Supporting Information

Trait-based approaches for predicting global impacts of pathogens in the genus *Phytophthora*

Supporting methods and results

Appendix S4

Pure latent variable models of trait syndromes

We specified a pure latent variable model (Hui, 2016) with two latent variables,

,

where is the appropriate link function for a given trait distribution. We assumed binomial responses for binary traits (chlamydospore production, oospore production, caducous sporangia, proliferating sporangia, root disease symptoms, foliar disease symptoms) and Gaussian responses for continuous traits (growth rate at optimum temperature, optimum/minimum temperature for growth and oospore wall index) and checked these assumptions visually using Dunn-Smythe residual plots. is the mean response for species *i* of trait *j*, is a trait-specific intercept. and are two latent variable vectors of length *n =* number of species = 179 and and are two vectors of coefficients corresponding to LV1 and LV2, each of length equal to the number of traits (*p* = 10). The coefficient represents the magnitude of responses of trait *j* to the LVs and can be interpreted in the same way as the loadings in an unconstrained ordination to illustrate covariance among traits. We use the LV estimates as our trait syndrome predictors. LVs also provide the axes of a biplot to visualise how species are positioned in trait-space and whether high impact species tend to share particular trait combinations. Mapping traits on to the latent variable axes can help to inform hypotheses about the ecological strategies or evolutionary processes captured by the LVs. We express the goodness-of-fit of these LV models as the proportion of deviance explained by two LVs relative to an intercept-only model of the species by trait matrix. To check convergence of estimated LVs and LV coefficients, we selected 20 of the 399 parameters at random and examined trace plots of the MCMC chains.

Results of trait syndrome models

The two LVs together explained 26% of the deviance in the intercept-only model of the species by trait matrix. Optimum temperature for growth and growth rate at optimum temperature had the strongest negative loadings on LV1, while oospore production and caducous sporangia loaded positively on LV1. LV2 explained the remaining 44.4 % of trait covariance. Proliferating sporangia, hyphal swellings and root disease loaded negatively on LV2 (Fig. S1).

Neither widespread nor generalist species clustered in any particular part of trait-space (Fig. S1), although the bottom right quadrant of trait-space contained predominantly host-specific species (Fig. S1b).

Both LVs had a negligible effect on geographic extent in all candidate trait syndrome models (Table S1). The best performing trait syndrome model of host range included only LV2 (Table S2) which always had a significant negative effect on host range. The interaction between LV1 and LV2 was not significant in models of geographic extent or host range. This is consistent with the lack of clustering of high impact species in one particular region of trait-space. The most important trait (root disease) in multiple trait models of geographic extent and host range is consistent with root disease loading negatively on LV2.



**Fig S1. Biplot of Phytophthora trait-space with species (points) coloured by the impact metrics (a) geographic extent and (b) host range with axes estimated from pure latent variable models. Open circles represent species for which impact data were not collated. Axes of trait-space were derived from multivariate latent variable models which capture trait covariance among species. The length and direction of arrows represents the strength of trait loadings onto each latent variable.**

Phylogenetic intra-class correlation in models of geographic was broadly similar between trait syndrome and individual trait-based models. Phylogenetic intra-class correlation in trait syndrome models of host range was much lower compared to multiple trait models of host range (Table S2, Fig. S1b).

LV2 was a significant predictor of host range. Lower values of LV2 were associated with root-attacking species with a combination of proliferating sporangia, hyphal swellings and greater oospore wall Fig. S1b. It is consistent with a soil-borne strategy that could be advantageous within the context of the global transport of plants-for-planting. Long-term survival structures can persist in soil and, under conducive conditions (generally warm and moist), can quickly produce many infective propagules on encountering naïve hosts. Analyses including minimum temperatures for growth also indicated covariance of cold-tolerance with root disease and oospore production. Although the trait syndromes we identified had inferior predictive power compared to multiple traits, they may still be of value for inferring ecological strategies driving *Phytophthora* invasions.

**Table S1 All candidate latent variable models of *Phytophthora* geographic extent ranked by out-of-sample predictive success using 10-fold cross-validation information criterion. Within-sample goodness-of-fit (R2) was partitioned into fixed effects (years known + traits, traits only) and phylogenetically structured residual error (phylogenetic intra-class correlation). NA values indicate the predictor was absent from the model. Asterisks indicate where the 95% credible intervals for a parameter estimate do not overlap zero.**

|  | **model rank** | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** |
| **parameter** | | | | |
| σ phylogenetic random intercept | 0.221\* | 0.237\* | 0.252\* | 0.246\* |
| years known to science | 2.276\* | 2.270\* | 2.266\* | 2.259\* |
| Latent variable 1 | NA | -0.063 | -0.058 | -0.055 |
| Latent variable 2 | 0.018 | NA | 0.001 | 0.002 |
| Latent variable 1 : Latent variable 2 | NA | NA | NA | 0.037 |
| **model performance** | | | | |
| adjusted phylogenetic intra-class correlation | 0.038 | 0.049 | 0.050 | 0.048 |
| conditional R2 (traits + phylogeny) | 0.027 | 0.032 | 0.039 | 0.045 |
| conditional R2 (years known + traits + phylogeny) | 0.529 | 0.531 | 0.533 | 0.533 |
| marginal R2 (traits only) | 0.003 | 0.004 | 0.009 | 0.015 |
| marginal R2 (years known + traits) | 0.493 | 0.490 | 0.488 | 0.492 |
| Δ IC (10-fold cross-validation) | 0.000 | 2.626 | 4.235 | 5.918 |

**Table S2 All candidate latent variable models of *Phytophthora* host range ranked by out-of-sample predictive success using 10-fold cross-validation information criterion. Within-sample goodness-of-fit (R2) was partitioned into fixed effects (years known + traits, traits only) and phylogenetically structured residual error (phylogenetic intra-class correlation). NA values indicate the predictor was absent from the model. Asterisks indicate where the 95% credible intervals for a parameter estimate do not overlap zero.** 

|  | **model rank** | | | |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** |
| **parameter** | | | | |
| σ phylogenetic random intercept | 0.234\* | 0.231\* | 0.247\* | 0.247\* |
| years known to science | 2.037\* | 2.038\* | 2.057\* | 2.042\* |
| Latent variable 1 | NA | -0.055 | -0.055 | 0.045 |
| Latent variable 2 | -0.472\* | -0.484\* | -0.484\* | NA |
| Latent variable 1 : Latent variable 2 | NA | NA | -0.072 | NA |
| **model performance** | | | | |
| adjusted phylogenetic intra-class correlation | 0.050 | 0.047 | 0.056 | 0.047 |
| conditional R2 (traits + phylogeny) | 0.111 | 0.117 | 0.129 | 0.034 |
| conditional R2 (years known + traits + phylogeny) | 0.561 | 0.564 | 0.570 | 0.497 |
| marginal R2 (traits only) | 0.077 | 0.083 | 0.092 | 0.004 |
| marginal R2 (years known + traits) | 0.519 | 0.523 | 0.523 | 0.448 |
| Δ IC (10-fold cross-validation) | 0.000 | 1.260 | 5.448 | 10.772 |

Appendix S5 Collation of global distribution data

*Phytophthora* reports were obtained from CAB Abstracts (n = 16247, CABI, 2018a) using the search term "title:(Phytophthora)". The geographic location was extracted from the CAB Abstract full record and the focal *Phytophthora* species were identified by text mining the CAB Abstract titles for species names. Country-level *Phytophthora* reports were also extracted from CABI Invasive Species Compendium datasheets (n = 387, CABI, 2018b) using the search term “Phytophthora&types=7,17,19”. Distribution data from the EPPO Global Database and EPPO Reporting Service (n = 1129, EPPO, 2018) were extracted from online distribution tables via R. Data from the GBIF database (n = 418, GBIF occurrence download <https://doi.org/10.15468/dl.dqsr1u>) was queried using the R package ‘rgbif’ (Chamberlain & Boettiger, 2017). Country-level distribution tables for *Phytophthora* species were extracted from online distribution tables at the DAISIE European Invasive Alien Species Gateway (DAISIE, 2018) and from NCBI Biosample (<https://www.ncbi.nlm.nih.gov/biosample>) records using the R package ‘rvest’ (Wickham, 2016).

Appendix S6

Bayesian multi-level models of global impacts

Phylogenetically-informed trait-based models of global impacts were specified as,

,

,

,

where g() is the appropriate link function relating the trait predictors to the impact metric on the linear predictor scale and is the expected impact of species *i*. The number of known host families for the *Phytophthora* species was modelled as a Poisson response with log link *y*i ~ *Poisson*(λi) where is the expected value on the log link scale. Geographic extent was modelled as a binomial response with logit link, where the number of trials was specified as the total number of countries reporting 1 or more *Phytophthora* species in our species-by-country occurrence database () and is the expected probability on the logit scale. is the overall intercept, is the estimated coefficient for trait *h* and is the observed value of trait *h* for species *i*. Phylogenetic signal in the residuals, is captured by a random species-level intercept with mean of zero and covariance, , derived from the phylogenetic tree. Over-dispersion relative to model assumptions can also lead to biased parameter estimates and associated standard errors (Harrison, 2014). To avoid overstating the importance of trait predictors and the amount of variance attributable to them, we included an observation-level random intercept term, , with mean 0 and covariance defined by an identity matrix. The remaining residual variance is denoted .

We specified weak normal priors with a mean of zero and a standard deviation of 10 for ,and a mean of zero and a standard deviation of 50 for . For the standard deviation of the phylogenetically structured random effects, , and the observation level random effect, , we used a weakly informative half student-t prior with 4 degrees of freedom and scale parameter of 1 to improve sampling time and to reflect that the standard deviation must be positive (Piironen & Vehtari, 2015).

We ranked the out-of-sample predictive performance of the 1284 models using ten-fold cross-validation, which randomly partitions the data into ten subsets and then refits the model ten times, each time leaving out a different subset with which to compare the observed and predicted impacts. An information criterion for each model is derived from the summed expected log pointwise predictive density (ELPD) for the predicted data (Gelman, Hwang, & Vehtari, 2014) and the effective number of parameters in the model. The ELPD is multiplied by -2 to place it on the deviance scale, akin to other information criteria like AIC and DIC.

Leave-one-out cross validation would have been preferable but was not computationally feasible, as each of the 1284 models would have to be refitted between 64 and 126 times, depending on the number of species with available trait and impact data.

Appendix S7

*Phytophthora* phylogenetic relationships

Phylogenetic non-independence in the residuals of fitted models can lead to under-estimates of confidence intervals for parameter estimates and a greater risk of type 1 errors in inference. To account for shared phylogenetic history when estimating the effects of traits on *Phytophthora* global impacts, we included a phylogenetically structured species-specific random intercept, constrained by the species-by-species phylogenetic covariance matrix. It also allows us to partition the variance explained by traits and phylogenetic relationships.

We infer species’ phylogenetic relationships from an ITS6-based phylogeny described in Barwell et al. (in. review) including all 179 species in our trait database. A single-gene phylogeny is typically less well-resolved, especially in the deeper nodes of the tree, and may have less statistical support than a phylogeny that considers phylogenetic relationships based on multiple loci (Gontcharov, Marin, & Melkonian, 2003). A version of the analyses was also performed using a multi-gene phylogeny for Phytophthora (Martin et al., 2014) obtained from Treebase (<http://purl.org/phylo/treebase/phylows/tree/TB2:Tr65776?format=nexml>). This phylogeny includes fewer of the species in our trait database, but should provide more robust estimates of the phylogenetic relationships among these species (see Appendix S3).

Multichotomies in the ITS phylogeny were resolved by transformation into a series of dichotomies each with one or more branches of length zero using the function *multi2di* in R package ape (Paradis, Claude, & Strimmer, 2004). The Martin *et al.*, (2014) phylogeny was rooted by removing one of the two outgroups. Both phylogenetic trees were converted from non-ultrametric to ultrametric, using a semi-parametric method based on penalized likelihood with a smoothing parameter of 1 (*chronopl* in R package ‘ape’: Paradis *et al.*, 2004).

Inverse covariance matrices were derived from both phylogenetic trees (*inverseA* in R package MCMCglmm: Hadfield, 2010) and then converted to covariance matrices for use in ‘brms’ models (Bürkner, 2017).

Results of analyses using a multi-gene phylogeny

The single gene phylogeny includes all species in the trait database and enabled us to incorporate all species with complete trait and impacts data into our analyses (n = 118 for geographic extent models; n = 113 for host range models). The analyses with the multi-gene phylogeny in Martin *et al.*, (2014) have substantially less power as the phylogeny includes only 48 species with complete trait data for modelling. However, the phylogenetic relationships among species should be better resolved.

Consistent with models using the single gene phylogeny, root disease and foliar disease appeared most frequently in the best-performing models of geographic extent and host range using the multi-gene phylogeny (Fig. S1). In models of host range, the signal of growth rate at optimum temperature was stronger compared to models using the single gene phylogeny, but optimum temperature for growth was a weaker predictor (Fig. S1b). This reversal of importance may be due to the covariance between these two predictors. Oospore wall index was present in four of the best-performing models, but was never significant, possibly due the restricted sample size when using the multigene phylogeny. By contrast, oospore wall index was present and significant in all but two of the best-forming models of host range using the single gene phylogeny.



**Fig. S2 Best performing trait-based models of Phytophthora (a) geographic extent and (b) host range impact metrics using a multigene phylogeny to account for shared phylogenetic history (number of observations = 48). 1280 candidate trait-based models were ranked using a 10-fold cross-validation information criterion. The subset of best models were selected based on difference in information criterion units, *Δ*IC, of less than 2 from the best model. Asterisks indicate where the 95% credible intervals for a parameter estimate do not overlap zero. Fixed effects are ordered from top to bottom by the number of times they were present in the top model subset. Grey shading indicates predictors were absent from that model. Red and blue colours indicate positive and negative parameter estimates, respectively. Deeper colours indicate stronger effect sizes. Within-sample goodness-of-fit (R2) was quantified by partitioning variance into fixed effects (years known + traits, traits only) and phylogenetically structured error (phylogenetic intra-class correlation).**

Appendix S8

Parameter estimates for phylogenetic random effects models and autocovariate regression models

Current methods for phylogenetic models fitted in the R package ‘brms’ (Bürkner, 2017) do not allow users to incorporate known phylogenetic position when predicting the response for a new species (Buerkner, pers. comm.). To circumvent this problem when using the trait-based modelling framework as a predictive tool, we instead modelled phylogenetic non-independence as a phylogenetic autocovariate (Dormann et al., 2007). We compared parameter estimates for the full models (all traits included) fitted using a phylogenetic autocovariate with those from models using a phylogenetic random intercept to account for phylogenetic non-independence to ensure they were consistent (Fig. S3).

**Fig. S3 Comparison of parameter estimates derived from models accounting for phylogenetic non-independence using a phylogenetic autocovariate and a phylogenetically structured random intercept.**

Appendix S9

Results of analyses using minimum temperature for growth as the temperature trait

Minimum temperature for growth is not well resolved for most *Phytophthora* species and is often extrapolated from refrigerator temperature and above to estimate minimum requirements for growth. Given this uncertainty, we use optimum temperature for growth as the temperature trait in the analyses in the main text. However, cold-tolerance is potentially an important adaptive trait for spread to higher latitudes. Therefore we also provide the results from analyses with minimum temperature for growth as the temperature trait.



**Fig. S4 Best performing trait-based models models of Phytophthora (a) geographic extent and (b) host range impact metrics using minimum temperature for growth instead of optimum as the temperature trait. 1280 candidate trait-based models were ranked using a 10-fold cross-validation information criterion. The subset of best models were selected based on difference in information criterion units, *Δ*IC, of less than 2 from the best model. Asterisks indicate where the 95% credible intervals for a parameter estimate do not overlap zero and indicate evidence for a significant effect of the trait in that model. Fixed effects are ordered from top to bottom by the number of times they were present in the top model subset. Grey shading indicates predictors were absent from that model. Red and blue colours indicate positive and negative parameter estimates, respectively. Deeper colours indicate stronger effect sizes. Within-sample goodness-of-fit (R2) was quantified by partitioning variance into fixed effects (years known + traits, traits only) and phylogenetically structured error (phylogenetic intra-class correlation).**

Appendix S10 Poorly predicted species from geographic extent and host range models

Poorly predicted species can be useful for understanding how future trait-based models can be improved. They may indicate a common, unmeasured trait that should be collated for future models. Poorly predicted species may also help to identify biases in the data. For example, they may represent disproportionately under- or over-reported species, globally.

**Table S3 The twenty worse predicted species from geographic extent models. Positive residuals indicate species impacts are higher than the model predicts.**

|  |  |  |  |
| --- | --- | --- | --- |
| **higher than predicted impacts** | | **lower than predicted impacts** | |
| **species** | **residuals** | **species** | **residuals** |
| Phytophthora ramorum | 3.12 | Phytophthora pini | -2.05 |
| Phytophthora x alni | 2.94 | Phytophthora erythroseptica | -1.16 |
| Phytophthora lacustris | 2.92 | Phytophthora trifolii | -1.15 |
| Phytophthora rubi | 2.92 | Phytophthora elongata | -1.01 |
| Phytophthora plurivora | 2.91 | Phytophthora aquimorbida | -1.01 |
| Phytophthora cinnamomi | 2.67 | Phytophthora humicola | -0.95 |
| Phytophthora quercina | 2.49 | Phytophthora pistaciae | -0.93 |
| Phytophthora hedraiandra | 2.49 | Phytophthora syringae | -0.93 |
| Phytophthora medicaginis | 2.01 | Phytophthora amaranthi | -0.89 |
| Phytophthora inundata | 1.91 | Phytophthora x incrassata | -0.68 |

**Table S4 The twenty worse predicted species from host range models. Positive residuals indicate species impacts are higher than the model predicts.**

|  |  |  |  |
| --- | --- | --- | --- |
| **higher than predicted impacts** | | **lower than predicted impacts** | |
| **species** | **residuals** | **species** | **residuals** |
| Phytophthora pachypleura | 3.74 | Phytophthora pini | -2.19 |
| Phytophthora cinnamomi | 3.73 | Phytophthora fragariae | -1.63 |
| Phytophthora ramorum | 3.43 | Phytophthora castaneae | -1.47 |
| Phytophthora chlamydospora | 2.31 | Phytophthora aquimorbida | -1.47 |
| Phytophthora cryptogea | 1.97 | Phytophthora syringae | -1.41 |
| Phytophthora niederhauserii | 1.95 | Phytophthora asparagi | -1.01 |
| Phytophthora multivora | 1.87 | Phytophthora amnicola | -0.93 |
| Phytophthora gregata | 1.69 | Phytophthora palmivora | -0.89 |
| Phytophthora lacustris | 1.61 | Phytophthora x heterohybrida | -0.85 |
| Phytophthora kernoviae | 1.57 | Phytophthora x incrassata | -0.80 |

**References**

Bürkner, P.-C. (2017). **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

CABI. (2018a). *CAB Direct*. Wallingford, UK: CAB International. Retrieved from https://www.cabdirect.org/cabdirect/about

CABI. (2018b). *Invasive Species Compendium*. Wallingford, UK: CAB International. Retrieved from www.cabi.org/isc.

Chamberlain, S. A., & Boettiger, C. (2017). R Python, and Ruby clients for GBIF species occurrence data. *PeerJ PrePrints*. https://doi.org/10.7287/peerj.preprints.3304v1

DAISIE. (2018). DAISIE European Invasive Alien Species Gateway, http://www.europe-aliens.org/. Retrieved from http://www.europe-aliens.org/

Dormann, C. F., Mcpherson, J. M., Arau, M. B., Bivand, R., Bolliger, J., Carl, G., … Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data : a review. *Ecography*, *30*, 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

EPPO. (2018). EPPO Global database, https://gd.eppo.int/. Retrieved from https://gd.eppo.int/

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gontcharov, A. A., Marin, B., & Melkonian, M. (2003). Are Combined Analyses Better Than Single Gene Phylogenies? A Case Study Using SSU rDNA and rbcL Sequence Comparisons in the Zygnematophyceae (Streptophyta). *Molecular Biology and Evolution*, *21*(3), 612–624. https://doi.org/10.1093/molbev/msh052

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, *33*, 1–22.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616. https://doi.org/10.7717/peerj.616

Hui, F. K. C. (2016). boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution*, *7*(6), 744–750. https://doi.org/10.1111/2041-210X.12514

Martin, F. N., Blair, J. E., & Coffey, M. D. (2014). A combined mitochondrial and nuclear multilocus phylogeny of the genus *Phytophthora*. *Fungal Genetics and Biology*, *66*. https://doi.org/10.1016/j.fgb.2014.02.006

Paradis, E., Claude, J., & Strimmer, K. (2004). Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290.

Piironen, J., & Vehtari, A. (2015). *Projection predictive variable selection using Stan+R*. Retrieved from https://groups.google.com/d/msg/stan-dev/

Wickham, H. (2016). rvest: Easily Harvest (Scrape) Web Pages. Retrieved from https://cran.r-project.org/package=rvest