RESEARCH ARTICLE

Predictive modeling of *Pseudomonas syringae* virulence on bean using gradient boosted decision trees

Renan N. D. Almeida[1], Michael Greenberg[1], Cedoljub Bundalovic-Torma[1], Alexandre Martel[1], Pauline W. Wang[1,2], Maggie A. Middleton[1,2], Syama Chatterton[3], Darrell Desveaux[1], David S. Guttman[1,2]*

1 Department of Cell & Systems Biology, University of Toronto, Toronto, Canada, **2** Centre for the Analysis of Genome Evolution & Function, University of Toronto, Toronto, Canada, **3** Agriculture and Agri-Food Canada, Lethbridge Research and Development Centre, Lethbridge, Canada

* david.guttman@utoronto.ca

Abstract

Pseudomonas syringae is a genetically diverse bacterial species complex responsible for numerous agronomically important crop diseases. Individual *P*. *syringae* isolates are assigned pathovar designations based on their host of isolation and the associated disease symptoms, and these pathovar designations are often assumed to reflect host specificity although this assumption has rarely been rigorously tested. Here we developed a rapid seed infection assay to measure the virulence of 121 diverse *P*. *syringae* isolates on common bean (*Phaseolus vulgaris*). This collection includes *P*. *syringae* phylogroup 2 (PG2) bean isolates (pathovar *syringae*) that cause bacterial spot disease and *P*. *syringae* phylogroup 3 (PG3) bean isolates (pathovar *phaseolicola*) that cause the more serious halo blight disease. We found that bean isolates in general were significantly more virulent on bean than non-bean isolates and observed no significant virulence difference between the PG2 and PG3 bean isolates. However, when we compared virulence within PGs we found that PG3 bean isolates were significantly more virulent than PG3 non-bean isolates, while there was no significant difference in virulence between PG2 bean and non-bean isolates. These results indicate that PG3 strains have a higher level of host specificity than PG2 strains. We then used gradient boosting machine learning to predict each strain's virulence on bean based on whole genome k-mers, type III secreted effector k-mers, and the presence/absence of type III effectors and phytotoxins. Our model performed best using whole genome data and was able to predict virulence with high accuracy (mean absolute error = 0.05). Finally, we functionally validated the model by predicting virulence for 16 strains and found that 15 (94%) had virulence levels within the bounds of estimated predictions. This study strengthens the hypothesis that *P*. *syringae* PG2 strains have evolved a different lifestyle than other *P*. *syringae* strains as reflected in their lower level of host specificity. It also acts as a proof-of-principle to demonstrate the power of machine learning for predicting host specific adaptation.

OPEN ACCESS

Citation: Almeida RND, Greenberg M, Bundalovic-Torma C, Martel A, Wang PW, Middleton MA, et al. (2022) Predictive modeling of *Pseudomonas syringae* virulence on bean using gradient boosted decision trees. PLoS Pathog 18(7): e1010716. https://doi.org/10.1371/journal.ppat.1010716

Editor: David A. Baltrus, University of Arizona, UNITED STATES

Received: January 6, 2022

Accepted: June 30, 2022

Published: July 25, 2022

Data Availability Statement: All accession numbers for sequence data are specified in S1 Table.

Funding: This study was funded by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to DSG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Author summary

Pseudomonas syringae is a genetically diverse Gammaproteobacterial species complex responsible for numerous agronomically important crop diseases. Strains in the *P. syringae* species complex are frequently categorized into pathovars depending on pathogenic characteristics such as host of isolation and disease symptoms. Common bean pathogens from *P. syringae* are known to cause two major diseases: (1) pathovar *phaseolicola* strains from phylogroup 3 cause halo blight disease, characterized by large necrotic lesions surrounded by a chlorotic zone or halo of yellow tissue; and (2) pathovar *syringae* strains from phylogroup 2 causes bacterial spot disease, characterized by brown leaf spots. While halo blight can cause serious crop losses, bacterial spot disease is generally of minor agronomic concern. Recently, statistical genetic and machine learning approaches have been applied to genomic data to identify genes underlying traits of interest or predict the outcome of host-microbe interactions. Here, we apply machine learning to *P. syringae* genomic data to predict virulence on bean. We first characterized the virulence of *P. syringae* isolates on common bean using a seed infection assay and then applied machine learning to the genomic data from the same strains to generate a predictive model for virulence on bean. We found that machine learning models built with k-mers from either full genome data or virulence factors could predict bean virulence with high accuracy. We also confirmed prior work showing that phylogroup 3 halo blight pathogens display a stronger degree of phylogenetic clustering and host specificity compared to phylogroup 2 brown spot pathogens. This works serves as a proof-of-principle for the power of machine learning for predicting host specificity and may find utility in agricultural diagnostic microbiology.

Introduction

Pseudomonas syringae is a genetically diverse Gammaproteobacterial species complex responsible for numerous agronomically important crop diseases [1–4]. Strains in the *P. syringae* species complex are frequently categorized into pathovars depending on pathogenic characteristics such as host of isolation and disease symptoms [5,6]. The species complex is also subdivided into phylogenetic groups (i.e., phylogroups, PGs) based on multilocus sequence typing or genomic analysis [1,7–10]. Currently, there are 13 recognized PGs [7], of which seven have been termed primary PGs based on their higher degree of genetic relatedness and the near universal presence of the canonical *P. syringae* type III secretion system (discussed below) [1,9]. In contrast, secondary PGs are genetically more diverse, include a larger fraction of environmental isolates, and are more likely to carry alternative type III secretion systems.

Strains in *P. syringae* complex have historically been considered to have high levels of host specificity [6,11,12]. This conclusion came from observed similarity of strains isolated off common hosts based on phenotypic or molecular typing and is the basis for the pathovar taxonomic system. The inherent assumption underlying this conclusion is that strains of the same pathovar should have higher fitness on one host than other hosts. The problem with this assumption is that it has rarely been rigorously and systematically tested. In fact, in the few cases where this has been tested, strains were found to show much more complex patterns of host specificity, with some having narrow ranges, while other are much more generalists [13,14]. In particular, PG2 strains seem to show the lowest degree of host specificity and be better adapted to the epiphytic environment than other *P. syringae* strains [1,3,11,13–16].

A particularly interesting host specificity pattern is when two or more evolutionarily distinct clades within the *P. syringae* complex have adapted to the same host. Phylogenetic analyses of *P. syringae* isolates suggest that this convergent host adaptation has occurred multiple times in the evolutionary history of the species complex. For example, cherry and plum pathogens are found in clades distributed in PG1, PG2, and PG3 [17,18], hazelnut pathogens are distributed among two distinct clades in PG1 and PG2 [19].

This study focuses on one of the most interesting examples of convergent host adaptation–*P. syringae* pathogens of common bean (including snap, green, kidney, and French bean). All *P. syringae* primary phylogroup bean isolates are found in either PG2 or PG3 [8,9]. The only other bean isolates reported in the *P. syringae* complex are a small number of *Pseudomonas viridiflava* strains in the much more divergent secondary phylogroups [20]. Common bean pathogens from *P. syringae* PG3 are generally classified as pathovar *phaseolicola* and are responsible for halo blight disease, which is characterized by large necrotic lesions surrounded by a chlorotic zone or halo of yellow tissue [21–23]. Bean pathogens of PG2 are generally classified as pathovar *syringae* and are responsible for bacterial spot disease, which is characterized by brown leaf spots [24,25]. While halo blight can cause serious crop losses, bacterial spot disease is generally of minor agronomic concern. The PG3 *phaseolicola* bean isolates show a high degree of phylogenetic clustering, with most strains sharing a relatively recent common ancestor that is closely related to a compact sister clade of soybean pathogens [9]. In contrast, PG2 *syringae* bean isolates show very little phylogenetic clustering and are frequently more closely related to non-bean isolates than other bean isolates [9].

Assuming that host specificity is a heritable trait, the exploitation of a common host by divergent lineages of strains can be explained by several different mechanisms, including: 1) evolution via shared, vertically transmitted host specificity factors; 2) convergent evolution via unrelated genetic mechanisms; or 3) convergent evolution via the horizontal acquisition of host specificity factors from divergent lineages. Another layer of complexity is that host specificity could come about either through the gain of genetic factors that promotes growth on a new host, or alternatively, by the loss of a factor that otherwise limits growth (e.g., by inducing a host immune response). In fact, the most thorough study of host convergence in *P. syringae* suggests that isolates can make use of multiple mechanisms simultaneously [17,18]. For example, diverse lineages of cherry pathogens have exchanged and lost key genes and used multiple mechanisms to successfully infect this host [17,18].

One of the most important and dynamic classes of *P. syringae* virulence and host specificity factors are type III secreted effectors (T3SEs). T3SEs are proteins translocated through the type III secretion system directly into the eukaryotic host cell where they interfere with host immunity or disrupt cellular homeostasis to promote the disease process. There are at least 70 distinct families of *P. syringae* T3SEs, and most strains carry a suite of T3SEs consisting of 12 to 50 T3SEs, with an average of ~30 [26]. Plants have responded to T3SEs by evolving immune receptors and complexes that trigger an effector-triggered immune (ETI) response when they detect the presence or activity of a T3SE [27,28]. Consequently, the outcome of any particular host-microbe interaction depends to a large degree on the specific T3SE profile of the pathogen and the complement of immune receptors carried by the host. The strong selective pressures imposed by the host-microbe arms race results in dynamic evolution of T3SEs in general, with frequent horizontal transmission, acquisition, and loss [9,26,29].

The suites of T3SEs carried by PG2 and PG3 strains vary in size, with PG2 strains carrying an average of ~19 T3SEs vs. ~27 for PG3 strains [26]. PG2 strains are also known to carry more phytotoxins, which contribute to virulence and niche competition via a variety of mechanisms such as membrane disruption and hormone mimicry [3,9,30]. These differences may help explain why PG2 strains show lower levels of host specificity and are better ability to survive on leaf surfaces (i.e., epiphytic growth) [1,3,11,13–16].

The application of statistical genetic and machine learning approaches to genomic data has greatly increased our power to identify genes underlying traits of interest, such as host specificity [31]. Statistical genetic approaches like genome-wide association studies (GWAS) are well developed for studying human traits and have more recently gained traction in the study of bacterial traits as statistical and phylogenetic methods have been developed to handle the shared evolutionary history of segregating genetic variants (i.e., population structure) [32–37]. While GWAS approaches have great power for finding genotype-phenotype associations, they generally measure associations on a locus-by-locus basis, and therefore can miss more complex interactions among loci that impact traits. An alternative approach for predicting genotype-phenotype associations is to use machine learning, which generally describes a large range of statistical approaches that create models derived from a dataset consisting of features (e.g., genetic variants) linked to a trait or outcome (e.g., host specificity). These models can be used to predict outcomes from new samples or to identify the feature(s) that carry the most importance in the model. Although machine learning approaches may be better suited for identifying interactions among genetic variants than GWAS, they are more limited in their ability to deal with complex evolutionary relationships among these variants [38,39].

Here, we implemented a rapid method for assessing *P. syringae* virulence on common bean. We used this screen to measure the virulence of 121 strains from nine phylogroups on bean, and then further expanded the dataset by imputing the virulence for an additional set of isolates based on their core genome relationship to the screened strains. We found that PG3 pathogens display a stronger degree of host specificity compared to PG2 pathogens. We then developed a gradient boosting regression model using k-mers derived from the whole genome sequence or virulence factors as features to predict the virulence of *P. syringae* isolates on bean. The model showed good performance and was able to predict the virulence of a set of test strains with high accuracy. This study acts as a proof-of-principle for the utility of machine learning to the prediction of plant-microbe interactions.

Results

Genome analysis

We characterized the genomic diversity of the 333 *P. syringae* isolates, including 46 newly sequenced bean isolates (18 PG2 pv. *syringae* and 28 PG3 pv. *phaseolicola*) collected from bean fields approximately 80 km east of Lethbridge, Alberta, Canada in 2012 via phylogenetic analysis (S1 Table). Core genome diversity was measured by synonymous substitution rates (Ks), while accessory genome diversity was measured by pairwise Jaccard distances. The collection includes 36 bean halo blight pathogens of pathovar *phaseolicola*, which all cluster in one closely related clade in PG3, with a core genome Ks of 0.0039 and an accessory genome Jaccard distance of 0.35 compared to the entirety of 142 PG3 strains, which had a core genome Ks of 0.047 and an accessory genome Jaccard distance of 0.64. While the 28 newly sequenced Canadian bean isolates from PG3 (pv. *phaseolicola*) all cluster, they are interspersed with other *phaseolicola* strain, indicating that they do not result in a biased assessment of PG3 bean isolate similarity. In contrast the 21 bean spot disease pathogens of pathovar *syringae* are broadly distributed throughout PG2 and had a core genome Ks of 0.1218 and an accessory genome Jaccard distance of 0.45 compared to the entirety of 66 PG2 strains, which had a core genome Ks of 0.1223 and an accessory genome Jaccard distance of 0.67 (Fig 1). Like what was found with the new PG3 Canadian bean isolates, the 18 newly analyzed PG2 (pv. *syringae*) isolates from Canada are interspersed with other PG2 bean isolates. Due to their clonal nature, PG3 bean

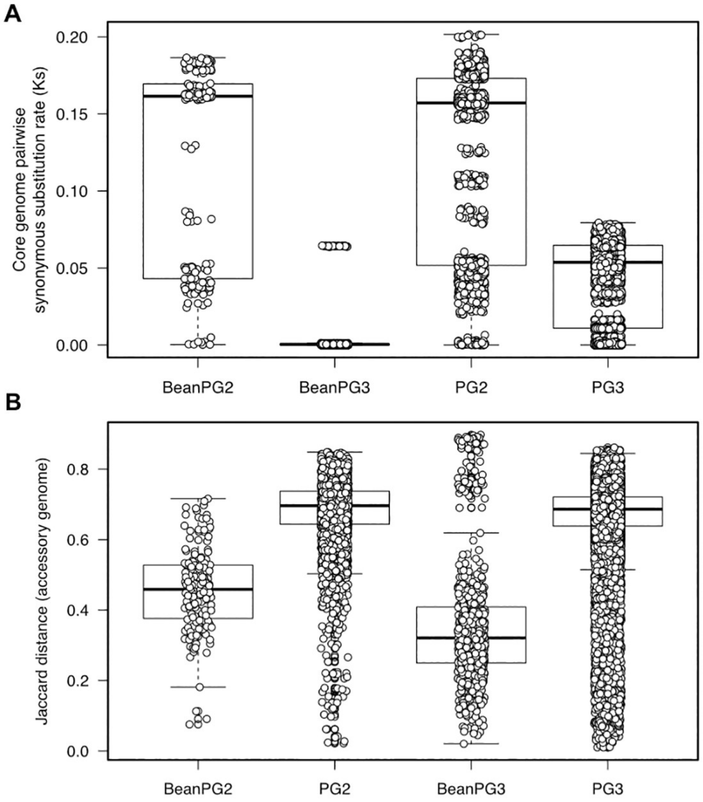

Fig 1. Core and accessory genome diversity. Comparison of (A) core genome synonymous substitution rate (Ks) and (B) accessory genome Jaccard distance for 21 PG2 bean isolates, 36 PG3 bean isolates, and all 66 isolates from PG2 and 142 isolates from PG3.

https://doi.org/10.1371/journal.ppat.1010716.g001

isolates were found to have a considerably higher number of gene families in the hard core (present in 100% of the isolates) and soft core (present in >95% of the isolates) genomes, as well as a lower number of singleton families (present in a single isolate) in comparison to PG2 bean isolates, despite the larger number of PG3 samples analyzed (Fig 2).

Virulence screen development

We developed a high-throughput seed infection assay to measure the virulence of *P. syringae* isolates on common bean. Given that contaminated seeds are a common inoculation source for bean infection, this assay provides a means to quantify host-pathogen interactions that closely reflects the 'natural' interaction [21,40–42]. For the screen, we soaked bean seeds (*P. vulgaris* var. Canadian Red) in a *P. syringae* suspension (~5x10^5 cells / ml) for 24 hours prior to planting, and measured plant fresh weight after 14 days. Bacterial virulence resulted in disease symptoms (S1 Fig) and reductions of overall plant health, which is reflected in lower plant fresh weight. We confirmed that virulence was type III dependent using a *hrcC* mutant of the bean pathogen *P. syringae* pv. *phaseolicola* 1448A (Pph1448A) (S2 Fig), and then assessed if plant weight was correlated with *in planta* bacterial load by comparing our seed infection assay to the traditional syringe infiltration virulence assay using 24 *P. syringae* isolates from 9 out of the 13 PGs (Fig 3). Well-established bean pathogens such as PG3 strain Pph1448A [21,23] and the PG2 strains *P. syringae* pv. *syringae* B728a (PsyB728a) [24,43] showed the highest levels of bacterial growth and lowest plant weights, while the other isolates from PGs 1–7, 11, and 13

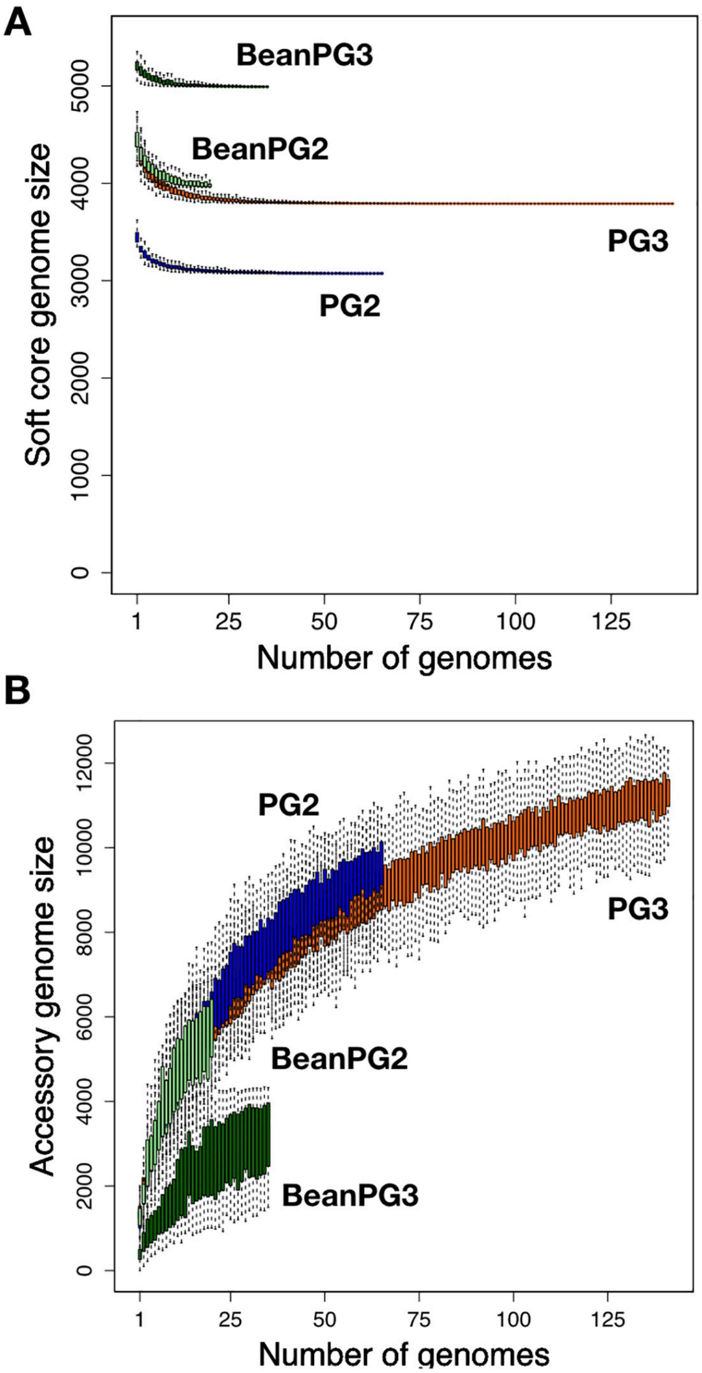

Fig 2. Rarefaction curves for the core and accessory genomes. Families present in 95% (soft core genome) of *P. syringae* isolates exponentially decay as each new genome is added to the analysis. The total number of gene families identified continues to increase indefinitely with the addition of new genomes when singletons (families only present in one isolate) are included.

https://doi.org/10.1371/journal.ppat.1010716.g002

showed a range of values. Overall, there was a significant negative association between bacterial growth and plant weight ($R^2 = 0.63$, $P = 5.0e-6$), supporting the use of seed infection and plant fresh weight to assess bacterial virulence. While the statistical relationship between bacterial growth and plant weight is strong, the moderate correlation emphasises that the former

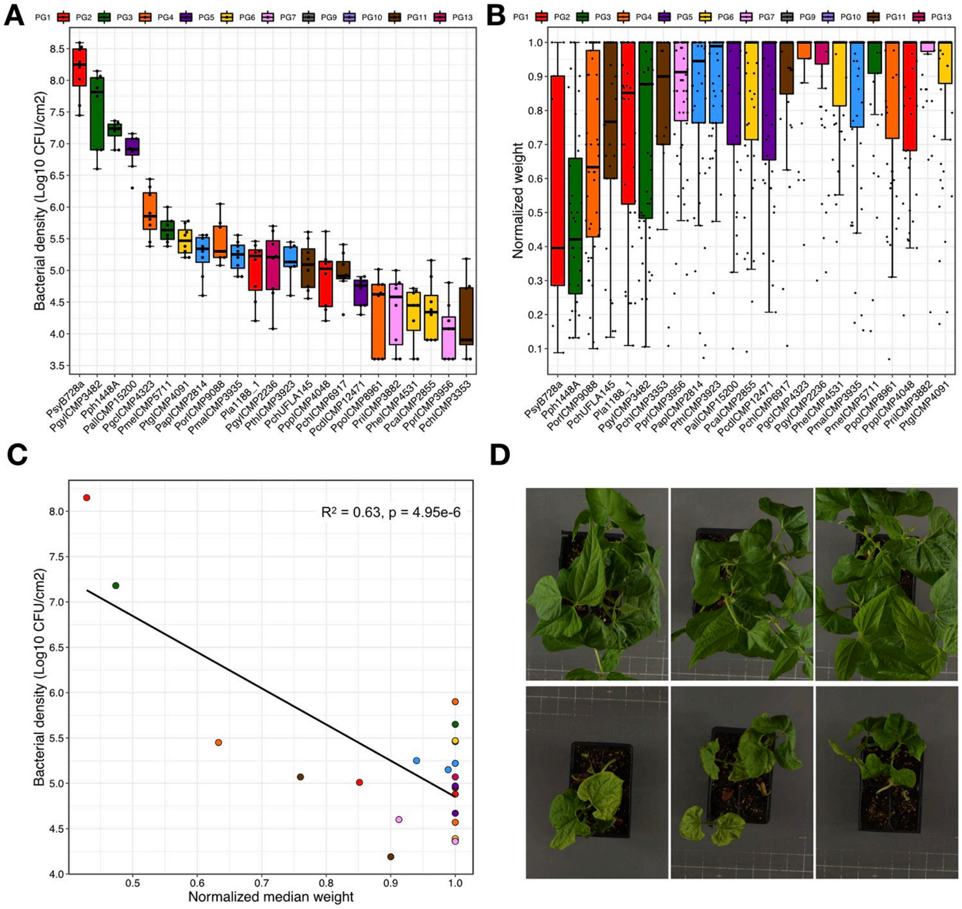

Fig 3. Correlation between bacterial load following pressure infiltration of mature bean leaves and plant weight following seed infection. (A) Bacterial density of 23 *P. syringae* isolates from 10 phylogroups in 14-day old, infected bean leaves (pressure infiltration 3 days post infection). (B) Normalized weight of 14-day old, infected bean plants (seed infected 14 days post infection) of 23 *P. syringae* isolates from 10 phylogroups. (C) Bacterial density in 14-day old, infected bean leaves (pressure infiltration 3 days post infection) as a function of normalized median weight of seed infected plants at 14 days old. There is a strong negative correlation between bacterial density and plant weight across 24 *P. syringae* isolates (linear regression; $F = 36.95$, df = 21, $p = 4.95e\text{-}06$, $R^2 = 0.62$). (D) Characteristic plant phenotypes of 14-day old plants following seedling infection. The photos at the top are for plants infected with $MgSO_4$, while the at the bottom show plants infected with the bean pathogen PsyAB2012-008_22.

https://doi.org/10.1371/journal.ppat.1010716.g003

measures bacterial fitness, while the latter measures host fitness. These two measures are certainly correlated during host-pathogen interactions, but there are many instances where the relationship breaks down, such as when the microbe is commensal or beneficial.

To determine the power of this assay, we performed initial seed infection trials with six *P. syringae* isolates and 50 or more replicate plants. We used a rarefaction analysis of normalized plant weights to determine the number of replicate plants required to distinguish pathogens from non-pathogens with >95% confidence (Tukey-HSD test) and found that the test power plateaued at ~20 replicates per treatment (S3 Fig). Therefore, we performed future seed infection assays using 30 replicate plants per treatment.

Virulence screen

We screened 121 non-clonal representative *P. syringae* isolates from nine PGs to assess the virulence potential as measured by reduced plant fresh weight in 14-day old bean plants after

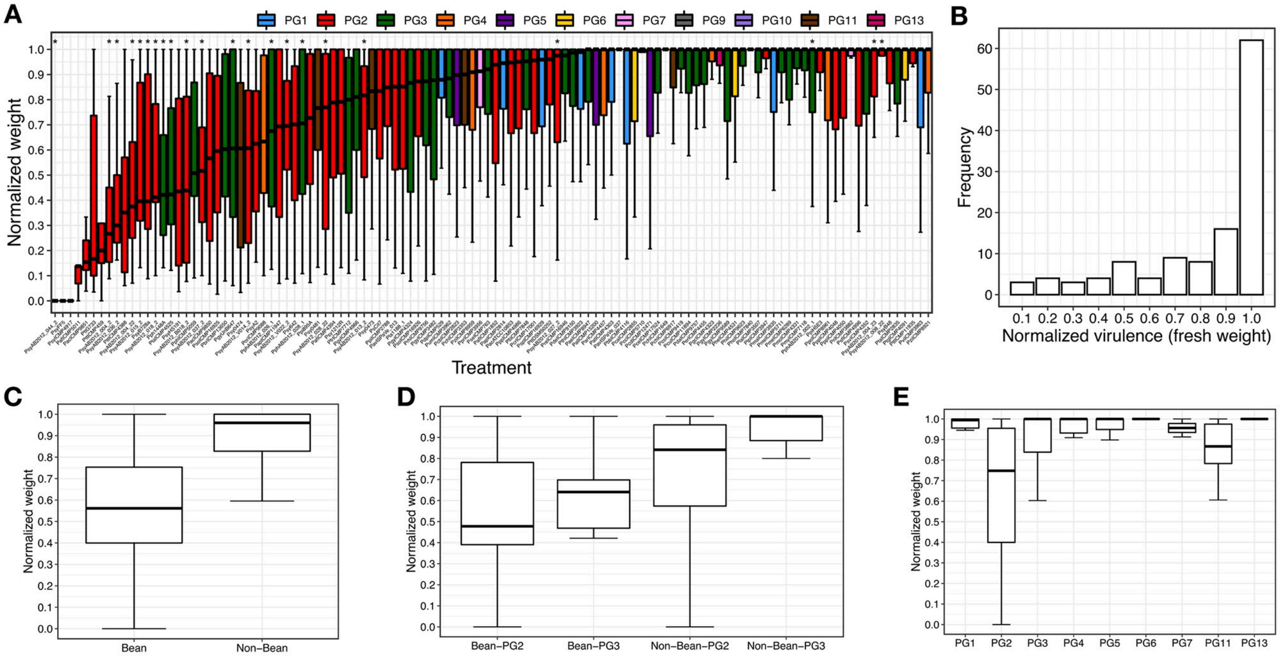

Fig 4. Virulence stratified by host and phylogroup. (A) Boxplots showing the distribution of virulence (i.e., normalized plant weight 14 days after seed infection) ordered by the median virulence. Colors correspond to phylogroups as shown along the top. Bean isolates are indicated with asterisks above their respective boxplots. (B) Frequency plot of virulence for the 121 screened *P. syringae* strains. Distribution of virulence values for (C) bean verses non-bean isolates, (D) bean and non-bean isolates stratified by phylogroup, and (E) virulence values stratified by phylogroup (PG) for the set of 121 screened strains.

https://doi.org/10.1371/journal.ppat.1010716.g004

seed infection (Fig 4). This screened set was subset of the non-clonal set selected to maximize coverage of the species complex while focusing on a manageable number for screening. The screened strains resulted in a highly skewed distribution of normalized fresh weights (i.e., virulence), with a mean of 0.78, median of 0.94, and standard deviation of 0.30. An examination of the 30 strains in the first quartile revealed normalized fresh weights between 0.00 and 0.61, with 56.7% (17 strains) being bean isolates. These 17 strains represent 58.6% of all 29 bean isolates screened.

Significant differences in virulence, as measured by normalized fresh weight, were observed when comparing the strain collection stratified by host of isolation and PG (Table 1). The 29 bean isolates had an average virulence (normalized fresh weight) of 0.59 compared to 0.85 for the 92 non-bean isolates (p = 4.2e-5, 2-tailed, heteroscedastic t-test, same for tests discussed below). As all the bean isolates are found in PG2 and PG3, we compared the virulence of bean isolates to non-bean isolates within these two PGs individually and found no significant difference for PG2 (p = 0.128) but a strong difference for PG3 (p = 8.9e-4). Additionally, there were no significant differences between PG2 and PG3 bean isolates (p = 0.460). We then looked for differences in virulence between strains from different PGs irrespective of their host of isolation (only comparing PGs with at least six tested strains, using 2-tailed, heteroscedastic t-tests, Bonferroni corrected for seven total tests), and found that strains in PG2 were significantly more virulent on bean than strains from PG1, PG3, and PG4 (p = 1.33e-07, 0.012, and 0.029 respectively), but not relative to PG6. In contrast, strains from PG3 were only significantly more virulent on bean than strains from PG1 (p = 0.006). No other significant pairwise PG comparisons were observed. Interestingly, we noticed that PG2 non-bean isolates showed higher virulence on bean than non-bean isolates

Table 1. Virulence and Germination Assay Summary.

	N	Virulence [2]			Germination Frequency (%)		
Group [1]	**Strains**	**Mean**	**Median**	**Stdv**	**Mean**	**Median**	**Stdv**
All Strains	121	0.789	0.913	0.271	60.10	63.33	24.31
PG1	8	0.972	0.995	0.043	72.99	79.17	15.71
PG2	50	0.658	0.748	0.320	46.74	53.33	26.41
PG3	42	0.841	0.964	0.218	68.15	67.50	16.53
PG4	6	0.924	1.000	0.147	83.61	82.50	4.52
PG5	3	0.966	1.000	0.059	73.33	78.33	10.14
PG6	3	1.000	1.000	0.000	80.56	81.67	3.47
PG7	2	0.956	0.956	0.062	85.00	85.00	7.07
PG11	6	0.851	0.867	0.151	43.89	50.00	20.75
PG13	1	1.000	1.000	NA	91.67	91.67	NA
All Bean Isolates	29	0.594	0.516	0.271	45.89	53.33	23.52
All Non-Bean	92	0.850	0.960	0.244	64.58	68.33	23.05
PG2 Bean	16	0.560	0.478	0.293	36.84	32.22	26.42
PG2 Non-Bean	34	0.704	0.841	0.327	51.41	56.67	25.47
PG3 Bean	13	0.635	0.606	0.246	57.02	58.89	13.21
PG3 Non-Bean	29	0.933	1.000	0.122	73.14	73.33	15.56

[1] PG = phylogroup

[2] Normalized fresh weight 2 weeks after seed infection

https://doi.org/10.1371/journal.ppat.1010716.t001

from other PGs (p = 2.87e-4). This indicates that PG2 isolates show greater virulence on bean *irrespective* of host of isolation, although the degree of virulence is relatively low. This pattern was reversed in PG3 where non-bean isolates had significantly lower virulence than non-bean isolates from all other PGs (p = 4.52e-3).

Germination screen

We then assessed whether the virulence of a strain also influenced the germination frequency of bean seeds. Pathogenic microbes are known to interfere with the seed germination both through the direct action of phytotoxins and the indirect action of immune activation [30,40,41,44,45]. In fact, seedling growth inhibition is a well-established assay for immune activation in *Arabidopsis thaliana* [45]. In general, the frequency distribution for bean germination inhibition was less skewed than the frequency distribution for virulence, with mean = 60.1%, median = 63.3%, and standard deviation = 24.3% (Fig 5 and Table 1). The average germination frequency for all bean isolates was 45.9% compared to 64.6% for non-bean isolates (p = 4.92e-4). When stratifying the bean isolates by PG, we found no significant difference in germination frequency between PG2 bean and non-bean isolates (p = 0.076), while the comparison was significant for PG3 (p = 0.002). However, in contrast to the virulence assays we observed a slightly significant difference between germination frequency for PG2 bean isolates and PG3 bean isolates (p = 0.014). Other inter-phylogroup comparisons were similar to what was found for the virulence assays, PG2 strains were significantly different from strains from PG1, PG3, and PG4 (p = 0.010, 6.23e-5, 8.53e-11 respectively 2-tailed, heteroscedastic t-test, Bonferroni corrected for seven tests), while PG3 strain were also significantly different from PG4 (p = 2.23e-4). Also similar to the virulence results, non-bean PG2 isolates resulted in a lower germination frequency than all other non-bean isolates (p = 9.60e-5), while non-bean PG3 strains had a higher germination frequency than non-bean isolates from all other PGs

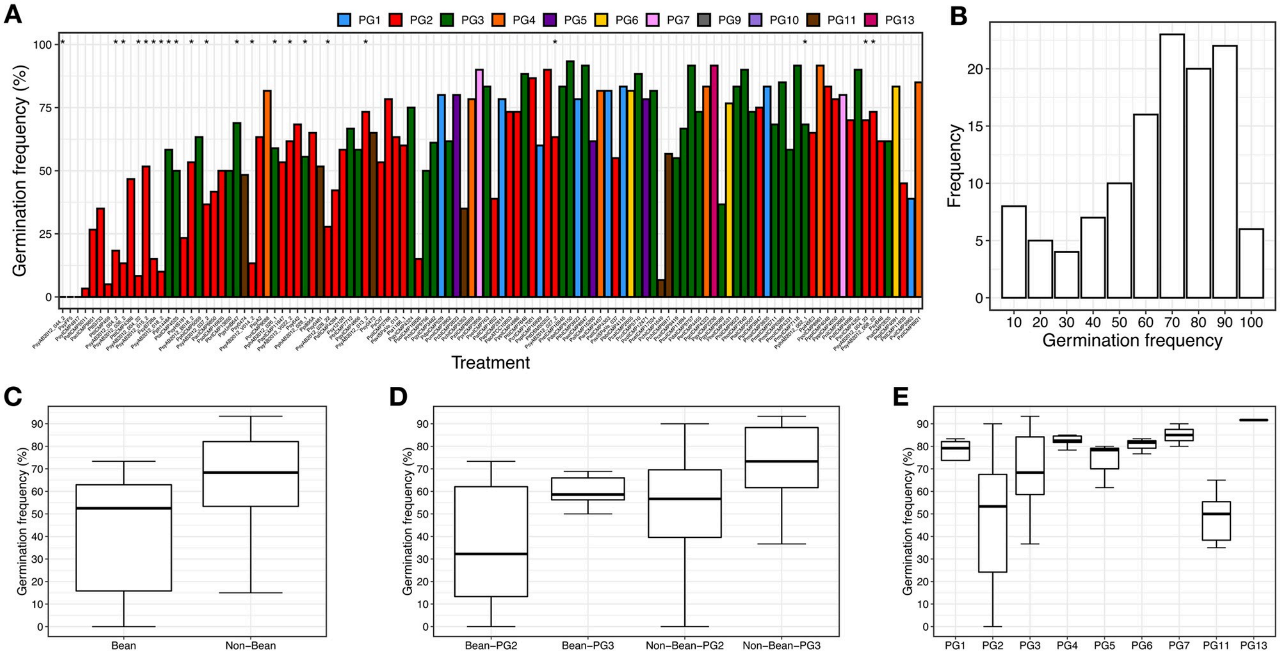

Fig 5. Germination frequencies by host and phylogroup. (A) Boxplots showing the germination frequencies frequency for 121 screened *P. syringae* strains with strains presented in the same manner and order as in Fig 6. (B) Frequency plot of germination for the 121 screened strains. Distribution of germination frequencies for (C) bean verses non-bean isolates, (D) bean and non-bean isolates stratified by phylogroup, and (E) virulence values stratified by phylogroup (PG) for the set of 121 screened strains.

https://doi.org/10.1371/journal.ppat.1010716.g005

(p = 0.004), although this pattern is absent when non-bean PG2 strains were removed from the analysis.

Finally, we measured the association between virulence (i.e., normalized fresh weight) and germination frequency and found a strong association between the two metrics for the full dataset (R^2 = 0.46, p = 2.2e-16). Stratifying by PG and bean isolates showed a strong association for PG2 bean isolates (linear regression; *F* = 28.91, df = 13, p = 0.0001, R^2 = 0.68), but no significant association for PG3 bean isolates (R^2 = 0.62, p = 0.06) (Fig 6).

Predictive modelling of *P. syringae* virulence on bean

We used two machine learning methods and three different genetic feature classes to predict *P. syringae* virulence on beans. The machine learning methods were gradient boosted decision tree regression models and random forest regression models. Here we only report the details of the gradient boosted models since they outperformed the random forest models (S4 Fig). The three genetic feature classes that were used in modeling were: 1) genomic k-mers; 2) T3SE k-mers; or 3) presence / absence of T3SEs and phytotoxins. T3SEs and phytotoxins are well-known virulence factors, with the former often strongly associated with host specificity. Plant weight 14 days after seed infection was used as the continuous outcome variable in our model. We could have also used seed germination frequency in this assay but felt that plant fresh weight more accurately reflected the virulence concerns of bean producers. The goal of analysis was to assess the power of machine learning to predict disease outcomes based on genome sequences and to predict the host specificity of new isolates based on their genome sequence.

We used two nested collections of strains to generate the model. The first collection was comprised of the 121 of the isolates directly screened for virulence, which was made of 29 bean isolates and 92 non-bean isolates, including 50 PG2 isolates (16 bean, 34 non-bean), and 42

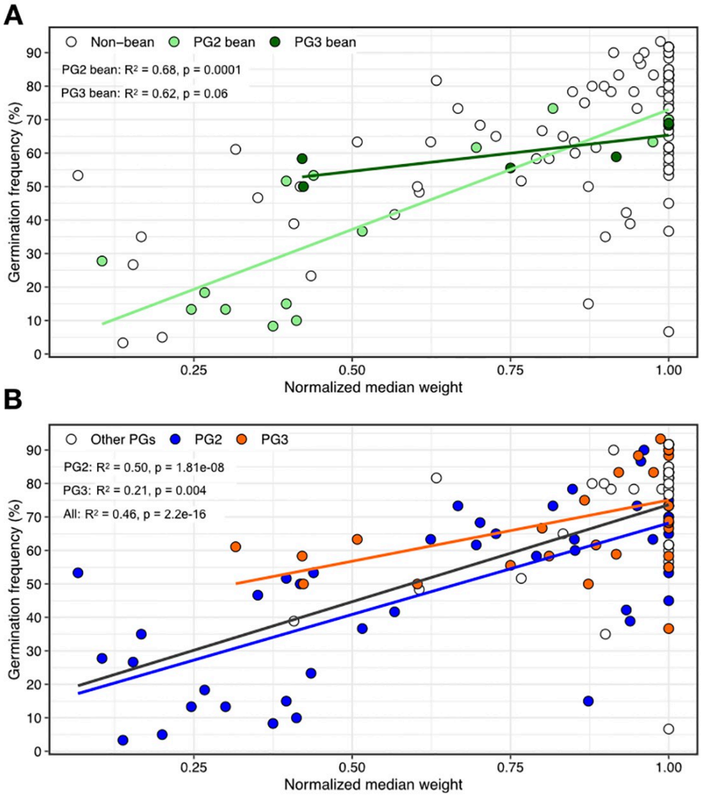

Fig 6. Correlation between germination frequency and virulence of seed infected bean plants. (A) Germination frequency vs virulence (i.e., normalized fresh weight) stratified by PG2 and PG3 bean isolates. A strong correlation is found between germination frequency and virulence of bean strains from PG2 (linear regression; *F* = 28.91, df = 13, *p* = 0.00012, r^2 = 0.66). (B) Germination frequency vs. virulence stratified by PG irrespective of host.

https://doi.org/10.1371/journal.ppat.1010716.g006

PG3 isolates (13 bean, 29 non-bean). This collection is slightly smaller than the full screened set since it does not include the additional PG3 bean isolates added to balance the experimental design (q.v., materials and methods). The second collection was an expanded strain set in which we imputed virulence values based on genomic similarity (S5 Fig). Imputation is the inference of the state of unknown or untested variant genetic loci based on their linkage to known variants. Imputation is very commonly used in many genetic applications (e.g., GWAS, epidemiology) to increase genetic marker density, and therefore, statistical power [46]. In this case we used what might be considered phylogenetic linkage, or simply, recent common ancestry. The imputation process involved identifying strains in our collection belonging to the same clonal lineage as those assayed in our virulence screen (i.e., having a core genome evolutionary distance of less than 0.001 and a T3SE Jaccard similarity of greater than 0.8 to a screened strain). Any strains meeting these criteria were assigned the same virulence as the corresponding screen strain. This imputation process almost tripled the size of our sample set, resulting in an expanded collection of 320 strains (Fig 7), which was made of 59 bean isolates and 261 non-bean isolates, including 66 PG2 isolates (19 bean, 47 non-bean), and 142 PG3 isolates (39 bean, 104 non-bean). We also trained a model on PG2 and PG3 strains separately since bean isolates from these PGs interact with their host very differently.

Our gradient boosted decision tree model showed a mean absolute error (MAE, absolute value of the difference between observed and expected values) between approximately 0.05 and 0.20 and root mean squared error (RMSE, standard deviation of the prediction error)

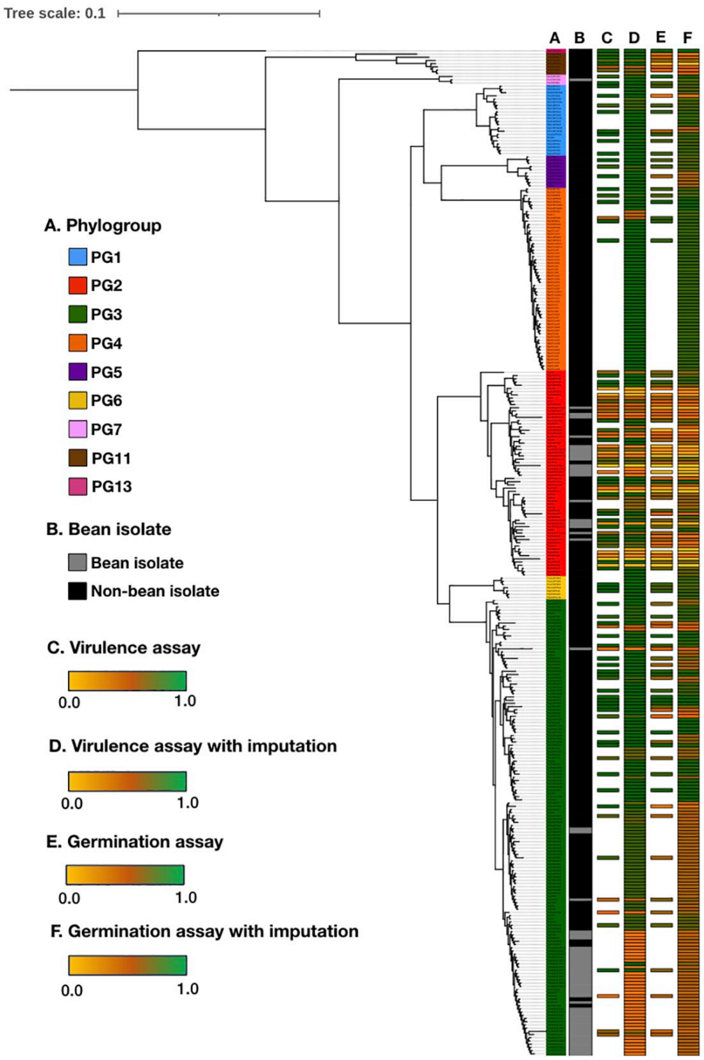

Fig 7. Core genome phylogeny of 320 *P. syringae* isolates. Leaf colors represent PG affiliation as shown in the legend. (A) Bean isolates in gray and non-bean isolates in black. (B) Normalized median bean weight after seed infection by the respective isolates. Green indicates a high weight, while yellow indicates a low weight as show in the legend. (C) Normalized median bean weight for the expanded strain collection based on phylogenetic imputation. (D) Normalized median germination frequency after seed infection by the respective isolates. Green indicates a high frequency, while yellow indicates a low frequency as show in the legend. (E) Normalized median germination frequency for the expanded strain collection based on phylogenetic imputation.

https://doi.org/10.1371/journal.ppat.1010716.g007

between approximately 0.10 and 0.26 (Fig 8 and Table 2). We assessed overall model performance via permutation tests, which were done by building 100 gradient boosted models using full genome k-mers on the extended strain collection in which the host of isolation were randomly assigned (i.e., permuting strain labels). The lowest RMSE out of the 100 permutated models was 0.266±.0021 (sd) compared to the observed RMSE value 0.140 for the same data structure (S6 Fig and Table 2). The fact that the observed RMSE is 64 standard deviations below the mean of the permutated models indicates that the model performs vastly better than random.

Overall, the model performed best on the PG3 strains, which is not surprising given their strongly phylogenetic clustering. The strong clonal separation of bean vs. non-bean pathogens

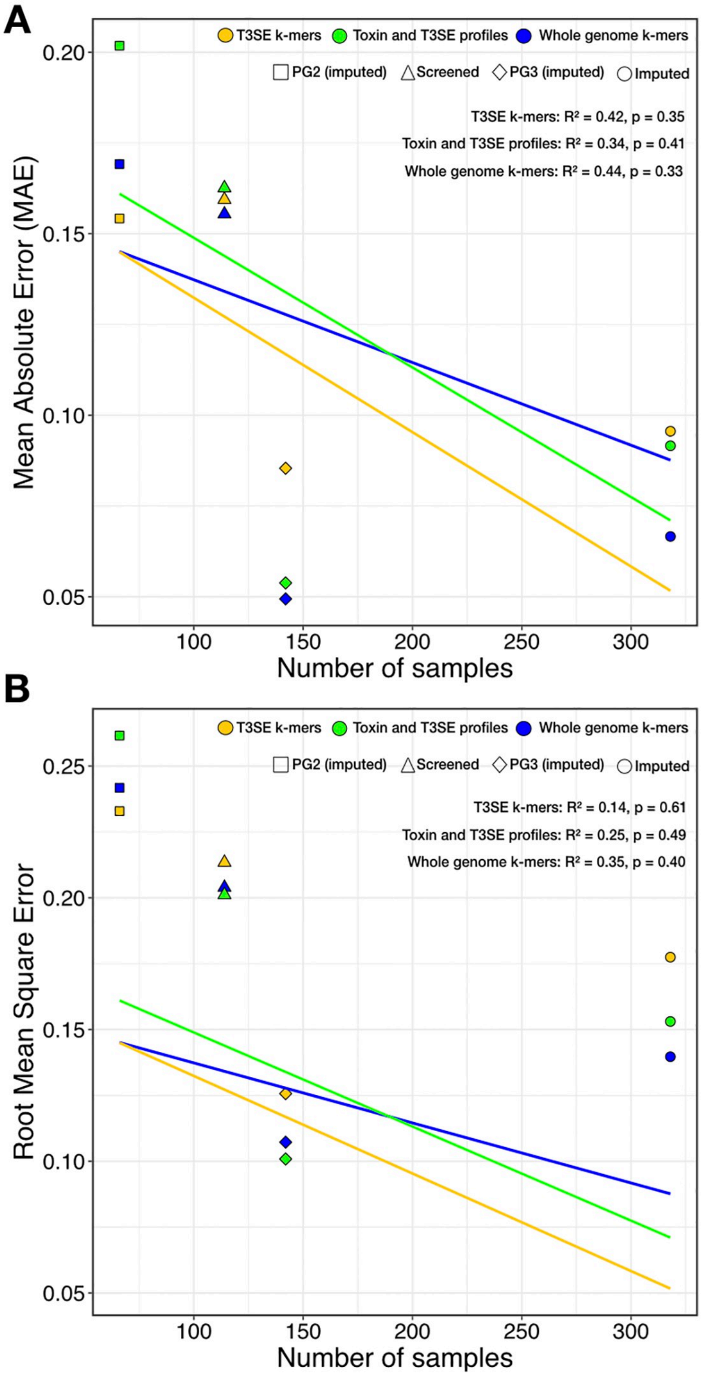

Fig 8. Performance of supervised machine learning models on virulence predictions with genome data. Shapes represent models trained with the screened collection (triangles), imputed expanded collection (circle), imputed expanded PG2 set (square), and imputed expanded PG3 set (rhombus). Colors represent models trained with a phytotoxin and T3SE binary matrix (green), T3SE k-mers (yellow), and whole genome k-mers (blue). (A) Mean Absolute Error (MAE) as a function of number of samples across 50 cross-validation splits. (B) Root Mean Square Error (RMSE) as a function of number of samples across 50 cross-validation splits.

https://doi.org/10.1371/journal.ppat.1010716.g008

Table 2. Machine Learning Model Performance Sorted by Decreasing MAE.

Group [1]	Samples	Dataset	MAE [2]	RMSE [3]
PG3	142	Whole genome k-mers	0.049	0.107
PG3	142	Toxin and T3SE binary matrix	0.054	0.101
Expanded	320	Whole genome k-mers	0.067	0.140
PG3	142	T3SE k-mers	0.085	0.126
Expanded	320	Toxin and T3SE binary matrix	0.092	0.153
Expanded	320	T3SE k-mers	0.096	0.177
PG2	66	T3SE k-mers	0.154	0.233
Screened	121	Whole genome k-mers	0.155	0.204
Screened	121	T3SE k-mers	0.159	0.213
Screened	121	Toxin and T3SE binary matrix	0.163	0.201
PG2	66	Whole genome k-mers	0.169	0.242
PG2	66	Toxin and T3SE binary matrix	0.202	0.262

[1] Group indicates PG strains (screened + imputed virulence used) for ML training. “Expanded” includes PG2 + PG3 strains.

[2] MAE, mean absolute error = mean of the absolute values of the observed value minus the expected value.

[3] RMSE, root-mean-square error = standard deviation of the regression residual (prediction error).

https://doi.org/10.1371/journal.ppat.1010716.t002

within this PG resulted in the statistical reinforcement of features with improved predictive power during model training. The models performed almost as well with the full expanded dataset (i.e., including screened & imputed strains) but surprisingly, there was no significant linear relationship between sample size and the overall performance of the models indicating that sample size what not the primary predictor of model performance (Fig 8, MAE linear regression $F = 1.04$, df = 2, p = 0.41, $R^2 = 0.3$). Perhaps not surprisingly, models built using the greatest number of genetic feature (i.e., whole genome k-mers) frequently showed the highest predictive power, with the best model (PG3 strains using whole genome k-mers) having a MAE of 0.05 (Table 2). An interesting finding was that within the PG2 dataset, the model built with T3SE k-mers outperformed whole genome k-mers, or the profile of toxins and T3SEs, which performed worst. This was unexpected given the small number of T3SEs carried by these strains relative to other strains in the *P*. *syringae* complex. These results support the general consensus that T3SEs play important roles in promoting or restricting host range (despite their relatively small numbers in this PG), while toxins have a more general, non-host specific role in host-microbe interactions.

Model functional validation

Finally, we evaluated the power of our gradient boosted decision tree regression model to predict the virulence of 16 strains that were not previously studied and that are not clonally related to any screened strain (i.e., have core genome evolutionary distance >0.001 and T3SE Jaccard similarity of a <0.8 compared to the screened strains), meaning that these validation strains can be viewed as completely naïve strain collection. We used whole genome k-mers to make virulence predictions and evaluated these against actual virulence measures obtained through the seed infection virulence assay. When comparing observed virulence to predicted virulence of the 16 strains in the functional validation set, we found a RMSE of 0.164, which compares favorably with the RMSE of 0.140 obtained for the whole genome k-mer model of the extended strain collection (Fig 9 and Table 3). Interestingly, the one strain that performed the most poorly was the PG2 strain PttDSM50252, which had an MAE of 0.437 while the other 15 strains had an average MAE of 0.112. If PttDSM50252 is removed from the calculation, the

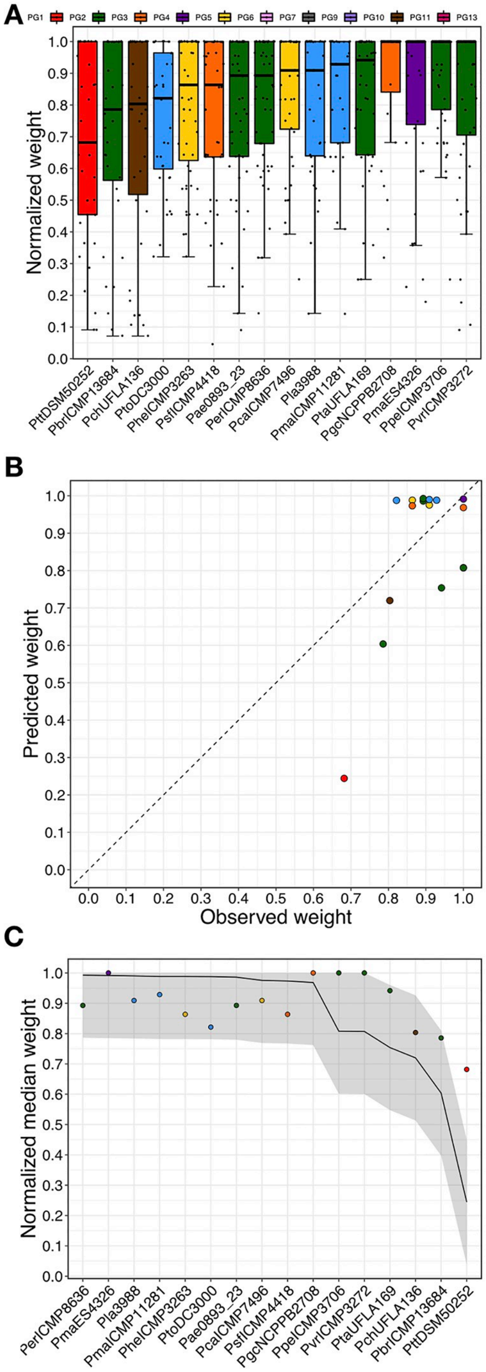

Fig 9. Correlation between predicted and observed weights of plants infected with isolates previously unseen by the model. (A) Normalized weight distribution of 16 isolates previously unseen by the model (not clonally related to any screened strain, i.e., having a core genome evolutionary distance >0.001 and a T3SE Jaccard similarity of <0.8 compared to the screened strains). (B) Predicted virulence as a function of observed virulence on seed infected bean plants. Color coding indicates PGs. Dashed line indicates 1:1 relationship. (C) Virulence for each isolate tested ordered by predicted virulence and colored by PG. The solid line represents virulence predictions for each isolate. The grey box represents the error margins for the predictions based on the MAE and RMSE values for the model.

https://doi.org/10.1371/journal.ppat.1010716.g009

Table 3. Model Functional Validation.

Treatment	PG	Host	Observed Weight[1]	Predicted Weight[1]	MAE
PttDSM50252	2	Wheat	0.682	0.244	0.437
PvrICMP3272	3	Kiwifruit	1.000	0.807	0.193
PpeICMP3706	3	Myrobalan Plum	1.000	0.808	0.192
PtaUFLA129	3	Coffee	0.942	0.754	0.188
PbrICMP13684	3	Paper Mulberry	0.786	0.604	0.182
PtoDC3000	1	Tomato	0.821	0.988	0.166
PheICMP3263	6	Sunflower	0.864	0.988	0.125
PsfICMP4418	4	Oat	0.864	0.973	0.110
PerICMP8636	3	Loquat	0.893	0.993	0.100
Pae0893_23	3	Horse Chestnut	0.893	0.986	0.093
PchUFLA136	11	Coffee	0.804	0.720	0.084
Pla3988	1	Cucumber	0.909	0.990	0.081
PcaICMP7496	6	Pawpaw	0.909	0.975	0.066
PmaICMP11281	1	Broccoli Raab	0.929	0.988	0.060
PgcNCPPB2708	4	Coffee	1.000	0.969	0.031
PmaES4326	5	Radish	1.000	0.991	0.009

[1] Z-score normalized fresh weight

https://doi.org/10.1371/journal.ppat.1010716.t003

RMSE value drops to 0.126. This finding further supports the hypothesis that PG2 show low degrees of host specificity.

Discussion

In this work we addressed whether host of isolation is a reliable predictor of host specific virulence and whether whole genome sequences can be used to predict the host specific virulence potential of individual strains. While host of isolation is a widely used surrogate for host specificity, this assumption has rarely been empirically tested [13,14], and the strength of this assumption is critical when viewed from the context of the virulence potential of emerging pathogens. Are strains isolated from one host only virulent on that host, or do they have the potential to move to other species? Are strains isolated from environmental sources, such as streams or soil, limited to those environments or can they ‘jump’ to a new host and potentially cause a significant outbreak?

Our first aim was to determine if infection of bean seeds by *P. syringae* recapitulated virulence responses seen in standard syringe inoculation virulence assays. We found a negative association between our virulence measure of normalized plant fresh weight after seed infection and *in planta* bacterial growth after syringe infiltration into leaf tissue, showing that the seed infection protocol effectively recapitulates standard methods. This finding is consistent with published and anecdotal reports that infected seed stocks are a significance source of bean disease [16,40,41,47]. In general, we found that bean isolates reduced mean plant fresh weight by 30.2% and median weight by 46.2% compared to non-bean isolates. While the PG2 bean isolates (leaf spot disease caused by pathovar *syringae*) had a normalized mean fresh weight of 0.56±0.293 (SD) compared to 0.64±0.246 for the PG3 bean isolates (halo blight disease caused by pathovar *phaseolicola*), this difference was not significantly different. A similar pattern was found when we examined seed germination frequencies, where bean isolates reduced the average germination frequency by 28.9% and median frequency by 22.0% compared to non-bean isolates, while the 38.8% mean germination frequency of PG2 bean pathogens was significantly lower than the 57.0% mean germination frequency of the PG3 bean pathogens (p = 0.014).

We find a striking difference when comparing bean to non-bean isolates found in the same phylogroup. As anticipated, bean seed infection with PG3 bean isolates resulted in significantly higher virulence and lower germination frequency than PG3 non-bean isolates, while in contrast, PG2 bean isolates did not differ significantly from PG2 non-bean isolates. PG2 strains generally (irrespective of host of isolation) show greater virulence on bean, indicating that strains from this phylogroup have lower host specificity, i.e., are host generalists. This is strongly supported when comparing non-bean isolates from PG2 to non-bean PG3 isolates (normalized fresh weight of 0.704 and 0.933, respectively; p = 4.67E-04). These findings are consistent with other studies that have found lower levels of host specificity among PG2 strains [13,14] and lends support to the hypothesis that PG2 strains may rely as much or more on toxins than T3SEs when compared to other *P. syringae* strains.

We expected that PG3 bean isolates would have higher virulence than PG2 bean isolates since halo blight caused by PG3 pathovar *phaseolicola* is a much more severe disease than spot disease caused by PG2 pathovar *syringae*, but this was not the case. There are several explanations for these data. First, is a simple experimental bias explanation driven by the fact that nearly all the PG3 bean isolates fall into one clonal group as defined by our clonality criteria of a core genome distance of <0.001 average substitutions per site and T3SE profile Jaccard similarity value >0.8. We attempted to address this issue by oversampling from the *phaseolicola* clonal group. But to ensure that we did not create another bias by adding too many very closely related strains, we only added seven additional strains to the original group of six PG3 bean isolates. Unfortunately, this still resulted in a small set that could easily be skewed by a few outlying measurements. Second, some of the PG3 strains likely elicit effector-triggered immunity in the cultivar of bean assayed, which would result in healthy plants. Given the small set of PG3 bean isolates, even a few ETI-eliciting strains will result in a large average decrease in virulence. And third, it is possible that the most severe symptoms of halo blight are only seen after leaf-to-leaf transmission caused by water splash rather than seed transmission [21,48].

While many genome-wide association studies have successfully identified strong genotype-to-phenotype linkages, we were unable to identify any loci significantly associated with bean isolation. Consequently, we shifted our focus to machine learning approaches as they can not only unravel genomic signatures associated with continuous phenotypes, but also predict the virulence potential of previously unseen isolates given their genome sequences. Regardless of PG affiliation, our model was able to predict the virulence of individual *P. syringae* isolates within reasonable error margins based solely on whole genome data. The fact that models trained on virulence factors alone could predict virulence with considerable accuracy supports the notion that T3SEs and phytotoxins play crucial roles in host adaptation processes. Nonetheless, the higher predictive power of models trained with whole genome k-mers suggests that factors other than canonically virulence-associated genes also play important roles on disease development and adaptation to beans.

While sample size is usually an important contributor to accurate model generation in machine learning, we only found an association between sample size and predictive power when we did not stratify by data types (i.e., whole genome k-mers, T3SE k-mers, and toxins and effector presence/absence). In this case, the screened strain collection providing a MAE of 0.153 while the larger imputed strain collection increased model performance to a MAE of 0.065. No association between sample size and predictive power was found when data types were not stratified, which may indicate that factors such as the phylogenetic structure of the sample outweigh the size of the sample. We also find poorer model performance for PG2 strains than PG3 strains. While this may partly reflect the differences in sample size between these two groups, it also likely reflects the underlying biology. The majority of bean isolates in PG3 are phylogenetically clustered, while there is little clustering of bean isolates in PG2. Consequently, the model may perform better on PG3 since it is essentially predicting phylogenetic structure. Another contributing factor is likely the finding discussed above, namely, that host specificity appears to be weaker in PG2. If PG2 strains are more generalists than specialists, then the host specificity signal would be weaker and any model trying to find this signal would perform more poorly.

In conclusion, we believe that this work demonstrates the potential utility of machine learning for predicting host-specific virulence. Future models would benefit from increased sample sizes, improved phenotyping capacity and accuracy, reliable metadata, and improved methods for controlling for population structure (i.e., non-independent evolutionary history). Given the relative ease of generating genomic data, it is likely that these models will play an increasingly important role in diagnostic microbiology, and hopefully provide a new and valuable tool for protecting crops from emerging pathogens in the future.

Materials and methods

Strain collection

Three hundred and thirty-three *Pseudomonas syringae* strains were used in this study (S1 Table). Forty-six *P. syringae* isolates were collected from bean fields approximately 80 km east of Lethbridge, Alberta, Canada, during the summer of 2012. Bean leaves with symptoms of bacterial diseases were collected from new growth during the vegetative growth stage. The remaining 288 isolates were previously published [9] and include 49 *P. syringae* type and pathotype strains [1,49]. A type strain is the isolate to which the scientific name of that organism is formally attached under the rules of prokaryote nomenclature, while a pathotype strain is similar but with the additional requirement that it has the pathogenic characteristics of its pathovar (i.e., a pathogen of a particular host) [5]. Out of the 333, 317 strains were used for comparative analyses and model training, while 16 were used for model functional validation. A subset of 267 non-clonal representative strains (discussed below) were selected for the predictive modeling to avoid clonal bias. A further subset of 121 isolates, including the type and pathotype strains, were selected for virulence assays.

Sequencing and quality control

DNA was extracted using the Gentra Puregene Yeast and Bacteria kit (Qiagen, Hilden, Germany). Illumina libraries with 300–400 bp inserts were generated using the Illumina Nextera XT kit according to the manufacturer's protocol (Illumina, CA, USA). Samples were multiplexed with the Illumina Nextera XT Index kit containing 96 indices. Samples were sequenced on the Illumina NextSeq 500 Mid Output v2 (300 cycle) kit with 150 base PE reads. All sequencing was performed at the University of Toronto's Centre for the Analysis of Genome Evolution and Function (CAGEF). Raw read quality was assessed with FastQC. Trimmomatic was used to remove adapters and trim raw sequencing reads based on a sliding window approach (window size = 4, required quality = 5).

De novo assembly

Paired-end reads were *de novo* assembled using the CLC Genomics Assembly Tool (CLC Genomics Workbench). Contigs shorter than 1kb were removed from the assemblies.
Low coverage contigs with matches to non-*Pseudomonas* genera and no matches to the

Pseudomonas genus that had a depth of coverage less than one standard deviation from the average assembly coverage were deemed contaminants and, therefore, removed from the final draft.

Pangenome analysis, gene prediction, annotation, and orthologous clustering

Gene prediction and annotation for all assemblies were performed with Prokka [50]. Prokka annotates inferred coding sequences by searching for sequence similarity in the UniProtKB [51] database and HMM libraries [52]. Additionally, all predicted genes were aligned against a custom T3SE database for the identification of potential T3Ses [26]. Pangenome analysis was performed via PIRATE [53], which iteratively clusters genes into orthologous groups by performing all-vs-all comparisons followed by MCL clustering given a certain percent identify threshold. Genes present in at least 95% of the genomes were classified as core. Core protein families were individually aligned with MUSCLE [54] and later concatenated into a single protein alignment. We used the FastTree2 approximate maximum-likelihood approach [55] to infer the phylogenetic relationships of all 320 isolates. Core genome synonymous substitution rates were estimated with MEGA7 [56] using the Nei-Gojobori method and Jukes-Cantor model. Jaccard distances were computed with R version 4.0.5 (42) using a binary matrix of presence and absence of accessory genes. Rarefaction curves were generated using a custom Python script.

Identification of non-clonal representative strains

We reduced the impact of phylogenetic bias in our predictive modeling by selecting only one representative strain from each clonal group (i.e., very closely related strains recently derived from a common ancestor) identified from the *P. syringae* core-genome phylogeny. We identified clonal groups by calculating the pairwise core genome evolutionary distance and the Jaccard similarity for T3SE profiles. We found the minimum pairwise core-genome evolutionary distance for isolates with identical T3SE profiles to be 0.001 average substitutions per site. We therefore pooled the 318 isolates if they had a core genome evolutionary distance of less than 0.001, resulting in 209 clusters. We further supplemented these clusters by adding back any strain that had a T3SE profile Jaccard similarity value less than 0.8, resulting in 267 non-clonal clusters. A single representative was selected out of each of these non-clonal clusters for downstream analyses. One exception was made to the strain selection process to balance our experimental design, which was skewed due to the fact that the vast majority of PG3 bean isolates (i.e., pathovar *phaseolicola*) fall into one clonal group. Initially, our selection criteria resulted in only six PG3 bean isolates compared to 16 PG2 bean isolates (i.e., pathovar *syringae*). We therefore added an additional seven *phaseolicola* strains to the screened set to better balance the number of bean isolates in PG2 and PG3. Evolutionary distances and Jaccard similarity scores were inferred with MEGA7 [56] and R version 4.0.5 [57].

Seed infection virulence assay

P. syringae strains were grown overnight at 30˚C in King's B media, re-suspended in 10 mM $MgSO_4$ and diluted to an OD_{600} of 0.001. *P. vulgaris* var. Canadian Red seeds were soaked for 24 hours in the bacterial suspension, planted in Sunshine Mix 1 soil with regular watering and grown for 14 days. Plant fresh weight and germination frequencies were measured and normalized to a control plant treated with 10 mM $MgSO_4$ sown on each flat. Trials were repeated three times.

Syringe infiltration virulence assays

P. syringae strains were grown overnight on appropriate antibiotics, re-suspended in 10 mM $MgSO_4$ and diluted to OD_{600} of 0.001. Two- to three-week-old *Phaseolus vulgaris* var. Canadian Red plants were syringe infiltrated and bacterial growth assays were carried out by harvesting eight leaf disks (1 cm^2) from each plant (two per each primary leaf) three days after infiltration. Disks were homogenized using a bead-beater in 200 μl sterile 10 mM $MgSO_4$, serially diluted in 96-well plates, and 5 μl from each dilution was spot plated on KB supplemented with rifampicin for positive and negative control strains. Plates were incubated for at least 24 hours at 30˚C and the resulting colony counts were used to calculate the number of CFUs per cm^2 in the leaf apoplast.

Predictive modeling of *P. syringae* virulence on bean

We used an implementation of gradient boosted decision trees to model the effect of *P. syringae* isolates on plant weights as a proxy for strain virulence using: 1) whole genome k-mers, 2) T3SE k-mers, and 3) a presence / absence matrix of T3Ses and phytotoxins. We split sequences into 31-mers with fsm-lite and generated a binary matrix for k-mers with identical distribution patterns using custom python scripts. Next, we used the Scikit-learn and the XGBoost python libraries [58] to generate a regression model for the prediction of normalized plant weights using all three datasets as input features. Given the relatively small size of our dataset, we used a cross-validation (CV) procedure to assess the performance of our model on 50 independent splits (S7 Fig). For each time, we randomly split the data into training (80%) and testing (20%) sets while maintaining the same plant weight distributions on both sets. Hyper parameters were fine-tuned using Scikit-learn's RandomizedSearchCV module and regression models were generated with XGBoost's XGBRFRegressor module.

Supporting information

S1 Fig. Characteristic plant phenotypes of 14-day old plants following seedling infection. Halo blight leaf symptoms on a plant infected with Phh1448A on the left, compared to healthy plants treated with $MgSO_4$ on the right.
(TIF)

S2 Fig. Normalized weights and bacterial densities for Pph1448A::HrcC. The mutant Pph1448A::HrcC is unable to deliver T3SEs into the host cell. Plants treated with this mutant therefore exhibit higher normalized fresh weights and lower bacterial densities in comparison to a wild-type treatment.
(TIF)

S3 Fig. Sample size selection via a simulated experimental setup. Bean plants were seed infected with 6 *P. syringae* isolates with a high number of replicates (>50). Plant weights were randomly selected according to various replicate sizes (8–32). Our ability to distinguish pathogens from non-pathogens using Tukey-HSD tests plateaus at 20 replicates per treatment.
(TIF)

S4 Fig. Performance statistics for random forest and gradient boosting machines. (A) Mean Absolute Error (MAE) as a function of number of samples across 50 cross-validation splits. (B) Root Mean Square Error (RMSE) as a function of number of samples across 50 cross-validation splits.
(TIF)

S5 Fig. Distribution of virulence values stratified by host and phylogroup. Boxplots showing the distribution of virulence (i.e., normalized plant weight 14 days after seed infection) values for (A) bean verses non-bean isolates for the set of 121 screened strains, and (B) for the 320 strains in the expanded dataset that includes both screened and imputed strains. (C) Distribution of virulence values for bean and non-bean isolates stratified by phylogroup for the screened strains and (D) for the expanded dataset. (E) Distribution of virulence values stratified by phylogroup (PG) for the screened strains, and the (F) expanded strain set.
(TIF)

S6 Fig. Distribution of RMSE statistics from gradient boosting models generated from 100 permutated dataset. Models were trained with whole-genome k-mers on the expanded strain collection by randomly swapping host of isolation labels. The average of the permuted distribution is 0.271±0.002 (sd). The observed RMSE using the equivalent model design was 0.140.
(TIF)

S7 Fig. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across 50 cross-validation splits. Model performance in terms of (A) MAE and (B) RMSE. Plant weight distributions were kept the same across splits.
(TIF)

S1 Table. List of *P. syringae* isolates used in this study.
(XLSX)

Author Contributions

Conceptualization: Renan N. D. Almeida, Michael Greenberg, Alexandre Martel, Maggie A. Middleton, Darrell Desveaux, David S. Guttman.

Data curation: Renan N. D. Almeida, Michael Greenberg, Pauline W. Wang, David S. Guttman.

Formal analysis: Renan N. D. Almeida, Cedoljub Bundalovic-Torma, David S. Guttman.

Funding acquisition: David S. Guttman.

Investigation: Renan N. D. Almeida, Michael Greenberg, Cedoljub Bundalovic-Torma, Pauline W. Wang.

Methodology: Renan N. D. Almeida, Michael Greenberg, Pauline W. Wang, Maggie A. Middleton, David S. Guttman.

Project administration: Pauline W. Wang, David S. Guttman.

Resources: Syama Chatterton, David S. Guttman.

Supervision: Pauline W. Wang, Darrell Desveaux, David S. Guttman.

Validation: Renan N. D. Almeida, David S. Guttman.

Visualization: Renan N. D. Almeida, David S. Guttman.

Writing – original draft: Renan N. D. Almeida, David S. Guttman.

Writing – review & editing: Renan N. D. Almeida, Michael Greenberg, Cedoljub Bundalovic-Torma, Pauline W. Wang, Syama Chatterton, Darrell Desveaux, David S. Guttman.

References

1. Baltrus DA, McCann HC, Guttman DS. Evolution, genomics and epidemiology of *Pseudomonas syringae*: Challenges in Bacterial Molecular Plant Pathology. Mol Plant Pathol. 2017; 18(1):152–68. https://doi.org/10.1111/mpp.12506 PMID: 27798954
2. Morris CE, Monteil CL, Berge O. The life history of *Pseudomonas syringae*: linking agriculture to earth system processes. Annu Rev Phytopathol. 2013; 51:85–104. https://doi.org/10.1146/annurev-phyto-082712-102402 PMID: 23663005
3. Xin XF, Kvitko B, He SY. *Pseudomonas syringae*: what it takes to be a pathogen. Nat Rev Microbiol. 2018; 16(5):316–28. https://doi.org/10.1038/nrmicro.2018.17 PMID: 29479077
4. Morris CE, Sands DC, Vinatzer BA, Glaux C, Guilbaud C, Buffiere A, et al. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. ISME J. 2008; 2(3):321–34. https://doi.org/10.1038/ismej.2007.113 PMID: 18185595
5. Bull CT, De Boer SH, Denny TP, Firrao G, Fischer-Le Saux M, Saddler GS, et al. Demystifying the nomenclature of bacterial plant pathogens. J Plant Pathol. 2008; 90(3):403–17.
6. Dye DW, Bradbury JF, Goto M, Hayward AC, Lelliott RA, Schroth MN. International standards for naming pathovars of phytopathogenic bacteria and a list of pathovar names and pathotype strains. Review of Plant Pathology. 1980; 59(4):153–68.
7. Berge O, Monteil CL, Bartoli C, Chandeysson C, Guilbaud C, Sands DC, et al. A user's guide to a data base of the diversity of *Pseudomonas syringae* and its application to classifying strains in this phylogenetic complex. PLoS One. 2014; 9(9):e105547. https://doi.org/10.1371/journal.pone.0105547 PMID: 25184292
8. Sarkar SF, Guttman DS. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. Appl Environ Microbiol. 2004; 70(4):1999–2012. https://doi.org/10.1128/AEM.70.4.1999-2012.2004 PMID: 15066790
9. Dillon MM, Thakur S, Almeida RND, Wang PW, Weir BS, Guttman DS. Recombination of ecologically and evolutionarily significant loci maintains genetic cohesion in the *Pseudomonas syringae* species complex. Genome Biol. 2019; 20(1):3. https://doi.org/10.1186/s13059-018-1606-y PMID: 30606234
10. Hwang MS, Morgan RL, Sarkar SF, Wang PW, Guttman DS. Phylogenetic characterization of virulence and resistance phenotypes of *Pseudomonas syringae*. Appl Environ Microbiol. 2005; 71(9):5182–91. https://doi.org/10.1128/AEM.71.9.5182-5191.2005 PMID: 16151103
11. Hirano SS, Upper CD. Population biology and epidemiology of *Pseudomonas syringae*. Annu Rev Phytopathol. 1990; 28:155–77.
12. Young JM. Taxonomy of *Pseudomonas syringae*. J Plant Pathol. 2010; 92(1):S5–S14.
13. Morris CE, Lamichhane JR, Nikolić I, Stanković S, Moury B. The overlapping continuum of host range among strains in the *Pseudomonas syringae* complex. Phytopathology Res. 2019; 1(1):4.
14. Morris CE, Moury B. Revisiting the Concept of Host Range of Plant Pathogens. Annu Rev Phytopathol. 2019; 57:63–90. https://doi.org/10.1146/annurev-phyto-082718-100034 PMID: 31082307
15. Hirano SS, Upper CD. Bacteria in the leaf ecosystem with emphasis on *Pseudomonas syringae*-a pathogen, ice nucleus, and epiphyte. Microbiol Mol Biol Rev. 2000; 64(3):624–53. https://doi.org/10.1128/MMBR.64.3.624-653.2000 PMID: 10974129
16. Upper CD, Hirano SS, Dodd KK, Clayton MK. Factors that affect spread of *Pseudomonas syringae* in the phyllosphere. Phytopathol. 2003; 93(9):1082–92. https://doi.org/10.1094/PHYTO.2003.93.9.1082 PMID: 18944091
17. Hulin MT, Armitage AD, Vicente JG, Holub EB, Baxter L, Bates HJ, et al. Comparative genomics of *Pseudomonas syringae* reveals convergent gene gain and loss associated with specialization onto cherry (*Prunus avium*). New Phytol. 2018; 219(2):672–96. https://doi.org/10.1111/nph.15182 PMID: 29726587
18. Hulin MT, Mansfield JW, Brain P, Xu X, Jackson RW, Harrison RJ. Characterization of the pathogenicity of strains of *Pseudomonas syringae* towards cherry and plum. Plant Pathol. 2018; 67(5):1177–93. https://doi.org/10.1111/ppa.12834 PMID: 29937581
19. O'Brien HE, Thakur S, Gong Y, Fung P, Zhang J, Yuan L, et al. Extensive remodeling of the *Pseudomonas syringae* pv. *avellanae* type III secretome associated with two independent host shifts onto hazelnut. BMC Microbiol. 2012; 12:141. https://doi.org/10.1186/1471-2180-12-141 PMID: 22800299
20. Lipps SM, Samac DA. *Pseudomonas viridiflava*: An internal outsider of the *Pseudomonas syringae* species complex. Mol Plant Pathol. 2021; https://doi.org/10.1111/mpp.13133 PMID: 34463014
21. Arnold DL, Lovell HC, Jackson RW, Mansfield JW. *Pseudomonas syringae* pv. *phaseolicola*: from 'has bean' to supermodel. Mol Plant Pathol. 2011; 12(7):617–27. https://doi.org/10.1111/j.1364-3703.2010.00697.x PMID: 21726364

22. 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. 2000; 19(13):3204–14. https://doi.org/10.1093/emboj/19.13.3204 PMID: 10880434

23. Joardar V, Lindeberg M, Jackson RW, Selengut J, Dodson R, Brinkac LM, et al. Whole-genome sequence analysis of *Pseudomonas syringae* pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol. 2005; 187(18):6488–98. https://doi.org/10.1128/JB.187.18.6488-6498.2005 PMID: 16159782

24. 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(31):11064–9. https://doi.org/10.1073/pnas.0504930102 PMID: 16043691

25. Marco ML, Legac J, Lindow SE. *Pseudomonas syringae* genes induced during colonization of leaf surfaces. Environ Microbiol. 2005; 7(9):1379–91. https://doi.org/10.1111/j.1462-2920.2005.00825.x PMID: 16104861

26. Dillon MM, Almeida RND, Laflamme B, Martel A, Weir BS, Desveaux D, et al. Molecular evolution of *Pseudomonas syringae* type III secreted effector proteins. Front Plant Sci. 2019; 10:418. https://doi.org/10.3389/fpls.2019.00418 PMID: 31024592

27. Buttner D. Behind the lines-actions of bacterial type III effector proteins in plant cells. FEMS Microbiol Rev. 2016; 40(6):894–937. https://doi.org/10.1093/femsre/fuw026 PMID: 28201715

28. Khan M, Seto D, Subramaniam R, Desveaux D. Oh, the places they'll go! A survey of phytopathogen effectors and their host targets. Plant J. 2018; 93(4):651–63. https://doi.org/10.1111/tpj.13780 PMID: 29160935

29. Martel A, Ruiz-Bedoya T, Breit-McNally C, Laflamme B, Desveaux D, Guttman DS. The ETS-ETI cycle: evolutionary processes and metapopulation dynamics driving the diversification of pathogen effectors and host immune factors. Curr Opin Plant Biol. 2021; 62:102011. https://doi.org/10.1016/j.pbi.2021.102011 PMID: 33677388

30. Bender CL, Alarcon-Chaidez F, Gross DC. *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. Microbiol Mol Biol Rev. 1999; 63(2):266–92. https://doi.org/10.1128/MMBR.63.2.266-292.1999 PMID: 10357851

31. Allen JP, Snitkin E, Pincus NB, Hauser AR. Forest and Trees: Exploring Bacterial Virulence with Genome-wide Association Studies and Machine Learning. Trends Microbiol. 2021; https://doi.org/10.1016/j.tim.2020.12.002 PMID: 33455849

32. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015; 25:17–24. https://doi.org/10.1016/j.mib.2015.03.002 PMID: 25835153

33. Falush D. Bacterial genomics: Microbial GWAS coming of age. Nat Microbiol. 2016; 1:16059. https://doi.org/10.1038/nmicrobiol.2016.59 PMID: 27572652

34. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006; 14(8):353–5. https://doi.org/10.1016/j.tim.2006.06.003 PMID: 16782339

35. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017; 18(1):41–50. https://doi.org/10.1038/nrg.2016.132 PMID: 27840430

36. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. Proc Natl Acad Sci U S A. 2013; 110(29):11923–7. https://doi.org/10.1073/pnas.1305559110 PMID: 23818615

37. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016; 1:16041. https://doi.org/10.1038/nmicrobiol.2016.41 PMID: 27572646

38. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the end-game for GWAS: Machine learning approaches for the prioritization of complex disease loci. Front Genet. 2020; 11:350. https://doi.org/10.3389/fgene.2020.00350 PMID: 32351543

39. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, et al. Current affairs of microbial genome-wide association studies: approaches, bottlenecks and analytical pitfalls. Front Microbiol. 2019; 10:3119. https://doi.org/10.3389/fmicb.2019.03119 PMID: 32082269

40. Darrasse A, Bureau C, Samson R, Morris CE, Jacques M-A. Contamination of bean seeds by *Xanthomonas axonopodis* pv. *phaseoli* associated with low bacterial densities in the phyllosphere under field and greenhouse conditions. Eur J Plant Pathol. 2007; 119(2):203–15.

41. Eyster HC. The cause of decreased germination of bean seeds soaked in water. Am J Bot. 1940; 27(8):652–9.

42. Hirano SS, Demars SJ, Morris CE. Survival, establishment, and dispersal of *Pseudomonas syringae* on snap beans (*Phaseolus vulgaris* L). Phytopathology. 1981; 71(8):881-.

43. Willis DK, Hrabak EM, Rich JJ, Barta TM, Lindow SE, Panopoulos NJ. Isolation and characterization of a *Pseudomonas syringae* pathovar *syringae* mutant deficient In lesion formation on bean. Mol Plant Microbe Interact. 1990; 3(3):149–56.

44. Chahtane H, Nogueira Füller T, Allard PM, Marcourt L, Ferreira Queiroz E, Shanmugabalaji V, et al. The plant pathogen Pseudomonas aeruginosa triggers a DELLA-dependent seed germination arrest in *Arabidopsis*. Elife. 2018; 7 https://doi.org/10.7554/eLife.37082 PMID: 30149837

45. Bredow M, Sementchoukova I, Siegel K, Monaghan J. Pattern-triggered oxidative burst and seedling growth inhibition assays in *Arabidopsis thaliana*. J Vis Exp. 2019;(147) https://doi.org/10.3791/59437 PMID: 31180345

46. Porcu E, Sanna S, Fuchsberger C, Fritsche LG. Genotype imputation in genome-wide association studies. Curr Protoc Hum Genet. 2013; Chapter 1:Unit 1.25. https://doi.org/10.1002/0471142905.hg0125s78 PMID: 23853078

47. Shade A, Jacques MA, Barret M. Ecological patterns of seed microbiome diversity, transmission, and assembly. Curr Opin Microbiol. 2017; 37:15–22. https://doi.org/10.1016/j.mib.2017.03.010 PMID: 28437661

48. Butterworth J, McCartney HA. The dispersal of bacteria from leaf surfaces by water splash. Journal of Applied Bacteriology. 1991; 71(6):484–96.

49. Thakur S, Weir BS, Guttman DS. Phytopathogen genome announcement: draft genome sequences of 62 *Pseudomonas syringae* type and pathotype strains. Mol Plant Microbe Interact. 2016; 29(4):243–6. https://doi.org/10.1094/MPMI-01-16-0013-TA PMID: 26883489

50. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30(14):2068–9. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063

51. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004; 32(Database issue):D115–9. https://doi.org/10.1093/nar/gkh131 PMID: 14681372

52. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013; 41(12):e121. https://doi.org/10.1093/nar/gkt263 PMID: 23598997

53. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. Gigascience. 2019; 8(10)

54. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004; 5:113. https://doi.org/10.1186/1471-2105-5-113 PMID: 15318951

55. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One. 2010; 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

56. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018; 35(6):1547–9. https://doi.org/10.1093/molbev/msy096 PMID: 29722887

57. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

58. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM; 2016.