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 Middleton[1], Syama Chatterton[3], Darrell Desveaux [1], and David S. Guttman[1,2,*]

[1]Department of Cell & Systems Biology, University of Toronto, 25 Willcocks St., Toronto, Ontario, Canada

[2]Centre for the Analysis of Genome Evolution & Function, University of Toronto, Toronto, Ontario, Canada

[3]Agriculture and Agri-Food Canada, Lethbridge Research and Development Centre, Lethbridge, Alberta, Canada

*To whom correspondence should be addressed. Email: david.guttman@utoronto.ca

Running title: Predicting *P. syringae* virulence on bean using a machine learning approach

Keywords: *Pseudomonas syringae*, *Phaseolus vulgaris*, bean, machine learning, random forest, gradient boosted decision trees, host specificity, virulence,

ABSTRACT

Pseudomonas syringae is a genetically diverse bacterial species complex responsible for numerous agronomically important crop diseases. Individual isolates of *P. syringae* are typically assigned pathovar names based on their host of isolation and the associated disease symptoms, and these pathovar designations are often assumed to reflect host specificity. Unfortunately, this assumption has rarely been rigorously tested, which poses a challenge when trying to identify genetic factors associated with host specificity. Here we develop a rapid seed infection assays to measure the virulence of 121 diverse *P. syringae* isolates on common bean (*Phaseolus vulgaris*). This collection includes *P. syringae* phylogroup 2 bean isolates (pathovar *syringae*) that cause bacterial spot disease and *P. syringae* phylogroup 3 bean isolates (pathovar *phaseolicola*) that cause the much more serious halo blight disease. We find that phylogroup 2 strains generally show lower levels of host specificity on bean, with the average level of virulence for all strains in this phylogroup (irrespective of host of isolation) being higher than the average level for all other *P. syringae* strains. We then use gradient boosted decision trees to model the *P. syringae* virulence weights using whole genome kmers, type III secreted effector kmers, and the presence / absence of type III effectors and phytotoxins. Our machine learning model performed best using whole genome data, and we were able to predict bean virulence with high accuracy (mean absolute error as low as 0.05). Finally, we functionally validated the model by predicting virulence for 16 strains and found that 15 (94%) of the strains had virulence levels within the bounds of estimated predictions given the calculated RMSE values (±0.20). This study further illustrates that *P. syringae* phylogroup 2 strains may have evolved a different lifestyle than other *P. syringae* strains and demonstrates the power of machine learning for predicting host specific adaptation.

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). Strains have also been subdivided into phylogenetic groups (i.e., phylogroups) based on multilocus sequence typing or genomic analysis (1, 7-10). Currently, there are 13 recognized phylogroups (7), of which seven have been termed primary phylogroups based on their genetic relatedness and the near universal presence of the canonical *P. syringae* type III secretion system (discussed below) (1, 9).

Phylogenetic analyses of *P. syringae* isolates suggests that adaptation to specific hosts has evolved multiple times in the evolutionary history of the species complex. For example, cherry and plum pathogens are found in clades distributed in phylogroups 1, 2, and 3 (11, 12), hazelnut pathogens are distributed among two distinct clades in phylogroups 1 and 2 (13), while pathogens of common bean (including snap, green, kidney, and French bean) are found in phylogroups 2 and 3 (9). Common bean pathogens from *P. syringae* phylogroup 3 are generally classified as pathovar *phaseolicola* and are responsible for halo blight disease, which is characterized by large necrotic lessons surrounded by a chlorotic zone (halo) of yellow tissue (14-16). Bean pathogens of phylogroup 2 are generally classified as pathovar *syringae* and are responsible for bacterial spot disease, which is characterized by brown leaf spots (17, 18). While halo blight can cause very serious crop losses, bacterial spot disease is generally of only minor agronomic concern. The phylogroup 3 *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. In contrast, phylogroup 2

syringae bean isolates show very little phylogenetic clustering and are frequently more closely related to a non-bean isolates than another 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 a number of 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. For example, diverse lineages of cherry pathogens use multiple mechanisms to cause disease and have exchanged and lost key genes that contribute to successful infections multiple times (11, 12).

One of the most important and dynamic classes of *P. syringae* virulence and host specificity factors are type III secreted effectors (T3SEs). T3SE are proteins secreted through the type III secretion system and translocated directly into the eukaryotic host cell where they can directly interfere with host immunity or disrupt cellular homeostasis to promote the disease process. There are 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 (19). 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 (20, 21). 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 very dynamic

evolution of T3SEs in general, with very frequent horizontal transmission, acquisition, and loss (9, 19, 22).

While nearly every strain in phylogroups 2 and 3 carries a substantial suite of T3SEs, the suites vary in size, with phylogroup 3 strains carrying an average of ~27 T3SEs vs. ~19 for phylogroup 2 strains (19). Phylogroup 2 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 mimicking of hormones (3, 9, 23). In general, phylogroup 2 strains are believed to show lower levels of host specificity and better ability to survive on leaf surfaces (i.e., epiphytic growth) (1-3, 24).

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 (25). Statistical genetic approaches such as 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 for dealing with the shared evolutionary history of segregating genetic variants (i.e., population structure) (26-31). 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. Machine learning approaches are better at

identifying suites of interacting genetic variants than GWAS, but they are more limited in their ability to deal with complex evolutionary relationships among these variants.

Here, we implemented a rapid, high-throughput method for screening a proxy of virulence for *P. syringae* isolates on common bean. We use this screen to assess virulence of 121 strains from nine phylogroups on bean, and then impute the virulence for an additional set isolates based on their core genome relationship to the screened strains. While the majority of bean pathogens come from phylogroups 2 and 3, we find that phylogroup 3 pathogens display a much stronger degree of host specificity in comparison to phylogroup 2 pathogens. Finally, we develop a random forest regression model using genomic kmers and virulence factors as features to predict the virulence of *P. syringae* isolates on bean.

MATERIALS AND METHODS

Strain collection. Three hundred and thirty-four *Pseudomonas syringae* strains were used in this study (Table S1). 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, 32). 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 334, 318 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 (33). Prokka annotates proteins by searching for sequence similarity in the UniProtKB (34) database and HMM libraries (35). Additionally, all predicted genes were aligned against a custom T3SE database for the identification of potential T3SEs (19). Orthologs were inferred with PIRATE (36), which 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 (37) and later concatenated into a single protein alignment. We used FastTree2 (38) to infer the phylogenetic relationships of all 318 isolates.

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). 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 in order to balance our experimental design, which was skewed due to the fact that the vast majority of phylogroup 3 bean isolates (i.e., pathovar *phaseolicola*) fall into one clonal group. Initially, our selection criteria resulted in only six phylogroup 3 bean isolates compared to 16 phylogroup 2 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 phylogroups 2 and 3. Evolutionary distances and Jaccard similarity scores were inferred with MEGA7 (39) and R version 4.0.5 (40).

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.

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, and rifampicin and kanamycin for strains harboring constructs. 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 using: 1) whole genome kmers, 2) T3SE kmers, 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 kmers with identical distribution patterns using custom python scripts. Next, we used the Scikit-learn and the XGBoost python libraries (41) 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. 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.

RESULTS

Virulence screen development. We developed a high-throughput seed infection assay to measure the virulence of *P. syringae* isolates on common bean. We used seed infection based on the understanding that many bean infections are caused by contaminated seeds (42, 43) (14, 44). These infections can reduce overall plant health, which is reflected in plant fresh weight. 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. Plant fresh weight was then determined after 14 days. We first 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 phylogroups (Fig. 1). Well-established bean pathogens such as *P. syringae* pv. *phaseolicola* 1448A (Pph1448A) (14, 16) and *P. syringae* pv. *syringae* B728a (PsyB728a) (17, 45) showed the highest levels of bacterial growth and lowest plant weights, while the other isolates from phylogroups 1-7, 11, and 13 showed a range of values. On the whole 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.

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 (Fig. S1). 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 phylogroups to assess the virulence potential as measured by reduced plant fresh weight in 14-day old bean plants after seed infection. The screened strains resulted 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 (Fig. 2). 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 in screened.

Significant differences in virulence, as measured by normalized fresh weight, were observed when comparing the strain collection stratified by host of isolation and phylogroup (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). All of the bean isolates are found in phylogroups 2 and 3, so we compared the virulence of bean isolates to non-bean isolates within these two phylogroups individually and found no significant difference for phylogroup 2 (p=0.128, 2-tailed, heteroscedastic T-test) but a strong difference for phylogroup 3 (p=8.9e-4, 2-tailed heteroscedastic T-test). Additionally, there were no significant differences between phylogroup 2 and phylogroup 3 bean isolates (p=0.460). We then looked for differences in virulence between strains from different phylogroups irrespective of their host of isolation (only comparing phylogroups with at least six tested strains, using 2-tailed, heteroscedastic T-tests, Bonferroni corrected for seven total tests), and found that strains in phylogroup 2 were significantly more virulent on bean than strains from phylogroups 1, 3, and 4 (p= 1.33e-07, 0.012, and 0.029 respectively), but not relative to phylogroup 6. In contrast,

strains from phylogroup 3 were only significantly more virulent on bean than strains from phylogroup 1 (p=0.006). No other significant pairwise phylogroup comparisons were observed. Interestingly, when we asked if non-bean phylogroup 2 isolates showed higher virulence on bean than non-bean isolates from all other phylogroups, we observed a significant difference (p=2.87e-4) indicating that while bean isolates specifically show greater virulence on bean, phylogroup 2 isolates in general also show greater virulence on bean. This pattern was actually reversed in phylogroup 3 where non-bean isolates had significantly lower virulence than non-bean isolates from all other phylogroups (p=4.52e-3).

Virulence and germination inhibition. 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 (42, 43, 46) both through the direct action of phytotoxins (23) and the indirect action of immune activation. In fact, seedling growth inhibition is a well-established assay for immune activation in *Arabidopsis thaliana* (47). In general, the frequency distribution for bean germination inhibition was much less skewed than the frequency distribution for virulence, with mean = 60.1%, media = 63.3%, and standard deviation = 24.3% (Fig 2, 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, 2-tailed, heteroscedastic T-test). When stratifying the bean isolates by phylogroup, we found no significant difference in germination frequency between phylogroup 2 bean and non-bean isolates (p=0.076), while there was for phylogroup 3 (p=0.002), and we did observe a significant difference between germination frequency for phylogroup 2 bean isolates and phylogroup 3 bean isolates (p=0.014). Similar to what was found for the virulence assays, phylogroup 2 strains were significantly different from strains from phylogroups 1, 3, and 4 (p=0.010, 6.23e-5, 8.53e-11 respectively 2-tailed, heteroscedastic T-test, Bonferroni corrected), while phylogroup three strain were also significantly different from phylogroup 4 (p=2.23e-4). Also similar to the virulence results, non-bean phylogroup 2 isolates

result in lower germination rates than all other non-bean isolates (p=9.60e-5), while non-bean phylogroup 3 strains had a higher germination rate than non-bean isolates from all other phylogroups (p=0.004), although this pattern disappears when non-bean phylogroup 2 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 phylogroup and bean isolates showed a strong association for phylogroup 2 bean isolates (linear regression; F = 28.91, df = 13, p=0.0001, R^2 = 0.68), but no significant association for phylogroup 3 bean isolates (R^2=0.62, p=0.06) (Fig 3).

Predictive modelling of *P. syringae* virulence on bean. We used a gradient boosted decision tree regression model to predict the virulence of *P. syringae* isolates on beans. The model used one of three input feature classes: 1) genomic kmers; 2) T3SE kmers; or 3) presence / absence of T3SEs and phytotoxins. T3SEs and phytotoxins are well-know virulence factors, with the former often strongly associated with host specificity. Plant weight 14 days after seed infection was used as a continuous outcome. The goal of this model 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 included the 114 of the isolates directly screened for virulence, which included 50 phylogroup 2 strains and 42 phylogroup 3 strains. This collection is slightly smaller than the full screened set since it does not include the additional phylogroup 3 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. The imputation process

identified strains in our collection that had a core genome evolutionary distance of less than 0.001 and a T3SE Jaccard similarity of greater than 0.8 compared to one of the screened strains. 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 318 strain, which included 67 phylogroup 2 strains and 142 phylogroup 3 strains. We also tested phylogroup 2 and phylogroup 3 strains separately since bean isolates from these phylogroups interact with their host very differently.

Our analysis revealed a significant linear relationship between sample size and the overall performance of the models (Fig. 5, MAE linear regression $F = 5.89$, df = 10, p=0.03, $R^2 = 0.3$). Hence, the data sets with the least number of samples consistently showed the poorest performance regardless of the input data (e.g., phylogroup 2 and the screened collection). Furthermore, our models achieved the highest predictive power when looking exclusively at isolates from PG3, which may be explained by the strong clonality patterns displayed by bean pathogens from this phylogroup. Perhaps not surprisingly, models built using the greatest diversity, with respect to number of strains (i.e., the expanded strain collection) and number of genetic feature (i.e., whole genome kmers) showed the highest predictive power, with the best model having a mean absolute error (MAE) of 0.06.

Model functional validation. We evaluate the power of our gradient boosted decision tree regression model on unseen strains by using the kmer profiles of *P. syringae* isolates not previously examined to predict virulence, and comparing these predictions to actual virulence measures obtained through the seed infection virulence assay. Importantly, none of the strains in the functional validation set were clonal related to any screened strain - they did not share a core genome evolutionary distance of >0.001 and a T3SE Jaccard similarity of <0.8 with any screened strain. When comparing actual to predicted virulence on the 16 strains in the

functional validation set, we found that 15 (94%) strains had virulence levels within the bounds of estimated predictions given the calculated RMSE values (±0.20) (Fig. 5).

DISCUSSION

This work addressed the question of whether host of isolation is a reliable predictor of host specific virulence, and conversely, can we use whole genome sequences to predict the host specific virulence potential of individual strains. *Pseudomonas syringae* host specificity has been studied in one form or another for many years. Some of the earliest molecular plant immunity work was done in *P. syringae* with the cloning of the first “avirulence” gene (i.e., an ETI-eliciting T3SE) in 1984 (48). Since those early studies, current technologies and methodologies now enable the genome sequencing and functional characterization of hundreds of strains (9, 24, 49). Despite this outstanding progress, we still have very limited ability to predict the virulence potential of individual strains, even when full genome sequences are available. The host of isolation is an extremely widely used surrogate for host specificity, and to a large extent the basis for the pathovar nomenclature system. Despite the importance of this assumption, it has rarely been systematically studied, and often fails when put to the test (24, 49).

Our first aim was to determine if infection of bean seeds by *P. syringae* recapitulated virulence responses seen in more standard syringe inoculation virulence assays of bacteria into leaves. We found a negative association between normalized plant fresh weight after seed infection and *in planta* bacterial growth after syringe infiltration into leaf tissue, showing that the alternative infection protocol works, and suggesting that virulence can systemically disrupt growth and development. This finding is consistent with published and anecdotal reports that infected seed stocks are a significance source of bean disease (42, 43, 50, 51). 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 phylogroup 2 leaf spot bean pathogens had a normalized mean fresh weight of 0.56±0.293 (stddev) compared to 0.64±0.246 for the phylogroup 3 halo blight bean pathogens, 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. But in this case, the 38.8% mean germination frequency of phylogroup 2 bean pathogens was significantly lower than the 57.0% mean germination frequency of the phylogroup 3 bean pathogens ($p=0.014$).

Phylogroup 2 strains in general showed lower levels of host specificity on bean. Neither our fresh weight virulence measure nor the germination rates were significantly different between the bean and non-bean isolates of phylogroup 2. This comparison was quite different for phylogroup 3 strains, where bean isolates showed significantly higher virulence and lower germination frequency compared to non-bean isolates from the same phylogroup. These findings are consistent with other studies that have found lower levels of host specificity among phylogroup 2 strains (24, 49) and lends support to the hypothesis that phylogroup 2 strains may rely as much or more on toxin than T3SEs than other *P. syringae* strains.

While many genome-wide association studies have successfully identified strong genotype to phenotype linkages to date, we were unable to identify any loci significantly associated with bean isolation (data not shown). 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 phylogroup affiliation, our model was able to predict the virulence of individual *P. syringae* isolates with low error margins, based solely on whole

genome data. However, 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 kmers suggests that factors other than virulence-associated genes also play important roles on disease development and adaptation to beans.

Sample size is one of the most important contributors to accurate model generation in machine learning. Simple models trained with large data sets can learn to distinguish label categories with much stronger accuracy than complex models trained with small datasets. Our results also found a correlation between sample size and model accuracy, with 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. We find much poorer model performance for phylogroup 2 strains than phylogroup 3 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 phylogroup 3 are phylogenetically clustered, while there is little clustering of bean isolates in phylogroup 2. Consequently, the model may perform better on phylogroup 3 since it is essentially predicting phylogenetic structure. Another contributing factor is likely the finding discussed above, namely, host specificity appears to be weaker in phylogroup 2. If phylogroup 2 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.

REFERENCES

1. Baltrus DA, McCann HC, Guttman DS. 2017. Evolution, genomics and epidemiology of *Pseudomonas syringae*: Challenges in Bacterial Molecular Plant Pathology. Mol Plant Pathol 18:152-168.
2. Morris CE, Monteil CL, Berge O. 2013. The life history of *Pseudomonas syringae*: linking agriculture to earth system processes. Annu Rev Phytopathol 51:85-104.
3. Xin XF, Kvitko B, He SY. 2018. *Pseudomonas syringae*: what it takes to be a pathogen. Nat Rev Microbiol 16:316-328.
4. Morris CE, Sands DC, Vinatzer BA, Glaux C, Guilbaud C, Buffiere A, Yan S, Dominguez H, Thompson BM. 2008. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. ISME J 2:321-34.
5. Bull CT, De Boer SH, Denny TP, Firrao G, Fischer-Le Saux M, Saddler GS, Scortichini M, Stead DE, Takikawa Y. 2008. Demystifying the nomenclature of bacterial plant pathogens. J Plant Pathol 90:403-417.
6. Dye DW, Bradbury JF, Goto M, Hayward AC, Lelliott RA, Schroth MN. 1980. International standards for naming pathovars of phytopathogenic bacteria and a list of pathovar names and pathotype strains. Review of Plant Pathology 59:153-168.
7. Berge O, Monteil CL, Bartoli C, Chandeysson C, Guilbaud C, Sands DC, Morris CE. 2014. 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 9:e105547.
8. Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. Appl Environ Microbiol 70:1999-2012.
9. Dillon MM, Thakur S, Almeida RND, Wang PW, Weir BS, Guttman DS. 2019. Recombination of ecologically and evolutionarily significant loci maintains genetic cohesion in the *Pseudomonas syringae* species complex. Genome Biol 20:3.

10. Hwang MS, Morgan RL, Sarkar SF, Wang PW, Guttman DS. 2005. Phylogenetic characterization of virulence and resistance phenotypes of *Pseudomonas syringae*. Appl Environ Microbiol 71:5182-91.

11. Hulin MT, Armitage AD, Vicente JG, Holub EB, Baxter L, Bates HJ, Mansfield JW, Jackson RW, Harrison RJ. 2018. Comparative genomics of *Pseudomonas syringae* reveals convergent gene gain and loss associated with specialization onto cherry (*Prunus avium*). New Phytol 219:672-696.

12. Hulin MT, Mansfield JW, Brain P, Xu X, Jackson RW, Harrison RJ. 2018. Characterization of the pathogenicity of strains of *Pseudomonas syringae* towards cherry and plum. Plant Pathol 67:1177-1193.

13. O'Brien HE, Thakur S, Gong Y, Fung P, Zhang J, Yuan L, Wang PW, Yong C, Scortichini M, Guttman DS. 2012. Extensive remodeling of the *Pseudomonas syringae* pv. *avellanae* type III secretome associated with two independent host shifts onto hazelnut. BMC Microbiol 12:141.

14. Arnold DL, Lovell HC, Jackson RW, Mansfield JW. 2011. Pseudomonas syringae pv. phaseolicola: from 'has bean' to supermodel. Mol Plant Pathol 12:617-27.

15. Tsiamis G, Mansfield JW, Hockenhull R, Jackson RW, Sesma A, Athanassopoulos E, Bennett MA, Stevens C, Vivian A, Taylor JD, Murillo J. 2000. Cultivar-specific avirulence and virulence functions assigned to avrPphF in *Pseudomonas syringae* pv. phaseolicola, the cause of bean halo-blight disease. EMBO J 19:3204-14.

16. Joardar V, Lindeberg M, Jackson RW, Selengut J, Dodson R, Brinkac LM, Daugherty SC, Deboy R, Durkin AS, Giglio MG, Madupu R, Nelson WC, Rosovitz MJ, Sullivan S, Crabtree J, Creasy T, Davidsen T, Haft DH, Zafar N, Zhou L, Halpin R, Holley T, Khouri H, Feldblyum T, White O, Fraser CM, Chatterjee AK, Cartinhour S, Schneider DJ, Mansfield J, Collmer A, Buell CR. 2005. Whole-genome sequence analysis of

Pseudomonas syringae pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J Bacteriol 187:6488-98.

17. Feil H, Feil WS, Chain P, Larimer F, DiBartolo G, Copeland A, Lykidis A, Trong S, Nolan M, Goltsman E, Thiel J, Malfatti S, Loper JE, Lapidus A, Detter JC, Land M, Richardson PM, Kyrpides NC, Ivanova N, Lindow SE. 2005. Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. Proc Natl Acad Sci U S A 102:11064-9.

18. Marco ML, Legac J, Lindow SE. 2005. Pseudomonas syringae genes induced during colonization of leaf surfaces. Environ Microbiol 7:1379-1391.

19. Dillon MM, Almeida RND, Laflamme B, Martel A, Weir BS, Desveaux D, Guttman DS. 2019. Molecular evolution of *Pseudomonas syringae* type III secreted effector proteins. Front Plant Sci 10:418.

20. Buttner D. 2016. Behind the lines-actions of bacterial type III effector proteins in plant cells. FEMS Microbiol Rev 40:894-937.

21. Khan M, Seto D, Subramaniam R, Desveaux D. 2018. Oh, the places they'll go! A survey of phytopathogen effectors and their host targets. Plant J 93:651-663.

22. Martel A, Ruiz-Bedoya T, Breit-McNally C, Laflamme B, Desveaux D, Guttman DS. 2021. The ETS-ETI cycle: evolutionary processes and metapopulation dynamics driving the diversification of pathogen effectors and host immune factors. Curr Opin Plant Biol 62:102011.

23. Bender CL, Alarcon-Chaidez F, Gross DC. 1999. *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. Microbiol Mol Biol Rev 63:266-92.

24. Morris CE, Moury B. 2019. Revisiting the concept of host range of plant pathogens. Annu Rev Phytopathol doi:10.1146/annurev-phyto-082718-100034.

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

26. Chen PE, Shapiro BJ. 2015. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol 25:17-24.

27. Falush D. 2016. Bacterial genomics: Microbial GWAS coming of age. Nat Microbiol 1:16059.

28. Falush D, Bowden R. 2006. Genome-wide association mapping in bacteria? Trends Microbiol 14:353-5.

29. Power RA, Parkhill J, de Oliveira T. 2017. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet 18:41-50.

30. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D. 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. Proc Natl Acad Sci U S A 110:11923-7.

31. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CC, Iqbal Z, Clifton DA, Hopkins KL, Woodford N, Smith EG, Ismail N, Llewelyn MJ, Peto TE, Crook DW, McVean G, Walker AS, Wilson DJ. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol 1:16041.

32. Thakur S, Weir BS, Guttman DS. 2016. Phytopathogen Genome Announcement: draft genome sequences of 62 *Pseudomonas syringae* type and pathotype strains. Mol Plant Microbe Interact 29:243-6.

33. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068-9.

34. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32:D115-9.

35. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41:e121.

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

37. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

38. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

39. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35:1547-1549.

40. R Development Core Team. 2020. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

41. Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining doi:10.1145/2939672.2939785. ACM, San Francisco, California, USA.

42. Darrasse A, Bureau C, Samson R, Morris CE, Jacques M-A. 2007. 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 119:203-215.

43. Eyster HC. 1940. The cause of decreased germination of bean seeds soaked in water. Am J Bot 27:652-659.

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

45. Willis DK, Hrabak EM, Rich JJ, Barta TM, Lindow SE, Panopoulos NJ. 1990. Isolation and characterization of a *Pseudomonas syringae* pathovar *syringae* mutant deficient In lesion formation on bean. Mol Plant Microbe Interact 3:149-156.

46. Chahtane H, Nogueira Füller T, Allard PM, Marcourt L, Ferreira Queiroz E, Shanmugabalaji V, Falquet J, Wolfender JL, Lopez-Molina L. 2018. The plant pathogen Pseudomonas aeruginosa triggers a DELLA-dependent seed germination arrest in Arabidopsis. Elife 7.

47. Bredow M, Sementchoukova I, Siegel K, Monaghan J. 2019. Pattern-Triggered Oxidative Burst and Seedling Growth Inhibition Assays in Arabidopsis thaliana. J Vis Exp doi:10.3791/59437.

48. Staskawicz BJ, Dahlbeck D, Keen NT. 1984. Cloned avirulence gene of *Pseudomonas syringae* pv. *glycinea* determines race-specific incompatibility on *Glycine max* (L.) Merr. Proc Natl Acad Sci U S A 81:6024-8.

49. Morris CE, Lamichhane JR, Nikolić I, Stanković S, Moury B. 2019. The overlapping continuum of host range among strains in the *Pseudomonas syringae* complex. Phytopathology Research 1:4.

50. Shade A, Jacques MA, Barret M. 2017. Ecological patterns of seed microbiome diversity, transmission, and assembly. Curr Opin Microbiol 37:15-22.

51. Upper CD, Hirano SS, Dodd KK, Clayton MK. 2003. Factors that Affect Spread of *Pseudomonas syringae* in the Phyllosphere. Phytopathol 93:1082-92.

TABLES

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

Table 2. Machine Learning Model Performance

Group	Samples	Dataset	MAE	RMSE
PG3	142	Whole genome kmers	0.049	0.107
PG3	142	Toxin and T3SE binary matrix	0.054	0.101
Imputed	318	Whole genome kmers	0.067	0.140
PG3	142	T3SE kmers	0.085	0.126
Imputed	318	Toxin and T3SE binary matrix	0.092	0.153
Imputed	318	T3SE kmers	0.096	0.177
PG2	66	T3SE kmers	0.154	0.233
Screened	114	Whole genome kmers	0.155	0.204
Screened	114	T3SE kmers	0.159	0.213
Screened	114	Toxin and T3SE binary matrix	0.163	0.201
PG2	66	Whole genome kmers	0.169	0.242
PG2	66	Toxin and T3SE binary matrix	0.202	0.262

591

FIGURE LEGENDS

Figure 1: Bacterial growth is inversely correlated with plant weight. Bacterial density in syringe infiltrated bean leaves as a function of normalized median weight of seed infected plants. There is a strong negative correlation between bacterial density and plant weight across isolates in the *P. syringae* species complex (linear regression; F = 36.95, df = 21, p = 4.95e-06, R^2 = 0.62).

Figure 2: Frequency distribution virulence and germination assays. Distributions of data from the 121 strains screened for virulence (plant fresh weight) and germination by the seed infection assay.

Figure 3: Normalized median plant fresh weights of seed infected bean plants. (A) Boxplots show virulence (normalized plant fresh weight) of 121 *P. syringae* strains after seed infection on bean. The boxplots are color-coded by the phylogroup of each strain, and asterisks above the boxplots denote bean isolates. Plants infected with bean isolates tend to have higher virulence (i.e., lower plant fresh weight) than plants infected with strains isolated from other hosts. (B) Germination frequency of seed infected plants. (C) Core genome phylogeny of 318 *P. syringae* isolates. Leaf colors represent phylogroup affiliation. Metadata row 1 shows isolation source; metadata row 2 shows normalized median weight of screened isolates, and metadata row 3 shows normalized median weight for the expanded strain collection based on phylogenetic imputation.

Figure 4: Correlation between germination frequency and virulence of seed infected bean plants. (A) Germination frequency vs virulence (i.e., normalized fresh weight) stratified by

phylogroup (PG) 2 and phylogroup 3 bean isolates. A strong correlation is found between germination frequency and virulence of bean strains from phylogroup 2 (linear regression; F = 28.91, df = 13, p = 0.00012, r^2 = 0.66). (B) Germination frequency vs. virulence stratified by phylogroup (PG) irrespective of host.

Figure 5: Performance of supervised machine learning models on virulence predictions with genome data. Models trained with the screed collection (yellow), imputed expanded collection (blue), imputed expanded phylogroup (PG) 2 set (orange), and imputed expanded phylogroup 3 set (green). Diamonds, circles, and triangles represent models trained with a phytotoxin and T3SE binary matrix, T3SE kmers, and whole genome kmers, respectively. (A) Mean Absolute Error (MAE) as a function of number of samples across 50 cross-validation splits. Increasing the number of samples seen by the model significantly increases model performance (linear regression; F = 5.89, df = 10, p = 0.03, r^2 = 0.37). (B) Root Mean Square Error (RMSE) as a function of number of samples across 50 cross-validation splits. Models trained on whole genome kmers with a higher number of samples consistently outperform other models.

Figure 6: Correlation between predicted and observed weights of plants infected with isolates previously unseen by the model. (A) Predicted virulence as a function of observed virulence on seed infected bean plants. Color coding indicates phylogroups (PG). Dashed line indicates 1:1 relationship. (B) Virulence for each isolate tested ordered by predicted virulence and colored by phylogroup. 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.

SUPPLEMENTARY MATERIAL

Figure S1: 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.

Figure S2: MAE and RMSE across 50 cross-validation splits. Model performance in terms of Mean Absolute Error and Root Mean Square Error. Plant weight distributions were kept the same across splits.

Table S1: List of *P. syringae* isolates used in this study.

Fig 1

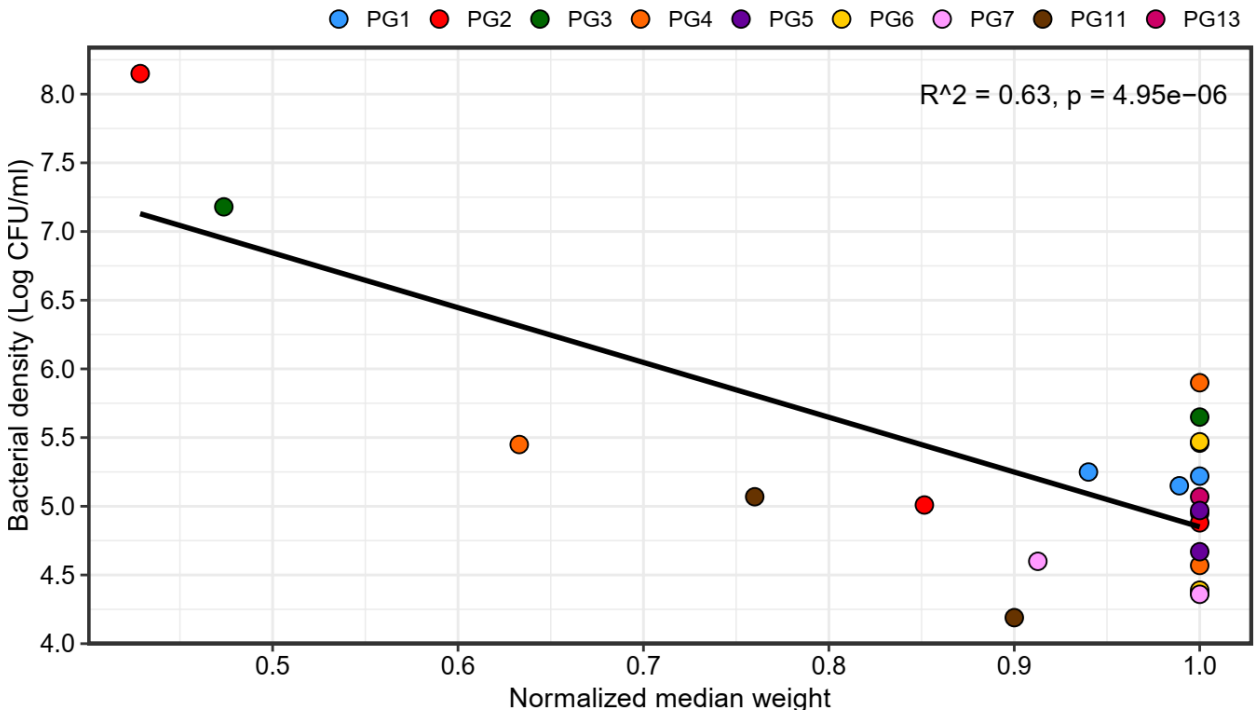

PG1
PG2
PG3
PG4
PG5
PG6
PG7
PG11
PG13
R^2 = 0.63, p = 4.95e−06
Bacterial density (Log CFU/ml)
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
0.5
0.6
0.7
0.8
0.9
1.0
Normalized median weight

Fig 2.

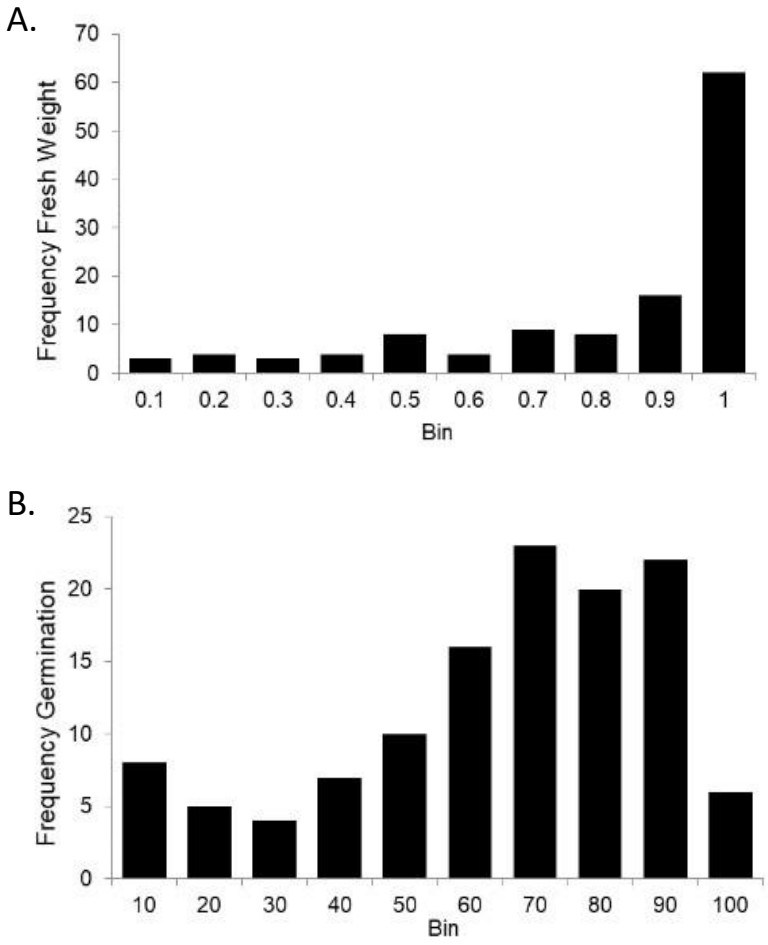

A.
Frequency Fresh Weight
70
60
50
40
30
20
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Bin
B.
Frequency Germination
25
20
15
10
5
0
10
20
30
40
50
60
70
80
90
100
Bin

Fig 3.

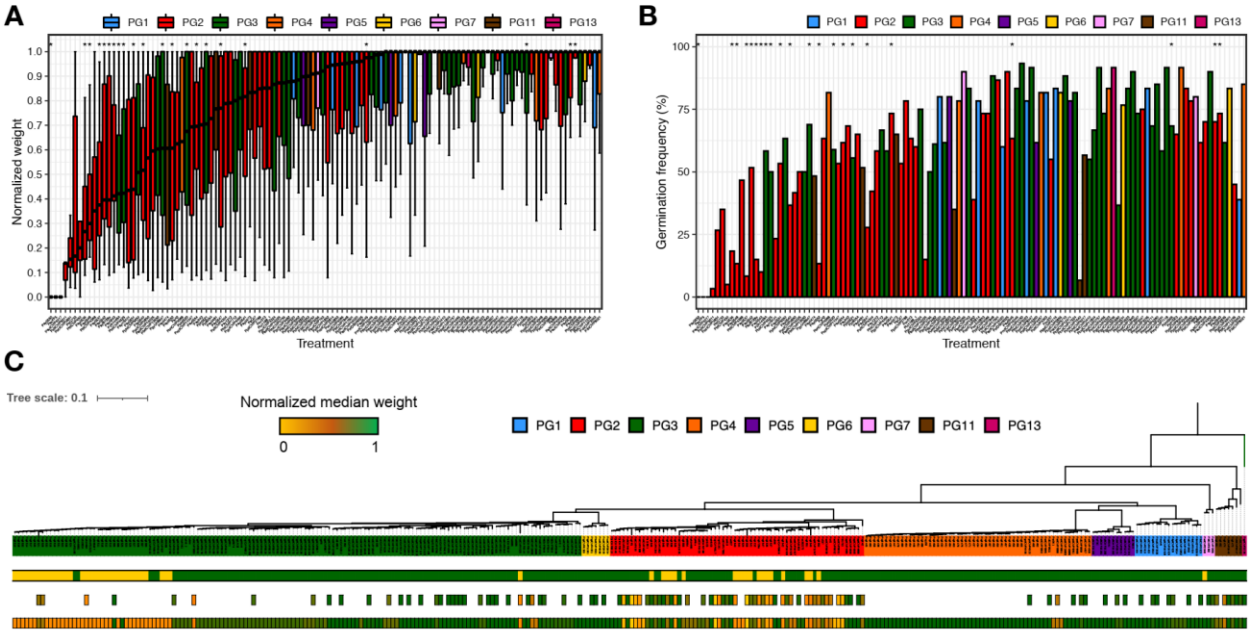

A
PG1 PG2 PG3 PG4 PG5 PG6 PG7 PG11 PG13
Normalized weight
Treatment
B
PG1 PG2 PG3 PG4 PG5 PG6 PG7 PG11 PG13
Germination frequency (%)
Treatment
C
Tree scale: 0.1
Normalized median weight
0
1
PG1 PG2 PG3 PG4 PG5 PG6 PG7 PG11 PG13

Fig 4.

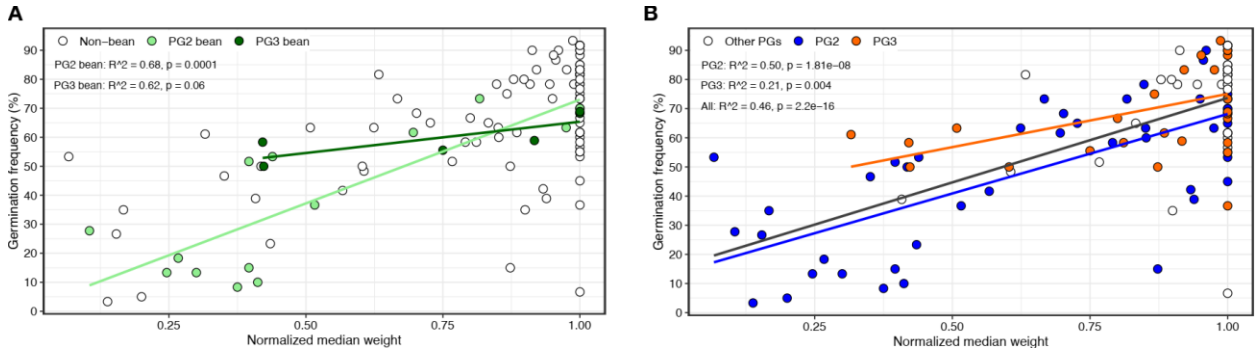

A
Non−bean
PG2 bean
PG3 bean
PG2 bean: R^2 = 0.68, p = 0.0001
PG3 bean: R^2 = 0.62, p = 0.06
Germination frequency (%)
Normalized median weight
B
Other PGs
PG2
PG3
PG2: R^2 = 0.50, p = 1.81e−08
PG3: R^2 = 0.21, p = 0.004
All: R^2 = 0.46, p = 2.2e−16
Germination frequency (%)
Normalized median weight

Fig 5.

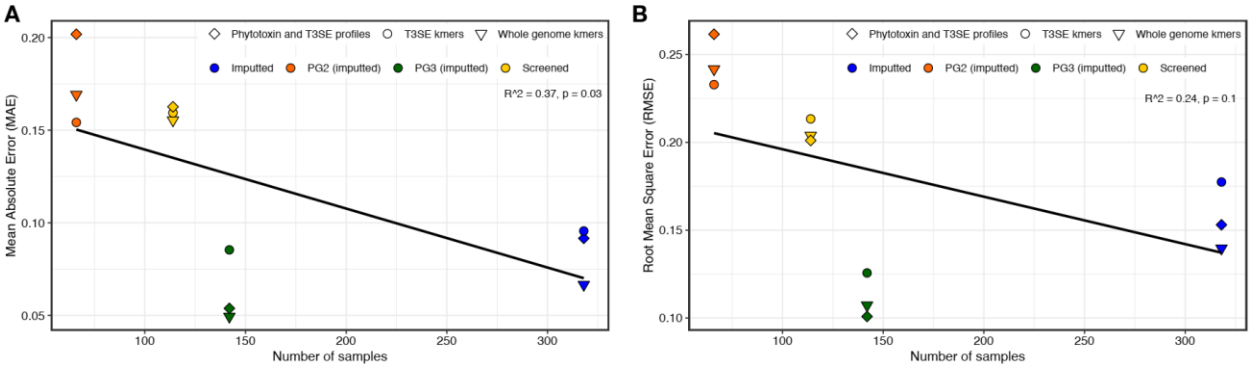

A
Phytotoxin and T3SE profiles
T3SE kmers
Whole genome kmers
Imputted
PG2 (imputted)
PG3 (imputted)
Screened
R^2 = 0.37, p = 0.03
Mean Absolute Error (MAE)
0.20
0.15
0.10
0.05
100
150
200
250
300
Number of samples
B
Phytotoxin and T3SE profiles
T3SE kmers
Whole genome kmers
Imputted
PG2 (imputted)
PG3 (imputted)
Screened
R^2 = 0.24, p = 0.1
Root Mean Square Error (RMSE)
0.25
0.20
0.15
0.10
100
150
200
250
300
Number of samples

Fig 6.

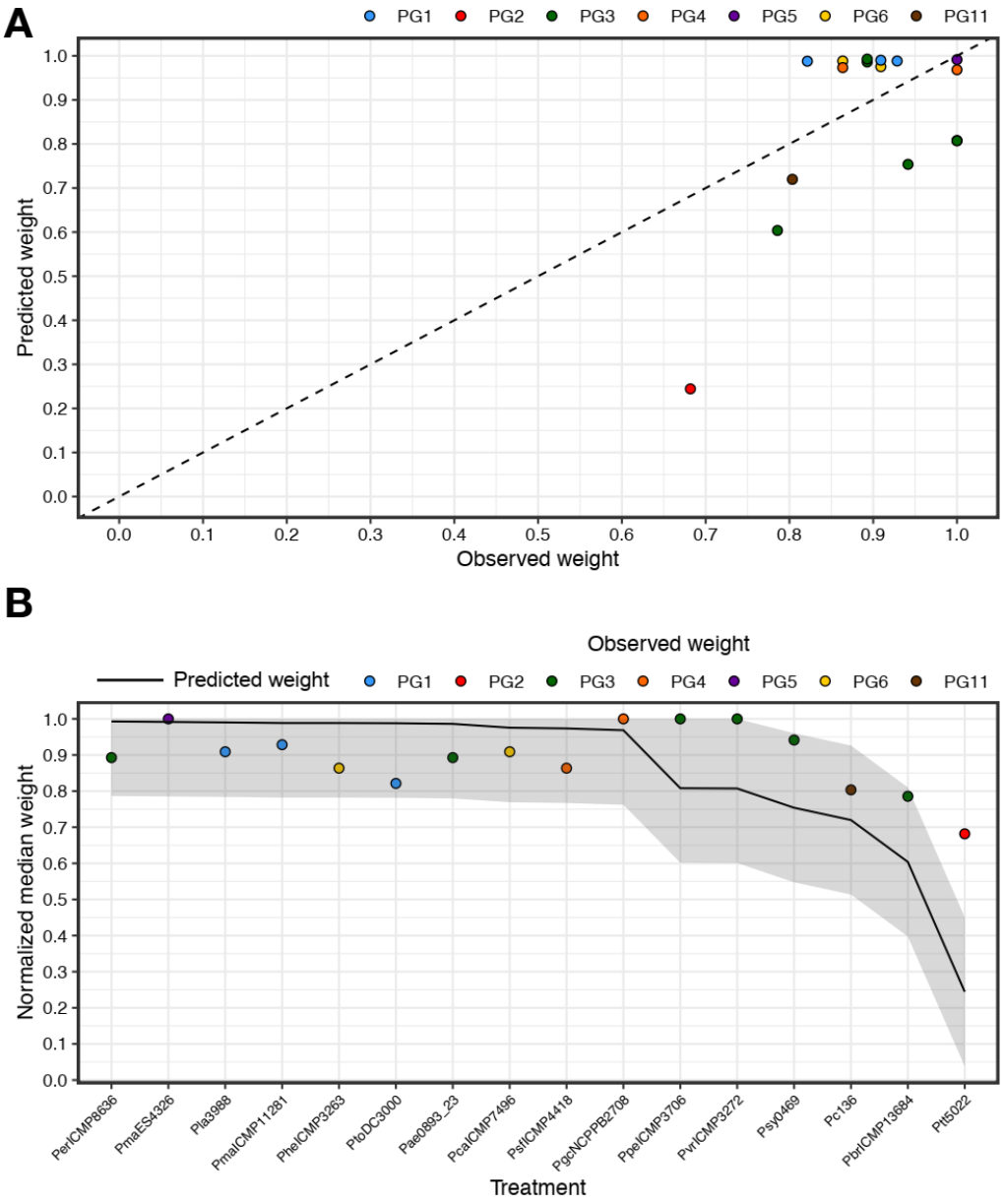

A
PG1
PG2
PG3
PG4
PG5
PG6
PG11
Predicted weight
Observed weight
B
Observed weight
Predicted weight
Normalized median weight
Treatment
PerICMP8636
PmaES4326
Pla3988
PmaICMP11281
PheICMP3263
PtoDC3000
Pae0893_23
PcaICMP7496
PsfICMP4418
PgcNCPPB2708
PpeICMP3706
PvrICMP3272
Psy0469
Pc136
PbrICMP13684
Pft5022

Fig S1

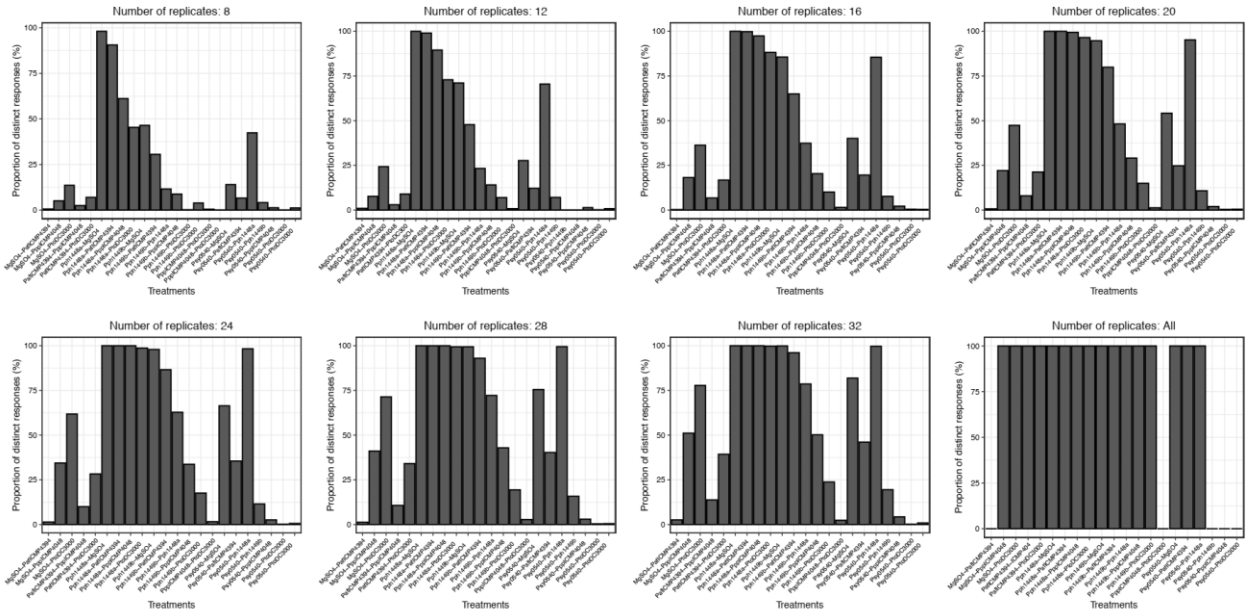
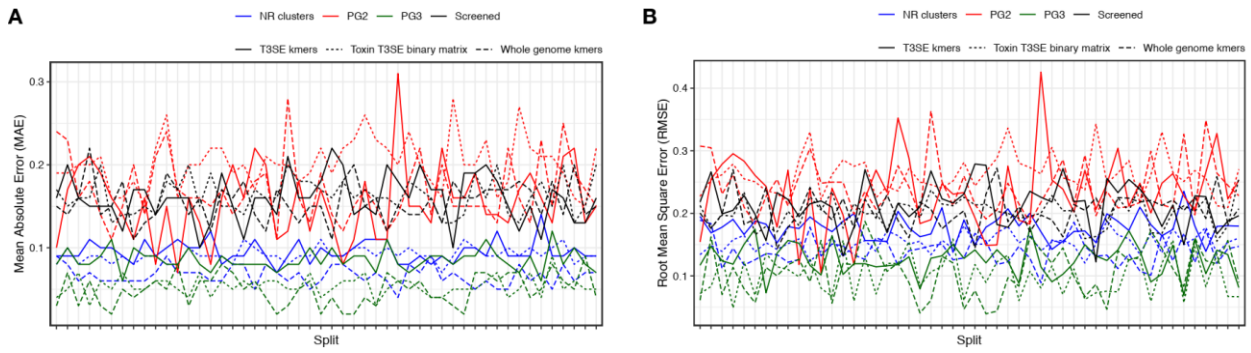

Number of replicates: 8
Number of replicates: 12
Number of replicates: 16
Number of replicates: 20
Number of replicates: 24
Number of replicates: 28
Number of replicates: 32
Number of replicates: All
Proportion of distinct responses (%)
Treatments

Fig S2.


A
NR clusters
PG2
PG3
Screened
T3SE kmers
Toxin T3SE binary matrix
Whole genome kmers
Mean Absolute Error (MAE)
0.3
0.2
0.1
Split
B
NR clusters
PG2
PG3
Screened
T3SE kmers
Toxin T3SE binary matrix
Whole genome kmers
Root Mean Square Error (RMSE)
0.4
0.3
0.2
0.1
Split