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.

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* pv. fragariae strains isolated from infected strawberries, 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.

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.

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* pv. fragariae strains isolated from infected strawberries, 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.
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 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
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          | Genomes | Strains | Non-Pathogenic | Pathogenic | Total |
|------------------|---------|---------|----------------|------------|-------|
| *X. arboricola*  | 33      | 110     | 26             | 26         | 54    |
| *X. campestris*  | 27      | 28      | 4              | 28         | 56    |
| *X. oryzae*      | 7       | 30      | 0              | 30         | 37    |
| *X. translucens* | 54      | 16      | 0              | 16         | 70    |
| *X. campestris*  | 40      | 52      | 0              | 52         | 92    |
| *X. albilineans* | 27      | 9       | 0              | 9          | 36    |
| *X. axonopodis*  | 20      | 12      | 0              | 12         | 32    |
| *X. vasicola*    | 6       | 6       | 0              | 6          | 12    |
| *X. cannabinus*  | 5       | 5       | 0              | 5          | 10    |
| *X. sonitii*     | 5       | 2       | 3              | 5          | 8     |
| *X. floridensis* | 4       | 8       | 0              | 8          | 12    |
| *X. maltensis*   | 4       | 0       | 4              | 4          | 8     |
| *X. fragariae*   | 3       | 3       | 0              | 3          | 6     |
| *X. evesicatoria*| 2       | 2       | 0              | 2          | 4     |
| *X. hororum*     | 2       | 2       | 0              | 2          | 4     |
| *X. nasturtii*   | 1       | 2       | 0              | 2          | 3     |
| *X. pseudoalbilineans* | 2 | 2 | 0 | 2 | 4 |
| *X. sacchari*    | 1       | 1       | 0              | 1          | 2     |
| *X. theicola*    | 2       | 2       | 0              | 2          | 4     |
| *X. spp.*        | 52      | 66      | 41             | 41         | 93    |

### Table 3. Annotation statistics of the collected genomes.

| Statistics | Min | Max | Average |
|------------|-----|-----|---------|
| Genomes    | 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

---

**De novo Genome Annotation:** Collected genomes were *de novo* annotated for Pfam domains using the SAPP platform (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. Overview of the Xanthomonas phenotype database

| Species          | Strains | +    | -    | Total  | Reference |
|------------------|---------|------|------|--------|-----------|
| *X. arboricola*  | 258     | 180  | 232  | 578    | (6, 8, 10, 13, 14, 21, 27–39) |
| *X. oryzae*      | 70      | 68   | 2    | 145    | (5, 9)    |
| *X. translucens* | 58      | 72   | 14   | 144    | (38, 40–42) |
| *X. campestris*  | 40      | 45   | 22   | 107    | (6, 9, 35, 36, 38, 43, 44) |
| *X. albilineans* | 27      | 18   | 9    | 54     | (45–47)   |
| *X. axonopodis*  | 20      | 12   | 14   | 46     | (28, 35, 38, 48) |

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

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 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 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 \( \sim 29 \) 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. vasiceola 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. thecicola 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
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 strains 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 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 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 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   |        |        |        |             |
| PF09483        | Type III secretion protein HpaP                        | 0.83  | 0.30   |        |        |        |             |
| * PF09486      | Type III secretion protein HrpB7                       | 0.83  | 0.30   |        |        |        |             |
| PF09487        | Type III secretion protein HrpB2                       | 0.83  | 0.30   |        |        |        |             |
| PF09502        | Type III secretion protein HrpB4                       | 0.83  | 0.30   |        |        |        |             |
| 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 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 (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 (*xyliloides*, *coryllina* 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 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 (60). Type III effector proteins are translocated into host cells where they target many host components, serving 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 important discriminator 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 AcrR 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 fulfill diverse roles from mechanical/environmental protection to cell adhesion 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 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 (64) suggesting an important role for suppressing the plant hosts immune system.
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 membrane-bound cellulose synthase complex. Mutations in other genes of the mlr family are related with morphological abnormalities in X. campesiris (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 RMySQLDB (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 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 based on a Pearson correlation of ρ > 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 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 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 α = 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 \times 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 fold 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](https://gitlab.com/wurssb/xanthomonas-phenotype-prediction).

Genome annotations in a binary RDF format (HDT) is available at [http://doi.org/10.4121/14546625](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 Bourseau, Sébastien Carrère, Sophie Cesbron, Nicolas WG Chen, Stéphane Cociancich, Armelle Darrasse, Nicolas Demaille, Jérôme Maes, 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 Bourseau, Mohammed Kerfsud, Fabien Chhel, Gilles Huault, Armelle Darrasse, Chrystelle Brin, Karine Durand, Ahmed Hajji, 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. Karinka Barsal, Amandeep Kaur, Samriti Midha, Sanjeev Kumar, Suresh Korpole, and Prabhu B Patil. Xanthomonas solitii sp. nov., a non-pathogenic bacterium isolated from healthy basmati rice (oryza sativa) seeds from india. bioRxiv, page 738047, 2019.
6. Salwa Essaifi, 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 b1ncontor bacterium isolated from the rice seed. Journal of biotechnology, 206:77–85, 2015.
8. Jerson Gatta-Cambronero, 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. Environmental microbiology reports, 8:573, 2017.
9. Thaës Wouters, Bart Cottyn, Liselot Detemmerman, S Dao, L Cueva-deazo, S Sanna, C Tekete, S Pousser, R Corral, I Trépied, et al. Analysis of xanthomonas oryzae pv. oryzae population in mal 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 Guimband, Karine Durand, Chrystelle Brin, Tristan Bourseau, 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 Bosisi, Céline Rousseau, Perrine Portier, Matthieu Barret, Marie-Agnès Jacques, and Marion Fischer-Le Saux. Ancient 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 P Hoffner, Valentin Pflüger, Brion Duffy, and Martine Maes. Xanthomonas arboricola pv. fragariae: what’s in a name? Phytopathology, 91(5):1123–1131, 2013.
14. Patricia Ferrante and Marco Scortichini. Xanthomonas arboricola pv. fragarieae: a confirmation of the pathogenicity of the pathotype strain. European journal of plant pathology, 150(3):825–828, 2018.
15. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, 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. Jonathan M Jacobs, Céline Pesce, Pierre Lefeuvre, and Ralf Koebnik. Comparative genomics of pathogenic and nonpathogenic strains of xanthomonas arboricola urvilei molecular and evolutionary events linked to pathoadaptation. Frontiers in plant science, 6:1126, 2015.

17. Jiaqian Bi, Thanh-Tien Bui, Benaysir Beja, Mihaela Deleanu, Hirokazu Iwamoto, et al. Pathogenicity islands in bacterial pathogenesis. Applied predictive modeling, 58. Max Kuhn, Kjell Johnson, et al.

18. Lijun Liu, Wei Sun, Yue Lu, Mark Davenport, et al. Comparative genomics of xanthomonas translucens pv. graminis. BMC genomics, 15(1):35–58, 2014.

19. Rogerio Rincón, João Henrique de Araújo, Suyash Lunawat, et al. Aerial contamination of sugarcane in guadeloupe by two strains of xanthomonas albilineans, the sugarcane leaf scald pathogen, in guadeloupe. Phytopathology, 96(1):1081–1091, 2006.

20. Valerian Meline, Wesley Delage, Christelle Brin, Camille Li-Marchetti, Daniel Sochard, Matthieu Aritula, Céline Rousseau, Armelle Damasse, Martial Braun, Guillaume Leteinture, 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.

21. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.

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, 6:435, 2015.

23. Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Conrad L Schoch, Stephen T Thwaites, Julian Smith, and Murray Grant. Draft genome sequences of xanthomonas plantarum pv. theaecola uehara & arai pv. nov. and xanthomonas translucens pv. graminis sp. nov. Antonie van Leeuwenhoek, 107(4):869–881, 2015.

24. 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, 6:435, 2015.

25. Sujan Timilsina, Neha Potnis, Eric A Newberry, Prabha Liyanapathiranage, Fernanda Sherry, and Ilene Karsch-Mizrachi. Genbank.

26. Daniel Sochard, Valérian Meline, Wesley Delage, Christelle Brin, Camille Li-Marchetti, Daniel Sochard, Matthieu Aritula, Céline Rousseau, Armelle Damasse, Martial Braun, Guillaume Leteinture, et al. Virulence and epidemiology of the phytopathogen xanthomonas mobil’omics. Frontiers in plant science, 11:4630, 2020.

27. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.

28. Jiaqian Bi, Thanh-Tien Bui, Benaysir Beja, Mihaela Deleanu, Hirokazu Iwamoto, et al. Pathogenicity islands in bacterial pathogenesis. Applied predictive modeling, 58. Max Kuhn, Kjell Johnson, et al.

29. José M Alonso-Alcalde, Marco Polo Zerlin, Jake Lefebvre, and Ned Tison. Genome sequences of xanthomonas arboricola pv. cynorhiza, isolated from tulip fields in colorado. Genome Announc., 3(1):e00246–13, 2015.

30. Juan F Pabón, Álvaro Díaz, Sergio Buitrago, Natalia Díaz, Juan Pablo Añón, et al. Comparative genomics of xanthomonas translucens pv. graminis and xanthomonas translucens pv. caricis: insights into pathoadaptation. Frontiers in plant science, 11:4630, 2020.

31. Valérie Arthu, Augustin Munisone, Alice Kabeja, Louis Butare, Floride Mukamuhirwa, et al. Comparative genomics of xanthomonas translucens pv. graminis and xanthomonas translucens pv. caricis: insights into pathoadaptation. Frontiers in plant science, 11:4630, 2020.

32. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.

33. Carolina Gonzalez, Silvia Restrepo, Joe Tóthme, and Valérie Verdière. Characterization of pathogenic and nonpathogenic strains of xanthomonas translucens pv. caricis: insights into the xanthomonas arboricola species context. Molecular plant pathology, 67(9):3645–3654, 2017.

34. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.

35. Louis Vauterin, Ping Yang, Anne Alvarez, Yuichi Takakawa, Don A Roth, Anne K Vidaver, Robert E Stahl, Karel Janda, and Jean Swings. Identification of non-pathogenic xanthomonas arboricola pv. caricis strains associated with walnut blight in chile. Genome Announc., 3(3):e0160–15, 2015.

36. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.

37. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.

38. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.

39. Shi-qi An, Neha Potnis, Max Dow, Frank-Jörn Vöhringer, Yong-qi He, Anke Becker, Doron Tepfer, Yi Li, Nian Wang, Leonidas Bleris, et al. Interproscan 5: genome-prokaryotic gene recognition and translation initiation site identification through a semantic framework using fair principles. BMC bioinformatics, 11(1):19, 2010.
Lin Wang, Li-Yan Yang, Yong-Liang Gan, Feng Yong, 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.

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.

Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Mattiob 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.

Ole Tange. Gnu parallel 2018. Lulu. com, 2018.

Lars Snipen and Kristian Hovde Liland. micropan: an r-package for microbial pan-genomics. BMC bioinformatics, 16(1):79, 2015.

Max Kuhn. Caret: classification and regression training. Astrophysics Source Code Library, 2015.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Oshen. Classification and regression trees. CRC press, 1984.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010.

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).