The Pennsylvania State University

The Graduate School

COMPETITION AND VIRULENCE IN *PSEUDOMONAS SYRINGAE*

A Dissertation in

Ecology

by

Chad Fautt

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2023

The dissertation of Chad Fautt was reviewed and approved by the following:

Estelle Couradeau
Assistant Professor of Soils and Environmental Microbiology
Co-Dissertation Advisor
Co-Chair of Committee

Kevin Hockett
Assistant Professor of Microbial Ecology
Co-Dissertation Advisor
Co-Chair of Committee

Francisco Dini Andreote
Assistant Professor of Phytobiomes

David Kennedy
Assistant Professor of Biology

Jason Kaye
Distinguished Professor of Soil Biogeochemistry
Chair of Ecology Intercollege Graduate Degree Program

ABSTRACT

The *Pseudomonas syringae* species complex (PSSC) is the most well-studied group of bacterial plant pathogens in the world, due in part to the large range of economically important crops they infect. While host-pathogen interactions are directly responsible for the disease caused by PSSC pathogens, the process of disease often requires pathogens to be successful plant epiphytes with the ability to outcompete other bacteria living on the surface of the leaf. In this dissertation, I present work that approaches PSSC not only as a capable pathogen, but as a deft microbial competitor, with the intent to improve our understanding of the strategies used by PSSC that allow it to succeed during the early stages of the disease process, and as a leaf epiphyte more generally. In chapter 2, I leverage 2,161 published PSSC genomes to analyze 16 polymerase chain reaction (PCR) primer sets for their ability to broadly identify PSSC strains, build novel classification models based on the best performing primer sets, show the potential of these classification models to predict virulence factors, and create a web portal incorporating the datasets and classifiers generated. In the next two chapters, I use genomics and simulation modeling to explore antimicrobial toxins produced by PSSC. In chapter 3, I use genomics to uncover new diversity of a potent toxin, tailocins, carried by PSSC, compare protein structures associated with this diversity, and suggest physiological features of the bacterial cell that might be driving the distribution of tailocins within the species complex. In chapter 4, I use agent-based modeling to explore the fitness costs of toxin production when the release of said toxin requires cell lysis and death. I show that such lysis might itself provide a benefit, if bacteriocin production is linked to intracellular damage. Overall, this dissertation explores the rich life of PSSC beyond plant-pathogen interactions and provides valuable insights into the microbial warfare that pathogens engage in outside of their host.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation wouldn't have been possible without the encouragement and support of more people than I am able to thank here.

For their financial and professional support during my time at Penn State, I would like to first and foremost thank my advisors Kevin Hockett and Estelle Couradeau, both of whom have given me incredible freedom over this dissertation, allowing me to pursue answers to questions I was interested in, using methods they had little experience with, and giving me time to figure things out when they inevitably went wrong. In addition, I want to thank my committee members Francisco Dini Andreote and David Kennedy for their thoughtful advice and help along the way, and Katriona Shea, whose *Ecological and Environmental Problem Solving* course was formative in my thinking about ecological processes and ultimately resulted in the work presented in Chapter 4 of this dissertation. I also would like to thank Ecology program chair, Jason Kaye, for having students' best interests at heart and actively working to make sure Ecology students have the best experience possible.

For their early encouragement and mentorship, I want to thank the professors at American River College. Community colleges represent a first step for many students, including me. The brilliant, thoughtful teachers I encountered at ARC not only provided me a strong academic foundation for future success but helped me consider a path forward I didn't know was possible.

In particular, I want to thank Professor Tami Hong, for your constant encouragement, and Professor Linda Zarzana, for your passion for both teaching and your students, and in particular for pushing me to apply for an REU research position when I had no idea what scientific research even looked like.

I also would like to acknowledge and deeply thank my undergraduate PI at UC Davis, Marie Jasieniuk, who in addition to mentoring me for two years, encouraged me to apply to graduate school and acted as a catalyst for my thinking that such a thing was a possibility for me.

Finally, I want to thank my partner, Sarah. Your constant faith in me and reassurance has helped me more than you can know. Sometimes you just need somebody to say, ‘stop being dramatic, you’ll be fine’, and I can’t thank you enough for being that person for me.

The material in chapter 2 is based upon work supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, through the Northeast Sustainable Agriculture Research and Education program under subaward number GNE20-232. The findings and conclusions do not necessarily reflect the view of the funding agency.

Additional support was provided by the College of Agricultural Sciences), the department of Ecosystem Science and Management, and the department of Plant Pathology and Environmental Microbiology through the mBiome initiative.

Chapter 1

Introduction

The *Pseudomonas syringae* species complex (PSSC) contains ca. 15 recognized species (1) that together cause disease in virtually every crop plant grown throughout the world, making it both one of the most common and economically significant bacterial plant pathogens. An illustrative example of the pathogenic potential contained within the PSSC is the 2008-2012 epidemic of bacterial canker of kiwifruit caused by *P. syringae* pathovar *actinidiae*, which at the height of the epidemic could be found on an estimated 81% of all kiwifruit trees in New Zealand, 10% of kiwifruit trees in France, and resulted in 10-50% crop losses in Italy (2). In New Zealand alone, the disease caused upwards of $120 million dollars in damage a year (3). Other notable outbreaks of disease caused by strains within the PSSC include Bleeding canker of European Horse Chestnut (4), bacterial speck of tomato (5), bacterial leaf spot of watermelon and squash (6) and bacterial blight of sweet onion(7)

PSSC is primarily a leaf, or foliar, pathogen (8).The process that leads to disease by foliar pathogens such as PSSC can be thought of as occurring in two main stages (9). In the first stage of the disease process, the epiphytic stage, a pathogen immigrates to a plant, often onto the leaf surface, and attempts to grow its population to a point that maximizes the chance of encountering an entry point to the interior of the leaf. Immigration is always a passive event (10), with bacterial cells being deposited onto the leaf by splashing of rain or irrigation water from nearby plants(11), by wind from nearby fields (12), or by rain (13). As the exterior of a leaf is exposed to large and frequent fluctuations in weather throughout the day, the niche breadth of the pathogen, including optimal temperature and humidity for growth along with resilience to desiccation stress and

ultraviolet radiation play a large role in survival at this stage (14), and ultimately in the ability to cause disease (9,14,15). To thrive on the leaf surface, the newly immigrated pathogen will also have to compete with the resident leaf microbiome for its share of the small pool of nutrients available, most of which are slowly leached out from the interior of the leaf in small, isolated patches where the waxy leaf cuticle is particularly thin or damaged (16–18). Despite all of the challenges faced at this stage of infection, a pathogen's ability to grow its population at this point is crucial, as epiphytic population size is correlated with both the probability of infiltrating the interior of the plant and the severity of disease it will eventually cause (14,19,20).

The second and final stage in the disease process is the endophytic stage. After accessing the plant interior through stomata, hydathodes, or other openings such as wounds in the leaf surface, success at the endophytic stage is determined primarily by plant-microbe interactions. The type III secretion system (T3SS) and associated type III effector (T3E) repertoires are the most significant and well-studied virulence factors. T3Es, after being injected into the plant cell via the T3SS, directly interfere with many molecular processes within the plant, resulting in disrupted plant defenses, the release of water and nutrients into the apoplast for use by the pathogen, and forced closure of stomata for increased humidity that is favorable for bacterial growth (21). Critically, however, when exposed to T3Es for enough time, plants commonly evolve the ability to detect specific effectors and use them as a signal to upregulate defense mechanisms, leading to the hypersensitive response (22). When this happens, the detected effector becomes a liability for the pathogen when trying to infect certain plants and becomes an *avirulence* factor. Given that pathogens in PSSC can carry anywhere from four to fifty T3Es, it is thought that the specific T3E repertoire a pathogen carries plays a large role in determining its host range (23). Crucially, however, many T3Es reside in a region of the genome called the exchangeable effector locus (EEL), which contains hallmarks of horizontal gene transfer including multiple transposable elements and a significantly lower GC content than the rest of the

genome (24). As the name suggests, this region acts as a repository for effectors which can be easily exchanged among pathogens throughout PSSC (25). Two consequences of the EEL for plant pathologists are that 1) host ranges and pathogenicity rarely coincide with phylogenetic signals (26) and 2) interactions between pathogens outside the plant that are difficult to study likely play a significant role in the evolution and diversification of pathogenicity within PSSC (27). Along with the T3SS, PSSC strains often produce plant hormone-like compounds that also play a key role in disease outcome (28,29), and metabolic genes that are tailored to living within the plant. One example of such metabolic genes is the recent discovery of the Woody Host and *Pseudomonas* region (WHOP) carried by several PSSC pathogens of woody plants (30–32). This region is thought to allow the pathogen to subsist on the complex carbohydrates present in woody tissue and is commonly found in pathogens that cause cankers in chestnut trees and kiwi vines. The elevated virulence in woody tissue is particularly significant as the economic loss associated with diseased perennials are often felt for multiple years, as opposed to the single season of losses associated with diseased annual crops (33).

Challenges at both stages of the disease process can hinder the ability of a pathogen to cause disease, and as such most plant pathologists evaluate pathogens within the framework of the Disease Triangle (34,35) - a conceptual model in which pathogen genetics, host genetics, and the environment interact to ultimately determine disease severity. This holistic approach recognizes the need to understand the entire ecosystem in which pathogens and hosts interact to make accurate disease predictions. However, one significant aspect often overlooked within this framework is the role of microbe-microbe interactions. Specifically, as resources are limited on the leaf surface, it's likely that competition plays a key role in survival of pathogens in the epiphytic stage and ultimately influences their ability to cause disease. Indeed, certain microbial leaf communities have been shown to have a protective effect for plants (36–39), demonstrating their role as an added layer of defense that pathogens must overcome. It has even been suggested

that the role of the host microbiome is so important for disease outcome that the longstanding Disease Triangle model should be modified to add a fourth leg representing this vital component (40). For a pathogen faced with an established community of epiphytic bacteria, there are generally two ways to improve competitive outcomes: 1) exploitative competition, where an organism tries to improve its ability to gather and use resources, and 2) interference competition, where an organism reduces the ability of competitors to gather and use resources. Strong exploitative competition between two organisms is typically thought to be a precursor to interference competition (41).

In this dissertation, I aimed to use computational methods to aid not only in our understanding of host-pathogen relationships and the end result of the disease process, but to also enrich our understanding of the strategies PSSC uses to outcompete other bacteria during the epiphytic stage of its life, with a particular focus on the ubiquitous anticompetitor toxins known as bacteriocins (42).

In chapter 2, to address the need for accurate diagnostic to accompany the marker gene based molecular assays currently used by plant pathologists, I use *in-silico* methods to evaluate and compare previously developed PCR primer sets for their ability to amplify broadly throughout the species complex. I further test if the amplicons produced by the primer sets accurately reflect the whole genome similarities between isolates, and if the classification resolution achieved by novel classification models allows for predictive insights in the functional potential of the PSSC isolate. In chapter 2.2, I describe the development of a web-based portal for rapid classification and T3E repertoire prediction in PSSC isolates from marker gene sequences, and the accompanying datasets created in chapter 2.1.

In chapter 3, I use genomic methods to investigate the diversity of tailocins in PSSC. As a potent bacteriocin found in ca. 85% of PSSC pathogens, tailocins are likely important mediators of competition on the leaf surface and beyond for PSSC (43). Tailocins are evolutionarily related

to bacteriophage (44), and like bacteriophage are very selective in their targeting, exhibiting narrow killing ranges that are determined by the ability of the tailocin's tail fibers to bind to specific cell surface molecules, with the primary target being sugar moieties found in the target cells' lipopolysaccharides (LPS) (45).To investigate the full diversity of tail fiber diversity and potential eco-evolutionary pressures acting on their distribution throughout PSSC, I screen 2,161 publicly deposited genomes for tailocin-associated tail fibers, describe new diversity not previously reported, and use AlphaFold-predicted protein structures. I show for the first time that tail fibers in PSSC tailocins can be divided into to two functionally distinct groups that are homologous to those found in bacteriophages: the short tail fibers and long tail fibers. Furthermore, I show that that distribution of tail fiber variants throughout the species complex correlates strongly with a known LPS synthesis gene, *rfbD*, suggesting that the tail fiber carried by a given PSSC isolate is dependent on its compatibility with LPS structure.

In chapter 4, I investigate bacteriocins more broadly, focusing on a striking feature of the toxins that has been described as a spiteful behavior (46). For many bacteriocin producers, the producing cell must kill itself to release toxin into the environment (47,48). In this chapter, I use agent-based modelling to highlight and present a possible solution to a paradox that has arisen in recent research regarding the regulation of bacteriocin production. Briefly, the paradox I address is that despite longstanding evidence that bacteriocin production only occurs in cells with DNA damage (49–53), recently there have been independent reports of a small portion of bacterial populations expressing and ultimately releasing bacteriocins under laboratory conditions with no addition of DNA damaging agents (54,55). In this chapter, I investigate a possible mechanism responsible for the constitutive bacteriocin production using an agent-based model, showing that the low-level constitutive bacteriocin production that has recently been observed is consistent with damage-induced bacteriocin production under genotoxic (i.e. DNA damaging) stress. Therefore, while the observations might be surprising, they are simply an

unexpected consequence of regulatory mechanisms long known to control expression of bacteriocin genes. I also show that in explicitly linking bacteriocin production to cellular damage, a secondary benefit arises from bacteriocin production in the culling of the most damaged population members and increasing resource use efficiency. This in turn allows a fitness advantage over competitors in some scenarios, even when competitors are resistant to the bacteriocin released by producers.

In chapter 5, I summarize the results of this dissertation, and suggest future research directions.

Chapter 2

Development of Syringae.org, a rapid LIN-based subspecific classification and virulence factor prediction tool for *Pseudomonas syringae*

2.1 - Evaluation of the taxonomic accuracy and pathogenicity prediction power of 16 primer sets amplifying single copy marker genes in the *Pseudomonas syringae* species complex

Abstract

The *Pseudomonas syringae* species complex is comprised of several closely related species of bacterial plant pathogens. Here, we used *in-silico* methods to assess 16 PCR primer sets designed for broad identification of isolates throughout the species complex. We evaluated their *in-silico* amplification rate in 2,161 publicly available genomes, the correlation between pairwise amplicon sequence distance and whole genome average nucleotide identity (ANI), and we trained naïve Bayes classification models to quantify classification resolution. Further, we show the potential for using single amplicon sequence data to predict an important determinant of host specificity and range, type III effector protein repertoires. The results presented here provide insights into the capabilities and limitations associated with amplicon sequencing for pathogen identification.

Introduction

The *Pseudomonas syringae* species complex (PSSC) consists of many closely related plant pathogens (56). With host ranges and symptomology that can overlap, accurate identification of isolates can be difficult (26). Aside from whole genome sequencing, which is costly and often impractical for routine identification, marker gene sequencing is the most effective method for specific and sub-specific classification of unknown PSSC isolates (57–59). This method has been used to aid in the identification of new pathogenic strains and species within the species complex, and detect known pathogens infecting novel hosts (60–62), highlighting the importance of amplicon sequencing for broadening our understanding of the species complex. However, although there have been many proposed PCR primer sets designed to amplify broadly within the species complex (56,63,64), there are open questions about the relative performance of each. Specifically, it is not clear if all primers allow for reliable amplification for all phylogroups in the species complex, or which primer sets allow for the greatest classification resolution. Further, while a primary goal of pathogen identification is often to predict pathogenic potential of the unknown isolate, it is not known if the classification resolution obtained by any of the currently published primers is sufficient to predict genomic features associated with host range and virulence.

While most of the primer sets evaluated in this study were originally designed for use with multi-locus sequence typing (MLST) (56,64), there has been continuous interest in classifying isolates with a single marker gene. In this regard, recombination rates and phylogenetic congruence have been used as metrics to suggest that genes such as citrate synthase (*CTS*) (59) and RNA polymerase sigma factor *rpoD* (65) can be used by themselves to accurately place unknown PSSC isolates into phylogroups. An in-depth comparison of these primer sets as

tools for classification has not been conducted, however, leaving it an open question as to which performs best.

Often, an implicit goal of bacterial isolate identification is to gain some insight into its functionality or ecological significance for the environment it was isolated from. In this vein, assuming a phylogenetic placement with sufficient resolution, prediction of an isolate's gene content can be made, providing insight into functional capacity. This concept has been demonstrated with Picrust2, in which improved functional predictions were achieved over Picrust1 solely from increased resolution in genome prediction (66). As PCR primer sets designed specifically for PSSC are used because they offer greater phylogenetic resolution over those targeting 16S rRNA genes (the marker gene used by Picrust2), we hypothesized that specific genes known to affect host range and virulence could be predicted in genomes based solely on amplicon sequences derived from commonly used PCR primer sets.

In PSSC, pathogenicity is determined in large part by the type III effector proteins (T3Es) carried by the pathogen (67) and therefore predicting T3E repertoires could provide valuable information about potential host range and specialization of an unknown isolate. While many pathogens in the species complex carry 30-40 T3Es, only effectors *AvrE*, *HopM*, and *HopAA* are considered part of the core PSSC genome of PSSC and are thought to confer general virulence to plants (68). Presence/absence of the other T3Es play a role in host adaptation and are more variable in the species complex, suggesting that if there is a taxonomic signature associated to their presence, an infraspecific classification is needed to accurately predict it. It is currently not known what phylogenetic resolution is needed to accurately predict T3E repertoires, or whether any published primer sets might allow high enough resolution to meet this threshold.

Here, we perform *in-silico* tests to assess the performance of 16 previously published PCR primer sets, targeting eight marker genes, against 2,161 PSSC genomes. Metrics used for assessment of phylogenetic classification were amplification rate, congruence of pairwise

amplicon distance with average nucleotide identity (ANI), and performance of naïve Bayes classifiers trained on *in-silico* amplicon data. We also investigate the potential for functional prediction from amplicon data by analyzing Jaccard similarity of T3E repertoires at the level of phylogenetic resolution achieved by each classifier and show that isolates not included in the training dataset can be accurately placed above phylogroup level and prediction of both T3E repertoire size and content is often possible, with presence/absence of 77 T3E subfamilies being correctly predicted with a median accuracy of 93% among 113 genomes in a test dataset consisting of recently sequenced PSSC genomes.

Overall, we found that some published primer sets may have substantial blind spots in the lineages they can amplify. However, many primers tested could both amplify broadly throughout the species complex and be used to classify isolates beyond the phylogroup level, allowing accurate prediction of the T3Es carried by unknown PSSC isolates. Our results suggest that for the highest classification resolution throughout the species complex, resulting in the most consistent T3E repertoire prediction accuracy, primer sets targeting the genes *gapA, gyrB* (63) and *PGI* (64) should be considered as the optimal primer sets.

Experimental Procedures

2,467 Genomes labeled as belonging to the *Pseudomonas syringae* group were obtained from the RefSeq database from NCBI in November 2021. Genomes were checked for completeness and assembly quality with BUSCO v 5.3.2 using default settings and the pseudomonadales_odb10 lineage. Genomes scoring > 99 made up the final dataset used for assessing primers. As the majority of genomes used were not assigned to a phylogroup, phylogroups were assigned based on average nucleotide identity (ANI) with phylogroup

reference genomes produced by (59). While Berge et al. suggest taking a simple nearest neighbor approach to assigning phylogroup, 173 genomes within our dataset shared less than 95% ANI with any phylogroup reference genome, indicating that they were either misclassified at the time of depositing into GenBank as belonging to PSSC, or that they might represent new phylogroups. As a result, these genomes were left unassigned to a phylogroup. Eventually a curated set of 2,161 genomes were used, with 1,988 assigned to a phylogroup (Supplementary data 2.1.1 contains the accession numbers of these genomes and assigned phylogroups, along with ANI clusters and T3E gene content described below, see Supplementary data 2.1.2 for detailed description of data contained within).

In-silico PCR was performed with *in_silico_PCR* (69) allowing for one mismatch per primer. The identity of amplicons was confirmed by visually inspecting multiple sequence alignments performed with MAFFT, using Geneious 2019.1.3 (https://www.geneious.com). The amplification rate reported is the percentage of the 2,161 genomes that resulted in successful amplification of the target gene fragment. Primers included in this study are given in Table 2.1.1.

Pairwise ANI values for all genomes were computed with fastANI v1.33 (70). For each primer set with amplification rate greater than 50%, amplicon sequences were aligned using MAFFT version 7 (71) with the options ‘globalpair’ and a ‘maxiterate’ of 1000. For amplicon sequence similarity, pairwise hamming distances for the aligned sequences were then computed with the ‘DistanceMatrix’ function in the R package DECIPHER (72). To quantify the correlation between amplicon sequence distances and ANI of the source genomes, the mean squared deviation of amplicon sequence similarity from ANI was computed as the sum of squared distances between the two values for each genome pair. As ‘DistanceMatrix’ reports distance in the range of 0-1, ANI values were normalized to the same range by dividing by 100.

Table **2.1.1**: Primer sets used in this study

Primer set	Forward sequence (5'-3')	Reverse Sequence (5'-3')	Original primer names	Source
gapA-H	TCGARTGCACSGGBCTSTTCACC	GTGTGRTTGGCRTCGAARATCGA	gapA+312s/gapA−874ps	Hwang et al., 2005
gyrB-H	TCBGCRGCVGARGTSATCATGAC	TTGTCYTTGGTCTGSGAGCTGAA	gyrB+271ps/gyrB−1022ps	Hwang et al., 2005
CTS-H	CCTGRTCGCCAAGATGCCGAC	CGAAGATCACGGTGAACATGCTGG	gltA+513s/gltA−1130s	Hwang et al., 2005
rpoD-H	GYGAAGGCGARATYGRAATCG	CCGATGTTGCCTTCCTGGATCAG	rpoD+364s/rpoD−1222ps	Hwang et al., 2005
CTS-SG	CCCGTCGAGCTGCCAATWCTGA	ATCTCGCACGGSGTRTTGAACATC	cts-Fs/cts-Rs	Sarkar and Guttman, 2004
gapA-SG	CGCCATYCGCAACCCG	CCCAYTCGTTGTCGTACCA	gapA-Fps/gapA-Rps	Sarkar and Guttman, 2004
gyrB-S	MGGCGGYAAGTTCGATGACAAYTC	TRATBKCAGTCARACCTTCRCGSGC	gyr-F/gyr-R	Sawada et al., 1999
rpoD-S	AAGCGTATCGAAGAAGGCATYCGTG	GGAACWKGCGCAGGAAGTCGGCACG	rpo-F/rpo-R	Sawada et al., 1999
acnB-Y	TGATGTTTGATGCCTTCCAC	TAAAACCCTTGGTGCTTTCG	acnB	Yan et al., 2008
Gap1-Y	CGTATCGCAATCAACGGTTT	GACTCTCCGTATCGCAATCA	gap-1	Yan et al., 2008
CTS-Y	WYTRACCGGYACMGTBGGY	TGGGCTGATSGGYTTRATYT	gltA	Yan et al., 2008
gyrB-Y	TGCVTTCGTTGARTACCTGA	ACGGAAGAAGAAGGTSAGCA	gyrB	Yan et al., 2008
PGI-Y	GCGTACTACCGYAMYCCBTC	CCACATMGGRAARATRTTYT	pgi	Yan et al., 2008
rpoD-Y	GAAGGCATCCGTGAAGTGAT	GCCACGGTTGGTGTACTTCT	rpoD	Yan et al., 2008
rpoB-T	TGGCCGAGAACCAGTTCCGCGT	CGGCTTCGTCCAGCTTGTTCAG	LAPS/LAPS27	Tayeb et al., 2005
rpoD-P	TGAAGGCGARATCGAAATCGCCAA	YGCMGWCAGCTTYTGCTGGCA	PsrpoDFNP1/PsrpoDnprpcr1	Parkinson et al., 2010

Using QIIME2's v2022.2 feature-classifier (73), naïve Bayes classifiers were trained on the unaligned *in-silico* amplicon sequences generated above. Primer sequences were left untrimmed from amplicons. Classification models require taxonomic descriptions of known genomes for training, however, nomenclature in PSSC is largely inconsistent (74), which can significantly reduce the predictive power of classification models. We therefore implemented instead a strict hierarchical taxonomy based on ANI, generated by the hierarchical clustering algorithm utilized by LINbase (75). Briefly, a randomly selected genome in the set is assigned to clusters representing twenty ANI values from 80-99%, each given a numeric signature of '0'. All other genomes are iteratively assigned to clusters based on the closest already-assigned genome, and given the same numeric signature for all ANI values up to the point which the two genomes differ, wherein the numeric signature is iterated (e.g. if genome #2 shares 97.5% ANI with genome #1, genome #2 will be assigned to cluster '0' along with genome #1 for all ANI values except 98%, where it will be assigned a numeric signature of '1'). The resulting taxonomy file consisting of twenty taxonomic levels representing ANI values from 80-99% (see Supplementary data 2.1.1).

Reference sequences for T3E subfamilies included in the *Pseudomonas syringae* type III effector compendium (PsyTEC) (76) were aligned using MAFFT with default settings, and the alignments input into the HMMER v3.1b2 (77) function HMMbuild to generate HMM profiles. Using HMMsearch, these 77 HMMs were run on the set of 2,161 genomes and an e-value of 10^{-20} was used as the threshold for considering a subfamily to be present in a genome.

Results and Discussion

As all primers used in this study were designed to work broadly on strains within the species complex, we first tested amplification rate of each primer set. Surprisingly, when tested on a comprehensive set of genomes representing the full known diversity of PSSC, 7/16 primer sets tested had an amplification rate of less than 50% (Fig. 2.1.1a). For the remaining 9 primers, performance was substantially better, with amplification rates ranging from 91.37% (rpoD-P) to 100% (rpoD-H). In all cases with successful amplification, only a single amplicon representing the desired region was generated, with no non-specific amplification detected. The large differences in amplification rate are likely due to the significant degeneracy built into the best performing primer sets (Table 1 and Supplementary data 2.2) and highlights the importance of considering the known diversity of PSSC when designing primers. It is worth noting, however, that the low *in-silico* amplification rates seen here do not necessarily translate to low amplification rates under laboratory conditions, as only a single mismatch was allowed per primer in our tests. In practice, successful amplification with two or more mismatches is not unreasonable to expect. Nonetheless, as primer sets with fewer mismatches are generally preferred, and the majority of the primer sets exhibiting low amplification rates targeted genes already targeted by better performing primer sets, primer sets with amplification rates <50% were removed from any further tests.

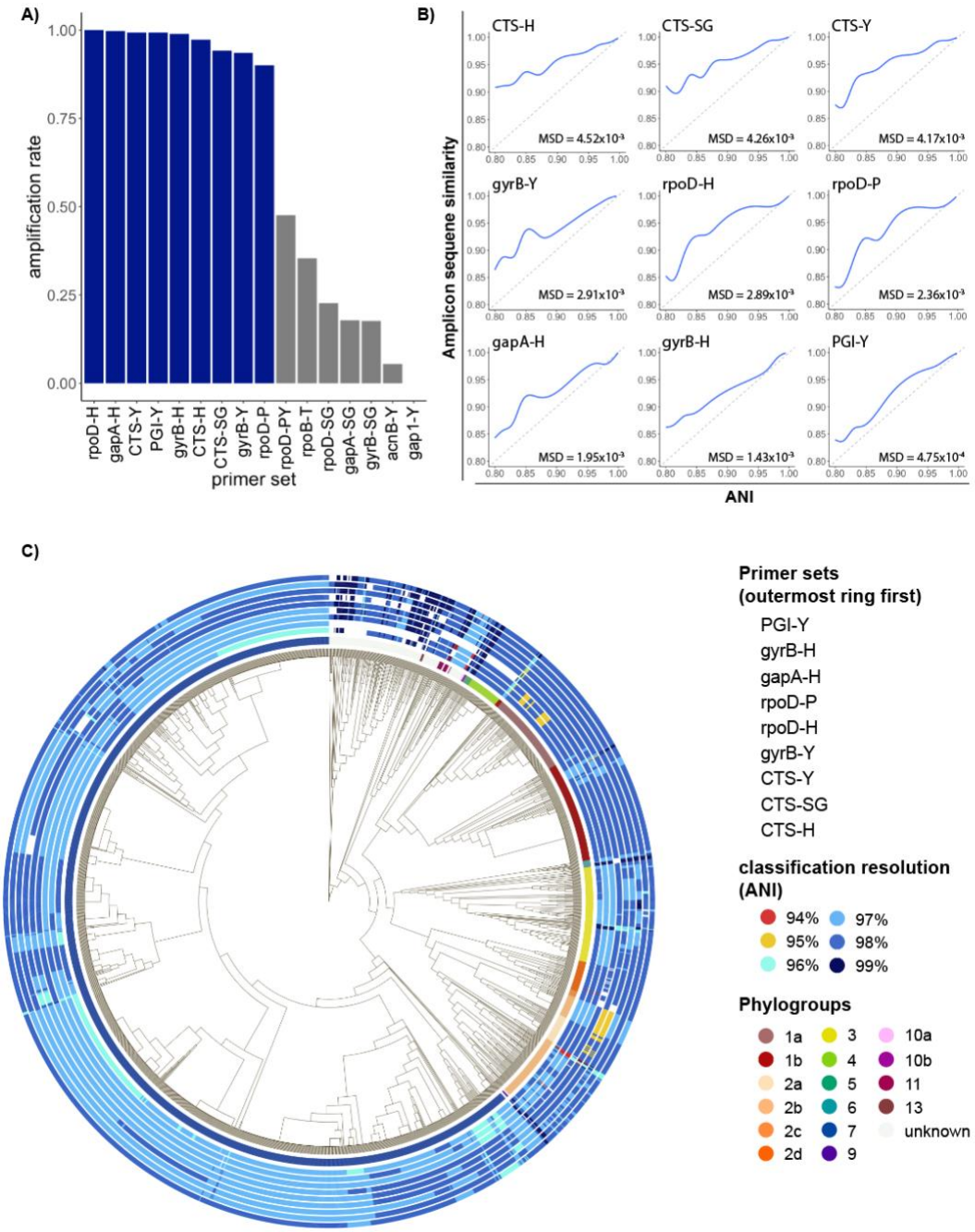


Figure **2.1.1**: Results of amplicon-based classification tests.

a) Proportion of genomes with successful amplification, allowing 1 mismatch per primer. Primer sets producing amplicons in more than 90% of genomes are highlighted in blue. Primer sets in grey were omitted from further analysis. **b**) Generalized additive models summarizing relationship between pairwise amplicon similarity and whole genome ANI. Mean squared deviation (MSD) of amplicon similarity from ANI is shown in the lower left corner, dashed grey line represents MSD = 0. **c)** Core genome phylogeny for all genomes used in this study. Innermost ring annotates phylogroups, while each outer ring represents the classification resolution obtained when amplicon sequences from each genome were identified with a Bayes classifier. White rectangles represent genomes from which in-silico amplification was unsuccessful.

When choosing a marker gene, or region of a marker gene, to use for identification purposes, an important consideration is the level of conservation found within the region, as regions that are too conserved result in reduced taxonomic resolution. To investigate the amount of conservation found within amplicons generated by each primer set, we compared pairwise amplicon similarity (represented by their Jaccard index) with ANI of the genomes from which the amplicons were derived (Fig. 2.1.1b). Of the primers tested, amplicons from citrate synthase showed the highest level of conservation, indicating they might not provide the best resolution when used for classification contrary to previous suggestions (Berge et al., 2014). On the other hand, amplicons generated by PGI-Y, targeting phosphoglucose isomerase, exhibited a mean squared deviation from ANI almost 10 times lower than any primer targeting CTS (Fig. 2.1.1b), indicating a very good congruence between the diversity at this locus and the one retrieved at the genome level.

To compare performance of primers in classifying individual strains throughout PSSC, classification models were trained on amplicons generated by each primer set, and then used to classify each strain in the training set. While using training data to evaluate the performance of classification models can at times lead to more accurate results than when testing with novel data, we observed that when classifying novel genomes (see Fig. 2.1.4a), the results are comparable to those discussed here, indicating no strong effect of using training data to evaluate our classification models. The relative performance of the classification models mirrored the amount of conservation observed above (Fig. 2.1.1b), although the practical differences in classification resolution were minimal (Table 2.1.2). Remarkably, every primer set allowed for classification beyond phylogroup level, with mean ANI of predicted clades ranging from 97.22% (CTS-H, CTS-Y & CTS-SG) to 97.93% (PGI-Y). Surprisingly, while the CTS gene has been suggested to be a particularly informative marker gene for PSSC (Berge et al., 2014), the three primers targeting this gene performed slightly below the other primers tested. In addition to training

classifiers from single amplicons, we also attempted to improve classification with concatenated amplicons generated from the three of the best performing primer sets (rpoD-P, PGI-Y, and gyrB-H), however mean ANI resolution from this classifier did not exceed 98% (data not shown), suggesting that the housekeeping genes used for amplicon sequencing have an inherent limit in the resolution they can provide. While mean performance of classification models suggest PGI-Y as the best primer set, it does not consider biases in the representation of each phylogroup in our dataset, and so we sought to analyze any discrepancies in primer performance among phylogroups (Fig. 2.1.1c).

Table **2.1.2**: Summary of classification test results

Primer set	Mean ANI resolution	SD
gyrB-H	97.5970987	0.59560943
CTS-SG	97.2226044	0.59653389
CTS-Y	97.2277597	0.65161671
gapA-H	97.5547818	0.5947916
rpoD-P	97.6722441	0.66541184
PGI-Y	97.9319664	0.43925883
gyrB-Y	97.5712166	0.57012883
rpoD-H	97.6307265	0.69616513
CTS-H	97.2234903	0.80563859

For strains belonging to Phylogroup 1, all primers performed well, amplifying every strain tested and classifying most to 98% ANI (Fig. 2.1.1c). An exception to the strong performance can be seen for 2 subclades which CTS-Y and CTS-SG were only able to classify at 95%. Overall, the best performing primers for phylogroup 1 were gapA-H and gyrB-H.

In Phylogroup 2, performance was more variable. Both rpoD-P and rpoD-H were unable to classify strains in a well sampled clade of phylogroup 2b above 95% ANI, and gyrB-Y was unable to amplify several strains within phylogroup 2d. As seen in phylogroup 1, the best performing primers for phylogroup 2 were gapA-H and gyrB-H. Phylogroup 3 strains were successfully amplified by every primer, with the exception of four strains that gyrB-H was unable to amplify from. Overall, PGI-Y and gyrB-Y were the best performing primers for strains in phylogroup 3. All primers performed equally well for phylogroup 4 strains, classifying to 98% ANI. gyrB-Y, however, failed to amplify from every strain in this phylogroup. Our dataset contained only nine phylogroup 5 strains, but every primer set was able to amplify and classify each one to 98-99% ANI. As with phylogroup 4, gyrB-Y was the only primer set unable to amplify and classify to 97-99% every strain in phylogroup 6. Within phylogroup 7, PGI-Y performed the best, classifying 89% (1,189/1,331) of strains to 98% ANI, while all other primers generally classified strains in this phylogroup to 97%. Phylogroups 9, 10, 11, and 13 were underrepresented in the dataset and so conclusions are difficult to draw about primer performance. Additionally, there were several strains not assigned to a phylogroup in our dataset. Among these strains, rpoD-H and gapA-H exhibited the highest amplification rates, and classification resolution ranging from 97-99% ANI. No single primer set universally outperformed the rest, and as such the suspected identity of an unknown isolate should be considered when choosing the appropriate set of primers to use for classification. However, PGI-Y, gapA-H, and gyrB-H generally performed well throughout the species complex and should be considered the default choices for highest classification resolution.

As classification based on single amplicon sequences resulted in fairly high genomic resolution (97-98% ANI), we sought to explore the possibility of predicting gene content of a target strain using gene content of predicted relatives. As type III effector proteins (T3Es) are important determinants of virulence in PSSC (68,78), we focused here on T3Es by first assessing

the prevalence of each T3E subfamily within genomic clusters sharing at least 98% ANI (Fig. 2.1.2). While there were several clusters that contained only a single genome (i.e., no genome in the dataset shared more than 98% ANI with them), even among better represented clusters, there was considerable similarity in T3E repertoires. This suggested that unknown isolates placed into these clusters should exhibit a predictable T3E repertoire. Perhaps not surprisingly, clusters representing phylogroups that contain most of the agricultural pathogens within PSSC exhibit many more T3Es as well as a greater diversity in repertoires between strains, indicated by the average Jaccard index within a given cluster (Fig. 2.1.2).

When the T3E repertoires of the 2,161 genomes were compared against the consensus repertoires (defined by taking the most common state of each T3E subfamily; absent or present) of their 98% ANI clusters, 75.3-100% of actual T3E states recapitulated the intra-cluster consensus (Fig. 2.1.3). It should be noted that T3E repertoires in our analysis were defined solely on presence or absence of gene subfamilies, and that even single amino acid changes within effector protein sequences can alter host compatibility. Therefore, T3E repertoire predictions presented here should be considered a useful starting point for generating hypotheses regarding host range of unknown isolates, and not any prediction of host range in itself.

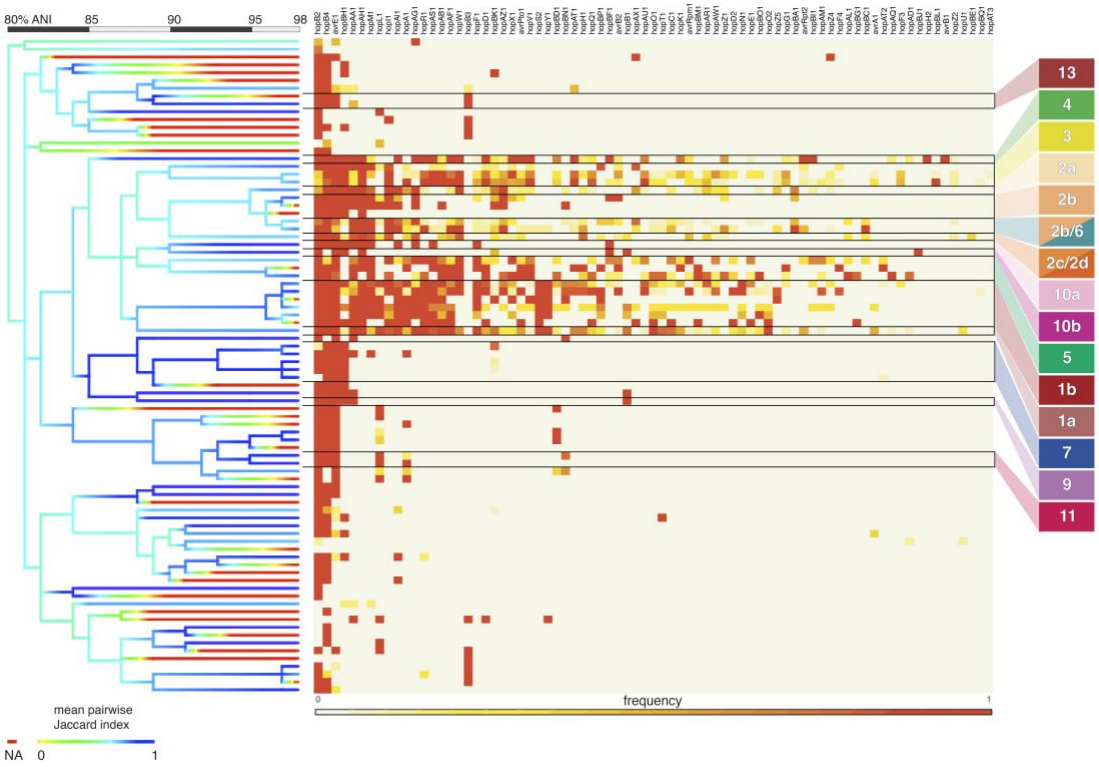


Figure **2.1.2**: Distribution of type III effector proteins throughout PSSC.

Heatmap shows frequency of each T3E subfamily, with each row as a cluster of genomes sharing 98% ANI. Black outlines and the associated labels on the right indicate phylogroups found in each row. Cladogram on left represents ANI-based clusters of genomes from 80-98% ANI and is colored by mean Jaccard index of genome pairs within each branch. Red branches indicate singletons, for which Jaccard index could not be calculated.

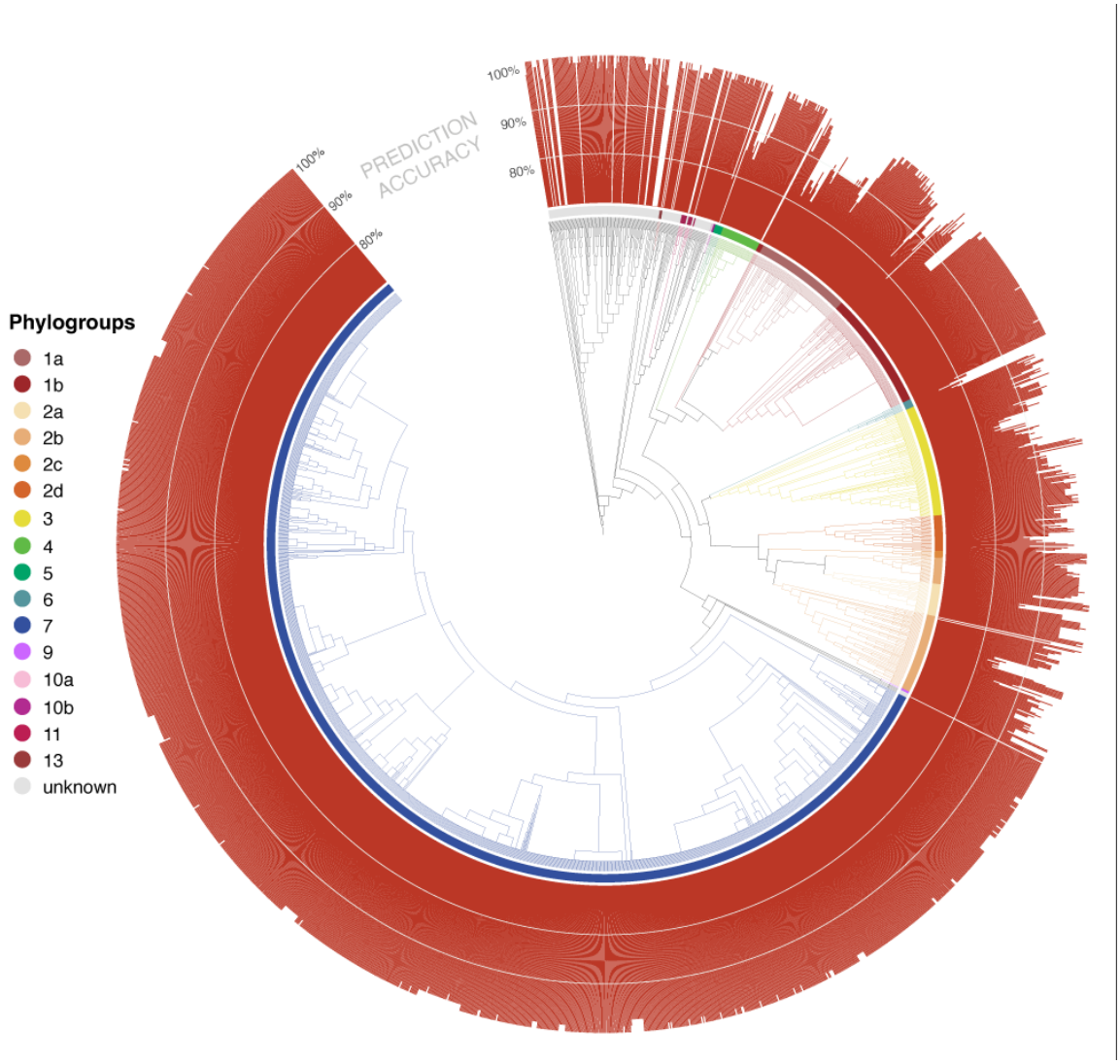


Figure **2.1.3**: Similarity between individual type III effector repertoires and the consensus repertoire at the 98% ANI level.

Shown is a approximately maximum-likelihood phylogenetic tree of PSSC built from nucleotide sequences of 120 core genes. Height of the red bars represent percent agreement between each genome's T3E repertoire and its 98% ANI cluster, scaled from 70-100%. Genomes with no bar represent singletons at the 98% level, and thus no consensus repertoire could be calculated. Innermost ring colors designate phylogroups, as seen in Fig. 2.1.1

To further test the feasibility of T3E repertoire prediction, 113 genomes not included in our initial training set underwent *in-silico* PCR for the nine primers tested above, classified using the trained naïve Bayes classifiers, and screened for T3Es. Actual T3E repertoires were then compared to the average repertoire within the predicted LIN group of the unknown strain, and prediction accuracy for the repertoire was assessed. While not every primer set was able to amplify from every genome (Fig. 2.1.4a), amplification rates were generally good at 92.9-100%. Classification of genomes from amplicon sequences resulted in the placement of most strains to 98% ANI or greater (Fig. 2.1.4a).

Overall accuracy of T3E repertoire predictions for all primers were nearly identical; each primer set allowed for a median prediction accuracy of 93.51% (Fig. 2.4b) but with individual prediction accuracies ranging widely from 64.9-100%. As some phylogroups are known to contain a greater number and diversity of effector proteins, we asked whether prediction accuracy varied significantly by phylogroup, and found a strong correlation between phylogroup and prediction accuracy (Fig. 2.1.4c). We also found that some Type 3 effectors were particularly difficult to predict (Fig. 2.1.4d) using our method. For example, for *HopA1*, a predicted absence of the subfamily was a false negative 62% of the time. Likewise, *AvrE1* was predicted to be present 100% of the time in our test genomes, resulting in false positives in 27% of genomes. These findings strongly suggest that at a minimum, any implementation of gene content prediction based on classification from marker gene sequences need to consider the appropriate testing errors to be of practical use.

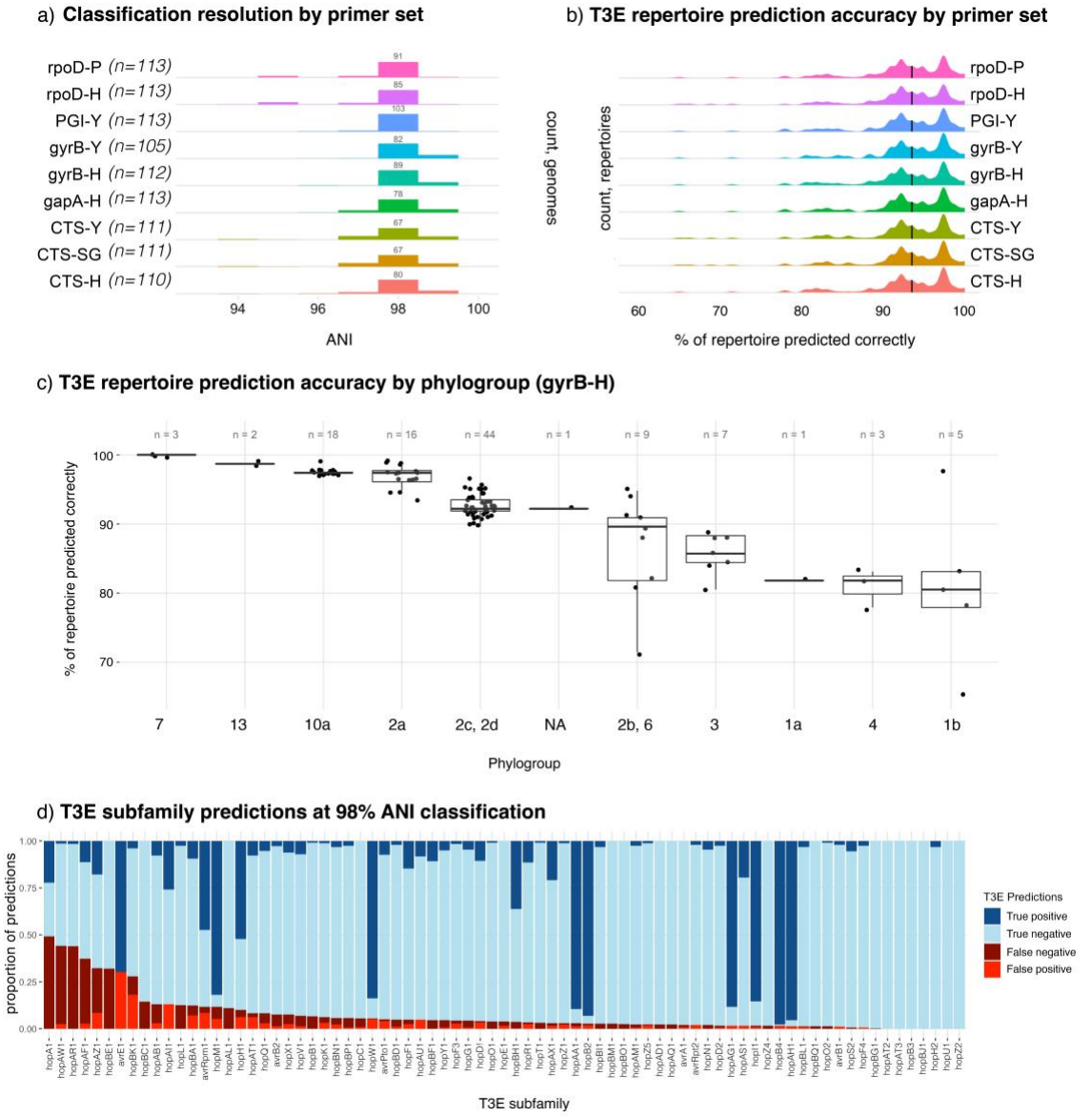


Figure **2.1.4**: Classification based on single marker genes allows for accurate prediction of type III effector repertoires.

a) density plots for classification resolution for each primer set. Number of successful in-silico amplifications from each primer set, out of 113, is indicated in parentheses. **b**) density plots for T3E repertoire prediction accuracy for each primer set. Prediction accuracy is defined as the percent of 77 effector protein subfamilies whose absence or presence in a genome was predicted correctly. Black vertical lines are median prediction accuracy. **c**) boxplots for T3E repertoire prediction accuracy by phylogroup of the classified genome. Each dot is a single genome, whose T3E repertoire was predicted based on classification with primer set gyrB-H. **d**) summary of T3E subfamily predictions in test genomes classified to at least 98% ANI, sorted from least to most correct predictions.

As more PSSC genomes are sequenced and deposited in public repositories, it is likely that our ability to predict T3E repertoires, as well as the presence of other important virulence factors, from amplicon sequencing will improve, due to increased sample sizes that more accurately reflect the true distribution of genes among closely related pathogens. This, coupled with research that indicates that host range can in part be inferred from virulence factors such as T3Es (67,79,80) suggests that amplicon sequencing remains a powerful method for studying disease dynamics, predicting pathogen spread, and rapidly detecting problematic PSSC strains.

Conclusion

In this study we set out to compare PCR primer sets designed to amplify broadly within PSSC. We found that there were significant differences in amplification rates that raise questions about the utility of some commonly used primers. However, we also found that classification resolution was relatively consistent between the primers tested, allowing placement of unknown genomes into clusters at the 98% ANI level.

The high resolution obtained from our classification models led us to investigate the potential of single amplicon sequences for prediction of type 3 effector protein subfamilies. we showed that with median accuracy of 93%, we were able to correctly predict effector repertoires of 113 recently sequenced PSSC strains, although the accuracy was dependent on phylogroup. These results highlight the importance of continued isolation and sequencing of plant pathogens as a source of data to be leveraged in the future for more efficient and informative screening assays. Based on our findings here, we currently recommend the primer sets gapA-H, gyrB-H, and PGI-Y for single amplicon sequence typing of isolated PSSC strains. In addition to these recommendations, we believe that the classifiers generated in the analysis of primer sets

presented here generate the highest resolution classification published to date and therefore should increase the utility of single amplicon sequencing used in the identification of PSSC isolates. Therefore, to make the classifiers and effector protein prediction more accessible to plant pathologists, in chapter 2.2, I describe these tools in more detail, as well as the creation of a web tool that allows rapid characterization of PSSC pathogens.

2.2 Description of syringae.org and the associated naïve Bayes classifiers for rapid *Pseudomonas syringae* isolate characterization

Abstract

The *Pseudomonas syringae* species complex (PSSC) is a diverse group of plant pathogens with a collective host range encompassing almost every food crop grown today. As a threat to global food security, rapid detection and characterization of epidemic and emerging pathogenic lineages is essential. However, phylogenetic identification is often complicated by an unclarified and ever-changing taxonomy, making practical use of available databases and the proper training of classifiers difficult. As such, while amplicon sequencing is a common method for routine identification of PSSC isolates, there is no efficient method for accurate classification based on this data. Here we present a suite of five Naïve bayes classifiers for PCR primer sets widely used for PSSC identification, trained on in-silico amplicon data from 2,161 published PSSC genomes using the life identification number (LIN) hierarchical clustering algorithm in place of traditional Linnaean taxonomy. Additionally, we include a dataset for translating classification results back into traditional taxonomic nomenclature (i.e. species, phylogroup, pathovar), and for predicting virulence factor repertoires. To aid in accessibility of our classifiers and associated datasets, we built syringae.org, allowing for rapid characterization of PSSC isolates from marker gene sequences.

Background & Summary

The *Pseudomonas syringae* species complex (PSSC) has been co-evolving with plants since before the emergence of angiosperms (9), and has diversified into one of the most economically important groups of plant pathogens in the world, with a collective host range spanning every major food crop grown today (81). Critically, while there are many pathogens within PSSC, there is also a wide range of virulence exhibited throughout the species complex, including non-pathogenic plant epiphytes and strains isolated from rain and snowpack with no known pathogenicity to plants (82,83). The ability to discriminate between lineages within the PSSC and rapidly predict potential pathogenicity of novel lineages is crucial for preventing epidemic outbreaks (84), detecting emerging pathogenic strains (85), and untangling correlations between virulence factors carried by a pathogen, its host range, and its virulence (86). Although the efforts to catalog PSSC diversity and to understand the molecular determinants of virulence have yielded great insights into their ecology and behavior (87), currently there is no efficient way to leverage these insights to efficiently predict the identity and pathogenicity of newly discovered PSSC strains. This is especially true for those researchers or labs that do not specialize on PSSC.

A major barrier to the characterization of PSSC strains is the inconclusive or inaccurate taxonomic identities of published genomes. By one estimate, 42% of all published PSSC genomes are misclassified at the species level, based on analysis of phylogenetic relationships described by average nucleotide identity (ANI) and multi-locus sequence analysis (MLSA) (74). As genomes deposited in databases such as GenBank often serve as reference sequences for identification of isolates found on or near diseased plants, the high rate of misclassification has a direct, negative impact on our ability to efficiently recognize pathogenic lineages. Specifically, one of the most effective methods for classification of amplicon sequences is the naïve Bayes

classifier (88), which heavily relies on accurate training data to generate accurate predictions. The designation of 13 phylogroups based on MLST has clarified phylogenetic relationships within PSSC (89), however most published genomes aren't ascribed to a phylogroup in public databases and thus their use in classification is limited. While Berge et al. 2014 (89) have addressed this shortcoming by providing a reference database of phylogroup type strains allowing classification based on the CTS gene, there has since been no broader effort to make the classification process more efficient. Yet another approach to circumvent the inaccurate taxonomy at the species level while allowing for placement into clades below the species and phylogroup level is the clustering of genomes by ANI (Average Nucleotide Identity) (90). This approach assigns a life identification number (LIN) to each unique genome in a database, creating hierarchical clusters of genomes that largely recapitulate traditional phylogenetic clades described by the core genome, and allow for higher resolution than traditional PSSC taxonomy (Figure 2.2.1). Using LINs to generate an ANI-based taxonomy, we trained high resolution naïve Bayes classifiers for commonly used PCR primer sets targeting gyrB, gapA, CTS, rpoD, (63) and pgi (64) (Table 2.2.1). As our classifiers report identity based on a difficult to interpret LIN, we also generated a comprehensive key describing key features for each of the 2,161 reference genomes in our training set along with their assigned LIN. This key allows for translation from classifier output to prediction of species, pathovar and phylogroups. As the vast majority of the genomes used in this study had no phylogroup assigned, we also provide new phylogroups assignments for over 2,000 publicly available PSSC genomes, based on previously suggested methods (89).

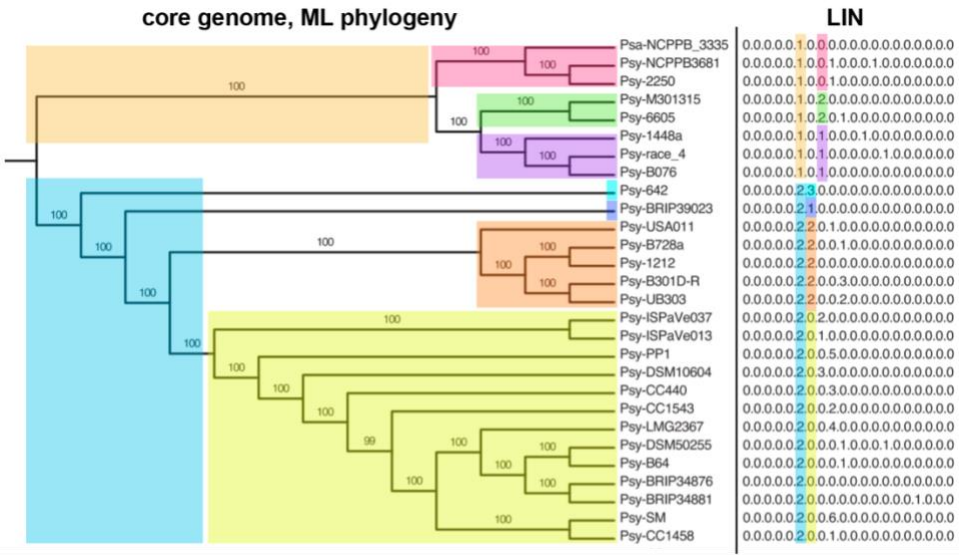


Figure **2.2.1**: LIN classification as an alternative to Linnean taxonomy

Comparison of clustering within PSSC that results from a maximum likelihood phylogenetic tree and LIN assigned based on ANI. Digits from left to right in each LIN correspond to inclusion of a strain in increasingly smaller clades within the phylogeny. Figure adapted from Vinatzer et al., 2017(90)

Table **2.2.1**: primer sets accepted by Syringae.org for isolate characterization

Target gene	**Forward sequence (5'-3')**	**Reverse Sequence (5'-3')**	**primer names**	**Source**
gapA	TCGARTGCACSGGBCTSTTCACC	GTGTGRTTGGCRTCGAARATCGA	gapA+312s/ gapA−874ps	Hwang et al., 2005
gyrB	TCBGCRGCVGARGTSATCATGAC	TTGTCYTTGGTCTGSGAGCTGAA	gyrB+271ps/ gyrB−1022ps	Hwang et al., 2005
CTS	CCTGRTCGCCAAGATGCCGAC	CGAAGATCACGGTGAACATGCTGG	gltA+513s/ gltA−1130s	Hwang et al., 2005
rpoD	GYGAAGGCGARATYGRAATCG	CCGATGTTGCCTTCCTGGATCAG	rpoD+364s/ rpoD−1222ps	Hwang et al., 2005
PGI	GCGTACTACCGYAMYCCBTC	CCACATMGGRAARATRTTYT	pgi	Yan et al., 2008

A second barrier to characterization of new PSSC isolates, even once identified, is the functional diversity exhibited throughout the species complex. Specifically, while PSSC strains are commonly grouped into different pathovars (pathogenic variants) based on host range and virulence, both of these traits can vary considerably among pathogenic strains belonging to the same pathovar, while strains belonging to different pathovars can nonetheless exhibit similar host ranges. These complex patterns stem, at least in part, from the formal definition of pathovar as 'a strain or set of strains with the same or similar characteristics, differentiated at infrasubspecific level from other strains of the same species or subspecies on the basis of distinctive pathogenicity to one or more plant hosts.'(91) This definition leaves room for broad interpretations of what should be considered a distinct pathovar. As such, some pathovars, such as pv. *avii*, have been delineated due to their ability to cause disease on a single host(92), while pathovars are defined based on their different host ranges among a small defined group of hosts (*P. savastanoi* pvs. *savastanoi*, *nerii*, *fraxini*, *mandevillae* and *retacarpa*)(93). Additionally, it has also been argued that pathogens sharing a wide common host range, regardless of a shared pathogenic potential for any single host, should also be considered as belonging to a single pathovar (94). Given the inconsistent criteria for delineating between pathovars, and recent evidence that host ranges in PSSC strains overlap with no discernable modularity (26), some groups have called into question the validity of pathovar designations for epidemiological and disease management purposes (95). Further, properly assigning a given isolate to an appropriate pathovar requires performing host range tests that are prohibitively laborious to many labs.

An alternative phylogenomic approach to predicting pathogenic potential would be beneficial, as others have demonstrated that comparative genomics can discriminate between strains known to have different host ranges (96) and correctly identify strains capable of infecting a given host (97). In both of the above cases, presence of virulence factors, particularly those associated with the type III secretion system (T3SS), were highly correlated with known

virulence patterns. Assuming T3SS effector proteins are conserved at some phylogenetic level, these results indicate that a phylogenomic signal may be present in PSSC that could be useful for assessing pathogenic potential without laborious experimental assays. In a recent contribution we showed the validity of such an approach by accurately predicting the presence of 77 type III effector (T3E) subfamilies in PSSC with a median accuracy of 80% using only single amplicon sequence data (98). We provide here a dataset for ANI based interpretation of taxonomy of PSCC, a HMMER-based survey of known virulence factors associated with the T3SS, type 3 effectors (T3E), and the Woody Host and *Pseudomonas* (WHOP) region associated with woody host infection (30) among our training set of genomes. With these data, we aim to provide a means for preliminary assessment and hypothesis generation regarding virulence traits from cost-effective amplicon sequencing data.

Methods

Reference PSSC genomes

All genome assemblies classified as '*Pseudomonas syringae* group' (taxid 136849) were downloaded from the GenBank via NCBI in November 2021, resulting in 2,468 RefSeq records recovered (99). Genomes were checked for completeness and assembly quality with BUSCO v5.3.1 using the pseudomonadales_odb10 lineage database (100), and genomes with a BUSCO score ≥ 99 were kept for further processing (Fig 2.2.2a). A CSV file (Supplementary data 2.1.1) summarizing each genome (and used as a backend database at www.syringae.org) was generated. Data included in this file are NCBI-submitted taxonomic data, type strain designations, phylogroups as assigned in this study, LIN clusters assigned for classification purposes, presence/absence of key virulence factors, and metadata found in each genome's Biosample record.

Assigning phylogroups to genomes

Phylogroup assignment of each genome was based on ANI shared with previously classified reference strains representing Phylogroups 1a,1b,2a,2b,2c,2d,3,4,5,6,7,9,10,11, and 13 (59) (Supplementary data 2.1.1). Reference strains for phylogroups 8 and 12 were not found among the 2,161 genomes characterized by SYRINGAE, either because they were not represented in the GenBank database or did not make it past the BUSCO quality check described above.

A genome was assigned to a given phylogroup if it was the most closely related to the reference strain for that phylogroup, based on ANI. To minimize inaccurate phylogroup assignments, 175 genomes sharing less than 95% ANI to any reference strain were left

unassigned. These genomes might reflect understudied groups within PSSC, or genomes mischaracterized as PSSC. Further work beyond the scope of this study would be needed to properly account for their true identity.

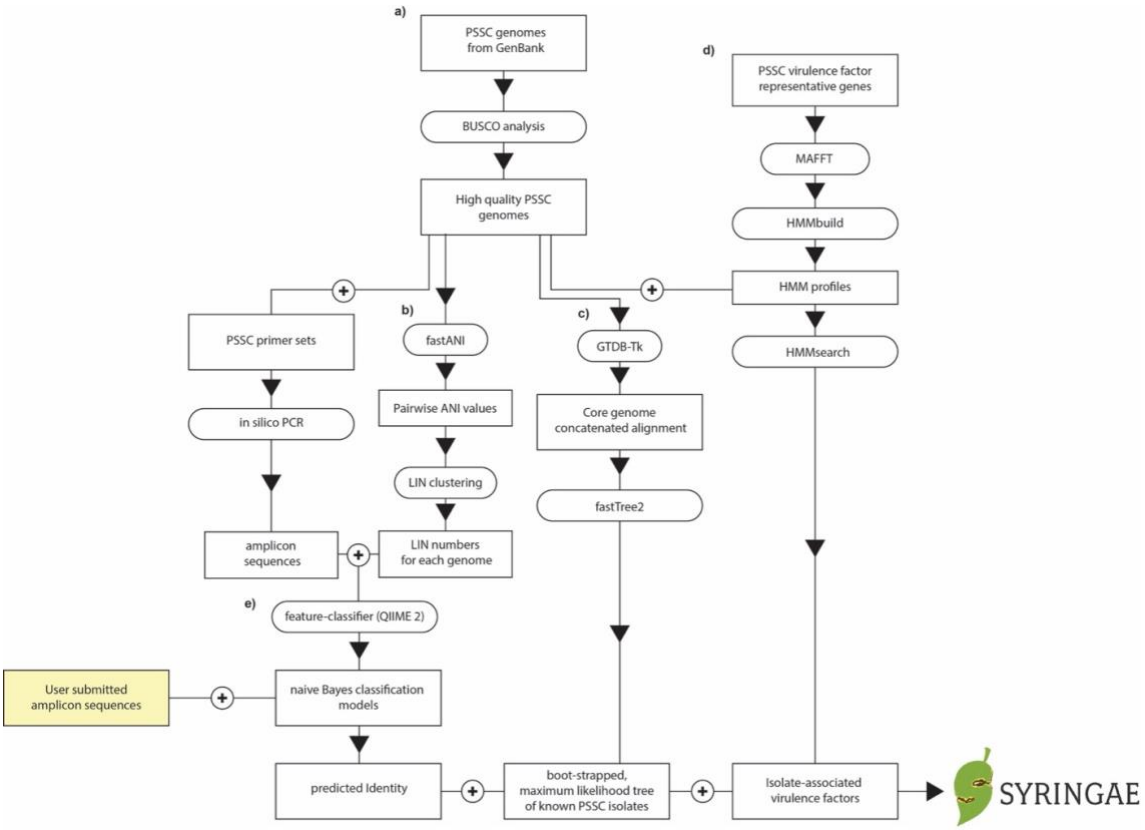


Figure **2.2.2**: Bioinformatic pipeline for syringae.org

Schematic of bioinformatic pipeline used for generating dataset and classifiers, including their incorporation into a web portal for accessing dataset and classifiers.

Assigning LIN clusters to genomes

A significant barrier to PSSC classification is unreliable and inconsistent taxonomic assignments. As such, SYRINGAE utilizes hierarchical clustering based on ANI values as an alternative to the Linnean taxonomy files typically used for Bayesian classification. Pairwise ANI between all genomes was calculated using fastANI v1.33 with default settings. Using the algorithm previously described, each genome was assigned to LIN cluster (fig 2.2.1). To describe the algorithm briefly, a random genome was designated as belonging to group '0' at every ANI bin (e.g. assuming ANI bins of 80, 90, and 95% would give a LIN number of '0.0.0'). Each subsequent randomly selected genome was assigned a LIN number based on the genome it has the highest ANI with among genomes already assigned a LIN number. If, for example, the second genome selected had an ANI of 92% with the first genome, its LIN number would be assigned as '0.0.1', as it meets the threshold for belonging to the same group as the first genome at the 80 and 90% ANI levels but differs from the first genome at the 95% level, and so a new group '1' is created for it. All genomes were sequentially assigned LIN numbers in this way. For SYINGAE, ANI bins at 1% increments between 80-99% were used.

A drawback to using LIN clustering for classification is that the LIN number assigned to a given genome is highly dependent on the order of genomes selected for clustering (i.e., unless the same set of genomes is used and the order that these genomes are selected for clustering is preserved, the genomes are assigned different LIN numbers every time). Thus, classification models built with our LIN 'taxonomy' will always return LIN numbers that can only be interpreted when used in conjunction with a database that explicitly describes the genome each LIN number represents. We overcome this limitation by first providing such an interpretive database in the file provided in Supplementary data 2.1.1, as well as by lowering the barrier to use with syringae.org, which uses metadata.csv to translate classification results automatically and the

display classification results to the user using traditional taxonomic nomenclature and an interactive phylogenetic tree.

Building the PSSC Phylogenetic tree

As a key component of visualizing and exploring the classifiers and dataset through the online portal hosted at www.syringae.org, a concatenated and masked gene alignment based on the core genome of PSSC was constructed using 120 bacterial marker genes within the BAC120 marker gene set with GTDB-TK 2.1.1 (using the 'identify' and 'align' commands) (101). From this alignment, FastTree2 (102) with default settings was used to construct an approximately maximum-likelihood phylogenetic tree from nucleotide sequences (fig 2.2.2c, Supplementary data 2.3).

Screening genomes for virulence factors of concern

Using the HMMER function HMMbuild {Citation}, we generated a single HMM file (supplementary data 2.4) containing HMMs for all virulence factors of concern (VFOC). As a first step, representative gene sequences were gathered as follows:

Canonical T3SS

nucleotide sequences from PSSC strains DC3000 (GCF_000007805.1) and B728a (GCF_000012245.1), as annotated by NCBI (and available in data record 'canonicalT3SS.fasta') were used as a database along with the 'annotate from database' tool within the Geneious prime 2019 software package (103), using 85% identity threshold for annotation of T3SS genes in all 2,161 genomes.

WHOP genes

previously annotated nucleotide sequences in strain NCPPB 3335 (30) were used as a database along with the 'annotate from database' tool within the Geneious prime 2019 software package (103), using 85% identity threshold for annotation of WHOP genes in all 2,161 genomes.

T3E genes

T3E nucleotide sequences contained in PsyTEC (76) were obtained from David Guttman on September 17th, 2021.

For each gene, nucleotide sequences were aligned with MAFFT (71) using default settings, and alignments were used as input for creation of HHM files using HMMER v3.3.2 (104) (fig 2.2.2d).

The VFOCs detailed in Supplementary data 2.5.1 and 2.6.1 are those that were found using the above HMM models. HMMER output files were manually inspected and filtered by E-value, with an E-value < 10^{-20} were considered to be significant hits. In instances where two genes were identified as more than one virulence factor (a common occurrence among closely related T3E subfamilies), the identification with the lowest E-value was chosen as the official annotation.

PSSC primer set selection

Over the last two decades, several PCR primers have been developed, often as part of MLST schemes, for building evolutionary accurate phylogenies and aiding in classification of unknown isolates. More recently, there has been interest in utilizing single amplicon sequences for these purposes. To investigate which primer sets provide the most value for classification using a single amplicon, we conducted a short but thorough in-silico investigation of 16

commonly used primer sets (98). Briefly, we assessed in-silico amplification in 2,161 genomes representing the full diversity of the species complex as we currently know it, investigated concordance between pairwise amplicon distance and whole genome ANI, and assessed resolution of naïve Bayes classifiers trained from amplicon data, as well as the potential for functional prediction based on the classification results. The best performing primer sets based on these metrics are represented in the classifiers presented here (see Table 2.2.1 above).

Training Naïve Bayes classification models

For each marker gene, a classification model was trained using the scikit-learn v0.24.1 feature-classifier plugin in QIIME 2 v2020.8.0. Training naïve Bayes classifiers requires both a list of sequences, and an associated taxonomy file for each sequence (typically in the format 'Order_Pseudomonadales; Family_Pseudomonadaceae; Genus_Pseudomonas; Species_syringae;'). LIN numbers assigned to each genome were used to construct a hierarchical taxonomy, with ANI bins within each LIN number acting as taxa levels, and groups acting as individual taxa (e.g., a taxonomy format of '80%_0; 90%_0; 95%_1) (fig 2.2.2e). Training data associated with all classifiers can be found in Supplementary data 2.7-2.12. Classifiers can be found in Supplementary data 2.13-2.17

Technical Validation

Genome records used for creation of this dataset were validated for assembly quality using BUSCO (105) and all genomes with a reported BUSO score < 99 were removed from the dataset. Accuracy of the classification models and functional predictions were investigated and published separately (98). Beyond the T3SS, T3E, and WHOP genes, which were annotated using HMM models built for this study, all gene annotations were taken directly from the NCBI Prokaryotic Genome Annotation Pipeline.

Syringae.org Usage Notes

The data described in this work are available as classifiers and accompanying datasets for incorporation into bioinformatic workflows. To aid in useability, however, all datasets are also accessible through Syringae.org, which provides three main ways to leverage the underlying data described here.

Identify

The primary functionality on syringae.org is the rapid characterization of *Pseudomonas sp.* isolates from single amplicon sequences. Input sequence(s) accepted by Syringae.org must be untrimmed amplicon sequences generated from primer sets described in table 2.2.2, in FASTA or multiFASTA format.

Outputs displayed to the user include:

1) A list of genomes predicted to be most closely related to the unknown isolate

2) Phylogenetic classification for each query sequence is also displayed as a phylogenetic tree rooted at the most recent common ancestor of all genomes in the Syringae database predicted

to be closely related to the unknown isolate. By default, the tree only shows the most closely related genomes, but users can toggle the ANI threshold scale to lower values, to “zoom out” and display more distant relatives. This extra functionality allows the ability to visualize the predicted placement of the unknown isolate in a larger phylogenetic context. The predicted shared ANI with the unknown isolate is displayed as a radial bar chart around the perimeter of the tree (Figure 2.2.3, external ring).

3) For the currently displayed tree, the abundance of represented species, phylogroups, and pathovars are to the left of the tree. (Figure 2.2.3)

4) A summary of the virulence factors found among the most closely related of the unknown isolate genomes given the selected ANI threshold is available as a second tab in the results section. (Figure 2.2.4)

Explore

An interactive visualization tool for exploring the phylogenetic and genetic diversity of PSSC. Users can filter the 2,161-genome phylogenetic tree Syringae uses for visualizing classification data by taxa (phylogroup, species, and pathovar), and annotate up to six features, including multiple taxa and presence/absence of up to 3 NCBI-annotated genes (Figure 2.2.5)

Search

A search tool for quickly finding metadata for any genome in SYRINGAE’s database. Search results include a list of the closest PSSC relatives in the dataset, as measured by fastANI (Figure 2.2.6).

Users can also search for any NCBI-annotated gene names or virulence factors annotated by us for Syringae. Search results include a list of all genomes in our database carrying the gene of interest, as well as each unique protein accession number associated with the gene name found within the species complex.

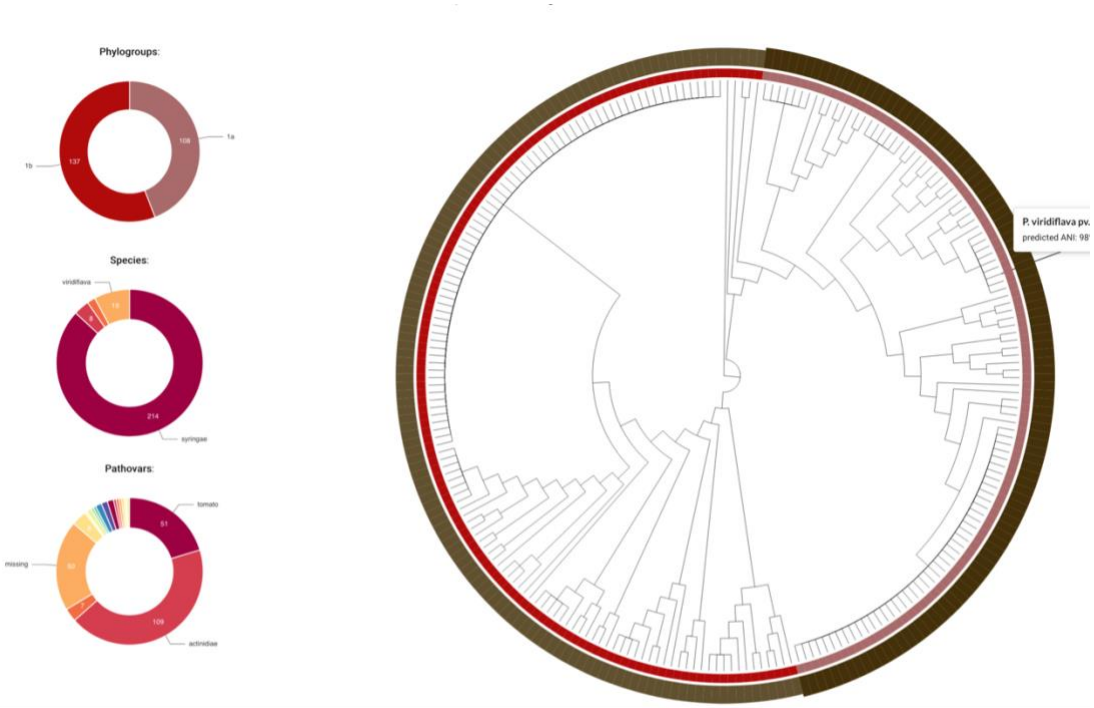


Figure **2.2.3**: A screenshot of classification results provided by Syringae.org

The tree is rooted at the most recent common ancestor of all genomes within the LIN cluster the unknown isolate has been placed into. The exterior ring shows the predicted ANI similarity between the unknown isolate and each reference genome. The interior ring shows phylogroup assigned to each reference genome. Relative abundance of phylogroups, species, and pathovars found within the unknown isolates predicted LIN cluster are shown to the left of the tree.

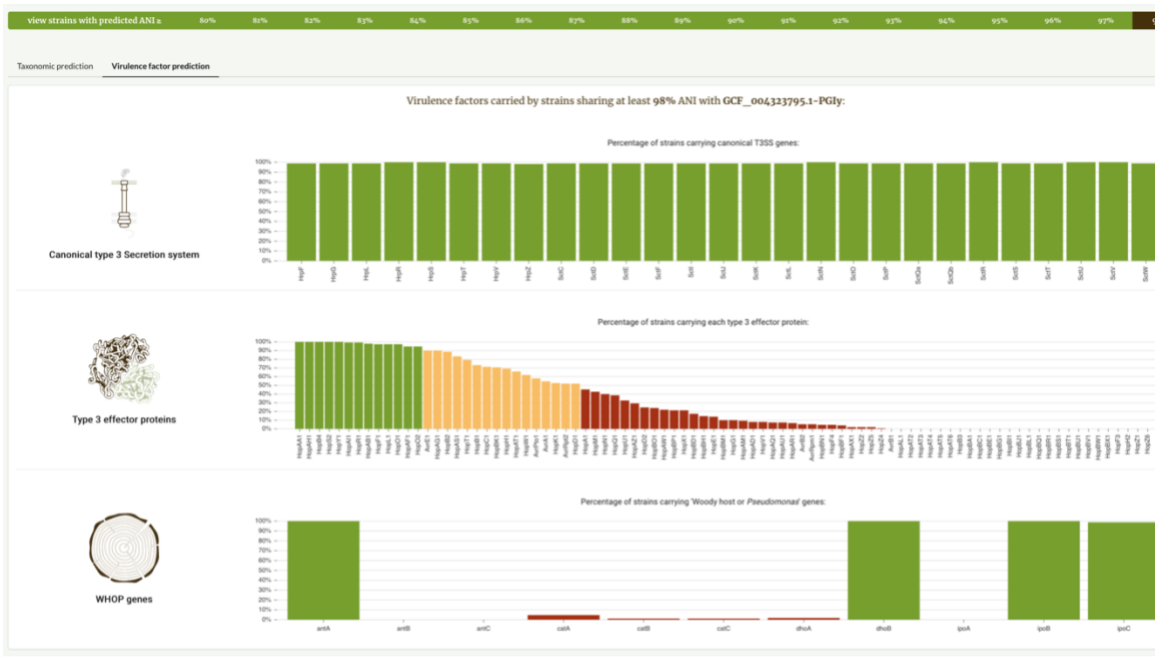


Figure **2.2.4**: A screenshot of virulence factor prediction provided by Syringae.org

The proportion of genomes within an unknown isolate's predicted LIN cluster that carry canonical type III secretion system genes, type III effectors, and WHOP genes are displayed. Green, yellow, and red bars denote virulence factors found in >90%, >50%, and <50% of related genomes, respectively.

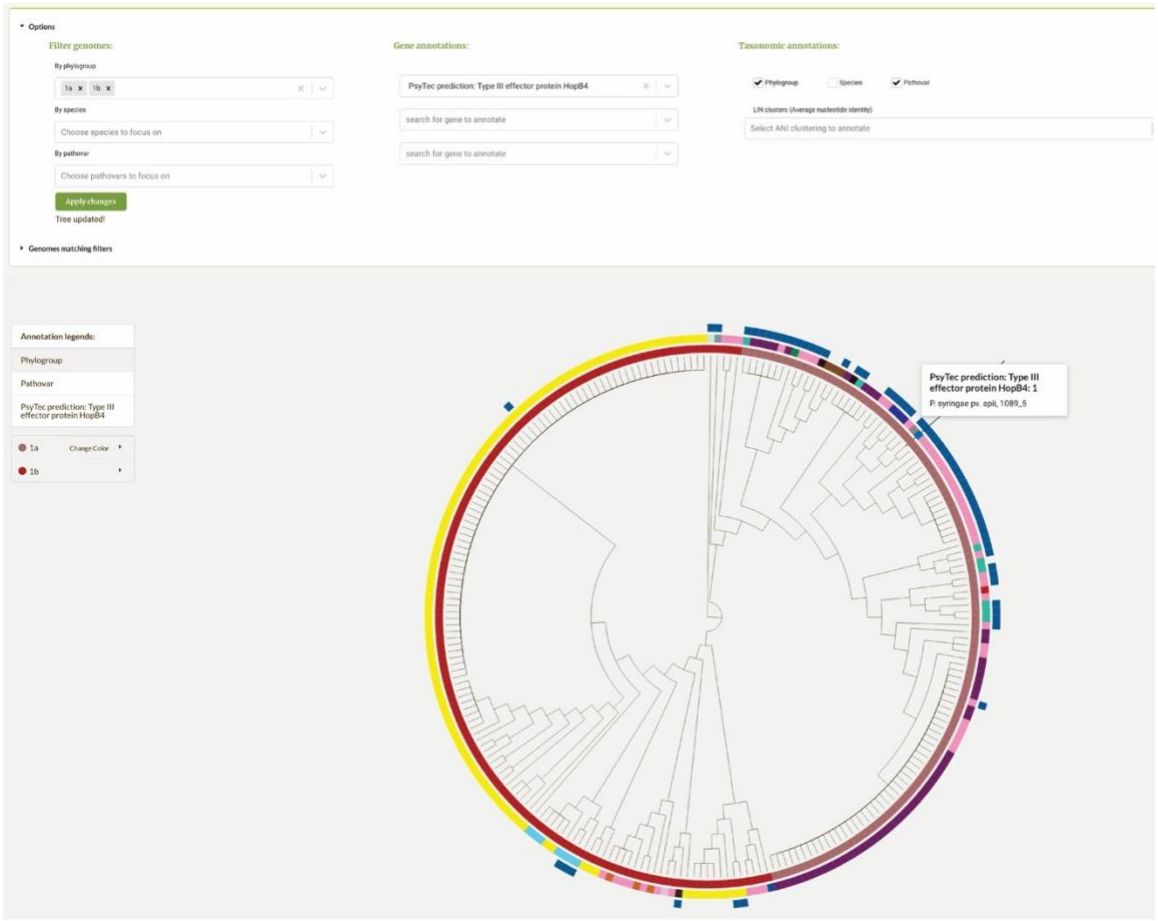


Figure **2.2.5**: A screenshot of the Explore functionality on Syringae.org

The form at top allows for filtering of genomes to display by taxa, and annotation by up to 3 simultaneous genes and 4 taxa. Annotation rings surrounding the phylogenetic tree are based on user-selected annotations; in this case showing, from innermost ring outward, phylogroups, pathovars, and absence/presence of effector protein subfamily HopB4.

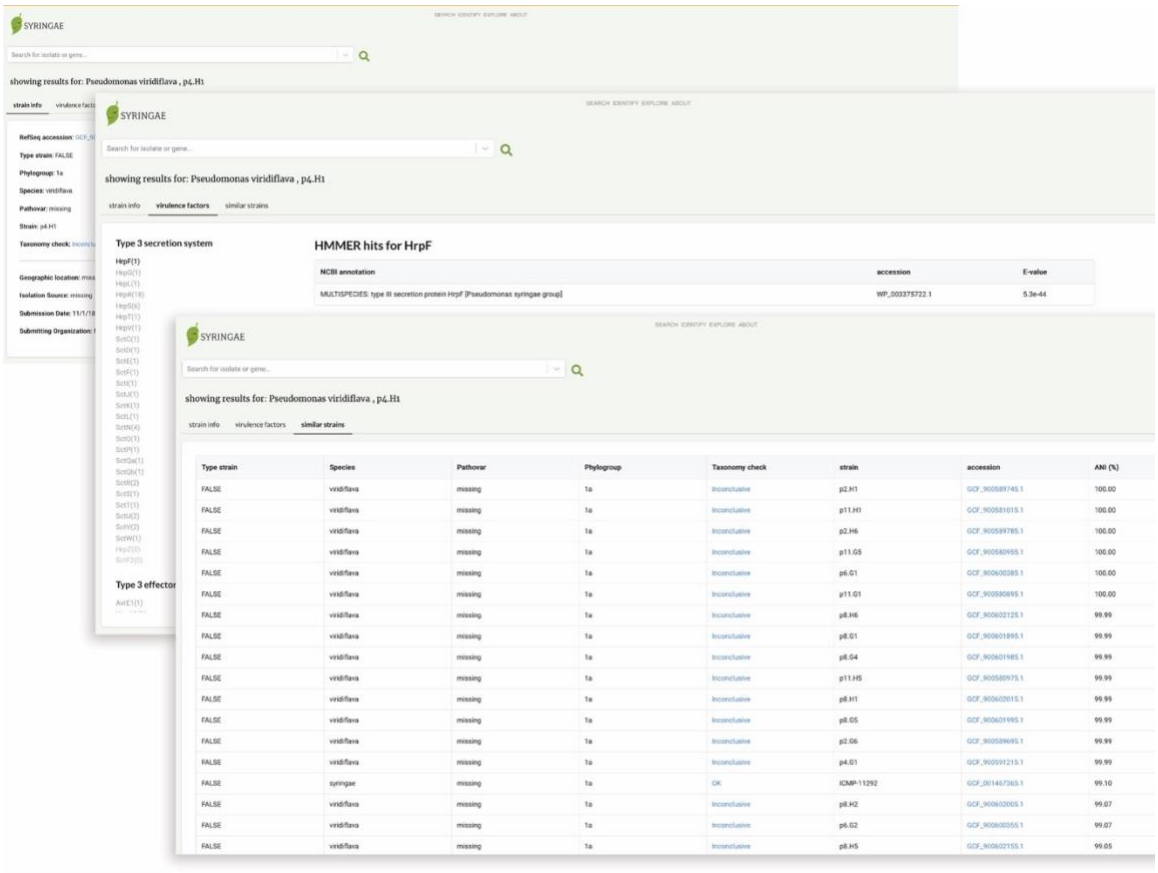


Figure **2.2.6**: A screenshot of search results for a PSSC genome on Syringae.org

displaying from top to bottom: metadata associated with the genome, predicted virulence factors, and most closely related genomes.

Summary

In this chapter I described the development of classification models, the curating of datasets for the prediction of virulence factors, and ultimately the building of a website for accessible characterization of PSSC isolates from amplicon sequences in syringae.org. The intended user base for syringae.org are researchers performing routine surveillance and diagnostics of suspected PSSC pathogenic lineages, with the goal of increasing the efficiency and utility of commonly used PCR primer sets. While the web tool and underlying bioinformatic pipelines will allow researchers to better monitor the spread and emergence of pathogenic lineages as they cause disease on plants, however, an ideal outcome for PSSC from a farmer's perspective would be to prevent disease before it occurs. With this in mind, in Chapter 3 I focus on how PSSC increases its fitness and competitive edge on the surface of the leaf upon immigration to the plant by focusing on the arsenal of anticompetitor toxins it uses to kill competing bacteria.

Code Availability

All scripts used in the generation of classifiers and dataset, as well as source code for the web app hosted at syringae.org are available on GitHub at https://github.com/cwf30/SYRINGAE (106). To aid in reproducibility, a conda environment YAML file (SYRINGAE_env.yml) containing all required packages and dependancies to run the pipeline described here, and a readme file outlining scripts used (README.txt) are also provided.

Supplementary data

All supplementary data are deposited at Zenodo (107). Data for Chapter 2 include:

Supplementary data 2.1.1

CSV file describing 2,161 PSSC genomes used in the evaluation of PCR primers in chapter 2.1 and used as reference genomes for classification of isolates at syringae.org in chapter 2.2. file contains RefSeq accession numbers for each genome, T3E family repertoires, and phylogroup and ANI clusters assigned to each genome (metadata.csv)

Supplementary data 2.1.2

Table describing columns contained in Supplementary data 2.1.1

Supplementary data 2.2

Excel file with in-silico amplification rates for all 16 primer sets evaluated in chapter 2.1, both overall and by phylogroup.

Supplementary data 2.3

Newick tree file containing the phylogenetic tree of 2,161 PSSC genomes used throughout chapters 2 and 3, with bootstrap values.

Supplementary data 2.4

HMM file containing hidden Markov models for all VFOCs identified in chapter 2.1 and 2.2.

Supplementary data 2.5.1

JSON file describing all canonical T3SS, T3E, and WHOP genes in the 2,161 genomes used in chapter 2.1 and 2.2, as detected by HMMER, with genome accessions as top-level keys. Additional keys found in each file are described in Table 2.2.3.

Supplementary data 2.5.2

Table containing description of JSON structure for Supplementary data 2.5.1

Supplementary data 2.6.1

JSON file describing all canonical T3SS, T3E, and WHOP genes in the 2,161 genomes used in chapter 2.1 and 2.2, as detected by HMMER, with protein accessions as top-level keys. Additional keys found in each file are described in Table 2.2.3.

Supplementary data 2.6.2

Table containing description of JSON structure for Supplementary data 2.6.1

Supplementary data 2.7

TSV file containing LIN numbers associated with each reference genome, used as input for training classifiers.

Supplementary data 2.8

FASTA file containing in-silico amplicons generated from primer set CTS_Hwang, used as input for training a Naïve Bayes classifier.

Supplementary data 2.9

FASTA file containing in-silico amplicons generated from primer set gapA_Hwang, used as input for training a Naïve Bayes classifier.

Supplementary data 2.10

FASTA file containing in-silico amplicons generated from primer set gyrB_Hwang, used as input for training a Naïve Bayes classifier.

Supplementary data 2.11

FASTA file containing in-silico amplicons generated from primer set pgi_Yan, used as input for training a Naïve Bayes classifier.

Supplementary data 2.12

FASTA file containing in-silico amplicons generated from primer set rpoD_Hwang, used as input for training a Naïve Bayes classifier.

Supplementary data 2.13

Naïve Bayes classifier for primer set CTS_Hwang

Supplementary data 2.14

Naïve Bayes classifier for primer set gapA_Hwang

Supplementary data 2.15

Naïve Bayes classifier for primer set gyrB_Hwang

Supplementary data 2.16

Naïve Bayes classifier for primer set pgi_Yan

Supplementary data 2.17

Naïve Bayes classifier for primer set rpoD_Hwang

Chapter 3

Genomic insights into tailocin diversity and evolution within the *Pseudomonas syringae* species complex

Abstract

To better understand the full diversity of tail fibers associated with tailocins in the *Pseudomonas syringae* species complex, we screened 2,161 publicly available genomes for their tailocin content, predicted protein structures that represent the diversity of fibers, and investigated forces possibly driving the distribution of fibers throughout the species complex. We uncovered previously unreported tail fiber clades formed through both vertical inheritance and large recombination events, and show for the first time, with structural and genetic evidence, that tailocins are associated with fibers homologous to both short and long tail fibers of bacteriophage. Furthermore, we find evidence that extant prophages in the species complex carry closely related tail fibers to those equipped by tailocins, and that distribution patterns of prophage fibers and closely related tailocin fibers are both correlated with the LPS synthesis gene *rfbD*. Our study provides important insight into the evolutionary forces that shape and maintain the diversity and distribution of tailocins within the *Pseudomonas syringae* species complex.

Introduction

It has been estimated that the global surface area of photosynthetic leaves is approximately 10^9 km^2, which harbors up to 10^{26} bacteria (108). Of these 10^{26} bacteria, a fraction will be foliar pathogens. In a variety of crops, it has been reported that potential foliar pathogens commonly make up anywhere from 5-60% of the epiphytic community, with the pathogenic *Pseudomonas syringae* being one of the most common inhabitants (14,109). For potential pathogens, the successful infection of a plant depends not only on the ability of the pathogen to evade plant defenses and manipulate plant metabolism (110) but also to proliferate on the leaf surface, thereby increasing the chances of infiltrating the leaf apoplast (14). Along with adaptation to frequent fluctuations in temperature, humidity, and exposure to ultra-violet (UV) radiation, the outcome of microbe-microbe interactions on the leaf surface has been consistently shown to impact pathogenic potential, leading to disease suppression when commensal epiphytes are able to reduce pathogenic populations (37,111,112). Competition between pathogens and other foliar epiphytes likely plays a large role in the distribution and dispersal of pathogenic populations, with implications for the prevalence of disease, as well as the rate of gene flow, adaptation, and the evolution of novel pathogenic strains (113,114). Therefore, gaining a better understanding of the mechanisms plant pathogens use to compete for limited resources in the phyllosphere would not only enhance our understanding of disease ecology but also have practical implications for the management of phytopathogens.

Phytopathogenic bacteria in the *Pseudomonas syringae* species complex (PSCC) are able to kill competitors with incredible selectivity through the production of bacteriophage derived toxins called tailocins (44,115). Many aspects of tailed phage (*Caudoviricetes*) and tailocin targeting are similar, including a highly specific and narrow killing spectra determined by tail fiber-cell surface receptor interactions. Specifically, in one of the most well studied phages, T4,

targeting and eventual infection of cells occurs through a multi-step process initiated by the binding of six long tail fibers to specific moieties in cell surface lipopolysaccharide (LPS) molecules of Gram-negative bacteria (116). Binding of long tail fibers is reversible, allowing the phage to 'walk' along the cell surface until a sufficient number of fibers are bound at the same time to trigger a conformational change in the phage baseplate that lowers the phage closer to the cell membrane (117). At this point, short tail fibers bind irreversibly to lipid A-inner core region of LPS, and further conformational changes in the baseplate trigger the piercing of the cell membrane (116). Thus, while phage tail fibers have no bactericidal activity themselves, they are essential for triggering a lethal chain reaction and ultimately determine a highly specific and narrow killing spectra (118,119). Tailocins carried by PSSC are also equipped with tail fibers that trigger the killing of the target cell by forming a pore in the cell's membranes through which cell contents pour out (120) and the proton-gradient is disrupted by an influx of protons from the environment (121). In contrast to bacteriophage, each tailocin particle is only associated with a single tail fiber gene (115).

Given the importance of tail fibers to the activity of tailocins, it is likely that they are under high selective pressure (122). This is true for bacteriophage, in which tail fibers show an increased rate of evolution compared to the rest of the genome (122) and there is frequent recombination of binding domains between phage families (123). It is likely that PSSC tailocins also experience heightened selection leading to diversification, as recombination in the tail fiber region has been detected that correlates with distinct killing spectra (124). It is widely believed that PSSC tailocin fibers bind to LPS (125–127), and a recent genome-wide association study expanded on this by showing that L- and D- rhamnose biosynthesis genes are highly correlated with the sensitivity of a cell to different tail fibers, suggesting that dTDP-4-dehydrorhamnose reductase (*rfbD)* is a particularly good genomic indicator of tailocin sensitivity (128). If rhamnose content in a cell's LPS determines its sensitivity to tailocins, it is reasonable to hypothesize that it

also imposes limitations on the tail fibers it can itself carry such that it does not kill itself. Thus, accounting for the full genetic and structural diversity of tail fibers associated with PSSC tailocins, and investigating compatibility between LPS structure and tailocin fiber types would not only provide insights into the ecological significance of tailocins to the species complex but could also suggest a framework for creating synthetic tail fibers capable of selectively targeting PSSC pathogens of interest (129).

Here, we report the results of a genetic and structural survey of tailocin-associated tail fibers found within PSSC. From a screen of 2,161 PSSC genomes, we found evidence for the circulation of three distinct recombinant tail fiber 'types' within PSSC, including one type not previously described. For all tailocin-associated fibers presented here, we show that tailocins were always found to carry only a single tail fiber type, with protein structure predictions of each fiber type suggesting one to be descended from long tail fibers, which in bacteriophage bind reversibly to the host cell, while the other two are descended from separate short tail fibers, which bind irreversibly, with possible implications for altering the killing efficiency of tailocins. Further, we found that all long tail fibers associated with PSSC tailocins are genetically and structurally related to tail fibers associated with intact prophages commonly found throughout the species complex. Finally, we show that the distribution of tailocin fiber types, as well as the distribution of related phage fibers, throughout PSSC is highly correlated with alleles of the LPS gene *rfbD*. This suggests the possibility that while *rfbD* predicts sensitivity to tailocins, it might also play a key role in maintaining the diversity of tailocin fibers in PSSC populations.

Methods

Genomes used in this study

All genomes labeled as '*Pseudomonas syringae* species group' were downloaded on November 17th, 2021 from NCBI. A total of 2,468 genomes were checked for completeness and assembly quality with BUSCO, and only genomes with a BUSCO score >= 99 were used in this study.

Phylogroups were assigned to all genomes as outlined in (1), with any genomes unassignable to a phylogroup due to sharing less than 95% ANI with any reference strain assumed to be outside PSSC. Although not considered as PSSC members here, these genomes were kept in our dataset, with the logic that they represent the outer boundaries of what might reasonably be considered to be the species group.

Genomic screens for genes of interest

Tail fibers

Strains in PSSC are thought to carry and produce a single tailocin, with each tailocin equipped with one of three distinct tail fibers that generally correlate with distinct killing spectra (124). A thorough search for all representatives of these three tail fibers within PSSC genomes was conducted using the following method. First, as tailocins in PSSC have been exclusively found directly downstream of the gene for Anthranilate synthase component 2 (*trpG*) (115), we extracted 25 kilobases in this region and annotated tail fiber genes contained within the extraction by first identifying open reading frames with GLIMMER3 (130) and then identifying tail fiber genes based on similarity to known tail fiber sequences (Table 3.1) using the 'Annotate from

Database' functionality in Geneious 2023.1.1 (131). 85% similarity was used a threshold for annotation.

Table **3.1**: Tailocin-associated tail fiber reference sequences used to annotate tail fiber genes

Type 1 tail fibers	
PSSC strain	**Protein accession**
CC1548	WP_024693886.1
B301D	WP_032656593.1
CC440	WP_024649699.1
USA007	WP_024658146.1
Type 2 tail fibers	
UB303	WP_024639976.1
Por 1_6	WP_005896706.1
CC1583	WP_024674765.1
CC1630	WP_005768002.1
Type 3 tail fibers	
UB246	WP_027901668.1, WP_027901669.1

Using the tail fiber genes annotated within Geneious, we then screened each genome for tailocin tail fibers that might be located outside of the expected genomic region. We built multiple sequence alignments for each of the three tail fiber types using MAFFT (71) under default settings, built HMM profiles with HHMbuild (HMM_1, HMM_2, and HMM_3, Supplementary data 3.1-3.3), and screened genomes with HMMscan in HMMER (77), with an E-value < 10^{-20} considered to be a significant hit. In instances where a single gene was identified as homologous to multiple tail fiber types, the prediction with the highest E-score was used to assign identity.

Tailocin- and prophage-associated genes

To help discriminate between tail fibers associated with tailocins or prophages, we relied on the presence of key genes in the same genomic region as the tail fibers. As mentioned above, tailocins in PSSC have previously only been found directly downstream of *trpG*, with *trpE* and *trpD* downstream of the tailocin region (115). Therefore, the presence of these genes suggest that tail fibers nearby belong to a tailocin. Additionally, while prophages require terminase and capsid genes for packaging and storing DNA (132), tailocins have no need for these genes. Thus, the presence of capsid and terminase genes suggests that any tail fibers nearby are associated with a prophage.

Genomic coding sequences (CDS) for each genome were downloaded from NCBI. Using an in-house script, 25 CDS upstream and downstream of each tail fiber identified above were searched for presence of the terms '*trpG*', '*trpE*', 'Anthranilate synthase component II', 'capsid', and 'terminase' in their gene names and descriptions. As only a small proportion of genomes in our dataset were fully circularized, it was expected that inevitably some of our indicator genes would not be on the same contig as the focal tail fiber, leading to false negatives. Therefore,

presence of any of the above genes were considered evidence of a tail fiber belonging the appropriate particle type (tailocin or prophage).

rfbD genes

Using the method described above, all CDS annotated as '*rfbD*' were extracted from genomes.

Tail fiber gene trees

Gene trees for tail fibers detected by each HMM used in this study were generated by first reducing each set of genes to non-redundant sequences. Amino acid sequences were then aligned using MAFFT under default settings, an alignment mask removing columns containing greater than 20% gaps was applied, and a phylogenetic tree was built using a Jukes-Cantor distance model and the neighbor-joining method. Both the alignment mask and phylogenetic tree were implemented with Geneious Prime 2023.1.1 (131).

AlphaFold protein structure prediction

Tail fibers equipped by both bacteriophage and tailocins are frequently homotrimeric (116,117,133). In accordance with this, we present predicted structures of all tail fibers as homotrimers.

Multimers were predicted with AlphaFold 2.3.0 (134), and the top ranked relaxed models were chosen for inclusion in the final manuscript. All predicted structures and amino acid sequences used for prediction are available in Supplementary data 3.4-3.13.

rfbD phylogenetic analysis

As the analysis presented was concerned with PSSC genomes only, any genomes for which we were unable to assign phylogroups were removed from the analysis. A gene tree for *rfbD* was built using the same method used to build the tail fiber gene trees above. *rfbD* genes found in all three genomes within phylogroup 13, which act as outgroup to all other phylogroups within PSSC, were used as an outgroup and subsequently removed from the tree for clarity in the final figure.

Results and Discussion

Diversity of R-type tailocin tail fibers within PSSC

Members of PSSC are commonly categorized into phylogroups, with a distinction between primary phylogroups (phylogroups1,2,3,4,5,6 and 10) (25) that are closely related to each other, and the more distant secondary phylogroups (7, 9, 11, and 13). While the secondary phylogroups do contain pathogens, most notably *P. viridiflava* from phylogroup 7 (135), the vast majority of highly virulent pathogens infecting common crops are found in the primary phylogroups (136). Among all genomes we surveyed, 62% were found to be carrying a tailocin directly downstream of the *trpG* gene, with no evidence of tailocins elsewhere in any PSSC strains. Tailocins were most common within the primary phylogroups (1,2,3,4,5,6 and 10), in which the proportion of genomes carrying tailocins increased to 84%. As the primary phylogroups are most highly represented in our dataset, contain the most agriculturally significant pathogens (25), and also carry tailocins most frequently, we decided to focus primarily on this group and only briefly examined patterns among secondary and unassigned phylogroup strains when relevant to patterns within the primary phylogroups.

We screened genomes for three previously described tail fiber types that are differentiated both by large recombination events in the coding sequences, and a marked difference in their killing spectra in tested against an array of PSSC strains (124). HMMs discussed here and subscripted with 1, 2, and 3 therefore correspond to three distinct clades of tail fibers described by Baltrus et al. (124). HMM_1 and HMM_2 returned 1,619 and 306 significant hits representing 290 and 63 nonredundant (unique) amino acid sequences, respectively. For tail fiber naming conventions, we strived to stay consistent with both the HMM that best matched the tail fiber and Baltrus et al.'s (124) existing numbering scheme for killing classes, such that e.g.,

tail fibers we designate as type 1 represent those detected by HMM_1 and also contain fibers in killing class 1.

Tailocin tail fibers found by HMM_1 form two monophyletic clades (Figure 3.1a), one of which contains fibers found exclusively in Phylogroup 7, and the other which contains those found throughout the primary phylogroups. Although these two clades of fibers are distinct evolutionarily, they appear to share a common ancestor, as they share strong homology throughout the length of the tail fibers (see pairwise alignment of type 1a and 1b, Figure 3.1b). Therefore, HMM_1 fibers associated with tailocins and found in the primary phylogroups and Phylogroup 7 are hereafter referred to as Type 1a and 1b, respectively.

Tail fibers identified with HMM_2 were comprised primarily of fibers closely related to those associated with killing class 2, with a small subclade showing evidence of a large recombination event replacing the latter half of the gene (see pairwise alignment between type 2 and type 3, Figure 3.1b). A representative fiber from this clade (found in strain CC1583) was among those investigated for killing activity by Baltrus et al. (124), but was dismissed as a truncated Type 2 fiber in the phylogenetic analysis. Nonetheless, CC1583's tailocin appeared to be functional and was described as belonging to killing class 3. We therefore refer to this clade of tail fibers as Type 3 (Figure 3.1a). Supplementary data 3.14 describes the full tail fiber contents of genomes screened in this study.

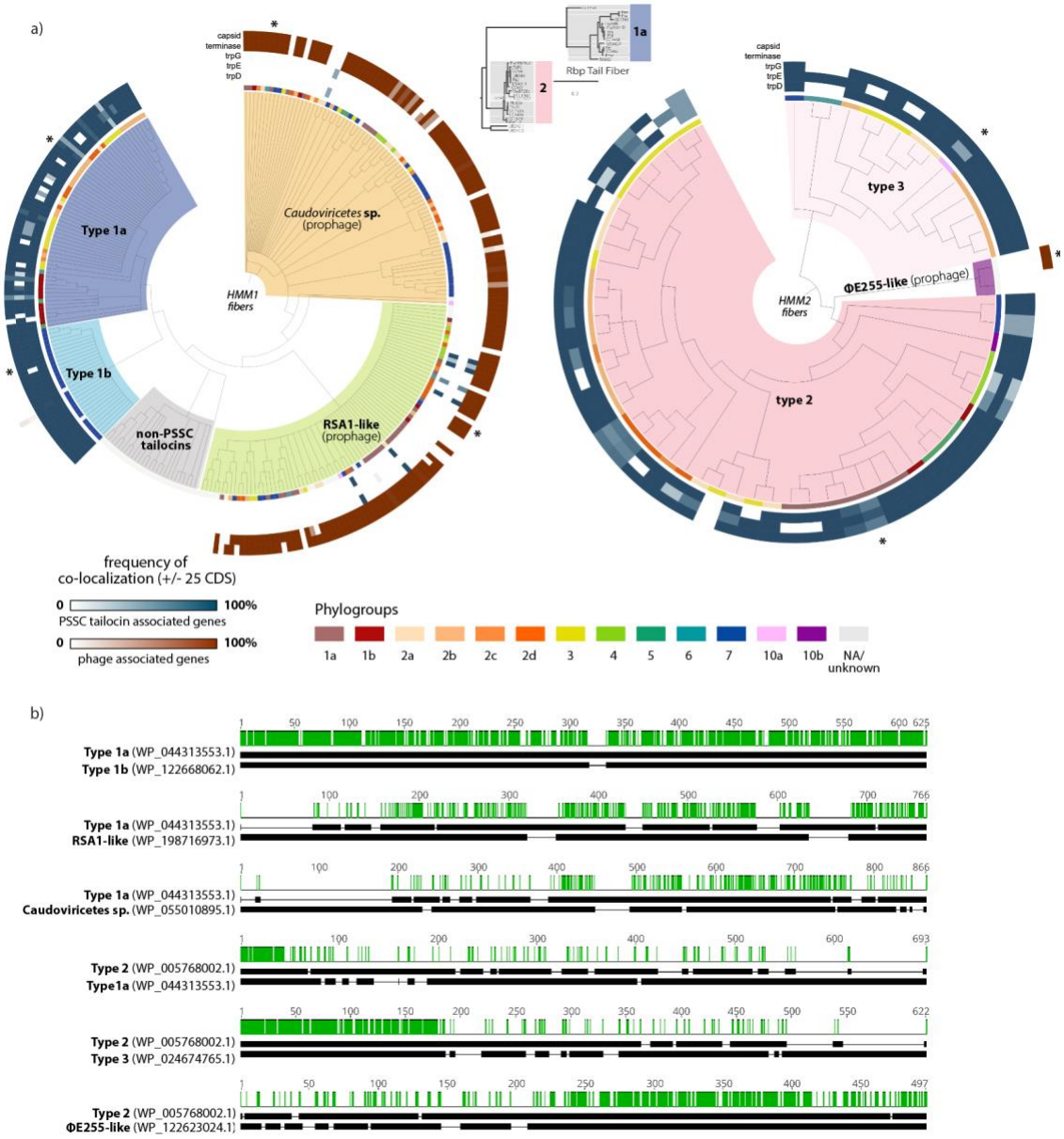


Figure **3.1**: Tail fibers associated with PSSC tailocins exhibited more diversity than previously reported and were closely related to prophage fibers found in the species complex.

a) gene trees for non-redundant tail fiber sequences detected by HMM_1 (left) and HMM_2 (right). Dendrograms are colored and labeled according to tail fiber types. Annotation rings from the inside out: 1) Phylogroup of the genomes containing each tail fiber; 2-4) blue heatmaps show the frequency of *trpD*, *trpE*, and *trpG* being detected within 25 CDS of the tail fiber; 5,6) red heatmaps show the frequency of terminase and capsid genes being detected within 25 CDS of the tail fiber. Asterisks denote tail fibers used for pairwise alignments in panel b. The central rectangular tree is a modified tail fiber gene tree from (124). Color and number annotations to the right of the tree correspond with the naming conventions used for fiber types in this study. **b)** pairwise amino acid sequence alignments for representative tail fibers, with green indicating conserved residues. Tail fiber group that the sequences are representing, along with the non-redundant protein accession numbers are provided to the left of each alignment.

Fibers identified by HMM_3 were found predominately in *Pseudomonas* strains outside any currently described phylogroup (Supplementary data 3.15), with thirteen instances in secondary phylogroups, and only a single instance within the primary phylogroups (ICMP-19589, phylogroup 3). It was determined that a proper accounting of the fibers detected by HMM_3 would require a broader screening of tailocins throughout the *Pseudomonas* genus, and further investigation of these fibers was deemed to be beyond the scope of this paper.

Tailocin fibers detected in PSSC (Type 1a, 1b, 2, and 3) all share strong homology of at least 60 n-terminal amino acid residues (Figure 3.1b), with total protein lengths ranging from ca. 400-800 residues. The conserved n-terminal region is thought to be important for attachment of phage tail fibers to the baseplate (124,133,137), and it is possible that 60 amino acids represents a minimally required domain for proper assembly and/or functioning of tail fibers with the PSSC tailocin. This is potentially significant for any future attempts to engineer synthetic versions of the PSSC tailocins, as it suggests the ability to fuse novel binding domains to the conserved n-terminal region to retarget killing spectra beyond any wild-type capabilities of the tailocin, as has been done with tailocins produced by *P. aeruginosa (118,129)*.

Extant prophages within PSSC genomes carry tail fibers closely related to Type 1 tailocin fibers

Our genomic screen revealed more diversity in tail fibers than previously expected; several genomes carried more than one tail fiber, despite inspection of tailocin regions in PSSC consistently revealing them to only carry a single fiber. We considered two possible explanations for this observation: 1) Additional tailocins were present in PSSC, or 2) there are prophages within PSSC that are equipped with tail fibers closely related to known tailocin tail fibers. In

attempt to clarify which tail fibers truly belonged to the tailocin of interest, and to determine the identity of any other tail fibers was to investigate the surrounding genes. Specifically, as the PSSC tailocin of interest is thought to always be located immediately adjacent to the trp operon, we looked for the presence of *trpG*, *trpE*, and *trpD* within +/-25 all tail fibers to indicate the fibers as being associated with the tailocin. Additionally, to assess whether any fibers might belong to prophages, we also searched this region for capsid and terminase genes – which are essential for tailed bacteriophages, but not for tailocins (138). We found that many tail fibers closely related to type 1 tailocins were adjacent to capsid and terminase genes, but none of the trp genes (red and blue annotation rings, Figure 3.1a).

Further analysis of the putative prophages with PHASTER suggested these are indeed intact prophages. As an illustrative example, the analysis of such a region found in *P. syringae* pv. tomato strain Pst-DC-98-1 being identified as an intact phage with a score of 91 (scores >90 are considered to be indicative of an intact phage by PHASTER), with 22/38 identified phage genes being most homologous to *Pseudomonas* phage Φ3. Specific identity varies considerably between phages characterized with PHASTER, particularly when focusing on the tail fiber gene specifically. Therefore, for practical purposes, we describe the origin of most of these fibers as simply belonging to the tailed bacteriophages, or *Caudoviricetes* phages (Figure 3.1a). An exception to this is a monophyletic clade of fibers that consistently BLAST as being closely related to RSA1 tail fibers, and thus this clade is described as RSA1-like (Figure 3.1a). RSA1-like fibers exhibit greater homology to both type 1a and 1b tailocin fibers than the aforementioned *Caudoviricetes* sp. fibers in all but the first 100 residues, although large deletions between RSA-1 like fibers and type 1 fibers are common (Figure 3.1b).

Nearly all fibers identified by HMM_2 were found to be associated with tailocins. Two putative prophages carrying partially homologous fibers were, however, found in two genomes – *Pseudomonas petrae* (GCF_900585705.1) and an additional strain (GCF_001698815.1) whose

identity is unclear, but shares 84% average nucleotide identity with *Pseudomonas foliumensis*, according to NCBI's taxonomy check. The regions surrounding these putative phage fibers were also determined by PHASTER to be intact prophages (score 140), with the fibers most closely related to phage ΦE255 (ΦE255-like, Figure 3.1a). Homology between ΦE255-like fibers and Type 2 fibers was restricted to the c-terminal region (Figure 3.1b). Given the consistent pattern of highly homologous c-terminal regions being found in circulating *Pseudomonas* phages, it is tempting to speculate that these phages were the tail fiber donors for type 1 and 2 c-terminal regions seen in PSSC tailocin fibers, although it is difficult if not impossible to confirm this.

Structural comparison of tail fibers and their functional implications

Given the extensive recombination evident among the tail fibers described above, we sought to better understand the impact such genetic diversity has had on protein structure. We used AlphaFold to predict structures for representatives from each clade of fibers and found that overall, type 2 and 3 tail fibers likely derive from short tail fibers while type 1a and 1b are likely long tail fibers with structural similarities to the *P. aeruginosa* R2 tailocin tail fiber (137). All fibers were modeled as homotrimers, in line with observations from various phage and tailocin tail fibers (116,139).

Type 1: the long tail fibers

The prototypical type 1 tail fiber is approximately 450 AA long, with an n-terminal baseplate attachment point, followed immediately by a large knob domain of unknown function. The majority of the tail fiber is comprised of three repeated units (Figure 3.2a), with

each repeat consisting of a shaft, an n-proximal knob (knob a) and a c-proximal knob (knob b) (Figure 3.2a). These repeats are not just structurally similar but share strong homology at the amino acid level across the length of the repeat (Figure 3.2b).

The general architecture of type 1 tailocin fibers bears a striking resemblance to the crystal structures for the c-terminal regions of R1 and R2 pyocin tail fibers, which contain a single pair of knobs followed by a c-terminus knob (133). Each of these knobs have been suggested to function as receptor binding domains, providing some evidence for the role of these structures in binding for type 1 fibers – although type 1 tail fibers and the R2 pyocin tail fiber from *P. aeruginosa* PA01 (PA0620) bare little similarity to each other at the amino acid level (Supplementary Figure 3.1).

Among type 1a fibers, amino acid conservation was the highest in the n-terminal attachment domain, and decreased steadily over the length of the fiber, with areas of low percent identity concentrated primarily in knobs found in repeats 2 and 3, as well as the c-terminal knob (fig. 3.2a). These results suggest that the distal binding domains might play a greater role in attachment to target cells than proximal knobs, and thus are under greater selection.

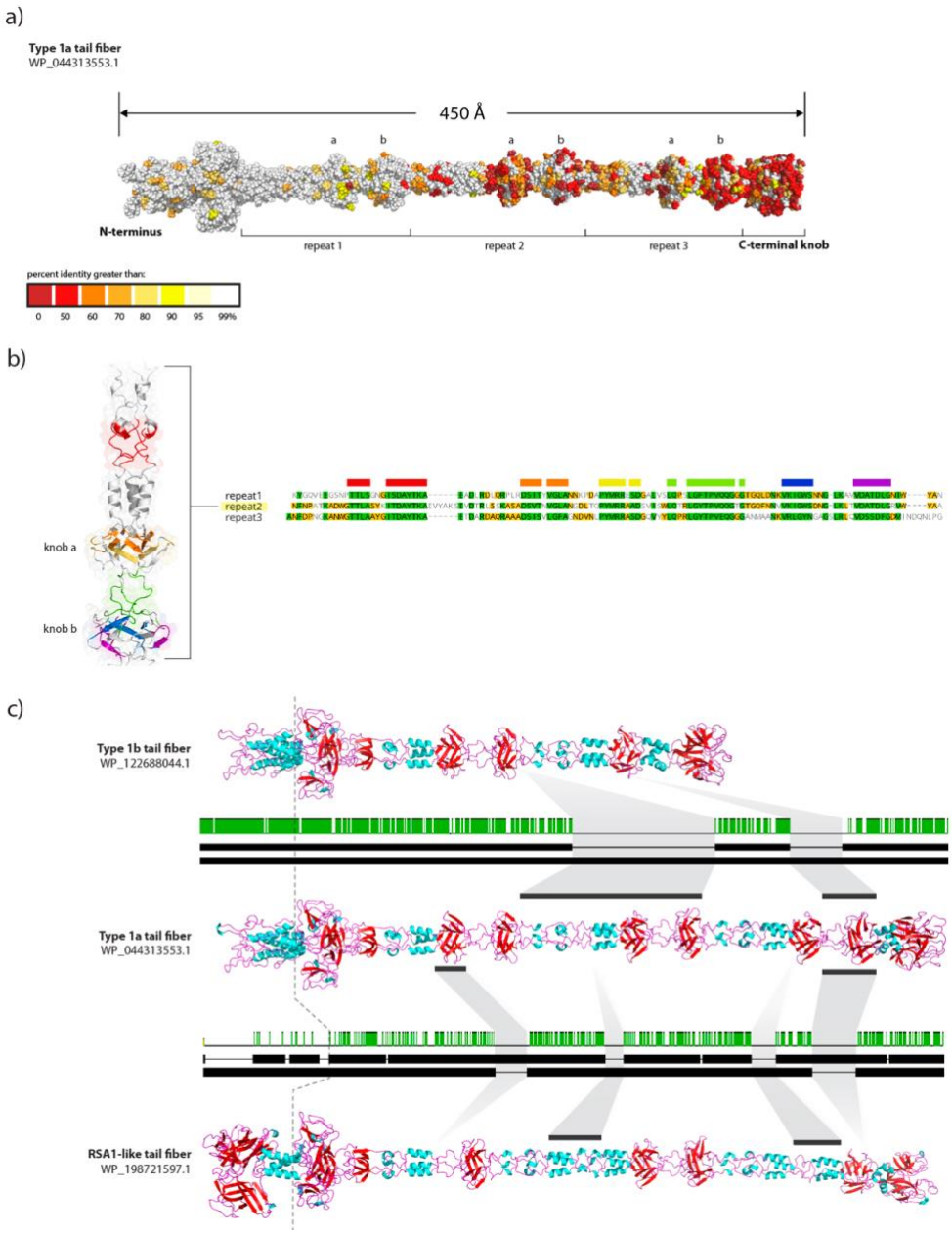


Figure **3.2**: Type 1 tail fibers and their phage relative display modular structure.

a) AlphaFold-predicted structure of a typical type 1a tail fiber, colored by conservation of residues. The n-terminus baseplate attachment point is far left, with Repeats 1, 2, and 3 labeled below the tail fiber. Knob domains a and b within each repeat re labeled above tail fiber. **b)** Amino acid multiple sequence alignment for repeats 1, 2, and 3 from the tail fiber in panel a. Cartoon ribbon structure for repeat 2 is to the left, colored in accordance with the highly conserved regions seen in the MSA. **c**) AlphaFold-predicted structures for tail fibers belonging to type 1b, 1a, and the RSA1-like phage. Between each pair of fibers is shown the accompanying amino acid pairwise alignment. Dotted line running through all fibers to the left denotes the large structural difference in the baseplate-attachment domain between tailocin-associated fibers (type 1a and 1b) and the RSA1-like fibers. The light grey annotations between fibers highlight large deletions present in some fibers tend to center around whole repeats or single knob domains.

The general architecture between type 1a, 1b, and RSA1-like fibers is strikingly similar, save for large deletions that commonly encompass entire knob domains. In an example of typical deletions found among tail fibers is highlighted in figure 3.2c, in which repeat 2 was deleted entirely in a small group of type 1b fibers, along with knob b in repeat 3. The result is a significantly shorter tail fiber with 3 internal knobs instead of the typical 6. Also shown in figure 3.2c is an instance where knob a in repeat 1 as well as knob b in repeat 3 of an RSA1-like fiber are missing, when compared to the prototypical type 1a fiber. The frequency of such large deletions (multiple sequence alignment of all type 1 tail fibers, supplementary Figures 3.2 and 3.3) strongly suggests there is significant functional importance for the predicted knob domains, and deletions of individual knobs might play a role in shaping killing spectra associated with a given tail fiber, although experimental work would be needed to determine what if any role these deletions play.

As noted above, the n-terminal region of RSA1-like fibers share no homology with tailocin-associated fibers (Figure 3.1b), and the predicted protein structures recapitulate these differences by exhibiting significant structural differences in this region (residues 1-156) compared to the otherwise similar type 1a fibers (Figure 3.2c). Functionally, this is a potentially significant difference, as it suggests that despite otherwise strong genetic and structural similarity, it is unlikely that RSA1-like prophage fibers are compatible with PSSC tailocin particles (i.e., a PSSC strain carrying both a tailocin and such a prophage is unlikely to be able to incorporate the prophage fibers into the structure of the tailocin).

Type 2: the short tail fibers

Structures of type 2 and 3 tail fibers are shorter and more globular than type 1 fibers (Figure 3.3a), suggesting both derive from short tail fibers. Reflecting the pattern of conservation seen in their amino acid sequences, a large segment (185AA) comprising the conserved baseplate attachment domain and a large knob of unknown function is shared in both fibers. The distal half of the fibers share no discernable similarities, and neither fiber appears to have any regions with particularly high sequence diversity, as opposed to type 1 fibers.

Sequences for both tail fiber types were scanned for known protein domains with InterProScan, and while no conserved domains were detected in the type 3 fibers, a large region in the distal half of the type 2 fibers was found to be homologous to the T4 short tail fiber binding domain (Figure 3.3b). This domain was also found in the ΦE255-like fibers, as was a DUF3751 domain at the baseplate attachment point of the fiber, which have been associated with both prophage and tailocin fibers in *P. putida* (44).

The significance of PSSC tailocins carrying either short or long tail fibers is unclear. In the infection cycle of bacteriophage, both of these fibers play distinct roles, with long tail fibers binding reversibly to the cell surface, allowing the phage to ‘walk’ around until a sufficient number of fibers have bound to trigger a conformational change in the baseplate (116). At this point, the phage lowers onto the cell surface and short tail fibers bind irreversibly to the target cell, allowing more efficient infection (116). Whether the nature of reversible vs. irreversible binding remains in tailocin fibers, or if these differences alter the killing efficiency of the particle remains to be seen.

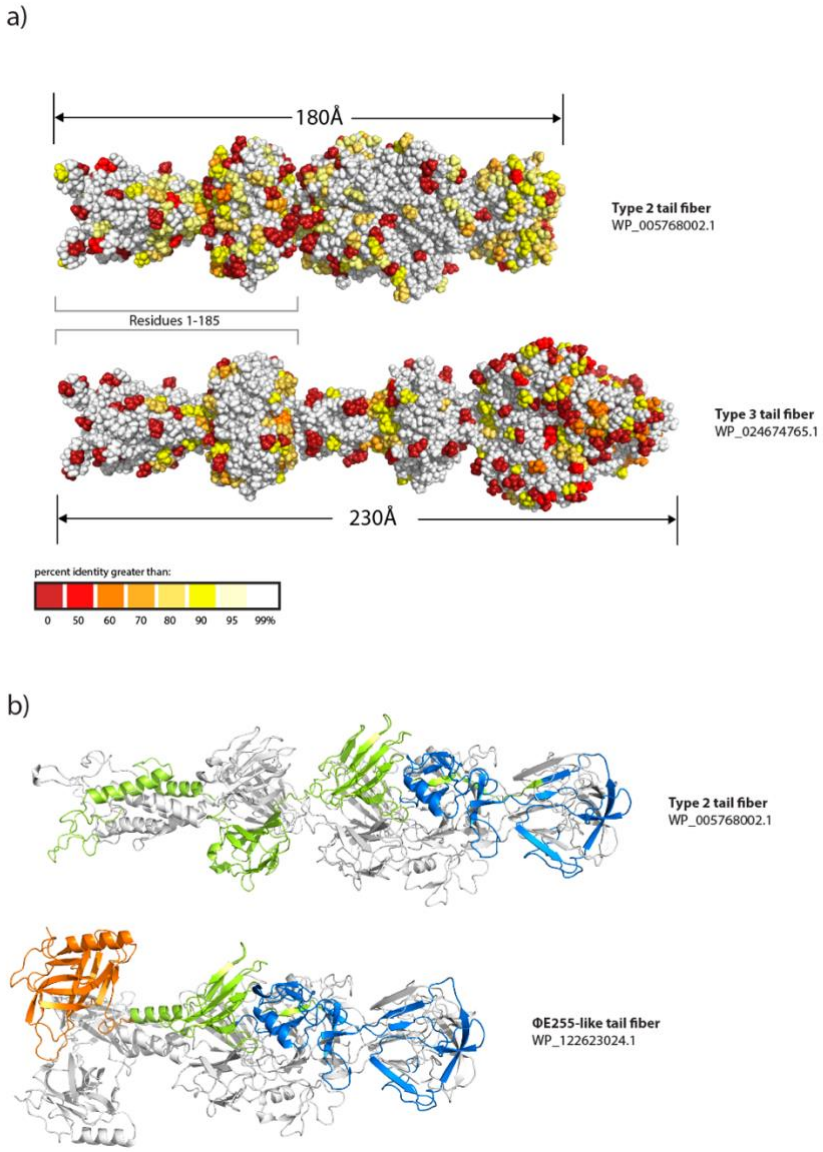


Figure **3.3**: Structural similarities between type 2, 3, and ΦE255-like fibers, as predicted by AlphaFold.

a) Structures of representative fibers belonging to type 2 and type 3, colored by residue conservation. The highly conserved 185 residue region seen in fig. 1b is highlighted with the open rectangles. **b)** Structures of a type 2 and ΦE255-like fiber are shown in ribbon form, with chain b and chain c colored white, and chain a colored to highlight the DUF3751 domain (orange), T4 short tail fiber binding domain (blue), and regions with no known function (green) according to InterProScan.

Distribution of Type 1 and 2 fibers is highly correlated with LPS gene *rfbD*

Having described the richness of tail fiber types carried by PSSC tailocins, we sought to better understand their distribution and relative abundance of each among the genomes in our dataset. The first interesting result was that among the primary phylogroups, type 1a and type 2 tail fibers were fairly evenly represented, at 38% and 42%, respectively, with only 4% of genomes carrying type 3 fibers. Surprisingly, however, the proportion of each fiber type varied greatly by phylogroup. For phylogroups 2, 3, and 6, which formed a monophyletic group in our core genome tree, both type 1a and 2 fibers were again evenly represented (table 3.2). However, phylogroups 1a, 1b, 4, and 5, which also formed a monophyletic group, tended to be dominated by one fiber type or the other (table 3.2). If these distributions represent genuine differences in tail fiber abundances between clades in PSSC, as opposed to being the result of sampling bias in some way, it suggests that perhaps tail fibers are under more heightened selection and undergo more frequent recombination in some phylogroups than others. For instance, the most virulent pathogens tend to reside in phylogroup 1, whereas phylogroup 2 is considered to contain strains that are more widespread, being better epiphytes and generally less virulent (136). In this instance, tailocins might provide a greater fitness advantage to strains that spend more time on the leaf surface than those that spend more time isolated from other microbial competitors inside the plant apoplast. However, it is also possible that sampling biases of highly clonal aggressive pathogens skew this dataset, leading to an underestimation of the true diversity of tail fibers among phylogroups 1a, 1b, 4, and 5.

Table **3.2**: Proportion of genomes carrying each tailocin-associated tail fiber.

PG = phylogroup, n = # of genomes

PG (n)	Type 1a	Type 1b	Type 2	Type 3
1a (108)	0	0	0.92	0
1b (137)	0.93	0	0.04	0
2a (36)	0.11	0	0.41	0.08
2b (118)	0.40	0	0.32	0.11
2c (8)	0.5	0	0.5	0
2d (40)	0.48	0	0.52	0
3 (125)	0.352	0	0.296	0.04
4 (44)	0	0	1	0
5 (9)	0.33	0	0.66	0
6 (9)	0.44	0	0	0.22
7 (1332)	0	0.589	0.002	0.003
9 (2)	0	0	0	0
10a (2)	0	0	0	1
10b (2)	0	0	1	0
11 (12)	0	0	0	0
13 (3)	0	0	0	0

We also looked at the co-distribution of tail fibers associated with tailocins and prophages (Figure 3.4a). We found a very strong correlation between genomes carrying tailocins equipped with type 2 fibers and those harboring prophages. Among genomes carrying type 2

fibers, 54% also carried at least one prophage fiber. In contrast, among genomes carrying type 1a fibers, only 4% of genomes carried any prophage fiber. Considering the structural the similarities between the prophage and type 1a fibers established above, and assuming structural similarities correspond with similar binding affinities to LPS motifs, this distribution pattern suggests that PSSC strains that are able to be infected by phages equipped with the RSA-1 like tail fibers might also be less likely to carry a type 1a tail fiber with their tailocin due to an increased chance of self-killing.

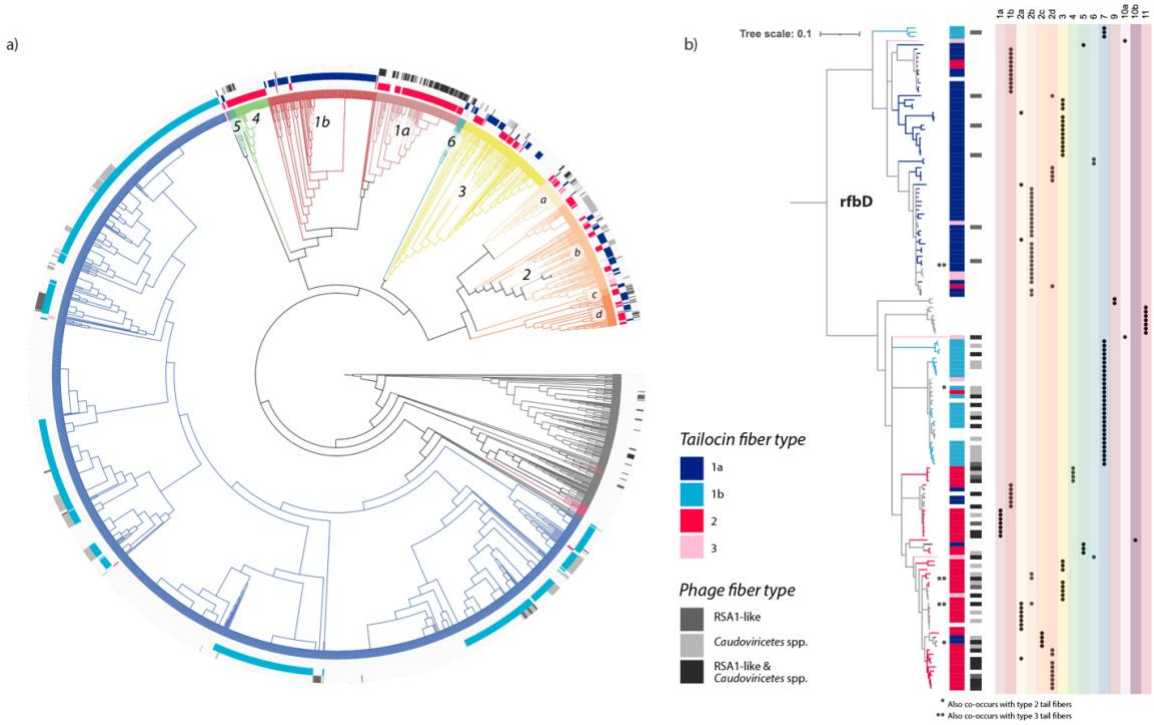


Figure **3.4**: Distribution of tail fiber types correspond better to presence of *rfbD* alleles than to phylogroup.

a) core genome phylogeny of PSSC, with branches colored according to phylogroup. Primary phylogroups are numbered. Grey branches are genomes that were too distant from reference strains to confidently assign to any known phylogroup. Three annotations rings represent, from inside to out, 1) presence of type 2 (dark pink) and 3 (light pink) tail fibers, 2) presence of type 1a (dark blue) and 1b (light blue) fibers, and 3) presence of phage fibers belonging to a broad group of *Caudoviricetes* phage (light grey), RSA1-like phages (mid-grey), or both (dark grey). **b)** a gene tree for *rfbD*, rooted at alleles found exclusively in phylogroup 13. Tree branches and the inner annotation strip are colored according to the tailocin-associated tail fiber types that were observed to be co-occurring in the same genome with each *rfbD* gene sequence. Instances where an *rfbD* sequence is seen co-occurring with multiple tail fiber types are marked with asterisks. The greyscale strip represents *rfbD* co-occurrences with phage-associated tail fibers. To the right, the phylogroup of origin for each *rfbD* sequence is shown.

It has recently been reported that sensitivity to tailocins in PSSC is strongly linked to the carriage of LPS-related genes, and that allelic variation in *rfbD* is a particularly robust predictor (128). Given that *rfbD* is such a strong predicter of tailocin sensitivity, we hypothesized that it might also reflect a barrier to carriage of certain tail fibers, as to avoid self-killing. Additionally, because we observed two dominant tail fiber types (type 1a and 2) and *rfbD* alleles within PSSC form two distinct clades associated with tailocin sensitivities, we hypothesized that the allele of *rfbD* carried by a genome would be strongly correlated with the type of tail fiber carried. Finally, we hypothesized that due to the significant structural similarities between type 1a and 1b fibers, the co-distribution pattern between type 1a fibers and *rfbD* alleles would resemble the pattern between type 1b fibers and *rfbD* alleles. To test these hypotheses, we built a gene tree of *rfbD*, and analyzed co-occurrence with each fiber type (1a, 1b, 2, and 3). For the sake of completeness, we also investigated the co-occurrence of *rfbD* alleles with RSA1-like and *Caudoviricetes* fibers.

Consistent with the results by Baltrus et al. (128), two major clades of rfbD were observed. Each clade exhibited distinct correlations with both tailocin-associated and prophage-associated fibers (Figure 3.4b). Within the first clade of *rfbD* genes, 84% of unique alleles were associated with type 1a fibers, only 9% were associated with any prophage fiber, and 5% were associated with type 2 fibers. In contrast, in the second clade of *rfbD* genes, 7% of unique alleles were associated with type 1a fibers, 58% were associated with at least one prophage-associated fiber, and 49% were associated with type 2 fibers. These results are consistent with the hypothesis that the distribution of tail fibers observed in PSSC is determined at least in part by aspects of LPS structure.

A surprising result from the analysis of tail fiber co-occurrence with *rfbD* alleles is that despite being both genetically and structurally very similar to type 1a tail fibers, type 1b fibers are

associated with *rfbD* alleles that are much more similar to those commonly found in strains carrying type 2 tail fibers than with type 1a fibers. Assuming that it is necessary for a tailocin-producing strain to carry a tail fiber that is compatible with its rhamnose-synthesizing gene *rfbD*, this result implies that despite the similarity between type 1a and 1b fibers, their binding activity and therefore killing spectra are rather distinct, or that despite phylogenetic similarity between *rfbD* alleles associated with type 2 and type 1b fibers, they produce LPS that are rather distinct, such that type 1b fibers are compatible with LPS that type 1a fibers would not be. Whichever is the case, the interplay between *rfbD* and tailocin activity warrants further investigation.

Finally, it is worth briefly commenting on the significance of the observed correlation between extant prophages in PSSC and alleles of *rfbD*. While it has been well documented that many *Pseudomonas* phages target LPS (125,140), and that these phages exhibit differential killing spectra (141,142), to our knowledge we are presenting here the first evidence for a rather striking and simple mechanism that might be driving phage-bacteria dynamics within PSSC: compatibility with one of two *rfbD*-dependent LPS structures that decorate PSSC outer membranes. If, like tailocins, PSSC phage populations turn out to be heavily biased toward certain strains based on the activity of *rfbD* within host strains, the implications for the evolution of PSSC could be significant, as phages have been implicated in the horizontal gene transfer of effector proteins between PSSC strains (143). From the perspective of management, the results here might also provide insight for more thoughtful production of phage therapy cocktails (144)

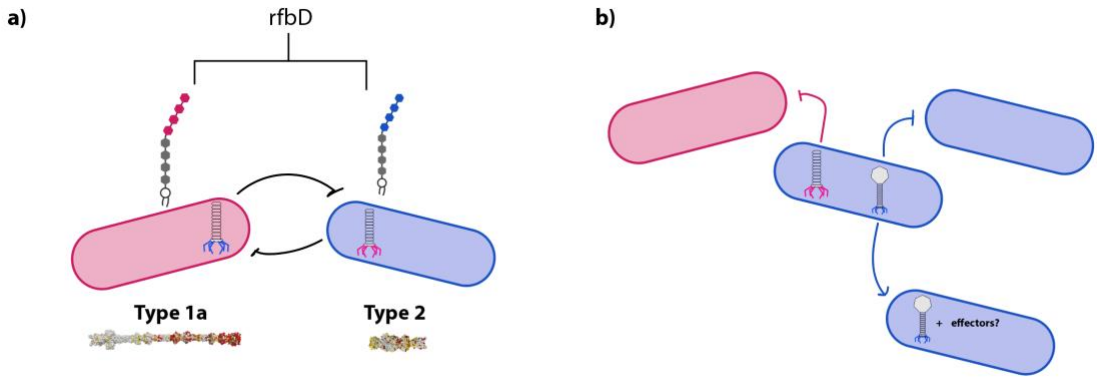


Figure **3.5**: Conceptual figure of main findings and implications

a) Two distinct clades of the LPS biosynthesis gene *rfbD* are likely to result in critical structural differences in LPS molecules decorating PSSC strains, denoted by pink and blue cells. Associated with these distinct *rfbD* clades is the co-occurrence of tail fiber types 1a and type 2 within the primary phylogroups of PSSC, which alongside previous data correlating *rfbD* with killing sensitivity, provides support for a conceptual model of reciprocal killing between strains carrying on or the other tail fiber. **b)** In strains carrying type 2 tail fibers (blue), it is common to find prophage equipped with tail fibers structurally similar to type 1a tail fibers, which likely target competitors with similar LPS structures as the current host. The circulation of prophages among a subset of strains based on *rfbD* not only implies killing of closely related competitors, however, as prophage often carry accessory genes and therefor might also be a source of gene flow among strains carrying similar *rfbD* genes and LPS structures.

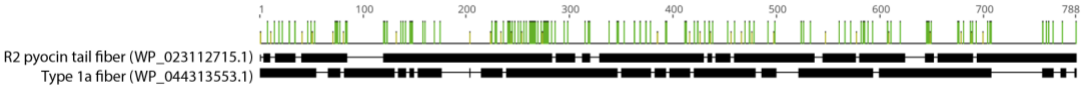
Summary

In this study, I screened 2,161 PSSC genomes for tailocin tail fiber genes with the aim to fully describe the diversity of tail fibers that associate with and determine killing spectra of tailocins carried by PSSC. The genomic screen confirmed that tail fiber types 1a and 2 are the most common tail fibers among tailocins in the primary phylogroups of PSSC. I further describe two previously undescribed tail fiber types: type 1b, a close relative of type 1a tail fibers, found exclusively in phylogroup 7, and type 3, a relatively rare tail fiber that shares ca. 50% homology with type 2 tail fibers. With AlphaFold-predicted protein structures, I showed that type 1 fibers are structurally similar to long tail fibers carried by bacteriophage, while type 2 and 3 tail fibers

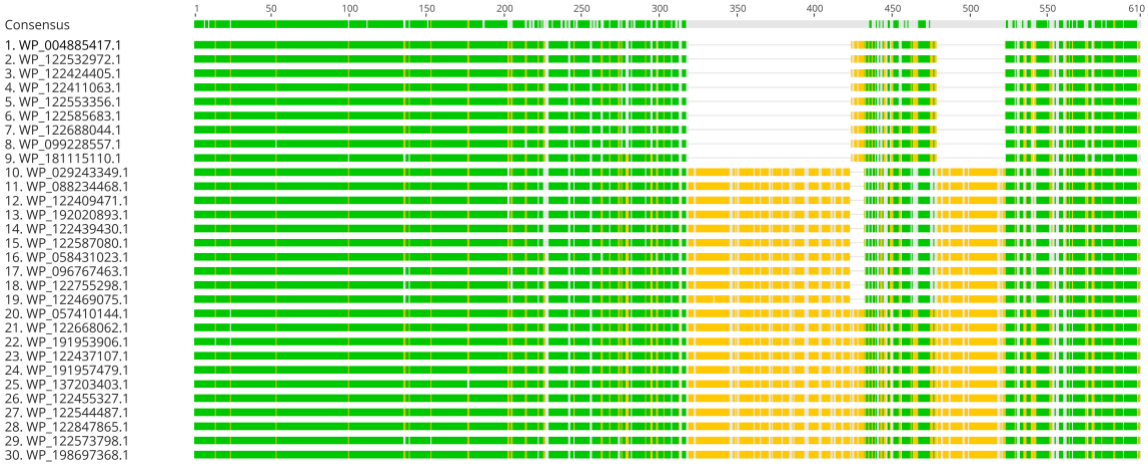
are structurally and genetically similar to short tail fibers. In addition to the tail fibers associated with tailocins, I also detected tail fibers that share strong homology and structural similarities to type 1 tail fibers but are carried by prophages within PSSC genomes. The distribution of the prophage tail fibers was found to have a strong positive correlation with the distribution of type 2 tail fibers, and the distribution of all tail fiber types described here, with the exception of type 3, were all found to be correlated with alleles of the L-Rhamnose biosynthesis gene *rfbD* (figure 3.5). The results presented here provide valuable insights into the diversity of tailocins in PSSC and clarifies the relationships and possible drivers of their distribution within the species complex.

To further interrogate the role tailocins might be playing in the fitness of PSSC, in chapter 4 I investigated the fitness costs and benefits associated with the production of tailocins and other proteinaceous toxins that require cell lysis for toxin release.

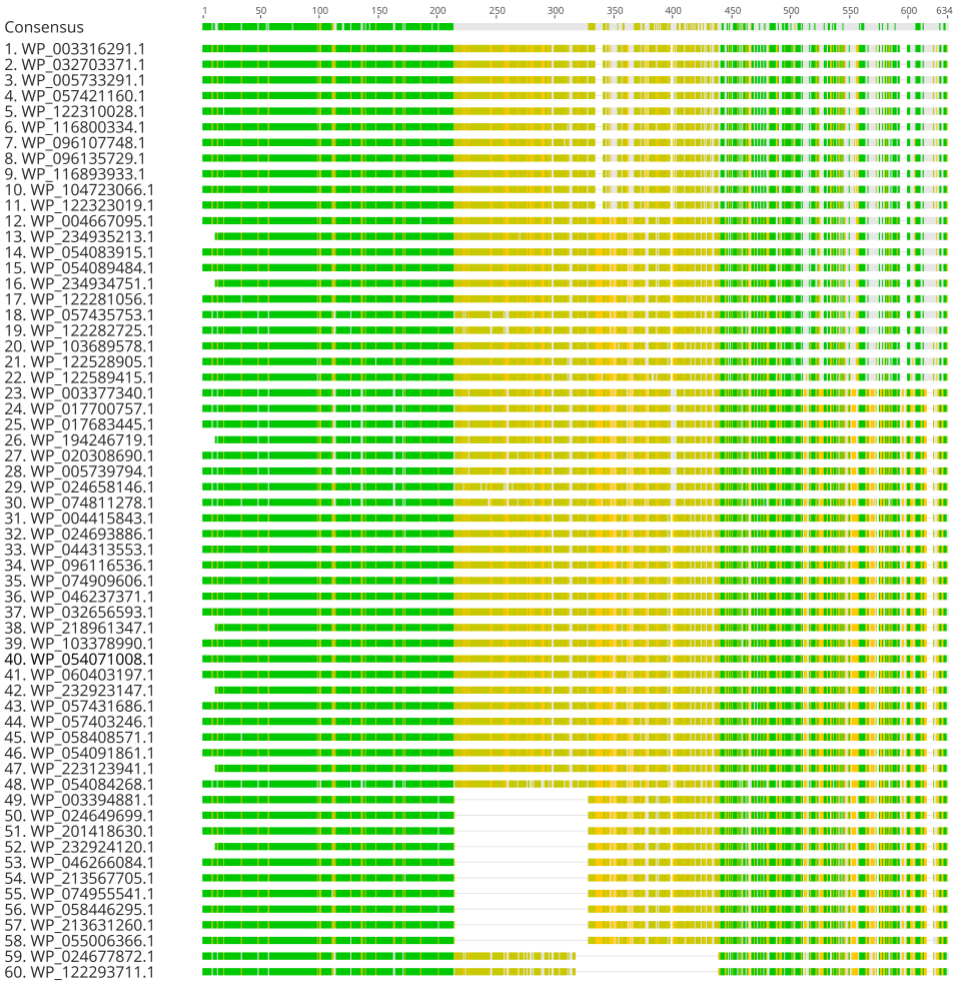
Supplementary Figures



Supplemental Figure **3.1**: Pairwise alignment of type 1a tailocin-associated tail fiber found in PSSC and an R2 pyocin fiber from *P. aeruginosa*



Supplemental Figure **3.2**: Multiple amino acid sequence alignment for all unique type 1b fibers found in this study



Supplemental Figure **3.3**: Multiple amino acid sequence alignment for all unique type 1a fibers found in this study

Supplementary data

All supplementary data are deposited at Zenodo (107). Data for Chapter 3 include:

Supplementary data 3.1

HMM_1, representing tailocin tail fibers associated with killing class 1

Supplementary data 3.2

HMM_2, representing tailocin tail fibers associated with killing class 2

Supplementary data 3.3

HMM_3, representing tailocin tail fibers from PSSC strain UB246

Supplementary data 3.4

Amino acid sequence for WP_044313553.1, representative of type 1a tailocin-associated tail fiber used for protein structure prediction in Supplementary data 3.5

Supplementary data 3.5

PDB file containing predicted structure of WP_044313553.1, representative of type 1a tailocin-associated tail fiber

Supplementary data 3.6

Amino acid sequence for WP_122688044.1, representative of type 1b tailocin-associated tail fiber used for protein structure prediction in Supplementary data 3.7

Supplementary data 3.7

PDB file containing predicted structure of WP_122688044.1, representative of type 1b tailocin-associated tail fiber

Supplementary data 3.8

Amino acid sequence for WP_005768002.1, representative of type 2 tailocin-associated tail fiber used for protein structure prediction in Supplementary data 3.9

Supplementary data 3.9

PDB file containing predicted structure of WP_005768002.1, representative of type 2 tailocin-associated tail fiber

Supplementary data 3.10

Amino acid sequence for WP_024674765.1, representative of type 3 tailocin-associated tail fiber used for protein structure prediction in Supplementary data 311

Supplementary data 3.11

PDB file containing predicted structure of WP_024674765.1, representative of type 3 tailocin-associated tail fiber

Supplementary data 3.12

Amino acid sequence for WP_198721597.1, representative of RSA1-like prophage-associated tail fiber used for protein structure prediction in Supplementary data 3.13

Supplementary data 3.13

PDB file containing predicted structure of WP_198721597.1, representative of RSA1-like prophage-associated tail fiber

Supplementary data 3.14

CSV file containing HMM_1 and HMM_2 genomic screen results, with accession numbers and identities of tail fibers detected in each genome.

Supplementary data 3.15

CSV file containing HMM_3 genomic screen results, with copy number of tail fibers detected in each genome and the phylogroup the genome belongs to

Chapter 4

A solution to the paradox of constitutive bacteriocin production

Abstract

Bacteriocins are proteinaceous antimicrobial toxins produced by many bacteria. For bacteriocins requiring cell lysis for release into the environment, expression is widely thought to be costly and thus tightly regulated by the SOS response to DNA damage. Recently, however, there has been evidence that under laboratory conditions, *E. coli* expresses bacteriocins stochastically in a small fraction of its population, resulting in constitutive expression at the population level, with no apparent inducing agent. Here, we use an agent-based model to provide a simple answer for these contradictory results: low constitutive expression can be recapitulated by assuming damage-induced bacteriocin production in a population growing under genotoxic stress. Further, we show that explicitly modeling damage-induced bacteriocin production reveals a secondary benefit of production in the form of culling damaged population members, allowing for more efficient resource utilization. Finally, we show that under moderate genotoxic stress, this secondary benefit allows bacteriocin producers to successfully invade populations of competitors that are 100% resistant to the bacteriocin released. Our results provide support for the prevailing framework in which bacteriocin production is tightly regulated by the SOS response and suggests that genotoxic stress is an important determinant for the competitive success in bacteriocin producers.

Introduction

Bacteria often find themselves competing for limited resources in harsh environments and have therefore evolved an extensive array of antimicrobial toxins to gain a competitive edge. One particularly ubiquitous group of toxins, found in an estimated 90% of bacterial species surveyed (145), are the ribosomally-encoded peptides characterized broadly as bacteriocins. Given their omnipresence, bacteriocins are thought to provide a substantial benefit to the bacteria carrying them (146), and likely play a key role in shaping many bacterial communities (147,148). The specifics of what these benefits and roles are, however, are still a matter of debate (149). Specifically, bacteriocins often require cell lysis for toxin release, which is thought to impose a significant fitness cost that must somehow be balanced with an equivalent benefit to the producer's kin to be advantageous (150,151). Theoretical and experimental work has identified nutrient availability (152), scale of competition (153), environment structure (154), and niche overlap of competitors (46) as potential factors altering favorability of bacteriocin production, but prediction of competitive outcomes still remains difficult at times (155,156). As such, improving our theoretical understanding of the costs and benefits of bacteriocin production might make the roles they play in community assembly and structure clearer.

It has long been known that lytic release of bacteriocins is connected to the SOS response to DNA damage (49,50,53). The SOS response in bacteria responds to excess single stranded DNA, which when bound by the protein RecA causes the autocleavage and inactivation of the transcriptional repressor LexA (157). In the absence of single stranded DNA, LexA binds to 16-basepair LexA boxes (158) located upstream of many genes associated with DNA repair (159), the arrest of the cell cycle (160), and eventually SOS-induced cell death (161), repressing their expression. To respond appropriately to the amount of DNA damage being experienced by the cell, bacteria control the order in which genes are released from LexA repression by altering the

DNA sequence of the LexA box (162), which in turn alters the binding efficiency of LexA (163). Bacteriocins produced by *E. coli* (colicins) are often regulated by one or two LexA boxes (164) and have been shown to be primarily upregulated when cells are exposed to genotoxic stress (165). Researchers routinely make use of this fact by inducing bacteriocin production on demand with the addition of the mutagen mitomycin C (166). The strict regulation of bacteriocins by the SOS response is thought to provide a strong ecological benefit, as models comparing the fitness of bacteria utilizing different production strategies suggest that stress-induced production consistently outperforms simple constitutive expression (167).

Despite the above evidence for the regulation of bacteriocins by the SOS response, recently two reports of low-level constitutive bacteriocin production under laboratory conditions have been published (54,55), with the suggestion that such behavior might act as a 'pre-emptive attack'(55) or be indicative of a bet-hedging strategy (54). In both reports, multiple colicins (bacteriocins produced by *E. coli*) exhibited the same expression patterns, albeit at production rates that varied widely between ~0.5-2%. These results are seemingly at odds with the predominant understanding of bacteriocin production and represent a sort of paradox. Why do we observe constitutive expression of bacteriocins under seemingly ideal conditions when multiple lines of evidence suggest their induction to be strongly linked with genotoxic stress?

Here, we present a simple solution to the paradox with an agent-based model by showing that low constitutive production of bacteriocins is consistent with cells growing under consistent, but low levels of genotoxic stress. Under such stress, particularly when the rate of damage overwhelms the cells' repair machinery, an efficient method of removing excess damage is dilution via replication (168,169). Recent work on bacterial senescence and immortality implicates asymmetrical damage partitioning during replication as an important process for allowing the rejuvenation of damaged cells, resulting in a heterogenous population of old daughter cells and new daughter cells, with the new daughter acquiring less damage and growing

faster than the old daughter cell. While it is unclear what type(s) of damage are responsible for the discrepancies in growth rate, asymmetry in the partitioning of inclusion bodies (170), proteins, and DNA damage (171) have been reported. We conceptualize bacteriocin production as simply an end stage in this process of bacterial aging and rejuvenation; when old daughter cells have accumulated too much damage to readily rejuvenate, utilizing the cell's remaining metabolic potential for bacteriocin production and release provides a larger benefit to its kin than persisting in the environment would. This conceptualization of early programmed death is consistent with data of SOS gene expression which shows that while many bacteriocins in *E. coli* have two LexA boxes controlling their expression, implying tight control of the genes, the binding affinity is relatively low, and often bacteriocin genes are expressed at higher levels than late SOS genes. In other words, bacteriocin production in cells often begins ramping up before the cell has fully been inundated by DNA damage.

Therefore, we hypothesized that the low level constitutive expression of bacteriocins that has been observed is the result of old daughter cells within a population undergoing programmed cell death and bacteriocin release as response to genotoxic stress in the environment.

To test our hypothesis, we built an agent-based model in which bacteria accumulate DNA damage from the environment at a rate that requires dilution via replication to remove from the cell, and in which lytic bacteriocin production is induced by the excess accumulation of damage beyond a set threshold. We show that such a model induces lytic bacteriocin release in about ~0.5-1% of the population depending on genotoxic stress in the environment, and that altering the threshold of DNA damage that triggers bacteriocin release alters the production rate, suggesting one mechanism for the diversity of constitutive production rates observed experimentally. Further, we show that explicitly linking bacteriocin production to cellular damage provides an additional benefit aside from killing competitors, as the active culling of the most damaged cells and subsequent redistribution of resources to nearby kin reduces mean damage in

the population, allowing bacteriocin producers to reach higher population levels when grown in isolation, and invade resistant populations. Finally, we show that the competitive outcomes for bacteriocin producers in our model are sensitive to the severity of genotoxic stressors in the environment, suggesting that genotoxic stress is an important determinant in the favorability of lytic bacteriocin production. Our results have large implications for how bacteriocin production might be influencing community dynamics in more subtle ways than previously thought and suggest that the suicidal release mechanism associated with bacteriocins might be a feature, not a bug.

Methods

Overview

Purpose

The purpose of this model is to test the hypothesis that constitutive bacteriocin production can arise from strict damage-induced production of the toxin, given that the population is experiencing genotoxic stress that results in the accumulation of damage.

State variables and scales

The model comprises three levels: individual, population, and environment. Individuals are characterized by the state variables of replication energy (E), cellular damage (D), energy threshold for replication (E*), and damage threshold resulting in cell death (D*) (Table 4.1), all of

which are adapted from and act in accordance with a model of bacterial aging proposed by Chao et al. (172).

Three populations of bacteria are considered: bacteriocin producers, bacteriocin sensitives, and bacteriocin resistants. Producers are entirely resistant to their own toxin along with the resistants, and sensitives are always killed by the toxin. While individuals in all populations are characterized by the same state variables, a key difference between populations is the parametrization and behavior associated with D*. For resistants and sensitives, $D^* = 1$, corresponding with an amount of accumulated damage that renders the cell incapable of increasing E and thus incapable of replication. Producers always have $D^* < 1$, and thus cell death occurs for individuals in this population while they are still capable of replicating. However, when $D = D^*$ for producers, bacteriocins are released into the nearby environment at the time of cell death.

The abiotic environment that the model takes place in is a 200x200 square lattice, with each lattice point capable of holding a single bacterial cell. Given an average bacterial size of 1-2 μm, the environmental scale considered here is ca. 200-400μm. Within this small patch, nutrients (N) are assumed to be homogeneous and are replenished fully at each timestep. The environment is also characterized by the rate of damage (Ω) it exerts on individuals, which is also homogeneous throughout the lattice. A conceptual overview of the model is depicted in figure 4.1.

Table **4.1**: State variables and model parameters

Parameter	**Values tested**
Damage rate **(Ω)**	0.005, 0.009
Damage threshold for death **(D*)**	0.8, 0.9, 0.95, 1
Starting replication energy **(E)**	0
Energy threshold for replication **(E*)**	18
Nutrients in lattice **(N)**	$3.6x10^{6}$ min^{-1}
Starting population of producers	4
Starting community size of non-producers	20
Sensitive competitors in community (%)	0, 50, 100
Bacteriocin killing range (neighborhood size)	15
Replication range (neighborhood size)	15

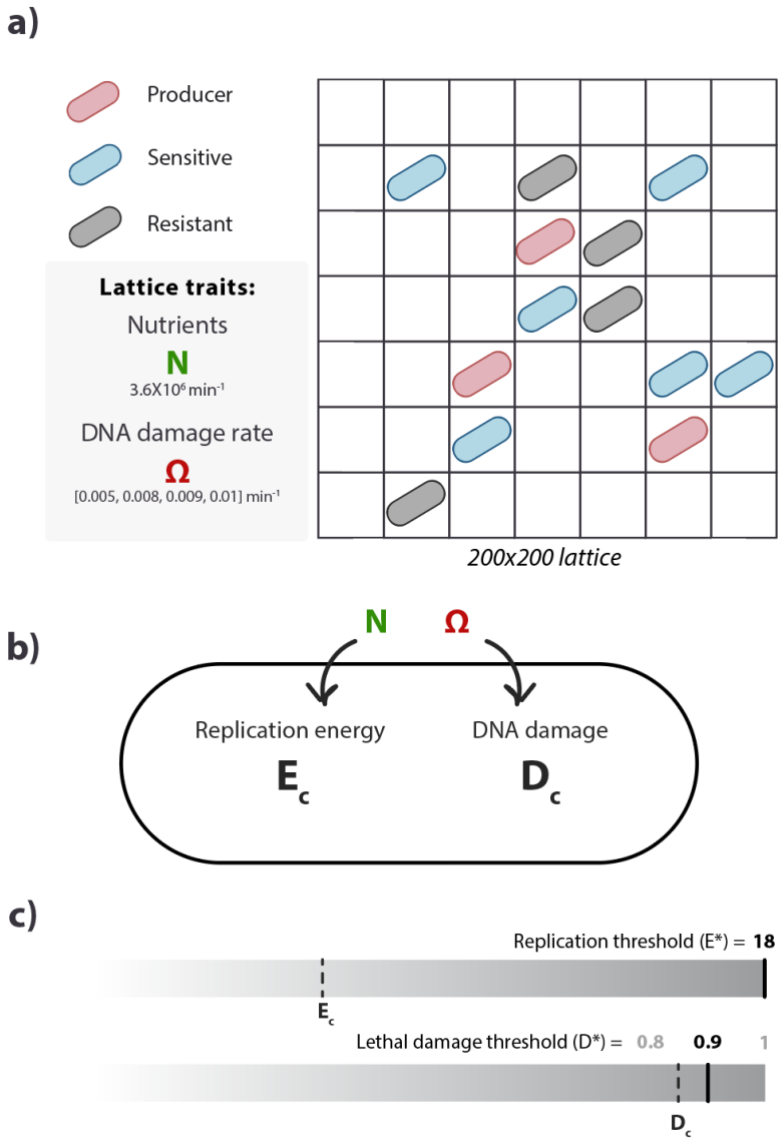


Figure **4.1**: Conceptual figure for the agent-based model

a) All simulations took place on a 200X200 lattice with three possible populations. Nutrients (N) in the lattice were homogeneously distributed, and replenished each timepoint. DNA damage rate (Ω) was homogeneous and constant, effecting all bacteria equally. **b)** Each bacterium accumulated energy for replication (E_c) and damage (D_c) from the environment, until either E_c or D_c reached their respective thresholds, E* and d*. **c)** cartoon representation of both cell states, E_c and D_c, shown as dashed lines, as they approach their thresholds. In the example shown, the cell is near D* and will thus will likely die soon.

Process overview and scheduling

The model proceeds in one minute time steps. Within each minute, or timestep, actions for each individual are processed in the following order: conversion of nutrients in the environment into replication energy, accumulation of cellular damage, death from cellular damage, and replication. Individuals are processed in random order.

Design concepts

Emergence: While bacteriocin production occurs at the level of the individual, the observed production rate of bacteriocin producers is measured at the population level. Thus, production rate is an emergent property of the interaction of individuals including competition for limited nutrients and space and the spatial organization of individuals from each population.

Observation: For analysis of simulations, population-level variables were recorded, including population size, summary statistics of cellular damage in each population, and the number of producers undergoing autolytic bacteriocin release (bacteriocin production rate). All variables were recorded at the end of each timestep.

Details

Initialization

Each simulation was initialized with a small number of bacteria that varied according to the experimental design. For each initial bacterium, E and D = 0.

Input

Genotoxic stress in the environment and D* for bacteriocin producers were the inputs that were varied in all simulations reported here.

Submodels

The submodels described below are taken directly from a model of aging and senescence developed by Chao et al. (172)*:*

Conversion of nutrients into replication energy

A bacterium acquires energy for replication by gathering nutrients from the environment. The ability to convert available nutrients to replication energy is directly dependent on the amount of damage accumulated, such that the increase in replication energy (ΔE) at each timestep is:

$$\Delta E = \min(1-D, N-D)$$

A salient feature of this formula is that when D = 1, a cell ceases to be able to convert any nutrients to replication energy, rendering it unable to replicate.

Damage accumulation

At each timestep, cellular damage increases by Ω. Damage accumulation is constant in both space and time during each simulation run.

Replication

When replication energy meets or exceeds the threshold for replication, a bacterium could replicate, assuming there was an empty lattice point within a 15x15 square neighborhood surrounding the cell.

If replication was possible, all bacteria between the replicating cell and the empty square were pushed toward the empty square, allowing space for a 'new' daughter cell to be produced immediately adjacent to the parent cell, which upon replication became an 'old' daughter cell. The differentiation between new and old daughter cells corresponds to the distribution of damage between each, as damage was asymmetrically distributed between daughter cells in bacteria. In line with empirical observations (172), the new daughter cell receives 48% of the parents accumulated damage, and the old daughter retains 52% of the damage.

The following submodels are additions to the above model that account for lethal levels of damage within a bacterial cell:

Death

When accumulated damage meets or exceeds the damage threshold for death, a bacterium would die, immediately converting the lattice point it was occupying into an uncolonized state capable of being utilized by other bacteria.

Bacteriocin release

For bacteriocin producers, when the bacterium died, it also released bacteriocins into the immediate 15x15 square neighborhood surrounding the cell. Any sensitive competitors in this neighborhood were killed, and bacteriocins were assumed to degrade immediately, with no further diffusion or persistence in the environment.

An overview of the submodels that make up the individual's life cycle is depicted in figure 4.2.

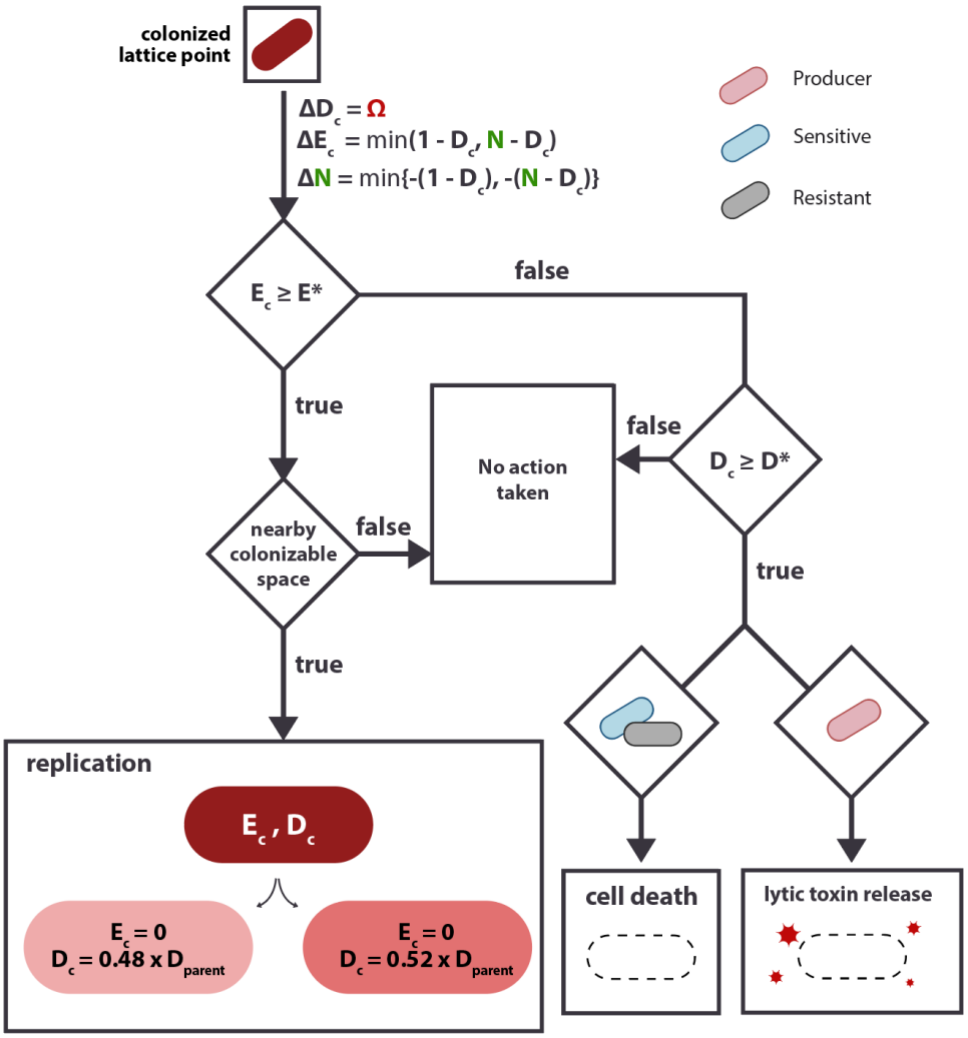


Figure **4.2**: submodels and their order of processing that determine the growth and death of bacteria

a) All simulations took place on a 200X200 lattice with three possible populations. Nutrients (N) in the lattice were homogeneously distributed and replenished each timepoint. DNA damage rate (Ω) was homogeneous and constant, effecting all bacteria equally. **b)** Each bacterium accumulated energy for replication (E_c) and damage (D_c) from the environment at a rate described in panel **(d)**, until either E_c or D_c reached their respective thresholds, E* and d*. **c)** cartoon representation of both cell states, E_c and D_c, shown as dashed lines, as they approach their thresholds. In the example shown, the cell is near D* and will thus will likely die soon.

Experiment 1: Bacteriocin producers grown in isolation

In an attempt to recapitulate the constitutive production of bacteriocins observed in *E. coli*, four populations of bacteria, 3 bacteriocin producers with increasing values for D* (0.80, 0.90, 0.95), and a non-bacteriocin producer with D* = 1 as a negative control, were grown under two levels of genotoxic stress, $\Omega = 0.005/min^1$ and $\Omega = 0.009/min^1$. Each population was grown in isolation, with 10 replicates for each simulation.

In each simulation, four bacterial cells were initiated in random squares throughout the lattice and monitored for 4,320 timesteps (72 hours). Population size and bacteriocin production rate at each timestep was recorded.

Experiment 2: Bacteriocin producers invading an established community

To investigate the implications of damage-induced bacteriocin production for competitive outcomes, three communities of competitors (100% resistants, 50% resistants and 50% senstitives, and 100% sensitives) were challenged with producers.

For each simulation, twenty competitors were initiated in random squares throughout the lattice and allowed to colonize the lattice for 1,440 timesteps (24 hours) before randomly placing four bacteriocin producers randomly in uncolonized lattice points and monitoring the community for an additional 4,320 timesteps. As with experiment 1, communities were grown under two levels of genotoxic stress, $\Omega = 0.005/min^1$ and $\Omega = 0.009/min^1$, with four populations of invaders (3 bacteriocin producers with increasing values for D* (0.80, 0.90, 0.95), and a non-bacteriocin producer with D* = 1 as a negative control) and 10 replicates were performed for each simulation.

Model assumptions

Nutrients were homogeneously dispersed, assuming global competition for resources. It's been shown that scale of resource competition is an important factor for favorability of bacteriocin production (46,153) and defining carrying capacity for local neighborhoods around the producer might alter the observed effects of carrying capacity to optimal production rates. Nonetheless, at the spatial scale modeled here (ca. 1-4mm^2), the rapid diffusion of resources makes global resource competition a reasonable assumption.

All competition is assumed to take place in a closed community. Immigration, especially frequent immigration, is likely to alter community dynamics described here.

Additionally, Assumptions about bacteriocin production are as follows:

Toxins do not persist in the environment beyond a single timestep

Producers are 100% resistant to toxin produced by their kin

Resistant cells are 100% resistant

All sensitive cells are 100% sensitive

Results

Damage-induced lytic bacteriocin release results in stable, constitutive production under sub-lethal genotoxic stress

To understand the basic behavior of bacteriocin production under our model, we first allowed isolated populations to grow from four cells to carrying capacity (14,173) . Every minute, we recorded the population size and the proportion of bacteria that underwent lytic bacteriocin

release since the last timepoint. The bacteriocin production rates reported are therefore in the units of lytic cells min^{-1}. We found that under both moderate ($\Omega = 0.005$) and strong ($\Omega = 0.009$) genotoxic stress, populations of bacteriocin producers exhibited rapid growth followed by stable oscillations around carrying capacity (figure 4.2a). As all populations were able to maintain their populations without crashing, we consider both genotoxic stress levels tested to be sub-lethal.

Bacteriocin production had two significant effects on carry capacity under moderate genotoxic stress (figure 4.2a, left). First, the effective carrying capacity varied significantly between all populations tested. Surprisingly, non-bacteriocin producers had the lowest mean population at carrying capacity ($2.7x10^4 \pm 38$, figure 4.3a), as measured from timepoint 1000 onward. Bacteriocin producers exhibited mean population sizes at carrying capacity of $2.9x10^4 \pm 8$, $2.9x10^4 \pm 25$, and $2.8x10^4 \pm 422$ when D* = 0.8, 0.9, and 0.95, respectively. The second major effect of bacteriocin production was in reducing the magnitude of oscillations around carrying capacity, particularly when D*=0.8, as reflected in the mean standard deviation from the mean population size from timepoint 1000 onward (D* = 0.8, $1.2x10^3 \pm 128$; non-bacteriocin producers, $4.3x10^3 \pm 117$). Both the increased carrying capacity and reduced population variance in bacteriocin producers suggest that in addition to killing sensitive competitors, damage-induced lysis can provide a significant benefit to bacteria by reducing population overshoots that lead to inevitable collapse and by increasing resource use efficiency. Under strong genotoxic stress, carrying capacity of all populations was significantly reduced, and any benefits of lysis to bacteriocin producers were largely negated, with differences in mean population size and variance being practically insignificant albeit still statistically significant (D* = 0.8, $2.1x10^4 \pm 10$, S.D. $1.4x10^3 \pm 11$; non-bacteriocin producers, $2.0x10^4 \pm 5$, S.D. $1.8x10^3 \pm 11$) (figure 4.2a, right).

Under both moderate and strong genotoxic stress, all bacteriocin producers exhibited stable oscillations in production rate between ~0.5-1% (figure 4.2b), consistent with the reported constitutive production rates for colicins of 0.5-2.6% (54,55). Perhaps counterintuitively, under

moderate genotoxic stress, the most aggressive threshold for lysis (D*=0.8) produced the lowest spikes in production rate – aside from a single large spike in production rate when the population first reached carrying capacity (figure 4.2b, left). Despite the more stable production rate with lower spikes when D*=0.8, mean production rates under moderate genotoxic stress consistently had a small but significant negative correlation with damage threshold (D* = 0.8, 6.3%; 0.9, 6.1%; 0.95, 5.7%) (figure 4.2c, left). Under high genotoxic stress, this correlation became more pronounced (D* = 0.8, 1.14%; 0.9, 1.01%; 0.95, 0.96%) (figure 4.2c, right), and the counterintuitive pattern in peak production rates was not observed. The results from the simulated growth of isolated populations suggest that an interaction between genotoxic stress and damage threshold for lytic bacteriocin release ultimately determine mean constitutive production rate and fitness (as defined by effective carrying capacity and magnitude of population oscillations) of bacteriocin producers.

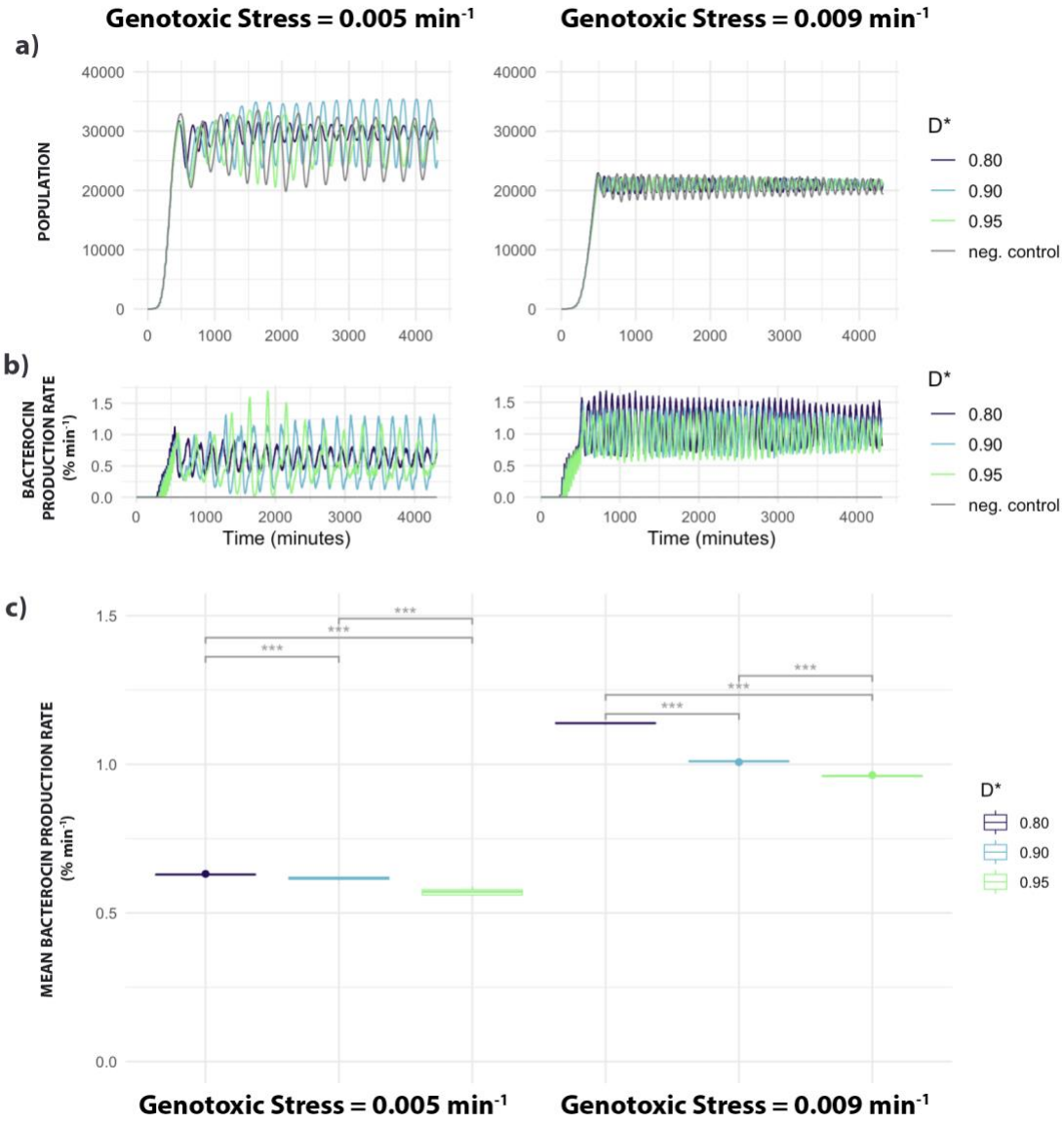


Figure **4.2**: Damage-induced bacteriocin production results in low constitutive bacteriocin production.

All data represents results from 10 replicates. **a)** population dynamics of single populations growing in isolation under moderate (left), and strong (right), genotoxic stress. **b)** bacteriocin production rate over the course of the simulation, calculated as the percentage of the population lysing themselves each minute, under moderate (left), and strong (right), genotoxic stress. **c)** boxplots for mean constitutive production rate, as measured from timepoints 1000-5760, under moderate (left), and strong (right), genotoxic stress.

Under moderate genotoxic stress, bacteriocin production improved competitive outcomes against both sensitive and resistant competitors

As we observed a strong interaction between Ω and D* in both the population dynamics and bacteriocin production rate, we sought to describe how these variables alter invasion dynamics of bacteriocin producers. We allowed three communities, comprised of 100% resistant, 50% resistant, and 100% sensitive populations, to establish for 24 hrs. (1440 timesteps) before challenging each with four bacteriocin producing cells – consistent with the number of immigrating bacteria on the surface of a leaf in a given hour (14,173) – and allowing the community to interact for 72 hrs., with population size and bacteriocin production rate collected every minute. For a negative control, we also challenged the same communities with non-bacteriocin producers.

As expected, under moderate genotoxic stress, increasing the proportion of sensitive competitors resulted in significantly larger populations of bacteriocin producers at the 72 hr. timepoint compared to the negative control (figure 4.3a). Surprisingly, both bacteriocin producing populations tested performed significantly better than the non-bacteriocin producers when invading a community of 100% resistant competitors (final population size, D*=0.8, 3563 ± 1119; 0.9, 1971 ± 802; non-bacteriocin producers, 285 ± 805) (figure 4.3a). These results reenforce those from the isolated growth experiments, and imply that when induced by DNA damage, bacteriocin production might provide a competitive advantage over even resistant competitors in some environmental conditions, assuming the resistant population's D* is greater than the producers. If we assume that D* for resistants is not greater than producers, due to evolution of the trait to optimize fitness, bacteriocin producers might no longer have a fitness advantage in competition with resistant competitors, but such a scenario would still suggest that programmed cell death accompanying bacteriocin release is not a burden to producers, but instead an optimal strategy with or without bacteriocins. Regardless of the value of D* for competitors,

the results presented here suggest that when coupled to cellular damage, bacteriocin production may not be as costly as is currently believed.

Compared to the isolated growth experiments, bacteriocin production rates exhibited greater fluctuations, with spikes in production rate up to ~4 and 7% when challenging sensitive and resistant communities, respectively (figure 4.3c and e). However, when challenging sensitive competitors (figure 4.3c), we observed a tendency for production rates to spike in the initial stage of invasion and stabilize to ~0.6% as sensitive competitors were killed off. Likewise, when challenging resistant populations (figure 4.3e) spikes in production rate lowered in intensity over the course of the simulations from ~6% to ~3%, as bacteriocin producers established themselves in the community. These results indicate that damage-induced bacteriocin production can be an effective strategy to sense the presence of non-kin, and that as relative abundance of bacteriocin producers increase, production rate tends to decrease.

Under strong genotoxic stress, the benefits of bacteriocin production were insignificant

When experiencing strong genotoxic stress, bacteriocin producers were not able to significantly increase their final population size against any community compared to non-bacteriocin producers (figure 4.3b). As observed under moderate genotoxic stress, production rates were dramatically higher than the isolated growth experiments, with frequent spikes in production rate >2% and >3% when challenging 100% sensitive and 100% resistant competitors, respectively (figure 4.3d and f). Unlike under moderate genotoxic stress, however, production rates remained high through these simulations. These results indicate that successful invasion of both sensitive and resistant communities is strongly dependent on the severity of genotoxic stress in the environment.

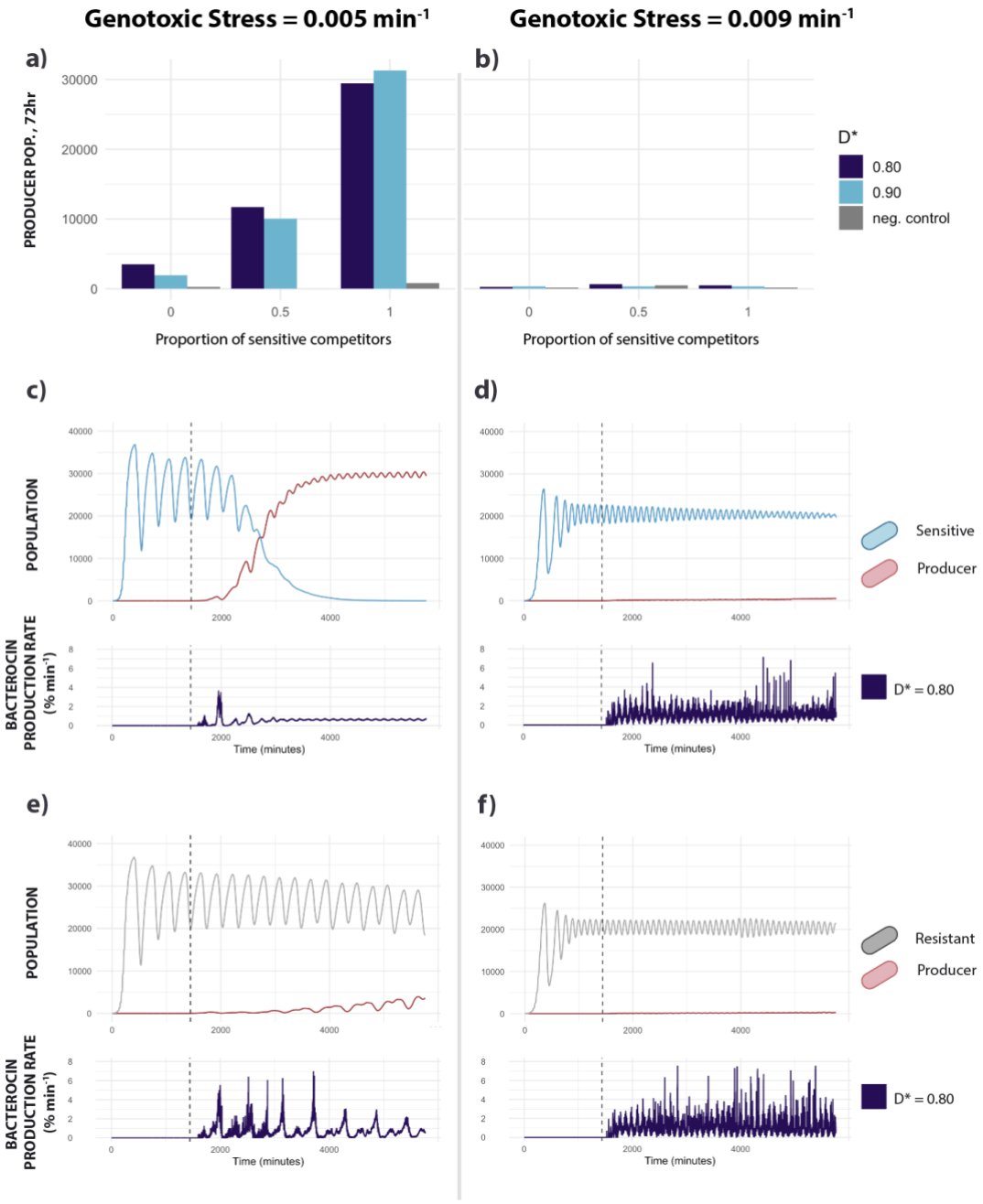


Figure **4.3**: Bacteriocin production improves invasion under moderate but not strong genotoxic stress.

Bar plots (**a)** and (**e**) show final invading population sizes (at timepoint 5,760) when invading communities composed of 0, 50, or 100% sensitive competitors. The negative control in each panel was a population of non-producers, with D*=1. Below, the population and bacteriocin production dynamics for bacteriocin producers with D*=0.8 invading 100% sensitive communities under **c)** moderate and **d)** strong genotoxic stress are shown, and the same bacteriocin producers invading 100% resistant competitors under **e)** moderate and **f)** strong genotoxic stress are at the bottom. In **c-f**, the vertical dashed line represents the timepoint (1,440) at which the invading population was added.

Discussion

Bacteriocin production, when induced by excess DNA damage, provides several advantages to a bacterial population. When grown in isolation, bacteriocin producers are able to sustain a higher carrying capacity and mitigate oscillations in population size cause by overshooting carrying capacity (figure 4.2). The exact mechanism allowing for these benefits are not clear, but a plausible explanation is that assuming some death in the population is inevitable due to constant damage accumulation and insufficient nutrients to support continued growth, early programmed death of the most damaged cells in a population allows for resources to be used more efficiently by healthier cells.

The benefits of damage-induced bacteriocin production seen in isolated growth simulations translated into increased invasion success, even when competing against resistant competitors (figure 4.3a and e). While this is an interesting result, it does not align with empirical observations (174) and interpretation should be taken with some caution. First, it is likely that along with the increased resource use efficiency noted above, localized competition for space in the model also contributed greatly to invasion success. Localized competition in our model arises via cells only being able to replicate into their surrounding 15x15 lattice points. As populations grew in spatially segregated colonies, when a producer lysed, it was most often other producers that were able to take advantage of the newly uncolonized lattice point. If all resources were under global competition (i.e., any bacterium in the lattice could replicate so long as there were any uncolonized lattice points) bacteriocin producers would likely no longer have an advantage over competitors. This is a known limitation on bacteriocin production that has already been explored theoretically (153) and known to apply even to invasion of sensitive competitors (154), however our model did not explore the sensitivity of successful invasion to altering the scale of competition, and it's possible that even small increases to the scale of competition could have

large effects on the competitive outcome. Another aspect of the model that might have played a role in the advantage of producers over resistant populations was the lack of response from the resistant population to nutrient stress when approaching carrying capacity. As a common environmental stress, most bacteria respond to starvation with dramatic shifts in gene expression and metabolic activity to increase survival (175). It is unclear how effective these responses are under the additional burden of genotoxic stress, but their incorporation into future models might result in a less dramatic overshoot and population collapse at carrying capacity than observed here, and in turn a produce less drastic difference in fitness under genotoxic stress between the producer and resistant populations.

Despite these caveats, the increased fitness that damage-induced production provided for bacteriocin producers in our model makes sense as a logical extension of apoptosis-like death, in which beyond a certain threshold of damage, cell death benefits the population more than continuing to use resources inefficiently does. In this vein, instead of bacteriocin production being considered an anticompetitor trait that confers a heavy fitness cost to the producer, it might be equally as reasonable to conceptualize it as a simple altruistic trait with the added bonus of killing sensitive competitors occasionally.

As expected, when sensitive competitors were challenged with bacteriocin producers, they were easily invaded and fully displace by the end of the simulation. However, this was only the case under moderate genotoxic stress. Under high genotoxic stress, producers failed to invade with any more success than no-bacteriocin producers. While this points to some trade-off between population loss from autolysis of producers and the resources gained by killing competitors, this result is likely highly exaggerated by a simplification in our model that removes all bacteriocin particles after the initial burst of killing. In reality, bacteriocins are likely able to persist in the environment and maintain their killing potential for much longer than the one-minute timestep used in the model. With the incorporation of bacteriocin particle diffusion and persistence in the

environment, it is reasonable to assume that the fitness benefits of bacteriocin production would be sufficient to overwhelm the sensitive competitors.

For PSSC and the bacteriocins it carries, including tailocins, the results of this study suggest that damage-induced bacteriocin production might offer an ideal solution to another stress faced in the epiphytic stage of the disease process. Leaves function most effectively when exposed to sunlight, and yet with sunlight comes high energy UV radiation that often results in severe decreases in leaf epiphytic populations (176). If programmed cell death is an effective strategy to manage such genotoxic stress, the addition of bacteriocin production would allow for removal of any sensitive cells that happened to be nearby, allowing a greater ability for PSSC populations to recover when the UV stress is relieved during the night, ultimately increasing the epiphytic population size of the pathogen and its ability to cause disease.

As for the paradox of constitutive expression, if we assume that the solution presented here is correct, that constitutive expression is the result of genotoxic stress, it implies that many experiments performed in an effort to understand the ecological role of bacteriocins have neglected to account for a key environmental factor when assessing competitive outcomes and measures of fitness. Specifically, while the nutrient-rich media and optimal temperatures used to grow bacteria in laboratory conditions might be ideal for producing large numbers of bacteria quickly, they might also be introducing stress due to the dense, fast-growing populations. For example, within the first hours of plating E. coli, upregulation of oxidative stress response genes can be observed (177), and at late log phase bacteriocin production increases significantly (47,165,178), indicating severe DNA damage and SOS response induction. Properly accounting for the genotoxic stress bacteria have experienced during and immediately prior to measuring production rate would be useful for providing support for or against the hypothesis presented here.

Summary

In this study, I present a possible explanation for recent observations that bacteriocin producers grown in isolation exhibit constitutive expression. I show that when modeled under a framework of damage dilution through replication, bacteriocin producers can both tightly regulate their lytic release of toxins and exhibit constitutive expression. Additionally, I show that damage-induced production provides a benefit for producers beyond killing sensitive competitors and increases fitness against even resistant competitors under moderate stress. These results have implications for both how we study bacteriocins in laboratory conditions and suggest that the fitness costs associated with bacteriocins may not be as severe as is often described (146,151). In turn, a lower fitness cost relaxes constraints on carrying the toxins and might explain why they are so ubiquitous in bacteria.

Code availability

The code for all simulations is available for review at github.com/cwf30/BacterioSim

Chapter 5

Conclusion

In Chapter 2, I describe the creation of Syringae.org, a web tool for PSSC isolate identification and prediction of virulence factors within that isolate. The goal of this website was to increase the efficiency and speed of routine surveillance of plant pathogens, providing:

- Infraspecific classification of PSSC isolates from a single marker gene, improving upon the performance of available classification systems.
- Prediction of whether the isolate has the canonical type III secretion system, a frequent tool of certain virulent bacterial pathogens.
- Prediction of type III secretion system effector families carried by the isolate, which ultimately play a significant role in determining host range
- Prediction of the Woody Pathogen and Pseudomonas (WHOP) metabolic gene cassette, which is associated with pathogens of woody hosts

While the website represents a useful tool as it stands, there are improvements that can be made to increase its utility and impact. First, a comprehensive report of all known virulence factors in closely related strains would ensure that the full potential of the isolate was covered by the functional prediction. There are many phytotoxins produced by PSSC that are not currently supported by Syringae.org, primarily due to my focus on those that are most likely to be informative with regard to the host range of the potential pathogen, and many phytotoxins are considered to be general virulence factors that do not impact host range as much as they do disease severity (28). Second, a long-term goal for Syringae.org is to collect spatiotemporal data for isolate detection throughout the world, as users submit marker genes for identification. At the moment, user activity is not recorded, and all marker genes submitted are forgotten as soon as the user leaves the site. By collecting marker gene sequences and their predicted identity, the website

would have a unique opportunity to track pathogens detected during routine surveillances. This data is often not published unless the isolate is of particular note, and thus it is likely that there is epidemiological data that is being underutilized.

In Chapter 3, I screened 2,161 PSSC genomes for tailocin-associated tail fibers and found new diversity previously unreported, along with many prophages within the species complex that are equipped with strikingly similar tail fibers to some known tailocin fibers. The major findings of this chapter were:

- The discovery of a rare tail fiber clade (type 3) associated with PSSC tailocins
- Type 1a and 1b fibers are long tail fibers that appear to have the same evolutionary origin, with 1a fibers found in primary phylogroups, and type 1b found in *P. viridiflava* and other genomes in phylogroup 7
- The long tail fibers are composed of several repeats of binding knobs, with instances of whole binding knob deletions observed
- Type 2 and 3 fibers are short tail fibers
- Several prophage tail fibers in PSSC have striking genetic and structural similarity to type 1 fibers, but are much more likely to be found in genomes with tailocins equipped with type 2 fibers
- The distribution of all tail fibers except for type 3 are strongly correlated with the LPS gene *rfbD*

There are many new questions and hypotheses generated by this work that will inform future work on tailocins in PSSC. The finding that *rfbD* and therefore likely LPS structure plays a large role in the tail fiber carried by a PSSC strain, along with this correlation also being observed in phage fibers suggests that LPS might be a signature of broader interactions within the species complex. Aside from the obvious killing activity by tailocins, phage also enhance the capabilities of pathogens by bringing with them many effectors and virulence factors that can alter host range

and virulence. A comprehensive description of prophages and their distribution throughout the species complex might provide support for the LPS-mediated influence on population dynamics.

Another question raised by this work is the significance of tailocins carrying a short *vs.* long tail fiber. In bacteriophage, long tail fibers bind reversibly and are used as the first point of contact between the particle and host. The reversibility allows the phage to find an ideal spot to settle before initiating infection. The short tail fibers, on the other hand, bind irreversibly and serve to stabilize the particle during infection (116). Whether this dichotomy in binding affinity is preserved in tailocin fibers is unknown. If it is, it likely has impacts on the killing efficiency of the tailocin, although it is difficult to predict in which direction.

In chapter 4, I built a simple agent-based model to investigate the potential of damage-induced bacteriocin production to recapitulate experimental observations of constitutive expression, which is seemingly at odds with the tight regulation of production by the SOS response that has decades of support. In this chapter I showed that:

- Under genotoxic stress, bacteriocin producers can exhibit constitutive expression.
- Damage-induced bacteriocin production provides an often-overlooked secondary benefit of removing the most damaged cells from the population, which when grown in isolation, increases carrying capacity through more efficient resource use, and in competition aids in invasion even of resistant competitors.
- The severity of genotoxic stress greatly impacted the invasion potential for bacteriocin producers, even when invading a population of sensitive competitors.

The results presented in this chapter offer an interesting framework to view bacteriocin production and suggest that the fitness costs of production might not be as severe as often assumed when autolytic release of the toxin is explicitly linked to cell aging and damage load. While many assumptions were made in the model about aging and the asymmetric partitioning of

DNA damage, I believe that the results align well with empirical observations of bacteriocin activity and are worth further investigation.

Overall, this dissertation highlights the multi-faceted challenges a plant pathogen faces during the disease process, and in turn the multi-faceted challenges we face when trying to predict the course of disease. In addition to the host manipulation that we often associate with pathogens, PSSC must also contend with abiotic stress in the environment and competition from leaf epiphytes. These challenges are not only critical for the pathogen to overcome, but they are intertwined. Damage from UV radiation, for example, slows the growth of the pathogen and reduces its population size (176). In chapter 4 I suggest that autolytic release of bacteriocins, in addition to killing competitors, aids in the response to genotoxic stress such as UV radiation through the culling of damaged cells from the population. Likewise, in chapter 3, I discuss the importance of LPS, a cell surface molecule that is critical for desiccation resistance (179), as a primary determinant not only of what tailocins can kill a cell, but also what tailocins can be carried by the cell. The ability of a pathogen to balance the response to abiotic and biotic stresses early on in the disease process, well before interactions with the host begin, are thus vital to consider when assessing a pathogen's potential virulence and represent promising opportunities for disrupting the disease process before the pathogen can begin establishing itself within the plant.

References

1. 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. Front Microbiol. 2017 Dec;8(DEC):2422.

2. Vanneste J. Pseudomonas syringae pv. actinidiae (Psa): a threat to the New Zealand and global kiwifruit industry. N Z J Crop Hortic Sci. 2012 Dec 1;40(4):265–7.

3. Donati I, Buriani G, Cellini A, Mauri S, Costa G, Spinelli F. New insights on the bacterial canker of kiwifruit (Pseudomonas syringae pv. actinidiae). J Berry Res. 2014;4(2):53–67.

4. Steele H, Laue BE, MacAskill GA, Hendry SJ, Green S. Analysis of the natural infection of European horse chestnut (Aesculus hippocastanum) by Pseudomonas syringae pv. aesculi. Plant Pathol. 2010;59(6):1005–13.

5. Şahin F. Severe outbreak of bacterial speck, caused by Pseudomonas syringae pv. tomato, on field-grown tomatoes in the eastern Anatolia region of Turkey. Plant Pathol. 2001;50(6):799–799.

6. Newberry EA, Babu B, Roberts PD, Dufault NS, Goss EM, Jones JB, et al. Molecular Epidemiology of Pseudomonas syringae pv. syringae Causing Bacterial Leaf Spot of Watermelon and Squash in Florida. Plant Dis. 2018 Mar;102(3):511–8.

7. Gitaitis RD. Bacterial Blight of Sweet Onion Caused by Pseudomonas viridiflava in Vidalia, Georgia. Plant Dis. 1991;75(11):1180.

8. Katagiri F, Thilmony R, He SY. The Arabidopsis Thaliana-Pseudomonas Syringae Interaction. Arab Book Am Soc Plant Biol. 2002 Mar 27;1:e0039.

9. Xin XF, Kvitko B, He SY. Pseudomonas syringae: What it takes to be a pathogen. Nat Rev Microbiol. 2018 May 1;16(5):316–28.

10. Upper CD, Hirano SS. Revisiting the roles of immigration and growth in the development of populations of Pseudomonas syringae in the phyllosphere. Phyllosphere Microbiol. 2002;69–79.

11. Arnold DL, Lovell HC, Jackson RW, Mansfield JW. Pseudomonas syringae pv. phaseolicola: from ‘has bean’ to supermodel. Mol Plant Pathol. 2011;12(7):617–27.

12. Upper CD, Hirano SS, Dodd KK, Clayton MK. Factors that Affect Spread of Pseudomonas syringae in the Phyllosphere. Phytopathology®. 2003 Sep;93(9):1082–92.

13. Petriccione M, Zampella L, Mastrobuoni F, Scortichini M. Occurrence of copper-resistant Pseudomonas syringae pv. syringae strains isolated from rain and kiwifruit orchards also infected by P. s. pv. actinidiae. Eur J Plant Pathol. 2017 Dec 1;149(4):953–68.

14. Hirano SS, Upper CD. Bacteria in the Leaf Ecosystem with Emphasis onPseudomonas syringae—a Pathogen, Ice Nucleus, and Epiphyte. Microbiol Mol Biol Rev. 2000 Sep;64(3):624–53.

15. Ichinose Y, Taguchi F, Mukaihara T. Pathogenicity and virulence factors of Pseudomonas syringae. J Gen Plant Pathol. 2013 Sep 1;79(5):285–96.

16. Lindow SE, Brandl MT. Microbiology of the Phyllosphere. Appl Environ Microbiol. 2003 Apr;69(4):1875–83.

17. Remus-Emsermann MNP, Tecon R, Kowalchuk GA, Leveau JHJ. Variation in local carrying capacity and the individual fate of bacterial colonizers in the phyllosphere. ISME J. 2012 Apr;6(4):756–65.

18. van der Wal A, Leveau JHJ. Modelling sugar diffusion across plant leaf cuticles: the effect of free water on substrate availability to phyllosphere bacteria. Environ Microbiol. 2011;13(3):792–7.

19. Rouse DI. A Model Relating the Probability of Foliar Disease Incidence to the Population Frequencies of Bacterial Plant Pathogens. Phytopathology. 1985;75(5):505.

20. Donati I, Cellini A, Sangiorgio D, Vanneste JL, Scortichini M, Balestra GM, et al. Pseudomonas syringae pv. actinidiae: Ecology, Infection Dynamics and Disease Epidemiology. Microb Ecol. 2020 Jul 1;80(1):81–102.

21. Cunnac S, Lindeberg M, Collmer A. Pseudomonas syringae type III secretion system effectors: repertoires in search of functions. Curr Opin Microbiol. 2009 Feb 1;12(1):53–60.

22. Klement Z, Goodman RN. The Hypersensitive Reaction to Infection by Bacterial Plant Pathogens. Annu Rev Phytopathol. 1967 Sep;5(1):17–44.

23. Bundalovic-Torma C, Lonjon F, Desveaux D, Guttman DS. Diversity, Evolution, and Function of Pseudomonas syringae Effectoromes. Annu Rev Phytopathol. 2022;60(1):211–36.

24. Deng WL, 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. J Bacteriol. 2003 Apr 15;185(8):2592–602.

25. 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 Biol. 2019 Jan 3;20(1):3.

26. Morris CE, Lamichhane JR, Nikolić I, Stanković S, Moury B. The overlapping continuum of host range among strains in the Pseudomonas syringae complex. Phytopathol Res. 2019 Dec 16;1(1):4.

27. Monteil CL, Cai R, Liu H, Mechan Llontop ME, Leman S, Studholme DJ, et al. Nonagricultural reservoirs contribute to emergence and evolution of Pseudomonas syringae crop pathogens. New Phytol. 2013;199(3):800–11.

28. Bender CL, Alarcón-Chaidez F, Gross DC. Pseudomonas syringae Phytotoxins: Mode of Action, Regulation, and Biosynthesis by Peptide and Polyketide Synthetases. Microbiol Mol Biol Rev. 1999;63(2):266–92.

29. Arrebola E, Cazorla FM, Durán VE, Rivera E, Olea F, Codina JC, et al. Mangotoxin: a novel antimetabolite toxin produced by Pseudomonas syringae inhibiting ornithine/arginine biosynthesis. Physiol Mol Plant Pathol. 2003 Sep 1;63(3):117–27.

30. Caballo-Ponce E, Van Dillewijn P, Wittich RM, Ramos C. WHOP, a Genomic Region Associated With Woody Hosts in the Pseudomonas syringae Complex Contributes to the Virulence and Fitness of Pseudomonas savastanoi pv. savastanoi in Olive Plants. Ornston. 2017;30(2):113.

31. Saint-Vincent PM, Ridout M, Engle NL, Lawrence TJ, Yeary ML, Tschaplinski TJ, et al. Isolation, Characterization, and Pathogenicity of Two Pseudomonas syringae Pathovars from Populus trichocarpa Seeds. Microorganisms. 2020 Aug;8(8):1137.

32. Turco S, Zuppante L, Drais MI, Mazzaglia A. Dressing like a pathogen: Comparative analysis of different Pseudomonas genomospecies wearing different features to infect Corylus avellana. J Phytopathol. 2022;170(7–8):504–16.

33. Lamichhane JR, Varvaro L, Parisi L, Audergon JM, Morris CE. Chapter Four - Disease and Frost Damage of Woody Plants Caused by Pseudomonas syringae: Seeing the Forest for the Trees. In: Sparks DL, editor. Advances in Agronomy [Internet]. Academic Press; 2014 [cited 2023 Oct 20]. p. 235–95. Available from: https://www.sciencedirect.com/science/article/pii/B9780128001325000043

34. Van der Plank JE. Plant diseases : epidemics and control. Academic Press; 1963. 349 p.

35. Xiang Q, Lott AA, Assmann SM, Chen S. Advances and perspectives in the metabolomics of stomatal movement and the disease triangle. Plant Sci. 2021;302:110697.

36. Phyllosphere bacterial strains Rhizobium b1 and Bacillus subtilis b2 control tomato leaf diseases caused by Pseudomonas syringae pv. tomato and Alternaria solani | Journal of Applied Microbiology | Oxford Academic [Internet]. [cited 2023 Sep 21]. Available from: https://academic.oup.com/jambio/article/134/7/lxad139/7221653

37. Ehau-Taumaunu H, Hockett KL. Passaging Phyllosphere Microbial Communities Develop Suppression Towards Bacterial Speck Disease in Tomato. Phytobiomes J. 2022 Jun 20;PBIOMES-05-22-0030-FI.

38. Morella NM, Weng FCH, Joubert PM, Jessica C, Lindow S, Koskella B. Successive passaging of a plant-associated microbiome reveals robust habitat and host genotype-dependent selection. Proc Natl Acad Sci U S A. 2020;117(2):1148–59.

39. Innerebner G, Knief C, Vorholt JA. Protection of Arabidopsis thaliana against Leaf-Pathogenic Pseudomonas syringae by Sphingomonas Strains in a Controlled Model System. Appl Environ Microbiol. 2011 May 15;77(10):3202–10.

40. Mannaa M, Seo YS. Plants under the Attack of Allies: Moving towards the Plant Pathobiome Paradigm. Plants. 2021 Jan;10(1):125.

41. Case TJ, Gilpin ME. Interference Competition and Niche Theory. Proc Natl Acad Sci. 1974 Aug;71(8):3073–7.

42. Darbandi A, Asadi A, Mahdizade Ari M, Ohadi E, Talebi M, Halaj Zadeh M, et al. Bacteriocins: Properties and potential use as antimicrobials. J Clin Lab Anal. 2021 Dec 1;36(1):e24093.

43. Weaver SL, Zhu L, Ravishankar S, Clark M, Baltrus DA. Interspecies killing activity of Pseudomonas syringae tailocins. Microbiology. 2022;168(11):001258.

44. Ghequire MGK, Mot RD. The Tailocin Tale: Peeling off Phage Tails. Trends Microbiol. 2015 Oct 1;23(10):587–90.

45. Carim S, Azadeh AL, Kazakov AE, Price MN, Walian PJ, Lui LM, et al. Systematic discovery of pseudomonad genetic factors involved in sensitivity to tailocins. ISME J. 2021;1–17.

46. Gardner A, West SA, Buckling A. Bacteriocins, spite and virulence. Proc R Soc B Biol Sci. 2004 Jul 22;271(1547):1529–35.

47. Martínez-Cuesta MC, Kok J, Herranz E, Peláez C, Requena T, Buist G. Requirement of Autolytic Activity for Bacteriocin-Induced Lysis. Appl Environ Microbiol. 2000 Aug;66(8):3174–9.

48. Cascales E, Buchanan SK, Duché D, Kleanthous C, Lloubès R, Postle K, et al. Colicin Biology. Microbiol Mol Biol Rev. 2007;71(1):158–229.

49. Gillor O, Vriezen JAC, Riley MA. The role of SOS boxes in enteric bacteriocin regulation. Microbiol Read Engl. 2008;154(Pt 6):1783.

50. Rebuffat S. Bacteriocins from Gram-Negative Bacteria: A Classification? In: Drider D, Rebuffat S, editors. Prokaryotic Antimicrobial Peptides: From Genes to Applications [Internet]. New York, NY: Springer; 2011 [cited 2023 Sep 22]. p. 55–72. Available from: https://doi.org/10.1007/978-1-4419-7692-5_4

51. Ogata S, Choi KH, Ikeda Y, Hongo M. BACTERIAL LYSIS OF *CLOSTRIDIUM* SPECIES. J Gen Appl Microbiol. 1974;20(3):153–68.

52. Bergan T, Ekström B, Nord CE. Ecological Impacts of Antibacterial Agents: Stockholm, March 7–8, 1986. Scand J Infect Dis. 1986 Aug 1;18(sup49):1–203.

53. Michel-Briand Y, Baysse C. The pyocins of Pseudomonas aeruginosa. Biochimie. 2002 May 1;84(5):499–510.

54. Bayramoglu B, Toubiana D, Van Vliet S, Inglis RF, Shnerb N, Gillor O. Bet-hedging in bacteriocin producing Escherichia coli populations: the single cell perspective. Sci Rep 2017 71. 2017;7(1):1–10.

55. Mavridou DAI, Gonzalez D, Kim W, West SA, Correspondence KRF. Bacteria Use Collective Behavior to Generate Diverse Combat Strategies. Curr Biol. 2018;28:345–55.

56. Sarkar SF, Guttman DS. Evolution of the Core Genome of Pseudomonas syringae, a Highly Clonal, Endemic Plant Pathogen. Appl Environ Microbiol. 2004 Apr;70(4):1999–2012.

57. Borschinger B, Bartoli C, Chandeysson C, Guilbaud C, Parisi L, Bourgeay JF, et al. A set of PCRs for rapid identification and characterization of Pseudomonas syringae phylogroups. J Appl Microbiol. 2016 Mar 1;120(3):714–23.

58. Guilbaud C, Morris CE, Barakat M, Ortet P, Berge O. Isolation and identification of Pseudomonas syringae facilitated by a PCR targeting the whole P. syringae group. FEMS Microbiol Ecol. 2016;92:146.

59. 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(9):e105547.

60. Dutta B, Gitaitis R, Agarwal G, Coutinho T, Langston D. Pseudomonas coronafaciens sp. nov., a new phytobacterial species diverse from Pseudomonas syringae. PLoS ONE. 2018 Dec 1;13(12).

61. Keshtkar AR, Khodakaramian G, Rouhrazi K. Isolation and characterization of Pseudomonas syringae pv. syringae which induce leaf spot on walnut. Eur J Plant Pathol 2016 1464. 2016 May 16;146(4):837–46.

62. Moretti C, Fakhr R, Buonaurio R. Calendula officinalis: A New Natural Host of Pseudomonas viridiflava in Italy. Httpsdoiorg101094PDIS-08-11-0691. 2012 Jan 11;96(2):285–285.

63. Hwang MSH, Morgan RL, Sarkar SF, Wang PW, Guttman DS. Phylogenetic characterization of virulence and resistance phenotypes of Pseudomonas syringae. Appl Environ Microbiol. 2005 Sep;71(9):5182–91.

64. 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. Appl Environ Microbiol. 2008 May;74(10):3171–81.

65. Parkinson N, Bryant R, Bew J, Elphinstone J. Rapid phylogenetic identification of members of the Pseudomonas syringae species complex using the rpoD locus. Plant Pathol. 2011 Apr 1;60(2):338–44.

66. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 2020 386. 2020 Jun 1;38(6):685–8.

67. 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 Phytol. 2018 Jul 1;219(2):672–96.

68. Dillon MM, Almeida RND, Laflamme B, Martel A, Weir BS, Desveaux D, et al. Molecular evolution of Pseudomonas syringae type iii secreted effector proteins. Front Plant Sci. 2019 Mar 22;10:418.

69. Ozer EA. in_silico_pcr. 2020.

70. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 2018 91. 2018 Nov 30;9(1):1–8.

71. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013 Apr 1;30(4):772–80.

72. Wright ES. DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics. 2015 Oct 6;16(1):1–14.

73. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019 Aug 1;37(8):852–7.

74. 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. Front Microbiol. 2017 Dec 7;8(DEC):2422.

75. Tian L, Huang C, Mazloom R, Heath LS, Vinatzer BA. LINbase: a web server for genome-based identification of prokaryotes as members of crowdsourced taxa. Nucleic Acids Res. 2020 Jul 2;48(W1):W529–37.

76. Laflamme B, Dillon MM, Martel A, Almeida RND, Desveaux D, Guttman DS. The pan-genome effector-triggered immunity landscape of a host-pathogen interaction. Science. 2020 Feb 14;367(6479):763–8.

77. Eddy SR. HMMER [Internet]. 2020 [cited 2022 Oct 31]. Available from: www.hmmer.org

78. Lindeberg M, Cunnac S, Collmer A. The evolution of Pseudomonas syringae host specificity and type III effector repertoires. Mol Plant Pathol. 2009 Nov 1;10(6):767–75.

79. Baltrus DA, Nishimura MT, Dougherty KM, Biswas S, Mukhtar MS, Vicente J, et al. The Molecular Basis of Host Specialization in Bean Pathovars of Pseudomonas syringae. Httpdxdoiorg101094MPMI-08-11-0218. 2012 Jun 5;25(7):877–88.

80. Ferrante P, Clarke CR, Cavanaugh KA, Michelmore RW, Buonaurio R, Vinatzer BA. Contributions of the effector gene hopQ1-1 to differences in host range between Pseudomonas syringae pv. phaseolicola and P. syringae pv. tabaci. Mol Plant Pathol. 2009 Nov 1;10(6):837–42.

81. Baltrus DA, McCann HC, Guttman DS. Evolution, genomics and epidemiology of Pseudomonas syringae: Challenges in Bacterial Molecular Plant Pathology. Vol. 18, Molecular Plant Pathology. Blackwell Publishing Ltd; 2017. p. 152–68.

82. 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. ISME J. 2008;2:321–34.

83. Morris CE, Kinkel LL, Xiao K, Prior P, Sands DC. Surprising niche for the plant pathogen Pseudomonas syringae. Infect Genet Evol. 2007 Jan 1;7(1):84–92.

84. Cunty A, Cesbron S, Poliakoff F, Jacques MA, Manceau C. Origin of the outbreak in France of Pseudomonas syringae pv. actinidiae biovar 3, the causal agent of bacterial canker of kiwifruit, revealed by a multilocus variable-number tandem-repeat analysis. Appl Environ Microbiol. 2015;81(19):6773–89.

85. Zhao M, Tyson C, Chen HC, Paudel S, Gitaitis R, Kvitko B, et al. Pseudomonas alliivorans sp. nov., a plant-pathogenic bacterium isolated from onion foliage in Georgia, USA. Syst Appl Microbiol. 2022 Jan 1;45(1):126278.

86. Preston GM. Pseudomonas syringae pv. tomato: the right pathogen, of the right plant, at the right time. Mol Plant Pathol. 2000 Sep 1;1(5):263–75.

87. Morris CE, Monteil CL, Berge O. The Life History of Pseudomonas syringae : Linking Agriculture to Earth System Processes . Annu Rev Phytopathol. 2013 Aug 4;51(1):85–104.

88. Ziemski M, Wisanwanichthan T, Bokulich NA, Kaehler BD. Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences. Front Microbiol. 2021;12.

89. 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(9).

90. Vinatzer BA, Weisberg AJ, Monteil CL, Elmarakeby HA, Sheppard SK, Heath LS. A proposal for a genome similarity-based taxonomy for plant-pathogenic bacteria that is sufficiently precise to reflect phylogeny, host range, and outbreak affiliation applied to Pseudomonas syringae sensu lato as a proof of concept. Phytopathology. 2017 Jan 1;107(1):18–28.

91. Young JM, Bull CT, De Boer SH, Firrao G, Gardan L, Saddler GE, et al. International Standards for Naming Pathovars of Phytopathogenic Bacteria. 2001 [cited 2023 Mar 28]. ISPP. Available from: https://www.isppweb.org/about_tppb_naming.asp

92. Ménard M, Sutra L, Luisetti J, Prunier JP, Gardan L. Pseudomonas syringae pv. avii (pv. nov.), the Causal Agent of Bacterial Canker of Wild Cherries (Prunus avium) in France. Eur J Plant Pathol 2003 1096. 2003 Jul;109(6):565–76.

93. Caballo-Ponce E, Pintado A, Moreno-Pcrez A, Murillo J, Smalla K, Ramos C. Pseudomonas savastanoi pv. mandevillae pv. nov., a clonal pathogen causing an emerging, devastating disease of the ornamental plant Mandevilla spp. Phytopathology. 2021 Aug 1;111(8):1277–88.

94. Young JM. An overview of bacterial nomenclature with special reference to plant pathogens. Syst Appl Microbiol. 2008 Dec 1;31(6–8):405–24.

95. Lamichhane JR, Messéan A, Morris CE. Insights into epidemiology and control of diseases of annual plants caused by the Pseudomonas syringae species complex. Vol. 81, Journal of General Plant Pathology. Springer Tokyo; 2015. p. 331–50.

96. Moreno-Pérez A, Pintado A, Murillo J, Caballo-Ponce E, Tegli S, Moretti C, et al. Host Range Determinants of Pseudomonas savastanoi Pathovars of Woody Hosts Revealed by Comparative Genomics and Cross-Pathogenicity Tests. Front Plant Sci. 2020 Jul 2;11:973.

97. Almeida RND, Greenberg M, Bundalovic-Torma C, Martel A, Wang PW, Middleton MA, et al. Predictive modeling of Pseudomonas syringae virulence on bean using gradient boosted decision trees. PLOS Pathog. 2022 Jul 1;18(7):e1010716.

98. Fautt C, Hockett KL, Couradeau E. Evaluation of the taxonomic accuracy and pathogenicity prediction power of 16 primer sets amplifying single copy marker genes in the Pseudomonas syringae species complex. Mol Plant Pathol. 2023;24(8):989–98.

99. NCBI assembly resource [Internet]. 2021. Available from: https://www.ncbi.nlm.nih.gov/assembly?term=all%5Bfilter%5D AND %28%22Pseudomonas syringae group%22%5BORGN%5D%29&cmd=DetailsSearch

100. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.

101. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2020 Mar 1;36(6):1925–7.

102. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE. 2010 Mar 10;5(3):e9490.

103. Dotmatics. Geneious [Internet]. 2022 [cited 2022 Oct 31]. Available from: https://www.geneious.com/

104. Eddy SR. HMMER [Internet]. 2020 [cited 2022 Oct 31]. Available from: www.hmmer.org

105. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct;31(19):3210–2.

106. Fautt C. cwf30/SYRINGAE: Official release 1.0 (Official). 2023;

107. Fautt C. Supplementary data for Competition and Virulence in Pseudomonas syringae [Internet]. [cited 2023 Oct 23]. Available from: https://zenodo.org/records/10035485

108. Vorholt JA. Microbial life in the phyllosphere. Nat Rev Microbiol. 2012 Dec;10(12):828–40.

109. Thompson IP, Bailey MJ, Fenlon JS, Fermor TR, Lilley AK, Lynch JM, et al. Quantitative and qualitative seasonal changes in the microbial community from the phyllosphere of sugar beet (Beta vulgaris). Plant Soil. 1993 Mar 1;150(2):177–91.

110. Rico A, McCraw SL, Preston GM. The metabolic interface between Pseudomonas syringae and plant cells. Curr Opin Microbiol. 2011 Feb 1;14(1):31–8.

111. Berg M, Koskella B. Nutrient- and Dose-Dependent Microbiome-Mediated Protection against a Plant Pathogen. Curr Biol. 2018 Aug;28(15):2487-2492.e3.

112. Morella NM, Zhang X, Koskella B. Tomato Seed-Associated Bacteria Confer Protection of Seedlings Against Foliar Disease Caused by Pseudomonas syringae. Phytobiomes J. 2019 Jan;3(3):177–90.

113. Baltrus DA, McCann HC, Guttman DS. Evolution, genomics and epidemiology of Pseudomonas syringae. Mol Plant Pathol. 2017;18(1):152–68.

114. Vinatzer BA, Monteil CL, Clarke CR. Harnessing population genomics to understand how bacterial pathogens emerge, adapt to crop hosts, and disseminate. Annu Rev Phytopathol. 2014;52:19–43.

115. Hockett KL, Renner T, Baltrus DA. Independent Co-Option of a Tailed Bacteriophage into a Killing Complex in Pseudomonas. mBio [Internet]. 2015 Aug 11 [cited 2023 Aug 13]; Available from: https://journals.asm.org/doi/10.1128/mbio.00452-15

116. Taslem Mourosi J, Awe A, Guo W, Batra H, Ganesh H, Wu X, et al. Understanding Bacteriophage Tail Fiber Interaction with Host Surface Receptor: The Key "Blueprint" for Reprogramming Phage Host Range. Int J Mol Sci. 2022 Oct 12;23(20):12146.

117. Islam MZ, Fokine A, Mahalingam M, Zhang Z, Garcia-Doval C, Raaij MJ van, et al. Molecular anatomy of the receptor binding module of a bacteriophage long tail fiber. PLOS Pathog. 2019 Dec 19;15(12):e1008193.

118. Dunne M, Prokhorov NS, Loessner MJ, Leiman PG. Reprogramming bacteriophage host range: design principles and strategies for engineering receptor binding proteins. Curr Opin Biotechnol. 2021 Apr 1;68:272–81.

119. Yehl K, Lemire S, Yang AC, Ando H, Mimee M, Torres MDT, et al. Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. Cell. 2019 Oct 3;179(2):459-469.e9.

120. Fernandez M, Godino A, Príncipe A, Morales GM, Fischer S. Effect of a Pseudomonas fluorescens tailocin against phytopathogenic Xanthomonas observed by atomic force microscopy. J Biotechnol. 2017 Aug 20;256:13–20.

121. Yao GW, Duarte I, Le TT, Carmody L, LiPuma JJ, Young R, et al. A Broad-Host-Range Tailocin from Burkholderia cenocepacia. Appl Environ Microbiol. 2017 May 1;83(10):e03414-16.

122. Veesler D, Cambillau C. A Common Evolutionary Origin for Tailed-Bacteriophage Functional Modules and Bacterial Machineries. Microbiol Mol Biol Rev [Internet]. 2011 Sep 1 [cited 2023 Aug 13]; Available from: https://journals.asm.org/doi/10.1128/mmbr.00014-11

123. Sandmeler H. Acquisition and rearrangement of sequence motifs in the evolution of bacteriophage tail fibres. Mol Microbiol. 1994;12(3):343–50.

124. Baltrus DA, Clark M, Smith C, Hockett KL. Localized recombination drives diversification of killing spectra for phage-derived syringacins. ISME J. 2019 Feb;13(2):237–49.

125. Jayaraman J, Jones WT, Harvey D, Hemara LM, McCann HC, Yoon M, et al. Variation at the common polysaccharide antigen locus drives lipopolysaccharide diversity within the Pseudomonas syringae species complex. Environ Microbiol. 2020 Dec;22(12):5356–72.

126. Kandel PP, Baltrus DA, Hockett KL. Pseudomonas Can Survive Tailocin Killing via Persistence-Like and Heterogenous Resistance Mechanisms. J Bacteriol [Internet]. 2020 Apr 20 [cited 2023 Aug 13]; Available from: https://journals.asm.org/doi/10.1128/jb.00142-20

127. Hockett K, Clark M, Scott S, Baltrus D. Conditionally Redundant Bacteriocin Targeting by<i>Pseudomonas syringae</i [Internet]. 2017 [cited 2023 Aug 13]. Available from: https://doi.org/10.1101/167593

128. Baltrus DA, Weaver S, Krings L, Nguyen AE. Genomic Correlates of Tailocin Sensitivity in Pseudomonas syringae [Internet]. bioRxiv; 2023 [cited 2023 Aug 13]. p. 2023.04.24.538177. Available from: https://www.biorxiv.org/content/10.1101/2023.04.24.538177v1

129. Williams SR, Gebhart D, Martin DW, Scholl D. Retargeting R-Type Pyocins To Generate Novel Bactericidal Protein Complexes. Appl Environ Microbiol. 2008 Jun;74(12):3868–76.

130. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007 Mar 15;23(6):673–9.

131. Dotmatics. Geneious [Internet]. 2022 [cited 2022 Oct 31]. Available from: https://www.geneious.com/

132. Sun S, Rao VB, Rossmann MG. Genome packaging in viruses. Curr Opin Struct Biol. 2010 Feb 1;20(1):114–20.

133. Buth SA, Shneider MM, Scholl D, Leiman PG. Structure and Analysis of R1 and R2 Pyocin Receptor-Binding Fibers. Viruses. 2018 Aug 14;10(8):427.

134. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

135. Sarris PF, Trantas EA, Mpalantinaki E, Ververidis F, Goumas DE. Pseudomonas viridiflava, a Multi Host Plant Pathogen with Significant Genetic Variation at the Molecular Level. PLoS ONE. 2012 Apr 27;7(4):e36090.

136. 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 Sep 3;9(9):e105547.

137. Salazar AJ, Sherekar M, Tsai J, Sacchettini JC. R pyocin tail fiber structure reveals a receptor-binding domain with a lectin fold. PLoS ONE. 2019 Feb 5;14(2):e0211432.

138. Patz S, Becker Y, Richert-Pöggeler KR, Berger B, Ruppel S, Huson DH, et al. Phage tail-like particles are versatile bacterial nanomachines – A mini-review. J Adv Res. 2019;19:75–84.

139. Weigele PR, Scanlon E, King J. Homotrimeric, β-Stranded Viral Adhesins and Tail Proteins. J Bacteriol. 2003 Jul;185(14):4022–30.

140. Holtappels D, Kerremans A, Busschots Y, Van Vaerenbergh J, Maes M, Lavigne R, et al. Preparing for the KIL: Receptor Analysis of Pseudomonas syringae pv. porri Phages and Their Impact on Bacterial Virulence. Int J Mol Sci. 2020 Apr 22;21(8):2930.

141. Pinheiro LAM, Pereira C, Frazão C, Balcão VM, Almeida A. Efficiency of Phage φ6 for Biocontrol of Pseudomonas syringae pv. syringae: An in Vitro Preliminary Study. Microorganisms. 2019 Sep;7(9):286.

142. Rombouts S, Volckaert A, Venneman S, Declercq B, Vandenheuvel D, Allonsius CN, et al. Characterization of Novel Bacteriophages for Biocontrol of Bacterial Blight in Leek Caused by Pseudomonas syringae pv. porri. Front Microbiol [Internet]. 2016 [cited 2023 Aug 14];7. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2016.00279

143. Hulin MT, Rabiey M, Zeng Z, Vadillo Dieguez A, Bellamy S, Swift P, et al. Genomic and functional analysis of phage-mediated horizontal gene transfer in Pseudomonas syringae on the plant surface. New Phytol. 2023 Feb;237(3):959–73.

144. Abedon ST, Danis-Wlodarczyk KM, Wozniak DJ. Phage Cocktail Development for Bacteriophage Therapy: Toward Improving Spectrum of Activity Breadth and Depth. Pharmaceuticals. 2021 Oct 3;14(10):1019.

145. Benítez-Chao DF, León-Buitimea A, Lerma-Escalera JA, Morones-Ramírez JR. Bacteriocins: An Overview of Antimicrobial, Toxicity, and Biosafety Assessment by in vivo Models. Front Microbiol. 2021 Apr 15;12:630695.

146. Riley MA, Wertz JE. Bacteriocins: evolution, ecology, and application. Annu Rev Microbiol. 2002;56:117–37.

147. Bruce JB, West SA, Griffin AS. Bacteriocins and the assembly of natural Pseudomonas fluorescens populations. J Evol Biol. 2017;30(2):352–60.

148. Zheng J, Gänzle MG, Lin XB, Ruan L, Sun M. Diversity and dynamics of bacteriocins from human microbiome. Environ Microbiol. 2015 Jun;17(6):2133–43.

149. Riley MA, Gordon DM. The ecological role of bacteriocins in bacterial competition. Elsevier Ltd; 1999.

150. Bucci V, Nadell CD, Xavier JB. The Evolution of Bacteriocin Production in Bacterial Biofilms. Am Nat. 2011;178(6):E162–73.

151. Inglis RF, Gardner A, Cornelis P, Buckling A. Spite and virulence in the bacterium Pseudomonas aeruginosa. Proc Natl Acad Sci. 2009 Apr 7;106(14):5703–7.

152. Frank SA. Spatial polymorphism of bacteriocins and other allelopathic traits. Evol Ecol. 1994;8(4):369–86.

153. Inglis RF, Roberts PG, Gardner A, Buckling A. Spite and the Scale of Competition in Pseudomonas aeruginosa. https://doi.org/101086/660827. 2015;178(2):276–85.

154. Chao L, Levin BR. Structured habitats and the evolution of anticompetitor toxins in bacteria. Proc Natl Acad Sci. 1981 Oct;78(10):6324–8.

155. Ehau-Taumaunu H, Hockett KL. The plant host environment influences competitive interactions between bacterial pathogens. Environ Microbiol Rep. 2022 Oct;14(5):785–94.

156. Ghoul M, Mitri S. The Ecology and Evolution of Microbial Competition. Trends Microbiol. 2016 Oct;24(10):833–45.

157. Maslowska KH, Makiela-Dzbenska K, Fijalkowska IJ. The SOS system: A complex and tightly regulated response to DNA damage. Environ Mol Mutagen. 2019 May;60(4):368.

158. Zhang APP, Pigli YZ, Rice PA. Structure of the LexA-DNA complex and implications for SOS box measurement. Nature. 2010 Aug 12;466(7308):883–6.

159. Friedman N, Vardi S, Ronen M, Alon U, Stavans J. Precise Temporal Modulation in the Response of the SOS DNA Repair Network in Individual Bacteria. PLOS Biol. 2005 Jun 21;3(7):e238.

160. Barshishat S, Elgrably-Weiss M, Edelstein J, Georg J, Govindarajan S, Haviv M, et al. OxyS small RNA induces cell cycle arrest to allow DNA damage repair. EMBO J. 2018 Feb;37(3):413–26.

161. Erental A, Kalderon Z, Saada A, Smith Y, Engelberg-Kulka H. Apoptosis-like death, an extreme SOS response in Escherichia coli. mBio. 2014 Jul 15;5(4):e01426-01414.

162. Goodman MF, McDonald JP, Jaszczur MM, Woodgate R. Insights into the complex levels of regulation imposed on Escherichia coli DNA polymerase V. DNA Repair. 2016 Aug 1;44:42–50.

163. Butala M, Žgur-Bertok D, Busby SJW. The bacterial LexA transcriptional repressor. Cell Mol Life Sci. 2008 Aug 26;66(1):82.

164. Lu FM, Chak KF. Two overlapping SOS-boxes in ColE operons are responsible for the viability of cells harboring the Col plasmid. Mol Gen Genet MGG. 1996 Jun 1;251(4):407–11.

165. Ghazaryan L, Tonoyan L, Ashhab AA, Soares MIM, Gillor O. The role of stress in colicin regulation. Arch Microbiol. 2014 Nov;196(11):753–64.

166. Mahony DE. Induction of Bacteriocins from Clostridium perfringens by Treatment with Mitomycin C. Antimicrob Agents Chemother. 1977 Jun;11(6):1067–8.

167. Niehus R, Oliveira NM, Li A, Fletcher AG, Foster KR. The evolution of strategy in bacterial warfare via the regulation of bacteriocins and antibiotics. Bitbol AF, Walczak AM, Kümmerli R, editors. eLife. 2021 Sep 7;10:e69756.

168. Erjavec N, Cvijovic M, Klipp E, Nyström T. Selective benefits of damage partitioning in unicellular systems and its effects on aging. Proc Natl Acad Sci. 2008 Dec 2;105(48):18764–9.

169. Kysela DT, Brown PJB, Huang KC, Brun YV. Biological Consequences and Advantages of Asymmetric Bacterial Growth. Annu Rev Microbiol. 2013;67(1):417–35.

170. Lindner AB, Madden R, Demarez A, Stewart EJ, Taddei F. Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. Proc Natl Acad Sci. 2008 Feb 26;105(8):3076–81.

171. Raghunathan S, Chimthanawala A, Krishna S, Vecchiarelli AG, Badrinarayanan A. Asymmetric chromosome segregation and cell division in DNA damage-induced bacterial filaments. Mol Biol Cell. 2020 Dec 15;31(26):2920–31.

172. Chao L, Rang CU, Proenca AM, Chao JU. Asymmetrical Damage Partitioning in Bacteria: A Model for the Evolution of Stochasticity, Determinism, and Genetic Assimilation. PLOS Comput Biol. 2016 Jan 13;12(1):e1004700.

173. Lindemann J, Upper CD. Aerial Dispersal of Epiphytic Bacteria over Bean Plants. Appl Environ Microbiol. 1985 Nov;50(5):1229–32.

174. Kerr B, Riley MA, Feldman MW, Bohannan BJM. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. Nat 2002 4186894. 2002;418(6894):171–4.

175. Nguyen D, Joshi-Datar A, Lepine F, Bauerle E, Olakanmi O, Beer K, et al. Active Starvation Responses Mediate Antibiotic Tolerance in Biofilms and Nutrient-Limited Bacteria. Science. 2011 Nov 18;334(6058):982–6.

176. Gunasekera TS, Paul ND. Ecological impact of solar ultraviolet-B (UV-B: 320–290 nm) radiation on Corynebacterium aquaticum and Xanthomonas sp. colonization on tea phyllosphere in relation to blister blight disease incidence in the field. Lett Appl Microbiol. 2007 May 1;44(5):513–9.

177. Cuny C, Lesbats M, Dukan S. Induction of a Global Stress Response during the First Step of Escherichia coli Plate Growth. Appl Environ Microbiol. 2007 Feb;73(3):885–9.

178. Zangeneh M, Khorrami S, Khaleghi M. Bacteriostatic activity and partial characterization of the bacteriocin produced by L. plantarum sp. isolated from traditional sourdough. Food Sci Nutr. 2020 Sep 13;8(11):6023–30.

179. Garmiri P, Coles KE, Humphrey TJ, Cogan TA. Role of outer membrane lipopolysaccharides in the protection of Salmonella enterica serovar Typhimurium from desiccation damage. FEMS Microbiol Lett. 2008 Apr 1;281(2):155–9.

VITA

Chad Fautt

2013-2016 American River College

2018 B.S. in Biotechnology, University of California, Davis

2023 PhD in Ecology, The Pennsylvania State University

Publications

Kandel, P. P., Naumova, M., Fautt, C., Patel, R. R., Triplett, L. R., & Hockett, K. L. (2022). Genome Mining Shows Ubiquitous Presence and Extensive Diversity of Toxin-Antitoxin Systems in *Pseudomonas syringae*. *Frontiers in microbiology*, *12*, 815911. https://doi.org/10.3389/fmicb.2021.815911

Fautt, C., Couradeau, E., & Hockett, K. L. (2022). SYRINGAE: A web-based application for *Pseudomonas syringae* isolate characterization (p. 2022.11.04.515192). bioRxiv. https://doi.org/10.1101/2022.11.04.515192

Fautt, C., Hockett, K.L. & Couradeau, E. (2023) Evaluation of the taxonomic accuracy and pathogenicity prediction power of 16 primer sets amplifying single copy marker genes in the *Pseudomonas syringae* species complex. *Molecular Plant Pathology*, 24, 989–998. https://doi.org/10.1111/mpp.13337