New hypotheses about the origin of *Pseudomonas syringae* crop pathogens

Rongman Cai

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Plant Pathology, Physiology and Weed Science

Boris A. Vinatzer, Chair
Scotland C. Leman
John G. Jelesko
Inyoung Kim

May 2, 2012
Blacksburg, Virginia

Keywords: *Pseudomonas syringae*, molecular evolution, host range evolution, population genetics, phylogenetic tree, ancestral state reconstruction, most recent common ancestor, microevolution, phylogeography, recombination, *HopM1*, *fliC*

New hypotheses about the origin of *Pseudomonas syringae* crop pathogens

Rongman Cai

Abstract

Pseudomonas syringae is a common foliar plant pathogenic bacterium that causes diseases on many crop plants. We hypothesized that today's highly virulent *P. syringae* crop pathogens with narrow host range might have evolved after the advent of agriculture from ancestral *P. syringae* strains with wide host range that were adapted to mixed plant communities. The model tomato and Arabidopsis pathogen *P. syringae* pv. *tomato* (*Pto*) DC3000 and its close relatives isolated from crop plants were thus selected to unravel basic principles of host range evolution by applying molecular evolutionary analysis and comparative genomics approaches. Phylogenetic analysis was combined with host range tests to reconstruct the host range of the most recent common ancestor of all analyzed strains isolated from crop plants. Even though reconstruction of host range of the most recent common ancestor of all analyzed strains was not conclusive, support for this hypothesis was found in some sub-groups of strains. The focus of my studies then turned to *Pto* T1, which was found to represent the most common *P. syringae* lineage causing bacterial speck disease on tomato world-wide. Five genomes were sequenced and compared to each other. Identical genotypes were found in North America and Europe suggesting frequent pathogen movement between these continents. Moreover, the type III-secreted effector gene *hopM1* was found to be under strong selection for loss of function and non-synonymous mutations in the *fliC* gene allowed to identify a region that triggers plant immunity. Finally, Pto *T1* was compared to closely related bacteria isolated from snow pack and surface water in the French Alps. Recombination between alpine strains and crop strains was inferred and

virulence gene repertoires of alpine strains and crop strains were found to overlap. Alpine strains cause disease on tomato and have relatively wider host ranges than Pto T1. The conclusion from these studies is that Pto T1 and other crop pathogens may have evolved from ancestors similar to the characterized environmental strains isolated in the French Alps by adapting their effector repertoire to individual crops becoming more virulent on these crops but losing virulence on other plants.

DEDICATION

This dissertation is dedicated to my family, all teachers and professors that taught, encouraged and inspired me, and all friends that helped me during my life. Without them, this dissertation would have never been possible.

ACKNOWLEDGEMENT

I would like to thank my advisor **Dr. Vinatzer**. During my Ph.D study, he had given me a lot of support, encouragement, suggestion, and training. I really enjoyed the time of research in his team. I also very appreciate my co-advisor **Dr. Scotland Leman** in the department of statistics who spent a lot of time to discuss with me about data analysis and papers concerned with my research. I also thank for the comments and discussions from other committee members that are **Dr. John Jelesko** and **Dr. Inyoung Kim**. In addition, I appreciate the help and assistant in the experiments from the lab technician **Haijie Liu**, graduated Ph.D students **Shuangchun Yan** and **Christopher Clarke**, visiting scholar **Marco Enrique Mechan Ilontop** and student worker **Tokia Goodman**.

I want to give my whole-hearted thanks to my husband **Zhanghan Wu**. During the last five years, he always stayed with me and encouraged me whenever and whatever difficulties we had met. Since last September, he took mega bus home from Washington DC every weekend after his graduation. Even though he was tired after taking 5-hour bus, he was glad to play with our daughter **Magnolia C. Wu**, teach her and cook when I have to work in the weekend. Thanks for our precious daughter. She is very pretty, sweet and smart. Her answers of “Thanks, mama” and “Mama work, daddy pei (means staying with daddy)” are great support for me in my life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature review

1. Bacterial Evolution

The majority of bacteria have both chromosomal DNA and relatively small circles of extra-chromosomal DNA called plasmids. Plasmids can be easily transferred between bacteria, greatly contributing to their genetic diversity. During the evolution, genes are passed down from generation to generation with small random differences. These differences are a result of mutation, recombination and acquisition or loss of genomic islands. The accumulation of such differences is responsible for the large quantity of existing bacterial species today (McVean, Awadalla et al. 2002).

1.1 Recombination

Recombination is one of the key evolutionary processes. Besides its functions in repair and disjunction of chromosomal DNA, recombination can lead to allelic variation. There are four common types of recombination (Figure 1.1), which are homologous recombination, site-specific recombination, transposition, and copy choice or strand transfer (Lewis-Rogers, Crandall et al. 2004). Homologous recombination can occur between any pair of homologous DNA molecules. DNA molecules exchange regions via breaking, invasion, ligation, branch migration, and holiday junction after synapsis

(Lodish, Berk et al. 2007). Site-specific recombination instead occurs only at specific sequences recognized by specific enzymes that perform recombination only at those specific sequences (Lodish, Berk et al. 2007). Bacteriophage DNA and bacterial plasmids can recombine with the main bacterial chromosome through attachment and crossing-over. Transposition is one of the site-specific recombination processes mediated by Transposable Elements (TEs), also called transposons, which include DNA-mediated transposons and RNA-mediated transposons. DNA-mediated transposons are able to move around and integrate into different positions in their host chromosome, whereas RNA-mediated transposons need transcription of DNA to mRNA and reverse transcription of mRNA to complementary DNA (cDNA) before moving and integrating. This last type of recombination often occurs in RNA viruses.

1.1 Mutation

Mutation is the second major evolutionary mechanism. Substitution is a one-nucleotide change, which can either change a codon to one encoding the same amino acid, a different amino acid, or an early stop codon (Madigan, Martinko et al. 2008). Insertions are mutations that add one or more extra base pair into DNA sequences, whereas deletions are mutations in which a section of DNA is lost. Both insertions and deletions can result in a frame shift mutation so that a number of nucleotides are not evenly divisible by three from a DNA sequence. Due to the triplet nature of gene expression by codons, insertions or deletions can change the reading frame, resulting in a completely different translation from the original (Madigan, Martinko et al. 2008).

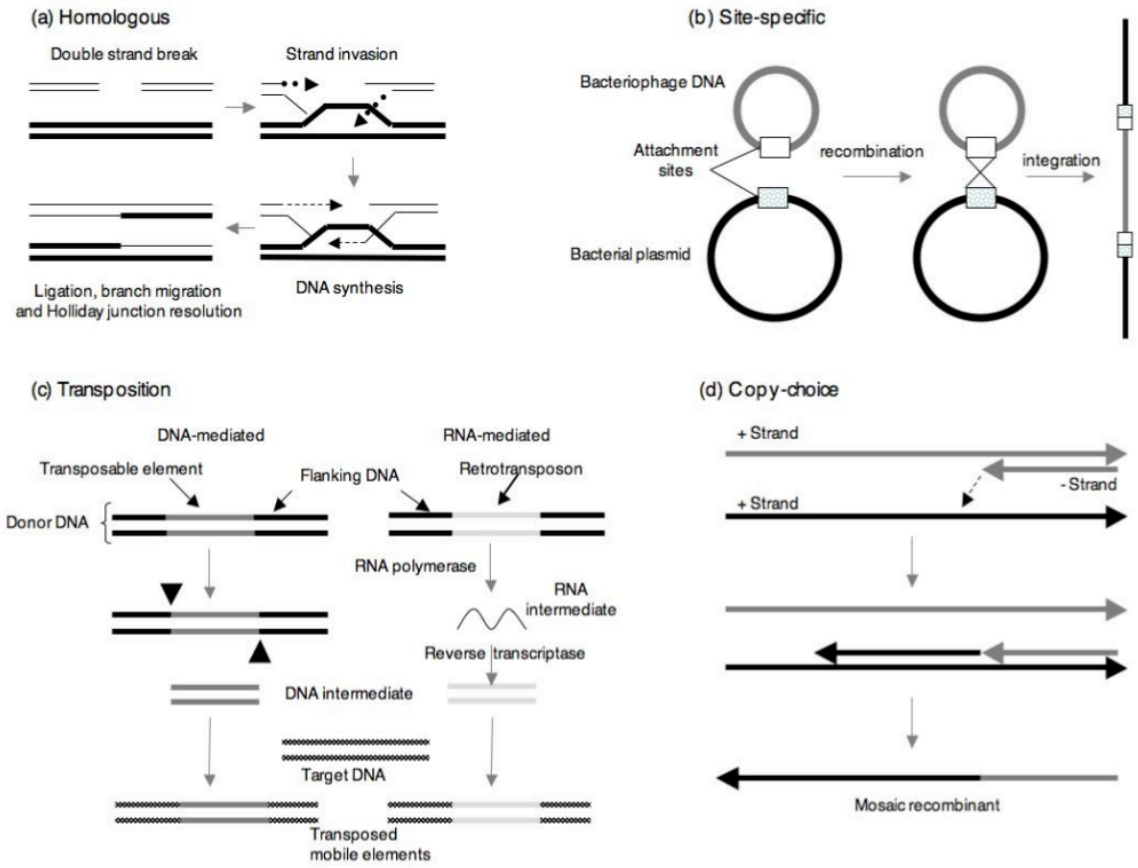


Figure 1.1 Four types of recombination (Lewis-Rogers, Crandall et al. 2004)

The estimation of synonymous (silent) to non-synonymous (non-silent) substitution rate is very important for the understanding of molecular evolution. In fact, the ratio of non-synonymous to synonymous substitution rate, that is, $w = dN/dS$, has been widely used as an indicator of natural selection acting on a protein (Yang and Nielsen 2000). There are three kinds of selection: positive selection (a new allele or mutant confers some increase in the fitness of the organism, or selection acts to favor this allele), negative selection (mutant confers some decrease in the fitness of organism, or selection acts to

remove this allele) and neutral selection (no selection due to random drift) (Wilson and McVean 2006). If the ratio of non-synonymous to synonymous substitution rate is much larger than 1, it indicates positive selection; if it is much less than 1, it implies negative selection. When the ratio is close to 1, it indicates neutral selection (Wilson and McVean 2006). Quantifying the ratio of non-synonymous to synonymous substitution rate needs to infer selection and requires explicit statistical models, which are incorporated into some well-developed programs, such as LDhat 2.1 (McVean, Awadalla et al. 2002), GARD (Sergei et al., 2006), PAML (Yang, 1997) and SplitsTree4 (Huson and Bryant, 2006).

2. Phylogeny Analysis

Evolutionary relationships among species have been studied to investigate their common ancestor. The construction of phylogenetic trees is one of the tools to visually represent the historic pattern of evolution. With the rapid accumulation of more and more DNA and protein sequences, the combination of phylogenies with new powerful statistical approaches has been developed for the study of evolution.

2.1 Coalescence Theory

The existing species obtained their genetic variation through evolution. Some methods, such as coalescence models, have been developed to trace back the genealogy of species in the backwards direction to represent their evolutionary history. The coalescence process (Figure 1.2) begins with current individuals and ends with a single

entity called Most Recent Common Ancestor (MRCA) into which all current individuals have coalesced (McVean, Awadalla et al. 2002).

The standard coalescence model is based on the following assumptions: mutations are neutral in which they do not affect the number of offsprings or the tendency of individuals to migrate bearing these mutations; the size of population is large and constant; recombination is negligible (Eldon and Wakeley 2006). Considering current individuals with the population size of N, which evolved according to the Wright-Fisher model (Fisher 1930) assuming that generations are discrete and that each new generation is formed by randomly sampling N parents with replacement from the current generation, two given sequences in any generation have descended from a common ancestor in previous generations with the probability of $1/N$. Since there are $\binom{k}{2}$ ways of selecting two sequences from objectives (k), the probability of coalescence in the previous generation is $\binom{k}{2}/N$. Therefore, for the number of generations (g) until coalescence, the probability of waiting g generations for coalescence can be specified as a geometric probability distribution with the parameter of $\binom{k}{2}/N$, which can be expressed as $P(G=g)=(1-\binom{k}{2}\Big/N)^{g-1}\binom{k}{2}\Big/N$. The expected time to coalescence through previous generation is $E(t_k)=N\Big/\binom{k}{2}$. The total coalescence time over the whole coalescence process is the sum of coalescence time in each generation, that is,

$E(T_{MRCA}) = E(\sum_{i=2}^{k} t_k) = 2N(1 - 1/k)$. Its deviation is a standard result of coalescence theory found in the work of Hudson (Hudson 1991) and Felsenstein (Felsenstein 2004).

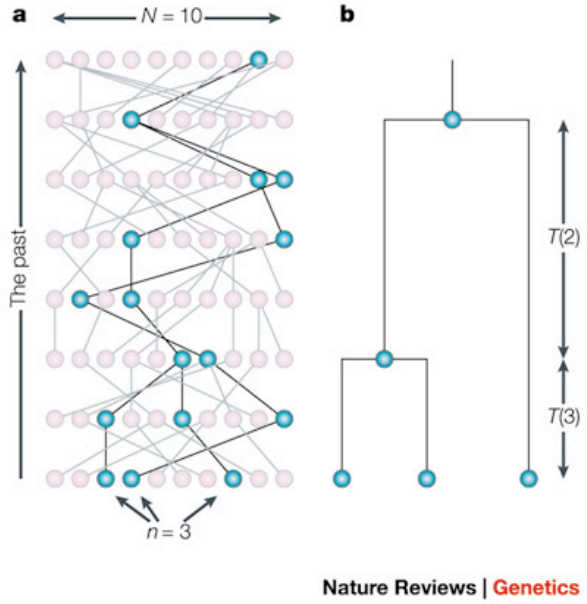


Figure 1.2 Coalescence process (Rosenberg and Nordborg 2002)

For the coalescence-based approach, it is necessary to include mutation and recombination occurring in nature into the model. The work of Hudson (Hudson 1991) and Felsenstein (Felsenstein 2004) states that the total number of mutations (D_t) during t generations approximates a Poisson distribution with mean value of ut, in which u is the mutation rate occurring in the single generation. Therefore, the expected number of differences between two sequences is the product of the expected coalescent time and mutation rate, that is, $E(\pi) = 4Nu$. N can be replaced with the scaled effective population size N_e, so that the amount of genetic diversity in the population is defined as $\theta = 4N_e u$. Because mutation and recombination could produce similar patterns of

genetic diversity (Figure 1.3), an approach based on the coalescent theory to estimate the population recombination rate (ρ) was suggested by Hudson (Hudson 2001), which is $\rho = 4N_e r$ and r is the rate of crossing-over per gene per generation. LDhat 2.1 is a program for the estimation of population mutation rate (u) and population recombination rate (ρ).

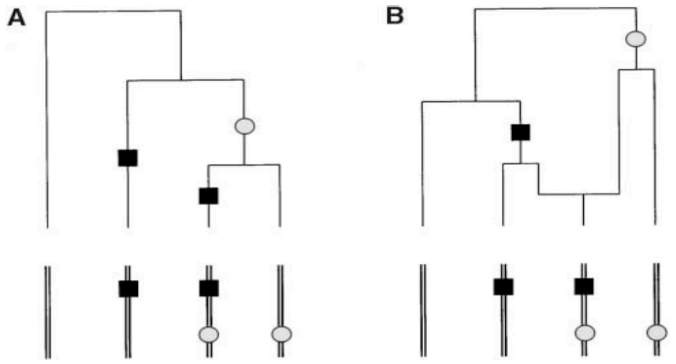


Figure 1.3 Recurrent mutation (A) and recombination (B) (McVean, Awadalla et al. 2002)

2.2 Phylogenetic Tree Construction

Evolutionary relationships among individuals are usually represented using a phylogenetic tree that consists of root, internal nodes, taxa or terminal nodes and branches. Taxa are the species, populations, or individuals at the tips of branches. Internal nodes represent hypothetical ancestors of nodes in the previous generations and the root is the ancestral node for all taxa. Branches connecting nodes indicate the relationship between ancestor and descendants. The branch pattern called the topology of tree characterizes the evolutionary relationships among all taxa.

2.2.1 Multilocus Sequencing Typing (MLST)

Multilocus Sequencing Typing (MLST) (Maiden, Bygraves et al. 1998) is a strain-typing system that focuses on the core genome of bacteria, which is the part of the genome encoding essential proteins, such as housekeeping genes. Besides the core genome, bacterial evolution also acts on the flexible genome that is responsible for the adaption to specific niches, hosts or environment (Sarkar and Guttman 2004). However, the flexible genome is usually not considered for MLST because it is too variable to reflect the evolutionary history of a species.

MLST is used to differentiate strains as well as to study the molecular evolution, population genetics, and epidemiology of a pathogen (Sarkar and Guttman 2004). MLST involves data collection, data analysis and multilocus sequence analysis. The nucleotide sequences of gene fragments from core genome loci are usually obtained by Polymerase Chain Reaction (PCR) and DNA sequencing of PCR products. Then, all unique sequences for each locus are assigned allele numbers and each strain is characterized by an allelic profile called Sequence Type (ST). Finally, the relationship of isolates is determined by analyzing allelic profiles, for example with the program eBurst (Feil, Li et al. 2004) or by analyzing the concatenated set of DNA sequences.

2.2.2 Models of Evolution

All phylogenetic methods make some assumptions about DNA substitution (transition and transversion) rates and other parameters. Some of the evolutionary models of DNA substitution include JC, K80, SYM, F81, HKY, and GTR (Felsenstein 2004). Their nested relationship is shown in Figure 1.4.

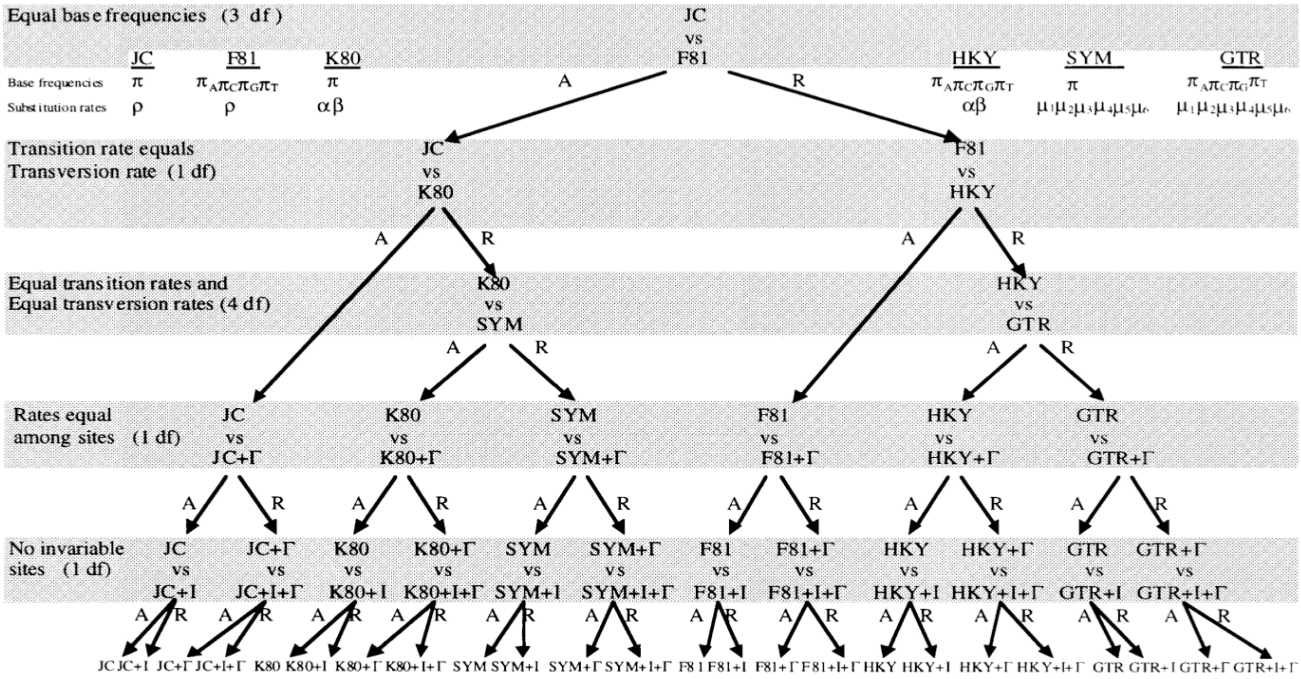


Figure 1.4 Hierarchical hypothesis testing in jModelTest (Posada and Crandall 1998)

To evaluate the above evolution models of DNA substitution, methods such as Likelihood Ratio (LR) tests, Akaike Information Criterion (AIC) and Bayes Factors can be used to select which model is the best to fit the data. The LR test is widely used to compare the maximized likelihoods of null (L_0) and alternative (L_1) models, that is, $\delta = 2\log\Lambda$ and $\Lambda = \frac{\max[L_0(NullModel)|Data]}{\max[L_1(AlternativeModel)|Data]}$ (Posada and Crandall 1998) . This method usually performs such pair-wise tests in a specific sequence until a final converged model cannot be rejected. When the compared models are nested and the null hypothesis is true, δ follows asymptotical χ^2 distribution with q degrees of freedom, which is the different number of free parameters between two models (Posada and Crandall 1998). If the associated P-value is smaller than the predefined Significant Level

(SL), usually SL=0.05, it indicates that the alternative model is significantly better than the null model to fit the data, and vice versa.

Another way of comparing different models without nested requirement is Akaike Information Criterion (AIC), that is $AIC = -2\ln(L) + 2n$, which is a function of maximum likelihood (L) and the number of independently estimated parameters within this model (n). Because AIC takes into account the number of parameters to be estimated achieving a particular degree of fit besides the statistical goodness of fit, it imposes a penalty for increasing the number of parameters (Posada and Buckley 2004). Since AIC represents the amount of information lost when one model is used to approximate another model, in the context of phylogeny, AIC can be taken as the amount of information lost when the model of General Time Reversible (GTR), for example, is used to approximate the real process of nucleotide substitution (Posada and Buckley 2004). Hence, lower values of the index indicate the preferred model.

Bayes factors can be taken as the Bayesian analogue of LR tests by comparing likelihoods, that is, $B_{ij} = \frac{P(D|M_i)}{P(D|M_j)}$, in which M_i and M_j are two competing models; D is dataset (Posada and Buckley 2004). If B_{ij} is larger than 150, Model j is considered to be very strong; If B_{ij} is in the range from 12 to 150, from 3 to 12 and from 1 to 3, Model j is considered to be strong, positive and barely worth mentioning, respectively; If B_{ij} is less than 1, Model j is considered to be negative. jModelTest (Posada and Crandall, 1998) is such a powerful program designed to compare different nested models of DNA

substitution (Figure 1.4) by choosing the above criteria. AIC is generally the right choice over BIC because the purpose of model selection is to identify how many parameters is useful for prediction rather than how many are non-zero. BIC is trying to tell us how many are non-zero.

2.2.3 Statistical Models of Phylogenetic Trees Construction

Phylogenetic trees are often constructed from molecular sequence data using methods of Neighbor-Joining (NJ), parsimony and likelihood. The distance methods such as NJ and Unweighted Group Method with Arithmetic Mean (UPGMA) are applied to construct the phylogentic tree after a distance matrix is computed. The parsimony approach is deterministic and tries to find the tree with the minimum number of changes required to explain data. More powerful methods, such as maximum likelihood, select the tree with the highest likelihood of generating data under an evolutionary model.

Bayesian inference of phylogeny is a well-developed approach, which produces a posterior probability ($\Pr[Tree|Data]$) of a tree through superposition of the prior probability of a phylogeny ($\Pr[Tree]$) with the likelihood ($\Pr[Data|Tree]$), that is, $\Pr[Tree|Data] = \frac{\Pr[Data|Tree]\Pr[Tree]}{\Pr[Data]}$ (Huelsenbeck, Ronquist et al. 2001). Usually, all trees are considered to have the equal prior probability, which means that the prior follows the uniform distribution, that is, $\Pr[Tree] \sim 1/N$, where N is the total number of tree topologies. Thus, the posterior probability is proportional to the likelihood, whose calculation involves the integration over all possible combination of branch lengths,

substitution parameters and gamma shape parameters representing rates across sites in the aligned matrix.

Because it is very difficult to directly calculate the posterior probability of a tree, Markov Chain Monte Carlo (MCMC), a powerful numerical method, can be used to approximately simulate the posterior probability. Firstly, a new tree is proposed by stochastically changing the branch length or topology of the current tree. Then, this tree is either accepted or rejected by calculating a probability described by Metropolis and Hasting (Chib and Greenberg 1995). If the new tree is accepted, then it is subjected to further perturbation. Thus, a sequence of such random variables constructs a Markov Chain, in which the transition probability between different values in the state space (the range of possible value of a random variable) only depends on the current state of the random variable (Gilks, Richardson et al. 1996). Finally, the Markov Chains would converge to a stationary distribution after running the chain long enough if it has the good mechanism for proposing new states.

2.3 Ancestral State Reconstruction

Ancestral state reconstruction is used to infer phenotypic character states at the interior hypothetic nodes and of the most recent common ancestor of a phylogeny if the characters for all terminal taxa in the phylogenetic tree are observed (Pagel 1994).

2.3.1 Continuous-time Markov Model of Character Evolution

The character states of observed terminal taxa can either be discrete or continuous. When describing discrete characters, the model of evolution adopts at least two states. Table 1.1 shows an example where this model adopts three states (0, 1 and 2), in which q_{ij} is the transition rate changing between any two states. The continuous-time Markov model describing the random evolution of discrete characters (Schluter, Price et al. 1997; Pagel and Meade 2006) has the following important features. The first is that the probability of state changes in time along the branches of the tree only depends on the current state at the time and does not depend on the previous state, which is also the principle of the Markov process. The second is that the change of states along all branches is independent, which means the state changes along one branch do not affect those along other branches. The third is that all states along one branch could repeatedly change between any two states. The last feature is that change rates are constant throughout time and along all branches.

Table 1.1 Example of the model of evolution for a trait that adopts three states

State	0	1	2
0	-	q_{01}	q_{02}
1	q_{10}	-	q_{12}
2	q_{20}	q_{21}	-

2.3.2 Inference of Character Evolution

Inference of character evolution can be performed using parsimony, Maximum Likelihood (ML) and Bayesian approaches. The parsimony method tries to find the reconstruction with the minimum amount of character changes in the tree, ignoring the uncertainty of mapping and phylogeny (Cunningham, Omland et al. 1998). When the evolution rates are quick and the probabilities of gains and losses are not equal, this approach can be misleading. The likelihood approach estimates the probabilities of all possible character state reconstructions at every node on the tree.

The Bayesian method is better than the other two approaches because it accounts for the uncertainty of both mapping and phylogeny and it also adds credibility to the reconstruction of evolutionary history. The work of Pagel (Pagel, Meade et al. 2004) gives us details of Bayesian estimation of ancestral character states on phylogeny. The posterior probability ($p(s_{ij}|D)_{i\in T}$) that character j is observed at node i (s_{ij}) in trees equals the probability of data integrated over parameters and trees given that node i adopts state j divided by the probability of data whether the tree contains node i or not. Here, Q denotes the matrix of rate coefficients (for example, matrix in Table 1.1) and D stands for the dataset of characters across the species in the tree (T).

$$p(s_{ij}|D)_{i\in T} = \frac{\int_T \int_Q p(D|s_{ij},Q,T)p(s_{ij})p(Q)P(T)dQdT}{\int_T \int_Q p(D|Q,T)p(Q)P(T)dQdT}$$

Software has been developed for the ancestral state reconstruction using the above three approaches, such as Mesquite (Maddison and Maddison 2001), MacClade

(Maddison and Maddison 2000) and BayesTraits (Pagel and Meade 2006). MacClade is restricted to parsimony reconstructions. Parsimony Reconstructions and Likelihood ancestral state reconstructions can be reported by Mesquite. Pagel's Multistate program of BayesTraits can also perform ancestral state reconstructions using global likelihoods for the states at the node. In addition, the Discrete program of BayesTraits can test correlations among characters using likelihood ratio tests.

2.3.3 Correlation of Phenotype with Phylogeny

Before the reconstruction of ancestral states, it is necessary to test whether phenotypic characters are correlated with their shared phylogeny or not. If there is no good correlation, it indicates that this particular phenotype may have evolved independently from the core genome. If there is good correlation, it indicates that this phenotype is the result of common ancestry from a single ancestral individual. Thus, it is meaningful to reconstruct ancestral states.

The association between characters at terminal taxa and phylogeny sometimes is very tight or completely interspersed, which makes it easy to visually determine whether there is association. However, when the distribution of characters is intermediate, the correlation with phylogeny is less clear. Bayesian Tip-association Significance Testing (BaTS) (Parker, Rambaut et al. 2008) is a program developed to determine the degree to which phenotypic characters are correlated with shared phylogeny through estimating Parsimony Score (PS), Association Index (AI) and Monophyletic Clade (MC) size, and providing the corresponding 95% Confidence Intervals (CI) for each estimate. PS

estimates the number of state changes in the phylogeny and low PS scores represent strong phylogeny-trait association due to the gain/loss of characters parsimoniously. AI ($AI = \sum_{i1}^{k} 1 - f_i / 2^{m_i - 1}$) is a sum across all the internal nodes in the phylogeny, in which k is the number of internal nodes; f_i is the frequency of the most common characters for each node i; m_i is the number of node subtended by node i. Low AI values indicate strong phylogeny-trait association. MC size for a particular character value of x is quantified by $MC(x) = \max_{i=1}^{k}(m_i I_i)$ with an indicator of I_i, in which I_i equals to 1 if characters subtended by node i have character value x, otherwise it equals to 0 (Parker, Rambaut et al. 2008). High MC values indicate strong phylogeny-tip association.

2.4 Whole Genome Trees

With the availability of complete genome sequences, we can obtain a large amount of phylogenetic information and study the evolution of genomes through the construction of whole genome trees. However, whole genome trees cannot simply be built like other phylogenetic trees using multi-sequence alignment because gene order rearrangement, gene loss and gene duplication at the whole genome level occur at a high rate (Snel, Huynen et al. 2005).

Genome trees are divided into five classes (Figure 1.5), which are alignment-free genome trees, gene content trees, gene order trees, genome trees based on average sequence similarity, and phylogenomics genome trees (Snel, Huynen et al. 2005).

Alignment-free genome trees are constructed through deriving a distance between genomes and then clustering them. This method is divided into two classes. One relies on word frequency, oligomers, K-strings (Qi, Wang et al. 2004) or n-mers in DNA or proteins. The second class is based on the information theory called “shared information”. Gene content trees based on presence or absence of genes are similar to a sequence-based phylogeny because a genome is simply treated as the sum of its genes. The choice of shared genes, orthology or homology, could result in different

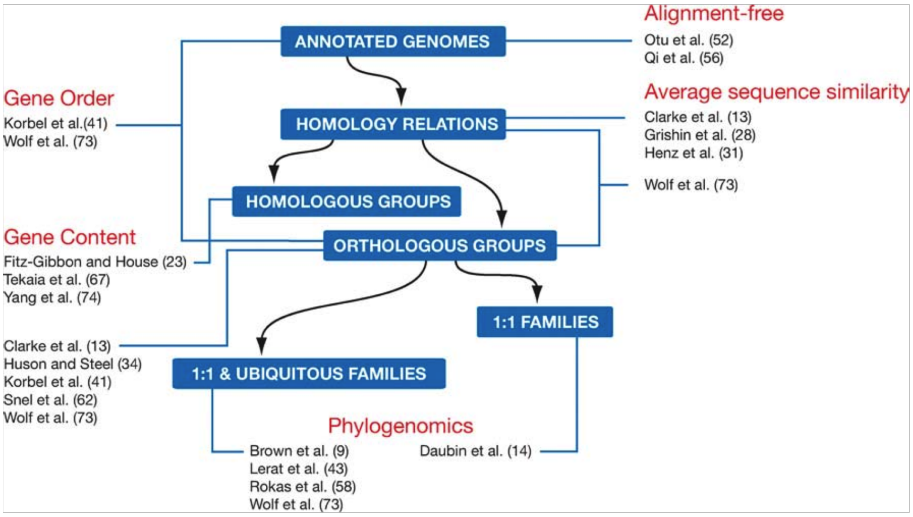


Figure 1.5 Classification of genome tree reconstruction methods (Snel, Huynen et al. 2005)

trees. Gene order trees are suitable for closely related species because gene order evolves faster than gene content. For genome trees based on the average sequence similarity, a distance matrix is obtained based on the average sequence similarity between genomes, without including any evolutionary models and without using maximum parsimony or likelihood when the phylogenetic tree is constructed.

Construction of the last class of trees, i.e., phylogenomics genome trees, is based either on the collection of phylogenetic trees derived from shared gene families or on a concatenated alignment of those families. For example, the homologous sequences from the different gene families are simply concatenated and aligned to produce a single tree (Snel, Huynen et al. 2005).

3. Pseudomonas syringae

Pseudomonas syringae is a plant pathogen that can causes various crop diseases, such as bacterial speck disease on tomato or bacterial blight disease on soybeans. According to the plant host from which strains were originally isolated, *P. syringae* strains are divided into approximately 50 pathovars (Sarkar and Guttman 2004). Some of them have been widely used as model organisms to study bacterial pathogenesis in plants, such as *P. syringae* pv. *syringae* (*Pss*), *P. syringae* pv. *tomato* (*Pst*), *P. syringae* pv. *phaseolicola* (*Psph*) (Hirano and Upper 2000).

3.1 PAMP and PAMP-triggered Immunity (PTI)

When plants are infected by *P. syringae,* plant membrane proteins called pattern recognition receptors can perceive molecular signatures characteristic of microbes termed Pathogen-Associated Molecular Patterns (PAMPs), so that a basal defense response is triggered. Faced with PAMP-Triggered Immunity (PTI), successful pathogens can secrete so-called effector proteins into plant cells to overcome these PAMP-induced defenses (Chisholm, Coaker et al. 2006). Some plants evolved

Resistance (R) proteins that are able to directly or indirectly detect these effectors (previously termed Avr proteins) and induce Effector-Triggered Immunity (ETI). This is often accompanied by the Hypersensitive Response (HR), which consists in death of infected plant cells within hours after initial contact with the pathogen (Agrios 1988). The protein flagellin and bacterial cell walls are important sources of PAMPs. Lipopolysaccharide (LPS), the principle component of the outer membrane of gram-negative bacteria, also acts as a PAMP (Nicaise, Roux et al. 2009).

3.2 Type III secretion System (T3SS) and Type III Effectors (T3E)-triggered Immunity

The Type III Secretion System (T3SS) plays a central role in virulence and host specificity in bacterial pathogens of both plants and animals, for example, *P. syringae*, *Ralstonia solanacearum*, *Xanthomonas* and *Erwinia* species (McCann and Guttman 2008). Many bacterial plant pathogens use a Type III Secretion System (T3SS) to inject effectors into plant cells (McCann and Guttman 2008). *P. syringae* utilizes the T3SS to deliver a collection of 15 to 30 effectors. In *P. syringae*, *hrp* (HR and pathogenicity) genes and *hrc* (HR and conserved) genes encode the T3SS. *avr* (avirulence) genes and *hop* (*Hrp*-dependent outer protein) genes both encode effector proteins (Collmer, Badel et al. 2000). However, plants evolved resistance (R) proteins that are able to directly or indirectly detect effectors (previously termed Avr proteins) and trigger a strong response termed Effector-Triggered Immunity (ETI). Such resistance responses can result in strengthening of the cell wall via callose deposition, expression of pathogenesis-related proteins, release of oxidative radicals and the HR (McCann and Guttman 2008). This

resistance is often accompanied by Hypersensitive Response (HR), which consists in death of infected plant cells within hours after initial contact with the pathogen (Agrios 1988). This kind of resistances is the base of gene-for-gene resistance between resistance genes in the host and avirulence (*avr*) genes in the pathogen that were first described by Flor (Flor 1942).

3.3 Evolution of Virulence and Host Specificity

Some of virulent pathogens do not release effectors that could promote ETI, and some of them do not have Microbe Associated Molecular Patterns (MAMP) alleles that minimize Microbe-Triggered Immunity (MTI). However, some virulent pathogens have evolved being able to suppress defenses in the relatively small quantity of their hosts. These pathogens can release some kinds of effectors or toxins, which could result in the suppression of MTI and ETI. Such evolutions of virulent pathogens also impose intensive selective pressure on their hosts so that they could evolve more effective and efficient defenses to pathogens in a host-specific manner (Nomura, Melotto et al. 2005).

In general, the host specificity is controlled by “gene-for-gene” interactions in which dominant allele in the host and dominant allele in the pathogen determines the outcomes of plant-pathogen interactions (Flor 1942 and Buell, Joardar et al. 2003). For example, in *P. syringae* the host-specific interactions with plants at least result from pathogens that encode proteins and trigger the resistance-gene-based plant innate immunity in resistant plants. The host specificity is also associated with the host range of pathogens. Sometimes a single effector released by the pathogen can restrict host

range. For example, hopQ1-1 excludes *N. benthamiana* form the host range of DC3000 while other times more than one effector and probably T3SS-independent factors also contribute to host range (Almeida, Yan et al. 2008).

Non-host resistance is the phenomenon that an entire plant species is immune to a particular pathogen (Thordal-Christensen 2003). On the non-host, some virulent pathogens are poorly adapted to the basic physiology of a plant and are unable to overcome the plant defense through suppressing MTI and ETI. Understanding the mechanisms that pathogens evolved to adapt to new hosts is fundamental for the study of the origin of diseases and of emerging pathogens to breed or engineer plants with durable disease resistance.

REFERENCES

Agrios, G., Ed. (1988). Plant Pathology. London, Academic Press.

Buddenhagen, I., L. Sequeira, et al. (1962). "Designation of races in Pseudomonas solanacearum." Phytopathology **52**(726).

Carver, T., M. Berriman, et al. (2008). "Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database." Bioinformatics **24**(23): 2672-2676.

Chen, F., A. J. Mackey, et al. (2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups." Nucleic Acids Research **34**(suppl 1): D363-D368.

Chib, S. and E. Greenberg (1995). "Understanding the Metropolis-Hastings Algorithm." The American Statistician **49**(4): 327-335.

Chisholm, S. T., G. Coaker, et al. (2006). "Host-Microbe Interactions: Shaping the Evolution of the Plant Immune Response " Cell **124**(4): 803-814.

Cunningham, C. W., K. E. Omland, et al. (1998). "Reconstructing ancestral character states: a critical reappraisal." Trends in Ecology & Evolution **13**(9): 361-366.

Draper, N. R. and H. Smith, Eds. (1966). Applied regrssion analysis. New York, NY.

Elahi, E. and M. Ronaghi (2004). Pyrosequencing. Bacterial Artificial Chromosomes. **255:** 211-219.

Eldon, B. and J. Wakeley (2006). "Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed." Genetics **172**(4): 2621-2633.

Feil, E. J., B. C. Li, et al. (2004). "eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data." The Journal of Bacteriology **186**(5): 1518-1530.

Felsenstein, J., Ed. (2004). Inferring Phylogenies.

Fisher, R. A., Ed. (1930). The Genetical Theory of Natrual Selection., Clarendon Press, Oxford.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." Syst. Zool **19**: 15.

Gabriel, D. W., C. Allen, et al. (2007). "Identification of Open Reading Frames Unique to a Select Agent: Ralstonia solanacearum Race 3 Biovar 2." Molecular Plant-Microbe Interactions **19**(1): 69-79.

Gilks, W. R., S. Richardson, et al., Eds. (1996). Markov Chain Monte Carlo in practice, Chapman & Hall/CRC.

Hamilton, J. D., Ed. (1994). Time Series Analysis New Jersey, Princeton University Press.

Hayward, A. C. (1991). "Biology and Epidemiology of Bacterial Wilt Caused by Pseudomonas Solanacearum." Annual Review of Phytopathology **29**(1): 65-87.

Hirano, S. S. and C. D. Upper (2000). "Bacteria in the Leaf Ecosystem with Emphasis on Pseudomonas syringae---a Pathogen, Ice Nucleus, and Epiphyte." Microbiology and Molecular Biology Reviews **64**(3): 624-653.

Hudson, R. R. (1991). "Gene genealogies and the coalescent process." Oxford Surveys in Evolutionary Biology **7**: 44.

Hudson, R. R. (2001). "Two-Locus Sampling Distributions and Their Application." Genetics **159**(4): 1805-1817.

Huelsenbeck, J. P., F. Ronquist, et al. (2001). "Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology." Science **294**(5550): 2310-2314.

Huson, D. H. and Bryant D., "Application of Phylogenetic Networks in Evolutionary Studies." Mol. Biol. Evol., 23(2):254-267, 2006.

Lambert, C. D. (2002). Agricultural Bioterrorism Protection Act of 2002: Possession, use, and transfer of biological agents and toxins; interim final rule. (7 CFR Part 331). Agricuture, Fed. Regist. **67:** 76908-76938.

Lewis-Rogers, N., K. A. Crandall, et al. (2004). "Evolutionary analysis of genetic recombination." Dynamical Genetics: 49-78.

Lewis, S., S. Searle, et al. (2002). "Apollo: a sequence annotation editor." Genome Biology **3**(12): research0082.0081 - 0082.0014.

Li, L., C. J. Stoeckert, et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome research **13**(9): 2178-2189.

Lin, J. and M. Gerstein (2000). "Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels." Genome Res **10**(6): 11.

Liu, C., T. I. Bonner, et al. (2003). "DNannotator: annotation software tool kit for regional genomic sequences." Nucleic Acids Research **31**(13): 3729-3735.

Lodish, H., A. Berk, et al., Eds. (2007). Molecular Cell Biology, 6th edition, Sara Tenney.

Maddison, D. and W. Maddison, Eds. (2000). MacClade 4: Analysis of Phylogeny and Character Evolution.

Maddison, W. and D. R. Maddison (2001). "Mesquite: a modular system for evolutionary analysis."

Madigan, M. T., J. M. Martinko, et al. (2008). "Brock Biology of Microorganisms, 12 edition."

Maiden, M. C. J., J. A. Bygraves, et al. (1998). "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms." Proceedings of the National Academy of Sciences of the United States of America **95**(6): 3140-3145.

McCann, H. C. and D. S. Guttman (2008). "Evolution of the type III secretion system and its effectors in plant-microbe interactions." New Phytologist **177**(1): 33-47.

Morelli, G., Y. Song, et al. (2010). "Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity." Nat Genet advance online publication.

McVean, G., P. Awadalla, et al. (2002). "A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences." Genetics **160**(3): 1231-1241.

Nicaise, V., M. Roux, et al. (2009). "Recent Advances in PAMP-Triggered Immunity against Bacteria: Pattern Recognition Receptors Watch over and Raise the Alarm." PLANT PHYSIOLOGY **150**(4): 1638-1647.

Pagel, M. (1994). "Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters." Proceedings of the Royal Society of London. Series B: Biological Sciences **255**(1342): 37-45.

Pagel, M. and A. Meade (2006). "BayesTraits." from http://www.evolution.rdg.ac.uk/BayesTraits.html.

Pagel, M., A. Meade, et al. (2004). "Bayesian Estimation of Ancestral Character States on Phylogenies." Systematic Biology **53**(5): 673-684.

Parker, J., A. Rambaut, et al. (2008). "Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty." Infection, Genetics and Evolution **8**(3): 239-246.

Posada, D. and T. R. Buckley (2004). "Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests." Systematic Biology **53**(5): 793-808.

Posada, D. and K. A. Crandall (1998). "JMODELTEST: testing the model of DNA substitution." Bioinformatics **14**(9): 817-818.

Poueymiro, M. and S. Genin (2009). "Secreted proteins from Ralstonia solanacearum: a hundred tricks to kill a plant." Current Opinion in Microbiology **12**(1): 44-52.

Poueymiro, M., S. Cunnac, et al. (2009). "Two Type III Secretion System Effectors from Ralstonia solanacearum GMI1000 Determine Host-Range Specificity on Tobacco." Molecular Plant-Microbe Interactions **22**(5): 538-550.

Qi, J., B. Wang, et al. (2004). "Whole Proteome Prokaryote Phylogeny Without Sequence Alignment: A K-String Composition Approach." Molecular Evolution **58**: 11.

Quail, M. A., I. Kozarewa, et al. (2008). "A large genome center's improvements to the Illumina sequencing system." Nat Meth **5**(12): 1005-1010.

Remenant, B., B. Coupat-Goutaland, et al. (2010). "Genomes of three tomato pathogens within the Ralstonia solanacearum species complex reveal significant evolutionary divergence." BMC Genomics **11**(1): 379.

Rosenberg, N. A. and M. Nordborg (2002). "Genealogical trees, coalescent theory and the analysis of genetic polymorphisms." Nat Rev Genet **3**(5): 380-390.

Sarkar, S. F. and D. S. Guttman (2004). "Evolution of the Core Genome of Pseudomonas syringae, a Highly Clonal, Endemic Plant Pathogen." Applied and Environmental Microbiology **70**(4): 1999-2012.

Schluter, D., T. Price, et al. (1997). "Likelihood of ancestor states in adaptive radiation." Evolution **51**(6): 1699-1711.

Sergei L. Kosakovsky Pond, David Posada, Michael B. Gravenor, Christopher H. Woelk, and Simon D.W. Frost. "GARD: a genetic algorithm for recombination detection." Bioinformatics (2006) 22(24): 3096-3098.

Snel, B., M. A. Huynen, et al. (2005). "Genome Trees and the Nature of Genome Evolution." Annual Review of Microbiology **59**(1): 191-209.

Stein, L. (2001). "Genome annotation: from sequence to biology." Nature Review Genetics **2**: 11.

Warren, A. and J. Setubal (2009). "The Genome Reverse Compiler: an explorative annotation tool." BMC Bioinformatics **10**(1): 35.

Wilson, D. J. and G. McVean (2006). "Estimating diversifying selection and functional constraint in the presence of recombination." Genetics **172**: 1411-1425.

Yan, S., H. Liu, et al. (2008). "Role of Recombination in the Evolution of the Model Plant Pathogen Pseudomonas syringae pv. tomato DC3000, a Very Atypical Tomato Strain." Applied and Environmental Microbiology **74**(10): 3171-3181.

Yang, Z. 1997. "PAML: a program package for phulogenetic analysis by maximum likelihood." Computer Application in BioSciences 13: 555-556.

Yang, Z. and R. Nielsen (2000). "Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models." Molecular Biology and Evolution **17**(1): 32-43.

Zerbino, D. R. and E. Birney (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." Genome Res. **18**: 821-829.

Chapter 2

Reconstructing Host Range Evolution of Bacterial Plant Pathogens using *Pseudomonas syringae* pv. *tomato* and Its Close Relatives as a Model

Rongman Cai[a], Shuangchun Yan[a], Haijie Liu[a], Scotland Leman[b], Boris A Vinatzer[a]

[a] Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Latham Hall, Ag Quad Lane, Blacksburg, VA-24061, USA;

[b] Department of Statistics, Virginia Tech, Blacksburg, VA-24061, USA

Corresponding author: Boris A Vinatzer, Latham Hall, Ag Quad Lane, Blacksburg, VA-24061, USA, phone +1 540 231 2126, Fax 540 231 3347, e-mail vinatzer@vt.edu

Abstract

Several lines of evidence suggest that highly virulent bacterial human pathogens evolved from less virulent wider host range animal pathogens since human migration out of Africa. To investigate evolution of host specificity of bacterial plant pathogens, here we report a molecular evolutionary analysis of the model plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000 and of close relatives that are pathogens of a diverse set of crop plants. Extensive host range tests on five different plant species were performed. Combining phylogenetic data with host range data, a reconstruction of host range of all putative ancestors was performed. In particular, the hypothesis was tested that highly virulent narrow host range pathogens of today's crops grown in monoculture evolved from ancestors with wider host range that were adapted to natural mixed plant communities of pre-agricultural times. We found support for this hypothesis in individual clades. However, reconstruction of host range of the most recent common ancestor of all analyzed strains was not conclusive. Based on the obtained results we stress the importance of including pathogens from wild plants when reconstructing the evolution of plant pathogenic bacteria.

Keywords

Host range evolution, bacterial plant pathogens, ancestral state reconstruction

1. Introduction

Human, animal, and plant pathogens can display different degrees of host specialization. Some cause disease on many host species while others are highly adapted to a single species. Molecular evolutionary analyses revealed that several highly specialized and highly virulent bacterial human pathogens that exist today evolved from relatively broad host range mild animal pathogens, for example, *Yersinia pestis* from an ancestor similar to *Y. pseudotuberculosis* (Achtman et al., 2004), *Salmonella enterica* serovar Typhi from an ancestor similar to *Salmonella enterica* serovar Typhimurium (Roumagnac et al., 2006), or *Bordetella pertussis* from a pathogen similar to *B. bronchiseptica* (Musser et al., 1986). Pathogen specialization to humans was probably favored by the high population densities that came with the beginning of human civilization (Achtman, 2008). Less is known about the evolution of host range in bacterial crop pathogens. Did today's highly specialized and virulent bacterial pathogens only recently adapt to crops grown at high density in agricultural monoculture by evolving from broad host range ancestors that were adapted to natural mixed – plant communities of pre-agricultural times? Evidence for such a scenario has been found for several fungal and oomycete plant pathogens by comparing specialized crop pathogens to closely related pathogens of wild plants as reviewed by Stukenbrock and McDonald (2008).

What we do know about bacterial plant pathogens is that they probably adapted relatively recently to different hosts since closely related pathogen strains with different host ranges do exist, see for example Yan et al. (2008). However, did these pathogens

with different host range restrict their host range during evolution or did ancestors simply switch from one host species to another?

It is believed that acquisition and loss of genes coding for type III secreted effector proteins are key events during host range evolution because effectors contribute to pathogen - host specificity by either triggering or suppressing plant immunity depending on the plant species that is being infected (Lindeberg et al., 2009). Immunity is triggered when an effector is directly or indirectly recognized by a resistance gene in the plant leading to “gene for gene resistance”, also called effector triggered immunity (ETI) while immunity is suppressed when an effector interferes with the activity of a component of the plant immune system (Chisholm et al., 2006). Also Pathogen associated molecular patterns (PAMPs), like flagellin, trigger plant immunity. Since it has been shown that different alleles of flagellin from strains with different host range greatly influence bacterial growth *in planta* (Takeuchi et al., 2003) and even closely related plant pathogen strains have different flagellin alleles (Cai et al., in press; Sun et al., 2006), also mutations in PAMP-encoding genes appear to contribute to host range evolution. However, the relative importance of effector loss and acquisition, allelic variation in PAMP-encoding genes and of possibly yet unknown factors, during host range evolution has not been investigated in any plant pathogen group yet.

Pseudomonas syringae is a great model pathogen to uncover the evolution of host range and the molecular basis of host specificity in bacterial plant pathogens because closely related strains with many different host ranges exist in this pathogen

group (Sarkar and Guttman, 2004). This makes it possible to use a comparative evolutionary genomics approach to find the genes that differ between closely related strains (Sarkar et al., 2006) with different host specificity and to infer how these genes were mutated, acquired, or lost during evolution. The closer the analyzed strains are related to each other, the fewer differences can be assumed to exist between them, and the easier it will be to identify how these differences arose during evolution.

Within *P. syringae*, the *Arabidopsis thaliana* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000 and its close relatives present a group of closely related strains particularly well suited for studies of host range evolution. Strain DC3000 is a rifampicin resistant derivative of strain NCPPB1106, the type strain of pathovar *tomato* (Cuppels, 1986). Pathovars are subspecific designations within *P. syringae* to indicate host range of particular strains (Dye et al., 1980). Strain NCPPB1106 was assigned to genomospecies 3 together with strains of pathovars *persicae*, *antirrhini*, *maculicola*, *viburni*, *apii*, *delphinii*, *passiflorae*, *philadelphi*, *ribicola*, and *primulae* based on DNA – DNA hybridization (DDH) indicating that strains of these pathovars are closely related (Gardan et al., 1999). Using multilocus sequence typing [MLST (Maiden et al., 1998)], Sarkar and Guttman (2004), Hwang et al. (2005), and Yan et al. (2008) confirmed that pv. *antirrhini*, *apii*, *lachrymans*, and *maculicola* are closely related to pv. *tomato* and thus recently evolved from a common ancestor. Yan et al. (2008) also found that strains belonging to these pathovars have different host ranges.

MLST allows inferring the relative contribution of mutation and recombination to the sequence differences found between strains in the analyzed housekeeping genes (Maiden, 2006). Recombination can be the result of transformation, transduction, or conjugation between related strains whereby recipient DNA is replaced with donor DNA through homologous recombination between similar DNA sequences. While Sarkar and Guttman (2004) concluded that recombination did not significantly contribute to sequence differences in *P. syringae* strains, Yan et al. (2008) found evidence that recombination played an important role in the evolution of pv. *tomato* strains and related strains from other pathovars.

While recombination affects all genes of a bacterial genome, horizontal gene transfer followed by site-specific recombination is limited to defined genomic regions that carry accessory genes implicated in adaptation to specific ecological niches (Dobrindt et al., 2004). In the case of pathogens, these genomic regions often carry virulence genes, like type III secreted effectors, and are called pathogenicity islands (Hacker et al., 1997). Genome comparisons (Lindeberg et al., 2008) and experimental evolution experiments (Lovell et al., 2009) have shown that also *P. syringae* strains acquire type III secreted effectors by horizontal gene transfer of pathogenicity islands. Moreover, Yan et al. (2008) found evidence that homologous recombination may also have contributed to differences in effector gene repertoires and, consequently, to host range differences between strains.

Once phylogenetic reconstruction and host range characterization is completed these data could be correlated with each other using “ancestral state reconstruction” (Pagel et al., 2004; Ronquist, 2004) to infer the host range of all hypothetical ancestors of the analyzed extant strains. We show and discuss here the results obtained for strain DC3000 and its relatives in regard to phylogenetic reconstruction, recombination analysis, host range, and ancestral state reconstruction of host range. We discuss the potential of our approach and its limitations and show how results here obtained provide the basis for future investigations into the molecular events that are at the basis of host switching in bacterial plant pathogens.

2. Materials and Methods

2.1 Strains, PCR, and DNA sequencing

Bacterial strains were either purchased from culture collections or generously provided by researchers around the world. Strains closely related to DC3000 were selected for further analysis. These are described in Table 2.1. Genomic DNA was prepared using the Puregene DNA purification system cell and tissue kit (Qiagen, Valenica, California). Primers used for PCR were described previously (Yan et al., 2008) besides primers for genes PSPTOT1_0038 (TTCCGGGATGTCCAGATG and ACCATCGTCGACCTGCAAC), PSPTOT1_2358 (CGACGTCGAAATTCAGCTC and AGGCACTGGTGCCCAACT), and PSPTOT1_1665 (CTCACCACCCAACACCAACT and ATCTCGTCGTAACGCTGGTT). PCR was performed, and PCR products were

prepared for sequencing as previously described (Yan et al., 2008). Sequencing was performed at the University of Chicago Cancer Research Center DNA sequencing facility.

2.2 Plant Infections

Plants of the species *Arabidopsis thaliana*, cauliflower (*Brassica oleracea* var. *botrytis*), celery (*Apium graveolens*), snapdragon (*Anthirrhinum majus*) and tomato (*Solanum lycopersicum*) were grown and infected with *P. syringae* strains as previously described (Yan et al., 2008). Concentration of inoculum was as follows (measured at an optical density of 600 nm): 0.01 for *A. thaliana*, 0.1 for cauliflower and celery, 0.001 for tomato, 0.02 for snapdragon. The surfactant Silwet L-77 was added at a final concentration of 0.02% to ensure uniform distribution of bacteria on the surface of leaves. Since the leaf surface of celery and snapdragon is waxy, 4g/L carborundum was added to the inoculum for these two plant species to create wounds facilitating bacterial invasion. Bacterial population sizes were determined 3 days post infection by dilution plating (7 days post infection for celery and snapdragon). For each strain/plant combination, infections were performed at least three times.

2.3 Phylogenetic Analysis

Chromatograms were reviewed and edited with SeqMan (Lasergene; DNAStar, Madison, WI). All sequences were uploaded to the PAMDB website (Almeida et al., 2010), which was used to concatenate sequences. Using MODELTEST (Posada and Crandall, 1998)

we determined the evolutionary models most appropriate for analyzing our concatenated gene sequences through application of the Likelihood Ratio (LR) test. Bayesian trees were constructed in MrBayes version 3.1 (http://mrbayes.csit.fsu.edu/), based on the evolutionary models suggested by MODELTEST. In order to assure convergence of our Bayesian trees, either $2*10^7$ (for concatenated genes) or $1*10^7$ (for individual genes) Markov Chain Monte Carlo (MCMC) iterations were used. Inspection of trace plots, from multiple runs, based on every 1,000th thinned sample demonstrated very good convergence.

To test whether there was phylogenetic congruence between MrBayes gene trees, the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) was performed in PAUP version 4.0. The program GARD (Genetic Algorithm for Recombination Detection) was used to localize recombination break points (Kosakovsky Pond et al., 2006). A phylogenetic network was constructed using the split decomposition algorithm (Huson, 1998) in SplitsTree version 4 (http://www.splitstree.org/) and applying the ParsimonySplits method, NeighborNet distance with 1000 bootstrap replicates.

2.4 Population Genetic Analysis

Genetic distances were calculated by using the methods of Jukes-Cantor, Tajima-Nei, p-distance and Maximum Composite Likelihood in MEGA version 4.0 (Tamura et al., 2007). 1000 bootstrap replicates were chosen for estimating the standard error of the mean genetic distance.

A number of recombination tests were performed in LDhat version 2.1 (Auton and McVean, 2007; McVean et al., 2002), which uses a coalescence-based method for detecting linkage disequilibrium. Within LDhat we used the program 'convert' to calculate the Watterson's infinite-sites estimator of the population-scaled mutation rate (θ) applied to the whole region and Tajima's D. We used the program 'pairwise' to estimate the population recombination rate (ρ).

2.5 Statistical analysis of host range

We relied on a logistic regression model for estimating the probability of each strain to be pathogenic, i.e., to cause disease, on each plant type. For our purposes, the binary response of the model represents the presence or the absence of disease on plant *i*, for each plant species in our study. Disease symptoms were recorded through visual inspection of photos taken of each plant and bacterial growth was determined by counting the number of colony forming units (cfu) present in samples of extracts from the same plant. For each observation of presence or absence of disease symptoms, we link the log-odds ratio to both the strain and plant types (categorical variables), and log-value of the cfu present in the plant extract of inoculated plants (continuous variable). Explicitly, the connection between the probability of disease, and explanatory covariates follows as:

$$\log\left(\frac{P_{i,j,k}}{1-P_{i,j,k}}\right) = \alpha_j^{(p)} + \alpha_k^{(s)} + \beta X_i \ ,$$

where $\Pr(Y_{i,j,k}=1)=P_{i,j,k}$ $\Pr(Y_{i,j,k}=1)=P_{i,j,k}$ is the probability that plant *i* of type *j*={*Arabidopsis* (Col, Mt), tomato (Chico III, Rio Grande), cauliflower, celery, snapdragon} and exposed to strain type *k* is diseased. $\alpha_j^{(p)}$ and $\alpha_k^{(s)}$ $\alpha_j^{(p)}$ and $\alpha_k^{(s)}$ represent both plant and strain specific intercepts. Additionally, X_i X_i represents the log-value of number of cfu.

The estimable parameters follow as: $\alpha_j^{(p)}$, $\alpha_k^{(s)}$ and β. We report the maximum likelihood estimates for each of these, along with their corresponding intervals. The full likelihood model follows as:

$$L(\alpha_1^{(p)},...,\alpha_J^{(p)},\alpha_1^{(s)},...,\alpha_K^{(s)},\beta)=\prod_i (P_{i,j,k})^{Y_{i,j,k}}(1-P_{i,j,k})^{1-Y_{i,j,k}}$$

$$L\left(\alpha_1^{(p)},...\alpha_J^{(p)},\alpha_1^{(s)},...,\alpha_K^{(s)},\beta\right)=\prod_i (P_{i,j,k})^{Y_{i,j,k}}(1-P_{i,j,k})^{1-Y_{i,j,k}}.$$

Based on our MLE estimates ($\hat{\alpha}_j^{(p)}$, $\hat{\alpha}_k^{(s)}$ and $\hat{\beta}$), we report estimates for the probability of plant *j*, getting diseased after exposure to strain *k*, by calculating:

$$\hat{P}_{j,k}=\frac{\exp(\hat{\alpha}_j^{(p)}+\hat{\alpha}_k^{(s)}+\hat{\beta}\overline{X}_{jk})}{1+\exp(\hat{\alpha}_j^{(p)}+\hat{\alpha}_k^{(s)}+\hat{\beta}\overline{X}_{jk})},$$

where $\overline{X}_{jk}$ $\overline{X}_{jk}$represents the average log cfu value across all plant *j*/strain *k* pairs.

Parameter estimates ($\hat{\alpha}_j^{(p)}$, $\hat{\alpha}_k^{(s)}$) which are far removed from 0 demonstrate the propensity for plant *j* to get diseased when exposed to strain *k*.

2.6 Ancestral State Reconstruction

The program BaTS (Parker et al., 2008) was used to determine the degree to which pathogenicity was correlated with shared phylogeny. BaTS can estimate the Association Index (AI), Fitch Parsimony Score (PS) and maximum exclusive single-state clade size (MC).

The program MultipleState implemented in BayesTraits (Pagel et al., 2004) was used to perform ancestral state reconstruction of pathogenicity. BayesTraits infers ancestral states at each node of a tree through a Bayesian framework taking as input character states of extant taxa and calculating posterior probabilities. Importantly, the uncertainty of the phylogeny is accounted for during ancestral state reconstruction by randomly selecting the posterior set of trees obtained through earlier Bayesian MCMC analysis of the data. In our case, we used 500 phylogenies that were randomly sampled from the converged Markov Chain obtained in Mr. Bayes.

3. Results

3.1 Phylogeny of P. syringae pv. tomato DC3000 and its relatives

An MLST analysis of relatives of the model plant pathogen *P. syringae* pv. *tomato* DC3000 (DC3000) was previously reported (Yan et al., 2008). A well supported phylogeny of strains belonging to pathovars *tomato (Pto)*, *maculicola (Pma)*, *antirrhini (Pan)*, and *apii* (*Pap*) was obtained. To further investigate evolution and mechanisms of

host specificity in this phylogenetic group, the phylogenetic analysis was extended to additional strains of above pathovars and to strains of pathovars *persicae (Ppe)*, *berberidis (Pbe)*, and *lachrymans (Pla)* that had also been reported to be closely related to pathovar *tomato* (Gardan et al., 1999; Sarkar and Guttman, 2004). When adding the new strains to the original set of strains (Table 2.1), the support for several basal branches in the trees built with the concatenated set of housekeeping genes used by (Yan et al., 2008) dropped. This was true for trees constructed using neighbor joining, maximum likelihood, and Bayesian inference (data not shown). Therefore, additional core genome genes, i.e. genes present exactly one time in all sequenced genomes of *P. syringae*, were concatenated with the original set of housekeeping genes (Table 2.2). These genes (PSPTOT1_0038, PSPTOT1_2358, and PSPTOT1_1665) encode predicted proteins of unknown function. Importantly, these additional three genes were all found to have a dN/dS ratio below 0.1 (0.019, 0.067, and 0.024, respectively) indicating that they are under purifying selection as were the genes previously used by Yan et al. (2008). These genes were thus suitable for phylogenetic analysis. However, even including the three additional genes, branch support for constructed trees was still low. Since genes that evolve at different rates may give different trees and consequently lower support for a tree based on the concatenated set of the same genes, we compared all loci for genetic distance between alleles (Table 2.2). Based on this analysis we excluded the *gapA* locus since it displayed the lowest genetic distance of all analyzed genes. Moreover, we had previously determined (Yan et al., 2008) that the three loci flanking the *avrPto1* genomic island, i.e., *kup*, PSPTO_3994 and

PSPTO_4019, were incongruent compared to all other analyzed loci (probably because of more frequent recombination events in this region due to selection acting on the *avrPto1* gene product that triggers or suppresses plant immunity depending on the genotype of the infected plant). Therefore, we decided to also exclude these three fragments. Using the remaining eight loci a bifurcating tree with high support for many branches was finally obtained. A Bayesian consensus tree is shown in Figure 2.1. Maximum Likelihood and Neighbor-joining trees with similar topology were also obtained (data not shown).

The tree in Figure 2.1 can be divided into a minimum of six clades with support values higher than 0.98: clade I contains five *Pap* strains isolated from celery and two *Pbe* strains [although the five *Pap* strains are identical in the genes used for tree construction they were analyzed as two distinct groups because they are different from each other in genes used previously by Yan et al. (2008)]; clade II contains fourteen strains isolated from various brassicaceae, Woolly Nightshade (*Solanum mauritianum*), and cucumber (*Cucumis sativus*) and belong to nine different genotypes (also called sequence types, STs, based on MLST terminology); clade III consists of 13 strains divided into two STs; four *Ppe* strains isolated from peach belong to two STs and form clade IV; six *Pan* strains isolated from snapdragon belong to two STs and form clade V; 26 *Pto* strains belonging to three STs and one *Pap* strain from celery make up clade VI; while two identical spinach mustard isolates (PmaF1 and PmaF7) do not cluster with any other strains.

3.2 Contribution of recombination and mutation to the evolution of DC3000 and its relatives

We previously reported (Yan et al., 2008) that the rate of recombination to mutation (ρ/θ) for DC3000 and its relatives was almost six based on the mean of the rates calculated for the individual gene fragments using the composite likelihood method of Hudson as implemented in LDhat version 2.1 (McVean et al., 2002). This result suggested that recombination played an important role in the evolution of DC3000 and its relatives. Using the same method, Sarkar and Guttman (2004) had reported that this rate for *P. syringae* was only 0.252 suggesting that *P. syringae* evolves mainly by vertical descent and should thus be considered a clonal species. We were interested to see if after adding more strains and using a slightly different gene set the ρ/θ ratio would increase or decrease for DC3000 and its relatives. As can be seen from Table 2.3, the ratio for individual genes and the calculated mean ratio (1.7) decreased compared to the ratio obtained by Yan et al. (2008) but still supports the conclusion that recombination played an important role in the evolution of this pathogen group (see also discussion below).

To further investigate the contribution of recombination to the evolution of DC3000 and its relatives, additional analyses were performed. The program GARD (Kosakovsky Pond et al., 2006)) was used to localize recombination break points. Applying this program to our data, recombination breakpoints were identified in half of the genes used for tree construction further confirming the importance of recombination

in the evolution of the DC3000 group (Table 2.3). The Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) was then used to determine if trees built on each individual gene fragment were compatible with the data from other gene fragments. Table 2.4 shows the P-values obtained with the SH test when comparing gene fragments with each other and with the concatenated data set. Almost all data from individual loci were found to be significantly incongruent with the trees built on the data from all other loci. Only the *acnB* data were found to be congruent with the *rpoD* and *pgi* trees and the PSPTOT1_2358 data were found to be congruent with the PSPTOT1_0038 tree. However, when comparing the data from the eight individual gene fragments with the Bayesian tree built from the concatenated data set they were found to be congruent. Therefore, while most individual loci are incongruent with the majority of the other loci (further suggesting that recombination occurred between the analyzed loci), the fact that the tree build on the concatenated gene set is congruent with the data from each individual locus gives us confidence that the tree shown in Figure 2.1 is an acceptable approximation of the evolutionary history of DC3000 and its relatives. However, to better account for the conflicting signals in the data, also a network was constructed using the ParsimonySplits method (Huson, 1998). The result is shown in Figure 2.2 Comparing this phylogenetic network in Figure 2.2 with the tree in Figure 2.1 suggests the presence of conflicting data possibly due to recombination during the evolution of clade II strains and clade V and clade VI strains. The long branches without any network structure for *Pbe*, *Pap*, and *Ppe* strains possibly suggest little recombination during the recent evolution of these strains. Alternatively, these strains

may have recombined during their recent evolution but with strains not included in our analysis (see more on sampling bias in the discussion below).

3.3 Host range analysis reveals different degree of host specialization

(Yan et al., 2008) had reported host range of DC3000 relatives mainly based on symptoms on three hosts: *A. thaliana*, cauliflower, and tomato. To obtain a more complete picture of host range, here host range tests were extended for one representative strain of each leaf of the tree shown in Figure 2.1 to the two phylogenetically diverse *A. thaliana* ecotypes Columbia [Col] and Martuba [Mt] (Innan et al., 1997), the two tomato cultivars 'Chico III' and 'Rio Grande', one cauliflower cultivar ('Precoce di Toscana'), one snapdragon cultivar ('First Ladies'), and one celery cultivar ('Verde Pascal'). To field conditions, tomato, cauliflower, and *A. thaliana* were inoculated by spraying bacteria on leaf surfaces. However, to obtain a homogeneous distribution of bacteria, a surfactant was added to bacterial suspension before inoculation. For celery and snapdragon, no disease was observed following the same method. Since Little et al. (1997) reported that outbreaks of *Pap* in California typically occur in greenhouses in which overhead irrigation causes damage to leaves, we decided to add carborundum to the inoculum for celery infections to facilitate bacterial invasion of leaves through carborundum-induced wounds. Since this allowed us to obtain disease symptoms and bacterial growth with most *Pap* strains on celery, we used this method with all strains on celery as well as on snapdragon, where it also allowed *Pan* strains to cause disease.

Representative pictures of disease symptoms for two strains/plant species are shown in Figure 2.3 while pictures of disease symptoms for all strain/plant species combinations are shown in Supplementary FigureS 2.1 through 2.5. Bacterial growth data are shown for each plant species in Figure 2.4. In most cases, there was a good correspondence between symptom severity and bacterial growth with the highest bacterial growth occurring with bacteria/plant combinations that gave the most severe symptoms and with the lowest bacterial growth occurring with bacteria/plant combinations that gave no disease symptoms at all. However, a small number of plants with mild disease symptoms sometimes showed high bacterial population densities and a small number of plants with severe disease symptoms sometimes had bacterial population densities only slightly higher than plants with no disease symptoms. In order to avoid arbitrarily choosing a cut off between "pathogenic" and "not pathogenic" assignments, we used logistic regression to calculate the probability for each isolate to either be pathogenic or non-pathogenic on each plant species based on symptoms and bacterial growth data (see materials and methods for details). Supplementary Table 2.1 shows our estimates of the explanatory variables (type of plants, strains and log-value of number of colony forming units (cfu). Probabilities of isolates to cause disease on each plant type are shown in Table 2.5. It is easy to see that combining symptoms and growth data in this way the probability of each strain for being pathogenic is either relatively low (0-0.40) or relatively high (0.60-1).

A summary of host range test results is reported in Figure 2.5 in which high

probability of pathogenicity is represented by a black square and low probability of pathogenicity is represented by a white square. Some important observations can be made: 1. host range test results for tomato and *A. thaliana* were only slightly different between the two tested cultivars/ecotypes (2 and 3 strains out of a total of 25 strains gave different results on the two *A. thaliana* ecotypes and two tomato cultivars respectively) suggesting that even using only one cultivar/plant species (as was the case for cauliflower, celery, and snapdragon) gave a good approximation of host range; 2. in many cases strains with similar host range cluster together but sometimes strains with very different host ranges are located next to each other in the same clade (for example, strains in clade II); 3. some strains have relatively wide host ranges, some strains have intermediate host ranges, while yet other strains can only cause disease on the plant species from which they were originally isolated (in one case, PapCFBP2103, a strain was not even pathogenic on the tested cultivar of the plant species of isolation), 4. While some plant species (tomato, cauliflower, and celery) are susceptible to many tested strains (even strains originally isolated from different plant species) other plant species (*A. thaliana* and even more so snapdragon) are susceptible only to a small subset of tested strains.

3.4 Ancestral state reconstruction of host range

We wanted to test the hypothesis that the extant highly virulent crop pathogens of narrow host range evolved from wider host range ancestors well adapted to mixed plant communities of pre-agricultural times. To do this, it was first necessary to determine if

there was any correlation between host range and phylogeny. Only if that is the case can it be attempted to infer host range of ancestors based on the host range of the extant strains. The program BaTS (Parker et al., 2008) can test the null hypothesis that a character state is randomly distributed among extant taxa on a phylogenetic tree. Results from BaTS are shown in Supplementary Table 2.2 and indicate non-random distribution of pathogenicity among strains for the two *A. thaliana* ecotypes. We thus focused on *A. thaliana* for ancestral state reconstruction.

The program BayesTraits was used for ancestral state reconstruction. BayesTraits calculates probabilities of character states at every node of a tree taking also into account the uncertainty of the tree itself using a Bayesian approach (see materials and methods for details). Figure 2.6 shows the results for *A. thaliana* 'Col' while Supplementary FigureS 2.6 shows the result for *A. thaliana* 'Mt' (results for ancestral state reconstruction for cauliflower, celery, snapdragon, and the tomato cultivar 'Chico III' are shown in the Supplementary FigureS 2.7 through 2.10). At each node the inferred pathogenicity is indicated as a pie chart. It is evident that pathogenicity or non-pathogenicity was assigned with high confidence to most nodes corresponding to ancestors of individual clades. However, approaching the basal node of the tree, i.e., the most recent common ancestor of all considered strains, probabilities of pathogenicity and non-pathogenicity in regard to all plant species approach 0.5. Therefore, using the strains considered here and the chosen approach it appears

impossible to come to any conclusion in regard to the host range of the most recent common ancestor of all analyzed strains (possible explanations are discussed below).

4. Discussion

4.1 Recombination and Phylogenetic reconstruction of strain DC3000 and its relatives

One of the main goals of MLST studies is to infer to what extent recombination contributed to the sequence diversity of the core genome existing within a group of bacteria. Comparing results obtained by Sarkar and Guttman (2004) and Yan et al. (2008) with the ones presented here highlights how changes in the set of analyzed strains and sequenced gene fragments strongly affect the outcome of such an analysis. Important questions that we cannot answer here are: did we omit from our analysis strains closely related to the analyzed strains that colonize wild plants or crops but were not sampled? Did the ancestors of these hypothetical strains recombine with the ancestors of the analyzed strains? These questions go to the heart of the problem of every population genetics analysis of human and plant pathogens when only a relatively small selection of strains from a selected group of hosts is being considered, i.e. sampling bias. In our study we had only access to pathogens isolated from a few crop species on which close relatives of DC3000 were reported to causes disease. However, the majority of DC3000 relatives may cause disease on wild plants. Therefore, we may have analyzed a highly biased sample of the population we are trying to characterize.

Therefore, we believe that it will only be possible to truly reconstruct the evolutionary history of a bacterial plant pathogen and assess the contribution of recombination to strain diversity if strains of that pathogen are extensively sampled from crops, wild plants, and the environment. Recent sampling of *P. syringae* from wild plants and the environment (Morris et al., 2010) will be an excellent opportunity to more realistically determine the role of recombination in the *P. syringae* species complex using a more representative sample of the pathogen population.

4.2 Host range differences between DC3000 relatives

The only possibility to compare host range of strains in controlled conditions is to infect plants under the same conditions of light, temperature, and humidity. However, different strains may have different optimal conditions under which they cause disease and environmental conditions under which a pathogen causes disease in the field may be difficult to simulate in the lab, for example, it is almost impossible to induce frost damage onto plants in the lab although frost damage greatly facilitates *P. syringae* infections in the field (Hirano and Upper, 2000). Therefore, it is difficult to confidently predict host range of plant pathogens based on results from infections in controlled conditions. However, aware of this limitation, we tried to use conditions of light, humidity, and temperature known to be most favorable for disease development by *P. syringae* in general. At the same time, we used the lowest inoculum dose that allowed strains to cause disease symptoms on the plant species from which they were originally isolated. Our basic assumption is that under the employed conditions and inoculum doses a

pathogen that does not cause disease on a plant species in the field will not be able to cause disease in the lab.

Another limitation of host range tests is that there are differences between genotypes of the same plant species in regard to resistance to individual strains because of different resistance gene alleles present in different plant genotypes of the same species. Therefore, we used two *A. thaliana* ecotypes and two tomato cultivars for host range determination. Interestingly, only few differences were found between ecotypes and cultivars: only two strains had different pathogenicity on the two *A. thaliana* ecotypes and only three strains had different pathogenicity on the two tomato cultivars. Therefore, we are confident that the results obtained with only one cultivar of cauliflower, snapdragon, and celery are an acceptable approximation of host range of the tested isolates on these species.

An expected result obtained from the host range tests was that in most cases strains with identical or similar host range cluster together. However, other times strains with very different host range are within the same clade (clade II is the most striking example). Moreover, in clade II closely related strains do not only have different host ranges but they represent extremes of wide and narrow host range. Based on the current knowledge of host range mechanisms that exist in *P. syringae* (Lindeberg et al., 2009), a likely explanation for these results is that relatively wide host range strains have a repertoire of type III-secreted effectors that does not include any effectors that trigger immunity on most of the analyzed plant species. These strains are probably

under selection to avoid acquisition of such effectors since a wide host range is advantageous in the ecological niche they adapted to. Possibly, these strains have a life style similar to the hypothetical life style of pathogens in pre-agricultural times when mixed-plant communities prevailed and having a wide host range allows these strains to find susceptible hosts relatively easily. On the other hand, strains with narrow host range may have one or more effectors that significantly increase virulence on one host but that are recognized by resistance genes in other plant species. In clade II, for example, PtoICMP3443 and PtoICMP3449 would represent wide host range strains that do not have effectors that trigger immunity in either *A. thaliana*, cauliflower, tomato, or celery. PmaM3 may have acquired one or more effectors contributing to virulence on cauliflower but triggering immunity in some *A. thaliana*, tomato, and celery genotypes. PmaF15 and Pla3988 may have acquired additional effectors favorable to growth on some Brassicaceae or cucumber but that trigger immunity on other plant species further restricting host range.

In other cases, narrow host range strains are located by themselves on a separate long branch. Possible explanations are: 1. these strains adapted to a single host a relatively long time ago, 2. adaptation to a single host was caused by recombination with another pathogen of that host increasing DNA sequence distances from other analyzed strains rapidly, 3. after adaptation to the new host the number of generations per year increased because the new host is more available for infections

than the previous hosts, or 4. closely related strains with wider host range exist but were not included in the analysis.

In one case, two distantly related strains were found to have the same host range: PtoJL1065 and Pap1089-5. Interestingly, Kunkeaw et al. (2010) found a strain with the same ST as Pap1089-5 on tomato in California. This finding suggests that *Pto* strains and *Pap* strains do not only have similar host range in the lab but also in the field. It will be very interesting to compare the effector repertoires of *Pap* and *Pto* strains to determine if similar effector repertoires are at the base of their similar host range.

4.3 Ancestral state reconstruction of host range

Considering that 1. undisturbed plant communities are often mixed communities consisting of different plant species, 2. the first agricultural fields of a single plant species appeared only about ten thousand years ago (Balter, 2010), 3. high density populations of genetically homogeneous crop species became available to pathogens only recently as a consequence of modern agriculture, the hypothesis that selection pressure on pathogens changed dramatically over the last few thousand years is an obvious one (Stukenbrock and McDonald, 2008). From selection for a wide host range that allowed pathogens to easily find hosts in mixed plant communities selection pressure probably shifted to selection for narrow host range accompanied by high virulence that allows pathogens to thrive in agricultural field that provide an almost unlimited supply of susceptible hosts.

Since we found closely related *P. syringae* strains with different degrees of host specialization we hoped to be able to find support for the above hypothesis. However, only in the case of *A. thaliana* was the ability of different strains to cause disease associated with phylogeny. For the other plant species random distribution of pathogenicity among strains could not be rejected. Possibly, host range changes in regard to these plant species occurred independently of the phylogeny based on housekeeping genes because of frequent horizontal gene transfer events of virulence genes. Alternatively, the number of strains included in our analysis could have been too low or to biased to be representative of the host range distribution existing in the pathogen population. Moreover, reconstructing pathogenicity in regard to *A. thaliana* (and in regard to the other plant species), it was impossible to make any interference for the most recent common ancestor (MRCA) of all analyzed strains. The MRCA had in fact approximately the same probability of being or not being a pathogen of *A. thaliana* (and of the other plants species that were considered). One problem with the ancestral state reconstruction of the MRCA could have been the low branch support at the base of the tree. However, even choosing a hypothetical support of 1 for all branches results did not improve (data not shown). Also assigning either pathogenicity or non-pathogenicity to our outgroup strain *Psy*B728a (or not assigning either) did not change results. Therefore, the quality of the tree does not seem to be a limitation. The problem instead may have been again the limited number of the strains that were considered and the bias in strain collection. One can assume that only a very small number of strains in the *P. syringae* species adapted to crops while the majority of strains are still

adapted to wild plants. These strains can be expected to be of relatively wide host range. Therefore, we believe that only including a representative sample of strains from wild plants will it be possible in the future to obtain meaningful results from ancestral state reconstruction of host range.

4.4 Molecular basis of host range evolution and mechanisms

Although we did not find support for host range evolution from wide to narrow, this study has provided a very useful resource to investigate the molecular basis of host range evolution and determination. Genome sequencing of the closely related strains with different host ranges characterized here will enable identification of gene differences between strains beyond what was accomplished using microarrays (Sarkar et al., 2006) and will allow the identification of all allelic differences in genes encoding effectors and PAMPS. Genome sequencing will help uncover the mechanisms, for example, acquisition and loss of genomic islands, that were involved in the evolutionary events that gave rise to these differences. Finally, it will be possible to experimentally test the genes and alleles that distinguish strains for their role in host range determination and their interaction with plant targets. Comparison of genomes of some of the analyzed strains is already under way (Almeida et al., 2009; Cai et al., in press) and has already provided new insight into plant pathogen evolution and pathogen – plant interactions.

Acknowledgements

Research in the Vinatzer lab was funded by the National Science Foundation (Award IOS 0746501)

References

Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol 62, 53-70.

Achtman, M., Morelli, G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., Vogler, A.J., Wagner, D.M., Allender, C.J., Easterday, W.R., Chenal-Francisque, V., Worsham, P., Thomson, N.R., Parkhill, J., Lindler, L.E., Carniel, E., Keim, P., 2004. Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci U S A 101, 17837-17842.

Almeida, N.F., Yan, S., Cai, R., Clarke, C.R., Morris, C.E., Schaad, N.W., Schuenzel, E.L., Lacy, G.H., Sun, X., Jones, J.B., Castillo, J.A., Bull, C.T., Leman, S., Guttman, D.S., Setubal, J.C., Vinatzer, B.A., 2010. PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. Phytopathology 100, 208-215.

Almeida, N.F., Yan, S., Lindeberg, M., Studholme, D.J., Schneider, D.J., Condon, B., Liu, H., Viana, C.J., Warren, A., Evans, C., Kemen, E., Maclean, D., Angot, A., Martin, G.B., Jones, J.D., Collmer, A., Setubal, J.C., Vinatzer, B.A., 2009. A Draft Genome Sequence of *Pseudomonas syringae* pv. *tomato* T1 Reveals a Type III Effector Repertoire

Significantly Divergent from That of Pseudomonas syringae pv. tomato DC3000. Mol Plant Microbe Interact 22, 52-62.

Auton, A., McVean, G., 2007. Recombination rate estimation in the presence of hotspots. Genome Research 17, 1219-1227.

Balter, M., 2010. The Tangled Roots of Agriculture. Science 327, 404-406.

Bull, C.T., Clarke, C.R., Cai, R., Vinatzer, B.A., Jardini, T.M., Koike, S.T., 2011. Multilocus Sequence Typing of *Pseudomonas syringae sensu lato* confirms previously described genomosoecies and permits rapid idenitfication of *P. syringae* pv. *coriandricola* and *P. syringae* pv. *apii* causing bacterial leaf spot on parsley. Phytopathology.

Cai, R., Lewis, J., Yan, S., Liu, H., Clarke, C.R., Campanile, F., Almeida, N.F., Studholme, D.J., Lindeberg, M., Schneider, D.J., Zaccardelli, M., Setubal, J.C., Morales-Lizcano, N.P., Bernal, A., Coaker, G., Baker, C., Bender, C.L., Leman, S., Vinatzer, B.A., in press. The plant pathogen *Pseudomonas syringae* pv. *tomato* is genetically monomorphic and under strong selection to evade tomato immunity. PLoS Pathog.

Chisholm, S.T., Coaker, G., Day, B., Staskawicz, B.J., 2006. Host-microbe interactions: shaping the evolution of the plant immune response. Cell 124, 803-814.

Cuppels, D.A., 1986. Generation and characterization of Tn5 insertion mutations in Pseudomonas syringae pv. tomato. Appl. Environ. Microbiol. 51, 323-327.

Cuppels, D.A., Ainsworth, T., 1995. Molecular and Physiological Characterization of *Pseudomonas syringae* pv. *tomato* and *Pseudomonas syringae* pv. *maculicola* Strains That Produce the Phytotoxin Coronatine. Appl. Environ. Microbiol. 61, 3530-3536.

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

Denny, T.P., 1988a. Differentiation of *Pseudomonas syringae* pv. *tomato* from *P. syringae* with a DNA hybridization probe. Phytopathology 78, 1186-1193.

Denny, T.P., 1988b. Genetic diversity and relationships of two pathovars of *Pseudomonas syringae*. Journal of General Microbiology 134, 1949-1960.

Dobrindt, U., Hochhut, B., Hentschel, U., Hacker, J., 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2, 414-424.

Dye, D.W., Bradbury, J.F., Goto, M., Hayward, A.C., Lelliott, R.A., Schroth, M.N., 1980. International standards for naming pathovars of phytopathogenic bacteria and a list of pathovar names and pathotype strains. Rev Plant Pathol 142, 153-158.

Gardan, L., Shafik, H., Belouin, S., Broch, R., Grimont, F., Grimont, P.A., 1999. DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (ex Sutic and Dowson 1959). Int J Syst Bacteriol 49 Pt 2, 469-478.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., Tschape, H., 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23, 1089-1097.

Hirano, S.S., 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.

Huson, D.H., 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68-73.

Hwang, M.S., Morgan, R.L., Sarkar, S.F., Wang, P.W., Guttman, D.S., 2005. Phylogenetic characterization of virulence and resistance phenotypes of *Pseudomonas syringae*. Appl Environ Microbiol 71, 5182-5191.

Innan, H., Terauchi, R., Miyashita, N.T., 1997. Microsatellite Polymorphism in Natural Populations of the Wild Plant *Arabidopsis thaliana*. Genetics 146, 1441-1452.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23, 1891-1901.

Kunkeaw, S., Tan, S., Coaker, G., 2010. Molecular and evolutionary analyses of *Pseudomonas syringae* pv. *tomato* race 1. Mol Plant Microbe Interact 23, 415-424.

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

Lindeberg, M., Cunnac, S., Collmer, A., 2009. The evolution of *Pseudomonas syringae* host specificity and type III effector repertoires. Mol Plant Pathol 10, 767-775.

Little, E.L., Koike, S.T., Gilbertson, R.L., 1997. Bacterial Leaf Spot of Celery in California: Etiology, Epidemiology, and Role of Contaminated Seed. Plant Disease 81, 892-896.

Lovell, H.C., Mansfield, J.W., Godfrey, S.A.C., Jackson, R.W., Hancock, J.T., Arnold, D.L., 2009. Bacterial Evolution by Genomic Island Transfer Occurs via DNA Transformation In Planta. Current Biology 19, 1586-1590.

Maiden, M.C., 2006. Multilocus Sequence Typing of Bacteria. Annu Rev Microbiol. 72, 7098-110.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95, 3140-3145.

McVean, G., Awadalla, P., Fearnhead, P., 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160, 1231-1241.

Morris, C.E., Sands, D.C., Vanneste, J.L., Montarry, J., Oakley, B., Guilbaud, C., Glaux, C., 2010. Inferring the Evolutionary History of the Plant Pathogen *Pseudomonas syringae* from Its Biogeography in Headwaters of Rivers in North America, Europe, and New Zealand. mBio 1.

Musser, J.M., Hewlett, E.L., Peppler, M.S., Selander, R.K., 1986. Genetic diversity and relationships in populations of *Bordetella* spp. J. Bacteriol. 166, 230-237.

Pagel, M., Meade, A., Barker, D., 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53, 673-684.

Parker, J., Rambaut, A., Pybus, O.G., 2008. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infect Genet Evol 8, 239-246.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817-818.

Ronald, P.C., Salmeron, J.M., Carland, F.M., 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.

Ronquist, F., 2004. Bayesian inference of character evolution. Trends Ecol Evol 19, 475-481.

Roumagnac, P., Weill, F.o.-X., Dolecek, C., Baker, S., Brisse, S., Chinh, N.T., Le, T.A.H., Acosta, C.J., Farrar, J., Dougan, G., Achtman, M., 2006. Evolutionary History of *Salmonella* Typhi. Science 314, 1301-1304.

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

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

Shenge, K.C., Mabagala, R.B., Mortensen, C.N., Stephan, D., Wydra, K., 2007. First Report of Bacterial Speck of Tomato Caused by *Pseudomonas syringae* pv. *tomato* in Tanzania. Plant Disease Note 91, 462.

Shimodaira, H., Hasegawa, M., 1999. log-likelihoods with application to phylogenetic inference. Mol Biol Evol 16, 1114-1116.

Stukenbrock, E.H., McDonald, B.A., 2008. The Origins of Plant Pathogens in Agro-Ecosystems. Annual Review of Phytopathology 46, 75-100.

Sun, W., Dunning, F.M., Pfund, C., Weingarten, R., Bent, A.F., 2006. Within-species flagellin polymorphism in *Xanthomonas campestris* pv *campestris* and its impact on elicitation of Arabidopsis FLAGELLIN SENSING2-dependent defenses. Plant Cell 18, 764-779.

Takeuchi, K., Taguchi, F., Inagaki, Y., Toyoda, K., Shiraishi, T., Ichinose, Y., 2003. Flagellin glycosylation island in *Pseudomonas syringae* pv. *glycinea* and its role in host specificity. J Bacteriol 185, 6658-6665.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24, 1596-1599.

Whalen, M.C., Innes, R.W., Bent, A.F., 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.

Yan, S., Liu, H., Mohr, T.J., Jenrette, J., Chiodini, R., Zaccardelli, M., Setubal, J.C., 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.

Zaccardelli, M., Spasiano, A., Bazzi, C., M., M., 2005. Identification and in planta detection of *Pseudomonas syringae* pv. *tomato* using PCR amplification of *hrpZ*. European Journal of Plant Pathology 111, 85-90.

Zhao, Y., Damicone, J.P., Demezas, D.H., Rangaswamy, V., Bender, C.L., 2000. Bacterial leaf spot of leafy crucifers in Oklahoma caused by *Pseudomonas syringae* pv. *maculicola*. *Plant Disease* 84, 1015-1020.

Tables

Table 2.1 *P. syringae* strains used in this study

pv.[a]	strain name	Host of isolation (common name)	Host of isolation (scientific name)	Location	Collected by	Year	Reference
Pto[PT]	DC3000	Tomato	*S. lycopersicum*	UK	R. A. Lelliott	1961	(Cuppels and Ainsworth, 1995)
Pto	Max 1	Tomato	*S. lycopersicum*	Italy	M. Zaccardelli	2002	(Zaccardelli et al., 2005)
Pto	Max 3	Tomato	*S. lycopersicum*	Italy	M. Zaccardelli	2003	(Zaccardelli et al., 2005)
Pto	Max 4	Tomato	*S. lycopersicum*	Italy	M. Zaccardelli	2002	(Zaccardelli et al., 2005)
Pto	Max 5	Tomato	*S. lycopersicum*	Italy	M. Zaccardelli	2002	(Zaccardelli et al., 2005)
Pto	Max 6	Tomato	*S. lycopersicum*	Italy	M. Zaccardelli	2002	(Zaccardelli et al., 2005)
Pto	Max 7	Tomato	*S. lycopersicum*	Italy	-	1996	(Zaccardelli et al., 2005)
Pto	Max 10	Tomato	*S. lycopersicum*	France	-	1998	(Zaccardelli et al., 2005)
Pto	Max 12	Tomato	*S. lycopersicum*	Italy	-	1991	(Zaccardelli et al., 2005)
Pto	LNPV17.41	Tomato	*S. lycopersicum*	France	-	1996	(Zaccardelli et al., 2005)
Pto	Max14	Tomato	*S. lycopersicum*	Spain	-		(Zaccardelli et al., 2005)
Pto	JL1065	Tomato	*S. lycopersicum*	CA, USA	J. Lindemann	1983	(Whalen et al., 1991)
Pto	kuzzen40	Tomato	*S. lycopersicum*	VA, USA	C. Waldenmaier	2005	(Yan et al., 2008)
Pto	PT13	Tomato	*S. lycopersicum*		Gitaitis		(Yan et al., 2008)
Pto	PT14	Tomato	*S. lycopersicum*		G. Bonn		(Yan et al., 2008)
Pto	PT21	Tomato	*S. lycopersicum*	FL, USA	T. Howe	1990	(Yan et al., 2008)
Pto	PT23	Tomato	*S. lycopersicum*		M. Ricker	1990	(Yan et al., 2008)
Pto	PT25	Tomato	*S. lycopersicum*		M. Ricker	1990	(Yan et al., 2008)
Pto	PT26	Tomato	*S. lycopersicum*		M. Ricker	1990	(Yan et al., 2008)
Pto	PT28	Tomato	*S. lycopersicum*	Mexico	J. Jones	1992	(Yan et al., 2008)
Pto	PT29	Tomato	*S. lycopersicum*	Mexico	J. Jones	1992	(Yan et al., 2008)
Pto	PT30	Tomato	*S. lycopersicum*	Mexico	J. Jones	1992	(Yan et al., 2008)
Pto	PT32	Tomato	*S. lycopersicum*	FL, USA	J. Jones	1993	(Yan et al., 2008)
Pma	M1	Cauliflower	*B. oleracea* var. *botrytis*	UK	R. Lelliott	1965	(Debener et al., 1991)
Pma[PT]	CFBP1657	Cauliflower	*B. oleracea* var. *botrytis*	New Zealand	D. Shackleton	1965	
Pma	M3	Cauliflower	*B. oleracea* var. *botrytis*	USA	W. Burkholder	1937	(Debener et al., 1991)
Pma	M6	Cauliflower	*B. oleracea* var. *botrytis*	UK	G. Jones	1965	(Debener et al., 1991)
Pma	M8	Kale	*B. oleracea* var. *acephala*	UK	J. Taylor		(Debener et al., 1991)
Pto	PST6	Tomato	*S. lycopersicum*	Canada	D. Cuppels	1980	(Denny, 1988a)
Pto	PT17	Tomato	*S. lycopersicum*	NJ, USA	D. Coplin		(Denny, 1988b)
Pto	PT2	Tomato	*S. lycopersicum*	GA, SA	S. McCarter		(Denny, 1988b)
Pto	PT18	Tomato	*S. lycopersicum*	CA, USA	C. Kado		(Denny, 1988b)
Pto	T1	Tomato	*S. lycopersicum*	Canada	G. Bonn	1986	(Ronald et al., 1992)
Pto	JL1031	Tomato	*S. lycopersicum*	CA, USA	J. Lindeman	1983	(Denny, 1988b)
Pto	B181	Tomato	*S. lycopersicum*	GA, USA	S. McCarter	1981	(Denny, 1988b)
Pma	mac1		*B. oleracea*	CA, USA	N. Keen		(Denny, 1988a)
Pto	PDDCC2844	Tomato	*S. lycopersicum*		R. Lelliott	1960	(Cuppels, 1986)
Pan	126	Snapdragon	*A. majus*		M. Moffett	1965	(Yan et al., 2008)
Pto	OH314	Horse nettle	*Solanum carolinense*	OH, USA	D. Coplin	1978	(Cuppels and Ainsworth, 1995)

Pto	NCPPB1108	Tomato	*S. lycopersicum*	UK	R. Lelliott	1960	(Yan et al., 2008)
Pto	ICMP3435	Woolly nightshade	*Solanum mauritianum*	New Zealand	D.R.W. Watson	1972	(Whalen et al., 1991)
Pto	ICMP3443	Woolly nightshade	*S. mauritianum*	New Zealand	D.R.W. Watson	1972	(Yan et al., 2008)
Pto	ICMP3449	Woolly nightshade	*S. mauritianum*	New Zealand	D.R.W. Watson	1972	(Yan et al., 2008)
Pto	ICMP3455	Woolly nightshade	*S. mauritianum*	New Zealand	D.R.W. Watson	1972	(Whalen et al., 1991)
Pto	ICMP9305	Woolly nightshade	*S. mauritianum*	New Zealand	D.R.W. Watson	1987	(Yan et al., 2008)
Pap	1089-5	Celery	*A. graveolens*	CA, USA	D. Cooksey	1989	(Bull et al., 2011)
Pan	ICMP4303	Snapdragon	*A. majus*	UK	G. Jones	1965	(Yan et al., 2008)
Pan	152E	Snapdragon	*A. majus*	UK	J. Taylor	1960	(Yan et al., 2008)
Pma	F1	Spinach mustard	*B. rapa* var. *perviridis*	OK, USA		1995	(Zhao et al., 2000)
Pma	F6	Kale	*B. oleracea* var. *acephala*	OK, USA		1995	(Zhao et al., 2000)
Pma	F7	Spinach mustard	*B. rapa* var. *perviridis*	OK, USA		1995	(Zhao et al., 2000)
Pma	F9	Spinach mustard	*B. rapa* var. *perviridis*	OK, USA		1995	(Zhao et al., 2000)
Pma	F10A	Turnip	*B. rapa* var. *rapifera*	OK, USA		1995	(Zhao et al., 2000)
Pma	F15	Turnip	*B. rapa* var. *rapifera*	OK, USA		1995	(Zhao et al., 2000)
Pma	F16	Turnip	*B. rapa* var. *rapifera*	OK, USA		1995	(Zhao et al., 2000)
Pma	F17	Spinach mustard	*B. rapa* var. *perviridis*	OK, USA		1995	(Zhao et al., 2000)
Pma	F18	Kale	*B. oleracea* var. *acephala*	OK, USA		1995	(Zhao et al., 2000)
Pma	F19	Turnip	*B. rapa* var. *rapifera*	OK, USA		1996	(Zhao et al., 2000)
Pap	BS252	Celery	*A. graveolens*	CA, USA	C. Bull	2001	
Pbe	ATCC13454	Barberry	*Berberidis sp.*	USA	W.H. Burkholder	1942	
Pbe^{PT}	CFBP1727	Barberry	*Berberidis sp.*	New Zealand	J.M. Young	1972	
Pma	ICMP4981	Cauliflower	*B. oleracea* var Botrytis	Zimbabwe		1970	
Pma	ICMP795	Cauliflower	*B. oleracea* var Botrytis	New Zealand	W.J. Kemp	1958	
Pma	ICMP2744	Mustard	*B. oleracea* var Botrytis	UK	J. D. Taylor	1968	
Pla^{PT}	ICMP3988	Cucumber	*Cucumis sativus*	USA		1935	(Hwang et al., 2005)
Ppe^{PT}	ICMP5846	Peach	*Prunus persicae*	France	J. P. Prunier	1974	
Ppe	CFBP5143	Peach	*P. persicae*	New Zealand	J.M. Young	1980	
Pap^{PT}	CFBP2103	Celery	*A. graveolens*	USA	W.H. Burkholder	1942	
Pma	NCPPB1766		*B. oleracea* var Botrytis	UK	G.E. Jones		(Cuppels and Ainsworth, 1995)
Pma	DC84-59		*B. oleracea*	Canada	D. Cuppels	1984	(Cuppels and Ainsworth, 1995)
Pto	DC84-1	Tomato	*S. lycopersicum*	Canada	D. Cuppels	1984	
Pto	CFBP1318	Tomato	*S. lycopersicum*	Switzerland	T. Burki	1969	(Cuppels and Ainsworth, 1995)
Pto	PST26L	Tomato	*S. lycopersicum*	South Africa		1986	(Cuppels and Ainsworth, 1995)
Pto	487	Tomato	*S. lycopersicum*	Greece		1979	(Cuppels and Ainsworth, 1995)
Pto	KS 127 M	Tomato	*S. lycopersicum*	Tanzania		2004	(Shenge et al., 2007)
Pto	KS 112 1r	Tomato	*S. lycopersicum*	Tanzania		2004	(Shenge et al., 2007)
Pap	BS329	Celery	*A. graveolens*	CA, USA		2002	(Bull et al., 2011)

Pap	BS546	Celery	*A. graveolens*	CA, USA	C. Bull	2003
Pap	CFBP1726	Celery	*A. graveolens*	USA	C. Wehlburg	1975
Ppe	CFBP1067	Peach	*P. persicae*	France	A. Vigouroux	1967
Ppe	CFBP3970	Peach	*P. persicae*	France	J. Luisetti	1994
Pan^{PT}	CFBP1620	Snapdragon	*A. majus*	UK	G. E. Jones	1965
Pan	CFBP1723	Snapdragon	*A. majus*	Australia	M. L. Moffett	1964
Pan	CFBP3715	Snapdragon	*A. majus*	UK	R. J. Roberts	1975
Pto	kuzzen100	Tomato	*S. lycopersicum*	VA, USA	C. Waldenmaier	2005

[a] pv. : pathovar, *Pto*: *P. syringae* pv. *tomato*, *Pma*: *P. syringae* pv. *maculicola, Pan*: *P. syringae* pv. *antirrhini, Ppe*: *P. syringae* pv. *persicae, Pap*: *P. syringae* pv. *apii, Pla*: *P. syringae* pv. *lachrymans, Pbe*: *P. syringae* pv. *berberidis*, PT: pathotype

Table 2.2 Length, number of polymorphisms, genetic distance, Tajima's D, and ratio of non-synonymous (d_N) to synonymous (d_S) mutations for all analyzed gene fragments.

Gene	length	Number of polymorphism	Jukes-Cantor Genetic distance	Tajima's D	d_N / d_S
gyrB	696	16	0.006(0.002)	0.074	0.00010
rpoD	636	21	0.012(0.003)	1.432	0.00991
pgi	567	13	0.005(0.002)	-0.447	0.02245
gltA	504	6	0.004(0.002)	0.536	0.00010
acnB	555	17	0.008(0.002)	-0.190	0.00010
PSPTOT1_0038	597	14	0.008(0.002)	0.935	0.01933
PSPTOT1_2358	498	17	0.010(0.003)	0.353	0.06697
PSPTOT1_1665	435	12	0.006(0.002)	-0.665	0.02400
gapA	600	7	0.002(0.001)	-1.102	0.07713
Mean (excluding gapA)	561	15	0.007	0.254	0.01787

Table 2.3 Length, estimates of population recombination rate (ρ) and population mutation rate (θ) and recombination breakpoints for all analyzed gene fragments.

Gene	length	ρ	per site ρ	θ	per site θ	ρ / θ	recombination breakpoints[a]
gyrB	696	6	0.00862	4.285	0.00616	1.400	Yes(1)
rpoD	636	26	0.04088	5.624	0.00884	4.623	Yes(1)
Pgi	567	0	0.00000	3.481	0.00614	0.000	No
gltA	504	6	0.01190	1.607	0.00319	3.734	Yes(1)
acnB	555	4	0.00721	4.552	0.00820	0.879	No
PSPTOT1_0038	597	6	0.01005	3.749	0.00628	1.600	Yes(1)
PSPTOT1_2358	498	5	0.01004	4.552	0.00914	1.098	No
PSPTOT1_1665	435	1	0.00230	3.213	0.00739	0.311	No
Mean		6.75	0.01138	3.883	0.00692	1.706	

[a] as determined with GARD (Kosakovsky Pond et al., 2006)

Table 2.4 Summary of Shimodaira-Hasegawa (SH) test results

	loci								
Mr.Bayes Tree	gyrB	rpoD	pgi	gltA	acnB	PSPTOT1 0038	PSPTOT1 2358	PSPTOT1 1665	Concatenated
gyrB		0.000	0.019	0.044	0.002	0.001	0.007	0.039	0.000
rpoD	0.000		0.032	0.015	**0.076**	0.001	0.007	0.048	0.000
pgi	0.000	0.000		0.020	**0.202**	0.000	0.004	0.007	0.000
gltA	0.001	0.000	0.031		0.009	0.000	0.005	0.013	0.000
acnB	0.000	0.000	0.040	0.029		0.000	0.006	0.044	0.000
PSPTOT1_0038	0.000	0.000	0.005	0.013	0.002		**0.068**	0.018	0.000
PSPTOT1_2358	0.001	0.000	0.006	0.018	0.003	0.000		0.022	0.000
PSPTOT1_1665	0.003	0.000	0.012	0.018	0.007	0.002	0.006		0.000
Concatenated	**0.145**	**0.156**	**0.356**	**0.334**	**0.730**	**0.556**	**0.728**	**0.535**	

P values are reported for each tree/data and data/tree comparison, P values >0.05 are in bold face.

Table 2.5 Probability with confidence intervals of each *P. syringae* isolate to cause disease on each plant species.

	Arabidopsis Col	Arabidopsis Mt	Tomato Chico III	Tomato Rio Grande	Cauliflower	Celery	Snapdragon
PsyB728a	0.06 (0.02,0.17)	0.09 (0.03,0.25)	0.08 (0.03,0.23)	0.1 (0.03,0.29)	0.2 (0.06,0.49)	0.15 (0.05,0.4)	0.29 (0.11,0.59)
Pap1089	0.12 (0.04,0.31)	0.1 (0.03,0.28)	**0.87** (0.62,0.97)	**0.62** (0.3,0.86)	**0.73** (0.41,0.92)	**0.76** (0.43,0.93)	0.08 (0.03,0.18)
PapBS252	0.16 (0.05,0.41)	0.13 (0.04,0.32)	**0.75** (0.43,0.92)	**0.73** (0.41,0.91)	**0.61** (0.28,0.86)	**0.88** (0.62,0.97)	0.01 (0.01,0.04)
PbeATCC13454	0.05 (0.02,0.15)	0.07 (0.02,0.21)	0.06 (0.02,0.17)	0.35 (0.13,0.66)	0.11 (0.04,0.30)	0.4 (0.15,0.74)	**0.63** (0.32,0.86)
PbeCFBP1727	0.09 (0.03,0.22)	0.09 (0.03,0.25)	0.21 (0.08,0.46)	0.01 (0,0.04)	0.01 (0.00,0.03)	0.38 (0.13,0.70)	0.02 (0.01,0.05)
Pma4981	**0.83** (0.55,0.95)	**0.86** (0.6,0.96)	**0.9** (0.7,0.97)	**0.92** (0.73,0.98)	**0.95** (0.82,0.99)	**0.87** (0.61,0.96)	0.1 (0.04,0.24)
PtoICMP3443	**0.87** (0.63,0.96)	**0.82** (0.54,0.95)	**0.85** (0.6,0.96)	**0.85** (0.59,0.95)	**0.94** (0.8,0.99)	**0.76** (0.46,0.92)	**0.7** (0.42,0.88)
PtoICMP3449	**0.85** (0.59,0.96)	**0.73** (0.41,0.91)	**0.93** (0.78,0.98)	**0.94** (0.78,0.98)	**0.87** (0.62,0.96)	**0.81** (0.51,0.94)	0.39 (0.17,0.65)
Pma795	0.13 (0.04,0.34)	0.31 (0.11,0.61)	0.4 (0.18,0.72)	**0.6** (0.26,0.82)	0.34 (0.13,0.64)	0.4 (0.18,0.75)	0.02 (0.01,0.06)
PmaM6	0.25 (0.09,0.53)	**0.6** (0.27,0.83)	0.4 (0.19,0.73)	0.13 (0.05,0.3)	**0.78** (0.48,0.93)	0.4 (0.21,0.78)	0.03 (0.01,0.07)
Pma2744	0.11 (0.04,0.30)	0.16 (0.05,0.38)	**0.84** (0.56,0.95)	**0.6** (0.24,0.81)	0.4 (0.18,0.75)	0.36 (0.13,0.67)	0.12 (0.05,0.28)
PmaF15	0.09 (0.03,0.25)	0.08 (0.02,0.22)	0.27 (0.09,0.56)	0.2 (0.07,0.47)	0.37 (0.13,0.69)	0.26 (0.09,0.57)	0.15 (0.05,0.34)
PmaM3	**0.7** (0.38,0.89)	0.17 (0.06,0.38)	**0.86** (0.62,0.96)	0.4 (0.19,0.71)	**0.89** (0.68,0.97)	**0.6** (0.29,0.84)	0.4 (0.21,0.69)
Pla3988	0.05 (0.01,0.15)	0.05 (0.01,0.15)	0.09 (0.03,0.25)	0.21 (0.07,0.49)	0.25 (0.08,0.56)	0.17 (0.05,0.43)	0.4 (0.17,0.72)
PtoDC3000	**0.84** (0.57,0.95)	**0.84** (0.57,0.95)	**0.96** (0.85,0.99)	**0.89** (0.68,0.97)	**0.91** (0.72,0.98)	**0.86** (0.59,0.96)	0.4 (0.19,0.67)
PmaF9	**0.77** (0.45,0.93)	**0.77** (0.45,0.93)	**0.82** (0.55,0.95)	0.4 (0.21,0.76)	**0.87** (0.62,0.97)	**0.6** (0.23,0.8)	0.04 (0.01,0.09)
PmaF1	**0.87** (0.64,0.96)	**0.82** (0.55,0.95)	**0.97** (0.9,0.99)	**0.93** (0.77,0.98)	**0.97** (0.87,0.99)	**0.8** (0.5,0.93)	0.4 (0.19,0.65)
Ppe5846	0.05 (0.01,0.14)	0.04 (0.01,0.13)	0.16 (0.05,0.4)	0.18 (0.06,0.45)	0.07 (0.02,0.22)	0.35 (0.11,0.69)	0.01 (0.00,0.03)
PpeCFBP5143	0.06 (0.02,0.17)	0.05 (0.01,0.16)	0.1 (0.03,0.27)	0.13 (0.04,0.35)	0.1 (0.03,0.28)	0.09 (0.03,0.26)	**0.6** (0.24,0.82)
Pan126	0.11 (0.03,0.31)	0.12 (0.04,0.35)	0.16 (0.05,0.4)	0.04 (0.01,0.13)	0.4 (0.15,0.76)	**0.6** (0.32,0.77)	**0.63** (0.31,0.87)
Pan4303	0.19 (0.07,0.44)	0.2 (0.07,0.46)	**0.7** (0.39,0.9)	**0.7** (0.39,0.9)	**0.64** (0.32,0.87)	**0.6** (0.28,0.85)	**0.93** (0.77,0.98)
PtoJL1065	0.09 (0.03,0.25)	0.2 (0.07,0.45)	**0.93** (0.75,0.98)	**0.9** (0.69,0.97)	0.19 (0.07,0.44)	**0.6** (0.22,0.8)	0.27 (0.11,0.52)
PtoM14	0.08 (0.03,0.23)	0.1 (0.03,0.26)	**0.7** (0.36,0.89)	**0.6** (0.24,0.81)	0.08 (0.03,0.22)	0.3 (0.11,0.62)	0.14 (0.05,0.32)
PtoT1	0.09 (0.03,0.25)	0.1 (0.03,0.27)	**0.83** (0.53,0.95)	**0.73** (0.39,0.92)	0.14 (0.04,0.36)	0.14 (0.04,0.36)	0.19 (0.07,0.42)
PapCFBP2103	0.04 (0.01,0.12)	0.09 (0.03,0.28)	0.26 (0.08,0.57)	0.1 (0.03,0.29)	0.19 (0.06,0.48)	0.14 (0.04,0.38)	0.01 (0.00,0.02)

Supplementary Table 2.1 Summary of parameter estimates for colony forming units, plants, and strains from the logistic regression model in section 2.5. Besides maximum likelihood estimates (MLE), standard errors and P-values are also provided to indicate plant/strain factor significance.

		MLE	standard error	P-value
Colony forming units ($\hat{\beta}$)		0.85	0.06	<0.001*
Plant ($\hat{\alpha}_j^{(p)}$)	Arabidopsis	-6.19	0.48	<0.001*
	Cauliflower	-5.24	0.53	<0.001*
	Celery	-5.59	0.53	<0.001*
	Tomato	-5.02	0.51	<0.001*
	Snapdragon	-4.51	0.45	<0.001*
Strain ($\hat{\alpha}_k^{(S)}$)	PsyB728a	-0.05	0.52	0.922
	Pap1089-5	0.52	0.46	0.258
	PapBS252	0.32	0.46	0.489
	PbeATCC13454	0.24	0.50	0.630
	PbeCFBP1727	0.67	0.54	0.217
	Pma4981	2.35	0.53	<0.001*
	PtoICMP3443	1.84	0.46	<0.001*
	PtoICMP3449	1.85	0.49	<0.001*
	Pma795	0.80	0.46	0.080
	PmaM6	1.05	0.44	0.016*
	Pma2744	0.28	0.44	0.533
	PmaF15	-0.06	0.47	0.893
	PmaM3	1.68	0.42	<0.001*
	Pla3988	0.79	0.58	0.176
	PtoDC3000	2.02	0.49	<0.001*
	PmaF9	1.23	0.46	0.007*
	PmaF1	2.40	0.47	<0.001*
	Ppe5846	-0.11	0.54	0.836
	PpeCFBP5143	0.05	0.55	0.927
	Pan126	0.00	0.55	<0.001*
	Pan4303	0.84	0.44	0.057*
	PtoJL1065	1.10	0.48	0.023*
	PtoMax14	0.30	0.46	0.516
	PtoT1	0.06	0.49	0.901
	PapCFBP2103	-0.48	0.52	0.352

Supplementary Table 2.2 Results of phylogeny-trait correlation obtained with the program BaTS

Species	Statistics	Single Best Mr. Bayes tree estimate	BaTS estimate (95% confidence interval)	p-value (BaTS null hypothesis test)
Arabidopsis (Col)	AI[a]	1.20	1.23 (0.65, 1.79)	0.01
	PS[b]	6.25	6.23 (5.0, 7.0)	0.01
	MC[c](disease)	1.58	1.60 (1.0, 2.35)	0.48
	MC (non-disease)	4.05	3.95 (2.3, 5.17)	0.02
Arabidopsis (Mt)	AI	1.21	1.23 (0.64, 1.80)	0.01
	PS	6.27	6.23 (5.0, 7.0)	0.009
	MC (disease)	1.58	1.60 (1.0, 2.35)	0.48
	MC (non-disease)	4.05	3.96 (2.3, 5.17)	0.02
Cauliflower	AI	1.48	1.50 (0.81, 2.20)	0.13
	PS	8.43	8.43 (6.07, 10.0)	0.13
	MC (disease)	2.22	2.23 (1.0, 3.96)	0.89
	MC (non-disease)	2.80	2.77 (2.0, 5.0)	0.15
Celery	AI	1.50	1.53 (0.84, 2.2)	0.01
	PS	8.63	8.64 (6.49, 10.81)	0.11
	MC (disease)	2.39	2.39 (1.99,4.0)	0.94
	MC (non-disease)	2.58	2.57 (2.0, 4)	0.97
Tomato (Chicoll)	AI	1.48	1.51 (0.83, 2.20)	0.14
	PS	8.45	8.44 (6.22, 10.0)	0.37
	MC (disease)	2.78	2.77 (2.0, 5.0)	0.26
	MC (non-disease)	2.21	2.22 (1.0, 3.96)	0.89
Tomato (Riogrande)	AI	1.49	1.52 (0.84, 2.20)	0.07
	PS	8.62	8.63 (6.56, 10.71)	0.11
	MC (disease)	2.58	2.58 (2.0, 4.0)	0.11
	MC (non-disease)	2.38	2.38 (1.99, 4.0)	0.94
Snapdragon	AI	0.95	0.97 (0.46, 1.43)	0.38
	PS	4.67	4.67 (3.96, 5.0)	0.25
	MC (disease)	1.30	1.31 (1.0, 2.0)	0.24
	MC (non-disease)	5.02	4.89 (2.58, 9.0)	0.47

[a] AI: association index, [b] PS: Fitch parsimony score , [c] MC: maximum exclusive single-state clade size

Figure Legends

Figure 2.1. Bayesian tree of eight concatenated core genome gene fragments. Each sequence type (ST) is identified by a representative isolate. The total number of analyzed isolates having the same ST is indicated in parenthesis. The genome-sequenced isolates *P. syringae* pv. *syringae* (Psy) B728a and *P. fluorescence* (Pf) 0 were used as outgroups. Clade credibility values are given above branches.

Figure 2.2 Split decomposition analysis of eight core genome gene fragments. Split decomposition analysis was performed in SplitsTree4 using the PasimonySplits method, NeighborNet distance and 1000 bootstrap replicates. Bootstrap values higher than 50 are shown.

Figure 2.3. *A. thaliana*, cauliflower, celery, snapdragon, and tomato leaves infected with examples of *P. syringae* strains that either do not cause disease (left panel) or cause disease (right panel). Photos were taken seven days post infection. Photos of other strain/plant combinations are shown in Supplementary Figures 1 to 5.

Figure 2.4. Bacterial growth on *A. thaliana* (ecotypes Mt and Col), tomato (cultivars Chico III and Rio Grande), cauliflower, celery, and snapdragon. Bacterial population sizes were determined three days post infection for all plants besides celery and snapdragon for which bacterial population sizes were determined 7 days post infection. Population sizes are indicated as colony forming units/cm^2 in the log scale.

Black bars indicate that the probability of the strain for causing disease was determined to be high (between 0.6-1, see Table 2.5); white bars indicate that the probability of the strain for causing disease was determined to be low (between 0-0.4, see Table 2.5).

Figure 2.5. Bayesian tree and experimentally determined host range of each isolate in regard to *A. thaliana*, cauliflower, celery, snapdragon and tomato. Host range was determined in regard to the plant species *A. thaliana* (ecotypes Col and Mt*)*, cauliflower, celery, snapdragon and tomato (cultivars Chico III and Rio Grande). The probability of each strain to be pathogenic on each plant was determined by applying logistic regression combining visually observed symptoms and bacterial growth. A black square indicates that the probability of the strain for causing disease was determined to be between 0.6-1 (see Table 2.5). A white square indicates that the probability of the strain for causing disease was determined to be between 0-0.4 (see Table 2.5).

Figure 2.6. Ancestral state reconstruction of host range for *A. thaliana* (Col). Pathogenicity of extant strains and inferred pathogenicty of hypothetical ancestors at nodes is indicated using red for pathogenicity and green for non-pathogenicity. Pie charts indicate the relative probability of an ancestor to be either pathogenic or non-pathogenic on *A. thaliana* (Col) using the program BayesTraits. Ancestral state reconstruction of host range for *A. thaliana* (Mt) is shown in Supplementary Figures 6.

Supplementary Figure 2.1. Disease symptoms on *A. thaliana*. Black squares indicate the presence of disease. White squares indicate the absence of disease.

Supplementary Figure 2.2 Disease symptoms on cauliflower. Black squares indicate the presence of disease. White squares indicate the absence of disease.

Supplementary Figure 2.3. Disease symptoms on celery. Black squares indicate the presence of disease. White squares indicate the absence of disease.

Supplementary Figure 2.4. Disease symptoms on snapdragon. Black squares indicate the presence of disease. White squares indicate the absence of disease.

Supplementary Figure 2.5. Disease symptoms on tomato cv. Chico III. Black squares indicate the presence of disease. White squares indicate the absence of disease.

Supplementary Figure 2.6. Ancestral state reconstruction of host range for *A. thaliana* (Mt).

Supplementary Figure 2.7. Ancestral state reconstruction of host range for cauliflower.

Supplementary Figure 2.8. Ancestral state reconstruction of host range for celery.

Supplementary Figure 2.9. Ancestral state reconstruction of host range for snapdragon.

Supplementary Figure 2.10. Ancestral state reconstruction of host range for tomato (Chico III).

Figures

Figure 2.1

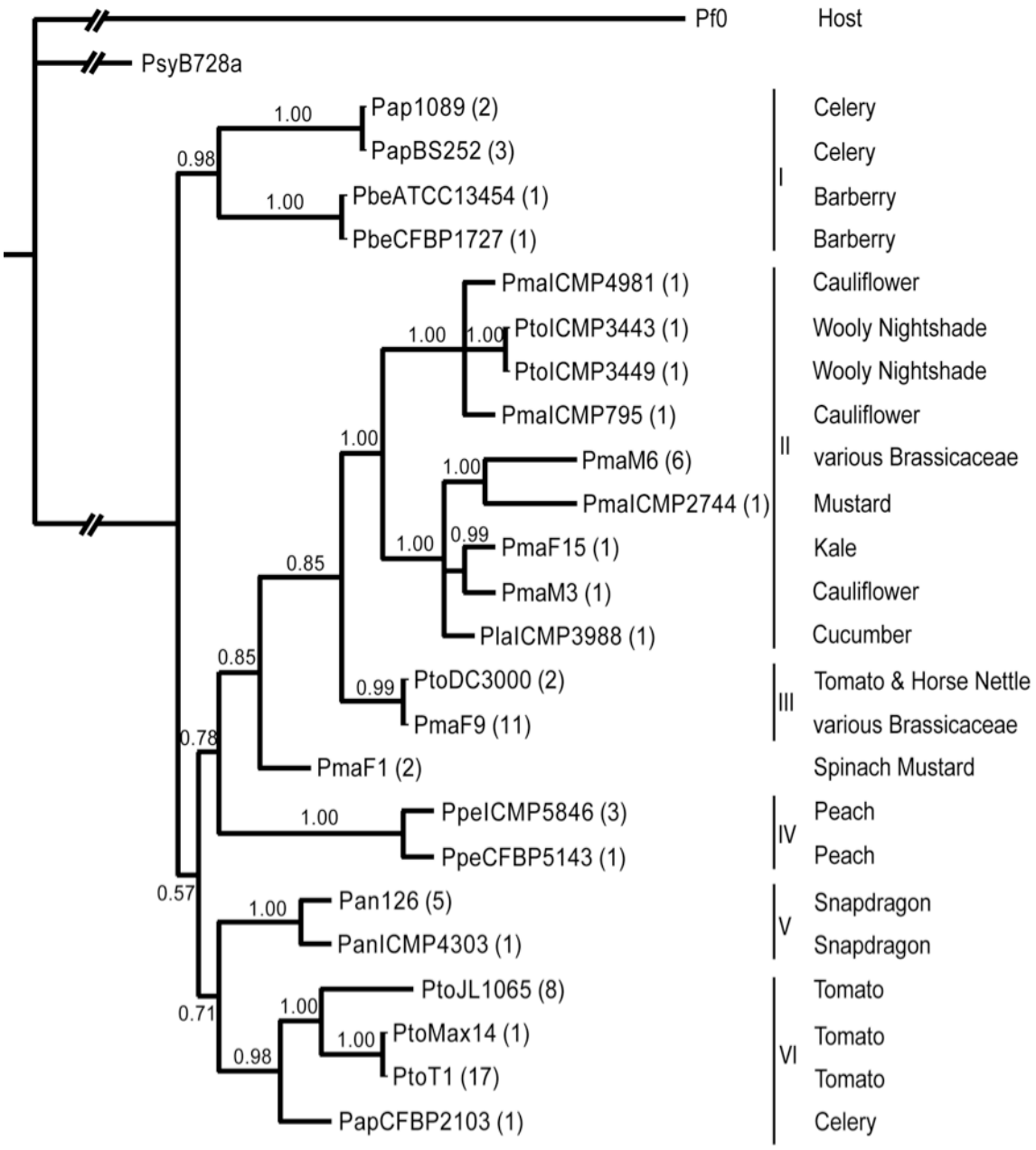


Figure 2.2

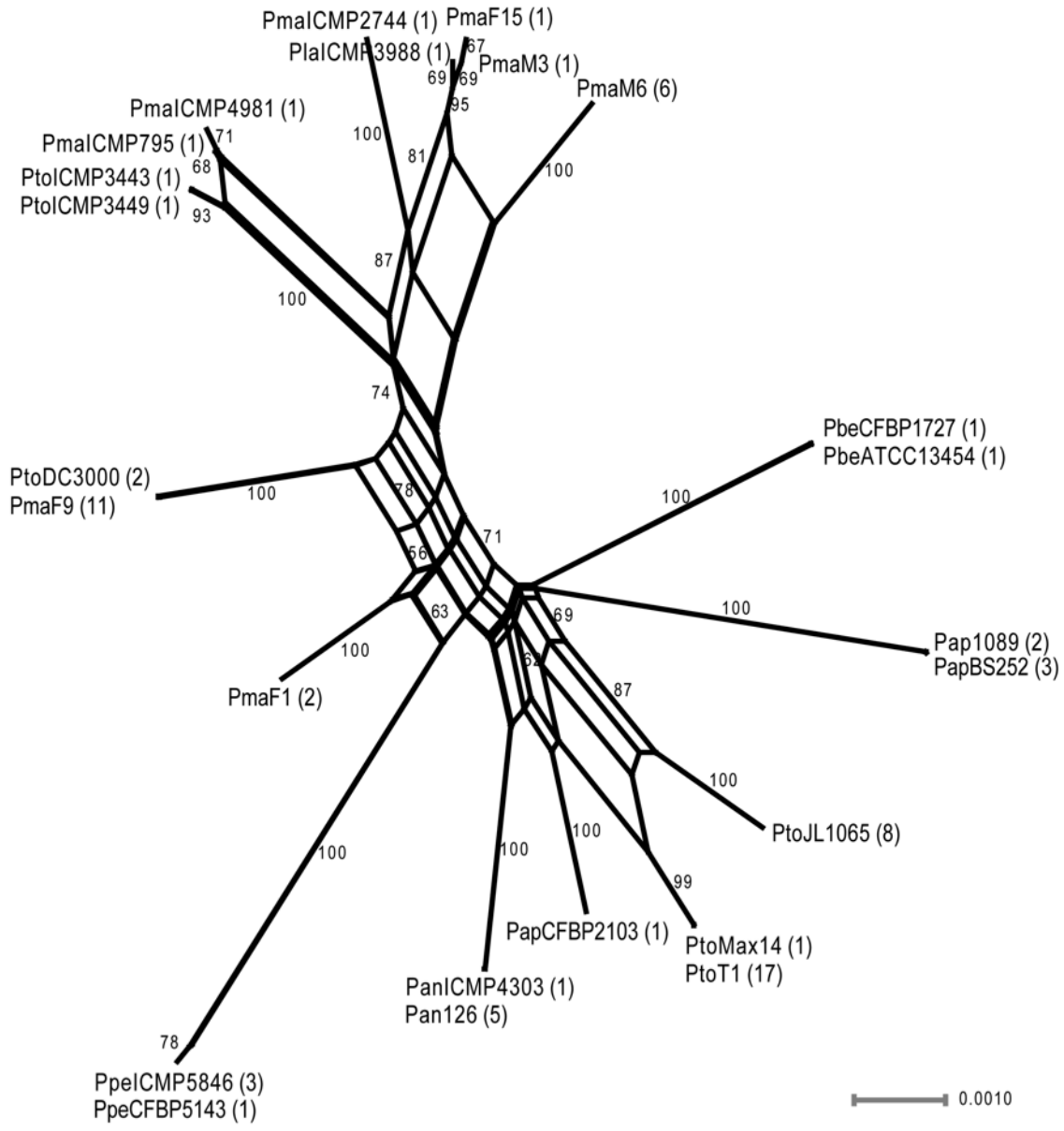

PmaICMP2744 (1)
PmaF15 (1)
PmaICMP3988 (1)
PmaM3 (1)
PmaM6 (6)
PmaICMP4981 (1)
PmaICMP795 (1)
PtoICMP3443 (1)
PtoICMP3449 (1)
PbeCFBP1727 (1)
PbeATCC13454 (1)
PtoDC3000 (2)
PmaF9 (11)
Pap1089 (2)
PapBS252 (3)
PmaF1 (2)
PtoJL1065 (8)
PapCFBP2103 (1)
PtoMax14 (1)
PtoT1 (17)
PanICMP4303 (1)
Pan126 (5)
PpeICMP5846 (3)
PpeCFBP5143 (1)
0.0010

Figure 2.3

A. thaliana (Col)

Pap1089 PtoICMP3443

Cauliflower

PpeCFBP5143 PtoICMP3443

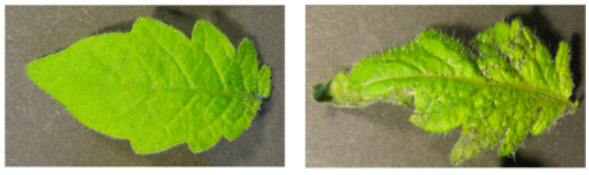
Celery

PbeCFBP1727 PapBS252

Snapdragon

PpeCFBP5143 Pan126

Tomato (Chico III)



PmaM6 PtoDC3000

Figure 2.4 Part1

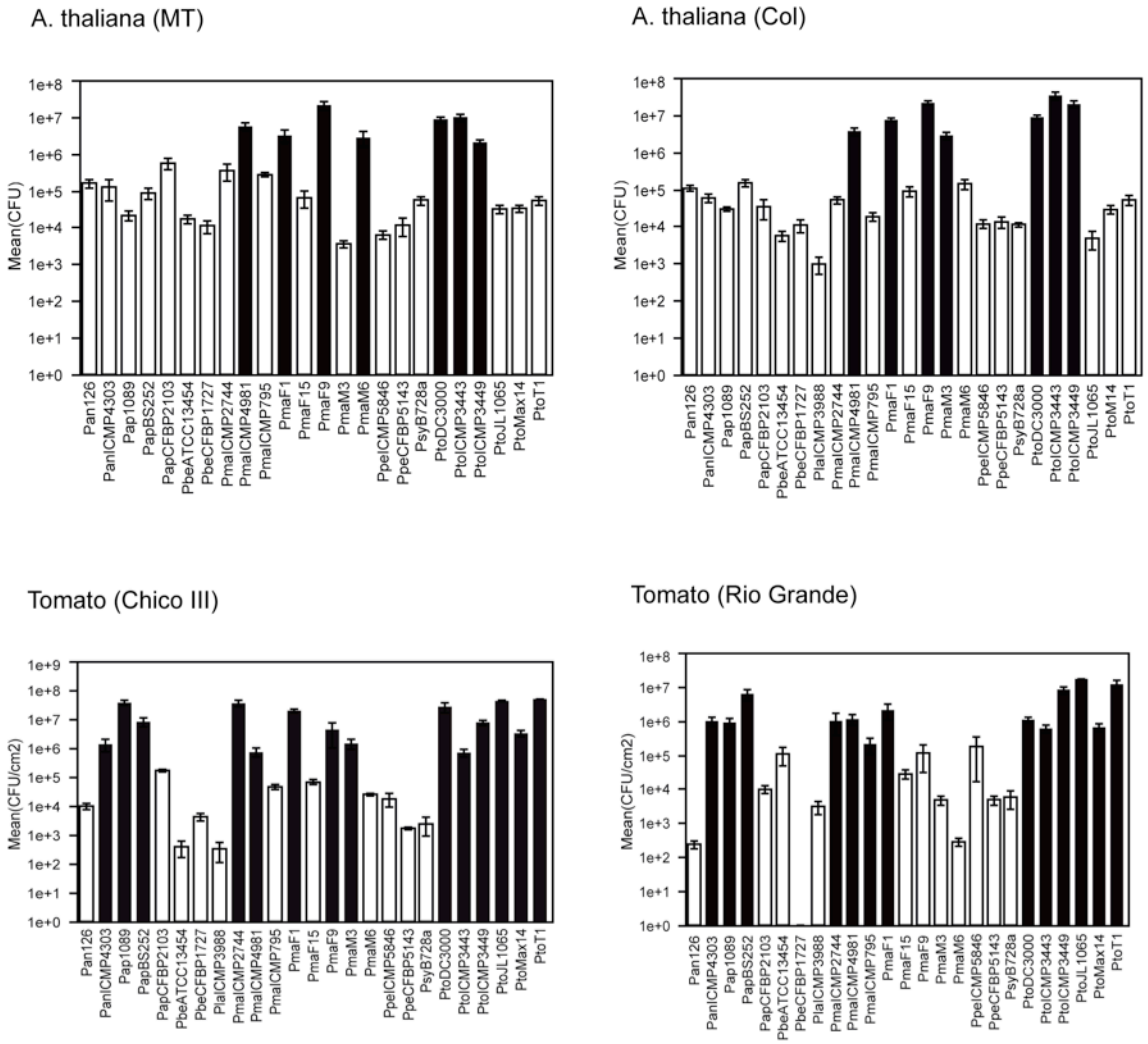


Figure 2.4 Part2

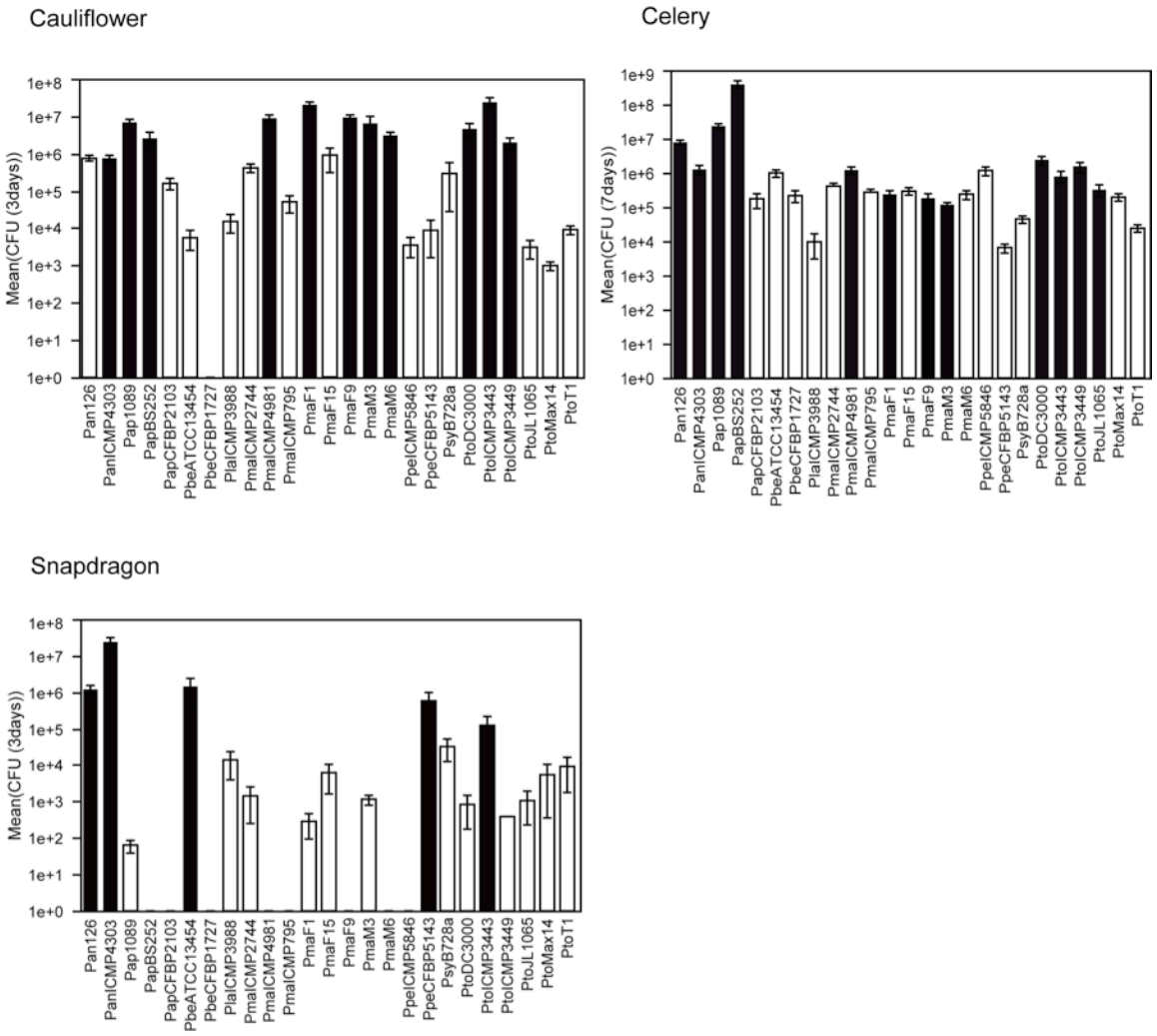


Figure 2.5

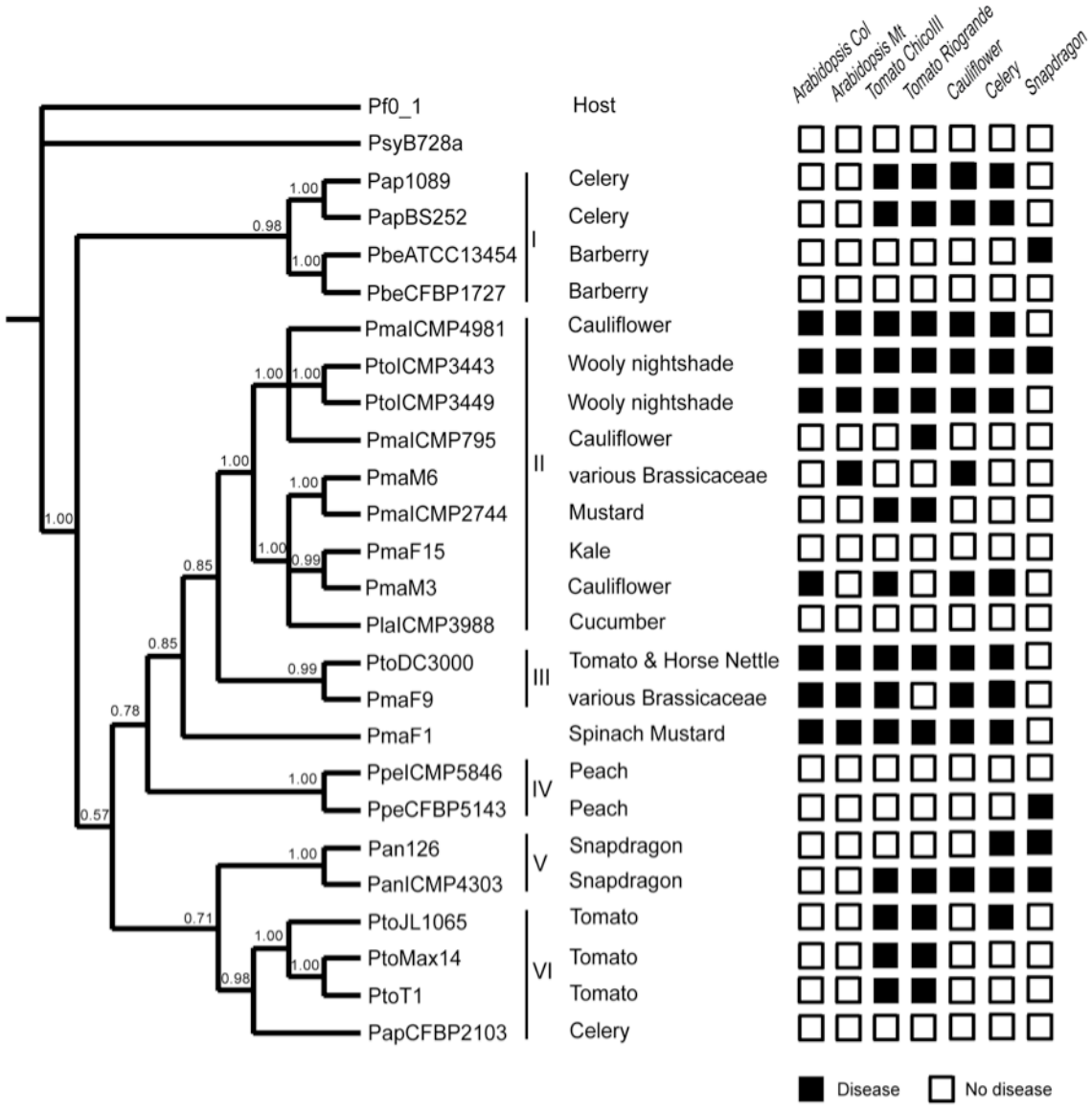
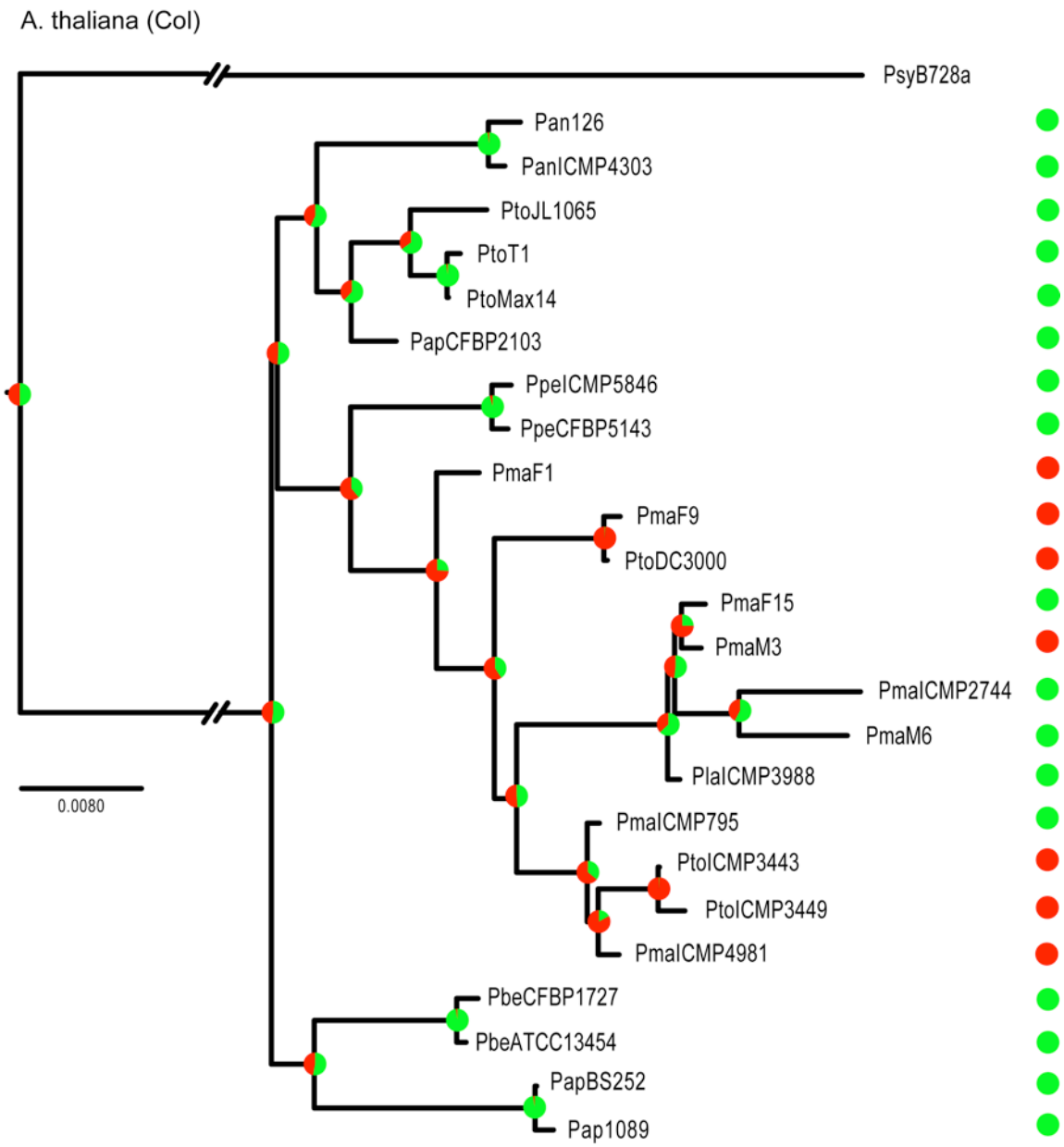
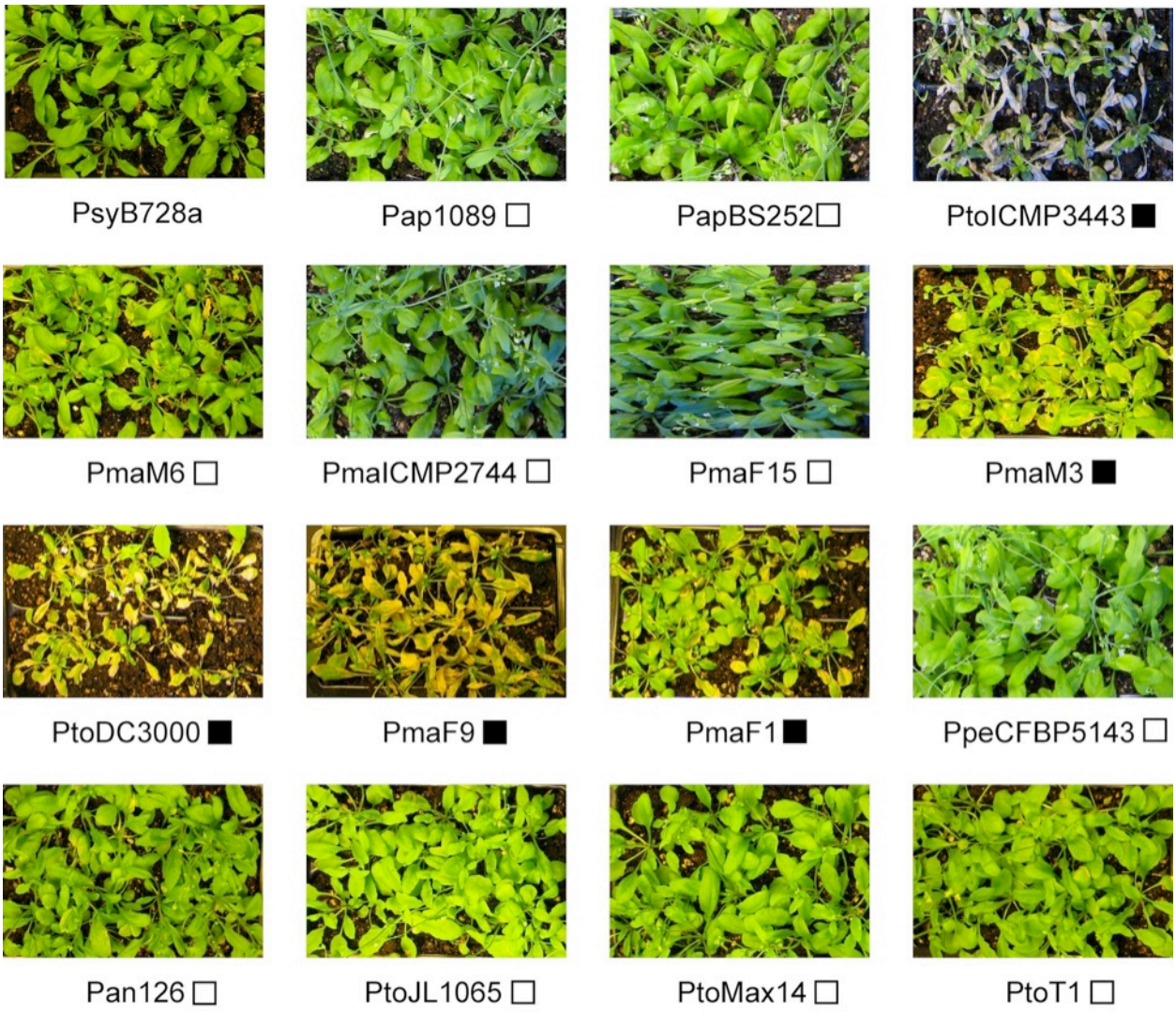

Arabidopsis Col
Arabidopsis Mt
Tomato ChicoIII
Tomato Riogrande
Cauliflower
Celery
Snapdragon
Host
Pf0_1
PsyB728a
Pap1089
PapBS252
PbeATCC13454
PbeCFBP1727
PmaICMP4981
PtoICMP3443
PtoICMP3449
PmaICMP795
PmaM6
PmaICMP2744
PmaF15
PmaM3
PlaICMP3988
PtoDC3000
PmaF9
PmaF1
PpeICMP5846
PpeCFBP5143
Pan126
PanICMP4303
PtoJL1065
PtoMax14
PtoT1
PapCFBP2103
Celery
Celery
Barberry
Barberry
Cauliflower
Wooly nightshade
Wooly nightshade
Cauliflower
various Brassicaceae
Mustard
Kale
Cauliflower
Cucumber
Tomato & Horse Nettle
various Brassicaceae
Spinach Mustard
Peach
Peach
Snapdragon
Snapdragon
Tomato
Tomato
Tomato
Celery
I
II
III
IV
V
VI
Disease
No disease

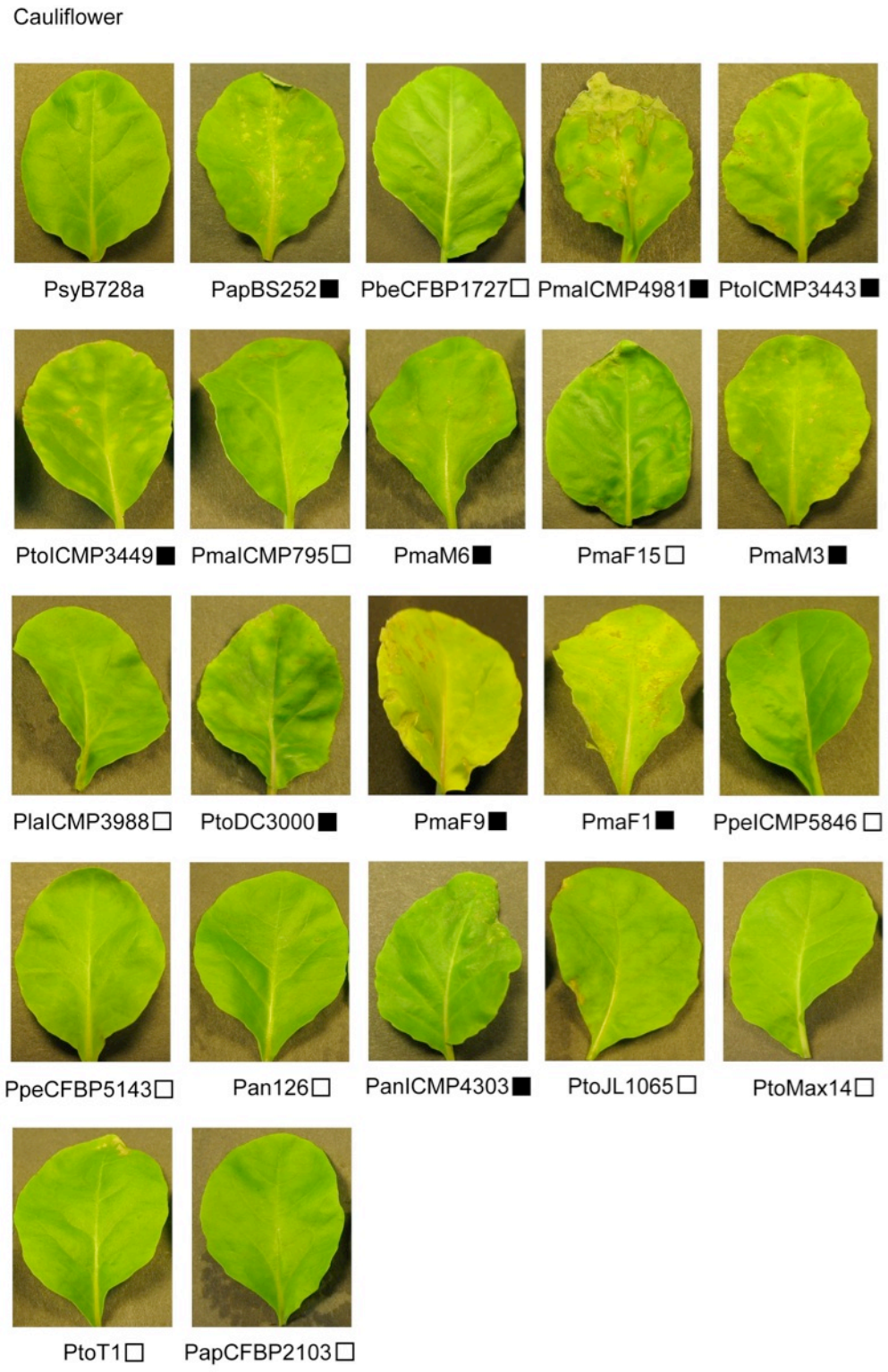
Figure 2.6


A. thaliana (Col)
PsyB728a
Pan126
PanICMP4303
PtoJL1065
PtoT1
PtoMax14
PapCFBP2103
PpeICMP5846
PpeCFBP5143
PmaF1
PmaF9
PtoDC3000
PmaF15
PmaM3
PmaICMP2744
PmaM6
PlaICMP3988
PmaICMP795
PtoICMP3443
PtoICMP3449
PmaICMP4981
PbeCFBP1727
PbeATCC13454
PapBS252
Pap1089
0.0080

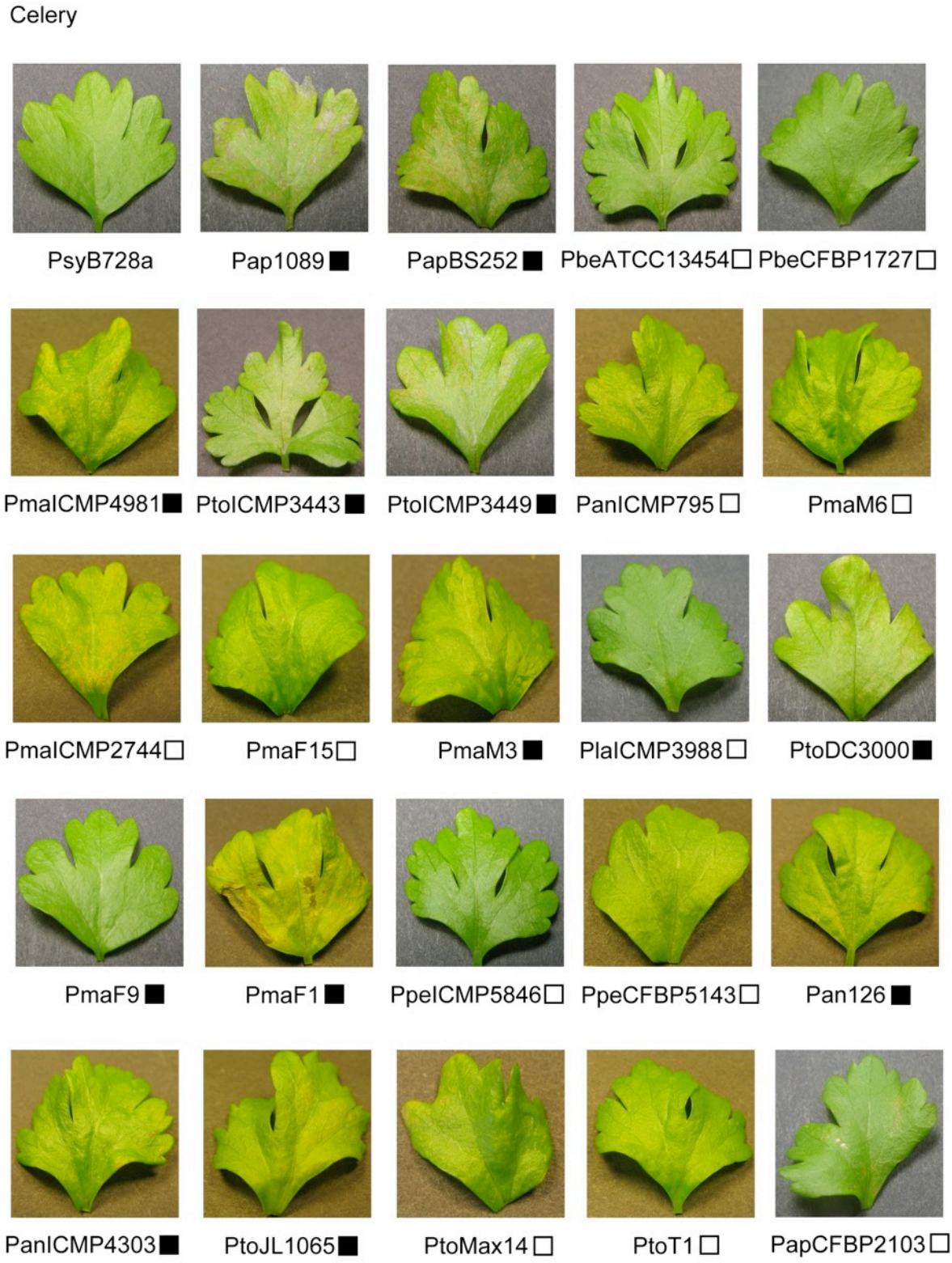
SFigure 2.1


PsyB728a
Pap1089 □
PapBS252□
PtoICMP3443 ■
PmaM6 □
PmaICMP2744 □
PmaF15 □
PmaM3 ■
PtoDC3000 ■
PmaF9 ■
PmaF1 ■
PpeCFBP5143 □
Pan126 □
PtoJL1065 □
PtoMax14 □
PtoT1 □

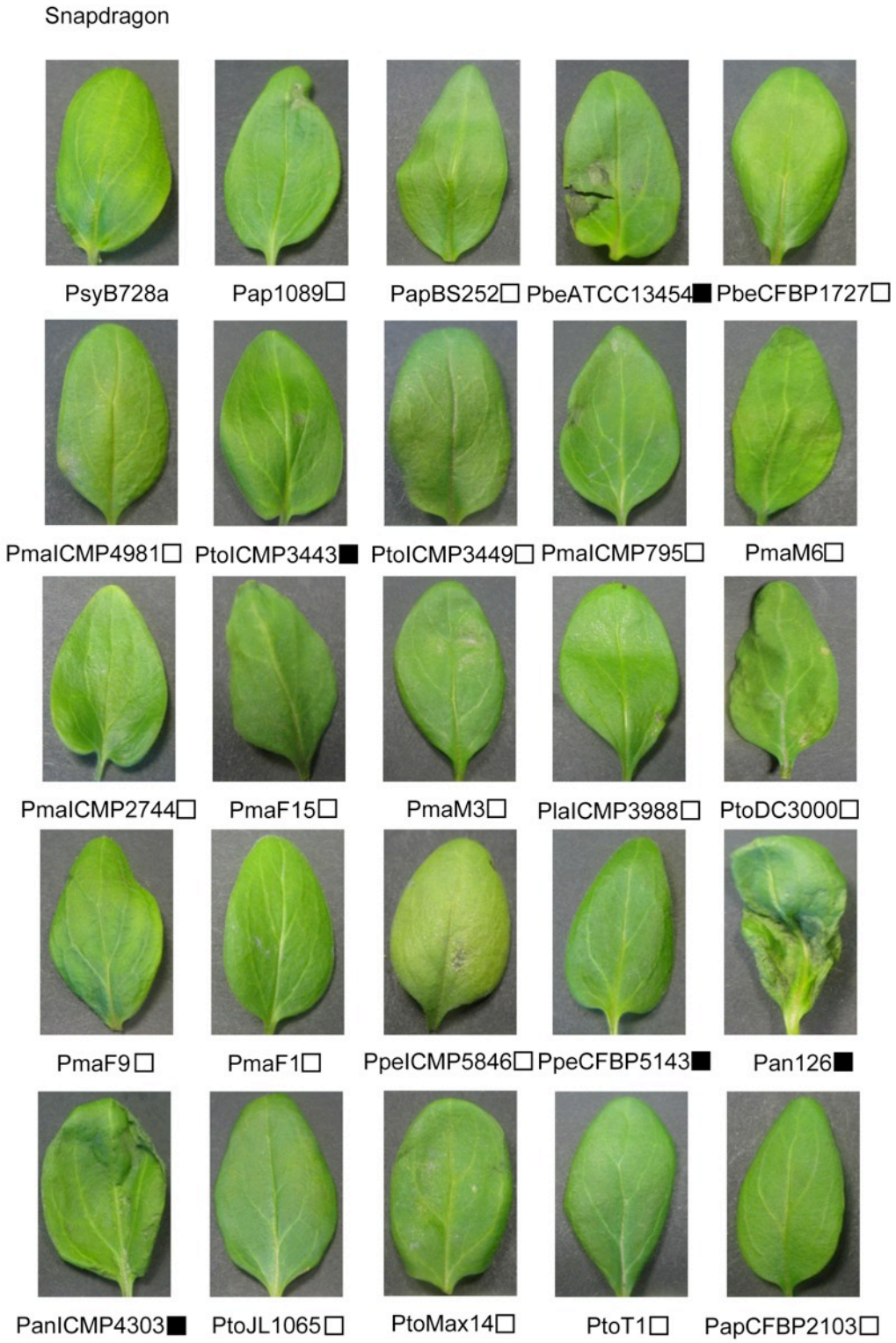
SFigure 2.2


Cauliflower
PsyB728a
PapBS252■
PbeCFBP1727□
PmaICMP4981■
PtoICMP3443■
PtoICMP3449■
PmaICMP795□
PmaM6■
PmaF15□
PmaM3■
PlaICMP3988□
PtoDC3000■
PmaF9■
PmaF1■
PpeICMP5846□
PpeCFBP5143□
Pan126□
PanICMP4303■
PtoJL1065□
PtoMax14□
PtoT1□
PapCFBP2103□

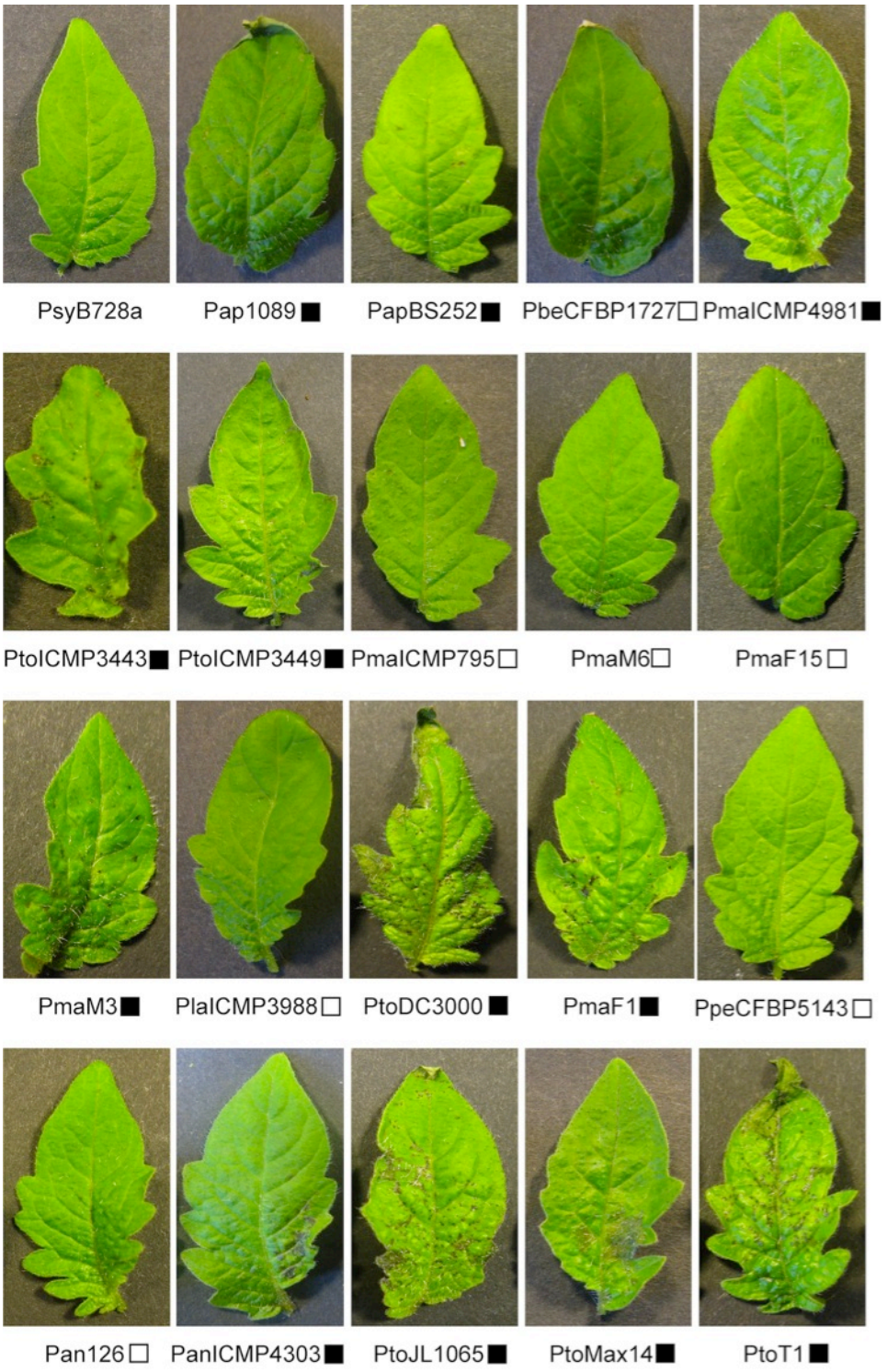
SFigure 2.3


Celery
PsyB728a
Pap1089 ■
PapBS252 ■
PbeATCC13454 □
PbeCFBP1727 □
PmaICMP4981 ■
PtoICMP3443 ■
PtoICMP3449 ■
PanICMP795 □
PmaM6 □
PmaICMP2744 □
PmaF15 □
PmaM3 ■
PlaICMP3988 □
PtoDC3000 ■
PmaF9 ■
PmaF1 ■
PpeICMP5846 □
PpeCFBP5143 □
Pan126 ■
PanICMP4303 ■
PtoJL1065 ■
PtoMax14 □
PtoT1 □
PapCFBP2103 □

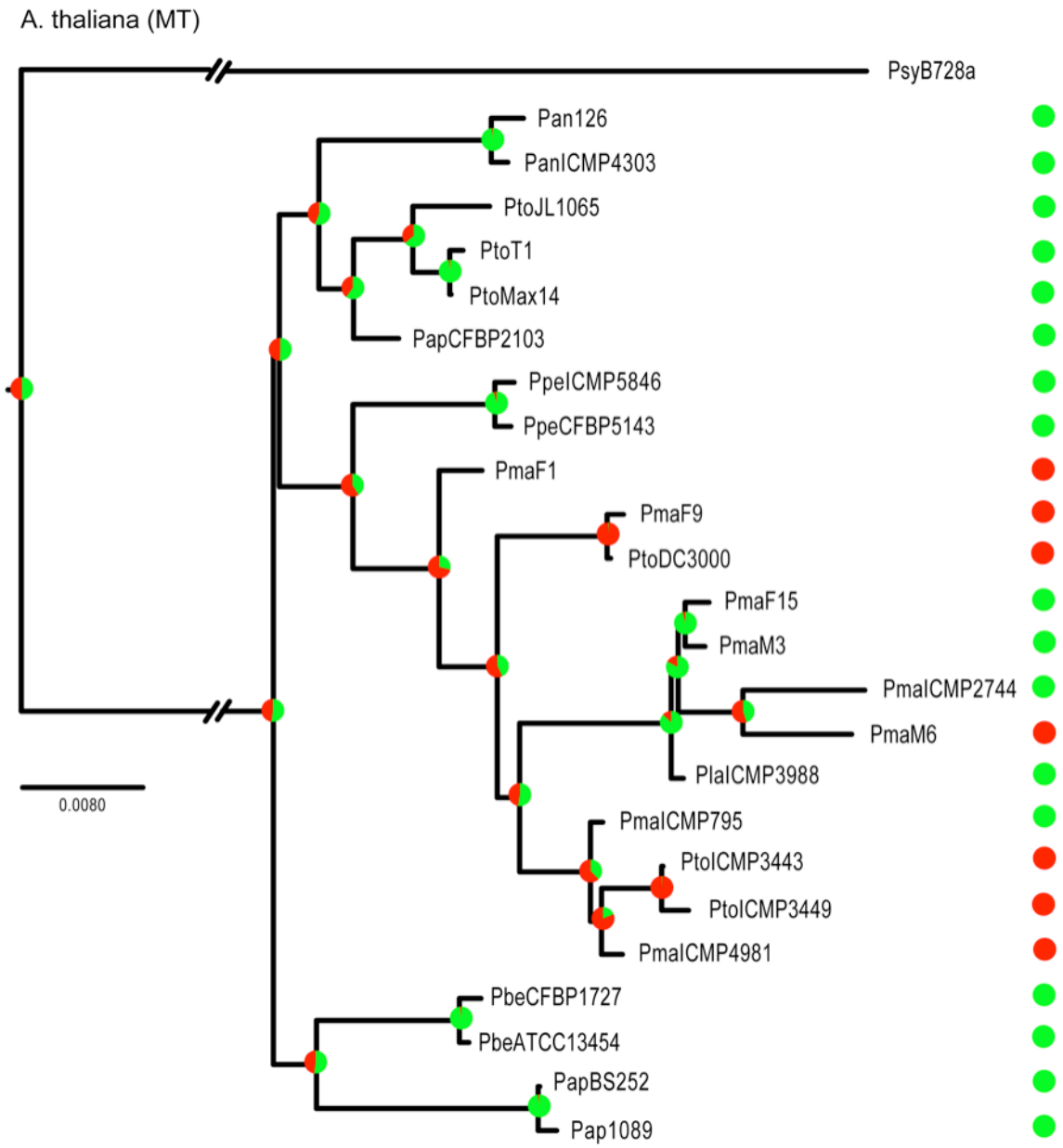
SFigure 2.4


Snapdragon
PsyB728a
Pap1089☐
PapBS252☐
PbeATCC13454■
PbeCFBP1727☐
PmaICMP4981☐
PtoICMP3443■
PtoICMP3449☐
PmaICMP795☐
PmaM6☐
PmaICMP2744☐
PmaF15☐
PmaM3☐
PlaICMP3988☐
PtoDC3000☐
PmaF9☐
PmaF1☐
PpeICMP5846☐
PpeCFBP5143■
Pan126■
PanICMP4303■
PtoJL1065☐
PtoMax14☐
PtoT1☐
PapCFBP2103☐

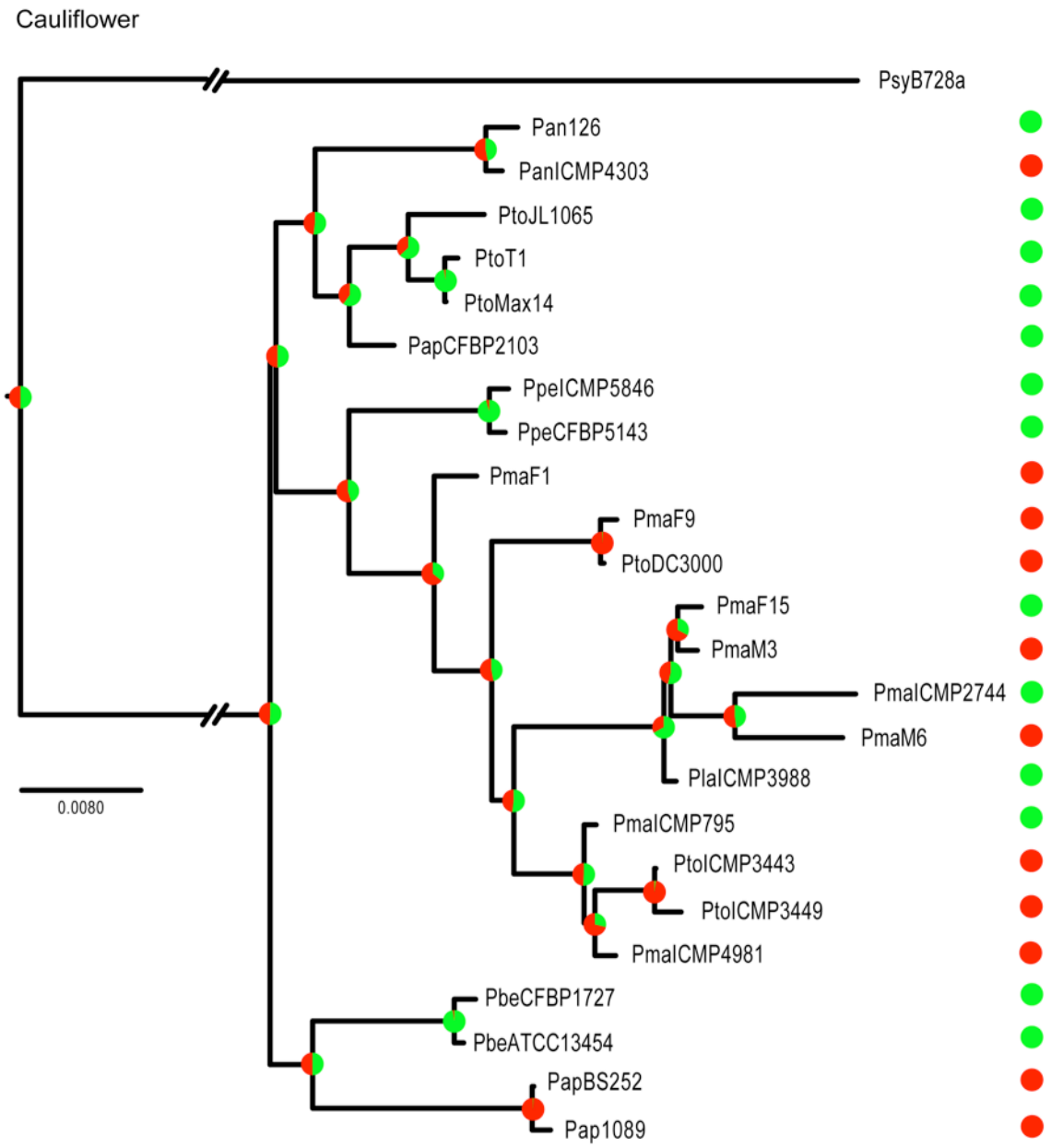
SFigure 2.5


PsyB728a
Pap1089■
PapBS252■
PbeCFBP1727□
PmaICMP4981■
PtoICMP3443■
PtoICMP3449■
PmaICMP795□
PmaM6□
PmaF15□
PmaM3■
PlaICMP3988□
PtoDC3000■
PmaF1■
PpeCFBP5143□
Pan126□
PanICMP4303■
PtoJL1065■
PtoMax14■
PtoT1■

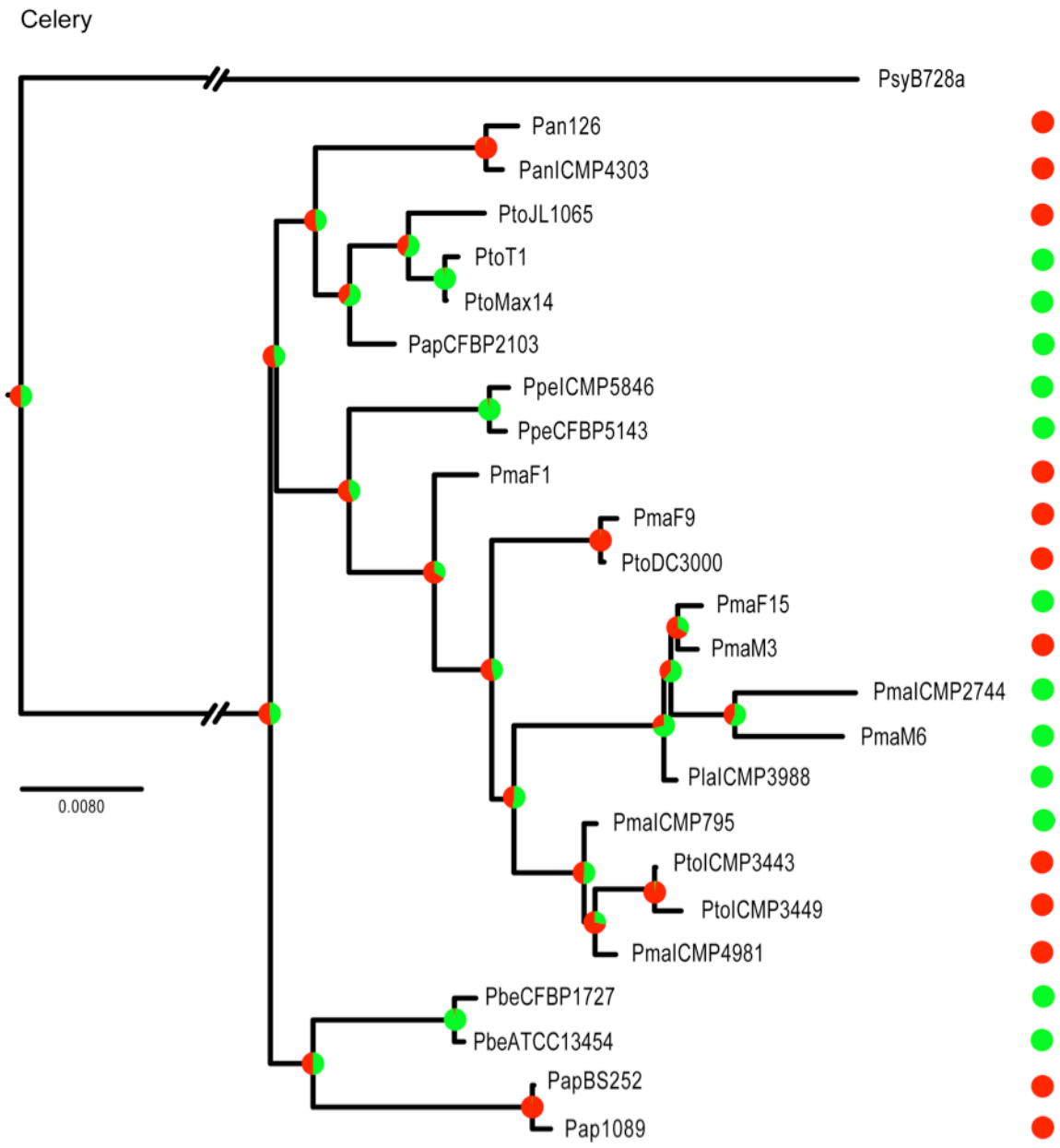
SFigure 2.6


A. thaliana (MT)
PsyB728a
Pan126
PanICMP4303
PtoJL1065
PtoT1
PtoMax14
PapCFBP2103
PpeICMP5846
PpeCFBP5143
PmaF1
PmaF9
PtoDC3000
PmaF15
PmaM3
PmaICMP2744
PmaM6
PlaICMP3988
PmaICMP795
PtoICMP3443
PtoICMP3449
PmaICMP4981
PbeCFBP1727
PbeATCC13454
PapBS252
Pap1089
0.0080

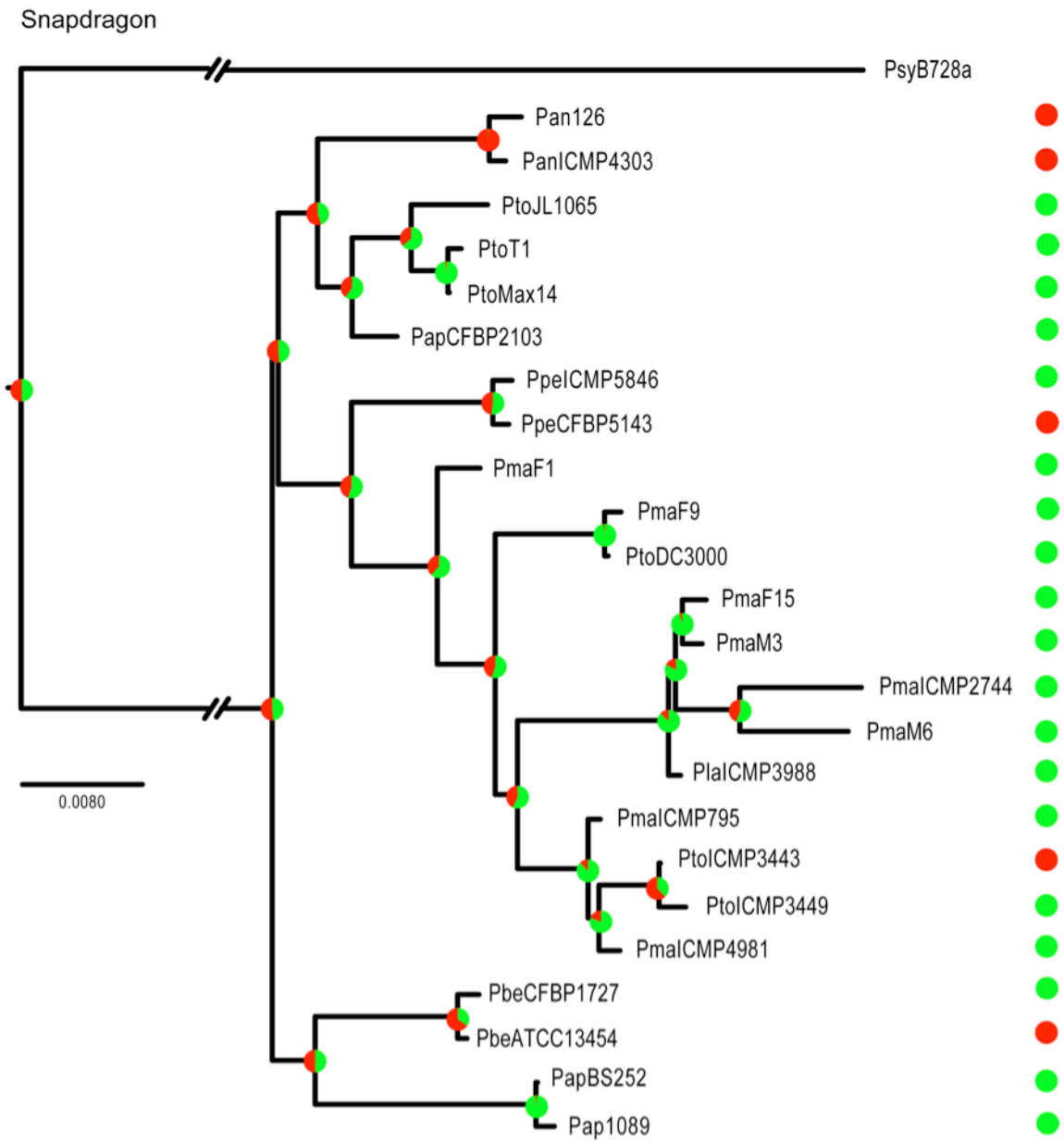
SFigure 2.7


Cauliflower
PsyB728a
Pan126
PanICMP4303
PtoJL1065
PtoT1
PtoMax14
PapCFBP2103
PpeICMP5846
PpeCFBP5143
PmaF1
PmaF9
PtoDC3000
PmaF15
PmaM3
PmaICMP2744
PmaM6
PlaICMP3988
PmaICMP795
PtoICMP3443
PtoICMP3449
PmaICMP4981
PbeCFBP1727
PbeATCC13454
PapBS252
Pap1089
0.0080

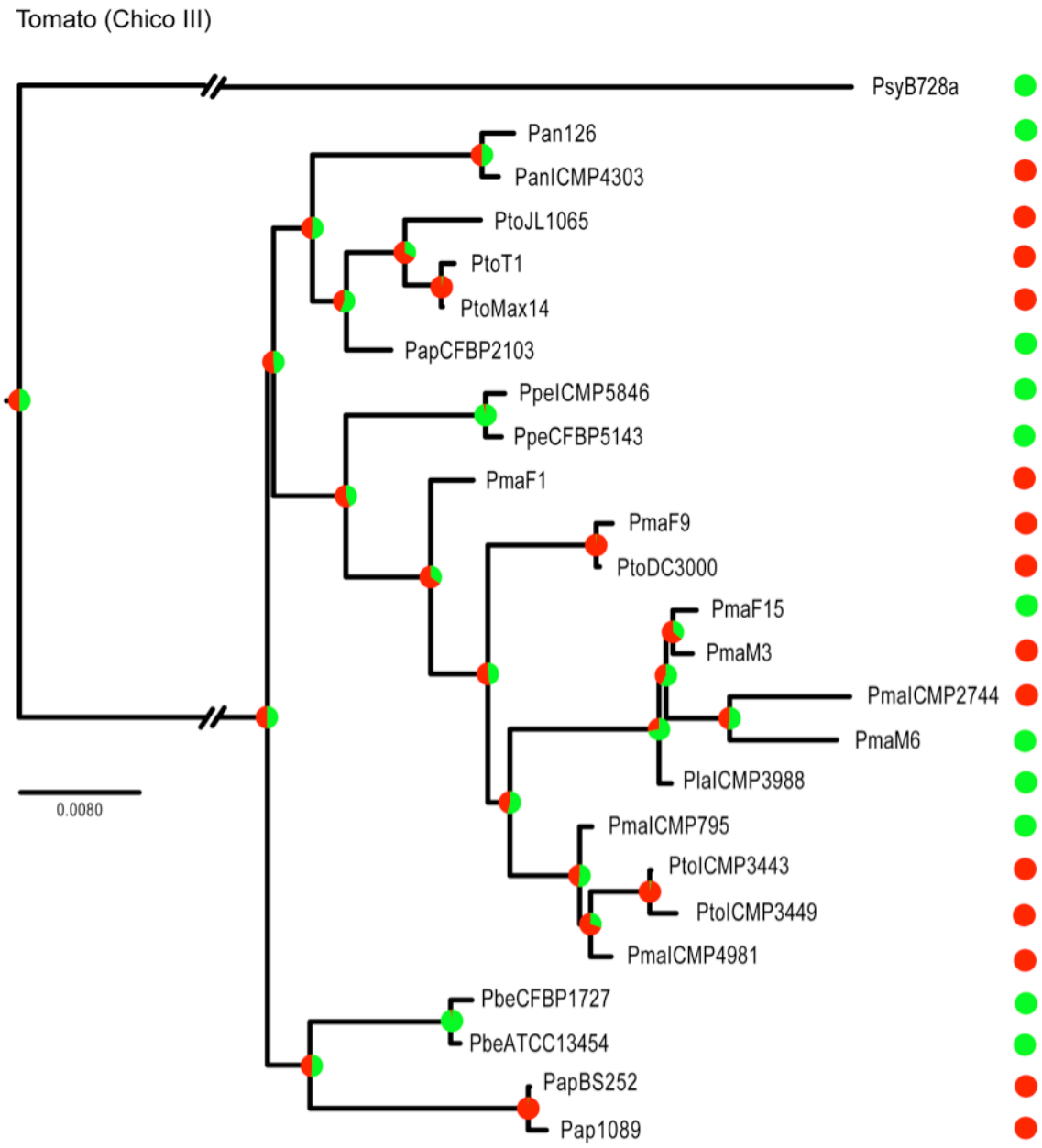
SFigure 2.8


Celery
PsyB728a
Pan126
PanICMP4303
PtoJL1065
PtoT1
PtoMax14
PapCFBP2103
PpeICMP5846
PpeCFBP5143
PmaF1
PmaF9
PtoDC3000
PmaF15
PmaM3
PmaICMP2744
PmaM6
PlaICMP3988
PmaICMP795
PtoICMP3443
PtoICMP3449
PmaICMP4981
PbeCFBP1727
PbeATCC13454
PapBS252
Pap1089
0.0080

SFigure 2.9


Snapdragon
PsyB728a
Pan126
PanICMP4303
PtoJL1065
PtoT1
PtoMax14
PapCFBP2103
PpeICMP5846
PpeCFBP5143
PmaF1
PmaF9
PtoDC3000
PmaF15
PmaM3
PmaICMP2744
PmaM6
PlaICMP3988
PmaICMP795
PtoICMP3443
PtoICMP3449
PmaICMP4981
PbeCFBP1727
PbeATCC13454
PapBS252
Pap1089
0.0080

SFigure 2.10


Tomato (Chico III)
PsyB728a
Pan126
PanICMP4303
PtoJL1065
PtoT1
PtoMax14
PapCFBP2103
PpeICMP5846
PpeCFBP5143
PmaF1
PmaF9
PtoDC3000
PmaF15
PmaM3
PmaICMP2744
PmaM6
PlaICMP3988
PmaICMP795
PtoICMP3443
PtoICMP3449
PmaICMP4981
PbeCFBP1727
PbeATCC13454
PapBS252
Pap1089
0.0080

Chapter 3

The plant pathogen *Pseudomonas syringae* pv. *tomato* is genetically monomorphic and under strong selection to evade tomato immunity

Rongman Cai[1*], James Lewis[1*], Shuangchun Yan[1], Haijie Liu[1], Christopher R. Clarke[1], Francesco Campanile[2], Nalvo F. Almeida[1, 3, 4], David J. Studholme[5], Magdalen Lindeberg[6], David Schneider[7], Massimo Zaccardelli[2], Joao C. Setubal[3, 8], Nadia P. Morales-Lizcano[9], Adriana Bernal[9], Gitta Coaker[10], Christy Baker[11], Carol L. Bender[11], Scotland Leman[12], Boris A. Vinatzer[1†]

[1] Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Latham Hall, Ag Quad Lane, Blacksburg, VA, USA

[2] CRA-Centro di Ricerca per l'Orticoltura, Sede di Battipaglia, Battipaglia (SA), Italy

[3] Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA

[4] Faculty of Computing, Federal University of Mato Grosso do Sul, Brazil

[5] Biosciences, University of Exeter, Exeter, Devon, UK

[6] Department of Plant Pathology and Plant – Microbe Biology, Cornell University, Ithaca, New Yor, USA

[7] U. S. Department of Agriculture Agricultural Research Service, Ithaca, New York, USA

[8] Department of Computer Science, Virginia Tech, Blacksburg, Virginia, USA

[9] Universidad de los Andes, Bogota, Colombia

[10] Department of Plant Pathology, University of California, Davis, California, USA

[11] Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK

[12] Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

*these authors' contribution was equal

[†] To whom correspondence should be addressed

Contact: Boris vinatzer, e-mail: vinatzer@vt.edu, phone: +1 540 231 2126, fax: +1 540 231 3347

Running title: Crop pathogen evolution and new MAMP

Abstract

Recently, genome sequencing of many isolates of genetically monomorphic bacterial human pathogens has given new insights into pathogen microevolution and phylogeography. Here, we report a genome-based micro-evolutionary study of a bacterial plant pathogen, *Pseudomonas syringae* pv. *tomato*. Only 267 mutations were identified between five sequenced isolates in 3,543,009 nt of analyzed genome sequence, which suggests a recent evolutionary origin of this pathogen. Further analysis with genome-derived markers of 89 world-wide isolates showed that several genotypes exist in North America and in Europe indicating frequent pathogen movement between these world regions. Genome-derived markers and molecular analyses of key pathogen loci important for virulence and motility both suggest ongoing adaptation to the tomato host. A mutational hotspot was found in the type III-secreted effector gene *hopM1*. These mutations abolish the cell death triggering activity of the full-length protein indicating strong selection for loss of function of this effector, which was previously considered a virulence factor. Two non-synonymous mutations in *fliC*, which encodes the flagellum subunit flagellin, in a region distinct from the known microbe associated molecular pattern (MAMP) flg22 allowed to identify a new MAMP. Interestingly, the ancestral allele of this MAMP induces a stronger tomato immune response than the derived alleles. The ancestral allele has largely disappeared from today's *Pto* populations suggesting that flagellin-triggered immunity limits pathogen fitness even in highly virulent pathogens. An additional non-synonymous mutation was identified in flg22 in Colombian isolates. Therefore, MAMPs are much more variable

than expected differing even between otherwise almost identical isolates of the same pathogen strain.

Author Summary

Our knowledge of the recent evolution of bacterial human pathogens has increased dramatically over the last five years. By comparison, relatively little is known about recent evolution of bacterial plant pathogens. Here, we analyze a large collection of isolates of the economically important plant pathogen *Pseudomonas syringae* pv. *tomato* with markers derived from the comparison of five genomes of this pathogen. We find that this pathogen likely evolved on a relatively recent time scale and continues to adapt to tomato by minimizing its recognition by the tomato immune system. We find that an allele of the flagellin subunit *fliC* that appeared in the pathogen population for the first time in the 1980s, and which is the most common allele of this gene in North America and Europe today, triggers a weaker tomato immune response than the *fliC* allele found in the 1960s and 1970s. These results not only impact our understanding of pathogen – plant interactions and pathogen evolution but also have important ramifications for disease prevention. Given the speed with which new pathogen strains spread and replace existing strains, limiting the movement of specific strains between geographic regions is critically important, even for pathogens known to have worldwide distribution.

Introduction

Most taxonomic descriptions of bacterial plant pathogens and studies of their life cycle were performed at a time when it was impossible to classify bacteria precisely. Therefore, it can be difficult to determine whether plant diseases affecting crops in the field today are caused by the same pathogens described in the literature as their causal agents. Moreover, in the absence of precise classification and identification of field isolates, new pathogen variants with increased virulence may spread around the globe unobserved, presenting a potential threat to biosecurity. Furthermore, model plant pathogen strains studied for their molecular interactions with plants in laboratories may not be representative of the pathogens that cause disease in the field and genes required for pathogen success in the field may not even impact bacterial growth or virulence when evaluated under laboratory conditions, which are generally optimized for disease development.

Several human diseases are caused by genetically monomorphic bacterial pathogens that evolved only after the human migration out of Africa. Genome sequencing of multiple strains belonging to each of these pathogens has elucidated their microevolution and their worldwide routes of dispersion. Examples include *Yersinia pestis* (Morelli, Song et al. 2010), *Bacillus anthracis* (Van Ert, Easterday et al. 2007), and *Salmonella* Typhi (Holt, Parkhill et al. 2008). Moreover, microevolution of clonal lineages within diverse pathogen species like *Escherichia coli*, *Staphylococcus aureus*, and *Clostridium difficile* have also been unraveled using single nucleotide

polymorphisms identified between genomes (Manning, Motiwala et al. 2008; Harris, Feil et al. 2010; He, Sebaihia et al. 2010). Similar studies have yet to be undertaken for plant pathogens. *Pseudomonas syringae* pv. *tomato* (*Pto*) is the causative agent of the bacterial speck disease of tomato (*Solanum lycopersicum*), a disease that occurs worldwide and causes severe reduction in fruit yield and quality, particularly during cold and wet springs, such as occurred in Florida and California in 2010. Three clonal lineages of *Pto* have been previously described based on multilocus sequence typing (MLST): T1, JL1065, and DC3000 (Yan, Liu et al. 2008). Housekeeping genes of JL1065 and T1 differ in DNA sequence by only 0.4% while DC3000 differs from JL1065 and T1 by 0.9%. JL1065 and T1 were found to be the common pathogenic agents of bacterial speck in the field worldwide. Although DC3000 is a derivative of the pathotype strain of *Pto* and the model pathogen most commonly used to investigate the molecular basis of bacterial speck disease (Buell, Joardar et al. 2003), this lineage is only rarely found on tomato (Yan, Liu et al. 2008). Comparing genomes of multiple isolates of the *P. syringae* pv. *tomato* (*Pto*) T1 lineage and performing a Single Nucleotide Polymorphism (SNP) analysis of a large collection of T1-like strains, we attempt here for the first time to unravel the microevolution and global spread of a bacterial plant pathogen.

Results and Discussion

T1-like strains are the most common Pto strains today

Extending our previous MLST analysis (Yan, Liu et al. 2008) to 112 *Pto* isolates collected worldwide between 1942 and 2009 (Table 3.1) we confirmed that T1 is the

most common *Pto* lineage, followed by JL1065 and DC3000. In fact, among all analyzed isolates only two DC3000-like strains and twenty-one JL1065-like strains were found while 89 isolates belonged to the T1 lineage. When plotting strain frequency over time (Figure 3.1A) and considering geographic origin of strains (Table 3.1 and Figure 3.1B), we observed an intriguing trend: DC3000-like and JL1065-like strains were the only *Pto* strains isolated until 1961 when the first T1-like strain was collected in the UK. T1-strains then quickly increased in frequency becoming the most common *Pto* lineage. Some JL-1065 strains were still isolated in the 1980s and 1990s but all strains in our collection isolated in Europe and North America after 1999 belong to the T1 lineage.

Genomes of five T1-like strains are extremely similar to each other

To investigate the recent evolution and virulence mechanisms of the T1 lineage, we obtained draft genome sequences using Illumina technology (Bentley 2006) of four T1-like strains in addition to the already sequenced genome of strain T1 (Almeida, Yan et al. 2009), which was collected in Canada in 1986. These four newly sequenced strains are: NCPPB1108 collected in the UK in 1961, LNPV17.41 collected in France in 1996, Max4 collected in Italy in 2002, and K40 isolated in the USA in 2005. These strains were chosen because they represent the diversity of our strain collection in regard to time of isolation and geographic location. The genomes of NCPPB1108, LNPV17.41, and K40 were assembled and submitted to the NCBI genome database (NZ_ADGA00000000, ADFZ00000000, NZ_ADFY00000000), annotated, and predicted protein repertoires were compared with other *P. syringae* genomes. The genome of Max4 was neither submitted to NCBI nor annotated owing to significantly higher fragmentation relative to

the other three genomes. A summary description of genomes can be found in Table 3.2 and predicted protein repertoires can be compared with additional *P. syringae* genomes online at genome.ppws.vt.edu .

Sequencing reads were aligned against the DC3000 genome and 11,145 high confidence single nucleotide polymorphisms (SNPs) were identified between DC3000 and the five T1-like genomes using the program MAQ (Li, Ruan et al. 2008). However, only a total of 157 SNPs were identified between any of the five T1-like strains, underscoring the close relationship among these strains (Supplementary Table 3.1). To validate the identified SNPs we also used a second approach. This time we called SNPs between the five T1-like genomes using the T1 genome as reference for alignment, used less stringent criteria, but limited SNP identification to *P. syringae* core genome genes (see details in regard to the differences between Maq settings used in the two approaches in the Materials and Methods section). Limiting SNP identification to the core genome allowed reliable SNP calls applying less stringent settings since genes in the core genome are present only in single copy, thus avoiding misalignment of reads typical with multigene families. 265 SNPs (listed in Supplementary Table 3.2) were identified in this way. Twenty-three of these SNPs were re-sequenced from PCR products using Sanger sequencing and all were confirmed (data not shown) giving us confidence in the reliability of this second approach. Since the total length of the core genome used for SNP identification in the second approach was 3,543,009 nt and the identified number of SNPs distinguishing pairs of genomes was found to be between 53 and 183 (based on the SNPs listed in Supplementary Table 3.2), the five T1-like core

genomes were determined to have pair-wise genetic distances between 0.000017 and 0.000098. This clearly shows that *Pto* is a genetically monomorphic pathogen similar to, for example, *Yersinia pestis* or *Salmonella* Typhi, both of which evolved only subsequent to human migration out of Africa (Achtman 2008). However, it is challenging to even estimate an approximate divergence time for the five sequenced T1-like strains since a yearly mutation rate has not yet been determined for any plant associated bacterium and data from the five genomes sequenced here are not sufficient to reliably infer a mutation rate based on the sequenced strains themselves and their time of isolation. Nonetheless, we attempted to get a rough estimate of divergence time assuming a minimum mutation rate of $3.4x10^{-9}$ per base pair per year as estimated for bacteria based on the *E. coli* and *Salmonella enterica* split (Ochman and Wilson 1987) and a maximum mutation rate of 5×10^{-6} per bp per year, which is similar to the maximum clock rates recently inferred for a clonal methicillin resistant *S. aureus* (MRSA) lineage (N√ºbel, Dordel et al. 2010) and for *Helicobacter pylori* (Morelli, Didelot et al. 2010) and similar to a maximum clock rate assumed previously for the plant pathogen *Clavibacter michiganensis* subsp. *sepedonicus* (Bentley, Corton et al. 2008). We then used the programs IMa2 (Hey and Nielsen 2007; Hey 2010) and BEAST (Drummond and Rambaut 2007) to calculate divergence times for each pair of strains. The obtained results suggest divergence times of around thousand years or less using the maximum mutation rate (Supplementary Table 3.3) or around one million years using the minimum mutation rate. However, (Ochman and Wilson 1987)considering that some of the T1-like genomes have a genetic distance from each other similar to that of the MRSA isolates

analyzed by Nübel and colleagues (N√obel, Dordel et al. 2010) for which a divergence time of only 20 years was inferred, we believe that T1-like strains have likely evolved from their most recent common ancestor after the domestication of tomato, which must have occurred sometime before the 15th century when tomatoes were first brought from Mexico to Europe (Peralta and Spooner 2007). To obtain a more reliable estimate of divergence times the yearly mutation rate for plant pathogens will need to be inferred in the future based on the genomes of many more strains isolated in different years from a geographic area, where the approximate year of a single recent introduction is known, as is the case for example for *P. syringae* pv *aesculi* recently introduced into the United Kingdom (Green, Studholme et al. 2010).

A phylogenetic tree was then constructed based on the SNPs indentified by aligning sequencing reads of the five T1-like strains against the DC3000 genome (Figure 3.2A). DC3000 was used as outgroup but only SNPs that distinguished the five T1-like strains from each other were considered (that is, SNPs that distinguished only DC3000 from all five T1-like strains were excluded because they were not informative with respect to evolution of T1-like strains). Trees with identical topology were obtained using only intergenic, intragenic, synonymous, or non-synonymous SNPs (data not shown), suggesting that selection did not significantly affect phylogenetic reconstruction. Typical for recently diverged bacterial genomes (Pearson, Okinaka et al. 2009), no homoplasies or recombination events were detected. Interestingly, strain NCPPB1108 isolated in 1961is located on the most basal branch of the tree, followed by T1 isolated in 1986 on the next branch, while the most recently isolated strains LNPV 17.41 (1996),

Max4 (2002), and K40 (2005) cluster together on the most derived branch. This could suggest that in the last 50 years we have witnessed an evolution of T1-like strains whereby the strains found on tomato today may have replaced their ancestors of the recent past and may be relatively more fit.

A SNP analysis suggests T1-like populations have replaced each other repeatedly over the last 50 years in North America and Europe

To address the question as to whether T1-like strains have evolved since 1961, we sequenced for all 89 T1-like strains in our collection the seven informative SNP loci distinguishing strains Max4, LNPV17.41, and K40 from strains T1, NCPPB1108, and DC3000 (which were identified in the alignment of the Max4, LNPV17.41, K40, and NCPPB1108 sequencing reads against the T1 genome). We also sequenced for all these strains four of the SNP loci distinguishing strains T1, Max4, LNPV17.41, and K40 from strains DC3000 and NCPPB1108. The analyzed SNPs are highlighted in the Supplementary Table 3.2. Eleven different genotypes were identified among the 89 analyzed strains based on these SNP loci and SNPs in the housekeeping genes used for the original MLST analysis. Genotype sequences are listed in Supplementary Table 3.3 and genotypes for each strain are listed in Table 3.1. A maximum likelihood tree was then constructed using DC3000 and JL1065 as outgroup (Figure 3.2B). When plotting frequency of the identified genotypes over time (Figure 3.3A), it becomes clear that genotype frequency has changed dramatically since 1961 with different genotypes peaking at different times. Moreover, genetic distance of genotypes appears to be correlated with time since the strains identified in the 1960s and 1970s are more similar

to the DC3000 outgroup than the strains isolated during the last 10 years (Figure 3.3B). This correlation between genetic distance and time was found to be statistically significant for strains collected in Europe, the only continent where strains were consistently sampled between 1961 and 2005. This suggests that genotypes may have evolved from each other. However, the strains from the most basal clade in the tree (Figure 3.2B) have either a 1 bp deletion or a 5 bp deletion in the gene coding for HopM1, a type III effector known to suppress plant immunity during infection of Arabidopsis by strain DC3000 (Badel, Nomura et al. 2003; Badel, Shimizu et al. 2006; Nomura, Debroy et al. 2006). These deletions cause frameshifts leading to truncated open reading frames that are respectively 636 and 1182 bp long compared to the full length *hopM1* gene in strain DC3000, which is 2139 bp long (Fig. 4A). In contrast, T1-like strains on all other branches of the tree have a *hopM1* allele with a nonsense mutation at bp 463 and the *hopM1* allele of strain JL1065 has a 180 bp long in-frame deletion starting at position 1379. Importantly, besides the 1 bp and 5 bp deletions and the premature stop codon all three *hopM1* alleles present in the T1-like strains have 100% DNA identity to each other including the up-stream promoter region and chaperone gene *shchopM1*. Therefore, three independent mutations truncated *hopM1* very recently in T1-like strains and not even one T1-like strain with the ancestral full-length *hopM1* allele is present in our strain collection. This suggests strong selection for loss of full-length HopM1 (see more below). Interestingly, only six strain out of 89 T1-like strains have the deletions causing frameshifts leading to premature stops at codon 212 and 394 while the other 83 T1-like strains have the *hopM1* allele with the early stop at codon

155. These 83 strains thus represent the main T1-lineage that has been causing bacterial speck since 1969, when the first member of this lineage was isolated in Switzerland. To distinguish the strains belonging to this most common T1 lineage from the other T1-like strains we call these strains from now on “T1-proper”.

The world map in Figure 3.3C shows that several genotypes within T1-proper are present in North America and Europe, suggesting that these strains have moved with relatively high frequency between continents, possibly within seed shipments. In fact, transmission of *Pto* via infested tomato seed has been documented (McCarter, Jones et al. 1983). Long distance movement of *Pto* through the atmosphere is also a possibility since *P. syringae* bacteria have been isolated from rain and snow (Morris, Sands et al. 2008). Moreover, as described above, genotypes with increasing genetic distance from the outgroup appear to have replaced one another in North America and Europe. However, members of more ancestral T1 lineages as well as JL1065-like strains have apparently persisted in developing countries in South America, Africa, and Asia (Table 3.1 and Figure 3.3). This suggests only occasional movement of *Pto* strains between Europe and North America on one hand and South America and Africa on the other. Moreover, the strains separated from the *Pto* population in North America and Europe seem to continue to adapt to tomato independently as evidenced by mutations found only in these strains (see also results for *fliC* alleles from strains isolated in Colombia below).

The truncated hopM1 alleles of T1-like strains do not cause cell death

Is it possible that the *hopM1* truncation of T1-proper strains contributed to the worldwide expansion of this lineage? Intriguingly, the full length HopM1 protein of strain DC3000 triggers cell death in several tomato cultivars and wild tomato relatives indicating that it may function as a so-called "avirulence" gene, the product of which is recognized by a plant resistance gene leading to activation of plant defenses including programmed cell death (Wroblewski, Caldwell et al. 2009). However, given that mutating *hopM1* $_{DC3000}$ reduced symptom development during tomato infection and did not increase bacterial population size *in planta*, HopM1$_{DC3000}$ has been considered a virulence factor on tomato (Badel, Nomura et al. 2003; Kvitko, Park et al. 2009). To determine if the truncated *hopM1* alleles that we identified in the T1 and JL1065 lineages lost the ability to trigger cell death in tomato, transient assays expressing all identified *hopM1* alleles directly in tomato leaves using *Agrobacterium*-mediated expression were performed. It was found that the *hopM1*$_{T1}$, *hopM1*$_{PT21}$, *hopM1*$_{NCPPB1108}$, and *hopM1*$_{JL1065}$ alleles do not trigger cell death while *hopM1*$_{DC3000}$ triggers cell death strongly (Figure 3.4B). However, when bacterial growth was compared under lab conditions between T1 and a T1 strain expressing *hopM1*$_{DC3000}$ ectopically, consistent differences were not observed (data not shown). We thus conclude that full-length HopM1 may be recognized by a tomato resistance gene leading to reduced bacterial growth in field conditions. Alternatively, the cell death triggered by *hopM1*$_{DC3000}$ in the *Agrobacterium*-mediated expression assay may not be due to recognition but may be correlated to the known role of *hopM1*$_{DC3000}$ in symptom formation (Badel, Nomura et al. 2003). If so, it is possible that the contribution of *hopM1* to disease symptoms may actually lead to an artificial selection against full

length *hopM1*: seedlings with severe disease symptoms infected with strains that carry the full length *hopM1* allele may be less likely to be sold to farmers for planting than seedlings with mild symptoms or no symptoms at all that are infected with strains that carry a disrupted *hopM1* allele. Thus, a gene like *hopM1* that increases symptom severity may actually render a plant pathogen less fit in an agricultural setting. Regardless, the obvious selection for inactivation of *hopM1* apparent upon analysis of multiple strains shows how characterization of pathogen populations beyond the study of a single model strain can provide new perspectives on the roles of individual virulence factors.

Allelic variation among T1-proper strains in the gene fliC

To assess other factors potentially contributing to the success of the T1-proper strains, two additional effector genes *avrRps4* and *avrPto1*, differing among the five sequenced T1 genomes were analyzed (see Supplementary Table 3.4 for results and Supplementary Table 3.5 for a list of all predicted effectors in the sequenced T1-like genomes). Neither effector was found to be consistently present or absent in T1-proper strains compared to other T1-like strains indicating that these effectors cannot explain the recent expansion of the T1-proper lineage. Nor was there a correlation with presence or absence of the gene cluster for the biosynthesis of the phytotoxin coronatine, which is known to play an important role in the pathogenesis of strain DC3000 on Arabidopsis (Melotto, Underwood et al. 2006), or *avrD1*, a gene specifying the production of defense inducing syringolides (Midland, Keen et al. 1993) (Supplementary Table 3.5). Also extending the search for differences in gene content

beyond known virulence genes did not lead to plausible hypotheses what could have determined the expansion of T1-proper strains compared to all other *Pto* strains. Only 27 gene families, mostly coding for hypothetical proteins or bacteriophage-related proteins, are present in each of the annotated draft genome sequences of the T1-proper strains T1, K40, and LNPV17.41 but absent from strains NCPPB1108, JL1065 and DC3000 (as determined by using the protein repertoire comparison tool at http://genome.ppws.vt.edu/orthologsorter/).

However, it was striking that one of the seven informative SNPs that distinguished LNPV17.41, K40, and Max4 from T1, NCPPB1108, and DC3000 was in the gene *fliC*, resulting in a S99F mutation (Fig. 5A). Intriguingly, the gene *fliC* codes for the flagellum subunit flagellin, well known to contain microbe associated molecular patterns (MAMPs) that trigger an innate immune response in plants and animals (Hayashi, Smith et al. 2001; Zipfel, Robatzek et al. 2004). The S99F mutation was found in a majority of T1-proper strains isolated from tomato after 1990 in North America and Europe (see genotypes IPV-CT28.31 and LNPV17.41 in Figure 3.3). Moreover, of all the mutations analyzed in the 89 *Pto* strains, only this particular SNP was incongruent with other SNPs: the S99F mutation is present in strains KSP53 and KS127M (both of genotype KSP53) from Tanzania, although their genetic background is different from all other strains that carry this mutation. This finding suggests a recombination or parallel evolution event involving *fliC* (which was not detected when sequencing the five T1-like genomes since the genomes of strains 632 and 633 were not completely sequenced) and further supports the idea of strong directional selection on the *fliC* gene.

Surprisingly, we even found two additional *fliC* mutations in T1-proper strains belonging to genotypes Colombia198 and Colombia338 isolated in different regions of Colombia in 2008 and 2009. Both mutations are non-synonymous with one of them (D39I) corresponding to a highly conserved amino acid in the middle of the flg22 peptide (Fig. 5A), a region of the FliC protein recognized by the tomato immune receptor LeFls2 (Robatzek, Bittel et al. 2007). The other mutation (A96V) is only two codons away from the *fliC* mutation described above (S99F). These findings suggest that even successful pathogens may be limited in their growth by the plant immune system and to be under selection pressure to further reduce induction of plant defenses. Moreover, the cluster of two mutations in a region apart from flg22 suggests a second region within flagellin besides flg22 that triggers a plant immune response. In fact, infiltrating 28 amino acid long peptides corresponding to the three alternative alleles of this region (denoted as flgII-28), we observed that the ancestral allele triggered induction of reactive oxygen species (ROS) indicative of a plant defense response while ROS triggered by the two derived alleles was significantly reduced and/or delayed depending on the tomato cultivar tested (Fig. 5B). The same trend was observed between the ancestral and derived flg22 alleles (Fig. 5B) Moreover, infiltration of the ancestral flgII-28 peptide into tomato leaves caused more stomatal closure than infiltration of the derived allele LNPV17.41 (Fig. 5C). Stomata are known to be important points of entry into the leaf apoplast for *Pto* (Melotto, Underwood et al. 2006). In fact, infiltration of tomato leaves with flgII-28 peptides in advance of spraying bacteria on leaf surfaces reduced apoplastic bacterial population sizes 24 hours after inoculation (Fig. 5D). Although the

effect of the three different alleles was not significantly different from each other, the ancestral allele consistently reduced population sizes slightly more than the two derived alleles in each of three independent experiments. Taken together, these finding suggest that the mutations in flgII-28 facilitate leaf invasion making strains that carry these mutations more competitive during this important phase of the pathogen life cycle. ROS were also induced by the ancestral flgII-28 allele in *Nicotiana benthamiana* but none of the flgII-28 alleles triggered ROS in Arabidopsis or bean (data not shown). This indicates that flgII-28 is a MAMP, which may be specifically recognized by Solanaceae species. Whether flgII-28 is recognized by the flg22-receptor LeFL2 (Robatzek, Bittel et al. 2007) or if it is recognized by a different receptor remains to be evaluated.

The almost complete worldwide replacement of strains having the ancestral flgII-28 with strains carrying the derived allele highlights how new pathogen variants can rapidly spread around the world. Therefore, reducing movement of plant pathogens between geographic regions represents an important strategy for avoiding spread of increasingly virulent pandemic strains - even in cases when strains or variants of the same pathogen are already present in these regions. Importantly, our data also reveal that MAMPs are more variable than expected, differing even among strains of the same clonal lineage. This finding questions the recently suggested durability of immunity triggered by other MAMPs (Lacombe, Rougon-Cardoso et al.). However, targeted gene engineering of the *FLS2* receptor gene, and possibly other yet uncharacterized flagellin receptors, may still have potential for strengthening the plant immune response against pathogens with mutated MAMPs.

Conclusion

We have shown how genome sequencing of multiple isolates of a crop pathogen and analysis of a large collection of isolates with genome-derived markers can yield new insights into plant pathogen evolution and molecular plant-pathogen interactions. We found that the typical bacterial speck pathogen of tomato, represented by the T1-proper lineage, is a recently evolved pathogen that rapidly spread around the world, similar to genetically monomorphic human pathogens like *Yersinia pestis* (Morelli, Song et al. 2010), *Bacillus anthracis* (Van Ert, Easterday et al. 2007), or *Salmonella* Typhi (Holt, Parkhill et al. 2008). This suggests that other bacterial plant pathogens may also have adapted to their hosts in recent history, possibly after domestication or - even more recently –after the advent of wide-spread cultivation in mono-culture of their hosts. Investigating microevolution of additional bacterial plant pathogens will make it possible to determine to what point the results obtained here for *Pto* are representative of bacterial plant pathogens in general. Inferring yearly mutation rates and divergence times will be essential for such studies. *P. syringae* pv *aesculi* (Green, Studholme et al. 2010) and *Ralstonia solanacearum* race 3 biovar 2 (Janse 1996) are examples of plant pathogens that have recently spread to a new world region and for which many isolates collected in recent years from different locations are available. Therefore, these pathogen will be excellent candidates for micro-evolutionary and phylogeographic studies.

Our results also highlight the value of assessing diversity in plant pathogen populations as an important complement to the study of model pathogen strains in lab conditions. This approach can lead to new hypotheses as to why some plant pathogens can cause disease and grow to high numbers on a plant species in lab conditions although they are rarely found on the same plant species in the field while other pathogens are successful both under lab conditions and in the field. Answering this question will be essential for gaining a better understanding of pathogen fitness in the field and to finding new avenues for successful control of plant diseases.

Materials and Methods

Bacterial strains, growth and DNA extraction

P. syringae pv. *tomato* strains listed in Table 3.1 were grown in King's Broth (KB) at 28°C and genomic DNA was extracted using the Gentra Puregene Yeast/Bacteria kit (Qiagen) following manufacturer's instructions.

Multilocus Sequence Typing

Fragments corresponding to the MLST loci *rpoD*, *pgi*, and *gapA* were PCR amplified and sequenced as previously described (Yan, Liu et al. 2008).

Genome sequencing

Genomic DNA of strains NCPPB1108, K40, and LNPV 17.41 was sequenced with Illumina technology (Bentley 2006) using the paired-end protocol with read length of

42nt at the University of Toronto Centre for the Analysis of Genome Evolution and Function (CAGEF). Genomic DNA of strain Max4 was also sequenced with Illumina technology but using the single read protocol as previously described for T1 (Almeida, Yan et al. 2009). Genomes of strains NCPPB1108, K40, and LNPV 17.41 were assembled using Velvet 0.7.55 (Zerbino and Birney 2008). Insert size for paired-end reads was set to 200; expected coverage was based on the number of reads used in the assembly and the expected genome size based on strain DC3000; coverage cutoff was set to 4; minimum contig length cut off was set to 100. A range of hash sizes was used to obtain the assembly with the highest N50 value and the lowest number of contigs for each genome. Scaffolding was turned off. Genomes were annotated using GRC (Warren and Setubal 2009).

SNP identification

SNPs between *Pto* strain T1 (Almeida, Yan et al. 2009) and the other four T1-like strains NCPPB1108, Max4, K40, and LNPV17.41 were identified by aligning Illumina sequence reads of T1, Max4, K40, NCPPB1108, and LNPV 17.41 against the DC3000 genome (Buell, Joardar et al. 2003) in MAQ (Li, Ruan et al. 2008). We only considered the 3,024,986 nucleotides in the DC3000 genome for which there was at least 20X depth of coverage by Illumina reads from each of the five Illumina datasets (i.e. T1, LNPV 17.41, K40, Max4, NCPPB1108) and for which there was at least 95% consensus between the aligned reads. The polymorphism states of the remaining 3,372,140 nt of the DC3000 chromosome were considered to be ambiguous and we made no attempt to detect SNPs there. We considered a SNP to be present at a given site if at least 95%

of the aligned reads at that site consistently called a different nucleotide from that in the reference sequence. We compared the position of each SNP against the positions of the predicted genes as specified in RefSeq:NC_004578 to determine whether it was intergenic or intragenic. For intragenic SNPs, we translated the open reading frame containing the SNP to check whether the SNP would result in a different amino acid sequence (i.e. whether it was a non-synonymous mutation). The process was automated using custom Perl scripts. SNPs that were not informative to distinguish T1-like strains from each other were not considered, i.e., all SNPs that distinguished DC3000 from the T1-like strains but that had the same nucleotide in all five T1-like strains. Only the SNP loci that distinguished T1-like strains from each other are shown in Supplementary Table 3.1 and were used for construction of the whole genome tree shown in Figure 3.2A (see below for details).

In a second independent search for SNPs between Pto strains T1, Max4, K40, NCPPB1108, and LNPV 17.41, Illumina sequence reads of the newly sequenced strains were aligned against the T1 draft genome using MAQ (Li, Ruan et al. 2008) using default parameters. The MAQ output was then parsed using a custom script eliminating all SNP calls that did not have the consensus A, C, G or T. A final list of core genome SNPs (Supplementary Table 3.2) was then assembled limiting SNPs to SNPs present in genes that were found to be present exactly one time in the *P. syringae* genomes T1 (Almeida, Yan et al. 2009), DC3000 (Buell, Joardar et al. 2003), B728a (Feil, Feil et al. 2005), and 1448A (Joardar, Lindeberg et al. 2005) using OrthoMCL (Li, Stoeckert et al. 2003). The total length of these genes is 3,543,009 nt.

Construction of whole genome trees

Based on silent, non-silent, intergenic, and intragenic sites, we constructed 5 bootstrapped (2000 replicates) Maximum Likelihood trees for the genomes of strains T1, Max4, LNPN17.41, K40 and NCPPB1108 using the genome of strain DC3000 as outgroup. The first four trees were based on each of the data features separately, and the remaining tree was based on the collection of all data features jointly, to which we refer to as the whole genome tree. Trees were constructed in PAUP version 4.0 (http://paup.csit.fsu.edu/) using parameters determined by jMODELTEST (Guindon and Gascuel 2003; Posada 2008). Non-silent, intragenic, and the whole genome data satisfied the GTR substitution model (Bos and Posada 2005); whereas, silent and intergentic data best fit the GTR+I and SYM models (Bos and Posada 2005), respectively. A Maximum parsimony tree was built using DNAPARS of the PHYLIP 3.69 package (http://evolution.gs.washington.edu/phylip.html).

SNP analysis

Primers were designed upstream and downstream of each of the seven SNPs that distinguished strains LNPV 17.41, K40, and Max4 from NCPPB1108 and T1. Four primer pairs were designed for additional five SNPs (two of them adjacent to each other) that distinguished LNPV 17.41, K40, Max4, and T1 from NCPPB1108 and DC3000. The 12 SNPs are highlighted in green in Supplementary Table 3.2 and primers are listed in Supplementary Table 3.7.

Construction of SNP tree

Based on the SNPs listed in Supplementary Table 3.3, 11 genotypes were identified among the T1-like strains listed in Table 3.1. Supplementary Table 3.4 lists the SNP genotype for each strain. jMODELTEST (Guindon and Gascuel 2003; Posada 2008) was used to determine the substitution model that best fit the data (SYM). A maximum likelihood tree was then built in PAUP version 4.0 (http://paup.csit.fsu.edu/). Bootstrap analysis was performed with 5000 replicates. A Maximum parsimony tree was built using DNAPARS of the PHYLIP 3.69 package (http://evolution.gs.washington.edu/phylip.html).

Molecular Evolutionary analysis

Based on a 10-year sliding window, we calculated the relative frequencies of T1-, JL1065- and DC3000-like strains, for the time period 1942-2009. Additionally, for the years 1961-2009, T1-like strains acquired across North America and Europe according to genotypes were also analyzed based on a 10-year sliding window. Each T1-like strain was uniquely classified based on a profile of 40 SNPs. Eight genotypes of T1-like strains were observed in North America and Europe. Frequency plots were generated for these genotypes using the statistical software language R (http://www.r-project.org/).

Genetic distances for all T1-like strains were calculated as compared to the DC3000 strain, under the Jukes-Cantor model. In order to investigate the relationship between these relative genetic distances and isolation year, we fit the regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where y_i is the relative genetic distance, x_i is the isolation year, and ε_i denotes independent normally distributed error. Values of β_1 which are distinguishable from zero indicate a linear temporal relationship between genetic distance (y_i) and time (x_i).

Estimation of divergence times

In order to estimate divergence times for the five sequenced T1-like stains (Max4, LNPV17.41, K40, T1 and NCPPB1108), we used IMa2 (Hey and Nielsen 2007; Hey 2010) and BEAST 1.6.1 (Drummond and Rambaut 2007). In both programs, we computed our estimates based on the nucleotides present at the concatenated SNP loci listed in Supplementary Table 3.1 and setting the mutational clock rate (μ) to 1.

IMa2 (Hey and Nielsen 2007; Hey 2010) was run in Markov Chain Monte Carlo (MCMC) mode. We considered our five strains to be derived from five populations, and assumed no migration in the model. The mutation model used for this analysis is the Hasegawa-Kishino-Yano (HKY) model. Prior distributions were selected as uniform distributions between zero and some upper bound. Upper bounds were chosen to be far removed from the maximum likelihood estimate: 300 for $t \times \mu$, and 200 for effective population size parameters. In order to reduce auto correlations in our MCMC samples, 20 million iterations were run, with samples stored 10,000 iterations after a 'burn-in' period of 2 million generations. Multiple runs of the algorithm produced nearly identical results.

In BEAST 1.6.1 (Drummond and Rambaut 2007), prior distributions were selected

as lognormal with units in % per million years. GTR was selected as substitution model. Since BEAST results are on a percent scale, results were converted to million years in order to compare to IMa2 results.

To rescale program outputs to an estimated clock rate and to the length of the genome used for SNP discovery, we used:

$$DT = \frac{t \times \mu \times L}{\hat{\mu} \times \kappa},$$

where DT is the rescaled divergence time in years; t is the estimated splitting time obtained from IMa2 or BEAST converted to years; $\hat{\mu}$ is the mutation rate per base pair (bp) per year; L is the length of SNPs used as input, which is 157 bp; and κ is the total length of the genome used for SNP discovery, which is 3,024,986 bp.

Effector prediction

Pseudomolecules were created from the draft genome sequences by concatenating contigs in the order from largest to smallest with the TIGR linker sequence "nnnnnttaattaattaannnnn" delimiting contig boundaries. Effectors were identified in the pseudomolecules using a combination of automated annotation generated by RAST (http://rast.nmpdr.org/), alignment of pseudomolecules with the DC3000 sequence visualized using the Artemis Comparison Tool, HrpL binding sites predicted as previously described (Ferreira, Myers et al. 2006), and PSI-BLAST of confirmed effector

sequences against the pseudomolecule sequences. Predicted effectors are listed in Supplementary Table 3.6.

HopM1 cloning and transient expression

The open reading frames including the ribosome binding site but not the stop codon of *hopM1* alleles were amplified by PCR from genomic DNA of *Pto* strains DC3000, JL1065, T1, NCPPB1108, and PT21 with the primer pairs listed in Supplementary Table 3.7 and with nested primers to add sequences for Gateway™ (Invitrogen) cloning using the protocol described previously (Vinatzer, Teitzel et al. 2006). The five PCR products were then cloned into the entry vector pDNOR207 (Invitrogen) using the Gateway™ BP cloning kit (Invitrogen). Recombined plasmids were confirmed by sequencing and cloned into the destination vector pBAV150 (Vinatzer, Teitzel et al. 2006) using the Gateway™ LR cloning kit (Invitrogen). *hopM1*-containing pBAV150 were mated from *Escherichia coli* into *Agrobacterium tumefaciens* C58C1 and used in transient assays of tomato leaves (at a concentration corresponding to an optical density at 600nm of 0.04) and in *Nicotiana benthamiana* leaves (corresponding to an optical density at 600nm of 0.4) using the same protocol as described previously for *Nicotiana benthamiana* (Vinatzer, Teitzel et al. 2006). Western blots were performed as described in (Vinatzer, Teitzel et al. 2006) also.

Characterization of MAMP-triggered immunity

Peptides corresponding to alleles of flg22 and flgII-28 were ordered from EZBiolab with >70% purity (see Figure 3.5 for peptide sequences). Peptides were resuspended in

sterile water and used to measure induction of reactive oxygen species (ROS) in the tomato cultivar Chico III. A luminol - horseradish peroxidase assay was used to quantify ROS induction as described by Chakravarthy and colleagues (Chakravarthy, Velasquez et al. 2010) with small modifications: 4-mm leaf discs were punched out with a cork borer and floated adaxial side up in 200μl ddH2O over night at room temperature in wells of a 96-well solid white plate. The ddH2O was then replaced with 100μl of ROS testing buffer containing 1uM of flg22 or flgII-28 peptide, 34μg/ml of luminol (Sigma), and 20μg of horseradish peroxidase (VI-A, Sigma). Luminescence was measured using a Bioteck, Synergy HT plate reader. Five leaf disks treated with the same peptide were tested in parallel. Leaf discs in testing buffer without addition of any flagellin peptide were used as a negative controls.

Analysis of stomatal closure after leaf infiltration with MAMPs

Leaves were treated with flg22 and flgII-28 peptides as described by Melotto and co-workers (Melotto, Underwood et al. 2006) with slight modifications. Briefly, 4 week-old tomato plants were sprayed with water, placed in transparent plastic bags, and transferred to a 28°C incubator exposed to light to induce stomatal opening. Whole leaves were detached from plants and placed on a glass slide. The leaves were immersed in 5uM of flagellin peptide dissolved in ddH20, or just ddH2O for mock treatment, and then covered with a cover slip. The mounted leaves were placed at room temperature for 2 hours and then viewed at 200x magnification using an Axio Imager M1 upright microscope (Zeiss). Pictures of stomata were taken using an

Axiocam MRm camera (Zeiss). Stomatal aperture of 20 stomata per test group per experiment were quantified using Axiovision v. 4.7.2 (Zeiss).

Leaf invasion assay

Leaves of 5-week-old tomato plants (cv. 'Chico III') were infiltrated with flg peptides at a 1 μM concentration via a blunt end syringe while still attached to the plant. Plants were placed in a high humidity container for 24 hours. Strain NCPPB1108 was then sprayed onto leaves at a concentration corresponding to an optical density at 600nm of 0.01 in 10mM $MgSO_4$ using a Preval® sprayer canister and placed back in the high humidity container. Bacterial invasion was assessed 24 hours after infection. 0.52mm sections were punched out of the infiltrated leaves and placed in a tube with 200μL 1% bleach with the leaf punch completely submerged. The tube was mildly vortexed for 5 seconds to remove epiphytic bacteria. The leaf punch was then removed from the 1% bleach solution, gently rinsed in ddH_2O, and then placed in a separate tube containing 200μL 10mM $MgSO_4$ and three 2mm glass beads. The tube was placed in a mini bead beater (Biospec Products, Inc.) and shaken for 90 seconds to grind the leaf and release endophytic bacteria into the solution. Colony forming units were counted after dilution plating.

Acknowledgements

We thank Sonia Gutierrez and Jorge Lopez for help with isolation of *P. syringae* from tomato in Colombia, DNA extraction, and PCR analysis.

Accession numbers

HQ992994 – *hopM1* operon of strain T1

HQ992995 – *hopM1* operon of strain NCPPB1108

HQ992993 – *hopM1* operon of strain PT21

JF268671 - *hopM1* operon of strain JL1065

JF261012 – *fliC* allele of strain K40

JF261011 – *fliC* allele of strain Col198

JF261013 – *fliC* allele of strain Col338

Supplemental Data

Supplemental Data consist in six tables.

References

Achtman, M. (2008). "Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens." Annu Rev Microbiol **62**: 53-70.

Almeida, N. F., S. Yan, et al. (2009). "A Draft Genome Sequence of Pseudomonas syringae pv. tomato T1 Reveals a Type III Effector Repertoire Significantly

Divergent from That of Pseudomonas syringae pv. tomato DC3000." Mol Plant Microbe Interact **22**(1): 52-62.

Badel, J. L., K. Nomura, et al. (2003). "Pseudomonas syringae pv. tomato DC3000 HopPtoM (CEL ORF3) is important for lesion formation but not growth in tomato and is secreted and translocated by the Hrp type III secretion system in a chaperone-dependent manner." Mol Microbiol **49**(5): 1239-1251.

Badel, J. L., R. Shimizu, et al. (2006). "A Pseudomonas syringae pv. tomato avrE1/hopM1 mutant is severely reduced in growth and lesion formation in tomato." Mol Plant Microbe Interact **19**(2): 99-111.

Bender, C. L. and D. A. Cooksey (1986). "Indigenous plasmids in Pseudomonas syringae pv. tomato: conjugative transfer and role in copper resistance." J Bacteriol **165**: 534-541.

Bentley, D. R. (2006). "Whole-genome re-sequencing." Curr Opin Genet Dev **16**(6): 545-552.

Bentley, S. D., C. Corton, et al. (2008). "Genome of the Actinomycete Plant Pathogen Clavibacter michiganensis subsp. sepedonicus Suggests Recent Niche Adaptation." J. Bacteriol. **190**(6): 2150-2160.

Bos, D. H. and D. Posada (2005). "Using models of nucleotide evolution to build phylogenetic trees." Developmental and Comparative Immunology **29**(2005): 211-227.

Buell, C. R., V. Joardar, 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**(18): 10181-10186.

Chakravarthy, S., A. C. Velasquez, et al. (2010). "Identification of Nicotiana benthamiana genes involved in pathogen-associated molecular pattern-triggered immunity." Mol Plant Microbe Interact **23**(6): 715-726.

Charity, J. C., K. Pak, et al. (2003). "Novel exchangeable effector loci associated with the Pseudomonas syringae hrp pathogenicity island: evidence for integron-like assembly from transposed gene cassettes." Mol Plant Microbe Interact **16**(6): 495-507.

Cuppels, D. A. and T. Ainsworth (1995). "Molecular and Physiological Characterization of Pseudomonas syringae pv. tomato and Pseudomonas syringae pv. maculicola Strains That Produce the Phytotoxin Coronatine." Appl. Environ. Microbiol. **61**(10): 3530-3536.

Cuppels, D. A., R. A. Moore, et al. (1990). "Construction and use of a nonradioactive DNA hybridization probe for detection of *Pseudomonas syringae* pv. *tomato* on tomato plants." Appl Environ Microbiol **56**: 1743-1749.

Denny, T. P. (1988). "Phenotypic diversity in *Pseudomonas syringae* pv. *tomato*." J Gen Microbiol **134**: 1939-1948.

Drummond, A. and A. Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees." BMC Evolutionary Biology **7**(1): 214.

Feil, H., W. S. Feil, 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**(31): 11064-11069.

Ferreira, A. O., C. R. Myers, 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**(11): 1167-1179.

Green, S., D. J. Studholme, et al. (2010). "Comparative Genome Analysis Provides Insights into the Evolution and Adaptation of <italic>Pseudomonas syringae</italic> pv. <italic>aesculi</italic> on <italic>Aesculus hippocastanum</italic>." PLoS ONE **5**(4): e10224.

Guindon, S. p. and O. Gascuel (2003). "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood." Systematic Biology **52**(5): 696-704.

Harris, S. R., E. J. Feil, et al. (2010). "Evolution of MRSA During Hospital Transmission and Intercontinental Spread." Science **327**(5964): 469-474.

Hayashi, F., K. D. Smith, et al. (2001). "The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5." Nature **410**(6832): 1099-1103.

He, M., M. Sebaihia, et al. (2010). "Evolutionary dynamics of Clostridium difficile over short and long time scales." Proceedings of the National Academy of Sciences **107**(16): 7527-7532.

Hey, J. (2010). "Isolation with Migration Models for More Than Two Populations." Molecular Biology and Evolution **27**(4): 905-920.

Hey, J. and R. Nielsen (2007). "Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics." Proceedings of the National Academy of Sciences **104**(8): 2785-2790.

Holt, K. E., J. Parkhill, et al. (2008). "High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi." Nat Genet **40**(8): 987-993.

Janse, J. D. (1996). "Potato brown rot in western Europe - history, present occurence and some remarks on possible origin." EPPO Bulletin(26): 17.

Joardar, V., M. Lindeberg, 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**(18): 6488-6498.

Kunkeaw, S., S. Tan, et al. (2010). "Molecular and evolutionary analyses of Pseudomonas syringae pv. tomato race 1." Mol Plant Microbe Interact **23**(4): 415-424.

Kvitko, B. H., D. H. Park, et al. (2009). "Deletions in the repertoire of Pseudomonas syringae pv. tomato DC3000 type III secretion effector genes reveal functional overlap among effectors." PLoS Pathog **5**(4): e1000388.

Lacombe, S., A. Rougon-Cardoso, et al. "Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance." Nat Biotechnol **28**(4): 365-369.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res.

Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-2189.

Manning, S. D., A. S. Motiwala, et al. (2008). "Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks." Proc Natl Acad Sci U S A **105**(12): 4868-4873.

McCarter, S. M., J. B. Jones, et al. (1983). "Survival of Pseudomonas syringae pv. tomato in Association with Tomato Seeds, Soil, Host Tissue, and Epiphytic Weed Hosts in Georgia." Phytopathology **73**(10): 1393-1398.

Melotto, M., W. Underwood, et al. (2006). "Plant stomata function in innate immunity against bacterial invasion." Cell **126**(5): 969-980.

Midland, S. L., N. T. Keen, et al. (1993). "The structures of syringolides 1 and 2, novel C-glycosidic elicitors from *Pseudomonas syringae* pv. *tomato*." Journal of Organic Chemistry **58**: 2940-2945.

Mitchell, R. E., C. N. Hale, et al. (1983). "Production of different pathogenic symptoms and different toxins by strains of *Pseudomonas syringae* pv. *tomato* not distinguishable by gel-immunodiffusion assay." Physiological and Molecular Plant Pathology **23**: 315-322.

Morelli, G., X. Didelot, et al. (2010). "Microevolution of <italic>Helicobacter pylori</italic> during Prolonged Infection of Single Hosts and within Families." PLoS Genet **6**(7): e1001036.

Morelli, G., Y. Song, et al. (2010). "Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity." Nat Genet **42**(12): 1140-1143.

Morris, C. E., D. C. Sands, et al. (2008). "The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle." ISME J **2**(3): 321-334.

N√obel, U., J. Dordel, et al. (2010). "A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant <italic>Staphylococcus aureus</italic>." PLoS Pathog **6**(4): e1000855.

Nomura, K., S. Debroy, et al. (2006). "A bacterial virulence protein suppresses host innate immunity to cause plant disease." Science **313**(5784): 220-223.

Ochman, H. and A. C. Wilson (1987). "Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes." Journal of Molecular Evolution **26**: 74-86.

Pearson, T., R. T. Okinaka, et al. (2009). "Phylogenetic understanding of clonal populations in an era of whole genome sequencing." Infect Genet Evol **9**(5): 1010-1019.

Peralta, I. E. and D. M. Spooner (2007). History, origin, and early cultivation of tomato (Solanaceae). Genetic improvement of solanaceous crops: tomato. M. K. Razdan and A. K. Mattoo. Enfield (NH), Science Publishers. **2:** 1-27.

Pernezny, K., V. Kudela, et al. (1995). "Bacterial diseases of tomato in the Czech and Slovak Republics and lack of streptomycin resistance among copper-tollerant bacterial strains." Crop Prot. **14**: 267-270.

Posada, D. (2008). "jModelTest: Phylogenetic Model Averaging." Molecular Biology and Evolution **25**(7): 1253-1256.

Robatzek, S., P. Bittel, et al. (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**(5): 539-547.

Shenge, K. C., R. B. Mabagala, et al. (2007). "First Report of Bacterial Speck of Tomato Caused by Pseudomonas syringae pv. tomato in tanzania." Plant Disease Note **91**: 462.

Van Ert, M. N., W. R. Easterday, et al. (2007). "Global genetic population structure of Bacillus anthracis." PLoS ONE **2**(5): e461.

Vinatzer, B. A., G. M. Teitzel, et al. (2006). "The type III effector repertoire of Pseudomonas syringae pv. syringae B728a and its role in survival and disease on host and non-host plants." Mol Microbiol **62**(1): 26-44.

Warren, A. S. and J. C. Setubal (2009). "The Genome Reverse Compiler: an explorative annotation tool." BMC Bioinformatics **10**: 35.

Whalen, M. C., R. W. Innes, et al. (1991). "Identification of Pseudomonas syringae pathogens of Arabidopsis and a bacterial locus determining avirulence on both Arabidopsis and soybean." Plant Cell **3**(1): 49-59.

Wroblewski, T., K. S. Caldwell, et al. (2009). "Comparative large-scale analysis of interactions between several crop species and the effector repertoires from multiple pathovars of pseudomonas and ralstonia." Plant Physiol **150**(4): 1733-1749.

Yan, S., H. Liu, et al. (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**(10): 3171-3181.

Zaccardelli, M., A. Spasiano, et al. (2005). "Identification and in planta detection of Pseudomonas syringae pv. tomato using PCR amplification of hrpZ." European Journal of Plant Pathology **111**: 85-90.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.

Zipfel, C., S. Robatzek, et al. (2004). "Bacterial disease resistance in Arabidopsis through flagellin perception." Nature **428**(6984): 764-767.

Figure legends

Figure 3.1. Strains of the T1-lineage have been the most common *Pto* strains since the 1960s and are present in all continents from which *Pto* strains were isolated.

(A) The lines indicate the frequency of T1-, JL1065-, and DC3000-like strains over time using a 10-year sliding window with a one-year step. Circles represent individual isolates and are placed in the graph in correspondence to the exact year at which isolates were collected. Full circles indicate isolates of which the genomes have been sequenced.

(B) World map with pie charts showing ratio of T1-, JL1065-, and DC3000-like strains for the continents from which *Pto* strains have been analyzed. Pie size is proportional to the total number of strains considered per continent.

Figure 3.2 Phylogenetic trees based on SNPs reveal the evolutionary relationship between T1-like *Pto* strains.

(A) Maximum likelihood tree based on 157 high quality SNPs identified between five genomes of T1-like strains by aligning sequencing reads against the DC3000 genome, which was used as an outgroup. The number of SNPs/branch are indicated underneath each branch and bootstrap values are indicated above each branch. A neighbor-joining tree and maximum parsimony tree were also constructed and had identical topology.

(B) Maximum likelihood tree based on twenty-four SNPs identified between DC3000-like, JL1065, and T1-like strains in the housekeeping genes *rpoD*, *pgi*, and *gapA* and based on 16 SNPs identified between T1-like strains in 11 fragments of *P. syringae* core genome genes (highlighted in Supplementary Table 3.2). Bootstrap values are indicated above each branch and number of strains that belong to each genotype are indicated in parenthesis. Clade-specific *fliC* and *hopM1* alleles are indicated below branches. The clade corresponding to strains called "T1-proper" is labeled as such. A maximum parsimony tree was also constructed and had identical topology. Since branch lengths of the tree are influenced by our selection of SNP loci, branch lengths are not scaled to evolutionary changes. Table 3.1 lists strains belonging to each genotype and Supplementary Table 3.3 lists DNA sequences of each genotype.

Figure 3.3. T1 genotypes change in frequency over time and genetic distances from the outgroup strain DC3000 increase over time. Several genotypes are present in both North America and Europe.

(A) The lines indicate the frequency of T1 genotypes over time using a 10-year sliding window with a one-year step. Circles represent individual isolates and are placed in the graph in correspondence to the exact year at which isolates were collected. Full circles indicate those isolates for which genomes have been sequenced.

(B) Genetic distance of strains from the out-group strain DC3000 plotted over time. Genetic distance was calculated based on the 24 MLST SNPs and the 16 genome

SNPs that were analyzed in all strains. When more than one isolate with the same genotype was collected during the same year, the total number of isolates is indicated next to the genotype symbol.

(C) World map with pie charts showing ratio of T1 genotypes for the continents from which T1-like strains have been analyzed. Pie size is proportional to the total number of strains considered per continent.

Figure 3.4. The *hopM1* gene is disrupted in all T1-like and JL1065-like strains. The encoded truncated proteins do not trigger cell death in tomato while the full-length protein encoded by the DC3000 *hopM1* gene does.

(A) Graphical presentation of *Pto hopM1* alleles. The stars indicate the position of deletions causing frameshifts in the PT21 and NCPPB1108 alleles. The PT21 allele is present in four strains of SNP genotype NCPPB1108 and in the only strain with SNP genotype CA315 while the NCPPB1108 allele is only present in strain NCPPB1108 (SNP genotype NCPPB1108). The T1 allele is present in all other T1-like strains, which are referred to as T1-proper in the text.

(B) Agrobacterium-mediated transient expression of *hopM1* alleles fused to *gfp* in the tomato cultivar “Chico III”. Only the $hopM1_{DC3000}$ allele triggered cell death. Similar results were obtained on the tomato cultivars “Rio Grande” and “Sunpride” in at least two independent experiments/cultivar. Leaf areas infiltrated with *Agrobacterium*

tumefaciens strains are traced in black. Strain names indicate which *hopM1::gfp* fusion construct was expressed in which leaf area. Agro EV: Agrobacterium carrying an empty vector control, T1-HA: in this leaf area the hopM1$_{T1}$ allele was expressed with an HA tag, CD: cell death.

(C) Western Blot analysis with GFP antibody of HopM1::GFP fusion proteins from extracts of *Nicotiana benthamiana* leaf disks infiltrated with the same *Agrobacterium tumefaciens* strains used in panel B. * indicate the bands of the expected size based on the sequence of the *hopM1* alleles in panel A. The Rubisco large subunit band from the Coomassie-stained gel is shown as loading control underneath the Western Blot.

Figure 3.5. The flagellin epitope flgII-28 triggers reactive oxygen species (ROS) in tomato leaves whereby derived alleles - typical of today's *Pto* strains - induce less ROS than the ancestral alleles - typical of strains isolated before 1985. Alleles of flgII-28 also induce stomatal closure and interfere with leaf invasion.

(A) Amino acid sequences of flg22 and flgII-28 alleles. The T1 alleles are identical to the DC3000 alleles and thus represent the ancestral states. The derived alleles are named after one of the genotypes in which they are present.

(B) Induction of reactive oxygen species (ROS) in tomato leaf disks of cultivar 'Chico III' after incubation with flg22 and flgII-28 peptides at a 1 μM concentration. ROS induction was significantly different at the 2 minutes time point in an unpaired Student's t-test at

the 0.05 level between $flg22_{T1}$ and $flg22_{Colombia338}$ and between $flgII\text{-}28_{T1}$ on one hand and $flgII\text{-}28_{LNPV17.41}$ and $flgII\text{-}28_{Colombia198}$ on the other. $flgII\text{-}28_{T1}$ and $flgII\text{-}28_{Colombia198}$ were also significantly different from each other at the 5 minutes time point. Similar results were obtained with three different tomato cultivars whereby experiments on each cultivar were repeated at least twice.

(C) Stomatal closure induced in tomato leaves of cultivar 'Chico III' after infiltration with flg22 and flgII-28 peptides at a 5 μM concentration or mock infiltration with sterile water. Similar results were obtained in three independent experiments. Different letters indicate significance at the 0.05 level in an unpaired Student's t-test.

(D) Leaves of tomato cultivar 'Chico III" were infiltrated with flg22 and flgII-28 peptides at a 1 μM concentration. Strain NCPPB1108 was then sprayed on leaf surfaces 24 hours later and apoplastic population sizes were measured another 24 hours later. Different letters indicate significance at the 0.05 level in an unpaired Student's t-test.

Tables

Table 3.1. *Pto* isolates used in this study sorted first by MLST genotype (GT) and then by year of isolation.

name	Country	Year	MLST GT	SNP GT[1]	HopM1 allele	obtained from	reference
ICMP 4325	Canada	1944	DC3000	-	DC3000	C. Bender, Oklahoma State U., USA	(Mitchell, Hale et al. 1983)
DC3000	UK	1961	DC3000	-	DC3000	J. Greenberg, U. of Chicago, USA	(Buell, Joardar et al. 2003)
NCPPB 1008	USA	1942	JL1065	-	JL1065	C. Bender, Oklahoma State U., USA	(Cuppels, Moore et al. 1990)
CFBP 1696	Denmark	1949	JL1065	-	JL1065	CFBP, France	this paper
NCPPB 880	Yugoslavia	1953	JL1065	-	JL1065	C. Bender, Oklahoma State U., USA	(Denny 1988)
ICMP 2846	USA	1956	JL1065	-	JL1065	C. Bender, Oklahoma State U., USA	(Mitchell, Hale et al. 1983)
CFBP 1319	Switzerland	1970	JL1065	-	JL1065	CFBP, France	this paper
CFBP 1785	Australia	1972	JL1065	-	JL1065	CFBP, France	this paper
ICMP 3647	Australia	1973	JL1065	-	JL1065	C. Bender, Oklahoma State U., USA	(Whalen, Innes et al. 1991)
ICMP 4355	Australia	1975	JL1065	-	JL1065	C. Bender, Oklahoma State U., USA	(Charity, Pak et al. 2003)
JL1065	USA	1983	JL1065	-	JL1065	R. Jackson, U. Reading, UK	(Whalen, Innes et al. 1991)
BS118	USA	1983	JL1065	-	JL1065	C. Bull, USDA ARS, Salinas, USA	this paper
BS120	USA	1983	JL1065	-	JL1065	C. Bull, USDA ARS, Salinas, USA	this paper
DC84-1	Canada	1984	JL1065	-	JL1065	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
PST26L	S. Africa	1986	JL1065	-	JL1065	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)

CFBP 3728	Yemen	1988	JL1065	-	JL1065	CFBP, France	this paper
PT 28	Mexico	1992	JL1065	-	JL1065	J. Jones, U. of Florida, USA	this paper
PT 29	Mexico	1992	JL1065	-	JL1065	J. Jones, U. of Florida, USA	this paper
CPST 147	Czek Rep.	1993	JL1065	-	JL1065	C. Bender, Oklahoma State U., USA	(Pernezny, Kudela et al. 1995)
56	USA	1995	JL1065	-	JL1065	G. Coaker, UC Davis, USA	this paper
Pst field 8	USA	1999	JL1065	-	JL1065	A. Bernal, U. de los Andes, Colombia	this paper
KS 112 lr	Tanzania	2004	JL1065	-	JL1065	M. Zaccardelli, CRA ORT, Italy	(Shenge, Mabagala et al. 2007)
KS 097 lr	Tanzania	2004	JL1065	-	JL1065	M. Zaccardelli, CRA ORT, Italy	(Shenge, Mabagala et al. 2007)
NCPPB 1108	UK	1961	T1	NCPPB1108	1108	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
CNBP 1318	Switzerland	1969	T1	CFBP1318	T1	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
NCPPB 2424	Switzerland	1969	T1	CFBP1318	T1	C. Bender, Oklahoma State U., USA	(Denny 1988)
CFBP 1321	Switzerland	1970	T1	CFBP1318	T1	CFBP, France	this paper
CFBP 1322	Switzerland	1970	T1	CFBP1318	T1	CFBP, France	this paper
CFBP 1323	France	1971	T1	NCPPB1108	PT21	CFBP, France	(Denny 1988)
CFBP 1426	France	1972	T1	CFBP1318	T1	CFBP, France	this paper
CFBP 1427	France	1972	T1	CFBP1318	T1	CFBP, France	this paper
DAR 31861	Australia	1975	T1	NCPPB1108	PT21	C. Bender, Oklahoma State U., USA	(Denny 1988)
PT 14	USA	1978	T1	PT14	T1	J. Jones, U. of Florida, USA	this paper
SM78-1	USA	1978	T1	T1	T1	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
DAR 30555	Australia	1978	T1	PT14	T1	C. Bender, Oklahoma State U., USA	(Denny 1988)

CFBP 1916	Canada	1978	T1	PT14	T1	CFBP, France	this paper
CFBP 1918	Canada	1978	T1	PT14	T1	CFBP, France	this paper
CFBP 2545	France	1978	T1	CFBP2545	T1	CFBP, France	this paper
487	Greece	1979	T1	CFBP1318	T1	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
CFBP 6876	France	1979	T1	CFBP2545	T1	CFBP, France	this paper
PST 6	Canada	1980	T1	PT14	T1	T. Denny U. of Georgia, USA	this paper
PT 18	USA	1980	T1	PT14	T1	T. Denny U. of Georgia, USA	this paper
AV80	USA	1980	T1	T1	T1	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
B181	USA	1981	T1	PT14	T1	T. Denny U. of Georgia, USA	(Denny 1988)
DCT6D1	Canada	1981	T1	PT14	T1	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
188B	Canada	1982	T1	T1	T1	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
BS117	USA	1982	T1	PT14	T1	C. Bull, USDA ARS, Salinas, USA	this paper
PT 17	USA	1983	T1	T1	T1	T. Denny U. of Georgia, USA	this paper
PT 2	USA	1983	T1	PT14	T1	T. Denny U. of Georgia, USA	this paper
CFBP 4408	France	1984	T1	CFBP1318	T1	CFBP, France	this paper
RG4	Venezuela	1985	T1	CFBP1318	T1	C. Bender, Oklahoma State U., USA	(Denny 1988)
T1	Canada	1986	T1	T1	T1	T. Denny U. of Georgia, USA	(Almeida, Yan et al. 2009)
CFBP 4409	France	1987	T1	CFBP1318	T1	CFBP, France	this paper
DC89-4H	Canada	1989	T1	PT14	T1	D. Cuppels, Agrifood Canada	(Cuppels and Ainsworth 1995)
PT 21	USA	1990	T1	NCPPB1108	PT21	J. Jones, U. of Florida, USA	this paper
PT 23	USA	1990	T1	LNPV17.41	T1	J. Jones, U. of Florida, USA	(Bender and Cooksey 1986)

PT 25	USA	1990	T1	LNPV17.41	T1	J. Jones, U. of Florida, USA	this paper
PT 26	USA	1990	T1	NCPPB1108	PT21	J. Jones, U. of Florida, USA	this paper
OMP-BO 407/91	Italy	1991	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	(Zaccardelli, Spasiano et al. 2005)
PT 32	USA	1993	T1	LNPV17.41	T1	J. Jones, U. of Florida, USA	this paper
CPST 236	Slovakia	1993	T1	PT14	T1	C. Bender, Oklahoma State U., USA	(Pernezny, Kudela et al. 1995)
IPV-CT 28.31	Italy	1995	T1	IPV-CT28.31	T1	M. Zaccardelli, CRA ORT, Italy	this paper
IPV-BO 2973	Italy	1996	T1	PT14	T1	M. Zaccardelli, CRA ORT, Italy	(Zaccardelli, Spasiano et al. 2005)
LNPV 17.41	France	1996	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	this paper
OMP-BO 443.1/96	Italy	1996	T1	PT14	T1	M. Zaccardelli, CRA ORT, Italy	(Zaccardelli, Spasiano et al. 2005)
A9	USA	1996	T1	LNPV17.41	T1	M. Davis, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
CFBP 5420	Macedonia	1996	T1	LNPV17.41	T1	CFBP, France	this paper
407	USA	1997	T1	LNPV17.41	T1	M. Davis, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
LNPV 18.76	France	1998	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	this paper
838-1	USA	1998	T1	LNPV17.41	T1	M. Davis, UC Davis, USA	this paper
315	USA	1998	T1	CA315	PT21	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
316	USA	1998	T1	LNPV17.41		G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
Pst field 1	USA	1999	T1	LNPV17.41	T1	A. Bernal, U. de los Andes, Colombia	this paper
Pst field 2	USA	1999	T1	LNPV17.41	T1	A. Bernal, U. de los Andes, Colombia	this paper
Pst field 3	USA	1999	T1	LNPV17.41	T1	A. Bernal, U. de los Andes, Colombia	this paper
Pst field 4	USA	1999	T1	LNPV17.41	T1	A. Bernal, U. de los Andes, Colombia	this paper

Pst field 5	USA	1999	T1	LNPV17.41	T1	A. Bernal, U. de los Andes, Colombia	this paper
Pst field 6	USA	1999	T1	PT14	T1	A. Bernal, U. de los Andes, Colombia	this paper
B98 or 57	USA	1999	T1	LNPV17.41		G. Coaker, UC Davis, USA	this paper
Max 1	Italy	2002	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	(Yan, Liu et al. 2008)
Max 4	Italy	2002	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	this paper
Max 5	Italy	2002	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	this paper
Max 6	Italy	2002	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	this paper
ISCI 181	Italy	2002	T1	IPV-CT28.31	T1	M. Zaccardelli, CRA ORT, Italy	this paper
ISCI 78	Italy	2003	T1	LNPV17.41	T1	M. Zaccardelli, CRA ORT, Italy	this paper
KS P 53	Tanzania	2004	T1	KSP53	T1	M. Zaccardelli, CRA ORT, Italy	(Shenge, Mabagala et al. 2007)
KS 127 M	Tanzania	2004	T1	KSP53	T1	M. Zaccardelli, CRA ORT, Italy	(Shenge, Mabagala et al. 2007)
ISCI 284	Italy	2004	T1	IPV-CT28.31	T1	M. Zaccardelli, CRA ORT, Italy	this paper
ISCI 286	Italy	2004	T1	IPV-CT28.31	T1	M. Zaccardelli, CRA ORT, Italy	this paper
ISCI 269	Italy	2004	T1	IPV-CT28.31	T1	M. Zaccardelli, CRA ORT, Italy	this paper
K40	USA	2005	T1	LNPV17.41	T1	C. Waldenmeier, VT, USA	this paper
K41	USA	2005	T1	LNPV17.41	T1	C. Waldenmeier, VT, USA	this paper
K100	USA	2005	T1	LNPV17.41	T1	C. Waldenmeier, VT, USA	this paper
838-4	USA	2005	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
838-16	USA	2005	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
836-2	USA	2005	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
838-8	USA	2005	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
838-9	USA	2005	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
838-6	USA	2005	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan

							et al. 2010)
1020	USA	2008	T1	LNPV17.41	T1	E. Bush, VT, USA	this paper
1021	USA	2008	T1	LNPV17.41	T1	E. Bush, VT, USA	this paper
410	USA	2008	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
16	USA	2008	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
20	USA	2008	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
21	USA	2008	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
22	USA	2008	T1	LNPV17.41	T1	G. Coaker, UC Davis, USA	(Kunkeaw, Tan et al. 2010)
338	Colombia	2009	T1	Colombia338	T1	A. Bernal, U. de los Andes, Colombia	this paper
196	Colombia	2009	T1	Colombia338	T1	A. Bernal, U. de los Andes, Colombia	this paper
198	Colombia	2009	T1	Colombia198	T1	A. Bernal, U. de los Andes, Colombia	this paper
199	Colombia	2009	T1	Colombia338	T1	A. Bernal, U. de los Andes, Colombia	this paper
201	Colombia	2008	T1	Colombia198	T1	A. Bernal, U. de los Andes, Colombia	this paper
204	Colombia	2009	T1	Colombia198	T1	A. Bernal, U. de los Andes, Colombia	this paper

[1] SNP genotype sequences are listed in Supplementary Table 3.3. SNP genotypes are only listed for T1-like strains (i.e., strains with MLST genotype T1).

Table 3.2. Summary of *Pto* draft genome sequences

Strain	Number of Contigs	N50	Largest Contig Size (bp)	Total Length (bp)	Illumina (X)[1]
NCPPB1108	304	46775	153603	6182607	42.6
K40	582	25354	104626	6254280	32.4
LNPV17.41	350	62385	239369	6157021	74.7
Max4	1176	12264	53242	6209056	27.5[2]

[1] Coverage was calculated based on total length of all reads used in each assembly.

[2] Assembled with a combination of both 454 and Illumina sequences (indicated coverage is based on Illumina reads only).

Supplementary Tables

Supplementary Table 3.1 SNPs identified between Max4, LNPV 17.41, T1, K40, and NCPPB1108 by aligning Illumina reads against the genome of Pto strain DC3000.

position in DC3000 genome	DC3000	Max4	LNPV 17.41	T1	K40	NCPPB 1108	genes genes	synonymous/non-synonympus or intergenic
210066	C	C	T	C	C	C	PSPTO_0186 D,D-heptose 1,7-bisphosphate phosphatase (209846-210397 +)	NS
327558	C	C	C	C	C	T	PSPTO_0301 4-aminobutyrate aminotransferase (327391-328671 +)	S
337095	G	T	T	G	T	G	PSPTO_0309 sulfate ABC transporter, permease protein CysT (336472-337293 +)	S
350787	C	C	C	C	C	A	PSPTO_0322 adenylate cyclase (349816-351231 -)	NS
363840	G	G	G	G	G	A	PSPTO_0334 alginate biosynthesis transcriptional regulatory protein AlgB (363589-364935 +)	S
374822	G	A	G	G	G	G	PSPTO_0344 DNA polymerase I (373243-376020 -)	NS
412163	G	C	C	C	C	G	PSPTO_0374 hypothetical protein (411649-412236 -)	NS
456705	T	G	G	G	G	T	intergenic	intergenic
467908	G	G	G	T	G	G	PSPTO_0422 virulence-associated protein, putative (467890-468183 -)	S
468768	T	C	C	C	C	T	PSPTO_0424 methyltransferase, putative (468737-469321 -)	NS
474967	A	A	A	A	G	A	PSPTO_0429 putative protein insertion permease FtsX (474625-475659 +)	NS
528525	C	A	A	C	A	C	PSPTO_0481 ACT domain-containing protein (528181-528699 +)	NS
555129	T	T	T	T	T	G	PSPTO_0506 acyl-CoA dehydrogenase family protein (553477-555255 +)	NS
560246	C	C	C	T	C	C	PSPTO_0512 pyrroloquinoline quinone biosynthesis protein PqqB (559887-560798 -)	NS
568123	T	T	T	T	T	C	PSPTO_0519 AsnC family transcriptional regulator (567763-568203 -)	S
605402	T	G	G	G	G	T	PSPTO_0551 dimethyladenosine transferase (604665-605471 -)	NS
607634	T	C	C	C	C	T	PSPTO_0553 peptidyl-prolyl cis-trans isomerase SurA (606454-607740 -)	NS
626469	C	C	A	C	C	C	PSPTO_0569 autotransporting lipase (625342-627264 -)	NS
654597	G	G	G	G	G	A	PSPTO_0595 hypothetical protein (653921-654616 +)	NS
665006	T	T	T	T	T	C	intergenic	intergenic
680050	A	A	A	G	A	A	PSPTO_0619 DNA-directed RNA polymerase subunit beta (678489-682562 +)	NS
700231	C	C	C	A	C	C	PSPTO_0644 50S ribosomal protein L30 (700111-700287 +)	NS
715737	A	A	A	A	A	G	PSPTO_0662 glycerol-3-phosphate ABC transporter, permease protein (715149-716042 -)	S
790022	A	A	A	A	C	A	PSPTO_0743 3-ketoacyl-(acyl-carrier-protein) reductase (789523-790878 -)	NS
798759	G	G	G	G	G	A	PSPTO_0750 copper-translocating P-type ATPase (796999-799197 -)	NS
862009	G	G	T	G	G	G	PSPTO_0795 hypothetical protein (861246-862010 +)	NS
970650	T	C	C	C	C	T	PSPTO_0892 autotransporter, putative (969369-972419 -)	S
1059630	G	T	T	T	T	G	intergenic	intergenic
1117119	G	T	T	T	T	G	intergenic	intergenic
1144313	A	T	T	T	T	A	intergenic	intergenic
1269060	C	T	C	C	C	C	PSPTO_1154 alpha/beta fold family hydrolase (1268283-1269062 -)	NS
1275046	A	A	A	A	G	A	PSPTO_1160 ABC transporter, ATP-binding protein (1273646-1275157 +)	S

1277566	C	T	C	C	C	C	PSPTO_1164 ompA family protein (1277310-1277999 +)	NS
1291142	G	A	G	G	G	G	PSPTO_1175 hypothetical protein (1290659-1291255 +)	NS
1325800	G	G	G	A	G	G	PSPTO_1209 RNA polymerase sigma factor (1325302-1325808 -)	S
1374099	G	T	G	G	G	G	PSPTO_1250 CsgG family protein (1373711-1374394 +)	NS
1433240	C	T	T	T	T	C	PSPTO_1304 hypothetical protein (1433113-1433583 +)	NS
1445080	C	C	C	C	C	A	PSPTO_1313 pili assembly chaperone (1444523-1445275 +)	NS
1480667	C	C	T	C	C	C	PSPTO_1348 sensory box protein (1480410-1482701 -)	NS
1524789	G	A	A	A	A	G	PSPTO_1382 type III helper protein HrpZ1 (1524678-1525790 +)	NS
1562412	A	G	A	A	A	A	intergenic	intergenic
1596159	G	G	T	G	G	G	PSPTO_1454 hypothetical protein (1595545-1596534 -)	NS
1656827	G	G	G	G	G	T	PSPTO_1499 response regulator/GGDEF domain-containing protein (1655919-1656923 +)	S
1657314	G	G	G	G	G	A	intergenic	intergenic
1759923	C	T	T	T	T	C	PSPTO_1605 sensory box histidine kinase (1759601-1760776 +)	NS
1780920	T	T	T	G	T	T	PSPTO_1624 acetate permease (1780575-1782233 +)	NS
1887415	G	G	G	G	G	T	PSPTO_1718 transporter, putative (1887368-1888576 +)	S
1970703	C	T	T	T	T	C	intergenic	intergenic
2007919	G	A	G	G	G	G	PSPTO_1838 succinylglutamate desuccinylase (2007120-2008127 +)	NS
2116354	C	C	C	C	C	T	PSPTO_1936 flagellar hook protein FlgE (2115552-2116877 +)	NS
2133148	G	G	G	G	G	T	PSPTO_1947 glycosyl transferase, group 2 family protein (2131534-2134440 +)	NS
2141061	G	C	C	C	C	G	PSPTO_1954 transcriptional regulator FleQ (2139870-2141345 +)	NS
2291032	G	G	G	T	G	G	intergenic	intergenic
2328677	G	T	T	G	T	G	PSPTO_2145 iron-regulated membrane protein, putative (2327858-2329093 -)	NS
2458284	T	T	T	C	T	T	PSPTO_2229 chaperone protein PapD (2457811-2458566 -)	NS
2499205	A	A	A	A	T	A	PSPTO_2256 Slt family transglycosylase (2499048-2500469 +)	NS
2552191	G	G	T	G	G	G	intergenic	intergenic
2553989	A	C	C	C	C	A	PSPTO_2305 levansucrase (2552817-2554064 +)	NS
2600803	C	C	C	C	T	C	intergenic	intergenic
2602500	G	G	G	A	G	G	PSPTO_2346 4-hydroxyphenylpyruvate dioxygenase, putative (2601132-2603039 +)	NS
2604676	A	G	G	G	G	A	intergenic	intergenic
2627801	C	C	C	C	C	T	PSPTO_2378 threonyl-tRNA synthetase (2627367-2629289 +)	S
2720409	T	T	A	T	T	T	PSPTO_2464 Mn2+/Fe2+ transporter (2720385-2721689 -)	NS
2821689	G	G	G	G	G	T	PSPTO_2554 hypothetical protein (2820042-2823548 +)	NS
2847502	G	G	G	A	G	G	PSPTO_2577 LysR family transcriptional regulator (2846878-2847771 +)	NS
2847503	G	G	G	A	G	G	PSPTO_2577 LysR family transcriptional regulator (2846878-2847771 +)	NS
2862816	C	C	C	T	C	C	PSPTO_2591 GGDEF domain-containing protein (2861816-2863402 +)	NS
2867703	C	C	T	C	C	C	PSPTO_2593 multidrug resistance protein AcrA/AcrE family (2866609-2867769 -)	NS
2884801	G	G	G	G	G	A	PSPTO_2602 yersiniabactin non-ribosomal peptide synthetase (2884461-2890634 -)	NS
2891659	A	A	A	A	T	A	PSPTO_2603 ABC transporter, ATP-binding/permease protein (2890695-2892440 -)	NS
2973068	C	C	C	C	C	G	PSPTO_2677 short chain dehydrogenase/reductase family oxidoreductase (2972883-2973743 +)	S
2978557	C	C	C	C	C	G	PSPTO_2682 hypothetical protein (2978050-2978649 -)	NS
3018728	A	C	C	C	C	A	PSPTO_2719 enoyl-CoA hydratase (3018617-3019429 +)	NS
3092752	C	C	T	C	C	C	PSPTO_2772 hypothetical protein (3092691-3093116 +)	NS
3113461	A	A	C	A	A	A	PSPTO_2795 error-prone DNA polymerase (3113215-3116310 +)	NS
3235859	C	C	C	T	C	C	PSPTO_2873 hypothetical protein (3235122-3236159 +)	S

3280249	C	C	C	C	C	T	PSPTO_2915 glutamine ABC transporter, permease protein (3279817-3280518 -)	S
3287551	C	C	C	C	C	G	PSPTO_2925 aspartate aminotransferase (3287487-3288695 +)	NS
3291168	G	A	G	G	G	G	PSPTO_2927 outer membrane porin OprE (3290338-3291669 +)	S
3302614	G	A	A	A	A	G	PSPTO_2939 hypothetical protein (3302230-3302742 -)	S
3334833	G	G	G	G	G	C	PSPTO_2964 methionyl-tRNA synthetase (3333375-3334859 -)	NS
3378908	G	A	A	A	A	G	PSPTO_3004 xylose transporter ATP-binding subunit (3377835-3379421 +)	NS
3432898	G	G	G	G	T	G	PSPTO_3051 2,4'-dihydroxyacetophenone dioxygenase (3432613-3433146 +)	NS
3437586	C	C	C	A	C	C	PSPTO_3057 MmgE/PrpD family protein (3437310-3438662 -)	S
3442554	G	A	A	A	A	G	PSPTO_3061 LysR family transcriptional regulator (3442333-3443256 +)	S
3463903	G	T	T	G	T	G	PSPTO_3081 LysR family transcriptional regulator (3463360-3464283 +)	NS
3498740	C	C	C	C	A	C	intergenic	intergenic
3499534	G	G	G	G	C	G	PSPTO_3112 glutathione reductase (3498764-3500122 -)	NS
3505372	C	C	T	C	C	C	intergenic	intergenic
3527804	C	T	C	C	C	C	PSPTO_3138 3-oxoadipate enol-lactone hydrolase (3527691-3528482 -)	NS
3559310	G	G	G	G	G	A	PSPTO_3166 hypothetical protein (3558705-3559667 +)	S
3639188	T	T	T	C	T	T	PSPTO_3229 filamentous hemagglutinin, intein-containing, putative (3629677-3648501 -)	NS
3659896	G	G	G	G	G	A	PSPTO_3238 tonB protein, putative (3659577-3660401 -)	NS
3674092	G	C	C	C	C	G	PSPTO_3249 dipeptide ABC transporter, permease protein DppB, putative (3673508-3674545 +)	S
3682333	C	C	G	C	C	C	PSPTO_3258 iron ABC transporter, ATP-binding protein (3682282-3683073 +)	NS
3719398	C	A	A	A	A	C	PSPTO_3291 methyl-accepting chemotaxis protein (3718924-3720552 -)	NS
3743986	C	G	G	G	G	C	PSPTO_3310 general secretion pathway protein L, putative (3743950-3745014 -)	S
3754462	C	C	C	C	C	G	PSPTO_3319 TetR family transcriptional regulator (3754165-3754809 -)	S
3757792	C	C	C	C	C	T	PSPTO_3323 aldehyde dehydrogenase family protein (3757178-3758761 +)	S
3762552	C	A	C	C	C	C	intergenic	intergenic
3821193	C	A	A	A	A	C	PSPTO_3379 methyl-accepting chemotaxis protein (3820066-3821745 -)	NS
3876030	G	T	T	G	T	G	intergenic	intergenic
3877617	G	A	A	A	A	G	PSPTO_3437 peptidase, M24 family protein (3876731-3878539 -)	NS
3913229	C	C	C	C	C	A	PSPTO_3466 alkanesulfonate monooxygenase (3912138-3913277 +)	S
3958408	C	C	C	C	C	T	PSPTO_3505 cytochrome b561, putative (3957995-3958534 +)	S
3989925	A	A	C	A	A	A	PSPTO_3534 glycosy hydrolase family protein (3989436-3990749 +)	NS
4053810	G	G	G	G	G	T	PSPTO_3594 beta-lactamase (4053747-4054898 +)	NS
4064806	C	C	A	C	C	C	PSPTO_3604 heavy metal sensor histidine kinase (4063509-4064897 +)	NS
4129644	C	C	T	C	C	C	PSPTO_3661 xanthine dehydrogenase, C-terminal subunit (4127968-4130346 +)	S
4139980	T	G	G	G	G	T	PSPTO_3671 hypothetical protein (4139347-4140132 +)	NS
4157976	G	G	G	G	A	G	PSPTO_3690 hypothetical protein (4157261-4158121 -)	NS
4189666	T	T	T	T	G	T	PSPTO_3712 ribonuclease H (4189457-4189909 -)	NS
4268997	T	T	T	C	T	T	PSPTO_3768 hypothetical protein (4268669-4269841 -)	NS
4274393	G	G	G	G	G	C	PSPTO_3772 dihydroxy-acid dehydratase (4273005-4274747 +)	NS
4445670	C	C	C	C	C	T	PSPTO_3932 hypothetical protein (4445527-4446045 -)	NS
4576906	C	C	C	C	C	A	PSPTO_4071 hypothetical protein (4576746-4577321 -)	NS
4642968	C	C	C	C	C	T	PSPTO_4119 carboxylesterase (4642495-4643643 -)	NS
4665696	C	T	T	T	T	C	PSPTO_4139 hypothetical protein (4665547-4666665 +)	S
4716863	T	T	T	C	T	T	intergenic	intergenic

4720805	G	G	G	G	G	T	PSPTO_4191 hypothetical protein (4719830-4722967 +)	NS
4758432	G	G	G	G	G	A	intergenic	intergenic
4835302	C	C	T	C	C	C	PSPTO_4292 sigma-54 dependent transcriptional regulator/response regulator (4835207-4836541 +)	S
4835668	G	G	G	G	G	A	PSPTO_4292 sigma-54 dependent transcriptional regulator/response regulator (4835207-4836541 +)	S
4884005	C	G	G	G	G	C	PSPTO_4333 moxR protein, putative (4883603-4884652 +)	NS
4894698	G	G	G	G	G	A	intergenic	intergenic
4974565	G	G	A	G	G	G	PSPTO_4407 UDP-N-acetylmuramate--L-alanine ligase (4973697-4975157 -)	NS
4975615	C	C	C	C	C	T	PSPTO_4408 undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase (4975150-4976220 -)	S
5003314	G	G	G	G	G	T	PSPTO_4435 trypsin domain-containing protein (5002982-5004142 +)	NS
5006146	G	G	G	G	G	A	PSPTO_4437 histidinol-phosphate aminotransferase (5005376-5006428 -)	NS
5098912	C	C	C	C	C	T	PSPTO_4519 non-ribosomal peptide synthetase, terminal component (5093815-5104113 +)	NS
5107113	G	G	G	G	G	T	PSPTO_4522 hypothetical protein (5106886-5108169 +)	S
5450732	G	G	G	G	G	T	PSPTO_4812 leucyl-tRNA synthetase (5449874-5452480 +)	NS
5452396	G	G	A	G	G	G	PSPTO_4812 leucyl-tRNA synthetase (5449874-5452480 +)	S
5490714	A	A	A	A	A	C	PSPTO_4845 putative lipoprotein (5485965-5490914 +)	NS
5505429	C	C	C	C	C	T	PSPTO_4860 acetyl-CoA carboxylase biotin carboxyl carrier protein subunit (5505070-5505525 +)	S
5531471	C	C	C	T	C	C	intergenic	intergenic
5542396	C	C	C	C	A	C	PSPTO_4892 phosphinothricin N-acetyltransferase, putative (5542269-5542811 +)	NS
5542402	A	A	A	G	A	A	PSPTO_4892 phosphinothricin N-acetyltransferase, putative (5542269-5542811 +)	NS
5548124	G	G	G	G	A	G	intergenic	intergenic
5564628	G	G	A	G	G	G	PSPTO_4916 high affinity branched-chain amino acid ABC transporter, ATP-binding protein (5564457-5565332 -)	S
5566009	C	C	A	C	C	C	PSPTO_4917 high-affinity branched-chain amino acid ABC transporter, permease protein BraE (5565329-5566642 -)	NS
5579978	T	T	T	T	T	C	PSPTO_4926 transglutaminase-like domain-containing protein (5576759-5580037 -)	S
5612608	T	T	T	C	T	T	PSPTO_4952 flagellar motor protein MotB (5612121-5613146 -)	NS
5716833	G	C	C	C	C	G	PSPTO_5022 hypothetical protein (5714337-5717465 -)	NS
5859927	G	G	A	G	G	G	PSPTO_5147 polyhydroxyalkanoate granule-associated protein PhaF (5859216-5859989 -)	S
5859928	T	T	C	T	T	T	PSPTO_5147 polyhydroxyalkanoate granule-associated protein PhaF (5859216-5859989 -)	NS
5867993	C	C	C	C	G	C	PSPTO_5159 methyl-accepting chemotaxis protein (5867181-5869097 +)	NS
5944827	G	A	A	A	A	G	PSPTO_5221 2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase (5943909-5945138 -)	S
6199364	A	G	G	G	G	A	PSPTO_5448 hypothetical protein (6198893-6201319 -)	S
6226396	C	C	C	C	G	C	PSPTO_5464 NAD(P) transhydrogenase, beta subunit (6225106-6226560 +)	NS
6290571	G	G	G	G	A	G	PSPTO_5523 cobalamin synthesis protein/P47K family protein (6289884-6290855 -)	S
6301559	C	C	C	C	C	A	PSPTO_5534 SPFH domain-containing protein (6301280-6302317 -)	NS
6305580	G	G	G	G	G	A	PSPTO_5537 hypothetical protein (6305268-6307025 +)	NS
6390327	G	G	T	G	G	G	PSPTO_5609 16S rRNA methyltransferase GidB (6389838-6390473 -)	S
6392694	T	T	T	T	T	C	intergenic	intergenic
6392755	A	A	G	A	A	A	intergenic	intergenic
6393695	C	C	C	T	C	C	PSPTO_5611 tRNA modification GTPase TrmE (6392951-6394321 -)	S

Supplementary Table 3.2 Core genome SNPs identified between Pto strains T1, Max4, NCPPB1108, K40, and LNPV17.41 by aligning Illumina reads against the T1 draft genome and only considering those SNPs located within core genome genes.

SNP	contig	position	**T1**	**Max4**	**NCPPB 1108**	**K40**	**LNPV 17.41**	locus_tag	strand	start	end	product
1	contig 00002	35461	**C**	**C**	**G**	**C**	**C**	PSPTOT1 _0038	-	35107	36849	dihydroxy-acid dehydratase
2	contig 00002	40857	**G**	**A**	**A**	**A**	**A**	PSPTOT1 _0042	+	40013	41185	6-phosphogluconolactonase
3	contig 00002	111360	**C**	**C**	**C**	**A**	**C**	PSPTOT1 _0101	-	110199	112166	ABC transporter, periplasmic substrate-binding protein
4	contig 00002	115770	**A**	**A**	**A**	**C**	**A**	PSPTOT1 _0105	+	115527	115979	ribonuclease HI
5	contig 00002	133353	**A**	**A**	**C**	**A**	**A**	PSPTOT1 _0118	-	133213	134781	methyl-accepting chemotaxis protein
6	contig 00002	159838	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _0138	-	159776	160708	membrane protein
7	contig 00003	3079	**G**	**G**	**G**	**G**	**A**	PSPTOT1 _0155	-	2266	4755	conserved hypothetical protein
8	contig 00003	49213	**G**	**G**	**G**	**G**	**T**	PSPTOT1 _0198	-	49122	50510	heavy metal sensor histidine kinase
9	contig 00003	49472	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _0198	-	49122	50510	heavy metal sensor histidine kinase
10	contig 00003	60209	**C**	**C**	**A**	**C**	**C**	PSPTOT1 _0208	-	59121	60272	beta-lactamase
11	contig 00003	122629	**T**	**T**	**T**	**T**	**G**	PSPTOT1 _0262	-	121805	123118	glycosyl hydrolase, family 5 PslG
12	contig 00003	141478	**A**	**A**	**G**	**A**	**A**	PSPTOT1 _0280	+	141216	143381	fatty oxidation complex, alpha subunit
13	contig 00003	152695	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _0291	-	152569	153108	cytochrome b561
14	contig 00003	163493	**A**	**G**	**G**	**G**	**G**	PSPTOT1 _0300	+	162912	163742	iolI protein
15	contig 00004	22870	**G**	**G**	**G**	**G**	**A**	PSPTOT1 _0324	-	22002	23462	UDP-N-acetylmuramate--alanine ligase
16	contig 00004	23920	**C**	**C**	**T**	**C**	**C**	PSPTOT1 _0325	-	23455	24525	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase
17	contig 00004	37960	**G**	**G**	**T**	**G**	**G**	PSPTOT1 _0336	+	36302	38116	lipoprotein
18	contig 00004	51626	**G**	**G**	**T**	**G**	**G**	PSPTOT1 _0352	+	51294	52454	trypsin domain protein
19	contig 00004	54456	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _0354	-	53686	54738	histidinol-phosphate aminotransferase
20	contig 00004	128194	**G**	**A**	**G**	**G**	**G**	PSPTOT1 _0426	-	128166	128480	conserved hypothetical protein
21	contig 00004	178564	**C**	**A**	**A**	**A**	**A**	PSPTOT1 _0459	+	177277	179100	GAF domain/GGDEF domain/EAL domain protein
22	contig 00004	195635	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _0474	+	194877	196499	dipeptide ABC transporter, periplasmic dipeptide-binding protein
23	contig 00004	197568	**C**	**C**	**T**	**C**	**C**	PSPTOT1 _0475	+	196829	198421	dipeptide ABC transporter, periplasmic dipeptide-binding protein
24	contig 00004	319696	**C**	**C**	**C**	**C**	**T**	PSPTOT1 _0590	+	319265	320548	glutamate-1-semialdehyde-2,1-aminomutase
25	contig 00004	326709	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _0597	+	325803	327323	apolipoprotein N-acyltransferase
26	contig 00004	329736	**G**	**G**	**T**	**G**	**G**	PSPTOT1 _0600	+	328878	331484	leucyl-tRNA synthetase
27	contig 00004	331400	**G**	**G**	**G**	**G**	**A**	PSPTOT1 _0600	+	328878	331484	leucyl-tRNA synthetase
28	contig 00005	48526	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _0660	-	46605	48704	outer membrane type III secretion hrcc
29	contig 00005	52823	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _0667	-	51804	52934	type III helper hrpZ1

30	contig 00005	130486	**G**	**G**	**T**	**G**	**G**	PSPTOT1_0733	-	130291	131043	pili assembly chaperone
31	contig 00005	141133	**A**	**A**	**G**	**A**	**A**	PSPTOT1_0741	-	140790	141266	conserved hypothetical protein
32	contig 00005	141467	**T**	**T**	**C**	**T**	**T**	PSPTOT1_0742	+	141432	141725	conserved hypothetical protein
33	contig 00005	152040	**T**	**C**	**T**	**T**	**T**	PSPTOT1_0752	-	151724	153010	glucose ABC transporter, periplasmic glucose-binding protein
34	contig 00005	160864	**C**	**C**	**C**	**C**	**T**	PSPTOT1_0760	+	160217	161122	regulatory protein
35	contig 00005	179806	**T**	**G**	**T**	**T**	**T**	PSPTOT1_0775	+	179670	181160	cytosol aminopeptidase
36	contig 00007	15345	**C**	**C**	**C**	**C**	**G**	PSPTOT1_0827	-	14924	16078	PQQ enzyme repeat domain protein
37	contig 00009	13293	**G**	**A**	**G**	**G**	**G**	PSPTOT1_0885	-	12063	13352	conserved hypothetical protein
38	contig 00009	27309	**T**	**T**	**C**	**T**	**T**	PSPTOT1_0898	-	26103	27620	aldehyde dehydrogenase family protein
39	contig 00009	61859	**G**	**G**	**A**	**G**	**G**	PSPTOT1_0931	-	60820	63444	alanyl-tRNA synthetase
40	contig 00009	63920	**A**	**A**	**G**	**A**	**A**	PSPTOT1_0932	-	63589	64629	threonine aldolase, low-specificity
41	contig 00009	65960	**C**	**T**	**C**	**C**	**C**	PSPTOT1_0934	-	65752	66759	succinylglutamate desuccinylase
42	contig 00009	71892	**A**	**A**	**G**	**A**	**A**	PSPTOT1_0939	-	71121	72152	arginine N-succinyltransferase, alpha subunit
43	contig 00009	83720	**T**	**T**	**G**	**T**	**T**	PSPTOT1_0949	-	82184	83788	phenylalanine hydroxylase transcriptional activator PhhR
44	contig 00009	83721	**C**	**C**	**G**	**C**	**C**	PSPTOT1_0949	-	82184	83788	phenylalanine hydroxylase transcriptional activator PhhR
45	contig 00009	115954	**A**	**A**	**G**	**A**	**A**	PSPTOT1_0986	-	115490	116395	transcriptional regulator, LysR family
46	contig 00009	116861	**T**	**T**	**C**	**T**	**T**	PSPTOT1_0987	+	116584	117306	conserved hypothetical protein
47	contig 00009	125102	**C**	**C**	**G**	**C**	**C**	PSPTOT1_0994	+	123954	125165	aspartate aminotransferase
48	contig 00009	147529	**A**	**C**	**A**	**A**	**A**	PSPTOT1_1020	-	147215	148192	UDP-glucose 4-epimerase
49	contig 00009	152569	**C**	**C**	**A**	**C**	**C**	PSPTOT1_1024	-	151105	152796	30S ribosomal protein S1
50	contig 00009	155126	**C**	**C**	**G**	**C**	**C**	PSPTOT1_1026	-	153604	155211	prephenate dehydrogenase/3-phosphoshikimate 1-carboxyvinyltransferase family protein
51	contig 00011	21911	**C**	**C**	**C**	**T**	**C**	PSPTOT1_1099	+	20799	21941	quinolinate synthetase
52	contig 00012	46585	**G**	**G**	**A**	**G**	**G**	PSPTOT1_1264	-	46234	46632	conserved hypothetical protein
53	contig 00015	14035	**G**	**G**	**A**	**G**	**G**	PSPTOT1_1354	+	11530	14640	AcrB/AcrD/AcrF family protein
54	contig 00015	20378	**A**	**C**	**A**	**A**	**A**	PSPTOT1_1359	-	17541	20387	type IV pilus-associated protein
55	contig 00018	40634	**G**	**T**	**G**	**T**	**T**	PSPTOT1_1456	+	40091	41014	transcriptional regulator, LysR family
56	contig 00018	48587	**A**	**A**	**C**	**A**	**A**	PSPTOT1_1463	+	47295	48923	nickel ABC transporter, periplasmic nickel-binding protein
57	contig 00018	65018	**C**	**G**	**C**	**C**	**C**	PSPTOT1_1477	+	65010	65786	outer membrane efflux protein
58	contig 00018	65019	**G**	**T**	**G**	**G**	**G**	PSPTOT1_1477	+	65010	65786	outer membrane efflux protein
59	contig 00018	69715	**A**	**A**	**G**	**A**	**A**	PSPTOT1_1483	-	69685	71247	alkyl hydroperoxide reductase, subunit F
60	contig 00018	93333	**C**	**C**	**T**	**C**	**C**	PSPTOT1_1503	+	92731	95514	glycosyl hydrolase, family 13
61	contig 00018	104601	**C**	**T**	**C**	**C**	**C**	PSPTOT1_1513	-	104488	105279	3-oxoadipate enol-lactone hydrolase
62	contig 00020	11135	**C**	**A**	**C**	**C**	**C**	PSPTOT1_1546	+	10926	11570	catabolite gene activator Crp
63	contig 00020	11601	**C**	**C**	**T**	**C**	**C**	PSPTOT1_1547	-	11582	12277	conserved hypothetical protein
64	contig 00020	34034	**G**	**G**	**G**	**G**	**T**	PSPTOT1_1574	+	33239	35161	autotransporting lipase, GDSL family

65	contig 00020	53133	**G**	**G**	**A**	**G**	**G**	PSPTOT1_1590	+	53027	54313	peptidyl-prolyl cis-trans isomerase SurA
66	contig 00020	55365	**C**	**C**	**A**	**C**	**C**	PSPTOT1_1592	+	55296	56102	dimethyladenosine transferase
67	contig 00020	63441	**C**	**C**	**G**	**C**	**C**	PSPTOT1_1599	+	63073	64302	tRNA nucleotidyltransferase
68	contig 00020	97801	**A**	**A**	**G**	**A**	**A**	PSPTOT1_1634	+	97694	98161	transcriptional regulator, AsnC family
69	contig 00020	104252	**A**	**G**	**G**	**G**	**G**	PSPTOT1_1638	+	103700	104611	coenzyme PQQ synthesis protein B
70	contig 00020	109369	**A**	**A**	**C**	**A**	**A**	PSPTOT1_1643	-	109243	111021	acyl-CoA dehydrogenase family protein
71	contig 00020	132608	**G**	**T**	**G**	**T**	**T**	PSPTOT1_1665	-	132434	132952	ACT domain protein
72	contig 00020	135006	**C**	**A**	**A**	**A**	**A**	PSPTOT1_1666	+	133275	135452	malate synthase G
73	contig 00020	160237	**A**	**G**	**G**	**G**	**G**	PSPTOT1_1687	+	157290	160310	sarcosine oxidase, alpha subunit
74	contig 00020	162709	**A**	**A**	**G**	**A**	**A**	PSPTOT1_1690	+	162109	163308	glutathione-independent formaldehyde dehydrogenase
75	contig 00020	180484	**A**	**A**	**T**	**A**	**A**	PSPTOT1_1707	+	180204	181262	conserved hypothetical protein
76	contig 00020	186262	**T**	**T**	**T**	**C**	**T**	PSPTOT1_1714	-	185570	186604	insertion permease FtsX protein
77	contig 00020	191521	**G**	**G**	**G**	**G**	**A**	PSPTOT1_1718	+	190415	191908	conserved hypothetical protein
78	contig 00020	192461	**G**	**G**	**A**	**G**	**G**	PSPTOT1_1719	+	191908	192492	methyltransferase
79	contig 00020	203207	**G**	**G**	**T**	**G**	**G**	PSPTOT1_1732	-	202538	203284	membrane protein
80	contig 00020	223828	**A**	**A**	**A**	**T**	**A**	PSPTOT1_1747	-	221993	223954	iron-sulfur cluster-binding protein
81	contig 00021	59192	**T**	**C**	**C**	**C**	**C**	PSPTOT1_1829	+	59037	59690	rna polymerase sigma 70 family
82	contig 00021	71898	**G**	**G**	**A**	**G**	**G**	PSPTOT1_1840	+	71492	72742	UDP-N-acetylglucosamine 2-epimerase
83	contig 00021	75577	**T**	**C**	**T**	**T**	**T**	PSPTOT1_1842	+	75136	77256	bacteriophage N4 adsorption protein B
84	contig 00021	99337	**G**	**A**	**G**	**G**	**G**	PSPTOT1_1867	-	98904	99593	ompA family protein
85	contig 00021	101858	**T**	**T**	**T**	**C**	**T**	PSPTOT1_1870	-	101747	103258	ABC transporter, ATP-binding protein
86	contig 00021	107844	**G**	**A**	**G**	**G**	**G**	PSPTOT1_1876	+	107842	108621	hydrolase, alpha/beta fold family
87	contig 00021	141597	**C**	**C**	**A**	**C**	**C**	PSPTOT1_1908	+	141414	142862	deoxyribodipyrimidine photolyase
88	contig 00021	265401	**A**	**C**	**C**	**C**	**C**	PSPTOT1_2021	+	265375	266082	conserved hypothetical protein
89	contig 00022	2203	**C**	**C**	**G**	**C**	**C**	PSPTOT1_2027	+	1490	2248	conserved hypothetical protein
90	contig 00022	8069	**G**	**G**	**A**	**G**	**G**	PSPTOT1_2032	+	7394	8542	carboxylesterase
91	contig 00022	14656	**G**	**A**	**G**	**G**	**G**	PSPTOT1_2038	-	14489	15190	ABC transporter
92	contig 00023	24301	**G**	**G**	**A**	**G**	**G**	PSPTOT1_2083	+	23696	24658	conserved hypothetical protein
93	contig 00023	49985	**A**	**A**	**A**	**A**	**T**	PSPTOT1_2110	-	49523	51130	metalloprotease
94	contig 00023	50674	**A**	**A**	**T**	**A**	**A**	PSPTOT1_2110	-	49523	51130	metalloprotease
95	contig 00023	51890	**G**	**G**	**G**	**G**	**C**	PSPTOT1_2111	+	51629	52945	sensor histidine kinase
96	contig 00023	121215	**C**	**C**	**C**	**C**	**G**	PSPTOT1_2159	+	121164	121955	iron ABC transporter, ATP-binding protein
97	contig 00024	2275	**C**	**C**	**A**	**C**	**C**	PSPTOT1_2176	-	2219	3559	glutamate synthase family protein
98	contig 00024	8484	**A**	**A**	**G**	**A**	**A**	PSPTOT1_2181	-	7595	8761	membrane protein
99	contig 00024	10743	**T**	**C**	**C**	**C**	**C**	PSPTOT1_2184	-	10475	11383	transcriptional regulator, LysR family
100	contig 00024	10744	**T**	**C**	**C**	**C**	**C**	PSPTOT1_2184	-	10475	11383	transcriptional regulator, LysR family
101	contig	10826	**A**	**A**	**G**	**A**	**A**	PSPTOT1	-	10475	11383	transcriptional regulator, LysR

	00024							_2184				family
102	contig 00025	9512	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _2234	+	9504	9896	hypothetical protein
103	contig 00025	48947	**T**	**T**	**A**	**T**	**T**	PSPTOT1 _2273	+	48481	49665	conserved hypothetical protein
104	contig 00025	153876	**T**	**T**	**T**	**T**	**G**	PSPTOT1 _2358	-	151027	154122	DNA polymerase III, alpha subunit
105	contig 00025	154983	**A**	**C**	**A**	**C**	**C**	PSPTOT1 _2359	-	154119	155534	conserved hypothetical protein
106	contig 00025	155700	**T**	**G**	**T**	**T**	**T**	PSPTOT1 _2360	-	155542	156162	conserved hypothetical protein
107	contig 00025	174516	**G**	**G**	**G**	**G**	**A**	PSPTOT1 _2383	-	174152	174577	hypothetical protein
108	contig 00025	205388	**C**	**C**	**T**	**C**	**C**	PSPTOT1 _2400	-	204272	205573	Secretion protein HlyD
109	contig 00025	212257	**G**	**T**	**G**	**G**	**G**	PSPTOT1 _2404	-	211025	212749	conserved hypothetical protein
110	contig 00025	222165	**G**	**C**	**G**	**G**	**G**	PSPTOT1 _2414	+	221750	223357	carboxyl transferase domain protein
111	contig 00025	234530	**T**	**C**	**C**	**C**	**C**	PSPTOT1 _2421	-	234488	235561	periplasmic sugar-binding domain protein
112	contig 00025	246507	**G**	**G**	**T**	**G**	**G**	PSPTOT1 _2433	-	245806	246705	enoyl-CoA hydratase/isomerase family protein
113	contig 00025	266898	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _2449	+	265455	266936	D-mannonate oxidoreductase
114	contig 00027	1325	**C**	**A**	**A**	**A**	**A**	PSPTOT1 _2471	-	128	1504	glutamine synthetase
115	contig 00027	9989	**C**	**G**	**G**	**G**	**G**	PSPTOT1 _2478	+	9501	10382	putrescine ABC transporter, permease protein
116	contig 00029	4291	**C**	**C**	**C**	**T**	**C**	PSPTOT1 _2555	+	3518	4396	nickel/cobalt transporter, high-affinity
117	contig 00029	9092	**C**	**A**	**C**	**C**	**C**	PSPTOT1 _2560	-	8182	9402	conserved hypothetical protein
118	contig 00029	17887	**C**	**C**	**C**	**C**	**T**	PSPTOT1 _2566	+	17792	19126	sigma-54 dependent transcriptional regulator/response regulator
119	contig 00029	18253	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _2566	+	17792	19126	sigma-54 dependent transcriptional regulator/response regulator
120	contig 00029	45528	**A**	**A**	**G**	**A**	**A**	PSPTOT1 _2588	+	45468	46940	phosphate transporter family protein
121	contig 00029	85531	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _2616	-	85026	86132	hydrolase
122	contig 00029	87792	**G**	**G**	**A**	**G**	**G**	PSPTOT1 _2618	+	87265	88071	acyltransferase domain protein
123	contig 00029	90523	**T**	**T**	**T**	**T**	**C**	PSPTOT1 _2621	+	90328	91377	oxidoreductase, FAD/FMN-binding
124	contig 00030	47069	**A**	**A**	**C**	**A**	**A**	PSPTOT1 _2685	-	45470	47920	glycogen phosphorylase
125	contig 00030	57948	**G**	**G**	**G**	**C**	**G**	PSPTOT1 _2691	-	56844	58760	**methyl-accepting chemotaxis protein**
126	contig 00030	66013	**A**	**A**	**A**	**A**	**G**	PSPTOT1 _2703	+	65952	66725	polyhydroxyalkanoate granule-associated protein PhaF
127	contig 00030	66014	**C**	**C**	**C**	**C**	**T**	PSPTOT1 _2703	+	65952	66725	polyhydroxyalkanoate granule-associated protein PhaF
128	contig 00030	66624	**C**	**C**	**A**	**C**	**C**	PSPTOT1 _2703	+	65952	66725	polyhydroxyalkanoate granule-associated protein PhaF
129	contig 00030	91833	**C**	**C**	**T**	**C**	**C**	PSPTOT1 _2724	+	90842	91945	3-dehydroquinate synthase
130	contig 00030	134895	**C**	**C**	**C**	**C**	**T**	PSPTOT1 _2760	+	134230	135906	malonate decarboxylase, alpha subunit
131	contig 00031	1502	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _2781	-	1068	1835	oxidoreductase, short-chain dehydrogenase/reductase family
132	contig 00031	22356	**G**	**G**	**C**	**G**	**G**	PSPTOT1 _2798	-	20898	22397	conserved hypothetical protein
133	contig 00036	55113	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _2879	-	54600	56186	D-xylose ABC transporter, ATP-binding protein
134	contig 00040	3224	**A**	**A**	**C**	**A**	**A**	PSPTOT1 _2919	+	2846	3391	conserved hypothetical protein
135	contig 00040	3584	**T**	**C**	**C**	**C**	**C**	PSPTOT1 _2920	-	3532	4587	conserved hypothetical protein
136	contig 00040	11872	**A**	**A**	**C**	**A**	**A**	PSPTOT1 _2931	-	11790	12698	trpBA operon transcriptional activator

137	contig 00040	13419	**C**	**C**	**G**	**C**	**C**	PSPTOT1_2932	+	12826	14055	tryptophan synthase, beta subunit
138	contig 00040	41052	**C**	**C**	**C**	**C**	**T**	PSPTOT1_2959	+	40832	41383	histidinol-phosphate phosphatase family protein
139	contig 00040	54668	**A**	**G**	**G**	**G**	**G**	PSPTOT1_2971	+	54042	55412	tRNA modification GTPase TrmE
140	contig 00040	58036	**C**	**C**	**C**	**C**	**A**	PSPTOT1_2974	+	57890	58525	glucose-inhibited division protein B
141	contig 00040	66885	**C**	**C**	**C**	**C**	**T**	PSPTOT1_2984	+	65635	67014	ATP synthase F1, beta subunit
142	contig 00040	106641	**G**	**G**	**T**	**G**	**G**	PSPTOT1_3017	+	105883	106920	SPFH domain / Band 7 family protein
143	contig 00040	106696	**G**	**A**	**A**	**A**	**A**	PSPTOT1_3017	+	105883	106920	SPFH domain / Band 7 family protein
144	contig 00040	114588	**G**	**G**	**A**	**G**	**G**	PSPTOT1_3025	+	114518	115729	cobalamin synthesis protein/P47K family protein
145	contig 00040	117094	**C**	**C**	**C**	**T**	**C**	PSPTOT1_3028	+	116810	117781	cobalamin synthesis protein/P47K family protein
146	contig 00040	123123	**C**	**C**	**G**	**C**	**C**	PSPTOT1_3034	-	122373	123215	conserved hypothetical protein
147	contig 00040	138494	**A**	**A**	**C**	**A**	**A**	PSPTOT1_3047	-	137760	138764	L-asparaginase I
148	contig 00040	149703	**G**	**G**	**A**	**G**	**G**	PSPTOT1_3060	+	149640	150929	General substrate transporter:Major facilitator superfamily
149	contig 00040	179161	**G**	**G**	**G**	**C**	**G**	PSPTOT1_3086	-	178997	180451	NAD(P) transhydrogenase, beta subunit
150	contig 00045	10259	**T**	**G**	**T**	**T**	**T**	PSPTOT1_3134	-	9367	10734	conserved hypothetical protein
151	contig 00045	10399	**T**	**C**	**T**	**T**	**T**	PSPTOT1_3134	-	9367	10734	conserved hypothetical protein
152	contig 00045	10465	**A**	**G**	**A**	**A**	**A**	PSPTOT1_3134	-	9367	10734	conserved hypothetical protein
153	contig 00047	7479	**G**	**G**	**G**	**T**	**G**	PSPTOT1_3202	+	7234	7491	ribosomal protein S16
154	contig 00047	43270	**G**	**G**	**T**	**G**	**G**	PSPTOT1_3228	+	42362	43366	response regulator/GGDEF domain protein
155	contig 00047	60237	**G**	**G**	**T**	**G**	**G**	PSPTOT1_3247	-	60185	60874	conserved hypothetical protein
156	contig 00047	142758	**A**	**G**	**G**	**G**	**G**	PSPTOT1_3322	-	141483	143198	prolyl-tRNA synthetase
157	contig 00047	160425	**A**	**A**	**G**	**A**	**A**	PSPTOT1_3343	+	160294	160983	exsB protein
158	contig 00048	38849	**T**	**T**	**T**	**T**	**A**	PSPTOT1_3387	+	37933	39351	major facilitator family transporter
159	contig 00048	49483	**G**	**G**	**A**	**G**	**G**	PSPTOT1_3397	-	49087	49995	glutaminase
160	contig 00048	68151	**G**	**A**	**A**	**A**	**A**	PSPTOT1_3412	-	67653	69638	**methyl-accepting chemotaxis protein**
161	contig 00052	11716	**T**	**T**	**C**	**T**	**T**	PSPTOT1_3445	+	10728	11762	twitching motility protein
162	contig 00052	28070	**G**	**G**	**A**	**G**	**G**	PSPTOT1_3460	+	23381	29344	sensor histidine kinase/response regulator
163	contig 00055	8029	**G**	**A**	**A**	**A**	**A**	PSPTOT1_3507	+	6472	10545	DNA-directed RNA polymerase, beta subunit
164	contig 00055	8033	**G**	**A**	**A**	**A**	**A**	PSPTOT1_3507	+	6472	10545	DNA-directed RNA polymerase, beta subunit
165	contig 00055	22146	**A**	**C**	**C**	**C**	**C**	PSPTOT1_3517	+	21428	22252	ribosomal protein L2
166	contig 00055	28214	**A**	**C**	**C**	**C**	**C**	PSPTOT1_3533	+	28094	28270	50S ribosomal protein L30
167	contig 00055	43720	**A**	**A**	**G**	**A**	**A**	PSPTOT1_3551	-	43132	44025	glycerol-3-phosphate ABC transporter, permease protein
168	contig 00055	84097	**T**	**A**	**T**	**T**	**T**	PSPTOT1_3598	+	83922	84626	O-antigen ABC transporter, ATP-binding protein
169	contig 00055	107858	**A**	**A**	**G**	**A**	**A**	PSPTOT1_3624	-	107333	109060	MORN repeat family protein
170	contig 00055	115467	**A**	**A**	**A**	**C**	**A**	PSPTOT1_3634	-	114968	116323	oxidoreductase, short chain dehydrogenase/reductase family
171	contig 00055	124204	**G**	**G**	**A**	**G**	**G**	PSPTOT1_3642	-	122444	124642	copper-translocating P-type ATPase
172	contig 00055	130608	**G**	**A**	**G**	**G**	**G**	PSPTOT1_3651	-	130315	131268	adenosine deaminase

173	contig 00055	139430	**T**	**T**	**T**	**G**	**T**	PSPTOT1_3659	-	138895	140250	hydroxydechloroatrazine ethylaminohydrolase
174	contig 00057	24030	**G**	**G**	**G**	**G**	**A**	PSPTOT1_3698	+	23833	25131	major facilitator family transporter
175	contig 00057	114183	**A**	**A**	**G**	**A**	**A**	PSPTOT1_3817	+	114138	114416	conserved hypothetical protein
176	contig 00057	124061	**G**	**T**	**G**	**T**	**T**	PSPTOT1_3828	+	123830	124147	conserved hypothetical protein
177	contig 00057	152478	**C**	**C**	**T**	**C**	**C**	PSPTOT1_3855	+	151936	154881	hypothetical protein
178	contig 00057	203308	**G**	**G**	**C**	**G**	**G**	PSPTOT1_3908	-	203024	204499	transcriptional regulator FleQ
179	contig 00057	208087	**G**	**A**	**G**	**A**	**A**	PSPTOT1_3913	-	207534	208382	flagellin
180	contig 00057	211216	**C**	**C**	**A**	**C**	**C**	PSPTOT1_3915	-	209924	212830	glycosyl transferase, group 2 family protein
181	contig 00057	236745	**T**	**T**	**G**	**T**	**T**	PSPTOT1_3937	+	236404	237150	conserved hypothetical protein
182	contig 00061	7249	**G**	**G**	**A**	**G**	**G**	PSPTOT1_3973	-	5761	7683	threonyl-tRNA synthetase
183	contig 00061	32551	**T**	**C**	**C**	**C**	**C**	PSPTOT1_4005	-	32012	33919	4-hydroxyphenylpyruvate dioxygenase
184	contig 00061	53390	**G**	**G**	**G**	**C**	**G**	PSPTOT1_4023	+	50340	53528	extracellular solute-binding protein/sensory box protein
185	contig 00064	35856	**G**	**T**	**G**	**G**	**G**	PSPTOT1_4123	-	35108	35935	hypothetical protein
186	contig 00064	35858	**C**	**G**	**C**	**C**	**C**	PSPTOT1_4123	-	35108	35935	hypothetical protein
187	contig 00064	35860	**G**	**C**	**G**	**G**	**G**	PSPTOT1_4123	-	35108	35935	hypothetical protein
188	contig 00064	66193	**G**	**G**	**A**	**G**	**G**	PSPTOT1_4154	-	64495	66234	biotin carboxylase/biotin carboxyl carrier protein
189	contig 00065	10957	**G**	**G**	**C**	**G**	**G**	PSPTOT1_4172	+	10839	12809	macrolide ABC efflux protein
190	contig 00065	17728	**A**	**T**	**A**	**A**	**A**	PSPTOT1_4176	+	17635	18918	aminotransferase, class V
191	contig 00065	21074	**C**	**C**	**T**	**C**	**C**	PSPTOT1_4178	+	19911	21575	pyoverdine ABC transporter, ATP-binding/permease protein
192	contig 00065	55242	**C**	**A**	**C**	**A**	**A**	PSPTOT1_4186	+	54847	56061	peptidase
193	contig 00067	49505	**C**	**T**	**C**	**C**	**C**	PSPTOT1_4259	+	48319	51084	DNA polymerase I
194	contig 00067	60487	**C**	**C**	**T**	**C**	**C**	PSPTOT1_4268	-	59392	60738	alginate biosynthesis transcriptional regulatory protein AlgB
195	contig 00067	73540	**G**	**G**	**T**	**G**	**G**	PSPTOT1_4280	+	73096	74511	adenylate cyclase
196	contig 00070	57366	**C**	**C**	**T**	**C**	**C**	PSPTOT1_4325	-	57182	58678	conserved hypothetical protein
197	contig 00070	75403	**C**	**A**	**A**	**A**	**A**	PSPTOT1_4340	-	74090	75748	sodium:solute symporter family protein
198	contig 00070	75584	**G**	**G**	**A**	**G**	**G**	PSPTOT1_4340	-	74090	75748	sodium:solute symporter family protein
199	contig 00070	81179	**G**	**A**	**G**	**G**	**G**	PSPTOT1_4348	-	81088	81996	transcriptional regulator, LysR family
200	contig 00070	96401	**A**	**A**	**G**	**A**	**A**	PSPTOT1_4361	-	95548	96723	sensory box histidine kinase
201	contig 00070	104694	**A**	**A**	**C**	**A**	**A**	PSPTOT1_4373	-	104203	105039	conserved hypothetical protein
202	contig 00070	134846	**C**	**C**	**A**	**C**	**C**	PSPTOT1_4408	-	134686	135261	conserved hypothetical protein
203	contig 00072	41689	**A**	**A**	**G**	**A**	**A**	PSPTOT1_4478	+	41221	42117	conserved hypothetical protein
204	contig 00072	44412	**G**	**G**	**A**	**G**	**G**	PSPTOT1_4480	+	43554	44510	conserved hypothetical protein
205	contig 00072	68388	**A**	**A**	**G**	**A**	**A**	PSPTOT1_4502	+	67206	69110	DNA topoisomerase IV, B subunit
206	contig 00072	84106	**G**	**A**	**A**	**A**	**A**	PSPTOT1_4513	+	83568	84593	chemotaxis motB protein
207	contig 00072	88379	**G**	**G**	**G**	**G**	**T**	PSPTOT1_4518	+	88295	89785	YjeF-related protein
208	contig 00072	116746	**A**	**A**	**G**	**A**	**A**	PSPTOT1_4539	+	116687	119965	transglutaminase-like superfamily domain protein

209	contig 00072	130698	**G**	**G**	**G**	**G**	**T**	PSPTOT1_4548	+	130065	131378	high-affinity branched-chain amino acid ABC transporter, permease protein BraE
210	contig 00072	132079	**C**	**C**	**C**	**C**	**T**	PSPTOT1_4549	+	131375	132250	high affinity branched-chain amino acid ABC transporter, ATP-binding protein
211	contig 00072	152988	**C**	**T**	**T**	**T**	**T**	PSPTOT1_4573	-	152579	153121	phosphinothricin N-acetyltransferase
212	contig 00072	152994	**G**	**G**	**G**	**T**	**G**	PSPTOT1_4573	-	152579	153121	phosphinothricin N-acetyltransferase
213	contig 00072	163898	**A**	**G**	**G**	**G**	**G**	PSPTOT1_4581	-	160468	163938	hypothetical protein
214	contig 00072	189946	**G**	**G**	**A**	**G**	**G**	PSPTOT1_4606	-	189850	190305	acetyl-CoA carboxylase biotin carboxyl carrier protein subunit
215	contig 00072	204902	**T**	**T**	**G**	**T**	**T**	PSPTOT1_4621	-	204702	209651	lipoprotein
216	contig 00072	221460	**G**	**G**	**C**	**G**	**G**	PSPTOT1_4633	+	220578	223277	sensory box histidine kinase
217	contig 00075	21989	**T**	**T**	**T**	**T**	**C**	PSPTOT1_4670	-	21989	22645	conserved hypothetical protein
218	contig 00079	14498	**C**	**C**	**T**	**C**	**C**	PSPTOT1_4709	+	12543	14969	conserved hypothetical protein
219	contig 00086	13556	**A**	**A**	**G**	**A**	**A**	PSPTOT1_4741	-	13154	14509	RNA methyltransferase, TrmA family
220	contig 00097	20412	**G**	**G**	**T**	**G**	**G**	PSPTOT1_4803	-	20364	21503	alkanesulfonate monooxygenase family protein
221	contig 00097	50381	**T**	**T**	**C**	**T**	**T**	PSPTOT1_4830	-	49702	50637	glycosyl transferase, group 2 family protein
222	contig 00097	56024	**T**	**T**	**C**	**T**	**T**	PSPTOT1_4834	+	55102	56910	peptidase, M24 family protein
223	contig 00098	10381	**T**	**G**	**T**	**T**	**T**	PSPTOT1_4846	+	8446	10488	cadmium-translocating P-type ATPase
224	contig 00098	10382	**G**	**T**	**G**	**G**	**G**	PSPTOT1_4846	+	8446	10488	cadmium-translocating P-type ATPase
225	contig 00098	21289	**G**	**G**	**A**	**G**	**G**	PSPTOT1_4855	-	20357	22180	urocanate hydratase
226	contig 00098	39807	**C**	**C**	**G**	**C**	**C**	PSPTOT1_4871	-	38794	39852	radical SAM domain protein
227	contig 00102	11047	**G**	**G**	**T**	**G**	**G**	PSPTOT1_4902	+	11000	12208	transporter
228	contig 00109	24783	**G**	**G**	**T**	**G**	**G**	PSPTOT1_4965	-	24708	25955	levansucrase
229	contig 00109	33995	**C**	**C**	**A**	**C**	**C**	PSPTOT1_4970	+	31318	34038	nitrate reductase
230	contig 00109	56723	**G**	**G**	**A**	**G**	**G**	PSPTOT1_4991	-	56081	56794	hydrolase, haloacid dehalogenase-like family
231	contig 00109	79308	**T**	**T**	**T**	**A**	**T**	PSPTOT1_5015	-	78044	79465	transglycosylase, SLT family
232	contig 00109	118907	**G**	**A**	**A**	**A**	**A**	PSPTOT1_5040	+	118625	119380	chaperone protein PapD
233	contig 00109	166164	**G**	**G**	**G**	**T**	**G**	PSPTOT1_5082	-	165918	166745	conserved hypothetical protein
234	contig 00112	15186	**C**	**C**	**T**	**C**	**C**	PSPTOT1_5116	+	15019	16299	4-aminobutyrate aminotransferase
235	contig 00112	24881	**G**	**T**	**G**	**T**	**T**	PSPTOT1_5124	+	24258	25079	sulfate ABC transporter, permease protein CysT
236	contig 00119	4636	**T**	**C**	**C**	**C**	**C**	PSPTOT1_5145	+	3561	5222	GGDEF domain protein
237	contig 00119	9523	**C**	**C**	**C**	**C**	**T**	PSPTOT1_5147	-	8429	9589	multidrug resistance protein, AcrA/AcrE family
238	contig 00119	52751	**T**	**T**	**G**	**T**	**T**	PSPTOT1_5176	-	52351	53256	transcriptional regulator, LysR family
239	contig 00119	63321	**T**	**G**	**T**	**T**	**T**	PSPTOT1_5185	+	62609	63328	cation ABC transporter, ATP-binding protein
240	contig 00119	93726	**G**	**G**	**C**	**G**	**G**	PSPTOT1_5213	+	93637	94515	transcriptional regulator, LysR family
241	contig 00119	109196	**C**	**C**	**G**	**C**	**C**	PSPTOT1_5230	+	109011	109871	oxidoreductase, short chain dehydrogenase/reductase family
242	contig 00121	24585	**G**	**G**	**G**	**T**	**G**	PSPTOT1_5274	-	24300	24731	acetyltransferase, GNAT family
243	contig 00121	40102	**C**	**T**	**C**	**C**	**C**	PSPTOT1_5288	+	38229	40633	cell division protein FtsK, truncated
244	contig	64047	**G**	**G**	**A**	**G**	**G**	PSPTOT1	-	63078	64661	aldehyde dehydrogenase family

	00121							_5312				protein
245	contig 00121	67377	**G**	**G**	**C**	**G**	**G**	PSPTOT1 _5316	+	67030	67674	transcriptional regulator, TetR family
246	contig 00121	77854	**C**	**C**	**G**	**C**	**C**	PSPTOT1 _5326	+	76826	77896	general secretion pathway protein L
247	contig 00121	80190	**T**	**G**	**T**	**T**	**T**	PSPTOT1 _5329	+	79006	81333	type II and III secretion system protein
248	contig 00121	95175	**C**	**C**	**G**	**C**	**C**	PSPTOT1 _5341	-	94772	95455	DNA-binding heavy metal response regulator
249	contig 00122	23348	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _5389	+	23037	24266	2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase
250	contig 00127	47397	**A**	**A**	**T**	**A**	**A**	PSPTOT1 _5476	-	47372	47950	tellurium resistance protein TerE
251	contig 00127	69221	**G**	**G**	**G**	**G**	**A**	PSPTOT1 _5503	+	69116	69649	MOSC domain protein
252	contig 00127	75440	**G**	**G**	**G**	**G**	**T**	PSPTOT1 _5508	+	73548	75650	chemotaxis sensor histidine kinase CheA
253	contig 00127	75990	**A**	**A**	**C**	**A**	**A**	PSPTOT1 _5509	+	75679	77334	methyl-accepting chemotaxis protein
254	contig 00127	75994	**G**	**G**	**T**	**G**	**G**	PSPTOT1 _5509	+	75679	77334	methyl-accepting chemotaxis protein
255	contig 00127	77143	**G**	**G**	**G**	**G**	**T**	PSPTOT1 _5509	+	75679	77334	methyl-accepting chemotaxis protein
256	contig 00127	77185	**T**	**T**	**T**	**T**	**A**	PSPTOT1 _5509	+	75679	77334	methyl-accepting chemotaxis protein
257	contig 00128	1054	**A**	**A**	**G**	**A**	**A**	PSPTOT1 _5548	+	454	1755	Mn2+/Fe2+ transporter, NRAMP family
258	contig 00128	1731	**A**	**A**	**A**	**A**	**T**	PSPTOT1 _5548	+	454	1755	Mn2+/Fe2+ transporter, NRAMP family
259	contig 00128	23702	**A**	**A**	**G**	**A**	**A**	PSPTOT1 _5567	+	23638	24984	monooxygenase, NtaA/SnaA/SoxA family
260	contig 00128	24767	**G**	**G**	**T**	**G**	**G**	PSPTOT1 _5567	+	23638	24984	monooxygenase, NtaA/SnaA/SoxA family
261	contig 00133	9968	**C**	**C**	**G**	**C**	**C**	PSPTOT1 _5614	-	9606	10028	protozoan/cyanobacterial globin family protein
262	contig 00133	9972	**C**	**C**	**A**	**C**	**C**	PSPTOT1 _5614	-	9606	10028	protozoan/cyanobacterial globin family protein
263	contig 00133	46974	**T**	**T**	**C**	**T**	**T**	PSPTOT1 _5650	+	46766	50446	urea amidolyase-related protein
264	contig 00134	8120	**C**	**C**	**A**	**C**	**C**	PSPTOT1 _5668	-	5958	9095	conserved hypothetical protein
265	contig 00134	61849	**A**	**A**	**G**	**A**	**A**	PSPTOT1 _5723	-	60880	61998	conserved hypothetical protein

Supplementary Table 3.3 Primers

F primer	**sequence**	**R primer**	**sequence**
SNP primers			
SNP6 F	CCATTATCCAAGGCACACAA	SNP6 R	AGGCAACCTGCCGAGCTTCT
SNP12 F	GGGTGAGTCCGTCAACAAGT	SNP12 R	GTACCGCCGAAACCTGGATA
SNP29 F	GCATCAGTTCGTTGCAAACC	SNP29 R	AAATTGTCGCCGAGCTTTTC
SNP55 F	GTACTGCCTCGGGTCGTG	SNP55 R	CCCGGGTAAACGACCAGTAA
SNP71 F	CTCACCACCCAACACCAACT	SNP71 R	ATCTCGTCGTAACGCTGGTT
SNP105 F	CGACGTCGAAATTCAGCTC	SNP105 R	AGGCACTGGTGCCCAACT
SNP176 F	AACATCAGATGGTTTCGATGC	SNP176 R	ATCAGATGGGCAGAGATGCT
SNP179 F	TTGCGCTCAGAGTCAAAGTG	SNP179 R	TCGACCTCGATGACTCGTCT
SNP192 F	GGCTTACGTCAACCCCTACA	SNP192 R	AGACCTGGAATCCGCTCTTT
SNP235 F	GGCTTCAAGATCGCCTACAC	SNP235 R	TGGAGCAGATTGATCAGCAG
SNP253/254 F	CAATTACCGGCTGAAGGAAA	SNP253/254 R	GAAACGGGTGAATTCCGATT
avrPto1 F	ACAACTCGGGTGACGAAGAT	avrPto1 R	CGGGCTAGGAGAAAGTGTTG
avrRps4 F	CAGTTAATTCCAGTTTCAACTACAC G	avrRps4R	AGGCGTCTCTGTAATTCAACAA
avrD1 F	ATCAAGGCCGGAAATAAACC	avrD1 R	CATCACGAGTCAAACCATCG
cor F	GACGACCAAGAAAGCCTCAG	cor R	CTTGCGCTGGTCTAGGGTAG
HopM1 cloning primers			
HopM1 F	AAAAAGCAGGCTC CTGGGAGATTCCAATGAT	HopM1R DC3000/JL1065	AGAAAGCTGGGT AACGCGGGTCAAGCAAGC
		HopM1R PT21	AGAAAGCTGGGT AGGAATAACCGTATTGGTATCCAC
		HopM1R NCPPB1108	AGAAAGCTGGGT AGGCTTCGCCGATTGCCTTG
		HopM1R T1	AGAAAGCTGGGT AGATCAGTTGCCCGACCTC

Supplementary Table 3.4 DNA sequences corresponding to the MLST and SNP genotypes listed in Table 3.1 (only nucleotides corresponding to SNPs are shown and were used for molecular evolutionary analyses, i.e. nucleotides identical in all analyzed strains were ignored).

MLST genotypes																		
	pgi	rpoD	gapA															
DC3000	TCTCCTGC	TATATCTTTGGAGG	GC															
JL1065	CCGCGCAT	CGCATCCCCTAGTG	AT															
T1	CTGTGCGC	CGTCCATCTGGAGA	AT															
SNP genotypes (column headers indicate position of each SNP in the contigs of the T1 genome assembly)																		
	C127_27,943	C25_154,983	C20_132,608	C20_132,657	C57_124,061	C18_40,632	C112_24,881	C65_55,242	C65_55,602	C57_208,087	C57_208,096	C57_208,267	C2_159,838	C3_141,478	C5_52,823	C127_75,990	C127_75,994
NCPPB1108	C	A	G	G	G	G	G	C	G	G	A	C	G	G	G	C	T
CA315	T	A	G	G	G	G	G	C	A	G	A	C	G	G	G	C	T
CFBP1318	C	A	G	G	G	G	G	C	G	G	A	C	A	A	G	C	T
Colombia198	C	A	G	G	G	G	G	C	G	G	A	T	A	A	G	C	T
Colombia338	C	A	G	G	G	G	G	C	G	G	T	T	A	A	G	C	T
KSP53	C	A	G	G	G	G	G	C	G	A[1]	A	C	A	A	G	C	T
T1	C	A	G	G	G	G	G	C	G	G	A	C	A	A	A	A	G
CFBP2545	C	A	G	T	G	G	G	C	G	G	A	C	A	A	A	A	G
PT14	C	C	T	G	T	G	G	C	G	G	A	C	A	A	A	A	G
IPV-CT28.31	C	C	T	G	T	G	G	C	G	A	A	C	A	A	A	A	G
LNPV17.41	C	C	T	G	T	T	T	A	G	A	A	C	A	A	A	A	G
1 homoplasy probably due to recombination or parallel evolution																		

Supplementary Table 3.5 List of strains with continent and year of isolation, MLST genotype, SNP genotype, and results for several virulence factors based on PCR (and sequencing of PCR products for hopM1)

ID	name	Continent	Year	MLST genotype	SNP genotype	hopM1 allele	Coronatine	avrRps4	avrD1	avrPto
1005	ICMP 4325	N. America	1944	DC3000	-	DC3000	+	-	+	+
1	DC3000	Europe	1961	DC3000	-	DC3000	+	-	-	+
1016	NCPPB 1008	N. America	1942	JL1065	-	JL1065	-	-	-	+
1347	CFBP 1696	Europe	1949	JL1065	-	JL1065	ND	ND	ND	ND
1007	NCPPB 880	Europe	1953	JL1065	-	JL1065	-	-	-	+
1009	ICMP 2846	N. America	1956	JL1065	-	JL1065	-	-	-	+
1346	CFBP 1319	Europe	1970	JL1065	-	JL1065	ND	ND	ND	ND
1348	CFBP 1785	Australia	1972	JL1065	-	JL1065	ND	ND	ND	ND
1015	ICMP 3647	Australia	1973	JL1065	-	JL1065	+	-	-	+
1017	ICMP 4355	Australia	1975	JL1065	-	JL1065	+	+	+	+
22	JL1065	N. America	1983	JL1065	-	JL1065	+	+	+	+
920	BS118	N. America	1983	JL1065	-	JL1065	+	+	+	+
922	BS120	N. America	1983	JL1065	-	JL1065	+	+	+	+
293	DC84-1	N. America	1984	JL1065	-	JL1065	+	+	+	+
300	PST26L	Africa	1986	JL1065	-	JL1065	+	-	+	+
1349	CFBP 3728	Asia	1988	JL1065	-	JL1065	ND	ND	ND	ND
49	PT 28	N. America	1992	JL1065	-	JL1065	-	-	+	+
50	PT 29	N. America	1992	JL1065	-	JL1065	+	+	+	+
1013	CPST 147	Europe	1993	JL1065	-	JL1065	-	-	+	+
1246	B64 or 56	N. America	1995	JL1065	-	JL1065	ND	ND	ND	+
1224	Pst field 8	N. America	1999	JL1065	-	JL1065	ND	ND	ND	ND
634	KS 112 lr	Africa	2004	JL1065	-	JL1065	+	+	+	+
635	KS 097 lr	Africa	2004	JL1065	-	JL1065	+	+	+	+
226	NCPPB 1108	Europe	1961	T1	NCPPB1108	1108	-	-	-	+
299	CFBP 1318	Europe	1969	T1	CFBP1318	T1	+	+	+	+
1008	NCPPB 2424	Europe	1969	T1	CFBP1318	T1	+	+	+	+
1333	CFBP 1321	Europe	1970	T1	CFBP1318	T1	ND	ND	ND	ND
1334	CFBP 1322	Europe	1970	T1	CFBP1318	T1	ND	ND	ND	ND
1335	CFBP 1323	Europe	1971	T1	NCPPB1108	PT21	ND	ND	ND	ND
1337	CFBP 1426	Europe	1972	T1	CFBP1318	T1	ND	ND	ND	ND
1338	CFBP 1427	Europe	1972	T1	CFBP1318	T1	ND	ND	ND	ND
1011	DAR 31861	Australia	1975	T1	NCPPB1108	PT21	+	+	+	+
297	SM78-1	N. America	1978	T1	T1	T1	+	+	+	+
1341	CFBP 2545	Europe	1978	T1	CFBP2545	T1	ND	ND	ND	ND
43	PT 14	N. America	1978	T1	PT14	T1	+	+	-	+
1012	DAR 30555	Australia	1978	T1	PT14	T1	+	+	+	+
1339	CFBP 1916	N. America	1978	T1	PT14	T1	ND	ND	ND	ND

1340	CFBP 1918	N. America	1978	T1	PT14	T1	ND	ND	ND	ND
301	487	Europe	1979	T1	CFBP1318	T1	+	-	+	+
1345	CFBP 6876	Europe	1979	T1	CFBP2545	T1	ND	ND	ND	ND
298	AV80	N. America	1980	T1	T1	T1	+	+	-	+
152	PST 6	N. America	1980	T1	PT14	T1	-	-	-	+
155	PT 18	N. America	1980	T1	PT14	T1	-	+	+	-
158	B181	N. America	1981	T1	PT14	T1	+	+	+	-
294	DCT6D1	N. America	1981	T1	PT14	T1	+	+	+	+
296	188B	N. America	1982	T1	T1	T1	+	+	+	-
919	BS117	N. America	1982	T1	PT14	T1	+	+	+	+
153	PT 17	N. America	1983	T1	T1	T1	+	+	+	+
154	PT 2	N. America	1983	T1	PT14	T1	+	+	+	+
1342	CFBP 4408	Europe	1984	T1	CFBP1318	T1	ND	ND	ND	ND
1006	RG4	S. America	1985	T1	CFBP1318	T1	+	-	+	+
156	T1	N. America	1986	T1	T1	T1	-	-	+	-
1343	CFBP 4409	Europe	1987	T1	CFBP1318	T1	ND	ND	ND	ND
295	DC89-4H	N. America	1989	T1	PT14	T1	+	+	+	+
45	PT 21	N. America	1990	T1	NCPPB1108	PT21	+	+	+	+
48	PT 26	N. America	1990	T1	NCPPB1108	PT21	+	+	+	+
46	PT 23	N. America	1990	T1	LNPV17.41	T1	+	+	+	+
47	PT 25	N. America	1990	T1	LNPV17.41	T1	+	+	+	+
19	OMP-BO 407/91	Europe	1991	T1	LNPV17.41	T1	+	+	+	+
1014	CPST 236	Europe	1993	T1	PT14	T1	-	-	+	+
52	PT 32	N. America	1993	T1	LNPV17.41	T1	+	+	+	+
854	IPV-CT28.31	Europe	1995	T1	IPV-CT28.31	T1	-	+	+	+
14	IPV-BO 2973	Europe	1996	T1	PT14	T1	+	+	-	+
638	OMP-BO 443.1/96	Europe	1996	T1	PT14	T1	+	+	+	+
20	LNPV 17.41	Europe	1996	T1	LNPV17.41	T1	+	+	+	+
824	A9	N. America	1996	T1	LNPV17.41	T1	+	+	+	-
1344	CFBP 5420	Europe	1996	T1	LNPV17.41	T1	ND	ND	ND	ND
825	407	N. America	1997	T1	LNPV17.41	T1	+	+	+	-
1244	CA315	N. America	1998	T1	CA315	PT21	+	+	+	+
17	LNPV 18.76	Europe	1998	T1	LNPV17.41	T1	+	+	+	+
826	838-1	N. America	1998	T1	LNPV17.41	T1	+	+	+	-
1245	316 or 55	N. America	1998	T1	LNPV17.41	T1	+	+	+	+
1223	Pst field 6	N. America	1999	T1	PT14	T1	+	+	+	ND
1218	Pst field 1	N. America	1999	T1	LNPV17.41	T1	+	+	+	ND
1219	Pst field 2	N. America	1999	T1	LNPV17.41	T1	+	+	+	ND
1220	Pst field 3	N. America	1999	T1	LNPV17.41	T1	+	+	+	ND
1221	Pst field 4	N. America	1999	T1	LNPV17.41	T1	+	+	+	ND
1222	Pst field 5	N. America	1999	T1	LNPV17.41	T1	+	+	+	ND
1247	B98 or 57	N. America	1999	T1	LNPV17.41	T1	+	+	+	+
640	ISCI 181	Europe	2002	T1	IPV-CT28.31	T1	+	+	+	+
8	Max 1	Europe	2002	T1	LNPV17.41	T1	-	-	-	+

11	Max 4	Europe	2002	T1	LNPV17.41	T1	+	+	+	+
12	Max 5	Europe	2002	T1	LNPV17.41	T1	+	+	-	+
13	Max 6	Europe	2002	T1	LNPV17.41	T1	-	+	-	+
10	ISCI 78	Europe	2003	T1	LNPV17.41	T1	+	+	-	+
632	KS P 53	Africa	2004	T1	KSP53	T1	-	-	+	+
633	KS 127 M	Africa	2004	T1	KSP53	T1	-	-	+	+
639	ISCI 284	Europe	2004	T1	IPV-CT28.31	T1	+	+	+	+
852	ISCI 286	Europe	2004	T1	IPV-CT28.31	T1	-	+	+	+
853	ISCI 269	Europe	2004	T1	IPV-CT28.31	T1	+	+	+	+
40	K40	N. America	2005	T1	LNPV17.41	T1	+	+	+	+
41	K41	N. America	2005	T1	LNPV17.41	T1	+	+	+	+
100	K100	N. America	2005	T1	LNPV17.41	T1	+	+	+	+
827	CA838-4	N. America	2005	T1	LNPV17.41	T1	+	+	+	-
828	CA838-16	N. America	2005	T1	LNPV17.41	T1	+	+	+	+ (mut.)
1178	CA836-2	N. America	2005	T1	LNPV17.41	T1	+	+	+	-
1181	CA838-8	N. America	2005	T1	LNPV17.41	T1	+	+	+	-
1182	CA838-9	N. America	2005	T1	LNPV17.41	T1	+	+	+	-
1183	CA838-6	N. America	2005	T1	LNPV17.41	T1	+	+	+	-
1020	-	N. America	2008	T1	LNPV17.41	T1	-	-	+	+
1021	-	N. America	2008	T1	LNPV17.41	T1	-	-	+	+
1186	CA410	N. America	2008	T1	LNPV17.41	T1	+	+	+	-
1191	CA20	N. America	2008	T1	LNPV17.41	T1	+	+	+	-
1192	CA21	N. America	2008	T1	LNPV17.41	T1	+	+	+	-
1193	CA22	N. America	2008	T1	LNPV17.41	T1	+	+	+	-
1188	CA16	N. America	2008	T1	LNPV17.41	T1	+	+	+	-
Col3	Colombia198	S. America	2009	T1	Colombia198	T1	ND	ND	ND	ND
Col6	Colombia201	S. America	2008	T1	Colombia198	T1	ND	ND	ND	ND
Col8	Colombia204	S. America	2009	T1	Colombia198	T1	ND	ND	ND	ND
Col1	Colombia338	S. America	2009	T1	Colombia338	T1	ND	ND	ND	ND
Col2	Colombia196	S. America	2009	T1	Colombia338	T1	ND	ND	ND	ND
Col4	Colombia199	S. America	2009	T1	Colombia338	T1	ND	ND	ND	ND
ND: not determined										

Supplementary Table 3.6 Predicted type III effector repertoires of T1-like strains (positions refer to whole genome shotgun sequences deposited at NCBI, besides Max4.) which was not depositied

	K40	K40	K40	LNPV17.41	LNPV17.41	LNPV17.41
	hrp box	**start coord**	**note**	**hrp box**	**start coord**	**note**
avrA1		6070707..6072638-Cterm, complement(6064535..6065341)-Nterm	at contig boundary	operon	complement(5970530..5973250)	
avrD1	unknown	contig_564(89..1024)		24.9	contig_42(487..1422)	
avrE1	16.2	contig_329(2039..6345) and contig_147(1..991)	disrupted by contig break	16.2	contig_58(12230..17531)	
avrPto	23.2	complement contig_134(2555..2979)		23.2	complement contig_218(23460..23884)	
avrRps4	yes-poor	contig_387(193..858)	hrp box below cutoff, but aligned with DC3000	ND	contig_318(1..362)	Likely disrupted by repetitive elements
avrRpt2	15.7	complement contig_540(383..1150)	Identical to T1 version		absent	
hopA1	operon	contig_95(2248..3390)	internal stop codon after 54aa	operon	contig_36(37588..38730)	internal stop codon after 54aa
hopB1	operon	contig_245(13877..14960)	truncated at Nterm but not due to sequence error, internal stop codon near Cterm	operon	contig_22(13881..14964)	truncated at Nterm but not due to sequence error, internal stop codon near Cterm
hopC1-1	21.3	complement contig_518(5054..5863)		21.3	complement contig_108(13877..14686)	
hopC1-2	22.1	complement contig_172(2445..3254)	early stop codon after 78aa	22.1	complement contig_44(2468..3277)	early stop codon after 78aa
hopD1	22.4	contig_549(199..2316)		22.4	contig_98(6944..9061)	
hopF2	operon	complement contig_245(18361..19008)		operon	complement contig_22(18365..19012)	
HopH1	25.3	complement contig_172(1137..1793)		25.3	complement contig_44(1160..1816)	
hopI1	17.5	contig_21(36542..37205) contig_476(1..523)	Split into two contigs	17.5	complement contig_27(1..916) and complement contig_(239030..239391)	Split into two contigs
hopK1		remanent type 1			remanent type 2 shorter than type 1	
hopM1	operon	complement contig_147(1507..3645)	internal stop codon after 154aa	operon	complement contig_58(18049..20187)	internal stop codon after 154aa
hopO1-1	operon?	K40_contig_556(382..1203)	Missing first 30 nt and no start codon	operon?	complement contig_201(1355..2176)	Missing first 30 nt and no start codon
hopO1-2	operon?	complement contig_394(11880..12776)	operon needs to be confirmed	operon?	complement contig_55(180293..181189)	operon needs to be confirmed
hopO1-3	no	complement contig_394(9453..10358)	Does not have the stop codon as DC3000 version, gains additional 133aa into intergenic region and into PSPTO_4591.	no	complement contig_55(177866..178771)	Does not have the stop codon as DC3000 version, gains additional 133aa into intergenic region and into PSPTO_4591.
hopP1	20.5	contig_179(5665..6639)		20.5	contig_93(18000..18974)	
hopQ1-1	24.2	complement contig_549(2432..3774)	frameshift after 353bp (deletion of one bp) resulted in early termination after 128aa, a lot of mutations	24.1	complement contig_98(9177..9950^), complement contig_259(1..477^), contig_261(1..272^), contig_338(347..474^)	fragments across four contigs, all at contig boundries
hopQ1-2		hits hopQ1-1?		no	hits hopQ1-1?	
hopR1	21.6	complement contig_71(24405..30278)		21.6	contig_207(9738..15611)	
hopS1	operon	complement contig_394(13323..13691)	has four additional aa, plus different 2 aa compared to the last two aa of DC3000 version	operon	complement contig_55(181758..182126)	has four additional aa, plus different 2 aa compared to the last two aa of DC3000 version
hopS2	operon	complement contig_394(7858..8391)		operon	complement contig_55(176271..176804)	
hopT1-1	operon	contig_556(1500..2348)	plasmid? Starts with GTG	operon	complement contig_201(210..1358)	plasmid?

hopT1-2	operon	complement contig_394(10705..11874)		operon	complement contig_55(179118..180287)	
hopT2	no	complement contig_394(9048..9431)		no	complement contig_55(177461..177844)	
hopW1	15.9	complement contig_403(1502..3823)	no longer has the stop codon in Pph1448A, became a 773 aa protein, cf. 94 aa	15.9	complement contig_40(91473..92336)	
hopY1	16.2	contig_43(10543..11406)		16.2	complement(2670565..2671428)	
hopAA1-1	17.9	complement contig_147(5695..7155)	different allele internal deletion	17.9	complement contig_58(22345..23805)	
hopAB2	20.2	contig_278(193..1932)	a little longer than DC3000 version (1740bp vs 1662bp)	20.2	complement contig_7(29356..31095)	a little longer than DC3000 version (1740bp vs 1662bp)
hopAE1	23	contig_292(31600..34368)		23	complement contig_55(192500..189732)	
hopAF1	21	complement contig_42(5601..6455)		21	contig_105(3007..3861)	
hopAG1	12.6	complement contig_71(1366..3507)	DC3000 version 459bp vs 2142bp		absent	
hopAH1	operon?	complement contig_71(30..1298)			absent	
hopAH2-1	no	contig_166(9921..11135)		no	contig_14(101430..102644)	
hopAH2-2	no	contig_166(11586..12836)		no	contig_14(99729..100979)	
hopAI1	operon?	contig_93(153..956)			absent	
hopAK1	16	contig_91(21821..23488)		16	complement contig_64(16438..18105)	
hopAN1	no	complement contig_416(8790..10004)		no	complement contig_129(5598..6812)	
hopAS1	15.6	complement contig_168(74813..78901)	identical to T1 version		contig_37(52684..56772)	

Continued:

	NCPPB 1108	NCPPB 1108	NCPPB 1108	T1	T1	T1	Pto Max4	Pto Max4	Pto Max4
	hrp box	**start coord**	**note**	**hrp box**	**start coord**	**note**	**hrp box- not determined for Max4**	**start coord**	**note**
avrA1	?	5974280..5977000	likely separated from hrp box by upstream contig boundary	yes	complement contig00107(20855..23575)	PSPTOT1_4933		contig_247 (63..2783)	
avrD1		absent		yes	complement contig00008(6365..7300)	PSPTOT1_0872		contig_468(333..1268)	
avrE1	16.2	complement config_6(82444..87747)	deletion	yes	contig00005(58109..63412)	PSPTOT1_0672		complement contig_261(1..4311) and contig_161(1..986)	disrupted by contig break
avrPto	23.2	contig_2(39541..39117)			absent			complement contig_241(2493..2917)	
avrRps4		absent			absent			complement contig_803(1627..2292)	
avrRpt2	?	contig_272(236..1003)	Identical to T1 version	yes	contig00026(16745..17512)	PSPTOT1_2469		complement contig_948(206..973)	Identical to T1 version
hopA1	operon	complement contig_96(10475..11617)	internal stop codon after 54aa	operon	contig00028(37689..38831)	PSPTOT1_2529. internal stop codon after 54aa		complement contig_108(7879..9021)	internal stop codon after 54aa
hopB1	operon	contig_279(13877..14960)	truncated at Nterm but not due to sequence error, internal stop codon near Cterm	operon	complement contig00005(30610..31692)	truncated at Nterm but not due to sequence error, internal stop codon near Cterm		contig_198(13858..14941)	truncated at Nterm but not due to sequence error, internal stop codon near Cterm
hopC1-1	21.3	complement contig_86(13873..14682)		yes	contig00011(17945..18754)	PSPTOT1_1097		contig_105(397..1206)	
hopC1-2	22.1	contig_138(256..1065)	early stop codon after 78aa	yes	complement contig00017(2328..2894)	early stop codon after 78aa		contig_609(1100..1909)	early stop codon after 78aa
hopD1	22.4	complement contig_179(1987..4104)		yes	complement contig00034(2254..4371)	PSPTOT1_2827		complement contig_371(89..2206)	frameshift results in early stop; or sequencing error

hopF2	operon	complement contig_279(18361..19008)		operon	complement contig00005(2254..4371)	PSPTOT1_0639. quite different from DC3000 version		complement contig_198 (18989..18342	
HopH1	25.3	contig_138(1717..2373)		yes	complement contig00017(1020..1676)	PSPTOT1_1413		contig_609(2561..3217)	
hopI1	17.5	contig_280(36546..37898)		yes	complement contig00004(297224..298690)	PSPTOT1_0569		complement contig_586(1..681) and complement contig_717(2804..3215)	Split into two contigs
hopK1		absent			absent			absent	
hopM1	operon	contig_6(79791..81928)	**internal frameshift resulted in early stop right after shift**	operon	complement contig00005(63928..66066)	internal stop codon after 154aa		complement contig_161(3642..1504)	internal stop codon after 154aa
hopO1-1	operon?	contig_176(927..1748)	Not annoated in draft genome. Missing first 30 nt and no start codon	operon?	complement contig00087(3985..4806)	Missing first 30 nt and no start codon		contig_485(1..2389)	Same as K40, on contig edge
hopO1-2	operon?	complement contig_158(23120..24016)	operon needs to be confirmed	operon?	complement contig00004(232420..232420)	operon needs to be confirmed		contig_327(1055..1947)	4bp deletion after 496bp resulted in early stop
hopO1-3	no	complement contig_158(20693..21598)	Does not have the stop codon as DC3000 version, gains additional 133aa into intergenic region and into PSPTO_4591.	no	complement contig00004(229993..230898)	PSPTOT1_0506. Does not have the stop codon as DC3000 version, gains additional 133aa into intergenic region and into PSPTO_4591.		contig_327(3469..4372)	Does not have the stop codon as DC3000 version, gains additional 133aa into intergenic region and into PSPTO_4591.
hopP1	20.5	complement contig_120(20370..21344)		yes	contig00119(110073..111047)	PSPTOT1_5231		complement contig_232(6928..7902)	Indentical to K40
hopQ1-1	24.2	contig_179(529..1871)	frameshift after 353bp (deletion of one bp) resulted in early termination after 128aa, a lot of mutations	yes	contig00034(795..2138)	PSPTOT1_2826		contig_869(1..200) and complement contig_694(1..1027)	spans two contigs or more; contig_869 starts at 31bp of DC3000 version
hopQ1-2		hits hopQ1-1?							
hopR1	21.6	contig_36(9713..15586)		yes	complement contig00127(107063..112936)	PSPTOT1_5535		complement contig_401(1935..7808)	identical to K40
hopS1	operon	complement contig_158(24563..24931)	has four additional aa, plus different 2 aa compared to the last two aa of DC3000 version	operon	complement contig00004(233863..234231)	PSPTOT1_0510		contig_327(140..508)	identical to K40
hopS2	operon	complement contig_158(19098..19631)		operon	complement contig00004(228398..228398)	PSPTOT1_0503		contig_327(5436..5969)	
hopT1-1	operon	contig_176(1745..2893)	plasmid?	operon	complement contig00087(2840..3988)	PSPTOT1_4762		contig_485(1215..2348)	identical to K40
hopT1-2	operon	complement contig_158(21945..23114)		possibly operpon	complement contig00004(231245..232414)	PSPTOT1_0507		contig_327(1953..3122)	identical to K40
hopT2	no	complement contig_158(20288..20671)		no	complement contig00004(229588..229971)	PSPTOT1_0505		contig_327(4396..4779)	identical to K40
hopW1	15.9	complement contig_243(4593..6914)		yes	contig00045(4436..6757)	PSPTOT1_3130		contig_525(1050..3371)	
hopY1	16.2	complement contig_161(73056..73919)		yes	complement contig00012(48432..49295)	PSPTOT1_1266		contig_11(782..1645)	identical to K40
hopAA1-1	17.9	contig_6(76173..77633)		yes	complement contig00005(68224..68224)	PSPTOT1_0677. present protein checked		complement contig_377(1^..1359) and complement contig_161(5754..5847^)	broken
hopAB2	20.2	complement contig_112(44653..46392)	a little longer than DC3000 version (1740bp vs 1662bp)	yes	complement contig00018(45202..46941)	PSPTOT1_1462. a little longer than DC3000 version (1740bp vs 1662bp)		complement contig_274(33151..34890)	
hopAE1	23	contig_155(32480..34308)		yes	complement contig00004(241859..244624)	PSPTOT1_0516		complement contig_351(8480..5712)	
hopAF1	21	complement		yes	contig00070(125303..	PSPTOT1_4		complement	identical to K40

		contig_34(..21756)			126157)	398		contig_145(12327..13181)	
hopAG1	12.6	contig_36(36490..38595)		yes	contig00070(125303..126157)	PSPTOT1_4398		complement contig_325(11771..13872)	stop after 701aa
hopAH1	operon	contig_36(38699..39967)		operon	complement contig00127(82682..83950)	PSPTOT1_5516		complement contig_325(10395..11663)	identical to K40
hopAH2-1	no	complement contig_127(34766..35980)		no	complement contig00121(101408..102622)	PSPTOT1_5346		contig_379(13981..15195)	identical to K40
hopAH2-2	no	complement contig_127(33065..34315)		no	complement contig00121(99707..100957)	PSPTOT1_5345		contig_379(15646..16896)	identical to K40
hopAI1	operon?	contig_36(40134..40937)		operon	complement contig00127(81712..82497)	PSPTOT1_5515		complement contig_325(9425..10228)	identical to K40
hopAK1	16	complement contig_91(16432..18099)		yes	complement contig00022(28216..29883)	PSPTOT1_2049		contig_87(8394..10061)	identical to K40
hopAN1	no	complement contig_43(11228..12442)		no	contig00053(5539..6753)	PSPTOT1_3476		contig_43(11228..12442)	identical to K40
hopAS1	15.6	contig_105(422..4174)	1250aa long, C terminus early stop and final few aa different	yes	contig00020(138971..143059)	PSPTOT1_1672		complement contig_330(14..2528) and complement contig_507(5560..7077)	

Figures

Figure 3.1

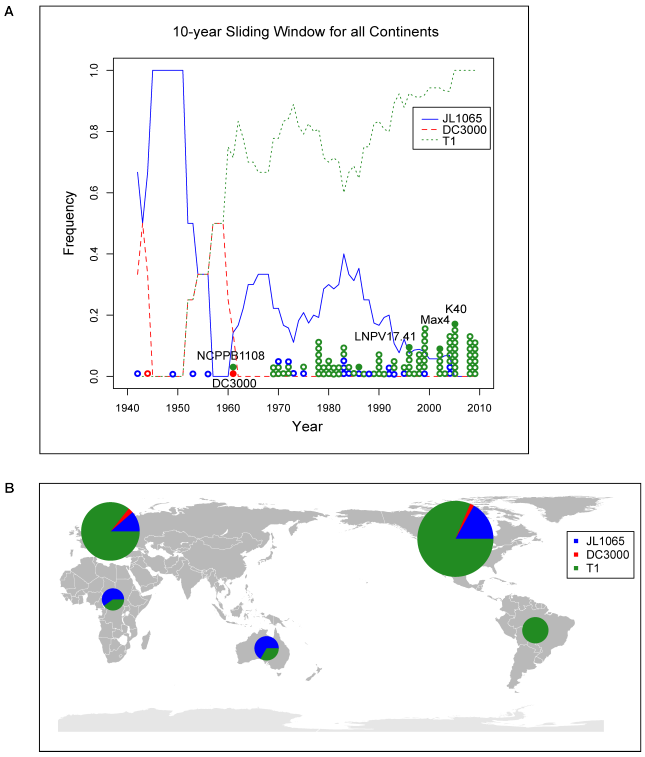


Figure 3.2

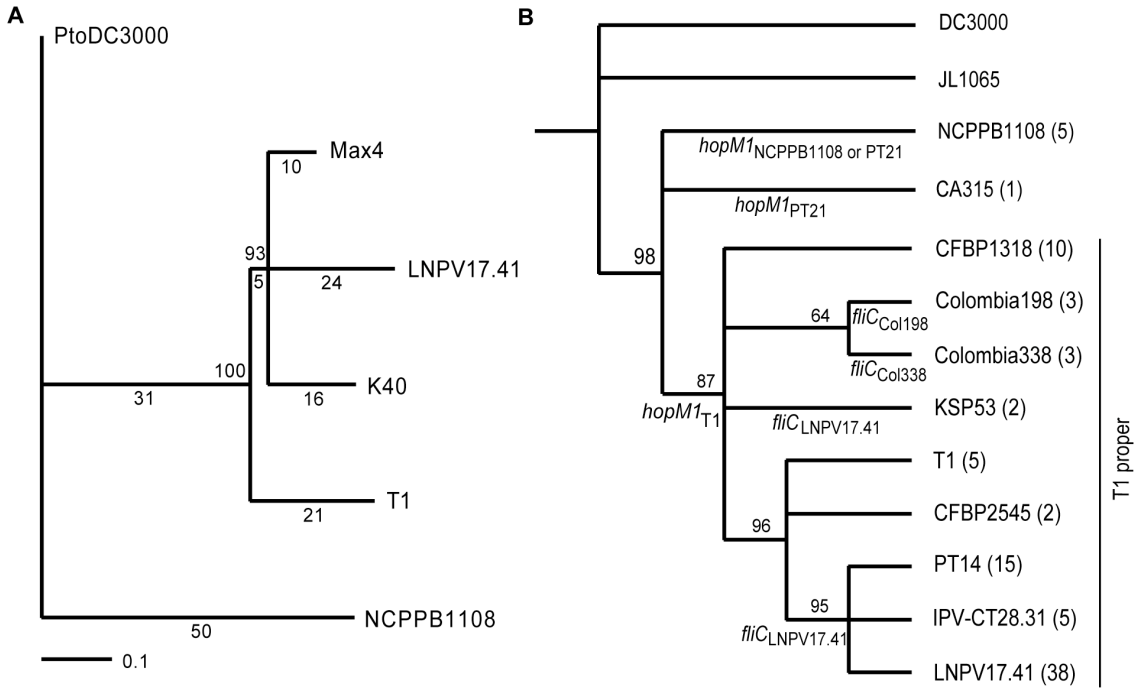

A
PtoDC3000
Max4
10
93
5
LNPV17.41
24
100
31
K40
16
T1
21
NCPPB1108
50
0.1
B
DC3000
JL1065
NCPPB1108 (5)
hopM1NCPPB1108 or PT21
CA315 (1)
hopM1PT21
98
CFBP1318 (10)
Colombia198 (3)
64
fliCCol198
Colombia338 (3)
fliCCol338
87
hopM1T1
KSP53 (2)
fliCLNPV17.41
T1 (5)
CFBP2545 (2)
96
PT14 (15)
95
fliCLNPV17.41
IPV-CT28.31 (5)
LNPV17.41 (38)
T1 proper

Figure 3.3

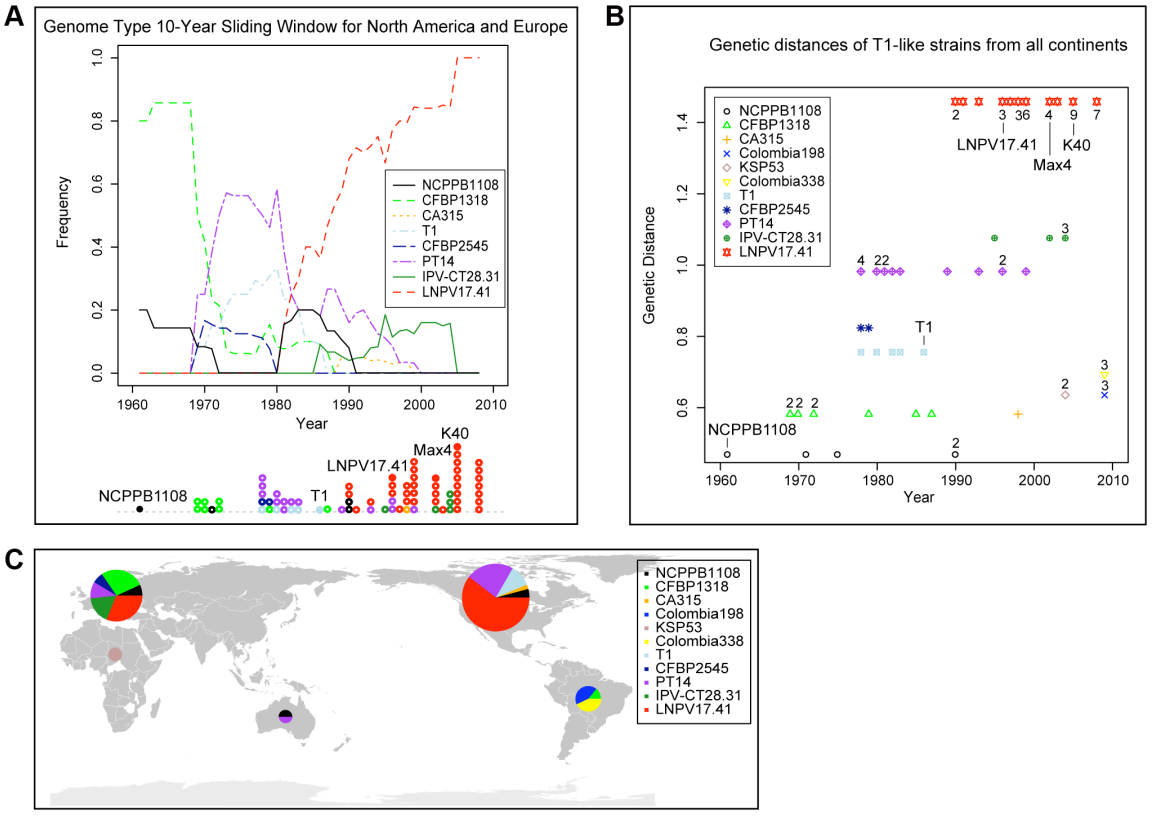

A
Genome Type 10-Year Sliding Window for North America and Europe
Frequency
Year
NCPPB1108
CFBP1318
CA315
T1
CFBP2545
PT14
IPV-CT28.31
LNPV17.41
K40
Max4
B
Genetic distances of T1-like strains from all continents
Genetic Distance
Year
C

Figure 3.4

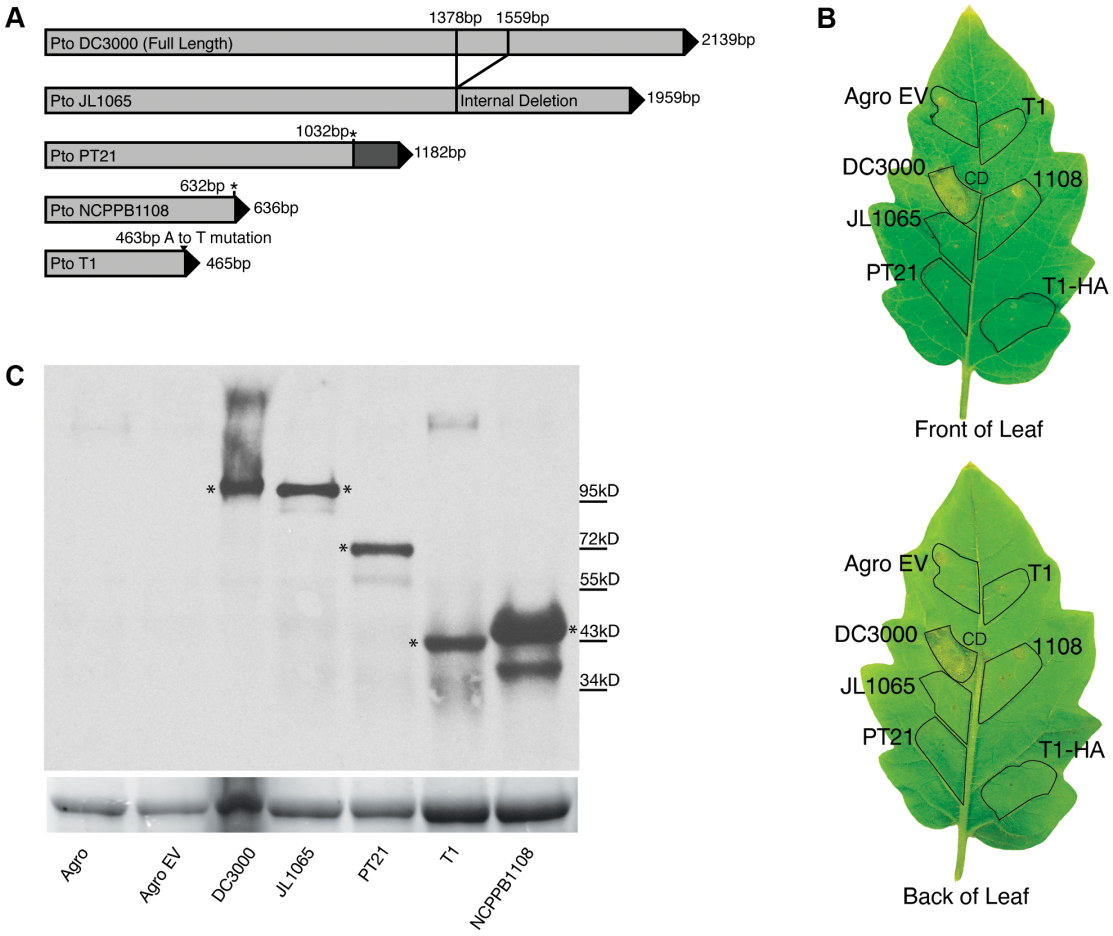
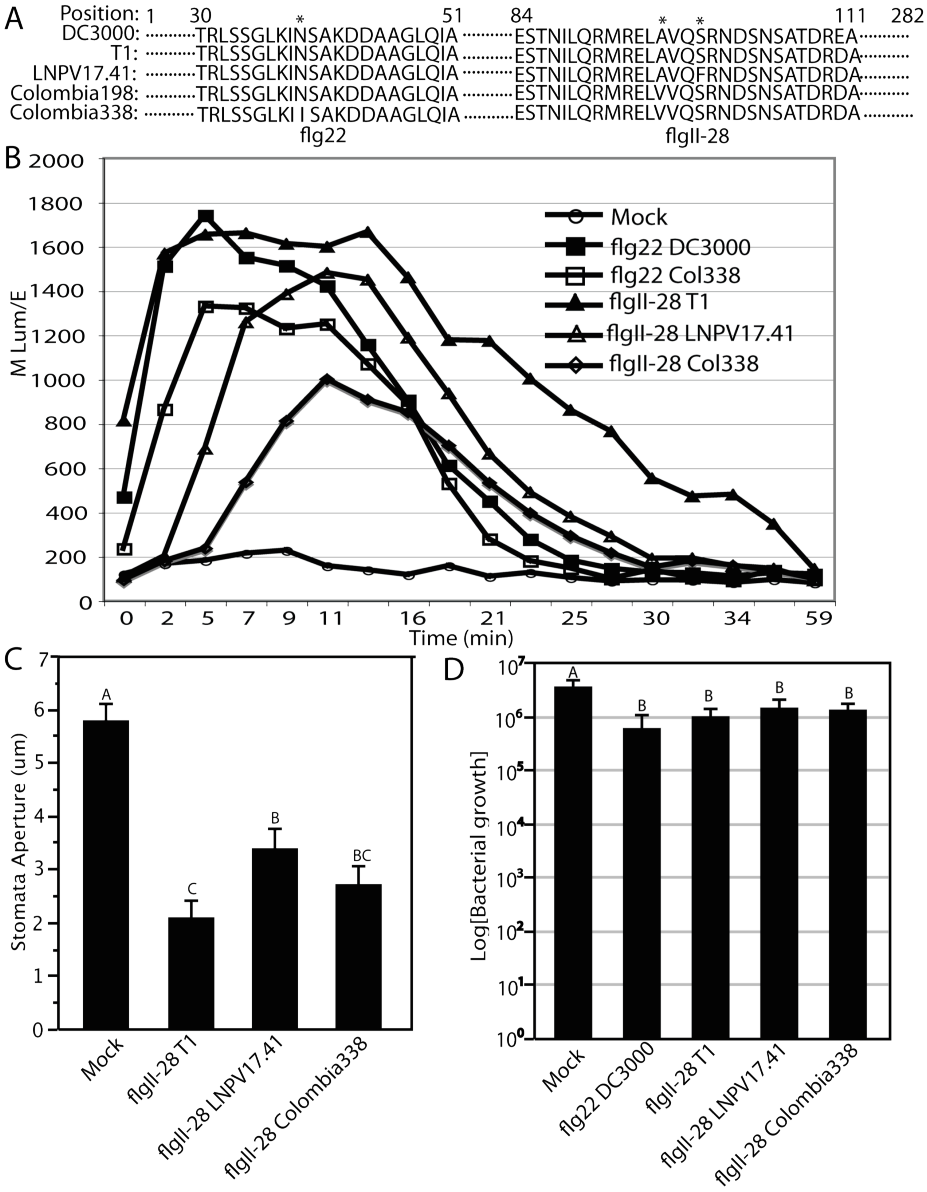

A
1378bp
1559bp
Pto DC3000 (Full Length)
2139bp
Pto JL1065
Internal Deletion
1959bp
1032bp
Pto PT21
1182bp
632bp
Pto NCPPB1108
636bp
463bp A to T mutation
Pto T1
465bp
B
Agro EV
T1
DC3000
CD
1108
JL1065
PT21
T1-HA
Front of Leaf
Agro EV
T1
DC3000
CD
1108
JL1065
PT21
T1-HA
Back of Leaf
C
95kD
72kD
55kD
43kD
34kD
Agro
Agro EV
DC3000
JL1065
PT21
T1
NCPPB1108

Figure 3.5


A
Position: 1 30 * 51 84 * * 111 282
DC3000: TRLSSGLKINSAKDDAAGLQIA ESTNILQRMRELAVQSRNDSNSATDREA
T1: TRLSSGLKINSAKDDAAGLQIA ESTNILQRMRELAVQSRNDSNSATDRDA
LNPV17.41: TRLSSGLKINSAKDDAAGLQIA ESTNILQRMRELAVQFRNDSNSATDRDA
Colombia198: TRLSSGLKINSAKDDAAGLQIA ESTNILQRMRELVVQSRNDSNSATDRDA
Colombia338: TRLSSGLKI I SAKDDAAGLQIA ESTNILQRMRELVVQSRNDSNSATDRDA
flg22
flgII-28
B
M Lum/E
Time (min)
Mock
flg22 DC3000
flg22 Col338
flgII-28 T1
flgII-28 LNPV17.41
flgII-28 Col338
C
Stomata Aperture (um)
Mock
flgII-28 T1
flgII-28 LNPV17.41
flgII-28 Colombia338
D
Log[Bacterial growth]
Mock
flg22 DC3000
flgII-28 T1
flgII-28 LNPV17.41
flgII-28 Colombia338

Chapter 4

Characterization of *Pseudomonas syringae* strains from snow pack and water in the French Alps suggests a critical role for the alpine ecosystem in crop pathogen emergence and evolution

Rongman Cai[1], Caroline L. Monteil[2], Marco E. Mechan Llontop[1], Haijie Liu[1], S. Leman[3], David J. Studholme[4], Cindy E. Morris[2], Boris A. Vinatzer[1†]

[1] Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Latham Hall, Ag Quad Lane, Blacksburg, VA, USA

[2] INRA Centre de Recherche en PACA, Plant Pathology Research Unit, Montfavet, France

[3] Department of Statistics, Virginia Tech, Blacksburg, Virginia, USA

[4] Biosciences, University of Exeter, Exeter, Devon, UK

[†] To whom correspondence should be addressed

Contact: Boris vinatzer, e-mail: vinatzer@vt.edu, phone: +1 540 231 2126, fax: +1 540 231 3347

Running title: Environmental recombination

Abstract

While it is well known that environmental reservoirs harbor emerging human pathogens, less is known about environmental reservoirs of emerging crop pathogens. Isolating bacteria belonging to the *Pseudomonas syringae* group of plant pathogens from snow pack and surface water in the French Alps we have identified genetic lineages never isolated from crops before that are as aggressive on tomato as the most aggressive *P. syringae* strains isolated from tomato. These lineages are closely related to clonal *P. syringae* pathogens isolated from tomato and have undergone extensive homologous recombination with each other and with crop pathogens revealing an epidemic population structure for *P. syringae*. A rich repertoire of type III-secreted effectors also points to environmental strains as potential source of novel virulence genes for already established crop pathogens. The analyzed environmental strains have a relatively wider host range than some of the closely related crop strains but are unable to cause disease on some of the crops on which closely related crop strains can cause disease. This suggests that crop pathogens may have evolved from ancestors similar to environmental strains adapting to specific crops by increasing their fitness on these crops but losing fitness on other plants.

Introduction

When new bacterial crop pathogens or new highly aggressive variants of established bacterial crop pathogens emerge they often spread around the world very fast causing significant economic damage but their evolutionary origins often remain elusive. Recent examples include *Pseudomonas syringae* pathovar (pv.) *aesculi*, causative agent of a devastating bleeding canker of horse chestnut epidemic in Europe [1], *P. syringae* pv. *actinidiae* causative agent of a dramatic kiwifruit canker epidemic in Europe and New Zealand [2], and *Xanthomonas arboricola*, strains of which are causing new diseases on various tree species in Europe [3]. While we have some hints in regard to the geographic origin of these pathogens, we don't know yet who their immediate ancestors were and how these ancestors evolved into highly aggressive crop pathogens. Because of recent advances in DNA sequencing we know that crop pathogens like *P. syringae* pv. *aesculi* or pv. *actinidiae* have large sets of virulence genes, in particular, coding for type III secreted effectors [4]. Comparing crop pathogen genomes with each other, we also know that at least some effector genes were acquired horizontally [5,6]. However, we do not know the identity of the bacteria from which they were acquired from and where and when they were acquired during crop pathogen evolution.

For several bacterial human pathogens, environmental reservoirs are important in regard to their evolution and disease epidemiology. For example, the causal agent of Legionnaire's disease, *Legionella pneumophila,* is ubiquitous in fresh water and has co-evolved with amoeba and the mechanisms for pathogenicity to humans appear to be derived from those to amoeba [7]. As another example, the type III secretion systems of

Pseudomonas aeruginosa and *Vibrio parahaemolyticus* once assumed to be only important for pathogenesis in humans are likely to have important roles for survival in the face of protists [8,9]. Also *Vibrio 174holera*, *Escherichia coli* and *Salmonella enterica*, for example, are well known for having environmental reservoirs important for disease emergence and epidemiology and the genetic and phenotypic diversity in these reservoirs exceeds by far the diversity found in the human hosts [10,11].

Surprisingly, the study of evolution and epidemiology of bacterial plant pathogens has so far almost exclusively focused on agricultural environments [12]. Only recently, characterization of bacteria belonging to the *P. syringae* species complex isolated from headwaters of rivers in North America, Europe, and New Zealand [13] and from precipitation, snow pack, and leaf litter [13,14], has revealed an immense genetic diversity within *P. syringae* outside of agricultural fields suggesting that *P. syringae* life history is connected to the water cycle and that these environments could be important reservoirs harboring emerging crop pathogens. Moreover, based on multilocus sequence typing (MLST) some of the environmental isolates were found to actually belong to the same genetic lineages as strains isolated from crops, in particular cantaloupe [15]. These environmental isolates are also highly aggressive on cantaloupe suggesting that precipitation and surface water may also be direct sources of inoculum for bacterial crop pathogens.

Here we compare environmental strains isolated from snow pack and surface water in the French Alps with very closely related crop pathogens including the genetically monomorphic tomato pathogen *P. syringae* pv. *tomato* (*Pto*) T1 [16]. We

found that some environmental strains are as virulent on tomato as the most virulent *Pto* strains, that their repertoire of genes coding for type III secreted effectors partially overlaps with that of strains isolated from crops, and that extensive recombination occurred between ancestors of crop pathogens and environmental strains. Moreover, host range of the environmental strains most similar to crop strains is relatively wider than that of crop strains. These results are interpreted and discussed in regard to how they may significantly change our current paradigm of bacterial crop pathogen emergence, evolution, and population structure.

Results

Isolation of P. syringae from snow pack and source water of rivers in the French Alps and New Zealand that are closely related to Pto and other crop pathogens

Pseudomonas syringae strains were isolated from snow pack and from headwaters of rivers in France and in New Zealand. The same strains were used already to infer the evolutionary history of the *P. syringae* species complex [13] and to study the contribution of *P. syringae* leaf litter populations to *P. syringae* snow pack populations [14]. Here we chose 71 out of 238 randomly selected strains isolated during these previous studies that, based on the *cts* locus [17], were most similar (over 97.7% DNA identity) to *Pto* strain DC3000, a rifampicin resistant derivative of strain NCPPB1106, the *Pto* type strain [18,19]. To further establish the relationship of these strains with PtoDC3000 the *gyrB* locus [20] of these strains was also sequenced. 26 of the 71 strains resulted over 99% identical at the *gyrB* locus to either PtoDC3000, PtoT1 (a

representative of the most common clonal lineage of *P. syringae* that causes bacterial speck disease of tomato; [16]), or *P. syringae* pv. *spinaceae* and pv. *apii*, which are other crop pathogens belonging to the same *P. syringae* clade as *Pto* [21,22]. Limiting the number of strains with identical *gyrB* sequence to one to four per sample (to avoid analysis of multiple isolates belonging to the same clone) fourteen of these strains were then selected for further analysis (**Table 4.1)**. Several of these strains show 100% DNA identity at the *gyrB* locus with either PtoDC3000 or PtoT1 giving a first indication that the sampled environments may represent reservoirs of current and/or emerging crop pathogens.

Environmental Pto relatives share alleles with clonal P. syringae crop pathogens at several core genome loci

Since we were most interested in the evolution of the most common tomato pathogen represented by the PtoT1 strain, out of the 14 strains listed in Table 4.1 we chose the seven strains that were most similar to T1 and analyzed them by multilocus sequence typing (MLST) [23] using all loci we previously described [21,24]. Three additional loci that we are investigating for their role in plant – *P. syringae* interactions were also sequenced: *fliC* coding for the flagellum subunit FliC, *cheA1* coding for a signaling component of the hypothetical chemotaxis pathway in *P. syringae*, and *cheA2* coding for a signaling component of a second hypothetical chemotaxis pathway in *P. syringae* (Clarke *et al*, in preparation). All loci are described in regard to sequence length, number and percentage of segregating sites, number of alleles, average pair-wise

genetic distance between alleles, Tajima'sD, and dN/S ratio in **Table 4.2**. Comparing loci in Table 4.2 for the reported parameters it becomes obvious that there are significant differences between some of the loci in regard to evolutionary rates and selection pressure. These differences were taken into account when constructing phylogenetic trees (see below).

Figure 4.1 shows an allele table for all loci for all analyzed strains isolated from crops (from now on referred to as "crop strains") and all analyzed strains isolated from snow pack and water (from now on referred to as "environmental strains"). Intriguingly, the two analyzed environmental strains from New Zealand are 100% identical to PtoDC3000 at all sequenced loci strongly suggesting that surface water is an inoculum source of this pathogen. Moreover, observing the distribution of alleles between strains in Figure 4.1 it can be seen that, for example, PtoT1 and PtoDC3000 each share alleles at three different loci with different environmental strains giving a first indication that recombination occurred between ancestors of crop strains and ancestors of environmental strains, which suggests a possibly important role of environmental strains during crop pathogen evolution.

Phylogenetic reconstruction and recombination analysis of environmental strains and crop strains suggests epidemic population structure of P. syringae

To further dig into the evolution of crop strains and environmental strains, we built several Bayesian trees [25] using either the concatenated set of all analyzed loci, a subset of loci, or individual loci. The tree based on all concatenated loci is shown in

Figure 4.2A. A second tree representing the consensus tree of all 13 individual gene trees is shown in **Supplementary Figure 4.1A**. The two trees have almost identical topologies. However, the question is to what degree do these trees reflect the evolutionary history of the analyzed strains? Therefore, we compared trees and data of individual loci with the tree based on the concatenated data set and the consensus tree of all loci using the Shimodaira-Hasegawa (SH) test [26] (**Supp. Table 4.1**). We also carefully compared average genetic distance, selection based on Tajima's D, and dN/dS ratios between genes (from **Table 4.2**). The SH-test clearly shows that *gyrB* data and *cheA1* data do not fit the concatenated tree and the consensus tree. The *fliC* gene is well known to be under selection for avoidance of recognition by the plant immune system [16,27] and, in fact, has a Tajima's D value exceeding 2 and a higher dN/dS ratio than all other genes confirming that this locus does not evolve neutrally (**Table 4.2**). The *gapA* gene on the other hand has a Tajima's D lower than -2 and evolves slower than all other loci. Therefore, we decided to build a second set of trees excluding *gapA*, *gyrB*, *cheA1*, and *fliC* (**Figure 4.2B** and **Supplementary Figure 4.1B**) hypothesizing that such trees might come closer to representing the evolutionary history of the core genome of the analyzed strains. To our surprise, almost all clades of the trees built on all loci (called from now on "13-gene trees") have very high support values while almost all clades of the trees built without *gapA*, *gyrB*, *cheA1*, and *fliC* (called from now on "9-gene trees") have very low support values. However, comparing the two trees with the individual trees for *gapA*, *gyrB*, *cheA1*, and *fliC* (**Supplementary Figure 4.2**) it became clear that the topology of the 13-gene trees is in fact heavily influenced by

cheA1, *fliC*, and *gyrB*. Each of these genes groups a sub-set of strains similarly to the 13-gene trees thereby probably adding statistical support to the clades containing these subsets of strains. For example, the *fliC* tree has a topology resembling the three main clades of the 13-gene trees, *gyrB* has a clade very similar to clade III of the 13-gene trees, and *cheA1* groups Pan126 with PapCFBP2103 as in the 13-gene trees. Other parts of the *cheA1* and *gyrB* tree instead have a very different topology from the 13-gene trees explaining why the *cheA1* and *gyrB* data are incongruent with the 13-gene trees based on the SH-test (**Supplementary Table 4.1**), for example, *cheA1* groups *P. syringae* pv. *spinaceae* ICMP16929 together with PtoT1 while these two strains are in different clades in the 13-gene trees.

Our overall conclusions from these tree comparisons is that the 13-gene trees are not a true representation of the phylogeny of the analyzed strains but are the result of strong (although partially conflicting) phylogenetic signals in *cheA1*, *fliC*, and *gyrB*. Taking these individual genes out of phylogenetic reconstruction, we are left with trees with very low statistical support (the 9-gene trees), which one would expect in the presence of recombination.

In order not to force a tree-like evolutionary history on the analyzed strains but to represent the phylogeny of strains in a way to allow for recombination, phylogenetic networks [28] based on either all genes, or leaving out *gap1*, *gyrB*, *cheA1*, and *fliC*, were also built (**Figure 4.3**). As expected, both networks show a high degree of reticulation indicative of recombination and, as in the phylogenetic trees in **Figure 4.2**,

several strains change their relative position to each other in the two networks highlighting again how *gyrB*, *cheA1*, and *fliC* influence phylogenetic reconstruction.

A population genetic test was then performed to infer the relative contribution of recombination and mutation to the sequence diversity existing at each locus in the combined group of crop and environmental strains using LDhat [29]. Since we had performed the same test previously for 23 out of the 24 analyzed crop lineages [24], we were also interested to see if adding the seven environmental strains changed the ratios of recombination to mutations rate found previously for crop strains only. As can be seen in **Figure 4.4**, for many genes the ratio of recombination to mutation rate is higher in the environmental strains compared to the crop strains and higher when considering crop and environmental strains combined compared to crop strains only. This result confirms that recombination definitely played an important role during the evolution of crop and environmental strains. However, because of the relatively small number of environmental strains so far analyzed the observed differences between crop strains alone, environmental strains alone, and the combined strains must be considered preliminary. Only analyzing a much larger number of environmental strains will it be possible to reliably compare the relative contribution of recombination and mutations to the diversity of the different groups of strains. However, what we feel confident concluding is the following: since the PtoT1 lineage that causes bacterial speck of tomato worldwide is a clonal, genetically monomorphic pathogen [16] but, as shown here, its ancestors clearly recombined with the ancestors of its closest relatives isolated from other crops and from the environment, *P. syringae* should be considered to have

an epidemic population structure consisting of clonal crop pathogens that emerge from recombining populations present in environmental reservoirs.

P. syringae strains from crops and environmental strains have partially overlapping repertoires of genes coding for type III secreted effectors

We sequenced a pool of twelve environmental strains (listed in **Table 4.1**) using Illumina technology [30], assembled all reads, and searched the so obtained assembly (called "environmental" assembly from now on) using all currently confirmed *P. syringae* type III-secreted effectors (http://www.pseudomonas-syringae.org/) as query. Since this environmental assembly contains a mix of twelve different genomes, it would be extremely difficult (not to say impossible) to determine which of these effectors are full-length or not. Therefore, all effector sequences found in the environmental assembly were then simply compared with the repertoire of effector sequences present in the PtoT1 lineage and in PtoDC3000 without considering if an effector sequence represents a functional full length gene or not (**Table 4.3**). Since two of the stains sequenced in bulk belonged to the same genetic lineage as the two strains from New Zealand identical to PtoDC3000 in all loci, it is not too surprising that all PtoDC3000 effectors were found in the environmental assembly. Four effectors present in PtoT1 but not in PtoDC3000 were also found in the environmental assembly (*avrRps4*, *hopAB3*, *hopAE1*, *hopW1*) while four of the effectors unique to the PtoT1 lineage (*avrA1*, *avrD1*, *avrRpt2*, and *hopAW1*) were not found. Even more interestingly, sequences similar to

eight effectors neither present in the PtoT1 lineage nor PtoDC3000 were found: *avrRpm2*, *hopAB1*, *hopAV1*, *hopAZ1*, *hopBB1*, *hopBD2*, *hopX2*, and *hopZ1*.

Since the effector gene *hopM1* plays a particularly important role in *P. syringae* – plant interactions [31], *hopM1* is disrupted in the PtoT1 lineage [16] and in several other sequenced *P. syringae* crop strains [4], and *hopM1* was found to be under diversifying selection possibly to avoid recognition by the plant immune system [4], the *hopM1* gene was PCR amplified and sequenced from all individual crop and environmental strains. Besides the tomato pathogens PtoT1 and PtoJL1065, which were already known not to carry full-length *hopM1* alleles [16], the two barberry strains Pbe ATCC13454 and Pbe CFBP1727 were found to be disrupted by a transposone insertion. Also one of the environmental strains appears to have an incomplete or disrupted *hopM1* gene since the 3' region of *hopM1* in the CSZ0914 strain could not be amplified by PCR. This possibly suggests that environmental strains are also subject to selection from the plant immune system (or from the immune system of yet unknown hosts).

Environmental Pto relatives are highly aggressive on tomato

After finding that environmental strains closely related to PtoT1 share alleles with PtoT1 at several loci, that PtoT1 ancestors and ancestors of environmental strains recombined, and that effector repertoires of environmental strains are overlapping (but not identical) with the effector repertoires of PtoT1 and PtoDC3000 the next big question was: do these strains cause disease on tomato? **Figure 4.5A** shows the answer: all seven environmental strains analyzed here grow to a population size in tomato leaves

within four days after being sprayed on leaf surfaces comparable to the most aggressive tomato strains of the PtoT1 lineage, for example, strain K40 isolated from a tomato seedling in a greenhouse on the Eastern Shore of Virginia in 2005 [16]. But not only population size, also disease symptoms induced by K40 and the environmental strains are similarly devastating (**Figure 4.5B**). **Supplementary Figure 4.3** shows similar results obtained with a second tomato cultivar. We conclude that environmental strains have the potential of becoming emerging tomato pathogens and could cause devastating bacterial speck epidemics in the future.

Environmental strains have a relatively wider host range than the PtoT1 lineage and other closely related crop strains

We had previously shown that PtoT1 and other highly aggressive crop strains have a relatively small host range and hypothesized that crop strains may have reduced their host range as a consequence of adaptation to crops grown in monoculture in agricultural fields while their ancestors may have had a relatively wider host range adapted to life in association with more natural mixed plant communities [24]. We were thus interested to determine the host range of the environmental strains.

Supplementary Figure 4.3 shows growth data and disease symptoms for all tested strains on all tested plant species. In summary, environmental strains cause disease on additional plant species (celery and cauliflower) compared to the PtoT1 lineage although they reach slightly lower population sizes than the adapted strains isolated from these crops. On Arabidopsis and snapdragon the environmental strains do not cause any

disease and grow to similarly low levels as crop strains non-adapted to these crops. These results show that environmental strains appear to have a relatively wider host range than some adapted crop strains like PtoT1 but are not generalists.

Discussion

We recently described the existence of an immense genetic diversity of *P. syringae* in precipitation, snowpack, leaf litter, and headwaters of rivers in North America, Europe, and New Zealand [13,14,15]. Most of the strains found in these environments had never been found on crops before but some were indistinguishable from crop pathogen strains based on MLST [15]. These findings revealed that a) *P. syringae* crop pathogens probably evolved from ancestors adapted to life in non-plant environments, in particular, life in compartments of the water cycle, b) precipitation and surface water may present so far neglected inoculum sources of disease outbreaks caused by *P. syringae*, c) *P. syringae* strains cycle at a yet undetermined rate through the compartments of the water cycle. On the other hand we found that almost all recent bacterial speck outbreaks on tomato around the world are caused by a single genetically-monomorphic lineage, PtoT1 [16], which causes disease on tomato but not on any other tested plant species [24]. We also developed the hypothesis that bacterial crop pathogens like PtoT1 evolved after the advent of agriculture (which allowed adaptation to a single host grown in monoculture) from pathogens with wider host range (adapted to natural mixed-plant communities) but we could not find any evidence for this

hypothesis comparing crop pathogens with each other [24]. However, evidence for a similar hypothesis was found for fungal crop pathogens by comparing crop pathogen isolates to pathogens isolated from wild relatives of crops [32].

Here we complemented and integrated previous results by analyzing environmental strains isolated from snow pack and surface water that are very closely related to PtoT1 and by comparing them with each other, with PtoT1, and with other closely related crop pathogen strains. Astoundingly, out of only 238 selected *P. syringae* strains isolated from only 20 environmental samples at sites in the Southern French Alps we already identified seven new genetic lineages of *P. syringae* that are different from each other but all share at least one allele with PtoT1 and other related crop pathogens at the sequenced MLST loci (**Figure 4.1**). While the same genetic lineages were sometimes found more than once within the same sample in no case were the same lineages identified in two different samples. Therefore, it is evident that these seven strains only represent a first tiny glimpse at the genetic diversity of strains closely related to PtoT1 that exists in the environment.

Even using only seven strains, we identified extensive recombination among environmental strains and between environmental strains and crop strains. At the other extreme, isolates of the PtoT1 lineage collected from cultivated tomato around the world are genetically monomorphic representing a single clonally-expanded genetic lineage. Therefore, we can confidently conclude that while crop strains of *P. syringae* appear to be in fact endemic clonal pathogens as previously proposed [17], *P. syringae* overall has an epidemic population structure similar to *Neisseria meningitidis* [33]: genetic

lineages recombine frequently (probably mostly in non-agricultural environments) and epidemic clones emerge occasionally out of this panmictic population and then expand on a crop and acquire an apparent clonal population structure. Our results here obtained for PtoT1 and related environmental *P. syringae* strains may very well turn out to exemplify what applies to many other bacterial crop pathogens not only within the *P. syringae* group but also within other plant pathogen groups like *Ralstonia solanacearum* and *Xanthomonas*, for which the possible existence of genetic diversity in non-agricultural substrates has been neglected so far.

One sequenced locus deserves particular attention. The gene *fliC* codes for flagellin, the structural subunit of bacterial flagellum, which is intensively studied for its plant immunity-triggering activity [16,27,34]. The fact that the number of alleles at the *fliC* locus is lower than at that at any other sequenced locus, the mean pair-wise genetic distance between alleles is the highest of all loci, and Tajima D for the *fliC* locus is larger than 2 is a clear indication for extensive and recent horizontal gene transfer of a few very different *fliC* alleles between strains and a strong selective advantage of recipients of these *fliC* alleles. Since groups of environmental strains and crop strains share identical *fliC* alleles, we conclude that these groups of strains are exposed to the same selection pressure. The intriguing question is if the different alleles trigger different strengths of plant immunity in different plant species and are the result of host specialization. What we do know is that all sequenced *fliC* alleles are identical in the plant-immunity triggering flg22 epitope [27] and in the key residues of the flgII-28 epitope ([16] and un published). Most differences between alleles are concentrated in a

region C-terminal to flgII-28 (approximately between amino acid positions 150 and 200), which has not yet been investigated for its possible role in triggering plant immunity. Moreover, this region may trigger immunity in other organisms with which *P. syringae* might interact with in the environment, for example, algae, amoeba or insects.

One other locus that attracted our attention is *pgi*. This locus does not stand out for its number of alleles, genetic distance, or selection regime when considering all environmental strains and crop strains together. Moreover, Pgi has no known role in plant-pathogen interactions. Intriguingly though, six of the seven environmental strains from France are identical at the *pgi* locus and the seventh strain has only one mutation compared to the other strains. Therefore, *pgi* has the smallest number of alleles and the lowest average genetic distance between alleles of all loci when considering only environmental strains from France. We hypothesized that a locus genetically linked to *pgi* possibly confers a selective advantage. If this were true, recombinants that acquired a specific *pgi* region including this unknown locus from an unknown donor by recombination would be selected for. Since we found that in the PtoDC3000 genome *pgi* is located only 44,000 bp away from the chemotaxis gene cluster containing *cheA1* and only approximately 50,000-90,000 bp away from a locus containing the III-secreted effector genes *hopAG1*, *hopAH1*, *hopAI1*, *hopR1*, *hopQ1*, and *hopD1*, we looked for linkage between *pgi* and *cheA1* and the presence or absence of these effectors, but we could not identify any (data not shown). We expect future sequencing of the seven individual environmental strains to reveal which locus/loci are linked to *pgi* and what predicted selective advantage they might confer.

The fact that the environmental strains sequenced in bulk have type III effector repertoires that overlap with the effector repertoires of PtoT1 and PtoDC3000 and are at least as aggressive on tomato as PtoDC3000 is intriguing. We were in particular surprised that some of the environmental strains grew to the same population densities as one of the most aggressive strains of the PtoT1 lineage, strain K40, although these strains have a full length *hopM1* gene and do not carry the effector *avrRpt2*, two differences that distinguish the environmental strains from K40 and all similarly aggressive strains of the PtoT1 lineage. Therefore, a disruption of *hopM1* and expression of *avrRpt2* are not required for full virulence in the genetic backgrounds of these environmental strains and disruption of *hopM1* and expression of *avrRpt2* might make these strains even more aggressive than strains of the PtoT1 lineage. Special precautions would need to be taken when experimentally testing this hypothesis since a real risk of creating *Pto* strains more dangerous to tomato production than any *Pto* strain currently existing in tomato fields worldwide would exist. On the other hand, considering that we found effector genes in the environmental bulk assembly that are absent from the PtoT1 genome and considering that recent ancestors of crop strains and environmental strains recombined with each other suggests that current crop strains might acquire new virulence genes from environmental strains through horizontal gene transfer in the future resulting in new variants with increased aggressiveness.

Since even without expressing *avrRpt2* and without disruption of *hopM1* some of the environmental strains are as aggressive on tomato as strains belonging to the PtoT1 lineage, why do we find strains belonging to the PtoT1 lineage on tomato fields around

the world but not the environmental strains? We speculate that these environmental strains may exist at relatively low frequency in the environment and that they may simply have never reached a tomato field at the right time when environmental conditions were favorable for a disease outbreak. Following the same rationale, the PtoT1 lineage may be nothing else than an environmental strain that by chance reached a tomato field when conditions for an outbreak were favorable and than expanded on tomato and spread worldwide becoming a pandemic strain (possibly further increasing in aggressiveness as suggested in our previous studies; [16]). However, there may be limiting factors to becoming a pandemic strain beyond being aggressive on a crop, for example, ability to colonize a crop in the first place, being efficiently transmitted between plants, and finally being efficiently transmitted long distance via seed or vegetative material.

While the environmental strains closely related to PtoT1 are as aggressive on tomato as typical PtoT1 strains, they have a somewhat wider host range than PtoT1 strains (they cause disease on celery and grow to higher population densities in cauliflower leaves). Neither PtoT1 nor the closely related environmental strains can significantly grow or cause any disease on *A. thaliana* or snapdragon. We thus conclude that the environmental strains already adapted to some plant species/families but not others. Moreover, it appears that crop strains might evolve from their ancestors by either restricting their host range when adapting to a single crop, with PtoT1 being such an example, or crop strains might acquire new hosts. For example, *P. syringae* pv. *antirrhini* causes disease on the ornamental snap dragon but none of the closely related

crop strains or environmental strains can grow to similar population densities or cause disease as severe as *P. syringae* pv. *antirrhini* on this plant (see also [24]). Comparing the genome sequence of *P. syringae* pv. *antirrhini* with its close relatives should answer the question of what occurred during the evolution of *P. syringae* pv. *antirrhini* that allowed this strain to acquire its high pathogenic potential on snapdragon.

Another yet unanswered question is why do we find highly aggressive tomato pathogens in the French Alps although the center of diversification of tomato is in South America? Possible hypotheses are that these strains co-evolved with other Solanaceae that are endemic to Europe, that they evolved in South America and traveled long distance through the atmosphere, or that they were imported together with solanaceous crops or even tomato. Only intensive international sampling of the genetic diversity related to PtoT1 that exists in the compartments of the water cycle will allow investigating these hypotheses in the future.

Since strains identical in all MLST loci to DC3000 were found in a creek in New Zealand and these strains are probably also almost identical to DC3000 in regard to their repertoire of type III secreted effectors (based on our bulk genome sequencing results), we conclude that surface water used for irrigation may very well be a significant inoculum source for plant disease caused by current *P. syringae* crop pathogens. When bacterial plant disease outbreaks occur, contaminated seed is usually the prime suspect. While this may very well often be the case, our results point to irrigation (and possibly precipitation) as additional suspects that cannot be neglected any more when investigating the epidemiology of diseases caused by *P. syringae*.

In conclusion, comparing a small number of environmental strains to closely related bacterial crop pathogens has opened our eyes to a genetic diversity that is starting to change our view of crop pathogen emergence, evolution, and epidemiology. More intensive environmental sampling over time and geographic spaces can be expected to answer many more questions. Extending this research to *R. solanacearum*, *Xanthomonas*, and other bacterial plant pathogens may reveal that what we have found for *P. syringae* may become a new paradigm for bacterial plant pathogens in general .

Materials and Methods

Bacterial strains

The environmental *Pseudomonas syringae* strains characterized in this study were isolated from samples described previously [13,14,15]. All but one of the crop strains were described previously [24]. The *P. syringae* pv. *spinaceae* pathotype strain CFBP 5524 was obtained from the French Collection of Plant Associated Bacteria (CFBP) (Angers, France).

Primers, PCR and sequencing of PCR products

Most primers used for PCR were described previously [21,24]. New primes were designed for the *cheA1* locus (cheA1-F: GAAGCCCGTGAGCTGTTG; cheA1-R: CAAATCCGTCAGCGAGAAG), the *cheA2* locus (cheA2-F: TTAGGGAGCACCCCATGA; cheA2-R: GCAACCCCGGATTCAAATA), and for the *fliC* lcous (fliC-F: GCCGGAAGCCACGTAGTA; fliC-R: TCATATCCATGACCATCACCTC). DNA extraction, PCR, and sequencing were performed as previously described [21,24].

Phylogenetic and population genetic analyses

Sequences were processed and uploaded to www.pamdb.org as previously described [35]. Evolutionary models for each individual gene fragments and for concatenated sequences were determined in jModelTest [36] applying the Likelihood Ratio (LR) test and are listed in **Supplementary Table 4.3**. Trees were constructed in MrBayes 3.1.2 [25,37]. Trees based on concatenated gene sequences were constructed using the evolutionary model determined for the entire concatenated sequence (we call so obtained trees "concatenated trees"). Trees based on all sequenced genes or a subset of genes were also constructed using partition-specific evolutionary models after creating partitions corresponding to each individual gene fragment (we called so obtained trees "partition trees"). The number of Markov Chain Monte Carlo (MCMC) iterations used for the concatenated sequences were 10^8 (with a sampling frequency of 1 every 5000 iterations) and for individual gene fragments 10^7 (with a sampling frequency of 1 every 2000 iterations). The burn-in was set to 25%. Convergence was evaluated by drawing the trace plot of recorded values of some parameters from MCMC run in R (http://www.r-project.org/). Consensus trees were obtained in MrBayes for concatenated trees and for individual gene trees and in PAUP for partition trees. To test phylogenetic congruence between trees the Shimodaira–Hasegawa (SH) test [26] was performed in PAUP 4.0 (http://paup.csit.fsu.edu/). Split decomposition analysis was performed using the Neighbornet method in SplitsTree 4 [38]. The same evolutionary models used for plylogenetic tree construction and 10,000 bootstrapping replicates were applied.

MEGA 4.0 was used to calculate genetic distances using the method of Jukes-Cantor [39]. The ratio of non-synonymous to synonymous substitutions (dN/dS) was estimated in PAML [40]. LDhat 2.1 [29,41] was used to estimate the population-scale mutation rate (θ), the population-scale recombination rate (ρ), and Tajima's D.

Plant infections

Plants were grown and infected witas previously described [21,24]. Infections were repeated at least three times for each strain/plant combination.

Bulk genome sequencing

Illumina sequencing was performed at the University of Exeter, Exeter, UK.

References

1. Green S, Studholme DJ, Laue BE, Dorati F, Lovell H, et al. (2010) Comparative Genome Analysis Provides Insights into the Evolution and Adaptation of *Pseudomonas syringae* pv. *aesculi* on *Aesculus hippocastanum*. PLoS ONE 5: e10224.
2. Mazzaglia A, Studholme DJ, Taratufolo MC, Cai R, Almeida NF, et al. (2012) *Pseudomonas syringae* pv. *actinidiae* (PSA) isolates from recent bacterial canker of kiwifruit outbreaks belong to the same genetic lineage. PLoS ONE in press.
3. Hajri A, Pothier JIF, Fischer-Le Saux M, Bonneau S, Poussier Sp, et al. (2012) Type Three Effector Gene Distribution and Sequence Analysis Provide New Insights into the Pathogenicity of Plant-Pathogenic Xanthomonas arboricola. Applied and Environmental Microbiology 78: 371-384.
4. Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, et al. (2011) Dynamic Evolution of Pathogenicity Revealed by Sequencing and Comparative Genomics of 19 *Pseudomonas syringae* Isolates. PLoS Pathog 7: e1002132.
5. Rohmer L, Guttman DS, Dangl JL (2004) Diverse evolutionary mechanisms shape the type III effector virulence factor repertoire in the plant pathogen Pseudomonas syringae. Genetics 167: 1341-1360.
6. Lindeberg M, Cunnac S, Collmer A (2009) The evolution of Pseudomonas syringae host specificity and type III effector repertoires. Mol Plant Pathol 10: 767-775.

7. Albert-Weissenberger C, Cazalet C, Buchrieser C (2007) <i>Legionella pneumophila</i> — a human pathogen that co-evolved with fresh water protozoa. Cellular and Molecular Life Sciences 64: 432-448.

8. Matz C, Moreno AM, Alhede M, Manefield M, Hauser AR, et al. (2008) Pseudomonas aeruginosa uses type III secretion system to kill biofilm-associated amoebae. ISME J 2: 843-852.

9. Matz C, Nouri B, McCarter L, Martinez-Urtaza J (2011) Acquired Type III Secretion System Determines Environmental Fitness of Epidemic <italic>Vibrio parahaemolyticus</italic> in the Interaction with Bacterivorous Protists. PLoS ONE 6: e20275.

10. Vezzulli L, Pruzzo C, Huq A, Colwell RR (2010) Environmental reservoirs of Vibrio cholerae and their role in cholera. Environmental Microbiology Reports 2: 27-33.

11. Winfield MD, Groisman EA (2003) Role of Nonhost Environments in the Lifestyles of Salmonella and Escherichia coli. Applied and Environmental Microbiology 69: 3687-3694.

12. Morris CE, Bardin M, Kinkel LL, Moury B, Nicot PC, et al. (2009) Expanding the Paradigms of Plant Pathogen Life History and Evolution of Parasitic Fitness beyond Agricultural Boundaries. PLoS Pathog 5: e1000693.

13. Morris CE, Sands DC, Vanneste JL, Montarry J, Oakley B, et al. (2010) Inferring the Evolutionary History of the Plant Pathogen *Pseudomonas syringae* from Its Biogeography in Headwaters of Rivers in North America, Europe, and New Zealand. mBio 1.

14. Monteil CL, Guilbaud C, Glaux C, Lafolie F, Soubeyrand S, et al. (2012) Emigration of the plant pathogen Pseudomonas syringae from leaf litter contributes to its population dynamics in alpine snowpack. Environmental Microbiology: no-no.

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

16. Cai R, Lewis J, Yan S, Liu H, Clarke CR, et al. (2011) The plant pathogen Pseudomonas syringae pv. tomato is genetically monomorphic and under strong selection to evade tomato immunity. PLoS Pathog 7: e1002130.

17. Sarkar SF, Guttman DS (2004) Evolution of the core genome of Pseudomonas syringae, a highly clonal, endemic plant pathogen. Appl Environ Microbiol 70: 1999-2012.

18. Cuppels DA (1986) Generation and characterization of Tn5 insertion mutations in Pseudomonas syringae pv. tomato. Appl Environ Microbiol 51: 323-327.

19. Buell CR, Joardar V, Lindeberg M, Selengut J, Paulsen IT, 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.

20. Hwang MS, Morgan RL, Sarkar SF, Wang PW, Guttman DS (2005) Phylogenetic characterization of virulence and resistance phenotypes of Pseudomonas syringae. Appl Environ Microbiol 71: 5182-5191.

21. Yan S, Liu H, Mohr TJ, Jenrette J, Chiodini R, et al. (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.

22. Bull CT, Clarke CR, Cai R, Vinatzer BA, Jardini TM, et al. (2011) Multilocus Sequence Typing of *Pseudomonas syringae sensu lato* confirms previously described genomosoecies and permits rapid idenitfication of *P. syringae* pv. *coriandricola* and *P. syringae* pv. *apii* causing bacterial leaf spot on parsley. Phytopathology.

23. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95: 3140-3145.

24. Cai R, Yan S, Liu H, Leman S, Vinatzer BA (2011) Reconstructing host range evolution of bacterial plant pathogens using Pseudomonas syringae pv. tomato and its close relatives as a model. Infection, Genetics and Evolution 11: 1738-1751.

25. Huelsenbeck J, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754 - 755.

26. Shimodaira H, Hasegawa M (1999) log-likelihoods with application to phylogenetic inference

. Mol Biol Evol 16: 1114-1116.

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

28. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254-267.

29. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160: 1231-1241.

30. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16: 545-552.

31. Nomura K, Debroy S, Lee YH, Pumplin N, Jones J, et al. (2006) A bacterial virulence protein suppresses host innate immunity to cause plant disease. Science 313: 220-223.

32. Stukenbrock EH, McDonald BA (2008) The Origins of Plant Pathogens in Agro-Ecosystems. Annual Review of Phytopathology 46: 75-100.

33. Maynard Smith J, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci U S A 90: 4384-4388.

34. Sun W, Dunning FM, Pfund C, Weingarten R, Bent AF (2006) Within-species flagellin polymorphism in Xanthomonas campestris pv campestris and its impact on elicitation of Arabidopsis FLAGELLIN SENSING2-dependent defenses. Plant Cell 18: 764-779.

35. Almeida NF, Yan S, Cai R, Clarke CR, Morris CE, et al. (2010) PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. Phytopathology 100: 208-215.

36. Posada D (2008) jModelTest: Phylogenetic Model Averaging. Molecular Biology and Evolution 25: 1253-1256.

37. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

38. Huson DH, Kloepper TH (2005) Computing recombination networks from binary sequences. Bioinformatics 21 Suppl 2: ii159-ii165.

39. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599.

40. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24: 1586-1591.

41. Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. Genome Research 17: 1219-1227.

Figure Legends

Figure 4.1. Allele composition at all sequenced loci for all analyzed crop strains and environmental strains.

Figure 4.2. Bayesian consensus trees based on the concatenated set of all thirteen gene fragments listed in Table 4.2 (A). The same gene fragments, but excluding *cheA1*, *fliC*, *gyrB* and *gap1*, were used to construct the tree shown in (B). *P. syringae* pv. *syringae* (Psy) B728a was used as outgroup. Clade credibility values are given above branches. The same gene fragments were also used to build consensus trees without concatenation by defining gene fragments as separate partitions with their own evolutionary models. (**Supplementary Figure 4.1**).

Figure 4.3. Split decomposition analysis of all thirteen concatenated core genome gene fragments from Table 4.2 (A) and of the same genes fragments but excluding *cheA1*, *fliC*, *gyrB* and *gap1* (B). Bootstrap values higher than 50 are shown.

Figure 4.4. The ratios of population-scale recombination rate (ρ) to population-scale mutation rate (θ) for crop strains (white), environmental strains (gray) and all crop and environmental strains (black) were estimated in LDhat 2.1. All numerical values used in Figure 4.4 are listed in Supplementary Table 4.2.

Figure 4.5. Bacterial growth (A) and disease symptoms (B) on tomato (cultivar 'Rio Grande'). (A) Bacterial population sizes were determined four days post infection. Population sizes are indicated as colony forming units/cm^2 on a log scale. Pto strains DC3000 and K40 were used as examples of aggressive strains and *P. syringae* pv. *maculicola* M6 was used as example of a strain that does not cause disease on tomato. (B) Pictures of disease symptoms were taken four days post infection.

Supplementary Figure 4.1. Bayesian consensus trees based on all thirteen gene fragments listed in Table 4.2 treating gene fragments as partitions with their own evolutionary models (A). The same gene fragments, but excluding *cheA1*, *fliC*, *gyrB* and *gap1*, were used to construct the tree shown in (B). *P. syringae* pv. *syringae* (Psy) B728a was used as outgroup. Clade credibility values are given above branches.

Supplementary Figure 4.2. Bayesian trees for each individual gene fragment listed in Table 4.2.

Supplementary Figure 4.3. Bacterial growth and disease symptoms on tomato (cultivar 'Sunpride'), *A. thaliana* (ecotypes 'Mt' and 'Col'), celery, cauliflower and snapdragon.

Tables

Table 4.1. Strains used in this study were collected between 2007 and 2010 in the indicated geographic locations and from the indicated substrates and showed high identity to either PtoT1, PtoDC3000, *P. syringae* pv. *spinaceae* or *P. syringae* pv. *apii*.

strain	Geographic location	substrate	% DNA identity in *gyrB* to strain	MLST	Bulk genome sequencing
SZ-003	France, Alpes-de-Haute, Sauze, east branch	surface water	99.8% spinaceae PT[1]	yes	yes
SZ-014	France, Alpes-de-Haute, Sauze, east branch	surface water	100% PtoT1	yes	yes
SZ-015	France, Alpes-de-Haute, Sauze, east branch	surface water	99.8% spinaceae PT	no	yes
SZ-135	France, Alpes-de-Haute, Sauze, source	surface water	99.8% apii &PtoT1	yes	yes
CSZ0223	France, Alpes-de-Haute-, Soudane, source	surface water	99.8% apii &PtoT1	yes	yes
CSZ0292	France, Alpes-de-Haute-, Sauze meadow	snow pack	100% apii PT	yes	yes
CSZ0295	France, Alpes-de-Haute, Sauze meadow	snow pack	100% apii PT	no	yes
CSZ0326	France, Alpes-de-Haute, Sauze meadow	snow pack	100% apii PT	no	yes
CSZ0914	France, Alpes-de-Haute, Sauze meadow	snow pack	100% PtoT1	yes	yes
CCV0611	France, Alpes-de-Haute, Col-de-Vars meadow	snow pack	100% PtoT1	yes	no
AI-001	New Zealand, Central Otago, South Island, Schoolhouse Creek	surface water	100% PtoDC3000	yes	yes
AI-056	Zealand, Central Otago, South Island, Schoolhouse Creek	surface water	100% PtoDC3000	no	yes
AI-088	Zealand, Central Otago, South Island, Schoolhouse Creek	surface water	100% PtoDC3000	no	yes
AI-103	Zealand, Central Otago, South Island, Schoolhouse Creek	surface water	100% PtoDC3000	yes	no

[1] PT: pathotype

Table 4.2. Sequenced loci, their length, number and % of segregating sites, number of alleles, average pairwise genetic distance calculated with the Jukes-Cantor method, Tajima's D, and ratio of non-synonymous (dN) to synonymous (dS) mutations. Alleles can be searched and downloaded from www.pamdb.org

Locus name	Length	Number of segregating sites	Ratio segregating sites	N. of alleles	Jukes-Cantor genetic distance	Tajima's D	dN/dS
acnB	555	17	0.03	13	0.007(0.002)	-0.201	0.0001
CheA1	597	40	0.07	13	0.018(0.003)	0.019	0.1780
CheA2	588	9	0.02	9	0.003(0.001)	-0.8	0.1137
fliC	846	47	0.06	7	0.023(0.004)	2.269	0.2208
gap1	600	19	0.03	10	0.003(0.001)	-2.02	0.0569
gltA	504	7	0.01	9	0.005(0.002)	0.789	0.0001
gyrB	696	18	0.03	14	0.007(0.002)	0.034	0.0164
kup	1059	55	0.05	21	0.009(0.002)	-1.09	0.0246
pgi	564	14	0.02	9	0.005(0.001)	-0.752	0.0207
PSPTOT1_0038	597	17	0.03	10	0.008(0.002)	0.466	0.0142
PSPTOT1_2359	498	19	0.04	15	0.010(0.003)	0.152	0.0864
PSPTOT1_1665	435	15	0.03	12	0.006(0.002)	-0.996	0.0392
rpoD	636	24	0.04	17	0.013(0.003)	1.124	0.0237
Concatenated genes	8175	301	0.04	31			

Table 4.3. Comparison of repertoires of DNA sequences with homology to type III-secreted effectors in twelve environmental strains (France and New Zealand) sequenced in bulk, PtoDC3000, and PtoT1 (including additional genomes of the same genetic lineage; ref)

Strains	Effector sequences present
Environmental strains, PtoDC3000, and PtoT1 lineage	*avrE1 avrPto1 hopA1 hopB1 hopC1 hopD1 hopF2 hopH1 hopI1 hopM1 hopO1-1 hopQ1-1 hopR1 hopS1 hopS2 hopT1 hopT2 hopY1 hopAA1 hopAF1 hopAG1 hopAH1 hopAI1 hopAS1*
Environmental strains and PtoDC3000 only	*hopE1 hopG1 hopK1 hopN1 hopU1 hopV1 hopX1 hopAB2 hopAD1 hopAM1 hopAO1 hopAT1*
Environmental strains and Pto T1 lineage only	*avrRps4 hopAB3 hopAE1 hopW1*
Environmental strains only	*avrRpm2 hopAB1 hopAV1 hopAZ1 hopBB1 hopBD2 hopX2 hopZ1*
PtoT1 lineage only	*avrA1 avrD1 avrRpt2 hopAW1*
PtoDC3000 only	none
PtoT1 lineage and PtoDC3000 only	none

Supplementary Table 4.1. Shimodaira–Hasegawa (SH) test [26].

MrBayes Trees	loci													
	acnB	gltA	gyrB	pgi	38	1665	2359	rpoD	gap1	kup	CheA1	CheA2	fliC	Concatenate
acnB	0.671	0.038	0.000	0.015	0.001	0.011	0.000	0.000	0.086	0.000	0.000	0.042	0.000	0.000
gltA	0.004	0.477	0.000	0.016	0.001	0.010	0.000	0.000	0.052	0.000	0.000	0.044	0.000	0.000
gyrB	0.001	0.083	**best**	0.005	0.001	0.027	0.000	0.000	0.081	0.000	0.000	0.031	0.000	0.000
pgi	0.153	0.045	0.000	**best**	0.001	0.002	0.005	0.005	0.053	0.005	0.000	0.040	0.000	0.000
PSPTOT1_0038	0.001	0.023	0.001	0.007	**best**	0.004	0.000	0.000	0.056	0.000	0.000	0.013	0.000	0.000
PSPTOT1_1665	0.006	0.039	0.000	0.007	0.001	**best**	0.000	0.000	0.088	0.000	0.000	0.030	0.000	0.000
PSPTOT1_2359	0.043	0.126	0.000	0.008	0.000	0.006	**best**	0.000	0.048	0.000	0.000	0.031	0.000	0.000
rpoD	0.006	0.080	0.000	0.017	0.001	0.025	0.001	**best**	0.047	0.000	0.000	0.034	0.000	0.000
gap1	0.006	0.090	0.000	0.012	0.001	0.016	0.000	0.000	**best**	0.000	0.000	0.034	0.000	0.000
kup	0.134	0.268	0.001	0.019	0.000	0.021	0.155	0.000	0.089	**best**	0.000	0.060	0.000	0.000
CheA1	0.018	0.558	0.000	0.032	0.000	0.009	0.000	0.000	0.052	0.001	**best**	0.045	0.000	0.000
CheA2	0.010	0.046	0.000	0.021	0.000	0.009	0.000	0.000	0.057	0.000	0.000	**best**	0.000	0.000
fliC	0.002	0.271	0.000	0.005	0.000	0.008	0.004	0.000	0.059	0.000	0.000	0.034	**best**	0.000
Concatenate	**best**	0.829	**0.001**	0.225	0.261	0.226	0.343	0.127	0.369	0.101	**0.000**	0.249	0.559	**best**
Partition	0.846	**best**	**0.001**	0.194	0.267	0.226	0.360	0.100	0.369	0.115	**0.000**	0.257	0.704	0.898

Supplementary Table 4.2. Estimates of population recombination rate (ρ) and population mutation rate (θ) for all loci calculated in LDhat [29,41].

genes	Crop strains			Environmental strains			Crop & environmental strains		
	per site ρ	per site θ	ρ/θ	per site ρ	per site θ	ρ/θ	per site ρ	per site θ	ρ/θ
acnB	0.0055	0.0082	0.7	0.1019	0.0022	46.1	0.0091	0.0077	1.2
gltA	0.0080	0.0032	2.5	0.0241	0.0041	5.9	0.0160	0.0035	4.6
gyrB	0.0073	0.0065	1.1	0.0160	0.0059	2.7	0.0131	0.0065	2.0
pgi	0.0000	0.0067	0.0	na[1]	na[1]	na[1]	0.0000	0.0062	0.0
PSPTOT10038	0.0085	0.0076	1.1	0.0034	0.0062	0.5	0.0220	0.0071	3.1
PSPTOT11665	0.0000	0.0080	0.0	0.0070	0.0056	1.2	0.0046	0.0086	0.5
PSPTOT12359	0.0101	0.0091	1.1	0.0000	0.0057	0.0	0.0122	0.0096	1.3
rpoD	0.0508	0.0093	5.5	0.0079	0.0109	0.7	0.0747	0.0095	7.9
gap1	0.0051	0.0031	1.6	0.0000	0.0095	0.0	0.0000	0.0079	0.0
kup	0.0048	0.0134	0.4	0.0239	0.0054	4.4	0.0095	0.0130	0.7
CheA1	0.0221	0.0158	1.4	0.0034	0.0158	0.2	0.0306	0.0169	1.8
CheA2	0.0000	0.0028	0.0	0.0297	0.0049	6.0	0.0105	0.0039	2.7
fliC	0.0000	0.0149	0.0	0.0000	0.0169	0.0	0.0000	0.0139	0.0
mean	0.0094	0.0084	1.2	0.0181	0.0078	5.7	0.0156	0.0088	2.0

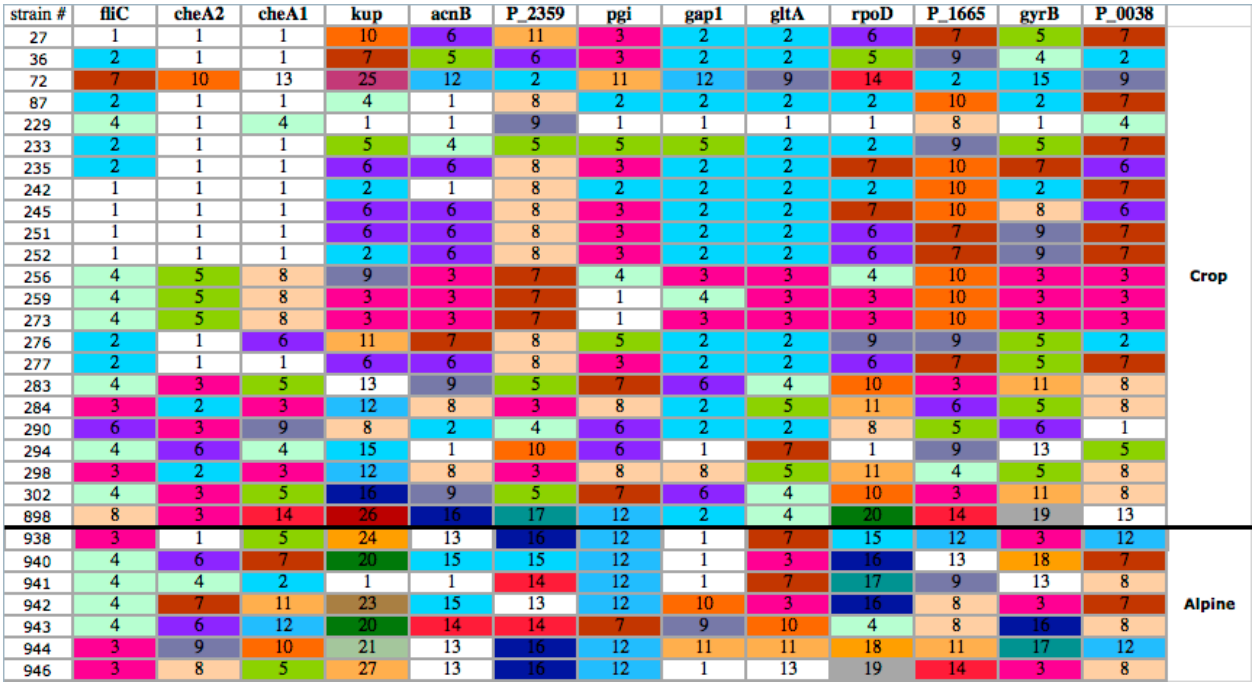
1 not possible to calculate because only one segregating site.

Supplementary Table 4.3. Molecular evolutionary models of each gene for all strains

Locus	Model	nst	rates
acnB	GTR+G	6	gamma
cheA1	HKY+G	2	gamma
cheA2	GTR+I	6	equal
fliC	GTR+I	6	equal
gap1	HKY+I	2	equal
gltA	GTR+G	6	gamma
gyrB	GTR+G	6	gamma
kup	GTR+I+G	6	gamma
pgi	GTR+I	6	equal
PSPTOT1_0038	GTR+I	6	equal
PSPTOT1_2359	HKY+G	2	gamma
PSPTOT1_1665	HKY+G	2	gamma
rpoD	GTR+I+G	6	gamma
Concatenated	GTR+G	6	gamma

Figures

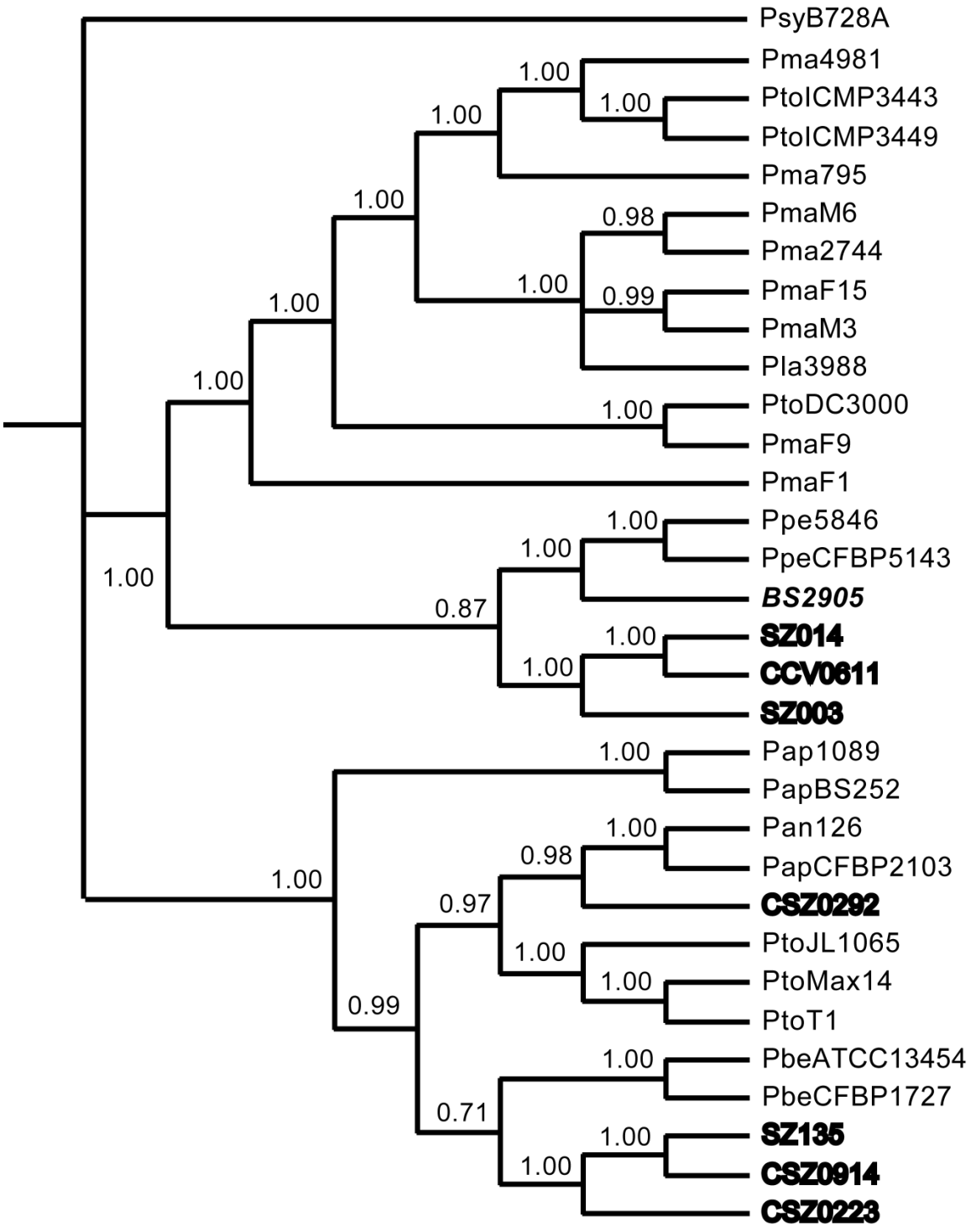
Figure 4.1



strain #	fliC	cheA2	cheA1	kup	acnB	P_2359	pgi	gap1	gltA	rpoD	P_1665	gyrB	P_0038	
27	1	1	1	10	6	11	3	2	2	6	7	5	7	
36	2	1	1	7	5	6	3	2	2	5	9	4	2	
72	7	10	13	25	12	2	11	12	9	14	2	15	9	
87	2	1	1	4	1	8	2	2	2	2	10	2	7	
229	4	1	4	1	1	9	1	1	1	1	8	1	4	
233	2	1	1	5	4	5	5	5	2	2	9	5	7	
235	2	1	1	6	6	8	3	2	2	7	10	7	6	
242	1	1	1	2	1	8	2	2	2	2	10	2	7	
245	1	1	1	6	6	8	3	2	2	7	10	8	6	
251	1	1	1	6	6	8	3	2	2	6	7	9	7	
252	1	1	1	2	6	8	3	2	2	6	7	9	7	
256	4	5	8	9	3	7	4	3	3	4	10	3	3	Crop
259	4	5	8	3	3	7	1	4	3	3	10	3	3	
273	4	5	8	3	3	7	1	3	3	3	10	3	3	
276	2	1	6	11	7	8	5	2	2	9	9	5	2	
277	2	1	1	6	6	8	3	2	2	6	7	5	7	
283	4	3	5	13	9	5	7	6	4	10	3	11	8	
284	3	2	3	12	8	3	8	2	5	11	6	5	8	
290	6	3	9	8	2	4	6	2	2	8	5	6	1	
294	4	6	4	15	1	10	6	1	7	1	9	13	5	
298	3	2	3	12	8	3	8	8	5	11	4	5	8	
302	4	3	5	16	9	5	7	6	4	10	3	11	8	
898	8	3	14	26	16	17	12	2	4	20	14	19	13	
938	3	1	5	24	13	16	12	1	7	15	12	3	12	
940	4	6	7	20	15	15	12	1	3	16	13	18	7	
941	4	4	2	1	1	14	12	1	7	17	9	13	8	
942	4	7	11	23	15	13	12	10	3	16	8	3	7	Alpine
943	4	6	12	20	14	14	7	9	10	4	8	16	8	
944	3	9	10	21	13	16	12	11	11	18	11	17	12	
946	3	8	5	27	13	16	12	1	13	19	14	3	8	

Figure 4.2

A


PsyB728A
Pma4981
PtoICMP3443
PtoICMP3449
Pma795
PmaM6
Pma2744
PmaF15
PmaM3
Pla3988
PtoDC3000
PmaF9
PmaF1
Ppe5846
PpeCFBP5143
BS2905
SZ014
CCV0611
SZ003
Pap1089
PapBS252
Pan126
PapCFBP2103
CSZ0292
PtoJL1065
PtoMax14
PtoT1
PbeATCC13454
PbeCFBP1727
SZ135
CSZ0914
CSZ0223
1.00
1.00
1.00
1.00
1.00
0.98
1.00
0.99
1.00
1.00
1.00
1.00
1.00
0.87
1.00
1.00
1.00
1.00
0.98
1.00
0.97
1.00
1.00
0.99
1.00
0.71
1.00
1.00

B

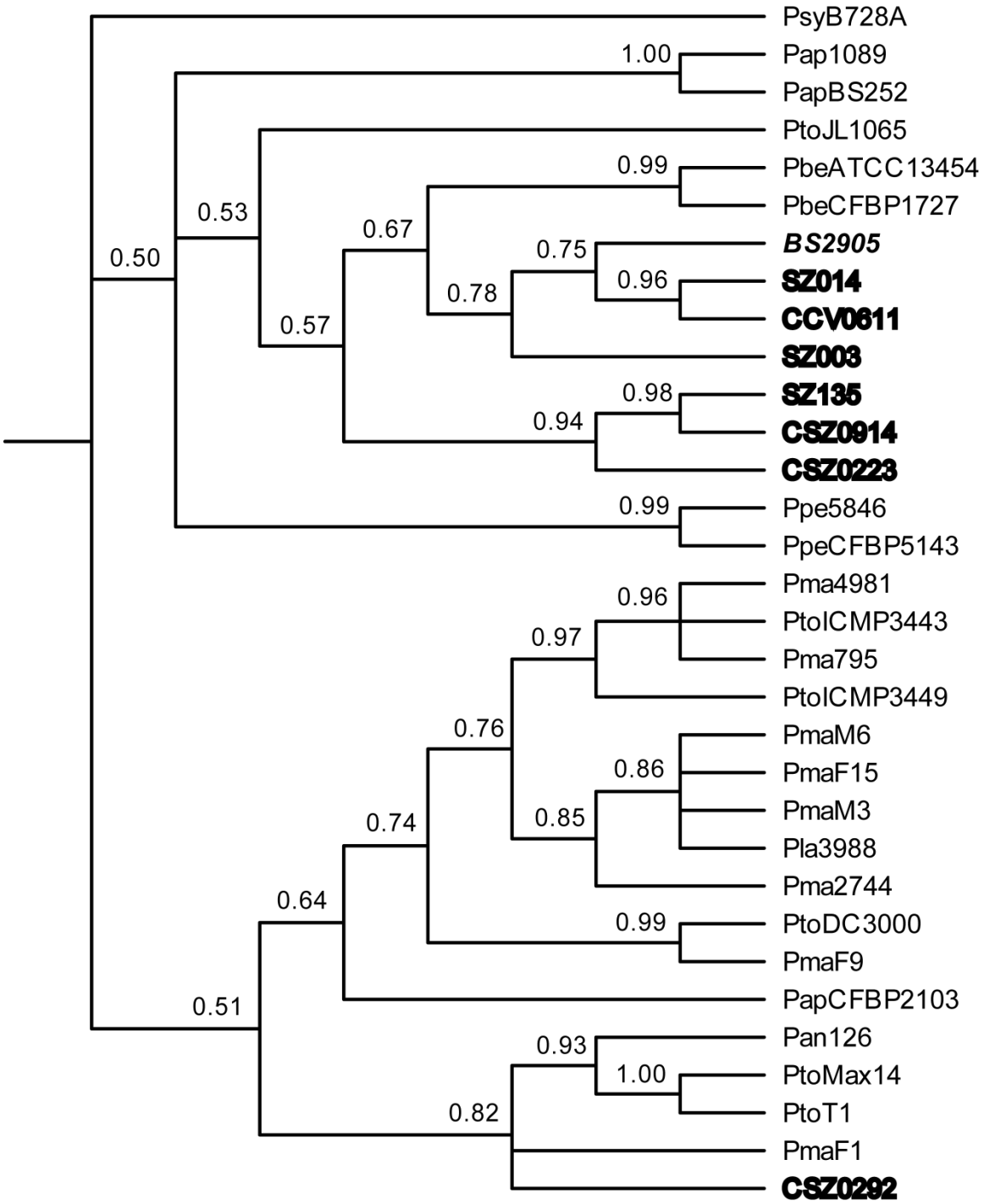
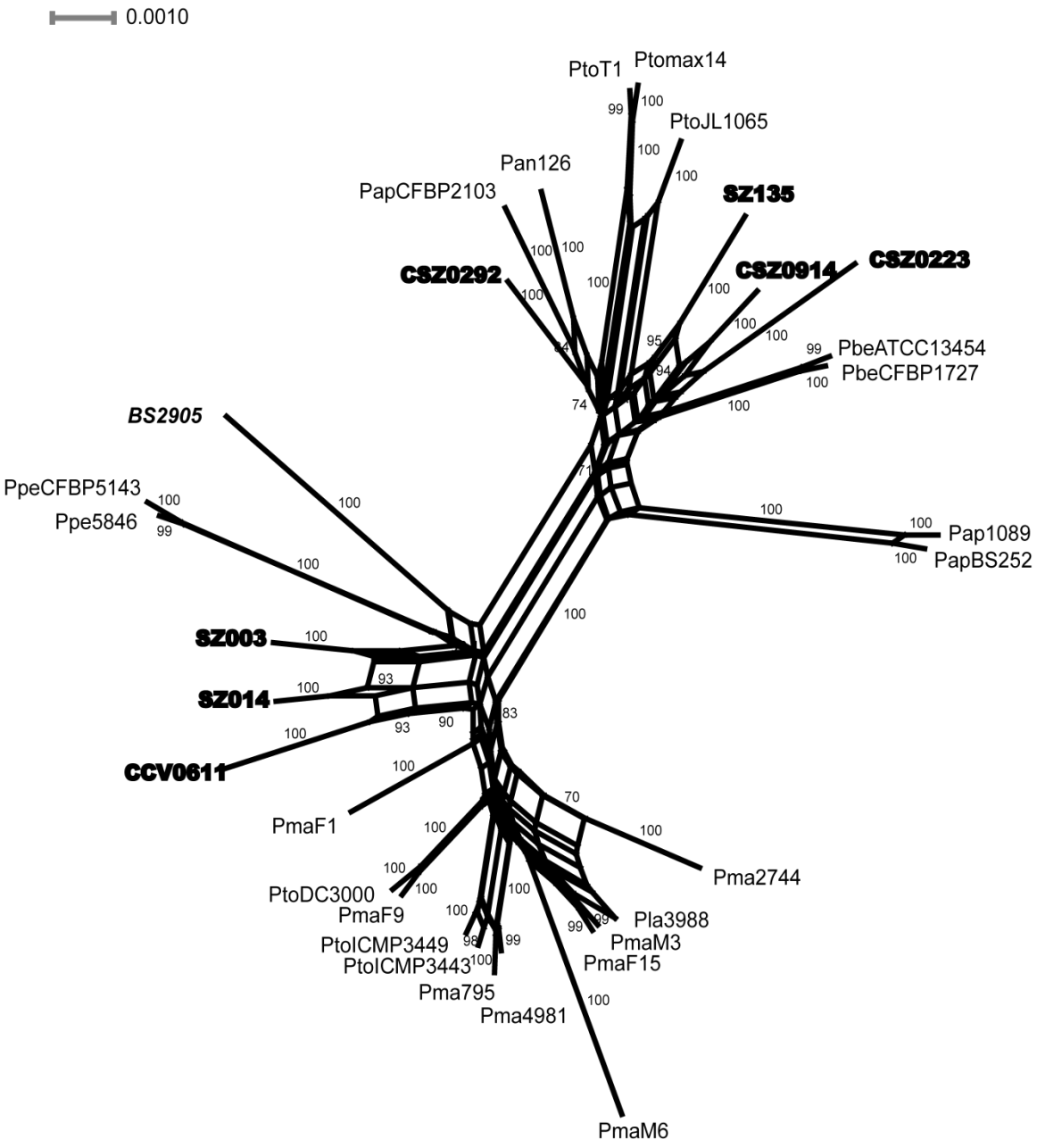

PsyB728A
1.00
Pap1089
PapBS252
PtoJL1065
0.99
PbeATCC13454
PbeCFBP1727
0.53
0.67
0.75
BS2905
0.50
0.96
SZ014
0.78
CCV0611
0.57
SZ003
0.98
SZ135
0.94
CSZ0914
CSZ0223
0.99
Ppe5846
PpeCFBP5143
Pma4981
0.96
PtoICMP3443
0.97
Pma795
PtoICMP3449
0.76
PmaM6
0.86
PmaF15
0.74
0.85
PmaM3
Pla3988
Pma2744
0.64
0.99
PtoDC3000
PmaF9
0.51
PapCFBP2103
0.93
Pan126
1.00
PtoMax14
0.82
PtoT1
PmaF1
CSZ0292

Figure 4.3

A


0.0010
PtoT1
Ptomax14
PtoJL1065
Pan126
PapCFBP2103
SZ135
CSZ0292
CSZ0914
CSZ0223
PbeATCC13454
PbeCFBP1727
BS2905
PpeCFBP5143
Ppe5846
Pap1089
PapBS252
SZ003
SZ014
CCV0611
PmaF1
Pma2744
PtoDC3000
PmaF9
Pla3988
PtoICMP3449
PtoICMP3443
PmaM3
PmaF15
Pma795
Pma4981
PmaM6

B

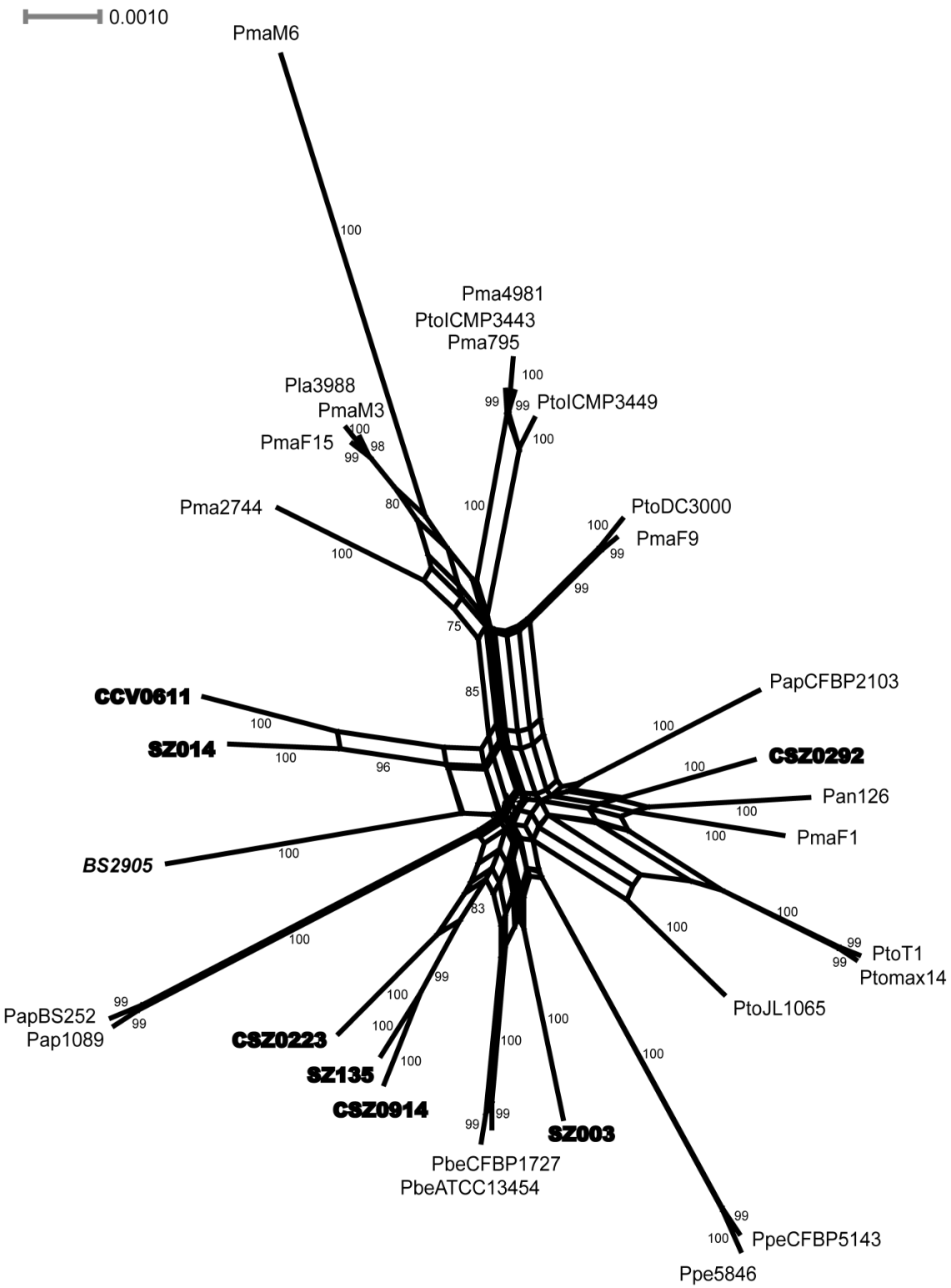

0.0010
PmaM6
Pma4981
PtoICMP3443
Pma795
Pla3988
PmaM3
PmaF15
PtoICMP3449
Pma2744
PtoDC3000
PmaF9
CCV0611
SZ014
PapCFBP2103
CSZ0292
Pan126
PmaF1
BS2905
PtoT1
Ptomax14
PtoJL1065
PapBS252
Pap1089
CSZ0223
SZ135
CSZ0914
SZ003
PbeCFBP1727
PbeATCC13454
PpeCFBP5143
Ppe5846

Figure 4.4

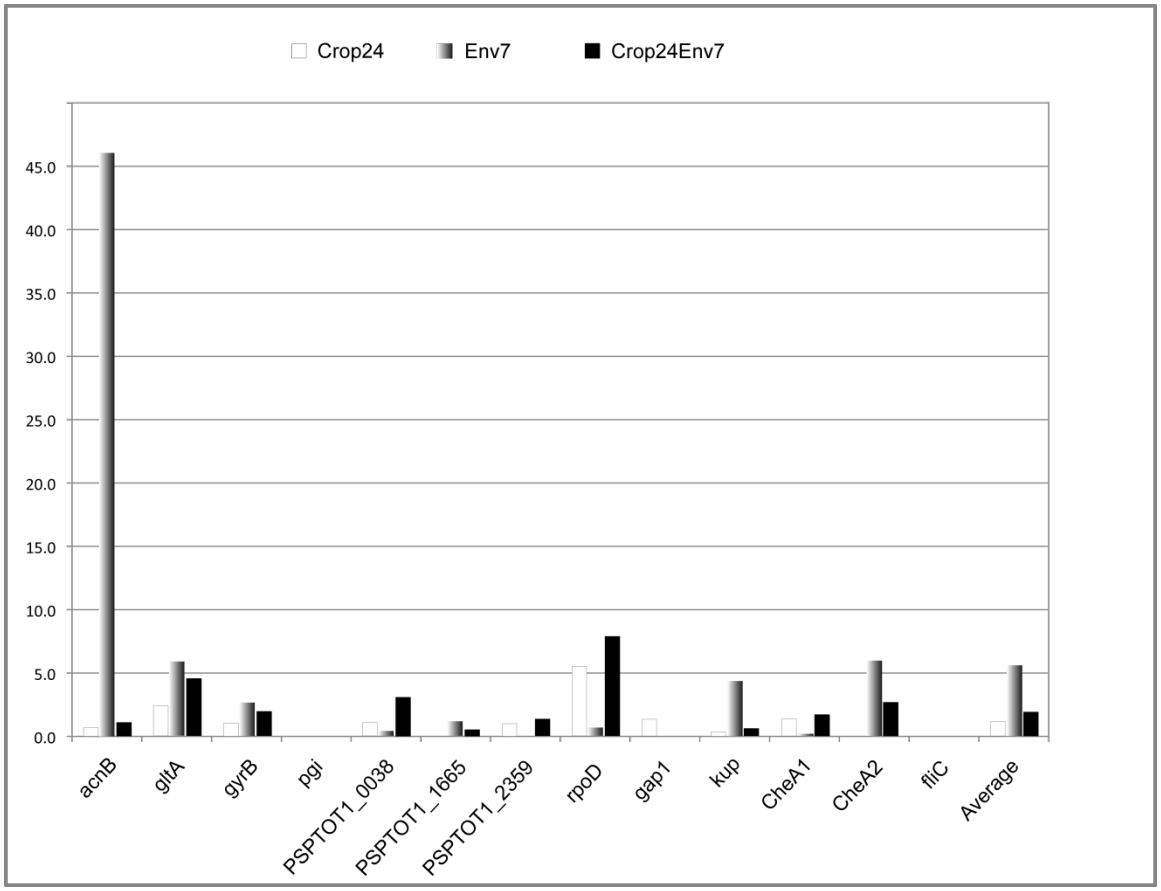
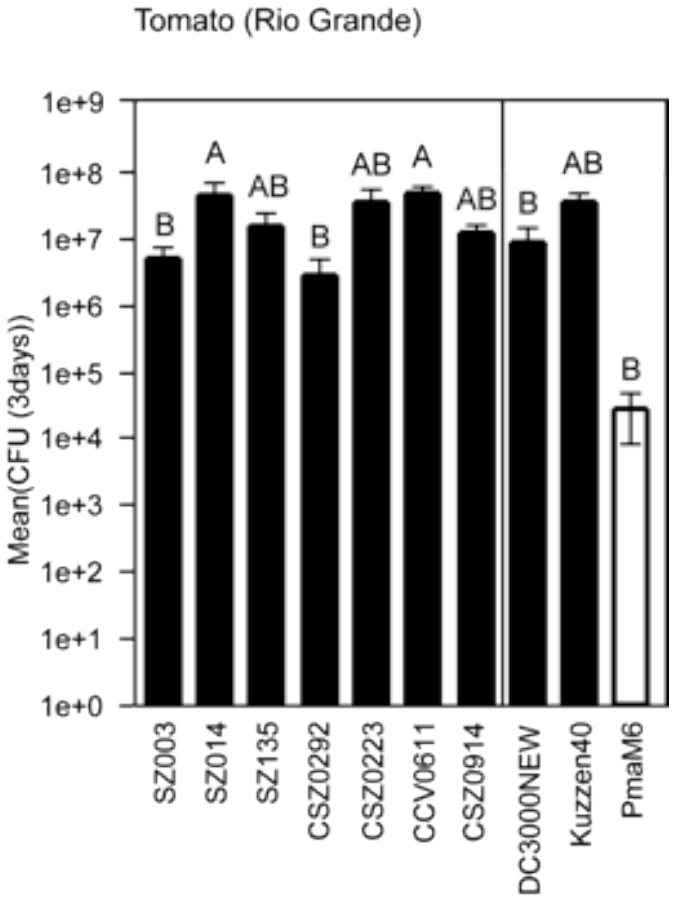
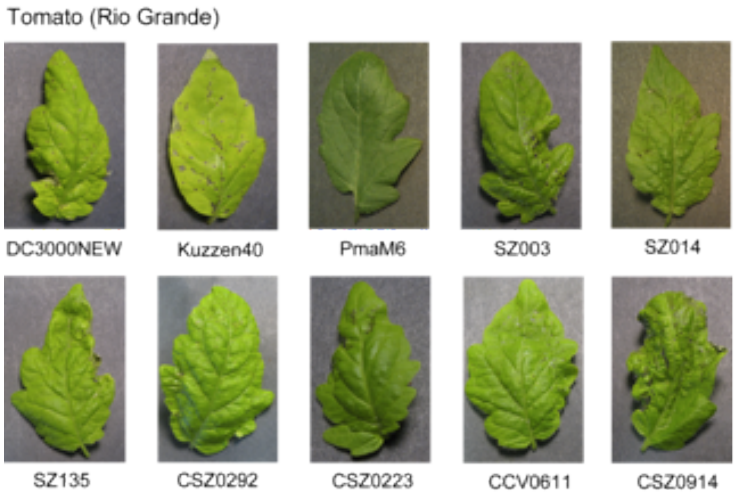

Crop24
Env7
Crop24Env7
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0
acnB
gltA
gyrB
pgi
PSPTOT1_0038
PSPTOT1_1665
PSPTOT1_2359
rpoD
gap1
kup
CheA1
CheA2
fliC
Average

Figure 4.5

A

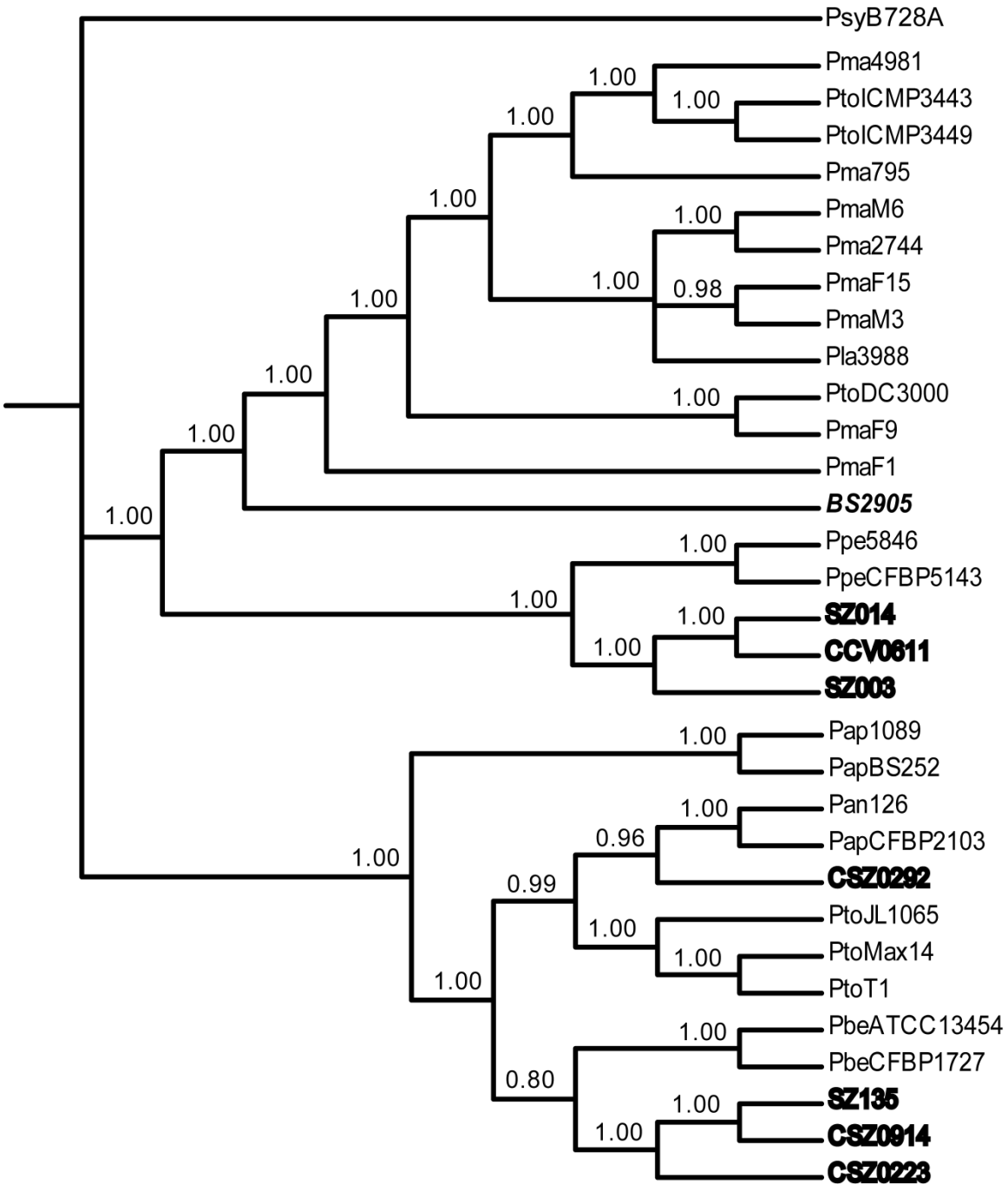

Tomato (Rio Grande)
Mean(CFU (3days))
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
B
A
AB
B
AB
A
AB
B
AB
B
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
DC3000NEW
Kuzzen40
PmaM6

B

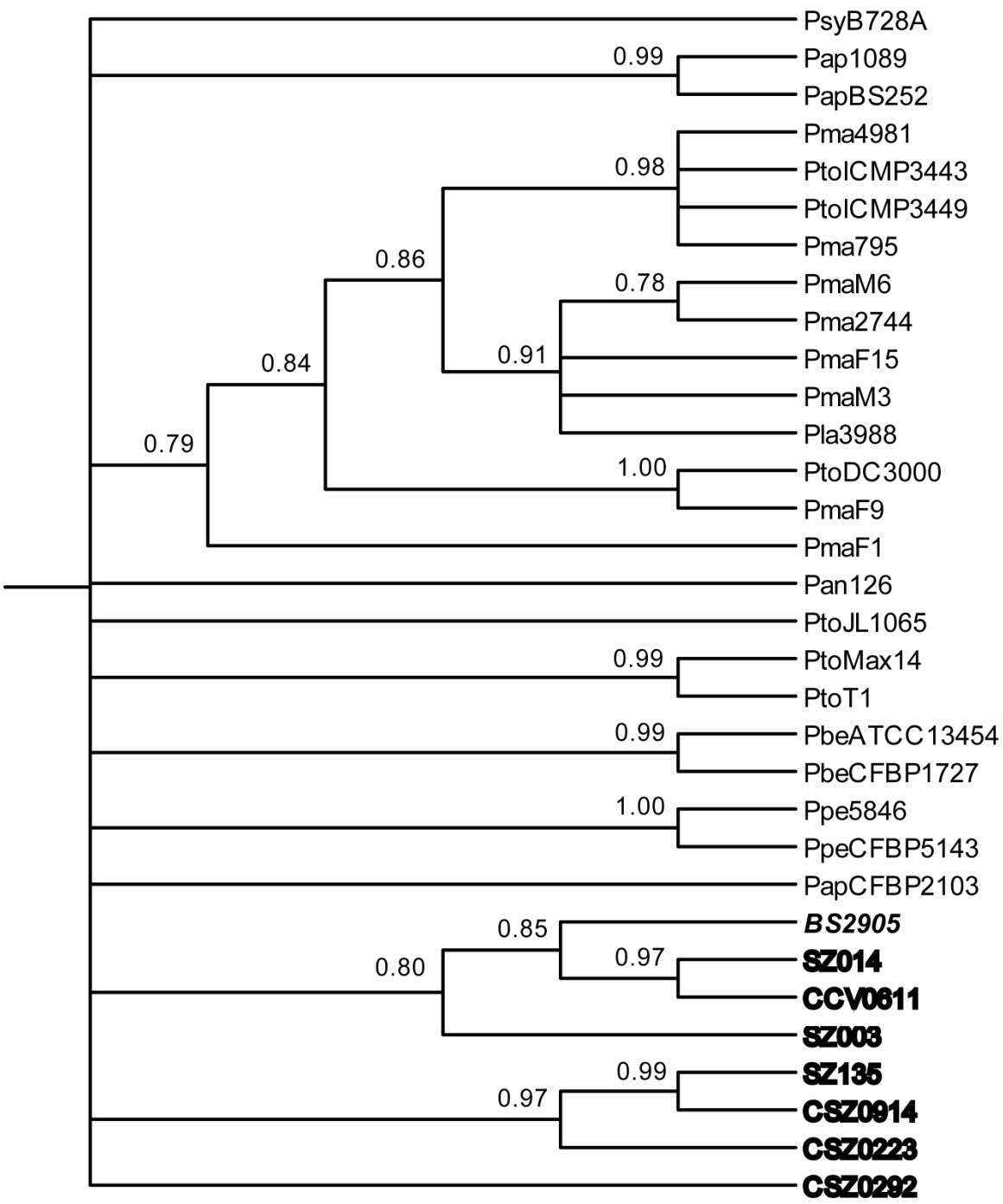

Tomato (Rio Grande)
DC3000NEW
Kuzzen40
PmaM6
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914

SFigure 4.1

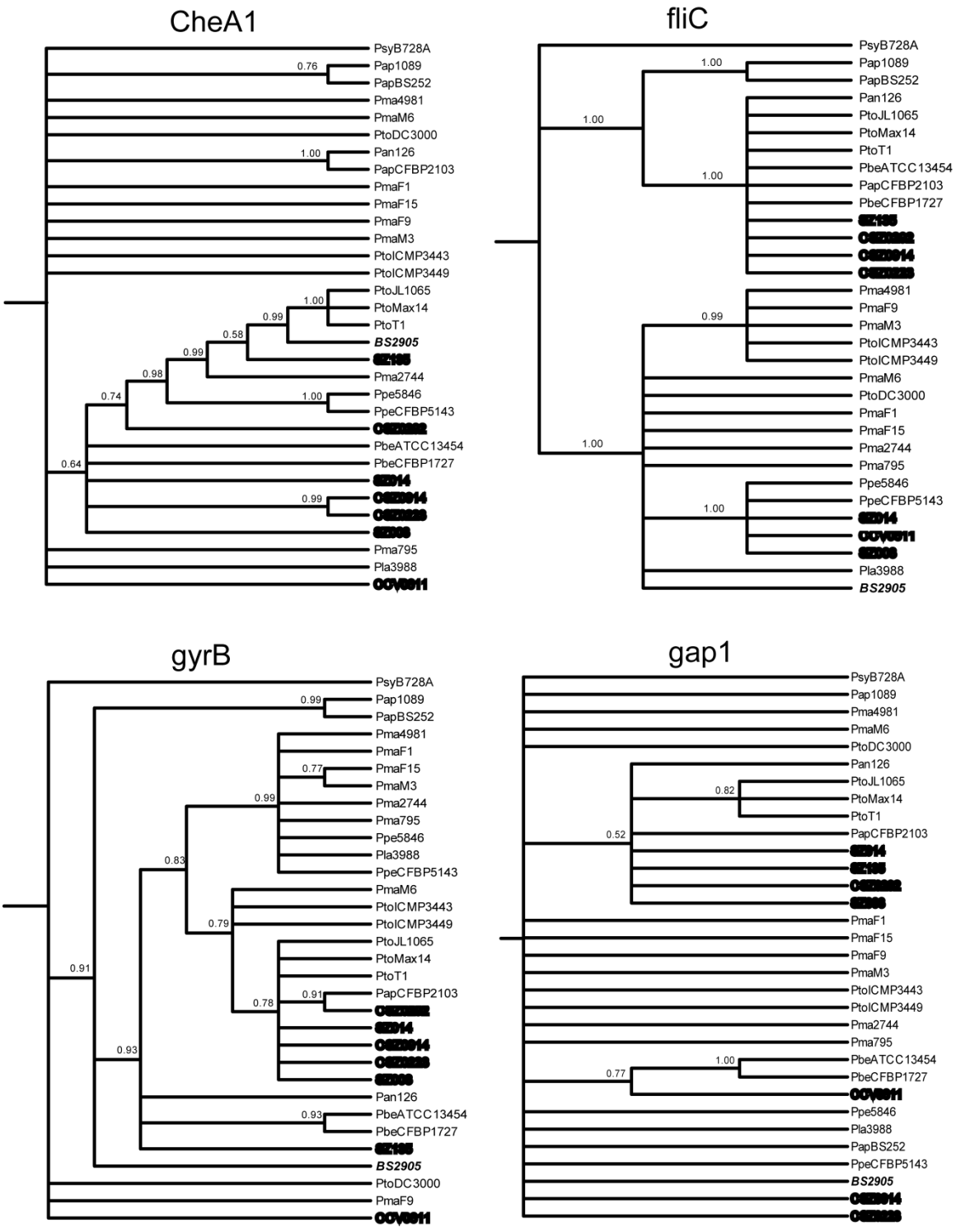
A


PsyB728A
Pma4981
PtoICMP3443
PtoICMP3449
Pma795
PmaM6
Pma2744
PmaF15
PmaM3
Pla3988
PtoDC3000
PmaF9
PmaF1
BS2905
Ppe5846
PpeCFBP5143
SZ014
CCV0611
SZ003
Pap1089
PapBS252
Pan126
PapCFBP2103
CSZ0292
PtoJL1065
PtoMax14
PtoT1
PbeATCC13454
PbeCFBP1727
SZ135
CSZ0914
CSZ0223
1.00
0.98
0.96
0.99
0.80

B

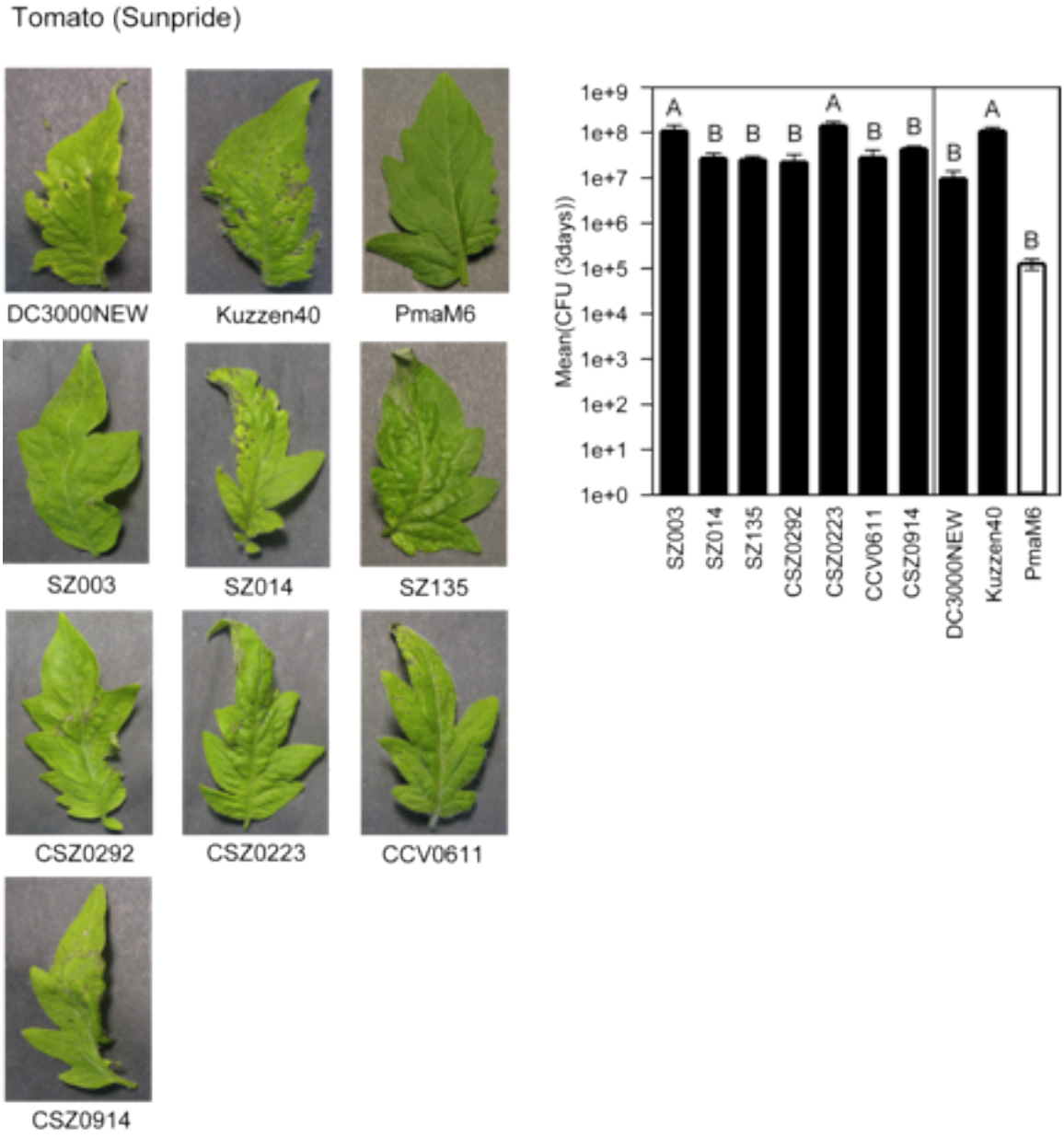

PsyB728A
0.99
Pap1089
PapBS252
Pma4981
0.98
PtoICMP3443
PtoICMP3449
Pma795
0.86
0.78
PmaM6
Pma2744
0.84
0.91
PmaF15
PmaM3
Pla3988
0.79
1.00
PtoDC3000
PmaF9
PmaF1
Pan126
PtoJL1065
0.99
PtoMax14
PtoT1
0.99
PbeATCC13454
PbeCFBP1727
1.00
Ppe5846
PpeCFBP5143
PapCFBP2103
0.85
BS2905
0.97
SZ014
0.80
CCV0611
SZ003
0.99
SZ135
0.97
CSZ0914
CSZ0223
CSZ0292

SFigure 4.2

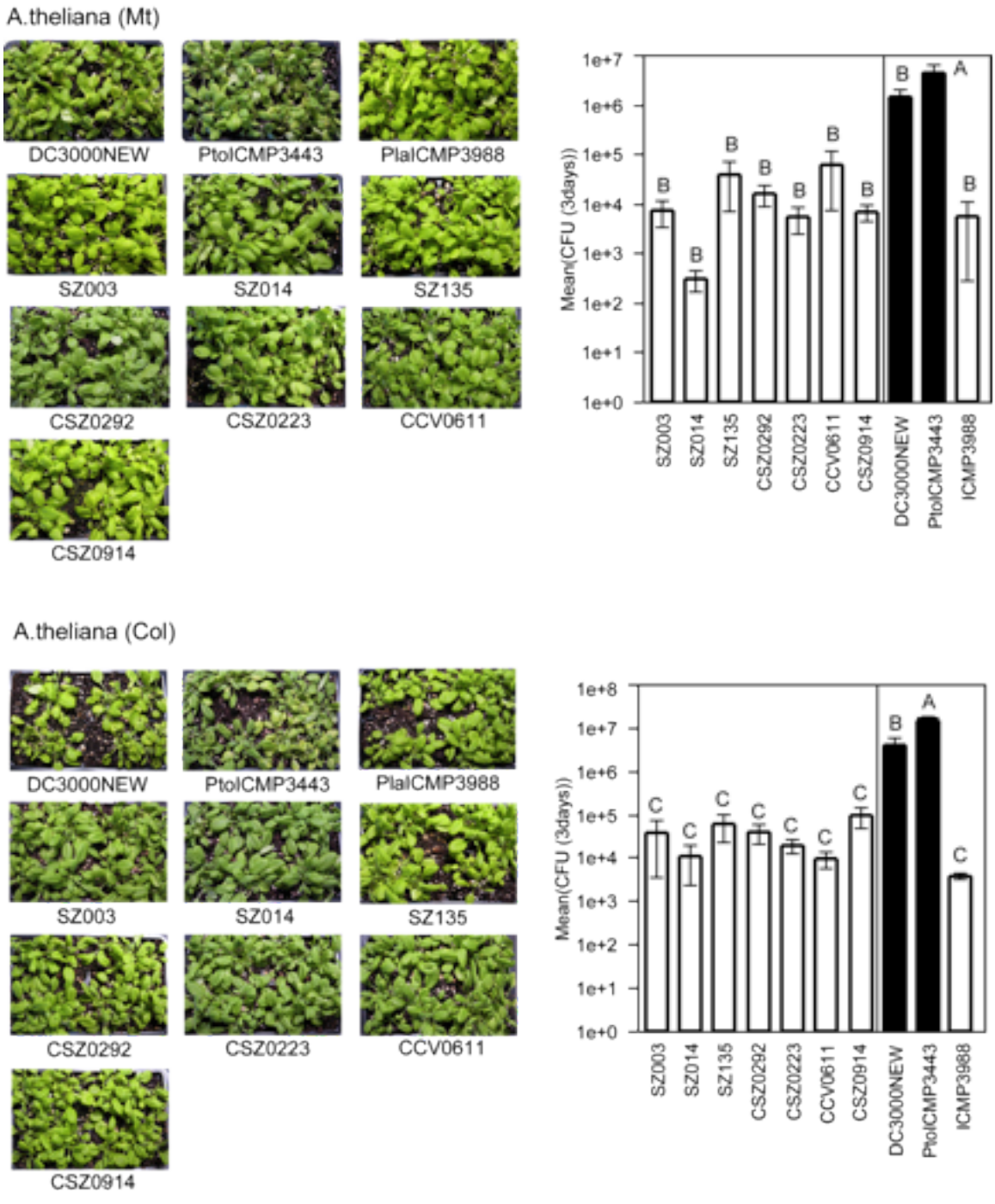

CheA1
PsyB728A
Pap1089
PapBS252
Pma4981
PmaM6
PtoDC3000
Pan126
PapCFBP2103
PmaF1
PmaF15
PmaF9
PmaM3
PtoICMP3443
PtoICMP3449
PtoJL1065
PtoMax14
PtoT1
BS2905
Pma2744
Ppe5846
PpeCFBP5143
PbeATCC13454
PbeCFBP1727
Pma795
Pla3988
fliC
PsyB728A
Pap1089
PapBS252
Pan126
PtoJL1065
PtoMax14
PtoT1
PbeATCC13454
PapCFBP2103
PbeCFBP1727
Pma4981
PmaF9
PmaM3
PtoICMP3443
PtoICMP3449
PmaM6
PtoDC3000
PmaF1
PmaF15
Pma2744
Pma795
Ppe5846
PpeCFBP5143
Pla3988
BS2905
gyrB
PsyB728A
Pap1089
PapBS252
Pma4981
PmaF1
PmaF15
PmaM3
Pma2744
Pma795
Ppe5846
Pla3988
PpeCFBP5143
PmaM6
PtoICMP3443
PtoICMP3449
PtoJL1065
PtoMax14
PtoT1
PapCFBP2103
Pan126
PbeATCC13454
PbeCFBP1727
BS2905
PtoDC3000
PmaF9
gap1
PsyB728A
Pap1089
Pma4981
PmaM6
PtoDC3000
Pan126
PtoJL1065
PtoMax14
PtoT1
PapCFBP2103
PmaF1
PmaF15
PmaF9
PmaM3
PtoICMP3443
PtoICMP3449
Pma2744
Pma795
PbeATCC13454
PbeCFBP1727
Ppe5846
Pla3988
PapBS252
PpeCFBP5143
BS2905

SFigure 4.3

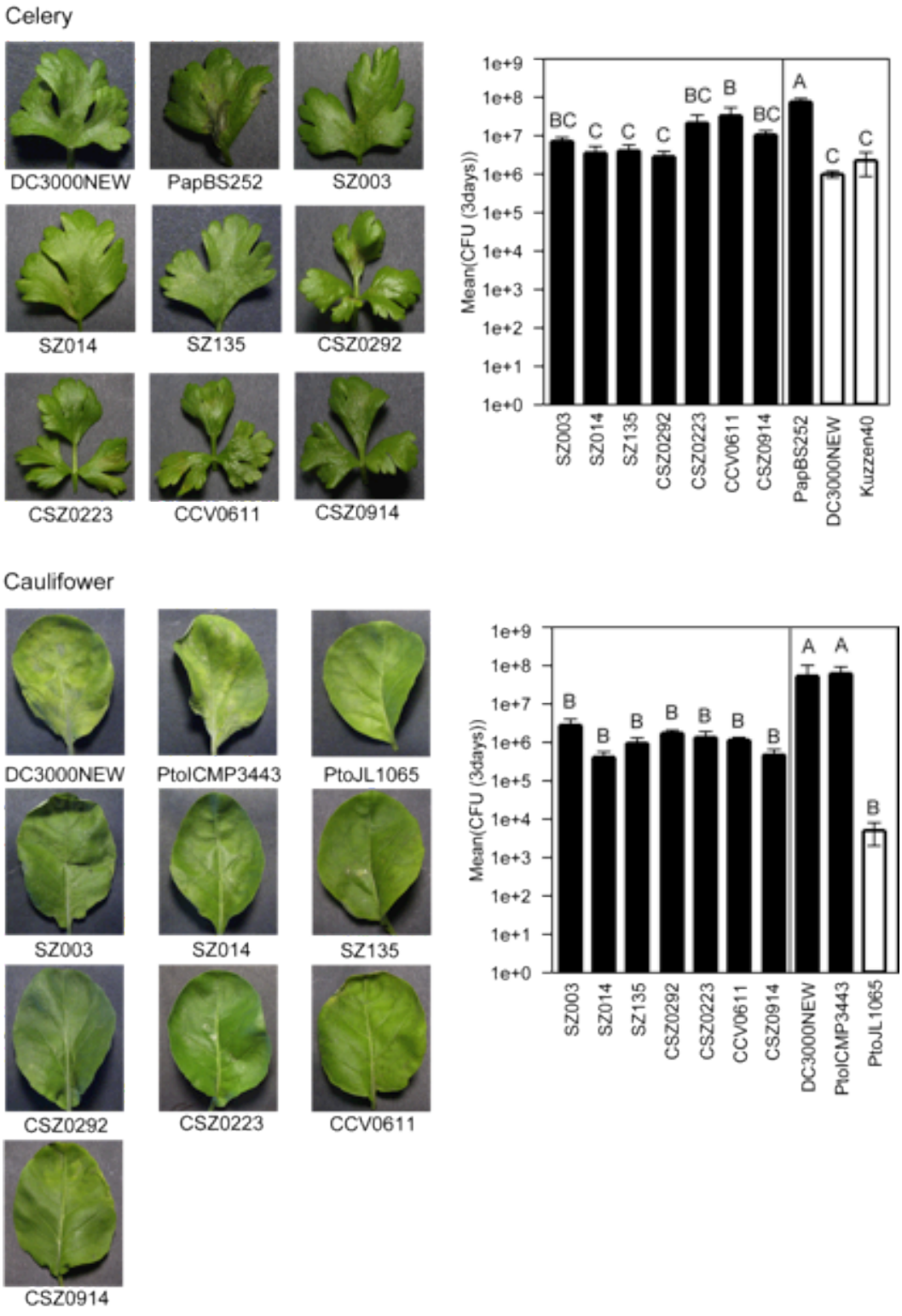
A


Tomato (Sunpride)
DC3000NEW
Kuzzen40
PmaM6
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
Mean(CFU (3days))
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
A
B
B
B
A
B
B
B
A
B
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
DC3000NEW
Kuzzen40
PmaM6

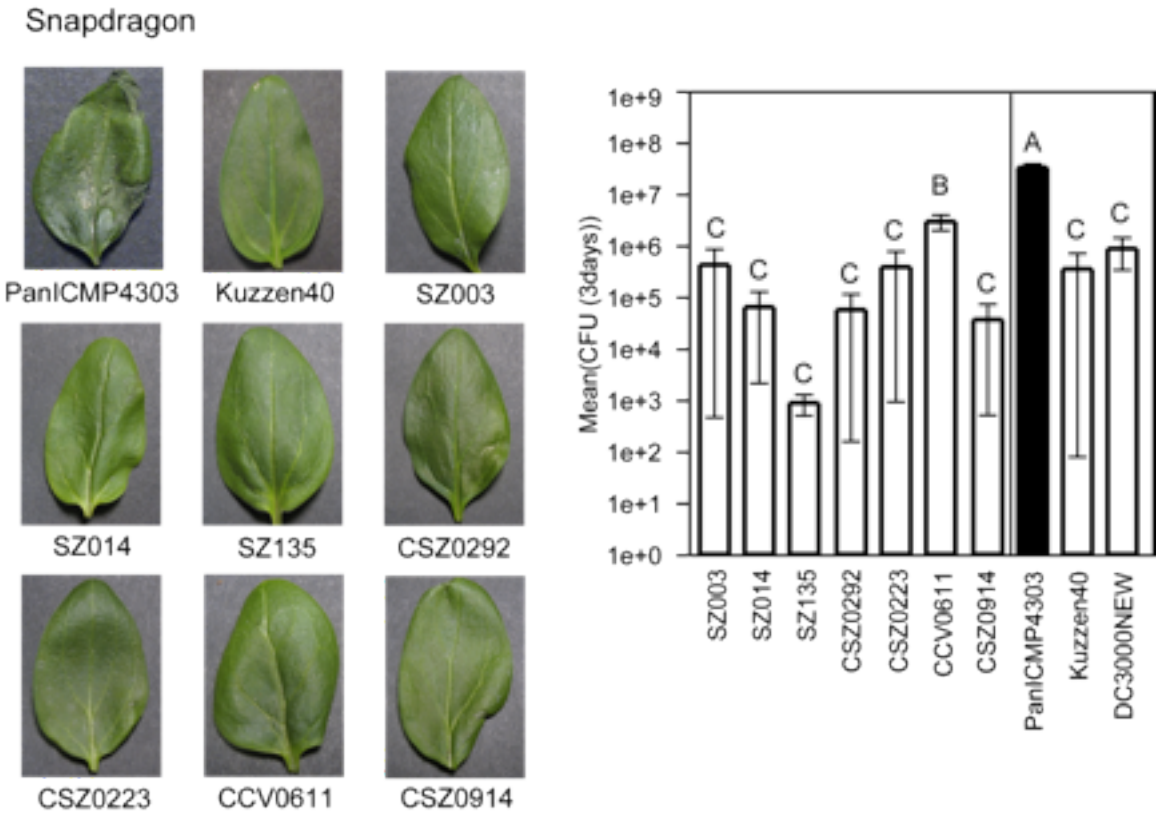
B


A.theliana (Mt)
DC3000NEW
PtoICMP3443
PlaICMP3988
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
Mean(CFU (3days))
1e+7
1e+6
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
DC3000NEW
PtoICMP3443
ICMP3988
A.theliana (Col)
DC3000NEW
PtoICMP3443
PlaICMP3988
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
Mean(CFU (3days))
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
DC3000NEW
PtoICMP3443
ICMP3988

C


Celery
DC3000NEW
PapBS252
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
Mean(CFU (3days))
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
BC
C
C
C
BC
B
BC
A
C
C
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
PapBS252
DC3000NEW
Kuzzen40
Caulifower
DC3000NEW
PtoICMP3443
PtoJL1065
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
Mean(CFU (3days))
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
B
B
B
B
B
B
B
A
A
B
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
DC3000NEW
PtoICMP3443
PtoJL1065

D


Snapdragon
PanICMP4303
Kuzzen40
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
Mean(CFU (3days))
1e+9
1e+8
1e+7
1e+6
1e+5
1e+4
1e+3
1e+2
1e+1
1e+0
C
C
C
C
C
B
C
A
C
C
SZ003
SZ014
SZ135
CSZ0292
CSZ0223
CCV0611
CSZ0914
PanICMP4303
Kuzzen40
DC3000NEW

Chapter 5

Summary and Conclusions

P.syringae is divided into approximately 50 different pathovar strains based on host range and symptoms. *P.syringae* evolves very fast, for example one recently emerging disease break caused by *P.syringae* pv. actinidiae (PAS) occurred in New Zealand in 2010 after it was first described in Japan in 1984. With the evolution of these pathogens, their host range has evolved, as well. Even though each strain of *P.syringae* is specific for a particular plant, some of them can develop pathogenicity in non-host as well. Especially farming practices such as monoculture results in the rapid production, infection and spread of pathogens. Some strains with the host of wild plants from a long distance could become virulent on genetically uniform crop plants by evolving their resistance mechanisms over thousands of years. We originally hypothesized that today's highly virulent *P. syringae* crop pathogens with narrow host range might have evolved from ancestral *P. syringae* strains with wide host range after the advent of agriculture.

To investigate the evolution and mechanisms of host range in *P. syringae*, *P. syringae* pv. *tomato* DC3000 and its close relatives isolated from different hosts were selected because it is feasible to identify the responsible genomic differences between vary similar strains to explain the mechanistic basis of host specificities. The technique of

Multilocus Sequence Typing (MLST) (Maiden et al., 1998) was applied to resolve the phylogenetic relationship based on the sequencing of several housekeeping genes under purifying selection and also to determine the contribution of homologous recombination. Since recombination and mutation shape the architecture of bacterial genomes and their relative contribution is important to understand virulence potential, results from NeighborNet construction, ratio of population-scale recombination to mutation and identification of recombination break points showed that recombination occurred among stains in this group. The host tests were combined with phylogeny to infer character ("diseased/healthy" and "relatively wide/narrow host range") evolution along branches of a phylogenetic tree for the ancestral state reconstruction if the character evolution was proved to be associated with phylogeny. Finally, the hypothesis was either confirmed or rejected because of a small selection of strains from a few crop species, lab conditions for plant infection, the low support on some branches in the tree and poor way of defining relatively narrow and wide host range and so on. With the recent development of new sequencing technology such as 454-Roche pyrosequencing (Elahi and Ronaghi, 2004) and illumina's Solexa (Quail, Kozarewa et al., 2008), whole genome sequencing has become inexpensive and time-efficient. It is feasible to sequence whole genome to identify the repertoire of type-III secreted effectors because strains with relatively wide host range may not have the effectors triggering immunity on most of tested plant species and strains with relatively narrow host range could have one or more effectors that significantly increase their virulence on the host.

It is clear that plant pathogen strains for the study of molecular-plant interactions in the lab may not truly represent the pathogens that could cause disease in the field. The human pathogens caused by genetically monomorphic bacterial pathogens, for example *Yersinia pestis* (Morelli, Song et al. 2010), had elucidated their microevolution and dispersion around the world. More study was concentrated on T1 clade in *P. syringae* group about the genome-based microevolution and phylogeography of *P. syringae* pv. *tomato* based on over 100 isolates collected in the last sixty years around the world (Cai et al., 2011). At first the phylogeny of five pathogen isolates from single nucleotide polymorphism (SNP) of whole genome sequenced using 454 and Illumina technology inspired our idea that *P. syringae* pv. *tomato* strains may have evolved recently. The comparison of genotypes among continents could identify how pathogen movement between continents. Besides genome-derived markers for sequence analysis, molecular analysis of key pathogen loci is very important for virulence and motility. When plants are infected by *P.syringae*, the receptor such as flagellin receptor on the host cell surface can perceive Pathogen-Associated Molecular Patterns (PAMPs) and a basal defense response is triggered. *P. syringae* has the type III secretion system to inject the type III secreted effector (T3SE) and some plants have the evolved resistant genes to recognize T3SE to lead immunity. It is necessary to compare how different on one of type III-secreted effector gene such as *hopM1* and fliC gene. All data of hopM1 showed a strong selection for loss of function of this effector and the data of fliC gene also showed that the ancestral alleles could trigger the stronger immune response than the derived alleles with mutated *fliC*.

Even though new crop pathogens emerge and often spread around the world very fast, their evolutionary origin remains unclear and inconclusive. It is well known that several bacterial human pathogens, such as *V.Cholerae* that causes a severe diarrheal disease on human, have the environmental reservoirs. *P. syringae* transmits a long distance through water cycle and transports to crop and wild plants somewhere (Morris et al., 2008). Recent sampling of *P.syringae* from rain, snow, alpine streams and lakes showed the abundance of *P.syringae* in the non-host environmental reservoirs (Morris et al., 2010). Based on the study of host range evolution (Cai et al., 2011), the environmental strains from snowpack and surface water in France and highly correlated with crop strains were added for the further analysis. To better understand the molecular evolution of *P. syringae* and its host specificity, our analysis was extended to five more genes. With more environmental strains adding, more reticulations were generated around Pbe and Ppe strains that were previously located at the end of long branches without any reticulations distributed along long branches (Cai et al., 2011). It indicated that more related strains were found from non-host environmental reservoir and that recombination events may have occurred among crop strains and environmental strains. Especially some reticulations generated among environmental strains and Pto strains may suggest the host change of tomato strains during recent evolution. The data of *PGI* allele suggested occurred recombination and frequent horizontal gene transfer due to the selection. The data of *FliC* allele showed some evidence of horizontal gene transfer due to the selection. The host range tests were performed on five species of plants. All stains were found to be as pathogenic as those from tomato. The overlapping repertoire

of genes coding for type III secretion effectors of whole genome between PtoT1 strains and environmental strains may suggest that environmental strains may be an important source for novel virulence genes of crop pathogens and that crop pathogens may share the similar ancestor as pathogens in the environmental. During the evolution, the *P. syringae* strains in non-plant environments had adapted to the specific crops and only virulent to these crops. It may indicate that strains from the nonhost environmental reservoirs have a wider host range than closely related crop pathogens.

References

Cai R, Lewis J, Yan S, Liu H, Clarke CR, et al. (2011) The Plant Pathogen *Pseudomonas syringae* pv. *tomato* Is Genetically Monomorphic and under Strong Selection to Evade Tomato Immunity. PLoS Pathog 7(8): e1002130. doi:10.1371/journal.ppat.1002130

Cai R, Yan S, Liu H, Leman S, Vinatzer A. B et al. (2011) Reconstructing host range evolution of bacterial plant pathogens using Pseudomonas syringae pv. tomato and its close relatives as a model, Infection, Genetics and Evolution, Volume 11, Issue 7, October 2011, Pages 1738-1751, ISSN 1567-1348

Elahi, E. and M.
Ronaghi (2004). Pyrosequencing. Bacterial Artificial Chromosomes. **255:** 211-219

Quail, M. A., I. Kozarewa, et al. (2008). "A large genome center's improvements to the Illumina sequencing system." Nat Meth **5**(12): 1005-1010

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95, 3140-3145.

Morelli, G., Y. Song, et al. (2010). “Yersinia pestis genome sequencing identified patterns of global phylogenetic diversity.” Nat Genet 42(12):1140-1143.

Morris, C.E., Sands, D.C., Vinatzer, B.A., et al. (2008) “The life history of theplant pathogen Pseudomonas syringae is linked to the water cycle”. ISME J 2: 321–334.Cai et al., 2011 plant pathogen

Morris, C. E., D. C. Sands, J. L. Vanneste, J. Montarry, B. Oakley, et al. (2010). Inferring the evolutionary history of the plant pathogen *Pseudomonas syringae* from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. mBio 1(3):e00107-10. doi:10.1128/mBio.00107-10