

Machine learning approaches to predict the plant-associated phenotype of *Xanthomonas* strains

Dennie te Molder ¹, Wasin Poncheewin ¹, Peter J. Schaap ^{1,2}, and Jasper J. Koehorst ^{1,2,✉}

¹Laboratory of Systems and Synthetic Biology | Department of Agrotechnology and Food Sciences, Wageningen University and Research, Wageningen, the Netherlands

²UNLOCK Large Scale Infrastructure for Microbial Communities, Wageningen University and Research, Wageningen, the Netherlands

The genus *Xanthomonas* has long been considered to consist predominantly of plant pathogens, but over the last decade there has been an increasing number of reports on non-pathogenic and endophytic members. As *Xanthomonas* species are prevalent pathogens on a wide variety of important crops around the world, there is a need to distinguish between these plant-associated phenotypes. To date a large number of *Xanthomonas* genomes have been sequenced, which enables the application of machine learning (ML) approaches on the genome content to predict this phenotype. Until now such approaches to the pathogenomics of *Xanthomonas* strains have been hampered by the fragmentation of information regarding strain pathogenicity over many studies. Unification of this information into a single resource was therefore considered to be an essential step. Mining of 39 papers considering both plant-associated phenotypes, allowed for a phenotypic classification of 578 *Xanthomonas* strains. For 65 plant-pathogenic and 53 non-pathogenic strains the corresponding genomes were available and *de novo* annotated for the presence of Pfam protein domains used as features to train and compare three ML classification algorithms; CART, Lasso and Random Forest. Recursive feature extraction provided further insights into the virulence enabling factors, but also yielded domains linked to traits not present in pathogenic strains.

pathogenicity | protein domains | Machine Learning | *Xanthomonas*

Correspondence: jasper.koehorst@wur.nl

Introduction

The genus of *Xanthomonas* is mostly known for its pathogenic behaviour, with significant economic and agricultural impact (1). *Xanthomonas* spp. infect a wide variety of plant crops, see Table 1 for examples, however individual *Xanthomonas* pathovars usually show a high degree of both host and tissue specificity (2). Whilst non-pathogenic xanthomonads have been reported as early as 1985 (3), during the last decade many new non-pathogenic strains have been discovered (4–9) and it has become apparent that these non-pathogenic strains form an integral part of the *Xanthomonas* epidemic population structure (10). Moreover, non-pathogenic strains show, in comparison to their pathogenic counterparts, a higher level of genetic diversity (1) suggesting that non-pathogenic xanthomonads are generalists that can epiphytically survive on a much wider host range and might play important roles in the microbiome of healthy, asymptomatic hosts (11, 12). While the relative abundance of these

non-pathogenic strains is still not known, their undisputable existence has raised the concern that diagnostic misidentifications might result in unnecessary control measures and/or high economic losses (8). The is-not-pathogenic label depends heavily on the tested conditions used. For example, a set of *X. arboricola* pv. *fragariae* strains isolated from infected strawberries did, when sprayed onto new plants, not cause symptoms (13), but a repetition of the same assay at increased humidity conditions did reveal the pathogenicity of these strains (14). Examples like this underline the importance of testing strains on a large range of hosts and conditions. However, as current tests are all aimed at establishing pathogenicity and given the large number of environmental parameters that impact infection, it is important to integrate extensive testing of strains with a pathogenomics framework providing insight in the relative importance of genome encoded virulence factors.

A vast array of genomic factors have already been found to impact virulence (reviewed in (15, 16)). Many are located in so called pathogenicity clusters such as the *hrp*-cluster expressing type III secretion systems (T3SS) and associated effectors (T3E) (17), the *xps*-cluster coding for a type II secretion system for secretion of host cell wall degrading enzymes (18), the *gum*-cluster responsible for production of the xanthan-based biofilm unique for *Xanthomonas* spp. (19) and the regulation of pathogenicity factors or *rpf*-cluster which positively regulates virulence (20). Elements of many of these clusters are also present in the genomes of the non-pathogens and currently it remains unclear what combinations of features drive the switch in life-style (15). Studies examining the genomic differences between pathogenic and non-pathogenic xanthomonads have focused on *X. arboricola* as a model (21) and as a result the majority of known non-pathogenic strains currently belong to this species. Using classical comparative genomic approaches, attempts have been made to understand what exactly separates pathogenic and non-pathogenic *X. arboricola* strains. These analyses revealed genome encoded differences in environmental sensing, flagellin protein sequences, and components of the type IV pilus, but the most remarkable differences were found in the T3SS and T3E genes content. Most non-pathogenic strains lacked parts of the T3SS and showed a more limited repertoire of T3Es and in extreme cases, non-pathogenic strains even lacked the entire T3SS (1, 8, 11, 22). How-

ever, these findings do not fully explain the differences between the two plant associated phenotypes in *X. arboricola*, as strains CFBP3122 and CFBP3123 were found to be pathogenic although they lacked T3SS and T3E genes (11). These findings also provide an incomplete framework for other species. For example, whilst for *X. maliensis* absence of the T3SS related genes appear to be strongly correlated with the non-pathogenic phenotype (23), several members of the *X. cannabis* species have been shown to be pathogenic to multiple different hosts even though they lack the entire T3SS.

Overall the results suggest that a more complete set of genome-encoded features is required to differentiate between phytopathogenic and non-pathogenic strains (22, 24). In the last ten years, the number of available *Xanthomonas* genomes has increased nearly 100-fold (2, 25). This large number of genomes makes it feasible to use machine learning (ML) approaches on the genome content to predict the pathogenicity of individual strains. Until now such approaches have been hampered by the fragmentation of information on the plant-associated phenotype of individual *Xanthomonas* strains over many literature sources. There are databases that track the pathogenicity of individual strains such as CIRM-CFBP (26), but they suffer from poor interoperability and a lack of provenance. Unification of plant-associated phenotype data into a single high quality resource was therefore considered an essential first step.

In this study, pathogenicity assays were retrieved from 39 recent studies that each took into account both pathogenic and non-pathogenic xanthomonads and unified into a single database. This database was then leveraged to extract available genome sequence which were *de novo* annotated for the presence of Pfam protein domains. These domains were subsequently used as feature input to train three different classifiers. The resulting models were examined for their ability to predict the pathogenicity of an individual strain. Important input features were extracted from these models providing new insights into the genome encoded factors contributing to the plant-associated phenotype. At the same time, these classifiers provide a way to cross-validate pathogenicity test results and conditions.

Results

The workflow is presented in Figure 1. Briefly, pathogenicity assays were obtained from literature and stored in an SQL database. From these assays strains were divided in two classes, available genomes were retrieved and *de novo* annotated for Pfam protein domains. Domain annotations were used as input to create a strain domain matrix as feature input to train three different classifiers: Classification And Regression Tree (CART), Lasso and Random Forest (RF).

Development of the *Xanthomonas* Phenotype Database: To manage literature derived information related to the plant-associated phenotype, an SQL database was created (Figure S1). This database, which we will be referred to as the phenotype database, was used to track

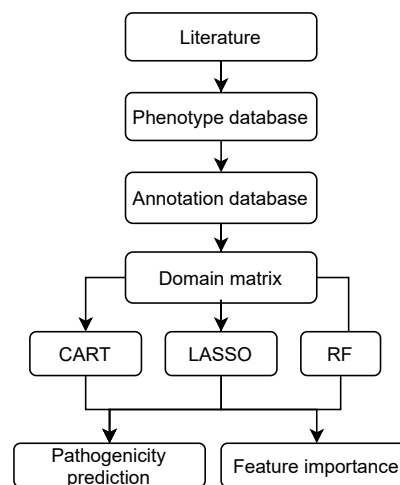


Fig. 1. Workflow. *Xanthomonas* pathogenicity assay data and strains were obtained from literature and stored in a SQL (phenotype) database. Available genomes were retrieved and *de novo* annotated for protein domains. Annotation results were stored in a Graph database. Strain specific domain content was used as input to train the classifiers. Resulting models were examined for their ability to predict pathogenicity and feature importance.

the outcome of individual pathogenicity assays. An assay was defined as the unique combination of a strain and host species as tested by a single source. This approach was favoured over tracking the pathogenicity of individual strains, as it enabled us to also track the criteria used to determine the plant-associated phenotype of a strain. To balance the phenotype database, the data collection effort was limited to recent studies that took both plant-associated phenotypes and their sources into account. This yielded a total of 895 distinct pathogenicity assays, extracted from 39 studies, describing 578 unique strains that were tested on 77 different plant host species (Table 1). Out of the collected 895 individual assays, 503 did and 392 did not indicate strain pathogenicity. From the 578 unique strains, 522 were assayed on their host of isolation (Supplementary File S1).

SQL queries were used to combine results from the various pathogenicity assays. From these combined results the plant-associated phenotype of each individual strain was inferred. Strains unable to induce symptoms on the isolation host after artificial inoculation under optimal conditions were labelled as non-pathogenic (10). Conversely, strains were considered pathogenic if they were able to induce symptoms on any of the tested hosts. Pathogenicity assays based on the "trunk incision" method (31) and pathogenicity assays on *Fragaria ananassa* were both considered to provide insufficient prove for non-pathogenicity, as the former is known to misclassify pathogenic strains that can only cause vertical oozing canker (6, 28) and for the latter the concern was raised that the host might be an unsuitable host to determine pathogenicity of *X. arboricola* strains (13).

Using these criteria, 158 strains were classified as non-pathogenic and 391 strains as pathogenic. Strain names and known aliases of these strains were cross-referenced with the GenBank sequence database (25), to obtain available matching genomes. This resulted in a set 65 pathogenic and 53 non-pathogenic *Xanthomonas* strains with a known genome

Table 1. Overview of the *Xanthomonas* phenotype database

Species	Strains	+	-	Top 5 Most Tested Hosts	Reference
<i>X. arboricola</i>	258	180	232	<i>Juglans regia</i> , <i>Prunus persica</i> , <i>Phaseolus vulgaris</i> , <i>Fragaria ananassa</i> , <i>Capsicum annuum</i>	(6, 8, 10, 13, 14, 21, 27–39)
<i>X. oryzae</i>	70	68	2	<i>Oryza sativa</i>	(5, 9)
<i>X. translucens</i>	58	72	14	<i>Lolium multiflorum</i> , <i>Asparagus virgatus</i> , <i>Hordeum vulgare</i> , <i>Anthurium andreaeanum</i>	(38, 40–42)
<i>X. campestris</i>	40	45	22	<i>Brassica oleracea</i> , <i>Musa acuminata</i> , <i>Saccharum officinarum</i> , <i>Zea mays</i> , <i>Prunus persica</i>	(6, 9, 35, 36, 38, 43, 44)
<i>X. albilineans</i>	27	18	9	<i>Saccharum officinarum</i> , <i>Zea mays</i>	(45–47)
<i>X. axonopodis</i>	20	12	14	<i>Manihot esculenta</i> , <i>Musa acuminata</i> , <i>Saccharum officinarum</i> , <i>Zea mays</i> , <i>Lycopersicon esculentum</i>	(28, 35, 38, 48)
<i>X. dyei</i>	13	12	27	<i>Dysoxylum spectabile</i> , <i>Laurelia novae-zelandiae</i> , <i>Metrosideros excelsa</i>	(39)
<i>X. vasicola</i>	6	12	6	<i>Musa acuminata</i> , <i>Saccharum officinarum</i> , <i>Zea mays</i>	(35)
<i>X. cannabidis</i>	5	5	2	<i>Phaseolus vulgaris</i> , <i>Capsicum annuum</i> , <i>Hordeum vulgare</i>	(22, 24, 44)
<i>X. sontii</i>	5	2	3	<i>Oryza sativa</i>	(5, 49)
<i>X. floridensis</i>	4	0	8	<i>Brassica oleracea</i> , <i>Nasturtium officinale</i>	(50)
<i>X. maliensis</i>	4	0	4	<i>Oryza sativa</i>	(9, 23)
<i>X. fragariae</i>	3	3	0	<i>Fragaria ananassa</i>	(13)
<i>X. euvesicatoria</i>	2	2	0	<i>Lycopersicon esculentum</i> , <i>Capsicum annuum</i>	(6)
<i>X. hortorum</i>	2	0	2	<i>Daucus carota</i>	(38)
<i>X. nasturtii</i>	2	2	2	<i>Brassica oleracea</i> , <i>Nasturtium officinale</i>	(50)
<i>X. pseudoalbilineans</i>	2	1	1	<i>Saccharum officinarum</i>	(46, 51)
<i>X. sacchari</i>	2	0	2	<i>Citrus sinensis</i> , <i>Oryza sativa</i>	(7, 38)
<i>X. hyacinthi</i>	1	2	0	<i>Hyacinthus orientalis</i> , <i>Scilla tubergeniana</i>	(52)
<i>X. melonis</i>	1	0	1	<i>Citrus sinensis</i>	(38)
<i>X. theicola</i>	1	1	0	<i>Camellia sinensis</i>	(53)
<i>X. spp.</i>	52	66	41	<i>Asparagus virgatus</i> , <i>Hordeum vulgare</i> , <i>Oryza sativa</i> , <i>Nicotiana tabacum</i> , <i>Phaseolus vulgaris</i>	(9, 13, 35, 38, 42)
Total	578	503	392	77 Distinct host species	39 Unique references

Strains: number of unique strains assayed; +: Number of positive tests; -: number of negative tests.

sequence encompassing a large majority of the observed genetic variation within this genus, with the exception of the of the *X. hortorum*, *X. gardneri*, *X. citri*, *X. perforans* and *X. vesicatoria* species (Table 2).

Table 2. Sequenced *Xanthomonads* with a known phenotype

Species	Non-Pathogenic	Pathogenic	Total
<i>X. arboricola</i>	28	26	54
<i>X. campestris</i>	2	6	8
<i>X. oryzae</i>	0	7	7
<i>X. translucens</i>	0	7	7
<i>X. cannabidis</i>	2	3	5
<i>X. vasicola</i>	0	5	5
<i>X. sontii</i>	3	1	4
<i>X. albilineans</i>	2	2	4
<i>X. axonopodis</i>	2	2	4
<i>X. sacchari</i>	2	0	2
<i>X. pseudoalbilineans</i>	1	1	2
<i>X. fragariae</i>	0	2	2
<i>X. dyei</i>	1	0	1
<i>X. floridensis</i>	1	0	1
<i>X. maliensis</i>	1	0	1
<i>X. melonis</i>	1	0	1
<i>X. hyacinthi</i>	0	1	1
<i>X. nasturtii</i>	0	1	1
<i>X. theicola</i>	0	1	1
<i>X. spp.</i>	7	0	7
Total	53	65	118

De novo Genome Annotation: Collected genomes were *de novo* annotated for Pfam domains using the SAPP plat-

form (54), which implements Prodigal for gene calling (55) and InterProScan for domain annotation (56). This was done to rule out technical differences due to the use of different annotation software or different versions of the same software. Genome annotations and provenance were stored in a separate database. The statistics are summarised in Table 3. More than 80% of the protein encoding genes code for at least one Pfam domain. The maximum number of genes is inflated by two outlier genomes of low assembly quality, resulting in the prediction of small incomplete genes. However these small genes did not code for protein domains. For annotation details see (Figure S2 and Supplementary File S2).

Table 3. Annotation statistics of the collected genomes.

	Min	Max	Average
Genome Size (bp)	3,534,477	5,333,718	4,773,005
Genes	2955	6207	4155
Domains	4013	5835	5233
Unique Domains	1964	2478	2355

Domain Matrix: For each strain the set of unique protein domains were extracted from the annotation database and combined into binary strain/domain presence matrix which was used for further analysis and model building. The resulting matrix contained 3609 unique Pfam protein domains distributed over the 118 *Xanthomonas* strains

As the ML models will be developed from the present *Xanthomonas* matrix, a "closed" domain representation of genus

is a prerequisite for reliable model building. When a genus is closed, it is assumed that majority of the observed genetic variation within is captured meaning that when new genomes are added, only limited number of new domains will be discovered. The Heaps law estimate was used to estimate the increase in the number of unique domains as a function of the increase in collection size. The decay parameter, alpha, was estimated to be 1.22 indicating that the pan-domainome captured in the matrix was closed (Figure 2). The captured phylogenetic diversity was additionally visualised using a binary domain distance tree (Figure S3 and Supplementary File S2).

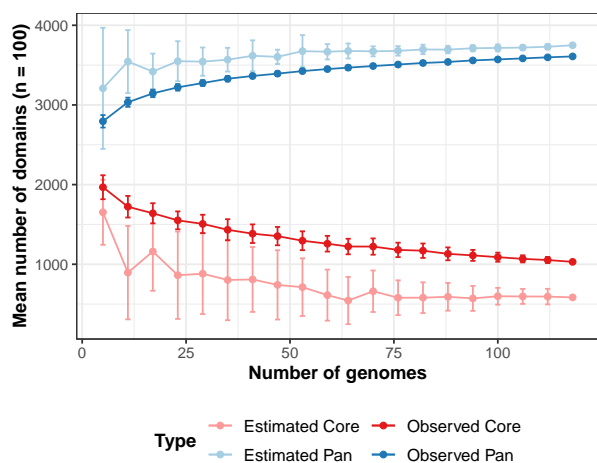


Fig. 2. The *Xanthomonas* Pan-domainome is closed. The decay parameter, alpha, was estimated to be 1.22 (see methods for details).

Matrix optimisation: To reduce the complexity of the data set, the domain matrix was filtered for domains with a low level of variability (present /absent in > 97.5% of samples). These domains either are part of the functional core and thus present in all strains (Figure 2) or represent rare domains containing little information about the general tendencies that discriminate pathogens from non-pathogens. Removal of these domains reduced the total number of domains to 1692. Next, highly correlated domains were treated as one by removal of highly correlated domains with a pair-wise Pearson correlation >0.8. This further reduced the complexity of the matrix, yielding a matrix of 871 unique domains used for training and further analysis.

Machine Learning Approaches: To find and visualise the covariance between the domain content and pathogenicity, Partial Least Squares Discriminant Analysis (PLS-DA) was applied to the data set. PLS-DA suggested that overall, the strain specific domain content provided a good way to discriminate between both classes, but also suggested that some strains might be misclassified. (Figure 3a & 3b).

To learn more about the relationship between the domain content and the plant-associated phenotype, three other types of ML approaches, selected for their high level of interpretability and their performance on data sets with a modest number of observations, were applied: CART, Lasso, and RF.

Model performances: To evaluate the performance of the respective models a 20x5-4x10 nested repeated Cross Validation (CV) was used, based on recommendations of Krstajic *et al.* (57) and Kuhn *et al.* (58). To allow for a better comparison all ML approaches were trained and tested on the identical data partitions. For all approaches, the test-set accuracy was highly variable with a difference in accuracy larger than 0.3 depending on the specific data training- and test-set partition (Figure 3c). This underlines the need to estimate the variation in the model performance using nested CV, if these estimates are to be used as an indication of real world performance. The CART model showed the lowest performance with a median accuracy of 0.750. The Lasso and RF models performed similarly better, with a median accuracy of 0.826 and 0.833 respectively. The RF model performed better on the prediction of non-pathogenic strains, with a sensitivity of 0.769 and specificity of 0.909 (considering pathogens as the positive class), whereas the Lasso's performance was more balanced, but slightly in favour of the pathogens with a sensitivity of 0.846 and specificity of 0.818. The differences between the training- and test-set performance were larger for the Lasso and RF models in comparison to the CART. The median selected tuning parameters indicated that all classifiers favoured relatively low complexity models; $cp = 0.068$ for CART (yielding 2–3 domains per tree), $\lambda = 7.85 \times 10^{-3}$ for Lasso (yielding ~ 29 domains with a non-zero coefficient), and $mtry = 89$ for RF (Supplementary File S3). The median precision-recall (PR) curve (Figure 3d) provided a more detailed representation of model performance. The PR-curves confirmed that the CART model underperforms in comparison to the other models. The drop in precision at low recall values of the CART is caused by the model attributing the highest probability of pathogenicity to two *X. axonopodis* strains that according to literature were non-pathogenic.

Species Specific Prediction Performance: To gain a better understanding of model behaviour, the *in silico* predicted class probabilities were compared with the *in vivo* predictions obtained from literature. To this end the median class-probabilities over the 20 repeats of the nested CV and the pathogenicity labels extracted from the database, were mapped onto the first two components of the previously created PLS (Figure 4).

Strains highlighted in red were predicted to be pathogenic with a high probability. Strains that scored consistently high across all approaches belonged to the *X. oryzae*, *X. vasicola* and *X. fragariae* species or to the *juglandis*, *corylina* and *pruni* pathovars of *X. arboricola*. For these species, with exception of *X. arboricola*, the dataset only contained genomes for strains that were pathogenic according to literature. The same was true for the *X. translucens* species, however only the Lasso and RF model predicted this species with high confidence, whereas the CART model showed mixed predictions for this species. The *X. hyacinthi*, *X. nasturtii* and *X. theicola* species also consisted entirely of pathogenic strains, but these species were predicted to have a lower probability of being pathogenic, which can likely be attributed to being represented with a single sequenced strain. The *X. nasturtii*

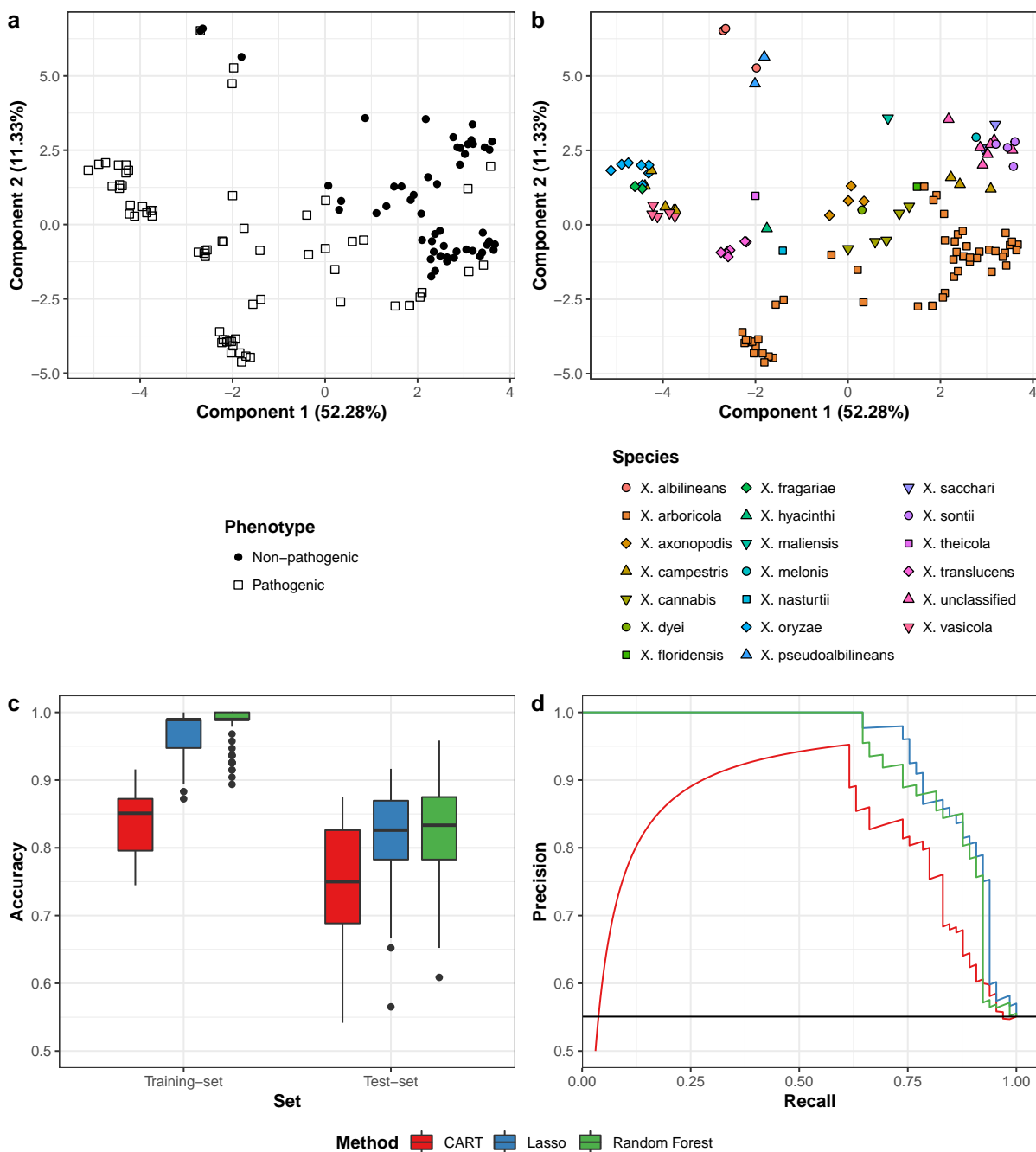


Fig. 3. Properties of the training data set and model prediction performances. Upper panel: Partial Least Squares Discriminant Analysis (a) *Xanthomonas* strains labelled by phenotype. (b) *Xanthomonas* strains labelled according to species classification. Lower panel: Performance of the ML classifiers. (c) Classifier accuracy using a nested cross-validation scheme for each classification method. (d) Precision-Recall curves of each classification method calculated from the median test-set probabilities.

strain was correctly predicted as pathogenic by all ML approaches, whereas the *X. hyacinthi* and *X. theicola* were incorrectly predicted by the median CART and Lasso models.

Strains highlighted in a blue/green colour were predicted to be non-pathogenic with a high probability. These strains formed two distinct clusters: one cluster in the lower-right containing the non-pathogenic *X. arboricola* strains and one cluster in the upper-right, containing strains from the non-pathogenic *X. sontii*, *X. sacchari* and *X. melonis* species and non-pathogenic strains with an undefined species taxonomy. The confidence in the predictions of the non-pathogenic

strains was lower compared to the strains that were consistently predicted as pathogenic, as indicated by a weaker deviation from the neutral score of 0.5. This increased uncertainty likely stems from four strains that are pathogenic according to literature, but are located close to the clusters of non-pathogens.

Strains located near the origin and top-middle of the PLS mainly consisted of pathogenic strains from the remaining *X. arboricola* pathovars and of the species with a small number of genomes with a mixed phenotype (*X. cannabis*, *X. axonopodis* and *X. (pseudo) albilineans*). All ML approaches

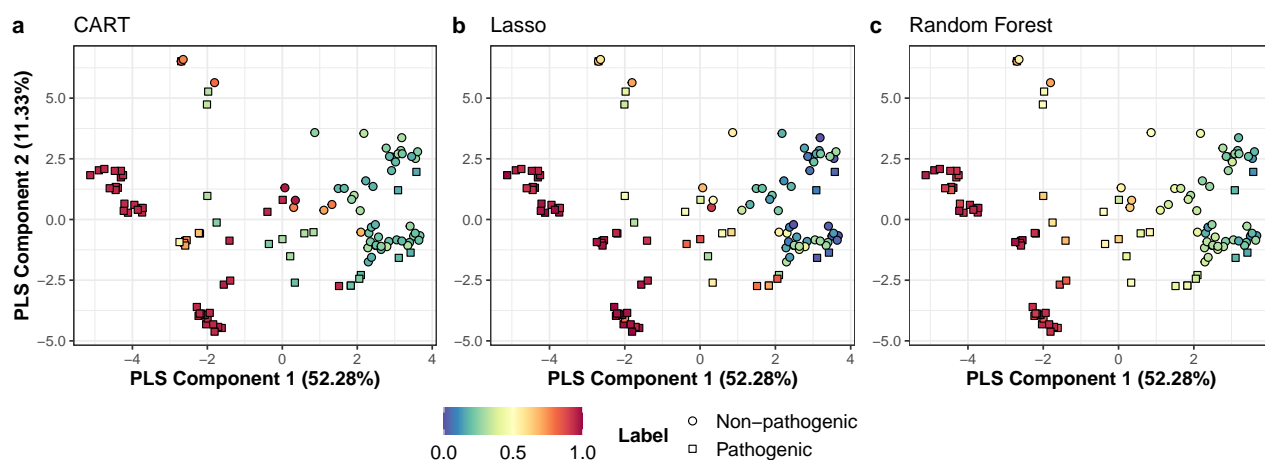


Fig. 4. Median predicted class probabilities mapped onto the PLS-DA. Colour scale (0–1) represents the median predicted probability for pathogenicity. Labels represent the phenotype according to literature. Class probabilities were predicted for the hold out samples in the 5-fold CV outer-loop and the median was taken over the results from the 20 repeats.

were uncertain about the *X. albilineans* species located at the top of the PLS, as indicated by the neutral probabilities. The species near the origin of the PLS showed a large difference between approaches. The CART model failed to correctly predict strains from both the *X. cannabis* and *X. axonopodis* species, the Lasso performed better on the *X. cannabis* species whilst still failing to reliably predict the *X. axonopodis* species and the RF gave neutral predictions for both species. The CART model also performed poorly on the less successful pathovars of the *X. arboricola* species. The Lasso slightly outperformed the RF on these *X. arboricola* pathovars, although the predictions made by the Lasso were more variable between strains.

Feature importance: All three ML approaches apply a form of feature selection, yielding variable importance scores for features and providing insights in the classifier’s decision making process. For each ML approach these scores were obtained in a different way. For CART, variable importance was obtained by tabulating reduction in loss function for all candidate variables considered at each split; for Lasso, variable importance was computed from coefficients using t-statistic; for RF the mean decrease in accuracy when a given variable was left out of bag was used as a measure of variable importance. Scores were summed over all folds of the repeated CV outer-loop and were scaled to have the most important domain at 100. For comparison purposes, domain enrichment was calculated for each class using a two-side Fisher exact test with Benjamini-Hochberg correction. The top 10 most important domains for each of the ML approaches were combined and sorted by enrichment score (Table 4). To provide additional context, the table also includes highly correlated domains removed by matrix filtering. The top 10 domains of CART and RF show a strong correlation, whereas Lasso used more less enriched domains.

Discussion

The *Xanthomonas* genus encompasses a diverse set of species able to infect a large variety of important crops. As non-

pathogenic strains seems to constitute a significant part of the *Xanthomonas* population, from a pest-control point of view the need arises to develop means to reliably distinguish this phenotype, while the genomic diversity may shed light to the *Xanthomonas* plant-associated lifestyles and contributing traits. Currently more than 1700 *Xanthomonas* genomes are available in public repositories, which enables the application of computational approaches for studying the genomic basis for pathogenicity. However, such approaches are hampered by the current publication bias towards pathogenic strains and fragmentation of information regarding the plant associated phenotypes of *Xanthomonas* specific strains. In order to obtain a reliable and balanced set of meta-data representing both plant-associated phenotypes, we collected pathogenicity assays from recent studies that included both plant-associated phenotypes in their study and stored this information in a relational database. Strains unable to induce symptoms on all tested hosts, including the isolation host, were considered to be non-pathogenic. Still, we expected that some mislabelling of the training data was inevitable: First, because the relative abundance of the non-pathogens in nature is unknown and second, due to a high dependency of host susceptibility on the *in vitro* conditions, while there is no complementary test for non-pathogenicity.

Available Genome sequences were retrieved from GenBank and *de novo* annotated for Pfam domains using the SAPP platform (54) A heap analysis of the resulting domain matrix estimated that the pan-domainome was closed, indicating that the data set contained the majority of the variable domains present in the *Xanthomonas* genus and despite some evidence for mislabelling, a partial least squares discriminant analysis indicated that a good discrimination between both phenotypes is possible when a diverse set of domains is considered. To enable inclusion of multiple domains into the decision making process, three different machine learning approaches were explored. CART and Lasso favour the lower-complexity models with the median models using 3 and 29 protein domains respectively. By design, RF used nearly all domains, but the median tuning parameter $m_{try} = 89$ was

Table 4. Top features used in the classifiers.

Top features enriched in pathogens								
Domain	Description	RF	Lasso	CART	P	NP	Enrichment	
	PF13855	Leucine-rich repeat	100.0	74.9	100.0	0.62	0.04	2.14e-08
*	PF09613	Type III secretion system, HrpB1/HrpK	52.1	14.3	75.6	0.83	0.30	3.83e-06
	PF05932	Tir chaperone protein (CesT) family				0.82	0.28	3.96e-06
	PF09483	Type III secretion protein HpaP				0.83	0.30	3.83e-06
	PF09486	Type III secretion protein HrpB7				0.83	0.30	3.83e-06
	PF09487	Type III secretion protein HrpB2				0.83	0.30	3.83e-06
	PF09502	Type III secretion protein HrpB4				0.83	0.30	3.83e-06
	PF05819	NolX				0.69	0.30	3.34e-03
	PF09994	Domain of unknown function DUF2235	19.0	13.0	33.3	0.54	0.09	8.84e-05
	PF13276	HTH-like domain	23.0	8.3	26.8	0.91	0.49	1.82e-04
	PF13333	Integrase, catalytic core	13.4	3.6	12.2	0.51	0.09	2.39e-04
	PF13579	Glycosyltransferase subfamily 4-like	16.0	100.0	4.4	0.88	0.53	2.89e-03
	PF14341	Type 4 fimbrial biogenesis protein PilX	15.2	18.3	4.5	0.85	0.49	4.25e-03
	PF01382	Avidin/streptavidin	6.3	33.6	0.0	0.26	0.04	2.74e-02
	PF10117	5-methylcytosine restriction system component	17.0	77.6	3.8	0.32	0.08	3.30e-02
*	PF12161	N6 adenine-specific DNA methyltransferase	5.5	27.7	0.1	0.88	0.62	4.15e-02
	PF01420	Restriction endonuclease, type I, HsdS				0.85	0.60	5.74e-02
Top features enriched in non-pathogens								
Domain	Description	RF	Lasso	CART	P	NP	Enrichment	
	PF12840	Helix-turn-helix domain	46.7	67.3	70.1	0.25	0.75	1.87e-05
	PF13570	Pyrrolo-quinoline quinone-like domain	12.7	4.3	18.2	0.60	0.98	1.01e-04
*	PF03552	Cellulose synthase	15.5	0.0	25.5	0.35	0.81	1.80e-04
	PF03170	Cellulose synthase BcsB, bacterial				0.37	0.83	1.01e-04
	PF05420	Cellulose synthase operon C, C-terminal				0.38	0.81	4.28e-04
*	PF01270	Glycoside hydrolase, family 8				0.37	0.79	7.33e-04
	PF13424	Tetratricopeptide repeat	16.9	26.0	26.1	0.51	0.91	4.28e-04
	PF12823	Domain of unknown function DUF3817				0.52	0.92	2.85e-04
	PF06629	MltA-interacting MipA	14.5	47.7	4.6	0.54	0.85	1.57e-02
	PF00656	Caspase domain	6.1	0.5	19.6	0.58	0.85	4.45e-02
	PF13391	HNH nuclease	8.5	55.6	1.4	0.08	0.30	5.35e-02
	PF10013	Uncharacterised conserved protein UCP037205	4.0	27.1	0.3	0.31	0.57	8.03e-02

Domain: Pfam accession number; RF: Random Forest scaled variable importance aggregated over all nested CV outer-loop models; Lasso: Lasso aggregated scaled variable importance; CART: CART aggregated scaled variable importance; P: domain persistence in pathogens; NP: domain persistence in non-pathogens; Enrichment: p-value domain enrichment based on a two-sided Fisher exact test with Benjamini-Hochberg multiple testing correction; * left square bracket: Highly correlated domains removed in matrix optimisation.

higher than default ($\sqrt{p} = 30$), indicating that there are only a limited number of important features.

Overall, non-pathogens were classified with a lower level of certainty by all ML approaches. A large part of this uncertainty seemed to stem from four pathogenic strains that showed a strong similarity to many of the non-pathogens according to the PLS-DA. Upon closer examination, the identification of these four strains as pathogens seemed doubtful: *X. sontii* strain ASD011, was the only pathogenic member of a species that has previously been defined by its non-pathogenicity (5). *X. campestris* NCPPB4393, actually belonged to the *X. sacchari* species, a species of which the pathogenicity is still ill defined (59) and this strain in particular was special, as it is the only *Xanthomonas* strain

known to be isolated from an insect host. *X. arboricola* LMG19145 belonged to the *X. arboricola* pv. *fragariae* subspecies, a subspecies with many conflicting pathogenicity reports, that have not been resolved to date (13, 14). *X. arboricola* 3004 was an aberrant strain that didn't cluster with any of the pathovars in the *X. arboricola* species and this strain was only known to be weakly pathogenic to barley. A closely related strain, *X. arboricola* CITA 44, was not able to cause any symptoms on this same host or any other tested host (29). This indicates that these strains are either mislabelled in literature, or belong to a class of very weak opportunistic pathogens. There also remains the possibility that strains identified as non-pathogens by literature symbiotically rely on co-infection with pathogenic species (15).

Given that infection assays typically use pure cultures, such lifestyle would go unnoticed.

Almost half of the tested strains belong to the *X. arboricola* species (Table 2). The predicted class probabilities correlated strongly with varying levels of pathogenicity reported in literature (10). Within this species the strains belonging to the highly pathogenic pathovars (*juglandis*, *corylina* and *pruni* pathovars) or non-pathogenic strains were correctly identified with a high probability. Strains belonging to the weakly pathogenic pathovars obtained more neutral probabilities (Figure 4) which might suggest that, at least for *X. arboricola*, these models can not only combine multiple features to discern phenotypes but are also able to score different levels of pathogenicity.

Important features enriched in the pathogens: Extraction and examination of the most important features used for prediction by the respective ML approaches, showed that the CART and RF considered mostly the same domains, with both favouring domains that were highly enriched in one of the classes. The Lasso behaved distinctly different with its variable importance scores showing a weaker correlation with the enrichment analysis.

Many of the highly important domains that were enriched in pathogens have already been related to pathogenicity in literature, indicating that the here used approach is capable of detecting biologically relevant features. Overall, the most important feature across models was a leucine-rich repeat (LRR) domain present in two *Xanthomonas* type III effector proteins (XopL and XopAE/HpaF). The domain was present in nearly two-thirds of the pathogens and absent in the non-pathogenic strains, with exception of two potentially mislabelled *X. axonopodis* strains. Similarly, a large group of correlated domains representing the Type III Secretion System (T3SS) was also found to be highly important, although these domains had a much broader representation amongst both phenotypes. The T3SS and its effectors are the most important and extensively studied aspect of *Xanthomonas* pathogenicity (16). The entire T3SS is encoded by the *hrp* cluster and consists of more than 20 proteins that form a needle-like syringe used to inject proteins into the host cytoplasm (60). Type III effector proteins are translocated into host cells where they target many host components, serving to suppress the host immune system, increase nutrient availability and facilitate the infection process (17). Specifically, mutations in HpaF were found to impact virulence in *X. axonopodis* (61).

Amongst the important predictors enriched in pathogens were two domains related to mobile genetic elements. The first, PF13276, was a helix-turn-helix-like domain that was found in several IS3 transposases and the second, PF13333, was an integrase domain that belonged to a putative OrfB transposase. Whilst both domains were enriched in pathogens, the helix-turn-helix-like domain was also well represented in the non-pathogens, whereas the integrase domain was only found in a few non-pathogens and was more scattered across the pathogens. Transposases are known to flank pathogenicity islands in genomes of *Xanthomonas* (62).

However, these domains are more commonly found within pathogenic islands, where, in some cases, they are assumed to have played a role in the initial mobilisation or subsequent rearrangement of the element (63).

The most important domain for the Lasso model was PF13579, a n-terminal glycosyl transferase 4-like domain of the RfaB family. This domain is most likely involved in LPS production, which is important for pathogenicity by providing a barrier against anti-microbial compounds, facilitating adhesion and preventing host recognition (15). Although the domain was enriched in pathogens, it was found pathogens and non-pathogens of all major species, with the exception of *X. albilineans*.

Next to domains that were already known to be involved in pathogenicity, a domain of unknown function (DUF2235), enriched in pathogens, was also found to be an important predictor. The domain, which represents a further uncharacterized alpha/beta hydrolase, was present in up to 16 proteins per genome (7 on average). Further analysis revealed that in one cluster this domain was fused with domain PF16014 (histone deacetylase complex subunit SAPI30 C-terminus domain). This domain is usually found to be part of transcriptional repressor Sin3 and the region containing this domain was also flagged as a superantigen-like protein SSL3 motif. The SSL3 protein is important for pathogenicity in the human pathogen *S. aureus*, where it is known to bind to the hosts Toll-Like Receptor 2 (TLR2), inhibiting stimulation by its ligands (64) suggesting an important role for suppressing the plant hosts immune system.

Important features enriched in the non-pathogens:

Domain enriched in non-pathogens or notably absent from pathogens were also highly important for model predictions. Many of these domains have an implied role in increasing resistance against environmental factors, which is in line with the idea that non-pathogenic strains are generalists that can survive in a much broader range of conditions than their pathogenic counterparts (11, 12). All methods agree that another Helix-Turn-Helix domain (PF12840) is an the important discriminant enriched in non-pathogens. This domain was found in all non-pathogens with exception for a subgroup of *X. arboricola* and in some pathogens of the *X. oryzae* and *X. campestris* species. The domain is found in DNA binding transcriptional regulators of the ArsR/SmtB, Arsenical Resistance Operon Repressor, family. In *X. campestris* 8004 it was found that the HTH ArsR containing gene is upstream of arsenite efflux pump Acr3 and a putative high-affinity Fe^{2+}/Pb^{2+} permease. In the same strain it was shown, via a knockout, that the *arsR* gene confers an increased resistance against arsenate (65).

Amongst the important predictors enriched in non-pathogens, was a group of correlated domains involved in cellulose synthesis. The model organism for studying bacterial cellulose synthase is *Acetobacter xylinum*. For this organism it is known that the complex produces and transports beta-1,4-glucan chains, creating rigid crystalline structures on the outer membrane. Bacterial cellulose can fulfil diverse roles from mechanical/environmental protection to cell adhesion

during symbiotic or pathogenic interactions (66). Cellulose synthesis is promoted by cyclic-di-GMP through the PilZ domain present in glycosyltransferase CeSA, which is part of the membrane-bound cellulose synthase complex (67). Cyclic di-GMP is also known to down-regulate biofilm formation, EPS production, extracellular enzyme production and *hrp* gene expression in *Xanthomonas* (15). Thus given that cellulose synthesis and production of pathogenicity factors, is inversely related we hypothesise that these domains might be involved in providing environmental protection when the bacterium is not shielded by the host homeostasis. The last domain that could be linked to environmental resistance was the MltA-interacting MipA domain (PF06629) was found in all *Xanthomonas* species, but had a lower presence in the highly pathogenic *X. arboricola* pathovars. MipA is a protein that mediates assembly of MltA into the PBP1B murin transglycosylase/transpeptidase complex. Mutations in other genes of the *mlt* family are related with morphological abnormalities in *X. campestris* (68). Given that the domain is widely present in both pathogenic and non-pathogenic strains, it seems unlikely that the domain is central to non-pathogenicity. However, the morphological changes induced by loss of genes containing this domain, could impair the bacterium's ability to resist mechanical stress.

Conclusion

The plant-associated phenotype of a *Xanthomonas* strain is the result of a multitude of non-persistent traits. Consequently a single genome encoded feature will have limited power to predict the plant-associated phenotype. By using machine learning methods that take into account an ensemble of domains, a better prediction can be obtained. Through recursive feature elimination, key domains related to plant-associated phenotype could be identified leading to the discovery of novel traits.

Methods

Data processing: Data processing, analysis and model building was done in *R* (v4.0.5). SQL was used to communicate with the phenotype database and SPARQL was used to communicate with the Graph annotation database. All *R* scripts were executed on a Windows 10 machine.

Phenotype Database: Pathogenicity assays of individual *Xanthomonas* strains were mined from literature that also considered non-pathogenicity. Relevant parameters and outcomes were stored in a relational database MySQL database (*MariaDB* v10.5.9). To create the database, the database model was forward-engineered into the *create_database.sql* script using *MySQL Workbench* (v8.0.24). A manually curated Excel form (*Supplementary File S1*) was used to fill the database and subsequently loaded into the database using the *input_data.R* script. Connection to the database was established using the base *R* *DBI* library and the *RMariaDB* (v1.1.0) driver. The script was also used to enforce additional constraints on the values of specific fields.

Data Retrieval: The *genbank.R* script was used to interface with the phenotype database. For the strains with a known pathogenicity, genomes were retrieved directly from the *NCBI Genbank* genome repository (accessed May 6th, 2020) using *RCurl* (v1.98-1.2). If multiple genomes were available for a single strain, the genome with the highest quality of assembly was taken.

Genome Annotation: Genomes were de novo annotated using the *SAPP* framework (54), running on a Linux machine (openSUSE leap 15.1) with *OpenJDK* v11.0.5. The retrieved genomes were converted to a HDT format using the *Conversion* module from *SAPP* (69). Protein encoding genes were identified using *Prodigal* (55) and annotated for Pfam protein domains (70) using *InterproScan* (v5.44-79.0) (56). Fully annotated genomes and their provenance, were uploaded to a linked data repository using *GraphDB* (v9.7.0) for further analysis. To speed up computation *SAPP* modules were run in parallel using the *GNU Parallel CLI* (71).

Data Analysis and Visualization: Annotation results were retrieved from the linked data repository using SPARQL queries and the *SPARQL* *R* package (v1.16). The binary Pfam domain presence/absence matrix was generated with (*sparql.R*). Dendrogram: Distances were calculated using the base *R* *dist* function with a Manhattan distance measure and hierarchical clustering was performed using the base *R* *hclust* function with average linkage. The *ape* (v5.4-1) *R* package was used to root the tree. The resulting dendrogram was visualised using the *dendextend* package (v1.14.0). The enrichment of single domains between and the two phenotypes was tested using a two-sided Fisher exact test with Benjamini-Hochberg multiple testing correction, using the *R* base *fisher.test* and *p.adjust*.

Heap analysis: The *micropan* (v2.1) package (72) was used. The effect of sample size on the estimated/observed pan and core domain sizes was explored by repeatedly ($n = 100$) sampling a fixed number of genomes using 20 different sample sizes equally distributed over the range spanning from 5 to 118, with 118 being the total number of genomes in this study. For each sample, the observed pan and core sizes were inferred directly from the data and the estimated pan and core sizes were obtained by fitting a binomial mixture model (with k ranging from $k = 3$ to $k = 17$) to the selected subset and taking the estimate with the lowest BIC.

Matrix optimisation: Domains with near-zero variance were removed using a threshold of present or absent in >97.5 genomes. For highly correlated domains a representing domain was chosen based on a Pearson correlation of $\rho > 0.8$.

Partial Least Squares Discriminant Analysis: A Partial Least Squares (PLS) Discriminant Analysis was performed by training a two-component PLS regression model on domain matrix using the "pls" (v2.7-3) *R* package. The first two-components were then extracted and visualised.

Model development: Model tuning and testing (*model_building.R*) was performed using the *Resample-Model* function, contained in custom a R library created for this analysis. This library provided an interface between the *rsample* package (v0.0.9), used for data partitioning, and the *Caret* package (v6.0-86), used for model building and tuning (73). The filtered pathogenicity dataset was partitioned using a nested Cross-Validation (CV) scheme consisting of a 20-times repeated 5-fold outer-loop and a 4-times repeated 10-fold inner-loop. Accuracy was used as the performance metric in both loops and additional statistics were generated using *Caret*'s *twoClassSummary* function. When calculating binary classification metrics pathogenicity was regarded as a 'positive' result and non-pathogenicity as a 'negative' result.

Using identical partitions, different types of models were built and tested: CART models from the *rpart* package (74), Lasso models from the *glmnet* package (by setting $\alpha = 1$) (75), and Random Forest models from the *RandomForest* package (76). For the CART models, the complexity parameter *cp* was varied across *Caret*'s default grid, with a size of 9. For the lasso model the parameter λ was optimised using a grid ranging from $1 * 10^{-4}$ to 1 with a size of 20 and an exponentially increasing step size. For the random forest model, the number of trees was fixed at 1000 and *mtry* was varied over *Caret*'s default grid of size 9.

Model Performance & Variable Importance: The performance of the respective approaches on the different genomes was examined by superimposing the median probability for a strain to be pathogenic over the previously created PLS-DA visualisation. For each approach, the variable importance scores for per domain were calculated using *Caret*'s *build-in varImp* function with scaling turned off. Results from multiple folds were combined by summing the unscaled variable importances for each domain, after which the final results were scaled to have the most important variable at 100.

Supporting information:

The code and SQL database are available at: <https://gitlab.com/wurssb/xanthomonas-phenotype-prediction>.

Genome annotations in a binary RDF format (HDT) is available at <http://doi.org/10.4121/14546625>.

Supplementary Figure S1 (PDF): Entity-Relationship diagram of the *Xanthomonas* phenotype database.

Supplementary Figure S2 (PDF): Annotation statistics of 118 sequenced *Xanthomonas* strains used in this study.

Supplementary Figure S3 (PDF): Domain based distance tree of the 118 *Xanthomonas* strains used in this study.

Supplementary File S1 (XLSX): Input and output of

the SQL database.

Supplementary File S2 (XLSX): Genome metadata, annotation statistics and the domain matrix.

Supplementary File S3 (XLSX): Model statistics and domain enrichment.

Funding

PS and JK acknowledge the Dutch national funding agency NWO, and Wageningen University and Research for their financial contribution to the Unlock initiative (NWO: 184.035.007). WP is financially supported by a Royal Thai Government Scholarship, Thailand.

Author contributions

D.M., W.P., P.S., J.K. participated in the conception and design of the study. D.M. was responsible for the code and design of the database. D.M., W.P., P.S., J.K. wrote sections of the manuscript. All authors critically revised the manuscript.

References

1. Marie-Agnès Jacques, Matthieu Arlat, Alice Boulanger, Tristan Boureau, Sébastien Carrère, Sophie Cesbron, Nicolas WG Chen, Stéphane Cociancich, Armelle Darrasse, Nicolas Denancé, et al. Using ecology, physiology, and genomics to understand host specificity in *xanthomonas*. *Annual review of phytopathology*, 54:163–187, 2016.
2. Robert P Ryan, Frank-Jörg Vorhölter, Neha Potnis, Jeffrey B Jones, Marie-Anne Van Sluys, Adam J Bogdanove, and J Maxwell Dow. Pathogenomics of *xanthomonas*: understanding bacterium–plant interactions. *Nature Reviews Microbiology*, 9(5):344, 2011.
3. JL Maas, MM Finney, EL Civerolo, and M Sasser. Association of an unusual strain of *xanthomonas campestris* with apple. *Phytopathology*, 75(4):438–445, 1985.
4. Tristan Boureau, Mohammed Kerkoud, Fabien Chel, Gilles Hunault, Armelle Darrasse, Chrystelle Brin, Karine Durand, Ahmed Hajri, Stéphane Poussier, Charles Manceau, et al. A multiplex-pcr assay for identification of the quarantine plant pathogen *xanthomonas axonopodis* pv. *phaseoli*. *Journal of microbiological methods*, 92(1):42–50, 2013.
5. Kanika Bansal, Amandeep Kaur, Samriti Midha, Sanjeet Kumar, Suresh Korpole, and Prabhu B Patil. *Xanthomonas sontii* sp. nov., a non-pathogenic bacterium isolated from healthy basmati rice (*oryza sativa*) seeds from india. *bioRxiv*, page 738047, 2019.
6. Salwa Essakhi, Sophie Cesbron, Marion Fischer-Le Saux, Sophie Bonneau, Marie-Agnès Jacques, and Charles Manceau. Phylogenetic and variable-number tandem-repeat analyses identify nonpathogenic *xanthomonas arboricola* lineages lacking the canonical type iii secretion system. *Appl. Environ. Microbiol.*, 81(16):5395–5410, 2015.
7. Yunxia Fang, Haiyan Lin, Liwen Wu, Deyong Ren, Weijun Ye, Guojun Dong, Li Zhu, and Longbiao Guo. Genome sequence of *xanthomonas sacchari* r1, a biocontrol bacterium isolated from the rice seed. *Journal of biotechnology*, 206:77–78, 2015.
8. Jerson Garita-Cambroner, Ana Palacio-Bielsa, María M López, and Jaime Cubero. Pan-genomic analysis permits differentiation of virulent and non-virulent strains of *xanthomonas arboricola* that cohabit prunus spp. and elucidate bacterial virulence factors. *Frontiers in microbiology*, 8:573, 2017.
9. Issa Wonni, Bart Cottyn, Liselot Detemmerman, S Dao, L Ouedraogo, S Sarra, C Tekete, S Poussier, R Corral, L Triplett, et al. Analysis of *xanthomonas oryzae* pv. *oryzicola* population in mali and burkina faso reveals a high level of genetic and pathogenic diversity. *Phytopathology*, 104(5):520–531, 2014.
10. Déborah Merda, Sophie Bonneau, Jean-François Guimbaud, Karine Durand, Chrystelle Brin, Tristan Boureau, Christophe Lemaire, Marie-Agnès Jacques, and Marion Fischer-Le Saux. Recombination-prone bacterial strains form a reservoir from which epidemic clones emerge in agroecosystems. *Environmental microbiology reports*, 8(5):572–581, 2016.
11. Déborah Merda, Martial Briand, Eran Bosis, Céline Rousseau, Perrine Portier, Matthieu Barret, Marie-Agnès Jacques, and Marion Fischer-Le Saux. Ancestral acquisitions, gene flow and multiple evolutionary trajectories of the type three secretion system and effectors in *xanthomonas* plant pathogens. *Molecular ecology*, 26(21):5939–5952, 2017.
12. Jun Zhang, Caiwen Zhang, Jing Yang, Ruijie Zhang, Jusheng Gao, Xia Zhao, Juanjuan Zhao, Dongfang Zhao, and Xiaoxia Zhang. Insights into endophytic bacterial community structures of seeds among various *oryza sativa* l. rice genotypes. *Journal of Plant Growth Regulation*, 38(1):93–102, 2019.
13. Joachim Vandroemme, Bart Cottyn, Joël F Pothier, Valentin Pflüger, Brion Duffy, and Martine Maes. *Xanthomonas arboricola* pv. *fragariae*: what's in a name? *Plant Pathology*, 62(5):1123–1131, 2013.
14. Patrizia Ferrante and Marco Scortichini. *Xanthomonas arboricola* pv. *fragariae*: a confirmation of the pathogenicity of the pathotype strain. *European journal of plant pathology*, 150(3):825–829, 2018.

15. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörg Vorhölter, Yong-qiang He, Anke Becker, Doron Teper, Yi Li, Nian Wang, Leonidas Bleris, et al. Mechanistic insights into host adaptation, virulence and epidemiology of the phytopathogen *Xanthomonas*. *FEMS Microbiology Reviews*, 2019.
16. Sujan Timilsina, Neha Potnis, Eric A Newberry, Prabha Liyanapathirane, Fernanda Iruegas-Bocardo, Frank F White, Erica M Goss, and Jeffrey B Jones. *Xanthomonas* diversity, virulence and plant-pathogen interactions. *Nature Reviews Microbiology*, pages 1–13, 2020.
17. Sabine Kay and Ulla Bonas. How *xanthomonas* type iii effectors manipulate the host plant. *Current opinion in microbiology*, 12(1):37–43, 2009.
18. Luciano Ariel Rigano, Caroline Payette, Geneviève Brouillard, Maria Rosa Marano, Laura Abramowicz, Pablo Sebastián Torres, Maximina Yun, Atilio Pedro Castagnaro, Mohamed El Oirdi, Vanessa Dufour, et al. Bacterial cyclic β -(1,2)-glucan acts in systemic suppression of plant immune responses. *The Plant Cell*, 19(6):2077–2089, 2007.
19. Adrian A Vojnov, Holly Slater, Michael J Daniels, and J Maxwell Dow. Expression of the gum operon directing xanthan biosynthesis in *xanthomonas campestris* and its regulation in planta. *Molecular plant-microbe interactions*, 14(6):768–774, 2001.
20. Lian-Hui Wang, Yawen He, Yunfeng Gao, Ji En Wu, Yi-Hu Dong, Chaozu He, Su Xing Wang, Li-Xing Weng, Jin-Ling Xu, Leng Tay, et al. A bacterial cell-cell communication signal with cross-kingdom structural analogues. *Molecular microbiology*, 51(3):903–912, 2004.
21. Sophie Cesbron, Martial Briand, Salwa Essakhi, Sophie Gironde, Tristan Boureau, Charles Manceau, Fischer-Le Saux, Marie-Agnès Jacques, et al. Comparative genomics of pathogenic and nonpathogenic strains of *xanthomonas arboricola* unveil molecular and evolutionary events linked to pathoadaptation. *Frontiers in plant science*, 6:1126, 2015.
22. Jonathan M Jacobs, Céline Pesce, Pierre Lefeuvre, and Ralf Koebnik. Comparative genomics of a cannabis pathogen reveals insight into the evolution of pathogenicity in *xanthomonas*. *Frontiers in plant science*, 6:431, 2015.
23. Lindsay R Triplett, Valérie Verdier, Tony Campillo, Cinzia Van Malderghem, Ilse Cleenwerck, Martine Maes, Loïc Deblais, Rene Corral, Ousmane Koita, Bart Cottyn, et al. Characterization of a novel clade of *xanthomonas* isolated from rice leaves in mali and proposal of *xanthomonas maliensis* sp. nov. *Antonie Van Leeuwenhoek*, 107(4):869–881, 2015.
24. Valente Aritua, Augustine Musoni, Alice Kabeja, Louis Butare, Floride Mukamuhirwa, Daphrose Gahakwa, Fred Kato, Mthec M Abang, Robin Buruchara, Melanie Sapp, et al. The draft genome sequence of *xanthomonas* species strain nyagatare, isolated from diseased bean in rwanda. *FEMS Microbiology*, 2015.
25. Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Conrad L Schoch, Stephen T Sherry, and Ilene Karsch-Mizrachi. Genbank. *Nucleic acids research*, 49(D1):D92–D96, 2021.
26. Cfbp - plant associated bacteria, Apr 2021.
27. Jorge Ibarra Caballero, Marcelo M Zerillo, Jacob Snelling, Christina Boucher, and Ned Tisserat. Genome sequence of *xanthomonas arboricola* pv. *corylina*, isolated from turkish filbert in colorado. *Genome Announc.*, 1(3):e00246–13, 2013.
28. Marion Fischer-Le Saux, Sophie Bonneau, Salwa Essakhi, Charles Manceau, and Marie-Agnès Jacques. Aggressive emerging pathogens of *xanthomonas arboricola* represent widespread epidemic clones distinct from poorly pathogenic strains, as revealed by multilocus sequence typing. *Appl. Environ. Microbiol.*, 81(14):4651–4668, 2015.
29. Jerson Garita-Cambronero, Ana Palacio-Bielsa, Maria M Lopez, and Jaime Cubero. Comparative genomic and phenotypic characterization of pathogenic and non-pathogenic strains of *xanthomonas arboricola* reveals insights into the infection process of bacterial spot disease of stone fruits. *PLoS one*, 11(8), 2016.
30. Jerson Garita-Cambronero, Ana Palacio-Bielsa, and Jaime Cubero. *Xanthomonas arboricola* pv. *pruni*, causal agent of bacterial spot of stone fruits and almond: its genomic and phenotypic characteristics in the *x. arboricola* species context. *Molecular plant pathology*, 19(9):2053–2065, 2018.
31. Ahmed Hajri, Damien Meyer, F Delort, J Guillaumes, Christelle Brin, and Charles Manceau. Identification of a genetic lineage within *xanthomonas arboricola* pv. *juglandis* as the causal agent of vertical oozing canker of persian (english) walnut in france. *Plant pathology*, 59(6):1014–1022, 2010.
32. James Harrison, Murray R Grant, and David J Studholme. Draft genome sequences of two strains of *xanthomonas arboricola* pv. *celebensis* isolated from banana plants. *Genome Announc.*, 4(1):e01705–15, 2016.
33. Gastón Higuera, Narjol González-Escalona, Camila Véliz, Francisca Vera, and Jaime Romero. Draft genome sequences of four *xanthomonas arboricola* pv. *juglandis* strains associated with walnut blight in chile. *Genome Announc.*, 3(5):e01160–15, 2015.
34. Alexander N Ignatov, Elena I Kyrova, Svetlana V Vinogradova, Anastasia M Kamionskaya, Norman W Schaad, and Douglas G Luster. Draft genome sequence of *xanthomonas arboricola* strain 3004, a causal agent of bacterial disease on barley. *Genome Announc.*, 3(1):e01572–14, 2015.
35. G Karamura, Julian Smith, David Studholme, Jerome Kubiriba, and E Karamura. Comparative pathogenicity studies of the *xanthomonas vasicola* species on maize, sugarcane and banana. *Afr. J. Plant Sci*, 9:385–400, 2015.
36. Akira Kawaguchi, Koji Inoue, and Yasuhiro Inoue. Biological control of bacterial spot on peach by nonpathogenic *xanthomonas campestris* strains az98101 and az98106. *Journal of general plant pathology*, 80(2):158–163, 2014.
37. Ulisses P Pereira, Hossein Gouran, Rafael Nascimento, James E Adaskaveg, Luiz Ricardo Goulart, and Abhaya M Dandekar. Complete genome sequence of *xanthomonas arboricola* pv. *juglandis* 417, a copper-resistant strain isolated from juglans regia l. *Genome Announc.*, 3(5):e01126–15, 2015.
38. Luc Vauterin, Ping Yang, Anne Alvarez, Yuichi Takikawa, Don A Roth, Anne K Vidaver, Robert E Stall, Karel Kersters, and Jean Swings. Identification of non-pathogenic *xanthomonas* strains associated with plants. *Systematic and applied microbiology*, 19(1):96–105, 1996.
39. JM Young, JP Wilkie, D-C Park, and DRW Watson. New zealand strains of plant pathogenic bacteria classified by multi-locus sequence analysis; proposal of *xanthomonas dyei* sp. nov. *Plant pathology*, 59(2):270–281, 2010.
40. Lena Hersemann, Daniel Wibberg, Jochen Blom, Alexander Goesmann, Franco Widmer, Frank-Jörg Vorhölter, and Roland Kölliker. Comparative genomics of host adaptive traits in *xanthomonas translucens* pv. *graminis*. *BMC genomics*, 18(1):35, 2017.
41. Roland Kölliker, Rolf Kraehenbuehl, Beat Boller, and Franco Widmer. Genetic diversity and pathogenicity of the grass pathogen *xanthomonas translucens* pv. *graminis*. *Systematic and applied microbiology*, 29(2):109–119, 2006.
42. JLW Rademaker, DJ Norman, RL Forster, FJ Louws, MH Schultz, and FJ De Buijn. Classification and identification of *xanthomonas translucens* isolates, including those pathogenic to ornamental asparagus. *Phytopathology*, 96(8):876–884, 2006.
43. Yung-An Lee, Pei-Yu Yang, and Shau-Chang Huang. Characterization, phylogeny, and genome analyses of nonpathogenic *xanthomonas campestris* strains isolated from brassic seeds. *Phytopathology*, pages PHYTO-08, 2020.
44. Valérian Meline, Wesley Delage, Chrystelle Brin, Camille Li-Marchetti, Daniel Sochard, Matthieu Arlat, Céline Rousseau, Armelle Darrasse, Martial Briand, Guillaume Lebreton, et al. Role of the acquisition of a type 3 secretion system in the emergence of novel pathogenic strains of *xanthomonas*. *Molecular plant pathology*, 20(1):33–50, 2019.
45. Patrice Champoiseau, J-H Daugrois, Isabelle Pieretti, Stéphane Cociancich, Monique Royer, and Philippe Rott. High variation in pathogenicity of genetically closely related strains of *xanthomonas albilineans*, the sugarcane leaf scald pathogen, in guadeloupe. *Phytopathology*, 96(10):1081–1091, 2006.
46. Jean-Heinrich Daugrois, V Dumont, Patrice Champoiseau, Laurent Costet, Rosiane Boisne-Noc, and Philippe Rott. Aerial contamination of sugarcane in guadeloupe by two strains of *xanthomonas albilineans*. *European journal of plant pathology*, 109(5):445–458, 2003.
47. Philippe C Rott, Laurent Costet, Michael J Davis, Roger Frutos, and Dean W Gabriel. At least two separate gene clusters are involved in albicidin production by *xanthomonas albilineans*. *Journal of bacteriology*, 178(15):4590–4596, 1996.
48. Carolina Gonzalez, Silvia Restrepo, Joe Tohme, and Valérie Verdier. Characterization of pathogenic and nonpathogenic strains of *xanthomonas axonopodis* pv. *manihotis* by pcr-based dna fingerprinting techniques. *FEMS microbiology letters*, 215(1):23–31, 2002.
49. S Ali Mirghasempour, Shiweng Huang, David J Studholme, and Carrie L Brady. A grain rot of rice in iran caused by a *xanthomonas* strain closely related to *x. sacchari*. *Plant Disease*, pages PDIS-01, 2020.
50. Joana G Vicente, Steve Rothwell, Eric B Holub, and David J Studholme. Pathogenic, phenotypic and molecular characterisation of *xanthomonas nasturtii* sp. nov. and *xanthomonas floridensis* sp. nov., new species of *xanthomonas* associated with watercress production in florida. *International journal of systematic and evolutionary microbiology*, 67(9):3645–3654, 2017.
51. G Orian. A disease of paspalum dilatatum in mauritius caused by a species of bacterium closely resembling *xanthomonas albilineans* (ashby) dowson. *Rev. Agric. Sucr. Ile Maurice*, 41:7–20, 1962.
52. JD Janse and HJ Miller. Yellow disease in scilla tubergeniana and related bulbs caused by *xanthomonas campestris* pv. *hyacinthii*. *Netherlands Journal of Plant Pathology*, 89(5):203–206, 1983.
53. K Uehara, K Arai, T Nonaka, I Sano, et al. Canker of tea, a new disease, and its causal bacterium *xanthomonas campestris* pv. *theaeicola* uehara & arai pv. nov. *Bulletin of the Faculty of Agriculture Kagoshima University*, 30:17–21, 1980.
54. Jasper J Koehorst, Jesse CJ van Dam, Edoardo Saccenti, Vitor AP Martins dos Santos, Maria Suarez-Diez, and Peter J Schaap. Sapp: functional genome annotation and analysis through a semantic framework using fair principles. *Bioinformatics*, 34(8):1401–1403, 2017.
55. Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):119, 2010.
56. Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.
57. Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):1–15, 2014.
58. Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.
59. David J Studholme, Arthur Wasukira, Konrad Paszkiewicz, Valente Aritua, Richard Thwaites, Julian Smith, and Murray Grant. Draft genome sequences of *xanthomonas sacchari* and two banana-associated *xanthomonas* reveal insights into the *xanthomonas* group 1 clade. *Genes*, 2(4):1050–1065, 2011.
60. Herbert Schmidt and Michael Hensel. Pathogenicity islands in bacterial pathogenesis. *Clinical microbiology reviews*, 17(1):14–56, 2004.
61. Jung-Gun Kim, Byoung Keun Park, Chang-Hyuk Yoo, Eunkyung Jeon, Jonghee Oh, and Ingyu Hwang. Characterization of the *xanthomonas axonopodis* pv. *glycines* hrp pathogenicity island. *Journal of Bacteriology*, 185(10):3155–3166, 2003.
62. Claudia B Monteiro-Vitorello, Mariana C De Oliveira, Marcelo M Zerillo, Alessandro M Varani, Edwin Civerolo, and Marie-Anne Van Sluys. Xylella and *xanthomonas mobil* omics. *Omic: a journal of integrative biology*, 9(2):146–159, 2005.
63. Jörg Hacker and James B Kaper. Pathogenicity islands and the evolution of microbes. *Annual Reviews in Microbiology*, 54(1):641–679, 2000.
64. Ryoosuke Yokoyama, Saotomo Itoh, Go Kamoshida, Takemasa Takii, Satoshi Fujii, Tsutomu Tsuji, and Kikuo Onozaki. Staphylococcal superantigen-like protein 3 binds to the toll-like receptor 2 extracellular domain and inhibits cytokine production induced by staphylococcus aureus, cell wall component, or lipopeptides in murine macrophages. *Infection and immunity*, 80(8):2816–2825, 2012.
65. Lian Zhou, Frank-Jörg Vorhölter, Yong-Qiang He, Bo-Le Jiang, Ji-Liang Tang, Yuquan Xu, Alfred Pühler, and Ya-Wen He. Gene discovery by genome-wide cds re-prediction and microarray-based transcriptional analysis in phytopathogen *xanthomonas campestris*. *BMC genomics*, 12(1):359, 2011.
66. Peter Ross, Raphael Mayer, and Moshe Benzman. Cellulose biosynthesis and function in bacteria. *Microbiology and Molecular Biology Reviews*, 55(1):35–58, 1991.
67. Takaaki Fujiwara, Keisuke Komoda, Naofumi Sakurai, Kenji Tajima, Isao Tanaka, and Min Yao. The c-di-gmp recognition mechanism of the pilz domain of bacterial cellulose synthase subunit a. *Biochemical and Biophysical Research Communications*, 431(4):802–807, 2013.

68. Lin Wang, Li-Yan Yang, Yong-Liang Gan, Feng Yang, Xue-Lian Liang, Wan-Lian Li, and Jiang Bo-Le. Two lytic transglycosylases of *Xanthomonas campestris* pv. *campestris* associated with cell separation and type iii secretion system, respectively. *FEMS microbiology letters*, 366(7):fnz073, 2019.
69. Javier D Fernández, Miguel A Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary rdf representation for publication and exchange (hdt). *Journal of Web Semantics*, 19:22–41, 2013.
70. Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2018.
71. Ole Tange. *Gnu parallel 2018*. Lulu. com, 2018.
72. Lars Snipen and Kristian Hovde Liland. micropan: an r-package for microbial pan-genomics. *BMC bioinformatics*, 16(1):79, 2015.
73. Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, 2015.
74. Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.
75. Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.
76. Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

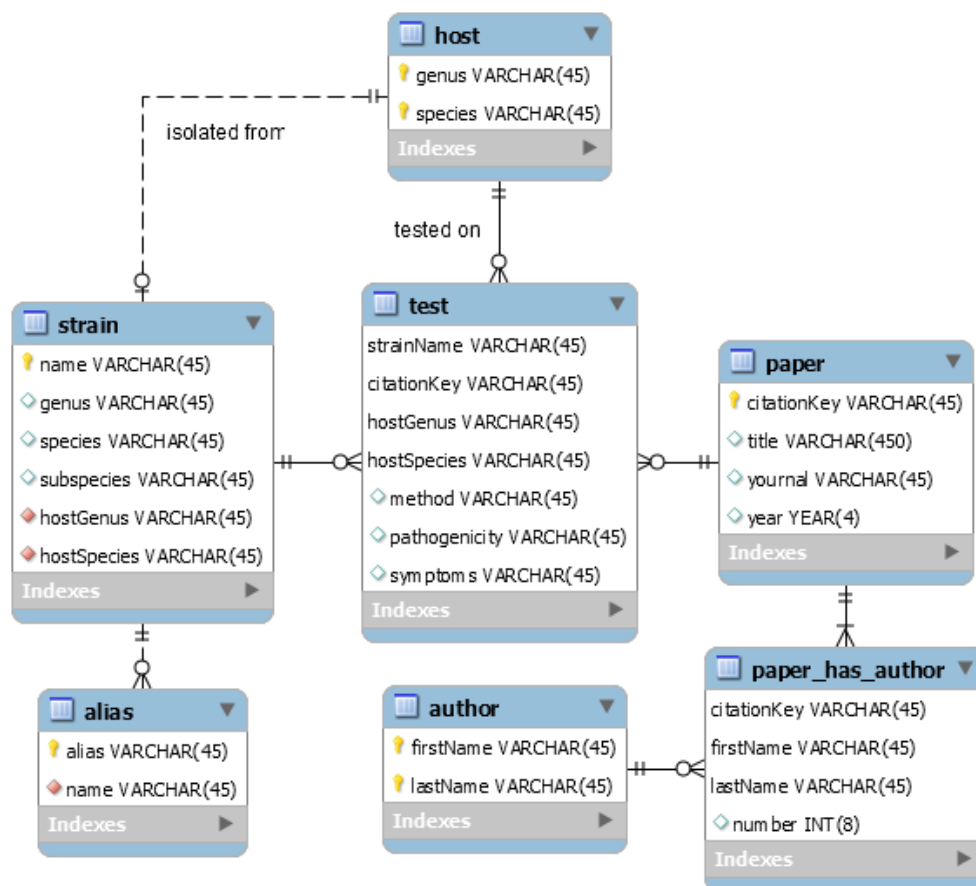


Fig. S1. Entity-Relationship diagram of the *Xanthomonas* phenotype database. Solid line: dependent relationship, dotted line: independent relationship.

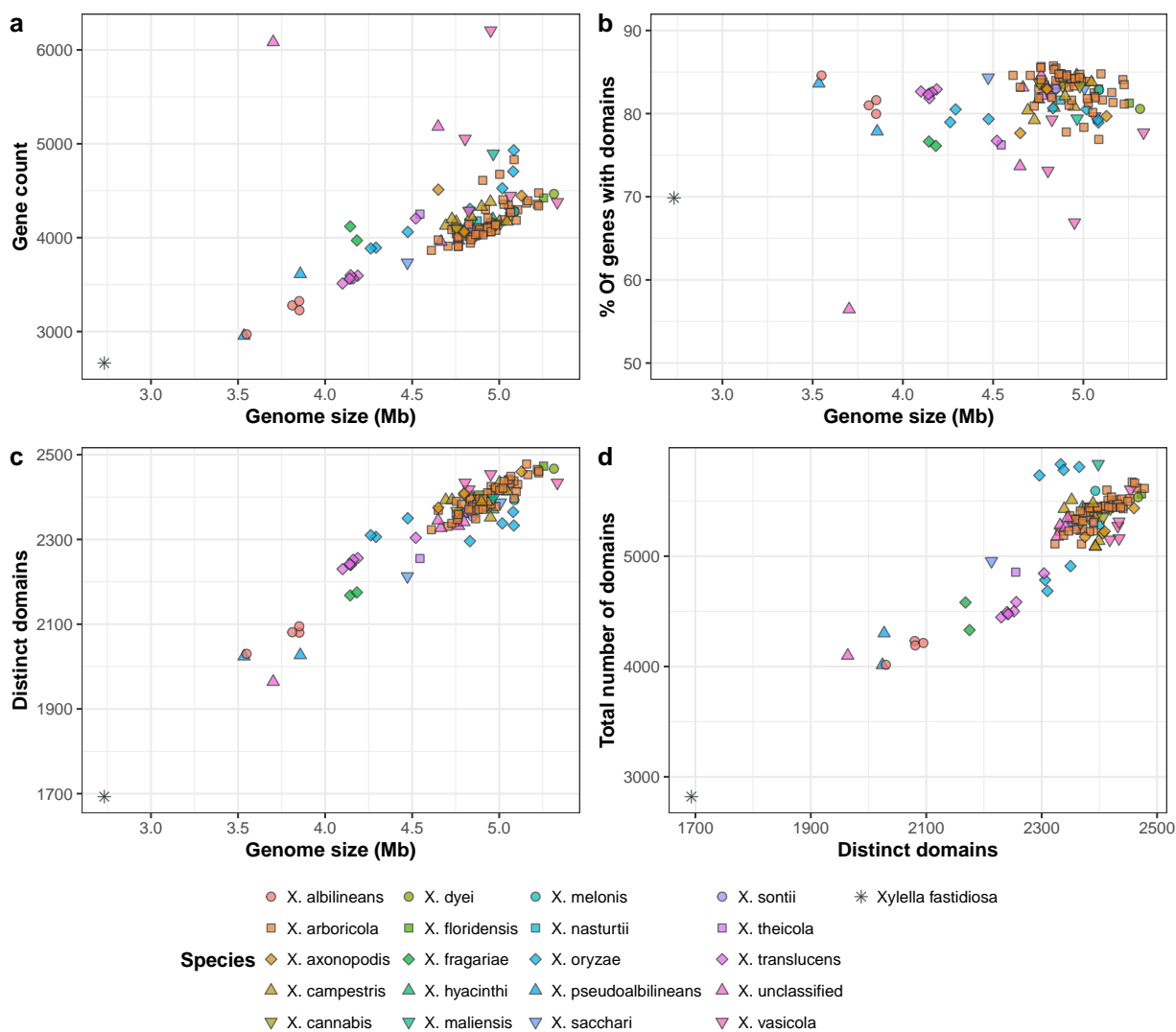


Fig. S2. Annotation statistics of 118 sequenced *Xanthomonas* strains used in this study. *Xylella fastidiosa* 9A5C (marked by *) was used as an out-group.

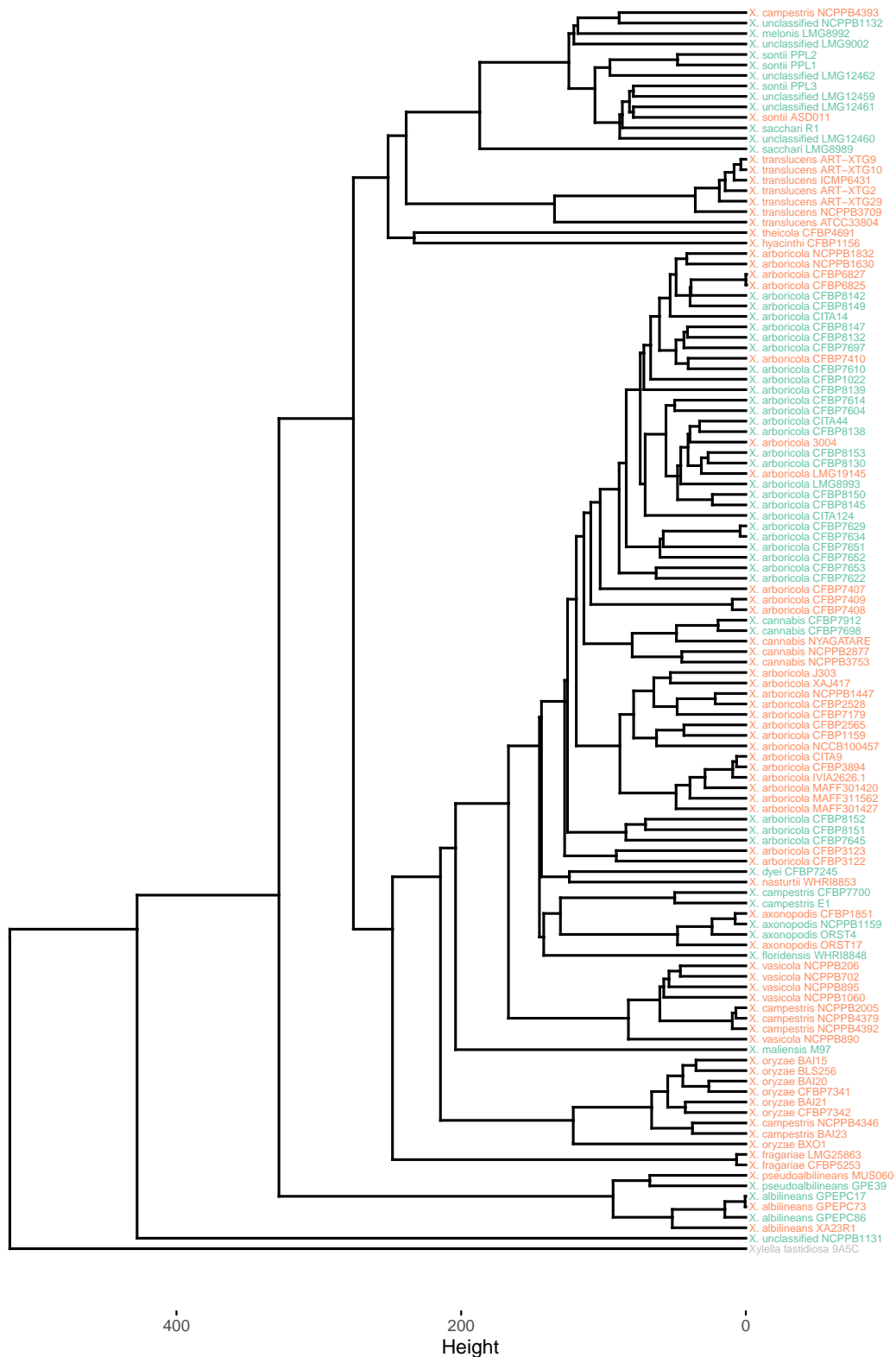


Fig. S3. Domain based distance tree of the 118 *Xanthomonas* strains used in this study. The tree was calculated on the binary domain presence/absence matrix using Manhattan distance. Colours indicate pathogenicity according to literature: red = pathogenic; green = non-pathogenic; *Xylella fastidiosa* is used as out-group (in grey).