas syringae* strain, *P. syringae* B728a (PsyB728a, Positive control [156]), *P. syringae* pv. *syringae* 508 (Pss508) and *P. syringae* TLP2 (PsyTLP2), to isolate single colonies. Each plate was incubated at 28 degrees Celsius overnight. From each streak plate, a single isolate colony was carefully selected and gently resuspended in 100ul of sterile 0.95% NaCl and then spread plated on King's B agar to generate a clonal bacteria for each strain. The plates were the incubated at 28 degrees Celsius for 24 hours. Afterwards, each bacteria lawn was gently scraped and suspended in 5ml of sterile 10mM MgCl_2 and then diluted to two concentrations (1×10^6 , and 1×10^8 cells/ml). The sterile buffer was used as the negative control. Each suspension was then directly injected in the leaves of 2-week old *Phaseolus vulgaris* cv. Tendergreen on the abaxial side using a blunt syringe. Inoculated sites were then monitored 24 hours after injection for target plant reaction. These reactions were assessed as no response (unchanged leaves), HR-like response (dry, dark necrotic spot), and successful infection (wet, necrotic spot).

Results and Discussion

Within 24 hours post-inoculation, the appearance and associated immune response of *P. vulgaris* differed between the inoculated bacterial strains (Figure A.1). The positive control, *P.*

syringae pv. *syringae* B728a, induced water-soaked lesions at all sites that expanded past inoculation boundaries in subsequent days, eventually leading to full leaf collapse (Table A.1). The onset of



Figure A.1 Representative images of the HR assay injection sites, at 1×10^8 mg/ml, after 24 hours. *P. syringae* injection sites on *P. vulgaris* cv. Tendergreen leaves are highlighted in red. (a) Water soaked lesions induced by *P. syringae* pv. *syringae* B728a (b) Dry lesions induced by *P. syringae* pv. *syringae* 508 (c) No lesions induced by *P. syringae* TLP2.

severe symptoms, including chlorosis of leaf tissue, was dependent on external humidity. On days with low humidity, severe symptoms of the plant tissue were typically delayed past 24 hours. The strain *P. syringae* pv. *syringae* 508 elicited a different response to PsyB728a, developing dry lesions at the inoculation sites that transition to dark necrotic spots (the hypersensitive response) that do not extend past the initial inoculation boundaries. Furthermore, no symptoms were observed for the lower initial concentrations of 1×10^6 cells/ml. The dry lesions observed at 1×10^8 cells/ml indicate that this strain is unable to outcompete the holobiont and generate a sufficiently aqueous environment for growth, in contrast to the water soaked lesions produced by PsyB728a infection [80]. Consequently, *P. vulgaris* is able to suppress and locally contain Pss508 within the infection site, preventing systemic spread [75,122]. Therefore, although it has been shown that Pss508 is capable of persisting in *P. vulgaris* (cv. Blue bush lake) [18], our data shows it is likely avirulent against *P. vulgaris* (cv. Tendergreen) [18].

In contrast to other two strains, PsyTLP2 did not trigger any response from *P. vulgaris* even though this strain is also capable of surviving in *P. vulgaris* (cv. Blue bush lake) [18]. Regardless of inoculation concentration, all infection sites appeared healthy with no indication of virulence, and PsyTLP2 appears to be possibly non-virulent in *Phaseolus vulgaris* cv. Tendergreen.

Table A.1 Variation in plant response in HR assay between three closely related *P. syringae* strains after 24 hour post-inoculation. Number of injection spots displaying a visible plant response (Figure A.1)

Sample concentration	Bacterial strain		
	<i>P. syringae</i> pv. <i>syringae</i> B728a [*]	<i>P. syringae</i> pv. <i>syringae</i> 508	<i>P. syringae</i> TLP2
0 mg/ml (Negative control)	0/24	0 [†] /22	0/24
1 x 10 ⁶ mg/ml	1 [‡] / 24	0/22	0/24
1 x 10 ⁸ mg/ml	24/24	22/22	0 [†] /24

^{*} - *P. syringae* reference strain. Known virulent strain against *Phaseolus vulgaris*.

[†] - 3 sites had visible spot that appears as light circular wounding from injection instead of a plant response

[‡] - All 24 sites developed the same symptoms as higher concentration sites in subsequent days

Appendix B: Chapter 3 Supplementary Materials

Table B.1 BUSCO assessment of the completeness of long-read assembly PsyTLP2 at different pipeline stages. Using the Gammaproteobacteria database, 452 universal orthologs was classified as either complete, duplicated, fragmented, or missing. The three genome sets used for correction are listed under post-correction.

Assembly	Status of universal single orthologs			
	<i>Complete</i>	<i>Duplicated</i>	<i>Fragmented</i>	<i>Missing</i>
Pre-correction				
Unitig	0	0	45	407
Consensus called	243	0	149	60
Signal-HMM polished	368	0	67	17
NN polished	373	0	61	18
Post-correction				
Equidistant Set	451	0	0	1
Psy Group 2 Set	435	0	15	2
Psy 441 Set	420	0	25	7

Table B.2 Number of corrections using different combinations of genome sets.

PsyTLP2 assembly				
<i>Genome sets</i> ¹				
Primary genome set	-	Set 1	Set 1	Set 1
Secondary genome set	-	-	Set 2	Set 3
<i>Detected Errors</i>				
Total	-	3464	3706	3683
Substitutions	-	1279	1361	1358
Deletions	-	1413	1511	1482
Insertions	-	772	834	817
<i>Indel Error Locations</i> ²				
Within gene	-	1487	1587	1565
Intergenic	-	697	757	733
<i>Number of genes</i>				
Complete	3717	4758 (+1072/-24)	4803 (+1128/-31)	4799 (+1120/-28)
Frameshifted	1202	173 (+6/-1034)	124 (+6/-1083)	130 (+6/-1077)
Incomplete	130	97 (+13/-46)	95 (+14/-49)	98 (+15 /-47)

¹The combination of genomes sets. Set one is the equidistant genome set. Set two is the Psy group 2 set. Set three is the Psy 441 set.

²Error locations were determined based on assembly annotations using NCBI's PGAP.

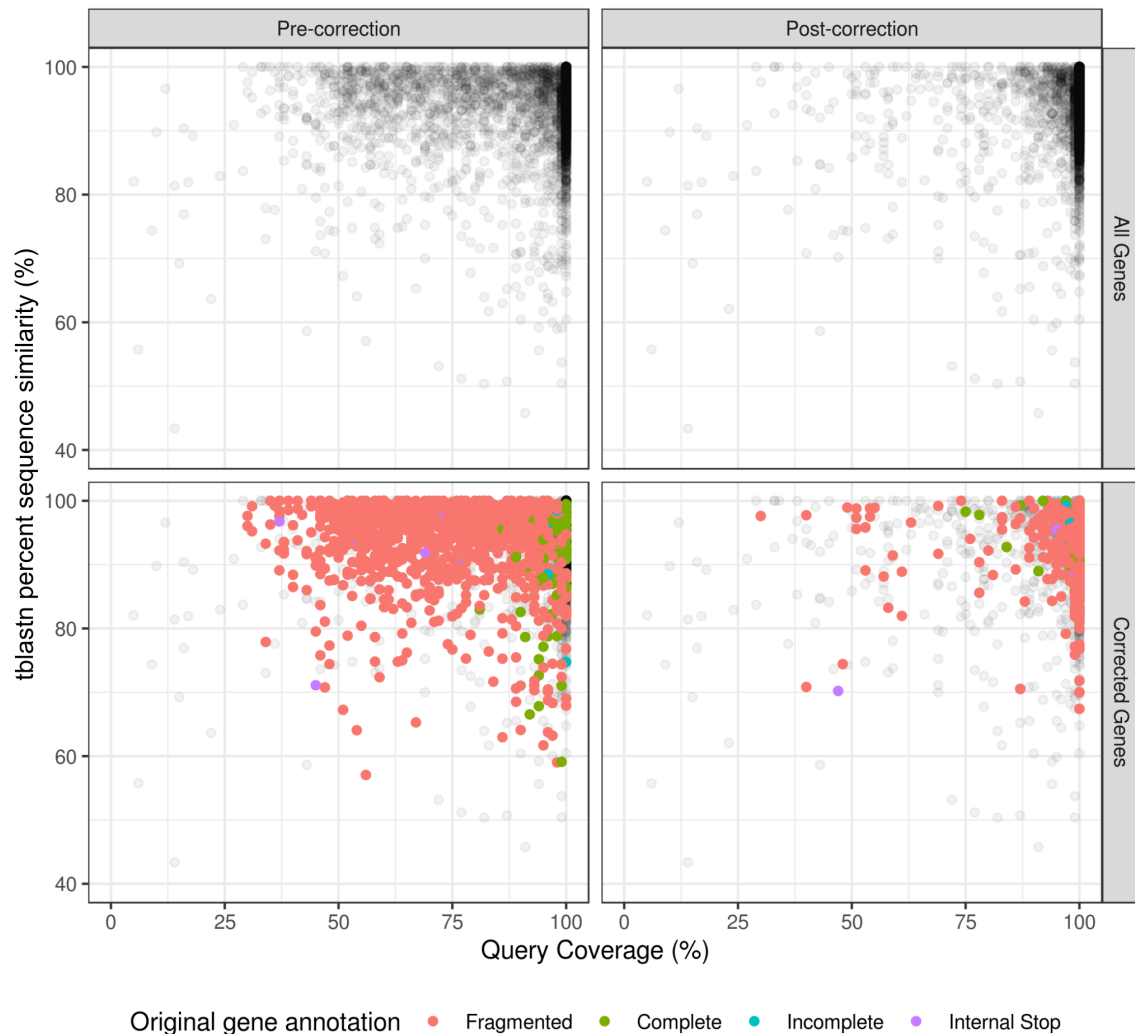


Figure B.1 Reduction of fragmentation in identified genes before and after Kastor correction. RefSeq proteins used in PGAP’s protein homology annotation of pre-corrected assemblies were mapped back to each assembly using tblastn. Corrected genes are coloured according to the original annotations while non-corrected genes are represented as grey. In detail, fragmented indicate genes where two or more frameshifted gene fragments were found; complete indicates complete genes; incomplete indicate truncated genes at the 5’ or 3’ end; and internal stop indicates genes containing a stop codon truncating the full gene. Gene fragmentation or incompleteness is inferred from lower query coverage when tblastn detects only a fragment of expected gene. Longer and more similar genes were found after correction. (Top) Overall profile of gene estimated from RefSeq proteins mapped back to PsyTLP2. (Bottom) Changes in genes found before and after correction. Genes containing corrections are highlighted by the original annotation.

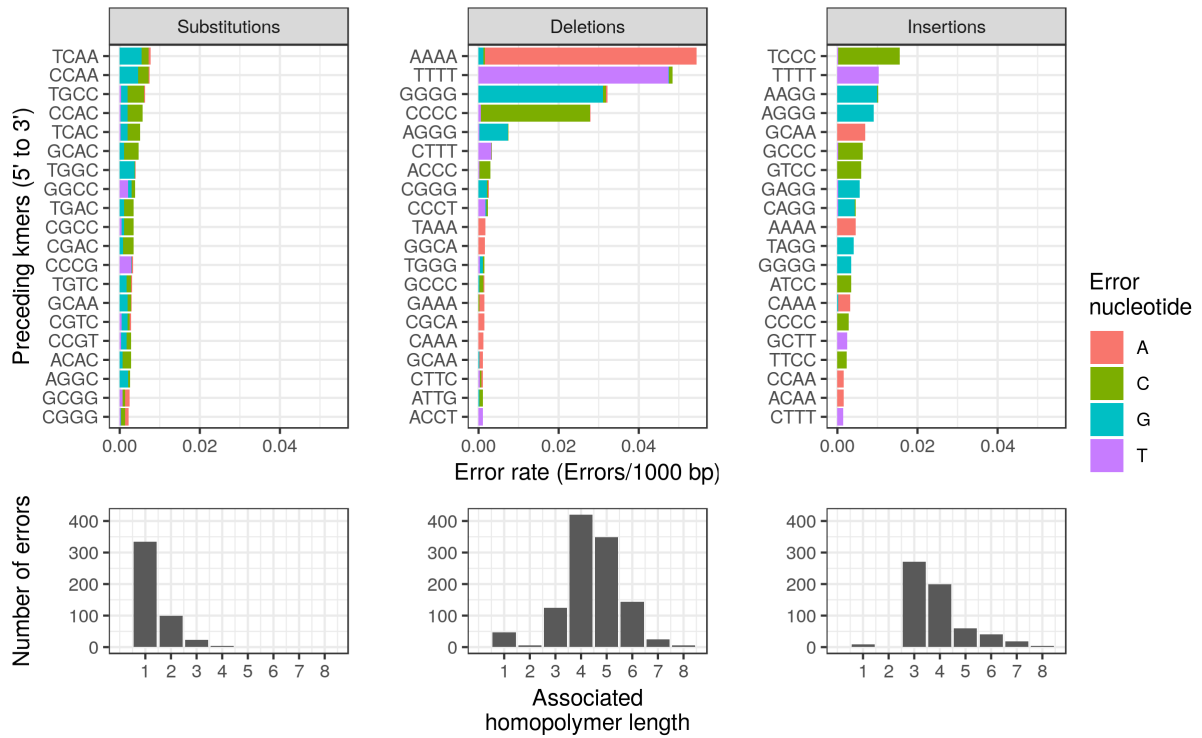


Figure B.2 Breakdown of the genomic sequence preceding detected errors in the PsyTLP2 assembly. Top plots: Top 20 k-mers ($k = 4$) upstream of detected errors separated by error type. Each bar is coloured with the associated erroneous nucleotide. Bottom plots: the distribution of the length of upstream homopolymers for the respective error type.

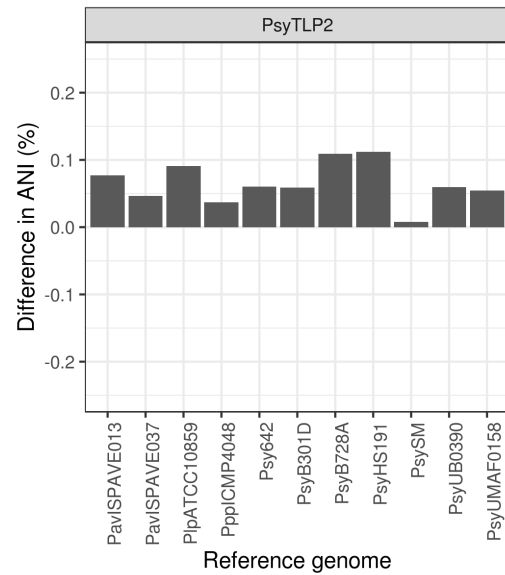


Figure B.3 Post Kastor correction Average Nucleotide Identity percent differences. ANI between PsyTLP2 and reference genomes was compared before and after error correction.

Table B.3 Reference genome nucleotide distance chart. Reference *P. syringae* genomes are labeled by strain name Top-right values represent gapped identity while bottom-left values represent symmetrical identity

	642	UB0390	B728a	B301D	SM	10859	HS191	UMAF0158	ICMP4048	ISPaVe013	ISPaVe037
642	-	98.73	94.35	94.28	94.42	94.5	94.34	94.04	94.11	94.73	94.61
UB0390	88.85	-	94.35	94.35	94.56	94.59	94.45	94.13	94.16	94.73	94.64
B728a	78.67	79.12	-	98.62	94.94	95.03	94.92	94.44	94.62	95.44	95.435
B301D	78.64	79.12	90.94	-	94.92	95.02	94.86	94.45	94.66	95.41	95.45
SM	79.09	79.11	83.22	82.27	-	98.04	97.08	94.48	94.53	96.48	96.43
10859	80.01	80.09	84.52	84.25	86.8	-	97.18	94.58	94.76	96.63	96.56
HS191	79.01	79.02	84.12	83.35	86.011	87.6	-	95.87	95.84	96.1	96.07
UMAF0158	80.57	80.03	82.23	82.5	81.89	83.52	86.06	-	98.67	95.15	95.14
ICMP4048	76.74	77.71	77.69	78.06	78.44	78.56	80.89	87.02	-	95.11	95.11
ISPaVe013	76.71	77.84	80.6	79.59	80.41	80.93	79.92	79.85	77.68	-	98.16
ISPaVe037	77.66	78.99	80.35	95.45	80.82	82.75	80.01	81.45	79.53	87.35	-

Appendix C: Chapter 5 Supplementary Materials

C.1 Incongruencies in core gene phylogeny with past research

Among the remaining strains, the constructed phylogeny is largely congruent with previous studies with only minor variations in the tree topology [43,44]. These differences include variation in strain placement within subphylogroups, particularly for groups 02a-d. However, these incongruencies are not unexpected and may arise from differences in core gene selection, influenced by the collection of genome in the study, and alignment during tree generation [50]. The core genes identified are shared between diverse group of *Pseudomonas* strains are different between this work and previous studies, resulting in slight changes in some strain placements in more sequence similar subphylogroups, compared to more divergent subphylogroups (e.g. subphylogroup 2c). Gene sets of highly diverged *P. syringae* strains, especially those in early branching phylogroups (e.g. group 13) would restrict the hard core genome limiting sequence diversity needed for very closely related strains [362]. This is exemplified in the heavily represented phylogroup 1, split here and in other studies [43,44] into subgroups 01a and 01b due to large number of strains (Figure 5.1). Many branches in this monophyletic group have SH-like support less than 70 indicating that there low support for specific topologies compared to other branches of a similarly size phylogroup (Figure C.1b-d). The lack of support reflects the high degree of similarity within group 1a and 1b strains of *P. syringae* [24,39,324]. However, the overall core genome is consistent with previous studies.

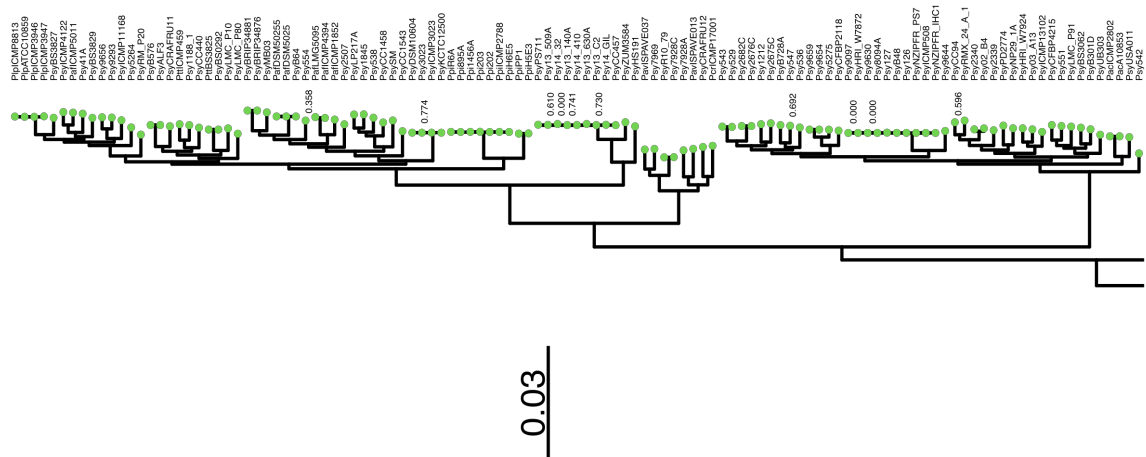


Figure C.1a Enlarged view of Phylogroup 2, excluding 2c, in Figure 5.1 phylogenetic tree generated from 2087 single copy gene families using FastTree with SH-like support below 0.8 are displayed. Strains in this figure depict the majority group 2 strains (the last branching phylogroup) after divergence with subphylogroup 2c strains.

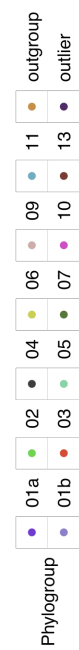
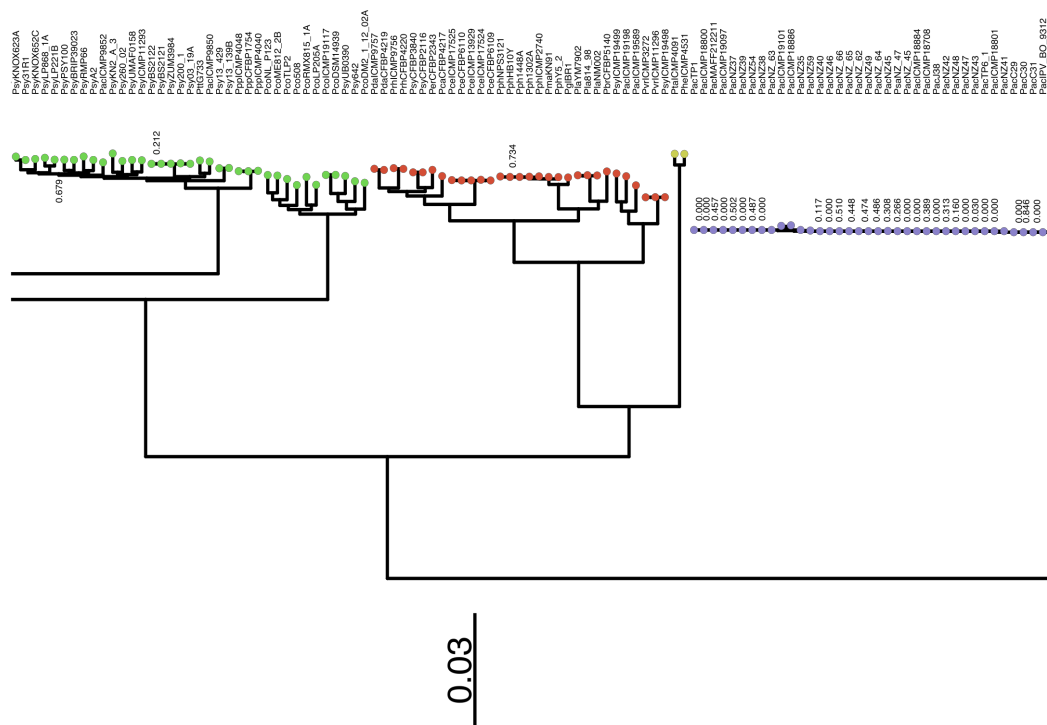


Figure C.1b Enlarged view of phylogroups 2c, 3, and 6 plus parts of phylogroups 1b and 2, from figure 5.1 phylogenetic tree generated from 2087 single copy gene families using FastTree with SH-like support below 0.8 are displayed. The low diversity of in phylogroup 1b strains is visible from the very short branch length and/or poor branch support. The subphylogroup 2c containing sequenced strains Pss508 and PsyTLP2 is the first subgroup to diverge in group 2.

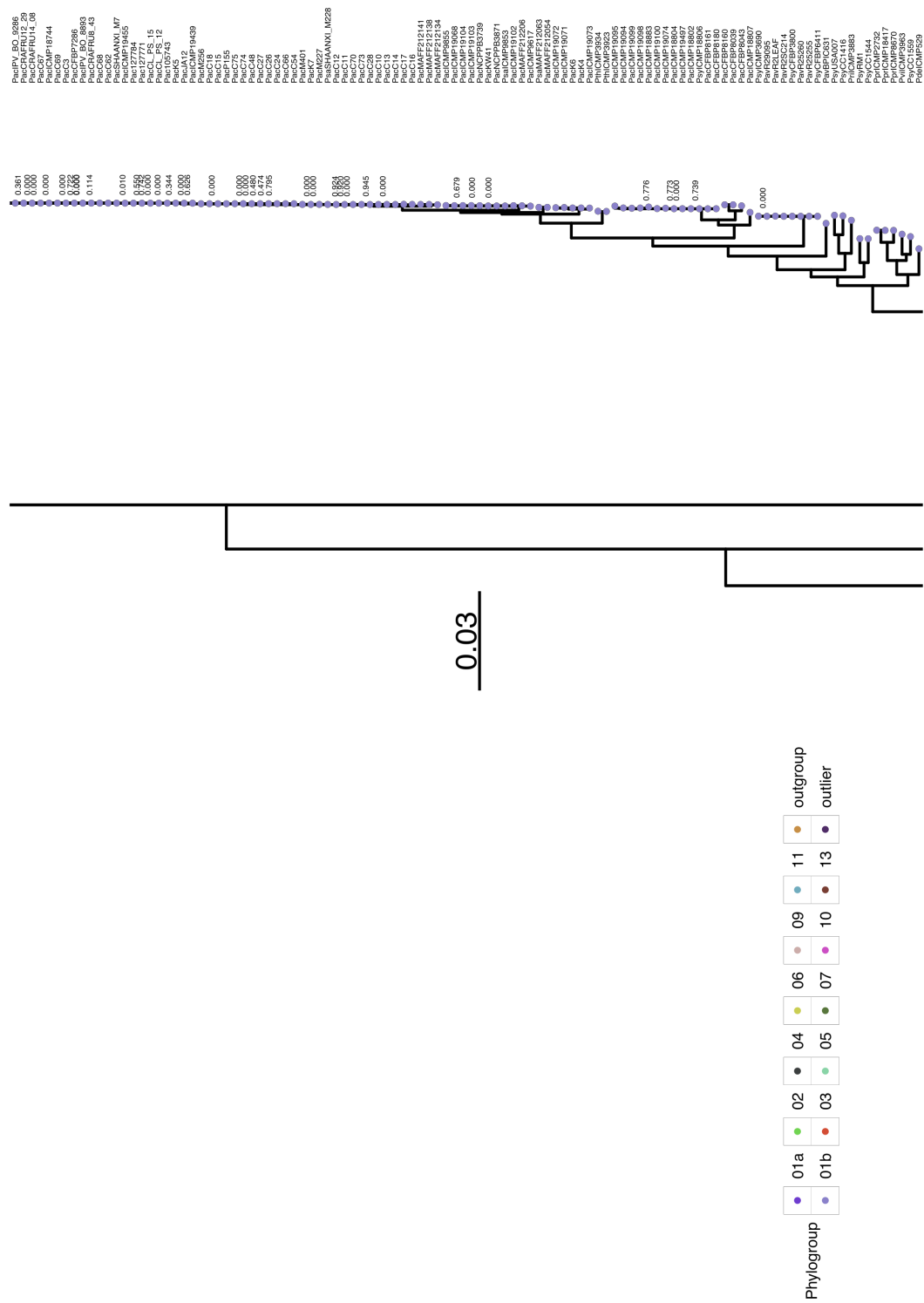


Figure C.1c Enlarged view of the figure 5.1 phylogenetic tree generated from 2087 single copy gene families using FastTree with SH-like support below 0.8 are displayed. Phylogroup 1b is among the least diverse phylogroup in the genome set; a large portion of these strains contain very short branch lengths and/or poor support.

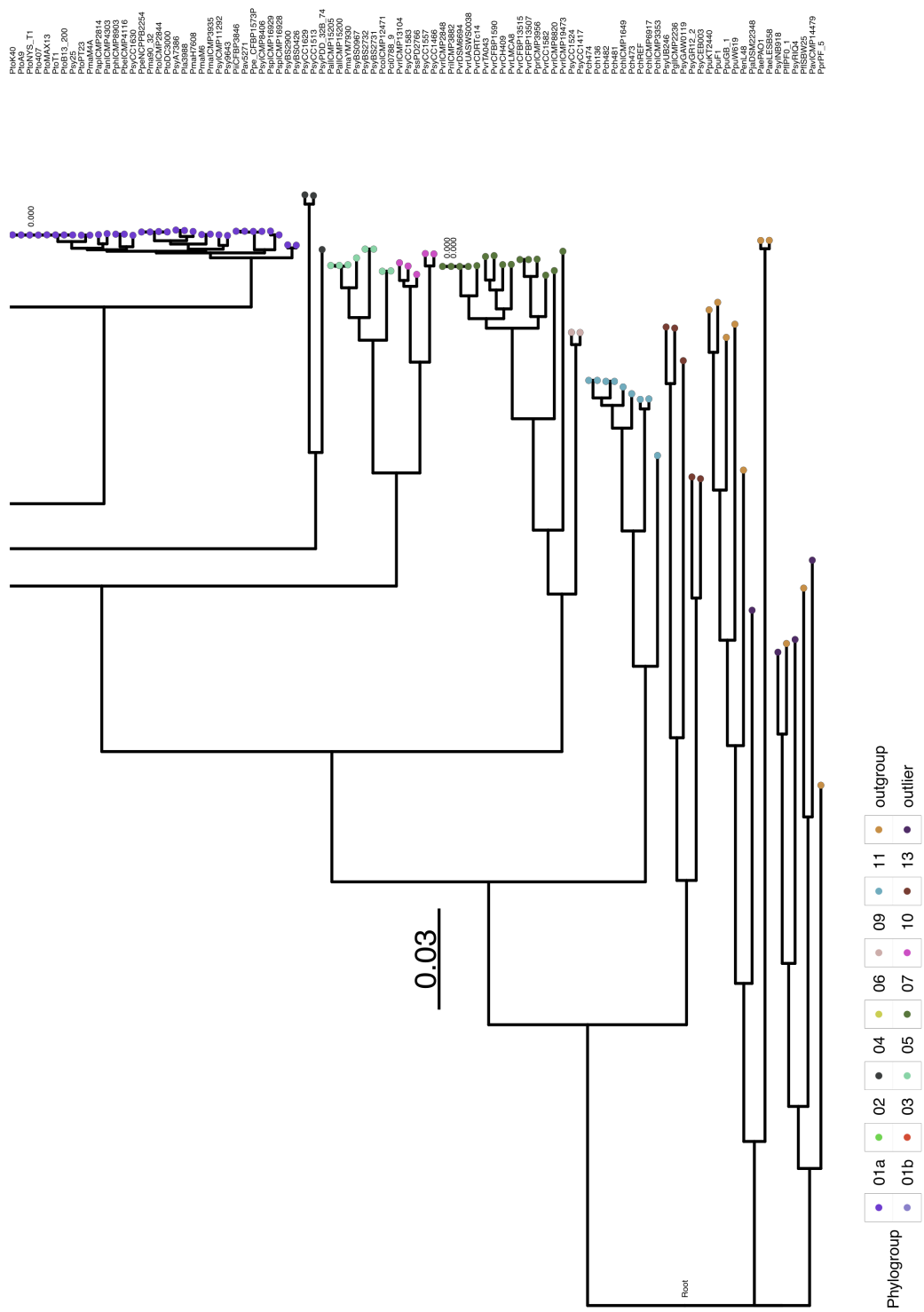


Figure C.1d Enlarged view of the figure 5.1 phylogenetic tree generated from 2087 single copy gene families using FastTree with SH-like support below 0.8 are displayed. The phylogeny was rooted with 10 reference *Pseudomonas* strains belonging to other pseudomonad species groups. *P. syringae* strains grouping with the outgroup were dropped from further analysis. Besides phylogroup 1a, the depicted phylogroups are considered the early branching strains (secondary phylogroup).

C.2 Imbalance in accessory genome from strain selection

The majority of gene families in the pangenome are a part of the accessory genome with 31220 gene families appearing in less than 10% of the strains. As expected, singletons are a large part of this group with 14393 of gene families (46.1%) found primarily in a single strain. The number of singleton genes is not uniformly distributed among the phylogroups, with less well-represented phylogroups containing proportionally more singleton gene families especially when the incidence of another strain containing a similar geneset increases with more representative strains from the same phylogroup. Thus, well-represented phylogroups have fewer unique gene families (singletons). [44].

This imbalance becomes apparent when the pangenome is assessed as phylogroups rather than bacterial strains (Figure C.2). To determine the unique genetic structures in subphylogroup 2c, we identified gene families that are shared between phylogroups with subphylogroup 2c as a distinct subgroup de-coupled from group 2. In this case, a gene family is attributed to a phylogroup when 95% of the group's strains harbour the gene family. Using this criterion, the phylogroups with greatest unique gene families are primarily from early branching phylogroups 9, 11, and 13. Similar to singleton genes, these gene families are isolated within specific phylogroups. Phylogroup 6 also contained an above average number of unique gene families, however it is only represented by two reference strains so phylogroup-specific genes are not readily assessed. Lower diversity in the sampled phylogroup would mask the more variable accessory gene families within these groups as unique to the phylogroup. Among the grouped sets, phylogroups 9, 11, and 13 plus the early branching phylogroup 7 share the most gene families (Figure C.2,) between each other than in comparison to other groups, sharing more gene families than the late branching strains.

In contrast, over-represented phylogroups contain fewer unique gene families than the earlier branching groups, with phylogroups 1,2, and 3 having the fewest unique gene families specific to these phylogroups (Figure C.2). Many of the genes in accessory genomes can also be found in other phylogroups suggesting a more “closed” phylogroup pangenome where additional genomes do not introduce new genes at a linear rate, but at a decreasing rate [399]. Late branching members within phylogroup 2 are one of the most densely sampled regions, and include one of the two first reference *P. syringae* strains, *P. syringae* pv. *syringae* B728a [158]. When subphylogroup 2c is analyzed separately, the remainder of group 2 was found to contain less than 50 unique gene families.

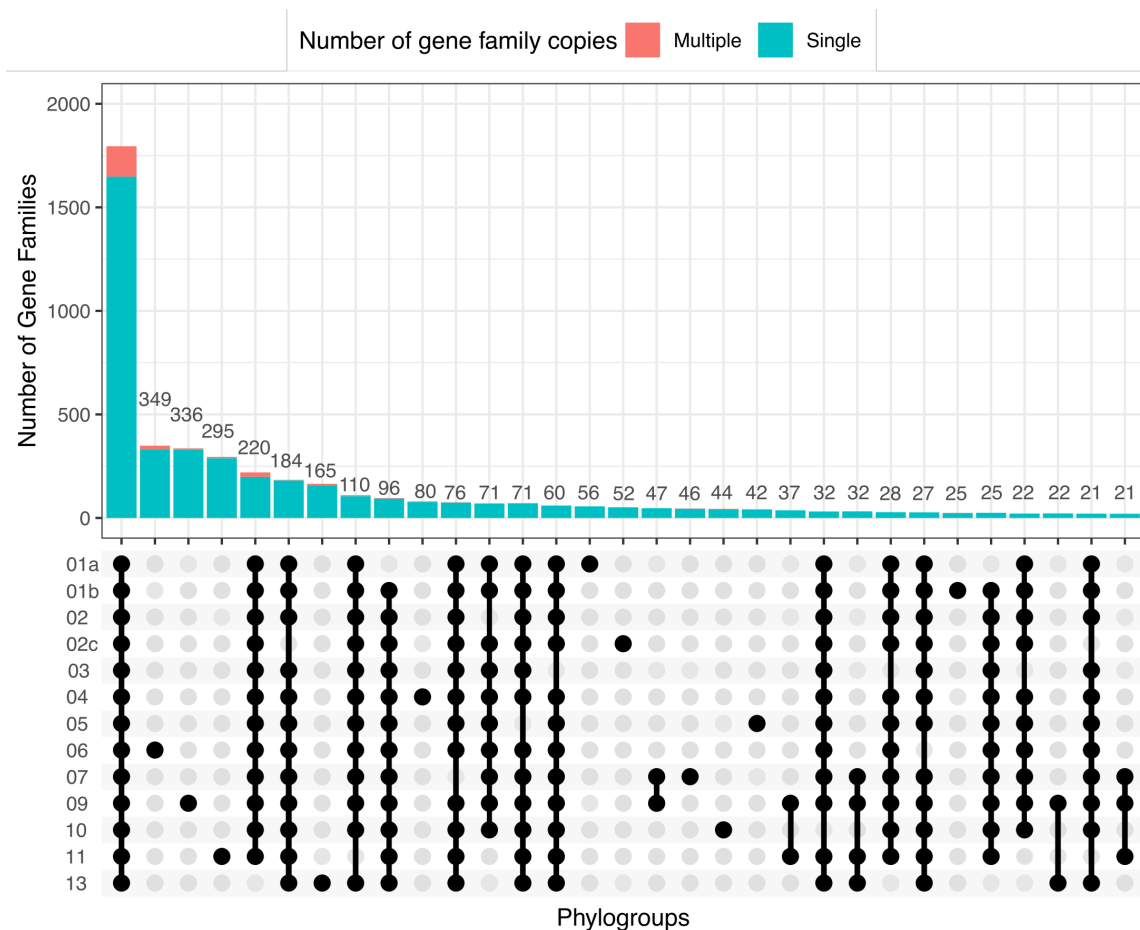


Figure C.2 An upset plot representing shared gene families between phylogroups (i.e., intersection in a Venn diagram). A gene is considered present in a phylogroup when it is present in 95% of the phylogroup strains. Joined black dots indicate phylogroups sharing the gene family. Subphylogroup 2c has been distinctly separated from phylogroup 2 to demonstrate the high number of unique gene families within this smaller subgroup.

C.3 Incongruence strains grouping with gene set clustering at different protein level

With hierarchical clustering, *P. syringae* strains groups were highly congruent with the core genome phylogeny with a few notable exceptions (Figure C.3, right cladogram). Strains within each phylogroup were strongly clustered together with very minimal strain re-arrangements within phylogroups. The most notable is the firm placement of PG-2c within the group 2 cluster, indicating

the subphylogroup shares the majority of its accessory genome with the larger phylogroup, supporting the recent divergence and a likelihood of subgroup 2c having continual gene exchange with virulent strains [338,395].

At the phylogroup level, phylogroups 10, 5, and 4 are clustered marginally closer to phylogroup 3 rather than group 1. The closer clustering of the more distant phylogroups indicate that in the full gene set union of these three phylogroups share closely related gene families with phylogroup 3 that are absent or divergent in phylogroup 1. Some of these gene families may represent mobile virulence genes due to infrequent inter-phylogroup recombination transfer events [43]. In fact, the grouping of these gene families may be attributed in part to the higher similarity of type III secretion related genes of phylogroup 5 with the phylogroup 3 variant. However, if the gene family threshold is reduced to 50% identity, the clustered phylogroups become highly congruent with the core genome phylogeny (Figure C.3). This suggests that the full gene set of phylogroups 5, 4, and 10 share more gene families with phylogroup 1, but sequence divergence has separated these families when a higher identity threshold is used. Overall, the observed low mobility of the pseudomonad genes during the gene set clustering supports an infrequent transfer between phylogroups.

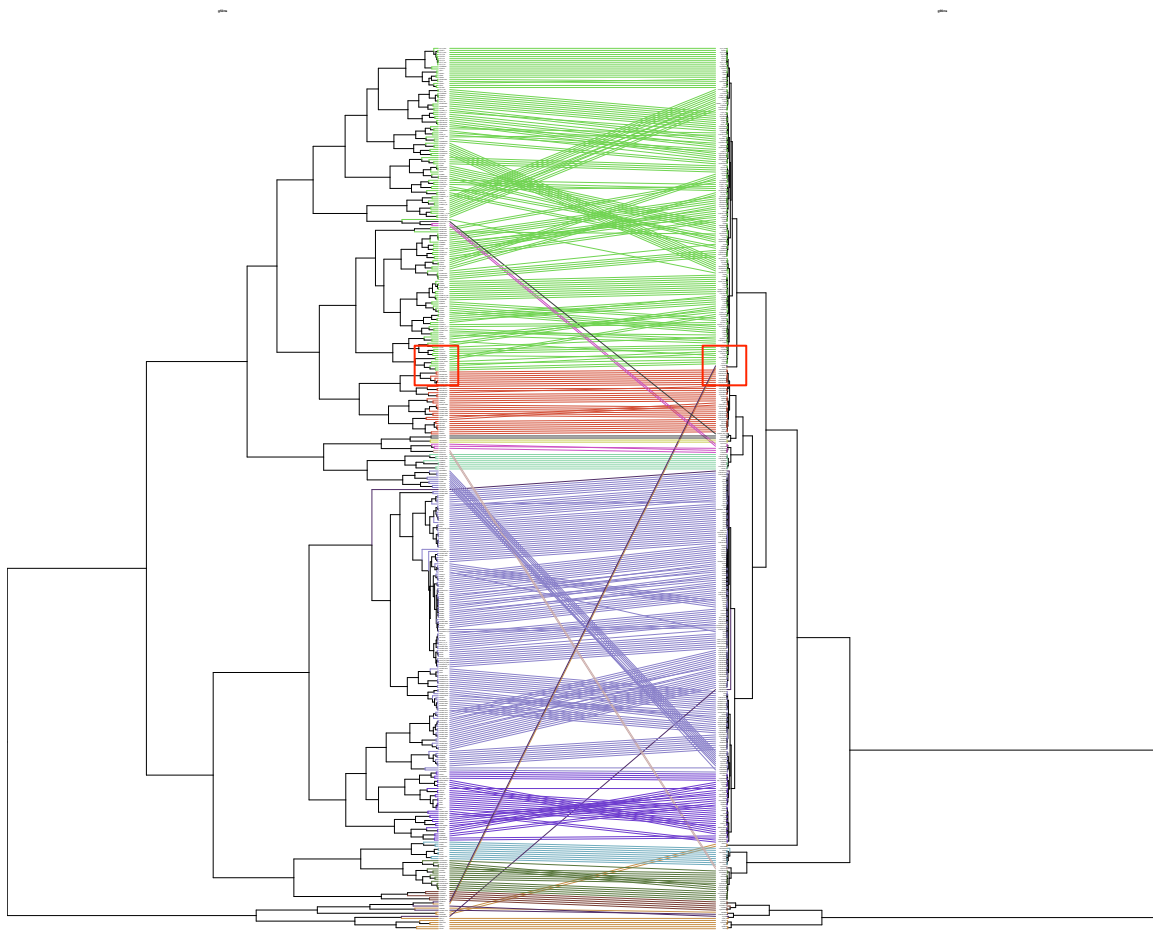


Figure C.3 Tanglegram comparison of the hierarchical clustering of shared gene families at 50% identity (left) compared to 90% identity (right). Slight changes to phylogeny grouping depending on the stringency of protein family identity. Identical strains for both trees are connected by a coloured line indicating the core-genome identified phylogroup. Subphylogroup 2c strains are highlighted in the red box (green) with distantly related group 13 (brown).