te Molder *et al. BMC Genomics* (2021) 22:848
https://doi.org/10.1186/s12864-021-08093-0

BMC Genomics

RESEARCH Open Access

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*]

Abstract

Background: 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 pathogenicity of individual strains over many studies. Unification of this information into a single resource was therefore considered to be an essential step.

Results: 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.

Conclusion: The literature resource in combination with recursive feature extraction used in the ML classification algorithms provided further insights into the virulence enabling factors, but also highlighted domains linked to traits not present in pathogenic strains.

Keywords: Pathogenicity, Protein domains, Machine learning, *Xanthomonas*

Background

The genus of *Xanthomonas* is mostly known trough its pathogenic members, 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 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 test conditions used. For example, a set of *X. arboricola* pv. fragariae strains isolated from infected

* Correspondence: jasper.koehorst@wur.nl
[1] Laboratory of Systems and Synthetic Biology, Wageningen University, Wageningen, the Netherlands
[2] UNLOCK, Wageningen University, Wageningen, the Netherlands

Table 1 Overview of the Xanthomonas phenotype database

Species	Strains	+	–	Top 5 Most Tested Hosts	Reference
X. arboricola	258	180	232	*Juglans regia, Prunus persica, Phaseolus vulgaris,* *Fragaria ananassa, Capsicum annuum*	[6, 8, 14, 21, 28, 29, 36, 54–56] [10, 13, 57–63]
X. oryzae	70	68	2	*Oryza sativa*	[5, 9]
X. translucens	58	72	14	*Lolium multiflorum, Asparagus virgatus, Hordeum vulgare, Anthurium andreanum*	[62, 64–66]
X. campestris	40	45	22	*Brassica oleracea, Musa acuminata, Saccharum officinarum, Zea mays, Prunus persica*	[6, 9, 59, 60, 62, 67, 68]
X. albilineans	27	18	9	*Saccharum officinarum, Zea mays*	[69–71]
X. axonopodis	20	12	14	*Maniholt esculenta, Musa acuminata, Saccharum officinarum, Zea mays, Lycopersicum esculentum*	[29, 59, 62, 72]
X. dyei	13	12	27	*Dysoxylum spectabile, Laurelia novae-zelandiae, Metrosideros excelsa*	[63]
X. vasicola	6	12	6	*Musa acuminata, Saccharum officinarum, Zea mays*	[59]
X. cannabis	5	5	2	*Phaseolus vulgaris, Capsicum annuum, Hordeum vulgare*	[23, 68, 73]
X. sontii	5	2	3	*Oryza sativa*	[5, 74]
X. floridensis	4	0	8	*Brassica oleracea, Nasturtium officinale*	[75]
X. maliensis	4	0	4	*Oryza sativa*	[9, 22]
X. fragariae	3	3	0	*Fragaria ananassa*	[13]
X. euvesicatoria	2	2	0	*Lycopersicum esculentum, Capsicum annuum*	[6]
X. hortorum	2	0	2	*Daucus carota*	[62]
X. nasturtii	2	2	2	*Brassica oleracea, Nasturtium officinale*	[75]
X. pseudoalbilineans	2	1	1	*Saccharum officinarum*	[70, 76]
X. sacchari	2	0	2	*Citrus sinensis, Oryza sativa*	[7, 62]
X. hyacinthi	1	2	0	*Hyacinthus orientalis, Scilla tubergeniana*	[77]
X. melonis	1	0	1	*Citrus sinensis*	[62]
X. theicola	1	1	0	*Camellia sinensis*	[78]
X. spp.	52	66	41	*Asparagus virgatus, Hordeum vulgare, Oryza sativa, Nicotiana tabacum, Phaseolus vulgaris*	[9, 13, 59, 62, 66]
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

strawberries did not cause symptoms when sprayed onto new plants [13], but a repetition of the same assay at an increased humidity 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 in vitro 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 genomic factors are also present in the genomes of the nonpathogens and currently it is 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 gene 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]. However, these findings do not fully explain the differences 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 appears to be strongly correlated with the non-pathogenic phenotype [22], several members of the *X. cannabis* species have been shown to be pathogenic to multiple different hosts even though they lack the entire T3SS [23].

Overall the results suggest that a more complete set of genome-encoded features is required to differentiate between phytopathogenic and non-pathogenic strains. In the last ten years, the number of available *Xanthomonas* genomes has increased nearly 100-fold [2, 24]. 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 [25]. Teper et al. successfully used machine learning to identify novel X euvesicatoria type III effector proteins [26]. Here we explored the applicability of three ML approaches to predict the plant-associated phenotype of Xanthomonas strains.

For *Xanthomonas* species such approaches have been hampered by the fragmentation of information on the plant-associated phenotype of individual *Xanthomonas* strains over many studies. There are databases that track the pathogenicity of individual strains such as CIRM-CFBP [27], 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 retrieved from 39 studies, that each took into account both pathogenic and non-pathogenic xanthomonads, were unified into a single database. This database was then leveraged to retrieve available genome sequence which were de novo annotated for the presence of Pfam protein domains. These domains were subsequently used as input to train three different classifiers. The resulting models were examined for their ability to predict the pathogenicity of individual strains (Fig. 1). Important features were extracted from these models providing new insights into the genome encoded factors contributing to the plantassociated phenotype. At the same time, these classifiers provide a way to crossvalidate pathogenicity test results and conditions.

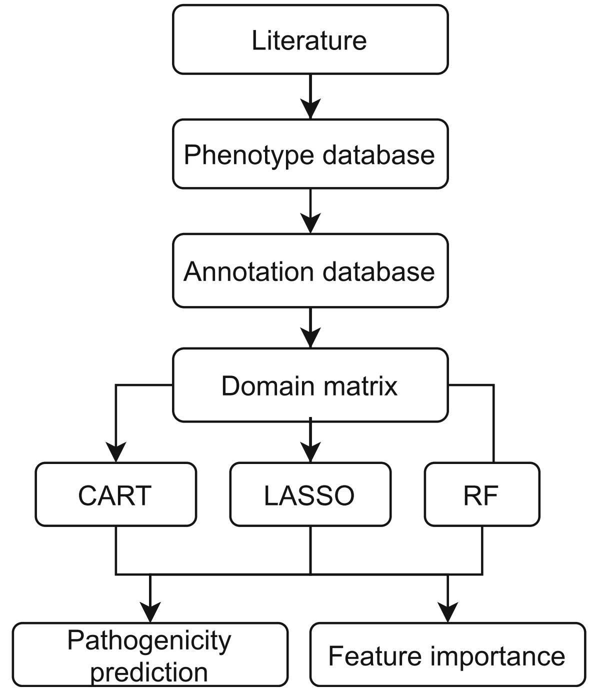

Fig. 1 Workflow. *Xanthomonas* pathogenicity assay data for different strains was obtained from literature and stored in a SQL (phenotype) database. Available genomes were retrieved and de novo annotated with protein domains. Annotation results were stored in a Graph database. Strain specific domain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

Results

Development of the *Xanthomonas* phenotype database

To manage literature derived information related to the strain specific plantassociated phenotype, an SQL database was created (Supplementary Fig. 5). This database, which will be referred to as the phenotype database, was used to track 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 track the criteria used to determine the plant-associated phenotype of a strain. Many studies considered only pathogenic strains and to correct for this imbalance in our database, the data collection effort was limited to studies that took both plant-associated phenotypes 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). From the 578 unique strains, 522 were assayed on their host of isolation (Supplementary File S1). Out of the collected 895 individual assays, 503 did and 392 did not observe symptoms indicative of pathogenicity.

Table 2 Sequenced Xanthomonads with a known phenotype

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

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 and all other hosts 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 [28] 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, 29] 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. For 29 strains the status was considered to be ambiguous and these were excluded from this study. Strain names and known aliases of these strains were cross-referenced with the GenBank sequence database [24], to obtain available matching genomes. This resulted in a set 65 pathogenic and 53 non-pathogenic *Xanthomonas* strains with a known genome 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).

De novo genome annotation

Collected genomes were de novo annotated for Pfam domains using the SAPP platform [30], which implements Prodigal for gene calling [31] and InterProScan for domain annotation [32]. 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 Fig. 6 and Supplementary File S2).

Domain matrix

For each strain the set of unique protein domains was extracted from the annotation database and combined into a binary strain/domain presence matrix used for further analysis and model building. The resulting

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

matrix contained 3609 unique Pfam protein domains distributed over the 118 *Xanthomonas* strains.

As the ML models will be developed from this present *Xanthomonas* matrix, a" closed" domain representation of genus is a prerequisite. When a genus is closed, it is assumed that the large majority of the observed genetic variation within is captured by the current data. 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 pandomainome captured in the matrix was closed (Fig. 2). The captured phylogenetic diversity was additionally visualised using a binary domain distance tree (Fig. 7 and Supplementary File S2).

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 are present in all strains 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 the domain with the highest absolute correlation from sets of domains with a pair-wise Pearson correlation > 0.8. This further reduced the complexity of the matrix, yielding a final matrix of 871 unique domains which was used for model building and further analysis.

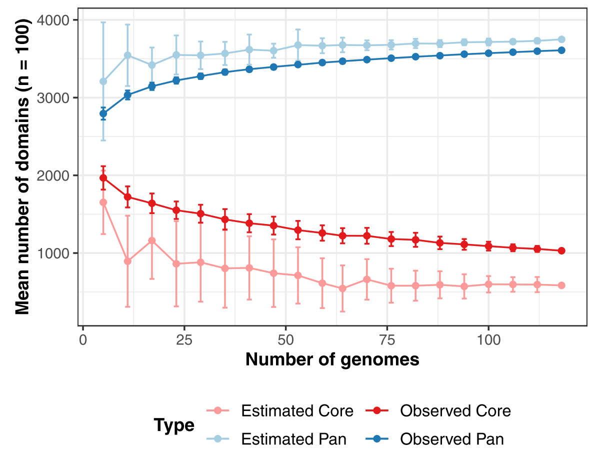

Fig. 2 The Xanthomonas Pan-domainome is closed. The decay parameter, alpha, was estimated to be 1.22 (see methods for details)

Machine Learning Approaches:

A Partial Least Squares Discriminant Analysis (PLS-DA) was applied to the data set to visualise covariance between the domain content and phenotype. The PLS-DA suggested that overall, the strain specific domain content provided a good way to discriminate between both classes, but also that some strains might be mislabelled. (Fig. 3a & 3b).

To learn more about the relationship between the domain content and the plantassociated phenotype, three 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 20 × 5-4 × 10 nested repeated Cross Validation (CV) was used, based on recommendations of Krstajic et al. [33] and Kuhn et al. [34]. 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 training- and test-set partition (Fig. 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 showed a similar improvement in performance, with a median accuracy of 0.826 and 0.833 respectively. The RF model performed better on the prediction of non-pathogenic strains, with an 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 median selected tuning parameters indicated that all classifiers favoured models of relatively low complexity; $cp = 0.068$ for CART (yielding 2–3 domains per tree), $\lambda = 7.85 * 10^{-3}$ for Lasso (yielding ~ 29 domains with a non-zero coefficient), and $mtry = 89$ for RF (Supplementary File S3). The median precision-recall (PR) curve (Fig. 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 is caused by the model attributing the highest probability of pathogenicity to two *X. axonopodis* strains

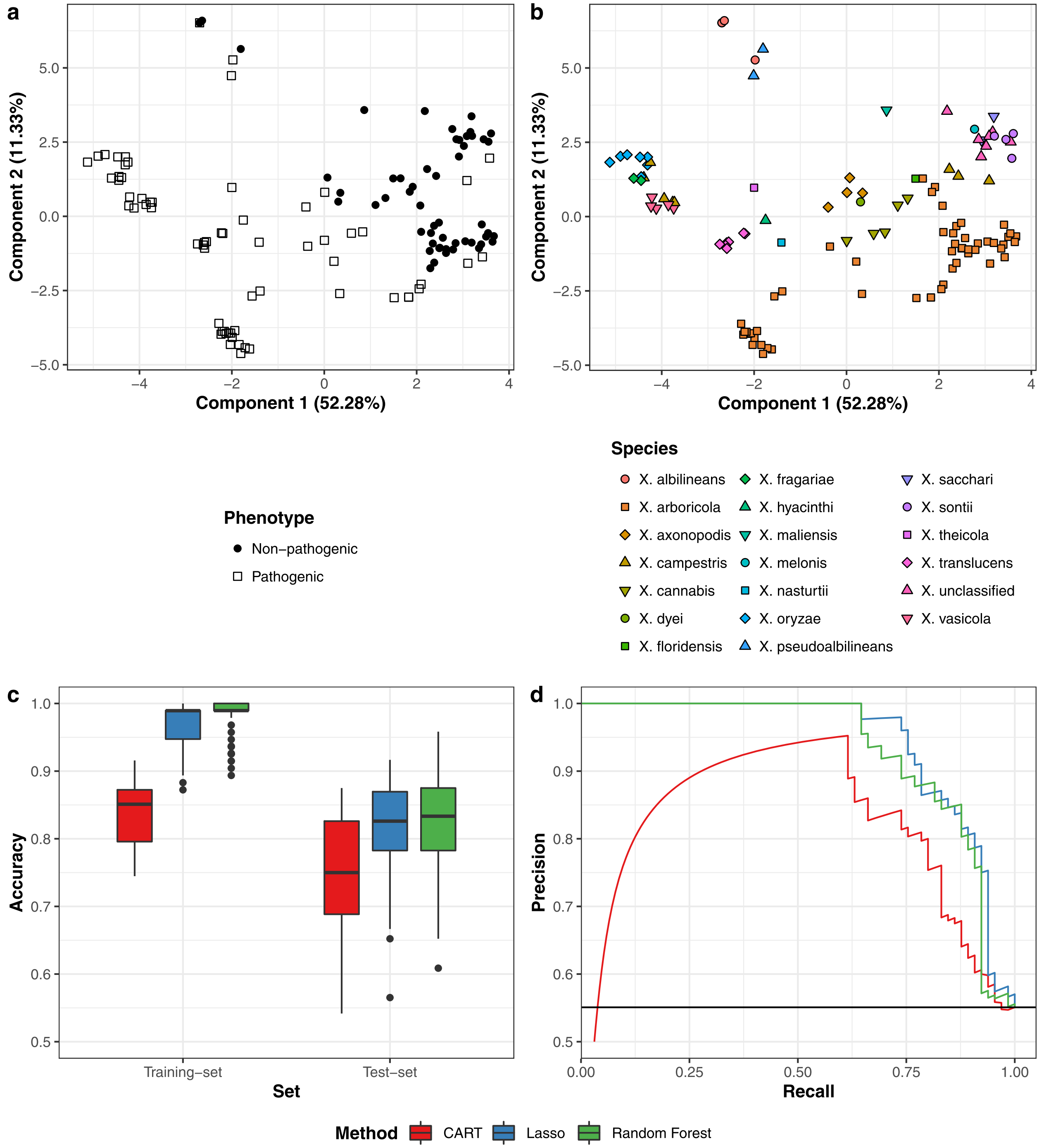

Fig. 3 Properties of the training data set and model prediction performances. Upper panel: Partial Least Squares Discriminant Analysis showing the discrimination between variables and phenotypic classes (a) *Xanthomonas* strains labelled by phenotype. (b) *Xanthomonas* strains labelled according to species classification. A standard PCA representation is available in supplementary file 8 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 predicted median test-set probabilities

that according to literature are non-pathogenic. The PR-curve also shows that there is no discernible difference in performance between the Lasso and RF.

Species specific prediction performance

To gain a better understanding of model behaviour, the in silico predicted class probabilities were compared

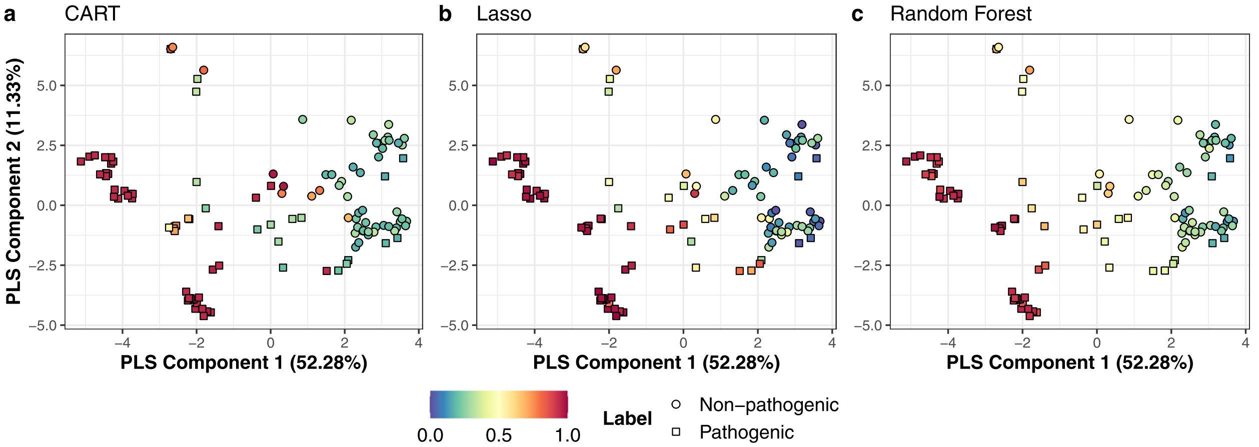

Fig. 4 Predicted median class probabilities mapped onto the PLS-DA. Colour scale (0–1) represents the probability of a strain being pathogenic. Labels represent the phenotype according to literature. A standard PCA representation is available in supplementary file 9

with the in vivo labels 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 (Fig. 4).

High confidence pathogen predictions labelled in red, that scored consistently high across all approaches, belonged to the *X. oryzae, X. vasicola* and *X. fragariae* species. For these species the data set only contained genomes from strains that, according to literature, were pathogenic. 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. A high confidence was also obtained for the *juglandis, corylina* and *pruni* pathovars of *X. arboricola*. Lower confidence scores were obtained for pathogenic strains belonging to species with a single sequenced genome (*X. hyacinthi, X. nasturtii* and *X. theicola*). The *X. nasturtii* 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.

High confidence non-pathogen predictions are labelled in a blue/green and 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. Overall lower confidence scores were obtained in comparison to the pathogens. This likely stems from four strains that are pathogenic according to literature, but are located close to the clusters of nonpathogens.

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 were uncertain about the *X. albilineans* species located at the top of the PLS, as indicated by the neutral scores. 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.

Feature importance

All three ML approaches apply a form of feature selection, by generating variable importance scores for the protein domains. For each ML approach these scores were obtained in a different way. For CART, variable importance scores was obtained by tabulating reduction in loss function for all candidate variables considered at each split; for Lasso, variable importance scores were computed from the 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 between the classes using a two-side Fisher exacted 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 enrichment scores of highly correlated domains removed in matrix filtering. The top 10 domains of CART and RF show a

Table 4 Top features used in theAQ5 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

strong correlation with enrichment, whereas Lasso relied more on 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 studying 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, enabling pathogenomics approaches. However, such approaches are hampered by the current publication bias towards pathogenic strains and fragmentation of information regarding the plant associated phenotypes of individual strains. In order to obtain a reliable balanced training set representing both plant-associated phenotypes, we collected pathogenicity assays from studies that included both plant-associated phenotypes in their study. 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 *abiotic* conditions used.

Available linked genome sequences were de novo annotated for Pfam domains. A heap analysis of the resulting domain matrix indicated that the pan-domainome was closed, and despite some evidence for mislabelling, a partial least squares discriminant analysis indicated that a good discrimination between both phenotypes is possible based on domain content. To enable the inclusion of multiple domains into the decision making process, three different machine learning approaches were explored. CART and Lasso favour 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 higher than default ($p\ = 30$), 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 it's non-pathogenicity [5]. *X. campestris* NCPPB4393, actually belonged to the *X. sacchari* species, a species of which the pathogenicity is still ill defined [35] 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 and a closely related strain, *X. arboricola* CITA 44, was not able to cause any symptoms on this same host or any other tested host [36]. This indicates that these strains are either misclassified, or belong to a class of very weak opportunistic pathogens. On the other hand, there also remains the possibility that strains identified as non-pathogens by the literature symbiotically rely on co-infection with pathogenic species [15]. Given that infection assays typically use pure cultures, such behaviour 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 non-pathogenic strains and strains belonging to the highly pathogenic pathovars (*juglandis*, *corylina* and *pruni* pathovars) were correctly predicted with high confidence. *X. arboricola* strains belonging to weakly pathogenic pathovars obtained more neutral probabilities (Fig. 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 pathogenic strains:

Extraction and examination of important features, showed that the CART and RF considered mostly the same domains, with both favouring domains that were highly enriched in one of the two 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. The role of other features that are predicted to be important for the plant-associated phenotype require experimental validation. Overall, the most important feature in the RF and CART 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 is 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 [37]. 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* [38].

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* [39]. 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 [40].

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 SAP130 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 [41] suggesting an important role for suppressing the plant hosts immune system.

Finally, the most important domain for the Lasso model was PF13579, a nterminal 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 antimicrobial 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*.

Important features enriched in non-pathogenic strains:

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 the Helix-Turn-Helix domain (PF12840) is an important discriminant enriched in non-pathogens. This domain was found in all non-pathogens with exception for a subgroup of the *X. arboricola* species 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 [42].

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 nteractions [43]. Cellulose synthesis is promoted by cyclic-di-GMP trough the PilZ domain present in glycosyltransferase CeSA, which is part of the membrane-bound cellulose synthase complex [44]. 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 MltAinteracting MipA domain (PF06629) which 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* [45]. 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 an accumulation of non-persistent traits. Consequently a single genome encoded feature will have limited power to correctly predict the plant-associated phenotype. By training machine learning methods that take into account an ensemble of domains, a better prediction can be obtained. However, databases that track the pathogenicity of individual species or strains are not harmonized leading to poor interoperability. Unification of phenotype data into a single interoperable resource was therefore considered an essential part of this study.

We explored the applicability of three ML approaches (CART, Lasso, and RF) to predict the plant-associated phenotype. Overall Lasso regression analysis and tree-based RF analysis performed best. Through recursive feature elimination, key domains related to the plant-associated phenotype could be identified, suggesting the involvement 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. 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 SQL 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 populate 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 and genome annotation:

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.

Genomes were de novo annotated using the *SAPP* framework [30], 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 [46]. Protein encoding genes were identified using *Prodigal* [31] and annotated for Pfam protein domains [47] using *InterproScan* (v5.44–79.0) [32]. To speed up computation, SAPP modules were run in parallel using the *GNU Parallel* CLI [48]. Annotated genomes and their provenance, were uploaded to a linked data repository using *GraphDB* (v9.7.0) for further analysis.

Genome 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 [49] was used. The effect of sample size on the estimated/observed pan and core domainome 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 by removing the domain with the highest absolute mean correlation from sets of domains with 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. From the resulting model the first two-components were extracted and visualised.

Model development

Model tuning and testing (*model building. R*) was performed using the *ResampleModel* 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 [50]. The filtered pathogenicity dataset was partitioned using a nested CrossValidation (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 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 [51], Lasso models from the *glmnet* package (by setting $\alpha = 1$) [52], and Random Forest models from the *RandomForest* package [53]. 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 and 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 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.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-08093-0.

Additional file 1. Input and output of the SQL database. Additional file descriptions text.

Additional file 2. Genome metadata, annotation statistics and the domain matrix Additional file descriptions text.

Additional file 3. Model statistics and domain enrichment Additional file descriptions text.

Additional file 4: Fig. S1. Entity-Relationship diagram of the Xanthomonas phenotype database. Solid line: dependent relationship, dotted line: independent relationship.

Additional file 5: Fig. S2. Annotation statistics of 118 sequenced Xanthomonas strains used in this study. *Xylella fastidiosa* 9A5C (marked by *) was used as an out-group.

Additional file 6: 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).

Additional file 7: Fig. S4. Properties of the training data set and model prediction performances. Principle Component Analysis showing the discrimination between variables and phenotypic classes (a) *Xanthomonas* strains labelled by phenotype. (b) *Xanthomonas* strains labelled according to species classification.

Additional file 8: Fig. S5. Predicted median class probabilities mapped onto the PCA. Colour scale (0–1) represents the probability of a strain being pathogenic. Labels represent the phenotype according to literature.

Acknowledgements

Text for this section.

Authors' contributions

D.M., W.P., P.J.S., J.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.J.S., J.J.K. wrote sections of the manuscript. All authors critically revised the manuscript.

Funding

P.J.S and J.J.K 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.

Availability of data and materials

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 https://doi.org/10.4121/14546625.

Declarations

Ethics approval and consent to participate

NA

Consent for publication

All authors consent to the publication of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 7 June 2021 Accepted: 15 October 2021
Published online: 23 November 2021

References

1. Jacques M-A, Arlat M, Boulanger A, Boureau T, Carr'ere S, Cesbron S, et al. Using ecology, physiology, and genomics to understand host specificity in xanthomonas. Annu Rev Phytopathol. 2016;54(1):163–87. https://doi.org/10.1146/annurev-phyto-080615-100147.
2. Ryan RP, Vorholter F-J, Potnis N, Jones JB, Van Sluys M-A, Bogdanove AJ, et al. Pathogenomics of xanthomonas: understanding bacterium–plant interactions. Nat Rev Microbiol. 2011;9(5):344.
3. Maas J, Finney M, Civerolo E, Sasser M. Association of an unusual strain of xanthomonas campestris with apple. Phytopathology. 1985;75(4):438–45. https://doi.org/10.1094/Phyto-75-438.
4. Boureau T, Kerkoud M, Chhel F, Hunault G, Darrasse A, Brin C, et al. A multiplex-pcr assay for identification of the quarantine plant pathogen *xanthomonas axonopodis* pv. phaseoli. J Microbiol Methods. 2013;92(1):42–50.
5. Bansal K, Kaur A, Midha S, Kumar S, Korpole S, Patil PB. Xanthomonas sontii sp. nov., a non-pathogenic bacterium isolated from healthy basmati rice (oryza sativa) seeds from india. bioRxiv. 2019;738047.
6. Essakhi S, Cesbron S, Fischer-Le Saux M, Bonneau S, Jacques M-A, Manceau C. Phylogenetic and variable-number tandem-repeat analyses identify nonpathogenic xanthomonas arboricola lineages lacking the canonical type iii secretion system. Appl Environ Microbiol. 2015;81(16):5395–410. https://doi.org/10.1128/AEM.00835-15.
7. Fang Y, Lin H, Wu L, Ren D, Ye W, Dong G, et al. Genome sequence of xanthomonas sacchari r1, a biocontrol bacterium isolated from the rice seed. J Biotechnol. 2015;206:77–8. https://doi.org/10.1016/j.jbiotec.2015.04.014.
8. Garita-Cambronero J, Palacio-Bielsa A, L'opez MM, Cubero J. Pan-genomic analysis permits differentiation of virulent and non-virulent strains of *xanthomonas arboricola* that cohabit prunus spp. and elucidate bacterial virulence factors. Front Microbiol. 2017;8:573.
9. Wonni I, Cottyn B, Detemmerman L, Dao S, Ouedraogo L, Sarra S, et al. Analysis of *xanthomonas oryzae* pv. oryzicola population in mali and burkina faso reveals a high level of genetic and pathogenic diversity. Phytopathology. 2014;104(5):520–31.

10. Merda D, Bonneau S, Guimbaud J-F, Durand K, Brin C, Boureau T, et al. Recombination-prone bacterial strains form a reservoir from which epidemic clones emerge in agroecosystems. Environ Microbiol Rep. 2016; 8(5):572–81. https://doi.org/10.1111/1758-2229.12397.
11. Merda D, Briand M, Bosis E, Rousseau C, Portier P, Barret M, et al. Ancestral acquisitions, gene flow and multiple evolutionary trajectories of the type three secretion system and effectors in xanthomonas plant pathogens. Mol Ecol. 2017;26(21):5939–52. https://doi.org/10.1111/mec.14343.
12. Zhang J, Zhang C, Yang J, Zhang R, Gao J, Zhao X, et al. Insights into endophytic bacterial community structures of seeds among various oryza sativa l. rice genotypes. J Plant Growth Regul. 2019;38(1):93–102.
13. Vandroemme J, Cottyn B, Pothier JF, Pfluger V, Duffy B, Maes M. Xanthomonas arboricola pv. fragariae: what's in a name? Plant Pathol. 2013; 62(5):1123–31.
14. Ferrante P, Scortichini M. Xanthomonas arboricola pv. fragariae: a confirmation of the pathogenicity of the pathotype strain. Eur J Plant Pathol. 2018;150(3):825–9.
15. An S-Q, Potnis N, Dow M, Vorholter F-J, He Y-Q, Becker A, et al. Mechanistic insights into host adaptation, virulence and epidemiology of the phytopathogen xanthomonas. FEMS Microbiol Rev. 2019.
16. Timilsina S, Potnis N, Newberry EA, Liyanapathiranage P, Iruegas-Bocardo F, White FF, et al. Xanthomonas diversity, virulence and plant–pathogen interactions. Nat Rev Microbiol. 2020:1–13.
17. Kay S, Bonas U. How xanthomonas type iii effectors manipulate the host plant. Curr Opin Microbiol. 2009;12(1):37–43. https://doi.org/10.1016/j.mib.2008.12.006.
18. Rigano LA, Payette C, Brouillard G, Marano MR, Abramowicz L, Torres PS, et al. Bacterial cyclic *β*-(1, 2)-glucan acts in systemic suppression of plant immune responses. Plant Cell. 2007;19(6):2077–89. https://doi.org/10.1105/tpc.106.047944.
19. Vojnov AA, Slater H, Daniels MJ, Dow JM. Expression of the gum operon directing xanthan biosynthesis in xanthomonas campestris and its regulation in planta. Mol Plant-Microbe Interact. 2001;14(6):768–74. https://doi.org/10.1094/MPMI.2001.14.6.768.
20. Wang L-H, He Y, Gao Y, Wu JE, Dong Y-H, He C, et al. A bacterial cell–cell communication signal with cross-kingdom structural analogues. Mol Microbiol. 2004;51(3):903–12. https://doi.org/10.1046/j.1365-2958.2003.03883.x.
21. Cesbron S, Briand M, Essakhi S, Gironde S, Boureau T, Manceau C, et al. Comparative genomics of pathogenic and nonpathogenic strains of xanthomonas arboricola unveil molecular and evolutionary events linked to pathoadaptation. Front Plant Sci. 2015;6:1126. https://doi.org/10.3389/fpls.2015.01126.
22. Triplett LR, Verdier V, Campillo T, Van Malderghem C, Cleenwerck I, Maes M, 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. 2015;107(4):869–81.
23. Jacobs JM, Pesce C, Lefeuvre P, Koebnik R. Comparative genomics of a cannabis pathogen reveals insight into the evolution of pathogenicity in xanthomonas. Front Plant Sci. 2015;6:431. https://doi.org/10.3389/fpls.2015.00431.
24. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. Genbank. Nucleic Acids Res. 2021;49(D1):92–6. https://doi.org/10.1093/nar/gkaa1023.
25. Allen JP, Snitkin E, Pincus NB, Hauser AR. Forest and trees: exploring bacterial virulence with genome-wide association studies and machine learning. Trends Microbiol. 2021;29(7):621–33. https://doi.org/10.1016/j.tim.2020.12.002.
26. Teper D, Sunitha S, Martin GB, Sessa G. Five xanthomonas type iii effectors suppress cell death induced by components of immunity-associated map kinase cascades. Plant Signal Behav. 2015;10(10):1064573. https://doi.org/10.1080/15592324.2015.1064573.
27. CIRM: CFBP - plant associated Bacteria (2021). https://www6.inrae.fr/cirm/CFBP-Bacteries-associees-aux-Plantes
28. Hajri A, Meyer D, Delort F, Guillaumes J, Brin C, Manceau C. 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 Pathol. 2010;59(6):1014–22.
29. Fischer-Le Saux M, Bonneau S, Essakhi S, Manceau C, Jacques M-A. Aggressive emerging pathovars of xanthomonas arboricola represent widespread epidemic clones distinct from poorly pathogenic strains, as revealed by multilocus sequence typing. Appl Environ Microbiol. 2015; 81(14):4651–68. https://doi.org/10.1128/AEM.00050-15.
30. Koehorst JJ, van Dam JC, Saccenti E, Martins dos Santos VA, Suarez-Diez M, Schaap PJ. Sapp: functional genome annotation and analysis through a semantic framework using fair principles. Bioinformatics. 2017;34(8):1401–3.
31. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010;11(1):119. https://doi.org/10.1186/1471-2105-11-119.
32. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. Interproscan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.
33. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of cheminformatics. 2014;6(1):1–15. https://doi.org/10.1186/1758-2946-6-10.
34. Kuhn M, Johnson K, et al. Applied Predictive Modeling vol. 26: Springer; 2013.
35. Studholme DJ, Wasukira A, Paszkiewicz K, Aritua V, Thwaites R, Smith J, et al. Draft genome sequences of xanthomonas sacchari and two banana-associated xanthomonads reveal insights into the xanthomonas group 1 clade. Genes. 2011;2(4):1050–65. https://doi.org/10.3390/genes2041050.
36. Garita-Cambronero J, Palacio-Bielsa A, Lopez MM, Cubero J. 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. 2016;11(8).
37. Schmidt H, Hensel M. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev. 2004;17(1):14–56. https://doi.org/10.1128/CMR.17.1.14-56.2004.
38. Kim J-G, Park BK, Yoo C-H, Jeon E, Oh J, Hwang I. Characterization of the *xanthomonas axonopodis* pv. glycines hrp pathogenicity island. J Bacteriol. 2003;185(10):3155–66.
39. Monteiro-Vitorello CB, De Oliveira MC, Zerillo MM, Varani AM, Civerolo E, Sluys M-AV. Xylella and xanthomonas mobil'omics. Omics: a journal of integrative biology. 2005;9(2):146–59. https://doi.org/10.1089/omi.2005.9.146.
40. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 2000;54(1):641–79. https://doi.org/10.1146/annurev.micro.54.1.641.
41. Yokoyama R, Itoh S, Kamoshida G, Takii T, Fujii S, Tsuji T, et al. 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. Infect Immun. 2012;80(8):2816–25. https://doi.org/10.1128/IAI.00399-12.
42. Zhou L, Vorholter F-J, He Y-Q, Jiang B-L, Tang J-L, Xu Y, et al. Gene discovery by genome-wide cds re-prediction and microarray-based transcriptional analysis in phytopathogen *xanthomonas campestris*. BMC Genomics. 2011;12(1):359.
43. Ross P, Mayer R, Benziman M. Cellulose biosynthesis and function in bacteria. Microbiol Mol Biol Rev. 1991;55(1):35–58. https://doi.org/10.1128/MMBR.55.1.35-58.1991.
44. Fujiwara T, Komoda K, Sakurai N, Tajima K, Tanaka I, Yao M. The c-di-gmp recognition mechanism of the pilz domain of bacterial cellulose synthase subunit a. Biochem Biophys Res Commun. 2013;431(4):802–7. https://doi.org/10.1016/j.bbrc.2012.12.103.
45. Wang L, Yang L-Y, Gan Y-L, Yang F, Liang X-L, Li W-L, et al. Two lytic transglycosylases of *xanthomonas campestris* pv. campestris associated with cell separation and type iii secretion system, respectively. FEMS Microbiol Lett. 2019;366(7):073.
46. Fern'andez JD, Mart'ınez-Prieto MA, Guti'errez C, Polleres A, Arias M. Binary rdf representation for publication and exchange (hdt). Journal of Web Semantics. 2013;19:22–41.
47. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The pfam protein families database in 2019. Nucleic Acids Res. 2018;47(D1):427–32. https://doi.org/10.1093/nar/gky995.
48. Tange O. Gnu Parallel 2018. Lulu. com; 2018.
49. Snipen L, Liland KH. micropan: an r-package for microbial pan-genomics. BMC bioinformatics. 2015;16(1):79.
50. Kuhn M. Caret: classification and regression training. Astrophysics Source Code Library. 2015.
51. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees: CRC press; 1984.

52. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22. https://doi.org/10.18637/jss.v033.i01.
53. Liaw A, Wiener M, et al. Classification and regression by randomforest. R news. 2002;2(3):18–22.
54. Caballero JI, Zerillo MM, Snelling J, Boucher C, Tisserat N. Genome sequence of *xanthomonas arboricola* pv. corylina, isolated from turkish filbert in colorado. Genome Announc. 2013;1(3):00246–13.
55. Garita-Cambronero J, Palacio-Bielsa A, Cubero J. 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. Mol Plant Pathol. 2018;19(9):2053–65.
56. Harrison J, Grant MR, Studholme DJ. Draft genome sequences of two strains of *xanthomonas arboricola* pv. celebensis isolated from banana plants. Genome Announc. 2016;4(1):01705–15.
57. Higuera G, Gonz'alez-Escalona N, V'eliz C, Vera F, Romero J. Draft genome sequences of four *xanthomonas arboricola* pv. juglandis strains associated with walnut blight in chile. Genome Announc. 2015;3(5):01160–15.
58. Ignatov AN, Kyrova EI, Vinogradova SV, Kamionskaya AM, Schaad NW, Luster DG. Draft genome sequence of xanthomonas arboricola strain 3004, a causal agent of bacterial disease on barley. Genome Announc. 2015;3(1):01572–14. https://doi.org/10.1128/genomeA.01572-14.
59. Karamura G, Smith J, Studholme D, Kubiriba J, Karamura E. Comparative pathogenicity studies of the xanthomonas vasicola species on maize, sugarcane and banana. African J Plant Sci. 2015;9(9):385–400. https://doi.org/10.5897/AJPS2015.1327.
60. Kawaguchi A, Inoue K, Inoue Y. Biological control of bacterial spot on peach by nonpathogenic xanthomonascampestris strains az98101 and az98106. J Gen Plant Pathol. 2014;80(2):158–63. https://doi.org/10.1007/s10327-014-0506-6.
61. Pereira UP, Gouran H, Nascimento R, Adaskaveg JE, Goulart LR, Dandekar AM. Complete genome sequence of *xanthomonas arboricola* pv. juglandis 417, a copper-resistant strain isolated from *juglans regia* l. Genome Announc. 2015;3(5):01126–15.
62. Vauterin L, Yang P, Alvarez A, Takikawa Y, Roth DA, Vidaver AK, et al. Identification of non-pathogenic xanthomonas strains associated with plants. Syst Appl Microbiol. 1996;19(1):96–105. https://doi.org/10.1016/S0723-2020(96)80016-6.
63. Young J, Wilkie J, Park D-C, Watson D. New zealand strains of plant pathogenic bacteria classified by multi-locus sequence analysis; proposal of *xanthomonas dyei* sp. nov. Plant Pathol. 2010;59(2):270–81.
64. Hersemann L, Wibberg D, Blom J, Goesmann A, Widmer F, Vorholter F-J, et al. Comparative genomics of host adaptive traits in *xanthomonas translucens* pv. graminis. BMC Genomics. 2017;18(1):35.
65. Kolliker R, Kraehenbuehl R, Boller B, Widmer F. Genetic diversity and pathogenicity of the grass pathogen *xanthomonas translucens* pv. graminis. Syst Appl Microbiol. 2006;29(2):109–19.
66. Rademaker J, Norman D, Forster R, Louws F, Schultz M, De Bruijn F. Classification and identification of xanthomonas translucens isolates, including those pathogenic to ornamental asparagus. Phytopathology. 2006;96(8):876–84. https://doi.org/10.1094/PHYTO-96-0876.
67. Lee Y-A, Yang P-Y, Huang S-C. Characterization, phylogeny, and genome analyses of nonpathogenic *xanthomonas campestris* strains isolated from brassica seeds. Phytopathology. 2020;08.
68. Meline V, Delage W, Brin C, Li-Marchetti C, Sochard D, Arlat M, et al. Role of the acquisition of a type 3 secretion system in the emergence of novel pathogenic strains of xanthomonas. Mol Plant Pathol. 2019;20(1):33–50. https://doi.org/10.1111/mpp.12737.
69. Champoiseau P, Daugrois J-H, Pieretti I, Cociancich S, Royer M, Rott P. High variation in pathogenicity of genetically closely related strains of xanthomonas albilineans, the sugarcane leaf scald pathogen, in Guadeloupe. Phytopathology. 2006;96(10):1081–91. https://doi.org/10.1094/PHYTO-96-1081.
70. Daugrois J-H, Dumont V, Champoiseau P, Costet L, Boisne-Noc R, Rott P. Aerial contamination of sugarcane in Guadeloupe by two strains of xanthomonas albilineans. Eur J Plant Pathol. 2003;109(5):445–58. https://doi.org/10.1023/A:1024259606468.
71. Rott PC, Costet L, Davis MJ, Frutos R, Gabriel DW. At least two separate gene clusters are involved in albicidin production by xanthomonas albilineans. J Bacteriol. 1996;178(15):4590–6. https://doi.org/10.1128/jb.178.15.4590-4596.1996.
72. Gonzalez C, Restrepo S, Tohme J, Verdier V. Characterization of pathogenic and nonpathogenic strains of *xanthomonas axonopodis* pv. manihotis by pcr-based dna fingerprinting techniques. FEMS Microbiol Lett. 2002;215(1):23–31.
73. Aritua V, Musoni A, Kabeja A, Butare L, Mukamuhirwa F, Gahakwa D, et al. The draft genome sequence of xanthomonas species strain nyagatare, isolated from diseased bean in Rwanda. FEMS Microbiology. 2015;362(4):1–4. https://doi.org/10.1093/femsle/fnu055.
74. Mirghasempour SA, Huang S, Studholme DJ, Brady CL. A grain rot of rice in iran caused by a xanthomonas strain closely related to x. sacchari. Plant Dis. 2020;01.
75. Vicente JG, Rothwell S, Holub EB, Studholme DJ. 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. Int J Syst Evol Microbiol. 2017;67(9):3645–54.
76. Orian G. A disease of paspalum dilatatum in Mauritius caused by a species of bacterium closely resembling xanthomonas albilineans (Ashby) dowson. Rev Agric Sucr Ile Maurice. 1962;41:7–20.
77. Janse J, Miller H. Yellow disease in scilla tubergeniana and related bulbs caused by *xanthomonas campestris* pv. hyacinthi. Neth J Plant Pathol. 1983;89(5):203–6.
78. Uehara K, Arai K, Nonaka T, Sano I, et al. Canker of tea, a new disease, and its causal bacterium xanthomonas campestris pv. Theaecola uehara & arai pv. Nov. Bulletin of the Faculty of Agriculture Kagoshima University. 1980;30:17–21.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.