PLOS ONE

RESEARCH ARTICLE

Similar levels of gene content variation observed for *Pseudomonas syringae* populations extracted from single and multiple host species

Talia L. Karasov[1,2], Luke Barrett[2,3], Ruth Hershberg[4☯], Joy Bergelson[1,2☯]*

1 Committee On Genetics Genomics & Systems Biology, University of Chicago, Chicago, Illinois, United States of America, **2** Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, United States of America, **3** CSIRO Agriculture, Canberra, ACT 2601, Australia, **4** Department of Genetics, the Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel

☯ These authors contributed equally to this work.
* jbergels@uchicago.edu

Citation: Karasov TL, Barrett L, Hershberg R, Bergelson J (2017) Similar levels of gene content variation observed for *Pseudomonas syringae* populations extracted from single and multiple host species. PLoS ONE 12(9): e0184195. https://doi.org/10.1371/journal.pone.0184195

Editor: Chih-Horng Kuo, Academia Sinica, TAIWAN

Received: March 14, 2017

Accepted: August 18, 2017

Published: September 7, 2017

Data Availability Statement: All relevant data are available from SUBID BioProject, accession numbers: SUB2715759 PRJNA387592 SAMN07158962 NHSR00000000 Pseudomonas syringae RM.P20 SUB2715759 PRJNA387592 SAMN07158961 NHSS00000000 Pseudomonas syringae RMX.24.a.1 SUB2715759 PRJNA387592 SAMN07158960 NHST00000000 Pseudomonas syringae RM.P66 SUB2715759 PRJNA387592 SAMN07158959 NHSU00000000 Pseudomonas syringae LP868.1a SUB2715759 PRJNA387592 SAMN07158958 NHSV00000000 Pseudomonas syringae LP221b SUB2715759 PRJNA387592

Abstract

Bacterial strains of the same species collected from different hosts frequently exhibit differences in gene content. In the ubiquitous plant pathogen *Pseudomonas syringae*, more than 30% of genes encoded by each strain are not conserved among strains colonizing other host species. Although they are often implicated in host specificity, the role of this large fraction of the genome in host-specific adaptation is largely unexplored. Here, we sought to relate variation in gene content between strains infecting different species to variation that persists between strains on the same host. We fully sequenced a collection of *P. syringae* strains collected from wild *Arabidopsis thaliana* populations in the Midwestern United States. We then compared patterns of variation observed in gene content within these *A. thaliana*-isolated strains to previously published *P. syringae* sequence from strains collected on a diversity of crop species. We find that strains collected from the same host, *A. thaliana*, differ in gene content by 21%, 2/3 the level of gene content variation observed across strains collected from different hosts. Furthermore, the frequency with which specific genes are present among strains collected within the same host and among strains collected from different hosts is highly correlated. This implies that most gene content variation is maintained irrespective of host association. At the same time, we identify specific genes whose presence is important for *P. syringae*'s ability to flourish within *A. thaliana*. Specifically, the *A. thaliana* strains uniquely share a genomic island encoding toxins active against plants and surrounding microbes, suggesting a role for microbe-microbe interactions in dictating the abundance within this host. Overall, our results demonstrate that while variation in the presence of specific genes can affect the success of a pathogen within its host, the majority of gene content variation is not strongly associated with patterns of host use.

SAMN07158957 NHSW00000000 Pseudomonas syringae LP217a SUB2715759 PRJNA387592 SAMN07158956 NHSX00000000 Pseudomonas syringae NP29.1a SUB2715759 PRJNA387592 SAMN07158954 NHSY00000000 Pseudomonas syringae Knox652c SUB2715759 PRJNA387592 SAMN07158953 NHSZ00000000 Pseudomonas syringae Knox623a SUB2715759 PRJNA387592 SAMN07158952 NHTA00000000 Pseudomonas syringae KN2.a.3 SUB2715759 PRJNA387592 SAMN07158951 NHTB00000000 Pseudomonas syringae LMC.P91 SUB2715759 PRJNA387592 SAMN07158950 NHTC00000000 Pseudomonas syringae LMC.P80 SUB2715759 PRJNA387592 SAMN07158949 NHTD00000000 Pseudomonas syringae LMC.P10 SUB2715759 PRJNA387592 SAMN07158965 Pseudomonas syringae LP205a SUB2715759 PRJNA387592 SAMN07158964 Pseudomonas syringae ME812.2b SUB2715759 PRJNA387592 SAMN07158963 Pseudomonas syringae RMX815.1a SUB2715759 PRJNA387592 SAMN07158955 Pseudomonas syringae NL.P123 SUB2715759 PRJNA387592 SAMN07158948 Pseudomonas syringae DM2.1.12.02A.

Funding: Funded by NSF345 MCB 0603515 to JB. https://www.nsf.gov/. TLK was supported by a Graduate Assistance in Areas of National Need (GAANN) training grant and NSF DDIG 1311515. https://www.nsf.gov/. RH was supported by an ERC FP7 CIG grant (https://erc.europa.eu/), by a Yigal Allon Fellowship (http://www.utdallas.edu/~DJL101000/DavidLary/Awards/Entries/1998/10/1_Alon_Fellowships_for_Outstanding_Young_Researchers.html), and by the Robert J. Shillman Career Advancement Chair.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Many microbial species colonize diverse biotic and abiotic ecological niches [1,2]. The traits that allow a microbe to survive in these varied environments are of wide ecological, clinical and agricultural relevance. For example, many pathogenic bacteria exhibit increased virulence in their host of isolation [3,4]. There is thus widespread interest in understanding the evolutionary and genetic mechanisms that allow strains to flourish in specific environments while perishing in others.

The plethora of bacterial genome sequences that are publicly available provides broad insight into the genes that are potentially adaptive for specific hosts and environments. Numerous studies have revealed that strains of the same species collected from disparate environments differ extensively in gene content [5–9], varying in the presence of dozens to even thousands of genes. At least a fraction of this variability underlies environment-specific adaptations, and the challenge has now become determining which genes are adaptive for which environments.

It is particularly important to understand the adaptive significance of gene content diversity in the ubiquitous plant pathogen *P. syringae*. *P. syringae* is a genetically diverse bacterial species complex encompassing lineages with both pathogenic and non-pathogenic lifestyles [2,10,11]. Particular strains of *P. syringae* are especially pathogenic on specific host species [12–15], and can cause extensive damage to crop populations [12,16]. Several of these strains have been classified as pathovars, or lineages with specific unifying pathogenicity characteristics that specialize them on specific hosts. Previous genome comparisons of host-specific *P. syringae* strains showed that over 30% of genes within a strain's genome are either unique to a single strain or are rare among other strains [7]. To date, however, only a few dozen of the thousands of variable genes have been shown to be adaptive in an environmentally specific manner [7,17,18].

P. syringae not only infects crop plant populations but also non-agricultural plant populations, including those of *A. thaliana*. Indeed, *P. syringae* is among the most abundant bacterial pathogens in *A. thaliana* leaves [19,20], causing reductions in fitness upon infection [21]. Interestingly, the evolution of *P. syringae* in *A. thaliana* populations appears to differ from that of *P. syringae* in crop populations. While crop-infecting strains frequently exhibit clonal, genetically monomorphic expansions and obvious disease symptoms [12,14,22], those *P. syringae* isolated from *A. thaliana* to date exhibit less obvious symptoms and more extensive genetic diversity [23], a diversity that is maintained even at a regional scale. The reason for this contrast in pathogenic genetic diversity between crops and a non-agricultural plant is unknown (though it is easy to speculate [24]). What is evident is that *P. syringae* colonizes *A. thaliana* populations successfully and frequently, and reduces yield [19,21,25].

The success and abundance of *P. syringae* in *A. thaliana* provides the opportunity to determine the evolution of *P. syringae* gene content within a single host species (*A. thaliana*) and to contrast this diversity with that observed among host species. Here, we begin to characterize gene content variability by genotyping 76 strains of *P. syringae* that reside on the same host species in the same geographical region of the Midwestern United States, and then fully sequencing 18 of these strains. We find that strains of *P. syringae* collected from *A. thaliana*, some from the same host population, exhibit variation in gene content similar to that observed between *P. syringae* strains collected from different crop host species. At the same time, the *P. syringae* strains collected from *A. thaliana* uniquely share a genomic island that encodes toxins active against a broad range of microbes, raising the possibility that these strains gain an advantage by suppressing other microbes. Combined, our findings reveal both extensive genetic turnover and a conserved genomic island that suggests the importance of microbe-microbe interactions in the evolution of a pathogen in natural plant populations.

Results

A specific lineage of *P. syringae* infects *A. thaliana*

To study the gene content diversity of *P. syringae* in Midwestern USA populations of *A. thaliana*, we first sought to characterize *P. syringae* phylogenetic diversity in this area. *P. syringae* is both abundant and pathogenic within these *A. thaliana* populations, inducing a fitness cost in infected individuals of up to 25% [19,21]. We first determined the phylogenetic distribution of the *P. syringae* that colonize *A. thaliana* by performing Multilocus sequencing analysis (MLSA) of six loci [26] in 99 strains. The MLSA indicated that *P. syringae* isolates from *A. thaliana* cluster nearly exclusively within one specific lineage of *P. syringae* (group II as represented in Fig 1A, despite the fact that pathogenic *P. syringae* is abundant in several other phylogenetic lineages). Note, however, that those lineages collected from *A. thaliana* are not genetically identical, and instead, several crop strains from phylogenetic group II are more closely related to several *A. thaliana* lineages than are other *A. thaliana* lineages (Figs 1A and 2B).

While this result suggests the preferential colonization of group II in *A. thaliana*, it is also possible that the skewed phylogenetic distribution is the result of population structure of *P. syringae* within the Midwestern USA. However, previous studies reveal the genetic diversity of *P. syringae* in the Midwest [28,29] to be higher than that maintained on *A. thaliana* [23]. To explicitly test whether *P. syringae* in the Midwestern USA are from only this one lineage, we sampled *Pseudomonas* from both *A. thaliana* populations and nearby agricultural tomato crops in the Midwestern USA in the fall of 2013. Through high-throughput genotyping of the

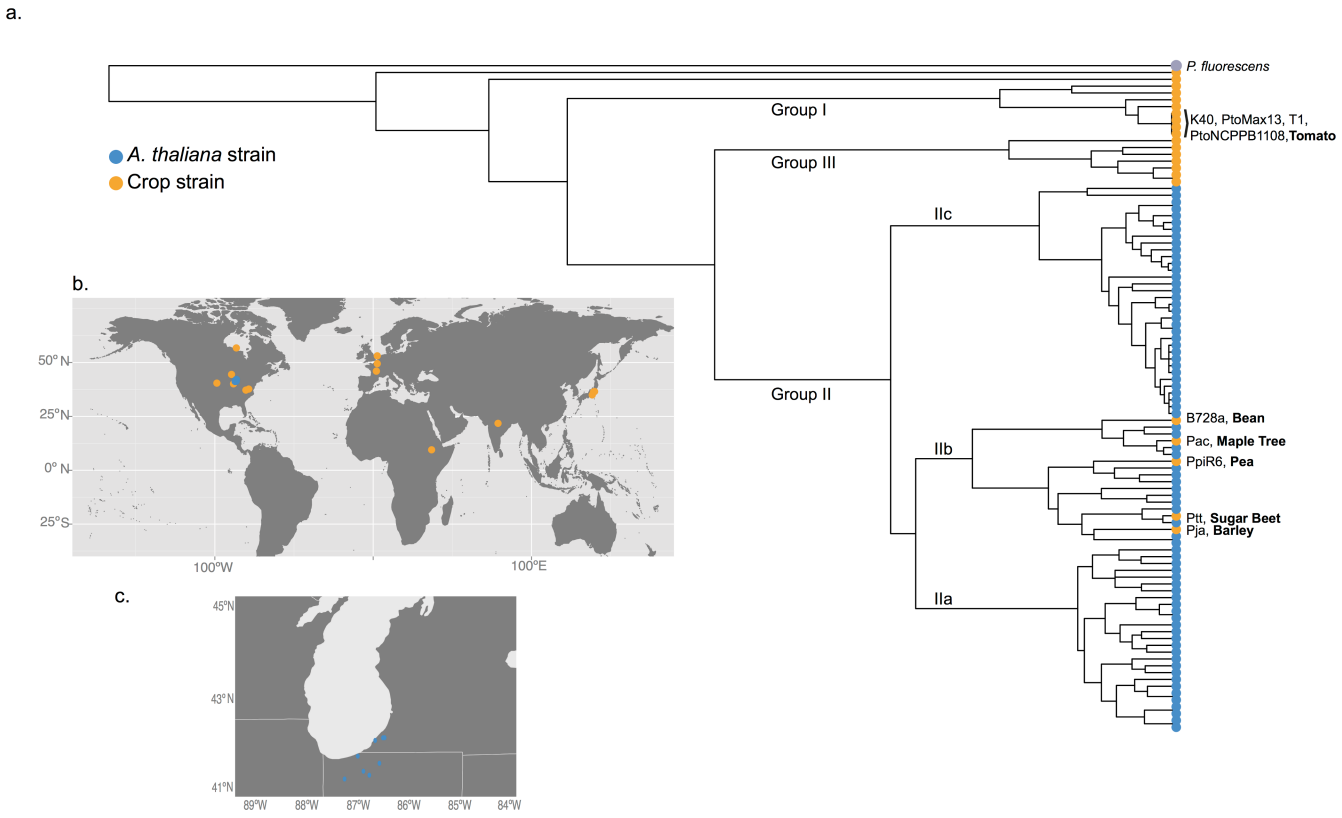

Fig 1. Composition of *P. syringae* in *A. thaliana* in the Midwestern USA. (a) Majority rule consensus tree based on six MLSA loci in *P. syringae*. Blue circles represent *P. syringae* strains collected from *A. thaliana* plants. Orange circles represent strains collected from crop plants. The names of select crop strains are provided next to the corresponding node along with the host of isolation in bold. The *A. thaliana* strains are restricted to group II while the crop strains span the *P. syringae* phylogenetic tree. (b) & (c) Geographic distribution of strains whose genomes were analyzed in this study.

https://doi.org/10.1371/journal.pone.0184195.g001

gyrb gene, we confirmed that group II [7,26,30] dominates within *A. thaliana* (Fig 1) yet another clade dominates within tomato crops (Fig 2A). The samples were collected within two weeks of one another, reducing the probability that differences in composition are due to temporal changes in microbial communities. These results suggest that a specific subset of *P. syringae* belonging to group II preferentially proliferates in *A. thaliana* populations.

P. syringae on *A. thaliana* exhibit variation in gene content

To assess gene content diversity in *A. thaliana P. syringae* strains, we sequenced the genomes of 18 randomly selected strains of *P. syringae* on *A. thaliana* [25] (Fig 2B) and compared them

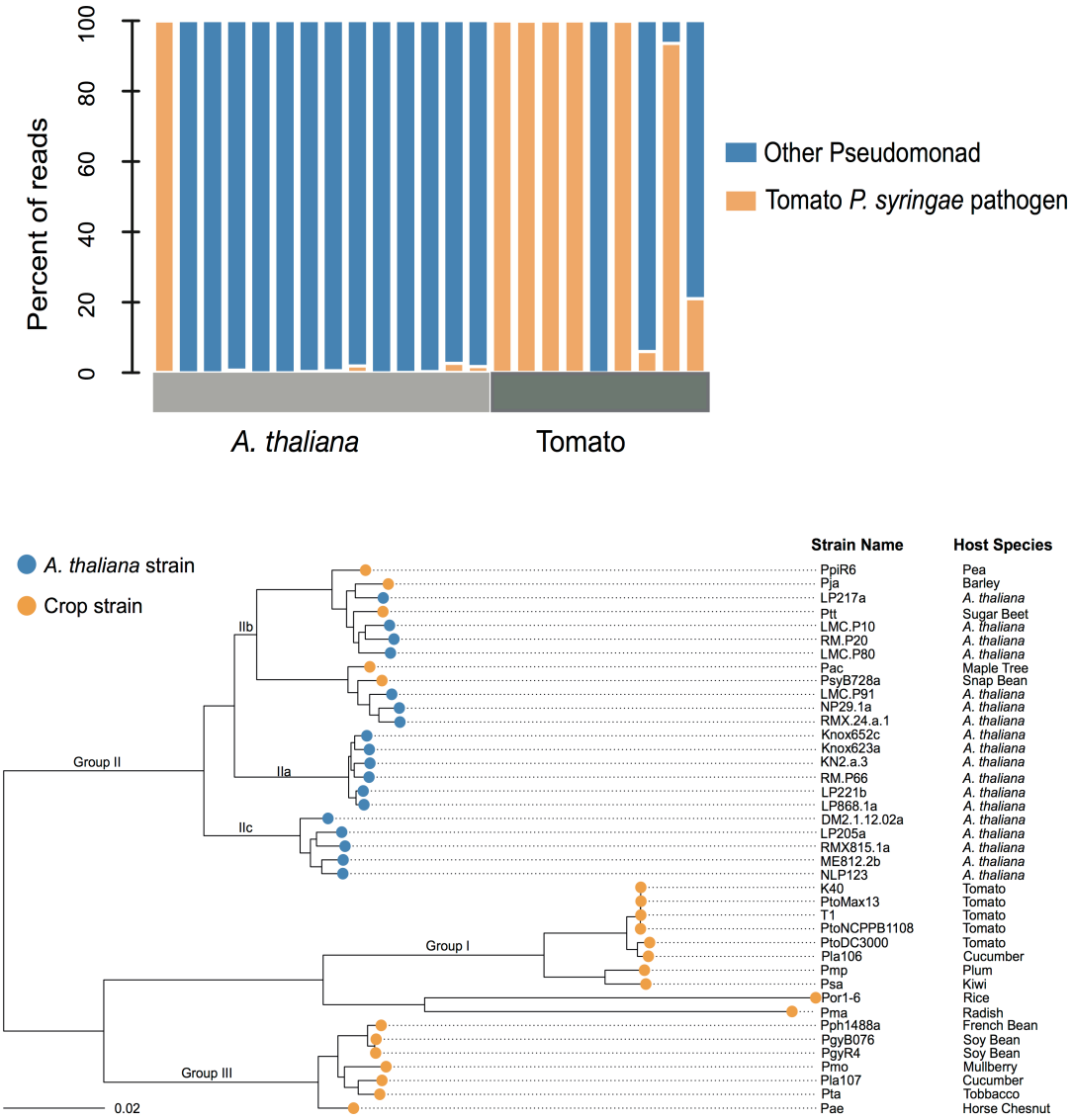

Fig 2. *P. syringae* from *A. thaliana* are genetically diverse, but are derived from one phylogenetic clade. (a) Pseudomonad composition of leaves of *A. thaliana* and tomato collected from the Midwestern USA was assessed via phylotyping of *gyrb*. The composition of Pseudomonads was significantly different between the leaves of tomato and *A. thaliana* (P = 0.008, Wilcoxon-rank-sum test), with the tomato leaves composed primarily of strains resembling the abundant tomato strains. (b) Maximum likelihood phylogeny of strains sequenced in this study based on 492 conserved genes. The scale bar indicates 2% sequence divergence. *A. thaliana* strains were primarily derived from phylogenetic group II, a result recapitulated in the wider MLSA (Fig 1). Group IIc lacks a canonical T3SS [27].

https://doi.org/10.1371/journal.pone.0184195.g002

to the genomes of 22 largely host-specific crop strains from diverse geographical locations (Fig 1B and 1C). Previous genome comparisons of *P. syringae* isolated from diverse crop species found that more than 30% of genes within a genome are unique or rare among crop strains [7]. In this study, we re-annotated all genomes for consistency of annotation methodology between studies and relaxed the definitions of conservation to accommodate differences in assembly qualities between strains and studies. We also verified that differences in assembly quality did not significantly influence the number of genes annotated per genome (S1 Fig). Our findings support the previous findings of diversity, observing that an average of 32% of the genes within a crop strain's genome (1710/5378 genes) varies in presence across crop strains. Although such large-scale differences in gene content have frequently been thought to result from host-specific adaptations [31] our analysis also revealed high levels of gene content variation among *A. thaliana* strains (Fig 3A). An analysis of the gene-frequency-spectrum of *P. syringae* within *A. thaliana* host populations revealed that the frequency distribution of genes among *A. thaliana* isolates closely resembles the gene-frequency-spectrum of *P. syringae* across diverse crop hosts: a similar number of genes are conserved across strains and a similar number of genes are distributed at intermediate frequency. An average of 79% of the genome of an *A. thaliana* strain is conserved (found in more than 90% of strains) across *A. thaliana* isolated strains (4014/5098 genes). Stated differently, 21% of an *A. thaliana* strain genome on average is not conserved across *A. thaliana* strains in comparison to the 32% that is not conserved across crop strains.

Crop strains encode more strain-specific genes

Perhaps the most prominent observable difference between the frequency spectrum of genes between *A. thaliana* and crop stains is that the pan genome of the crop strains contains an increased number of singleton genes (roughly three times more strain specific genes per crop strain with an average of 311 vs. 110 strain specific genes per crop and *A. thaliana* strain respectively), present in only a single sequenced strain (Fig 3A). Such increased numbers of singletons could reflect host-specific adaptations (but see [8] and discussion in Methods section for alternatives). An alternative hypothesis for this excess of singletons is the greater phylogenetic distance among crop strains. It has been shown that gene content differences frequently correlate with phylogenetic relatedness [32]. Sampling of more closely related strains (such as *A. thaliana* strains) may result in deeper sampling of rare genes making the identification of singletons less likely. To explore this possibility, we tested the relationship between nucleotide divergence between strains and the number of singletons observed (Fig 3C and 3D). When we considered only those crop strains that were as closely related to other strains as are the *A. thaliana* strains, we still found a significantly higher number of strain-specific genes per crop strain (Fig 3D) (Wilcoxon-rank-sum test, $P = 0.003$). These results suggest that the increased number of singletons in crop strains is not fully explained by phylogenetic distance.

Whether most singletons are functional and, furthermore, whether they are adaptive is unclear. We could not detect any functional differences between the genes that exist as singletons in our two populations: in a comparison of gene functions for these singletons, we found significant enrichment only for genes of unknown function (Fisher's exact test q-value<0.01, [33]) but not for any other functional category.

The overlap of pan genomes between and within hosts

When considering particular genes that are variable in their presence, we find the frequency of a gene's presence in strains collected from different hosts to be a good predictor of its frequency in co-occurring strains in *A. thaliana* populations (Fig 3B; Pearson correlation

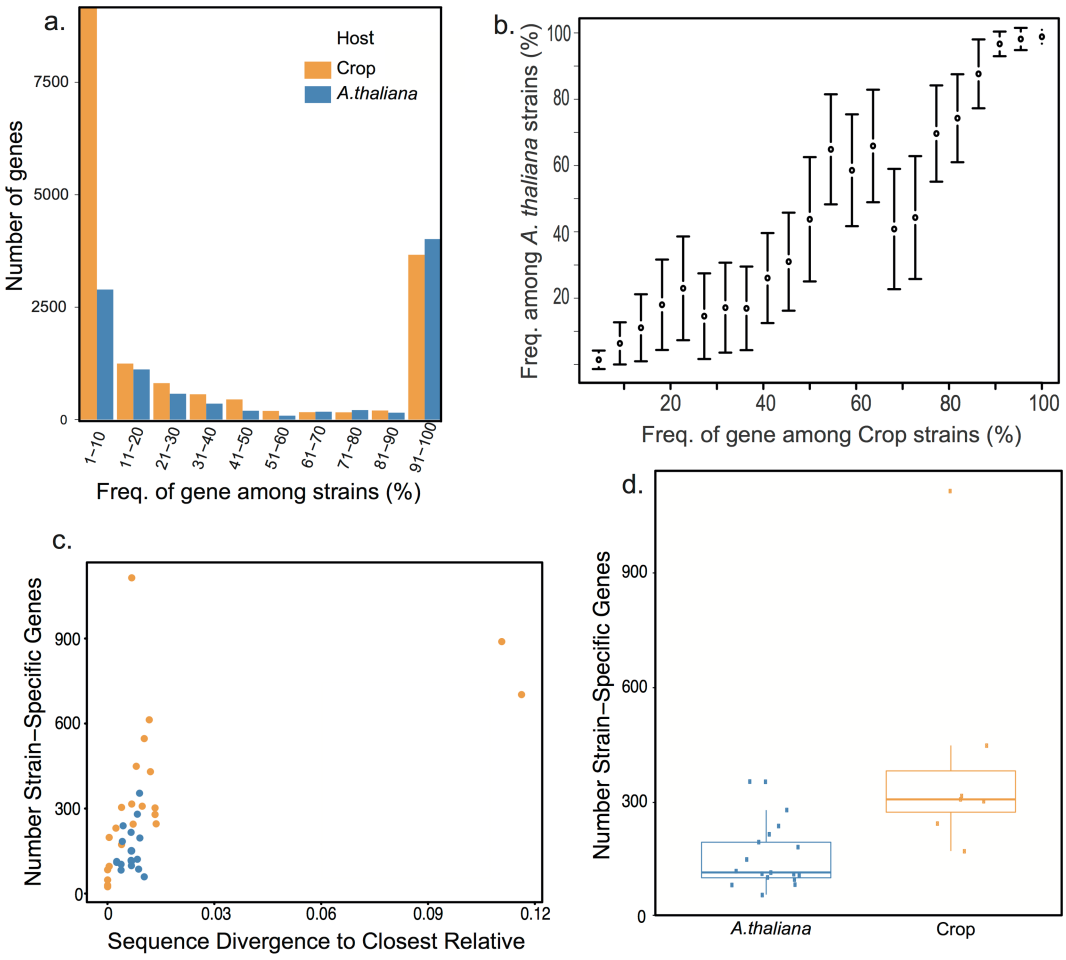

Fig 3. Extensive gene content variation in *A. thaliana* strains mirrors variation between strains from different hosts. (a) The frequency of genes among the 18 *A. thaliana* strains was compared to the frequency of genes across 22 strains collected from different host species, in different locations. Values were binned at 0.1 frequency increments with the number on the x-axis denoting the frequency of a gene across strains. *A. thaliana* and crop strains have a similar number of fixed and high frequency genes, suggesting that both groups have similar numbers of genes important for survival. However, crop strains have significantly more singleton genes (Welch's t-test, P = 0.003). (b) Correlation between the frequency of a gene among crop strains and the frequency of the gene among *A. thaliana* strains. Results are shown as the mean +/ the standard error. (c) Correlation between sequence divergence between a strain and its closest relative with the number of strain-specific genes in that strain (d) The effect observed in (c) is significant also when comparing only those crops strains with divergences similar to those of *A. thaliana* strains (P = 0.003, Wilcoxon-rank-sum test). Results are presented as a box plot, with the mean, 5,25,75 and 95th percentiles illustrated.

https://doi.org/10.1371/journal.pone.0184195.g003

r = 0.94, P < 0.001). That is, the same genes are variable within and between host species. This result suggests that the molecular and/or evolutionary processes that generate and maintain presence/absence polymorphisms are recapitulated within and between hosts and geographic regions. This result is robust to the exclusion of core genes (Pearson correlation r = 0.50, P < 0.001).

The similarity of the within and between host core genomes

The core genome of a microbial taxon group, defined as those genes conserved across isolates of that group, is comprised primarily of genes that have been vertically inherited. The core

genome is therefore thought to be enriched for genes essential for survival [34]. The core genome of *A. thaliana* and crop strains is similar in gene number (4014 genes vs. 3665 genes respectively, defining core genes to be those that are present in >90% of strains within each group). Furthermore, largely the same genes are found in the core genomes of the *A. thaliana* and crop strains of *P. syringae*; 88% (3551 genes) of the core for *A. thaliana* strains overlap with the core for all crop strains. This suggests that the majority of genes fundamental to survival within the relatively constrained Midwestern USA *A. thaliana* environment are the same as those required for survival across multiple host species and geographic locations.

A genomic island encoding toxins active against plants and microbes is enriched in *A. thaliana* strains

While patterns of gene-content variation are similar between strains associated with the same and different hosts, 436 genes are specifically conserved among *A. thaliana* strains (and vary in presence across crop strains). This host-specific conservation could be the result of host specific adaptive conservation. Comparison of the gene functions enriched in the *A. thaliana*-specific suite of genes to those genes conserved across all strains reveals substantial enrichment for functions associated with phytotoxin production (S2 Fig) (322-fold enrichment, Fisher's exact test q-value<0.001 [33]) and minor enrichment for transcriptional regulation (2-fold, Fisher's exact test q-value<0.001). Included in the enriched genes associated with phytotoxin production are those encoding the biosynthetic pathway for syringomycin and syringopeptin, toxins which exhibit broad host-range plant virulence and antifungal properties [35]. Numerous studies have demonstrated that syringomycin and related non-ribosomal lipodepsipeptides can suppress other microbes and increase growth of the pathogen within agricultural plant species [35,36].

The relevance of syringomycin-syringopeptin biosynthesis to *A. thaliana*-associated success is further supported by the phylogenetic distribution of toxin-associated genes. Because *A. thaliana* isolates derive almost exclusively from a single phylogenetic clade, phylogenetic group II (Figs 1 and 2), we aimed to identify genomic regions specific to this clade. We filtered gene content for those genes that were both unique to, and at high frequency (>90%) within, phylogenetic group II. Fifty-six genes met these criteria (Fig 4A, S2 Table), 29.0% (16/56) of which lie in the gene cluster that encodes the proteins necessary for syringomycin-syringopeptin biosynthesis (Fig 4A) [17,35]. Indeed, functional tests of phylogenetic group II strains have confirmed their fungal-suppressive capacity [26]. The remaining phylogenetic group II specific genes are distributed throughout the genome (S2 Table).

Both the functional enrichments and the phylogenetic distribution reveal the tight correlation between strain abundance in *A. thaliana* and the presence of the syringomycin-syringopeptin biosynthetic cluster. Future functional work should investigate whether these broad host-range toxins are necessary for *P. syringae* persistence in *A. thaliana* populations as well as natural plant populations more generally.

A. thaliana-associated *P. syringae* encode few effectors

The Type-III secretion system (T3SS) and its related effectors are prime *a priori* candidates for involvement in adaptation to different host environments. Effectors and the T3SS are central to the capacity of *P. syringae* to infect plants, and the presence or absence of specific effectors can dictate the success of an infection [37]. Effector composition differs between crop strains, and is not tightly correlated with phylogeny, suggesting the rapid gain and loss of these genes [7]. Through computational annotation of the effector content in our *P. syringae* strains, we found that the effector composition of *A. thaliana*-associated *P. syringae* strains also varied,

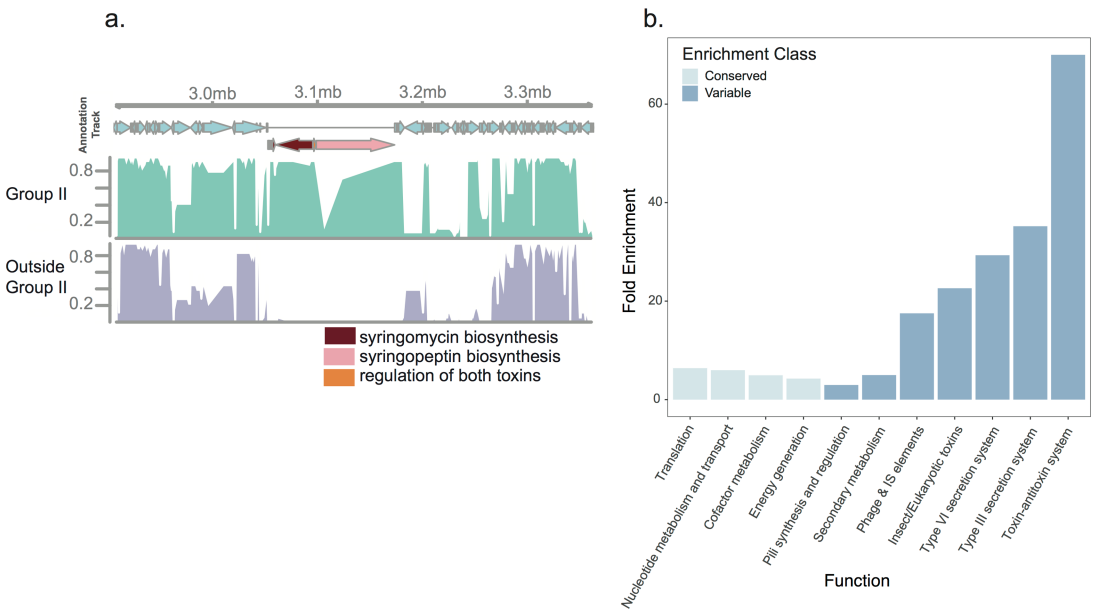

Fig 4. Genes involved in biotic interactions evolve quickly amongst *A. thaliana* strains. The genomes of *P. syringae* from *A. thaliana* contain genes whose products have the ability to suppress a diversity of microbes. (a) A view of gene content conservation surrounding the syringomycin-syringopeptin genomic island. The x-axis denotes the position of the gene in the *P. syringae* strain B728a. The y-axis shows the percent conservation across crop and *A. thaliana P. syringae*. A genomic island encoding the syringomycin biosynthetic cluster is conserved across strains of *P. syringae* from phylogenetic group II. The genes responsible for synthesis of syringomycin and syringopeptin are colored cranberry and pink, respectively. The genes annotated within the orange arrow are involved in the regulation of both toxins. (b) Fold enrichment of gene functions in the conserved and variable portions of the *P. syringae* genome. Only gene categories that were significantly enriched in either the conserved or variable gene sets were included in this figure. Significance was assessed via Fisher's exact test, and a false discovery rate significance level of 0.01.

https://doi.org/10.1371/journal.pone.0184195.g004

with each strain containing homologous sequence to an average of 6.8 effectors. Note that we consider all homologous sequences (50% identity over 50% of the length), including effectors that may also be truncated. The crop *P. syringae* genomes contain an average number of effectors more than four times higher (29.4) than the stains associated with *A. thaliana*. Five of the sequenced *A. thaliana*-associated strains lack a canonical T3SS entirely and encode 0–1 effectors, with only hopAH2 found in these strains (S3 Fig). Note that the classification of hopAH2 as an effector has recently been questioned, due to the absence of the hrp box, or the N-terminus of the T3SS translocator.

The dearth of effectors in the phylogenetic group II that encompasses *A. thaliana* strains was previously noted [7], and postulated to be the result of a physiological trade-off between effector and toxin- mediated interactions with the host [38]. The *A. thaliana*-associated strains may, however, encode other effectors that have yet to be identified in *P. syringae*.

Genes involved in biotic interactions are more likely to exhibit presence/absence polymorphisms

Many of the genes at the interface of host-pathogen interactions evolve quickly, exhibiting both presence/absence and nucleotide variability [7,26]. We determined the gene ontology classifications most frequently found in the variable vs. core genomes for the *A. thaliana P. syringae* sequences. As expected, the core genome is significantly enriched for basic cellular functions such as cell division and translation (Fig 4B). In contrast, the variable genome exhibits

more than a 60-fold enrichment of genes involved in specific plant-microbe interactions (e.g., T3SS and pilus) and genes involved in microbe-microbe interactions (the type six secretion system) (Fig 4B). Virulence factors such as the T3SS and its myriad effectors frequently cluster in genomic islands that exhibit presence/absence polymorphisms (S4 Fig). Phage and insertion sequence (IS) elements are also enriched in the variable component of the genome (Fig 4B). These results indicate that those genes involved in biotic interactions exhibit among the highest rates of gain and loss and also reflect the clade-specific distribution of the syringomycin-syringopeptin synthesis. Whether this increased rate of acquisition and loss is the result of selection, or the result of particular molecular mechanisms (such as heightened rates of uptake and excisions [39,40]) is currently unknown.

Discussion

Horizontal gene transfer and gene loss enable bacterial species to evolve rapidly. A consequence of this genetic malleability is that different strains of the same species can quickly become divergent in the genes they encode. Numerous studies have demonstrated that strains that have likely diverged phenotypically (such as those collected from different hosts) differ extensively in gene content. Whether this variation in gene content is the result of genome-wide adaptation to the colonization of specific environments remains a largely unanswered question. A key to identifying environment-specific genetic adaptations is determining gene content variation both within and between environments.

Here we characterized the extent of gene content diversity present in co-occurring strains of the ubiquitous plant pathogen *P. syringae*. Our comparison of whole genome sequences of *P. syringae* resident on a single host, *A. thaliana*, has revealed patterns of gene content variation very similar to patterns found for crop strains that specialize on diverse host species from diverse geographic locations. Specifically, gene content variation of strains collected from *A. thaliana* is 2/3 that of the variation of strains collected from disparate hosts. This variable portion of the genome consists of genes strongly enriched for microbe-host and microbe-microbe interactions. This raises the possibility that the majority of the variable genome may not be the result of adaptation to alternative hosts but, rather, a more general armament to survive in dynamic and variable host environments. The excess of strain-specific genes among crop strains is an intriguing observation, albeit with several possible explanations. Future work should characterize the function of candidate strain-specific genes in *in planta* survival.

There are also several plausible explanations for the maintenance of gene content diversity within co-occurring strains of *P. syringae*. *A. thaliana* is an annual weedy species, found at relatively low densities among several other plant species [41]. *P. syringae* that propagate in *A. thaliana* can do so only for a portion of the year (due to plant senescence), and must then disperse to other host species or non-host environments [2,25]. This variation in host occupation is likely to differ from the strains isolated from crops, several of which have evolved to specialize on their host of isolation [12,24,42]. The strains isolated from crops have repeatedly if not continuously infected crop host populations [12]

While *P. syringae* on *A. thaliana* may propagate in hosts other than *A. thaliana*, there is no evidence that these *P. syringae* predictably colonizes any other host. The extensive diversity of *P. syringae* strains in non-host environments suggests that many *P. syringae* opportunistically colonize a diversity of environments. It is also likely that the "*A. thaliana*" strains colonize other hosts, though at present we do not have evidence to support this. The different composition of strains infecting *A. thaliana* could be the result of differences in selective pressures in alternative non-host and host environments.

Another possibility is that genetic diversity observed between *A. thaliana* strains is a consequence of intraspecific differences in *A. thaliana* plants. The genes underlying the detection and response to microbes are among the most variable in *A. thaliana* populations [43], and differences in resistance traits could drive the diversification of pathogen genomes. Intraspecific differences in the host environment extend beyond host genetics. *A. thaliana* leaves and roots harbour thousands of microbial species [44–46]. These microbial communities differ between plants of the same genotype, plants of different genotypes and plants in different soils. Consequently, plant-associated *P. syringae* evolve both in response to interactions with the host, but also in response to interactions with other microbial species. The rapid turnover of genes involved in interactions with other microbes (such as the secretion systems) suggests that microbe-microbe interactions contribute to gene content variation within host populations.

It is important to consider the possibility that we observe similar levels of variation among *A. thaliana* strains and among strains collected from various crop hosts because novel genes do not drive adaptation to particular host environments [32,47–49]. That is, it is possible that there exists a pool of genes that will be more readily lost independent of environment, because their maintenance is not strongly favored in any environment [49]. It is also possible that there are genes that, across environments, are more readily acquired via HGT. This, in turn, may lead to the observed pattern by which most genes will be found at similar frequencies within and between environments. Indeed, our finding that the frequency of genes tends to be similar within and between hosts highlights the caution that should be taken when attributing the absence or presence of specific genes to the occupation of specific environments.

Despite the extensive gene content variation we observe among strains residing within *A. thaliana*, these resident strains of *P. syringae* are nevertheless characterized by the presence of a genomic island that encodes for the production of two toxins, syringomycin and syringopeptin. While the full breadth of their toxicities is not well characterized, they are effective against fungi, plants, and gram positive bacteria, pointing towards a potential ability of the *A. thaliana P. syringae* strains to suppress and outcompete other microbes and to infect diverse plant host species.

P. syringae in Midwestern populations of *A. thaliana* do not show evidence of specific adaptation to *A. thaliana* and instead exhibit features suggestive of a generalist lifestyle [50]. Perhaps as a consequence of this generalist lifestyle, the *A. thaliana P. syringae* employ novel tactics to promote their success. The lack of a T3SS in one third of the *P. syringae* that colonize *A. thaliana* allows them to avoid detection, although these strains pay a penalty in terms of virulence when T3SS+ pathogens are not present [23]. Indeed, in a broad GWAS analysis of the host factors shaping microbial communities, there is only a modest contribution from plant *R* genes, the genes classically suspected of engaging in arms races [45]. Clearly, an understanding of microbe-microbe interactions is essential for understanding the structure and distribution of microbial communities *in planta*, as well as the spread of disease.

Materials and methods

Description of strains

P. syringae strains genotyped and sequenced in this study were originally isolated from *A. thaliana* populations in the Midwestern USA between 2000 and 2014. The 18 *P. syringae* strains that underwent full genome sequencing were isolated from nine *A. thaliana* populations residing in agricultural fields in Northwestern Indiana and Southwestern Michigan, USA. These populations are separated by an average distance of 28km and a maximum distance of 98km. The 22 crop strains were isolated from 22 distinct locations, separated by an

average distance of 9648km (Fig 1, S1 Table). The 18 strains isolated from *A*. *thaliana* span the genetic diversity of strains from the *A*. *thaliana* environment, encompassing groups IIa-c [25].

Genomic DNA extraction, DNA sequencing, assembly and annotation

Total DNA was extracted using the Puregene (Illumina) extraction kit from a single colony that was picked and grown overnight in 5mL of King's B media. Colonies were diluted 1:10 in the morning after 12-16hrs, and grown 2–6 hours to an OD600>0.1, which was then followed by DNA extraction. Mate-pair libraries were constructed at Argonne National laboratories and paired end libraries constructed at Beijing Institute for Genomics.

100-base pair reads were generated from the sequencing of mate-pair libraries for 18 *A*. *thaliana* strains on a Genome Analyzer II (Illumina). 75-bp un-paired reads were also generated for each of the strains. *De novo* assembly was performed using Velvet 1.1.05 [51]. Minimus2 from the Amos package [52] was used to merge contigs generated from the different sequencing methods. The genomes were annotated with Rapid Annotation using Subsystem Technology (RAST) server [53]. The 22 previously annotated crop strains were re-annotated using the RAST server for consistency in annotation between all strains. The draft genomes used in this study consisted of single up to thousands of contigs. Due to unavoidable errors in contig assembly, multiple contigs may overlap the same genomic region, resulting in the duplication of genomic regions in an assembly. In consideration of this possibility, duplicate genes (100% identity) within a genome were removed from the annotated dataset, and not considered in the analyses. The genome assembly statistics are presented in Table 1.

Phylogeny construction

To determine the phylogenetic relationship between the strains isolated from *A*. *thaliana* we performed multi-locus-sequencing-analysis (MLSA). The MLSA-based phylogeny evaluating the relationship between 76 strains of *P*. *syringae* isolated from *A*. *thaliana*, 22 crop strains, and a *Pseudomonas fluorescens* outgroup was constructed from six housekeeping genes [54] *cts*, *cgi*, *gyrb*, *can*, *gap A* and *rpoD* using ClonalFrame, a software optimized for estimating phylogenies in the face of bacterial recombination. ClonalFrame [55] was run for 10000 burn-in iterations, 10000 post-burn in iterations, with sampling on every 10th generation. The phylogeny presented represents the 50% majority rule consensus tree. To corroborate the division on the tree, we also constructed a maximum-likelihood phylogeny from the concatenated MLSA sequences. This phylogeny supported the restriction of *A*. *thaliana* strains to phylogenetic group II. The phylogeny for the 40 genomes used for calculations of sequence divergence in this study was constructed using dnaml [56]. Two hundred sixty four genes that were reciprocal best hits were concatenated and used in the dnaml analysis. Orthologs used for phylogenetic analyses were defined as those genes that were reciprocal best hits, and aligned across 100% of the sequence in a global alignment comparison using fasta36 [57]. We use this stringent criterion for determining orthologs when estimating the species phylogeny because we aim to minimize the possible effects of horizontal gene transfer (HGT) on the phylogenetic tree we generate. Highly conserved genes, both in presence and in sequence length are more likely to be vertically inherited than those genes that vary between strains. The 436 genes that met the described criteria and that were observed in all 40 *P*. *syringae* strains and the outgroup *P*. *fluorescens* Pfo-1 strain were concatenated for each strain, and a maximum-likelihood tree for concatenated genes was determined using dnaml [56,58]. While it is possible that some of these 436 conserved genes have been subject to HGT in some of the studied strains, the signal these relatively rare HGT events introduce should be weak considering most concatenated genes are likely to have been inherited vertically along the tree.

Table 1. Assembly information for strains analyzed. The N50 for the strains sequenced in this study is for unscaffolded contigs. Several of the previously sequenced assembled genomes are scaffolded, however, increasing the N50 for these genomes.

Strain	Gene Num.	Num. Contigs	Size (bp)	N50
Pph1448a	5454	3	6112448	Complete
PsyB728a	5254	1	6093698	Complete
PtoDC3000	5661	3	6538260	Complete
PtoMax13	5525	349	6105073	62407
PtoNCPPB1108	5482	304	6082048	47802
Pmp	5327	969	6039297	15161
Pla107	6248	791	6759945	22550
Pmo	5693	3414	6392728	5634
Pja	5674	4661	6380619	4021
PpiR6	5872	5099	6520586	3003
Pma	5474	878	6221751	17222
Psa	5245	941	5849032	14086
Pla106	5293	798	5895184	15738
Ptt	5262	3776	6243278	4753
Pac	5498	1179	6183769	12409
Pta	5541	1613	6158837	16098
PgyB076	5652	104	6236653	202511
PgyR4	5355	108	5905212	3723
Pae	5308	915	5960467	16806
T1	5587	122	6145942	150139
K40	5557	582	6153658	26013
Por1-6	5290	2855	6704257	10037
DM2.1.12.02a	5180	157	5914114	88439
LMC.P10	5226	172	6238204	83072
LMC.P80	5136	137	6189597	120774
LMC.P91	5374	243	6324900	59429
KN2.a.3	5201	133	5804868	120373
Knox623a	5391	554	5946364	22175
Knox652c	5071	160	5905380	95163
NL.P123	5030	220	5855665	61758
NP29.1a	5088	208	6088841	76982
LP217a	5242	382	6224113	47360
LP221b	4967	122	5939644	95628
LP868.1a	5153	123	5907756	111176
RM.P66	5290	70	7081112	284403
RMX.24.a.1	5367	224	6271973	69103
RM.P20	5153	159	6229891	99197
RMX815.1a	5077	167	5565846	67363
ME812.2b	4888	334	5636659	56761
LP205a	4964	478	5698429	30345

https://doi.org/10.1371/journal.pone.0184195.t001

Determining Pseudomonad composition of Midwestern tomato and *A. thaliana* leaves

Leaves from tomato and *A. thaliana* in agricultural and natural populations respectively were collected in November 2013. Single leaves were removed from tomato plants irrespective of plant disease state and frozen immediately in liquid Nitrogen. Whole *A. thaliana* rosettes were

collected, and frozen immediately in liquid Nitrogen. *A. thaliana* roots were removed from the rosette after which *A. thaliana* and tomato leaves were processed identically. The plant material was lyophilized overnight until complete dehydration. The powerplant pro DNA extraction was then used to extract DNA from the lyophilized material. Two sequential extractions were performed on the tomato tissue to remove residual secondary metabolite contamination, which inhibited subsequent polymerase-chain reactions. Bar-coded PCR amplification of the different extractions was then performed to amplify a fragment of *gyrase b* using primers modified from [26] to be able to anneal to an Illumina flow cell. The barcoded samples were then sequenced on the MiSeq using a 500 cycle kit. Fifty samples, 25 from *A. thaliana* and 25 from tomato produced viable reads mapping to *gyrase b*. The resulting reads were then clustered with the usearch-global algorithm to a library of *gyrase b* sequences generated from sequences within the PAMDB [59]. Those clusters that mapped with 75% identity were considered for further analysis.

Identification of conserved vs. variable genes

The pan genome of *A. thaliana* strains and all 40 strains for which full genome sequence data were available were determined using a method similar to that described in [7] but with slight amendments including the alignment algorithm used, and the parameters for determining homologs. In brief, we determined the presence and absence of genes within a genome using an iterative global alignment with the software fasta36 [57]. We compared the translated sequences using a cutoff of 50% homology over 50% the length of one of the two genes. Starting with the ORFs in Pph1448a as the initial pan genome, we compared the translated ORFs of each subsequent draft genome to the pan genome, and determined which genes in the draft genome had not been observed in the compiled pan genome. These unique genes were then added to the pan genome for the next iteration of sequence comparison with the next draft genome. After these iterations had been completed for all 40 strains, we compared the ORFs of each genome to the completed pan genome. This step was necessary to properly incorporate information for genes that were homologous to more than one gene in the pan genome. This method for characterizing the content of the pan genome scores a gene as present if homology is exhibited over either of the genes in a comparison, fragmented coding sequences will align with the intact ortholog, and these fragments will not inflate the number of genes in the pan genome as described [8].

Calculation of sequence divergence

For the calculation of sequence divergence between two strains, we considered only those genes that were found in all *P. syringae* strains, that were reciprocal best hits, and that aligned across 100% of their sequence (fasta36 [57]). These 436 genes were concatenated for each strain and sequence divergence was calculated using the method of [56].

Identification of effector repertoire

Annotations of previously identified *P. syringae* effectors were obtained from [7]. These annotations included 79 effectors, with several effectors represented with more than one allele. The presence of an effector in the *A. thaliana*- associated *P. syringae* genomes was assessed by using tblastn [60] to identify protein homologs in the genomes. A homolog was considered to be present if it matched the previous annotation with 50% identity at the protein level over 50% of the protein. Seven effectors are found in more than 50% of the group II strains. Three of these lie in the conserved effector locus cluster (*avrE*, *hopA1*, *hopM1*) and three lie in another operon (*hopAH1*, *hopAG1*, *hopAI1*). Interestingly, a homolog of *hopAH2* is found in

every *P. syringae* genome (except RMX815.1a) in every phylogenetic clade, including the *P. syringae* strains that lack the canonical T3SS. *hopAH2* lacks a hrp box, and its status as an effector has come into question (http://www.pseudomonas-syringae.org/). It is important to note that in this study we identified only effectors that had previously been annotated, and that our results do not preclude the presence in these strains of effectors that have yet to be annotated as such.

Determining functional enrichments

A recent study [61] manually annotated the gene functions of a *P. syringae* strain closely related to the strains we sequenced here, annotating genes as one of 63 functional categories. These annotations, while not covering the entire genome, are more specific than those provided by RAST and other automated annotation method. Two-thousand eight hundred and fifty genes were included in the enrichment annotations. We classified 2404 of these genes as conserved (present in at least 17 of the 18 genomes or 94.4% of the genomes), and 446 as variable. Using the 63 categories of annotation, we determined enrichments for functions within the conserved and variable categories. Statistical significance was assessed using fisher's exact test, and a False Discovery Rate of 0.01 [33].

Comparisons of the functional enrichments of the strain specific genes were performed using the functional annotations of all genomes generated in the RAST pipeline [53]. Subsystem annotations were compared across genomes with the R package 'TopGO' (created by Adrian Alexa and Jorg Rahnenfuhrer), and significant enrichment was assessed using fisher's exact test, and a False Discovery Rate of 0.01 [33].

Supporting information

S1 Fig. Relationship between genome assembly N50 and number of genes identified per genome. The relationship between genome assembly quality and the number of genes (non-duplicates) annotated per genome is not significantly different from zero (linear regression, P = 0.291). The opaque gray shows the 95% confidence interval for the predicted relationship.
(PDF)

S2 Fig. Fold enrichment of gene functions enriched among genes conserved specifically in *A. thaliana* strains in comparison to genes conserved across all strains. Only gene categories that were significantly enriched in either the conserved or variable gene sets were included in this figure. Significance was assessed via Fisher's exact test, and a false discovery rate of 0.01.
(PDF)

S3 Fig. *A. thaliana* associated strains encode few canonical *P. syringae* effectors. The x-axis details the 79 effectors (or alleles of effectors) that were identified previously in *P. syringae* genomes [7]. The y-axis shows the 18 genomes sequenced in this study, and whether an orthologue of these effectors was identified (dark blue indicates presence in genome).
(PDF)

S4 Fig. Gene content variation at the conserved effector locus. Conservation of genes along the conserved effector locus (shaded in red). Approximately one third of strains collected from *A. thaliana* lack the canonical T3SS [23].
(PDF)

S1 Table. Location of isolation for strains analyzed in this study. Information was not available for several strains (listed as "NA"). When non-specific isolation information was provided by collector (such as country of origin), the latitude of country is provided. LMC is an

abbreviation for Lake Michigan College.
(PDF)

S2 Table. Genes unique to and conserved in group II. This table shows the position in the B728a genome of each of the genes unique to group II. Fifty-six genes were conserved in 90% or more of the strains in group II, but absent outside group II. 29% of these genes lie in the syringomycin/syringopeptin biosynthetic cluster, highlighted in Cyan.
(PDF)

Acknowledgments

We thank David Baltrus for genomic DNA and providing plasmids used for cloning in addition to useful consultation on the analyses. We are also grateful to Dennis Gross for sharing bacterial strains and Tim Morton, Allie Kreitman and Eric Laderman who aided in experiments. Jean Greenberg and Michael Werner further provided thoughtful comments on the manuscript.

Author Contributions

Conceptualization: Talia L. Karasov, Luke Barrett, Ruth Hershberg, Joy Bergelson.

Data curation: Talia L. Karasov, Luke Barrett.

Formal analysis: Talia L. Karasov, Ruth Hershberg.

Funding acquisition: Joy Bergelson.

Investigation: Luke Barrett, Ruth Hershberg, Joy Bergelson.

Methodology: Talia L. Karasov, Luke Barrett, Ruth Hershberg.

Resources: Talia L. Karasov, Luke Barrett, Joy Bergelson.

Software: Talia L. Karasov, Ruth Hershberg.

Supervision: Luke Barrett, Ruth Hershberg, Joy Bergelson.

Validation: Talia L. Karasov.

Visualization: Talia L. Karasov.

Writing – original draft: Talia L. Karasov, Luke Barrett, Joy Bergelson.

Writing – review & editing: Talia L. Karasov, Joy Bergelson.

References

1. Savageau MA. Escherichia coli Habitats, Cell Types, and Molecular Mechanisms of Gene Control. Am Nat. [University of Chicago Press, American Society of Naturalists]; 1983; 122: 732–744.
2. Morris CE, Sands DC, Vinatzer BA, Glaux C, Guilbaud C, Buffière A, et al. The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle. ISME J. 2008; 2: 321–334. https://doi.org/10.1038/ismej.2007.113 PMID: 18185595
3. Thrall PH, Burdon JJ. Evolution of virulence in a plant host-pathogen metapopulation. Science. 2003; 299: 1735–1737. https://doi.org/10.1126/science.1080070 PMID: 12637745
4. Thrall PH, Burdon JJ, Young A. Variation in resistance and virulence among demes of a plant host–pathogen metapopulation. J Ecol. 2001; 89: 736.
5. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, et al. Genotypic diversity within a natural coastal bacterioplankton population. Science. 2005; 307: 1311–1313. https://doi.org/10.1126/science.1106028 PMID: 15731455

6. Loper JE, Hassan KA, Mavrodi DV, Davis EW 2nd, Lim CK, Shaffer BT, et al. Comparative genomics of plant-associated Pseudomonas spp.: insights into diversity and inheritance of traits involved in multitrophic interactions. PLoS Genet. 2012; 8: e1002784. https://doi.org/10.1371/journal.pgen.1002784 PMID: 22792073
7. Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, Cherkis K, et al. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. PLoS Pathog. 2011; 7: e1002132. https://doi.org/10.1371/journal.ppat.1002132 PMID: 21799664
8. Nowell RW, Green S, Laue BE, Sharp PM. The extent of genome flux and its role in the differentiation of bacterial lineages. Genome Biol Evol. 2014; 6: 1514–1529. https://doi.org/10.1093/gbe/evu123 PMID: 24923323
9. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009; 5: e1000344. https://doi.org/10.1371/journal.pgen.1000344 PMID: 19165319
10. Bartoli C, Berge O, Monteil CL, Guilbaud C, Balestra GM, Varvaro L, et al. The Pseudomonas viridiflava phylogroups in the P. syringae species complex are characterized by genetic variability and phenotypic plasticity of pathogenicity-related traits. Environ Microbiol. 2014; 16: 2301–2315. https://doi.org/10.1111/1462-2920.12433 PMID: 24612372
11. Baltrus DA, McCann HC, Guttman DS. Evolution, Genomics and Epidemiology of Pseudomonas syringae. Mol Plant Pathol. Wiley Online Library; 2016; https://doi.org/10.1111/mpp.12506 PMID: 27798954
12. Cai R, Lewis J, Yan S, Liu H, Clarke CR, Campanile F, et al. The plant pathogen Pseudomonas syringae pv. tomato is genetically monomorphic and under strong selection to evade tomato immunity. PLoS Pathog. 2011; 7: e1002130. https://doi.org/10.1371/journal.ppat.1002130 PMID: 21901088
13. Baltrus DA, Nishimura MT, Dougherty KM, Biswas S, Mukhtar MS, Vicente J, et al. The molecular basis of host specialization in bean pathovars of Pseudomonas syringae. Mol Plant Microbe Interact. 2012; 25: 877–888. https://doi.org/10.1094/MPMI-08-11-0218 PMID: 22414441
14. Butler MI, Stockwell PA, Black MA, Day RC, Lamont IL, Poulter RTM. Pseudomonas syringae pv. actinidiae from recent outbreaks of kiwifruit bacterial canker belong to different clones that originated in China. PLoS One. journals.plos.org; 2013; 8: e57464. https://doi.org/10.1371/journal.pone.0057464 PMID: 23555547
15. Nowell RW, Laue BE, Sharp PM, Green S. Comparative genomics reveals genes significantly associated with woody hosts in the plant pathogen Pseudomonas syringae. Mol Plant Pathol. 2016; https://doi.org/10.1111/mpp.12423 PMID: 27145446
16. Tsiamis G, Mansfield JW, Hockenhull R, Jackson RW, Sesma A, Athanassopoulos E, et al. Cultivar-specific avirulence and virulence functions assigned to avrPphF in Pseudomonas syringae pv. phaseolicola, the cause of bean halo-blight disease. EMBO J. EMBO Press; 2000; 19: 3204–3214. https://doi.org/10.1093/emboj/19.13.3204 PMID: 10880434
17. Feil H, Feil WS, Chain P, Larimer F, DiBartolo G, Copeland A, et al. Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci U S A. 2005; 102: 11064–11069. https://doi.org/10.1073/pnas.0504930102 PMID: 16043691
18. Sarkar SF, Gordon JS, Martin GB, Guttman DS. Comparative genomics of host-specific virulence in Pseudomonas syringae. Genetics. 2006; 174: 1041–1056. https://doi.org/10.1534/genetics.106.060996 PMID: 16951068
19. Jakob K, Goss EM, Araki H, Van T, Kreitman M, Bergelson J. Pseudomonas viridiflava and P. syringae—natural pathogens of Arabidopsis thaliana. Mol Plant Microbe Interact. 2002; 15: 1195–1203. https://doi.org/10.1094/MPMI.2002.15.12.1195 PMID: 12481991
20. Kniskern JM, Traw MB, Bergelson J. Salicylic acid and jasmonic acid signaling defense pathways reduce natural bacterial diversity on Arabidopsis thaliana. Mol Plant Microbe Interact. 2007; 20: 1512–1522. https://doi.org/10.1094/MPMI-20-12-1512 PMID: 17990959
21. Gao L, Roux F, Bergelson J. Quantitative fitness effects of infection in a gene-for-gene system. New Phytol. Wiley Online Library; 2009; 184: 485–494. https://doi.org/10.1111/j.1469-8137.2009.02959.x PMID: 19659661
22. Ferrante P, Scortichini M. Molecular and phenotypic features of Pseudomonas syringae pv. actinidiae isolated during recent epidemics of bacterial canker on yellow kiwifruit (Actinidia chinensis) in central Italy. Plant Pathol. Blackwell Publishing Ltd; 2010; 59: 954–962.
23. Barrett LG, Bell T, Dwyer G, Bergelson J. Cheating, trade-offs and the evolution of aggressiveness in a natural pathogen population. Ecol Lett. 2011; 14: 1149–1157. https://doi.org/10.1111/j.1461-0248.2011.01687.x PMID: 21951910
24. Karasov TL, Horton MW, Bergelson J. Genomic variability as a driver of plant-pathogen coevolution? Curr Opin Plant Biol. 2014; 18: 24–30. https://doi.org/10.1016/j.pbi.2013.12.003 PMID: 24491596

25. Kniskern JM, Barrett LG, Bergelson J. Maladaptation in wild populations of the generalist plant pathogen Pseudomonas syringae. Evolution. 2011; 65: 818–830. https://doi.org/10.1111/j.1558-5646.2010.01157.x PMID: 21044058

26. Hwang MSH, Morgan RL, Sarkar SF, Wang PW, Guttman DS. Phylogenetic characterization of virulence and resistance phenotypes of Pseudomonas syringae. Appl Environ Microbiol. 2005; 71: 5182–5191. https://doi.org/10.1128/AEM.71.9.5182-5191.2005 PMID: 16151103

27. Clarke CR, Cai R, Studholme DJ, Guttman DS, Vinatzer BA. Pseudomonas syringae strains naturally lacking the classical P. syringae hrp/hrc locus are common leaf colonizers equipped with an atypical type III secretion system. Mol Plant Microbe Interact. Am Phytopath Society; 2010; 23: 198–210. https://doi.org/10.1094/MPMI-23-2-0198 PMID: 20064063

28. Hirano SS, Upper CD. Bacteria in the Leaf Ecosystem with Emphasis onPseudomonas syringae—a Pathogen, Ice Nucleus, and Epiphyte. Microbiol Mol Biol Rev. 2000; 64: 624–653. PMID: 10974129

29. Renick LJ, Cogal AG, Sundin GW. Phenotypic and Genetic Analysis of Epiphytic Pseudomonas syringae Populations from Sweet Cherry in Michigan. Plant Dis. 2008; 92: 372–378.

30. Morris CE, Sands DC, Vanneste JL, Montarry J, Oakley B, Guilbaud C, et al. Inferring the evolutionary history of the plant pathogen Pseudomonas syringae from its biogeography in headwaters of rivers in North America, Europe, and New Zealand. MBio. 2010; 1. https://doi.org/10.1128/mBio.00107-10 PMID: 20802828

31. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005; 15: 589–594. https://doi.org/10.1016/j.gde.2005.09.006 PMID: 16185861

32. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics. 2011; 12: 32. https://doi.org/10.1186/1471-2164-12-32 PMID: 21232151

33. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. [Royal Statistical Society, Wiley]; 1995; 57: 289–300.

34. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 2005; 102: 13950–13955. https://doi.org/10.1073/pnas.0506758102 PMID: 16172379

35. Guenzi E, Galli G, Grgurina I, Gross DC, Grandi G. Characterization of the syringomycin synthetase gene cluster. A link between prokaryotic and eukaryotic peptide synthetases. J Biol Chem. 1998; 273: 32857–32863. PMID: 9830033

36. Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JHM, et al. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science. 2011; 332: 1097–1100. https://doi.org/10.1126/science.1203980 PMID: 21551032

37. Jones JDG, Dangl JL. The plant immune system. Nature. 2006; 444: 323–329. https://doi.org/10.1038/nature05286 PMID: 17108957

38. Hockett KL, Nishimura MT, Karlsrud E, Dougherty K, Baltrus DA. Pseudomonas syringae CC1557: a highly virulent strain with an unusually small type III effector repertoire that includes a novel effector. Mol Plant Microbe Interact. 2014; 27: 923–932. https://doi.org/10.1094/MPMI-11-13-0354-R PMID: 24835253

39. Jackson RW, Mansfield JW, Arnold DL, Sesma A, Paynter CD, Murillo J, et al. Excision from tRNA genes of a large chromosomal region, carrying avrPphB, associated with race change in the bean pathogen, Pseudomonas syringae pv. phaseolicola. Mol Microbiol. 2000; 38: 186–197. PMID: 11069647

40. Godfrey SAC, Lovell HC, Mansfield JW, Corry DS, Jackson RW, Arnold DL. The stealth episome: suppression of gene expression on the excised genomic island PPHGI-1 from Pseudomonas syringae pv. phaseolicola. PLoS Pathog. 2011; 7: e1002010. https://doi.org/10.1371/journal.ppat.1002010 PMID: 21483484

41. Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M. Arabidopsis thaliana: a model plant for genome analysis. Science. 1998; 282: 662, 679–82. PMID: 9784120

42. McCann HC, Rikkerink EHA, Bertels F, Fiers M, Lu A, Rees-George J, et al. Genomic analysis of the Kiwifruit pathogen Pseudomonas syringae pv. actinidiae provides insight into the origins of an emergent plant disease. PLoS Pathog. 2013; 9: e1003503. https://doi.org/10.1371/journal.ppat.1003503 PMID: 23935484

43. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science. 2007; 317: 338–342. https://doi.org/10.1126/science.1138632 PMID: 17641193

44. Bodenhausen N, Horton MW, Bergelson J. Bacterial communities associated with the leaves and the roots of Arabidopsis thaliana. PLoS One. 2013; 8: e56329. https://doi.org/10.1371/journal.pone.0056329 PMID: 23457551

45. Horton MW, Bodenhausen N, Beilsmith K, Meng D, Muegge BD, Subramanian S, et al. Genome-wide association study of Arabidopsis thaliana leaf microbial community. Nat Commun. 2014; 5: 5320. https://doi.org/10.1038/ncomms6320 PMID: 25382143

46. Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S, et al. Defining the core Arabidopsis thaliana root microbiome. Nature. 2012; 488: 86–90. https://doi.org/10.1038/nature11237 PMID: 22859206

47. Haegeman B, Weitz JS. A neutral theory of genome evolution and the frequency distribution of genes. BMC Genomics. 2012; 13: 196. https://doi.org/10.1186/1471-2164-13-196 PMID: 22613814

48. Bolotin E, Hershberg R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol. 2015; 7: 2173–2187. https://doi.org/10.1093/gbe/evv135 PMID: 26163675

49. Bolotin E, Hershberg R. Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes. Sci Rep. 2016; 6: 35168. https://doi.org/10.1038/srep35168 PMID: 27734920

50. Woolhouse ME, Taylor LH, Haydon DT. Population biology of multihost pathogens. Science. 2001; 292: 1109–1112. PMID: 11352066

51. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18: 821–829. https://doi.org/10.1101/gr.074492.107 PMID: 18349386

52. Arbuckle JL. Amos (version 7.0)[computer program]. Chicago: SpSS. 2006;

53. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008; 9: 75. https://doi.org/10.1186/1471-2164-9-75 PMID: 18261238

54. Sarkar SF, Guttman DS. Evolution of the Core Genome of Pseudomonas syringae, a Highly Clonal, Endemic Plant Pathogen. Appl Environ Microbiol. 2008; 74: 1961–1961.

55. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics. 2007; 175: 1251–1266. https://doi.org/10.1534/genetics.106.063305 PMID: 17151252

56. Hasegawa M, Kishino H, Saitou N. On the maximum likelihood method in molecular phylogenetics. J Mol Evol. 1991; 32: 443–445. PMID: 1904100

57. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988; 85: 2444–2448. PMID: 3162770

58. Felsenstein J. Phylogenies and the Comparative Method. Am Nat. [University of Chicago Press, American Society of Naturalists]; 1985; 125: 1–15.

59. Almeida NF, Yan S, Cai R, Clarke CR, Morris CE, Schaad NW, et al. PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. Phytopathology. 2010; 100: 208–215. https://doi.org/10.1094/PHYTO-100-3-0208 PMID: 20128693

60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

61. Yu X, Lund SP, Scott RA, Greenwald JW, Records AH, Nettleton D, et al. Transcriptional responses of Pseudomonas syringae to growth in epiphytic versus apoplastic leaf sites. Proc Natl Acad Sci U S A. 2013; 110: E425–34. https://doi.org/10.1073/pnas.1221892110 PMID: 23319638