IDENTIFICATION OF VIRULENCE FACTORS IN THE PHYTOPATHOGEN *P. SYRINGAE* BY MOLECULAR EVOLUTION

Emily Jane Fisher

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology.

Chapel Hill
2008

Approved by

Professor Jeff Dangl

Professor Corbin Jones

Professor Jason Reed

Professor Matthew Wolfgang

Professor Janne Cannon

ABSTRACT

Emily Fisher

IDENTIFICATION OF VIRULENCE FACTORS IN THE PHYTOPATHOGEN *P. SYRINGAE* BY MOLECULAR EVOLUTION

(Under the direction of Jeff Dangl)

The interaction between pathogen and host is often likened to an evolutionary arms race in which the pathogen evolves a novel mechanism to infect the host and the host, in turn, evolves better defenses against the pathogen. This process repeats throughout the history of the pathogen-host interaction as new virulence and defense mechanisms arise. My work used this evolutionary interplay between pathogen and host to identify novel virulence genes in the model bacterial pathogen *Pseudomonas syringae*. My hypothesis was that genes in the *P. syringae* genome that determine pathogen virulence may be rapidly evolving due to selective pressure from the plant host immune system. Therefore, rapidly-evolving genes in *P. syringae* may be required for virulence. I identified genes purported to be rapidly-evolving in the genome of *P. sryingae* by two methods: codon volatility and comparison of homologs. The codon volatility method identified 31 candidate virulence genes based on the proportion of volatile codons in each gene. Three of these 31 have previously-predicted roles in virulence, though only one has been demonstrated to function in pathogen growth on plants. For the second method, I compared the

genomes of three isolates of *P. syringae* and identified 20 genes with high numbers of non-synonymous (dN) mutations compared to silent, synonymous mutations (dS). High dN/dS ratios are consistent with positive selection on these genes and we therefore anticipated that these genes would function in bacterial virulence. Mutational analysis of 9 of these candidates revealed that two candidate genes had newly-discovered roles in virulence. Two additional genes were previously-identified to function in virulence, and two others are required for growth in minimal medium. Including only those genes with direct functions in bacterial virulence during infection, my candidate list based on dN/dS ratios gives a 4-fold enrichment in virulence genes compared to the genome as a whole. We therefore conclude that rapid evolution rate as evaluated by dN/dS is a predictor of virulence function and propose that this method could be useful in other systems to identify novel virulence determinants.

ACKNOWLEDGEMENTS

Because it would be overwhelming to sufficiently acknowledge everyone who contributed to my work and professional development during my 6 years of grad school at UNC, I would like to briefly thank a few individuals who have contributed to my graduate career. First, I would like to thank Jeff Dangl for his support and for giving me a home for the last 6 years. I have learned about science and about life from Jeff, and have no regrets about my choice of lab and mentor. With Jeff comes a larger family of scientists beginning with Sarah Grant, who also advised me and supported my work throughout my tenure at UNC. The Dangl/Grant lab includes many individuals who aided my progress, but I would like to especially acknowledge Petra Epple for her invaluable help on my dissertation, David Hubert for being a friend and a model for me throughout grad school, Youssef Belkhadir and Gopal Subramanian for being great friends with a passion for good science, Zack Nimchuk for introducing me to the lab and for understanding that I had to abandon AvrPphE, Marc Nishimura for reminding me of the importance of controls during my 4th year, Mindy Roberts for keeping me sane in lab and at home, and David Baltrus for rounding out our Tyler's trivia team and for being my comparative genomics cohort. My best friend in lab and in life is Dr. Ben Holt III who taught me how to be passionate about science, politics, and life all at once. Thanks, Ben, for being an exemplary human being and for helping me remember that scientists are human.

Outside of lab, I received help from many members of the UNC community. Thanks to my committee, all of whom were incredibly patient with my project and my legendarily-long

committee meetings. Thanks most of all to Corbin Jones who is a collaborator and constant supply of enthusiasm for this project. Corbin also brought with him Erin Friedman and Josie Reinhardt whose help was invaluable on the work in Chapter 3. Thanks to Matt Wolfgang for technical help with Pseudomonas knockouts and motility, along with his lab members Nan Fulcher and Ryan Heineger. Thanks to Jason Reed for being so thoughtful and constructive during committee meetings. Thanks to Janne Cannon for advice early in my career and for bringing an outside perspective to my work. From the Biology Department, I'd like to thank Vicki Bautch, Jason Lieb, Greg Copenhaver, and Steve Matson for support throughout my tenure here.

I'd like to thank my colleagues Jan LaRocque, Danny Monroe, Willow Gabriel, Jen-Yi Lee, Deirdre Tatomer, Kyle Gaulton, Evan Merkhoffer, Kim Peters, Ben Harrison, and Mindy Roberts for their friendship.

I'd like to thank my family for their support and love and for understanding what a crazy and fulfilling experience this has been for me. Thanks also to Brian Robertson for love and support throughout.

TABLE OF CONTENTS

LIST OF TABLES……..……..……vii

LIST OF FIGURES……...viii

Chapter

I. An Introduction to The Evolution of Virulence in *P. syringae*…….…..…..……1

A. Introduction…….…….…….…….……2

B. References…….……..…….……19

II. Codon Volatility Identifies Candidate Virulence Factors……..……..……....26

A. Abstract…….…….……..……27

B. Results……..…….……...…….28

C. Discussion…….…….……..……53

D. Methods and Materials…….……...…….....……60

E. References……..…….…….…….....63

III. The Comparative Method Identifies Candidate Virulence Factors ……..….....68

A. Introduction……..…….……..…….....69

B. Results…….…….…….…….……..70

C. Discussion…….…….……109

D. Methods and Materials…….…….……119

E. References…….……123

IV. Conclusions and Future Directions ……....……...……...……...128

A. References……..……139

LIST OF TABLES

Table 2.1 Transition/transversion rates used to calculate volatility *p*-values are within known parameters……..……...……...….32

Table 2.2 Codon frequencies in Pto_{DC3000} favor highly volatile serine codon AGC……...…….…….…...…..34

Table 2.3 Conservation of most and least volatile genes across *Pseudomonads*…….……...……..……...44

Table 2.4 Candidate genes with high volatility and conservation only in the genomes of plant-associated bacteria...…….……46

Table 2.5 Volatility and bacterial lifestyle enrich resultant candidate gene pools in purported virulence genes…….……..…….52

Table 3.1 Summary of Pto_{DC3000} Candidate Gene Identification and Experimental Validation.…….……...……..……....77

Table 3.2 Candidate genes and their conservation in genomes or organisms in which saturating mutageneses have been performed…….……81

Table 3.3 Summary of mutant phenotypes …….……..……87

Table 3.4 Summary of swarming motility phenotypes …...……..……..…….100

Table 3.5 Only motility controls had defects in swimming motility ……..……107

Table 3.6 Gene selection based on high dN/dS enriches the resultant candidate gene pool in purported virulence genes…….……108

Table 4.1 Low dN/dS Genes from 3-way comparison.....137

LIST OF FIGURES

Figure 2.1 Volatility of codons……..…....…30

Figure 2.2 Genes in all volatility categories have equal representation in the amino acids involved in volatility…….……...35

Figure 2.3 Highly volatile genes include an overabundance of very long proteins……...…..……..…….……..38

Figure 2.4 Conservation of genes in other bacteria varies with volatility.……...……..…..…..40

Figure 2.5 Divergence between genes of different volatility categories based on predicted subcellular localization ……..……...42

Figure 2.6 Volatile gene *hopL1* is conserved in diverse bacteria……...…….....48

Figure 2.7 Volatile gene *corR* is required for virulence of *Pto*$_{DC3000}$ on Arabidopsis and tomato……...……...…..50

Figure 3.1 Seven-gene MLST phylogeny of strains used for the dN/dS comparison ……...……..72

Figure 3.2 Venn diagram showing the distribution of high dN/dS candidate genes from *Pto*$_{DC3000}$ in other Pseudomonads …….…….……..……....…76

Figure 3.3 No correlation between volatility and dN/dS ranks.……...……78

Figure 3.4 An insertional mutation strategy was used to create mutations in candidate genes…….……...……...79

Figure 3.5 Mutations in *flaA, motA*, and *gacA* cause growth defects *in planta*.……...……..…84

Figure 3.6 Mutations in *flaA* and *motA* cause loss of swarming and swimming motility …….……...……..……..86

Figure 3.7 *pspto5557* is a duplicate of the *cysC* gene found in two loci in the *Pto*$_{DC3000}$ genome …….……...……90

Figure 3.8 CysC proteins are more similar to one another than to CysNC bifunctional proteins in the same genomes ……..…….……...…..91

Figure 3.9 Phenotypes of *pspto5557* mutant and complemented construct (pEF109)……...……...….92

Figure 3.10 The operon containing candidate genes *pspto5537*, *pspto5538*, and *pspto5539* is conserved in Pseudomonad and Xanthomonas genomes, but not in Psy_{B728a} or Pph_{1448A}……..…….....96

Figure 3.11 The *pspto5537-5540* operon is required for full Pto_{DC3000} virulence on Arabidopsis and tomato and possibly in swarming motility …..……...…….….98

Figure 3.12 The *pspto2999* mutant has a swarming defect but no defect in swimming motility ……..……..……..…..101

Figure 3.13 *pspto4043* and *pspto5566* are required for growth in minimal medium …….…….....…102

Figure 3.14 The *avrPto avrPtoB* double mutant is deficient in swarming motility but not growth in minimal medium …….…….104

Figure 4.1 Additional *P. syringae* genome sequences add to our understanding of Pto_{DC3000}……..……..……...…..134

Chapter 1

An Introduction to the Evolution of Virulence in *P. syringae*

This project addresses the hypothesis that patterns of evolution can identify genes important to pathogen virulence. This hypothesis is based on the fact that pathogen genes are under selection imposed by the host immune system. To address this hypothesis, I used comparative genomics and evolutionary genetics tools to identify evolving candidate virulence genes. Furthermore, I moved beyond the typical *in silico* definition of such candidate gene lists to experimentally test 9 of 20 candidate virulence genes so defined in a model pathosystem. Several genes were validated as both evolving and functionally relevant for virulence. In this chapter I will introduce the model system used to evaluate that hypothesis and provide background and justification for this investigation.

The Pseudomonas-Arabidopsis Interaction Is a Model for Pathogen-Host Interaction

Pseudomonas syringae isolates can cause bacterial speck or blight diseases on diverse plant hosts. *P. syringae* has been found to infect seeds, roots, and fruits, but the main site of bacterial proliferation is within the intercellular spaces of the leaf, termed the apoplast (Hirano and Upper, 2000). *P. syringae* strains have been isolated from many different plant hosts (as diverse as tomato, maple trees, and wheat) each strain is given a pathovar name indicating the plant from which it was originally isolated. Pathovar names do not necessarily indicate the limits of host range, though, as some pathovars infect several different hosts (Tsiamis et al., 2000; Yan et al., 2008). The host ranges of some *P. syringae* pathovars can also be quite narrow (Yan et al., 2008). Two pathovars were found to be virulent on the genetic model plant *Arabidopsis thaliana*: pathovars tomato

(*Pto*) and maculicola (*Pma*) (Crute et al., 1994). One strain, *Pto* strain DC3000 (Pto_{DC3000}) was found to be particularly virulent on Arabidopsis, where the bacteria proliferate causing chlorosis and necrotic spots on leaf tissue, and has thus become a model pathogen used to dissect the genetics of plant disease resistance and susceptibility (Katagiri et al., 2002; Yan et al., 2008). Pathovar maculicola has also been used for similar studies, but here I will focus on only $Pto_{DC3000.}$ Tomato plants have also been used as a model host for Pto_{DC3000} infection. In tomato, Pto_{DC3000} infection is characterized by chlorosis and necrosis on leaves as well as raised black lesions on the fruit (Goode and Sasser, 1980).

The Type III Secretion System Is Required for Pto_{DC3000} virulence

Early genetic screens for non-virulent *P. syringae* mutants implicated the type III secretion system (TTSS) in plant disease (Lindgren et al., 1986). The TTSS is a needle-like apparatus enabling injection of bacterial proteins directly into the host cell and is a common virulence mechanism in many pathogens. The type III secretion apparatus is well conserved among diverse bacteria and the structure of the apparatus (the injectisome) is highly similar to the structural components of flagellum (Cornelis, 2006). The TTSS injectisome includes the same basic basal structure as the flagellum, but lacks the motor proteins MotA and MotB as well as the flagellar hook and the filament itself. In place of the flagellar hook and filament, the basal structure is attached to the type III pilus (HrcC in *P. syringae*). It is through this pilus that the type III effector proteins are injected into the host cell where they manipulate the host cell machinery to the benefit of

the bacterial pathogen (Grant et al., 2006; Jin et al., 2003). *P. syringae* lacking the type III pilus, *hrcC*, are non-virulent on both resistant and susceptible plants. These bacteria are unable to elicit a form of programmed cell death around the site of infection, called the hypersensitive response (HR), in resistant plants, since no type III effectors are delivered, but they are also unable to suppress host defenses through the function of type III effectors. This non-virulent phenotype of *hrcC* mutant bacteria will be demonstrated throughout the following chapters and is an important control in my studies.

In many cases, regulation of TTSS genes is induced upon host contact and is not part of the bacterial lifestyle outside of the host environment (He et al., 2004; Rahme et al., 1992; Rico and Preston, 2008). Further refinement in the evolved utilization of type III secretion systems is observed in the human gut pathogen *Salmonella enterica*. *Salmonella enterica* encodes two distinct TTSS (SPI1 and SPI2), each of which is differentially regulated: SPI is induced in the lumen of the small intestine and this can be mimicked in the lab using high osmolarity, low oxygen, and pH8, while SPI2 is induced only in macrophages which is mimicked in the lab using an acidic minimal medium and low-oxygen conditions. TTSS expression in *Pto*$_{DC3000}$ is induced upon contact with the plant and it has been shown that fluid extracted from the leaf apoplast is sufficient to induce expression of *hrpL*, which controls expression of the TTSS (Rico and Preston, 2008; Xiao et al., 1994).

Initial studies of host range for *Pto*$_{DC3000}$ focused on induction of the HR in resistant plants (Beers and McDowell, 2001). The HR is strongly correlated with plant disease resistance, though it is not absolutely required for a successful disease resistance response (Aviv et al., 2002). It was shown that induction of the HR was dependent both

on bacterial genes, termed *avr* genes because they render the bacteria avirulent, and plant-encoded disease resistance genes (*R*-genes). This led to the so-called gene-for-gene hypothesis which proposed that *avr* gene products (Avr proteins) were directly recognized by corresponding *R*-gene products (Flor 1971). Avr proteins were subsequently shown to be directly injected into the plant cytoplasm via the TTSS and were also shown to be required for full pathogen virulence (Lee, 1997; Mudgett and Staskawicz, 1999). For instance *avrRpm1*, a gene from *P. syringae* pathovar maculicola strain M2, triggers the HR on Arabidopsis plants expressing the corresponding *R* gene *RPM1*, but *avrRpm1* is also required for full virulence on the susceptible Arabidopsis accessions (inbred genotypes) Mt-0 and Fe-1 which lack *RPM1* (Debener et al., 1991; Ritter and Dangl, 1995).

While individual pathogen isolates have varying suites of *avr* genes, the TTSS is a conserved virulence mechanism employed by bacteria infecting a wide range of hosts (Grant et al., 2006). Bacteria that infect mammals as well as those infecting plants encode the TTSS and include *E. coli, Salmonella, Yersinia, Pseudomonas, Xanthomonas, Ralstonia* and many others (Grant et al., 2006). In each of these bacteria, a membrane-embedded needle-like apparatus is constructed that forms a conduit between the bacterial cell and the host cell cytoplasm through which "type III effector" proteins are transported. The full suite of type III effector proteins are thought to be required for full virulence on susceptible hosts, but since Avr proteins are among these TTSS effector proteins, single type III effectors can negatively influence pathogen fitness on genotypically resistant hosts. While the proteins that make up the TTSS apparatus are well-conserved throughout diverse bacteria, the effector proteins encoded by a particular bacterial species, and

different pathovars within a given species, differ greatly (Grant et al., 2006; Mudgett, 2005). As noted above, Avr proteins are type III effectors and the distribution of *R*-genes in a plant population will limit the abundance of the corresponding *avr*-genes in the bacterial population (Rose et al., 2007). However, in the absence of a functional R-protein, Avr proteins have been shown to contribute to virulence (Belkhadir et al., 2004; Lorang et al., 1994; Ritter and Dangl, 1995). This virulence function of the *avr* genes is part of the balancing selection that maintains these genes in the bacterial population even though, in the context of a specific R-gene-expressing host, they can be detrimental to bacterial proliferation (Van der Hoorn et al., 2002).

Moreover, the virulence function of TTSS effectors has been demonstrated by expressing the effector gene transgenically in plants (Chen et al., 2004; Hauck et al., 2003; Underwood et al., 2007). For instance, the He lab (Michigan State) showed that the expression of effector genes *avrPto* and *hopAO1 in planta* enable a normally non-virulent *Pto*$_{DC3000}$ mutant to grow to high levels on the transgenic leaf. Similarly, *avrRpt2*, an effector gene from *Pto* strain T1, was shown to add virulence to *Pma* on plants lacking the corresponding R protein RPS2 (Chen et al., 2000; Whalen et al., 1991). These virulence effects explain why the TTSS is a conserved virulence mechanism despite the fact that some effectors are detrimental to bacterial growth on certain *R*-gene encoding plants. Indeed, *Pto*$_{DC3000}$ TTSS mutants are unable to proliferate on plant leaves including Arabidopsis. (Lindgren et al., 1986; Peet et al., 1986). The TTSS mutant *hrcC* is used throughout this work as a negative control for bacterial virulence. The role of the type III effector AvrPto will be highlighted in chapter 3 along with evidence that it is a rapidly evolving gene, possibly due to direct host contact.

Bacterial Infection Triggers Non-Specific (Basal) Defenses

The TTSS effectors are not the only triggers of plant defenses. Conserved proteins common to many bacteria, such as flagellin, induce disease resistance in Arabidopsis (Gómez-Gómez and Boller, 2000). A specific 22-amino acid peptide derived from flagellin interacts with the Toll-like receptor FLS2 and induces defense responses including callose deposition, a localized build-up of the plant cell wall, resulting in reduced bacterial growth (Chinchilla et al., 2006; Zipfel et al., 2004). Conversely, plants encoding non-functional *fls2* are more susceptible to bacterial infection. FLS2 appears to be a major factor in basal resistance and is targeted by several TTSS effector proteins (Abramovitch et al., 2006; Xiang et al., 2008). The role of FLS2 in determining bacterial virulence will be addressed in Chapter 3 when I discuss the requirements for motility during infection and the consequences of recognition by FLS2.

A second strategy to suppress basal defense is to disrupt plant hormone signaling via salicylic acid (SA), which is involved in *R*-gene mediated resistance and basal defense (Nawrath and Métraux, 1999). Plants lacking isochorismate synthase activity due to a mutation in the gene *eds16/sid2* cannot synthesize SA and are more susceptible to *P. syringae* (both *Pto*$_{DC3000}$ and *Pma* strain ES4326), the oomycete pathogen *Peronospora parasitica*, and the biotrophic fungus powdery mildew *Erysiphe cichoracearum* (Dewdney et al., 2000; Nawrath and Métraux, 1999; Wildermuth et al., 2001). Enhanced susceptibility to *Pto*$_{DC3000}$ was found in the *eds16/sid2* mutants infected with wild-type *Pto*$_{DC3000}$ and *Pto*$_{DC3000}$ expressing the avr gene *avrRpt2* (from *Pst* T1), suggesting that SA is required for R-protein-mediated disease resistance as well as basal

resistance (Nawrath and Métraux, 1999). Another plant hormone, jasmonic acid (JA) is a natural antagonist of SA signaling and acts to down-regulate SA-dependent resistance (Schilmiller and Howe, 2005). *Pto*$_{DC3000}$ produces a toxin, coronatine, that is structurally highly similar to jasmonic acid conjugated to isoleucine (JA-Ile). Coronatine disrupts SA signaling, suppresses plant defenses, and contributes to virulence of this pathogen. Coronatine is required for full virulence and symptom induction by *Pto*$_{DC3000}$ on both Arabidopsis and tomato (Bender et al., 1999; Bender et al., 1987). The role of coronatine in *Pto*$_{DC3000}$ virulence will be discussed in Chapter 2 along with evidence that the response regulator CorR is encoded by a gene rich in so-called volatile codons.

Evolution of Virulence

This project aims to identify novel virulence factors via analyses of the evolutionary pressure on the genes of *Pto*$_{DC3000}$ and especially pressure potentially exerted by the plant host during this interaction. The evolutionary relationship between pathogen and host is often likened to an arms race wherein a pathogen acquires or evolves the ability to cause disease and the host, in turn, evolves a strategy to halt infection (Dawkins and Krebs, 1979). This process repeats as the host immune system acts as a force of natural selection on the genes of the pathogen and the disease, in turn, acts as a selective agent on the genes of the host. Other groups have shown that proteins on the exposed surface of pathogens, those directly in contact with the host immune system, are often rapidly evolving (Hughes and Nei, 1992). These rapidly evolving proteins were also shown to be important components of pathogen virulence (Hughes and

Nei, 1992). Based on this knowledge, I hypothesized that the rate of evolution could act as a predictor of virulence function. Identification of virulence factors in this way may reveal subtle modulators of pathogen virulence previously unidentified in genetic screens, which tend to identify only genes causing drastic changes in pathogen virulence (Huynh et al., 1989).

Bacterial genomes evolve in two major ways:

i) Acquisition of novel genes via horizontal transfer

ii) Mutation.

Horizontal gene transfer is a major mechanism shaping bacterial genomes. While vertical gene transfer refers to genetic material obtained directly from the most recent ancestor, horizontal gene transfer occurs when a bacterium (or any organism, but I will focus on bacterial horizontal gene transfer) acquires foreign genetic material and replicates this new DNA along with its own genome. This ability of many bacteria to acquire genes from the environment (transformation), other bacteria (conjugation) or through bacteriophages (transduction) enabled some of the seminal studies that identified DNA as the genetic material. Early work by Avery, MacLeod, and McCarty demonstrated that heat-killed virulent Pneumococcal bacteria could transform non-virulent bacteria into virulent bacteria (Avery et al., 1944). They concluded that the non-virulent bacteria must acquire genetic material from the heat-killed samples and further investigations showed that DNA was that genetic material.

Horizontal gene transfer complicates the concept of bacterial species because large portions of the genome can have patterns of evolution inconsistent with the "core" genome, which composes most of the genome (70-80%) and encodes most essential

cellular functions (Gal-Mor and Finlay, 2006). Traditionally, bacterial species have been defined by phenotypic similarity, while recent genomic studies have caused a reevaluation of these species divisions. For instance, *E. coli* and *S. flexneri* are phylogenetically indistinguishable based on genome sequencing, but they cause different disease outcomes and have historically been treated as separate species (Wei et al., 2003). Similarly, the tumor-inducing plant pathogen *Agrobacterium tumefaciens* was recently found to be genetically indistinguishable from strains of Rhizobia, root symbionts, based on 16s rRNA sequences (Young et al., 2001). Debates have arisen over the proposal to reclassify several Agrobacterium strains as Rhizobia (Farrand et al., 2003; Young et al., 2001) and the debate continues as to whether genomic similarity based on the "core" genome or phenotypic characteristics should determine the identity of a bacterial species (Gevers et al., 2005).

Horizontal gene transfer is also a common mechanism by which bacteria acquire virulence factors (Gal-Mor and Finlay, 2006). Two of the main virulence mechanism described above, encoding the TTSS and genes for coronatine biosynthesis, were obtained by *Pto*$_{DC3000}$ via horizontal transfer (Buell et al., 2003; Collmer et al., 2000). The genes required for coronatine biosynthesis, including the *cma* and *cfa* gene clusters, are located near type III effector genes and are bordered by IS elements (Buell 2003). This observation along with the observation that these regions are non-syntenic with analogous regions in other *Pseudomonas* genomes suggests that these genes were acquired by horizontal transfer. The TTSS apparatus genes are part of a mobile pathogenicity island characterized by aberrant GC content compared to the rest of the genome and the proximity to mobile elements (Buell et al., 2003). Similarly, the CEL,

(conserved effector locus) encoding several TTSS effectors, is embedded in a mobile genetic element characterized by its association with $tRNA^{Leu}$ (Deng et al., 2003). Comparative genomic hybridization also identified several type III effectors acquired by horizontal transfer by showing that the phylogeny of the effector alleles did not match the phylogeny of the core genome created by housekeeping gene analysis (Sarkar et al., 2006). This incongruity between the phylogenies suggests that these genes were not acquired vertically and therefore are likely products of horizontal transfer.

Genome wide comparisons both within and across bacterial species have revolutionized our thinking about the "core" and "plastic" genomes that comprise a species "pan-genome". In the frame of my work, for example, genes conserved among all *P. syringae* isolates would make up the minimal *P. syringae* genome required for plant infection and survival in niches common to all *P. syringae* lifestyles, while genes conserved among only, for example, the tomato pathovars would give us insight into genes required to disarm tomato-specific defenses.

Several examples are of note. First, hybridization-based comparisons of the suites of genes encoding the TTSS and TTSS effectors, quorum sensing, the type IV pilus, flagellin, toxin production, and other processes suggested a link between the presence of these genes and plant host range (Sarkar et al., 2006). Sarkar et al. (2006) found host-specific profiles of virulence-associated genes, as determined by comparative genomic hybridization (CGH) of DNA from 91 strains of *P. syringae* to a microarray of Pto_{DC3000} genes. They deemed genes involved in the general secretion apparatus, flagella, and alginate biosynthesis to be "conserved" virulence-associated genes because they are found in >95% of strains tested and concluded that these genes must be required for all *P.*

syringae lifestyles regardless of host plant. They also showed that moderately variable genes included those involved in quorum sensing, type IV pili, TTSS regulation, polysaccharide synthesis and siderophores. This is in contrast to earlier work in *P. aeruginosa* that demonstrated high conservation of virulence-associated genes in approximately 97% of isolates regardless of whether the site of isolation was clinical or environmental (Wolfgang et al., 2003).

Sarkar et al. (2006) also found clade-specific and host-specific virulence-associated genes by focusing the distribution of known virulence factors in multiple *P. syringae* isolates. For instance, they showed that *shcA*, encoding a type III chaperone discussed further in Chapter 2, is very common in soybean pathogens, indicating that it may be required for virulence on this host. One important drawback to the CGH technique is that it is unable to differentiate between "present" and functional alleles, since the alleles in the 91 compared strains were not sequenced. Furthermore, presence of hybridization cannot discern the level of nucleotide diversity, potentially of functional relevance, occurring in genes shared among the 91 strains. Importantly, the stringency of the CGH study will influence whether a homolog will be identified. For instance, if a gene is rapidly evolving in one strain, this sequence diversification may eliminate detection of the rapidly evolving homolog. Finally, CGH will not be able to identify genes of interest that are not present in the set of genes that constitute the template for the hybridization.

If the conclusions of these CGH studies are correct, though, there is much to be learned about the determinants of bacterial virulence from surveying conserved and non-conserved genes from bacterial pathogens of different hosts and different environmental

niches (Sarkar et al., 2006; Wolfgang et al., 2003). Similar studies have identified host-specific virulence factors in *V. cholera, M. tuberculosis, N. meningitides, P. aeruginosa, and S. epidermidis* (Dorrell et al., 2005).

Previous work compared the genome of Pto_{DC3000} to the fully-sequenced genomes of the soil microbe *P. putida* KT2440 and *P. aeruginosa* PA01 (Joardar, Lindeberg et al. 2005). This work identified lineage-specific regions of the Pto_{DC3000} genome and suggested that these regions might be enriched for genes required for lifestyles in the history of *P. syringae* that are not shared with the histories of *P. putida* or *P. aeruginosa* (or were present in the history but have since been lost by the non-phytopathogens). Recent work compared the three fully-sequenced *P. syringae* pathovars tomato strain DC3000 (Pto_{DC3000}), syringae strain B728a (Psy_{B728a}), and phaseolicola strain 1448A (Pph_{1448A}) based on known virulence factors from diverse bacterial pathogens and regions of non-synteny in these three *P. syringae* strains (Lindeberg, Myers et al. 2008). They compiled recent functional and genomic data from these three model pathogen isolates and list orthologs from all three genomes, noting the synteny or lack thereof. Synteny, conservation of gene order and genomic location, implies vertical inheritance, whereas divergent genomic locations imply distinct acquisition by horizontal transfer. They report 121 variable genomic regions and list 281 genes in Pto_{DC3000} suggested to function in virulence. Because my work aims to identify novel virulence factors based on the rate of molecular evolution, these two comparative reviews, one with a broad view of the differences between phytopathogenic and non-phytopathogenic Pseudomonads (Joardar, Lindeberg et al. 2005), the other focused on variation among closely-related genomes (Lindeberg, Myers et al. 2008), provide useful resources for correlating candidate

virulence factors with my comparative evolution work. As expected, several genes identified in my work are in genomic regions identified as lineage-specific or orthologs of virulence factors in these papers (Joardar, Lindeberg et al. 2005; Lindeberg, Myers et al. 2008). These results are discussed in Chapters 2 and 3.

Molecular evolution and its application in this work

Mutation is the most common way for genomes to evolve. The two major classes of mutation are non-synonymous (those that cause a change in the protein coding sequence) and synonymous (those with no effect on protein sequence). Synonymous mutations are mostly immune from natural selection because they have no functional consequence with the exception of the codon use preference of the organism. Because natural selection does not act against synonymous mutations, these silent substitutions become fixed in a population very rapidly. In contrast, non-synonymous mutations have functional consequences on the encoded protein and are often lost (Kimura, 1983). However, non-synonymous mutations can become fixed in a population of their impact is beneficial or neutral. The rate of non-synonymous mutations (dN) compared to the rate of silent, synonymous mutations (dS) is a common method to identify molecular evolution in a given gene. If a gene incurs more non-synonymous mutations than silent synonymous mutations, this is consistent with positive selection, therefore a $dN/dS > 1$ is indicative of positive selection. Genes in which there is no selection, such as pseudogenes, have a dN/dS approximately = 1, while genes under functional constraint

against sequence diversifications have dN/dS <1 (Nei and Gojobori, 1986). However, this method requires comparable genes from related species.

Previous work demonstrated that known virulence factors are rapidly evolving in *Pto*$_{DC3000}$. Rohmer et al. (2004) looked at the rate of molecular evolution using the dN/dS ratio in type III effector gene families where at least three homologs per family were present (Rohmer et al., 2004). They further analyzed specific sites that were polymorphic among these homologs to identify amino acids that may be critical for protein function. This yielded evidence of positive selection on 5 TTSS effectors from *Pto*$_{DC3000}$: *hopX (avrPphE), hopAF1 (holPtoN), hopQ1 (holPtoQ), hopI1(hopPmaI)*, and *hopAB3 (hopPmaL)*, as well as diversification in the N-terminus of the harpin *hrpW*, a domain known to be sufficient for induction of the HR in non-host plants (Charkowski et al., 1998). These findings support my expectation that identification of novel rapidly evolving genes in *Pto*$_{DC3000}$ will reveal novel virulence factors. On the other hand, Rohmer et al. also noted that many type III effectors appeared to be under purifying selection. They postulated that genes under purifying selection are constrained against diversification because the function of these genes is required for survival or infection.

When this project was initiated, only one genome sequence was completed for the *P. syringae* pathovars, that of *Pto*$_{DC3000}$ (Buell et al., 2003). We therefore initially employed what was at the time a new and unproven method allowing single-genome consideration of molecular evolution rates, called "codon volatility" which is discussed in Chapter 2 (Plotkin et al., 2004). Volatility describes the proportion of a given gene rich in so-called "volatile codons", which were proposed to be footprints of recent gene evolution (see Chapter 2). This approach yielded a rank order list of candidate virulence

genes encoding proteins predicted to localize to the bacterial outer membrane, where they could come in direct contact with the host cell. Several candidates are annotated as hypothetical and seem to be unique to the sequenced strain *Pto*$_{DC3000}$.

Before comprehensive mutational analysis of the volatile candidate genes had begun, though, two important things happened. First, several critiques were leveled against the use of codon volatility as a method (discussed in Chapter 2 and (Chen et al., 2005; Hahn et al., 2005; Nielsen and Hubisz, 2005). Second, and most importantly, two additional *P. syringae* genomes became available, allowing me to use more thoroughly vetted and widely accepted dN/dS approach both to generate candidate lists of evolving genes for mutation and to cross reference to the volatility analysis (Feil et al., 2005; Joardar et al., 2005a). If the volatility analysis accurately identifies evolving genes, the candidate lists from the two methods should contain the same genes.

Each method has advantages and drawbacks. For instance, volatility is evaluated assuming constant GC content and transition-transversion (κ) rates across the genome, which excludes evaluation of horizontally transferred genes which often have aberrant GC and κ, depending on the donor genome. Similarly, the comparative method requires identifiable homologs and some knowledge of the relatedness of species containing the homolog, and therefore excludes genes that are strain-specific. Excluding these two categories of genes, though, the most volatile genes were expected to overlap to some degree with the genes with the highest evolution rate based on the comparative method.

The comparative method uses homologous sequences and evaluates the ratio of non-synonymous mutations to synonymous mutations (Kimura, 1977). A dN/dS ratio between 0 and 1 indicates negative selection on a gene, while a dN/dS greater than 1

indicates positive selection (Miyata and Yasunaga, 1980). Positive selection can indicate diversification, possibly due to selective pressure imposed by the host immune system and may indicate some function in pathogen virulence. As mentioned above, Rohmer et al (2004) used dN/dS to identify type III effector genes under positive selection in Pto_{DC3000} (Rohmer et al., 2004). They used multiple homologous sequences to carefully dissect domains and even codons that are under positive selection, possibly due to host recognition. Our study evaluates the overall dN/dS of the entire gene and requires that the gene as a whole have a high dN/dS.

Several studies have used this method to identify genes under positive selection in *E .coli* (Chen et al., 2006; Petersen et al., 2007). In both of these studies, which focused on different sets of *E. coli* and, in the case of Petersen et al., *S. flexneri*, identified genes required for pathogen virulence (iron uptake genes, toxins, etc.) and genes encoding proteins exposed on the outer membrane where they may come into contact with the host immune system. This project is similar to the studies in *E. coli* in that it utilizes closely-related strains, but one major difference is that many of the comparisons in this project included only two homologous sequences, whereas the *E. coli* studies required a minimum of 3 homologues. My comparative approach yielded a list of 20 candidate genes with high dN/dS ratios which are discussed in Chapter 3.

In this Introduction, I have stated my hypothesis that genes relevant to bacterial virulence and plant disease will be under selection due to pressure from the plant host. The interaction between host and pathogen can be described as an arms-race in which detection by the host plant induces the pathogen to evolve mechanisms to avoid detection. This process repeats over time, causing rapid evolution of the bacterial genes required for

virulence. I have introduced the model host-pathogen system and the basic major virulence mechanism used by *P. syringae* to suppress host basal defense. My goal is to find more subtle virulence factors using the tools of comparative genomics and molecular evolution rates. I outlined two different theoretical approaches, codon volatility and dN/dS ratios, as tools for this work. In the subsequent chapters, I will discuss my results using these methods to determine whether

1) these two methods lead to an overlapping set of candidate virulence factors, and/or
2) candidates identified by either method are required for virulence of Pto_{DC3000} during infection of Arabidopsis or tomato plants.

My focus, as noted at the beginning, was to generate rank order candidate gene lists for *in vivo* studies. Quite beyond the typical analysis of merely identifying those genes that might be evolving under host defense pressure, I sought to experimentally confirm roles in virulence for any or all of my candidate genes.

References

Abramovitch, R. B., Anderson, J. C., and Martin, G. B. (2006). Bacterial elicitation and evasion of plant innate immunity. Nat Rev Mol Cell Biol *7*, 601-611.

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transfomrations of Pneumococcal Types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. Journal of Experimental Medicine *79*, 137-158.

Aviv, D. H., Rusterucci, C., Holt III, B. F., Dietrich, R. A., Parker, J. E., and Dangl, J. L. (2002). Runaway cell death, but not basal disease resistance, in *lsd1* is SA- and *NIM1/NPR1*-dependent. Plant J *29*, 381-391.

Beers, E. P., and McDowell, J. M. (2001). Regulation and execution of programmed cell death in response to pathogens, stress and developmental cues. Curr Opin Plant Biol *4*, 561-567.

Belkhadir, Y., Nimchuk, Z., Hubert, D. A., Mackey, D., and Dangl, J. L. (2004). Arabidopsis RIN4 negatively regulates disease resistance mediated by RPS2 and RPM1 downstream or independent of the NDR1 signal modulator, and is not required for the virulence functions of bacterial type III effectors AvrRpt2 or AvrRpm1. Plant Cell *16*, 2822-2835.

Bender, C. L., Alarcon-Chaidez, F., and Gross, D. C. (1999). Pseudomonas syringae phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. Microbiol Mol Biol Rev *63*, 266-292.

Bender, C. L., Stone, H. E., and Cooksley, D. A. (1987). Reduced pathogen fitness of *Pseudomonas syringae* pv. *tomato* Tn5 mutants defective in coronatine production. Physiol Molec Plant Pathol *30*, 273-283.

Buell, C. R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I. T., Gwinn, M. L., Dodson, R. J., Deboy, R. T., Durkin, A. S., Kolonay, J. F.*, et al.* (2003). The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. Proc Natl Acad Sci U S A *100*, 10181-10186.

Charkowski, A. O., Alfano, J. R., Preston, G., Yuan, J., He, S. Y., and Collmer, A. (1998). The Pseudomonas syringae pv. tomato HrpW protein has domains similar to harpins and pectate lyases and can elicit the plant hypersensitive response and bind to pectate. J Bacteriol *180*, 5211-5217.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R. R., Ozersky, P.*, et al.* (2006). Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. Proc Natl Acad Sci U S A *103*, 5977-5982. Epub 2006 Apr 5973.

Chen, Y., Emerson, J. J., and Martin, T. M. (2005). Evolutionary genomics: codon volatility does not detect selection. Nature *433*, E6-7; discussion E7-8.

Chen, Z., Kloek, A. P., Boch, J., Katagiri, F., and Kunkel, B. N. (2000). The *Pseudomonas syringae avrRpt2* gene product promotes pathogenicity from inside the plant cell. Mol Plant-Microbe Interact *13*, 1312-1321.

Chen, Z., Kloek, A. P., Cuzick, A., Moeder, W., Tang, D., Innes, R. W., Klessig, D. F., McDowell, J. M., and Kunkel, B. N. (2004). The Pseudomonas syringae type III effector AvrRpt2 functions downstream or independently of SA to promote virulence on Arabidopsis thaliana. Plant J *37*, 494-504.

Chinchilla, D., Bauer, Z., Regenass, M., Boller, T., and Felix, G. (2006). The Arabidopsis receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception. Plant Cell *18*, 465-476. Epub 2005 Dec 2023.

Collmer, A., Badel, J. L., Charkowski, A. O., Deng, W. L., Fouts, D. E., Ramos, A. R., Rehm, A. H., Anderson, D. M., Schneewind, O., van Dijk, K., and Alfano, J. R. (2000). Pseudomonas syringae Hrp type III secretion system and effector proteins. Proc Natl Acad Sci U S A *97*, 8770-8777.

Cornelis, G. R. (2006). The type III secretion injectisome. Nat Rev Microbiol *4*, 811-825.
Crute, I. R., Beynon, J., Dangl, J. L., Holub, E. B., Mauch-Mani, B., Slusarenko, A., Staskawicz, B. J., and Ausubel, F. M. (1994). Microbial pathogenesis of Arabidopsis. In Arabidopsis, E. M. Meyerowitz, and C. R. Somerville, eds. (Cold Spring Harbor, Cold Spring harbor Laboratory Press), pp. 705-748.

Dawkins, R., and Krebs, J. R. (1979). Arms races between and within species. Proc R Soc Lond B Biol Sci *205*, 489-511.

Debener, T., Lehnackers, H., Arnold, M., and Dangl, J. L. (1991). Identification and molecular mapping of a single *Arabidopsis thaliana* locus determining resistance to a phytopathogenic *Pseudomonas syringae* isolate. Plant J *1*, 289-302.

Deng, W. L., Rehm, A. H., Charkowski, A. O., Rojas, C. M., and Collmer, A. (2003). Pseudomonas syringae exchangeable effector loci: sequence diversity in representative pathovars and virulence function in P. syringae pv. syringae B728a. J Bacteriol *185*, 2592-2602.

Dewdney, J., Reuber, T. L., Wildermuth, M. C., Devoto, A., Cui, J., Stutius, L. M., Drummond, E. P., and Ausubel, F. M. (2000). Three unique mutants of Arabidopsis identify eds loci required for limiting growth of a biotrophic fungal pathogen. Plant J *24*, 205-208.
Dorrell, N., Hinchliffe, S. J., and Wren, B. W. (2005). Comparative phylogenomics of pathogenic bacteria by microarray analysis. Curr Opin Microbiol *8*, 620-626.

Farrand, S. K., Van Berkum, P. B., and Oger, P. (2003). Agrobacterium is a definable genus of the family Rhizobiaceae. Int J Syst Evol Microbiol *53*, 1681-1687.

Feil, H., Feil, W. S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., Lykidis, A., Trong, S., Nolan, M., Goltsman, E.*, et al.* (2005). Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci U S A *102*, 11064-11069. Epub 12005 Jul 11025.

Gal-Mor, O., and Finlay, B. B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. Cell Microbiol *8*, 1707-1719. Epub 2006 Aug 1724.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., and Swings, J. (2005). Opinion: Re-evaluating prokaryotic species. Nat Rev Microbiol *3*, 733-739.

Gómez-Gómez, L., and Boller, T. (2000). FLS2: An LRR receptor like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. Mol Cell *5*, 1003-1011.

Goode, M. J., and Sasser, M. (1980). Prevention-The key to controlling bacterial spot and bacterial speck of tomato. Plant Disease *64*, 831-834.

Grant, S. R., Fisher, E. J., Chang, J. H., Mole, B. M., and Dangl, J. L. (2006). Subterfuge and manipulation: type III effector proteins of phytopathogenic bacteria. Annu Rev Microbiol *60*, 425-449.

Hahn, M. W., Mezey, J. G., Begun, D. J., Gillespie, J. H., Kern, A. D., Langley, C. H., and Moyle, L. C. (2005). Evolutionary genomics: codon bias and selection on single genomes. Nature *433*, E5-6; discussion E7-8.

Hauck, P., Thilmony, R., and He, S. Y. (2003). A Pseudomonas syringae type III effector suppresses cell wall-based extracellular defense in susceptible Arabidopsis plants. Proc Natl Acad Sci U S A *100*, 8577-8582.

He, S. Y., Nomura, K., and Whittam, T. S. (2004). Type III protein secretion mechanism in mammalian and plant pathogens. Biochim Biophys Acta *1694*, 181-206.

Hirano, S. S., and Upper, C. D. (2000). Bacteria in the leaf ecosystem with emphasis on Pseudomonas syringae-a pathogen, ice nucleus, and epiphyte. Microbiol Mol Biol Rev *64*, 624-653.

Hughes, A. L., and Nei, M. (1992). Models of host-parasite Interaction and MHC polymorphism. Genetics *132*, 863-864.

Huynh, T. V., Dahlbeck, D., and Staskawicz, B. J. (1989). Bacterial blight of soybean: Regulation of a pathogen gene determining host cultivar specificity. Science *245*, 1374-1377.

Jin, Q., Thilmony, R., Zwiesler-Vollick, J., and He, S. Y. (2003). Type III protein secretion in Pseudomonas syringae. Microbes Infect *5*, 301-310.

Joardar, V., Lindeberg, M., Jackson, R. W., Selengut, J., Dodson, R., Brinkac, L. M., Daugherty, S. C., Deboy, R., Durkin, A. S., Giglio, M. G.*, et al.* (2005a). Whole-genome sequence analysis of Pseudomonas syringae pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol *187*, 6488-6498.

Joardar, V., Lindeberg, M., Schneider, D. J., Collmer, A., and Buell, C. R. (2005b). Lineage-specific regions in Pseudomonas syringae pv tomato DC3000. Molecular Plant Pathology *6*, 53-64.

Katagiri, F., Thilmony, R., and He, S. Y. (2002). The Arabidopsis Thaliana-Pseudomonas Syringae Interaction. In The Arabidopsis Book, American Society of Plant Biologists, Rickville, MD.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature *267*, 275-276.
Kimura, M. (1983). Rare variant alleles in the light of the neutral theory. Mol Biol Evol *1*, 84-93.

Lee, C. A. (1997). Type III secretion systems: machines to deliver bacterial proteins into eukaryotic cells? Trends Microbiol *5*, 148-156.

Lindeberg, M., Myers, C. R., Collmer, A., and Schneider, D. J. (2008). Roadmap to New Virulence Determinants in Pseudomonas Syringae: Insights from Comparative Genomics and Genome Organization. Molecular Plant-Microbe Interaction *21*, 685-700.

Lindgren, P. B., Peet, R. C., and Panapoulos, N. J. (1986). Gene cluster of *Pseudomonas syringae* pv. *phaseolicola* controls pathogenicity on bean plants and hypersensitivity on nonhost plants. J Bacteriol *168*, 512-522.

Lorang, J. M., Shen, H., Kobayashi, D., Cooksey, D., and Keen, N. T. (1994). *avrA* and *avrE* in *Pseudomonas syringae* pv. *tomato* PT23 play a role in virulence on tomato plants. Mol Plant-Microbe Interact *7*, 208-215.

Miyata, T., and Yasunaga, T. (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol *16*, 23-36.
Mudgett, M., and Staskawicz, B. (1999). Characterization of the Pseudomonas syringae pv. tomato AvrRpt2 protein: demonstration of secretion and processing during bacterial pathogenesis. Mol Microbiol *32*, 927-941.

Mudgett, M. B. (2005). New insights to the function of phytopathogenic bacterial type III effectors in plants. Annu Rev Plant Biol *56*, 509-531.

Nawrath, C., and Métraux, J.-P. (1999). Salicylic acid induction-deficient mutants of Arabidopsis express PR-2 and PR-5 and accumulate high levels of camalexin after pathogen attack. Plant Cell *11*, 1393-1404.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol *3*, 418-426.

Nielsen, R., and Hubisz, M. J. (2005). Evolutionary genomics: detecting selection needs comparative data. Nature *433*, E6; discussion E7-8.

Peet, R. C., Lindgren, P. B., Willis, D. K., and Panopoulos, N. J. (1986). Identification and cloning of genes involved in phaseolotoxin production by Pseudomonas syringae pv. "phaseolicola". J Bacteriol *166*, 1096-1105.

Petersen, L., Bollback, J. P., Dimmic, M., Hubisz, M., and Nielsen, R. (2007). Genes under positive selection in Escherichia coli. Genome Res *17*, 1336-1343. Epub 2007 Aug 1333.

Plotkin, J. B., Dushoff, J., and Fraser, H. B. (2004). Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. Nature *428*, 942-945.

Rahme, L. G., Mindrinos, M. N., and Panapoulos, N. J. (1992). Plant and environmental sensory signals control the expression of *hrp* genes in *Pseudomonas syringae* pv. *phaseolicola*. J Bacteriol *174*, 3499-3507.

Rico, A., and Preston, G. M. (2008). Pseudomonas syringae pv. tomato DC3000 uses constitutive and apoplast-induced nutrient assimilation pathways to catabolize nutrients that are abundant in the tomato apoplast. Mol Plant Microbe Interact *21*, 269-282.

Ritter, C., and Dangl, J. L. (1995). The *avrRpm1* gene of *Pseudomonas syringae* pv. *maculicola* is required for virulence on Arabidopsis. Mol Plant-Microbe Interact *8*, 444-453.

Rohmer, L., Guttman, D. S., and Dangl, J. L. (2004). Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen Pseudomonas syringae. Genetics *167*, 1341-1360.

Rose, L. E., Michelmore, R. W., and Langley, C. H. (2007). Natural variation in the Pto disease resistance gene within species of wild tomato (Lycopersicon). II. Population genetics of Pto. Genetics *175*, 1307-1319. Epub 2006 Dec 1318.

Sarkar, S. F., Gordon, J. S., Martin, G. B., and Guttman, D. S. (2006). Comparative genomics of host-specific virulence in Pseudomonas syringae. Genetics *174*, 1041-1056. Epub 2006 Sep 1041.

Schilmiller, A. L., and Howe, G. A. (2005). Systemic signaling in the wound response. Curr Opin Plant Biol *8*, 369-377.

Tsiamis, G., Mansfield, J. W., Hockenhull, R., Jackson, R. W., Sesma, A., Athanassopoulos, E., Bennett, M. A., Stevens, C., Vivian, A., Taylor, J. D., and Murillo, J. (2000). Cultivar-specific avirulence and virulence functions assigned to *avrPphF* in *Pseudomonas syringae* pv. *phaseolicola*, the cause of bean halo-blight disease. EMBO J *19*, 3204-3214.

Underwood, W., Zhang, S., and He, S. Y. (2007). The Pseudomonas syringae type III effector tyrosine phosphatase HopAO1 suppresses innate immunity in Arabidopsis thaliana. Plant J *52*, 658-672. Epub 2007 Sep 2018.

Van der Hoorn, R. A., De Wit, P. J., and Joosten, M. H. (2002). Balancing selection favors guarding resistance proteins. Trends Plant Sci *7*, 67-71.

Wei, J., Goldberg, M. B., Burland, V., Venkatesan, M. M., Deng, W., Fournier, G., Mayhew, G. F., Plunkett, G., 3rd, Rose, D. J., Darling, A.*, et al.* (2003). Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect Immun *71*, 2775-2786.

Whalen, M. C., Innes, R. W., Bent, A. F., and Staskawicz, B. J. (1991). Identification of *Pseudomonas syringae* pathogens of *Arabidopsis* and a bacterial locus determining avirulence on both Arabidopsis and soybean. Plant Cell *3*, 49-59.

Wildermuth, M. C., Dewdney, J., Wu, G., and Ausubel, F. M. (2001). Isochorismate synthase is required to synthesize salicylic acid for plant defence. Nature *414*, 562-565.

Wolfgang, M. C., Kulasekara, B. R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C. G., and Lory, S. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of Pseudomonas aeruginosa. Proc Natl Acad Sci U S A *100*, 8484-8489. Epub 2003 Jun 8418.

Xiang, T., Zong, N., Zou, Y., Wu, Y., Zhang, J., Xing, W., Li, Y., Tang, X., Zhu, L., Chai, J., and Zhou, J. M. (2008). Pseudomonas syringae effector AvrPto blocks innate immunity by targeting receptor kinases. Curr Biol *18*, 74-80. Epub 2007 Dec 2027.

Xiao, Y., Heu, S., Yi, J., Lu, Y., and Hutcheson, S. W. (1994). Identification of a putative alternate sigma factor and characterization of a multicomponent regulatory cascade controlling the expression of *Pseudomonas syringae* pv. *syringae* Pss61 *hrp* and *hrmA* genes. J Bacteriol *176*, 1025-1036.

Yan, S., Liu, H., Mohr, T. J., Jenrette, J., Chiodini, R., Zaccardelli, M., Setubal, J. C., and Vinatzer, B. A. (2008). Role of recombination in the evolution of the model plant pathogen Pseudomonas syringae pv. tomato DC3000, a very atypical tomato strain. Appl Environ Microbiol *74*, 3171-3181. Epub 2008 Mar 3131.

Young, J. M., Kuykendall, L. D., Martinez-Romero, E., Kerr, A., and Sawada, H. (2001). A revision of Rhizobium Frank 1889, with an emended description of the genus, and the inclusion of all species of Agrobacterium Conn 1942 and Allorhizobium undicola de Lajudie et al. 1998 as new combinations: Rhizobium radiobacter, R. rhizogenes, R. rubi, R. undicola and R. vitis. Int J Syst Evol Microbiol *51*, 89-103.

Zipfel, C., Robatzek, S., Navarro, L., Oakeley, E. J., Jones, J. D., Felix, G., and Boller, T. (2004). Bacterial disease resistance in Arabidopsis through flagellin perception. Nature *428*, 764-767.

Chapter 2

Codon Volatility Identifies Candidate Virulence Factors

Abstract:

Bacteria become virulent pathogens by acquiring or evolving genes beneficial to their interaction with specific hosts. We are interested in the specific interactions between the host plant *Arabidopsis thaliana* and the bacterial pathogen *Pseudomonas syringae* pathovar tomato strain DC3000 (*Pto*$_{DC3000}$), as discussed in Chapter 1. It has been demonstrated in other pathogens that genes necessary for successful infection and disease are sometimes rapidly evolving, possibly as part of a mechanism to evade evolving host defenses. Utilizing bioinformatics and comparative genomics tools, I attempted to identify novel *Pto*$_{DC3000}$ virulence genes based on their purported likelihood of recent molecular evolution as measured by codon volatility. This bioinformatic approach requires only a single whole genome sequence and evaluates the percentage of so-called "volatile" codons in each gene. Volatile genes, rich in volatile codons, are predicted to be rapidly evolving compared to the rest of the genome. I identified candidate genes that are both highly volatile and specific to plant pathogenic *Pseudomonads*. This generated a rank order candidate list of 12 genes. Most candidate genes (70%) have no predicted function. Three of the top 12 candidate genes have been previously shown to function in *Pto*$_{DC3000}$ virulence supporting the hypothesis that volatility may be useful in identifying virulence factors.

Results:

I performed a bioinformatics screen to identify evolving genes that might be implicated in plant pathogenesis. The screen was based on the concept of codon volatility (Plotkin et al., 2004). Volatility measures the distribution of certain "volatile" codons throughout a gene and compares the proportion of volatile codons to that of other genes in the genome. Plotkin et al. reported that volatile codons are indicative of recent molecular evolution because they are likely the result of a recent nucleotide substitution (Figure 2.1). The volatile codons encode amino acids serine, glycine, leucine, and arginine. Each codon is assigned a volatility score based on the likelihood that it arose as the result of a single nucleotide change that alters the amino acid coding sequence. For instance, consider two codons encoding arginine: CGA and AGA. CGA has 4 non-synonymous possible progenitor codons, 4 synonymous possible progenitors, and one stop codon progenitor. This gives a volatility score of 4/8, as stop codons are excluded from consideration because it is unlikely that a functional gene would arise from one that was non-functional due to the presence of a stop codon. AGA has 6 non-synonymous progenitors, 2 synonymous progenitors, and one stop codon progenitor. This leaves a volatility score of 6/8; therefore AGA is more volatile than CGA (Figure 2.1).

When volatility was proposed as a method to identify rapidly evolving genes, Plotkin et al. (2004) reported correlative data that genes with high percentages of volatile codons, termed "volatile genes" here, were also undergoing rapid molecular evolution as evaluated by a comparative method (Plotkin et al., 2004). This correlation between the novel volatility method and the established comparative data was provided for two organisms: the bacterial pathogen *Mycobacterium tuberculosis* and the eukaryotic

pathogen *Plasmodium falciparum*, and demonstrated that highly volatile genes included genes previously shown to have high rates of molecular evolution. Genes with known virulence functions were found among the most volatile in both pathogens. These authors concluded that gene volatility could act as a proxy for a high rate of molecular evolution and further could identify genes required for pathogen virulence. I therefore used the volatility assessment as one approach to identifying genes that are rapidly evolving and may also function in pathogenesis in the model phytopathogen *Pto*$_{DC3000}$.

The volatility assessment compares each gene to others within the same genome and has no inherent value outside of the native genome. While this is not ideal, volatility is unique in that it offered evolutionary insight for organisms where only one genome is sequenced. When this project began, only the *P. syringae* strain *Pto*$_{DC3000}$ genome was complete (Buell et al., 2003), making cross-isolate comparisons impossible using traditional methods such as those described in Chapter 3, below. New complete genome sequences for two *P. syringae* isolates that are bean pathogens, *Psy*$_{B728a}$ and *Pph*$_{1448A,}$ became available later and the comparative approach representing these three pathogens is discussed in Chapter 3 (Feil et al., 2005; Joardar et al., 2005).

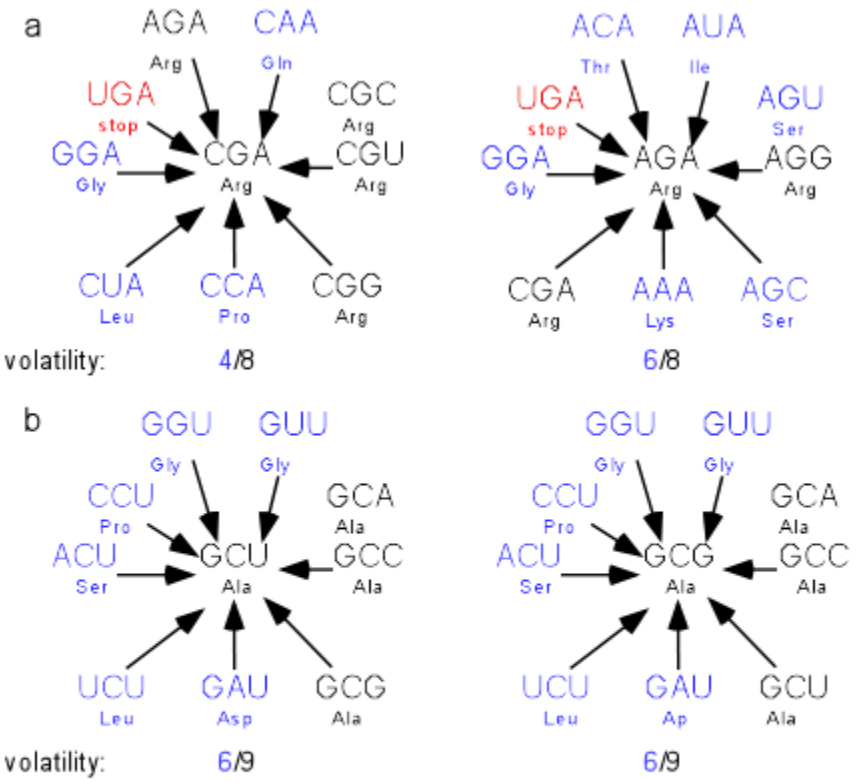


Figure 2.1. Volatility of Codons: Two example codons and their progenitor codons are shown for an amino acids involved in volatility a) and one not involved in volatility b). Synonymous mutations are shown in black and non-synonymous substitutions are in blue. Volatility score for each codon is shown below. Amino acids not involved in volatility have identical volatility scores for all codons while those involved in volatility correspond to codons with differential volatilities.

The annotated open reading frames from the sequenced model pathogen *P. syringae* pv. tomato strain DC3000 (*Pto*$_{DC3000}$) were used as input into the volatility algorithm (Buell et al., 2003; Plotkin et al., 2004). The volatility algorithm assumes a constant GC content and codon bias across the entire genome, but bacterial genomes often evolve by acquisition of genes by horizontal transfer, which is characterized by regions of the genome with aberrant GC content and codon usage patterns that deviate significantly from the core genome (Hacker and Kaper, 2000). In the case of *Pto*$_{DC3000}$,

7% of protein coding genes were found to be part of mobile elements purported to have been gained by horizontal transfer (Buell et al., 2003).

Because the volatility algorithm only works on genomes with constant GC and codon bias, genes with aberrant GC content must be removed from the analysis. The percentage of codons with a G or C as the third nucleotide (GC3) is characteristic of a given genome (Lawrence and Ochman, 1997). In the case of *Pto*$_{DC3000}$, the mean GC3 content is 70.25 with a standard deviation of 8.145% (Rohmer et al., 2004). Any gene with GC3 content outside of 1 standard deviation from the mean was removed from the volatility analysis.

The volatility algorithm also requires a known, or set, transition/transversion mutation rate (κ), which was unknown for *Pto*$_{DC3000}$. I therefore performed the analysis at 9 different κ -values (1.1-3.5). These 9 κ values include two experimentally-determined κ from *P. aeruginosa* PA01 and E. *coli* K12 that have similar GC content to that of *Pto*$_{DC3000}$ (Perna et al., 2001; Spencer et al., 2003). Therefore this is a conservative method to rank genes based on volatility (Table 2.1). Each value of κ resulted in a different volatility *p*-value from which I then obtained the average as the volatility *p*-value for that gene. As reported for *M. tuberculosis*, *p*-values obtained using different κ values correlated ($R^2 = 0.884$ for κ = 1 or 2). While the absolute rank for a given gene may change at different values of κ, the average of the *p*-values represents the consensus of the 9 measurements. A low *p*-value indicates high levels of volatile codons, therefore high volatility. This analysis yielded a list of *Pto*$_{DC3000}$ genes rank ordered from most volatile to least volatile.

Genome	% GC	κ
E. coli K12	50	2.5
Pto_{DC3000}	58.4	unknown
Pph_{1448A}	58	unknown
Psy_{B728a}	59	unknown
P. aeruginosa PA01	66.6	3

Table 2.1: Transition/transversion rates used to calculate volatility *p*-values are within known parameters. GC content and transition/transversion rate (κ) of genomes used as a guide for the κ estimate for Pto_{DC3000}.

Three groups of genes were selected for further analysis: genes of highest volatility, of middle volatility, and lowest volatility. From the raw data generated by volatility analysis of the entire genome, 200 genes from each volatility p-value range were selected. After removal of genes with GC3 content outside of 1 standard deviation from the mean, 121 most volatile, 163 of middle volatility, and 129 least volatile genes remained. Lists of these genes are available on the Dangl lab server in (Users>Emily>legacy). This result demonstrates that genes with aberrant GC3 percentage compared to the mean of the genome are distributed across the volatility spectrum.

Highly volatile genes are not enriched in amino acids encoded by codons with differential volatility

One concern with the codon volatility assessment is that it relies on the genome-wide distribution of codons which encode only four of the twenty amino acids: glycine, serine, leucine, and arginine (Plotkin et al., 2004). All other amino acids are encoded by codons with equivalent volatilities. Chen et al. re-evaluated Plotkin's volatility data for *P. falciparum* and *M. tuberculosis* and showed that codons encoding serine play the largest role in volatility in these strains because the distribution of volatility values of serine codons most closely matched the volatility distribution for the genome as a whole (Chen et al., 2005). Additionally, Pillai et al determined that abundance of arginine was strongly correlated with volatility in the HIV-1 genome (Pillai et al., 2005). To determine whether an overabundance of serine, arginine and leucine codons in the most volatile genes biased our prediction of volatility and possible positive selection, I evaluated the frequency of each of these codons in each of the 3 volatility categories. As a control, I also examined the prevalence of alanine and proline, two amino acids that do not contribute to volatility but are highly represented in the *Pto*$_{DC3000}$ genome (10.6% and 4.7% of encoded amino acids, respectively, compared to an average of 4.3%) (Benson et al., 2006).

As shown in Figure 2.2, the high volatility codons encoding ser, arg, and leu are slightly more prevalent in proteins encoded by highly volatile genes than those of the least volatile genes. The differences in ser, arg, and leu between the three categories of volatile genes are not expected to be significant, though since they are within the 95% confidence interval for the percentages (Figure 2.2). This result, particularly the enrichment is serine, could be the defining feature that makes these genes display high volatility. This is a particular concern in the genome of *Pto*$_{DC3000}$ because the most common serine tRNA (22.6 per thousand codons) in this strain corresponds to the ACG

codon, which is also the most volatile serine codon (volatility of 8/9) (Table 2.1). The other volatile serine codon, AGU (volatility of 8/9) is far less abundant (6.6 per thousand codons) (Table 2.1).

UUU 13.0	UCU 3.8	UAU 9.2	UGU 2.5
UUC 23.2	UCC 9.5	UAC 16.3	UGC 7.3
UUA 2.2	UCA 4.5	UAA 0.8	UGA 1.9
UUG 18.0	UCG 14.4	UAG 0.4	UGG 13.9
CUU 8.8	CCU 7.8	CAU 9.3	CGU 15.5
CUC 14.8	CCC 10.3	CAC 13.6	CGC 30.4
CUA 2.3	CCA 5.9	CAA 12.6	CGA 4.4
CUG 67.4	CCG 23.9	CAG 32.9	CGG 8.4
AUU 14.4	ACU 5.9	AAU 9.5	AGU 6.6
AUC 32.3	ACC 27.3	AAC 22.3	AGC 22.6
AUA 2.9	ACA 5.0	AAA 15.7	AGA 2.0
AUG 23.3	ACG 12.8	AAG 21.8	AGG 2.8
GUU 10.0	GCU 13.7	GAU 20.9	GGU 17.6
GUC 24.1	GCC 43.4	GAC 33.5	GGC 44.5
GUA 6.3	GCA 16.3	GAA 30.5	GGA 3.9
GUG 30.6	GCG 32.5	GAG 25.8	GGG 10.3

Table 2.2. Codon frequencies in *Pto*$_{DC3000}$ favor highly volatile serine codon AGC. Codons are listed followed by their frequency per 1000 codons (NCBI). Codons contributing to volatility are blue, control codons encoding alanine and proline are purple, and stop codons are in red. The most common serine codon, AGC, is highlighted in grey. This information is found at http://www.kazusa.or.jp/codon/

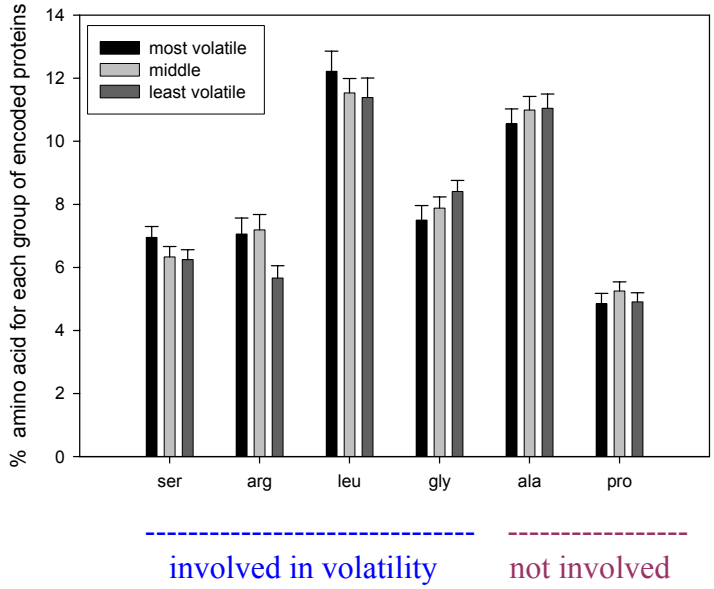


Figure 2.2: Genes in all volatility categories have equal representation in the amino acids involved in volatility. The percentage of the given amino acid in proteins encoded by genes from three volatility score categories does not differ significantlybetween categories with the exception of arg. Serine, arginine, leucine, and glycine contribute to the evaluation of volatility, while alanine and proline do not. Error bars represent 2x standard error or 95% confidence intervals.

Volatility separates predicted proteins based on extreme sizes

51% of the 121 highly volatile genes are hypothetical proteins with no predicted function. This is potentially interesting for subsequent experimental mutational analysis because it may allow functional characterization of novel protein families. It is worrisome, though, because many short predicted ORFs are categorized as "hypothetical" when, in fact, they are pseudogenes (Salzberg et al., 1998). I therefore compared the lengths of ORFs in each volatility score category and found that the lengths are essentially the same (Figure 2.3). When each volatility score category was separated into bins of different protein lengths, the coefficient of variance (cv) was small (<25%) for all

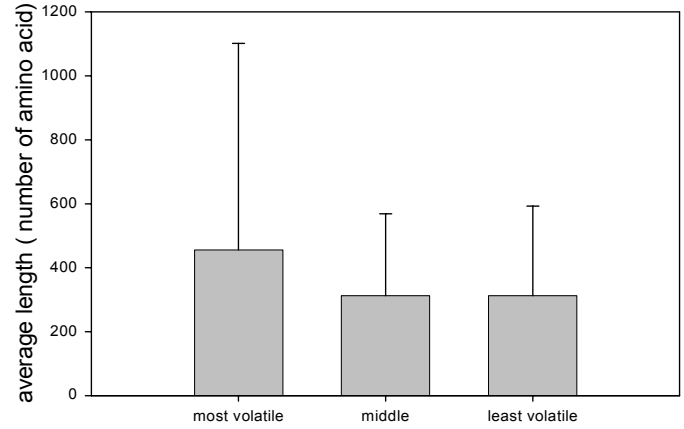
protein lengths except for very large proteins longer than 1000 amino acids in length (cv = 85%), indicating that the standard deviation of open reading frame lengths in the high volatility >1000 a.a. bin is large compared to the mean length in this category. This high coefficient of variance indicates that a few extremely large ORFs are among the most volatile genes. Hence, highly volatile genes were skewed in favor of extremely long ORFs, but were not predominantly short ORFs. Short ORFs, fewer than 400 bases in length, were excluded from potential future mutational analysis because they are potential pseudogenes, while longer predicted ORFS were included because the probability of an open reading frame of 400 bases occurring by chance alone is < 1/10200. Short ORFs predicted to be under 400 bases in length do not include an overabundance of the highly volatile genes (not shown).

Volatility differentiates genes based on homology and predicted localization

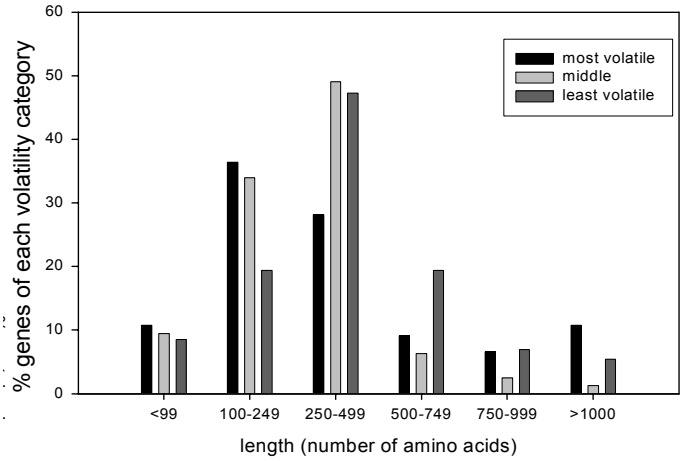
Genes from the three volatility categories were compared based on their homology to other genes. Nucleotide sequences of 121 most volatile genes, 163 middle volatility genes, and 129 least volatile genes were used in a blastx search of genomes of other bacteria. Strains were chosen for comparison based on their evolutionary relationship to *Pto*$_{DC3000}$ (other Pseudomonads) and their ecological niche (plant-associated bacteria). *Pto*$_{DC3000}$ sequences were compared to the genomes of other Pseudomonads (*P. fluorescens* Pf0-1*, P. putida* KT2440*, P. aeruginosa* strain PA01, *Psy*$_{B728a}$, and *Pph*$_{1448a}$), other plant pathogens (*Agrobacterium tumefaciens* C58, *Ralstonia solanacearum* GMI1000*, Xanthomonas campestris* pv. campestris str. ATCC 33913*,* and

Xanthomonas axonopodis pv. campestris strain 301), and two plant symbiotic bacteria (NGR234 and *Sinorhizobium meliloti* strain 1021) (Bell et al., 2004; da Silva et al., 2002; Feil et al., 2005; Freiberg et al., 1997; Galibert et al., 2001; Goodner et al., 2001; Joardar et al., 2005; Nelson et al., 2002; Salanoubat et al., 2002; Copeland, Richardson NCBI submission NC_007492).

a



b



c

	<99	100-249	250-499	500-749	750-1000	>1000
average length (aa)	79.0	174.4	340.6	627.5	851.4	1689.2
Standard deviation	13.8	43.8	73.8	85.4	98.3	1400.1
coefficient of variance	17.4	25.1	21.7	13.6	11.5	82.9

Figure 2.3: Highly volatile genes include an overabundance of very long proteins. a) The average length of highly volatile genes is longer than that of other volatility categories. Error bars represent 2x standard error. b) Histogram showing that a high percentage of the most volatile genes are longer than 1000 amino acids long. c) Table showing average lengths of most volatile proteins in each length bin from b. The coefficient of variance (standard deviation / mean) in the largest protein category is high, meaning that the standard deviation is large compared to the mean. In other words, a few extremely long proteins are included in this group.

I found that the most volatile genes had significantly fewer homologs in these strains than the genes in the middle or low volatility groups. The least volatile genes have on average homologs in 8 of these organisms while the most volatile average only 4.3 homologs and those of middle volatility average 5.6 strains. The distribution of most volatile and least volatile across these genome sequences also generate distinctly different patterns (see Figure 2.4) and these two distributions are significantly different by chi-squared test ($p < 0.001$). This initial analysis was performed using search criteria ($e = 10^{-10}$, Figure 2.4a), and the differential trend was also maintained between the high and low volatility categories at a more-stringent value ($e = 10^{-15}$) (Figure 2.4b).

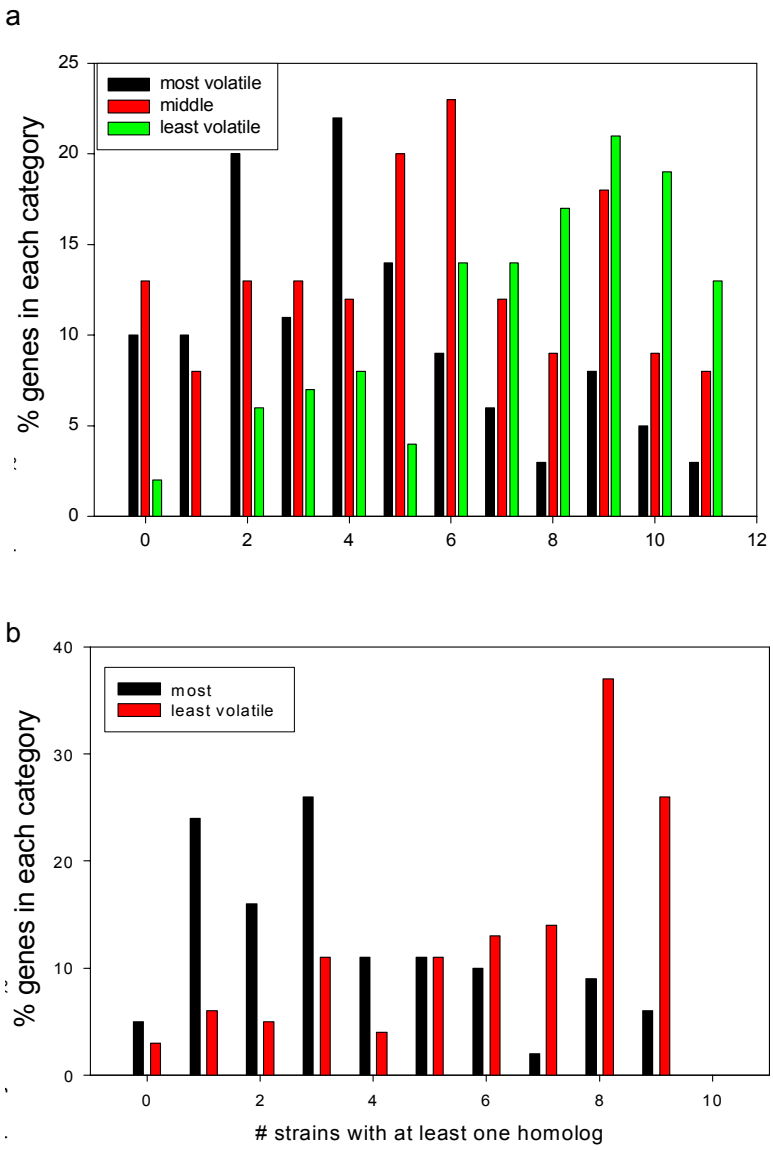


Figure 2.4: Conservation of genes in other bacteria. Genes of the three three volatility categories were used to search 11 bacterial genomes for homologs at two levels of stringency a) $e = 10^{-10}$, b) $e = 10^{-15}$.

The three groups of genes were also compared based on their predicted subcellular localization. Plotkin (2004) showed that highly volatile genes in *M. tuberculosis* were enriched for those involved in virulence, and these were located at the bacterial outer membrane where they directly contact the host immune system (Plotkin et al., 2004). I expected to find the same trend in the volatile genes of Pto_{DC3000}. Analysis

of predicted protein localization, based on homology to known proteins as well as motif signatures, using the program PsortB, revealed that the most-volatile genes included more predicted outer membrane-localized proteins (~12%) than the least-volatile (0%) and more than the genome as a whole (2%) (Gardy et al., 2005).

The distributions throughout various subcellular compartments for the most-volatile and least-volatile gene groups were significantly different based on the Kolmogorov-Smirnov two-sample test evaluating differences in distributions of two samples of continuous observations. In this test, the null hypothesis is that the two distributions (for example, percent genes in each category for most-volatile vs. least volatile) are indistinguishable. In this case (most volatile vs. least volatile) I can reject the null hypothesis because the D value obtained for the data is greater than the D calculated with a *p*-value of 0.01. These *p*-values are based on the number of genes in each category and are described in Biometry (Sokal and Rohlf, 1994). While the statistical test shows that the predicted overall distributions of proteins among subcellular compartments are significantly different for most volatile and least volatile categories, the histogram (Figure 2.5) shows that the most volatile genes have a high percentage of outer membrane proteins compared to the least volatile genes and compared to the genome as a whole. This agrees with Plotkin's findings (2004) that proteins exposed on the outer surface of *M. tuberculosis* are highly volatile (Plotkin et al., 2004).

Another category that is highly represented in the most volatile *Pto*$_{DC3000}$ genes is those encoding proteins of unknown localization. This may be due to the fact that many of the most-volatile genes are of unknown function. Genes annotated as encoding proteins of "unknown function" often lack characterized domains, which are required for

function prediction as well as cellular localization prediction by PsortB (Gardy et al., 2005).

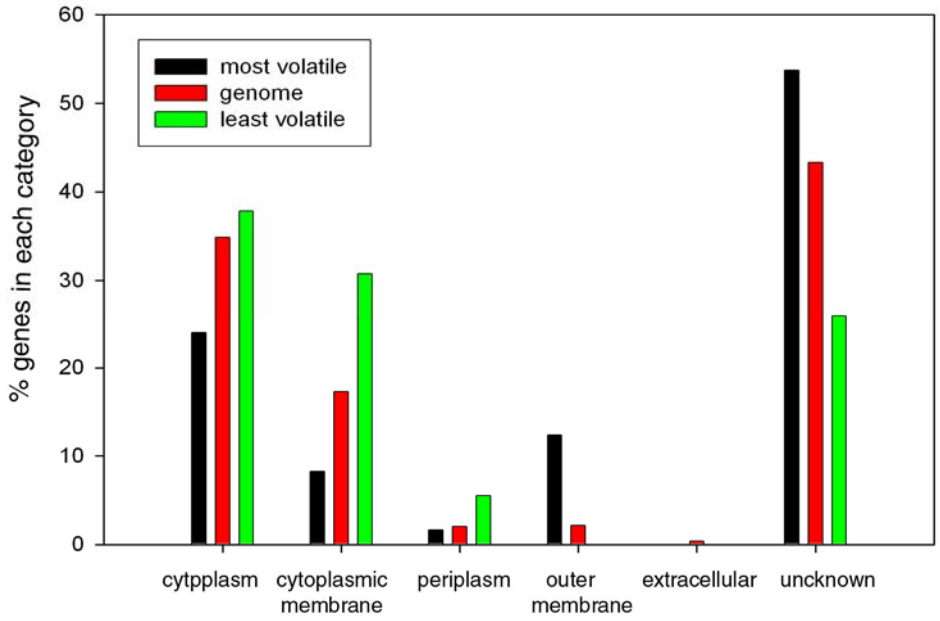


Figure 2.5. Divergence between genes of different volatility categories based on predicted subcellular localization. Most-volatile genes are overrepresented in predicted outer membrane localization.

Volatile genes merit further investigation.

The above analyses were designed to determine whether volatility was likely to predict molecular evolution and potential virulence function, or was a statistical anomaly in which so-called volatile codons correlated with virulence factors in the two genomes initially reported (Plotkin et al., 2004). After rejecting the hypothesis that the most volatile genes of *Pto*$_{DC3000}$ are determined purely due to an overabundance of the amino acids encoded by volatile codons (Figure 2.2), and rejecting the hypothesis that volatile genes of unknown function are short non-gene ORFs (Figure 2.3), I decided that highly

volatile genes were worth pursuing as potential virulence factors. Further analysis demonstrating that the most volatile genes have few homologs (Figure 2.4) and contain an abundance of genes encoding proteins predicted to localize to the bacterial outer membrane supported the hypothesis that the most volatile genes may be involved in sensing the external environment, and perhaps in pathogen virulence (Figure 2.5).

Candidate genes chosen based on conservation in other bacteria

The volatility algorithm ranks each gene in the genome based on the volatility *p*-value that has no reference point outside of the analyzed genome. Because this value has no intrinsic meaning, we prioritized candidates for further study based on other comparative genomics criteria. The purpose of the volatility screen was to identify rapidly evolving genes that might be important for virulence, therefore we chose candidate genes based on their conservation among pathogens and their exclusion from non-pathogens (Table 2.2). Homology searches using blastp ($e=10^{-10}$) identified those genes from the most volatile class that have homologs in either the soil bacterium *P. putida* or the human pathogen *P. aeruginosa* strain PA01. It is important to note that this search was done against the *P. aeruginosa* strain PA01 since it is unable to infect plants, unlike PA14, a strain that is a successful plant pathogen (Plotnikova et al., 2000; Stover et al., 2000). The 89 genes with homologs in these two non-phytopathogens were excluded from our candidate list. Additionally, 1 highly volatile gene (*pspto1585*) was rejected because of its length. Encoding only 63 amino acids, this gene is unlikely to be

real and may be a mistake of the gene identification program, therefore this gene was not considered further.

From *Pto*:		absent from *Ppu* and *PA but present in:*			
	Ppu or *PA*	*Pto* only	*Pto Psy*	*Pto Pph*	*Pto Pph Psy*
whole genome *n* = 3106	2257 (72%)	224 (7%)	87 (3%)	98 (3%)	440 (14%)
most volatile *n* = 121	79 (65%)	13 (11%)	6 (5%)	5 (4%)	18 (15%)
least volatile* *n* = 126	119 (94.4%)	1 (0.8%)	0 (0%)	1 (0.8%)	5 (4%)

Table 2.3: Conservation of most and least volatile genes across *Pseudomonads*. Genes with GC3 content one standard deviation or more from the mean were removed from this comparison and the number of remaining genes is *n*. Homology searches done using BLASTp at e = 10^{-10}. Number of genes in each category is listed with the percentage of each gene group in parentheses. Both the numbers in each category and the percentages for least volatile are significantly different from the whole genome by chi squared test (p = 0.0013). Note that the least volatile genes are, as expected, highly conserved across the Pseudomonads.

The remaining 30 candidate genes are described in Table 2.4. Further homology searches were performed with these candidates against the genomes of plant symbiotic bacteria (*Sinorhizobium meliloti* strain 1021), plant pathogenic bacteria (*Xanthomonas campestris* pv. campestris strain ATCC 33913, *Xanthomonas axonopodis* pv. citri strain 301, *Ralstonia solnacearum* strain GMI1000, *Erwinia caratovora* subspecies atroseptica strain SCRI1043, and *Agrobacterium tumefaciens* strain C58, and *P. syringae* pathovars phaseolicol*a strain* 1448A and syringae strain B728a) (Bell et al., 2004; da Silva et al., 2002; Feil et al., 2005; Galibert et al., 2001; Goodner et al., 2001; Joardar et al., 2005; Salanoubat et al., 2002). The results of these homology searches (blastp e = 10^{-10}) revealed that 11 of the 30 candidates are conserved in phytopathogens and the symbiont, 11 are phytopathogen specific, 3 are conserved only in *P. syringae* genomes, and 5 are found only in the genome of Pto_{DC3000}.

I also performed the volatility analysis on the genome of *Psy*$_{B728a}$ and found that 6 of the 31 candidate genes from *Pto*$_{DC3000}$ have homologs that are also among the most volatile in the *Psy*$_{B728a}$ genome. While volatility does not have meaning across genomes, it is interesting that these genes are among the most volatile in two different genomes of related pathogens. If volatility predicts virulence function, we would expect that these genes are critical for virulence in at least these two pathovars of *P. syringae*.

Candidate gene	Virulence-associated (Lindeberg et al. 2008)	volatile in *Psy*$_{B728a}$ (rank)	predicted function	predicted cellular localization
pspto1045		yes (25)	virulence-associated protein	Cytoplasmic
pspto1330			glycosyl transferase	Unknown
pspto1345			gluconolactonase	Unknown
pspto2062			conserved hypothetical protein	Unknown
pspto2474			DNA/RNA non-specific endonuclease	Unknown
pspto3053			conserved domain protein	Unknown
pspto3383			conserved hypothetical protein	Unknown
pspto4319			conserved hypothetical protein	CytoplasmicMembrane
pspto4325			hypothetical protein	Cytoplasmic
pspto4655			hypothetical protein	Unknown
pspto4832			conserved hypothetical protein	OuterMembrane
pspto0657			hypothetical protein	Cytoplasmic
pspto0714		yes (13)	autotransporter, putative	OuterMembrane
pspto3078			conserved hypothetical protein	Unknown
pspto3200			hypothetical protein	Unknown
pspto3293			hypothetical protein	Unknown
pspto4287	yes	yes (42)	binary cytotoxin	Unknown
pspto4289			conserved hypothetical protein	Unknown
pspto4346			hypothetical protein	Unknown
pspto4611		yes (76)	conserved domain protein	OuterMembrane
pspto4870		yes (85)	hypothetical protein	Unknown
pspto5108			conserved hypothetical protein	Cytoplasmic
pspto0871			macrolide efflux protein	CytoplasmicMembrane
pspto2872		yes (108)	type III effector HopPtoL	CytoplasmicMembrane
pspto3907			conserved hypothetical protein	Cytoplasmic
pspto0011			conserved hypothetical protein	Unknown
pspto2682			hypothetical protein	Unknown
pspto3684			hypothetical protein	Unknown
pspto4704	yes		DNA-binding response regulator CorR	Cytoplasmic
pspto5353	yes		type III chaperone protein SchA	Unknown

Table 2.4. Candidate genes with high volatility and conservation only in the genomes of plant-associated bacteria. Genes listed in blue are conserved in plant pathogens and the symbiotic bacterium S. meliloti, in green are genes conserved in plant pathogens but not the symbiont, purple indicates genes that are specific to *P. syringae* genomes, and black indicates genes specific to *Pto*$_{DC3000}$.

Four highly volatile genes were previously implicated in *P. syringae* pathogenesis

Four of 121 highly volatile genes have been implicated in virulence of *P. syringae*. HopL1 was originally identified as a potential type III effector because of its sequence homology to SrfC, a *Salmonella enterica* protein that is co-regulated with the TTSS encoded by the pathogenicity island SPI-2 (Petnicki-Ocwieja et al., 2002; Worley and Heffron, 2000). SrfC was considered a putative effector protein because its expression is induced by the SPI-1 response regulator SsrB and contains a coiled-coil motif (Petnicki-Ocwieja et al., 2002). HopL1 was found in the genomes of diverse pathogens including *Pectobacteriua* (previously called *Erwinia*), *Salmonella*, and *Yersinia*, as well as the symbiotic bacterium *Mesorhizobium loti* (Grant et al., 2006). This conservation may imply a role for HopL1 in suppression of general host defenses.

HopL1 is secreted into culture media in TTSS-inducing media, but is not translocated into host cells using an HR-inducing *avrRpt2* fusion protein (Chang et al., 2005; Petnicki-Ocwieja et al., 2002). Proteins that are secreted but not translocated may still function as helper proteins with the TTSS by facilitating the penetration of the host cell with the type III pilus. The *Pto*$_{DC3000}$ *hopL1* gene does not contain a *hrp* box in its promoter and is not induced by the TTSS-specific sigma factor HrpL, as shown in both a promoter-trap screen and by microarray expression analysis (Chang et al., 2005; Ferreira et al., 2006). Though no mutational analysis has been performed on this gene, the consensus in the field of *P. syringae* pathogenesis is that HopL1 is not a type III effector (Lindeberg et al., 2006). The potential role of HopL1 as a secreted virulence factor remains unexplored.

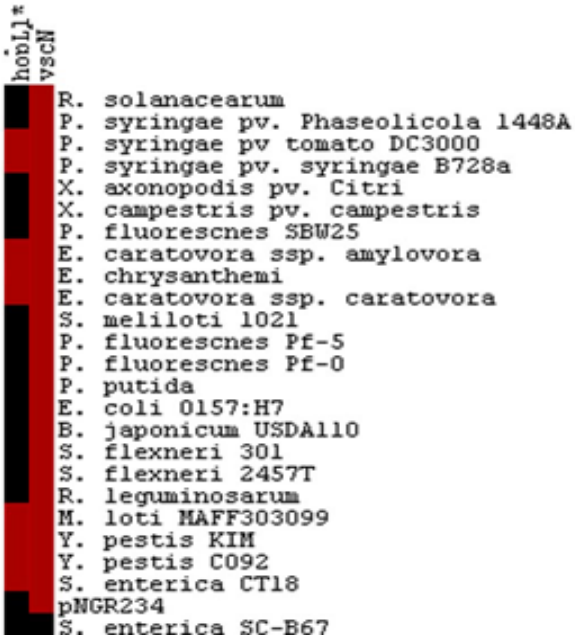


Figure 2.6. Diagram shows presence (red) or absence (black) of genes for *hopL1* and *vscN* based on blastp ($e = 10^{-5}$). The *vscN* gene encodes part of the type III secretion apparatus and is a positive control for the presence of the type III secretion system. Please note that *E. caratovora* subspecies amylovora is now called *Pectobacterium atroceptica* and subspecies carotovora is now called *Pectobacterium carotovora*.

Another of the high volatility candidates found in all three pathovars of *P. syringae* is *shcA*, which encodes a chaperone protein involved in type III secretion. ShcA is required for secretion and translocation of the effector HopA1 into host plant cells in *P. syringae* pv. syringae strain 61 (van Dijk et al., 2002). HopA1 translocation, and therefore ShcA function, is required to elicit the hypersensitive response in *Nicotiana tobacum* plants, demonstrating a role for ShcA in plant resistance. However, a direct role for HopA1 and ShcA in bacterial virulence has not been investigated. Chaperones associated with the TTSS unfold type III effector proteins so that they can be threaded through the type III pilus. This function requires a specific interaction with a specific

type III effector protein, in this case HopA1, but one aspect of chaperone function that may be evolving is the client specificity. *P. syringae* delivers many type III effectors to the host cell during infection and this may be ordered or coordinated via specific interactions with chaperones (Chang et al., 2005). While the chaperone function of ShcA may be functionally constrained against diversification, the client binding region may be diversifying to create a more promiscuous chaperone. If the order of coordination of type III effector delivery is important for pathogenesis, sequence changes in chaperone proteins are expected to be critical to pathogen virulence.

The *corR* gene is unique to *Pto*$_{DC3000}$ and was previously identified as a response regulator required for induction of coronatine biosynthetic genes. Coronatine is a phytotoxin which is required for full virulence in *Pto*$_{DC3000}$ (Brooks et al., 2004; Melotto et al., 2006). Coronatine was shown to be required for full virulence, as coronatine biosynthetic mutants *cfa6* and *cmaA* showed 100-fold less growth than wild-type *Pto*$_{DC3000}$ after 4 days in 4-week old Arabidopsis leaves (Brooks et al., 2004). CorR protein was later shown to bind to *cmaA* and *cfa6* promoters and is required for expression of these genes (Sreedharan et al., 2006). The *corR* mutant also showed 100-fold less growth than wild type *Pto*$_{DC3000}$ 10 days after spray-inoculation of tomato leaves (Sreedharan et al., 2006). Consistent with these findings, I found that the *corR* mutant grew 5-fold less after 3 days in Arabidopsis seedlings (Figure 2.7). I also found that the *corR* mutant grew 100-fold less than wild-type *Pto*$_{DC3000}$ in tomato seedling apoplasts after 6 days (Figure 2.7) (Moneymaker variety Park Seed, Greenville SC). Tomato seedlings inoculated with *corR* mutant bacteria also showed less symptoms than those inoculated with wild-type *Pto*$_{DC3000}$ (Figure 2.7c).

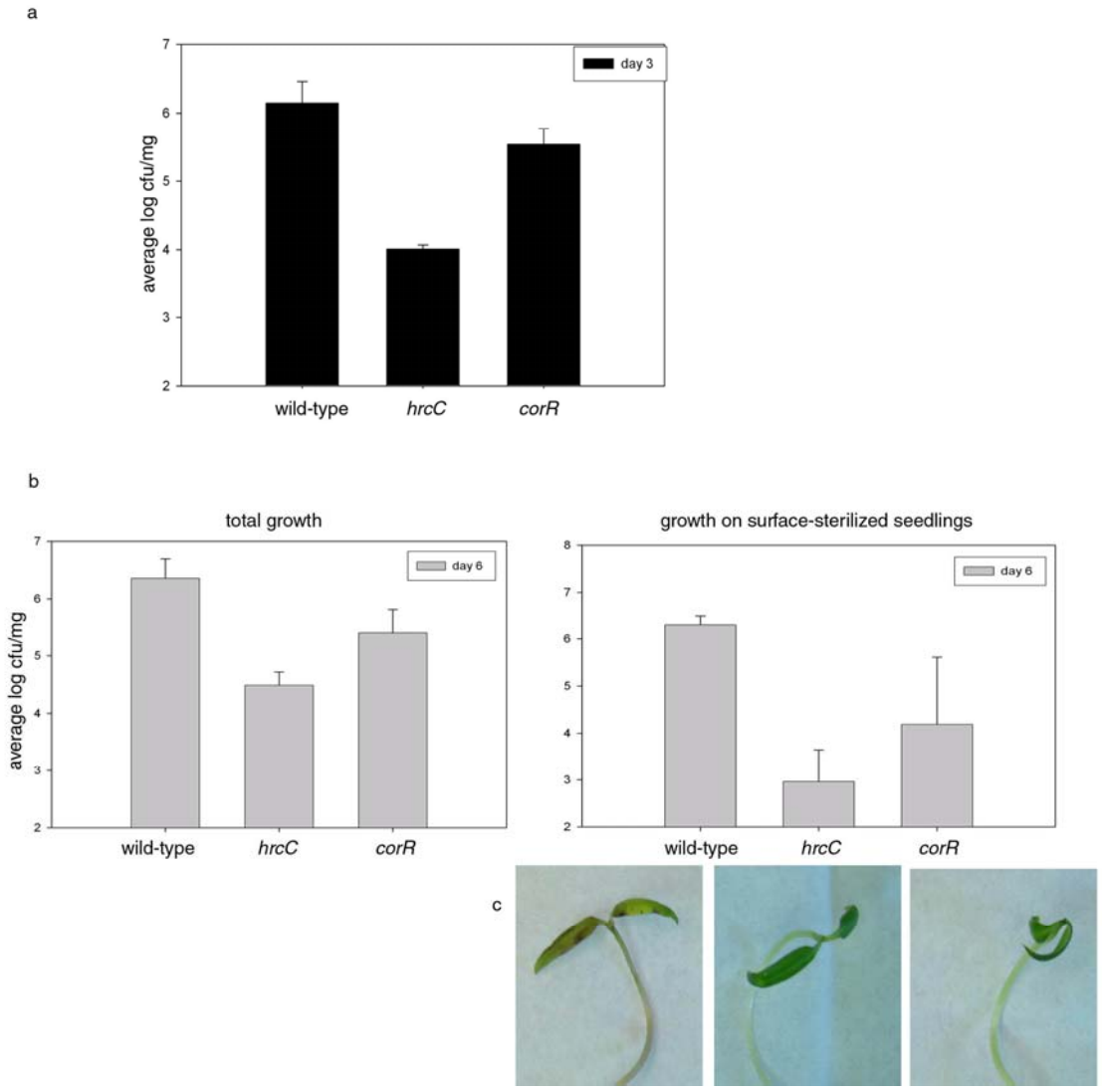


Figure 2.7: Volatile gene *corR* is required for virulence of Pto_{DC3000} on Arabidopsis and tomato. a) *corR* mutant shows 5-fold less growth than wild-type after 3 days in Arabidopsis. This difference is significant by t-test ($p = 0.006$) and is representative of three experiments. b) *corR* mutant grows 10-fold less than wild-type on tomato seedlings and 100-fold less in the tomato seedling apoplast. Error bars represent 2x standard error. c) Photos of representative tomato seedlings on day 6 after inoculation.

Finally, *pspto4287* is homologous to *xaxB* from the bacterium *Xenorhabdus nematophila*. *X. nematophila* is a nematode pathogen that expresses the binary cytotoxin Xax. Xax has been shown to induce apoptosis in both insect and mammalian cells but its effect on plant cells in unknown (Vigneux et al., 2007). However, this gene is included

in a recent review by Lindeberg et al. (2008) reviewing genes potentially relevant to *P. syringae* pathogenesis (as indicated in Table 2.4) and is included in the discussion of these results below (Lindeberg et al., 2008).

Highly volatile genes are enriched in putative virulence factors.

Lindeberg et al. (2008) recently catalogued potential virulence genes in *P. syringae*. These genes included genes with known virulence effect in the three sequenced strains of *P. syringae* (*Pto*$_{DC3000}$, *Psy*$_{B728a}$, and *Pph*$_{1448A}$), genomic regions that are variable among the three sequenced strains, and *P. syringae* homologs of genes required for virulence in other bacteria which included 281 *Pto*$_{DC3000}$ genes (Lindeberg et al., 2008). These 281 genes represent 5.1% of the genes of *Pto*$_{DC3000}$. I used these genes as a guide to show that the highly volatile candidates (Table 2.4) accurately identify genes previously identified as potential virulence factors. I performed two sequential bioinformatic screens to identify the candidate virulence genes listed in Table 2.4. First, I selected 121 genes with high volatility, and second I narrowed my focus to genes from this set that are found in the genomes of plant-associated bacteria but are excluded from *P. putida* KT2440 and *P. aeruginosa* PA01. These are 30 candidate virulence factor genes (Table 2.4). Three of these 30 candidates (*pspto4287*, *pspto4704*, and *pspto5353*) are also considered virulence factor candidates according to Lindeberg et al. (2008). This means that my two screens, together, identified potential virulence factors that are both novel in that many have not been previously identified by any means, and enriched with respect to candidate genes from the genome survey (Lindeberg et al., 2008). In principle,

this could be due entirely to one step in this screen or the other. I found that 9.0% of highly volatile genes alone (using the 121 most volatile genes described above) were among those identified by Lindeberg et al. (2008). Similarly, only excluding genes present in either *P. putida* KT2440 or *P. aeruginosa* PA01 showed that 8.5% of these remaining genes are implicated in virulence. Both of these enrichments are significant, as shown by Fisher's Exact Test and described in Table 2.5.

candidate identification method	number of putative virulence factors	total number of genes in this pool	% of putative virulence factors in pool	*p*-value Fisher's exact test
Lindeberg et al. (2008)	281	5475	5.1%	--
volatility alone	11	121	9.0%	0.03
lifestyle alone	124	1456	8.5%	< 0.001
volatility + lifestyle	3	30	10%	0.0661

Table 2.5: Volatility and bacterial lifestyle enrich resultant candidate gene pools in purported virulence genes. Both highly volatile genes and genes specific to plant-associated bacteria are enriched in potential virulence genes as reported by Lindeberg et al. (2008). *p*-values are the results of Fisher's Exact Test comparing the indicated candidate identification method to the results of Lindeberg et al. (2008). The candidate gene list described in Table 2.4 is enriched in potential virulence factors based on percentage, however, this is only marginally significant, likely as a result of the small number of genes used in this comparison.

Discussion:

I found that 1 of the 31 highly volatile genes I selected for further study were previously shown to function in *Pto*$_{DC3000}$ virulence, while 3 others have a predicted role in virulence. Volatile genes *hopL1* and *shcA* were identified as part of the type III secretion system, a major determinant of *Pto*$_{DC3000}$ virulence on plants (Deng et al., 1998) (Figure 2.7). I also confirmed that the highly-volatile gene *corR*, encoding a response regulator which induces coronatine toxin production, is required for full *Pto*$_{DC3000}$ virulence on Arabidopsis seedlings, consistent with other publications (Rangaswamy and Bender, 2000; Sreedharan et al., 2006).

The concept of using codon volatility and the distribution of codons within a single genome to evaluate evolution is a controversial one. I derived sample sets of genes from 3 places in the volatility spectrum (high, middle, and low), removed those genes with aberrant GC3 to accommodate the limitations of the volatility calculation, and used these three gene sets to address some criticisms of the volatility method.

Plotkin's initial paper received immediate criticism because it relies on only a few amino acids, is heavily influenced by codon usage bias, and because random assignation of volatility values for codons recapitulates several patterns reported by Plotkin. Chen et al. (2005) found that in *M. tuberculosis*, volatility correlated strictly with the codons encoding serine (Chen et al., 2005). In *Pto*$_{DC3000}$, the most common serine codon, AGA, is also the codon with the highest volatility. I addressed this criticism by comparing the frequency of serine, leucine, glycine, and arginine, the amino acids contributing to volatility, in three groups of genes from the top, middle and bottom of the volatility

spectrum. In *Pto*$_{DC3000}$, the percent of serine among these sets is approximately equal, therefore we do not believe that genes are highly volatile due simply to an overabundance of serines (Figure 2.2).

If volatility correlates with genes undergoing rapid molecular evolution, I expect those genes on the low end of the volatility rank order to include genes under purifying selection such as housekeeping genes that have essential function and are therefore constrained against sequence diversification. I showed that the least volatile genes were highly conserved throughout diverse bacteria compared to the most volatile genes. On the other hand, I expect genes that are diversifying in one genome would be more difficult to detect by BLAST homology searches, therefore I expect genes evolving away from homology would not identify homologs in other organisms. I found that the most volatile genes have fewer homologs in other genomes than the least volatile genes, which is consistent with the prediction that the most volatile genes should be under diversifying selection (Figure 2.4).

Plotkin (2004) showed that highly volatile genes of *M. tuberculosis* and *P. falciparum* included genes that encode proteins on the outer surface of the pathogen, where they may interact directly with the host immune system (Plotkin et al., 2004). I found that the most volatile genes of *Pto*$_{DC3000}$ were also likely to encode predicted outer membrane proteins compared to the least-volatile genes. This also supports our underlying hypothesis that proteins encoded by the virulence genes of *Pto*$_{DC3000}$ will be in contact with the environment, and potentially under selective pressure from the host plant immune system. Proteins on the outer membrane of the bacteria have the potential to interact directly with the host cell and trigger host defense responses, thus exposing them

to selective pressure by the host immune system. Flagellin, for example, comes into contact with host cells during the *Pto*$_{DC3000}$-Arabidopsis interaction and triggers host basal defense via the FLS2 receptor (Gómez-Gómez and Boller, 2000). Flagellin, though, is a protein conserved throughout many bacteria and is required for swimming motility (Felix et al., 1999; Shimizu et al., 2003). In bacteria for which swimming motility is required for survival, there will be functional constraints that limit diversification of flagellin-encoding and flagellin assembly genes. Volatility has the potential to identify novel proteins directly contacted by the host cell and, indeed, identified many outer membrane proteins in *Pto*$_{DC3000}$ as well as in *M. tuberculosis* and *P. falciparum (Plotkin et al., 2004)*.

A major drawback of volatility as a method to identify genes undergoing molecular evolution is that it assumes a fairly homogenous genome and requires constant transition/transversion (κ) mutation rate as well as codon use throughout the genome (Plotkin et al., 2004). Bacterial genomes, though, evolve partly through horizontal transfer and therefore include regions with variable κ, GC3 content, and codon usage. Many *Pto*$_{DC3000}$ genes (1610/5726 or 28% of genes) were ignored in this discussion of volatility because they have GC3 content outside of 1 standard deviation from the mean GC3. Among those genes excluded from analysis are many type III effectors and the TTSS itself, which are known virulence factors (Grant et al., 2006). Though one could argue that it is logical to consider potentially horizontally transferred genes independently as possible virulence factors, eliminating these known virulence determinants from my analysis with volatility means that positive controls with characterized roles in virulence are missing from this analysis.

Another major criticism of volatility in identifying genes under positive selection is that the rank order of *Pto*$_{DC3000}$ genes according to volatility *p*-values within a genome has no value outside of that genome. I attempted to add weight to the volatility assessment within *Pto*$_{DC3000}$ by choosing to focus on genes that are volatile in other genomes (Table 2.3) or are unique to *Pto*$_{DC3000}$ (Table 2.3), since genes with no identifiable homolog are considered by volatility but not by the comparative method comparing the rate of non-synonymous mutations (dN) to synonymous mutations (dS) as discussed in Chapter 3.

Other groups evaluated volatility by comparing it to genome-wide dN/dS (Pillai et al., 2005; Stoletzki et al., 2005). Correlation varied based on the organism of interest and volatility accurately predicted molecular evolution in four *Saccaromyces* isolates, but not in the HIV-1 genome (Pillai et al., 2005; Stoletzki et al., 2005). The correlation in yeast genomes was later shown to be an artifact of the codon use bias in those organisms, casting doubt on the meaning of volatility as it relates to measurable sequence diversification (Friedman and Perrimon, 2004). It became clear that the comparative evaluation of evolution would be more robust as a means to select genes for functional testing in order to understand the selective effect of the plant host immune system on *Pto*$_{DC3000}$ genes. As more *P. syringae* genome sequences became available, I used the comparative genomics based strategy to identify genes undergoing molecular evolution in *Pto*$_{DC3000}$ and those results are presented in Chapter 3.

The most striking result of the volatility analysis is described in Table 2.2. Genes with very low volatility ranks were nearly all conserved in related Pseudomonads. This supported our hypothesis that volatility rank would correlate inversely with the rate of

evolution in these genes. Genes with low evolution rates often have essential function (Jordan et al., 2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria and are therefore found in many bacteria. If volatility predicts evolution rate, those with low rank should be essential and are expected to be conserved in other bacteria. However, of potential interest is the concept that the few genes with low volatility that are only found in *P. syringae* isolates may therefore have essential functions required only in these bacterial lineages.

Another striking result of the volatility analysis of *Pto*$_{DC3000}$ is that a high proportion of volatile genes are predicted to localize to the bacterial outer membrane where they could interact directly with the environment and with host plants (Figure 2.5). Both the environment and the host plant immune system were presumed to act as selective forces on the genome of phytopathogens, therefore this finding agrees with that hypothesis. If one were interested specifically in outer membrane or extracellular proteins, there are 138 of these in the *Pto*$_{DC3000}$ genome based on PsortB (118 outer membrane and 20 extracellular). Creating 138 directed mutations is a large undertaking, especially in a strain that is not particularly amenable to high-throughput directed insertion mutations (discussed in Chapter 3). I used volatility to focus our attention on proteins that may be under selective pressure, but evaluation of the efficacy of volatility would require knockout mutations in all 15 volatile outer membrane proteins and a control set of mutations in 15 randomly-chosen outer membrane proteins. Alternatively, the comparative method evaluating dN/dS in these genes can be used to evaluate evolution throughout the *Pto*$_{DC3000}$ genome by a second method (Chapter 3) in order to

support the hypothesis that volatile genes are evolving due to selective pressure exerted by the plant immune system.

Volatility may represent some characteristic of the genome that is independent of real molecular evolution as measured by dN/dS after comparison of homologs, but is still relevant to bacterial virulence. Some volatile genes from both *P. falciparum* and *M. tuberculosis* are rapidly-evolving and relevant to disease. For example, in *M. tuberculosis*, PPE and PE-PGRS genes are highly volatile and have high dN/dS as well as a role in virulence in macrophages (Plotkin et al., 2004; Ramakrishnan et al., 2000; Sassetti et al., 2003). In *P. falciparum*, several known antigens were among the most volatile genes, though a subset of EMP1 antigens that was shown not to bind host-specific receptors was not volatile, suggesting that volatility preferentially identifies proteins in direct contact with the host immune system. Even if volatility does not predict evolution, per se, it may be useful in identifying virulence factors.
Table 2.5 shows that volatility alone significantly enriches the candidate gene pool for potential virulence factors.

Potential virulence factors were defined by Lindeberg et al. (2008) as genes located in non-syntenic regions in the *P. syringae* genomes, genes previously-reported to function in virulence, and homologs in *P. syringae* of genes required for virulence in other systems. This collection of 281 genes includes many genes whose role in virulence has not been empirically demonstrated; therefore is an inclusive catalogue of possible virulence determinants. The two successive screens used to create my candidate gene list (Table 2.4) were based first on high volatility and second on bacterial lifestyle. That is, a gene must be highly volatile and absent from genomes of Pseudomonads that do not

interact with plants in order to be considered a candidate. Both high volatility and plant-associated lifestyle alone successfully enrich for potential virulence factors (Table 2.5). The combination of the two screens, resulting in the 30-gene candidate list, is only marginally significantly enriched in virulence factors according to the Fisher's Exact Test (Table 2.5), but I believe this is due to the small number of genes on this list, since both steps of the selection process have a significant impact on the proportion of virulence factors on the resulting candidate pools. Furthermore, the genes described by Lindeberg et al. (2008) will not include any novel virulence genes that are not regularly horizontally transmitted. In contrast, the goal of my study was to find novel virulence factors regardless of origin. The enrichment in reported virulence factors in my candidate list acts as a positive control for identification of virulence genes. Future mutational analysis is required to determine whether volatility and bacterial lifestyle, used to generate the candidate list in Table 2.4, can identify novel virulence factors not addressed in previous work (Lindeberg et al., 2008).

Materials and Methods

Volatility in the *Pto*$_{DC3000}$ genome

Volatility was evaluated at 9 different values of κ using the algorithm described by Plotkin et al. http://mathbio.sas.upenn.edu/volatility/cgi-bin/volatility.pl *p*-values were calculated using the following κ values: 1.1, 1.4, 1.7, 2.0, 2.3, 2.6, 2.9, 3.2, and 3.5, as recommended by Plotkin (personal communication). The mean of the resulting *p*-values was used as the "volatility *p*-value" in creating rank order of genes.

Sequence homology searches

The Basic Linear Alignment Search Tool from NCBI was used to find homologs of the genes of *Pto*$_{DC3000}$.

Prediction of subcellular localization

PsortB version 2.0 is available at www.psort.org/psortb

Bacterial Growth in Arabidopsis

Dip inoculations: Bacteria were suspended in 10mM $MgCl_2$ containing 200µl/l silwet at an OD_{600} = 0.05 which is 2.5^x10^7 cfu/ml. *Pto*$_{DC3000}$ wild-type and a *Pto*$_{DC3000}$ strain containing a disruption of *hrcC* (He lab) were used as positive and negative controls, respectively. 2-week old Arabidopsis seedlings were dipped in the suspension and plants were harvested 1 hour (day 0), 1 day (day 1), and 3 days (day 3) after dip inoculation. Each bar represents 4 tubes containing 1mL of 10mM $MgCl_2$ containing 200µl/L silwet and 3 seedlings after 1 hour shaking at 28°C. Samples were serially diluted, plated on KB Rif, and counted after 1 day at 28°C. Another set of plants was surface-sterilized in 70% Ethanol at each timepoint described above. After 30 seconds in 70% ethanol, seedlings were plunged into 1 liter of water to remove ethanol. Plants were

blotted with paper towels and added to 1ml 10mM $MgCl_2$, shaken, diluted, and quantified as described above for the non-sterilized samples.

Bacterial Growth in Tomato

Bacterial growth on tomato seedlings was performed by adapting the protocol of {Uppalapati, 2008 #4788}. Tomato seeds of the Moneymaker variety (Park Seed Wholesale, Greenwood, SC) were sterilized by shaking in 70% ethanol for 20 minutes followed by shaking in 100% bleach for 20 minutes. Seeds were then washed 4 times in excess sterile water and plated on plates containing MS medium with Gamborg vitamins and 0.8% agar. These plates were placed in the dark for approximately 7 days until the hypocotyls emerged. Bacteria were suspended in 10mM $MgCl_2$ plus 200ul/L silwet at an OD = 0.1 ($5 * 10^7$ cfu/ml). Seedlings were inoculated using a 10ml pipet so that each leaf got at least one drop of inoculum. Bacterial suspension was removed with a sterile pipet 5 minutes after inoculation. Plates were sealed with parafilm and placed in an incubator with 12 hour light. Samples were taken 3 and 6 days after inoculation. Leaves were removed from seedling stems and were surface-sterilized in 70% ethanol for 30 seconds before being immersed in 1L water to remove ethanol. Leaves were blotted and moved to sterile tubes containing 1ml 10mM MgCl2 with 200µl/l silwet. Tubes were shaken for 1 hour then 200µl were removed, serially diluted, and plated as for Arabidopsis dip growth curves. Bacterial colonies were counted 24-48 hours later and expressed as cfu/mg of plant tissue. Error bars represent 2x standard error or approximately 95% confidence.

Statistics

χ^2 tests were performed in SigmaPlot (Systat Software).

Kolmogorov-Smirnov two-sample test evaluating differences in distributions of two samples of continuous observations were performed as described in Biometry {Sokal, 1994 #4783}. D values were obtained from box 13.9 pg 435.

Fisher's Exact Test was used in Table 2.5 because it is preferred over χ^2 when the numbers are small, as with the number of volatile genes previously reported to be potential virulence factors. This test was performed using GraphPad available online: http://www.graphpad.com/quickcalcs/index.cfm

References:

Bell, K. S., Sebaihia, M., Pritchard, L., Holden, M. T., Hyman, L. J., Holeva, M. C., Thomson, N. R., Bentley, S. D., Churcher, L. J., Mungall, K.*, et al.* (2004). Genome sequence of the enterobacterial phytopathogen Erwinia carotovora subsp. atroseptica and characterization of virulence factors. Proc Natl Acad Sci U S A *101*, 11105-11110.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2006). GenBank. Nucleic Acids Res *34*, D16-20.

Brooks, D. M., Hernandez-Guzman, G., Kloek, A. P., Alarcon-Chaidez, F., Sreedharan, A., Rangaswamy, V., Penaloza-Vazquez, A., Bender, C. L., and Kunkel, B. N. (2004). Identification and characterization of a well-defined series of coronatine biosynthetic mutants of Pseudomonas syringae pv. tomato DC3000. Mol Plant Microbe Interact *17*, 162-174.

Buell, C. R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I. T., Gwinn, M. L., Dodson, R. J., Deboy, R. T., Durkin, A. S., Kolonay, J. F.*, et al.* (2003). The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. Proc Natl Acad Sci U S A *100*, 10181-10186.

Chang, J. H., Urbach, J. M., Law, T. F., Arnold, L. W., Hu, A., Gombar, S., Grant, S. R., Ausubel, F. M., and Dangl, J. L. (2005). A high-throughput, near-saturating screen for type III effector genes from Pseudomonas syringae. Proc Natl Acad Sci U S A *102*, 2549-2554.

Chen, Y., Emerson, J. J., and Martin, T. M. (2005). Evolutionary genomics: codon volatility does not detect selection. Nature *433*, E6-7; discussion E7-8.

da Silva, A. C. R., Ferro, J. A., Reinach, F. C., Farah, C. S., Furian, L. R., Quaggio, R. B., Monteiro-Vitorello, C. B., and al., e. (2002). Comparison of the genomes of two *Xanthomonas* pathogens with differeing host specificities. Nature *417*, 459-463.

Deng, W.-L., Preston, G., Collmer, A., Chang, C.-J., and Huang, H.-C. (1998). Characterization of the *hrpC* and *hrpRS* operons of *Pseudomonas syringae* pathovars syringae, tomato, and glycinea and analysis of the ability of *hrpF, hrpG, hrcC, hrpT* and *hrpV* mutants to elicit the hypersensitive response and disease in plants. J Bacteriol *180*, 4523-4531.

Feil, H., Feil, W. S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., Lykidis, A., Trong, S., Nolan, M., Goltsman, E.*, et al.* (2005). Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci U S A *102*, 11064-11069. Epub 12005 Jul 11025.

Felix, G., Duran, J. D., Volko, S., and Boller, T. (1999). Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. Plant J *18*, 265-276.

Ferreira, A. O., Myers, C. R., Gordon, J. S., Martin, G. B., Vencato, M., Collmer, A., Wehling, M. D., Alfano, J. R., Moreno-Hagelsieb, G., Lamboy, W. F.*, et al.* (2006). Whole-genome expression profiling defines the HrpL regulon of Pseudomonas syringae pv. tomato DC3000, allows de novo reconstruction of the Hrp cis clement, and identifies novel coregulated genes. Mol Plant Microbe Interact *19*, 1167-1179.

Freiberg, C., Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal, A., and Perret, X. (1997). Molecular basis of symbiosis between Rhizobium and legumes. Nature *387*, 394-401.

Friedman, A., and Perrimon, N. (2004). Genome-wide high-throughput screens in functional genomics. Curr Opin Genet Dev *14*, 470-476.

Galibert, F., Finan, T. M., Long, S. R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P.*, et al.* (2001). The composite genome of the legume symbiont Sinorhizobium meliloti. Science *293*, 668-672.

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

Gómez-Gómez, L., and Boller, T. (2000). FLS2: An LRR receptor like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. Mol Cell *5*, 1003-1011.

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M., Halling, C.*, et al.* (2001). Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58. Science *294*, 2323-2328.

Grant, S. R., Fisher, E. J., Chang, J. H., Mole, B. M., and Dangl, J. L. (2006). Subterfuge and manipulation: type III effector proteins of phytopathogenic bacteria. Annu Rev Microbiol *60*, 425-449.

Hacker, J., and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol *54*, 641-679.

Joardar, V., Lindeberg, M., Jackson, R. W., Selengut, J., Dodson, R., Brinkac, L. M., Daugherty, S. C., Deboy, R., Durkin, A. S., Giglio, M. G.*, et al.* (2005). Whole-genome sequence analysis of Pseudomonas syringae pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol *187*, 6488-6498.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res *12*, 962-968.

Lawrence, J. G., and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol *44*, 383-397.

Lindeberg, M., Cartinhour, S., Myers, C. R., Schechter, L. M., Schneider, D. J., and Collmer, A. (2006). Closing the circle on the discovery of genes encoding Hrp regulon members and type III secretion system effectors in the genomes of three model Pseudomonas syringae strains. Mol Plant Microbe Interact *19*, 1151-1158.

Lindeberg, M., Myers, C. R., Collmer, A., and Schneider, D. J. (2008). Roadmap to New Virulence Determinants in Pseudomonas Syringae: Insights from Comparative Genomics and Genome Organization. Molecular Plant-Microbe Interaction *21*, 685-700.

Melotto, M., Underwood, W., Koczan, J., Nomura, K., and He, S. Y. (2006). Plant stomata function in innate immunity against bacterial invasion. Cell *126*, 969-980.

Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Martins dos Santos, V. A., Fouts, D. E., Gill, S. R., Pop, M., Holmes, M.*, et al.* (2002). Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440. Environ Microbiol *4*, 799-808.

Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A.*, et al.* (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature *409*, 529-533.

Petnicki-Ocwieja, T., Schneider, D. J., Tam, V. C., Chancey, S. T., Shan, L., Jamir, Y., Schechter, L. M., Janes, M. D., Buell, R., Tang, X.*, et al.* (2002). Genomewide identification of proteins secreted by the Hrp type III protein secretion system of *Pseudomonas syringae* pv. tomato DC3000. Proc Natl Acad Sci USA *99*, 7652-7657.

Pillai, S. K., Kosakovsky Pond, S. L., Woelk, C. H., Richman, D. D., and Smith, D. M. (2005). Codon volatility does not reflect selective pressure on the HIV-1 genome. Virology *336*, 137-143.

Plotkin, J. B., Dushoff, J., and Fraser, H. B. (2004). Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. Nature *428*, 942-945.

Plotnikova, J. M., Rahme, L. G., and Ausubel, F. M. (2000). Pathogenesis of the human opportunistic pathogen Pseudomonas aeruginosa PA14 in Arabidopsis. Plant Physiol *124*, 1766-1774.

Ramakrishnan, L., Federspiel, N. A., and Falkow, S. (2000). Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family. Science *288*, 1436-1439.

Rangaswamy, V., and Bender, C. L. (2000). Phosphorylation of CorS and CorR, regulatory proteins that modulate production of the phytotoxin coronatine in Pseudomonas syringae. FEMS Microbiol Lett *193*, 13-18.

Rohmer, L., Guttman, D. S., and Dangl, J. L. (2004). Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen Pseudomonas syringae. Genetics *167*, 1341-1360.

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J. C., Cattolico, L.*, et al.* (2002). Genome sequence of the plant pathogen Ralstonia solanacearum. Nature *415*, 497-502.

Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. Nucl Acids Res *26*, 544-548.

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol *48*, 77-84.

Shimizu, R., Taguchi, F., Marutani, M., Mukaihara, T., Inagaki, Y., Toyoda, K., Shiraishi, T., and Ichinose, Y. (2003). The DeltafliD mutant of Pseudomonas syringae pv. tabaci, which secretes flagellin monomers, induces a strong hypersensitive reaction (HR) in non-host tomato cells. Mol Genet Genomics *269*, 21-30. Epub 2003 Feb 2013.

Sokal, R. R., and Rohlf, F. J. (1994). Biometry, 3 edn, W. H. Freeman).
Spencer, D. H., Kas, A., Smith, E. E., Raymond, C. K., Sims, E. H., Hastings, M., Burns, J. L., Kaul, R., and Olson, M. V. (2003). Whole-genome sequence variation among multiple isolates of Pseudomonas aeruginosa. J Bacteriol *185*, 1316-1325.

Sreedharan, A., Penaloza-Vazquez, A., Kunkel, B. N., and Bender, C. L. (2006). CorR regulates multiple components of virulence in Pseudomonas syringae pv. tomato DC3000. Mol Plant Microbe Interact *19*, 768-779.

Stoletzki, N., Welch, J., Hermisson, J., and Eyre-Walker, A. (2005). A dissection of volatility in yeast. Mol Biol Evol *22*, 2022-2026. Epub 2005 Jun 2015.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M.*, et al.* (2000). Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. Nature *406*, 959-964.

van Dijk, K., Tam, V. C., Records, A. R., Petnicki-Ocwieja, T., and Alfano, J. R. (2002). The ShcA protein is a molecular chaperone that assists in the secretion of the HopPsyA

effector from the type III (Hrp) protein secretion system of Pseudomonas syringae. Mol Microbiol *44*, 1469-1481.

Vigneux, F., Zumbihl, R., Jubelin, G., Ribeiro, C., Poncet, J., Baghdiguian, S., Givaudan, A., and Brehelin, M. (2007). The xaxAB genes encoding a new apoptotic toxin from the insect pathogen Xenorhabdus nematophila are present in plant and human pathogens. J Biol Chem *282*, 9571-9580. Epub 2007 Jan 9517.

Worley, M. J., and Heffron, F. (2000). Identification of exported bacterial proteins via gene fusions to Yersinia pseudotuberculosis invasin. Methods Enzymol *326*, 97-104.

Chapter 3

The Comparative Method Identifies Candidate Virulence Factors

Introduction:

While the promise of identifying evolving genes using a single genome and a simple algorithm was enticing, the value of the volatility has been questioned (discussed in Chapter 2 (Chen et al., 2005; Hahn et al., 2005). Additionally, two additional finished *P. syringae* whole genome sequences became available as my volatility analysis was ongoing. Therefore, we employed a different method to identify evolving *P. syringae* genes, one that was both thoroughly vetted, accepted, and would provide a contrast to my volatility candidate gene list.

Comparing homologous sequences in related organisms, Nei and Gojobori (1986) demonstrated how to evaluate the rate of non-synonymous changes in DNA sequence (dN: those nucleotide changes that alter the encoded protein sequence) to synonymous changes (dS: those that leave the protein sequence unchanged) (Nei and Gojobori, 1986). Synonymous, or silent, changes happen in a manner that is virtually free of selection, though they are affected by the codon bias used by the organism. These silent mutations act as a rough molecular clock and are assumed to occur at a constant rate. Non-synonymous mutations, on the other hand, change the protein coding sequence and are therefore subject to selection based on the resulting protein function. Therefore, a sequence wherein non-synonymous mutations outnumber synonymous mutations is said to be undergoing rapid, likely adaptive, evolution. A protein in which many silent mutations occur but very few (or no) non-synonymous muatations occur is assumed to be functionally constrained against sequence divergence. The dN/dS ratio, therefore, indicates the rate of molecular evolution: $dN/dS > 1$ implies that a gene is under positive

selection and dN/dS < 1 indicates negative selection, while dN/dS approximately equal to 1 represents neutrality (Nei and Gojobori, 1986).

Results:

To identify rapidly evolving genes in the model plant pathogen *P. syringae* pathovar tomato strain DC3000 (Pto_{DC3000}), I utilized for comparisons two additional finished genome sequences: *P. syringae* pathovar syringae strain B728a (Psy_{B728a}), and pathovar phaseolicola strain 1448A (Pph_{1448A}) (Feil et al., 2005; Joardar et al., 2005). In Chapter 2, I discussed the results of volatility analysis of genes from Pto_{DC3000}. I hypothesized that volatile genes would be those undergoing molecular evolution, based on the original publication of this method, which correlated the most volatile genes in two pathogns to genes with previously noted high rates of molecular evolution, as determined by dN/dS (Plotkin et al., 2004). If the volatility method actually identifies genes with high rates of molecular evolution (as measured by dN/dS), the second, classical comparative method should identify an overlapping set of candidate genes for further experimentation. However, the comparative method will be unable to identify selection on the Pto_{DC3000} genes that do not have homologs in either Psy_{B728a} or Pph_{1448A}, so the Pto_{DC3000}-specific volatile candidate genes described in Chapter 2 (Table 2.3), are not expected to be identified in the dN/dS analysis.

***Pph* and *Psy* $_{B728a}$ are more closely related to each other than to *Pto*$_{DC3000}$**

Protein sequences of seven housekeeping genes (*acnB, gapA, gltA, gyrB, pfk, pgi*, and *rpoD*) from each genome were concatenated and used to build a phylogeny in Phylip (Felsenstein, 1989; Sarkar and Guttman, 2004). These proteins are highly conserved throughout bacteria and exhibit few nucleotide substitutions. However, the concatenated sequence file including all 7 proteins includes enough substitutions that a phylogeny can be made. Figure 3.1a shows that the two bean pathogens, *Psy* $_{B728a}$ and *Pph*$_{1448A}$, are more closely related to each other than they are to *Pto*$_{DC3000}$. This information is critical for the three-way comparison to evaluate dN/dS described later in this Chapter, because it enables extrapolation of an ancestoral sequence from each set of homologs based on the evolutionary distance between the three strains.

I used the same technique to place the plant pathogenic *P. syringae* strains into the larger context of related bacteria and this is shown in Figure 3.1b. This larger phylogeny has no scale bar representing substitutions per site, but represents the relative relatedness of the strains and shows that *P. syringae* isolates are more closely related to one another than they are to other types of bacteria, including other Pseudomonads (*P. putida*, and *P. fluorescens*). This shows that the core genomes of the three *P. syringae* strains are highly similar and that any variation among gene sequences in these strains should be recent and potentially due to host-specific niche invasion. It is also interesting to note that this phylogeny separates animal pathogens (Salmonella, Escheria, Shigella, and Yersinia) in to a clade distinct from strains that are phytopathogens, plant-associated bacteria, or plant symbionts.

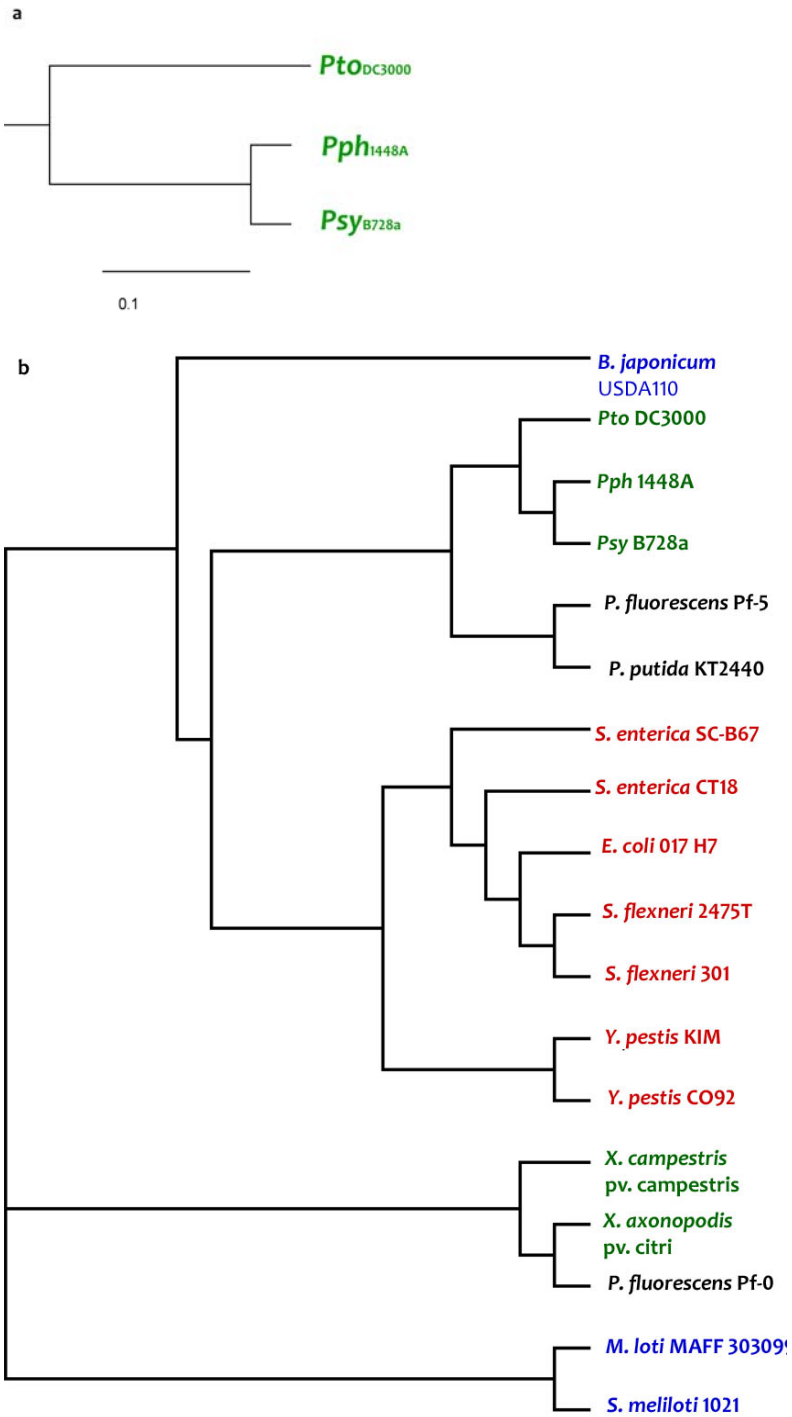


Figure 3.1: 7-gene MLST phylogeny of strains used for the dN/dS comparison. a) tree showing relatedness of the three strains used for dN/dS analysis. Bar represents 0.1 substitutions per site. b) relatedness of P. syringae to other related bacteria. Phytopathogens are listed in green, plant symbionts are in blue, mammalian pathogens are in red, and soil bacteria are in black.

Pairwise comparisons between *Pto* and other genomes

Molecular evolution rate, expressed as dN/dS, can be evaluated using as few as two homologous sequences (Goldman and Yang, 1994). To determine the rate of molecular evolution of *Pto*$_{DC3000}$ genes, we compared each annotated gene from *Pto*$_{DC3000}$ to the genomes of *Psy* $_{B728a}$ and *Pph*$_{1448A}$ using the blastp algorithm with an e- value of 10^{-5} (McGinnis and Madden, 2004). Of the 5492 genes predicted for *Pto*$_{DC3000}$, 4650 had homologs in *Pph*$_{1448A}$ and 4472 had homologs in *Psy* $_{B728a}$. Each of these pairs of sequences was entered into the PAML program (Yang, 1998) to calculate the dN/dS using the yn00 module of PAML. The yn00 module is designed to evaluate dN/dS for pairwise comparisons based on the work of Yang and Nielsen (Yang et al., 2000). This module counts synonymous and non-synonymous sequence changes, corrects for the observed transition/transversion ratio, and estimates the likelihood of multiple successive mutations at the same site which may lead to any base substitution.

Three-way comparison of *Pto*$_{DC3000}$, *Psy* $_{B728a}$, and *Pph*$_{1448A}$

The dN/dS value gains statistical significance as homologous sequences from more strains are included because it incorporates the pattern of sequence divergence (Yang, 1997). The codeml module of PAML uses a phylogeny that represents the evolutionary history of the strains in question and extrapolates ancestor homologs of the current sequences based on that phylogeny (Figure 3.1a). With this predicted ancestor sequence information, the program can identify differential diversification on one branch of the phylogeny compared to the others. There are 3809 genes in *Pto*$_{DC3000}$ that have homologs in both *Psy* $_{B728a}$, and *Pph*$_{1448A.}$ For the three-way analysis, the MLST

phylogeny representing strain relatedness using housekeeping genes (Figure 3.1a) is provided along with each example of the three homologous sequences of interest to the codeml module of PAML. The dN/dS is then calculated by comparing the current *Pto*$_{DC3000}$ sequence to the predicted ancestor sequence. This comparison can reveal evolution specific to the *Pto*$_{DC3000}$ branch of the phylogeny that is not seen in any pairwise comparison. The three-way comparison has more statistical power than the pairwise calculations and this is enhanced as more homologs are added (discussed in chapter 3).

20 candidate virulence genes identified by homology and length

Genes with dN/dS greater than 1.0 are generally considered to be under positive selection and may be diversifying in sequence. I chose genes with dN/dS > 0.8 and < 999 for further study. I included genes with dN/dS less than 1.0 in my candidate list because I was interested in diversification that may have occurred very recently or are still occurring and would therefore have only a slightly enlarged dN/dS value from neutrality. A few genes yielded a dN/dS of 999.0 and these were excluded from my analysis because the infinitesimal rate of silent substitutions (dS) casts doubt on the measurement. Most bacterial genes are approximately 1000 bp in length (*Pto*$_{DC3000}$ open reading frames average 995 nucleotides), so 999 non-synonymous mutations in 1000 bases would yield a gene sequence that could not be identified as a homolog. For instance, the gene *pspto4675* had a dN/dS = 999 in the three-way comparison even though it encodes only 257 amino acids. In both pairwise comparisons, the dN/dS for this gene is low,

indicating that an alignment error or some other error resulted in the over-inflated dN/dS value of 999.

Ninety one genes of *Pto*$_{DC3000}$ had dN/dS > 0.8 in at least one comparison. Many of these "genes" were annotated as insertional sequences from transposons. Since these are unlikely to function in pathogen virulence, these viral and transposon remnants were removed from the candidate list, leaving 31 candidate genes. To avoid inclusion of non-genes in our analyses, we removed all genes shorter than 300 nucleotides in length. Most of these are annotated as unknown proteins and short unknown proteins are likely to be sequences inaccurately identified as genes during annotation with GLIMMER (Delcher et al., 1999).

Twenty genes remained in the candidate list after the removal of short sequences shorter than 300 nucleotides in length (Table 3.1 and Figure 3.2). Most (**14**) candidates appeared to be evolving compared to the *Psy* $_{B728a}$ homolog. Only **one** candidate was evolving compared to the *Pph*$_{1448A}$ homolog**, two** were evolving in both pairwise (*Pto*$_{DC3000}$- *Psy* $_{B728a}$ and *Pto*$_{DC3000}$- *Pph*$_{1448A}$) comparisons, and **three** were evolving based on the 3-way analysis. It might seem counterintuitive that a gene would have high dN/dS in the 3-way comparison, but not in either pairwise comparison. This is because the 3-way comparison evaluates dN/dS between the *Pto*$_{DC3000}$ allele and the predicted ancestor allele instead of either of the other current alleles and therefore would be detected in the 3-way analysis but neither of the two pairwise analyses if the two alleles are not evolving quickly. These candidates are summarized in Table 3.1.

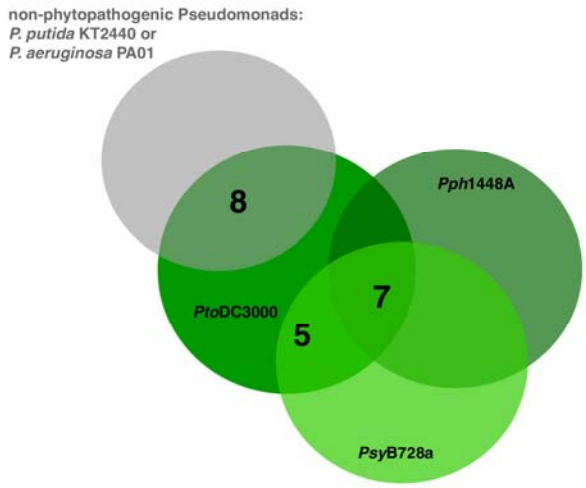


Figure 3.2. Venn diagram showing the distribution of high dN/dS candidate genes from Pto_{DC3000} in representative whole genome sequences of other Pseudomonads. Blastp at e = 10^{-5} shows that 12 candidates are specific to *P. syringae* isolates, while 8 are conserved in non-phytopathogenic Pseudomonads.

Thirteen of these 20 are conserved hypothetical genes with unknown function (Table 3.1). Of the 20 candidate genes, **8** are also conserved in either *P. putida* strain KT2440 or *P. aeruginosa* strain PA01 (Figure 3.2). Conservation in non-phytopathogens may indicate an evolving function outside of pathogenesis, for example in environmental sensing. Seven candidate genes are absent from the non-phytopathogenic Psuedomonad genomes, but are conserved in all three *P. syringae* pathovars and **5** are found only in Pto_{DC3000} and Psy_{B728a} (Figure 3.2).

virulence phenotype	candidate gene *Pto*$_{DC3000}$	relevant comparison	dN/dS	homolog
newly-identified	*pspto5537*	*Psy*$_{B728a}$	3.8	*psyr1314*
	pspto5538	*Psy*$_{B728a}$	2.5	*psyr1849*
	pspto5557	*Psy*$_{B728a}$	1.07	*psyr4126*
	pspto4043	*Psy*$_{B728a}$	1.18	*psyr0670*
	pspto5566	*Psy*$_{B728a}$	7.34	*psyr1051*
	pspto2999	*Psy*$_{B728a,}$ *Pph*$_{1448A}$	0.83, 0.96	*psyr2880, pspph2359*
previously-identified	*avrPto*	*Psy*$_{B728a}$	0.816	*psyr4919*
	pspto0834	*Pph*$_{1448A}$	0.82	*pspph3759*
wild-type phenotype	*pspto3932*	*Psy*$_{B728a,}$ *Pph*$_{1448A}$	1.2, 1.02	*psyr2677, pspph2397*
no mutant obtained	*pspto1269*	*Psy*$_{B728a}$	1.08	*psyr1087*
	pspto1442	*Psy*$_{B728a}$	1.59	*psyr1255*
	pspto1585	3-way	1.08	*psyr4891, pspph4921*
	pspto2512	*Psy*$_{B728a}$	1.9	*psyr2324*
	pspto2020	3-way	0.8	*psyr1829, pspph1789*
	pspto3320	*Psy*$_{B728a}$	0.85	*psyr2181*
	pspto3491	3-way	1.42	*psyr1680, pspph2103*
	pspto3623	*Psy*$_{B728a}$	0.95	*psyr4726*
	pspto4655	*Psy*$_{B728a}$	0.93	*psyr4726*
	pspto5539	*Psy*$_{B728a}$	0.95	*psyr0141*
	pspto5588	*Psy*$_{B728a}$	19.76	*psyr5109*

Table 3.1: Summary of *Pto*$_{DC3000}$ Candidate Gene Identification and Experimental Validation. Genes in blue are those for which we were able to generate loss of function mutation and present functional data. Genes in black are those in which no mutant was obtained, despite repeated attempts. Note that the homologs listed in column 4 are those for which the relevant, high dN/dS, comparison was made and do not include homologs that might have been found in other bacterial species.

dN/dS does not correlate with volatility

If codon volatility accurately reflects DNA sequence evolution, I would expect the candidate lists from the two methods to contain overlapping sets of genes. This was not the case. Further, none of the candidate genes from the dN/dS analysis were among the 200 most volatile genes. Because both candidate lists are short (20 for dN/dS and 200 for volatility), I extended this analysis to the entire genome. Each gene was given a rank based on volatility p-value and a rank based on dN/dS value. Genes with the same p- or

dN/dS value as another gene were given the same rank. Figure 3.3 shows that the volatility and 3-way dN/dS ranks do not correlate with one another. The R^2 value of 0.07 indicates no significant correlation between these ranks.

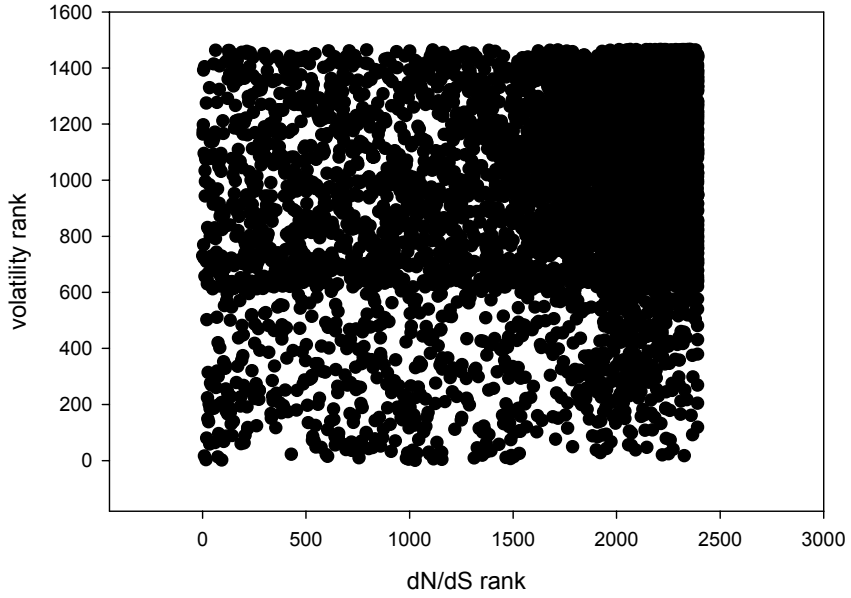


Figure 3.3: No correlation between volatility and dN/dS ranks. Graph shows the volatility and 3-way dN/dS ranks for 5499 genes. Genes with equivalent dN/dS were given the same rank, as were genes with equivalent volatility *p*-values (discussed in Chapter 2).

Mutational Analysis of 9 candidates

We investigated whether the genes identified as having high dN/dS values affected pathogenicity by attempting to mutate all 20 *Pto*$_{DC3000}$ candidate genes from Table 3.1. To this end, we employed single cross-over homologous recombination. When this work was initiated, mutagenesis in *Pto*$_{DC3000}$ via homologous recombination was difficult and rare (Sheng Yang He personal communication), though recently several labs have published improved methods for creating directed mutations and these methods are currently in use in our lab (Wei et al., 2007). In this work, an internal fragment of the

candidate gene was cloned into a suicide vector (House et al., 2004) (Invitrogen). The resulting vector was conjugated into *Pto*$_{DC3000}$ and putative mutants were isolated on selective and confirmed by PCR and sequencing of the plasmid-genome border.

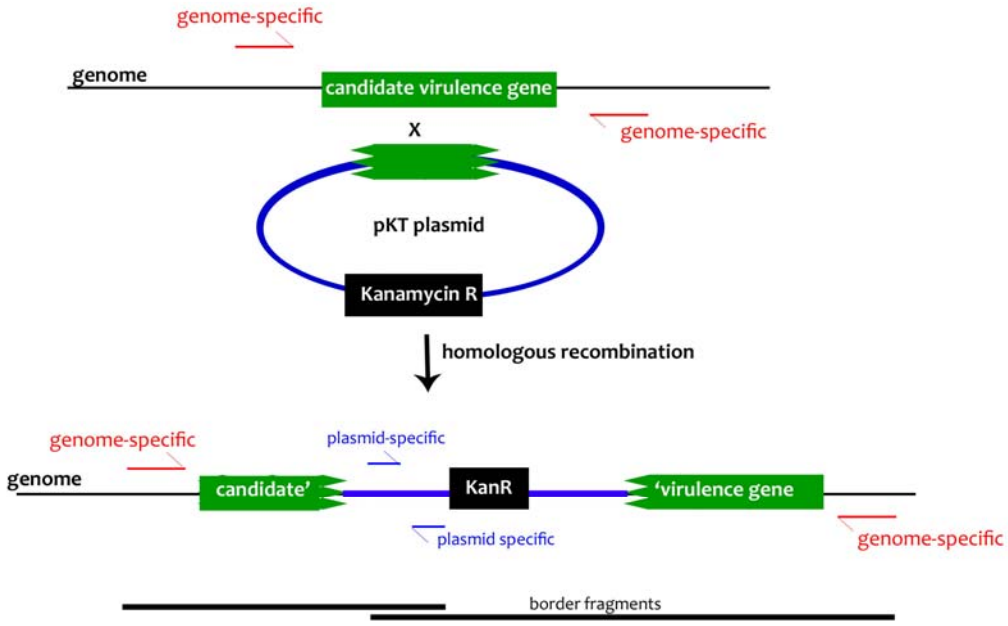


Figure 3.4. An insertional mutation strategy was used to create mutations in candidate genes. The pKT plasmid contains a selectable Kanamycin resistance marker and an internal fragment of the candidate gene. Homologous recombination creates a disruption of the candidate gene that can be detected by PCR using genome-specific and plasmid-specific oligonucleotides (diagramed as half-arrows).

We created 7 novel mutants in this study and analyzed two mutants in additional candidate genes previously-published based on plant interaction phenotypes that were not related to their evolutionary trajectory (described in Table 3.1). Because my mutation strategy required cloning of a fragment wholly-internal to the target gene, short genes necessitated short inserts, which decreases the likelihood of homologous recombination. It is important to note that mutations in short genes with inserts fewer than 350 bases long were difficult or impossible for me to obtain.

Another explanation for failure to recover mutations in some genes is that they may be essential. While a comprehensive analysis of the essential genes in *Pto*$_{DC3000}$ has not been performed, two of our candidate genes were identified as likely to be essential via comparison to other related bacteria. *pspto2020* was shown to be essential for survival of *P. aeruginosa* strain PA14 in a saturating mutagenesis screen (Liberati et al., 2006). Similarly, *pspto2512* was shown to be essential in *E. coli* strain K12 substrain MG1655 (Gerdes et al., 2003). Because the homologs are essential in other bacteria, and these genes are not duplicated in the *Pto*$_{DC3000}$ genome, it is likely that these genes are essential in *Pto*$_{DC3000}$ as well and understanding the virulence function of these genes through knockout mutation is likely impossible.

virulence function	candidate gene	PA14 homolog	*E. coli* K12 homolog	gene length	insert length	integration attempts or mutant from	PA14 lethal?	E. coli lethal?
newly-identified	*pspto5537*	31720	*ynbC*	1757	437	1	no	no
	pspto5538	31730	*ynbD*	1355	510	3	no	no
	pspto5557	57710 *cysC*	*cysC*	668	393	1	no	no
	pspto4043	none	none	602	396	1		
	pspto5566	none	none	1691	507	4		
	pspto2999	none	none	752	450	2		
previously-identified	*avrPto*	none	none		------	SY He		
	pspto0834	10900 *ydjL*	*tdh*		------	AR Collmer	no	no
wild-type phenotype	*pspto3932*	none	none	518	375	7		
no mutant obtained	*pspto1269*	14450	none	425	225	1	no	
	pspto1442	38260	none	536	350	7	no	
	pspto1585	67860	none	747	510	1	no	
	pspto2020	44170	none	333	261	4	**yes**	
	pspto2512	01400	*miaA*	521	340	6	no	**yes**
	pspto3320	none	none	456	373	0 (PCR)		
	pspto3491	none	none	368	327	3		
	pspto3623	none	*yidX*	365	327	0 (PCR)		no
	pspto4655	none	none	869	424	0 (PCR)		
	pspto5539	31730	none	455	353	2	no	
	pspto5588	18320	*arnT*	1730	447	4	no	no

Table 3.2. Candidate genes and their conservation in genomes or organisms in which saturating mutageneses have been performed. Gene order is identical to that in Table 3.1 Blue indicates that a mutant was obtained. Red indicates the genes shown to be essential in other genomes. Black indicates that no mutant was obtained. PCR indicates that an internal fragment of the gene was never amplified by PCR.

Phenotypic characterization of evolving, putative virulence genes

Mutants were characterized for three major phenotypes: growth in liquid medium, growth in plants (Arabidopsis and tomato), and motility. These results are summarized in Table 3.3.

Pto$_{DC3000}$ uses three types of motility: (1) swimming motility, a flagellin-based motility (Figure 3.6) (Shimizu et al., 2003) (2) swarming motility, coordinated movement involving quorum sensing and flagellin-based movement (Chatterjee et al., 2003) and potentially (3) twitching (discussed below), a type IV pilus-dependent motility (Roine et

al., 1998). Any of these types of movement could be required for proper host colonization. Notably, before this work began, among motility mutants, only the swarming-deficient *gacA* mutant in *Pto*$_{DC3000}$, had been shown to be defective in growth within the leaf apoplast, while *pilA*, which is required for twitching motility in *P. aeruginosa* (Whitchurch et al., 2004) was shown to affect *Pto*$_{DC3000}$ survival on the leaf surface, and *flaA*, required for swimming and swarming motility (this work) was previously uncharacterized (Chatterjee et al., 2003). However, GacA is part of a two-component regulatory system which controls many aspects of *Pto*$_{DC3000}$ biology including coronatine biosynthesis genes and the TTSS, both of which are required for virulence of this strain. Hence, it remains unclear whether swarming motility defect observed in a *gacA* mutant is also required for full virulence. Loss of the TTSS, coronatine, quorum sensing, or some combination of all three could affect virulence (Bender et al., 1999; Bender et al., 1987; Chatterjee et al., 2003; Lindgren et al., 1986; Peet et al., 1986). Because of the lack of previous data regarding the contribution of the various *Pto*$_{DC3000}$ motility systems (swimming, swarming and twitching) in plant colonization, I assessed mutants deficient in each class of *Pto*$_{DC3000}$ movement: *flaA, gacA*, and *pilA*. I did so because I felt it was necessary to understand whether any of these motility systems was a reasonable proxy assay for movement that was relevant to infection.

I first found that the flagellin-deficient *flaA* (gift of the SY He lab) mutant of *Pto*$_{DC3000}$, which is deficient in swimming and swarming motility (Figure 3.6a, 3.6b), grows 10-fold more than wild type bacteria in Arabidopsis, despite its inability to swim or swarm (Figure 3.5a, 3.5b). I further demonstrated that the increased growth of *flaA* is due to evasion by *flaA* of FLS2-dependent host defenses, since growth of both *flaA* and

wild-type bacteria on the *fls2* mutant is equivalent to growth of *flaA* in *FLS2* plants (Figure 3.5a). Hence, the use of *flaA* to control for requirements for swimming as a virulence function is confounded by the recognition of flagellin peptide (Gomez-Gomez and Boller, 2002).

I therefore constructed a second swimming-defective mutant, *motA*, which still expresses and assembles flagellin in other systems, but is deficient in motility (Toutain et al., 2005). The *motA* mutant strain grew 5-fold less than wild-type in dip inoculations of Arabidopsis (Figure 3.5b; $p = 0.01$ over 3 experiments). Additionally, I observed enhanced growth of *flaA* mutant as compared to wild-type in flood inoculated 7 day old tomato seedlings of cultivar Moneymaker, while *motA* growth was decreased relative to wild type (Figure 3.5c). These results are presumably due to the presence of an active flagellin receptor in tomato and a conserved requirement for MotA-dependent swimming for colonization of the tomato apoplast. Both *flaA* and *motA* mutants are also defective in swarming, as flagellar function is required for swarming, as seen in Figures 3.6a and 3.6b (Kaiser, 2007) (Robatzek et al., 2007). Figure 6d shows the results of a mixed infection wherein wild-type and motA mutant were co-inoculated onto tomato seedlings at equal concentrations (see Materials and Methods). In a co-infection, *motA* grows 5-fold less than the total growth, meaning that it is slightly outcompeted by wild-type in this assay (Figure 3.6d). My results with *motA* thus indicate that flagellar-based motility, whether swimming or swarming, is required for full virulence. Hence, this mutant is a useful control with which to compare novel mutants for defects in virulence-associated motility.

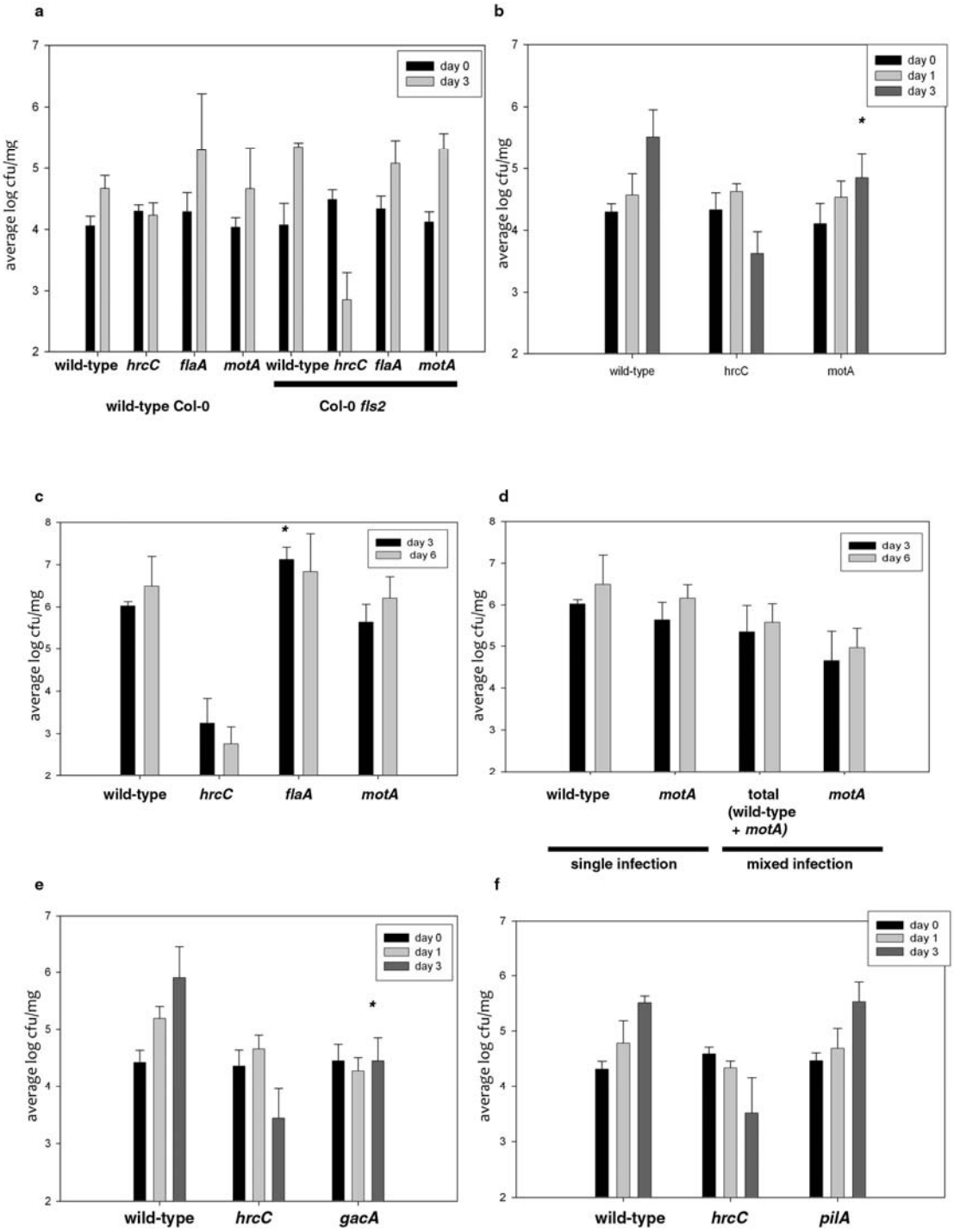


Figure 3.5. Mutations in *flaA, motA*, and *gacA* cause growth defects in planta.

a) Arabidopsis dip growth curves (see Methods) on wild-type *FLS2* and *fls2* mutant plants. The bacterial *flaA* mutant has a growth advantage on wild type Arabidopsis that is eliminated in the *fls2* mutant plants. Error bars represent 95% confidence interval.

b) Growth of *motA* mutant on wild-type Arabidopsis. The growth defect in this mutant is small but significant (indicated by * $p = 0.0126$ in 3 experiments)

c) Growth on tomato seedling leaves also shows enhanced growth of *flaA* mutant as compared to wild-type(indicated by * $p = 0.018$ black bars), while *motA* growth is decreased relative to wild type, especially at day 3

d) The defect in *motA* growth is evident in mixed infections. Single infection data presented is the same as in part b. The growth defect is small, and not significant due to the error in these assays, but the difference is evident especially at day 3 (black bars).

e) *gacA* mutant is defective in growth on Arabidopsis (indicated by * $p < 0.001$).

f) Growth of *pilA* mutant on Arabidopsis is identical to wild-type Pto_{DC3000}.

The role of the type IV pilus in virulence of Pto_{DC3000} is not clear. Roine et al. (1998) showed that the type IV mutant *pilA* was defective in epiphytic growth on the tomato leaf surface, but had no effect when infiltrated directly into tomato leaf apoplasts (MPMI 1998). I found no defect of *pilA* in epiphytic or apoplastic growth in Arabidopsis dip assays, indicating that the type IV pilus is not required for leaf colonization or growth within the Arabidopsis apoplast (Figure 3.5f). Additionally, I found that the swimming motility of Pto_{DC3000} masked twitching motility, since both wild-type and *pilA* mutant Pto_{DC3000} were motile on the surface of the twitching indicator plates (described in materials and methods) and did not allow reliable measurement of twitching using a plate assay (Whitchurch et al., 2004).

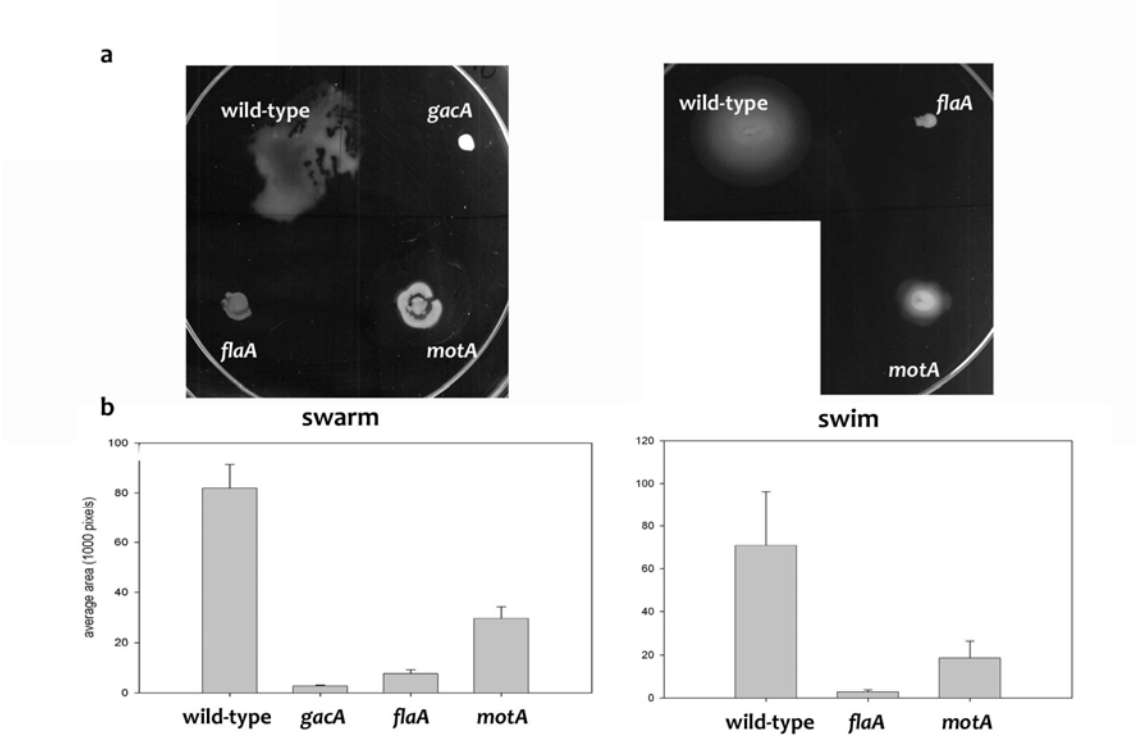


Figure 3.6. Mutations in *flaA* and *motA* cause loss of both swarming and swimming motility. a) Swarming (left) and swimming (right) motility of the listed motility mutants on a representative plate. *flaA* and *motA* are non-motile in both assays. C) Quantification of motility from the experiments represented above. Statistics are summarized in Table 3.4 (swarming) and Table 3.5 (swimming).

The role of swimming and swarming motility in plant infections by *Pto*$_{DC3000}$ was previously unknown beyond the role of flagellin peptide in triggering FLS2-dependent recognition by Arabidopsis and tomato (Robatzek et al., 2007). Decreased growth of *motA* mutant bacteria in Arabidopsis (Figure 3.5b) shows that flagellin-based motility is, in fact, required for full virulence of *Pto*$_{DC3000}$. However, the *motA* mutant is deficient in both swimming and swarming motility (fig. 6) (Tables 3.4 and 3.5), so it remains unclear whether swarming per se is required for *Pto*$_{DC3000}$ virulence. In the following analysis of candidate genes identified based on high dN/dS, both swimming and swarming phenotypes will be reported, and the implications of these phenotypes will be discussed.

Several dN/dS candidates function in virulence

virulence phenotypes	Mutant gene	*Ath* growth	Tomato growth	Swimming	Swarming	Liquid growth
newly-identified	*pspto5537*	5-fold less	10-fold less	wt	10-fold less	wt
	pspto5538	5-fold less	10-fold less	wt	wt	wt
	pspto5557	5-fold less	wt	wt	2-fold less	wt
	pspto4043	wt	wt	wt	wt	less
	pspto5566	wt	wt	wt	wt	less
	pspto2999	wt	wt	wt	2-fold less	wt
previously-identified	*avrPto*	wt	wt (Lin and Martin 2005)	wt	2-fold less	wt
	pspto0834	5-fold less (Ferriera et al. 2006)	wt	wt	wt	wt
wild-type phenotypes	*pspto3932*	wt	wt	wt	wt	nd

Table 3.3. Summary of mutant phenotypes. "wt" indicates wild-type growth. Descriptions of growth are compared to wild-type *Pto*$_{DC3000}$.

Novel genes found to function in virulence

pspto5557 was found to have a dN/dS value of 1.1 compared to the *Psy*$_{B728a}$ allele, *psyr4126,* in that pairwise comparison. *pspto5557* is predicted to encode an adenylyl sulfate kinase *cysC*. In *E. coli* and many other bacteria, *cysC* is part of an operon containing *cysN* and *cysD*, two sulfyl adenylate transferases. These three proteins function together to remove sulfyryl groups from donor molecules and transfer them to recipients. They are required for proper amino acid metabolism (Leyh et al., 1988). In many bacteria, including *Pto*$_{DC3000}$, the *cysC* and *cysN* genes are combined and encode a single protein, CysNC, with both the kinase and transferase activities (Shen et al., 2002).

This *cysNC* gene remains in an operon with *cysD*. In *Pto*$_{DC3000}$, an additional copy of *cysC* is found elsewhere in the genome and this copy is *pspto5557*.

Figure 3.7a shows the genomic context of *pspto5557* located immediately downstream of a transposon insertional sequence (IS element purple in Figure 3.7a) and upstream of a 6-gene operon, part of which encodes iron-dependent ABC transporters (*pspto5560-5563*), and this operon is also directly upstream of a second IS element. These IS boundaries indicate that this segment of the genome is likely to have been obtained by horizontal transfer. This hypothesis is supported by the fact that the genes bounded by these two IS elements are conserved in relative localization in the *Pph*$_{1448a}$ genome, which contains 2 copies of *cysC* (*pspph4301*and *pspph5079)* in addition to *cysNC* (*pspph4132*), (line 2 of Figure 3.7a) and this synteny is disrupted by the IS elements. Additionally, pspto5557 has low GC3 content (60%) which is outside of one standard deviation of the mean GC3 content for the genome. This further suggests that the *pspto5557* gene was acquired by horizontal transfer. Figure 3.7b shows the genomic context of *pspto4432* (*cysNC*) in the *Pto*$_{DC3000}$ genome. Here, the adenylyl sulfate kinase and sulfur transferase functions encoded by *cysNC* are coregulated with *cysD* which encodes a second sulfur transferase. The cysNC operon is not bounded by IS sequences and the *cysNC* gene has GC3 content consistent with that of the core genome (74% GC3). Together, these data indicate that *cysNC* was not acquired by horizontal transfer, but rather is part of the core genome. The co-regulation of genes involved in sulfur metabolism is recapitulated in the *E. coli* genome and others shown in Figure 3.7c, where *cysC* is a gene encoding an adenylyl sulfate kinase that is co-regulated with *cysN* and *cysD*, all separate genes in a common operon. Figure 3.8 shows a phylogenetic tree of

the distances between CysC and CysNC genes in *Pto*$_{DC3000}$, *Psy*$_{B728a}$, and *Pph*$_{1448A}$. The duplicate CysC proteins are each much more distantly-related than the CysNC sequences are, indicating that these are diversifying in sequence while the tightly-clustered CysNC sequences are not (Figure 3.8).

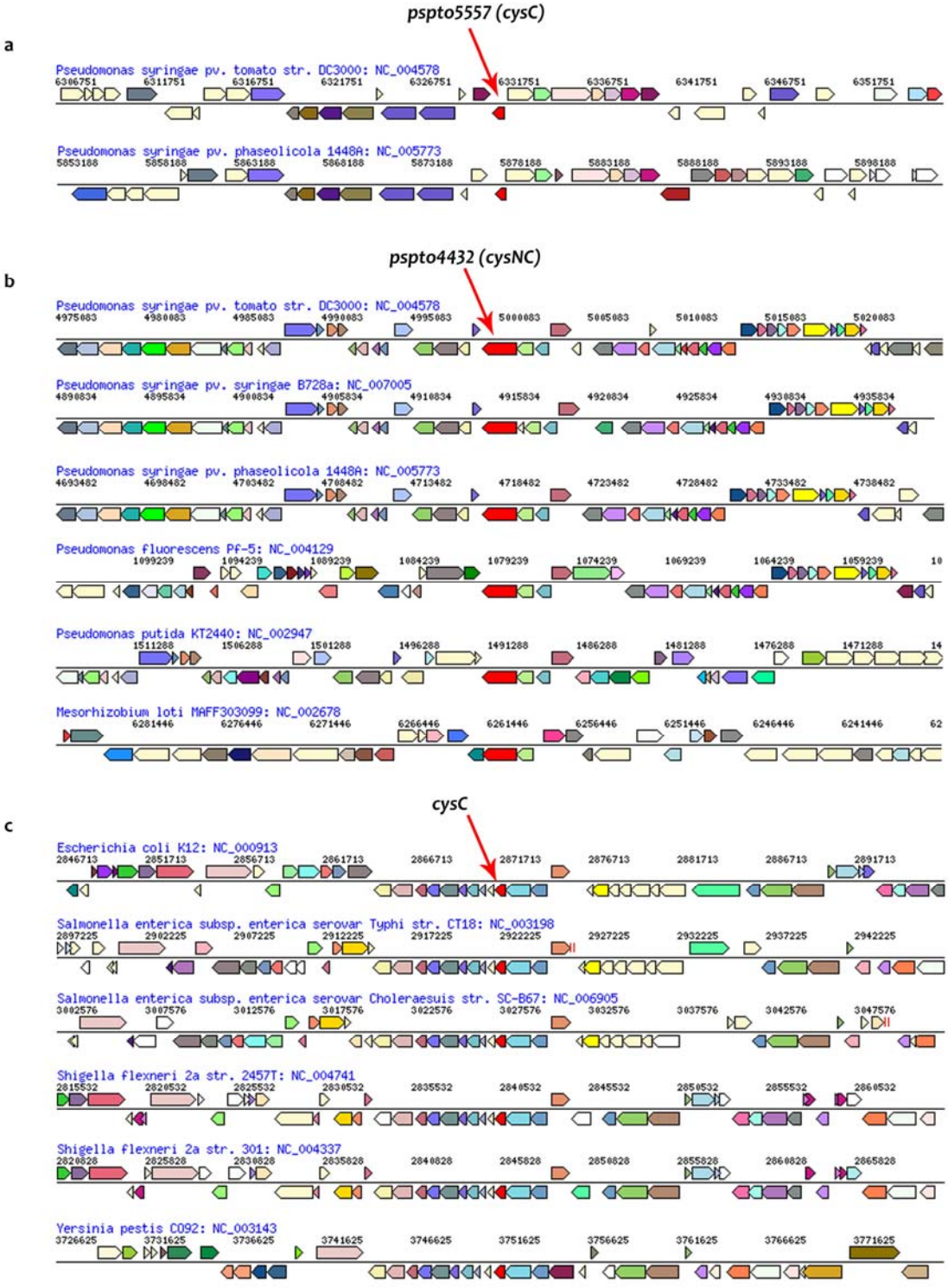

a
pspto5557 (cysC)
Pseudomonas syringae pv. tomato str. DC3000: NC_004578
Pseudomonas syringae pv. phaseolicola 1448A: NC_005773
b
pspto4432 (cysNC)
Pseudomonas syringae pv. tomato str. DC3000: NC_004578
Pseudomonas syringae pv. syringae B728a: NC_007005
Pseudomonas syringae pv. phaseolicola 1448A: NC_005773
Pseudomonas fluorescens Pf-5: NC_004129
Pseudomonas putida KT2440: NC_002947
Mesorhizobium loti MAFF303099: NC_002678
c
cysC
Escherichia coli K12: NC_000913
Salmonella enterica subsp. enterica serovar Typhi str. CT18: NC_003198
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67: NC_006905
Shigella flexneri 2a str. 2457T: NC_004741
Shigella flexneri 2a str. 301: NC_004337
Yersinia pestis CO92: NC_003143

Figure 3.7. *pspto5557* represents an additional copy of the *cysC* gene found also as *cysNC* Pto_{DC3000} genome. a) Genomic regions surrounding *pspto5557* and *pspph5179*, both in red. b) Genomic regions surrounding *cysNC* (red) in *Pto, Psy, Pph, P. fluorescens, P. putida*, and *M. loti*. c) Genomic regions surrounding *cysC* (red) from *E. coli* and related bacteria. Color codes indicate common function within comparison, but not between a, b, and c.

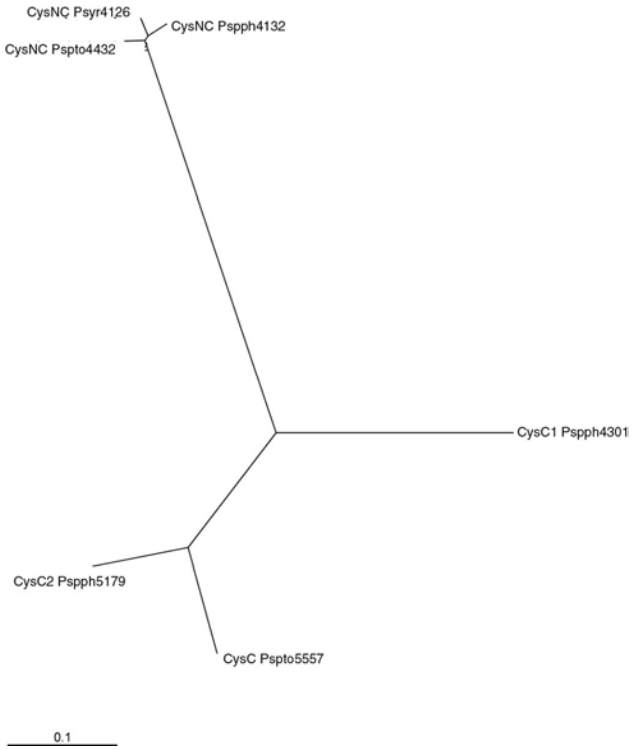
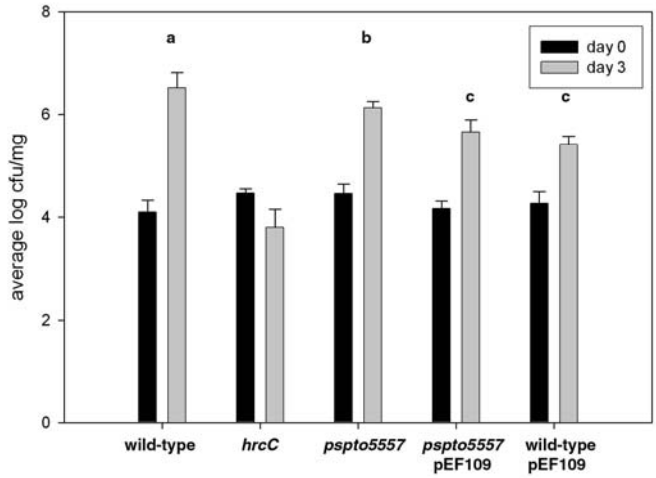


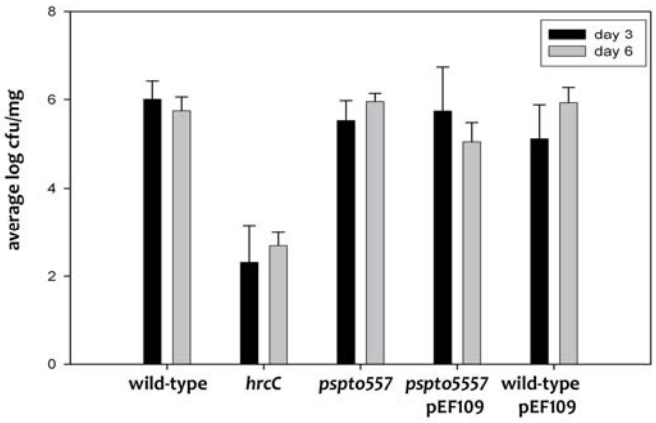
Figure 3.8. CysC proteins are more similar to one another than to CysNC bifunctional proteins in the same genomes. Neighbor-joining tree of CysC and CysNC protein sequences. The CysNC proteins cluster together. Bar represents 0.1 substitutions per site.

The *pspto5557* mutant shows reduced growth on Arabidopsis in dip seedling assays, as well as a swarming motility defect (Figure 3.9 a and c). Importantly, since *cysC* is likely to be involved in metabolism, the *pspto5557* mutant shows no growth defect in liquid minimal media (not shown). Additionally, the swimming motility of *pspto5557* is comparable to wild-type Pto_{DC3000}, indicating that the swarming defect I observed is not due to a flagellar defect.

a


8
6
4
2
0
average log cfu/mg
a
b
c
c
day 0
day 3
wild-type
hrcC
pspto5557
pspto5557 pEF109
wild-type pEF109

b


8
6
4
2
0
average log cfu/mg
day 3
day 6
wild-type
hrcC
pspto557
pspto5557 pEF109
wild-type pEF109

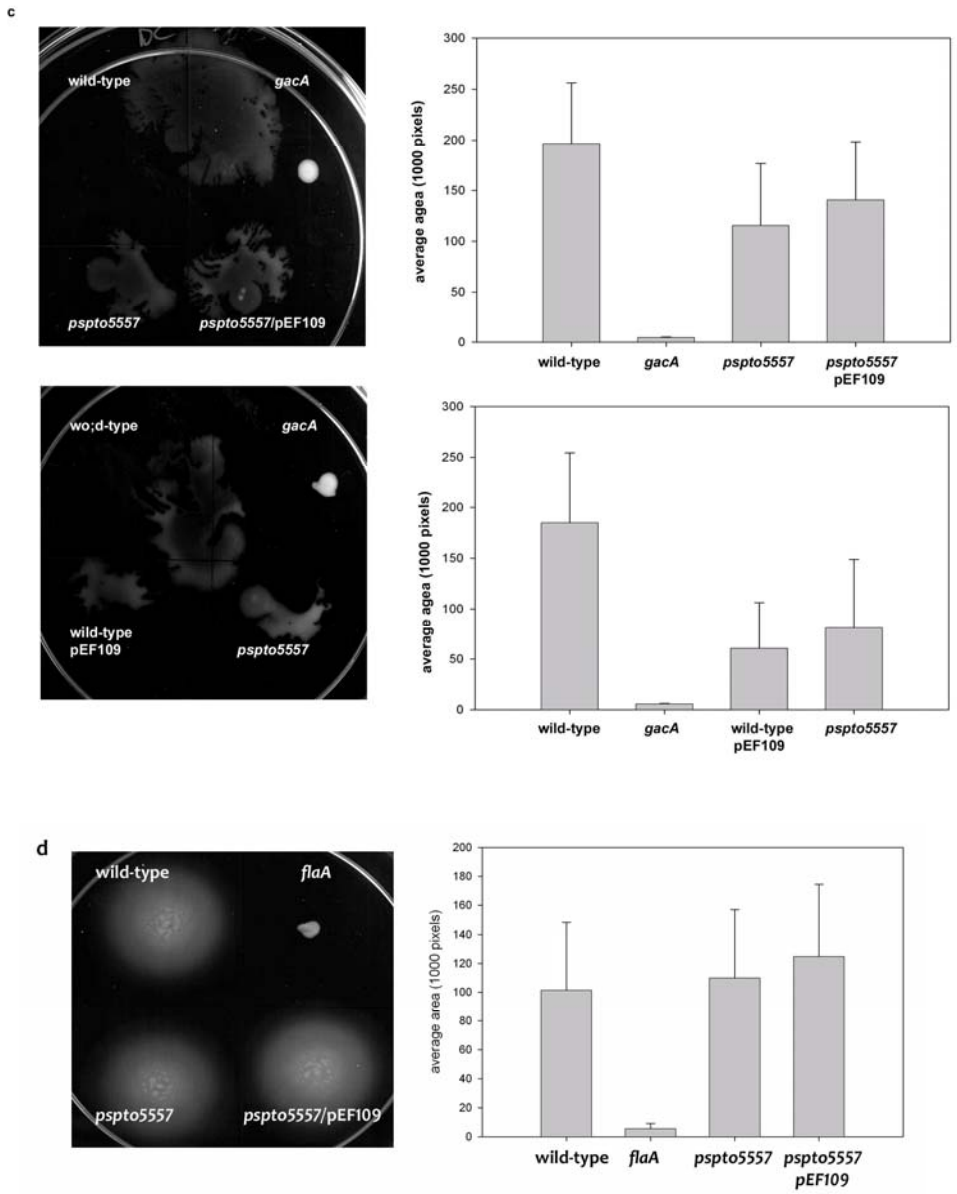


Figure 3.9. Phenotypes of the *pspto5557* mutant and the complemented mutant carrying construct (pEF109).

a) Growth in Arabidopsis dip assay shows that *pspto5557* is required for full virulence. Letters indicate significantly different categories ($p < 0.001$ based on students' t-test).

b) Growth in tomato seedling leaves shows that *pspto5557* is not required for growth in tomato.

c) Two representative swarming motility assays comparing *pspto5557* and a complementation construct (pEF109) in either the mutant or wild-type. *pspto5557* and the pEF109 complemented mutant strain are statistically different from wild type ($p = 0.0092$ over 6 experiments and 85 replicates for *pspto5557*, $p < 0.001$ over 3 experiments and 37 replicates of *pspto5557*/pEF109).

d) *pspto5557* and the complementation construct have no effect on swimming motility p = 0.072 over 4 experiments and 47 replicates for pspto5557, p = 0.2 over 3 experiments and 13 replicates).

The complementation construct pEF109 contains full-length *pspto5557* under the control of the *trp* promoter in the plasmid pDLtrpgw (thanks to Derek Lundberg). Addition of pEF109 to the *pspto5557* mutant exacerbates the growth defect of the *pspto5557* mutant (Figure 3.9a) and does not complement the swarming defect. Addition of pEF109 to wild-type *Pto*$_{DC3000}$ also induces a growth defect in Arabidopsis (Figure 3.9a, far right), indicating that the likely over-expression of *pspto5557* has the same phenotype as the mutation in this gene. Since the sequence, and presumably the function of *pspto5557* is duplicated in the *Pto*$_{DC3000}$ genome, it is possible that feedback between the two alleles results in over-expression of *cysNC* in the *pspto5557* mutant and that this is detrimental to growth in Arabidopsis. pEF109 was also sufficient to induce a swarming defect in wild-type *Pto*$_{DC3000}$ (Figure 3.9c, bottom panel), indicating that over-expression of this *pspto5557* interferes with motility and/or quarum sensing.

pspto5538 and ***pspto5537*** are part of an operon consisting of 4 genes (Figure 3.10). An additional gene (*pspto5539*) located in the same operon was also identified. All three of these genes, but not the downstream gene *pspto5540*, were found to have high dN/dS ratios compared to the *Psy*$_{B728a}$ alleles (3.8, 2.5, and 0.95 respectively, see Table 3.1) in that pairwise comparison. Unfortunately, we were unable to obtain an integration mutant in *pspto5539,* a short gene of only 455 bp.

Note that the homologs in *Psy*$_{B728a}$ are not part of an operon, but rather are in distinct genomic locations, suggesting that this operon configuration is not conserved throughout *P. syringae* strains, though it is conserved in other plant pathogens such as Xanthomonas (Figure 3.10). This indicates that in *Psy*$_{B728a}$, these genes are perhaps not co-regulated and that their gene products, therefore, may not function together, whereas they are under the control of a common promoter in *Pto*$_{DC3000}$ and may function in a common pathway. Most of this operon is conserved, however, in two phytopathogenic *Xanthamonas* strains (*X. axonopodis* pathovar citri strain 306 and *X. campestris* pathovar campestris strain ATCC33913), the mammalian pathogen *P. aeruginosa* PA01, and the soil bacterium *P. fluroescens* Pf-5. Further, this operon is found in three different genomic contexts, that basically mirror the three larger clades in the phylogeny in Figure 3.1b, indicating that this operon may function some aspect of the overlapping lifestyles of these strains (Figure 3.10). The *pspto5537-pspto5540* operon is flanked by IS sequences 29 genes upstream and 2 genes downstream (plum colored gene in Figure 3.10 row 1), but all genes in this operon have GC3 content consistent with the core genome, therefore it is not clear whether they were acquired by horizontal gene transfer.

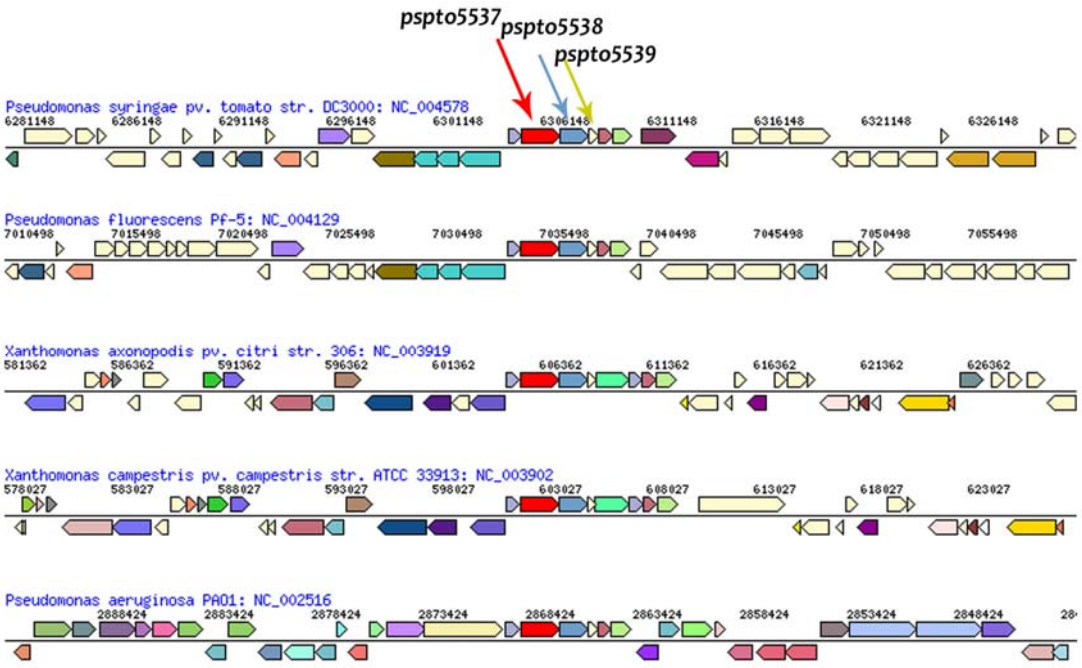


Figure 3.10. The operon containing candidate genes *pspto5537*, *pspto5538*, and *pspto5539* is conserved in Pseudomonad and Xanthomonas genomes, but not in Psy_{B728a} or Pph_{1448A}. Diagram shows *pspto5537* (red), *pspto5538* (blue), and *pspto5539* (white) and the rest of the operon. Color codes indicate common functional categories. These three candidate genes are coregulated in the genomes in the diagram, but not in Psy_{B728a}, the source of the dN/dS comparison.

The *pspto5538* mutant grows 5-fold less than wild type Pto_{DC3000} after 3 days in Arabidopsis and tomato leaves (Figure 3.11a and b). *pspto5538* encodes a predicted dual-specificity phosphatase, which in mammalian systems is responsible for regulating mitogen-activated protein kinases (MAPKs) by removing phosphate groups from tyrosine or threonine residues (Keyse, 2000). Bacteria do not encode MAPK proteins, but dual-specificity phosphatases are common in bacteria that associate with eukaryotes, for example Xanthomonas and *P. aeruginosa* strains diagramed in Figure 3.10. Disruption of MAPK function by bacteria is a common strategy to subvert host defense by both animal and plant pathogens (Shan et al., 2007). Indeed, a type III effector from Pto_{DC3000},

hopAO1 (*hopPtoD2*), was shown to have protein tyrosine phosphatase activity after delivery into eukaryotic cells. This activity was required for full virulence on tomato leaves and to suppress HR triggered by a constitutively active MAPK, $MEK2^{DD}$ in *N. benthamiana* leaves (Espinosa et al., 2003). Protein threonine- or tyrosine-phosphatase activity of *hopAO1* is required for both suppression of plant cell death and for full virulence on tomato leaves. Protein tyrosine phosphatase activity was shown to be dependent on a catalytic cysteine at position 178. The catalytic cysteine, C178, is conserved in Pspto5538, indicating that these two bacterial proteins may function similarly but probably modify different plant substrates. Importantly, we have not shown translocation of Pspto5538 into plant cells. *pspto5538* does not encode the N-terminal sequences associated with delivery by the TTSS, and is it not regulated by the TTSS sigma factor HrpL, so its proposed function modifying plant host proteins is hypothetical. If Pspto5538 and HopAO1 targeted the same plant substrates, we would expect the growth defect only in the double mutant, but I showed that the mutation in the *pspto5538* operon is sufficient to cause a growth defect (Figure 3.11a and b). Therefore, if Pspto5538 functions by directly targeting host proteins, I expect them to be distinct from the host proteins targeted by HopAO1.

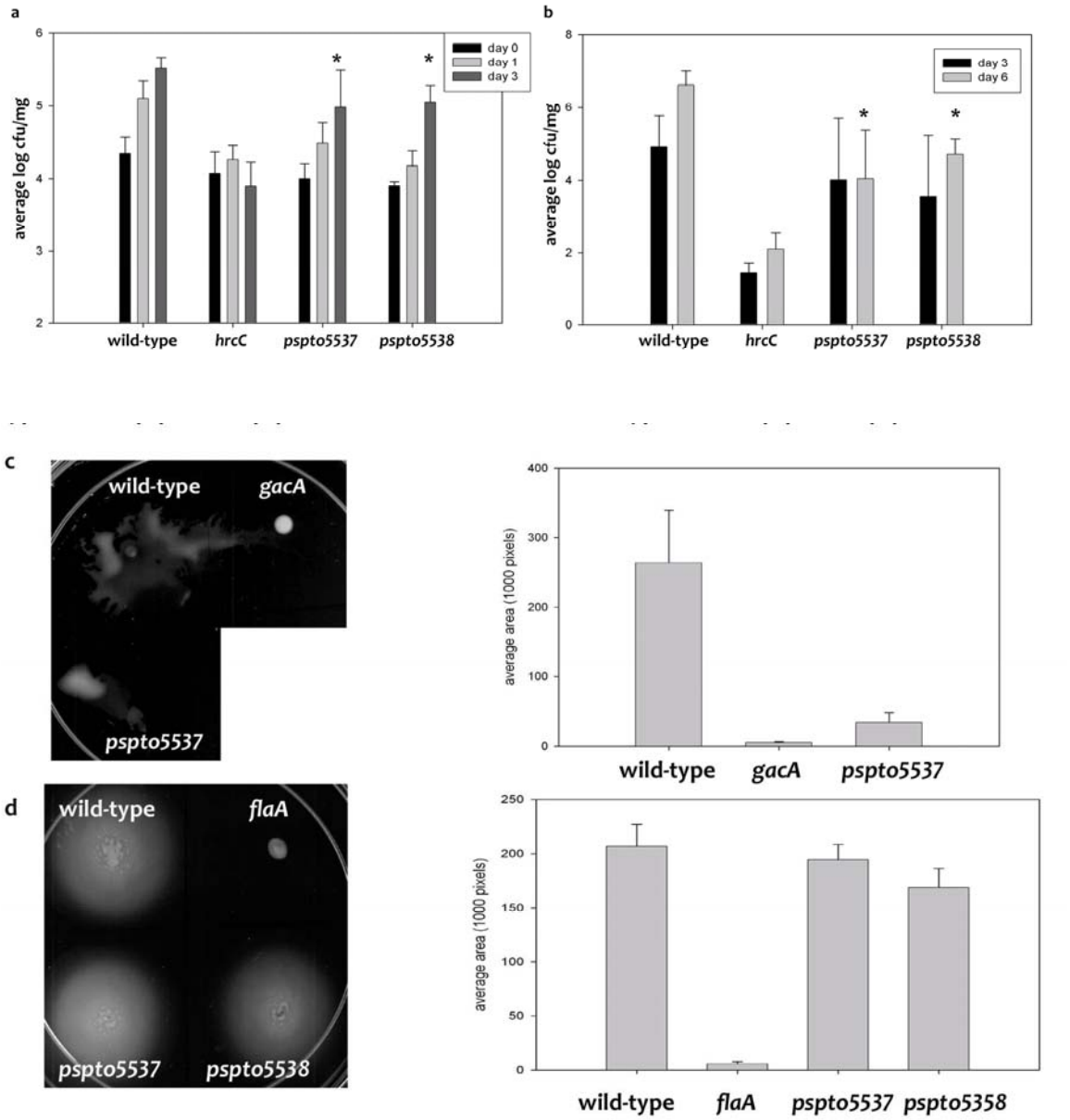


Figure 3.11. The *pspto5537-5540* operon is required for full PtoDC3000 virulence on Arabidopsis and tomato and possibly in swarming motility.

a) Growth of *pspto5537* and *pspto5538* mutants in Arabidopsis dip growth assays is 5-fold less than wild-type bacteria. * indicates statistical difference from wild-type bacteria ($p = 0.0317$ over 9 experiments).

b) Growth of *pspto5537* and *pspto5538* mutants is less than wild-type at day 6 in tomato seedling growth assays. Error bars represent 2x standard error or 95% confidence intervals. * indicates statistical difference from wild-type ($p = 0.014$ for pspto5537 and $p = 0.0004$ for pspto5538.

c) The *pspto5537* mutant has a swarming defect. A representative plate is shown and that particular experiment is quantified in the graph to the right.

d) Mutations in *pspto5537* and *pspto5538* have no effect on swimming motility. A representative plate is on the left and that experiment is quantified on the right. Overall statistics for motility experiments are in Tables 3.4 and 3.5.

The *pspto5538* mutant has no defect in swimming or swarming (Tables 3.4 and 3.5).

pspto5537 has phenotypic defects on Arabidopsis and tomato that mirror *pspto5538*, and additionally shows a defect in swarming motility on indicator plates (Figure 3.11c). Since the plasmid insertion causes a polar mutation in the operon from *pspto5537* to *pspto5540*, this *pspto5537* mutation can be considered a second allele of the *pspto5538* mutation. Hence, I conclude that whichever gene in that operon is required for full virulence, these can be considered two independent isolates of that common phenotype. Complementation data will confirm whether either of these genes is required for virulence and swarming motility.

Several candidates have defects in swarming motility

Several mutants derived from this work have a defect in swarming motility that may indicate an inability to sense the environment. *pspto2999*, *pspto5537* (mentioned above), and *avrPto* (mentioned below) all have swarming defects but show no growth defect in either liquid medium or on plants. Swarming phenotypes are summarized in Table 3.4. Importantly, none of the mutants described here have defects in swimming motility

(summarized in Table 3.5) with the exception of the controls used explicitly to test this phenotype: *motA, gacA,* and *flaA*. Hence, it is likely that they are impaired in response to environmental signals and not only in motility.

virulence phenotypes	genotype	different from wild-type?	p value	number of experiments	number of total replicates
newly-identified	*pspto5557*	yes	0.0092	6	85
	pspto5557/pEF109	yes	<0.001	3	37
	wild-type/pEF109	yes	0.013	1	4
	pspto5537	yes	< 0.001	3	39
	pspto5538	no	0.1196	3	45
	pspto4043	no	0.7704	3	31
	pspto5566	no	0.1608	4	53
	pspto2999	yes	< 0.0001	4	55
previously-identified	*avrPto avrPtoB*	yes	0.0335	3	43
	pspto0834	no	0.4086	3	33
wild-type phenotype	*pspto3932*	no	0.803	3	45
controls	*motA*	yes	< 0.0001	3	37
	hrcC	yes	0.0215	3	39

Table 3.4. Summary of swarming motility phenotypes. *p* values are based on the students t-test combining data from all experiments.

One of these, *pspto2999*, encodes a hypothetical protein encoded upstream of a cysteine protease in a 2-gene operon. The relevance of this mutation to Pto_{DC3000} virulence, though is not known since the *pspto2999* mutant grew to wild-type levels in both the Arabidopsis dip and tomato flood assays.

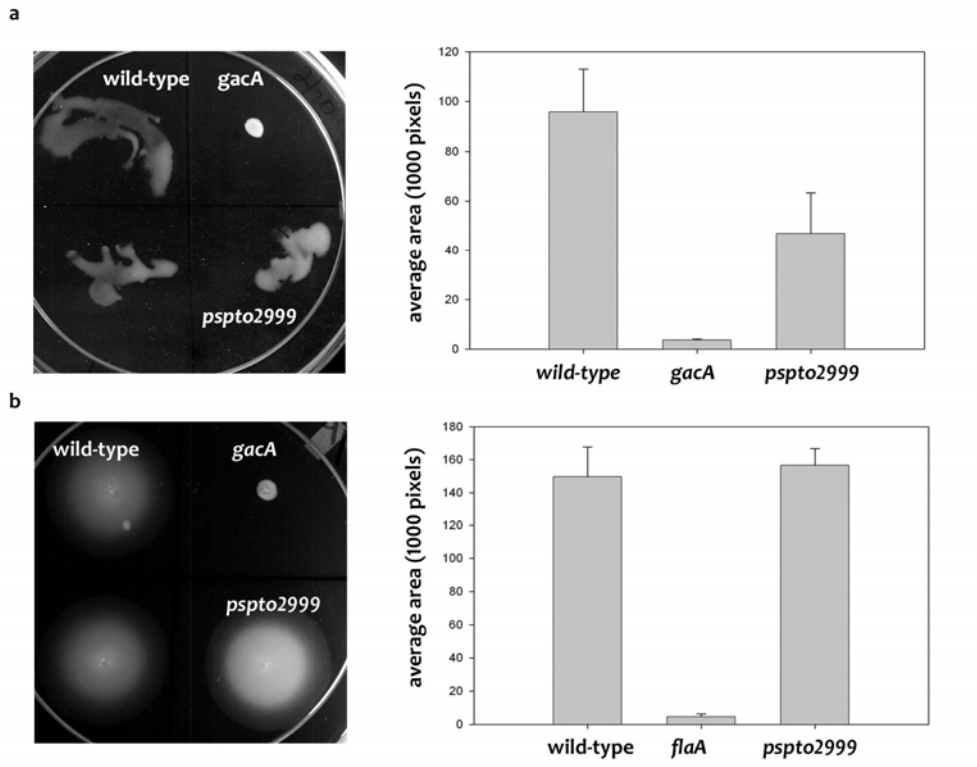


Figure 3.12. The *pspto2999* mutant has a swarming defect but no defect in swimming motility. a) A representative plate and quantification of swarming motility b) a representative plate and quantification of swimming motility from a representative experiment. Error bars represent 2x standard error or 95% confidence.

Two candidate genes are required for wild type growth in liquid medium.

pspto4043 encodes a hypothetical protein of unknown function. This mutant grows to wild-type levels on Arabidopsis and tomato leaves and shows no defect in swimming or swarming motility, but grows more slowly than wild-type in liquid minimal medium (Figure 3.13) (Huynh et al., 1989).

pspto5566 encodes a TnsD-like transposition protein. Like *pspto4043*, the *pspto5566* mutant shows no growth defect on Arabidopsis or tomato plants, but grows more slowly in minimal medium (Figure 3.13).

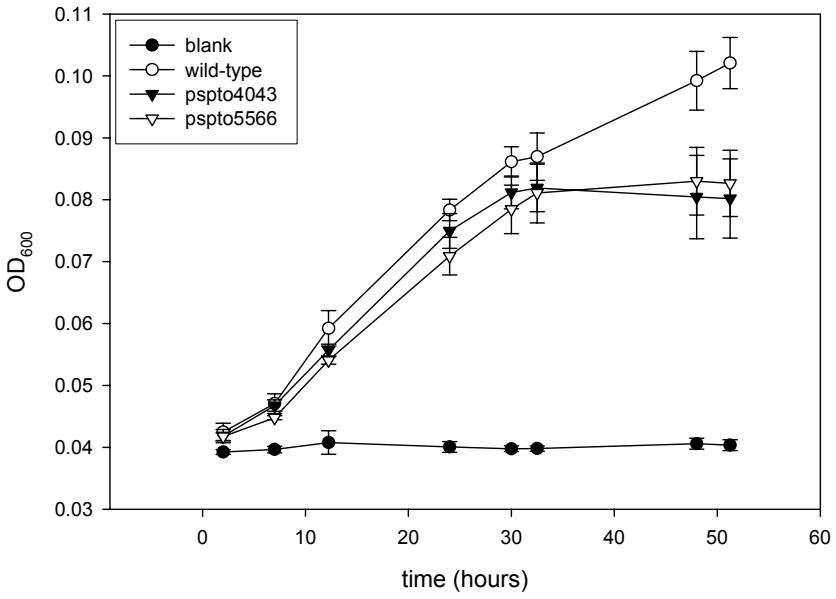


Figure 3.13. Growth in liquid minimal medium. *pspto4043* and *pspto5566* grew significantly more slowly than wild-type in minimal medium.

The known type III effector *avrPto* has high a dN/dS ratio

avrPto is a type III effector gene encoding a protein that is injected directly into the host cell during infection (Schechter et al., 2004). It was originally identified by Ronald et al (1992) based on its avirulence function: its ability to trigger *R*-gene-mediated resistance in tomato.

avrPto is present in Pto_{DC3000} and Psy_{B728A} (*psyr4919)* and had a dN/dS value of 0.816 in that pairwise comparison. Note that the atypical GC3 content of *avrPto* (47.9%) excluded this gene from the volatility analysis discussed in Chapter 2.

We obtained an *avrPto avrPtoB* double mutant from Sheng Yang He (Michigan State University). I found that this mutant has wild-type growth on Arabidopsis ($p = 0.08$ over 3 experiments) but has reduced swarming motility compared to wild-type ($p = 0.0314$ in 4 experiments; Figure 3.14a). While a single *avrPto* mutant is preferable, the double mutant was not available in our lab. *avrPto* is required for full virulence in *P. syringae* pathovar tomato strain T1 (Shen et al., 2001) but is not required for virulence on either Arabidopsis or susceptible tomato plants in the strain Pto_{DC3000} (Lin and Martin, 2005).

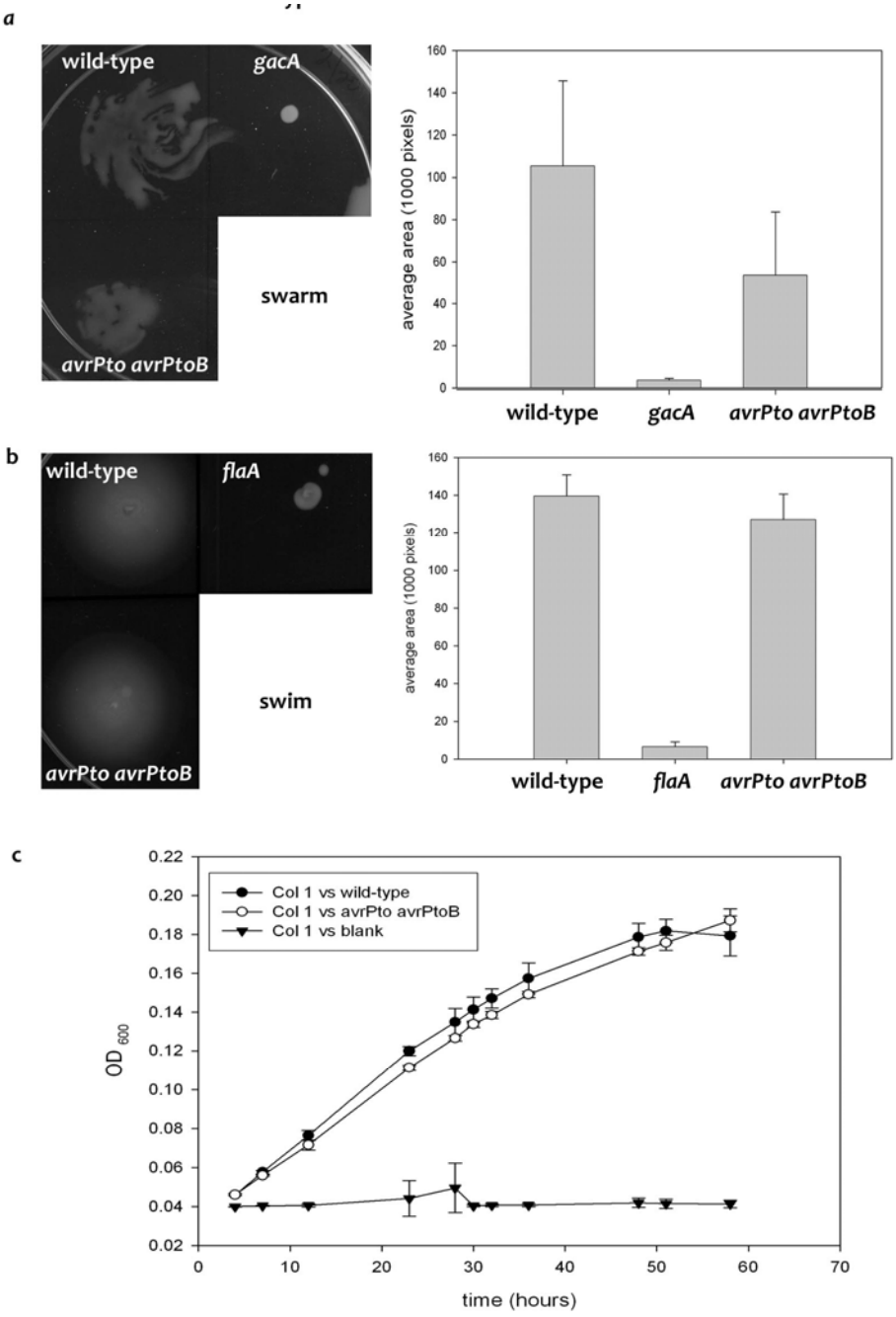


Figure 3.14. The *avrPto avrPtoB* double mutant is deficient in swarming motility but not growth in minimal medium.

a) Swarming motility of *avrPto avrPtoB* on a representative plate and the quantification of motility from the same experiment. Swarming motility is significantly different in 3 experiments (n = 43) $p = 0.0335$.

b) *avrPto avrPtoB* swimming motility in a representative plate and the quantification of that experiment (n = 33, p =0.467).

c) Growth of *avrPto avrPtoB* in minimal medium is not different from wild-type. Error bars represent 2x standard error, approximately 95% confidence.

***pspto0834* is a previously-identified virulence factor**. Another gene that I identified by its high dN/dS was previously shown to function in Pto_{DC3000} virulence. Vencato et al. (2006) found the predicted alcohol dehydrogenase-encoding gene *pspto0834* to be co-regulated with the type III secretion system (Vencato et al., 2006). Ferreira et al. (2006) found that *pspto0834* expression was dependent on the TTSS alternative sigma factor HrpL and that a *pspto0834* mutant grew 5-fold less than wild-type Pto_{DC3000} on 5-week old Arabidopsis ecotype Col-0 in vacuum infiltration assays (Ferreira et al., 2006). Hence, this rapidly evolving gene is required for full virulence of Pto_{DC3000} (Ferreira et al., 2006).

We were unable to repeat the published growth defect in our assays. This may be due to the difference in the age of the plants at time of inoculation (2 weeks vs. 5 weeks old) or due to assay differences: our assay involves dipping plants in a bacterial suspension in the absence of surfactant, while Ferreira et al. used vacuum infiltration or older leaves in the presence of surfactant. I approximated the conditions of Ferriera et al. using hand-inoculation of *pspto0834* in the presence of the surfactant (silwet) and again saw no growth defect (Ferreira et al., 2006). We also found no motility defect in swimming or swarming in this mutant. This discrepancy points to the fact that many genes of interest in virulence or host association may have weak phenotypes that are observed only under specific conditions.

Epiphytic growth was not affected by these mutants.

I separated epiphytic and apoplastic growth of bacterial strains by surface sterilizing Arabidopsis seedlings before bacterial growth was measured. The surface-sterilized samples represent apoplastic growth only, while unsterilized samples represent the total growth on the surface and inside the leaf. In all cases, the mutants survived on the surface as well as wild-type, which is almost none at all. By day 3, the surface-sterilized growth was identical to the total growth (data not shown). The only strains that persisted on the leaf surface were *hrcC* and *gacA* controls. Neither of these strains have functional type III secretion systems. Based on these experiments, I conclude that none of the tested candidate genes is required for invasion of the apoplast or for aberrant survival on the leaf surface.

None of the mutants have defects in swimming motility

While several mutants were found to have defects in swarming (Figures 3.11c, 3.12a, Table 3.4), none had defects in swimming motility. These results are summarized in Table 3.5.

virulence phenotype	genotype	different from wild-type?	p value	# experiments	# total replicates
newly-identified	*pspto5537*	no	0.56	3	44
	pspto5538	no	0.181	3	39
	pspto5557	no	0.072	4	47
	pspto4043	no	0.911	3	29
	pspto5566	no	0.32	3	29
	pspto2999	no	0.927	3	27
previously-identified	*avrPto avrPtoB*	no	0.467	3	33
	pspto0834	no	0.3387	3	31
wild-type phenotype	*pspto3932*	no	0.534	3	41
controls	*motA*	yes	< 0.001	3	35
	hrcC	no	0.947	4	47
	gacA	yes	<0.001	3	43

Table 3.5. Only motility controls had defects in swimming motility. Swimming phenotypes of all mutants discussed in this study show that no candidate genes of the 7 tested are required for swimming motility.

Rapidly evolving genes are enriched in potential virulence factors

In this work, I have described virulence-specific mutant phenotypes for 4 genes (or operons, in the case of *pspto5537-39*) that have elevated dN/dS. Two (*pspto5557* and the *pspto5537-41* operon) have newly-identified roles in virulence and two (*pspto0834* and *avrPto*) have previously-identified functions in virulence and disease resistance, respectively. When Lindeberg et al. (2008) catalogued *P. syringae* candidate virulence factors in a recent review (Lindeberg et al., 2008), they found 281 genes in Pto_{DC3000} that could be implicated in virulence. This represents 4.9% of the genes of $Pto_{DC3000.}$ My candidate gene list includes 2 of these previously-identified virulence factors (*pspto0834* and *avrPto*), meaning that the dN/dS analysis identifies a candidate list in which 10% are shared with a larger list derived from comparison and literature curation. A stringent

accounting of my results showing that two new loci (*pspto5557* and the *pspto5537-pspto5540* operon) are required for full virulence in Arabidopsis and tomato plants increases the proportion of virulence factors in this list to 20%. After addition of these two newly-discovered virulence genes to the 281 described by Lindeberg et al. (2008), I found that this enrichment in possible virulence factors was significant, as represented by p-value of 0.0182 in the Fisher's Exact Test (Table 3.6).

candidate identification method	number of putative virulence factors	total number of genes in this pool	% of virulence factors in pool	*p*-value Fisher's exact test
Lindeberg et al. (2008)	283*	5475	5.1%	--
high dN/dS candidates	4	20	20%	0.0182

Table 3.6: Gene selection based on high dN/dS enriches the resultant candidate gene pool in purported virulence genes. *Two novel virulence genes identified in this study (*pspto5557* and the *pspto5537-pspto5540* operon) were added to the putative virulence genes reported by Lindeberg et al. (2008). The proportion of genes in the dN/dS candidate list is significantly higher than that of the genome as a whole, as represented by the significant *p*-value in Fisher's Exact Test.

Discussion

I identified novel virulence factors from 7 of the 9 mutants reported here.

pspto5557 is a gene whose predicted function is duplicated in the genome of *Pto*$_{DC3000}$. *Pto*$_{DC3000}$ encodes a bifunctional CysNC protein called *pspto4432* which functions as an adenylyl sulfate kinase (CysC) and as a sulfur transferase (CysN), while *pspto5557* encodes the CysC protein alone. Duplication of gene function is one of the classic models allowing changes in coding sequence and potential changes in function (Ohno, 1970). In *Pto*$_{DC3000}$, the core genome encodes adenylylsulfate kinase function which is co-regulated with sulfur transferases in the *cysNC cysD* operon (Figure 3.7b). These genes are constrained against mutation due to a potential requirement for function. Indeed, *cysNC* (*pspto4432*) has very low dN/dS in the 3-way and pairwise analyses (approximately 0.01 in all analyses). So long as the *cysNC* gene function remained intact, *pspto5557* was free to mutate without functional consequence. If the *pspto5557* encoded a non-functional protein, I would expect to observe no phenotype in the *pspto5557* mutant. Further, if *pspto5557* encoded a protein of no function, I would not expect that overexpression from the complementation construct, pEF109, to have any phenotypic consequences. Both the mutant and overexpression phenotypes suggest that *pspto5557* encodes a functional protein, though the specific function of the Pspto5557 protein is still unknown.

The *Psy*$_{B728a}$ genome, in contrast, only contains the *cysNC* gene and does not include a *cysC* duplicate. In *Pph*$_{1448A}$, there is one *cysNC* gene and two additional *cysC*

copies, *pspph4301* and *pspph5179*. Figure 3.8 shows that the CysC proteins encoded by *pspto5557*, *pspph4301* and *pspph5179* are more closely related to one another than any are to the CysNC proteins.

Evolutionary analysis also identified the *pspto5537-5541* operon as containing 3 genes with high dN/dS, *pspto5537, pspto5538,* and *pspto5539*. These three genes were therefore identified as virulence gene candidates. Our inelegant method of site-directed insertion mutagenesis resulted in polar mutations in *pspto5537* and *pspto5538*. No mutant was found in *pspto5539*, probably due to its short length (455 nucleotides). However, the *pspto5537* and *pspto5538* mutants had the same phenotype when grown on Arabidopsis and tomato. This indicated either that these two genes have identical roles in virulence or that something downstream of *pspto5537* is required for full virulence in these assays. We are currently addressing this question using complementation constructs of the individual ORFs from this operon alone, and the operon as a whole. The fact that two genes in an operon give the same phenotype is, however, essentially the same as having an independent second mutant allele, giving me confidence in these results.

If the gene that complements this phenotype is *pspto5538* or *pspto5539*, this leads me to conclude that identification of virulence factors by high dN/dS ratio is a viable strategy. If *pspto5540* or *pspto5541* complements the phenotypes, we could conclude that the dN/dS analysis can identify operons of functional relevance. If it is the operon, and not the specific identified gene, that is important, we could expand our mutational analysis to the operons identified by this screen.

Evolutionary analysis identified two known virulence factors.

The list of candidate genes having high dN/dS when compared to closely related strains includes two previously identified virulence factors. Ronald et al (1992) originally identified *avrPto* based on its avirulence function: its ability to trigger *R*-gene-mediated resistance in tomato (Ronald et al., 1992). *avrPto* is a type III effector gene encoding a protein that is injected directly into the host cell during infection (Schechter et al., 2004). The genetic interaction of *avrPto* with the corresponding *R*-gene *PTO* in tomato has become the canonical example of "gene-for-gene" resistance in plants. Addition of the *avrPto* gene from Pto_{DC3000} to strain Pto_{T1} rendered Pto_{T1} avirulent on formerly susceptible tomato plants, meaning that bacteria expressing *avrPto* are unable to grow on tomatoes expressing the genes *PTO* and *PRF*. Conversely, *avrPto* is also able to enhance the virulence of Pto_{T1} when inoculated onto tomatoes lacking the *PTO* gene (Shan et al., 2000). However, a deletion of *avrPto* in Pto_{DC3000} had no effect on virulence. The authors postulated that *avrPto* does function in strain Pto_{DC3000} but that the high virulence of this strain masks the effect of a single effector (Ronald et al., 1992). Our studies showed that *avrPto* mutation has no effect on growth of Pto_{DC3000} on Arabidopsis or tomato plants, confirming previous results. However, *avrPto* in Pto_{DC3000} was previously shown to be required for symptom development, but was not required for full virulence on tomato (Lin and Martin, 2005).

pspto0834 was shown by Vencato et al. (2006) to be co-regulated with the TTSS and to be required for full virulence on Arabidopsis in vacuum infiltration assays (Vencato et al., 2006). I was unable to repeat that result in dip assays and in hand-

infiltration infections, which should closely mimic the vacuum infiltration protocol. It is likely that different growth or infection conditions masked the weak growth defect reported by Vencato et al (Vencato et al., 2006). I also showed that *pspto0834* is not required for full virulence on tomato, indicating that it may have an Arabidopsis-specific function.

Evolutionary analysis identified two novel genes required for growth in liquid medium.

Both *pspto5566* and *pspto4043* mutants had growth defects in liquid minimal medium (fig 13). Neither of these mutants had any growth defect on Arabidopsis or tomato. These genes may be under selective pressure during times of low nutrient availability in the environment, but not on plants.

P. syringae **has several rapidly-evolving genes, some of which are required for full virulence.**

My candidate list of evolving genes is comparable in length to candidates from similar analyses performed in *E. coli*. (Chen et al., 2006; Petersen et al., 2007). Two different evaluations of evolution across *E. coli* genomes have been published. One analysis focused on several strains of uropathogenic *E. coli* and identified 29 evolving genes, 11 of which had a dN/dS of 999.00 and would have been excluded from our candidate list, leaving 18 genes that would have passed our criteria (Chen et al., 2006). The second study identified 22 candidates that are evolving in a more diverse set of *E. coli* and *S.*

flexneri strains. However, only 18 are non-viral remnants and only 15 are non-viral and have dN/dS < 999. In these two studies, only two genes are found in both candidate lists: *ompF* and *ompC*, which encode membrane-embedded proteins known to function in Salmonella virulence (Negm and Pistole, 1999). There are no *ompF* or *ompC* homologs in Pto_{DC3000}, so these genes cannot be used as a positive control for the evaluation of evolution rate of Pto_{DC3000} genes. It is interesting, though, that pathogen genomes of similar size to Pto_{DC3000} have approximately the same number of genes with high dN/dS. This may indicate that selective pressures exerted by diverse hosts such as humans and plants have similar effects on their respective pathogens. It also places my results within the boundaries established in previous publications based on well-studied bacteria, suggesting that my candidates are not false-positives.

Our initial hypothesis that genes under selection would be in direct physical contact with the host cell or host immune system predicted that genes with high dN/dS would encode proteins localized to the bacterial outer membrane or secreted by the bacteria. Petersen et al. found that 8/18 genes with high dN/dS (after excluding viral genes and genes with dN/dS = 999 as discussed earlier) were predicted to localize to the outer membrane (Petersen et al., 2007). Their analysis included 2 uropathogenic strains of *E. coli*, 2 enterohemorragic strains of *E. coli*, and 4 strains of *Shigella flexneri* compared to the laboratory strain *E. coli* K12. When Chen et al. compared uropathogenic E. coli strains only, they found that 3/18 genes with high dN/dS (again, ignoring viral genes and those with dN/dS = 999) encoded proteins predicted to localize to the outer membrane (Chen et al., 2006). In our study, none of the candidate virulence factors are predicted to localize to the Pto_{DC3000} outer membrane. However, 4/20 candidate genes

with high dN/dS encode proteins that are predicted to localize to the cytoplasmic membrane as determined by PsortB (Gardy et al., 2005). One of these is *pspto5538*, which is required for full virulence during infection of both Arabidopsis and tomato (Figure 3.11a and b). None of the other genes whose products exhibit predicted membrane localization (*pspto3320*, *pspto5539*, and *pspto5588*) were analyzed in this work, though it is possible that the growth defects of the *pspto5537* and *pspto5538* mutants are truly phenotypes that result in a disruption of expression of the downstream gene *pspto5539*. Complementation analysis is required before those conclusions are fully resolved. One gene that demonstrates the limits of PsortB as a predictive tool is *avrPto*, which was shown experimentally to be transported via the TTSS but has unknown localization according to the PsortB program. With this limitation of the PsortB program in mind, I can currently conclude that the high dN/dS genes of *Pto*$_{DC3000}$ are not highly enriched in outer membrane proteins.

It is also important to note that the previously-identified evolving type III effectors present in *Pto*$_{DC3000}$ *hopX, hopAF1, hopQ1, hopI1*, and *hopAB3* are not on my list (Rohmer et al., 2004). The initial study used different homolog sets, across the GenBank accessions available at that time, to evaluate dN/dS (Rohmer et al., 2004). *hopQ1* was, for example, compared to homologs from Xanthamonas and Ralstonia. *hopAF1* was compared to homologs from *Pgy* and Xanthamonas. *hopI1* and *hopAB3* were compared to homologs from *Pgy* and *Psy*. *hopX* is highly conserved throughout *P. syringae* and the *Pto*$_{DC3000}$ allele is identical to that of *Pph*$_{1448a}$ at the nucleotide sequence level (Nimchuk et al., 2007). Rohmer et al. found *hopX* to have a high dN/dS compared

to 14 other alleles from *Pph* and *Pma* isolates. It is likely that the alleles found in Psy_{B728a} and Pph_{1448a} are too similar to the Pto_{DC3000} allele to recapitulate the result.

I also expect to lose any genes that have evolution only at a few sites. The method I used was to look at the overall dN/dS of the entire open reading frame. Therefore, if a gene has a long region with zero non-synonymous mutations and a short region of rapid evolution, the high dN/dS of the evolving region may be masked by the domain(s) under functional constraint against diversification. I therefore consider this a first pass at assessing the role of evolution in identifying virulence factors. A more detailed analysis would be possible with the aid of additional homologous sequences. If a given residue is highly variable among many homologous sequences, this site may be a target of selection. However, with only a maximum of 3 homologs compared, as in this study, evaluation of the relative importance of a single site over another is impossible. This is discussed further in Chapter 4.

Motility and Virulence

I found that motility is required in a limited sense, for full virulence using our infection methods. I attempted to mimic a natural infection by dipping Arabidopsis seedlings into a bacterial suspension in the absence of surfactant, normally used to ensure equal distribution of bacteria throughout the apoplast. This results in a lower effective dose of bacteria, as approximately 100-fold fewer bacteria enter the apoplast during the first hour of infection in our method compared to bacterial suspensions using the typical

dose of silwet surfactant. Non-motile *motA* mutant bacteria were only slightly less virulent than wild-type Pto_{DC3000} (Figure 3.5b, c, and d). The non-motile *flaA* mutant grew better than wild type bacteria in some assays. I was able to attribute this extra growth to flagellin perception of wild-type bacteria by the plant FLS2 receptor, since the growth enhancement of *flaA* bacteria was eliminated when grown on *fls2* mutant Arabidopsis.

The *pilA* mutant does not make a wild-type type IV pilus, but also grew to wild-type levels in our dip inoculation assays (Roine et al., 1998) and Figure 3.5f). The question of whether *pilA* is required for twitching motility remains unclear for Pto_{DC3000}, but this is the case in *P. aeruginosa* (Roine et al., 1998). We attempted to measure twitching motility using protocols developed for *P. aeruginosa* which measures motility between a thin layer of agar and petri plate (Corning nonpyrogenic polystyrene). Though occasionally some twitching was visible on the bottom surface of the agar, bacterial motility on top of the growth medium masked twitching and prevented accurate its measurement. *P. aeruginosa* PAK control strains do not swim on LB medium and twitching in these strains was easily observed.

Swarming motility does appear to be intertwined with virulence in a manner that is not directly obvious. I obtained 3 mutants (*avrPto, pspto5537, and pspto5557*) that grow less than wild type on either Arabidopsis or tomato, or both (see Table 3.3) and do not swarm as well as wild type. However, I also obtained several mutants with swarming defects, but no defect in growth on either Arabidopsis or tomato. None of these mutants had a defect in swimming.

For example, the *pspto5537* mutant had a major swarming defect but the *pspto5538*, which had an identical virulence profile, did not. Also, the pEF109 complementation construct for *pspto5557* did not complement the swarming defect of the *pspto5557* mutant. These data together suggest that swarming defects might be incidental, and may be a result of our mutation strategy. Because the mutation strategy requires days of bacterial proliferation, ample time for these mutations to occur in the population is provided along with the mutation construct. We found that homologous recombination in Pto_{DC3000} is rare and it is possible that these non-swarming variants of Pto_{DC3000} are somehow more amenable to acceptance or recombination of the integration plasmid. Mutants obtained from other labs including two independent *hrcC* mutants and the *avrPto avrPtoB* double mutant also showed clear swarming defects, indicating that such proposed incidental mutations could be common in strains that have been manipulated in the lab.

Nine other candidates yet to be explored

I was not able to obtain mutations in nine remaining candidates. I believe that much of this is due to the short length of these candidates which necessitated an even shorter region available for homologous recombination between the pKT plasmid and the target gene. I believe that an approach similar to that of (Wei et al., 2007) will be effective in creating these mutations for future analysis. This approach is different because it creates a non-polar deletion of the target gene that does not impact other genes

in the operon. This approach also utilizes long regions of homology (over 1kb) to allow homologous recombination at the target locus.

Rate of molecular evolution identifies novel virulence factors.

At present, I have described mutant phenotypes for 7 genes (or operons, in the case of *pspto5537-39*) that have elevated dN/dS. Two (*pspto5557* and the *pspto5537-41* operon) have newly-identified roles in virulence, two (*pspto0834* and *avrPto*) have previously-identified functions in virulence and disease resistance, respectively, and three (*pspto2999*, *pspto4043*, and *pspto5566*) have functions relevant to motility or growth but no identified function in growth on plants. Even if limited to only those loci directly implicated in bacterial growth in plants (*pspto5557* and the *pspto5537-pspto5540* operon), the proportion of virulence factors in my candidate gene list increases to 20%, which is 4-fold enrichment relative to our random expectation (Table 3.6). Thus, my results indicate that the rate of molecular evolution can be used as a predictor of virulence function. Further analysis of my uncharacterized candidate genes may push this proportion even higher and future analyses investigating the rapidly evolving genes of related bacteria can extend this conclusion beyond the genome of *Pto*$_{DC3000.}$

Materials and Methods

Bacterial mutants

Internal fragments of candidate genes comprising approximately 2/3 of the gene length were amplified from *Pto*$_{DC3000}$ and cloned by BP clonase into the suicide vector pMK2010 (House et al. 2004). These constructs are called pKT. pKT vectors were transformed to mating-competent *E. coli* strain S17.1 {Prifer, 1983 #4787}. Cultures of *E. coli* containing the pKT vector of interest and *Pto*$_{DC3000}$ were grown overnight. The following day, 1mL of each culture was washed three times in sterile water. Cultures were mixed at a 5:1 *E .coli*: *Pto*$_{DC3000}$ ratio on nitrocellulose squares, placed on KB plates with no antibiotics and incubated for 48 hours at 28°C. The nitrocellulose squares were removed to 2mL sterile water in sterile test tubes and vortexed to free the bacteria from the nitrocellulose. The entire 2mL of bacterial suspension was plated in 500µl aliquots to large plates containing KB Rf50 Km 50 Cyc50 to select for *Pto*$_{DC3000}$ and the integrated plasmid. Putative mutants were screened using primers outside of the reading frame of the target gene in combination with plasmid-specific primers. Clones that yielded a PCR product representing the gene fragment plus the plasmid fragment (calculated independently for each gene) were considered mutants. The PCR was performed using the MTN-Pfx program and Accuprime Pfx polymerase. The PCR products were sequenced using gene-specific primers to confirm the mutation.

Bacterial Growth in Arabidopsis

Dip inoculations: Bacteria were suspended in 10mM $MgCl_2$ at an OD_{600} = 0.05 which is $2.5^{x}10^{7}$ cfu/ml. *Pto*$_{DC3000}$ wild-type and a *Pto*$_{DC3000}$ strain containing a

disruption of *hrcC* (He lab) were used as positive and negative controls, respectively. 2-week old Arabidopsis seedlings were dipped in the suspension and plants were harvested 1 hour (day 0), 1 day (day 1), and 3 days (day 3) after dip inoculation. Each bar represents 4 tubes containing 1mL of 10mM $MgCl_2$ containing 200µl/L silwet and 3 seedlings after 1 hour shaking at 28°C. Samples were serially diluted, plated on KB Rif, and counted after 1 day at 28°C. Another set of plants was surface-sterilized in 70% Ethanol at each timepoint described above. After 30 seconds in 70% ethanol, seedlings were plunged into 1L of water to remove ethanol. Plants were blotted with paper towels and added to 1ml 10mM $MgCl_2$, shaken, diluted, and quantified as described above for the non-sterilized samples.

Hand inoculations: Bacteria were suspended in 10mM $MgCl_2$ at an OD = .002 (10^6 cfu/ml). 4 week old Arabidopsis leaves were inoculated on the underside of leaves with the bacterial suspension using a needle-less syringe. Plants were covered with a plastic dome for 8-12 hours after inoculation. Bacterial growth was measured at 1 hour (day 0) and 3 days (day 3) after inoculation. Leaves were removed and a round core was taken as a representative sample. Three cores were combined in a single 1.5ml eppendorf tube containing 200µl 10mM $MgCl_2$. Leaf cores were ground using a drill bit and the volume was brought up to 1ml 10mM $MgCl_2$. 200µl were removed, serially diluted, and plated on selective KB medium. Colonies were counted by eye 24-48 hours after plating. Bacterial growth is expressed as cfu/cm^2 since 3 cores represents 1cm^2.

Bacterial Growth in Tomato

Bacterial growth on tomato seedlings was performed by adapting the protocol of {Uppalapati, 2008 #4788}. Tomato seeds of the Moneymaker variety were sterilized by shaking in 70% ethanol for 20 minutes followed by shaking in 100% bleach for 20 minutes. Seeds were then washed 4 times in excess sterile water and plated on plates containing MS medium with Gamborg vitamins and 0.8% agar. These plates were placed in the dark for approximately 7 days until the hypocotyls emerged. Bacteria were suspended in 10mM MgCl2 plus 200ul/L silwet at an OD = 0.1 ($5 * 10^7$ cfu/ml). Seedlings were inoculated using a 10ml pipet so that each leaf got at least one drop of inoculum. Bacterial suspension was removed with a sterile pipet 5 minutes after inoculation. Plates were sealed with parafilm and placed in an incubator with 12 hour light. Samples were taken 3 and 6 days after inoculation. Leaves were removed from seedling stems and were surface-sterilized in 70% ethanol for 30 seconds before being immersed in 1L water to remove ethanol. Leaves were blotted and moved to sterile tubes containing 1ml 10mM MgCl2 with 200ul/l silwet. Tubes were shaken for 1 hour then 200ul were removed, serially diluted, and plated as for Arabidopsis dip growth curves. Bacterial colonies were counted 24-48 hours later and expressed as cfu/mg of plant tissue.

Error bars represent 2x standard error or approximately 95% confidence.

Bacterial Motility

Swimming motility was evaluated on LB plates containing 0.3% agar. Plates were thick, containing approximately 50mL LB agar each. Bacteria were suspended at equivalent OD_{600} and stabbed into a marked spot on a plate using a toothpick. The *flaA*

(He lab) mutant was used as the negative control for this type of motility. Plates were scanned after 48 hours at room temperature (approximately 23°C).

Swarming motility is assessed on KB plates containing 4% agar. Bacteria were suspended to OD_{600} = 2.0 and 1µl was spotted on each plate. The *gacA* (Chatterjee lab) mutant is the negative control for this. Plates were scanned 24 hours after inoculation.

Importantly, wild-type, negative control, and experimental (mutant) bacteria were monitored on a single plate. This reduced plate-to-plate variation. Both types of motility were quantified based on the area of bacterial movement. Scanned images (black-and-white jpegs) were opened in the ImageJ program. The threshold function creates a binary image where bacterial growth is red and the medium beneath is black. Red pixels were quantified and averages were graphed (as seen in Figure 3.6). *p*-values are the result of students t-tests comparing the motility of wild-type and mutant Pto_{DC3000} across many plates and several different experiments. Error bars represent 2x standard error or approximately 95% confidence.

Statistics

χ^2 tests were performed in SigmaPlot (Systat Software).

Fisher's Exact Test was used in Table 3.5 because it is preferred over χ^2 when the numbers are small, as with the number of volatile genes previously reported to be potential virulence factors. This test was performed using GraphPad available online: http://www.graphpad.com/quickcalcs/index.cfm

References

Bender, C. L., Alarcon-Chaidez, F., and Gross, D. C. (1999). Pseudomonas syringae phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. Microbiol Mol Biol Rev *63*, 266-292.

Bender, C. L., Stone, H. E., and Cooksley, D. A. (1987). Reduced pathogen fitness of *Pseudomonas syringae* pv. *tomato* Tn5 mutants defective in coronatine production. Physiol Molec Plant Pathol *30*, 273-283.

Chatterjee, A., Cui, Y., Yang, H., Collmer, A., Alfano, J. R., and Chatterjee, A. K. (2003). GacA, the response regulator of a two-component system, acts as a master regulator in Pseudomonas syringae pv. tomato DC3000 by controlling regulatory RNA, transcriptional activators, and alternate sigma factors. Mol Plant Microbe Interact *16*, 1106-1117.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R. R., Ozersky, P.*, et al.* (2006). Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. Proc Natl Acad Sci U S A *103*, 5977-5982. Epub 2006 Apr 5973.

Chen, Y., Emerson, J. J., and Martin, T. M. (2005). Evolutionary genomics: codon volatility does not detect selection. Nature *433*, E6-7; discussion E7-8.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res *27*, 4636-4641.

Espinosa, A., Guo, M., Tam, V. C., Fu, Z. Q., and Alfano, J. R. (2003). The Pseudomonas syringae type III-secreted protein HopPtoD2 possesses protein tyrosine phosphatase activity and suppresses programmed cell death in plants. Mol Microbiol *49*, 377-387.

Feil, H., Feil, W. S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., Lykidis, A., Trong, S., Nolan, M., Goltsman, E.*, et al.* (2005). Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci U S A *102*, 11064-11069. Epub 12005 Jul 11025.

Felsenstein, J. (1989). Phylogeny Inference Package (Version 3.2). Cladistics *5*, 164-166.

Ferreira, A. O., Myers, C. R., Gordon, J. S., Martin, G. B., Vencato, M., Collmer, A., Wehling, M. D., Alfano, J. R., Moreno-Hagelsieb, G., Lamboy, W. F.*, et al.* (2006). Whole-genome expression profiling defines the HrpL regulon of Pseudomonas syringae pv. tomato DC3000, allows de novo reconstruction of the Hrp cis clement, and identifies novel coregulated genes. Mol Plant Microbe Interact *19*, 1167-1179.

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

Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balazsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S.*, et al.* (2003). Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol *185*, 5673-5684.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol *11*, 725-736.

Gomez-Gomez, L., and Boller, T. (2002). Flagellin perception: a paradigm for innate immunity. Trends Plant Sci *7*, 251-256.

Hahn, M. W., Mezey, J. G., Begun, D. J., Gillespie, J. H., Kern, A. D., Langley, C. H., and Moyle, L. C. (2005). Evolutionary genomics: codon bias and selection on single genomes. Nature *433*, E5-6; discussion E7-8.

House, B. L., Mortimer, M. W., and Kahn, M. L. (2004). New recombination methods for Sinorhizobium meliloti genetics. Appl Environ Microbiol *70*, 2806-2815.

Huynh, T. V., Dahlbeck, D., and Staskawicz, B. J. (1989). Bacterial blight of soybean: Regulation of a pathogen gene determining host cultivar specificity. Science *245*, 1374-1377.

Joardar, V., Lindeberg, M., Jackson, R. W., Selengut, J., Dodson, R., Brinkac, L. M., Daugherty, S. C., Deboy, R., Durkin, A. S., Giglio, M. G.*, et al.* (2005). Whole-genome sequence analysis of Pseudomonas syringae pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol *187*, 6488-6498.

Kaiser, D. (2007). Bacterial swarming: a re-examination of cell-movement patterns. Curr Biol *17*, R561-570.

Keyse, S. M. (2000). Protein phosphatases and the regulation of mitogen-activated protein kinase signalling. Curr Opin Cell Biol *12*, 186-192.

Leyh, T. S., Taylor, J. C., and Markham, G. D. (1988). The sulfate activation locus of Escherichia coli K12: cloning, genetic, and enzymatic characterization. J Biol Chem *263*, 2409-2416.

Liberati, N. T., Urbach, J. M., Miyata, S., Lee, D. G., Drenkard, E., Wu, G., Villanueva, J., Wei, T., and Ausubel, F. M. (2006). An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A *103*, 2833-2838. Epub 2006 Feb 2813.

Lin, N. C., and Martin, G. B. (2005). An avrPto/avrPtoB mutant of Pseudomonas syringae pv. tomato DC3000 does not elicit Pto-mediated resistance and is less virulent on tomato. Mol Plant Microbe Interact *18*, 43-51.

Lindgren, P. B., Peet, R. C., and Panapoulos, N. J. (1986). Gene cluster of *Pseudomonas syringae* pv. *phaseolicola* controls pathogenicity on bean plants and hypersensitivity on nonhost plants. J Bacteriol *168*, 512-522.

McGinnis, S., and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res *32*, W20-25.

Negm, R. S., and Pistole, T. G. (1999). The porin OmpC of Salmonella typhimurium mediates adherence to macrophages. Can J Microbiol *45*, 658-669.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol *3*, 418-426.

Nimchuk, Z. L., Fisher, E. J., Desveaux, D., Chang, J. H., and Dangl, J. L. (2007). The HopX (AvrPphE) family of Pseudomonas syringae type III effectors require a catalytic triad and a novel N-terminal domain for function. Mol Plant Microbe Interact *20*, 346-357.

Ohno, S. (1970). Evolution by Gene Duplication, Springer-Verlag).

Peet, R. C., Lindgren, P. B., Willis, D. K., and Panopoulos, N. J. (1986). Identification and cloning of genes involved in phaseolotoxin production by Pseudomonas syringae pv. "phaseolicola". J Bacteriol *166*, 1096-1105.

Petersen, L., Bollback, J. P., Dimmic, M., Hubisz, M., and Nielsen, R. (2007). Genes under positive selection in Escherichia coli. Genome Res *17*, 1336-1343. Epub 2007 Aug 1333.

Plotkin, J. B., Dushoff, J., and Fraser, H. B. (2004). Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. Nature *428*, 942-945.

Robatzek, S., Bittel, P., Chinchilla, D., Kochner, P., Felix, G., Shiu, S. H., and Boller, T. (2007). Molecular identification and characterization of the tomato flagellin receptor LeFLS2, an orthologue of Arabidopsis FLS2 exhibiting characteristically different perception specificities. Plant Mol Biol *64*, 539-547. Epub 2007 May 2025.

Rohmer, L., Guttman, D. S., and Dangl, J. L. (2004). Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen Pseudomonas syringae. Genetics *167*, 1341-1360.

Roine, E., Raineri, D. M., Romantschuk, M., Wilson, M., and Nunn, D. N. (1998). Characterization of type IV pilus genes in Pseudomonas syringae pv. tomato DC3000. Mol Plant Microbe Interact *11*, 1048-1056.

Ronald, P. C., Salmeron, J. M., Carland, F. M., and Staskawicz, B. J. (1992). The cloned avirulence gene *avrPto* induces disease resistance in tomato cultivars containing the *Pto* resistance gene. J Bacteriol *174*, 1604-1611.

Sarkar, S. F., and Guttman, D. S. (2004). Evolution of the core genome of Pseudomonas syringae, a highly clonal, endemic plant pathogen. Appl Environ Microbiol *70*, 1999-2012.

Schechter, L. M., Roberts, K. A., Jamir, Y., Alfano, J. R., and Collmer, A. (2004). Pseudomonas syringae type III secretion system targeting signals and novel effectors studied with a Cya translocation reporter. J Bacteriol *186*, 543-555.

Shan, L., He, P., and Sheen, J. (2007). Intercepting host MAPK signaling cascades by bacterial type III effectors. Cell Host Microbe *1*, 167-174.

Shan, L., He, P., Zhou, J., and Tang, X. (2000). A cluster of mutations disrupt the avirulence but not the virulence functions of AvrPto. Molec Plant-Microbe Interact *13*, 592-598.

Shen, Y., Chern, M., Silva, F. G., and Ronald, P. (2001). Isolation of a Xanthomonas oryzae pv. oryzae flagellar operon region and molecular characterization of flhF. Mol Plant Microbe Interact *14*, 204-213.

Shen, Y., Sharma, P., da Silva, F. G., and Ronald, P. (2002). The Xanthomonas oryzae pv. lozengeoryzae raxP and raxQ genes encode an ATP sulphurylase and adenosine-5'-phosphosulphate kinase that are required for AvrXa21 avirulence activity. Mol Microbiol *44*, 37-48.

Shimizu, R., Taguchi, F., Marutani, M., Mukaihara, T., Inagaki, Y., Toyoda, K., Shiraishi, T., and Ichinose, Y. (2003). The DeltafliD mutant of Pseudomonas syringae pv. tabaci, which secretes flagellin monomers, induces a strong hypersensitive reaction (HR) in non-host tomato cells. Mol Genet Genomics *269*, 21-30. Epub 2003 Feb 2013.

Toutain, C. M., Zegans, M. E., and O'Toole, G. A. (2005). Evidence for two flagellar stators and their role in the motility of Pseudomonas aeruginosa. J Bacteriol *187*, 771-777.

Vencato, M., Tian, F., Alfano, J. R., Buell, C. R., Cartinhour, S., DeClerck, G. A., Guttman, D. S., Stavrinides, J., Joardar, V., Lindeberg, M*., et al.* (2006). Bioinformatics-enabled identification of the HrpL regulon and type III secretion system effector proteins of Pseudomonas syringae pv. phaseolicola 1448A. Mol Plant Microbe Interact *19*, 1193-1206.

Wei, C. F., Kvitko, B. H., Shimizu, R., Crabill, E., Alfano, J. R., Lin, N. C., Martin, G. B., Huang, H. C., and Collmer, A. (2007). A Pseudomonas syringae pv. tomato DC3000 mutant lacking the type III effector HopQ1-1 is able to cause disease in the model plant Nicotiana benthamiana. Plant J *51*, 32-46. Epub 2007 Jun 2008.

Whitchurch, C. B., Leech, A. J., Young, M. D., Kennedy, D., Sargent, J. L., Bertrand, J. J., Semmler, A. B., Mellick, A. S., Martin, P. R., Alm, R. A.*, et al.* (2004). Characterization of a complex chemosensory signal transduction system which controls twitching motility in Pseudomonas aeruginosa. Mol Microbiol *52*, 873-893.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci *13*, 555-556.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol *15*, 568-573.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics *155*, 431-449.

Chapter 4

Conclusions and Future Directions

The *P. syringae*-Arabidopsis interaction is a model for pathogen-host interactions (Katagiri et al., 2002). Much is known about the molecular and genetic interplay between these organisms (Ausubel et al., 1995; Grant et al., 2006). Work in our field has used this system to study questions of innate immunity, programmed cell death and bacterial virulence, mostly focused on major virulence mechanisms like the type III secretion system (TTSS) and toxin production (Abramovitch et al., 2006; Grant et al., 2006). My work aimed to expand our understanding of virulence factors to include previously-overlooked genes with function in pathogenesis. My hypothesis was that genes involved in pathogen virulence would be under selective pressure imposed by the host. Therefore, identification of rapidly evolving genes in the pathogen genome would yield a list of potential virulence factors. I took two approaches to identify genes under selection in the genome of *P. syringae* pathovar tomato strain DC3000 (Pto_{DC3000}).

The first approach, described in Chapter 2, assessed codon volatility in the Pto_{DC3000} genome. This approach measured the distribution of so-called volatile codons, previously reported to indicate recent positive selection at the site of a volatile codon (Plotkin et al., 2004). The volatility approach, refined with comparative genomics to identify genes specific to plant-associated bacteria and exclude those genes found in the genomes of soil microbes or pathogens exclusively associated with mammalian hosts, resulted in a list of 31 candidate virulence genes (Table 2.4). Among these highly volatile candidates were 3 genes with previously implicated in virulence: *corR, hopL1*, and *shcA*. These are discussed in detail in Chapter 2.

The second approach to evaluate molecular evolution of genes in the Pto_{DC3000} genome relied on established methods comparing homologous genes in closely related *P.*

syringae isolates, both of which are pathogens of bean plants: *P. syringae* pathovar syringae strain B728a (Psy_{B728a}) and pathovar phaseolicola strain 1448A (Pph_{1448A}). These two *P. syringae* genomes were compared to Pto_{DC3000} in a pairwise manner as well as together in a 3-way comparison involving most (3809 of 5724) of the genes in the Pto_{DC3000} genome. Candidate virulence genes were selected if they had a high ratio of non-synonymous mutations (dN) to synonymous mutations (dS), referred to as dN/dS, indicative of positive selection (Nei and Gojobori, 1986). I identified genes with a dN/dS value between 0.8 and 999 (as discussed in Chapter 3) and this yielded a list of 20 candidate virulence genes, of which I obtained mutations in 9. My work on the mutational analysis of these, combined with previously-published results shows that 4 / 9 assayed have growth defects in Arabidopsis or tomato (Table 3.3), 2 have growth defects in minimal medium, and one additional gene had a defect in swarming motility, though the relevance of that phenotype to pathogenesis and to the disruption of the target gene has not yet been fully established.

While the list of volatile genes includes intriguing candidates based on their predicted cellular localization in the outer membrane, their prevalence in pathogen genomes and their exclusion from non-pathogen genomes, volatility as a method to identify genes under selection has been challenged statistically (see Discussion Chapter 2). Nevertheless, it remains to be seen whether mutation of the genes on my list would identify relevant virulence phenotypes, since volatility as a method to identify interesting new virulence genes has not been thoroughly tested. In this Chapter, I will discuss the future direction of the dN/dS analysis presented in Chapter 3 especially as it relates to current genome sequencing projects in the labs of Jeff Dangl and Corbin Jones.

Additional genome sequences add to our understanding of *P. syringae* evolution.

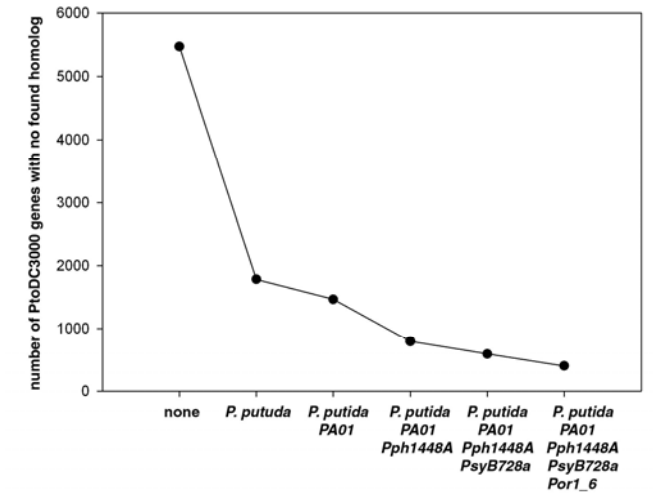
As discussed in Chapter 3, the list of candidate rapidly-evolving genes between *Pto*$_{DC3000}$, *Psy*$_{B728a}$, and *Pph*$_{1448A}$ is short (19 genes) but of comparable length to that of evolving genes in *E. coli*. The *E. coli* studies used 6 (Petersen et al., 2007) and 7 (Chen et al., 2006) sequenced strains, while our analysis only used 3 genomes. We would like to extend the dN/dS analysis to include additional genomes from diverse strains of *P. syringae*. *P. syringae* isolates have been collected from geographically distant locations (Canada, UK, Germany, Russia, etc.) and from diverse host plant species (beans, tomato, cucumber, chestnut, wheat, barley, etc.). Distinct geographical and host niches are expected to produce evolutionarily divergent species with specific host ranges (Yan et al., 2008). Expanding of the breadth of sequenced *P. syringae* genomes will enable a better understanding of the pan-genome diversity throughout this species.

The multi-locus sequence type (MLST) phylogeny based on conserved housekeeping genes of *P. syringae* isolates shows 5 distinct clades (Sarkar and Guttman, 2004). These clades roughly correspond to the host from which the strain was isolated. All tomato isolates are in clade 1, kidney bean isolates are clade 3, etc. The fully-sequenced strains *Pph*$_{1448A}$, *Psy*$_{B728a}$, and *Pto*$_{DC3000}$ represent three of the 5 clades, but a more detailed examination of evolution between clades may indicate different evolutionary pressures based on different host ranges (Yan et al., 2008). Conversely, a detailed analysis of evolution in a collection of strains with the same host may show how geography influences selective pressure within a clade (Petersen and Chen refs). dN/dS analyses performed with at least 3 homologous sequences have more statistical power and significance than those comparing only 2 strains (Goldman and Yang, 1994). dN/dS

values obtained using the codeml module of PAML include consideration of strain relatedness based on the MLST phylogeny provided to the program and therefore are considered more reliable than pairwise comparisons (Goldman and Yang, 1994). I expect that more genome sequences will increase the number of genes with at least two homologs for which codeml can be used to determine the dN/dS and this is demonstrated in Figure 4.1b. If the *E. coli* analyses are an appropriate guide, these further dN/dS comparisons may not increase the number of candidate genes, but their selection will have more meaning. One might also be able to identify parts of the *P. syringae* phylogeny that are under more intense selective pressure than the rest of the strains.

Additional homologous sequences and the codeml program can also identify specific regions or individual codons that are under positive selection. This was useful when Rohmer et al. showed that the N-terminus or HrpW had a dN/dS = 9.5. Rohmer et al. found that the gene as a whole had an overall dN/dS = 0.2 but when the M3 model was applied in the codeml program, the dN/dS = 9.5. These complex models that identify molecular evolution in specific lineages, or in specific parts of the coding sequence, are only possible with several homologous sequences and are lost when comparing only 2 homologs. For instance, in our analysis, *hrpW* had dN/dS of 0.08 in the Psy_{B728a} pairwise comparison, 0.1 in the Pph_{1448A} pairwise comparison, and 0.14 in the 3-way comparison. None of these analyses were sensitive enough to detect rapid molecular evolution in the N-terminus of the HrpW protein. A repeated dN/dS using many homologs will enable more careful dissection of dN/dS that was missed in my analysis of the overall molecular evolution rate of an entire gene.

Whatever enlarged set of strains selected for genome sequencing, these new *P. syringae* genomes will increase the number of genes that can be compared for dN/dS. For example, BLAST searches of *Pto*$_{DC3000}$ genes against distantly-related Pseudomonads (*P. putida* and *P. aeruginosa*) show that most genes belong in the conserved "core" genome shared among diverse bacteria. Subsequent BLAST searches using *Pto*$_{DC3000}$ genes with no homologs in the distantly-related Pseudomonads against the genomes of closely-related strains of *P. syringae* reduce the number of *Pto*$_{DC3000}$-specific genes. Our lab is presently sequencing the genome of *P. syringae* pathovar oryzae strain 1_6 (*Por*1_6). Josie Reinhardt and I found that blast searches against this new *Por*1_6 genome further reduced the number of *Pto*$_{DC3000}$-specific genes to less than 10% of the genome (Figure 4.1a). As new genomes are added, the number of truly *Pto*$_{DC3000}$-specific genes is expected to level-off. While some genes may prove to be truly unique to *Pto*$_{DC3000}$, BLAST searches against additional genomes will add depth to our understanding of many genes in *Pto*$_{DC3000}$ as homologs are added. Truly strain specific genes may represent those required for niche specialization.



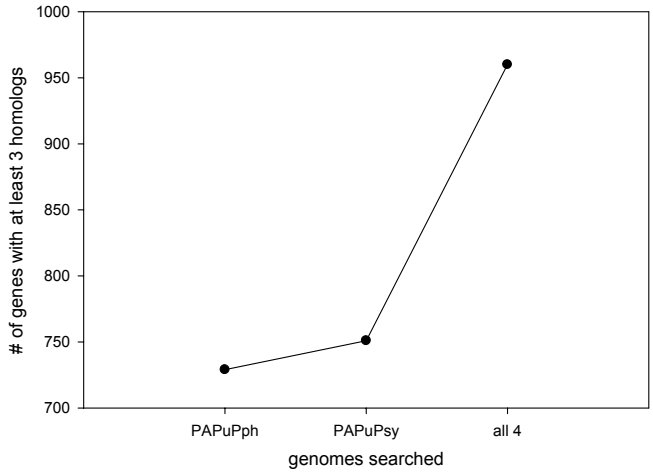


Figure 4.1. Additional *P. syringae* genome sequences add to our understanding of Pto_{DC3000}. a) Subsequent blast homology searches (blastp e = 10^{-5}) reduce the number of Pto_{DC3000}genes with no known homologs. The x-axis shows genomes used as sequential homology searches. b) Pto_{DC3000} genes with at least three homologs in searched genomes increase as the number of genomes increases.

Another mechanism of bacterial evolution is to mix-and-match genome composition created by horizontal transfer of genes between species (reviewed by(Gal-Mor and Finlay, 2006). Genome sequences of diverse *P. syringae* isolates will give us a better understanding of the core or minimal *P. syringae* genome. Comparison of the minimal *P. syringae* genome to the genomes of sequenced soil microbes that occupy

similar ecological niches as *P. syringae*, such as water and soil samples, but do not directly interact with host plants may tell us which genes are specifically required for leaf colonization by *P. syringae* species.

Evolution rate can identify novel virulence factors.

I showed that identification of rapidly evolving genes can also identify novel virulence factors in *Pto*$_{DC3000}$ such as *pspto5557* and *pspto5537/8/9* (Chapter 3) as well as previously-identified virulence factors *pspto0834* and *avrPto*. Chapter 3 described the role of these novel genes in virulence on Arabidopsis and tomato plants. Because these genes are conserved in other well-studied *P. syringae* strains, this project could continue through investigation of the role of *pspto5557* and *pspto5537/8/9* homologs in either *Psy*$_{B728a}$ or *Pph*$_{1448A}$. The *Pph*$_{1448A}$ genome contains two duplicate copies of *cysC*: *pspph4301* and *pspph5179*. Mutation of either one of these may have an effect in *Pph*$_{1448A}$ virulence on bean plants, where *Pph*$_{1448A}$ is an effective pathogen, or they may have redundant virulence function. *pspto5537/8/9* also have homologs in *Psy*$_{B728a}$ that may function in bean pathogenesis of that strain.

A similar mutation strategy (perhaps even the same plasmid based on the *Pto*$_{DC3000}$ sequence) could be used to test the requirement of *Psy*$_{B728a}$ alleles in bean pathogenesis. Another way to approach this question is to use the *Psy*$_{B728a}$ / *Pph*$_{1448A}$ alleles to complement the *Pto*$_{DC3000}$ mutation. If the *Pph*$_{1448A}$ *cysC1* gene complements the *Pto*$_{DC3000}$ *pspto5557* deletion, we know that sequence diversification has not abolished function of this allele. If it does not complement, we may conclude that the sequence

changes have either abolished function of the protein or that they have added host-specificity.

Genes with evolutionary constraint may be required for virulence.

Very few genes in *Pto*$_{DC3000}$ have high dN/dS (Table 3.1). While this is consistent with similar studies in *E. coli*, it left me with very few genes on which to focus. If we are trying to understand the evolutionary pressures on *P. syringae* bacteria, it may also be useful to examine those genes that are under strong evolutionary constraint and thus have very low dN/dS. Of the 200 genes with the lowest dN/dS rank in the 3-way analysis, only 13 have no homologs in *P. putida* or *P. aeruginosa.* They are described in Table 4.1. These 13 genes may play important roles in virulence. Because they are not conserved in *P. putida* (soil microbe) or *P. aeruginosa* (mammalian pathogen, at least the PA01 strain used here), they may function specifically in plant pathogenesis. Because they are conserved in all three *P. syringae* strains, they may be part of the minimal *P. syringae* genome or have some indispensible function in causing disease on the plant leaf and/or fruit. Eleven of the 13 are also conserved in *P. syringae* pv. glycinea (*Pgl*) and pv. lacrymans strain 106 (*Pla*106) based on preliminary genomic sequence results (J. Reinhardt personal communication). Again, it is apparent that as we generate more diverse *P. syringae* genome sequences, the resolution power of this particular hypothesis will be strengthened.

candidate gene Pto_{DC3000}	*Pla106* and *Pgl* conservation	dN/dS	homolog
pspto1012	yes	0.0052	*psyr0874, pspph0912*
pspto1114	yes	0.005	*psyr0954, pspph1003*
pspto2179	yes	0.0023	*psyr1989, pspph1958*
pspto2581	yes	0.0056	*psyr2272, pspph2917*
pspto2584	yes	0.0112	*psyr2275, pspph2914*
pspto2585	yes	0.0048	*psyr2276, pspph2913*
pspto2640	yes	0.0077	*psyr2373, pspph2508*
pspto3315	yes	0.0075	*psyr3149, pspph3062*
pspto3488	yes	0.0097	*psyr3263, pspph3183*
pspto3493	yes	0.0104	*psyr3269, pspph3187*
pspto3675	no	0.0085	*psyr1801, pspph1761*
pspto4203	yes	0.003	*psyr3937, pspph3934*
pspto5062	no	0.007	*psyr0464, pspph0455*

Table 4.1: Low dN/dS Genes from 3-way comparison.

If these genes are indeed specific to *P. syringae* strains, I expect they will be required for plant pathogenesis or for some other aspect of *P. syringae* lifestyle.

Evolutionary analysis of open reading frames cannot predict all virulence factors.

Many known virulence factors are not candidate genes in either the rapidly-evolving candidate list discussed in Chapter 3 or the list of evolutionarily-constrained *P. syringae*-specific candidates listed in Table 4.1 above. Notably, the TTSS and its associated effectors were not identified by this bioinformatic screen with the exception of *avrPto*. This is partly because the TTSS and type III effectors are horizontally transferred (Buell 2003), so sequence diversification may not be a major mechanism controlling this important virulence factor. During infection, expression of the TTSS is tightly regulated (Mole et al., 2007) and it is likely that regulation of expression of other virulence factors is regulated as well. If the precise timing or protein level of a virulence factor is critical

to successful infection, we expect that the promoter region and not the coding region of that virulence factor will be evolving. Because promoters do not encode proteins, we cannot differentiate between synonymous and non-synonymous mutations in these regions, therefore this type of analysis will not work. Another barrier to a bioinformatic analysis of promoters is that promoter regions are often difficult to identify and only the promoters for genes of known function and consequence have been carefully mapped. Questions of the evolution of promoters would be better considered using a comparative transcriptional profiling technique if possible. This is underway in *Pectobacterium* to address regulation of the TTSS by the sigma factor *hrpL* in our lab.

While evolutionary analysis does not yield a comprehensive list of virulence factors, it can contribute to our understanding of pathogenesis by identifying novel genes involved in virulence.

References

Abramovitch, R. B., Anderson, J. C., and Martin, G. B. (2006). Bacterial elicitation and evasion of plant innate immunity. Nat Rev Mol Cell Biol *7*, 601-611.

Ausubel, F. M., Katagiri, F., Mindrinos, M., and Glazebrook, J. (1995). Use of Arabidopsis thaliana defense-related mutants to dissect the plant response to pathogens. Proc Natl Acad Sci U S A *92*, 4189-4196.

Chen, S. L., Hung, C. S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R. R., Ozersky, P.*, et al.* (2006). Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. Proc Natl Acad Sci U S A *103*, 5977-5982. Epub 2006 Apr 5973.

Gal-Mor, O., and Finlay, B. B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. Cell Microbiol *8*, 1707-1719. Epub 2006 Aug 1724.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol *11*, 725-736.

Grant, S. R., Fisher, E. J., Chang, J. H., Mole, B. M., and Dangl, J. L. (2006). Subterfuge and manipulation: type III effector proteins of phytopathogenic bacteria. Annu Rev Microbiol *60*, 425-449.

Katagiri, F., Thilmony, R., and He, S. Y. (2002). The Arabidopsis Thaliana-Pseudomonas Syringae Interaction. In The Arabidopsis Book, American Society of Plant Biologists, Rickville, MD.

Mole, B. M., Baltrus, D. A., Dangl, J. L., and Grant, S. R. (2007). Global virulence regulation networks in phytopathogenic bacteria. Trends Microbiol *15*, 363-371. Epub 2007 Jul 2012.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol *3*, 418-426.

Petersen, L., Bollback, J. P., Dimmic, M., Hubisz, M., and Nielsen, R. (2007). Genes under positive selection in Escherichia coli. Genome Res *17*, 1336-1343. Epub 2007 Aug 1333.

Plotkin, J. B., Dushoff, J., and Fraser, H. B. (2004). Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. Nature *428*, 942-945.

Sarkar, S. F., and Guttman, D. S. (2004). Evolution of the core genome of Pseudomonas syringae, a highly clonal, endemic plant pathogen. Appl Environ Microbiol *70*, 1999-2012.

Yan, S., Liu, H., Mohr, T. J., Jenrette, J., Chiodini, R., Zaccardelli, M., Setubal, J. C., and Vinatzer, B. A. (2008). Role of recombination in the evolution of the model plant pathogen Pseudomonas syringae pv. tomato DC3000, a very atypical tomato strain. Appl Environ Microbiol *74*, 3171-3181. Epub 2008 Mar 3131.