INFORMATICS APPROACHES EXPLORING BOTH VIRULENCE FACTORS OF *PANTOEA ANANATIS* AND THE HOST-RESISTANCE TO CENTER ROT IN *ALLIUM* SPP.

by

BRENDON K. MYERS

(Under the Direction of Bhabesh Dutta)

ABSTRACT

Pantoea ananatis (PA) is a significant bacterial pathogen responsible for onion center rot (OCR) in bulb onions (*Allium cepa*). The pathogen's virulence in onions is attributed to two major virulence factors: the chromosomal HiVir gene cluster that causes phosphonate toxin-mediated necrosis and the plasmid-borne allicin tolerance (*alt*) gene cluster, which allows the bacterium to overcome onion-derived antimicrobial thiosulfinates. However, the genetic factors involved in infecting non-bulb *Allium* species like leeks (*A. porrum*), Welsh onions (*A. fistulosum*), and chives (*A. schoenoprasum*) remain poorly understood, leading to an incomplete understanding of the PA-*Allium* pathosystem. In our initial objective, 92 PA strains were screened for pathogenicity on *A. fistulosum* × *A. cepa* and *A. porrum*. Our results revealed higher aggressiveness in the hybrid Allium species. At the same time, genome-wide association studies (GWAS) identified 835 genes linked to pathogenicity on *A. fistulosum* × *A. cepa* and 243 genes associated with infection on *A. porrum*. This suggests that PA may utilize a shared set of virulence genes for *Allium* infection but requires host-specific adaptations for non-bulb species. We further validated that the HiVir gene cluster is the primary pathogenicity factor across *A. fistulosum* x *A.*

cepa and *A. porrum*. To explore the diversity of thiosulfinate tolerance gene clusters, we employed Natural Language Processing (NLP)-like deep learning techniques to identify *alt*-like gene clusters across 238,362 bacterial genomes. The model discovered 47 novel *alt*-like clusters, 15 of which we experimentally validated in PA strains. The results demonstrate the utility of deep learning and language-like processing in uncovering diverse, difficult-to-work-with gene clusters, enabling a greater capacity to investigate PA-Allium interactions. Finally, we screened 982 *Allium* genotypes and identified a resistant *A. cepa* genotype, DPLD 19-39, consistently exhibiting reduced foliar necrosis and bulb rot. Transcriptomic analysis indicated that potential host resistance against PA is mediated by cell wall fortification, reactive oxygen species (ROS) regulation, and programmed cell death, potentially blocking PA from invading the tissues instead of an aggressive immune response. These defense mechanisms provide key targets for breeding programs to develop optimized PA-resistant onion genotypes.

INDEX WORDS: *Pantoea ananatis,* Onion center rot, *Allium*s, Allicin tolerance, Genome-wide association (GWAS), Natural Language Processing (NLP), Disease resistance, Transcriptomics

INFORMATICS APPROACHES EXPLORING BOTH VIRULENCE FACTORS OF *PANTOEA ANANATIS* AND THE HOST-RESISTANCE TO CENTER ROT IN *ALLIUM* SPP.

by

BRENDON K. MYERS

B.S., Washington State University, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2024

INFORMATICS APPROACHES EXPLORING BOTH VIRULENCE FACTORS OF *PANTOEA ANANATIS* AND THE HOST-RESISTANCE TO CENTER ROT IN *ALLIUM* SPP.

by

BRENDON K. MYERS

Major Professor:	Bhabesh Dutta
Committee:	Brian Kvitko Stephanus Venter James Leebens-Mack

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2024

Dedication

To my feline companion, whom I have dragged across the country and the state of Georgia multiple times, Revan Myers.

Acknowledgements

I am deeply grateful to my academic advisor, Dr. Bhabesh Dutta, whose unwavering support and insightful discussions were invaluable throughout my doctoral studies. His dedication refined my scientific and professional acumen and significantly enhanced the quality of my research and personal life.

I am equally thankful to each member of my doctoral committee: Dr. Brian Kvitko, Dr. Stephanus Venter, and Dr. James Leebens-Mack for their critical insights and timely responses which greatly enriched my work.

I am grateful to both past and present members of the Dutta lab. Their collaboration in experimental design, data collection, and insightful advice was instrumental.

Further, I acknowledge the support and resources provided by the Georgia Advanced Computing Resource Center (GACRC), which have been pivotal in advancing my research capabilities. The GACRC staff's expertise and readiness to assist with computational challenges greatly enhanced my data analysis efforts, and for this, I am profoundly thankful.

Lastly, my heartfelt appreciation goes to my family for their constant encouragement and support during this journey. Special thanks must be extended to my mother, Ena, whose resilience and support remain unparalleled.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Onion center rot (OCR) presents a significant economic threat to onion growers locally, nationally, and globally. In the U.S. state of Georgia, onions are a valuable vegetable commodity, with an average farm gate value of $145 million (USDA, 2020). *Pantoea ananatis* (*PA*)-induced OCR is a consistent and significant threat to onion production (Brannen et al., 2017). Despite the substantial economic impact of OCR, the breadth and depth of PA's pathogenicity and virulence mechanisms still need to be better understood (Stumpf et al., 2017; Asselin et al., 2018). Moreover, worldwide reports of both bacterial onion blights and bulb rots caused by the *Pantoea* complex suggest that other species like *Pantoea agglomerans*, *Pantoea dispersa*, and *Pantoea stewartii* contribute to similar issues across onion-growing regions (Brady et al., 2011; Edens et al., 2006; Stumpf et al., 2018; Chang et al., 2018). Previous research has identified the importance of some universal genetic factors in PA pathogenesis, yet host-specific virulence mechanisms remain elusive (Morohoshi et al., 2007; Asselin et al., 2018). By examining other *Allium* species, this research will improve our understanding of the impact of virulence factors across the genus, indirectly aiding the management of OCR in the economically significant *Allium cepa*. Additionally, the recent discovery of *alt* gene clusters across various bacterial genera demands investigation regarding their prevalence and genetic diversity (Stice et al., 2020). Typical sequence-based methodologies struggle to identify these gene clusters, and manual curation of bacterial genomes is inefficient and vulnerable to investigator bias. We aim to overcome these limitations by employing NLP-like computational techniques to vectorize variables inherent to bacterial gene clusters

and enhance the detection and classification of *alt*, *alt-like*, and pseudo-*alt* clusters. Finally, addressing the long-overdue need for resistance screening in *Allium* species against PA is crucial. Genetic resistance is considered the ideal control strategy for this disease, as it would reduce the need for chemical inputs, which are costly and environmentally harmful (Rice et al., 2006; Stumpf et al., 2021). Identifying resistance traits will allow us to discover and characterize resistant onion genotypes. Incorporating these findings into onion breeding programs will provide a sustainable solution for crop protection and the future of the onion industry (Stumpf et al., 2017).

***Pantoea ananatis*-center rot of onion: A Linguistic and Mathematical Framework for Plant Pathology**

PA is a Gram-negative, rod-shaped, facultative anaerobe like other members of the Enterobacteriaceae family (Gitaitis & Gay, 1997; Coutinho & Venter, 2009; De Maayer et al., 2014; Weller-Stuart et al., 2017). Colonies are yellow-pigmented, and the cells utilize glucose in an oxidative and fermentative manner. The bacterium tests positive for the following biochemical assays: β-D-galactosidase and catalase, citrate utilization, and acetoin and indole production. PA tests negative for ornithine decarboxylase, lysine decarboxylase, urease, and oxidase (Gitaitis & Gay, 1997; De Maayer et al., 2014; Weller-Stuart et al., 2017). This bacterium is motile due to peritrichous flagella (Coutinho & Venter, 2009; De Maayer et al., 2014) and is commonly found in soil, water, and plants and is noted for both its pathogenicity in plants and its biotechnological applications (Hara et al., 2012; De Maayer et al., 2014; Weller-Stuart et al., 2017).

We can also view the complex, intricate pathosystem interactions through a language framework, applying Natural Language Processing (NLP) techniques to analyze

biological systems (Oikonomou et al., 2024; Wagner et al., 2022). The pathosystem of PA-derived OCR can be likened to an anthology of languages, where cross-communication between the pathogen PA and the host *Allium cepa* results in a biological reality (Stice et al., 2021; Wagner et al., 2022; Choi & Lee, 2023). Restructuring the simplified "string" of categorical variables as a genome allows more straightforward comparative calculations; however, maintaining the conceptual pairing between the categorical description and its string representation enhances the applicability for downstream analysis in much the same way English words represent conceptual intent, genes represent biological intent and are effectively analogous. The biological intent of PA's OCR is visually represented in symptoms that primarily affect the foliar tissues and internal scales of onions, resulting in internal bulb rot and significantly reducing marketability (Gitaitis & Gay, 1997; Agarwal et al., 2019).

Multiple inoculum sources (seed, weeds, thrips) have been demonstrated to initiate OCR epidemics in fields, all extensions of biological intent as various phenotypes (Walcott et al., 2002; Gitaitis et al., 2002). Studies have detected *P. ananatis* in naturally infested onion seeds from symptomless mother plants, meaning visual inspection alone is insufficient to prevent seed contamination (Walcott et al., 2002). Seed infestation has been confirmed by immunomagnetic separation and polymerase chain reaction (IMS-PCR) assays with species-specific primers (Walcott et al., 2002). The primary mode of transmission, however, involves insect vectors—specifically tobacco thrips (*Frankliniella fusca*) and onion thrips (*Thrips tabaci*), which are common in Georgia's Vidalia onion region (Gitaitis et al., 2013). These thrips acquire and transmit *P. ananatis*, which persists in their gut and is detected in their feces (Gitaitis et al., 2013). The actual infection of the

onion occurs through thrips defecation near feeding wounds, enabling the pathogen to spread into the plant through the wounds (Gitaitis et al., 2013; Santos et al., 2020). PA can survive epiphytically and endophytically on various host plants, including 25 weed species in Georgia; these alternative hosts serve as local inoculum sources (Gitaitis et al., 2002). Environmental factors such as optimal temperatures and leaf wetness influence bacterial persistence and spread in onion fields (Gitaitis et al., 2002). There are many points within this pathosystem where biological intent, genes, and their downstream phenotypes interact in complex ways, all of which must interact to make OCR the problem it is today. Effectively, OCR is a dramatic flashpoint that is just the extension of the biological language of PA, *Alliums*, and the anthropocentric/economic variables that drive human interest in the first place. This concept is not exactly new; it is just a rewording of genomics; however, reframing the mechanism through the lens of information flow, which language is, makes understanding and applying computational methods far easier.

Genomics of *Pantoea ananatis* and *Allium* spp.

The genomic landscape of PA offers a vast and complex repository, with genomes ranging from 4.3 Mb to 5.25 Mb, encoding over 4,000 genes (De Maayer et al., 2014; Weller-Stuart et al., 2017). The genomic architecture typically comprises a large circular chromosome, the large *Pantoea* plasmid (LPP-1), and variable accessory plasmids that provide genomic plasticity (De Maayer et al., 2014). The LPP-1 plasmid ranges from 280.8 to 352.8 Kb and encodes 200-300 genes, some of which house mobile genetic elements, antibiotic resistance genes, and other determinants of host-microbe interactions (Weller-Stuart et al., 2017). The mobile accessory genome underpins the diversity and phenotypic variation across PA strains, particularly in pathogenicity and

environmental fitness (Shyntum et al., 2015; Weller-Stuart et al., 2017). Although it lacks Type III and Type II secretion systems, PA relies on its Type VI secretion system to compete with other bacteria and establish infection in onion leaves (Shyntum et al., 2015).

Onion-pathogenic PA strains utilize a gene cluster known as "HiVir" to produce a phosphonate toxin, which has been linked to onion pathogenicity. Initially, the toxin was indirectly associated with the necrosis on onion leaf and bulb, but later studies confirmed its direct involvement by characterizing the purified toxin (Asselin et al., 2018; Polidore et al., 2021). The HiVir cluster contains core genes, such as *pepM* (*hvrA*), which are essential for phosphonate biosynthesis. This process mimics phosphonate and causes competitive inhibition of critical metabolic enzymes, leading to cell death and the hallmark symptoms of OCR (Asselin et al., 2018; Polidore et al., 2021). Deletion of *hvrC* results in the loss of PA's pathogenicity, whereas deletion of genes such as *hvrK* only leads to modulation of the severity of symptoms. In response to cell death *Allium* species, including onions, employ defense mechanisms based on thiosulfinate phytoanticipins, sulfur-containing compounds that disrupt microbial cell membranes and interfere with their metabolic processes, causing oxidative stress and microbial cell death (Curtis et al., 2004; Barbu et al., 2023). This defense response introduces an additional layer of complexity for the pathogen, which must counteract the oxidative stress induced by thiosulfinates while simultaneously producing phosphonate toxins to infect host cells (Curtis et al., 2004; Barbu et al., 2023). Successful *Allium* pathogens possess mechanisms to reduce sensitivity to thiosulfinates. The *alt* gene cluster confers resistance to thiosulfinates produced by *Allium* species (Stice et al., 2020). In PA, the *alt* cluster contains 11 genes associated with sulfur metabolism, although the exact biochemical

mechanism remains unknown. Similar gene clusters, re-termed as thiosulfinate tolerance gene (TTG) clusters, have been detected in other pathogens, including *Burkholderia* spp. and *Pseudomonas fluorescens* (Borlinghaus et al., 2020; Paudel et al., 2024). However, TTG clusters between these groups are distinct and share little gene sequence similarity or synteny (Stice et al., 2020; Borlinghaus et al., 2020; Paudel et al., 2024).

Additional factors such as quorum sensing and motility were reported to be associated with onion pathogenicity (Morohoshi et al., 2007; Weller-Stuart et al., 2017). In addition to its role in plant pathology, PA has been noted for producing valuable compounds such as exopolysaccharides, carotenoids, and a wide range of catabolic enzymes. PA also demonstrated the ability to degrade environmental pollutants and generate biohydrogen under anaerobic conditions, making it useful for bioremediation, sustainable agriculture, bioenergy production, and industrial applications (Choi et al., 2021; Usuda et al., 2022).

The genomic architecture of *Allium cepa* stands out due to its large size, approximately 16 Gb, one of the largest among cultivated crops (Fu et al., 2019). Unlike PA, which relies on mobile elements and plasmids for genomic plasticity, the complexity of *A. cepa's* genome is driven by its size and a high proportion of repetitive sequences (Fu et al., 2019; Kuhl et al., 2004). Genes within *A. cepa* govern metabolic pathways involved in both the flavor profile and defense against pests and pathogens (Havey & Ghavami, 2018; Kuhl et al., 2004). Repetitive elements, such as long-terminal repeats (LTRs), contribute to the genome's plasticity, facilitating gene duplications that enhance flavor diversity and disease resistance mechanisms (Fu et al., 2019; Chalbi et al., 2023). This genetic variability is crucial for *A. cepa's* adaptation to various environments and the potential for breeding programs to improve traits like disease resistance and crop yield

(Shigyo et al., 2018). The *Allium* genus is a rich repository of biochemical compounds like and organosulfur compounds, which have been proposed as alternatives for antibiotics and pesticides (Bastaki et al., 2021; Iwar et al., 2024). These compounds also have antidiabetic, hepatoprotective, and antiplatelet activity (Bastaki et al., 2021; Iwar et al., 2024).

As mentioned previously, the mechanisms inherent in these systems that govern information flow in PA and *Allium cepa* open opportunities for a wide range of advanced computational analysis (Weller-Stuart et al., 2017; Shigyo et al., 2018).

Genomic Self-Regulation in *P. ananatis* and *Allium* spp.

Autopoiesis is an elegant, though debated, word that reinterprets biological principles through a systems and self-organizational framework, emphasizing how living systems maintain and regenerate themselves (Ruiz-Mirazo & Moreno, 2012). While "biological self-programming" is not a widely accepted term, the author of this thesis would argue that it provides a valuable alternative for conceptualizing these mechanisms through a computational perspective, where the information flow required to maintain and regenerate living systems can be analogous to adaptive programming (De la Fuente, 2021; Bich & Moreno, 2015). This framing simplifies the exploration of how organisms like PA and members of the *Allium* genus manage and adapt their genetic information in response to environmental changes and, by extension, how to take advantage of their properties for downstream problem-solving (Ruiz-Mirazo & Moreno, 2012). Interestingly, the argument made here mirrors a common critique of autopoiesis, which is sometimes considered redundant as a term (Moreno, 1987). However, if we were to accept the definition of autopoiesis as "a network of processes, which produces all the components

whose internal production is necessary to maintain the network operating as a unit" then the author of this thesis would argue that biological self-programming refers more directly to "the mechanisms of information flow responsible for the actualization of biological intent" wherein biological intent refers to the genomic corpus of an organism and the fluid mechanisms that influence and express it; this framework does not rely on organizationally closed systems and may simplify cognitive processing for mechanisms under the umbrella of genetics (Bich & Moreno, 2015). This definition aligns more closely with the concepts of biological autonomy and information processing in biological systems but ideally conveys the problem-solving utility utilized by the author of the thesis (De la Fuente, 2021).

Bacterial genomic adaptability is driven by horizontal gene transfer and high mutation rates (Soucy et al., 2015; Baltrus, 2013). Due to their limited storage capacity for genetic information, bacteria employ rigorous, selective retention and avoidance of specific gene pairings (Croucher & Didelot, 2015). Bacteria tend to evolve through dynamic, modular genomes, where the presence or absence of specific genes in proximity to one another can significantly affect their fitness and virulence (Soucy et al., 2015). This modularity within bacterial genomes creates an inherent vulnerability to methods that exploit gene co-occurrence or avoidance patterns or their presence and absence from a genome (Whelan et al., 2020). Synergistic genes are often co-retained because they optimize the bacterium's survival in specific environments, such as conferring an advantage in host-pathogen interactions (Whelan et al., 2020). Redundant or deleterious gene pairs are consistently avoided and filtered out of the genome due to evolutionary pressures, as their combination reduces overall fitness and genetic efficiency (Whelan et al., 2020).

These are examples of how genomic patterns in bacterial systems self-regulate to balance adaptability with stability (Baltrus, 2013). PA is no different from any other bacterium in this regard, as its genome is also highly plastic and constantly optimizing, making it easier for the analysis of coincidence patterns; by tracing which gene pairs are consistently preserved or avoided across different genomes, crucial genetic interactions are inferred that may underly pathogenicity, antimicrobial resistance or environmental fitness (Whelan et al., 2020; Alyssa & Stavrinides, 2015). This interplay of gene-pair dynamics reflects a form of "biological self-programming," where the bacterium fine-tunes its genomic architecture in what is effectively string redundancy filtering.

In contrast, the various members of the far more genetically complex *Allium* genus, with their slower genomic evolution, exhibit a more stable, long-term form of programming (Van de Peer et al., 2021; Adams, 2007). Genetic traits in *Allium* species are less likely to shift rapidly in response to environmental pressures, with adaptation occurring over extended periods (Sattler et al., 2016). The slower evolution mechanisms of *Alliums* are underpinned by gene duplication, polyploidy, and epigenetic regulation, all of which contribute to the plant's genetic diversity and resilience (Qiao et al., 2019). Gene duplication expands gene families through the altered function of duplicated genes, allowing for functional redundancy in unaltered genes and potentially enhancing a plant's ability to adapt to environmental stresses through metabolic flexibility (Adams, 2007). Polyploidy duplicates entire sets of chromosomes, increasing the overall genetic content; polyploid plants are often more robust and capable of exhibiting greater variation in traits such as growth rates, stress tolerance, and nutrient uptake (Sattler et al., 2016). Epigenetic regulation allows *Alliums* to regulate gene expression without altering the

underlying DNA sequence through mechanisms such as DNA methylation and histone modification, which can be triggered by environmental stimuli and provide a flexible mechanism for the plant to adjust its physiological and biochemical responses (Lämke & Bäurle, 2017). Epigenetic changes are heritable, meaning they can contribute to longer-term adaptation, yet they can also be reversible if environmental conditions change (Lämke & Bäurle, 2017).

The dynamic flow of genetic information within *Pantoea's* genome offers computational opportunities to identify critical vulnerabilities in its genetic architecture; by mapping gene pairings that drive virulence or adaptation, bioinformatics tools can predict which regulatory nodes, if disrupted, would weaken the bacterium's ability to thrive (Whelan et al., 2020; Guevarra et al., 2021). Taking advantage of these mechanisms opens the door to targeted strategies that could interfere with *Pantoea's* pathogenic or virulence programming, reducing its potential; in contrast, computational approaches for *Allium* species can focus on mapping the slower, more stable flow of genetic information related to stress responses and defense pathways (Dahiya et al., 2024). Taking advantage of these slower but more stable genomic mechanisms allows genomic tools such as QTL mapping and marker-assisted selection to optimize resilience and stress tolerance traits without requiring direct genetic modification (Guevarra et al., 2021).

Algorithmic Learning in Bioinformatics

It is difficult to find patterns in large biological datasets (Ernst & Kellis, 2012; Domingos, 2012; Alpaydin, 2020). Unfortunately, biology does not fit neatly into our anthropocentric organizational schemes, so recognition strategies based on human-designed rules can fall short (Domingos, 2012; Bishop, 2006). However, some algorithms can learn their own

rules (Hastie et al., 2009; Svetnik et al., 2004). Machine learning is a subset of artificial intelligence that focuses on developing algorithms to learn from and make predictions or decisions based on data (Bishop, 2006). There are three main phases when utilizing a machine learning model. The training phase occurs when the model learns from a dataset, usually with labeled examples (Alpaydin, 2020). The model will make predictions and adjust the weight parameters to minimize the difference between predicted and "true" output (Hastie et al., 2009). The validation step involves fine-tuning the model parameters to prevent overfitting, otherwise known as when a model trains too well and cannot be generalized (Domingos, 2012). Finally, the model is tested on the new dataset to evaluate its performance, which helps gauge how generalizable the model is to real-world scenarios (Shmueli, 2010).

However, some datasets are too complex for a simple machine-learning architecture to handle (Haykin, 2009). For answers on handling such datasets, we turn to the most advanced computational hardware on the planet: the brain's structure. The human brain has the capacity to solve problems due in part to its compartmentalization and prodigious application of biological neural networks (Domingos, 2012; Hinton & Salakhutdinov, 2006). Computational neural networks are a subclass of machine learning designed to mimic this biological blueprint, and much like neural associations, training, and exercise can strengthen or delete "synapses" (Haykin, 2009; Hinton & Salakhutdinov, 2006). A typical neural network comprises layers with interconnected "neurons" as nodes (LeCun et al., 2015). Each connection has an associated weight, adjusted during training (Hornik et al., 1989). The input layer receives data, the hidden layers are intermediate layers where different representations of the input data are formed, and the output layer

produces the final prediction (LeCun et al., 2015). Data flows from input to output, where each neuron will process the input via performing a weighted sum, followed by a non-linear operation (Goodfellow et al., 2016).

There are several kinds of neural networks. The simplest are feedforward neural networks (FNN) that do not form a cycle between nodes. FNNs are typically used for pattern recognition or classification (Bishop, 2006). Convolutional neural networks (CNN) tend to be used for processing grid-like data, such as images, and convolutional layers are applied to filter spatial hierarchies in data (He et al., 2015). Recurrent neural networks (RNN) are designed to work with sequence data such as text, gene sequences, or speech; for RNNs, connections between the nodes form a cycle, allowing a data memory of previous inputs that enables RNNs to be useful for tasks where context and sequential order are important (Hochreiter & Schmidhuber, 1997). Autoencoders are used for unsupervised learning tasks by compressing input into a latent-space representation and reconstructing the output from this representation (Trabelsi et al., 2018). Generative adversarial networks (GANs) comprise two neural networks: a generator and a discriminator (Goodfellow et al., 2014). These two networks compete against each other, where the generator learns how to create data resembling the training set while the discriminator learns to distinguish generated data from real data (Goodfellow et al., 2014).

Despite the attempt to mimic our systems for thinking, AI and ML computers are only capable of formal thinking, where a set of rules and symbols must be used to generate statements and proofs (Turing, 1936; Gödel, 1931). Gödel proved that in any formal system with sufficient resources, there are statements that are true but cannot be proven within the system (Gödel, 1931). Turing further showed that it is impossible for a

general algorithm to determine whether any given program will run indefinitely or halt (Turing, 1936). With these results, formal systems have inherent limitations in self-reflection and self-understanding and, by extension, are incapable of a true holistic understanding of a given problem (Bostrom, 2014). Humans, in contrast, can engage in non-formal thinking where logical reasoning, intuition, emotions, and previous experiences can cross-communicate to produce more holistic thinking (Damasio, 1994).

Further, this non-formal thinking leads to self-reflection, or the ability to think about our thoughts and reflect on our mental processes to comprehensively understand limitations and capacity in a way that formal systems cannot replicate (Kahneman, 2011). As such, we can examine the routine strategies for formal thinking machines, including methods like pattern recognition, image processing, and feature extraction (Russell & Norvig, 2009). These models can also use their predictive power on the rules of biology to make predictions for scientific hypotheses that we have not yet been able to test (Shmueli, 2010). Conversely, bioinformatic analysis into our biological systems also informs machine learning systems, allowing them to utilize their structures that mimic biological systems in the first place (Auslander et al., 2021). Ultimately, this is where an understanding of "biological self-programming" would help determine the appropriate architecture and feature selection for a given problem, where a wide range of specialized genomics concepts may be generalized but still interrogated for downstream problem-solving with a soft lens of formal thinking.

Contemporary Management for OCR

Contemporary management strategies for OCR rely on more established cultural methods. In Georgia, the climate creates an optimal environment for bacterial

proliferation, making early maturing, short-day onion varieties a recommended option for reducing the risk of infection (Agarwal et al., 2019; Penn State Extension, 2024). Early maturing varieties can escape the conditions that favor bacterial disease development and prevent PA from migrating into the bulb. Late-maturing onion genotypes provide an extended window for infection; this is due to high thrips pressure, warm and humid weather, and the absence of effective bactericides (Gitaitis et al., 2007; Dutta et al., 2014).

Thrips, as the primary vectors for *P. ananatis*, necessitate consistent management throughout the growing season due to the bacterium's non-persistent nature; however, more than thrips management is required for comprehensive OCR management (Dutta et al., 2014). For example, cultural methods such as switching from overhead to drip irrigation minimize leaf wetness, reducing bacterial spread. Proper plant spacing improves air circulation and lowers humidity around the plants, which helps limit bacterial growth (Agarwal et al., 2019; University of Georgia, 2023). Further, the interaction between onion genotypes and growth stages significantly manages PA infections (Stumpf et al., 2017). Sanitation measures through the destruction of plant debris and the timely removal of weeds are recommended to reduce alternate hosts for *Pantoea* spp. (Agarwal et al., 2019). Controlling weeds has successfully reduced the initial inoculum, as demonstrated by reducing the spread of *Pseudomonas viridiflava* in Georgia onion farms (Penn State Extension, 2024). Chemical control measures, particularly the preventive use of copper-based bactericides, are traditionally employed but have shown limited effectiveness due to the thrips' tendency to localize in hard-to-reach areas of the onion neck and the evidence of copper tolerance among *P. ananatis* (Gitaitis et al., 2007).

This tolerance could be mediated by efflux systems or mechanisms similar to or involving the "cop operon," observed in other bacteria, though these mechanisms remain unconfirmed in PA (Dutta et al., 2014). Currently, there are no commercially available onion genotypes resistant to *Pantoea* spp., making the use of certified disease-free seeds and transplants a core component of preventing the introduction of the bacterium into production fields (University of Georgia, 2023). An integrated approach targeting inoculum sources and vectors, combined with cultural, biological, and chemical strategies, is critical for effective OCR management. Recommendations from the University of Georgia Cooperative Extension provide growers with tailored recommendations for managing OCR (University of Georgia, 2023).

Justification for Research

OCR presents a significant economic threat to onion growers locally, nationally, and globally, and PA's virulence mechanisms remain not well understood. Leveraging the "biological self-programming" concept inherent in PA genetics, we aim to enhance data mining techniques by integrating two different association methodologies, genome-wide association studies with gene-pair coincidence analysis, to uncover genomic content responsible for disease phenotypes. We will examine other *Allium* species, improving our understanding of the impact of virulence factors across the genus and indirectly improving our overall management of OCR in *A. cepa*. Secondly, the recent discovery of *alt* gene clusters across various bacterial genera demands investigation regarding their prevalence and the diversity of their genetic content. Frustratingly, typical sequence-based methodologies struggle to identify *alt* clusters, and manual curation of bacterial genomes is unacceptably inefficient. We aim to overcome these limitations by employing

NLP-like computational techniques to vectorize abstract categorical variables, such as the seemingly guilt-by-association "gene sentence structure" inherent to a bacterial gene cluster, which will enhance the detection and classification of *alt*, *alt*-like, and pseudo-*alt* clusters in a manageable format. Finally, we propose to address the long-overdue need for resistance screening in *Allium* species against PA. Identifying resistance traits will allow us to discover and characterize resistant onion genotypes; more importantly, incorporating these findings into onion breeding programs, providing a much-needed solution for sustainable crop protection and the future of the onion industry.

Objectives

1. Utilize a pan-genome genome-wide association study, with an additional layer of gene-pair coincidence, in various strains of PA in *A. fistulosum* x *A. cepa* and *A. porrum* to determine potential virulence factors.
2. Utilize Natural Language Processing (NLP)-like deep learning to identify and validate thiosulfinate tolerance clusters in diverse bacteria.
3. Determine host-resistance and their mechanism in Allium genotypes against PA.

References

1. USDA (United States Department of Agriculture Agricultural Statistics Service-National). (2020). Vegetables 2019 summary. Retrieved from https://downloads.usda.library.cornell.edu/usdaesmis/files/02870v86p/0r967m63g/sn00bf58x/vegean20.pdf

2. Brannen, P., Brock, J., Dutta, B., Jagdale, G., Jogi, A., Kemerait, B., et al. (2017). 2017 Georgia plant disease loss estimates. Athens, GA: University of Georgia Cooperative Extension.
3. Stumpf, S., Gitaitis, R., Coolong, T., Riner, C., & Dutta, B. (2017). Interaction of onion cultivar and growth stages on incidence of Pantoea ananatis bulb infection. Plant Disease, 101(10), 1616–1620.
4. Asselin, J. E., Bonasera, J. M., & Beer, S. V. (2018). Center rot of onion (Allium cepa) caused by Pantoea ananatis requires pepM, a predicted phosphonate-related gene. Molecular Plant-Microbe Interactions, 31(12), 1291–1300.
5. Brady, C. L., Goszczynska, T., Venter, S. N., Cleenwerck, I., De Vos, P., Gitaitis, R. D., & Coutinho, T. A. (2011). Pantoea allii sp. nov., isolated from onion plants and seed. International Journal of Systematic and Evolutionary Microbiology, 61(5), 932–937.
6. Edens, D. G., Gitaitis, R. D., Sanders, F. H., and Nischwitz, C. 2006. First Report of Pantoea agglomerans Causing a Leaf Blight and Bulb Rot of Onions in Georgia. Plant Dis. 90:1551–1551
7. Stumpf, S., Kvitko, B., & Dutta, B. (2018). Isolation and characterization of novel Pantoea stewartii subsp. indologenes strains exhibiting center rot in onion. Plant Disease, 102(4), 727–733. https://doi.org/10.1094/PDIS-07-17-1107-RE
8. Chang, C. P., Sung, I. H., & Huang, C. J. (2018). Pantoea dispersa causing bulb decay of onion in Taiwan. Australasian Plant Pathology, 47(6), 609–613.
9. Morohoshi, T., Nakamura, Y., Yamazaki, G., Ishida, A., Kato, N., & Ikeda, T. (2007). The plant pathogen Pantoea ananatis produces N-acylhomoserine lactone

and causes center rot disease of onion by quorum sensing. Journal of Bacteriology, 189(22), 8333–8338.

10. De Maayer, P., Aliyu, H., Vikram, S., Blom, J., Duffy, B., & Cowan, D. A. (2017). Phylogenomic, pan-genomic, pathogenomic, and evolutionary genomic insights into the agronomically relevant enterobacteria Pantoea ananatis and Pantoea stewartii. Frontiers in Microbiology, 8, 1755.
11. Rice, P. J., Harman-Fetcho, J. A., Heighton, L. P., McConnell, L. L., Sadeghi, A. M., & Hapeman, C. J. (2006). Environmental fate and ecological impact of copper hydroxide: Use of management practices to reduce the transport of copper hydroxide in runoff from vegetable production. Crop Protection Products for Organic Agriculture, ACS Symposium Series, 917, 217–230.
12. Stumpf, S., Leach, L., Srinivasan, R., Coolong, T., Gitaitis, R., & Dutta, B. (2021). Foliar chemical protection against Pantoea ananatis in onion is negated by thrips feeding. Phytopathology, 111(2), 258–267.
13. Gitaitis, R. D., & Gay, J. D. (1997). First report of a leaf blight, seed stalk rot, and bulb decay of onion by Pantoea ananas in Georgia. Plant Disease, 81, 1096.
14. Coutinho, T. A., & Venter, S. N. (2009). Pantoea ananatis: An unconventional plant pathogen. Molecular Plant Pathology, 10, 325–335.
15. De Maayer, P., Chan, W. Y., Rubagotti, E., Venter, S. N., Toth, I. K., Birch, P. R., et al. (2014). Analysis of the Pantoea ananatis pan-genome reveals factors underlying its ability to colonize and interact with plant, insect, and vertebrate hosts. BMC Genomics, 15:404.

16. Weller-Stuart, T., De Maayer, P., & Coutinho, T. (2017). Pantoea ananatis: Genomic insights into a versatile pathogen. Molecular Plant Pathology, 18, 1191-1198.
17. Hara, Y., Kadotani, N., Izui, H., et al. (2012). The complete genome sequence of Pantoea ananatis AJ13355, an organism with great biotechnological potential. Applied Microbiology and Biotechnology, 93(1), 331-341.
18. Stice, S. P., Thao, K. K., Khang, C. H., Baltrus, D. A., Dutta, B., & Kvitko, B. H. (2020). Thiosulfinate tolerance is a virulence strategy of an atypical bacterial pathogen of onion. Current Biology, 30(16), 3130-3140.e6.
19. Curtis, H., Hilgenfeldt, U., Noll, U., & Slusarenko, A. J. (2004). Broad-spectrum activity of the volatile phytoanticipin allicin in extracts of garlic (Allium sativum L.) against plant pathogenic bacteria, fungi, and oomycetes. Physiological and Molecular Plant Pathology, 65(2), 79-89.
20. Barbu, I. A., Cristea, O., Chelariu, E. L., & Ciofu, C. (2023). Phytochemical characterization and antimicrobial activity of several Allium extracts. Molecules, 28(10), 3980.
21. Polidore, A. L. A., Furiassi, L., Hergenrother, P. J., & Metcalf, W. W. (2021). A phosphonate natural product made by Pantoea ananatis is necessary and sufficient for the hallmark lesions of onion center rot. mBio, 12, e03402-20.
22. Shyntum, D. Y., Venter, S. N., Moleleki, L. N., Toth, I. K., & Coutinho, T. A. (2015). Pantoea ananatis utilizes a type VI secretion system for pathogenesis and bacterial competition. Molecular Plant-Microbe Interactions, 28(5), 420-431.

23. Borlinghaus, J., Albrecht, F., Gruhlke, M. C. H., Nwachukwu, I. D., & Slusarenko, A. J. (2020). Genetic and molecular characterization of multicomponent resistance of Pseudomonas against allicin. Life Science Alliance, 3(5), e202000670.
24. Paudel, S., Zhao, M., Stice, S. P., Dutta, B., & Kvitko, B. H. (2024). Thiosulfinate tolerance gene clusters are common features of Burkholderia onion pathogens. Molecular Plant-Microbe Interactions, 37(3), 298-312.
25. Morohoshi, T., Nakamura, Y., Yamazaki, G., Ishida, A., Kato, N., & Ikeda, T. (2007). Quorum sensing in bacterial systems. Trends in Microbiology, 15, 185-192.
26. Choi, S. R., Lee, M. S., & Kim, J. Y. (2021). Sustainable bioenergy production using Pantoea. BioEnergy Research, 14(2), 295-312.
27. Usuda, H., Fukuda, A., Kuroiwa, T., Hasegawa, M., & Kondo, K. (2022). The potential of Pantoea in anaerobic biohydrogen production. Journal of Bioscience and Bioengineering, 134(5), 413-423.
28. Fu, J., McCallum, J., Baldwin, S., & Deng, Y. (2019). The impact of genome plasticity on flavor diversity and defense in Allium crops. Genome Biology, 20, 153.
29. Chalbi, K., Voigt, K., Eshetu, F., & Schmülling, T. (2023). Large-scale sequencing of the Allium cepa genome reveals insights into LTR regulation. The Plant Journal, 114(2), 321-335.
30. Shigyo, M., Fujito, S., Ohara, T., & Kik, C. (2018). Comprehensive analysis of Allium genome plasticity. Theoretical and Applied Genetics, 131(7), 1541-1552.
31. Bastaki, A. S., Al-Ammar, G. M., & Al-Essa, S. M. (2021). Biological activity of Allium extracts as alternative antibiotics. Journal of Ethnopharmacology, 265, 113319.

32. Iwar, R., Matsuura, H., & Harada, H. (2024). Antidiabetic potential of Allium-derived phytochemicals. Journal of Medicinal Chemistry, 67(2), 245-258.
33. Ruiz-Mirazo, K., & Moreno, A. (2012). Self-organization in biological systems: Concepts and models. Physics of Life Reviews, 9(3), 288-320.
34. Bich, L., & Moreno, A. (2015). Autopoiesis and its challenges: From a biological concept to an inter-organizational one. Philosophy of Science, 82(2), 240-260.
35. De la Fuente, I. (2021). Information processing in biological networks: A model for self-regulation in biological systems. Frontiers in Bioengineering and Biotechnology, 9, 582311.
36. Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: Building the web of life. Nature Reviews Genetics, 16(8), 472-482.
37. Baltrus, D. A. (2013). Exploring the costs of horizontal gene transfer. Trends in Ecology and Evolution, 28(8), 489-495.
38. Whelan, F. J., Rusilowicz, M., & McInerney, J. O. (2020). Coinfinder: Detecting significant associations and dissociations in pangenomes. Microbial Genomics, 6(3), e000338.
39. Alyssa, M. W., & Stavrinides, J. (2015). Pantoea: Insights into a highly versatile and diverse genus within the Enterobacteriaceae. FEMS Microbiology Reviews, 39(6), 968-984.
40. Adams, K. L. (2007). Evolution of duplicate gene expression in polyploid and hybrid plants. Journal of Heredity, 98(2), 136-141.
41. Van de Peer, Y., Mizrachi, E., & Marchal, K. (2021). Polyploidy: An evolutionary and ecological force in stressful times. The Plant Cell, 33(1), 11-26.

42. Sattler, M. C., Carvalho, C. R., & Clarindo, W. R. (2016). The polyploidy and its role in plant breeding: A review. Planta, 243(2), 281-296.
43. Qiao, X., Li, Q., Yin, H., & Wang, M. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome Biology, 20, 38.
44. Lämke, J., & Bäurle, I. (2017). Epigenetic and chromatin-based mechanisms in environmental stress adaptation and stress memory in plants. Genome Biology, 18, 124.
45. Guevarra, R. B., Shin, J., & Lee, J. (2021). Comprehensive genomic analysis reveals virulence factors and antibiotic resistance genes in Pantoea agglomerans KM1. PLoS One, 16(1), e0239792.
46. Dahiya, P., Kumar, P., Rani, S., et al. (2024). Comparative genomic and functional analyses for insights into Pantoea agglomerans strains adaptability in diverse ecological niches. Current Microbiology, 81, 254.
47. Brauer, E. K., Lee, H., Svircev, A. M., Castle, A. J., & Pauls, K. P. (2018). Integrative network-centric approach reveals signaling pathways associated with plant resistance and susceptibility to Pseudomonas syringae. PLOS Biology.
48. Ichinose, Y., Taguchi, F., Mukaihara, T., & Murata, K. (2013). Pathogenicity and virulence factors of Pseudomonas syringae. Journal of General Plant Pathology, 79(4), 285-296.
49. Misra, B. B., Langefeld, C. D., Olivier, M., & Cox, L. A. (2018). Integrated omics: Tools, advances, and future approaches. Journal of Molecular Endocrinology. (PMID: 30006342).

50. Ernst, J., & Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. Nature Methods, 9(3), 215-216.
51. Domingos, P. (2012). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.
52. Alpaydin, E. (2020). Introduction to machine learning (4th ed.). MIT Press.
53. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
54. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
55. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2004). Random forest: A classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences, 43(6), 1947-1958.
56. Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3), 289-310.
57. Haykin, S. (2009). Neural networks and learning machines (3rd ed.). Prentice Hall.
58. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.
59. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
60. Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359-366.
61. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

62. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).
63. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
64. Trabelsi, C., Bilaniuk, O., Zhang, Y., Serdyuk, D., Subramanian, S., Santos, J. F., Mehri, S., Rostamzadeh, N., Bengio, Y., & Pal, C. J. (2018). Deep complex networks. International Conference on Learning Representations (ICLR).
65. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672-2680.
66. Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, 42(2), 230-265.
67. Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik, 38, 173-198.
68. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.
69. Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. G.P. Putnam.
70. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
71. Russell, S., & Norvig, P. (2009). Artificial intelligence: A modern approach (3rd ed.). Prentice Hall.

72. Auslander, N., Gussow, A. B., & Koonin, E. V. (2021). Incorporating machine learning into established bioinformatics frameworks. International Journal of Molecular Sciences, 22(6), 2903.
73. Dutta, B., Barman, A. K., Srinivasan, R., Avci, U., Ullman, D. E., Langston, D. B., & Gitaitis, R. D. (2014). Transmission of Pantoea ananatis and P. agglomerans, causal agents of center rot of onion (Allium cepa), by onion thrips (Thrips tabaci) through feces. Phytopathology, 104(8), 812-819.
74. Gitaitis, R., & Walcott, R. (2007). The epidemiology and management of seedborne bacterial diseases. Annual Review of Phytopathology, 45, 371-397.
75. Penn State Extension. (2024). Rotten to the core: The center rot disease of onion. Pennsylvania State University. Retrieved from https://extension.psu.edu/rotten-to-the-core-the-center-rot-disease-of-onion.
76. University of Georgia Cooperative Extension. (2023). Onion production guide (Publication No. B1198). Retrieved from https://extension.uga.edu/publications/detail.html?number=B1198.

Chapter 2

Genome-wide association and dissociation studies in *P. ananatis* reveal potential virulence factors affecting *Allium porrum* and *A. fistulosum* x *A. cepa* hybrid[1]

---

[1] Myers, B. K., Shin, G. Y., Agarwal, G., Stice, S. P., Gitaitis, R. D., Kvitko, B. H., & Dutta, B. Genome-wide association and dissociation studies in *Pantoea ananatis* reveal potential virulence factors affecting *Allium porrum* and *Allium fistulosum* × *Allium cepa* hybrid. *Frontiers in Microbiology*. 1/6/2021.

Abstract

Pantoea ananatis is a member of a *Pantoea* species complex that causes center rot of bulb onions (*A. cepa*) and also infects other *Allium* crops like leeks (*Allium porrum*), chives (*Allium schoenoprasum*), bunching onion or Welsh onion (*Allium fistulosum*), and garlic (*Allium sativum*). This pathogen relies on a chromosomal phosphonate biosynthetic gene cluster (HiVir) and a plasmid-borne thiosulfinate tolerance cluster (*alt*) for onion pathogenicity and virulence, respectively. However, pathogenicity and virulence factors associated with other *Allium* species remain unknown. We used phenotype-dependent genome-wide association (GWAS) and phenotype-independent gene-pair coincidence (GPC) analyses on a panel of diverse 92 *P. ananatis* strains, which were inoculated on *A. porrum* and *A. fistulosum* × *A. cepa* under greenhouse conditions. Phenotypic assays showed that, in general, these strains were more aggressive on *A. fistulosum* × *A. cepa* as opposed to *A. porrum*. Of the 92 strains, only six showed highly aggressive foliar lesions on *A. porrum* compared to *A. fistulosum* × *A. cepa*. Conversely, nine strains showed highly aggressive foliar lesions on *A. fistulosum* × *A. cepa* compared to *A. porrum*. These results indicate that there are underlying genetic components in *P. ananatis* that may drive pathogenicity in these two *Allium* spp. Based on GWAS for foliar pathogenicity, 835 genes were associated with *P. ananatis*' pathogenicity on *A. fistulosum* × *A. cepa* whereas 243 genes were associated with bacterial pathogenicity on *A. porrum*. The Hivir as well as the *alt* gene clusters were identified among these genes. Besides the 'HiVir' and the *alt* gene clusters that are known to contribute to pathogenicity and virulence from previous studies, genes annotated with functions related to stress responses, a potential toxin-antitoxin system, flagellar-motility, quorum sensing,

and a previously described phosphonoglycan biosynthesis (pgb) cluster were identified. The GPC analysis resulted in the identification of 165 individual genes sorted into 39 significant gene-pair association components and 255 genes sorted into 50 significant gene-pair dissociation components. Within the coincident gene clusters, several genes that occurred on the GWAS outputs were associated with each other but dissociated with genes that did not appear in their respective GWAS output. To focus on candidate genes that could explain the difference in virulence between hosts, a comparative genomics analysis was performed on five *P. ananatis* strains that were differentially pathogenic on *A. porrum* or *A. fistulosum* × *A. cepa*. Here, we found a putative type III secretion system, and several other genes that occurred on both GWAS outputs of both *Allium* hosts. Further, we also demonstrated utilizing mutational analysis that the *pepM* gene in the HiVir cluster is important than the *pepM* gene in the pgb cluster for *P. ananatis* pathogenicity in *A. fistulosum* × *A. cepa* and *A. porrum*. Overall, our results support that *P. ananatis* may utilize a common set of genes or gene clusters to induce symptoms on *A. fistulosum* × *A. cepa* foliar tissue as well as *A. cepa* but implicates additional genes for infection on *A. porrum*.

Introduction

Pantoea ananatis is one of the several species of bacteria within the *Pantoea* genus that causes onion center rot (OCR). The OCR can cause considerable losses in both yield and quality in *Allium*s, particularly in bulb onions (*Allium cepa*) in the southeastern United States (Stumpf et al*.,* 2018; Gitaitis *et al*., 2003; Stice *et al*., 2018). There is currently no known resistance to *P. ananatis* in commercial onion cultivars and resistance in other *Allium* spp. is yet to be evaluated. *P. ananatis* invades the plant through foliar wounds

leading to water-soaked lesions, blighting, and wilting of the leaf. Foliar colonization can lead to bulb invasion that often results in further post-harvest losses (Carr *et al*., 2013). While *P. ananatis* can be seedborne and seedling transmitted, thrips (*Frankliniella fusca, Thrips tabaci*) mediated transmission seems to be more common and epidemiologically important, particularly in Georgia, United States. Several published reports indicate that these thrips species can acquire epiphytic *P. ananatis* populations from various environmental host plants including weeds and can transmit the pathogen to healthy onion seedlings (Dutta *et al*., 2016; Dutta et al., 2014). *P. ananatis* collectively has a broad reported host range as it can cause disease in a diverse range of crops including maize (*Zea mays L*.), pineapple (*Ananas comosus*), rice (*Oryza sativa*), and Sudan grass (*Sorghum bicolor x S. bicolor var. sudanese*) (De Maayer, *et al*., 2014).

Based on observations made by Stice *et al.,* (2018), *P. ananatis* is pathogenic on a variety of *Allium* spp.; however, the authors observed that the strains varied greatly in their pathogenic potential and aggressiveness on onion, shallot (*A. cepa* var. *aggregatum),* chives *(A. schoenoprasum),* and leeks *(A. porrum).* The bunching onion or Welsh onion (*A. fistulosum*) has been shown to be a pathogenic host for *P. ananatis* (Kido *et al*., 2010; Wang, Lin and Huang. 2018). Symptoms observed on these hosts were comparable to those observed on typical bulb-forming onion seedlings. *P. ananatis* is unusual when compared with other Gram-negative plant pathogenic bacteria in that it does not utilize either the type II secretion system (T2SS) or the type III secretion system (T3SS) to secrete cell-wall degrading enzymes and deliver virulence effectors into the target host cell, respectively (Chang, Desveaux and Creason (2014). Asselin *et al*. (2018) and Takikawa *et al*. (2018) independently identified a chromosomally located “HiVir,” or

High **Vir**ulence gene cluster. This gene cluster has been demonstrated to be a critical pathogenicity factor for *P. ananatis* in onions (Asselin *et al*. 2018. Takikawa *et al*. 2018). The HiVir encodes for a phosphonate phytotoxin “Pantaphos” that was shown to be critical for bulb necrosis (Asselin *et al*. 2018. Polidore, *et al.* 2021.). Another cluster of importance was recently characterized by Stice, *et al.* and is a plasmid-borne virulence gene cluster coined as “*alt*” or thiosulfinate (**Al**licin) **t**olerance (Stice *et al.,* 2018. Stice *et al.,* 2020). The *alt* cluster in *P. ananatis* imparts tolerance to thiosulfinates allowing *P. ananatis* to colonize the thiosulfinate-rich environment in necrotic onion bulbs (Stice *et al.,* 2018. Stice *et al.,* 2020.). Despite these advances in understanding pathogenicity and virulence mechanisms in *P. ananatis*-onion pathosystem, the mechanisms behind bacterial capacity to colonize other *Allium* spp. such as *A. porrum* and *A. fistulosum x A. cepa* remain unknown.

Whole-genome sequencing (WGS) of bacteria is routinely performed in many laboratories for diagnostics, understanding host-pathogen interactions, and for ecological studies (Chewapreecha *et al.,* 2014. Laabel *et al.,* 2014. Sheppard *et al.,* 2013. Desjardins *et al.,* 2016. Farhat *et al.,* 2013. Earie *et al.,* 2016. Hall *et al.,* 2014. Holt *et al.,* 2015. Lees *et al.,* 2016.). Despite generation of large informatics data sets, the primary challenge when handling these methodologies is managing an appropriate strategy to go from raw data to detailed, biologically relevant information. One strategy commonly used to relate genotype to phenotype is the genome-wide association study (GWAS). This strategy was first adopted in human-based medicine, but soon after gained general popularity in analyzing bacterial genomes to answer various questions related to pathogenicity, antibiotic resistance, and bacterial survival (Chewapreecha *et al.,* 2014.

Laabel *et al.,* 2014. Sheppard *et al.,* 2013. Desjardins *et al.,* 2016. Farhat *et al.,* 2013. Earie *et al.,* 2016. Hall *et al.,* 2014. Holt *et al.,* 2015. Lees *et al.,* 2016.). GWAS has proven to be an excellent tool in correlating genomic variations with observed phenotypes such as virulence factors, antibiotic resistance, and tolerance to abiotic and biotic stresses (Chewapreecha *et al.,* 2014. Laabel *et al.,* 2014. Sheppard *et al.,* 2013. Desjardins *et al.,* 2016. Farhat *et al.,* 2013. Earie *et al.,* 2016. Hall *et al.,* 2014. Holt *et al.,* 2015. Lees *et al.,* 2016.). In this study we utilized the WGS strategy to build a "pan-genome" and compare the combined genomes of a bacterial population to pathogenicity phenotype on two *Allium* spp. (*A. porrum* and *A. fistulosum x A. cepa).*

The complete complement of the total genes within a genomic set is termed as "pan-genome" (Medini *et al.,* 2005. Tettelin, *et al., 2005.*). A pangenome consist of "core" genes that are common across all bacterial strains of a species and the "accessory" genes that are specific/present in only some strains (Tettelin, *et al.,* 2005). Accessory genes are responsible for key differentiation among strains and have been associated with pathogenicity islands or with niche adaptation (Brockhurst *et al.,* 2019*).* Ideally, pan-GWAS can also be used to identify associations between genotypic traits and observed phenotype. This may aid in determining potential gene or gene clusters that are responsible for the observed phenotype with a pre-defined set of statistical criteria (Brynildsrud *et al., 2016*). In the current study, we utilized a pangenome-wide association study (pan-GWAS) to identify presence and absence variants in *P. ananatis* strains (*n*=92) that are associated with foliar symptoms in *A. porrum* and *A. fistulosum x A. cepa*. Using pangenome-GWAS, we report a set of potential *P. ananatis* virulence factors associated with these *Allium* hosts including the "HiVir" cluster. These include the

phosphonoglycan biosynthesis (pgb) cluster, a type III secretion system, a putative toxin/antitoxin region, and several other virulence-associated genes. Further, we also demonstrated that the *pepM* gene in the HiVir cluster is more important than the *pepM* gene in the "pgb" cluster for *P. ananatis* pathogenicity in *A. fistulosum* x *A. cepa* and *A. porrum*.

A recent study tested the hypothesis that genes generally co-occur (associate) or avoid each other (dissociate) based on the fitness consequences in a particular set of genomes (Whelan, Rusilowicz and McInerney, 2019). For example, in our case, we presume that genes, which allow necrosis in *Allium* spp. and confer thiosulfate tolerance should associate as these traits are co-beneficial for survival in niche of an *Allium* host. However, genes that, in combination, result in the production of a toxic byproduct (as has been observed associated with siderophore biosynthesis in *Salinispora* spp.) (Bruns *et al.,* 2018), perform some redundant function, or trigger an immune response, should dissociate with each other as co-expression may reduce strain fitness. Therefore, we analyzed genomic interactions of accessory genes in the pangenome derived from 92 *P. ananatis* genomes to determine genes associated with virulence, with a premise that virulence genes should associate with other virulence genes throughout the accessory pangenome and that redundant virulence genes should naturally dissociate.

Materials and methods

Bacterial strains, identification, culturing, and mutagenesis

Pantoea ananatis strains (*n*=92) used in this study were isolated from diverse sources; weeds, thrips, and onion tissue (foliage, bulbs, and seeds) in Georgia from 1992-2019.

The metadata for each strain such as the source, year of isolation, and county of origin in Georgia for these strains and their distribution within each category are listed (Table 1). Among the strains used, 55 strains (59.8%) were isolated from onion foliage or bulb tissue, which constituted majority of the strains and remaining 38 strains (40.2%) were isolated from other diverse sources including weeds, and thrips. This is followed by the weeds; *Richardia scabra L.* (8.7%; 8/92), *Digitaria spp.* (6.5%; 6/92), and *Verbena bonariensis* (4.3% ;4/92). The strains from various plant sources constituted 8.7% (8/92) of the total strains studied (Table 1; Figure 1B). Strains from thrips (*Frankliniella fusca* and *F. occidentalis*) constituted the remaining 13% (12/92). These curated strains were initially identified as *P. ananatis* by their colony morphology and physiological characteristics such as being: Gram-negative, facultatively anaerobic, positive for indole production, and negative for nitrate reductase and phenylalanine deaminase. Further confirmation was done using *P. ananatis*-specific PCR assay as described earlier (Walcott *et al.,* 2002).

Inoculum was prepared by transferring single colonies of each bacterial strain from 24 h-old cultures on nutrient agar (NA) medium to nutrient broth (NB). The broth was shaken overnight on a rotary shaker (Thermo Scientific, Gainesville, FL) at 180 rpm. After 12 h of incubation, 1 ml of each bacterial suspension were centrifuged at 5,000 × g (Eppendorf, Westbury, NY) for 2 mins. The supernatant was discarded, and the pellet was re-suspended in deionized water. Inoculum concentration was adjusted using a spectrophotometer (Eppendorf, Westbury, NY) to an optical density of 0.3 at 600 nm [≈1 × 10^8 colony forming unit (CFU)/ml].

Deletion of the *P. ananatis* PANS 02-18 $pepM_{pgb}$ $pepM_{HiVir}$ genes was conducted by two-step allelic exchange as described by Stice et al. (2020). In brief, approximately 400-bp flanks to the targeted genes were directly synthesized as a single joined sequence by Twist bioscience and cloned via BP clonase II using primer-introduced *att*B1/2 recombination sites into the pR6KT2G Gateway® compatible sacB-based allelic exchange vector. These deletion constructs were introduced into PANS 02-18 via biparental conjugation with the RHO5 *E. coli* strain and single crossover events were recovered via gentamicin selection. Second crossover events were recovered via liquid sucrose counter-selection and identified by screening for backbone eviction based on loss of gentamicin resistance and the formation of white colonies on X-gluc. Deletion mutants were identified and confirmed based on PCR and amplicon sequencing using independent primers designed to amplify from genomic regions adjacent to the 400-bp deletion flank regions.

Phenotypic assessment of *P. ananatis*: red onion scale necrosis, foliar pathogenicity and aggressiveness assay of on *A. porrum* and *A. fistulosum x A. cepa.*

Pathogenic potential of *P. ananatis* strains were initially phenotyped on onion scale using previously described red scale necrosis (RSN) assay (Stice *et al.,* 2018). Red onions (cv. Red Burgundy) were surface sterilized with 70% ethanol and the outermost scale sliced to approximately 3 cm×4 cm. The resulting scales were set on a sterile petri dish or on sterilized microtube trays, with the bottom covered with sterilized paper towels pre-moistened with distilled water. Each onion scale was then wounded via direct penetration with a sterilized needle and inoculated with 10 µl of approximately $1x10^6$ CFU/mL

inoculum of *P. ananatis*. A known onion-pathogenic strain (PNA 97-1) was used as a positive control. Sterile water was used as a negative control. The resulting petri dishes were then laid in an aluminum tray (46 cm × 25 cm × 10 cm) and covered with a plastic lid. These onion scales were then incubated for 5 days in the dark. The area of pigment clearing, and necrotic lesions were measured at 7 days post-inoculation. Strains that did not clear the red anthocyanin pigment or developed necrotic lesions, were declared non-pathogenic. Strains that caused necrosis along with pitting, with a visible zone of pigment clearing were considered pathogenic. Three replications were performed for each strain and in total two experiments were conducted.

Foliar pathogenicity and aggressiveness of *P. ananatis* strains (*n*=92) were determined on *A. porrum* (cv. King Richard) and *A. fistulosum* x *A. cepa* (cv. Guardsman) under controlled greenhouse conditions. *P. ananatis* strain (PNA 97-1) was used as a positive control for both *Allium* species (14,31). Seedlings were established in plastic pots (T.O. plastics, Clearwater, MN) with dimension of 9 cm × 9 cm × 9 cm (length × breadth × height) containing a commercial potting mix (Sta-green, Rome, GA). The seedlings were maintained under greenhouse condition at 25-28°C and 70-90% relative humidity with a light:dark cycle of 12h:12h. Osmocote smart release plant food (The Scotts Company, Marysville, Ohio) was used for periodic fertilization. Bacterial strains were maintained on NA plates and inoculum was generated as described above. Once the primary leaf of each *Allium* spp. reached 9 cm, seedlings were inoculated using a cut-tip method as described previously (Dutta *et al.,* 2014). Briefly, a wound was created by cutting the central leaf (2 cm from the apex) with a sterile pair of scissors. Using a micropipette, a 10 µl drop of a bacterial suspension containing 1×10^8 CFU/ml (1×10^6

CFU/leaf) was deposited at the cut-end. Seedlings inoculated with sterile water as described above were used as negative control. Three replications per strain per host were used for one experiment and a total of two independent experiments were conducted.

The seedlings were observed daily for symptom development until 5 days post-inoculation (DPI) and were compared with the foliar symptoms displayed by the positive control on each *Allium* species. The aggressiveness of *P. ananatis* strains was determined based on the lesion length on each *Allium* spp. For *A. porrum*, strains that caused a lesion length of 0.2-0.5 cm were considered less aggressive, 0.5-0.9 cm moderately aggressive, and >1 cm highly aggressive. For *A. fistulosum x A. cepa*, strains were considered highly aggressive when a lesion length of >1.4 cm was observed. Lesion lengths ranging from 0.7-1.4 cm were considered as moderately aggressive and strain with lesion length <0.7 cm was regarded as less aggressive. Bacterial strains that did not display any lesion were considered as non-pathogenic. To confirm if the symptoms were caused by *P. ananatis*, bacteria were isolated from the region adjoining the symptomatic and healthy tissue on PA-20 semi-selective medium and incubated for 5-7 days at 28°C (Goszczynska, Venter and Couthino. 2006). Presumptive colonies were further confirmed using *P. ananatis*-specific assay as mentioned above (Goszczynska, Venter and Couthino. 2006). Further, strain identity from randomly isolated colonies from *A. porrum* and *A. fistulosum x A. cepa* were confirmed by their DNA fingerprints using repetitive extragenic palindrome (rep)-PCR as previously described (Dutta *et al.,* 2014).

Genome sequencing: Data filtering, draft genome assembly and annotation

Genomic DNA was extracted utilizing the E.Z.N.A bacterial DNA kit Omega Bio-Tek (Norcross, GA). A 50 µl of DNA (50 ng/µl) per sample was used for library preparation as per the manufacturer's instructions at Novogene Bioinformatics Technology Co. Ltd. (Beijing, China). Genomic DNA of each sample was randomly sheared into short fragments of about 200-400 base pairs (bp). The obtained fragments were subjected to library construction using the NEBNext® DNA Library Prep Kit. After end repairing, dA-tailing, and further ligation with NEBNext adapter, the required fragments (in 200-400 bp size) were PCR enriched by P5 and indexed P7 oligos. The library was subsequently sequenced on Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA). Pair-end sequencing were performed with the read length of PE150 bp at each end. The raw fastq reads obtained were quality filtered. FastQC was used to assess the raw fastq files. Reads were filtered utilizing Trimmomatic (v. 0.36). The read data were filtered to remove low quality reads/bases and trimmed for reads containing primer/adaptor sequences using Trimmomatic's paired end mode (Bolger, Lohse and Usadel. 2014). Further, all 5' and 3' stretches of ambiguous 'N' nucleotides were clipped to ensure high quality reads. Trimmed data were re-assessed using FastQC and further used for genome assembly followed by pan-genome analyses. Further, all contigs ≤ 500bp were removed using Seqtk (1.3). The cleaned reads were assembled using SPAdes (v. 3.15.3) (Bankevich *et al.,* 2012). Both the paired and unpaired data were used in assembly at default settings. The scaffolds of the respective 92 *P. ananatis* strains were annotated using Prokka (v. 1.14.5) (Seemann, 2014). The resulting .gff files were used in the downstream pan-genome analysis. Average nucleotide identity was determined using FastANI (Jain, *et al.,* 2018). KEGG gene ontology (GO) assignment was conducted using BioBam BLAST2GO

pipeline (BioBam 2022, Götz *et al.,* 2008). A phylogenetic tree of single nucleotide polymorphisms (SNP's) of core genome was generated utilizing PanSeq at default settings and RAxML alignment with 10,000 bootstrap replicates (Laing *et al.,* 2010. Stamatakis, 2014). The RAxML Boostrap random number used was 9595, with parsimony random seed of 5959.For Coinfinder and Roary plots, a tree was generated using FastTree 2.1.11 at default settings utilizing Roary's core gene alignment (Price, Dehal and Arkin. 2009).

Pan-Genome, genome-wide associate studies (GWAS), and gene coincidence of *P. ananatis* (N = 92 strains)

The annotated genomes that passed quality control were used as inputs in ROARY (v. 3.12.0) at default settings that aided in generating a pan-genome with core and accessory genes. The complete pan-genome matrix in the form of presence and absence variant was used as inputs in SCOARY (v. 1.6.16). The SCOARY program conducted GWAS analysis that determined association between genomes and observed phenotypes (Brynildsrud, *et al.,* 2016. Page *et al.,* 2015.). This program was operated twice separately on each host plant of interest, once at default parameters, and a second time with a forced maximum p-value 0.05 across all testing parameters. For the gene association/dissociation analysis, complete pangenome of 92 *P. ananatis* strains were used as inputs for Coinfinder (v. 1.0.1). This program generated both gene-pair associations and dissociations with modification to association significance increased to a p-value of 0.1, and default settings (p=0.05) for dissociation as previously described (Whelan, Rusilowicz and McInerney, 2019). Direct comparisons of genetic sequences

were performed using the Clustal Omega online server at default settings (Madeira *et al.* 2022).

Tobacco infiltration assay for *P. ananatis* strain (PNA 15-3) with putative Type III secretion system (T3SS)

A single colony of *P. ananatis* strain PNA 15-3 (with putative T3SS) and a strain of *Pantoea stewartii* subsp. *indologenes* (20GA0713; positive control for T3SS) was suspended and grown overnight in modified Coplin medium (Asselin, Bonosera and Beer. 2018). Approximately 100 µl of the overnight culture was syringe-infiltrated into the tobacco leaf and the resulting infiltrated area was marked with a black marker. A sterile Coplin lab medium was used as a negative control. The symptom was observed at 48-hours post inoculation (hpi) when the image was taken. This experiment was repeated twice.

Results

Phenotypic assessment of *P. ananatis*: red onion scale necrosis, foliar pathogenicity and aggressiveness assay of on *A. porrum* and *A. fistulosum x A. cepa.*

Phenotyping of 92 *P. ananatis* strains displayed variability in the level of aggressiveness on both *Allium* spp. (Table 1.1). Variations in *P. ananatis* pathogenicity and aggressiveness on two *Allium* spp. were considerable (Figure 1.1 A and B and Table 1.1). Strains screened in this study belonged to different isolation sources (Figure 1.1 C and Table 1). Using the RSN disease phenotyping assay, we observed 61.3% (57/92) of *P. ananatis* strains displayed typical necrosis of red onion scale whereas 38.7%% (36/92) of strains did not cause necrosis (Figure 1.1 D).

When *P. ananatis* strains were screened on *A. fistulosum x A. cepa*, 20.4% (19/92) and 44.1% (41/92) were found to be non-pathogenic and mildly aggressive, whereas 25.8% (24/92) and 9.7% (9/92) of the strains were identified as moderately aggressive and highly aggressive, respectively (Figure 1.1 E). In contrast, on *A. porrum*, 45.7% (42/92) and 37% (34/92) of the strains were non-pathogenic and mildly aggressive, respectively. Interestingly, a much lower proportion of the strains; 14.1% (13/92) and 3.3% (3/92) identified as moderately aggressive and highly aggressive, respectively on *A. porrum* (Figure 1.1 F). The percentage of strains that were pathogenic on *A. porrum* but non-pathogenic on *A. fistulosum x A. cepa* was only 2.1% (2/92). In contrast, 27% (25/92) of the strains that were pathogenic on *A. fistulosum x A. cepa* were non-pathogenic on *A. porrum*. Interestingly, 4.3% (4/92) of strains were highly aggressive on *A. porrum* but less aggressive on *A. fistulosum x A. cepa* whereas 9.7% (9/92) of strains were highly aggressive on *A. fistulosum x A. cepa* but less aggressive on *A. porrum*. Percentage of strains that were moderately to highly aggressive on both *Allium* sp. was 4.3% (4/92) whereas 18.8% (17/92) of the strains were non-pathogenic on both hosts tested. All the strains isolated from symptomatic *A. porrum* or *A. fistulosum x A. cepa* were identified as *P. ananatis* by recovery on PA-20 semi-selective medium and a *P. ananatis*-specific PCR assay as described above. *P. ananatis* colonies were not recovered from any of the negative control seedlings on PA-20 medium indicating no potential cross-contamination among the inoculated strains.

Interestingly, among the 57 RSN-positive strains, 63.2% (36/57) of strains were pathogenic on both *A. porrum* and *A. fistulosum x A. cepa* whereas 0% (0/57) and 26.3% (15/57) of strains were only pathogenic on *A. porrum* or *A. fistulosum x A. cepa*,

respectively. The remaining RSN-positive strains were non-pathogenic in the leaf tip necrosis assay on both hosts 10.5% (6/57). Among the RSN-negative strains, 41.6% (15/36) strains were non-pathogenic on both hosts, whereas 25% (9/36) were pathogenic on both hosts. Also, 5.5% (2/36) and 27.7% (10/36) of strains were pathogenic on only *A. porrum* and *A. fistulosum x A. cepa*, respectively.

The *P. ananatis* pan-genome, architecture, and annotation

Post-sequencing, 1,594,092,228 raw reads were obtained and after stringent quality-filtering and trimming nearly 87% of the total reads (1,413,144,772 quality reads) were retained. The FastQC results indicated the sequence quality "passed," as the majority of per-nucleotide and sequence qualities achieved high scores with no issues reported. For example: a good score can be ascertained with an average quality score of 30 to 40, with an exponentially increasing quality score distribution. Sequences that failed to incorporate into the final pangenome, or showed signs of contamination, were removed from the study. All *P. ananatis* sequences used in this study were submitted to NCBI (bio project PRJNA825576). Their corresponding accession numbers are listed in the supplementary table 1.1.

An overview of the final pangenome shows a core genome (occurs in 99% or more genomes, N >= 91) of 2914 genes, a soft-core (occurs in 95% to 99% of genomes, N = 87 to 91) of 687 genes, a shell genome of 1833 genes (occurs in 15% to 95% of the genomes, N = 14 to 87), and a cloud genome (occurs in 0% to 15% of genomes, N = 0 to 14) of 9,196 genes for a total of 14,630 genes (Figure 1.2 A). Details of the number of core and accessory genes contributed by each strain are shown in Figure 1.2 B. A visual representation of the total presence and absence variants of the pangenome where

genomic differences in accessory components as well as the homogeneity of the core genome across the strains can be observed (Figure 2C). The phylogenetic tree produced that was used as input for both the gene-pair coincidence (GPC) analysis and the ROARY plots script is shown in supplementary figure 1.1.

The assignment of GO terms resulted in 19,323 annotations (Figure 1.3, Supplementary figure 1.2). Among the genes annotated and assigned to biological processes (BP) within the pangenome, 3,828 are dedicated to cellular processes, 3109 to metabolic processes, 930 to localization, 761 to biological regulation, 726 to the regulation of biological processes, 514 to the response to stimulus, 191 for signaling, 152 for the interspecies interaction between organisms, 101 for locomotion, 65 to viral processes, 46 for the negative regulation of biological processes, 43 for detoxification, 42 to developmental processes, 35 for positive regulation of biological processes, 36 for reproduction, 12 for nitrogen utilization, 5 for carbon utilization, 4 for multicellular organismal process, and one for immune system process (Figure 1.3 A).

Among the genes that are assigned to molecular functions (MF), 3182 are for catalytic activity, 2713 for binding activity, 676 for transporter activity, 263 ATP-dependent activity, 252 with transcription regulator activity, 89 with molecular transductor activity, 82 with structural molecule activity, 37 with small molecule sensor activity, 34 for antioxidant activity, 30 with toxin activity, 23 with translocation regulation activity, 15 with molecular function activity, 12 for cytoskeletal motor activity, 6 for molecular carrier activity, and finally one assigned with nutrient reservoir activity (Figure 13 B). Among the genes assigned to cellular components, 2,894 are assigned as a cellular anatomical entity, 210

are protein-containing complexes, and 2 are virion components (Figure 1.3 C). Further break downs of these groups are available in the supplementary figure 1.2.

To determine the relationship between phylogeny of bacterial strains and their pathogenicity on *Allium* hosts, a phenotypic tree based on SNPs of core genes was constructed using RAxML and PanSeq (Figure 4) and is visually represented using the Interactive Tree of Life online tool (Letunic and Bork. 2016). When assessing the phylogenetic tree in its totality, it is difficult to determine a precise pattern except for strains from the same year of isolation tend to group together. This potentially indicates that these strains in the same group are genetically closely related. Despite this lack of obvious pattern in the overview of the phylogenetic tree, there are several clades where the terminal taxa are sorted based on their pathogenicity. Overall, these results provide support that changes in pathogenicity are not the result of strain lineage, but rather an expansive accessory genome.

Genome wide association studies identify potential pathogenicity and virulence factors associated with *P. ananatis* affecting *A. porrum* and *A. fistulosum x A. cepa*. A pangenome utilizing ROARY was built, and the strength of gene association to the pathogenic phenotype on seedlings (*A. porrum* and *A. fistulosum x A. cepa*) was calculated using SCOARY. A total of 836 genes were found associated with RSN phenotype in *A. fistulosum* x *A. cepa* ($p<=0.05$). Further, to avoid false positive associations, Benjamini-Hochberg correction (BHC) and Bonferroni correction were relied upon ($p<=0.05$) and as a result only 50 genes were found significantly associated with the pathogenic phenotype. Within this set of 836 genes, we found two divergent copies of phosphoenolpyruvate mutase (*pepM*) genes, annotated as “phosphonopyruvate

hydrolase.” The first gene *hvrA/pepM* belongs to the previously described ‘HiVir’ cluster. Among the top twenty significantly associated genes in *P. ananatis* affecting *A. cepa* x *A. fistulosum*, the HiVir cluster genes were found to be prominent (Supplementary file 1.1).

The second *pepM* gene belonged to a separate gene cluster (pgb) previously described by Polidore *et al*. (2021), which ranked at 451 based on naïve significance value. Five phosphatase genes, one gene related to chemotaxis, three related to virulence-region associated *virB*, and several genes in the previously described *alt* gene cluster (Stice *et al.* 2018. Stice *et al.,* 2020. Jain *et al.,* 2018). Two copies of the *fliC* gene, which encodes the flagellin monomer, were also found to be related with *P. ananatis’ pathogenicity* on *A. fistulosum x A. cepa* (He *et al.,* 2012. Macnab, 2003). Flagellar motility has been previously observed to be important for onion leaf virulence (Weller-Stuart *et al.* 2017). Among the associated genes, we also screened for genes (annotated or hypothetical) to assess if they occur in clusters. Within the top 50 significantly associated genes we found at least five hypothetical gene clusters (group_4714-4719, group_3715-3726, group_5180-5182, group_5653-5660, group_5704-5778) as well as the HiVir gene cluster (Supplementary file 1.1).

Using *A. porrum* pathogenic strains, a total of 243 genes were found associated (naïve significance of p <= 0.05) with the pathogenic phenotype. However, none of the predicted genes were associated with the phenotype when the Bonferroni correction, and BHC were applied. When selectively screened for previously described genes known for pathogenicity and virulence in onion (HiVir genes, *alt*), we observed the HiVir cluster to be significantly associated with the phenotype; however, it ranked lower (rank: 91-101) compared to other annotated or hypothetical genes.

We also found significantly associated genes (*n*=123), which were shared between the two hosts, with 48 of the top 50 associated genes in *A. fistulosum x A. cepa* occurred in both GWAS results (Supplementary file 1.1). Some of the known genes that were shared between the *A. fistulosum x A. cepa* and *A. porrum* include the entire HiVir gene cluster, *pemK_2* (mRNA interferase), *soj* (sporulation initiation inhibitor protein), *parM* (Plasmid segregation protein), *umuD* (protein UmuD), *tibC* (glycosyltransferase), *ycaD* (uncharacterized MFS transporter), *dadA* (D-amino acid dehydrogenase 1), *frbC* (2-phosphonomethylmalate synthase), *amiD* (N-acetylmuramoyl-L-alanine amidase), and *rfbB* (dTDP-glucose 4,6-dehydratase). The genes that constitute the thiosulfinate tolerance cluster (*alt*) only appeared in *A. fistulosum x A. cepa* association with the phenotype with their annotations following annotations; *xerC* (tyrosine recombinase XerC), *altA*/*nemA* (N-ethylmaleimide reductase), *gor* (glutathione reductase), *altJ/osmC* (peroxiredoxin OsmC), and *altD/trxA* (thioredoxin) (Supplementary file 1.1). Genes that are members of the larger OVRA region, but not *alt*-specific genes were also associated with the pathogenicity phenotype, and they include *rbsC* (ribose import permease protein RbsC), *rbsB* (ribose import binding protein), *rbsA* (ribose import ATP-binding), and *altD/trxA* (thioredoxin) (Supplementary file 1.1).

Use of gene-pair coincidence (GPC) for phenotype independent determination of pathogenicity and virulence factors

Gene-pair association of the *P. ananatis* pan-genome resulted in a total of 165 genes separated into 39 individual groups (Table 1.2, Supplementary file 1.2). Of the 165 genes, 45 genes are shared with the genes that are predicted based on GWAS for pathogenic phenotype on *A. fistulosum x A. cepa* and only two genes are shared with the genes that

are predicted via GWAS for *A. porrum* pathogenicity (Table 1.2, Supplementary file 1.2). An overview of the associative Coinfinder output can be seen in Figure 1.5. Of the groups that also occurred on the GWAS analysis, 9/10 are saturated with genes that only associate with the pathogenic phenotype for *A. fistulosum* x *A. cepa,* and only one group is saturated with genes that associate with the pathogenic phenotype on *A. porrum.* These results indicate that in our pangenome there is a stronger associative pressure on genes that are specific to one host or the other, and there is no evidence for gene association between genes that associate for both hosts.

Gene-pair dissociation of the *P. ananatis* pangenome resulted in 255 genes separated into 50 groups of dissociated genes (Table 1.3, Supplementary file 1.2). Of the 255 genes, twenty-two are shared with the genes associated with the pathogenic phenotype on *A. fistulosum* x *A. cepa* as predicted by GWAS, whereas only three genes are shared with the pathogenic phenotype on *A. porrum*. Components with dissociating gene-pairs that also occur on the GWAS output are summarized in table 3. A full summary of the genes, their coincidence values, and their groups can be found in supplementary file 2. An overview of the dissociative Coinfinder output can be seen in figure 1.6. Among these groups, 4/6 show a dissociative relationship between genes that associate with the *A. fistulosum* x *A. cepa* disease phenotype and genes that do not associate with the pathogenic phenotype on either host. There is one group of genes that with the *A. porrum* pathogenic phenotype and genes that do not associate with the disease phenotype on either host. Finally, there is one group where there is a dissociative relationship between genes associated with the pathogenic phenotype in *A. fistulosum* x *A. cepa* or *A. porrum*,

but not both (Table 1.3). These results indicate that there may be selective pressure for virulence factors that are host specific, rather than generally useful like the HiVir cluster.

Comparative genomics of strains with *Allium* species-specific pathogenicity.

Five strains were chosen based on their species-specific pathogenicity on *Allium* hosts. The strains PNA 15-3, PANS 99-14 are pathogenic on *A. porrum*, but non-pathogenic on *A. fistulosum x A. cepa* and are all RSN-negative. The strains PNA 07-10, PNA 07-1, and PNA 05-1 are all pathogenic on *A. fistulosum x A. cepa* and RSN-positive, but non-pathogenic on *A. porrum* (Table 1.1). Gene presence and absence were compared manually among these strains. All strains possessed HiVir cluster except for the strain PNA 15-3. The absence of this cluster is the likely cause for the strain's inability to cause foliar lesions on *A. fistulosum x A. cepa*, and necrosis on onion scale.

The strain PNA 15-3, however, carries genes that indicate the presence of a type III secretion system, a virulence pathway that uncommon in *P. ananatis* (Supplementary table 1.2). When comparing this type-III secretion system to those found in Kirzinger et al. (2015), it appears to show similarities with PSI-1b. Attempts to align this sequence to the type III secretion system found in *P. stewartii* subsp. *indologenes* indicated low sequence similarity. When PNA 15-3 was inoculated into tobacco leaf panels, no hypersensitive response was observed (Supplementary figure 1.3).

We found 43 genes proximal to each other surrounding the *stcC* gene in PNA 15-3. A total of 35 genes in the cluster were annotated as hypothetical proteins. The other eight genes were annotated as: *oleC* (olefin beta-lactone synthetase), *gacA* (response regulator GacA), *mxiA* (protein MxiA), *hrcN* (type III secretion ATP synthase HrcN), *spaP* (surface presentation of antigens protein SpaP), *spaQ* (surface

presentation of antigens protein SpaQ), *yscU* (yop proteins translocation protein U), *sctC* (type 3 secretion system secretin), and *dctD* (C4-dicarboxylate transport transcriptional regulatory protein DctD). Further, we utilized NCBI database nucleotide BLAST to query the type III secretion system sequence at default values for the *P. anantis* taxid in the WGS database. We observed an 88% similarity with over 98% query coverage for PANS 99-23, PANS 99-26, PANS 200-1, PNA 86-1, UMFG54 (JACAFO010000015.1), NRRL B-14773 (JACEUA010000002.1), and DE0584 (VDNR01000019.1). Using NCBI database nucleotide blast at default values for the *P. anantis* taxid in the nr nucleotide collection, we observed an 88% identity with LCFJ-001 (CP066803.1) -and FDAARGOS_680 (CP054912.1) chromosomal sequences.

The strains PNA 07-10, PNA 07-1, and PNA 05-1 shared several genes that do not occur in PNA 15-3, or PANS 99-14. Apparent gene clusters shared by these *A. fistulosum* x *A. cepa* pathogenic strains are described in the following paragraph. The first major gene cluster shared only by PNA 07-10, PNA 07-1, and PNA 05-1 strains is the *alt* cluster, followed by a hypothetical gene cluster consisting of 9 hypothetical genes, and 3 annotated genes: *argT_3* (lysine/arginine/ornithine-binding periplasmic protein), group_2282 (ureidoglycolate lyase), and *dapL* (LL-diaminopimelate aminotransferase). Then, there is a small gene cluster consisting of *bepF* (efflux pump periplasmic linker BepF), group_5443 (adaptive-response sensory-kinase SasA), *phoP_3*, (alkaline phosphatase synthesis transcriptional regulatory protein PhoP), and group_5445 (hypothetical protein) and another small gene collection consisting of *parA_2* (plasmid partition protein A), *yedK_1* (SOS response-associated protein YedK), *ppaC* (putative manganese-dependent inorganic pyrophosphatase). Another small cluster with *crcB_2*

(putative fluoride ion transporter CrcB) and two hypothetical proteins. Finally, they share a moderate gene cluster of 10 hypothetical genes and the annotated genes of: *virB_2* (virulence regulon transcriptional activator VirB), *uspA_2* (universal stress protein A), *galE_2* (UDP-glucose 4-epimerase), *ybjJ_2* (inner membrane protein YbjJ), *nudK_3* (GDP-mannose pyrophosphatase NudK), group_5271 (phosphorylated carbohydrates phosphatase), *mtnP*(S-methyl-5'-thioadenosine phosphorylase), *gph_2* (phosphoglycolate phosphatase), *arnB_2* (UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase), *arnB_3* (UDP-4-amino-4-deoxy-L-arabinose—oxoglutarate aminotransferase), *perA* (GDP-perosamine synthase), *iolG_5* (inositol 2-dehydrogenase/D-chiro-inositol 3-dehydrogenase). Of these 60 genes, none of them are associated with the *A. porrum* disease phenotype and 58 genes are associated with the *A. fistulosum* x *A. cepa* disease phenotype. The two genes that are not associated with the *A. fistulosum* x *A. cepa* disease phenotype are annotated as hypothetical genes. Of these 60 genes, only one appears in the gene pair dissociation component 30 as the hypothetical gene “tar” dissociating with group_397 (Supplementary table 1.2).

Apparent gene clusters shared by the *A. porrum* pathogenic PNA 15-3 and PANS 99-14 are described in the following paragraph. The first moderate gene cluster has 7 hypothetical genes and *caf1M* (chaperone protein Caf1M). The second gene cluster is composed of 7 hypothetical genes as well as *amiD_3* (N-acetylmuramoyl-L-alanine amidase AmiD), *yraI_2* (putative fimbrial chaperone YraI), *htrE_2* (outer membrane usher protein HtrE), *fimC* (chaperone protein FimC). For the third gene cluster, only half is actually shared between the two strains, with 99-14 taking *oatA_1* (O-acetyltransferase OatA), group_5288 (hypothetical protein), group_5289 (virulence regulon transcriptional

activator VirB) and group_5290 (hypothetical protein), with the remaining genes group_5494 (hypothetical protein), group_5495 (HTH-type transcriptional regulator PgrR), group_5496 (hypothetical protein), *iolS_2* (aldo-keto reductase IolS), and *ywrO_2* (general stress protein 14) shared between the two strains. The following cluster is a completely hypothetical cluster of 7 genes, followed by a cluster of three genes and group_7087 (replicative DNA helicase). Of these 39 genes, none appeared in the *A. porrum* GWAS output whereas 28 genes did appear in the *A. fistulosum* x *A. cepa* GWAS output. Furthermore, of these genes, none appeared in any gene-pair dissociation output. However, 19 of these genes appeared in the gene-pair coincidence output within components 9 (N =1), 11 (N = 5), 31 (N = 3), 37 (N = 2), 38 (N = 3), and 39 (N = 4) (Supplementary table 1.2).

The HiVir gene cluster found in the RSN-positive, *A fistulosum* x *A. cepa* pathogenic strains PNA 07-10, PNA 07-1, and PNA 05-1 contained no single nucleotide polymorphism (SNP) compared to that of the RSN-positive, wild type *P. ananatis* PNA 97-1 (Figure 7). However, three unique SNPs that resulted in missense mutations were identified in the RSN-negative, *A. porrum* pathogenic strain PANS 99-14. These mutations included alanine (A) to valine (V) change in amino acid position 7 in *hvrA* (*pepM*) gene, glutamine (Q) to lysine (K) change in amino acid position 352 in *hvrB* gene and, lysine (K) to arginine (R) change in amino acid position 11 in *hvrH* gene. These variant SNPs could be associated with disruption of the Pantaphos pathway and loss of necrosis-associated phenotypes (Figure 1.7).

Description of a "pgb" gene cluster in *P. ananatis*

Comparative genome analysis focusing on only strains that are pathogenic on *A. porrum*, but non-pathogenic on *A. fistulosum* x *A. cepa* (PNA 15-3, PANS 99-14) vs. strains (PANS 99-11, PANS 99-12, PNA 06-4, PNA 99-9) that are pathogenic and highly aggressive on both hosts identified another *pepM* gene, which appeared to be a member of a secondary phosphonate biosynthetic cluster (Table 1.4). In 8/92 of *P. ananatis* strains (PANS 02-12, PANS 99-31, PANS 200-2, PANS 2-5, PANS 2-7, PANS 2-8, PANS 99-11, PANS 99-12) there was a putative phosphonate biosynthetic cluster with 14 genes and a total length of approximately 19,000 bp (Table 1.2 and Figure 1.8). This cluster shows high sequence similarity to the pgb-cluster mentioned in Polidore *et a*l. (2021), which was not responsible for generating onion-bulb rot symptoms. In our annotations, the left-flank of the cluster begins with a prophage integrase *intS* and is followed by the *pepM* phosphoenolpyruvate mutase (5' to 3' 900bp long). The following *cpdA* is a 3',5'-cyclic adenosine monophosphate phosphodiesterase *cpdA* (3' to 5' 771bp). The third component of the cluster is *fabG* encoding 3-oxoacyl-[acyl-carrier-protein] reductase FabG3 (765 bp 3' to 5'). The fourth gene is a phosphonopyruvate decarboxylase, *aepY* (1176bp, 3' to 5'). The fifth gene of the cluster is *phnW* encoding for 2-aminoethylphosphonate-pyruvate transaminase (1095, 5' to 3'). The sixth component is *asnB1* encoding putative asparagine synthetase [glutamine-hydrolyzing] (1,758bp, 3' to 5'). Following *asnB2* is spsI1, encoding for Bifunctional IPC transferase and DIPP synthase (771 bp 5' to 3'). The gene *asd1* follows *spsl1* that encodes aspartate-semialdehyde dehydrogenase (1056 bp 5' to 3'). The ninth component of the cluster is the MFS 1 transporter (1227bp, 5' to 3'). The tenth component is *glyA1*, a serine hydroxymethyltransferase (1359bp, 5' to 3'). The

CDP-alcohol phosphatidyltransferase is the eleventh gene (600bp 5' to 3'). The twelfth gene is *spsI2*, a glucose-1-phosphate adenylyl/thymidylyltransferase Bifunctional IPC transferase and DIPP synthase (720bp, 5' to 3'). The thirteenth gene is *aspC1*, aspartate aminotransferase (1173 bp, 5' to 3'). The fourteenth gene is UDP-2,3-diacylglucosamine diphosphatase (762bp, 3' to 5'). Following the UDP-2,3-diacylglucosamine diphosphatase is another transposase. An interesting observation of this phosphonate cluster is the inclusion of phosphonopyruvate decarboxylase directly within the set of genes. This characteristic is unique when comparing it to the HiVir. The phosphonopyruvate decarboxylase has been described to play a critical role in the generation of phosphonates via the stabilization of the PEP mutase reaction (Supplementary table 1.3) (Kirzinger, Butz and Stavrinides. 2015).

Presence and absence of *alt*, HiVir, pgb, gene clusters

Of the 92 tested strains, thirty-five do not contain a complete *alt* gene cluster while the remaining fifty-seven do possess the entire gene cluster. Of the 92 tested strains, twenty-two lacked a complete HiVir cluster and the remaining seventy-strains possessed the entire gene cluster. Forty-five strains that produced foliar lesions had both *alt* and HiVir clusters, whereas 20 strains had only the HiVir gene cluster. Three strains only had *alt* whereas seven strains lacked both gene clusters. Some foliar lesions were formed by the seven strains (PANS 99-22, PANS 99-26, PANS 200-1, PANS 99-36, PNA 98-3, PNA 11-1, and PNA 15-3) that lacked both clusters, however the lesions varied in size and consistency between replicates. The strains that lacked both gene clusters were also RSN-negative. Six out of seven strains showed some degree of foliar lesions on *A. cepa x A. fistulosum*; however, PNA 15-3 showed moderately aggressive lesion length on *A.*

porrum but did not produce any foliar lesion on *A. cepa x A. fistulosum*. Of the 92 tested strains, only 8 strains contained the pgb cluster, while 84 strains lacked it (Figure 1.8).

Comparison of phosphonate biosynthetic clusters, pgb vs. HiVir in *P. ananatis*

When comparing the pgb cluster against the HiVir cluster, only the annotated *pepM* and MFS transporter are found to be common features (Table 1.4). In both clusters, the phosphoenolpyruvate mutase occurs first, with the MFS transporter being at the center of the cluster (5' to 3': 9th gene in pgb and 5' to 3': 9th in HiVir). There are no other shared annotated genes between the clusters. However, sequence alignment using Clustal Omega revealed 48.3% similarity between *pepM* from HiVir and pgb cluster. Similarly, the MFS transporters from HiVir and pgb clusters displayed 47.6% sequence similarity.

Role of *pepM* gene in the pgb biosynthetic cluster

Based on the RSN assay, wild-type strain (PANS 02-18) and the single *pepM* mutant strain in the pgb cluster (Δ*pepM*$_{pgb}$) produced considerably large necrotic areas on red-onion scale compared to the single *pepM* mutant strain in the HiVir cluster (Δ*pepM*$_{HiVir}$) (Figure 1.9 A). Based on the seedling pathogenicity assay, the single *pepM* mutant strain in the HiVir cluster (Δ*pepM*$_{HiVir}$) had significantly lower necrotic lesion length on both *Allium* hosts compared to the wild-type strain (PANS 02-18) and the single *pepM* mutant strain in the pgb cluster (Δ*pepM*$_{pgb}$) (Figure 1.9 B and C). In both hosts, the deletion of *pepM* gene in the pgb cluster did not significantly affect the foliar lesion length compared to the wild-type strain (Figure 1.9 D and E). While the double mutant strain where *pepM* genes were deleted in both the HiVir and the pgb clusters (Δ*pepM*$_{HiVir}$Δ*pepM*$_{pgb}$) appears to have a higher average lesion length than that of the single mutant strain (Δ*pepM*$_{HiVir}$) on *A. porrum* (Figure 1.9 B and D). In *A. fistulosum* × *A. cepa*, the lesion lengths did not

differ significantly between the double mutant strain and the single mutant strain (Δ*pepM*$_{HiVir}$).

Discussion

Pathogenicity and aggressiveness of *P. ananatis* phenotyping on *A. porrum* and *A. fistulosum*

Phenotyping of 92 *P. ananatis* strains displayed variability in the level of aggressiveness in both *Allium* spp. A considerably higher percentage of strains were either non-pathogenic or less aggressive on *A. porrum* (82.6%) than on *A. fistulosum x A. cepa* (65.1%). Also, a considerable percentage of strains were moderately or highly aggressive on *A. fistulosum x A. cepa* (34.9%) compared to *A. porrum* (17.4%). Interestingly, when a subset of less-aggressive strains (on both *Allium* spp.) was previously assayed on onion seedlings they were moderately-to-highly aggressive on onion (Stice *et al.,* 2018. Agarwal *et al.,* 2021). These observations potentially indicate that both *A. porrum* and *A. fistulosum x A. cepa* are inherently less susceptible to *P. ananatis* compared with the typical bulb onion. Also, when *A. porrum* and *A. fistulosum x A. cepa* were compared to each other, the former tends to show less severe symptoms compared to the later. However, we acknowledge that only one cultivar of each *Allium* spp. was evaluated, and it is possible that other cultivars or varieties of these hosts might show a range of susceptibility to *P. ananatis*. One aspect of these observations might be in part explained by the genetic nature of the *A. fistulosum x A. cepa*. cv. Guardsman itself. Due to being a hybrid between bunching onions and the typical bulb onion, it may be reasonable to expect *A. fistulosum x A. cepa* to be more susceptible to *P. ananatis* strains that were collected from symptomatic *A. cepa* tissues. However, without deeper genetic investigation of this hybrid

we cannot predict for certain if any susceptibility-related genes or phenotypes were inherited. The *P. ananatis* culture collection used in this study also favors areas where onions are grown such as the Vidalia region and Tift County, which may provide a bias for aggressive *P. ananatis* strains on cultivars that are hybridized with *A. cepa*. Despite this, we recovered some strains that were more aggressive on *A. porrum* than *A. fistulosum x A. cepa*. Examples of these are PANS 99-11, PANS 99-12, and PNA 06-4 (Table 1.1).

HiVir gene cluster, previously identified as critical for red onion scale necrosis and *A. cepa* pathogenicity, is also important for foliar pathogenicity in *A. porrum*, *A. fistulosum* x *A. cepa*

Of the tested strains 56 showed a positive reaction to the RSN assay, while 36 did not. Most of the strains that were pathogenic on *A. fistulosum* × *A. cepa* were also able to cause necrosis on red onion scale. Based on the previous reports, HiVir is important for RSN-positive phenotype and foliar necrosis in onion and it is likely that foliar lesions on *A. fistulosum* × *A. cepa* is also governed by the same gene cluster. Similarly, in *A. porrum,* a trend between RSN-positive phenotype and foliar pathogenicity was observed with majority of the strains. Further mutational analysis also indicated that HiVir is important for foliar pathogenicity on both *A. fistulosum* × *A. cepa* and *A. porrum*. However, we identified several strains that did not follow this trend. For example, while having a complete HiVir cluster and being RSN-positive, the PANS 19-8, and PANS 19-10 strains were unable to inflict foliar lesions on *A. porrum* or *A. fistulosum* x *A. cepa*. This indicates that either mutation in their nucleotides/SNPs or other/*alt*ernative pathogenicity factors might be involved with this group of strains. Consistent with prior observations by Polidore

et al. [13], we also observed that the pgb cluster is not important for foliar pathogenicity in these *Allium* spp. The PANS 02-18 Δ*pepM*$_{pgb}$ and wild-type strains displayed similar foliar pathogenicity in *A. porrum* or *A. fistulosum* x *A. cepa.*

Pan-Genome of *P. ananatis*

In this study, we generated a pan-genome of 92 *P. ananatis* strains where we identified a conserved core genome of 2,914 genes, with a larger accessory genome of 9,196 genes. Earlier pan-genome reports identified similar values of core genes, with varying numbers of *P. ananatis* strains used for the analysis (Stice *et al.,* 2018. De Maayer *et al.,* 2014. Agarwal *et al.,* 2021. Sheibani-Tezerji *et al.,* 2015). In this work, however, we have a larger set of accessory genes compared to previously observed pan-genome study by Agarwal *et al.,* (2021). The authors observed 6,808 cloud genes compared to 9,196 cloud genes in our current study. This discrepancy is likely due to use of larger number of diverse strains in this study compared to Agarwal *et al.* (2021) and is a reasonable increase for an open pangenome (Costa *et al.,* 2020). Due to the cosmopolitan nature of *P. ananatis* and the extensive host range, it is entirely plausible that the bacterium would have a sizable pangenome when comparing populations from diverse hosts (De Maayer *et al.,* 2014). It may be prudent to utilize a larger collection of strains from non-*Allium* hosts for further pan-genomic assessment, where accessory genes may discriminate strains for sub-species delineation. Comparisons of these resulting sub-species may likely be more informative for the detection of novel virulence factors. In addition, it is also likely that there may be multiple copies of similar annotated genes (PPC, PPC_1) (Supplementary table 1.4), which may artificially inflate the true nature of the pangenome. Even if the pangenome was artificially inflated, the GWAS results found significant

association of genes that are known virulence factors (such as the HiVir cluster), as well as putative virulence factors that were found previously to be associated with foliar pathogenicity in *A. cepa* (Agarwal *et al.,* 2021).

We identified 244 genes that were significantly associated (naïve) with foliar pathogenicity in *A. porrum* whereas 836 total genes were associated with foliar pathogenicity in *A. fistulosum x A. cepa,* with 77 genes displaying significance agreement between naïve, Bonferroni, and BHC corrections. Among the genes associated for two hosts, 123 genes were shared. For both hosts, the HiVir cluster was found within the top-100 significantly associated genes. The occurrence of this cluster grants some additional credibility to other genes that show stronger significance with phenotypic association. However, most of the genes that occurred in the GWAS output apart from the HiVir cluster are annotated as hypothetical and would require further characterization to determine their relevance. In addition to this, among these 123 genes, 48 of the top 50 strongly statistically significant genes associated with the pathogenic phenotype in *A. fistulosum x A. cepa* are also found to be associated with the pathogenic phenotype in *A. porrum*. While this alone is not enough to declare relevancy, it lends some credibility that these hypothetical genes may be useful as general virulence factors for the *Allium*s *spp.* and should undergo downstream mutational analysis to assess their functions.

Some of the genes with non-hypothetical annotations that were shared between the *A. fistulosum x A. cepa* and *A. porrum* include the entire HiVir gene cluster (some listed as hypothetical), *pemK_2* (mRNA interferase), *soj* (sporulation initiation inhibitor protein), *parM* (Plasmid segregation protein), *umuD* (UmuD, translesion DNA polymerase subunit), *tibC* (glycosyltransferase), *ycaD* (uncharacterized MFS transporter), *dadA* (D-

amino acid dehydrogenase 1), *frbC* (2-phosphonomethylmalate synthase), *amiD* (N-acetylmuramoyl-L-alanine amidase), and *rfbB* (dTDP-glucose 4,6-dehydratase). Upon manual investigation of the local *pemK_2* region there seems to be a repeating pattern of seven genes, three flanking to the left, and four on the right. Utilizing BLAST for these gene sequences against the total gene sequences available for our *P. ananatis* strains shows that the pemK_2 region appears frequently throughout the entirety of the pangenome (Supplementary table 1.5). The *pemK* gene is a known factor in toxin/antitoxin systems that are vital for bacterial competition and function (Poluktova et al., 2017. Lee, Rogers and Stenger, 2012. Klimina, et al., 2013). Unfortunately, there is little information within the literature pertaining to the potential diversity and utility of *P. ananatis'* toxin/antitoxin systems, including *pemK*. Without functional analysis, it is not possible to determine whether it is an *Allium*-specific virulence factor as opposed to a coincidental gene cluster, or if the annotation provided is correct. We also found an antitoxin gene, *higB* that was associated with *A. fistulosum* x *A. cepa* pathogenicity. Despite the lack of information, the region may be a valuable target for a toxin/antitoxin system within our *P. ananatis* strains from Georgia. Another annotated gene of interest includes *tibA*, an adhesin/invasion autotransporter. The sporulation initiation inhibitor protein, soj, is noted as having a "centromere-like function involved in forespore chromosome partitioning inhibition of Spo0A activation" in *Bacillus subtilis (*Kunst *et al.* 1997). *Pantoea ananatis* is not a spore forming bacteria, however the gene's inclusion with other genes that are similarly annotated for DNA manipulation and management may indicate that there is a requirement for maintaining genetic stability. The genes that constitute the *alt* thiosulfinate tolerance *alt* cluster, only appeared on the *A. fistulosum x*

A. cepa GWAS output with their old annotations of *xerC* (tyrosine recombinase XerC), *altA*/*nemA* (N-ethylmaleimide reductase), *gor* (glutathione reductase), *altJ*/*osmC* (peroxiredoxin OsmC), and *altD*/*trxA* (thioredoxin). Non-*alt* members of the OVRA region include *rbsC* (ribose import permease protein RbsC), *rbsB* (ribose import binding protein), *rbsA* (ribose import ATP-binding).

Overall, GWAS was able to determine genes associated with foliar necrosis in *A. fistulosum x A. cepa* and *A. porrum* hosts. Follow-up experiments will test the validity of these hypothetical and annotated gene clusters for their relevance in pathogenicity and virulence in *A. fistulosum x A. cepa* and *A. porrum* hosts.

Comparative genomics of strains pathogenic on *A. porrum* but non-pathogenic on *A. fistulosum* x *A. cepa* against strains that are pathogenic on *A. fistulosum* x *A. cepa* but non-pathogenic on *A. porrum*

By comparing strains that were pathogenic on only *A. fistulosum* x *A. cepa* or *A. porrum* we hoped to significantly reduce the background noise (non-relevant genes from accessory) that may potentially result from the extensive *P. ananatis* pan-genome. Here we found several genes that belonged to strains that were only pathogenic to one host or the other, as well as a few interesting gene clusters.

One of the gene clusters of interest appears to be saturated with genes that were annotated to have some role for a type III secretion system. When using NCBI's nucleotide blast against the nr and whole genome sequence database, there were several other *P. ananatis* strains that shared a high consensus to the sequence (N = 11; in NCBI). Most of the strains in NCBI with an annotated type III secretion-system were isolated in GA (N = 7; in NCBI) with two strains from onions and five strains from weeds. The

remaining four strains with potential type III secretion-system were not isolated from GA. Further work is required to assess if the annotations are correct and investigate their role in onion pathogenicity.

Among the clusters shared by the *A. fistulosum x A. cepa* pathogenic strains PNA 07-10, PNA 07-1, and PNA 05-1, we found the *alt* cluster, two larger gene clusters (N=12 genes, N=22 genes), and two small gene clusters (N=7 genes, N=3 genes). Of the total 60 shared genes, none of them are on the *A. porrum* GWAS output, and 58 appeared on the *A. fistulosum* x *A. cepa* GWAS output. Only one gene appears in the gene pair dissociation component 30 as the hypothetical gene *tar* that dissociates with group_397 (Supplementary file 1.4). Among the clusters shared by the *A. porrum* pathogenic PNA 15-3 and PANS 99-14 there are three gene clusters with 7, 8, and 11 genes, a small gene cluster (N=4 genes) and a third cluster that is only partially shared between the two strains (N=9 genes). Of these 39 genes, 28 were identified through GWAS as associated with the disease phenotype for *A. fistulosum x A. cepa,* but not associated with the disease phenotype for *A. porrum* (Supplementary table 1.2). None of these genes appeared in the gene-pair dissociation output. However, 19 of these genes were found within components 9 (N =1), 11 (N = 5), 31 (N = 3 genes), 37 (N = 2 genes), 38 (N = 3 genes), and 39 (N = 4 genes) in the gene-pair association analysis (Supplementary table 2).

While the inclusion of the *alt* cluster in the *A. fistulosum* x *A. cepa* strains PNA 07-10, PNA 07-1, and PNA 05-1 is unsurprising, as they were isolated from symptomatic onions, its absence from PNA 15-3 is unexpected. The strain PNA 15-3 was isolated from symptomatic onion bulbs, and we would expect the presence of *alt* cluster as it aids in colonization of the bulb (Stice *et al.,* 2020). The strain PANS 99-14 was isolated from an

asymptomatic *Digitaria* spp. and may not need to rely on an *alt* cluster to survive in this environment. Despite this, both strains were able to generate a lesion on the *A. porrum* foliar tissue, and both strains failed to produce a positive result for red-onion scale necrosis assay. These results indicate that some of the shared genes between these strains may enable an increased fitness for the *A. porrum* foliar environment that is not present in the *A. cepa* bulb tissue, or the *A. fiustulosum* x *A. cepa* foliar tissue. The large percentage of these shared genes (19/39) occur on the gene-pair association output, seems to suggest that these genes occur together at a higher frequency than others. Much like how the *alt* cluster provides protection to *P. ananatis* in the thiosulfinate-rich in bulb, it is possible that some of these clusters provide protection to bacteria in the diverse *Allium* foliar environments.

We also aligned the HiVir gene clusters of *A. fistulosum* x *A. cepa* pathogenic (PNA 05-1, PNA 07-1 and PNA 07-10) and *A. porrum* pathogenic strains (PNA 15-3, and PANS 99-14) against each other and against the wild type *P. ananatis* strain PNA 97-1. No single nucleotide polymorphism (SNP) leading to missense mutation was identified in the RSN positive, *A. fistulosum* x *A. cepa* pathogenic PNA 05-1, PNA 07-1 and PNA 07-10 strains. However, several missense mutations were present in the *hvr*A, *hvrB* and *hvrH* genes of the RSN negative, *A. porrum* pathogenic PANS 99-14 strain. These mutations were found only in PANS 99-14 *hvr* genes and not in the *hvr* genes of the RSN positive PNA 97-1. According to Polidore et al. (2021) *hvrA* and *hvrB* genes encode enzymes that are essential for the proposed phosphonate-toxin ‘Pantaphos’ biosynthesis pathway. It is thus possible that the production of phosphonate toxin is compromised by these mutations. However, functional analysis needs to be conducted to confirm the impact of

these mutations. In the case of RSN negative PNA 15-3, the strain lacks HiVir cluster but was still able to cause foliar lesions on *A. porrum*. It is possible that the pathogenicity of *A. porrum* in RSN negative PNA 15-3, and PANS 99-14 strains may be mediated by the genes other than of the HiVir cluster.

Gene-Pair Coincidence

In this work, we utilized gene-pair coincidence as a supporting methodology to predict genes in *P. ananatis* that are relevant for pathogenicity. Hypothetically, genes that are important for survival in pathogenic bacteria should associate throughout a pangenome as their co-occurrence is beneficial for survival. Likewise, gene combinations that compromise survival in specific environments should dissociate with each other as natural selection selects against non-optimized populations. Likewise, as bacteria have limited genomes this extends to redundant genes. In this work, we hypothesized that the utilization of gene-pair coincidence should provide a phenotype-independent method of validation for the predicted genes from the pangenome via the phenotype-dependent GWAS methodology.

Our gene-pair association analysis generated 39 networks with a total of 165 individual genes, where 2 common genes associated with the disease phenotype for *A. porrum* and 45 common genes associating with the disease phenotype for *A. fistulosum x A. cepa*. In gene networks with genes that associated with the disease phenotype (networks: 1, 3, 5, 11, 15, 16, 22, 24, 31, 32), the entire component is found on the GWAS result, indicating that these pathogenicity-associating genes are evolutionarily associating with each other in our pangenome. Of these associative gene pair networks, none of the genes are associated with the disease phenotype for both hosts, only one or the other.

Component 24 is unique in that only genes that with the disease phenotype for *A. porrum* were present. It is possible that these genes provide a unique advantage to overcome *A. porrum* host resistance. Unfortunately, both genes are annotated as hypothetical. The genes found in components 1, 3, 5, 11, 15, 16, 22, 24, 31, and 32 may be more useful survival in the *A. fistulosum x A. cepa* as opposed to being general virulence factors. Gene-pair association has provided support for further investigation of several potential gene groups that would have been harder to distinguish utilizing GWAS alone. Surprisingly, neither the HiVir gene cluster nor the *alt* cluster appear on the associative network. We would assume that they should co-occur as they have been shown to be relevant in *A. cepa* pathogenicity and are quite prevalent in this pangenome. It may be due to the small number of samples used in this study compared to number of strains/genomes needed to close the pan-genome. Another explanation may be due to how ROARY/Coinfinder organize gene information and these known virulence factors were omitted from pairwise analysis due to a lack of orthologous gene families. To determine if the issue was caused by noise due to genes not present in gene clusters, we conducted the analysis after removing single genes from the ROARY csv file. However, this only strengthened resulting p-values, but not the overall result. As such a larger sample size of strains with a more comprehensive accessory genome could be included to better support GPC and GWAS results.

Our gene-pair dissociation analysis generated 50 gene-pair networks with a total of 255 genes, where there are three genes associating with the pathogenic phenotype for *A. porrum* and 22 associating with the disease phenotype for *A. fistulosum x A. cepa*. Here, dissociation is dominated by networks where only a fraction of the dissociating

genes also associates with the disease phenotype for either host. Dissociation networks 10, 17, 23, and 35 contain genes that associate with the pathogenic phenotype for *A. fistulosum x A. cepa* that dissociate with genes that do not associate with the disease phenotype for *A. fistulosum x A. cepa*. Dissociation network 37 is particularly interesting in that it shows dissociation between one hypothetical gene that associates with the pathogenic phenotype for *A. porrum*, and two genes that associate with the disease phenotype for *A. fistulosum x A. cepa.* Whether these hypothetical genes have antagonistic function with each other is unknown; however, it is worth investigating to understand their roles and mechanisms in host-pathogen-environment interactions. Dissociation network 21 is the only network where 2 genes that associate with the pathogenic phenotype in *A. fistulosum x A. cepa* dissociate with two genes that do not associate with the disease phenotype for either host. These results are expected as genes that associate with the foliar pathogenic phenotype should dissociate with genes that do not associate with the same phenotype. Here we do not find HiVir or *alt* dissociating with other genes, indicating that these clusters do not have a competitive interaction that may result in a loss of fitness within our pangenome.

Again, these observations provide some level of confidence that the genes being predicted in the GWAS output are playing a role that enable them to be associated with the foliar pathogenicity phenotype on both *Allium* hosts. The diversity of GPC and their occurrence on the phenotype-dependent analysis enforce the assumption made previously that there could be several mechanisms of causing symptoms in *Allium* species other than phosphonate-based toxins.

References

1. Stumpf S, Kvitko B, Gitaitis R, & Dutta B (2018) Isolation and Characterization of Novel Pantoea stewartii subsp indologenes Strains Exhibiting Center Rot in Onion. Plant Dis 102(4):727-733.
2. Gitaitis RD, Walcott RR, Wells ML, Perez JCD, & Sanders FH (2003) Transmission of *Pantoea ananatis*, Causal Agent of Center Rot of Onion, by Tobacco Thrips, Frankliniella fusca. Plant Dis 87(6):675-678.
3. Stice, S. P., Stumpf, S. D., Gitaitis, R. D., Kvitko, B. H., & Dutta, B. (2018). *Pantoea ananatis* genetic diversity analysis reveals limited genomic diversity as well as accessory genes correlated with onion pathogenicity. Frontiers in Microbiology. https://doi.org/10.3389/fmicb.2018.00184
4. Carr EA, Zaid AM, Bonasera JM, Lorbeer JW, & Beer SV (2013) Infection of Onion Leaves by *Pantoea ananatis* Leads to Bulb Infection. Plant Dis 97(12):1524-1528.
5. Dutta B, et al. (2014) Transmission of *Pantoea ananatis* and P. agglomerans, causal agents of center rot of onion (*Allium cepa*), by onion thrips (Thrips tabaci) through feces. Phytopathology 104(8):812-819.
6. Dutta B, et al. (2016) Interactions Between Frankliniella fusca and *Pantoea ananatis* in the Center Rot Epidemic of Onion (*Allium cepa*). Phytopathology 106(9):956-962.
7. De Maayer P, et al. (2014) Analysis of the *Pantoea ananatis* pan-genome reveals factors underlying its ability to colonize and interact with plant, insect and vertebrate hosts. BMC Genomics 15:404.

8: Kido, K., Hasegawa, M., Matsumoto, H., Kobayashi, M., & Takikawa, Y. (2010). *Pantoea ananatis* strains are differentiated into three groups based on reactions of tobacco and welsh onion and on genetic characteristics. Journal of General Plant Pathology. https://doi.org/10.1007/s10327-010-0230-9

9: Wang, Y. J., Lin, C. H., & Huang, C. J. (2018). Occurrence, identification, and bactericide sensitivity of *Pantoea ananatis* causing leaf blight on welsh onion in Taiwan. Journal of Plant Pathology. https://doi.org/10.1007/s42161-018-0067-1

10: Chang, J. H., Desveaux, D., & Creason, A. L. (2014). The ABCs and 123s of bacterial secretion systems in plant pathogenesis. Annual Review of Phytopathology.

11. Asselin JE, Bonasera JM, & Beer SV (2018) Center rot of onion (*Allium cepa*) caused by *Pantoea ananatis* requires pepM, a predicted phosphonate-related gene. Mol Plant Microbe Interact.

12. Takikawa Y, and Kubota, Y. (2018) A genetic locus determining pathogenicity of *Pantoea ananatis* (Abstr.). Phytopathology.

13. Polidore, A. L. A., Furiassi, L., Hergenrother, P. J., Metcalf, W. W., & Handelsman, J. (2021). A Phosphonate Natural Product Made by *Pantoea ananatis* is Necessary and Sufficient for the Hallmark Lesions of Onion Center Rot Downloaded from. https://doi.org/10.1128/mBio.03402-20

14. Stice, S. P., Thao, K. K., Khang, C. H., Baltrus, D. A., Dutta, B., & Kvitko, B. H. (2020). Thiosulfinate Tolerance Is a Virulence Strategy of an Atypical Bacterial Pathogen of Onion. Current Biology, 30(16), 3130-3140.e6. https://doi.org/10.1016/j.cub.2020.05.092

15. Chewapreecha, C., Harris, S.R., Croucher, N.J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D.M., Mather, A.E., Page, A.J., Salter, S.J., Harris, D., Nosten, F., Goldblatt, D., Corander, J., Parkhill, J., Turner, P. and Bentley, S.D. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. Nature Genetics, 46(3), pp.305–309. doi:10.1038/ng.2895.

16. Laabei, M., Recker, M., Rudkin, J.K., Aldeljawi, M., Gulay, Z., Sloan, T.J., Williams, P., Endres, J.L., Bayles, K.W., Fey, P.D., Yajjala, V.K., Widhelm, T., Hawkins, E., Lewis, K., Parfett, S., Scowen, L., Peacock, S.J., Holden, M., Wilson, D. and Read, T.D. (2014). Predicting the virulence of MRSA from its genome sequence. Genome Research, [online] 24(5), pp.839–849. doi:10.1101/gr.165415.113.

17. Sheppard, S.K., Didelot, X., Meric, G., Torralbo, A., Jolley, K.A., Kelly, D.J., Bentley, S.D., Maiden, M.C.J., Parkhill, J. and Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proceedings of the National Academy of Sciences, 110(29), pp.11923–11927. doi:10.1073/pnas.1305559110.

18. Desjardins, C.A., Cohen, K.A., Munsamy, V., Abeel, T., Maharaj, K., Walker, B.J., Shea, T.P., Almeida, D.V., Manson, A.L., Salazar, A., Padayatchi, N., O'Donnell, M.R., Mlisana, K.P., Wortman, J., Birren, B.W., Grosset, J., Earl, A.M. and Pym, A.S. (2016). Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D -cycloserine resistance. Nature Genetics, [online] 48(5), pp.544–551. doi:10.1038/ng.3548.

19. Farhat, M.R., Shapiro, B.J., Kieser, K.J., Sultana, R., Jacobson, K.R., Victor, T.C., Warren, R.M., Streicher, E.M., Calver, A., Sloutsky, A., Kaur, D., Posey, J.E., Plikaytis,

B., Oggioni, M.R., Gardy, J.L., Johnston, J.C., Rodrigues, M., Tang, P.K.C., Kato-Maeda, M. and Borowsky, M.L. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nature Genetics, 45(10), pp.1183–1189. doi:10.1038/ng.2747.

20. Earle, S.G., Wu, C.-H., Charlesworth, J., Stoesser, N., Gordon, N.C., Walker, T.M., Spencer, C.C.A., Iqbal, Z., Clifton, D.A., Hopkins, K.L., Woodford, N., Smith, E.G., Ismail, N., Llewelyn, M.J., Peto, T.E., Crook, D.W., McVean, G., Walker, A.S. and Wilson, D.J. (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nature Microbiology, 1(5). doi:10.1038/nmicrobiol.2016.41.
21. Hall, B.G. (2014). SNP-Associations and Phenotype Predictions from Hundreds of Microbial Genomes without Genome Alignments. PLoS ONE, 9(2), p.e90490. doi:10.1371/journal.pone.0090490.
22. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. (2015) Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. Proc Natl Acad Sci.;112(27):E3574–81.
23. Lees, J.A., Vehkala, M., Välimäki, N., Harris, S.R., Chewapreecha, C., Croucher, N.J., Marttinen, P., Davies, M.R., Steer, A.C., Tong, S.Y.C., Honkela, A., Parkhill, J., Bentley, S.D. and Corander, J. (2016). Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nature Communications, [online] 7(1), p.12797. doi:10.1038/ncomms12797.

24. Medini, D., Donati, C., Tettelin, H., Masignani, V. and Rappuoli, R. (2005). The microbial pan-genome. Current Opinion in Genetics & Development, [online] 15(6), pp.589–594. doi:10.1016/j.gde.2005.09.006.

25. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proc Natl Acad Sci U S A. 2005;102(39):13950–5

26. Brockhurst, M.A., Harrison, E., Hall, J.P.J., Richards, T., McNally, A. and MacLean, C. (2019). The Ecology and Evolution of Pangenomes. Current Biology, 29(20), pp.R1094–R1103. doi:10.1016/j.cub.2019.08.012.

27. Brynildsrud, O., Bohlin, J., Scheffer, L. and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biology, 17(1). doi:10.1186/s13059-016-1108-8.

28. Whelan, F. J., Rusilowicz, M., & McInerney, J. O. (2019). Coinfinder: Detecting Significant Associations and Dissociations in Pangenomes. BioRxiv, 859371. https://doi.org/10.1101/859371

29. Bruns, H., Crüsemann, M., Letzel, A. C., Alanjary, M., McInerney, J. O., Jensen, P. R., Schulz, S., Moore, B. S., & Ziemert, N. (2018). Function-related replacement of bacterial siderophore pathways. ISME Journal. https://doi.org/10.1038/ismej.2017.137

30. Walcott, R. R., Gitaitis, R. D., Castro, A. C., Sanders, F. H., & Diaz-Perez, J. C. (2002). Natural infestation of onion seed by *Pantoea ananatis*, causal agent of center rot. Plant Disease.

31. Gitaitis, R.D. and Gay, J.D. (1997). First Report of a Leaf Blight, Seed Stalk Rot, and Bulb Decay of Onion by Pantoea ananas in Georgia. Plant Disease, 81(9), pp.1096–1096. doi:10.1094/pdis.1997.81.9.1096c.

32. Dutta, B., Barman, A. K., Srinivasan, R., Avci, U., Ullman, D. E., Langston, D. B., & Gitaitis, R. D. (2014). Transmission of *Pantoea ananatis* and P. agglomerans, Causal agents of center Rot of Onion (*Allium cepa*), by onion thrips (thrips tabaci) through feces. Phytopathology. https://doi.org/10.1094/PHYTO-07-13-0199-R

33. Goszczynska, T., Venter, S. N., & Coutinho, T. A. (2006). PA 20, a semi-selective medium for isolation and enumeration of *Pantoea ananatis*. Journal of Microbiological Methods. https://doi.org/10.1016/j.mimet.2005.05.004

34. Dutta, B., Gitaitis, R., Smith, S., & Langston, D. (2014). Interactions of seedborne bacterial pathogens with host and non-host plants in relation to seed infestation and seedling transmission. PLoS ONE. https://doi.org/10.1371/journal.pone.0099215

35. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu170

36. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. https://doi.org/10.1089/cmb.2012.0021

37. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, [online] 30(14), pp.2068–2069. doi:10.1093/bioinformatics/btu153.

38. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications, 9(1), 1–8. https://doi.org/10.1038/s41467-018-07641-9

39. BioBam. (2022). OmicsBox Bioinformatics Software | Biobam. [online] Available at: https://www.biobam.com/omicsbox [Accessed 2 Nov. 2022].

40. Götz S., Garcia-Gomez JM., Terol J., Williams TD., Nagaraj SH., Nueda MJ., Robles M., Talon M., Dopazo J. and Conesa A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 36(10), 3420-35.

41. Laing, C., Buchanan, C., Taboada, E.N., Zhang, Y., Kropinski, A., Villegas, A., Thomas, J.E. and Gannon, V.P. (2010). Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics, 11(1), p.461. doi:10.1186/1471-2105-11-461.

42. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), pp.1312–1313. doi:10.1093/bioinformatics/btu033.

43. Price, M.N., Dehal, P.S. and Arkin, A.P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution, 26(7), pp.1641–1650. doi:10.1093/molbev/msp077.

44. Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv421

45. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A. and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Research. doi:10.1093/nar/gkac240.

46. Jo Ann E. Asselin, Jean M. Bonasera, and Steven V. Beer. (2018) Center Rot of Onion (*Allium cepa*) Caused by *Pantoea ananatis* Requires pepM, a Predicted Phosphonate-Related Gene. Molecular Plant-Microbe Interactions. 31:12, 1291-1300

47. Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research, 44(W1), W242–W245. https://doi.org/10.1093/nar/gkw290

48. Stice, S. P., Shin, G. Y., De Armas, S., Koirala, S., Galván, G. A., Siri, M. I., Severns, P. M., Coutinho, T., Dutta, B., & Kvitko, B. H. (2021). The Distribution of Onion Virulence Gene Clusters Among Pantoea spp. Frontiers in Plant Science, 12(March), 1–14. https://doi.org/10.3389/fpls.2021.643787

49. He, Y., Xu, T., Fossheim, L.E. and Zhang, X.-H. (2012). FliC, a Flagellin Protein, Is Essential for the Growth and Virulence of Fish Pathogen Edwardsiella tarda. PLoS ONE, 7(9), p.e45070. doi:10.1371/journal.pone.0045070.

50. Macnab, R.M. (2003). How Bacteria Assemble Flagella. Annual Review of Microbiology, 57(1), pp.77–100. doi:10.1146/annurev.micro.57.030502.090832.

51. Weller-Stuart, T., Toth, I., De Maayer, P., and Coutinho, T. (2017). Swimming and twitching motility are essential for attachment and virulence of *Pantoea ananatis* in onion seedlings. Mol Plant Pathol 18, 734-745.

52. Kirzinger, M. W. B., Butz, C. J., & Stavrinides, J. (2015). Inheritance of Pantoea type III secretion systems through both vertical and horizontal transfer. Molecular Genetics and Genomics, 290(6), 2075–2088. https://doi.org/10.1007/s00438-015-1062-2

53. Metcalf, W. W., & Van Der Donk, W. A. (2009). Biosynthesis of phosphonic and phosphinic acid natural products. In Annual Review of Biochemistry (Vol. 78, pp. 65–94). NIH Public Access. https://doi.org/10.1146/annurev.biochem.78.091707.100215

54. Agarwal, G., Choudhary, D., Stice, S. P., Myers, B. K., Gitaitis, R. D., Venter, S. N., Kvitko, B. H., & Dutta, B. (2021). Pan-Genome-Wide Analysis of *Pantoea ananatis* Identified Genes Linked to Pathogenicity in Onion. Frontiers in Microbiology, 12(August), 1–19. https://doi.org/10.3389/fmicb.2021.684756

55. Sheibani-Tezerji, R., Naveed, M., Jehl, M. A., Sessitsch, A., Rattei, T., & Mitter, B. (2015). The genomes of closely related *Pantoea ananatis* maize seed endophytes having different effects on the host plant differ in secretion system genes and mobile genetic elements. Frontiers in Microbiology, 6(MAY), 1–16. https://doi.org/10.3389/fmicb.2015.00440

56. Costa, S. S., Guimarães, L. C., Silva, A., Soares, S. C., & Baraúna, R. A. (2020). First Steps in the Analysis of Prokaryotic Pan-Genomes. Bioinformatics and Biology Insights. SAGE Publications Inc. https://doi.org/10.1177/1177932220938064

57. Poluektova, E.U., Yunes, R.A., Epiphanova, M.V., Orlova, V.S. and Danilenko, V.N. (2017). The Lactobacillus rhamnosus and Lactobacillus fermentum strains from human biotopes characterized with MLST and toxin-antitoxin gene polymorphism. Archives of Microbiology, 199(5), pp.683–690. doi:10.1007/s00203-017-1346-5.

58. Lee, M. W., Rogers, E. E., & Stenger, D. C. (2012). Xylella fastidiosa plasmid-encoded PemK toxin is an endoribonuclease. Phytopathology. https://doi.org/10.1094/PHYTO-05-11-0150

59. Klimina, K. M., Kjasova, D. K., Poluektova, E. U., Krügel, H., Leuschner, Y., Saluz, H. P., & Danilenko, V. N. (2013). Identification and characterization of toxin-antitoxin systems instrains of lactobacillus rhamnosus isolated from humans. Anaerobe. https://doi.org/10.1016/j.anaerobe.2013.05.007

60. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessières P, Bolotin A, et al. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature. Nov 20;390(6657):249-56. doi: 10.1038/36786. PMID: 9384377.

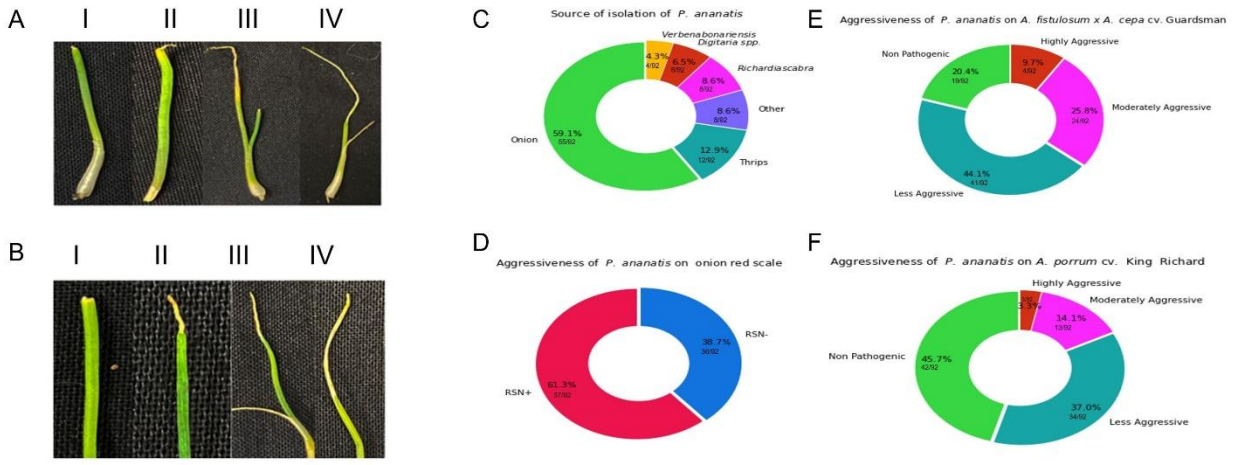


Figure 1.1: Visual representation of spectrum of foliar symptoms caused by *P. ananatis* inoculation on *Allium porrum* (cv. King Richard) and *A. fistulosum* x *A. cepa* (cv. Guardsman) as well as source of isolation and results the red-onion scale necrosis (RSN) assay. 1A shows examples of foliar lesion severity and strain aggressiveness on *A. fistulosum* x *A. cepa.* The panels 1A (i-iv) indicate symptoms associated with increasing level of aggressiveness. 1B shows examples of foliar lesion severity and strain aggressiveness on *A. porrum*. The panels 1B (i-iv) indicate symptoms associated with increasing level of aggressiveness. 1C is the breakdown of the source of isolation of strains, where *A. cepa* (onions) contributes majority of the strains (59.1%) and strains from non-onion sources comprise of 40.9% (*Richardia scabra*: 8.6%, *Digitaria spp*. : 6.5%, *Verbena bonariensis*: 4.3%, Thrips: 12.9%, and other sources: 8.6%). 1D displays the percentage of strains that can cause red-onion scale necrosis (61.3%) and those that could not (38.7%). Panels 1E and 1F visualize the breakdown of strains that are either highly- or moderately- or less-aggressive or non-pathogenic on both hosts.

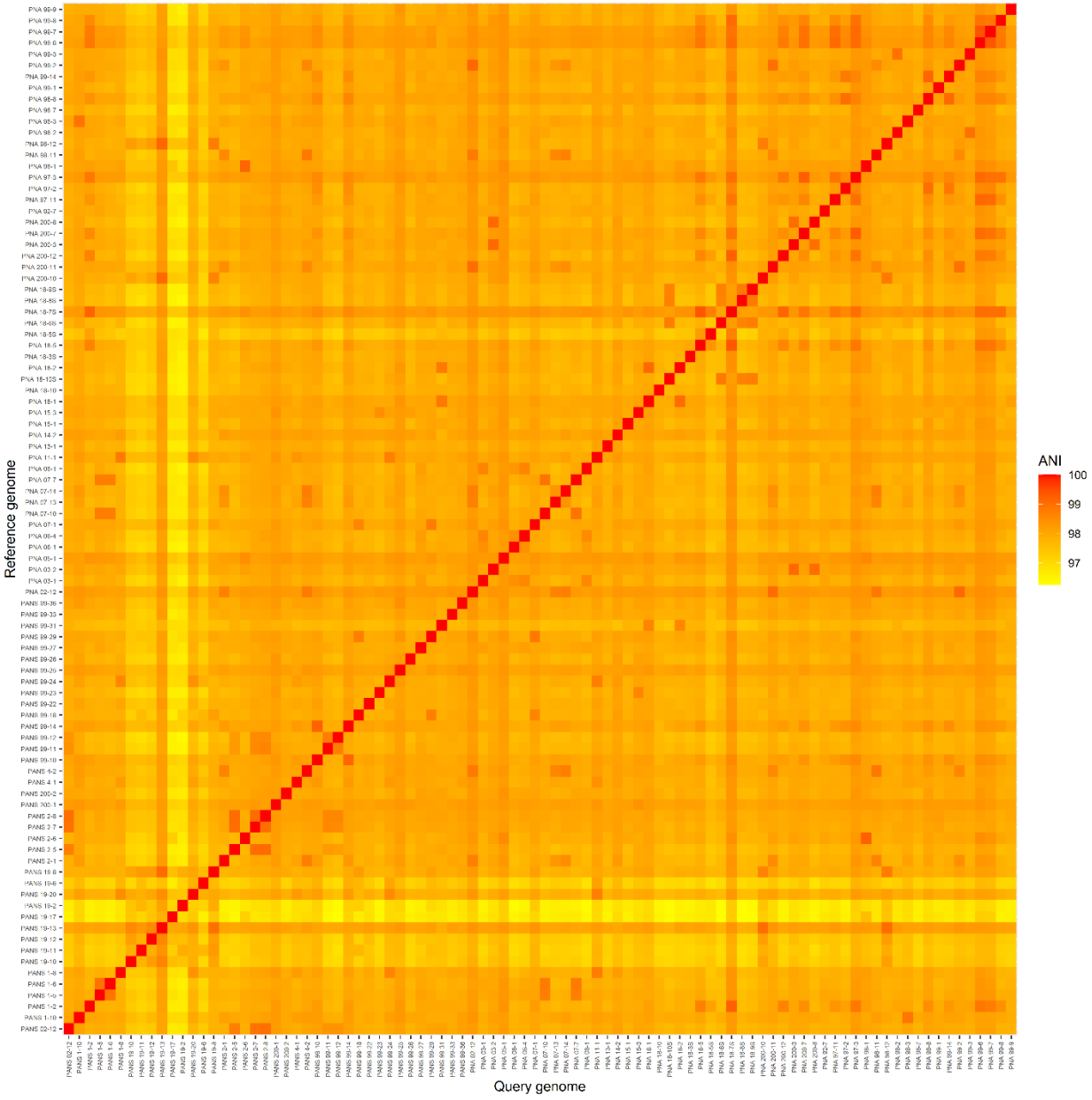


Figure 1.2: A heatmap of average nucleotide identity (ANI) comparisons between every *Pantoea ananatis* strain used in this study.

The smallest ANI comparison was between PNA 18-6S and PANS 19-17 with an ANI score of 96.25.

The highest ANI comparison non-self strain comparison was between PNA 99-6 and PNA 99-7 with a score of 99.41.

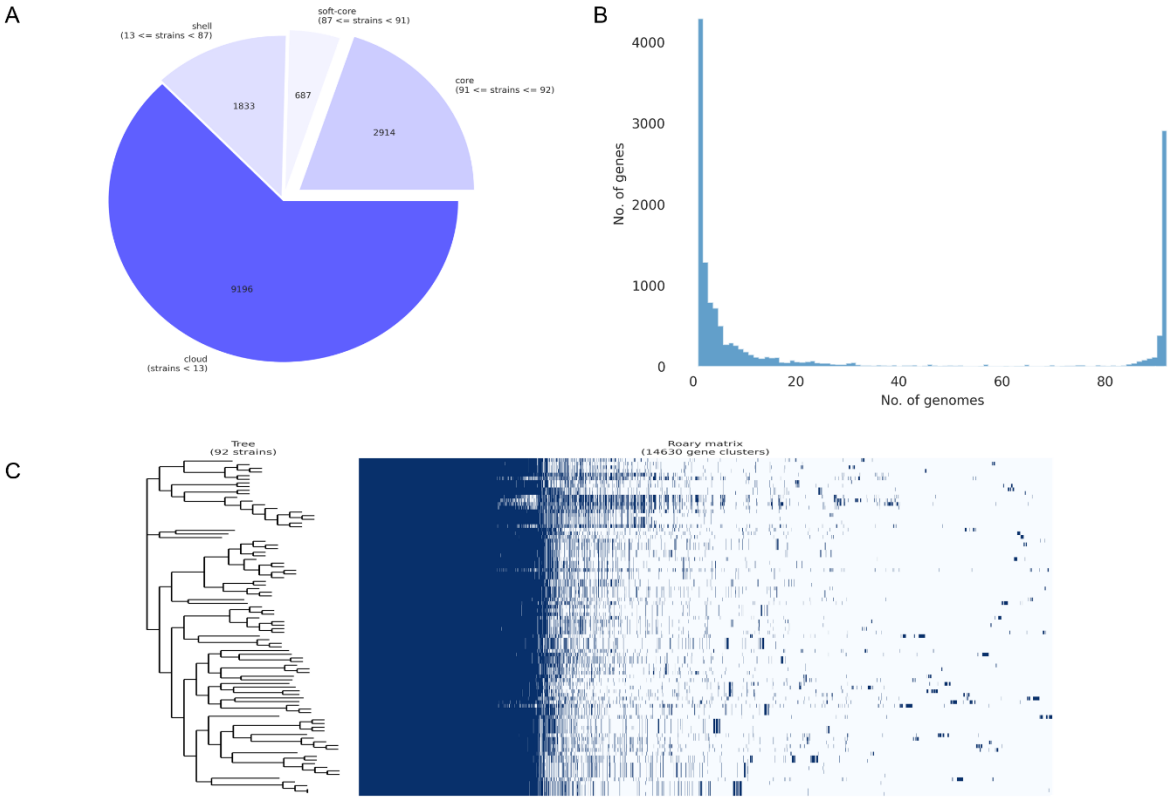


Figure 1.3: Pan-genome analysis of 92 *Pantoea ananatis* genomes.

(A) Pie chart representation of pan-genome composition of *P. ananatis*. The core genome consists of 2914 genes, the soft core 687 genes, the shell 1833 genes, and the cloud 9,196 genes for a total of 14,630 genes; (B) distribution of gene (cluster) sizes as a function of the number of genomes they contain displaying the partition of pan-genomic matrix into shell, cloud, soft-core and core compartments using ROARY outputs; and (C) pan-genome gene presence and absence matrix for 92 *P. ananatis* genomes and associated phylogeny.

Figure 1.4: Phylogenetic tree based on core single nucleotide polymorphism (SNP) variants of genes among the *Pantoea ananatis* strains. Strains color coded in green are non-pathogenic on *Allium porrum* (cv. King Richard) and *A. fistulosum* x *A. cepa* (cv.

Guardsman); strains that are color-coded in purple are pathogenic on both hosts; blue coded strains are pathogenic on *A. porrum* only; and pink coded strains are pathogenic on *A. fistulosum* x *A. cepa* only. Bootstrap values are shown on each branch after 10,000 iterations.

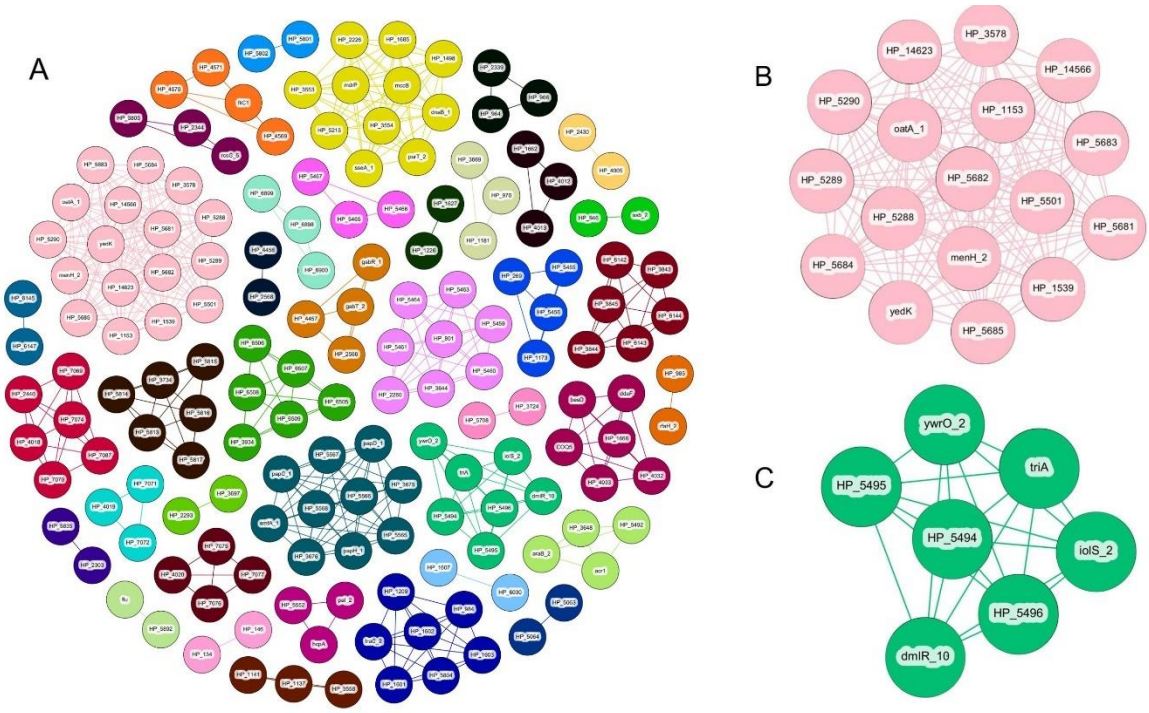


Figure 1.5: Coinfinder derived association of gene pairs from *Pantoea ananatis* genomes (N = 92) displaying various networks (A). Gene-pair association networks for components 31 (B) and 11 (C) are extracted due to their association with foliar pathogenicity for *A. fistulosum* x. *A. cepa*. Gephi was used to apply the Fruchtermann Reingold layout to the network (https://gephi.org/).

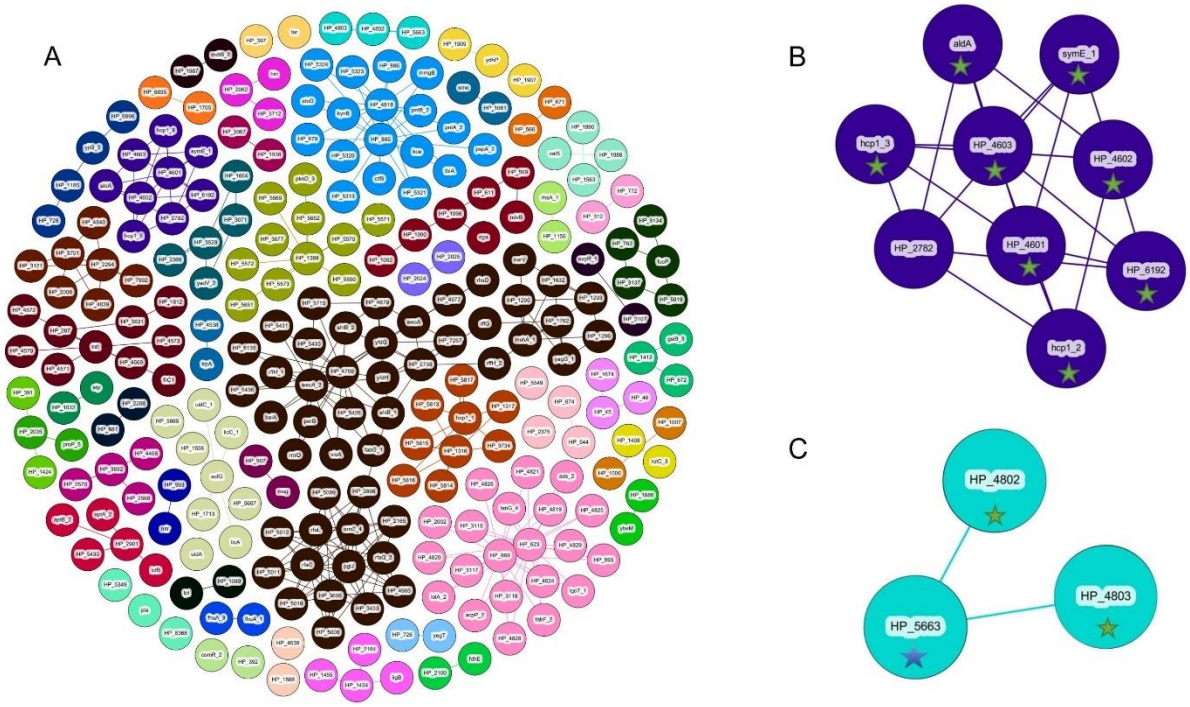


Figure 1.6: Coinfinder derived dissociation of gene pairs from *Pantoea ananatis* genomes (*N* = 92) displaying various networks (A). Gene-pair dissociation networks for components 23 (B) and 37 (C) are extracted due to their with foliar pathogenicity *for A. fistulosum x. A. cepa* (green star) or association with foliar pathogenicity for *A. porrum* (blue star). Gephi was used to apply the Fruchtermann Reingold layout to the network (https://gephi.org/).For all node labels, “group_” was replaced by “HP_” (hypothetical protein) for legibility.

Strain	*hvrA*	*hvrB*	*hvrC*	*hvrD*	*hvrE*	*hvrF*	*hvrG*	*hvrH*	*hvrI*	*hvrJ*	*hvrK*
PNA 97-1											
PANS 99-14	A7V	Q352K						K11R			
PNA 07-10											
PNA 07-1											
PNA 05-1											

Figure 1.7: Graphical representation of the HiVir cluster with single nucleotide polymorphisms (SNPs) determined via direct comparison with the PNA 97-1 "wildtype." Unique missense SNPs are coded with a pink pin. A short table below shows SNP comparisons between the the *A. fistulosum* x *A. cepa* pathogenic strains (PNA 07-10, PNA 07-1 and PNA 05-1) against the *A. porrum* pathogenic strains (PNA 15-3 and PANS 99-14).

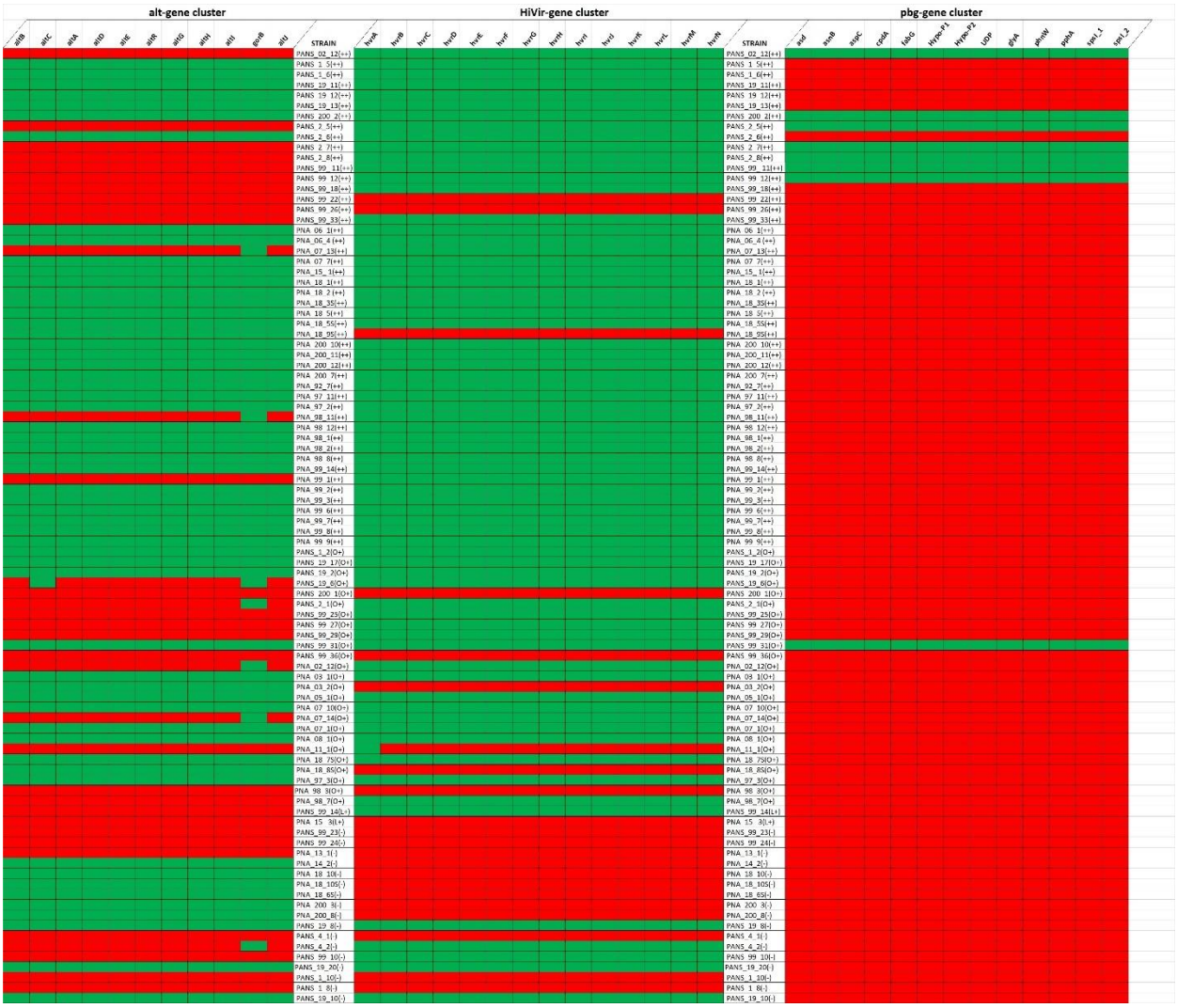


Figure 1.8: Presence and absence of the Alt (left), HiVir (middle), and the pgb cluster (right) genes within the genome of *Pantoea ananatis* based on their foliar pathogenicity on *Allium porrum* (cv. King Richard) and *A. fistulosum* x *A. cepa* (cv. Guardsman). Green and red represent presence and absence of gene, respectively for each gene cluster evaluated. The foliar pathogenicity phenotype and associated aggressiveness is represented as "++" if the strain is pathogenic on both hosts, "L+" if pathogenic only on *A. porrum*, "O+" if pathogenic only on *A. cepa* x *A. fistulosum*, and "-" if pathogenic on neither host.

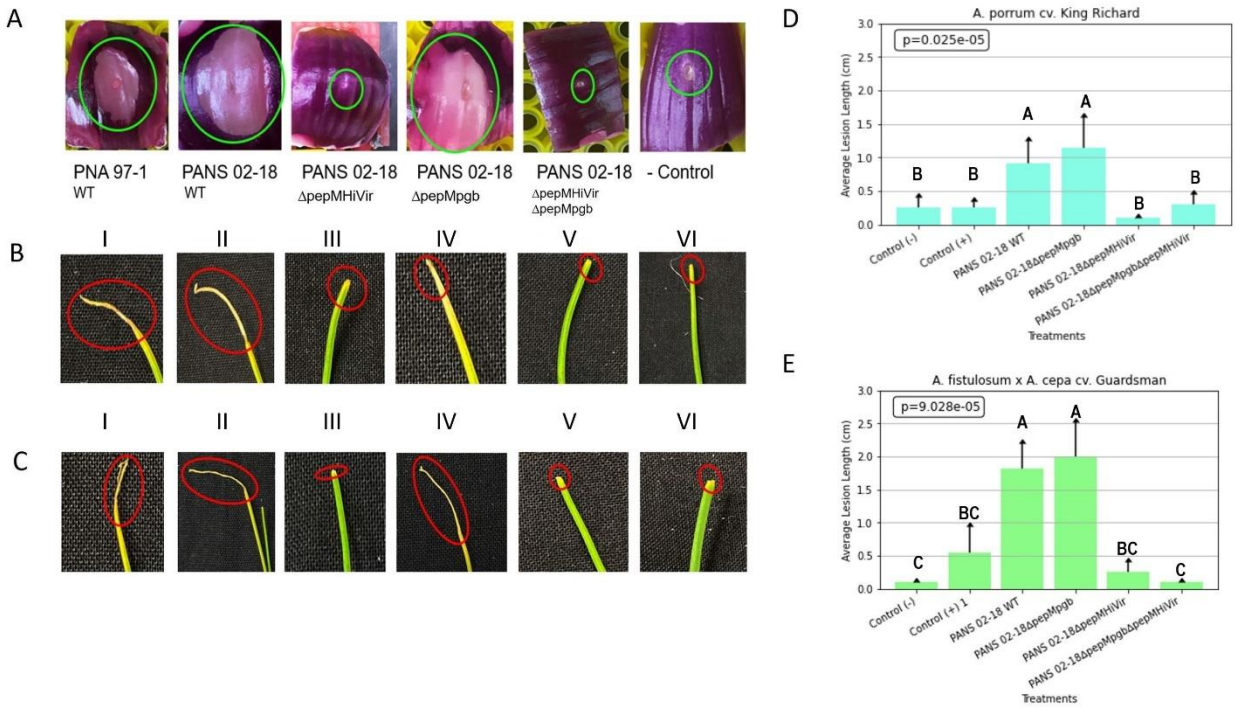


Figure 1.9: Role of HiVir and pgb biosynthetic clusters in foliar pathogenicity on *Allium* species. (A) Red scale necrosis phenotype observed with the wild-type and mutant strains of *P. ananatis.* (B) Foliar pathogenicity assay on leek (*Allium porrum* cv. King Richard) seedlings with the wild-type and mutant strains of *P. ananatis*; and (C) Foliar pathogenicity assay on Japanese bunching onion (*A. fistulosum* x *A. cepa* cv. Guardsman) seedlings with the wild-type and mutant strains of *P. ananatis*. Green and red circles indicate red scale necrosis and foliar necrosis, respectively. Panel D and E display the results of the mean lesion length with standard error in *A. porrum* and *A. fistulosum* x *A. cepa* upon inoculation with the wild-type and mutant strains of *P. ananatis*. The strains utilized in the experiments mentioned above include; positive control strain PNA 97-1, positive control PANS 99-11 (a known aggressive strain on leek), PANS 02-18ΔpepMHiVir,

PANS 02-18ΔpepMpgb, and PANS 02-18ΔpepMΔpepMpgb. Seedlings and red-scale inoculated with sterile water comprised negative controls. Data presented here is the mean of two independent experiments. Letters on the bars indicate mean separation with LSD $P<0.05$.

Strain	Source	Location (county, state)	Leek (Allium porrum cv. King Richard)[a]	Bunching onion (A. fistulosum cv. Guardsman)[f]	Red onion scale assay[c]
PNA_15_3[d]	Onion	Tattnall Co. GA	++*	-	-
PANS_99_14[d]	Digitaria spp.	Tift Co. GA	+	-	-
PANS_99_11[e]	Digitaria spp.	Tift Co. GA	+++	++	+
PANS_99_12[e]	Digitaria spp.	Tift Co. GA	+++*	+	+
PNA_06_4[e]	Onion	Wayne Co. GA	+++	++	+
PNA_97_1[e]	Onion	Tift Co. GA	+++*	++	+
PNA_99_9[e]	Onion seedlings	Tattnall Co. GA	++	+++	+
PNA_99_7[e]	Onion leaf	Tattnall Co. GA	+	+++	+
PNA_99_2[e]	Onion leaf	Tattnall Co. GA	+	+++	+
PNA_99_14[e]	Onion seedlings	Toombs Co. GA	+*	+++	+
PNA_98_1[e]	Onion	Tattnall Co. GA	+	+++	+
PNA_97_11[e]	Onion	Toombs Co. GA	+	+++*	+
PANS_02_7[e]	Thrips from peanut blossoms	Tift Co. GA	+	+++	+
PANS_02_6[e]	Thrips from peanut blossoms	Tift Co. GA	+*	+++	+
PNA_99_3[e]	Onion seedlings	Tift Co. GA	+	+++	+
PNA_07_10[f]	Onion	Toombs Co. GA	-	+	+
PNA_07_1[f]	Onion	Tattnall Co. GA	-	+	+
PNA_05_1[f]	Onion	Vidalia Region, GA	-	+	+
PNA_03_2[f]	Onion	Tift Co. GA	-	+*	-
PNA_03_1[f]	Onion	Tift Co. GA	-	+	+
PNA_02_12[f]	Onion	Tift Co. GA	-	+*	+
PANS_99_36[f]	Richardia scabra L.	Terrell Co. GA	-	+*	-
PANS_99_31[f]	Urochloa texana	Tattnal Co. GA	-*	+*	+
PANS_99_29[f]	Digitaria spp.	Tift Co. GA	-*	+	+
PANS_99_27[f]	Desmodium tortuosum	Vidalia Region, GA	-	+	+
PANS_99_25[f]	Acanthospermum hispidum	Vidalia Region, GA	-	+*	+
PANS_200_1[f]	Slender amaranth	Reidsville, GA	-	+*	-
PNA_18_8S[f]	Onion	Vidalia Region, GA	-	+	-
PNA_18_7S[f]	Onion	Vidalia Region, GA	-	+	+
PNA_97_3[f]	Onion	Toombs Co. GA	-	+	+
PNA_98_7[f]	Onion	Tift Co. GA	-	+	-
PNA_98_3[f]	Onion	Dougherty, GA	-*	+	-
PNA_11_1[f]	Onion	Vidalia Region, GA	-	+*	-
PNA_08_1[f]	Onion	Tattnall Co. GA	-*	++	+
PNA_07_14[f]	Onion	Toombs Co. GA	-	+*	-
PANS_02_1[f]	Adult tobacco thrips from peanut	Tift Co. GA	-	+*	-
PANS_01_2[f]	Thrips from Onion leaf	Tift Co. GA	-	+*	+
PANS_19_2[f]	Digitaria spp.	Tift Co. GA	-	+	+
PANS_19_6[f]	Richardia scabra L.	Tift Co. GA	-	+	+
PANS_19_17[f]	Richardia scabra L.	Tift Co. GA	-	+*	-
PNA_18_2[g]	Onion	Vidalia Region, GA	++	+	+
PNA_15_1[g]	Onion	Tattnall Co. GA	++	+	+
PANS_200_2[g]	Portulaca spp.	Reidsville, GA	++*	+*	+
PANS_01_6[g]	Adult tobacco thrips	Tift Co. GA	++	+	+
PANS_01_5[g]	Adult tobacco thrips	Tift Co. GA	++*	+*	+
PANS_19_12[g]	Verbena bonariensis	Tift Co. GA	++*	+	-
PANS_19_13[g]	Verbena bonariensis	Tift Co. GA	++	+*	-
PANS_02_5[g]	Thrips from peanut blossoms	Tift Co. GA	++	++	+
PNA_97_2	NA	NA	+*	+	-
PNA_99_8	Onion leaf	Wheeler Co. GA	+*	++	+
PNA_99_6	Onion leaf	Toombs Co. GA	+	++	+
PNA_99_1	Onion	MT Vernon, GA	+	++	+
PNA_98_8	Onion	Vidalia Region GA	+	++	+
PNA_98_2	Onion	Tift Co. GA	+*	++	+
PNA_98_12	Onion	Toombs Co. GA	+	++	+
PNA_98_11	Onion	Evans Co. GA	+*	++*	-
PNA_92_7	Onion	Vidalia Region GA	+*	+	+
PNA_200_7	Onion	Tift Co. GA	+*	+	+
PNA_200_12	Onion	Tift Co. GA	+*	+	+
PNA_200_11	Onion	Tift Co. GA	+	++*	+
PNA_200_10	Onion	Tift Co. GA	+*	+	+
PNA_18_9S	Onion	Vidalia Region, GA	+*	+*	+
PNA_18_5S	Onion	Vidalia Region, GA	+	+*	+
PNA_18_5	Onion	Vidalia Region, GA	+*	+	+
PNA_18_3S	Onion	Vidalia Region, GA	+*	+*	+
PNA_18_1	Onion	Vidalia Region, GA	+	+	+
PNA_07_7	Onion	Toombs Co. GA	+*	+	+
PNA_07_13	Onion	Toombs Co. GA	+*	+	-
PANS_99_33	Richardia scabra L.	Coffee Co. GA	+	++*	+
PANS_99_26	Euphorbia hyssopifolia	Vidalia Region, GA	+	+	-
PANS_99_22	Digitaria spp.	Tift Co. GA	+	+	-
PANS_02_8	Thrips from peanut leaf	Tift Co. GA	+*	++	+
PANS_99_18	Richardia scabra L.	Tift Co. GA	+*	+	+
PANS_02_12	Peanut Leaf	Tift Co. GA	+	+	-
PANS_19_11	Richardia scabra L.	Tift Co. GA	+	+	+
PNA_06_1	Onion	Vidalia Region, GA	+	+	-
PANS_04_1	Thrips	Tift Co. GA	-	-*	-
PANS_99_24	Onion Seedlings	Vidalia region	-	-*	-
PANS_19_8[h]	Richardia scabra L.	Tift Co. GA	-	-	+
PANS_19_10[h]	Richardia scabra L.	Tift Co. GA	-	-	+
PNA_200_8[h]	Onion	Tift Co. GA	-*	-	-
PNA_200_3[h]	Onion	Tift Co. GA	-	-	-
PNA_18_6S[h]	Onion	Vidalia Region, GA	-	-	-
PNA_18_10S[h]	Onion	Vidalia Region, GA	-*	-	-
PNA_18_10[h]	Onion	Vidalia Region, GA	-*	-	-
PNA_14_2[h]	Onion	Lyons, GA	-	-*	-
PNA_13_1[h]	Onion	Lyons, GA	-	-*	-
PANS_99_23[h]	Cyperus esculentus	Vidalia Region, GA	-	-	-
PANS_04_2[h]	Adult tobacco thrips from peanut	Tift Co. GA	-	-	-
PANS_01_8[h]	Adult tobacco thrips	Tift Co. GA	-	-*	-
PANS_01_10[h]	Thrips feces from peanut leaf	Tift Co. GA	-	-*	-
PANS_99_10[h]	Verbena bonariensis	Tift Co. GA	-	-*	-
PANS_19_20[h]	Verbena bonariensis	Tift Co. GA	-	-	-

Table 1. Pantoea ananatis strains, their source of isolation, and their associated pathogenicity and aggressiveness on leek (Allium porrum) and Japanese bunching onion (Allium fistulosum x Allium cepa).

Table 1.1: *Pantoea ananatis* strains, their source of isolation, and their associated

pathogenicity and aggressiveness on leek (*Allium* porrum) and Japanese bunching onion (*Allium fistulosum x Allium cepa*).

a) Foliar lesion rating of *P. ananatis* strains on Leek (A. porrum cv. King Richard). Strains with a lesion length 0.2-0.5 cm, 0.5-0.96 cm and >1 cm were considered as less aggressive (+), moderately aggressive (++), and highly aggressive (+++), respectively.

b) Foliar lesion rating of *P. ananatis* strains on bunching onion (*A. fistulosum* x *A. cepa* cv. *Guardsman*). Strains with lesion lengths of <0.7 cm, 0.7-1.4 cm and >1.4 cm were regarded as less aggressive (+), moderately aggressive (++), and highly aggressive (+++), respectively.

c) Ability of strain to clear red anthocyanin pigment and cause pitting on onion scales.

d) Strains that are highly aggressive on leeks but non-pathogenic on bunching onion.

e) Strains that are highly aggressive on leeks and bunching onions and are able to cause necrosis on red-onion scale.

f) Strains that are non-pathogenic on leeks and less-aggressive on bunching onion.

g) Strains that are moderately aggressive on leeks, and are less-aggressive on bunching onions.

h) Non-pathogenic strains.

* Lesion phenotype was inconsistent among the six replicates.

TABLE 2 List of gene-pair association components that contain genes shared with the predicted genes from the genome wise association studies (GWAS) results.

Gene-pair association and dissociation analysis				
Associated gene-pairs				
Component	GWAS correlation	Gene	Molecular function	Biological function
1	*A. fistulosum* × *A. cepa*	*group_3805*	–	–
		rcsC_5	ATP binding	Two-component regulatory system
		group_2344	–	–
3	*A. fistulosum* × *A. cepa*	*group_964*	–	–
		group_966	–	–
		group_2339	–	–
5	*A. fistulosum* × *A. cepa*	*rfaH_2*	DNA binding	Transcription/transcription antitermination
		group_985	–	–
11	*A. fistulosum* × *A. cepa*	*group_5494*	–	–
		group_5495	–	–
		dmlR_10	DNA-binding transcription factor	Regulation of transcription
		ywrO_2	NADPH dehydrogenase	Positive regulation of ion transport
		group_5496	–	–
		iolS_2	Oxidoreductase	Aldo-keto reduction
		triA	Hydrolase	Deamination
15	*A. fistulosum* × *A. cepa*	*group_1662*	–	–
		group_4013	–	–
		group_4012	–	–
16	*A. fistulosum* × *A. cepa*	acr1	Oxidoreductase	Lipid metabolic process
		araB_2	Ribulokinase	L-arabinose catalytic process
		group_5492	–	–
		group_3648	–	–
22	*A. fistulosum* × *A. cepa*	*group_3724*	–	–
		group_5708	–	–
24	*A. porrum*	*group_2568*	–	–
		group_4458	–	–
31	*A. fistulosum* × *A. cepa*	*oatA_1*	Acyltransferase	Lipopolysaccharide biosynthesis
		group_3578	–	–
		group_1153	–	–
		group_14566	–	–
		group_5501	–	–
		group_5290	–	–
		yedK	Peptidase/single-strand DNA binding	SOS response
		group_5681	–	–
		group_5289	–	–
		group_14623	–	–
		group_5685	–	–
		group_5684	–	–
		group_5288	–	–
		menH_2	Hydrolase activity	Menaquinone biosynthetic process
		group_5682	–	–
		group_5683	–	–
		group_1539	–	–
32	*A. fistulosum* × *A. cepa*	*flu*	Binding	Cell adhesion
		group_5892	–	–

Table 1.2: List of gene-pair association components that contain genes shared with the predicted genes from the genome wide association studies (GWAS) results.

TABLE 3 List of gene-pair disassociation components that contain genes shared with the predicted genes from the genome wise association studies (GWAS) results.

Gene-pair association and dissociation analysis				
Dissociated gene-pairs				
Component	GWAS correlation	Gene/group	Molecular function	Biological function
10	None	group_4840, group_3121, group_3701, group_2294	–	–
	A. cepa × *A. fistulosum*	group_7992, group_2008, group_4839		
17	None	group_611, group_1090, group_1096, group_1092	–	–
		cga	Carbohydrate binding	Carbohydrate metabolism
	A. cepa × *A. fistulosum*	group_569	–	–
		ndvB	Carbohydrate binding	Carbohydrate metabolism
21	*A. porrum*	group_2568, group_4458	–	–
	None	group_3692, group_2570		
23	None	group_2782	–	–
	A. cepa × *A. fistulosum*	group_4603, group_4602, group_4601, group_6192		
		hcp1_2	Family type IV secretion system effector	Toxin
		symE_1	DNA binding/RNA binding	RNA degradation
		hcp1_3	Family type IV secretion system effector	Toxin
		aldA	Aldehyde dehydrogenase	Varied catabolic processes
35	None	group_1295, group_1632, group_1293, group_1296, group_1762, group_7257, group_4977, group_5436, group_5426, group_2896, group_5435, group_5431, group_6135, group_2165, group_5008, group_3433, group_5012, group_5010, group_5009, group_5011, group_3715 ,group_4679, group_3606	–	–
		yagG_1	Symporter activity	Carbohydrate/sodium transport
		mshA_1	D-Inositol-3-Phosphate Glycosyltransferase activity	Mycothiol biosynthetic process
		rffG	dTDP-glucose 4,6-dehydratase Activity	Varied Biosynthetic Processes
		rffH_2	Glucose-1-phosphate Thymidylyltransferase activity	Extracellular polysaccharide biosynthetic process
		narV	Nitrate reductase activity	Aerobic respiration/nitrate assimilation
		rmlD	dTDP-4-dehydrorhamnose reductase activity	dTDP-rhamnose biosynthetic process/polysaccharide biosynthesis/O-Antigen biosynthesis
		perB	DNA binding	Regulator
		baiA	NAD+ binding	Protein homotetramerization
		vioA	dTDP-4-amino-4,6-dideoxy-D-glucose transaminase activity	Lipopolysaccharide biosynthetic process
		fabG_1	3-Oxoacyl-[acyl-carrier-protein] reductase (NADH) Activity	Fatty acid elongation

Table 1.3 Part 1: List of gene-pair dissociation components that contain genes shared with the predicted genes from the genome wide association studies (GWAS) results (part 1).

Gene-pair association and dissociation analysis				
Dissociated gene-pairs				
Component	GWAS correlation	Gene/group	Molecular function	Biological function
		rffH_1	Glucose-1-phosphate Thymidylyltransferase Activity	Extracellular polysaccharide biosynthetic process
		arnC_4	Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose Transferase Activity	Polysaccharide biosynthetic process
		rfaQ_2	Glycosyltransferase activity	–
		rfaC	Lipopolysaccharide heptosyltransferase activity	Lipopolysaccharide core region biosynthetic process
		rfaL	Ligase activity	Lipopolysaccharide core region biosynthetic process
		pglJ	Hexosyltransferase activity	Protein N-linked glycosylation via asparagine
		yfdH	Glycosyltransferase activity	–
		shlB_2	–	Protein transport
		rfaQ_2	Glycosyltransferase activity	–
		rfaC	Lipopolysaccharide heptosyltransferase activity	Lipopolysaccharide core region biosynthetic process
		rfaL	Ligase activity	Lipopolysaccharide core region biosynthetic process
		pglJ	Hexosyltransferase activity	Protein N-linked Glycosylation via Asparagine
		yfdH	Glycosyltransferase	–
		wecA	Glycosyltransferase activity	O Antigen biosynthetic process
	A. fistulosum × A. cepa	group_4905, group_2739, group_4759	–	
		shlB_1	Putative exported adhesin activator	Protein transport
		rhsD	–	Cellular response to sulfur starvation
		wecA_2	Glycosyltransferase activity	O Antigen biosynthetic process
37	*A. porrum*	group_5663	–	–
	A. fistulosum × A. cepa	group_4803, group_4802		

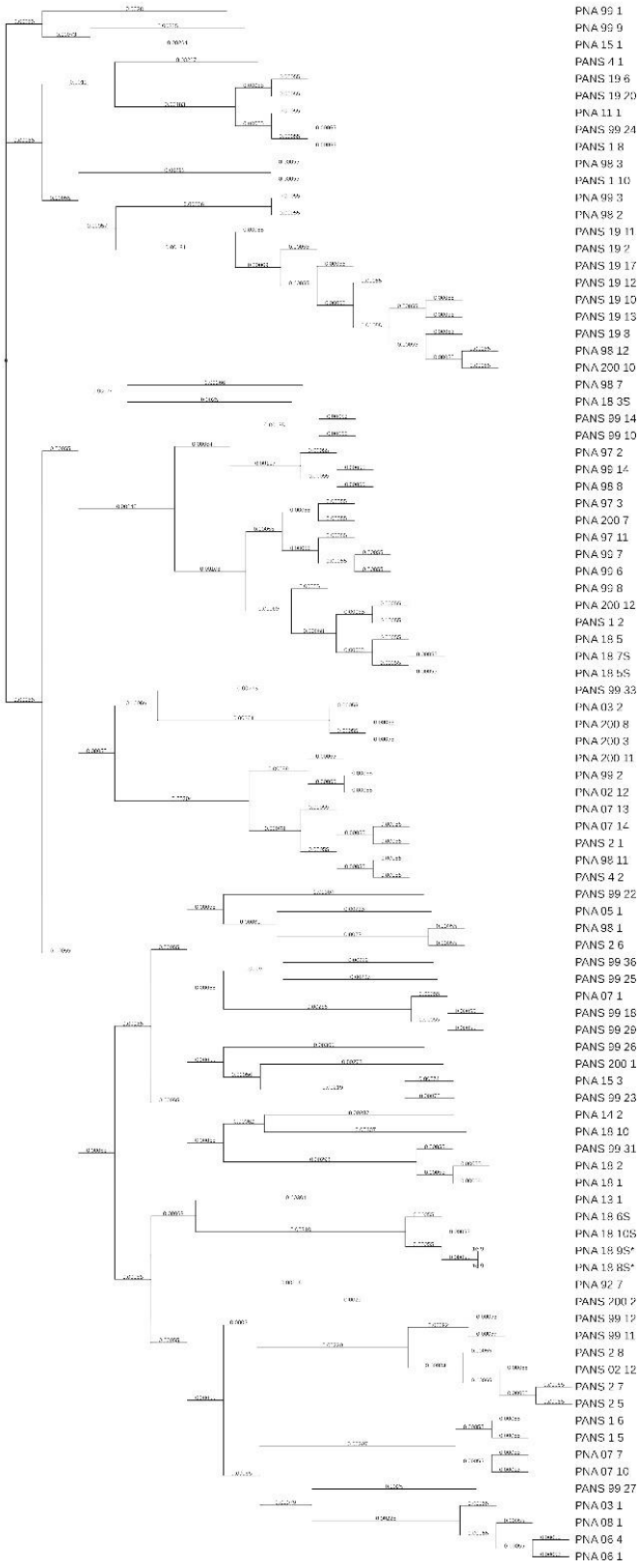
Table 1.3 Part 2 (continued): List of gene-pair dissociation components that contain genes shared with the predicted genes from the genome wide association studies (GWAS) results (part 2).

TABLE 4 Composition and annotation of the HiVir and the pgb gene clusters in *Pantoea ananatis*.

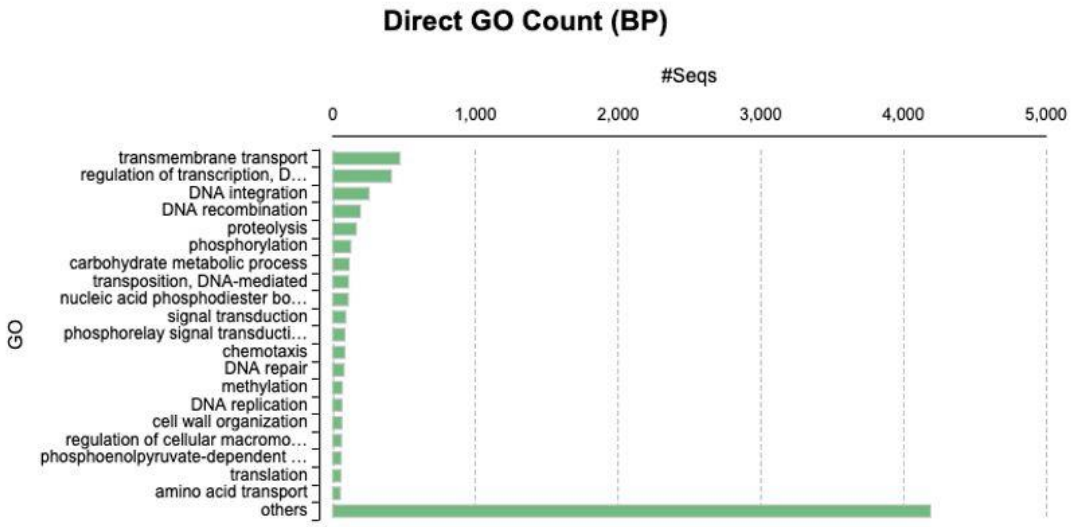
HiVir gene cluster	Annotations	pgb gene cluster	Annotations2
hvrA	**Phosphoenolpyruvate phosphomutase**	pepM	**Phosphoenolpyruvate phosphomutase**
hvrB	FMN-dependent oxidoreductase (nitrilotriacetate monooxygenase family)	cpdA	3',5'-cyclic adenosine monophosphate phosphodiesterase CpdA
	Monooxygenase	fabG	3-oxoacyl-[acyl-carrier-protein] reductase FabG
hvrC	Homocitrate synthase NifV	aepY	phosphonopyruvate decarboxylase
	2-Isopropylmalate synthase	phnW	2-aminoethylphosphonate--pyruvate transaminase
hvrD	3-Isopropylmalate/(R)-2-methylmalate dehydrataselarge subunit	asnB	Putative asparagine synthetase [glutamine-hydrolyzing]
hvrE	3-Isopropylmalate dehydratase small subunit	spsI1	Bifunctional IPC transferase and DIPP synthase
hvrF	Methyltransferase	asd	Aspartate-semialdehyde dehydrogenase
hvrG	Acetyltransferase (GNAT) family protein		**MFS 1 transporter YcxA**
hvrH	ATP-grasp domain containing protein	glyA	Serine hydroxymethyltransferase
hvrI	**MFS transporter**		CDP-alcohol phosphatidyltransferase
	Macrolide efflux protein A	spsI2	glucose-1-phosphate adenylyl/thymidylyltransferase
hvrJ	Hypothetical protein		Bifunctional IPC transferase and DIPP synthase
hvrK	Flavin reductase	aspC	Aspartate aminotransferase
			UDP-2,3-diacylglucosamine diphosphatase

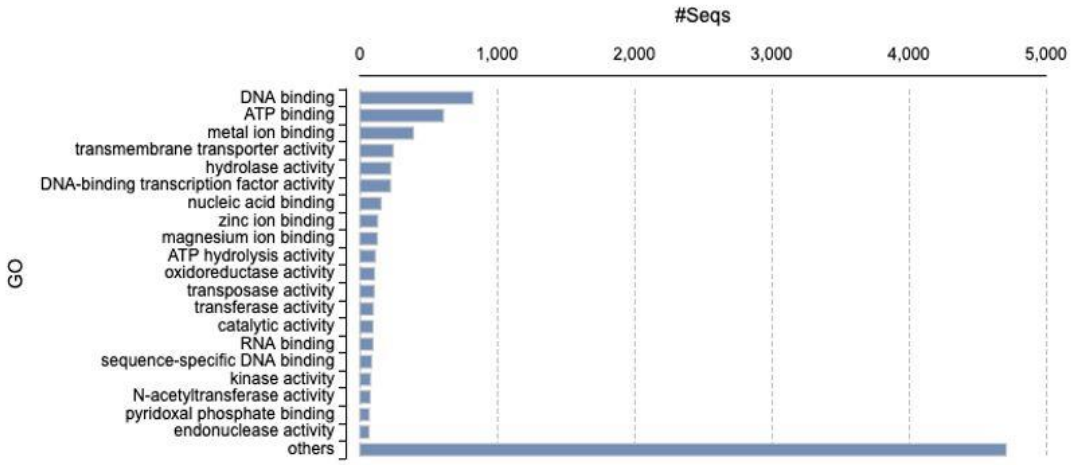
Table 1.4: Composition and annotation of the HiVir and the pgb gene clusters in *Pantoea ananatis*.

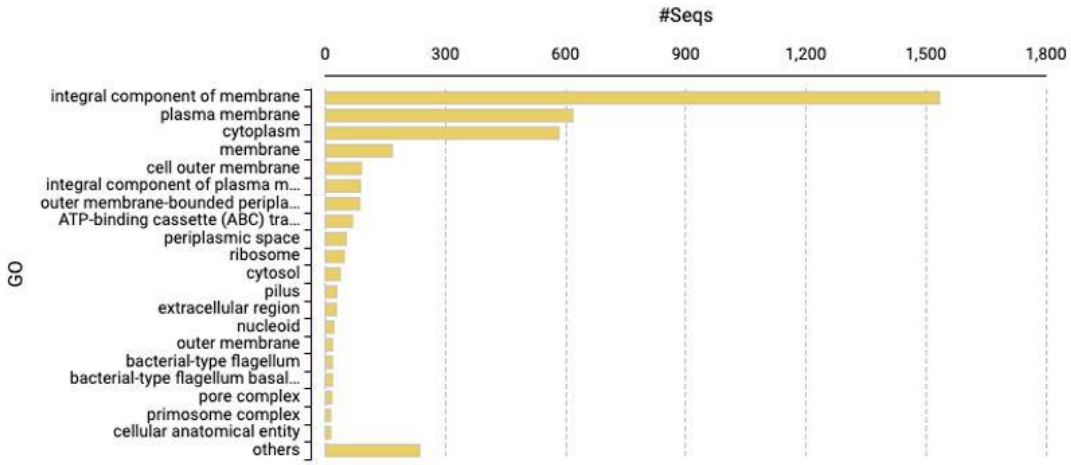
Tree scale: 0.001



Supplementary Figure 1.1: Phylogenetic tree of gene-presence and absence across *Pantoea ananatis* genomes (n=92) derived from FastTree analysis. Strains with an asterisk (*) produced branch lengths of 0.


Direct GO Count (BP)
#Seqs
0
1,000
2,000
3,000
4,000
5,000
GO
transmembrane transport
regulation of transcription, D...
DNA integration
DNA recombination
proteolysis
phosphorylation
carbohydrate metabolic process
transposition, DNA-mediated
nucleic acid phosphodiester bo...
signal transduction
phosphorelay signal transducti...
chemotaxis
DNA repair
methylation
DNA replication
cell wall organization
regulation of cellular macromo...
phosphoenolpyruvate-dependent ...
translation
amino acid transport
others


Direct GO Count (MF)
#Seqs
0
1,000
2,000
3,000
4,000
5,000
GO
DNA binding
ATP binding
metal ion binding
transmembrane transporter activity
hydrolase activity
DNA-binding transcription factor activity
nucleic acid binding
zinc ion binding
magnesium ion binding
ATP hydrolysis activity
oxidoreductase activity
transposase activity
transferase activity
catalytic activity
RNA binding
sequence-specific DNA binding
kinase activity
N-acetyltransferase activity
pyridoxal phosphate binding
endonuclease activity
others


Direct GO Count (CC)
#Seqs
0
300
600
900
1,200
1,500
1,800
GO
integral component of membrane
plasma membrane
cytoplasm
membrane
cell outer membrane
integral component of plasma m...
outer membrane-bounded peripla...
ATP-binding cassette (ABC) tra...
periplasmic space
ribosome
cytosol
pilus
extracellular region
nucleoid
outer membrane
bacterial-type flagellum
bacterial-type flagellum basal...
pore complex
primosome complex
cellular anatomical entity
others

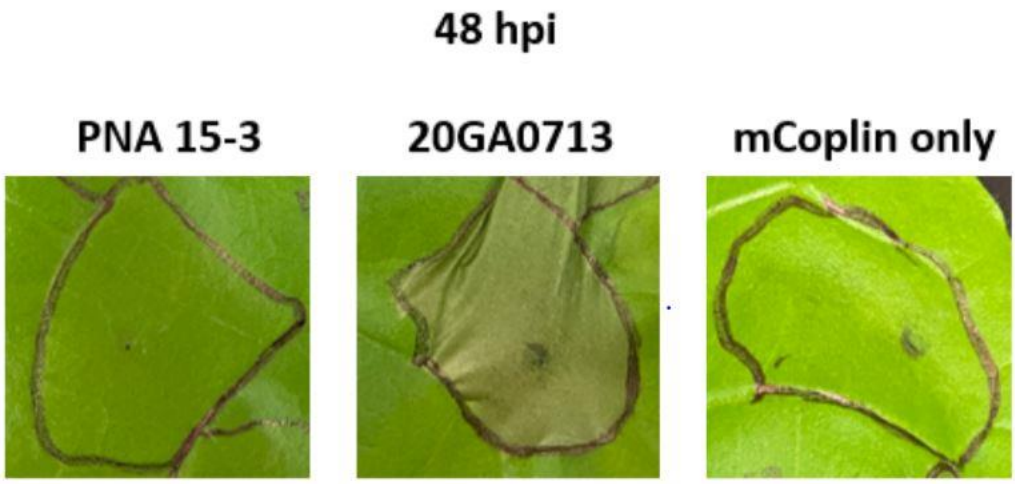
Supplementary Figure 1.2: Distribution of gene ontology (GO) terms: annotations of core, soft-core, shell and cloud genes of *Pantoea ananatis* pan-genome.

Section A shows the distribution of GO terms related to biological processes (BP).

Section B shows the distribution of GO terms related to molecular function (MF).

Section C shows the distribution of GO terms related to cellular component (CC).

Supplementary Figure 1.3. Additional break down of GO distribution per term category of biological process (BP), molecular function (MF), and cellular components (CC). Results shown here are limited to top 20 categories.



Supplementary Figure 1.4. Results of tobacco infiltration assay after 48 hours with PNA 15-3 (left), positive control 20GA0713 (center), and mCoplin negative control (right). Hypersensitive response is marked with black marker.

PNA 15-3 putatative T3SS does not induce a hypersensitive response, whereas 20GA713 does.

Chapter 3

NLP-like Deep Learning Aided in Identification and Validation of Thiosulfinate Tolerance Clusters in Diverse Bacteria[1]

---

[1] Myers, B. K., Lamichhane, A., Kvitko, B. H., & Dutta, B. (2024). NLP-like deep learning aided in identification and validation of thiosulfinate tolerance clusters in diverse bacteria. Submitted to *mBio*, 9/24/2024.

Abstract

Allicin tolerance (*alt*) clusters in phytopathogenic bacteria, which provide resistance to thiosulfinates like allicin, are challenging to find using conventional approaches due to their varied architecture and the paradox of being vertically maintained within genera despite likely being horizontally transferred. This results in significant sequential diversity that further complicates their identification. Natural language processing (NLP) - like techniques, such as those used in DeepBGC, offers a promising solution by treating gene clusters like a language, allowing for identifying and collecting gene clusters based on patterns and relationships within the sequences. We curated and validated *alt*-like clusters in *Pantoea ananatis* 97-1R (PA), *Burkholderia gladioli* pv. *gladioli* FDAARGOS 389 (BG), and *Pseudomonas syringae* pv. tomato DC3000 (PTO). Leveraging sequences from the RefSeq bacterial database, we conducted comparative analyses of gene synteny, gene/protein sequences, protein structures, and predicted protein interactions. This approach enabled the discovery of several novel *alt*-like clusters previously undetectable by other methods, which were further validated experimentally. Our work highlights the effectiveness of NLP-like techniques for identifying underrepresented gene clusters and expands our understanding of the diversity and utility of *alt*-like clusters in diverse bacterial genera. This work demonstrates the potential of these techniques to simplify the identification process and enhance the applicability of biological data in real-world scenarios.

Introduction

Plants deploy an impressive array of small molecules to defend themselves against herbivory and pathogen-mediated infection. Thiosulfinates such as allicin are charismatic small molecules known for their role as antifeedants and antimicrobials in *Allium* species (1, 2). These small molecules are reactive organosulfur compounds responsible for several *Allium* species' characteristically pungent flavor and smell (3, 4). Allicin is produced when the enzyme alliinase acts on alliin, transforming it into a thiol-reactive compound. It interacts with cellular thiols, leading to allyl-mercapto modifications in proteins that deactivate enzymes and cause protein aggregation (5). Additionally, allicin reacts with reduced glutathione, converting it to S-allylmercaptoglutathione and thus depleting the cellular glutathione pool (5, 6). Thiosulfinates have been demonstrated to be inhibitory to a wide range of microorganisms both *in vitro* and *in vivo* (7, 8, 9, 10). Recently, gene clusters associated with allicin tolerance were identified not only in the onion pathogens *Pantoea ananatis* (PA) and *Burkholderia gladioli* (BG) but also in the garlic saprophyte *Pseudomonas fluorescens* (10, 11, 12). These genes were named allicin tolerance (*alt*) genes and are enriched for genes involved in thiol redox reactions. The *alt* clusters increased onion virulence capacity in PA and BG strains and conferred increased allicin tolerance to *E. coli* (12, 13). The *alt* gene cohort appears to function additively for managing cellular thiol stresses, with multiple genes conferring partial tolerance phenotypes (10). In their 2018 study, Stice et al. (10) data mined the NCBI GenBank database to identify *Pantoea* spp. with *alt* clusters, using the *altG* gene as an indicator. In doing so, the authors observed that several strains isolated from *Allium* hosts and some *Brassica* species carry *alt* clusters. In contrast, strains isolated from non-

thiosulfinate-producing hosts did not. Inspired by an intuitive understanding of the characteristics defining an *alt* cluster, manual curation led to discovering a unique cluster within BG and other *Burkholderia* spp. (12), supported by multigene BlastX analysis (12). Considering the importance of *alt* clusters in thiosulfinate tolerance and plant-microbe interactions, identifying the variety of *alt* clusters and their presence in bacterial species is crucial. The *alt* clusters that were characterized and validated share little sequence or gene synteny similarity. Typical gene-mining techniques, such as NCBI BLAST or multigene BLAST, do not translate well between *alt* clusters localized within different bacterial genera. Although the *alt* cluster is potentially horizontally transferred as it is localized on plasmids, it seems to be maintained vertically within individual bacterial genera. This makes identifying *alt* clusters within genera comparatively easier; however, their identification among distinct genera is quite challenging. Isolating thiosulfinate-tolerant bacteria from a thiosulfinate-producing host and then manually curating the annotations list for a conspicuous gene cluster has been the *modus operandi* for *alt* gene cluster discovery to date. However, it is a time-consuming process that requires in-depth training and a reliable annotation pipeline. Even in optimal conditions, individual researchers might develop personal biases towards which annotations they deem more reliable or questionable, potentially resulting in misidentification of *alt* clusters. To formalize an *alt*-identification and recovery method independent of the issues caused by gene sequence and gene synteny, we used NLP-like techniques for mining putative *alt*-like gene clusters.

The methodology employed here is similar to those used with genome mining for secondary metabolite biosynthetic gene clusters. These pipelines must overcome a

challenging task that requires careful consideration of gene content. For example, bacteria tend to organize genes into localized clusters to make metabolite synthesis more efficient (14,15,16). While manual curation and BLAST are effective for similar biosynthetic gene clusters (BGCs) in closely related organisms, they fall short when sequence data alone is inadequate or manual efforts are impractical due to time or cost (17). In such cases, more rigid, 'hard-coded' algorithms are used, though they require predefined gene and protein data rules, limiting their use with less-defined gene clusters (18, 19). Machine learning is the natural next step in algorithmic complexity to solve these problems, autonomously allowing for a more generalizable "learning" of input content. This allows for discovering more novel BGC's as the algorithm generates its own rules during training for further downstream applications. An example of this is ClusterFinder (20). ClusterFinder utilizes a Hidden Markov Model (HMM) approach rather than sequence alignment, allowing for greater freedom of discovery. However, HMM does not preserve position dependency effects or any potential higher-order information that may be relevant for BGC discovery (21, 22, 23).

To address the need for higher-order information in BGC discovery, a deep learning approach using Recurrent Neural Networks (RNNs) with the addition of vector representations of protein family tags (Pfam) was designed, which improved the capacity for algorithmically detecting novel BGCs (24). DeepBGC utilizes an NLP strategy for identifying and even extracting novel BGCs from bacterial genomes via a clever use of a Bidirectional Long Short-Term Memory (BiLSTM) RNN (25,26) and a word2vec-like word embedding skip-gram neural network that the authors named pfam2vec (24).

In this work, we trained DeepBGC on our small collection of validated *alt* clusters to determine the potential for more complex artificial intelligence methods to accelerate the discovery process. Although the *alt* cluster does not represent a typical BGC where each gene collaboratively synthesizes a molecule, the organization and perceived additive function of these genes for the *alt* phenotype renders the cluster amenable to methodologies like those used in BGC discovery. The new *alt* model was then utilized to data mine the entire RefSeq bacterial database for potential *alt*-like clusters. Representative clusters were selected and refined through manual curation and sequencing data analysis to produce representative sequences of *alt*-like gene clusters. The genes, proteins, and predicted binding potential for selected genes from each cluster were compared to identify potentially valuable methodologies for differentiating *alt*-like gene clusters. Finally, chosen *alt*-like gene clusters were validated by expression of synthesized gene pairs in *alt*-gene cluster defective strain of PA (PNA 97-1 *Δalt*) and screened for increased thiosulfinate tolerance based on the improved ability of strains to grow in thiosulfate-rich onion extract.

Results

alt Seed Cluster Gene and Protein Sequence Comparisons Show Low Sequence Similarity

Onion-associated bacteria like PA and BG possess *alt*-clusters that impart the ability to survive and propagate in thiosulfinate-rich-environments (Figure 2.1 A). In some cases, this may lead to bulb rot symptoms (Figure 2.1 A). In the genomic comparison of PA, PTO, and BG, the total gene counts are 11, 16, and 7, respectively. Shared genes across these strains include *altA*, *altB*, *altC*, *altE*, *altR*, *altJ*, and *altI*. When evaluating synteny,

among PA, PTO, and BG there appears to be little in common between the three sequences. Between PA and BG, *altA* and *altC* do localize; however, their order is inverted. Further, the *altR* and *altE* are adjacent between both PA and BG. The *altJ* and *altB* are adjacent but inverted between BG and PTO. The *altE* and *altA* are also adjacent but inverted between BG and PTO (Figure 2.1 B).

We observed high degrees of dissimilarity when comparing the total gene cluster sequence similarity among our original three validated *alt* clusters. Additionally, when analyzing individual genes with annotations shared across all three clusters, the similarity percentages exhibit a range between 21.9 and 74.1%. Specifically, *altI* sequences show similarities from 39.1 to 52.1%, *altA* from 66 to 69.9%, and *altC* from 47.3 to 50.6%. Sequences of *altE* vary from 62 to 69.4%, *altR* from 46.5 to 51.2%, and *altJ* from 41.1 to 70.5%. The *altB* gene maintains high consistency around 74% across all comparisons. A second *altR* gene in the PTO cluster displays 47.5 to 54.5% similarity. For genes only shared between PA and PTO, the lowest similarity is noted in *altJ* at 21.9%, with other genes like *altD*, *altH*, and gor displaying up to 52.4% similarity (Figure 2.1, Table 2.1).

A broad range of dissimilarities is observed in assessing protein sequence similarity across the three validated alt clusters, with percentages ranging from 18.1% to 82.1%. Notably, *altI* shows significant variation, with 48.2% similarity between PA and BG, dropping to 18.1% when comparing BG vs. PTO. The protein sequences in *altA* range from 67.9 to 74.3% across comparisons, while *altC* varies from 38.5 to 43.8%. The *altE* sequences are relatively similar, ranging from 62.5 to 71.8%. The *altR* protein sequences vary from 35.4 to 43.9%, and *altJ* from 27.2 to 76.6%. The *altB* exhibits high consistency, with similarities ranging from 78.5 to 82.1%. A second *altR* in the PTO cluster shows

similarities between 36.5% to 47.3%. Similarities for proteins exclusively shared between PA and PTO are notably lower, with *altJ* at 25.5%, *altD* at 30.5%, *altH* at 45.5%, and *gor* at 43.9% (table 2.2).

DeepBGC Data Mining of the NCBI RefSeq Database and Filtering for Autonomous Collection of *alt*-like Gene Clusters

DeepBGC training on the three validated *alt* cluster sequences was repeated 15 times, and the reports were compared to assess model performance. The average loss across runs was minimal at 0.00, indicating steady performance. However, a maximum loss value of 0.40 suggests some performance variability. In assessing model performance for DeepBGC, accuracy was consistent across all tests, averaging 1.00 with a standard deviation of 0.00 and a minimum accuracy of 0.98, highlighting the model's reliability. Precision and recall were both low, averaging 0.01, indicating a challenge in accurately identifying and capturing true positives from the dataset. This variability was reflected in the AUC-ROC scores, which averaged 0.82, suggesting good discriminatory ability with room for improvement. Statistical analysis confirmed significant variability in precision (F-value: 3.78, p-value: $\sim 2\times10^{-6}$), recall (F-value: 5.17, p-value: $\sim 7.7\times10^{-10}$), and AUC-ROC (F-value: 16.09, p-value: $\sim 7.64\times10^{-43}$). In contrast, differences in loss and accuracy were not statistically significant (F-values: 0.86 and 0.39, p-values: 0.60 and 0.98, respectively), indicating stable performance in these areas. The detailed statistical insights underscore the need for further refinement to enhance precision, recall, and overall model robustness. These results are expected with the small training dataset we

can access and are overcome with manual inspection of DeepBGC extractions for validity.

Upon completion of the DeepBGC-enabled data mining of 238,362 bacterial genomes from RefSeq, we extracted 12,280 gene clusters. These were reduced to 1,777 sequences post MMseqs2 redundancy filtering with an average GC% of 53.5% (median 55.1%, max. 76.6% and min. 25.7%), an average sequence length of 8,800 (max 61,726, min 1,215, and median of 6,424), and finally an average file size of 23KB (max 114KB, min 9KB, and a median of 19KB). After further manual curation to remove all gene sequences that appear split by the end of contigs, only four genes in total length, or do not have at least 3 unique *alt*-like pfam tags, we chose 47 representative *alt*-like sequences. These 47 representative clusters contained an average GC% of 51.7% (max 69.9%, min 32.4%, and median 53.5%), an average sequence length of 7,931 (max 30,170, min 3,109, and median 6,316), and finally an average file size of 28 KB (max 114 KB, min 12 KB, and median 23 KB). When screening for clusters that are representative of our initial three *alt* clusters, the *Pantoea alt* cluster is represented by an *alt*-like cluster from *Duffyella gerundensis* (NZ_LN907829.1) with 94% sequence identity and identical values of assigned Pfam domains, the *Burkholderia alt* cluster is represented by a truncated *alt*-like gene cluster from *Paraburkholderia graminis* (NZ_CP024936.1) with 74% sequence similarity. The *Pseudomonas alt* cluster is represented by itself as PTO (NC_004578.1). For all downstream gene cluster comparisons, we used the PA and BG *alt* clusters for comparison as references (Figure 2.2).

Gene sequence similarity among these 49 clusters is low for alignment-based comparison methods. To minimize this, we color-coded genes based on their known

relevance and converted the color code into strings for Levenshtein comparisons. These gene clusters are separated into 4 distinct groups (Figure 2). The total counts for *alt*-like genes among these representative clusters are as follows, *altR* (N=41), *altC* (N=38), *altJ* (N=36), *altE* (N=36), *altA* (N=33), *altB* (N=28), *altG* (N=11), *altI* (N=8), *altD* (N=8), *PSPTO_4258* (N=7), *altH* (N=6), *PSPTO_4257* (N =6), *gor* (N =2), *kefC* (N=2), *PSPTO_5268* (N=1). Among these, *altR* has the highest count per gene cluster (figures 2.2 and 2.3).

BLAST of DeepBGC-Mined *alt*-like Clusters and NCBI GenBank for Representative Sequence-Species Diversity Shows Wide Diversity of *alt*-like Gene Clusters Among Bacterial Genera

To compare the diversity of bacterial species represented by recovered *alt*-like gene clusters, we employed BLAST to retrieve clusters from both the sequences obtained through DeepBGC-enabled data mining of RefSeq and NCBI GenBank. Due to the varying selection of available sequences between NCBI's RefSeq and GenBank, cross-comparison between the two databases may offer a more comprehensive understanding of species diversity compared to solely re-screening NCBI RefSeq with BLAST. Notably, *Klebsiella pneumoniae* emerged as a predominant species, constituting 56% of the recovered sequences in one instance and demonstrating significant representation across multiple samples. Conversely, specific sequences lacked a single dominant species, particularly those associated with *Stenotrophomonas maltophilia*. Detailed analysis of biodiversity using Shannon-Wiener indices unveiled varying levels of diversity among samples. For instance, sequences attributed to *S. maltophilia* exhibited higher diversity, representing 63-84% of recovered sequences. In contrast, sequences linked to

Pseudomonas aeruginosa displayed lower diversity, comprising 77-82% of sequences. Additionally, GenBank BLAST analysis yielded taxonomic insights into the retrieved sequences. While *Klebsiella pneumoniae* was prevalent, other species, such as *Pseudomonas fluorescens* and *Escherichia coli*, were also prominently featured. Specific genera exhibited species-specific enrichment, with *Pseudomonas* and *Paenibacillus* showing pronounced representation in the sequences. These findings underscore the wide distribution of *alt*-like gene clusters across bacterial species and highlight their potential ecological importance (Figure 2.2). These results are summarized in the supplementary table (Supplementary table 2.1).

3D Superimposition of Predicted Protein Models is Insufficient for Differentiating Between *alt*, and Unrelated Proteins

Due to the complexity inherent in classifying *alt* clusters by sequence and gene synteny, we investigated potential discrepancies in predicted 3D structures. Our analysis began with *altR*, a *tetR*-family regulator within *alt*-like gene clusters, revealing high structural similarities between BG and PA and BG and PTO, with zeal scores of 0.93 and 0.94, respectively. Further examination of secondary *altR* variants from PTO showed similar congruence, with scores ranging from 0.95 to 0.96. Extending our analysis to other genes such as *altA*, *altB*, *altC*, *altE*, and *altI*, we consistently observed high zeal scores (0.91 to 0.97) indicative of substantial structural similarity across different organisms. However, *altI* presented some structural discordance, with lower zeal scores down to 0.67, suggesting potential functional diversity. We expanded our study to include multiple sequence alignments of the five most frequently identified *alt*-like genes post-DeepBGC detection, followed by ITASSER-based 3D structural predictions. These comparisons

involved a broad set of sequences, with resulting zeal scores ranging from 0.40 to 1.00, reflecting a wide diversity in structural similarity among the *altC* variants. Although the structural comparisons generally supported the structural resemblance across these genes, they did not provide a clear distinction between the datamined gene clusters. (Figures 4 and 5; Supplementary figures 1 and 2; Supplementary files 2.1 and 2.2).

Crosstree Comparisons between Protein Sequence Similarity and Gene Synteny Indicate vertical transmission and divergence of *alt* and *alt*-like Genes

To determine if there is any grouping of *alt*-like genes based on protein sequence similarity, we utilized RAxML to generate phylogenetic trees based on sequence similarity. Further, we used phytools to compare trees for pattern similarity. We utilized further R scripting to label the connecting lines with colors representing the terminal group these sequences belong to and their validation results. Bootstrap values for the trees appear low on several edges, indicating difficulty organizing groups effectively based on sequence. However, the comparison of the two trees together shows that the "core" *alt* proteins are primarily concordant with each other. While there are some potential notable exceptions, such as the *altC* from NZ_JACXQ010000006.1, the *Rahnella aqualitis* representative *alt*-like cluster, this appears to be due to the rotation of the tree as opposed to a biological reality. This opinion is further supported by both overlaying the validation data on these tree comparisons, where proteins with similar *alt* tolerance appear to be grouping together, and other comparison trees place the sequence much closer to the other gene synteny groups. These trees suggest that these collections of *alt*-like proteins appear to have independent evolutionary histories as vertically maintained genes despite being horizontally transferred. Further, the concordance of the validation data and these

proteins seem to suggest specialization is occurring with the more robust *alt* phenotypes consistently grouping (Supplementary folder 2.1).

Phenotypic testing with Synthesized *altC*/*altE* Pairs Provides Evidence for Thiosulfinate Tolerance Functionality in Predicted *alt*-like Gene Clusters

To evaluate whether predicted *alt*-like clusters were legitimate and capable of conferring increased thiosulfinate tolerance phenotypes, we heterologously expressed synthesized *altC*/*altE* gene pairs representing key phylogenetic nodes in PA strain PNA 97-1R Δ*alt*, which lacks the functioning *alt* cluster and has poor thiosulfinate tolerance. Strains were grown in 50:50 LB onion extract as in Stice et al., and the mean area under the growth curve (AUC) was determined. Growth was compared against a thiosulfinate-sensitized PNA 97-1R Δ*alt* GFP expressing strain as a control and the PA wild-type strain (PNA 97-1R). Across all experiments, expression of GFP in PA PNA 97-1R Δ*alt* consistently showed the lowest growth of the inoculated OJ, indicating poor thiosulfinate tolerance.

In contrast, our positive control, PA PNA 97-1R WT, showed robust growth. Irrespective of *altC/altE* pairs from different bacteria, heterologous expression in PA PNA 97-1R Δ*alt* resulted in increased tolerance to thiosulfinate in our onion-juice (OJ) growth assay. The *altC/altE* pairs for bacteria that are closer to PA phylogenetically (*Erwinia persicina* CFBP8795, *Rahnella aquatillis* Ra9-2) tended to result in improved restoration of thiosulfinate tolerance to PNA 97-1R Δ*alt* compared with those that were phylogenetically distant (*Paenibacillus nuruki* TI45-13ar, *Burkholderia gladioli* BCC1802, and *Novosphingobium sp.* Chol11). However, an exception in this trend was observed with *Gluconobacter kondonii* (Dm-54). Despite its relative closedness with PA

phylogenetically, the *altC/altE* heterologous expression in PA PNA 97-1R Δ*alt* did not result in consistent growth in the OJ growth assay, indicating comparatively lower tolerance to thiosulfinates (Figure 2.6 A-C and E). In addition, *Cronobacter dubliensis* (cro910B3) showed weaker tolerance but more robust tolerance than that of *Gluconobacter kondonii* (Dm-54) despite its relative closeness with PA phylogenetically. Overall, while all *altC/altE* pairs conferred increased thiosulfinate tolerance, the quantitative performance of individual *altC*/*altE* pairs is not easily predicted based solely on their phylogenetic similarity (Figure 2.6 D).

A hierarchical clustering analysis was conducted based on the Euclidean distance of the growth curves from the experiments to provide a comprehensive view of the growth response patterns across different bacterial strains. This analysis categorized the bacterial strains into clusters based on their growth response to thiosulfinate exposure. The hierarchical clustering dendrogram revealed distinct clusters, with each branch representing a similarity in growth responses among the strains. A distinct cluster formed by *G. kondonii* (Dm-54), *C. dublinensis* (cro910B3), and *P. nuruki* (TI45-13ar) indicates unique growth response profiles, which is supported by the unexpectedly poor performance of *G. kondonii* (Dm-54), and variability in responses from both *C. dublinensis* (cro910B3), and *P. nuruki* (TI45-13ar). A second significant cluster includes *E. persicina* (CFBP8795) and *P. ananatis* 97-1R WT, showing more similar growth curves when compared to the remaining strains. These results are supported by the consistently high performance of the *E. persicina* (CFBP8795) *altC*/*altE* pair. The next group consists of similarly performing strains with *altC*/*altE* pairs from *Pseudomonas sp.* (Root569), *R. aquatilis* (Ra9-2), *Novosphingobium spp.* (Chol11), *P. aryabhattai* (LAD), and *S.*

maltophilia (CV_2003_STM1) with *P. ananatis* (PNA 97-1R) placed in an intermediate rating with the previous group. These results are supported by the consistently high, but not as high, performance of *P. ananatis* (PNA 97-1R) when compared to *E. persicina* (CFBP8795), but not as variable as the remaining members of the group. *V. coralliilyticus* (09-121-3), and *B. gladioli* (BCC1802), *P. syringae* pv. tomato (DC3000), and *P. fluorescens* (PS838) are the final group. As expected, the GFP strain is positioned near the negative control, reinforcing its minimal growth and low tolerance to thiosulfinates because it lacked *alt* genes. The hierarchical clustering analysis provides a comprehensive view of the growth response patterns across different bacterial strains. It aligns with the tolerance experiments and phylogenetic analysis findings, demonstrating their similar growth profiles and tolerance mechanisms independently of protein sequence content or lineage.

Binding Affinity Prediction with AI-BIND of *altR* Demonstrates NLP-like Techniques Are Effective for Predicting and Classifying *alt* and *alt*-like Proteins

To determine if NLP-like techniques for binding affinity prediction could be used to help differentiate between functional *alt* clusters and possible pseudo clusters, we utilized AI-Bind to screen our *altR* protein sequences against a library of small molecules collected from PubChem focusing on sulfur compounds (Supplementary file 2.2). Due to the likelihood of noise among most of these binding predictions, rows within .001 similarity were extracted for individual assessment. Among the values extracted, several similarities among the columns can be seen, and these 28,481 predictions may be the primary drivers for the separations seen with the Levenshtein distance matrix. When the distance matrix is overlaid with the gene synteny plot and compared to the *altR* RAxML

tree, it appears that the results generated from AI-Bind are capable of sorting *altR* proteins into their appropriate gene synteny groups. In addition, when integrating the findings with those from the experimental validation experiments, there is a pronounced division between *alt* clusters with robust phenotypes and those exhibiting weaker phenotypes. These results indicate that the results from AI-Bind could also sort *altR* proteins into groups that reflect the thiosulfinate tolerance of their respective *altC*/*altE* pair. As such, the binding predictions that AI-Bind produced may be helpful in further methodologies to automate the detection and distinction of *alt*, *alt*-like, and pseudo-*alt* proteins. This pattern further supports the notion that most, if not all, of the DeepBGC-identified representative *alt*-like clusters in this study are capable of functioning similarly to *alt*, a conclusion reinforced by the experimental validation results (Figure 2.7).

Discussion

Identifying *alt* and *alt*-like clusters poses challenges concerning variable gene synteny and divergent sequence similarities across bacterial genera. Specifically, *'alt* clusters' refer to gene clusters experimentally validated to exhibit the thiosulfinate tolerance phenotype. Meanwhile, *'alt*-like clusters' resemble these gene clusters in genetic composition but lack experimental validation for the associated phenotype. 'Pseudo *alt* clusters,' on the other hand, have been experimentally shown to not possess the phenotype despite their similarity in appearance to *alt* clusters. The three gene clusters utilized in our training set show little gene and protein sequence similarity and do not share overall gene synteny. Our analysis found a limited set of seven genes shared among our lab-validated gene clusters, with notable variations in gene and protein sequence similarities—highlighting the *altB* reductase in the SDR family oxidoreductase

family as the most conserved element across these clusters. The observed sequence similarities range significantly, suggesting a nuanced spectrum of conservation and divergence within these gene clusters. Interestingly, despite the diversity in sequence similarity, the predicted protein structures demonstrated a surprising level of uniformity according to I-TASSER system evaluations. This uniformity, especially in the context of different pathogens from onion, underscores a potentially ancient divergence and pseudo-vertical transmissibility for this horizontally transferred region.

Current *alt* clusters have been identified experimentally or predicted intuitively based on gene co-localization and annotation. However, this approach is difficult to rigorously codify and could lead to significant discrepancies between investigators. Further, we do not have a collection of pseudo-*alt* clusters to provide a comparison, exacerbating the difficulty in describing an actual *alt* cluster. Computational strategies, such as artificial intelligence methodologies like machine learning or deep learning, offer more sophisticated ways to "digitize" intuition for dissemination. In this work, we utilized an NLP-like method to generate a model capable of data mining these complex gene clusters with an unconventional training set of only 3 divergent validated gene clusters and, by extension, make a transition from bespoke manual curation of *alt* clusters into a streamlined process.

NLP in biology is becoming a valuable tool in studying gene function. There is a significant volume of genes that have an unknown function. By extension, we cannot access these genes' full potential for biotechnology, agriculture, and medicine. In prokaryotes, for example, genes with a complementary function tend to group into biosynthetic gene clusters (14, 15, 16, 27). Relevant biosynthetic gene clusters can be

detected and datamined by focusing on higher-order information and gene proximity. In this work, we used DeepBGC to train a model of our three previously validated *alt* clusters to overcome the limitations with more traditional sequence-only methods for data mining gene clusters, as well as explore the utilization of these techniques for identifying patterns that can be useful for identifying these clusters more robustly. DeepBGC utilizes Pfam information rather than the amino acid sequence to classify BGCs, with the additional caveat of understanding the importance of gene localization in gene clusters (24). Utilizing vectorized Pfam domains and gene localization elegantly simulates our intuitive process to curate gene clusters and produces a tangible model that is more appropriate for rigorous scientific evaluation. Sequence-based methods, such as BLAST, had been insufficient for data mining these clusters across multiple genera due to low sequence similarity. However, the methodology utilized by DeepBGC produced a model that can successfully detect *alt* clusters reproducibly from diverse genera of bacteria. In an ideal scenario, a bioinformatician would have access to thousands of examples for their training set. In this study, we only had access to three validated examples of *alt*-clusters from PA, BG, and PTO. Despite this, we were able to successfully detect, retrieve, and validate several *alt* clusters that were previously undetectable. This methodology also alleviates the immense effort required to screen this expansive list of bacterial genomes. Utilizing NLP technologies in a biological context is a powerful tool for "standardizing" the intuitive extraction process.

We acknowledge that using the model provided in the supplementary materials, with a training set of only three clusters, is too broad to filter out background noise effectively. For example, our analysis incorrectly identified several gene clusters simply

due to the presence of several tandem thioredoxin-related genes. Additionally, other clusters were mistakenly detected due to the presence of multiple copies of *tetR*-family repressor genes, leading to false classifications. Further, some datamined *alt* clusters would lose genes on the terminal ends of their gene cluster, but the same cluster from another genome would contain the entire expected sequence. This issue is resolved by running the model multiple times and determining the “average” cluster sequence. However, these types of errors are commented upon in the DeepBGC manuscript and are to be expected (24). As always, manual curation should be employed to ensure that AI models behave appropriately. Despite the occasional error in incredibly diverse genomes, when the model is run on the genome of an onion pathogen with a known *alt* cluster, deepBGC always performed the expected extraction.

We utilized another text-comparison technique to compare gene synteny. The complexity of the *alt* clusters often overwhelmed traditional DNA-sequence-based methods, frequently leading to system failures in organizing the information. However, by converting from one language to another and calculating the Levenshtein distance matrix, we successfully organized gene clusters into gene synteny groups quickly and reliably. The Levenshtein distance matrix is the "edit distance" between two strings. These are insertions, deletions, and substitutions (28, 29). We cut down the computational time and simplified the visualization process by converting our gene clusters into a color code and then a string representing these color codes. The application of language processing techniques is not limited to complex AI modeling or requires expensive computational equipment to be helpful. By applying the Levenshtein distance matrix, we can initiate the classification of *alt* clusters based on higher-order information, such as guilt-by-

association syntax, in a comprehensive manner. Although the *alt* cluster exhibits many characteristics typical of horizontally transferred gene clusters, it appears to be maintained vertically within several bacterial genera. Despite this, gene synteny is not unique across bacterial genera, as specific gene patterns recur across multiple genera even if their sequence content is different.

When comparing the gene synteny to the experimental validation of *altC*/*altE* pairs, it appears that the *altC*/*altE* pairs from the first terminal group have more robust thiosulfinate tolerance restoration phenotypes than *altC*/*altE* pairs from other terminal groups. These *alt* clusters are also represented among many members of the enterobacteria; however, *alt*-like clusters from *Erwinia*/*Pantoea* displayed the strongest phenotype. A notable caveat with this methodology is that the strongest phenotype is observed when these clusters are expressed in *Pantoea*, potentially due to interactions and dependencies with other endogenous host factors. Based on the current information, gene cluster synteny alone is insufficient for comprehensively categorizing *alt*-like clusters. These results are unsurprising, as genes with distinct evolutionary histories can independently form gene clusters with similar synteny. However, repeating motifs among several gene clusters is a strong indication of collaboration for a phenotype, and we would argue that the guilt-by-association of these shared genes is still a substantial factor in identifying *alt*-like clusters, even if their motifs are not perfect indicators of *alt*-like phenotype performance in OJ.

In drug discovery, the conformation and 3D structure of molecules are critical, as small molecules must fit into a binding pocket of a target protein with a favorable reaction. Similarly, proteins that yield similar phenotypic functions are expected to have

comparable shapes, regardless of their sequence similarity (30). This work explored the potential for predicted protein conformation to indicate the *alt* phenotype. The results of the *altC*/*altE* pair validation suggest that the 3D superimposition of our putative *alt*-like proteins to the *alt*-verified proteins may indicate qualitative but not quantitative phenotype. Perhaps proteomic profile to compare proper *alt* and pseudo-*alt* proteins are necessary, as our validation experiment showed all selected *altC*/*altE* pairs provided thiosulfinate tolerance, with some exception to the pair derived from *Gluconobacter kondii* (Dm-54). For example, when interpreting our phenotypic validation results in the context of our 3D superimposition, it is essential to note that the *altC*/*altE* pair from *Priestia aryabhattai* LAD (NZ_CP072478.1) exhibited a more robust thiosulfinate tolerance restoration phenotype compared to *Pseudomonas sp.* Root569 (NZ_LMGQ01000029.1), despite the latter with higher Zeal scores. This observation suggests that while structural similarities generally correlate with functional outcomes, exceptions highlight the complexity of phenotype-genotype relationships. Furthermore, *altC* variants with Zeal scores greater than 0.98 consistently supported more robust bacterial growth in our thiosulfinate tolerance growth assay, implying a potential threshold effect where high structural fidelity may enhance certain functional capabilities. Conversely, *altE* adds another layer of complexity; *Priestia aryabhattai* LAD (NZ_CP072478.1), with a lower Zeal score of only 0.69, showed a slightly stronger thiosulfinate tolerance restoration phenotype than *Novosphingobium sp*. Chol11 (NZ_OBMU01000004.1), which had a higher Zeal score of 0.94.

Additionally, our comparison of the *E. coli nemR* repressor with the four *altR* sequences in the genomes used for our training dataset revealed high similarity in their 3D protein structures. The *nemR* repressor in *E. coli* shows ranges from 0.88 to 0.94

similarity. In contrast, the other four exhibit similarities ranging from 0.92 to 0.98. This level of resemblance is expected, given that they all are annotated as *tetR* repressors. The *nemR* repressor in *E. coli* is assumed to be responsive to reactive chlorine (bleach) and nitrogen species [31]. As such, we find the protein shape to be helpful in providing a secondary opinion for the protein predictions, as apparent outliers can be screened independent of annotations but alone cannot be used to separate functional *alt* proteins from possible pseudo-*alt* proteins. These findings underscore the limitations of relying solely on structural predictions to infer functional characteristics, highlighting the need for more complex integrated approaches to classify *alt* and pseudo-*alt* proteins. This opinion is reinforced by the results of the AI-Bind screen, where we assessed if screening potential binding affinity of proteins to a set of organo-sulfur molecules could differentiate *alt*-like clusters.

We utilized AI-Bind to evaluate the predicted binding affinity of *altR* sequences against a library of 381,350 small molecules. We then calculated the string differences from a concatenation of the resulting scores to determine if the output could be informative for classification. Initially, the matrix generated from the AI-Bind average scores seemed discordant compared to the trees derived from protein sequence similarity. However, overlaying the AI-Bind prediction matrix with the data from the gene synteny matrix, as well as the result of the experimental validation, shows that binding affinity predictions from AI-Bind are capable of sorting *altR* proteins into groups that are reflective of our other screening methods, independently. These findings suggest that using average binding predictions may be an effective tool for further classifying *alt* clusters and separating *alt* and pseudo-*alt* proteins. It is important to note that AI-Bind, however, is not

in and of itself utilized for the classification of proteins in this way and is only NLP-like in that it could classify based on sequence data rather than utilizing more traditional NLP-like systems.

Conclusions

NLP-like technologies are powerful tools to assist in the discovery and classification of gene clusters. Here, we generated a model capable of detecting and extracting *alt* clusters, validating the phenotype in transformed bacteria that previously lacked it. Despite its limited training set, the NLP-like algorithm used here demonstrated its capacity to identify several biologically relevant gene clusters. A model that quickly and accurately discovers and extracts *alt* clusters proves beneficial for diagnostic plant pathology and environmental bacteriology, particularly as the *alt* cluster is crucial for effectively colonizing *Allium* species or other thiosulfinate-producing hosts. The distribution of *alt* clusters beyond plant pathogens aligns with these secondary metabolites, shaping their microbial communities, as observed with the benzoxazinoid tolerance of maize root colonizers. Employing sophisticated NLP-like tools may revolutionize our understanding of critical gene clusters that facilitate complex host-microbe interactions, potentially leading to breakthroughs in several multidisciplinary fields. In developing a more robust *alt* cluster detection system, integrating models that encompass Pfam domains, gene localization, and predicted binding affinity might be sufficient to distinguish between *alt* clusters—those experimentally validated to function—and pseudo *alt* clusters, which appear similar but are experimentally validated not to possess the phenotype.

Materials and Methods

alt Gene Cluster Seed Sequences

For this work, we used three validated *alt* clusters for DeepBGC training. Each cluster is distinct in both gene sequence and gene synteny. The 11-gene *Pantoea alt* cluster was used from *Pantoea ananatis* strain PNA 97-1R plasmid unamed2 (NCBI accession: PRJNA384061). The 7-gene *Burkholderia alt* cluster was used from *Burkholderia gladioli* pv. *gladioli* strain FDAARGOS_389, plasmid unnamed (NCBI accession: PRJNA231221). The *Pseudomonas alt* cluster was used from *Pseudomonas syringae* pv. *tomato* str. DC3000, complete genome (NCBI accession: PRJNA57967) (supplementary folder 2).

Gene/Protein sequence comparisons for validated *alt* clusters

To understand sequence similarity between validated *alt* genes, we performed multiple sequence alignments of protein and nucleic acid sequences at default settings using the Clustal Omega online server (32).

DeepBGC training and RefSeq screening

To determine if AI trained on higher-order information can assist in efficiently datamining *alt*-like clusters from a collection of genomes, we trained the DeepBGC model on our small sample size of 3 validated *alt* clusters. The *alt* detection model was trained using the author's supplied negative dataset "GeneSwap_Negatives.pfam.tsv" and ran with the default provided "deepbgc.json" with DeepBGC version 0.1.27. DeepBGC training on the initial three *alt* sequences was repeated 15 times and the reports were compared to assess model performance. DeepBGC options on the database data mining included a minimum protein count of 4 and a minimum score of .9. RefSeq genomes were separated

into 48 sub-directories of 5,000 genomes, and DeepBGC jobs were submitted to the UGA GACRC via an array element on the batch partition. We scanned 238,362 genomes using this model from the NCBI bacterial refseq database. The genomes were downloaded via the NCBI FTP service, and the assembly list is provided (Supplementary files 3 and 4) (24).

Filtering gene cluster representation via MMseqs2

To compress the DeepBGC extractions into a smaller representation for analysis, we used MMseqs2 release 13-45111 to generate representative sequences. The options used were a query coverage of 90%, sequence homology of 75%, and connected component clustering. These options allow for a "core" representative sequence with leniency for small changes in gene presence or absence (33).

BLAST of NCBI GenBank for Representative Sequence Diversity

The final selection of 47 representative *alt*-like clusters was utilized as the query sequence for both the collection of putative *alt*-like gene clusters from DeepBGC and NCBI GenBank, following similar rules to the MMSEQ2 redundancy filtering. Recovered species were then counted and organized into a list for the calculation of the Shannon-Wiener Index via our own Python script.

Gene synteny comparisons

During our manual curations of *alt*-like clusters we noticed a pattern where gene synteny was conserved among bacterial genera. For those clusters, we generated cluster comparisons with sequence data, as the method for data mining was determinate upon them. However, post-DeepBGC screening of RefSeq, we found many *alt*-like clusters that share low enough sequence similarity between genes of similar annotation that several

methods for cluster comparison would fail. To overcome this barrier, we assigned a color code to *alt*-like genes that received pfam tags like our test run of the original test sequences. To optimize the human capacity to read the information and remove unintended bias between colors, we assigned several shades of green to *alt*-like genes and grey color to genes that are not relevant. We then used an Excel script to convert these colors into color codes. These codes were then concatenated into strings and underwent a Levenstein distance matrix calculation using the Levenstein and Dendropy Python packages [34, 35]. After initial tree construction, further manual curation was applied to finalize *alt*-like representatives by selecting gene clusters with at least 3 unique *alt*-like pfam tags that match those applied to the seed clusters and the removal of clusters that were split into separate contigs.

Generating 3D protein models and Zeal score comparisons

While the previous methods make gene comparisons primarily on sequence or trained guilt-by-association with higher-order information, we wanted to compare *alt*-like proteins for potential structure diversity or abnormalities directly. Models for select *alt*-like proteins were generated using I-TASSER 5.2 with the -LBS option set to true. Predicted protein 3D models were then compared using the Zeal GUI with global alignment.
(36, 37).

Protein-ligand binding prediction

The *alt*, and *alt*-like mechanisms of action are currently unknown. However, it is reasonable to suspect that the binding interactions between chemicals and proteins would be essential in defining *alt*, *alt*-like, and pseudo-*alt* gene cluster classes. Due to the computationally expensive nature of drug-target binding predictions we turned to using

AI-Bind, a deep-neural network designed for a more generalizable prediction of binding between proteins and small molecules. A comprehensive list of small molecule SMILES and InChiKeys were downloaded from the PubChem database with the following search terms: “allyl, cysteine sulfoxide, disulfide, polysulfide, S-Nitrosothiol, sulfenic acid, sulfenic, Sulfimide, sulfinic acid, silfinic, sulfone, Sulfonic acid, Sulfonic, Sulfonium, sulfoxide, sulfoximide, Sulfurane, thiolaldehyde, thioamide, thiocarbonyl, thiocarboxylic acid, thioester, thio, thiosulfinate, 316263-glutamylcysteine, and s-Allylmercaptoglutathione." The results were concatenated, and duplicate entries were removed for a final list size of 381,349 small molecules. In our final representative dataset, these were then screened against the 53 *altR*-like proteins that received Pfam tags from deepBGC. Binding results from the *altR*-like proteins were converted into strings and compared via the calculation of the Levenstein distance matrix above to produce a neighbor-joining tree for ease of comparison (38).

altC/*altE* Validation

To validate the representative *alt*-like clusters produced by DeepBGC, we conducted an onion-juice (OJ) growth assay. Previous research has demonstrated that the presence of *alt*C alone is sufficient to determine an *alt* phenotype. As such, we selected 14 *altC* genes (supplementary table 2) for validation, along with their potential *altE* partner if present. These 14 gene pairs (*altC*/*altE*) were inserted into Twist Bioscience’s pENTR plasmid and inserted into *P. ananatis* PNA 97-1^{R} *Δalt* [10] by the following method.

Electroporation and Confirmation

The recipient strain's electrocompetent (e-comp) cells (*P. ananatis* PNA 97-1R *WT* and *Δalt*) were prepared using standard methods. Plasmid constructs were electroporated

into the recipient cells at 1.8kV. Transformed cells were mixed in 1 ml LB and left for incubation at 28°C for an hour. Post-incubation, cells were pelleted, resuspended in LB, and plated onto LB+Km plates. Individual transformed recipient cell colonies were grown overnight in LB+Km broth. Plasmids were extracted and sequenced to confirm the insertion of the pENTR plasmid constructs into recipient cells.

The plasmid pENTR::GFP served as an empty vector and was inserted into both PNA 97-1^{R} *WT* and *Δalt* strains, which acted as positive and negative controls for the onion-juice growth assay. The remaining 14 plasmid inserts were transformed into PNA 97-1^{R} *Δalt* strains.

Preparation of Onion Juice Extract

Onion juice was extracted using Juicer method (10). One yellow onion bulb (400-500 g) was processed through an industrial strength juicer, resulting in 300-400 mL of crude onion extract. The extract was then centrifuged at 10,000 g for 1.5 hours at 4°C. After centrifugation, the supernatant was carefully removed and filtered through a Nalgene disposable vacuum filter sterilization unit. The onion juice was then stored at -20^{o}C for future use.

Liquid Growth Assay

The growth assay used 100-well honeycomb plates with the BioScreen C system (Lab Systems Helsinki, Finland). Seven-day-old OJ was utilized for the assay. 16 bacterial strains culture was started on LB+Km plates, and overnight cultures were prepared the following day in LB+Km broth from single colonies. On the third day, the growth assay was conducted for 48 hours with low agitation at 28°C. The growth media consisted of LB supplemented with an equal volume of onion juice. The experiment included 16 test

strains (figure 6), including PNA 97-1 WT pENTR::GFP and *Δalt* pENTR::GFP strains, and a negative control (LB+OJ). Each well contained 400 µL of a mixture comprising 360 µL of growth media (LB+OJ) and 40 µL of a bacterial suspension with an OD600 of 0.5 in sterile dH2O, with a minimum of 5 well replicates. Absorbance values were measured every 30 minutes for 48 hours. Each experimental assay was repeated twice for biological replication.

Euclidean Distance Comparison of Treatment Groups

CSV files generated from three growth phases from our experiment, the lag, log, and stationary phases, were used to compare Euclidean distances between the growth curves of each treatment group via a Python script. For each phase, mean trendlines were calculated by averaging the sample data for each treatment group. If trendlines varied in length, shorter sequences were padded with NaN values to match the most extended sequence in each phase. Euclidean distance matrices were generated using the pdist function from the scipy.spatial.distance module, with NaN values imputed using the SimpleImputer with a mean strategy. For each phase, pairwise distances were calculated between all samples, and the resulting distance matrices were saved as CSV files. The average distances within and between treatment groups were computed to create symmetric group-level distance matrices that represented the average pairwise Euclidean distances for each treatment group in each phase. Hierarchical clustering was performed on these group-level matrices using the linkage method with 'average' linkage, and dendrograms were generated to visualize the clustering relationships between treatment groups. Dendrogram branches were color-coded according to predefined treatment groups to reflect their clustering pattern. The aggregated group average distances across

all phases were calculated to compare the treatment groups' growth patterns. A neighbor-joining tree for the combined dataset was generated and saved for visualizing the clustering results (supplementary figure 3).

Tree Comparisons

Maximum likelihood trees were compared against each other using the phytools R package (39). For figure 7, edges between the nodes were organized based on their color code.

References

1) Lancaster JE, Collin H. 1981. Presence of alliinase in isolated vacuoles and of alkyl cysteine sulphoxides in the cytoplasm of bulbs of onion (*Allium cepa*). Plant Sci Lett. 22(2):169–176.
2) Rose P, Whiteman M, Moore PK, Zhu YZ. 2005. Bioactive S-alk(en)yl cysteine sulfoxide metabolites in the genus *Allium*: the chemistry of potential therapeutic agents. Nat Prod Rep. 22(3):351–368.
3) Eady CC, et al. 2008. Silencing onion lachrymatory factor synthase causes a significant change in the sulfur secondary metabolite profile. Plant Physiol. 147(4):2096–2106.
4) Leontiev R, Hohaus N, Jacob C, Gruhlke MCH, Slusarenko AJ. 2018. A comparison of the antibacterial and antifungal activities of thiosulfinate analogues of allicin. Sci Rep. 8(1):6763.
5) Muller A, et al. 2016. Allicin induces thiol stress in bacteria through S-allylmercapto modification of protein cysteines. J Biol Chem. 291(22):11477–11490.

6) Borlinghaus J, Albrecht F, Gruhlke MC, Nwachukwu ID, Slusarenko AJ. 2014. Allicin: chemistry and biological properties. Molecules. 19(8):12591–12618.
7) Jones MG, Hughes J, Tregova A, Milne J, Tomsett AB, Collin HA. 2004. Biosynthesis of the flavour precursors of onion and garlic. J Exp Bot. 55(404):1903–1918.
8) Curtis H, Noll U, Störmann J, Slusarenko AJ. 2004. Broad-spectrum activity of the volatile phytoanticipin allicin in extracts of garlic (*Allium* sativum L.) against plant pathogenic bacteria, fungi, and oomycetes. Physiol Mol Plant Pathol. 65(2):79–89.
9) Wallock-Richards D, Doherty CJ, Doherty L, Clarke DJ, Place M, Govan JR, Campopiano DJ. 2014. Garlic revisited: antimicrobial activity of allicin-containing garlic extracts against Burkholderia *cepa*cia complex. PLoS One. 9(12)
10) Stice SP, Thao KK, Khang CH, Baltrus DA, Dutta B, Kvitko BH. 2020. Thiosulfinate tolerance is a virulence strategy of an atypical bacterial pathogen of onion. Curr Biol. 30(16):3130-3140.e6.
11) Stice SP, Stumpf SD, Gitaitis RD, Kvitko BH, Dutta B. 2018. *Pantoea ananatis* genetic diversity analysis reveals limited genomic diversity as well as accessory genes correlated with onion pathogenicity. Front Microbiol. 9:184.
12) Paudel S, Zhao M, Stice SP, Dutta B, Kvitko BH. 2023. Thiosulfinate tolerance gene clusters are common features of Burkholderia onion pathogens. Mol Plant Microbe Interact. 36(5):355-375.
13) Borlinghaus J, Bolger A, Schier C, Vogel A, Usadel B, Gruhlke MC, Slusarenko AJ. 2020. Genetic and molecular characterization of multicomponent resistance of Pseudomonas against allicin. Life Sci Alliance. 3(8)

14) Hopwood DA, Merrick MJ. 1977. Genetics of antibiotic production. Bacteriol Rev. 41:595–635.

15) Martin JF. 1992. Clusters of genes for the biosynthesis of antibiotics: regulatory genes and overproduction of pharmaceuticals. J Ind Microbiol. 9(2):73–90.

16) Martín MF, Liras P. 1989. Organization and expression of genes involved in the biosynthesis of antibiotics and other secondary metabolites. Annu Rev Microbiol. 43:173–206.

17) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

18) Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39–W346.

19) Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, Wohlleben W. 2009. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol. 140(1-2):13–17.

20) Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, et al. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 158(2):412–421.

21) Yoon BJ. 2009. Hidden Markov models and their applications in biological sequence analysis. Curr Genomics. 10(6):402–415.

22) Choo KH, Tong JC, Zhang L. 2004. Recent applications of Hidden Markov Models in computational biology. Genomics Proteomics Bioinformatics. 2(2):84–96.

23) Eddy SR. 2004. What is a hidden Markov model? Nat Biotechnol. 22(10):1315–1316.

24) Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, Durcak J, Al-Mahfoudh R, et al. 2019. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 47(18)

25) Hochreiter S, Schmidhuber J. 1997. Long short-term memory. Neural Comput. 9(8):1735–1780.

26) Schuster M, Paliwal KK. 1997. Bidirectional recurrent neural networks. IEEE Trans Signal Process. 45(11):2673–2681.

27) Fischbach M, Voigt CA. 2010. Prokaryotic gene clusters: a rich toolbox for synthetic biology. Biotechnol J. 5(12):1277-1296.

28) Berger B, Waterman MS, Yu YW. 2021. Levenshtein distance, sequence comparison and biological database search. IEEE Trans Inf Theory. 67(6):3287-3294.

29) Levenshtein VI. 1965. Binary codes capable of correcting deletions, insertions, and reversals. Dokl Akad Nauk SSSR. 163(4):845-848.

30) Klebe G. 2013. Protein modeling and structure-based drug design. In: Klebe G, ed. Drug Design. Berlin, Heidelberg: Springer.

31) Jo I, Chung IY, Bae HW, Kim JS, Song S, Cho YH, Ha NC. 2015. Does the transcription factor NemR use a regulatory sulfenamide bond to sense bleach? J Biol Chem. 290(41):24826-24842.

32) Sievers F, Higgins DG. 2021. Clustal Omega. European Bioinformatics Institute. Available from https://www.ebi.ac.uk/Tools/msa/clustalo/.

33) Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 35(10):1026-1028.

34) Levenshtein Python Package. 2023. Available from https://pypi.org/project/python-Levenshtein/.

35) Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 26(12):1569-1571.

36) Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: Protein structure and function prediction. Nat Methods. 12(1):7-8.

37) Ljung F, André I. 2021. ZEAL: protein structure alignment based on shape similarity. Bioinformatics. 37(5):676-677.

38) Chatterjee A, Walters R, Shafi Z, Ahmed OS, Sebek M, Gysi D, Yu R, Eliassi-Rad T, Barabási AL, Menichetti G. 2021. Improving the generalizability of protein-ligand binding predictions with AI-Bind. Nat Commun. 12(1):4847. https://doi.org/10.1038/s41467-021-27668-9.

39) Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 3(2):217-223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

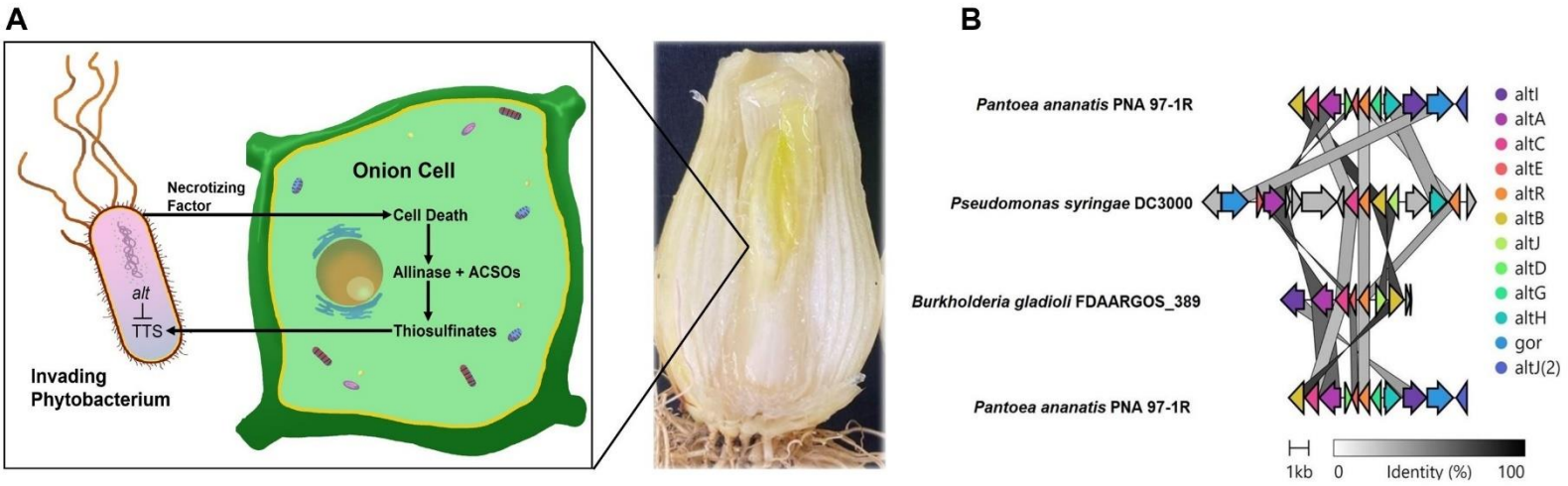


Figure 2.1: Overview of the importance of thiosulfinate tolerance with *Pantoea ananatis* as a model example.

(A) Pictorial representation of the chemical arms race between an invading phytobacterium and its *Allium* host, depicted as *Allium cepa*. When the phytopathogen utilizes its necrotizing factors to kill the host cells, it, in turn, becomes challenged with toxic thiosulfinate stress (TTS) that is managed by the allicin tolerance (*alt*) cohort. In addition, an example of bulb-rot symptoms due to *Pantoea ananatis* compatible interactions in *A. cepa* in onion bulbs is included. The provided example is a longitudinal section of an infected bulb displaying rotten water-soaked center scales with visible bacterial growth.

(B) Gene cluster synteny comparisons of *alt* clusters used as the input sequences for DeepBGC training. These comparisons were generated with Clinker. The arrows represent coding sequences along with their directionality. Shaded lines

reflect the degree of similarity between the gene clusters, with darker shades indicating higher similarity. Arrows are colored based on their *alt* annotations.

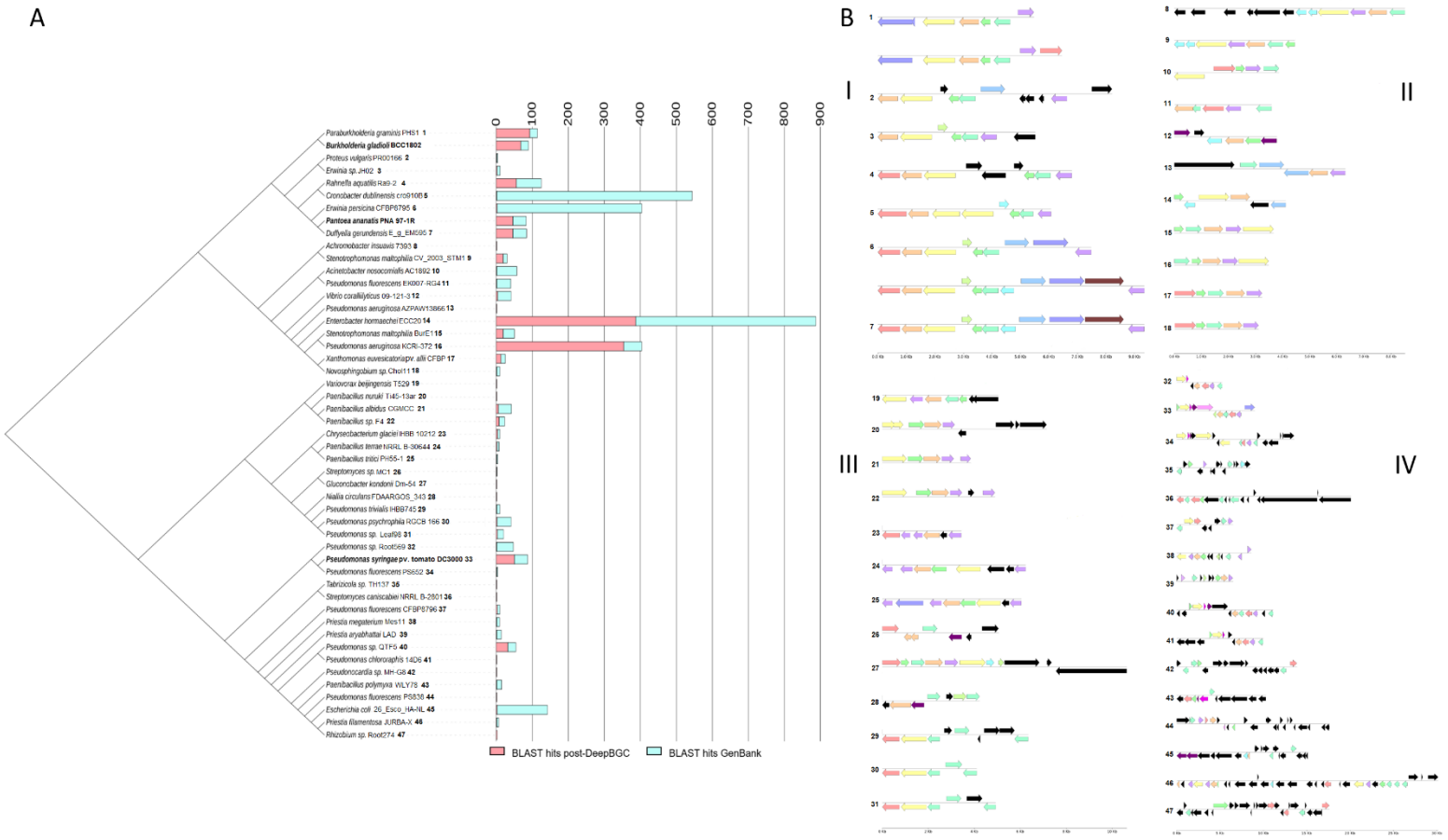


Figure 2.2: Allicin tolerance (*alt*)-like representative clusters post-DeepBGC extraction of the RefSeq bacterial database.

(A) A comprehensive insight into the distribution of *alt*-like clusters within the NCBI system using the Levenshtein distance matrix of color-coded Pfam domain tags. The resulting BLAST hits of representative *alt*-like clusters on all extracted *alt*-like gene clusters from the RefSeq database are shown in pink, while the resulting BLAST hits of representative *alt*-like clusters from the online GenBank bacterial database are shown in blue. Each line indicates 50 sequences. This tree compares the pattern of Pfam tags in gene clusters and should not be misconstrued as a phylogenetic tree.

(B) Color-coded examples of selected representative *alt*-like Pfam domain tags in the deepBGC-extracted clusters; (I) represents the first terminal group between

Paraburkholderia graminis (PHS1) to *Duffyella gerundensis* (E_g_EM595); (II) represents the terminal group between *Achromobacter insuavis* (7393) to *Novosphingobium sp.* (Chol11); (III) represents the terminal group between *Variovorax bejingensis* (T529) to *Pseudomonas sp.* (Leaf98); (IV) represents the terminal group between *Pseudomonas spp.* (Root569) to *Rhizobium sp.* (Root274). These are unrooted neighbor-joining trees based on the Levenshtein differences between a color code conversion of Pfam tags into text strings using the Levenshtein python package. Gene clusters are numbered for ease of comparison. Gene clusters with similar Pfam annotation synteny, but different sequence content was listed as separate clusters (see clusters 17, 18).

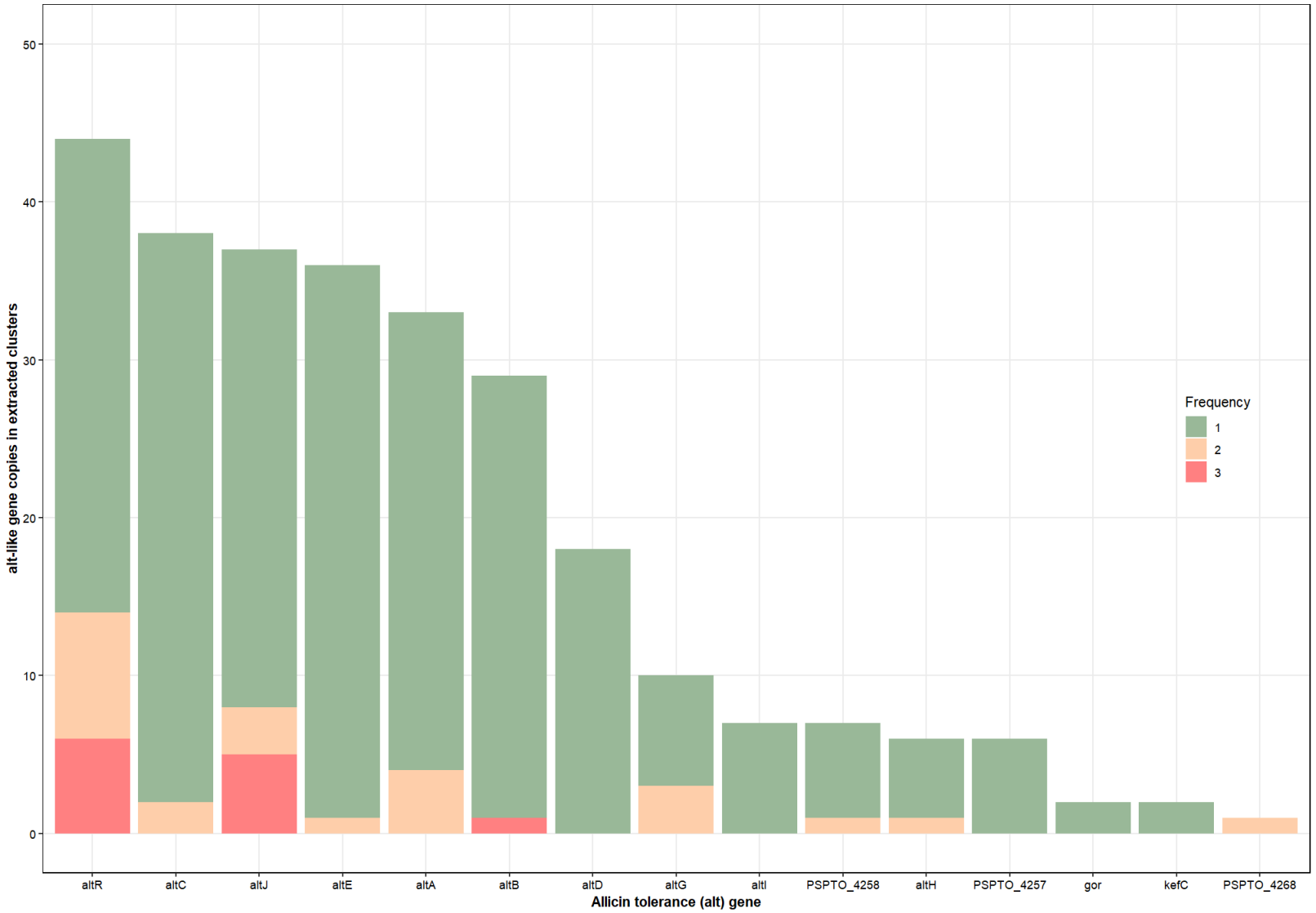


Figure 2.3: Frequency of allicin tolerance (*alt*)-like genes in DeepBGC extracted gene clusters from bacterial RefSeq database. Here, *alt*-like gene frequency was calculated from each of the representative gene clusters and added input clusters to determine the number of copies of alt-like genes present in each gene cluster. The graph is organized based on total gene count, with *altR* being the highest and *PSPTO_5268* the lowest. Green colors indicate the *alt*-like gene appeared only once in the extracted gene cluster. Yellow indicates the gene appeared twice in certain gene clusters. Pink indicates the gene

appeared three times in certain gene clusters. Red indicates the gene appeared four times in certain gene clusters. *altC*, *altE*, *altA*, *altG*, *altH*, and *PSPTO_4258* appear once or twice in certain genomes. *PSPTO_5268* appeared twice in one extracted genome. *altI*, *altD*, *gor*, and *kefC* all appeared only once in their extracted genomes.

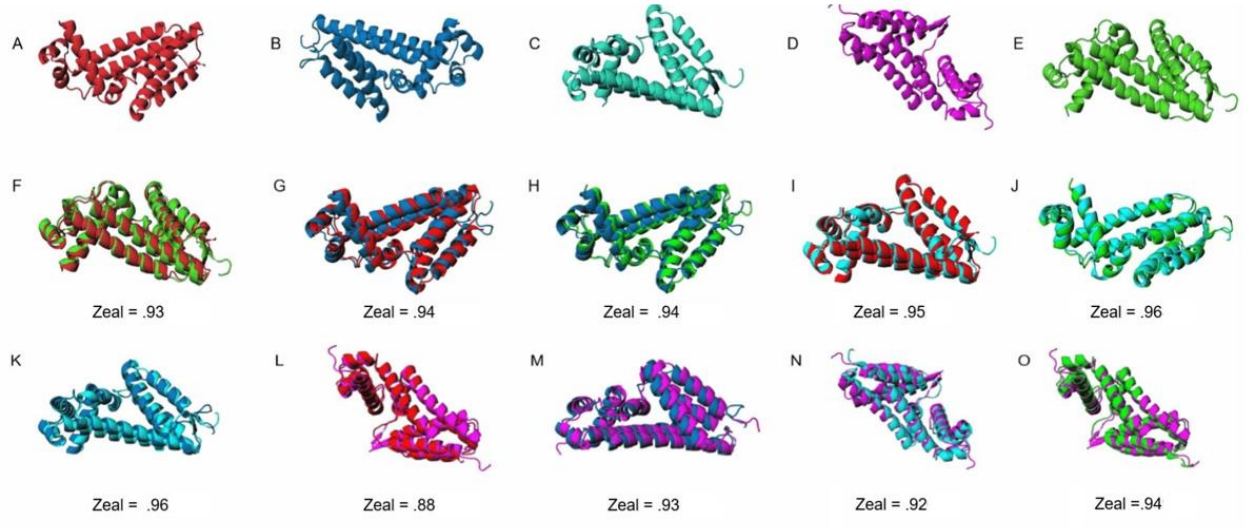


Figure 4.4: Comparative 3D superimposition of I-TASSER predicted *altR* repressors between *Burkholderia gladioli* pv*. gladioli* FDAARGOS_389 (BG), *Pantoea ananatis* PNA 97-1R (PA), *Pseudomonas syringae* pv. tomato DC3000 (PTO), and an unrelated repressor from *Escherichia coli*, *nemR* (ECN). The initial row (A-E) represents all predicted protein structures used for downstream comparisons. The predicted model proteins are displayed as BG *altR* (A), PTO *altR* (B), PTO *out altR* (C), ECN (D), and PA *altR* (E). The 3D superimposition comparisons are shown in panels F-O. Values below each comparison refer to the Zeal score as predicted by the Zeal GUI (https://andrelab.lu.se/) and are an indication of shape similarity. For example, a zeal score of "1" indicates the same 3D protein shape. The "F" compares the predicted *altR* protein from BG vs. PA, while panels G and H represent the comparison of *altR* between BG vs. PTO and PA vs. PTO, respectively. The panels I, J, and K compare *altR* between BG, PA, and PTO vs. the PTO out_*altR* as indicated in Figure 1, respectively. The panels L, M, N, and O compare BG *altR*, PTO *altR*, PTO *out_altR*, and PA *altR* against ECN, respectively.

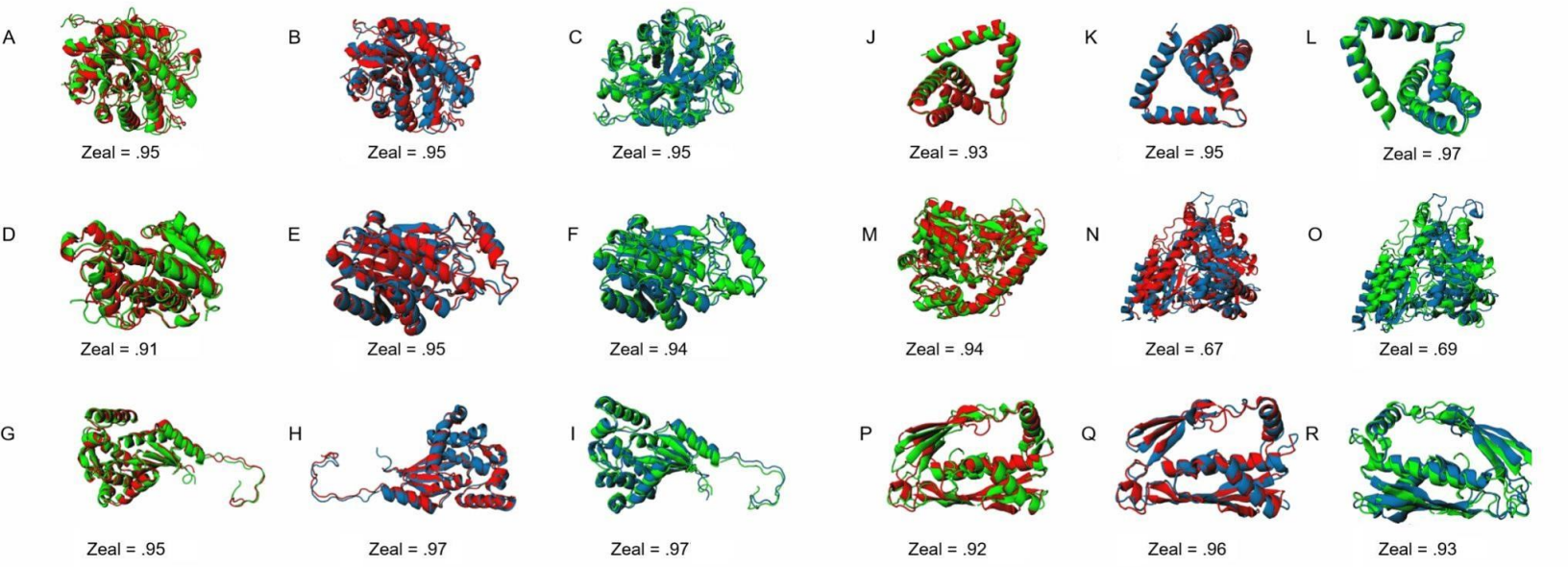


Figure 2.5: Comparative 3D superimposition of I-TASSER predicted allicin tolerance (*alt*) proteins between *Burkholderia gladioli* pv. *gladioli* FDAARGOS_389 (BG), *Pantoea ananatis* PNA 97-1R (PA), and *Pseudomonas syringae* pv. tomato DC3000 (PTO). BG proteins are colored red, PA proteins are colored green, and PTO proteins are colored blue for ease of visualization. Values below each comparison refer to the Zeal score as predicted by the Zeal GUI (https://andrelab.lu.se/) and are an indication of shape similarity. For example, a zeal score of "1" indicates the same shape. Each alt protein prediction is organized into three groups. The A, B, and C are comparisons of *altA* between BG vs. PA, BG vs. PTO, and PA vs. PTO, respectively. The panels D, E, and F are comparisons of *altB* between BG vs. PA, BG vs. PTO, and PA vs. PTO, respectively. The panels G, H, and I are comparisons of *altC* between BG vs. PA, BG vs. PTO, and PA vs. PTO, respectively. The panels J, K, and L are comparisons of *altE* between BG vs. PA, BG vs. PTO, and PA vs. PTO, respectively. The panels M, N, and O are comparisons of *altI* BG vs. PA, BG vs. PTO, and PA vs. PTO, respectively. The panels

P, Q, and R are comparisons of *altJ* BG vs. PA, BG vs. PTO, and PA vs. PTO, respectively.

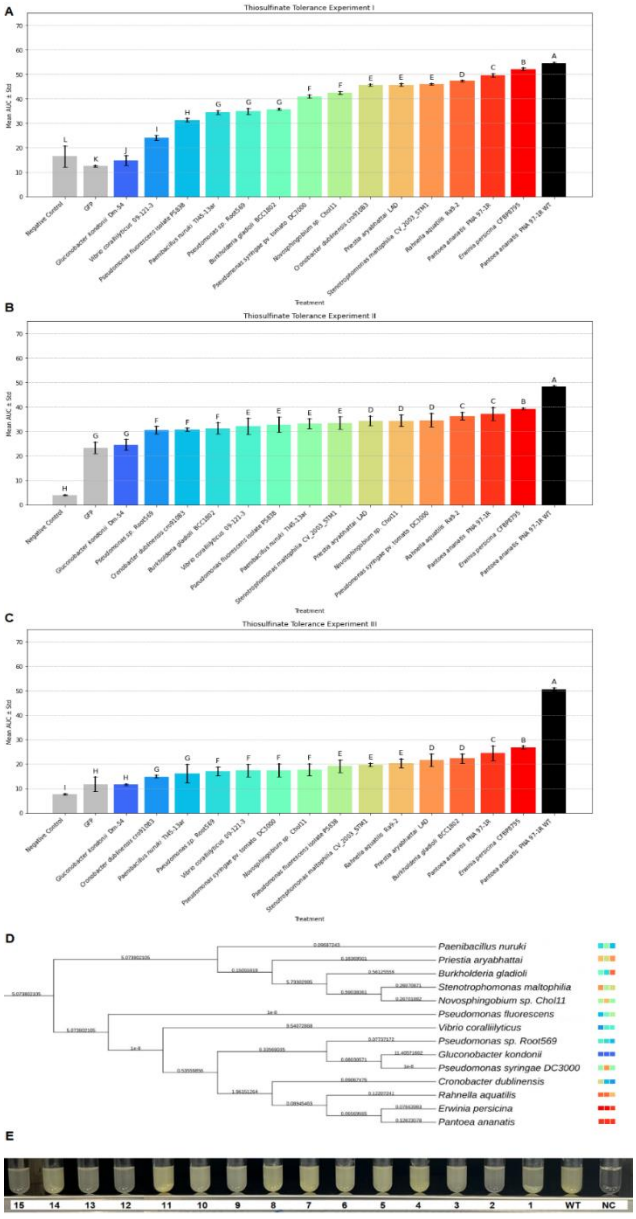


Figure 2.6: Thiosulfinate Tolerance was enhanced in *Pantoea ananatis* when transformed with *altC*/*altE* Pairs from diverse bacterial genera and their phylogenetic relationship among each other. Thiosulfinate tolerance of *Pantoea ananatis* PNA97-1 Δ*alt* was enhanced when transformed with *altC*/*altE* pairs representative of different bacterial genera and species conducted across three experiments, as well as their phylogenetic relationships. The top three bar charts in figure 6A, B, and C, represent three independent

experiments, I, II, and III, respectively, with mean and standard error bars. The x-axis represents *P. ananatis* PNA 97-1R Δ*alt* transformed with *altC*/*altE* pairs from representative of different bacterial genera and species and controls (empty vector and water), while the y-axis shows the mean area under the curve (AUC) values. Statistical groupings are denoted by letters above the bars, with 'A' representing the group with the highest tolerance. Subsequent letters (B, C, D, E…) indicate progressively lower tolerance, based on Tukey's HSD test results, with differences considered statistically significant at $P<0.05$. The negative control and GFP consistently show the lowest tolerance (labeled "E"), whereas *P. ananatis* PNA 97-1R WT and its variants exhibit the highest tolerance (labeled "A"). Figure 6D depicts the phylogenetic relationships among the tested *altC*/*altE* pairs in transformed *P. ananatis* from diverse bacterial genera and species, with branch lengths representing gene sequence distances. The color coding of the transformed *P. ananatis* strains with different *altC*/*altE* pairs matches the bars in the bar charts displayed above, providing a visual correlation between genetic similarity and thiosulfinate tolerance. Strain names were truncated for ease of visualization. For an alternative analysis of growth curve patterns across all experiments, please refer to the Euclidian distance tree in supplementary figure 3. Figure 6E shows a visual comparison of bacterial growth (in terms of turbidity) *P. ananatis* PNA 97-1R Δ*alt* transformed with *altC*/*altE* pairs from the following bacterial strains: 1: *Priestia aryabhattai* LAD; 2: *Novosphingobium* sp. Chol11; 3: GFP; 4: *Pseudomonas spp.* Root569; 5: *Stenotrophomonas maltophilia* CV_2003_STM1; 6: *Burkholderia gladioli* BCC1802; 7: *Pseudomonas fluorescens* PS838; 8: *Rahnella aquatilis* Ra9-2; 9: *Vibrio coralliilyticus* 09-121-3; 10: *P. ananatis* PNA 97-1 (WT, non-transformed); 11: *Pseudomonas syringae*

DC3000; 12*: Paenibacillus nuruki* T145-13ar; 13: *Gluconobacter kondonii* Dm-54; 14: *Erwinia persicina* CFBP8795; 15*: Cronobacter dublinensis* cro91083, and NC: Negative control.

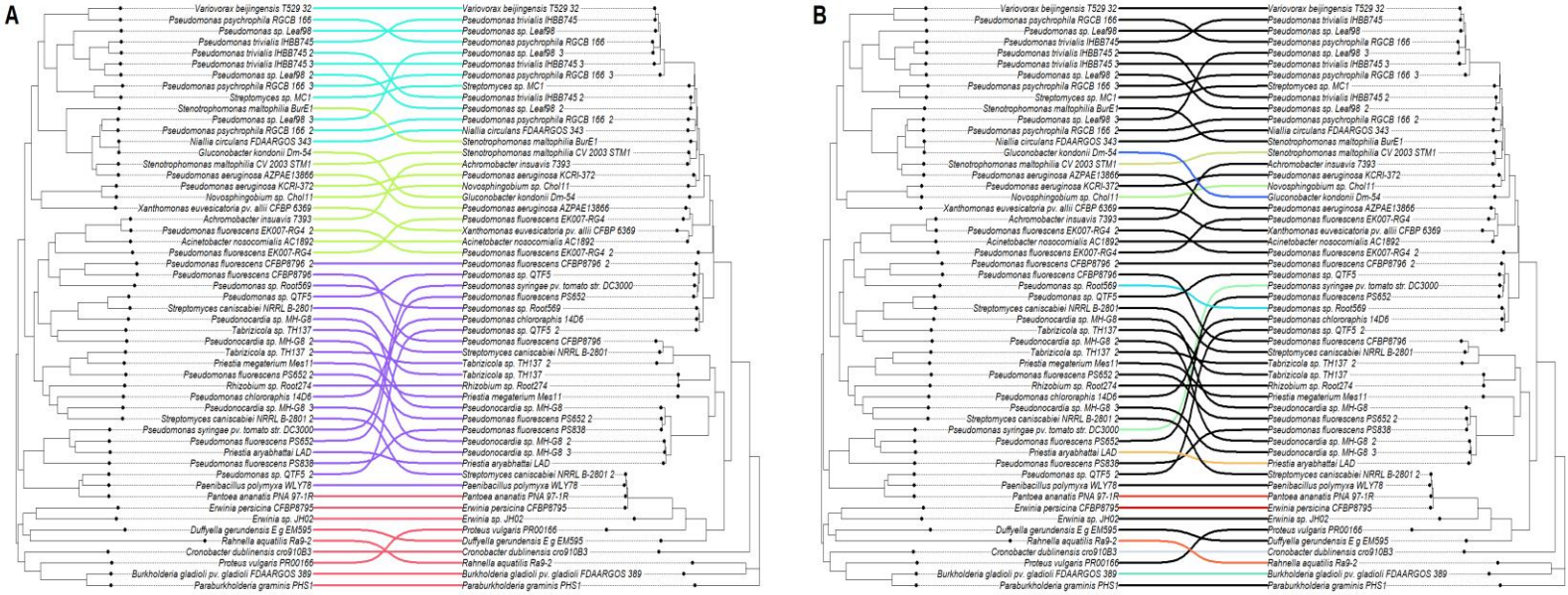


Figure 2.7: Comparison between gene synteny, protein similarity, and binding affinity prediction. Here are two schematic representations (A and B) that map the relationship between gene synteny alongside protein sequence similarity (Panel A) and between protein binding affinity predictions with corresponding protein sequences (Panel B). In Panel A, the branching lines are color-coded to distinguish between different gene synteny groups, which are identified as follows: pink for group I, green for group II, blue for group III, and purple for group IV, as previously defined in figure 2. Panel B contrasts the predicted binding affinities, as calculated by AI-Bind, of *altR* proteins against the similarity of the *altR* sequence. The coloration corresponds to the phenotypic data obtained from follow-up experimental validation, as described in figure 6, with the applied color serving as the "average" RGB value of the three colors.

alt Gene Sequence Similarity[A]	*altI*[B]	*altA*[C]	*altC*[D]	*altE*[E]	*altR*[F]	*altJ*[G]	*altB*[H]	*altR* Gene Sequence Comparisons[I]	*altR*[I]	*gor*	*altJ*	*altH*	*altD*
BG[K] vs PA[L]	52.1	66.0	50.6	62.0	51.0	42.0	74.1	PTO *out_altR* vs PA *altR*	47.5	NA*	NA	NA	NA
PA vs PTO[M]	39.1	66.4	48.8	62.8	46.5	45.1	73.9	PTO *out_altR* vs PTO *altR*	51.1	52.4	21.9	52.3	46.1
PTO vs BG	40.8	69.9	47.3	69.4	51.2	70.5	74.1	PTO *out_altR* vs BG *altR*	54.5	NA	NA	NA	NA

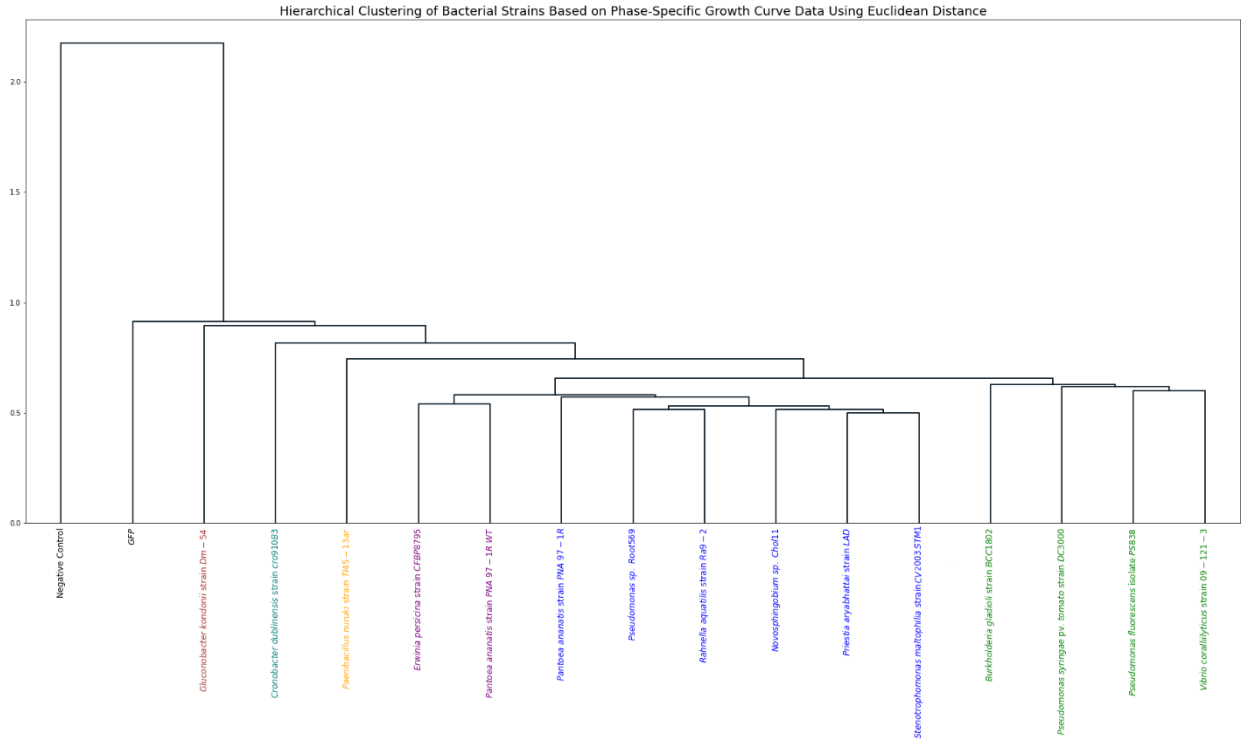
A. Gene multi-sequence alignment similarity calculated by Clustal Omega for shared *alt* genes among the three gene clusters used as the inputs for DeepBGC
B. *altI* gene sequence comparisons with output as percentages
C. *altA* gene sequence comparisons with output as percentages
D. *altD* gene sequence comparisons with output as percentages
E. *altE* gene sequence comparisons with output as percentages
F. *altR* gene sequence comparisons with output as percentages
G. *altJ* gene sequence comparisons with output as percentages
H. *altB* gene sequence comparisons with output as percentages
I. Gene multi-sequence alignment similarity calculated by Clustal Omega for the PTO out_altR gene and the shared *altR* genes among the three gene clusters used as inputs for deepBGC
H. *out_altR* gene sequence comparisons with output as percentages
K. *Burkholderia gladioli* pv. *gladioli* FDAARGOS_389 (BG)
L. *Pantoea ananatis* PNA 97-1R (PA)
M. *Pseudomonas syringae* pv. tomato DC3000 (PTO)
*Not present

Table 2.1: Multi-sequence alignment (MSA) comparison of allicin tolerance (*alt*) genes among shared genes in the *alt* clusters used as the inputs for training by DeepBGC. Sequences were retrieved from *Burkholderia gladioli* pv. *gladioli* FDAARGOS_389 (BG), *Pantoea ananatis* PNA 97-1R (PA), and *Pseudomonas syringae* pv. tomato DC3000 (PTO). MSA was calculated using Clustal Omega in the default settings at https://www.ebi.ac.uk/Tools/msa/clustalo/. Comparisons were made between shared alt genes (*altI*, *altA*, *altC*, *altE*, *altR*, *altJ*, and *altB*) as well as an additional comparison between a secondary *altR*-like gene in *P. syringae* pv. tomato DC3000 (*out_altR* as in figure 1) between the three bacterial strains that were used as input data in DeepBGC.

alt Protein Sequence Similarity[A]	*altI*[B]	*altA*[C]	*altC*[D]	*altE*[E]	*altR*[F]	*altJ*[G]	*altB*[H]	altR Protein Sequence Comparisons[I]	*altR*[J]	*gor*	*altJ*	*altH*	*altD*
BG[K] vs PA[L]	48.2	67.9	38.5	62.5	40.1	27.8	81.3	PTO *out_altR* vs PA *altR*	36.5	NA	NA	NA	NA
PA vs PTO[M]	22.9	69.6	39.2	64.6	35.4	27.2	82.1	PTO *out_altR* vs PTO *altR*	42.9	44.0	25.6	45.6	30.6
PTO vs BG	18.1	74.3	43.8	71.8	43.9	76.6	78.5	PTO *out_altR* vs BG *altR*	47.3	NA	NA	NA	NA

A. Protein multi-sequence alignment similarity calculated by T-Coffee for shared *alt* proteins among the three gene clusters used as inputs for DeepBGC
B. *altI* protein sequence comparisons with output as percentages
C. *altA* protein sequence comparisons with output as percentages
D. *altD* protein sequence comparisons with output as percentages
E. *altE* protein sequence comparisons with output as percentages
F. *altR* protein sequence comparisons with output as percentages
G. *altJ* protein sequence comparisons with output as percentages
H. *altB* protein sequence comparisons with output as percentages
I. Protein multi-sequence alignment similarity calculated by T-Coffee for the PTO out_*altR* protein and the shared *altR* proteins among the three protein clusters used as inputs for deepBGC
H. out_*altR* protein sequence comparisons with output as percentages
K. *Burkholderia gladioli* pv. *gladioli* FDAARGOS_389 (BG)
L. *Pantoea ananatis* PNA 97-1R (PA)
M. *Pseudomonas syringae* pv. tomato DC3000 (PTO)

Table 2.2: Multi-sequence alignment (MSA) comparison of allicin tolerance (*alt*) proteins among shared proteins in the *alt* clusters used as the inputs for training by DeepBGC. Sequences were retrieved from *Burkholderia gladioli* pv. *gladioli* FDAARGOS_389 (BG), *Pantoea ananatis* PNA 97-1R (PA), and *Pseudomonas syringae* pv. tomato DC3000 (PTO). MSA was calculated using T-Coffee in the default settings at https://www.ebi.ac.uk/Tools/msa/tcoffee/. Comparisons were made between shared *alt* proteins (*altI*, *altA*, *altC*, *altE*, *altR*, *altJ*, and *altB*) as well as an additional comparison between a secondary *altR*-like protein in *P. syringae* pv. tomato DC3000 (out_*altR* as in figure 1) between the three bacterial strains that were used as input data in DeepBGC.



Supplementary Figure 2.3. Hierarchical clustering of bacterial strains based on phase-specific growth curve data using Euclidean distance. To provide a comprehensive view of the growth response patterns across different bacterial strains, a hierarchical clustering analysis was conducted based on the Euclidean distance of the growth curves from the experiments. This analysis categorized the bacterial strains into clusters based on their growth response to thiosulfinate exposure. The hierarchical clustering dendrogram revealed distinct clusters, with each branch representing a similarity in growth responses among the strains. A distinct cluster formed by *G. kondonii* (Dm-54), *C. dublinensis* (cro910B3), and *P. nuruki* (TI45-13ar) unique growth response profiles, which is supported by the unexpectedly poor performance of *G. kondonii* (Dm-54), and variability in reponses from both *C. dublinensis* (cro910B3), and *P. nuruki* (TI45-13ar). A second significant cluster includes *E. persicina* (CFBP8795) and *P. ananatis* 97-1R WT showing more-similar growth curves when compared to the remaining strains. These results are

supported by the consistently high performance of the *E. persicina* (CFBP8795) *altC*/*altE* pair. The next group consists of similarly performing strains with *altC*/*altE* pairs from *Pseudomonas sp.* (Root569), *R. aquatilis* (Ra9-2), *Novosphingobium spp.* (Chol11), *P. aryabhattai* (LAD), and *S. maltophilia* (CV_2003_STM1) with *P. ananatis* (PNA 97-1R) placed in an intermediate rating with the previous group. These results are supported by the consistently high, but not as high, performance of *P. ananatis* (PNA 97-1R) when compared to *E. persicina* (CFBP8795), but not as variable as the remaining members of the group. *V. coralliilyticus* (09-121-3), and *B. gladioli* (BCC1802), *P. syringae pv. tomato* (DC3000), and *P. fluorescens* (PS838) are the final group. The GFP strain is positioned near the negative control, reinforcing its minimal growth and low tolerance to thiosulfinates, as expected because they lacked *alt* genes. Overall, the hierarchical clustering analysis provides a comprehensive view of the growth response patterns across different bacterial strains, aligning with the findings from the tolerance experiments and phylogenetic analysis, demonstrating their similar growth profiles and tolerance mechanisms independently of protein sequence content or lineage.

Chapter 4

Insights into host-resistance of *Allium* genotypes against *Pantoea ananatis*

Myers *et al.* To be submitted to *Frontiers in Plant Science*.

Abstract

Onion (*Allium cepa* L.) is a widely cultivated crop that suffers from substantial losses due to *Pantoea ananatis* (PA), a bacterial pathogen responsible for onion center rot (OCR). Severe outbreaks of OCR have been reported globally, leading to significant economic impacts, particularly in onion-producing regions like Georgia, USA. The pathogen's virulence is driven by the chromosomally located HiVir gene cluster, which produces the phytotoxin pantaphos, causing extensive necrosis in infected tissues. Despite its economic importance, *Allium* genotypes with resistance against PA are unknown. In this study, we conducted a comprehensive screening across 982 *Allium* genotypes to evaluate resistance against PA. Only one *A. cepa* genotype, DPLD 19-39, demonstrated a consistent resistant phenotype by exhibiting lower foliar necrosis and bulb rot. Transcriptomic analysis identified that resistance may be associated with enhanced cell wall reinforcement, oxidative stress regulation, and programmed cell death (PCD). Our findings indicate a mechanism for resistance against PA in *A. cepa* and suggest that future efforts should focus on these defense pathways to develop PA-resistant onion cultivars.

Introduction

Onion (*Allium cepa* L.) belongs to the family Amaryllidaceae and is a biennial plant primarily cultivated annually for its edible bulb (Rabinowitch & Brewster, 1989). *Allium cepa* genotypes are highly susceptible to *Pantoea ananatis* (PA), a bacterium that causes onion center rot (OCR). Severe infections of OCR in major onion-producing regions, like Vidalia onion fields in Georgia, have led to significant economic losses, sometimes

amounting to hundreds of thousands to millions of dollars in revenue (Gitaitis & Gay, 1997; Schwartz & Mohan, 2008; Coutinho & Venter, 2009; University of Georgia, 2024, Penn State Extension, 2024). In addition to other members of the *Allium* genus, PA infects a wide range of economically important crops globally. It was first reported with fruitlet rot on pineapple in the Philippines in 1928 (Serrano, 1928), and since then, it has been identified as an epiphyte or endophyte on both dicots and monocots, distributed across regions such as Georgia, Colorado, Michigan, New York, and Pennsylvania in the United States (Wells et al., 1987; Gitaitis & Gay, 1997), as well as internationally on crops like honeydew melon, cantaloupe, onion, sudangrass, eucalyptus, rice, netted melon, maize, and sorghum in countries including Ecuador, Italy, Japan, Argentina, Poland, Brazil, and South Africa (Bruton et al., 1991; Coutinho & Venter, 2009; Kido et al., 2008; Alippi & López, 2010; Cota et al., 2010). PA can be seed-borne and seedling-transmitted, but *Thrips tabaci*-mediated transmission is more common and epidemiologically significant, particularly in regions like the southeastern United States (Gitaitis et al., 2002; Dutta et al., 2014). These thrips species can acquire epiphytic PA populations from various environmental host plants and transmit the pathogen to healthy onion seedlings. PA invades the plant through foliar wounds, leading to water-soaked lesions, blighting, and wilting of the leaves. Foliar colonization can eventually invade the bulb, causing post-harvest losses (Schwartz & Mohan, 2008; Stice et al., 2018). The virulence of PA is attributed to the chromosomally located "HiVir" gene cluster, which encodes the phosphonate phytotoxin pantaphos (Asselin et al., 2018; Polidore et al., 2021). Pantaphos disrupts metabolic processes in the plant, resulting in cell death and necrosis (Asselin et al., 2018; Polidore et al., 2021). Cell death in *Allium* tissues leads to a significant

challenge, however, as the tissues are rich in thiosulfinate compounds, which serve as a natural antimicrobial; the plasmid-borne thiosulfinate tolerance (alt) gene cluster allows PA to thrive and proliferate in disrupted *Allium* tissues by reducing toxic thiol stress (Stice et al., 2020, 2021). There is no known mechanism or *Allium* genotype for host resistance against PA or its pantaphos toxin.

In this study, we conducted a comprehensive genotype screen among various *Allium* genotypes for resistance against PA. In addition, through transcriptome analysis between PA susceptible (Sweet Harvest) vs. resistant genotype (DPLD 19-39), we identified differentially expressed transcripts potentially involved in pathogen resistance mechanisms.

Materials and Methods

Bacterial strain, identification, culturing

PA PNA 97-1 used in this study was isolated from *A. cepa* in 1997 and is a well-characterized pantaphos-producing aggressive bacterial strain (Gitaitis, R. D., & Gay, J. D. 1997). Inoculum was prepared by transferring single colonies from 24 h-old cultures on nutrient agar (NA) medium to nutrient broth (NB). The broth was shaken overnight on a rotary shaker (Thermo Scientific, Gainesville, FL) at 180 rpm. After 12 h of incubation, 1 ml of each bacterial suspension was centrifuged at 5,000 × g (Eppendorf, Westbury, NY) for 2 mins. The supernatant was discarded, and the pellet was re-suspended in PBS. Inoculum concentration was adjusted using a spectrophotometer (Eppendorf, Westbury, NY) to an optical density of 0.3 at 600 nm [≈1 × 10^8 colony forming unit (CFU)/ml].

Phenotypic assessment of *PA* PNA 97-1 on *Allium* genotypes

Foliar pathogenicity and aggressiveness of PA 97-1 were determined under field and controlled greenhouse conditions. Infested onion seeds of *Allium* genotypes were used in the field experiment. This was done to ensure maximum exposure of the pathogen to the *Allium* host, starting from the seed and seedling stages. Infested seeds were generated separately for each *Allium* genotype by exposing them to inoculum (at a concentration stated above) via vacuum infiltration per the manufacturer's instruction for 1 minute. An additional cycle of vacuum infiltration for 1 minute was also conducted. Ten seeds in three replicates for each *Allium* genotype were planted in a row at a 10-cm spacing. These seeds were allowed to germinate and grow to at least the four true-leaf stage. The tallest leaf of each *Allium* genotype was inoculated using a cut-tip method as described previously (Dutta et al.,2014). Briefly, a wound was created by cutting the central leaf (2 cm from the apex) with a sterile pair of scissors. 10 µl drop of a bacterial suspension containing 1×108 CFU/ml was injected at the cut end. One plant at each end of the row was inoculated with sterile water as a negative control for foliar inoculation for each replicate/*Allium* genotype. The rest of the plants were inoculated in between for each plot. A susceptible *Allium* genotype, Sweet Harvest, was used in this experiment. The field was left without management against weeds and thrips to further pathogen spread and disease development. Field plants were assessed for foliar symptoms at least 4 times (1-day post-inoculation, 1-week post-inoculation, 19 days post-inoculation, and one-month post-inoculation).

For the greenhouse studies, seedlings were established in plastic pots (TO plastics, Clearwater, MN) with dimensions of 9 cm × 9 cm × 9 cm (length × breadth ×

height) containing a commercial potting mix (Sta-green, Rome, GA). The seedlings were maintained at 25-28°C and 70-90% relative humidity with a light:dark cycle of 12h:12h. Bacterial strain (PNA 97-1) was maintained on NA plates, and inoculum was generated as described above. Once the primary leaf of each *Allium* genotype reached 9 cm, seedlings were inoculated using a cut-tip method as described previously (Dutta et al., 2014). Seedlings were inoculated with sterile water using the same methodology as above for negative control. A susceptible *Allium* genotype, Sweet Harvest, was also used in this experiment. One experiment used ten replications per genotype, and two independent experiments were conducted with selected genotypes.

Based on the greenhouse experiments, one *Allium* genotype (DPLD 19-39) was selected for growth chamber assessment and was compared with a susceptible genotype (Sweet Harvest). Seedlings for these two genotypes were inoculated at a 4-true-leaf growth stage using a protocol described above. Disease assessments were done according to the protocol stated below.

The pathogenicity and aggressiveness of PA 97-1 were determined based on the lesion length on each *Allium* genotype, measured with the ruler at different assessment periods. The lesion length was recorded weekly for three weeks after foliar inoculation for the field experiment. The lesion size was analyzed using the rating scale for the field evaluations for *Allium fistulosum*, where lesion size was categorized from 0 (no lesion) to 6 (>20.1 cm or dead), and for *Allium cepa*, where the lesion size was categorized from 0 (no lesion) to 10 (>40 cm or dead).

For greenhouse and growth chamber evaluations, the percent lesion length relative to the average length of the leaf for that genotype was calculated at 12 days post-

inoculation. The area under the lesion progress curve (AULPC) was calculated for each genotype and compared between each other and the controls. Analysis of variance (ANOVA) was determined for percent lesion length in R (R version 4.3.0), and Tukey's honest significant difference (HSD) test was used to determine the mean separation for different genotypes.

Phenotypic assessment of bulb infection on selected *Allium* genotypes against PA PNA 97-1 invasion

Bulbs of DPLD 19-39 and Sweet Harvest were harvested after three months of growth under controlled conditions in a growth chamber. The seedlings were maintained at 25-28°C and 70-90% relative humidity with a light:dark cycle of 12h:12h. The outer tunic layer was carefully removed, and the surface was sterilized by wiping the surface with a sterilized paper towel soaked with 70% ethanol. Bulbs were further kept for air drying. After air-drying, scales were carefully removed, and using a sterile inoculation loop, bacterial ooze was scooped, ten-fold serially diluted, and spread-plated onto a semi-selective medium, PA-20 (Goszczynska et al.. 2006). After a period of incubation (7 days), small colonies were enumerated. Representative colonies were also assayed with PA HiVir-specific PCR assay (Shin et al., 2024).

In addition to the evaluation of inter-scale bacterial colonization, remaining healthy appearing bulbs from both genotypes were surface sterilized with 70% ethanol after the removal of tunic layers. Each bulb was placed on a plate containing two layers of paper towel pre-moistened with sterile water. Onion bulbs were inoculated longitudinally at the shoulder with a syringe and a sterile needle containing a volume of 400 µl (1×108 CFU/ml) (Schroeder et al., 2010). Special attention was given to the uniformity of depth of

inoculation into each bulb, which was ascertained by placing a thin rubber stopper in the needle. Following inoculation, bulbs were incubated at 25°C in an aluminum tray. After a week of incubation, bulbs were sliced vertically alongside the inoculation site, and the weight of the whole bulb and symptomatic scales with necrotic lesions (and visual rot) were measured and recorded.

Transcriptome Analysis of DPLD 19-39 vs Sweet Harvest

Host plants were grown under greenhouse conditions, as previously described. Plants were inoculated with PA, with two treatments for each genotype: PA-positive and a PBS buffer (negative control). After 24 hours, foliar tips were excised 1 cm below a visible lesion and immediately frozen in liquid nitrogen for RNA extraction. According to the manufacturer's protocol, total RNA was extracted using the Qiagen RNeasy Plant Mini Kit (Qiagen). RNA sequencing was outsourced to Azenta Life Sciences, where library preparation was performed using the Illumina PolyA selection value package. Sequencing was carried out using Illumina's 2x150 bp paired-end (PE) technology, generating approximately 350 million PE reads per sample (Qiagen citation, Azenta citation).

Differential Gene Expression Analysis

The RNA-seq data was analyzed following a standard bioinformatics pipeline to perform quality control of raw sequencing initially reads using FastQC (Andrews, 2010), which assesses sequencing quality metrics related to per-base quality scores, GC content, sequence contaminants, and adapter presence. The low-quality reads are then removed using Cutadapt (Martin, 2014). Next, aligned reads are processed using the STAR aligner (Dobin et al., 2013) against the *Allium cepa* genome and corresponding gene models (Finkers et al., 2021). Mapped fragments contained in binary bam files were then

processed using featureCounts software (Liao et al., 2014) to create a file corresponding to a matrix of gene expression levels.

The matrix of absolute read counts for each gene is then subjected to a normalized gene expression analysis with statistical significance using the Bioconductor (Gentleman et al., 2004) package DESeq2 (Love et al., 2014). In this step, the read counts data is transformed into a DESeq2 dataset for differential expression analysis, including pre-filtering low-count genes to identify significant results based on adjusted p-values (Benjamini-Hochberg statistical method). Genes with adjusted p-values of less than 0.05 are considered significant in this study. This threshold indicates less than a 5% chance that the observed results are due to random variation alone.

Gene ontology and pathway analysis

Gene ontology annotation and pathway investigation were performed using gene expression data with statistical significance (DGE) subjected to the software clusterProfiler (Xu et al., 2024). Briefly, clusterProfiler internally uses a biological knowledge database, including Gene Ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG), by performing over-representation and gene set enrichment analyses. This analysis allows for the investigation of the association of specific gene lists or sets with biological functions, pathways, and classifications.

This analysis of DGE data determines which functionalities or pathways appear at a higher frequency than expected in the entire reference transcriptome set of *Allium cepa*, making it most suitable for analyzing genes with substantial effects, including the ones related to host-pathogen interaction, plant defense, immune system, and others.

Primer design: Primers were designed using Geneious Prime based on the RNA-seq results, focusing on differentially expressed genes with statistical significance. Five genes that were either highly up- or down-regulated based on transcriptomic analysis were chosen. Specificity was confirmed through BLASTn against the *A. cepa* genome (ASM3076508v1, GCA_030765085.1). The list of genes, their primer sequences, and conditions are listed in supplementary table 3.7.

cDNA synthesis

To validate the differential gene expression results from RNA-seq, quantitative PCR (qPCR) was performed. According to the manufacturer's instructions, the first-strand cDNA was synthesized from total RNA using the Bio-Rad iScript cDNA Synthesis Kit (Bio-Rad Laboratories).

qPCR amplification

The qPCR reactions were carried out using the Bio-Rad iTaq Universal SYBR Green Supermix according to the manufacturer's instructions. Reactions were conducted in triplicate, using methods per the manufacturer's instruction. The qPCR was performed using a 96-well 0.2mL block. Real-time PCR was conducted for five target genes in a QuantStudio 3 (Thermo Fisher Scientific, Waltham, MA) with an amplification program that included an initial denaturation at 95℃ for 10 min, followed by 40 cycles of denaturation at 95℃ for 10 sec, annealing at 55℃ for 30 sec and extension at 72℃ for 30 sec. The amplification program used for the PR1 gene was the same. The cycle threshold (Ct) values thus obtained were converted into relative fold differences of marker genes in treated samples compared with the water-control samples (negative control) and relative to the endogenous control gene (PR1) using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen,

2001; Schmittgen and Livak, 2008). This was done after verifying the stability of the endogenous control genes and that the primer pairs had high, comparable PCR efficiencies (Schmittgen and Livak, 2008). Relative fold changes of target genes were calculated and compared.

Results

Field evaluations confirm that resistance to *Pantoea ananatis* PNA 97-1 is rare in *Allium* genotypes

A panel of 982 *Allium* genotypes were screened against PA PNA 97-1, an aggressive pantaphos-containing strain isolated in Georgia from symptomatic onion. Considerable variations in disease severity were observed across *A. cepa*, *A. cepa* var *cepa*, and *A. fistulosum*. For the *A. cepa* that survived screening with enough replicates for our cut-off, most genotypes were classified as susceptible, accounting for 92.5% of the total genotypes screened. Some genotypes that displayed considerably high foliar disease severity and corresponding high AULPC values include New Mexico Yellow Grano, Linea 139, and Portuguesa Tardia. A smaller proportion, 3.2%, displayed significantly lower disease severity and AULPC values, including the following genotypes: DPLD 19-39, California Red, 1607 Super Sleeper F1, Red Bermuda, Glory, A5718, and Saturn. Only a tiny fraction of the genotypes displayed resistance to the pathogen. Thirty-one *A. fistulosum* genotypes were screened in two sets (Set 1: N=10 genotypes with 5 replicates per genotype; Set 2: N=21 genotypes with seven replicates per genotype) (supplementary table 3.1). The phenotypic screen revealed significant variation in disease severity across *A. fistulosum* genotypes, with the highest disease severity AULPC observed in Japanese Bunching Hikari and Hardy Long White (supplementary table 3.2,

3.3). Genotypes such as Feast, Kannon Hosonegi, and Winter Snow Foot, Aigarshu displayed considerably low disease severity and AULPC. Genotypes like Shounai Nebuka Negi, Yakko, Big Buncher, YatabeYaty:50, and Koshizu Nebuka displayed significantly lower disease severity and AULPC than other genotypes. In the *A. cepa* subsp. *cepa* genotypes, most genotypes displayed moderate to high levels of disease severity. In Set 1 (supplementary table 4), which included ten varieties with six replicates each, the highest AULPC values were observed for Sweet Spanish Los Animas Special, Yellow Ebenezer, and No. 8656 compared with Yellow Grano, which had significantly lower AULPC values. Similarly, in Set 2 (supplementary table 3.5, 3.6), which assessed eleven genotypes with three replicates each, significant differences in AULPC values among the screened genotypes were not observed (2935B, White Portugal, Stuttgarter, Yellow Sweet Spanish Utah Strain, 607 Ebenezer, Calred, Early Crystal 281, Giolla di Rovato da Scttaceto, Early Crystal, White Sweet Spanish, and White Lisbon).

Greenhouse screening identifies consistently resistant *Allium* genotypes against *Pantoea ananatis* PNA 97-1

In two independent greenhouse experiments (GH-1 and GH-2), four *Allium* genotypes, DPLD-19-39, Sweet Harvest, Zhang Qiu Da Cong, and Koshizu Nebuka, were evaluated for foliar disease severity. Since there was a significant interaction between the two experiments, both experiments were analyzed separately. The main effects, genotypes ($P < 0.001$), treatment ($P < 0.001$), and the interaction term (genotype x treatment), were significant ($P < 0.001$) (table 2). In greenhouse experiment 1 (GH-1), significant effects were observed for genotype ($P < 0.001$), treatment ($P < 0.001$), and their interactions (P

< 0.001) (Table 3.2). The comparison of foliar disease severity among four *Allium* genotypes (Table 3.3) showed that Sweet Harvest exhibited the highest mean AULPC value (60.5) compared to DPLD-19-39, Koshizu Nebuka and Zhang Qiu Da Cong. Regarding treatment effects, inoculated plants showed significantly higher AULPC values than the control plants (AULPC; inoculated=61.9 vs. control=17.2). The significant genotype x treatment interactions ($P < 0.01$) underscore the differential responses of the genotypes to inoculation with PA strain PNA 97-1. Specifically, Sweet Harvest had the highest mean AULPC value when inoculated (105.5), while Zhang Qiu Da Cong had the lowest (29.7). DPLD-19-39 (62.2) and Koshizu Nebuka (50.1) exhibited intermediate values but were not significantly different. In greenhouse experiment 2 (GH-2), similar trends were observed, with significant effects for genotype ($P < 0.01$), treatment ($P < 0.01$), and their interactions ($P < 0.01$) (table 2). Again, Sweet Harvest had the highest mean AULPC value (130.7), compared with DPLD-19-39 (68.2), Koshizu Nebuka (36.2), and Zhang Qiu Da Cong (93.6). In terms of treatment effects, inoculated plants again showed significantly higher AULPC values than controls (AULPC; inoculated = 104.9 vs. control = 56.31) (Figure 3.1).

Growth chamber evaluation continues to support higher disease severity in Sweet Harvest compared to DPLD 19-39 following *Pantoea ananatis* PNA 97-1 inoculation

In a controlled growth chamber experiment, we evaluated the foliar disease severity of *A. cepa* genotypes, DPLD 19-39, and Sweet Harvest against PA PNA 97-1. Disease severity was assessed by measuring the AULPC over two weeks, with disease progression recorded every other day. Disease severity and AULPC values were significantly higher for bacterial inoculated Sweet Harvest than DPLD 19-39. Although inoculation with PBS

also resulted in some necrosis in seedlings of both genotypes, AULPC values were significantly lower than the AULPC observed for the inoculated seedlings of both genotypes (Figure 3.2).

Evaluation of bulb rot symptoms in DPLD 19-39 confirms reduced severity against *P. ananatis*

Further, we analyzed the ability of 97-1 to penetrate the onion bulb for both DPLD 19-39 and Sweet Harvest. First, we examined the outer scale of each onion bulb once the foliar lesion experiment concluded. Sweet Harvest inoculated leaves consistently led to the rotting of the attached bulb scale. In contrast, their negative controls were consistently asymptomatic (Figure 3.3A). PA was isolated from the inner scales in 100% of the replicates/samples assayed, which were later confirmed using PA-HiVir specific PCR assay as mentioned above. Isolations made from the inner scales of asymptomatic DPLD 19-39 resulted in bacterial recovery but were not *P. ananatis*, as confirmed by the above PCR assay. To assess if DPLD 19-39 or Sweet Harvest would develop further bulb-rot symptoms, we inoculated PNA 97-1 directly into the bulb and observed that Sweet Harvest consistently developed internal rot symptoms (Figure 3.3B). The Sweet Harvest negative control and DPLD 19-39 treatments did not result in internal bulb rot (Figure 3.3B).

Transcriptome analysis of DPLD 19-39 vs. Sweet Harvest genotypes

To investigate whether DPLD 19-39 and Sweet Harvest genotypes present different transcriptional responses during the disease, the plants inoculated by PNA 97-1 were

subjected to RNA extraction and transcriptome analysis. Three biological replicates per treatment were used. The RNA-seq data generated an average of 20,394,265 million sequenced paired-end reads. After quality analysis and mapping to reference genome, sequenced libraries produced an average of approximately 90% of successfully mapped reads subjected to differential expression analysis with statistical significance (DEASS) genes.

In the analysis of susceptible Sweet Harvest control (HC) compared to Sweet Harvest inoculated (HI), it was found that 998 DEASS genes, with 598 genes down-regulated in HI and 400 genes up-regulated in HI when compared to respective HC controls. Between the top down-regulated genes in HI, there are genes associated with nucleic acid binding, membrane functions, catalytic activities, DNA binding, nucleotide binding, mitochondrial functions, protein binding, enzyme regulation, signal transduction, catabolic processes, energy generation, carbohydrate metabolic processes, and protein metabolic processes. Conversely, among the top-upregulated genes in HI, there are genes associated with DNA binding, nucleotide binding, signal transduction, DNA metabolic processes, energy generation, nuclear functions, transferase activity, structural molecule activity, response to endogenous stimuli, protein and RNA binding, secondary metabolic processes, plasma membrane functions, catalytic and hydrolase activities, nuclear envelope functions, lipid metabolic processes, nuclease activity, and multicellular organism development.

Next, we also performed gene ontology (GO) enrichment analysis of the 998 DEASS genes in HI. In the GO category of biological process, it identified ten terms with significant alteration, including response to the bacterium, defense response to the

bacterium, defense response to other organism, hormone metabolic process, response to water, response to acid chemical, response to water deprivation, lipid homeostasis, phytoesteroid biosynthetic process and brassinosteroid biosynthetic process (Figure 3.4A). In the GO category of cellular component, it identified five terms with significant alterations, including secretory vesicle, cell wall, apoplast, plant-type cell wall and external encapsulating structure (Figure 3.4B). In the GO category of molecular function, it identified ten terms with significant alterations, including UDP−glucosyltransferase activity, glucosyltransferase activity, quercetin 3−O−glucosyltransferase activity, quercetin 7−O−glucosyltransferase activity, "oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen," monooxygenase activity, heme binding, tetrapyrrole binding, "oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen" and "iron ion binding" (Figure 3.4C). The KEGG-based biological pathway enrichment analysis of all DEASS genes also found two significantly altered pathways in HI compared to HC, galactose metabolism, and zeatin biosynthesis (Figure 3.4D).

In the analysis of the resistant DPLD 19-39 control (DC) compared to DPLD-19-39 inoculated (DI), it found 57 DEASS genes, with 27 down-regulated in DI and 30 up-regulated genes in DI compared to respective controls DC. Between the top down-regulated genes in DI, there are genes associated with nucleic acid binding, membrane functions, cytoplasmic activities, catalytic activities, DNA binding, nucleotide binding, mitochondrial functions, enzyme regulation, signal transduction, catabolic processes, energy generation, carbohydrate and protein metabolic processes. Conversely, among

the top-upregulated genes in DI, there are genes associated with cytoplasmic functions, DNA binding, cytoskeleton functions, nucleotide binding, signal transduction, biosynthetic processes, nuclear functions, lipid binding, protein and RNA binding, enzyme regulator activity, plasma membrane functions, nucleic acid binding, catalytic and hydrolase activities, protein metabolic processes, membrane functions, transport activities, translation factor activity with RNA binding, multicellular organism development, binding, plastid functions, and cellular homeostasis. Next, we also performed GO enrichment analysis of the 57 DEASS genes in DI and only found a significant alteration in the molecular function category (Figure 3.5) in the following GO terms: carbohydrate derivative binding, adenyl ribonucleotide binding, ribonucleotide binding, ATP binding, adenyl nucleotide binding. The KEGG-based biological pathway analysis also did not detect any significant alterations considering the DEASS genes of DI samples.

In the analysis of DI compared to HI, the resistant and susceptible genotypes inoculated by PNA 97-1 found 1577 DEASS genes, with 879 downregulated and 698 up-regulated genes in HI compared to the genotype DI. Between the top down-regulated genes in HI, there are genes associated with nucleic acid binding, membrane functions, catalytic activities, DNA binding, nucleotide binding, mitochondrial functions, enzyme regulation, signal transduction, catabolic processes, energy generation, and carbohydrate and protein metabolic processes. Between the top up-regulated genes in HI, there are genes associated with cytoplasmic functions, DNA binding, nucleotide binding, biosynthetic processes, nuclear functions, transferase activity, cellular processes, structural molecule activity, protein and RNA binding, secondary metabolic processes, plasma membrane functions, nucleic acid binding, vacuole functions, catalytic

activity, protein and lipid metabolic processes, membrane functions, transport activities, and multicellular organism development. Next, we also performed gene ontology (GO) enrichment analysis of the 1577 DEASS genes between DI and HI. In the GO category of cellular component, it was identified 4 terms with significant alteration, including apoplast, plant−type cell wall, cell wall, external encapsulating structure (Figure 3.6A). In the GO category of molecular function, it was identified 6 terms with significant alteration, including "tetrapyrrole binding", "heme binding", "protein dimerization activity", "structural constituent of chromatin", "oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen", "iron ion binding" (Figure 3.6 B). The KEGG-based biological pathway analysis did not detect any significant alterations considering the DEASS genes of HI samples compared to DI.

In the last analysis of the transcriptome sequencing data, we compared both control genotypes, HC with DC, and found 1254 DEASS genes, with 597 downregulated genes and 657 up-regulated genes in DC compared to the HC genotype. Between the top down-regulated genes in DC, there are genes associated with nucleic acid binding, membrane functions, catalytic activities, DNA binding, nucleotide binding, mitochondrial functions, enzyme regulation, signal transduction, catabolic processes, energy generation, carbohydrate and protein metabolic processes. Between the top up-regulated genes in DC, there are genes associated with cytoplasmic functions, response to stress, DNA binding, nucleotide binding, nuclear functions, transferase activity, structural molecule activity, lipid binding, protein and RNA binding, catalytic and hydrolase activities, membrane functions, transport activities, and multicellular organism development. Next, we also performed GO enrichment analysis of the 1254 DEASS genes in DC, and it was

only found a significant alteration in the molecular function category (Figure 3.7) in the following GO terms: hydrolase activity, acting on glycosyl bonds, sodium:proton antiporter activity, transmembrane receptor protein serine/threonine kinase activity, monooxygenase activity, glucosyltransferase activity, iron ion binding, heme binding, tetrapyrrole binding, molecular transducer activity, UDP−glucosyltransferase activity. The KEGG-based biological pathway analysis also did not detect any significant alterations considering the DEASS genes of DC samples compared to HC.

We used qPCR to calculate the average cycle threshold (CT) values and validate responses and expression observed in the above transcriptome analysis for the two genotypes under control and inoculated conditions. While the relative expression values show some variability among replicates, the trends of the direction of expression (up- or down-regulated) for five selected genes were exact (Supplementary figure 1). While none of the five of these transcripts are significantly different between the control and inoculated host when considering adjusted p-values, all but one are significantly different when comparing the two host genotypes. Ultimately, the trends seen here match the expected values from the DGE data. All transcriptome pathways are summarized via mapman analysis in figure 3.8.

Discussion

Our extensive phenotypic assessment of 982 *Allium* genotypes, including various species from the *Allium* genus, *A. cepa*, *A. cepa* var. *cepa*, and *A. fistulosum*, indicated substantial phenotypic differences in resistance to PA. Both field and greenhouse evaluations indicated that specific genotypes consistently displayed lower disease severity, as measured by the AULPC; notably, the *A. fistulosum* genotypes "Zhang Qiu Da Cong" and

A. cepa genotype "DPLD 19-39" demonstrated moderate to high levels of resistance, with significantly lower AULPC values compared to highly susceptible genotype (Sweet Harvest). These results align with previous studies indicating that *Allium* species' resistance is considerably variable; for instance, resistance to pathogens such as *Fusarium oxysporum* varies across species like onion (*Allium cepa*) and its wild relatives (*A. roylei*, *A. fistulosum*) (Khrustaleva et al., 2000; Havey et al., 2004). In our study, *A. fistulosum* had the largest number of resistant genotypes (n = 18) against PA, which is still a startlingly small value of the overall tested genotypes screened.

These results were consistent with the various greenhouse and the growth chamber experiments, where *P. ananatis* did not cause systemic infection in DPLD 19-39 but did in Sweet Harvest. The consistency in resistance phenotype in DPLD 19-39 is particularly exciting, as it indicates potential utility of this genotype in future breeding efforts.

The transcriptomic analysis of two *A. cepa* genotypes, DPLD 19-39 (resistant) and Sweet Harvest (susceptible) revealed possible mechanisms responsible for the noted resistance to PA in *Allium cepa*. In the resistant DPLD 19-39, genes associated with cell-wall structural molecule activity and nucleotide binding were significantly upregulated; these plants likely fortify their cell walls in pathogen-infected conditions. This upregulation suggests resistant genotype possesses stronger cell walls, which serve as a barrier against PA. Inoculated resistant plants also showed increased expression of genes linked to membrane functions and enzymatic activity, supporting the hypothesis that resistant plants actively reinforce their cell walls in response to pathogen-derived CWDEs like pectate lyase (Flors et al., 2008; Ponce de León & Montesano, 2013; Wang et al., 2021).

In addition to cell-wall mediated defense, however, there was upregulation of genes involved in oxidoreductase activity and ROS regulation in both control and inoculated conditions. ROS manipulation suggests that resistant plants are better equipped to regulate ROS production, possibly preventing the intended necrosis induced by pantoxin produced by PA. This modulation involves ROS-scavenging enzymes like catalases and peroxidases, which are necessary for signaling defense responses while minimizing oxidative damage (Torres et al., 2006).

Resistant genotype also showed upregulation of genes related to hormonal pathways, particularly those involved in developmental processes and enzyme regulation. This indicates that JA and ET signaling plays a crucial role in coordinating PCD and further reinforcing the cell wall to limit pathogen-induced necrosis, consistent with previous research on defense mechanisms against necrotrophic pathogens (Bolwell & Daudi, 2009; Ali et al., 2024).

The transcriptome data also highlighted the significant upregulation of genes involved in extracellular matrix organization and PCD, a key mechanism for containing pathogen spread by localizing necrotic tissue. This may be particularly important in PA infections, where the pantaphos toxin causes extensive necrosis in susceptible plants (Asselin et al., 2018; Polidore et al., 2021). DPLD 19-39 exhibited upregulation of genes related to secondary metabolic processes, suggesting they may produce phenolic compounds or other metabolites that reinforce the cell wall and potentially neutralize toxins like pantaphos. While the direct role of secondary metabolites in detoxifying pantaphos remains speculative, their contribution to cell wall fortification is likely critical (Flors et al., 2008; Ponce de León & Montesano, 2013). Importantly, we were unable to

detect any significant upregulation of genes related to C-P lyase activity, which breaks down phosphonates, indicating that resistant plants do not degrade pantaphos through this pathway if they are degrading it, or at least the genes responsible for that activity are not annotated. Instead, resistance likely relies more on strengthening physical defenses, such as cell wall fortification, and managing the toxin's downstream effects, rather than directly neutralizing the phosphonate via metabolic breakdown (Coutinho & Venter, 2009; Polidore et al., 2021).

Identifying the significantly different genes involved in cell wall reinforcement, ROS modulation, hormonal signaling, and PCD in resistant onion genotypes provides valuable targets for breeding programs to enhance resistance to PA by incorporating resistant traits from genotypes like "Zhang Qiu Da Cong" and "DPLD 19-39," it is possible to develop new onion genotypes with improved resistance to pantaphos and reduced susceptibility to PA-induced OCR. Future research must further characterize the genetic basis of resistance to PA, whether through immune responses to PA or insensitivity to pantaphos; continued transcriptomic analysis of resistant and susceptible genotypes will be crucial for identifying resistance. Such insights are essential for developing more durable and effective resistance strategies in *Allium* crops, allowing for more precise predictions of disease incidence and contributing to the long-term sustainability of onion production (Khandagale et al., 2022; Prajapati et al., 2023).

Conclusions

In this study, we attempted to find an *A. cepa* genotype that shows promising resistant phenotype (both in foliar and bulb assays) against PA (PNA 97-1). DPLD 19-39 did not show symptoms typical of PA inoculation in foliar or bulb tissue under field, greenhouse,

and growth chamber experiments. The transcriptome response of DPLD 19-39, especially compared to the susceptible genotype Sweet Harvest, indicates that the primary defensive strategy against PA may depend on cell-wall mediated fortification response.

References

1. Gitaitis, R. D., & Gay, J. D. (1997). First report of a leaf blight and bulb rot of onion caused by *Pantoea ananatis* in Georgia. Plant Disease, 81(9), 1096. https://doi.org/10.1094/PDIS.1997.81.9.1096C
2. Schwartz, H. F., & Mohan, S. K. (2008). Compendium of Onion and Garlic Diseases and Pests. American Phytopathological Society (APS Press), 2nd edition.
3. Coutinho, T. A., & Venter, S. N. (2009). *Pantoea ananatis*: An unconventional plant pathogen. Molecular Plant Pathology, 10(3), 325–335. https://doi.org/10.1111/j.1364-3703.2009.00542.x
4. Serrano, F. B. (1928). Bacterial fruitlet brown rot of pineapples in the Philippines. Philippine Journal of Science, 36, 271-305.
5. Wells, J. M., Ceponis, M. J., & Chen, T. A. (1987). Isolation and characterization of strains of Erwinia ananas from honeydew melons. Phytopathology, 77, 511-514.
6. Bruton, B. D., Wells, J. M., Lester, G. E., & Patterson, C. L. (1991). Pathogenicity and characterization of Erwinia ananas causing post-harvest disease of cantaloupe fruit. Plant Disease, 75, 180-183.
7. Kido, K., Adachi, R., Hasegawa, M., & Hikichi, Y. (2008). Internal fruit rot of netted melon caused by *Pantoea ananatis* in Japan. Journal of General Plant Pathology, 74(4), 302-312.

8. Alippi, A. M., & López, A. C. (2010). First report of leaf spot disease of maize caused by *Pantoea ananatis* in Argentina. Plant Disease, 94, 487.
9. Cota, L. V., Costa, R. V., Silva, D. D., Parreira, D. F., Lana, U. G. P., & Casela, C. R. (2010). First report of pathogenicity of *Pantoea ananatis* in sorghum (Sorghum bicolor) in Brazil. Australasian Plant Disease Notes, 5, 120-122.
10. Gitaitis, R. D., Walcott, R., Culpepper, S., Sanders, H., Zolobowska, L., & Langston, D. (2002). Recovery of *Pantoea ananatis*, causal agent of center rot of onion, from weeds and crops in Georgia, USA. Crop Protection, 21, 983-989.
11. Dutta, B., Barman, A. K., Srinivasan, R., Avci, U., Ullman, D. E., Langston, D. B., & Gitaitis, R. D. (2014). Transmission of *Pantoea ananatis* and P. agglomerans, causal agents of center rot of onion (*Allium cepa*), by onion thrips (Thrips tabaci) through feces. Phytopathology, 104(8), 812-819.
12. Stice, S. P., Stumpf, S. D., Gitaitis, R. D., Kvitko, B. H., & Dutta, B. (2018). *Pantoea ananatis* genetic diversity analysis reveals limited genomic diversity as well as accessory genes correlated with onion pathogenicity. Frontiers in Microbiology, 9, 184.
13. Asselin, J. A. E., Bonasera, J. M., & Beer, S. V. (2018). Center rot of onion (*Allium cepa*) caused by *Pantoea ananatis* requires PepM, a predicted phosphonate-related gene. Molecular Plant-Microbe Interactions, 31, 1291-1300.
14. Polidore, A. L. A., Furiassi, L., Hergenrother, P. J., & Metcalf, W. W. (2021). A phosphonate natural product made by *Pantoea ananatis* is necessary and sufficient for the hallmark lesions of onion center rot. mBio, 12.

15. Stice, S. P., Thao, K. K., Khang, C. H., Baltrus, D. A., Dutta, B., & Kvitko, B. H. (2020). Thiosulfinate tolerance is a virulence strategy of an atypical bacterial pathogen of onion. Current Biology, 30(16), 3130-3140.e6.
16. Flors, V., Ton, J., van Doorn, R., Jakab, G., García-Agustín, P., & Mauch-Mani, B. (2008). Interplay between JA, SA and ABA signalling during basal and induced resistance against Pseudomonas syringae and Alternaria brassicicola. Plant Journal, 54(1), 81-92.
17. Ponce de León, I., & Montesano, M. (2013). Activation of defense mechanisms against pathogens in mosses and flowering plants. International Journal of Molecular Sciences, 14(2), 3178-3200.
18. Wang, Y., Li, X., Fan, B., Zhu, C., & Chen, Z. (2021). Regulation and function of defense-related callose deposition in plants. International Journal of Molecular Sciences, 22(5), 2393.
19. Ninkuu, V., Yan, J., Fu, Z., Yang, T., Ziemah, J., Ullrich, M. S., Kuhnert, N., & Zeng, H. (2022). Lignin and its pathway-associated phytoalexins modulate plant defense against fungi. Journal of Fungi, 9(1), 52.
20. Ma, Q.-H. (2024). Lignin biosynthesis and its diversified roles in disease resistance. Genes, 15(3), 295.
21. Bolwell, G. P., & Daudi, A. (2009). Reactive oxygen species in plant–pathogen interactions. In L. Rio & A. Puppo (Eds.), Reactive oxygen species in plant signaling (pp. 153–173). Springer.
22. Torres, M. A., Jones, J. D. G., & Dangl, J. L. (2006). Reactive oxygen species signaling in response to pathogens. Plant Physiology, 141(2), 373–378.

23. Lee, D. H., Lal, N. K., Lin, Z. J. D., Ma, S., Liu, J., Castro, B., Toruño, T., Dinesh-Kumar, S. P., & Coaker, G. (2020). Regulation of reactive oxygen species during plant immunity through phosphorylation and ubiquitination of RBOHD. Nature Communications, 11(1), 1838.
24. Ali, S., Tyagi, A., & Mir, Z. A. (2024). Plant immunity: At the crossroads of pathogen perception and defense response. Plants, 13(11), 1434.
25. Lukan, T., & Coll, A. (2022). Intertwined roles of reactive oxygen species and salicylic acid signaling are crucial for the plant response to biotic stress. International Journal of Molecular Sciences, 23(10), 5568.
26. Khandagale, K., Roylawar, P., Kulkarni, O., Khambalkar, P., Ade, A., Kulkarni, A., Singh, M., & Gawande, S. (2022). Comparative transcriptome analysis of onion in response to infection by Alternaria porri (Ellis) Cifferi. Frontiers in Plant Science, 13, 857306. https://doi.org/10.3389/fpls.2022.857306
27. Prajapati, M. R., Singh, J., Kumar, P., et al. (2023). De novo transcriptome analysis and identification of defensive genes in garlic (*Allium* sativum L.) using high-throughput sequencing. Journal of Genetic Engineering and Biotechnology, 21, 56.
28. Gao, S., Wang, F., Niran, J., Li, N., Yin, Y., et al. (2021). Transcriptome analysis reveals defense-related genes and pathways against Xanthomonas campestris pv. vesicatoris in pepper (Capsicum annuum L.). PLOS ONE, 16(3), e0240279.
29. Kim, M. Y., Han, J. W., Dang, Q. L., Kim, J.-C., Kim, H., & Choi, G. J. (2022). Characterization of Alternaria porri causing onion purple blotch and its antifungal compound magnolol identified from Caryodaphnopsis baviensis. PLOS ONE, 17(1), e0262836.

30. Slavokhotova, A., Korostyleva, T., Shelenkov, A., Pukhalskiy, V., Korottseva, I., Slezina, M., Istomina, E., & Odintsova, T. (2021). Transcriptomic analysis of genes involved in plant defense response to the cucumber green mottle mosaic virus infection. Life, 11(10), 1064. https://doi.org/10.3390/life11101064
31. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B. M., Dettling, M., Dudoit, S., ... & Irizarry, R. A. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology, 5(10), R80.
32. Finkers, R., van Kaauwen, M., Ament, K., Burger-Meijer, K., Egging, R., Huits, H., Kodde, L., Kroon, L., Shigyo, M., Sato, S., et al. (2021). Insights from the first genome assembly of Onion (*Allium cepa*). G3, 11, jkab243. https://doi.org/10.1093/g3journal/jkab243
33. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), 10-12.
34. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data.
35. Dobin, A., et al. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21.
36. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923-930.
37. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.

38. Xu, S., Hu, E., Cai, Y., Xie, Z., Luo, X., Zhan, L., Tang, W., Wang, Q., Liu, B., Wang, R., Xie, W., Wu, T., Xie, L., & Yu, G. (2024). Using clusterProfiler to characterize multiomics data. Nature Protocols.
39. Shin, G. Y., Asselin, J. A., Smith, A., Aegerter, B., Coutinho, T., Zhao, M., Dutta, B., Mazzone, J., Neupane, R., Gugino, B., Hoepting, C., Khanal, M., Malla, S., Nischwitz, C., Sidhu, J., Burke, A. M., Davey, J., Uchanski, M., Derie, M. L., ... Kvitko, B. (n.d.). Plasmids encode and can mobilize onion pathogenicity in Pantoea agglomerans. bioRxiv. https://doi.org/10.1101/2023.03.07.531382
40. Havey, M. J., Galmarini, C. R., Gökçe, A. F., & Henson, C. (2004). QTL affecting soluble carbohydrate concentrations in stored onion bulbs and their association with flavor and health-enhancing attributes. Genome, 47(3), 463–468.
41. Khrustaleva, L. I., & Kik, C. (2000). Introgression of *Allium fistulosum* into *A. cepa* mediated by A. roylei. Theoretical and Applied Genetics, 100, 17–26.
42. Ipek, M., & Simon, P. (2001). Genetic diversity in garlic (*Allium* sativum L.) as assessed by AFLPs and isozymes. American Society of Horticultural Science Annual Conference, 98(3), 22–25

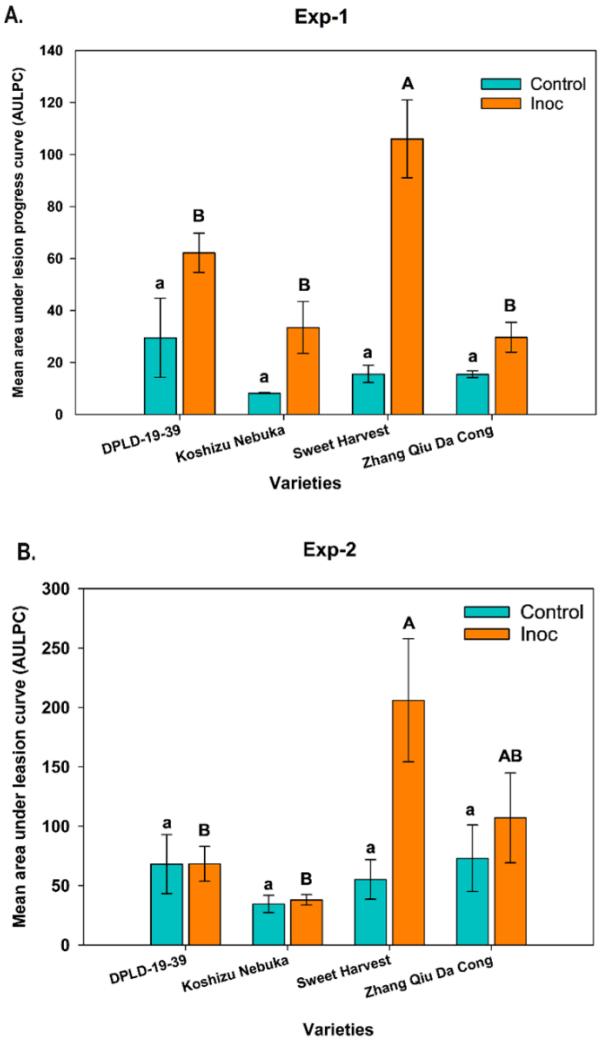


Figure 3.1: Foliar lesion progression in *Allium cepa* and *Allium fistulosum* genotypes inoculated with *Pantoea ananatis* 97-1 under greenhouse conditions.

Four genotypes were tested: *Allium cepa* (DPLD-19-39 and Sweet Harvest) and *Allium fistulosum* (Koshizu Nebuka and Zhang Qiu Da Cong). Plants were inoculated with 10 µL of a bacterial suspension (10^8 CFU/mL) and incubated for 7 days, with phosphate-buffered saline solution -treated plants served as controls. The area under the lesion progress curve (AULPC) was calculated from lesion measurements. Panels A and B: AULPC for experiments 1 and 2, respectively.

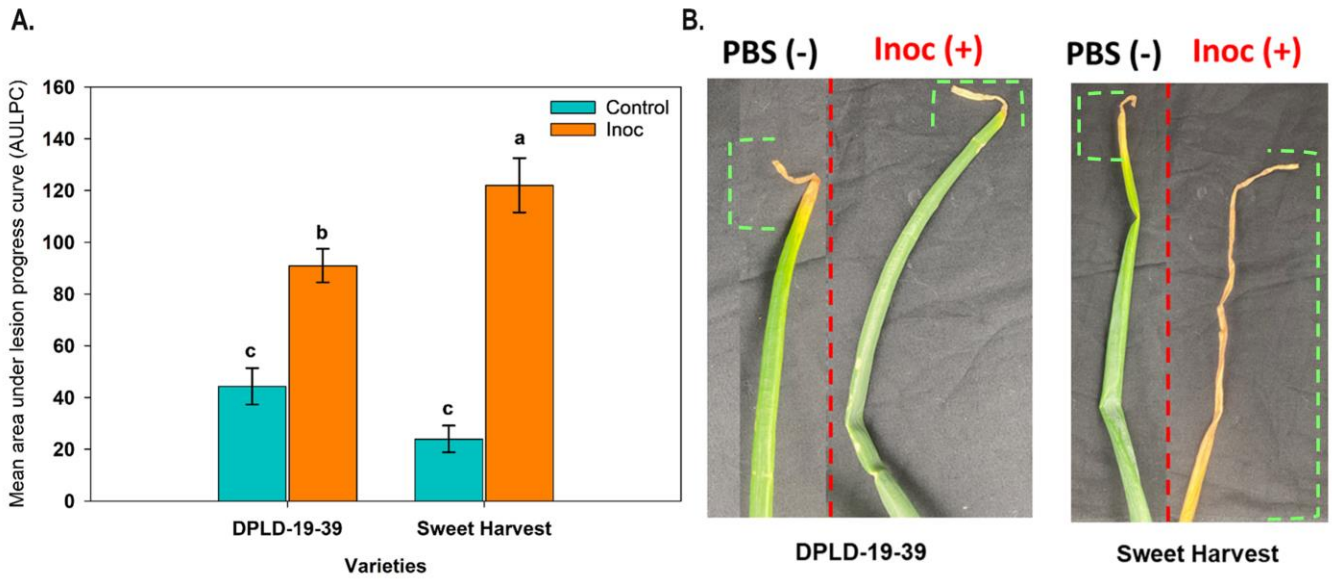


Figure 3.2: Area under lesion progression curve **(**AULPC) for DPLD-19-39 and Sweet Harvest in a growth chamber experiment. Onion seedlings were inoculated with *Pantoea ananatis* strain 97-1 with a bacterial suspension of 10µl containing 10^8 colony-forming units/ml and incubated for 14 days. Seedlings inoculated with phosphate-buffered saline (PBS) solution served as negative controls. The data for lesion length progression over time (every other day for 12 time points post inoculation), and AULPC was determined. Twenty replicates per genotype were used in this experiment. The bars followed by the same letters are not significantly different according to Tukey's honestly significant difference (P<0.05) test. Panel B indicates the foliar lesion in both the negative control (PBS) and inoculated (Inoc (+)) treatments in DPLD-19-39 and Sweet Harvest, respectively. The green brackets in the image indicate the degree or severity of necrotic lesions observed during the experiment.

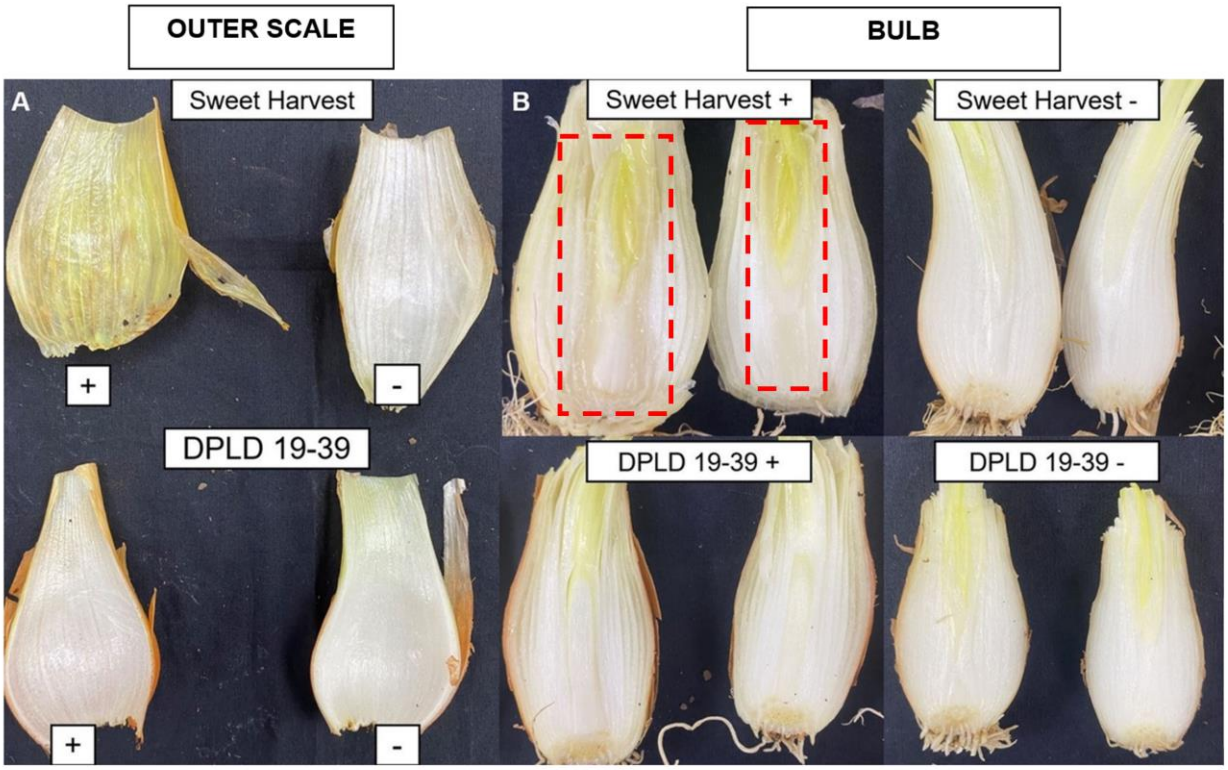


Figure 3.3: Comparison of onion bulb symptoms in *Allium cepa* genotypes "Sweet Harvest" and "DPLD 19-39" following inoculation with *Pantoea ananatis* PNA 97-1. Panel A: Outer scale images of two *Allium cepa* genotypes, "Sweet Harvest" and "DPLD 19-39", taken after 3-months under growth-chamber conditions. Outer scales were removed and assessed for water-soaked bacterial lesion or associated slime. Greasy greenish-yellow indicates water soaked bacterial slime associated lesion. Panel B: Internal bulb rot symptoms in response to PA inoculation to *Allium cepa* genotypes "Sweet Harvest" and "DPLD 19-39. After a week of incubation, bulbs were sliced vertically alongside the inoculation site and the weight of the whole bulb and symptomatic scales with necrotic lesions (and visual rot) was measured and recorded. The signs "+" and "-" denote PA-inoculated and PBS-inoculated onion tissues, respectively. Red dotted lines indicate water-soaked lesions.

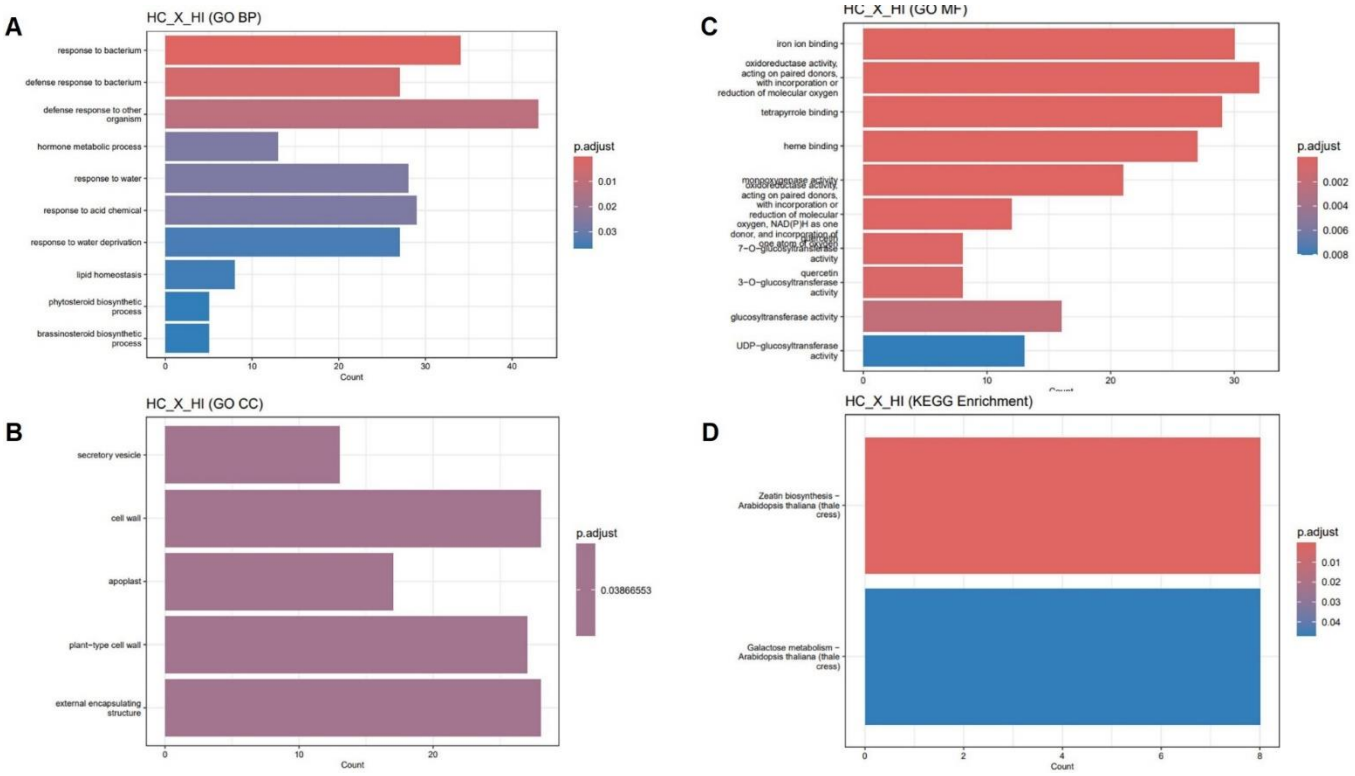


Figure 3.4: Gene Ontology (GO) enrichment analysis for differentially expressed genes between Sweet Harvest control (HC) vs. Sweet Harvest inoculated (HI) plants. Inoculation was done with *Pantoea ananatis* (PNA 97-1). Panel A: GO enrichment analysis for Biological Processes (BP) shows significant upregulation of defense-related processes, including response to bacterium and defense against other organisms, alongside processes related to water and hormone metabolism. Panel B: GO enrichment for Cellular Components (CC) highlights the involvement of genes associated with the cell wall, apoplast, and external encapsulating structures, suggesting structural reinforcement during infection. Panel C: GO enrichment for Molecular Functions (MF) indicates upregulation of genes involved in iron ion binding, oxidoreductase activity, and transferase activities, reflecting oxidative stress responses. Panel D: KEGG pathway enrichment analysis reveals significant activation of zeatin biosynthesis and galactose metabolism, pointing to hormonal regulation and cell wall remodeling efforts.

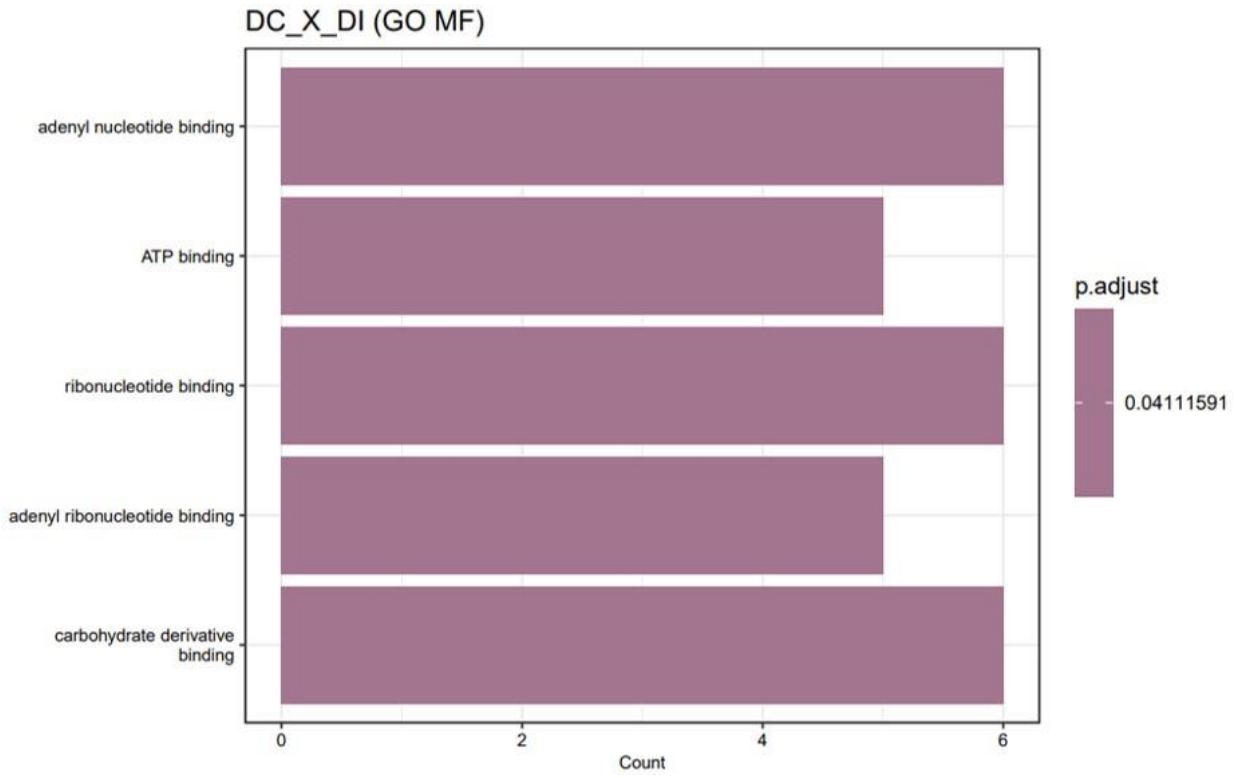


Figure 3.5: Gene Ontology (GO) enrichment analysis for differentially expressed genes between DPLD-19-39 control (DC) and DPLD-19-39 inoculated (DI) plants, with a focus on the molecular functions (MF) that are enriched in response to inoculation with *Pantoea ananatis* (PNA 97-1). GO enrichment analysis for MF reveals significant enrichment in genes associated with binding activities, including adenyl nucleotide binding, ATP binding, ribonucleotide binding, and carbohydrate derivative binding; these molecular functions are essential in energy transfer, nucleotide metabolism, and carbohydrate interactions, all of which are crucial during the plant's response to pathogen stress.

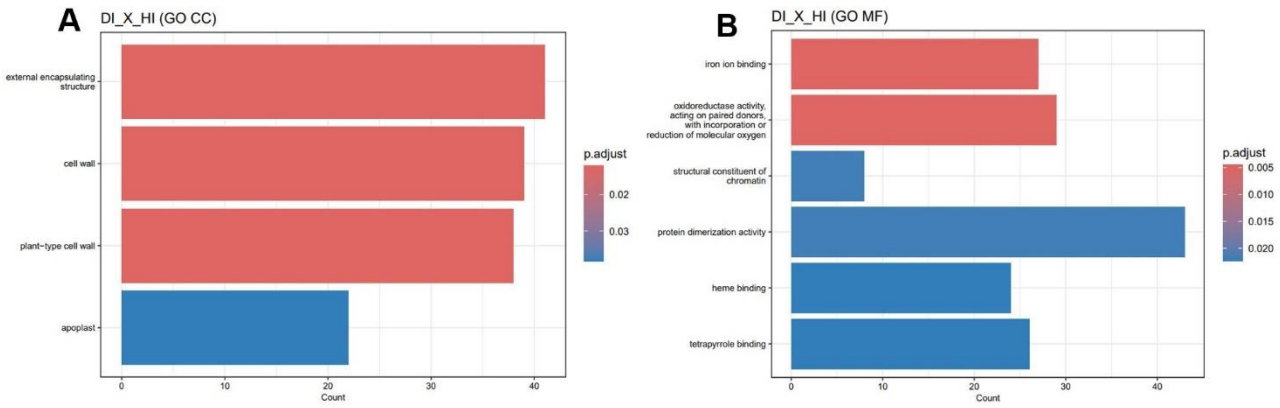


Figure 3.6: Gene Ontology (GO) enrichment analysis for differentially expressed genes between DPLD 19-39 inoculated (DI) and Sweet Harvest inoculated (HI) plants with *Pantoea ananatis* (PNA 97-1), highlighting cellular components and molecular functions that differ between the resistant and susceptible genotypes. Panel A: GO enrichment analysis for Cellular Components (CC) shows significant upregulation in genes associated with the external encapsulating structure, cell wall, plant-type cell wall, and apoplast in DPLD-19-39 compared to Sweet Harvest. This suggests that the resistant genotype, DPLD-19-39, enhances structural defenses during infection, potentially limiting pathogen spread by reinforcing its cell walls and apoplastic barriers. Panel B: GO enrichment analysis for Molecular Functions (MF) indicates enrichment in genes related to iron ion binding, oxidoreductase activity, and chromatin structural constituents, alongside protein dimerization and heme binding.

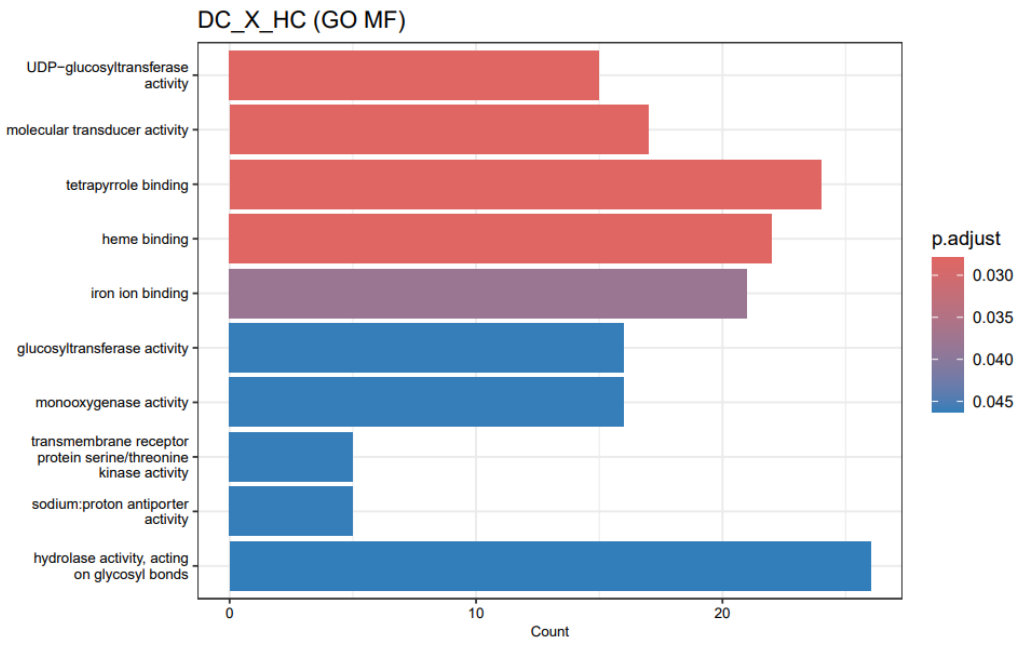


Figure 3.7: Gene Ontology (GO) enrichment analysis for differentially expressed genes between DPLD 19-39 control (DC) and Sweet Harvest control (HC) plants, highlighting molecular functions enriched in the resistant and susceptible genotypes under non-infected conditions. GO enrichment analysis for Molecular Functions (MF) shows significant upregulation in *DPLD 19-39* of genes related to UDP-glucosyltransferase activity, molecular transducer activity, and iron ion binding.

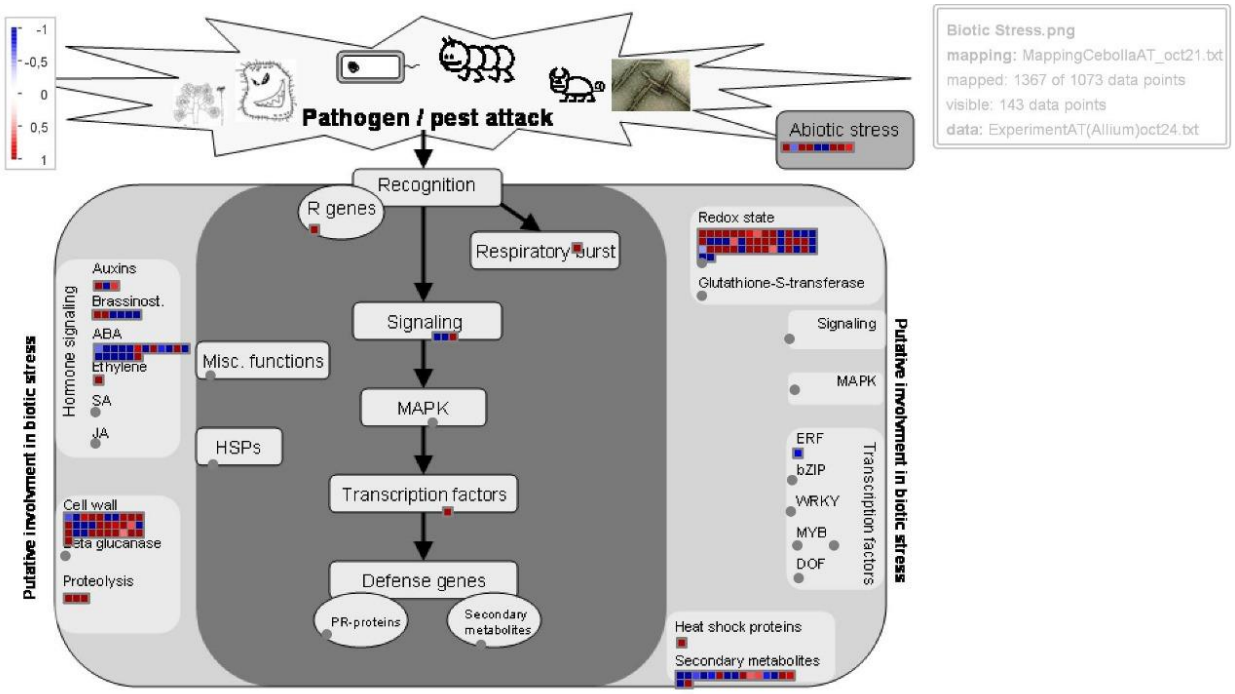


Figure 3.8: Main biotic stress pathways identified in resistant *Allium cepa* using MapMan software (Thimm et al. 2004). This molecular pathway analysis compares differentially expressed genes with statistical significance (DEASS) between DPLD-19-39 (resistant) and Sweet Harvest (susceptible) in response to PNA 97-1 infection. Red boxes indicate up-regulated genes, while blue boxes indicate down-regulated genes. Light gray boxes indicate specific molecular pathways, with the total genes belonging to that specific pathway.

Allium genotypes	AULPC*	Groups**
New Mexico Yellow Grano	149.2	a
Linea 139	147.3	ab
Portuguesa Tardia	144.7	abc
Big Ben	142.0	abc
Swat Selection	140.1	abc
Poona Red	139.7	abc
WHITE IMPERIAL SPANISH	139.3	abc
ELSOMS DOMINATOR	138.1	abc
Malakoff GRU	138.1	abc
Brown Spanish Creamflesh	138.0	abc
Violet De Galmi	137.2	abc
Siohu	136.8	abc
Stuttgart Giant	136.3	abc
Kanda	135.6	abc
Karacabey	135.6	abc
No 9767	135.3	abc
L 365	135.2	abc
Brown Beauty	134.0	abc
Kaba Maikopskaya	134.0	abc
Markovskij Mestnyj	134.0	abc
Odourless	134.0	abc

Pionier	134.0	abc
Reliance GRU B	133.6	abc
Senshyu	132.8	abc
1620 Pedro	132.7	abc
Utah Valencia	132.7	abc
Kille	132.0	abc
Porters Early Globe h 771U	132.0	abc
Exhibition	131.5	abc
Yellow Sweet Spanish Sumida	131.5	abc
UDAIPUR 101	131.3	abc
G 29178	131.3	abc
RED SYNTHETIC	130.5	abc
SAJOVAMOS	130.5	abc
Scanion	130.1	abc
Borrettana	130.0	abc
Rawska	130.0	abc
A7728	129.9	abc
Golden Globe	129.8	abc
Hiberna Vsetatska Ozima	129.2	abc
Rijnsburger Augusta	128.7	abc
Rijnsburger Robusta	128.4	abc
No 10545	128.3	abc
Rossa Di Firenze	128.3	abc

PUKEKOHE LONGKEEPER M+R EARLY	128.1	abc
MAKOVSKI GRU	127.5	abc
Rijnsburger	127.2	abc
Henderson Early Golden	127.1	abc
STAMME 4	126.9	abc
Rijnsburger Wijbo	126.8	abc
Brigham Yellow Globe	126.0	abc
KAIZUKA	126.0	abc
619 Southport White Bunching	125.8	abc
Indian Queen	125.6	abc
PUKEKOHE LONGKEEPER ULTRA	124.8	abc
White Sweet Spanish Utah	124.3	abc
B5718	124.3	abc
Ideal	124.1	abc
NUMEX BR 1	123.5	abc
STURON	123.2	abc
Gelbe Wiener	123.1	abc
Emerald Isle	123.0	abc
YALOVA 1	122.3	abc
AGRIFOUND ROSE GRU	121.9	abc
Bianca Pompei Italy	121.9	abc
Early Yellow Globe	121.4	abc
Sweet Spanish Valencia	121.3	abc

UDAIPUR 102	121.1	abc
Imai Yellow	120.0	abc
CANDY F1	119.0	abc
Perfecto Blanco	118.7	abc
Yellow Multiplier	117.2	abc
Rijnsburger Envee	116.8	abc
PATNA RED	116.6	abc
Texas EARLY GRANO 502 GRU	116.6	abc
Wellving 76522 71	116.5	abc
Besszonovskij	116.3	abc
Q75	116.0	abc
Excel 986	115.9	abc
Rossvale	115.8	abc
Ptujska	115.3	abc
HUNTER RIVER WHITE	115.2	abc
Blanca Grande	114.9	abc
DE LA REINE	114.9	abc
No 8362	114.6	abc
ZWIJNDRECHTSE	114.6	abc
UDAIPUR 103	114.3	abc
No 9491	114.0	abc
Japanese Seketcan	113.9	abc
TORRENS WHITE	113.9	abc

Early Yellow Grano Tex 502	113.7	abc
ASHTON	113.0	abc
617 Southport White Globe Onion	112.3	abc
CEBOLE DE BARCELOS	112.3	abc
L T Medicina	112.3	abc
SENSHUU KOUDAKA	112.3	abc
Dorata Di Palma	111.4	abc
Japanese semi globe yellow	111.2	abc
DOWNING’S YELLOW GLOBE TRAPP STRAIN	110.8	abc
KISHUU HIRAGATA KI	110.8	abc
Ebenezer	110.3	abc
No Kp 13	109.8	abc
WHITE EBENEZER	109.5	abc
Z025	108.7	abc
620 Yellow Sweet Spanish	107.9	abc
Porters Early Globe H 771N	107.4	abc
Caledon Globe	106.6	abc
Early Yellow Sweet Spanish	106.3	abc
Presto	106.0	abc
White Sweet Spanish	104.9	abc
Shamrock	104.5	abc
Rio Grande	103.9	abc

SerraNA	103.9	abc
No Kp S	103.8	abc
No Known Plant ID Mol	103.6	abc
No 259	103.1	abc
No 8543	103.0	abc
White Creole	101.8	abc
Extra Early Kaizuka	101.5	abc
GOLDRUSH YELLOWGLOBE	101.5	abc
Onion Red 2	101.3	abc
376R Yet	101.0	abc
Morada de Amposta	100.8	abc
N 53	100.7	abc
Borrettana_2	100.2	abc
Portuguesa Amarilla Tardia	99.8	abc
Lord Howe Island	97.8	abc
LOWSHAN	97.8	abc
GIANT ROCCA BROWN	97.4	abc
Vertus	96.4	abc
White Creole PRR PVP	96.4	abc
Moravanka	96.3	abc
Pungent	95.9	abc
Yalova 9	95.3	abc
Giant Zittau	95.3	abc

TREBONS	94.7	abc
Kilsurski	93.7	abc
Caldera 1028	92.7	abc
Radar	92.7	abc
Imperial 48	91.3	abc
Hybrid Elite	90.5	abc
HYDURO F1	90.5	abc
Malakoff France	90.5	abc
Early Texas Yellow Grano	89.4	abc
Senshyu 234	89.2	abc
No K1417	89.0	abc
Pios	88.9	abc
616 Southport Red Globe Onion	88.1	abc
DOWNY MILDEW RES SELECTION	87.3	abc
Red Wethersfield	86.7	abc
LIASKOVSKI 58	86.6	abc
Arpadzik	85.9	abc
Shiraz A	85.6	abc
TAK NO 747	84.4	abc
White Portugal	83.3	abc
Obrovska Zluta Germany	83.1	abc
HYGRO F1	83.0	abc
lowa Yellow Globe 51	82.8	abc

No 9526	81.8	abc
Jaune Paille Des Vertus	81.8	abc
G 29180	81.6	abc
No 366	80.5	abc
Cebola Valencia	79.5	abc
Odourless Green Leaf	79.3	abc
No Known Plant ID Aus	79.2	abc
M8155A	78.2	abc
Hunter River Brown	77.3	abc
No K93	76.5	abc
White Sweet Spanish Valencia	76.0	abc
Yellow Sweet Spanish Utah Jumbo	76.0	abc
Colorado De Amposta	74.7	abc
Cardinal	74.6	abc
No 6656	74.6	abc
PYRAMID GRU	72.8	abc
No 6819	72.0	abc
Fruhi Blassrote	71.1	abc
B 12132 B	70.3	bc
ASIMER ADVANCE	70.2	bc
Siohu PBR 3	66.9	bc
303 R	66.8	bc
Sogan	65.9	bc

Yellow Sweet Spanish UT Jumbo	65.5	bc
California Red	65.3	c
1607 Super Sleeper F1	61.8	c
Red Bermuda	59.7	c
Glory	58.4	c
A5718	52.8	c
Saturn	65.0	c

*Data for lesion progression over time were taken and the area under lesion progress curve (AULPC) was determined. Seven replicates per genotypes were used in this experiment. AULPC was calculated from the lesion area of the foliar necrosis, which was recorded with a measuring scale at 1-, 7-, 19-, and 32-days post-inoculation. Day 1 all lesions were effectively 0 in length.

**Mean AULPC with same letters are not significantly different according to Tukey's honest significant difference (P<0.05) test.

Supplementary table 3.1. Foliar disease severity of *Allium cepa* genotypes under field conditions.

Allium fistulosum genotype	Mean AULPC*	Group**
Koshizu Nebuka	36.6	b
Feast	42.8	ab
Kannon Hosonegi	53.8	ab
Winter Snow Foot	57.5	ab
Aigarsyu	72.9	ab
Matsumoto ippon Futo	74.5	ab
Crystal White Wax L303	77.2	ab
JAPANESE BUNCHING	78.3	ab
Japanese Bunching Asage	85.0	ab
JAPANESE BUNCHING HIKARI	100.6	a

*Data for lesion progression over time were taken and the area under lesion progress curve (AULPC) was determined. Five replicates per genotypes were used in this experiment. AULPC was calculated from the lesion area of the foliar necrosis, which was recorded with a measuring scale at 7 days post-inoculation.

**Mean AULPC with same letters are not significantly different according to Tukey's honest significant difference ($P<0.05$) test.

Supplementary table 3.2: Foliar disease severity of *Allium fistulosum* genotypes under greenhouse conditions [Set-1 (10 genotypes with 5 replicates data)].

Allium fistulosum genotypes	Mean AULPC*	Group**
HARDY LONG WHITE	120.0	a
Wakamidori	114.7	ab
Q9	99.2	abc
Kuronobori	89.7	abcd
Matsumoto	88.5	abcde
Koshizu Nebuka	76.7	abcdef
Shiobara-bansei	71.1	bcdefg
Ciboule C8379	69.0	cdefg
Kujyo-futo	60.8	cdefg
Prolific Twin	60.8	cdefg
Miya Negi	60.1	cdefg
lwatsuki	52.8	cdefg
Zhang Qiu Da Cong	52.1	defg
Prolific Buncher	51.7	defg
Asagikei-kujyo	47.3	defg
Kincho Long White	47.2	defg
VIR3139	46.3	efg
Shounai Nebuka Negi	31.5	fg

Yakko	31.5	fg
Big Buncher	27.1	g
Yatabe Yaty:50	25.9	g

*Data for lesion progression over time were taken and the area under lesion progress curve (AULPC) was determined. Seven replicates per genotypes were used in this experiment. AULPC was calculated from the lesion area of the foliar necrosis which was recorded with measuring scale every other day for 38 days post inoculation.

**Mean AULPC with same letters are not significantly different according to Tukey's honest significant difference (P<0.05) test.

Supplementary table 3.3: Foliar disease severity of *Allium fistulosum* genotypes under greenhouse conditions [Set-2 (21 genotypes with 7 replicates data)]

***Allium cepa* subsp. *cepa* genotypes**	**Mean AULPC***	**Group****
Sweet Spanish Los Animas Special	130.8	a
Yellow Ebenezer	127.3	a
No 8656	123.0	a
White Sweet Spanish Jumbo	110.8	ab
Grano 502	101.4	ab
White Sweet Spanish California	101.4	ab
2935A	96.7	ab
Early Texas White Grano	86.8	ab
NM899	84.6	ab
Yellow Grano	64.6	b

*Data for lesion progression over time were taken and the area under lesion progress curve (AULPC) was determined. Six replicates per genotypes were used in this experiment. AULPC was calculated from the lesion area of the foliar necrosis which was recorded with measuring scale at 7 days post inoculation.

**Mean AULPC with same letters are not significantly different according to Tukey's honest significant difference ($P<0.05$) test.

Supplementary table 3.4: Foliar disease severity of *Allium cepa* var. *cepa* genotypes under greenhouse conditions [Set-1 (10 genotypes with 6 replicates data)].

Allium cepa subsp. *cepa* **genotypes**	**Mean AULPC***	**Group****
2935B	118.0	a
White Portugal	107.7	a
Stuttgarter	97.2	a
Yss Utah Str	97.2	a
607 Ebeniezer	92.8	a
Cal Red	80.5	a
Early Crystal 281	73.8	a
Cipolla di Rovato	65.8	a
Early Crystal	65.7	a
White Sweet Spanish	59.3	a
White Lisbon	54.5	a

*Data for lesion progression over time were taken and the area under lesion progress curve (AULPC) was determined. Three replicates per genotypes were used in this experiment. AULPC was calculated from the lesion area of the foliar necrosis which was recorded with measuring scale at 7 days post inoculation.

**Mean AULPC with same letters are not significantly different according to Tukey's honest significant difference ($P<0.05$) test.

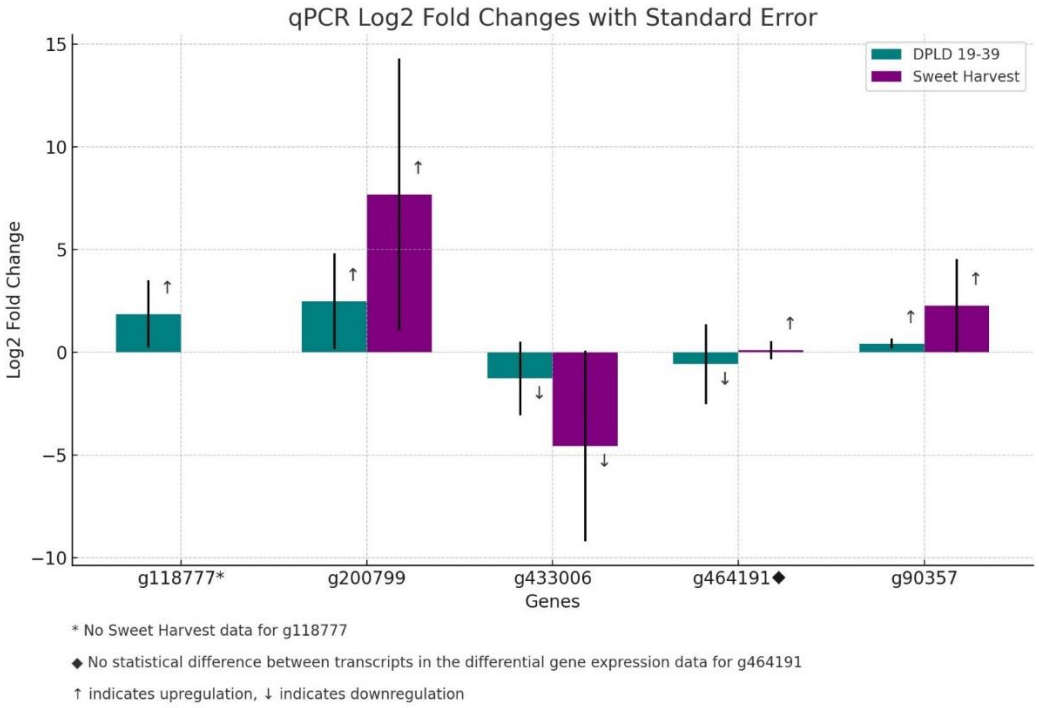
Supplementary table 3.5: Foliar disease severity of *Allium cepa* var. *cepa* genotypes under greenhouse conditions [Set-2 (11 genotypes with 3 replicates data)]

Sequence Name	Sequence	Denaturation Temperature	Annealing Temperature	Extension Temperature
g118777__1,103 R	GCAGATCTTTCAGGCGCATG	95°C	57°C	72°C
g118777__456 F	TGCTTGCGGATACATGGGTT	95°C	55°C	72°C
g200799__190 F	GAGAACGTGATTCGCGATGC	95°C	57°C	72°C
g200799__304 R	CCAAAGGAGCCCTGACACTT	95°C	57°C	72°C
g433006__1,121 R	ATACAACGCAAACCCAACGC	95°C	55°C	72°C
g433006__145 F	GGCAAGCCCATGTCTCTCTT	95°C	57°C	72°C
g464191__1,170 R	TTTCTCACCATCCGGCGTAG	95°C	57°C	72°C
g464191__612 F	GCCTAGTTTCCCAGCCAGTT	95°C	57°C	72°C
g90357__844 F	CAAGCTGATATGGCCGTTGC	95°C	57°C	72°C
g90357__2,779 R	TGCTTCCTCCATTGCGTGAT	95°C	55°C	72°C

PR1_Forwa rd	TTCTTCCCTCGAA AGCTCAA	95°C	53°C	72°C
PR1_Rever se	CGCTACCCCAGG CTAAGTTT	95°C	57°C	72°C

*This table lists the primer sequences and corresponding PCR conditions used to amplify specific genes in *Allium* genotypes. Each entry includes the sequence name, primer sequence, and the denaturation, annealing, and extension temperatures for the PCR reactions. All reactions were conducted with a denaturation temperature of 95°C, varying annealing temperatures between 53°C and 57°C, and an extension temperature of 72°C.

Supplementary table 3.6: Primer sequences and PCR conditions for qPCR analysis*.



Supplementary figure 3.1: RNA-seq validated log2 fold changes for selected genes in DPLD 19-39 and Sweet Harvest genotypes, presented with standard error bars. Genes labeled on the x-axis correspond to differentially expressed genes validated through RNA-seq analysis. Green bars represent fold changes in DPLD 19-39, while purple bars represent those in Sweet Harvest. For gene g118777, no expression data was available for Sweet Harvest (*); this is due to the incredibly small number of reads and is not an unexpected result. Genes g200799, g433006, g464191, and g90357 display varying degrees of upregulation (↑) or downregulation (↓) across the two genotypes. All genes are significantly different when compared across genotypes except for g464191. These results highlight distinct transcriptional responses between the two genotypes responding to PA strain 97-1 infection, and all match the predicted log2 change comparisons predicted by the RNA seq analysis.

CHAPTER 5

CONCLUSIONS

The research outlined in this dissertation aimed to utilize a wide range of informatics approaches to explore the virulence factors of PA and the potential host resistance to center rot in *Allium A. cepa*.

Our first objective aimed to identify potential virulence factors in PA strains contributing to pathogenicity in non-traditional *Allium* species, such as *Allium porrum* (leek) and *A. fistulosum x A. cepa* (bunching onion hybrid). We utilized pan-Ggenomic genome-wide association studies (GWAS) and gene-pair coincidence analyses to explore genetic variations and their correlation with pathogenic phenotypes. This approach emphasized the utility of combining phenotype-dependent and phenotype-independent methodologies better to understand bacterial virulence mechanisms in diverse plant pathosystems, as previously demonstrated in similar studies, by leveraging the constant genomic filtering inherent to bacterial genomics (De Maayer et al., 2014; Agarwal et al., 2021). We observed significant variability in the pathogenicity of PA across *A. porrum* and *A. fistulosum x A. cepa;* notably, strains pathogenic on one host species were not necessarily pathogenic on another, suggesting the presence of host-specific virulence factors. Our pan-genomic analysis identified a core genome of approximately 2,914 genes, consistent with findings from De Maayer et al. (2014) and Agarwal et al. (2021). The HiVir cluster, was responsible for virulence in bulb onions (Asselin et al., 2018; Polidore et al., 2021), was also found to play a role in virulence across different *Allium* species, suggesting a conserved pathogenic mechanism. Additionally, our GWAS

analysis uncovered genes associated with thiosulfinate tolerance, specifically the *alt* gene cluster, which was correlated with infection in *A. fistulosum* x *A. cepa* and higher aggressiveness. This finding is supported by Stice et al. (2020), who demonstrated that the *alt* cluster is critical for thiosulfinate resistance, a key factor in allowing PA to colonize necrotic onion tissue. Interestingly, some strains that appeared pathogenic on *A. porrum* lacked the HiVir and *alt* clusters, indicating alternative virulence pathways that warrant further investigation. The co-occurrence and dissociation of virulence-related genes, as revealed by gene-pair association analyses, further emphasize the evolutionary pressures that shape the bacterial genome for host-specific pathogenicity and should be used in tandem with GWAS analysis to provide a deeper level of insight into the genomics at hand (Whelan et al., 2020). The results of this objective contribute significantly to understanding the genetic basis of PA virulence in non-traditional *Allium* species, laying the groundwork for developing targeted management strategies and leveraging genetic diversity for disease-resistance breeding.

Data mining genomic variants in PA on *Allium* crops is difficult, even when using well-reviewed and accepted methods. The allicin tolerance (*alt*) gene cluster is a particularly irritating case study where traditional sequence-based, and potentially even gene-pair and GWAS methodologies, may fail you due to the inherent biology of the gene cluster itself. The second objective of this research was to develop an autonomous method for identifying *alt* gene clusters in diverse bacterial genera, leveraging deep learning techniques. The *alt* gene cluster has been manually detected in several bacterial species, including *P. ananatis* (Stice et al., 2020), *Pseudomonas syringae*, and *Burkholderia gladioli* (Paudel et al., 2024). Building on the DeepBGC platform, we trained

a BiLSTM RNN to identify *alt* gene clusters based on gene proximity and protein domain details (Pfam domains). This approach enabled us to mine large bacterial datasets for *alt*-like clusters effectively. Our model identified over 12,000 *alt*-like gene clusters across 238,000 bacterial genomes, with 47 representative clusters selected post-further filtering and 15 of those validated experimentally. Interestingly, and yet another discrepancy with the *alt* cluster, the synteny of *alt* genes was not always predictive of level of tolerance, suggesting the involvement of additional factors like protein-protein interactions or some level of protein specialization. Naturally, it would be unsurprising for *Allium* pathogens to have specialized *alt* proteins, but in-depth comparisons between the datamined groups will be necessary to describe those differences.

Advanced protein structure analyses using I-TASSER revealed striking structural similarities between *alt*-like gene products across bacterial genera, even in the absence of sequence conservation (Yang et al., 2015). This indicates that *alt* gene functionality is driven by conserved motifs rather than strict sequence similarity, which is ultimately unsurprising; the 3D structure of a protein is not the end-all-be-all of protein function. However, the 3D structure of a protein should logically be similar to proteins with the same function, though ultimately, this was unhelpful for differentiating *alt* clusters.

AI-Bind analysis further demonstrated the potential of machine learning to predict the binding affinity of *altR* proteins to sulfur compounds, allowing us to, with some degree, cluster functional *alt* clusters based on the phenotypic data (Chatterjee, et al. 2023). These findings are helpful for exploring potential downstream classification techniques and, by extension, improve our ability to identify and validate complex gene clusters in bacterial genomes autonomously. This work contributes a robust framework for

accelerating gene discovery and understanding bacterial virulence in plant pathosystems, especially for abnormal, novel gene clusters.

The third objective of this study was to identify an *A. cepa* genotype with resistance to PA. Through comprehensive phenotypic and transcriptomic analyses, we identified the genotype DPLD 19-39 as exhibiting significantly reduced disease symptoms compared to the susceptible genotype, Sweet Harvest. Across multiple experimental conditions: field, greenhouse, and growth chamber trials, DPLD 19-39 consistently demonstrated lower lesion severity and delayed pathogen progression to the PA strain PNA 97-1.

Transcriptomic profiling of DPLD 19-39 revealed that genes involved in cell wall reinforcement, nucleotide binding, and reactive oxygen species (ROS) regulation were significantly upregulated in response to pathogen infection (Torres et al., 2006; Flors et al., 2008; Ponce de León & Montesano, 2013; Wang et al., 2021). These findings suggest that resistant plants strengthen their cell walls to limit pathogen entry and actively modulate ROS production to mitigate the stress caused by pantaphos, the primary phytotoxin produced by PA. The upregulation of genes associated with hormonal pathways, such as those for jasmonic acid (JA) and ethylene (ET), further supports the potential role of programmed cell death (PCD) and cell wall fortification in conferring resistance to PA, though these mechanisms need validation (Bolwell & Daudi, 2009; Ali et al., 2024). The lack of upregulation of genes annotated to be involved in any C-P lyase activity in DPLD 19-39 suggests that its resistance mechanism does not rely on the direct degradation of pantaphos but on managing its downstream effects through enhanced physical and biochemical defenses (Coutinho & Venter, 2009; Polidore et al., 2021).

Future research will have to further characterize the genetic basis of resistance in DPLD 19-39 and other genotypes, such as *A. fistulosum* (*genotype:* Zhang Qiu Da Cong). DPLD 19-39 holds significant potential as a genetic resource for breeding onion varieties with enhanced resistance to PA. By integrating the resistant traits of genotypes like DPLD 19-39 and Zhang Qiu Da Cong, breeding programs could improve crop resilience, reduce economic losses from OCR, and promote the long-term sustainability of onion production.

References

1. De Maayer, P., Chan, W. Y., Rubagotti, E., Venter, S. N., Toth, I. K., Birch, P. R., et al. (2014). Analysis of the Pantoea ananatis pan-genome reveals factors underlying its ability to colonize and interact with plant, insect, and vertebrate hosts. BMC Genomics, 15, 404.
2. Agarwal, G., Choudhary, D., Stice, S. P., Myers, B. K., Gitaitis, R. D., Venter, S. N., Kvitko, B. H., & Dutta, B. (2021). Pan-genome-wide analysis of Pantoea ananatis identified genes linked to pathogenicity in onion. Frontiers in Microbiology, 12, 684756.
3. Asselin, J. A. E., Bonasera, J. M., & Beer, S. V. (2018). Center rot of onion (Allium cepa) caused by Pantoea ananatis requires PepM, a predicted phosphonate-related gene. Molecular Plant-Microbe Interactions, 31, 1291–1300.
4. Polidore, A. L. A., Furiassi, L., Hergenrother, P. J., & Metcalf, W. W. (2021). A phosphonate natural product made by Pantoea ananatis is necessary and sufficient for the hallmark lesions of onion center rot. mBio, 12, e03402–20.

5. Stice, S. P., Gitaitis, R. D., Kvitko, B., & Dutta, B. (2021). The distribution of onion virulence gene clusters among Pantoea spp. Frontiers in Plant Science, 12, 643787.
6. Whelan, F. J., Rusilowicz, M., & McInerney, J. O. (2020). Coinfinder: Detecting significant associations and dissociations in pangenomes. Microbial Genomics, 6(3), e000338.
7. Paudel, S., Zhao, M., Stice, S. P., Dutta, B., & Kvitko, B. H. (2024). Thiosulfinate tolerance gene clusters are common features of Burkholderia onion pathogens. Molecular Plant-Microbe Interactions, 37(3), 298–312.
8. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER suite: Protein structure and function prediction. Nature Methods, 12(1), 7–8.
9. Chatterjee, A., Walters, R., Shafi, Z., et al. (2023). Improving the generalizability of protein-ligand binding predictions with AI-Bind. Nature Communications, 14, 1989.
10. Torres, M. A., Jones, J. D. G., & Dangl, J. L. (2006). Reactive oxygen species signaling in response to pathogens. Plant Physiology, 141(2), 373–378.
11. Flors, V., Ton, J., van Doorn, R., Jakab, G., García-Agustín, P., & Mauch-Mani, B. (2008). Interplay between JA, SA, and ABA signalling during basal and induced resistance against Pseudomonas syringae and Alternaria brassicicola. Plant Journal, 54(1), 81–92.

12. Ponce de León, I., & Montesano, M. (2013). Activation of defense mechanisms against pathogens in mosses and flowering plants. International Journal of Molecular Sciences, 14(2), 3178–3200.
13. Wang, Y., Li, X., Fan, B., Zhu, C., & Chen, Z. (2021). Regulation and function of defense-related callose deposition in plants. International Journal of Molecular Sciences, 22(5), 2393.
14. Bolwell, G. P., & Daudi, A. (2009). Reactive oxygen species in plant–pathogen interactions. In L. Rio & A. Puppo (Eds.), Reactive oxygen species in plant signaling (pp. 153–173). Springer.
15. Ali, S., Tyagi, A., & Mir, Z. A. (2024). Plant immunity: At the crossroads of pathogen perception and defense response. Plants, 13(11), 1434.
16. Coutinho, T. A., & Venter, S. N. (2009). Pantoea ananatis: An unconventional plant pathogen. Molecular Plant Pathology, 10(3), 325–335.