1-1-2020

Taxonomic resolution affects host-parasite association model performance

Tad A. Dallas
Louisiana State University

Daniel J. Becker
University of Oklahoma

Recommended Citation

Dallas, T., & Becker, D. (2020). Taxonomic resolution affects host-parasite association model performance. *Parasitology* https://doi.org/10.1017/S0031182020002371

Parasitology

cambridge.org/par

Taxonomic resolution affects host–parasite association model performance

Tad A. Dallas[1] and Daniel J. Becker[2]

[1]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70802, USA and [2]Department of Biology, University of Oklahoma, Norman, OK 73019, USA

Research Article

Cite this article: Dallas TA, Becker DJ (2011). Taxonomic resolution affects host–parasite association model performance. *Parasitology* **148**, 584–590. https://doi.org/10.1017/S0031182020002371

Received: 19 October 2020
Revised: 7 December 2020
Accepted: 9 December 2020
First published online: 21 December 2020

Key words:
Boosted regression trees; parasite macroecology; phylogenetic scale

Author for correspondence:
Tad A. Dallas, E-mail: tad.a.dallas@gmail.com

Abstract

Identifying the factors that structure host–parasite interactions is fundamental to understand the drivers of species distributions and to predict novel cross-species transmission events. More phylogenetically related host species tend to have more similar parasite associations, but parasite specificity may vary as a function of transmission mode, parasite taxonomy or life history. Accordingly, analyses that attempt to infer host–parasite associations using combined data on different parasite groups may perform quite differently relative to analyses on each parasite subset. In essence, are more data always better when predicting host–parasite associations, or does parasite taxonomic resolution matter? Here, we explore how taxonomic resolution affects predictive models of host–parasite associations using the London Natural History Museum's database of host–helminth interactions. Using boosted regression trees, we demonstrate that taxon-specific models (i.e. of Acanthocephalans, Nematodes and Platyhelminthes) consistently outperform full models in predicting mammal-helminth associations. At finer spatial resolutions, full and taxon-specific model performance does not vary, suggesting tradeoffs between phylogenetic and spatial scales of analysis. Although all models identify similar host and parasite covariates as important to such patterns, our results emphasize the importance of phylogenetic scale in the study of host–parasite interactions and suggest that using taxonomic subsets of data may improve predictions of parasite distributions and cross-species transmission. Predictive models of host–pathogen interactions should thus attempt to encompass the spatial resolution and phylogenetic scale desired for inference and prediction and potentially use model averaging or ensemble models to combine predictions from separately trained models.

Introduction

Host–parasite associations are structured by complex and interrelated constraints, including geographic range overlap, evolutionary relatedness and life-history traits (Cooper *et al.*, 2012*b*; Dallas *et al.*, 2016; Olival *et al.*, 2017; Albery *et al.*, 2020). Parasite species do not interact with a random subset of available host species but instead are involved in complex tradeoffs related to the number of host species they can potentially infect relative to the efficacy with which they can exploit host resources for survival, growth and reproduction (Krasnov *et al.*, 2004; Leggett *et al.*, 2013). Estimating the factors that shape host–parasite associations is of fundamental importance to parasitologists and ecologists, as understanding the drivers and distribution of species diversity is a central aim of ecology. From an applied perspective, predicting host–parasite associations is an increasing priority, given the global homogenization of biota and changing land use (Dornelas *et al.*, 2014, 2019; Borremans *et al.*, 2019) that brings novel host species – including humans – into close contact.

Host species that are more similar to one another tend to have more similar parasite communities (Dallas and Poisot, 2018), and more phylogenetically similar parasite species tend to infect a phylogenetically similar set of host species (Dallas *et al.*, 2016). These simple phylogenetic rules suggest that data on other host and parasite species can inform predictive models of host–parasite associations. However, host–parasite associations may be formed dependent on parasite type, meaning that host traits important to one parasite group could have null or opposing effects on another parasite group. For example, if a gastrointestinal parasite and an ectoparasite have similar host species, we can use this to predict potential hosts for either parasite; however, we presume these two parasites are responding in a similar manner to some qualities of those host species. Although the host community of the two parasites may be quite similar, the underlying responsible host life-history characteristics are likely quite different. Aspects of skin thickness may be important in determining infection success for ectoparasites (Moorhouse and Tatchell, 1966) but may be irrelevant for a gastrointestinal parasite reliant on food-borne transmission. Such an example is more broadly illustrative of the encounter and compatibility filters that different parasites experience when infecting distinct host species (Combes, 2001).

The complexities of parasite life history and parasite specificity create an odd tradeoff for the development of predictive models. On one hand, the inclusion of more records of host–parasite association may allow better estimating the importance of host or parasite traits that result in a given interaction between host and parasite species. For example, a fast life

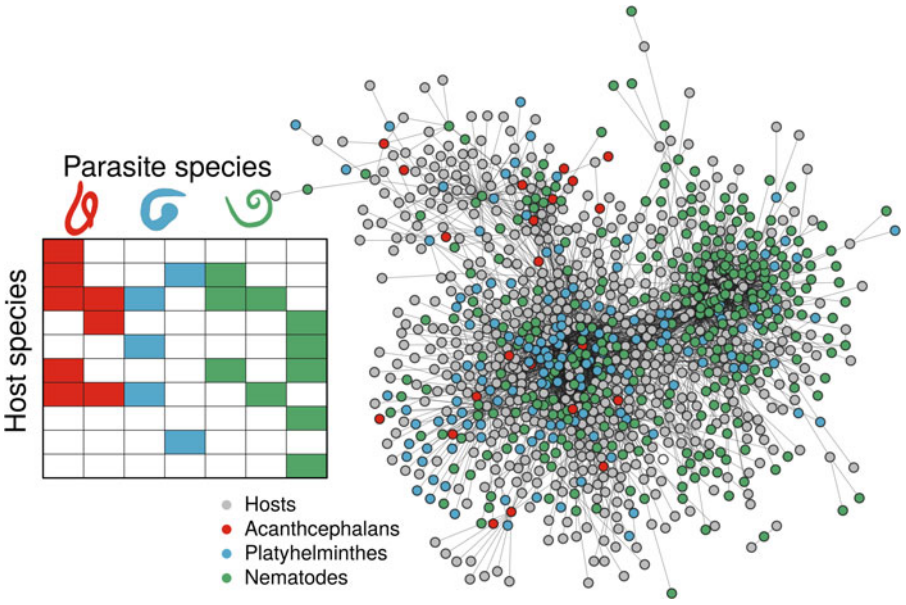

Fig. 1. Associations of different helminth parasite taxa (indicated with colour) and their mammal host species, using both a simple representation of the interaction matrix (left panel), and the real data as a network plot (right panel), where nodes represent host or parasite species (indicated by colour) and links between them represent instances of recorded host–helminth interactions. Host–helminth associations were modeled as a function of both host and helminth variables, using either all the data available or data on a specific helminth taxon (Platyhelminthes, Acanthocephalans or Nematodes). By using data on all associations, it may expand the available host and helminth covariate space, enhancing the discrimination capacity of the model. However, by subsetting to specific parasite taxa, we constrain the host and helminth covariate space to include only the most relevant information to the modeling task.

history strategy is one of the key traits shaping the richness of all zoonotic pathogens in rodents (Han *et al.*, 2015). Similarly, even in trait-free approaches, more associations could help to more precisely estimate the similarity between host species or parasite species in their interaction with one another. This suggests that including more data would improve model performance. On the other hand, the inclusion of parasite taxa that are distributed independently – or differently than other parasite taxa – across host species could reduce model performance by adding noise to any host trait or phylogenetic signal present in each parasite taxon. Because traits of hosts or parasites have evolved along lineages, life history may be confounded with taxonomy in models using all possible host–parasite associations (Washburne *et al.*, 2018; Albery and Becker, 2020), and this challenge may not simply be overcome by including taxonomy as a predictor. Lastly, a model that includes all data could perform well at predicting higher-order associations between host and parasite species but perform worse relative to a more taxon-specific model. Although ecology and evolution more broadly have increasingly considered how 'phylogenetic scale' shapes general patterns of species distributions (Cruz *et al.*, 2005; Cavender-Bares *et al.*, 2006), consideration of how host–parasite associations vary across or within taxonomic ranks remains unexplored. However, such consideration of the phylogenetic scale of host and/or parasite species could have important implications for predictive models of infectious disease.

Here, we explore how predictive models of host–parasite associations differ in terms of performance and the importance of both host and parasite traits when we compare models leveraging the full range of parasites to taxon-specific models (Fig. 1). To do this, we used the London Natural History Museum's (LNHM) database of host–helminth associations (Gibson *et al.*, 2005), the most extensive database of host–parasite interaction data currently in existence. Restricting our analyses to all mammals and their helminth parasites, we developed predictive models of helminth associations, comparing the full model to helminth taxonomic subsets consisting of Acanthocephalans, Nematodes and Platyhelminthes. We find taxon-specific models consistently outperformed the full model in predicting mammal-helminth associations, despite the taxon-specific submodels generally identifying the same set of important life-history characteristics of host and parasite species.

Methods

Host–helminth interaction data

Records of helminth parasite occurrences on host species were obtained from the parasite database of the LNHM (Gibson *et al.*, 2005) and accessed programmatically using the *helminthR* package (Dallas, 2016). These data currently represent one of the largest sources of host–parasite interaction data (Gibson *et al.*, 2005; Dallas *et al.*, 2018), despite being restricted to helminth parasites. Host–helminth interaction data are georeferenced only to the geopolitical location (e.g. France), large water body (e.g. Lake Michigan), or coast (e.g. coast of Argentina) for over 400 terrestrial and aquatic locations. We removed host–helminth interactions from locations that were aquatic/marine, overly vague, or locations nested within other locations (e.g. 'Western Europe'). This resulted in data on a total of 555 host species and their associations with 151 Platyhelminthes (trematodes and cestodes), 22 Acanthocephalans and 327 Nematodes (Fig. 1). LNHM data are reported per host–helminth species and are not stratified by parasite life stage, which could obscure how different hosts maintain helminths across complex life cycles.

From these data, we considered three different geographic scales of analysis; global, USA and a state within the USA (Texas). Geographic scale influences the number of host and helminth species, limiting the available interaction and covariate space that the model is trained upon and thus possibly further affecting the performance of helminth taxonomic subgroup models relative to the full model. The amount of data lost by focusing on smaller geographic scales was pronounced, with the number of host species decreasing from 555 in the global network to 122 in the US network and a mere 18 in the state of Texas. Similar patterns were observed for parasite richness, which decreased from 500 helminth species in the global network to 234 in the USA and 62 in the state of Texas. Note that these numbers reflect those helminth species for which we had parasite trait data available and not the absolute diversity of helminth parasites in a given location.

Host and parasite species variables

Host species variables were obtained from PanTHERIA (Jones *et al.*, 2009), restricting the analyses to mammal host species.

https://doi.org/10.1017/S0031182020002371

These data included covariates describing host morphology, life history, geography and taxonomy (see Supplemental Material), which are generally considered as aspects of host species that affect host–parasite interactions (Sears *et al.*, 2015; Dallas *et al.*, 2016) and zoonotic spillover potential (Olival *et al.*, 2017). We only considered host variables with more than 80% data coverage, resulting in a total of 19 host covariates. Further, host or parasite species for which no covariate data were available were removed from the analysis. The fraction of missing data ranged from 9% for species adult body mass to 15% for species litter size. Host data were obtained independently from the interaction data, such that data on geography and environmental covariates are estimated from host spatial distribution data independent of the host occurrences at geopolitical scales in the LNHM data.

Parasite variables were compiled from the species description of each parasite from five helminth taxonomy references (Petrochenko and Skrjabin, 1971; Yamaguti, 1971; Levine, 1980; Crompton and Nickol, 1985; Anderson, 2000). These data were aggregated by Dr Alyssa Gehman and have been published previously elsewhere (Dallas *et al.*, 2019). A total of 18 helminth parasite covariates were considered, including taxonomic information (phylum and class), infection sites (e.g. 'intestine'), and numerous measures of helminth morphology (e.g. length and width of different life stages). A full list of host and parasite variables is provided in Supplemental Materials.

Boosted regression tree models

Boosted regression tree (BRT) models were used to estimate the suitability of an association between mammal host and helminth parasite species using the *gbm* R package (Ridgeway, 2006). This is a flexible regression approach that allows for non-linear responses, variable interactions and missing data (Elith *et al.*, 2008). The response we modelled is a binary variable representing the existence of a known interaction between a host and helminth species, which is determined by the combination of aspects of both host and helminth species; absences are those host–parasite interactions not observed in the LNHM dataset. BRTs are well suited for capturing complex patterns in data and for identifying important predictor variables and they often outperform parametric models (e.g. GLMs; Pichler *et al.*, 2020). However, we acknowledge that other predictive approaches (e.g. random forests) might generate different results.

We trained models separately for each parasite taxonomic phylum (Platyhelminthes, Acanthocephalans and Nematodes), and for all helminth parasites combined. Accordingly, parasite phylum was included in the full model but not in the taxonomic subset models. Further, we also trained models for the entire global host–helminth interaction data and for geographic subsets of the USA and the state of Texas. These geographic locations were selected out of convenience and data availability. It is important to note that this geographic restriction affects model performance in two different ways. First, models trained on fewer interactions will likely be less accurate. Second, models trained on fewer species will likely be unable to determine the host and parasite traits most important to the link prediction task. That is, geographic extent, in a large part, is simply a measure of data quantity.

For each parasite group and geographic scale combination, we trained 50 BRT models on 80% of the available data and estimated interaction suitability on the remaining 20% test set. Models were trained using a maximum of 50 000 trees, with a learning rate of 0.001 (Elith *et al.*, 2008), binomial error structure and an interaction depth of 3, which allows for interactions between covariates. All models were internally cross-validated (5-fold) to determine the optimal number of regression trees.

Comparing model performance, predictions, and variable importance

Model performance was estimated using area under the curve (AUC; Bradley, 1997), accuracy (Sing *et al.*, 2005), and the true skill statistic (TSS; Allouche *et al.*, 2006). We measured accuracy as the maximum fraction of true positive and true negative values in the test set divided by the total number of cases, with decision threshold separating the cases set to maximize accuracy. Each measure of model performance was calculated for each model, resulting in a total of 50 estimates of model performance per parasite phyla submodel and full model. Further, predictions from the full model were subset by parasite taxa, in order to evaluate the ability of the full model to capture taxa-specific variation in host–helminth interactions. We used Welch's two-sample *t*-tests to compare model performance between the full model and each parasite taxa subset.

Estimated host–helminth interaction relative suitability values for each submodel were related to the relative suitability values for the same interactions generated from the full model. This was performed in order to examine the ability of each model to accurately distinguish host–helminth parasite interactions using all host–helminth association data available, or a perhaps more relevant helminth taxonomic subset of interactions.

Finally, we estimated average variable importance measures across models trained on parasite taxonomic subsets and the full model to determine if restricting parasite taxonomic resolution changes the relative importance of variables used in predictive models. Variable importance was estimated for each model by quantifying the relative improvement to model fit as a result of the inclusion of a given covariate into the model, weighted by the number of trees in which the covariate occurred (De'Ath, 2007; Elith *et al.*, 2008). These values are then scaled to sum to 100, with larger numbers corresponding to higher variable importance (Elith *et al.*, 2008).

R code and data to reproduce the analyses is provided at https://figshare.com/s/bfcc1fd78168edebd09f.

Results

Model performance

BRT models performed well when trained on the full dataset ($\overline{AUC} = 0.90 \pm 0.005$; $\overline{Accuracy} = 0.87 \pm 0.005$; $\overline{TSS} = 0.64 \pm 0.01$), Acanthocephalans ($\overline{AUC} = 0.89 \pm 0.027$; $\overline{Accuracy} = 0.91 \pm 0.021$; $\overline{TSS} = 0.69 \pm 0.06$), Nematodes ($\overline{AUC} = 0.92 \pm 0.006$; $\overline{Accuracy} = 0.90 \pm 0.005$; $\overline{TSS} = 0.71 \pm 0.02$), and Platyhelminthes ($\overline{AUC} = 0.85 \pm 0.011$; $\overline{Accuracy} = 0.83 \pm 0.009$; $\overline{TSS} = 0.55 \pm 0.03$). This would suggest that all models performed decently well at estimating out-of-sample host–helminth associations. However, when we considered the ability of the full model to estimate host–helminth associations for each helminth taxonomic subgroup, we found that helminth parasite subsets generally performed better than the full model (Table 1 and Fig. 2), except for platyhelminth parasites when model performance was estimated using AUC and TSS. This effect was sensitive to geographic scale – and subsequently the amount of data – as parasite submodels generally performed no different from full models when predicting host–helminth interactions in the USA and in the state of Texas (Fig. 2). However, with the exception of nematodes within Texas when performance was estimated using accuracy, the full model never significantly outperformed a parasite taxon-specific model at either scale (Tables S3 and S4).

Variable importance

Interestingly, the full model and taxonomic subset models tended to agree on which host and helminth parasite covariates were

https://doi.org/10.1017/S0031182020002371

Table 1. Model performance – quantified using AUC, accuracy and TSS – declined when the full model was used to predict on helminth taxonomic subsets when considering the global set of interactions between hosts and helminth parasites

Performance	Helminth taxa	*t*	df	*P* value
AUC	Acanthcephalans	4.65	92.4	**0.0001**
	Platyhelminthes	0.76	97.8	0.45
	Nematodes	1.79	97.5	0.077
Accuracy	Acanthcephalans	10.44	94.5	**0.0001**
	Platyhelminthes	2.96	94.9	**0.003**
	Nematodes	2.00	96.4	**0.049**
TSS	Acanthcephalans	5.44	97.5	**0.0001**
	Platyhelminthes	1.21	97.1	0.229
	Nematodes	2.28	97.8	**0.025**

Model performance was compared to helminth group subset predictions from the full model using Welch's two-sample *t*-tests across the 50 trained boosted regression models. Bold P-values indicate significance assessed at $\alpha = 0.05$.

most important to model performance – regardless of geographic scale considered (Figs S2 and S3). We found that host family, the dominant infection site of the helminth parasite, and the maximum latitude of the host species consistently appeared in the top five predictor variables (Fig. 3a). Parasite class had relatively weaker importance across models and parasite phylum was unimportant when included in the full model ($\overline{\text{importance}} = 0.04$). Ranking the host and parasite variables by their relative importance and calculating rank correlations across different models further supports the conclusion that models tended to have similar variable importance values (Fig. 3b). Further, models trained at different geographic scale found the same variables were important, based on Pearson's correlations of mean variable importance values comparing the global model to submodels of host–helminth interactions within the USA and the state of Texas (Fig. S1). Finally, the relative importance of host and helminth parasite species covariates to model performance was generally balanced across the different trained models (Fig. 3c). Similarly, although effects of predictors such as host family and dominant

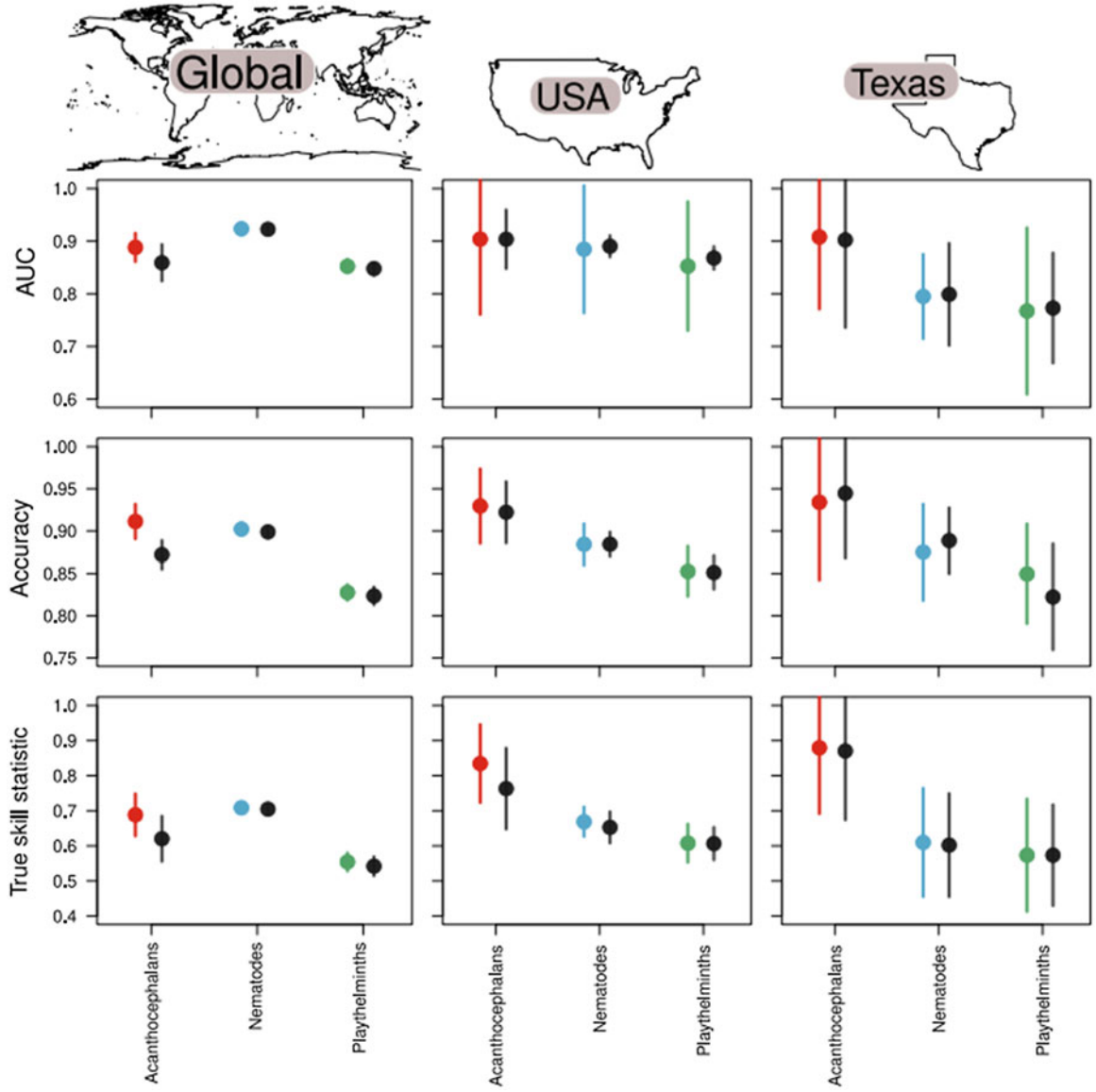

Fig. 2. The full model (black points) performed worse than each taxa-specific helminth submodel (coloured points) in terms of discrimination ability (AUC), accuracy and TSS when considering the global model (left column). Here, values closer to 1 indicate improved model performance. The relative improvement of taxa-specific models over full models declines as the geographic scale considered becomes smaller, evidenced by the models trained on host–helminth interactions from the USA (middle column) and a state within the USA (Texas; right column). This suggests that both taxonomic and geographic scale of host–parasite associations are important to consider when developing predictive models.

https://doi.org/10.1017/S0031182020002371

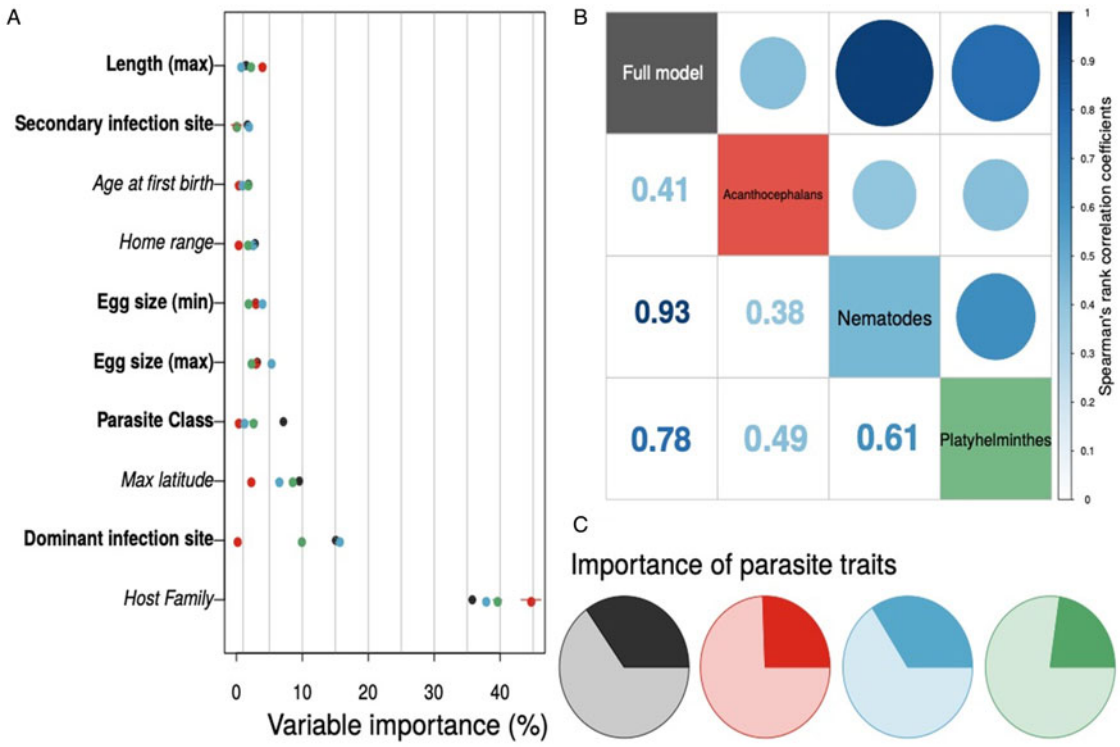

Fig. 3. Variable importance for each global host–helminth interaction model – with helminth taxonomic group denoted by point colour – tended to be conserved, with host family and the site of infection as dominant predictors across models (panel a; host variables are italicized, helminth parasite covariates are bolded; only the top 10 predictor variables are shown here). The rank order of mean variable importance tended to be positively correlated among models as well (panel b). Finally, while important variables tended to be the same across models, the relative importance of helminth parasite covariates (darker colours in the pie charts in panel c) compared to host covariates (lighter shaded regions) did show variation.

infection site were difficult to distinguish owing to a large number of levels, we did find that maximum latitude was a consistent indicator of host–parasite associations across models (Fig. S4).

Discussion

Predictive models can help identify aspects of species that shape host–parasite associations and forecast likely interactions, yet their inference may depend upon phylogenetic scale. Here, all trained BRT models performed fairly well at estimating host–helminth associations. However, by targeting specific helminth groups, taxonomically restricted models outperformed the full model almost universally, even though this full model included parasite taxonomy (class and phylum) as predictors. This suggests that using taxonomic subsets specific to parasite groups of direct interest may lead to more accurate predictions. This is counter to the idea that including a diverse set of parasite taxa might enhance model performance by training the model on a broader range of traits and associations (Wisz *et al.*, 2008). That is, a model might be able to detect higher-level variation in host–parasite associations, leveraging information across parasite taxa, leading to enhanced predictive performance. Interestingly, despite the sizable differences in helminth taxa in body size and life history, models trained on helminth taxonomic subsets identified many of the same important variables (Fig. 3). This suggests that the same set of host and helminth covariates were important to estimating host–parasite associations, despite the differences in the ability to rank host–helminth associations among models. Together, this suggests predictive models should attempt to encompass the phylogenetic scale desired for inference and potentially use model averaging or ensemble models to combine predictions from separately trained models.

The importance of taxonomic scale (or phylogenetic scale more generally) is increasingly acknowledged in ecology and evolution (Graham *et al.*, 2018; Washburne *et al.*, 2019) and likely also has important implications for studies of host–parasite interactions. Our results suggest that inference about particular taxonomic groups is maximized for models trained on that particular host or parasite clade. Other related work has shown that the zoonotic potential of viruses or host species differs by the phylogenetic scale considered and, in some cases, restricting models by particular taxa can alter findings of trait-based analyses (Washburne *et al.*, 2018; Crowley *et al.*, 2020). However, such analyses have not been extended to broad bipartite host–parasite associations and thus questions of host or parasite specificity. Closer attention to the phylogenetic scale could have important implications for broader models of parasite sharing. Within particular host–parasite systems, such as bacterial pathogens in rodents (Withenshaw *et al.*, 2016) and bats (Becker *et al.*, 2020), finer resolutions of parasite sharing (e.g. using parasite genetics) can reveal covert host specificity. Similar taxon-specific approaches to parasite sharing could assess how the centrality of host species to transmission networks varies based on particular parasite groups or more conservative parasite species resolutions.

The strong specificity and co-evolutionary relationships present in host–helminth associations may ultimately lead to taxonomic subset models outperforming a full model, which would be especially true if fine-scale parasite species life-history variation was key to determining host–parasite associations for given parasite taxa. The importance of taxonomic resolution and such life-history traits was evident through the predominance of the host family and dominant helminth infection site as key predictive covariates across helminth taxonomic subset models. Given the overall consistent finding that parasite sharing is generally restricted by phylogenetic processes (Streicker *et al.*, 2010;

https://doi.org/10.1017/S0031182020002371

Cooper *et al*., 2012*a*; Albery *et al*., 2020; Shaw *et al*., 2020), taxonomic subset models may more broadly improve predictions about host–parasite interactions. In our analysis, we stratified models by helminth phyla (Acanthocephala, Nematoda and Platyhelminthes); however, other taxonomic resolutions may be desirable, especially for parasites that display finer-scale lineage or genotype variation. In general, there is no single scale at which ecological phenomena such as host–parasite interactions should be studied (Levin, 1992), although several tools are increasingly available to identify the most important phylogenetic scales for describing such data (Washburne *et al*., 2019). Application of methods such as phylogenetic factorization can allow researchers to identify clades that best capture variation in host–parasite interactions and accordingly apply predictive models to those taxa.

While phylogenetic scale is an important consideration when enough data are available, smaller geographic scales – consisting of a subset of the global host–helminth network – may not predispose taxonomic subset models to perform better than the full model. This was evident in our analyses of host–helminth interactions at two finer geographic scales (i.e. country and state). Here, despite selecting the same host and helminth covariates in top-performing models, the difference in performance between the full model and helminth taxonomic subset models was essentially null. This still argues against the idea that more data generate a better model, which would require the full model to instead outperform the helminth group-specific models; however, full models generally did not show improved performance regardless of phylogenetic or geographic scale. Together, this finding highlights the importance of both phylogenetic *and* geographic scale when constructing predictive models of species interactions. Additionally, this result suggests that restricting analyses phylogenetically trades off in terms of model performance by reducing the amount of available data.

Although host family and parasite infection site were key predictors across both full and taxonomic subset models, we also identified a host's maximum latitude as a general predictive variable. The effects of this covariate were largely consistent across models (Fig. S4), with a greater likelihood of host–parasite association at higher latitudes. This pattern supports previously observed latitudinal gradients in parasite richness further from the equator (Lindenfors *et al*., 2007) or could reflect geographic biases in sampling (Dallas *et al*., 2018). Alternatively, greater likelihood of host–parasite association could represent possibly weaker resistance of hosts to infection at more extreme geographic range margins (Becker *et al*., 2019), thereby facilitating parasite establishment (Briers, 2003).

While the LNHM's helminth data represent one of the most extensive host–parasite databases to date, especially considering the taxonomic scope of the parasites considered (Gibson *et al*., 2005; Dallas *et al*., 2018), it is important to acknowledge that the recorded interactions between host and helminth parasite are not exhaustive. Host species that are more abundant, more conspicuous, or easier to sample may be over-represented in the host–helminth association data (Carlson *et al*., 2020). If this differential sampling is associated with host taxonomy, this bias could inflate the importance of host taxonomy in estimating host–helminth associations. However, it is difficult to imagine a situation where this would lead to the full model consistently performing worse than models on taxonomic subsets of parasites.

Our finding on the importance of taxonomic resolution may extend to models aimed at estimating associations between species more broadly. For instance, if the importance of phylogenetic scale translates to other systems, then the understanding and prediction of plant-pollinator, consumer-resource and site–species interactions may all be affected by taxonomic resolution. One interesting outcome of this is that training models on particular subsets of interactors in these networks may improve model performance. With the increased availability of data on species interactions, it may be tempting to include all data on species interactions in a model, under the idea that more information will improve the model's ability to discriminate (van Proosdij *et al*., 2016). Similar ideas are developing in species distribution modeling, where joint species distribution models use community-scale data to forecast species distributions by leveraging information on species shared environmental responses. However, the appropriate scale at which to subset species or habitats to enhance model predictive performance is presently unknown but a pressing research need.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0031182020002371

Data. R code is available on figshare at https://figshare.com/s/bfcc1fd78168edebd09f.

Author contributions. TAD performed the analysis. All authors contributed to manuscript writing.

Financial support. This work was performed under the Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with support of the EC Research Innovation Action under the H2020 Programme; in particular, the authors gratefully acknowledge the support of the Barcelona Supercomputing Centre. This work was also supported by funding to the Viral Emergence Research Initiative (VERENA) consortium including NSF BII 2021909. Lastly, the Macroecology of Infectious Disease Research Coordination Network (funded by NSF/NIH/USDA DEB 131223) provided useful discussions and support for this study.

Conflict of interest. The authors have no conflicts of interest to declare.

Ethical standards. Not applicable.

References

Albery GF and Becker DJ (2020) Fast-lived hosts and zoonotic risk. *Trends in Parasitology*.

Albery GF, Eskew EA, Ross N and Olival KJ (2020) Predicting the global mammalian viral sharing network using phylogeography. *Nature Communications* **11**, 1–9.

Allouche O, Tsoar A and Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* **43**, 1223–1232.

Anderson RC (2000) *Nematode Parasites of Vertebrates: Their Development and Transmission*. Wallingford, Oxon (UK): Cabi Publishing.

Becker DJ, Nachtmann C, Argibay HD, Botto G, Escalera-Zamudio M, Carrera JE, Tello C, Winiarski E, Greenwood AD, Méndez-Ojeda ML, Loza-Rubio E, Lavergne A, de Thoisy B, Czirják GÁ, Plowright RK, Altizer S and Streicker DG (2019) Leukocyte profiles reflect geographic range limits in a widespread neotropical Bat. *Integrative and Comparative Biology* **59**, 1176–1189.

Becker DJ, Speer KA, Brown AM, Fenton MB, Washburne AD, Altizer S, Streicker DG, Plowright RK, Chizhikov VE, Simmons NB and Volokhov DV (2020) Ecological and evolutionary drivers of haemoplasma infection and bacterial genotype sharing in a neotropical bat community. *Molecular Ecology* **29**, 1534–1549.

Borremans B, Faust C, Manlove KR, Sokolow SH and Lloyd-Smith JO (2019) Cross-species pathogen spillover across ecosystem boundaries: mechanisms and theory. *Philosophical Transactions of the Royal Society B* **374**, 20180344.

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30**, 1145–1159.

Briers RA (2003) Range limits and parasite prevalence in a freshwater snail. *Proceedings of the Royal Society of London B: Biological Sciences* **270**, S178–S180.

Carlson CJ, Dallas TA, Alexander LW, Phelan AL and Phillips AJ (2020) What would it take to describe the global diversity of parasites? *Proceedings of the Royal Society B: Biological Sciences* **287**, 20201841.

Cavender-Bares J, Keen A and Miles B (2006) Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. *Ecology* **87**, S109–S122.

Combes C (2001) *Parasitism: The Ecology and Evolution of Intimate Interactions*. Chicago, IL: University of Chicago Press.

Cooper N, Griffin R, Franz M, Omotayo M and Nunn CL (2012a) Phylogenetic host specificity and understanding parasite sharing in primates. *Ecology Letters* **15**, 1370–1377.

Cooper N, Kamilar JM and Nunn CL (2012b) Host longevity and parasite species richness in mammals. *PLoS ONE* **7**, e42190.

Crompton DWT and Nickol BB (1985) *Biology of the Acanthocephala*. Cambridge, UK: Cambridge University Press.

Crowley D, Becker D, Washburne A and Plowright R (2020) Identifying suspect Bat reservoirs of emerging infections. *Vaccines* **8**, 228.

Cruz FB, Fitzgerald LA, Espinoza RE and Ii JAS (2005) The importance of phylogenetic scale in tests of Bergmann's and Rapoport's rules: lessons from a clade of South American lizards. *Journal of Evolutionary Biology* **18**, 1559–1574.

Dallas T (2016) Helminthr: an R interface to the London Natural History Museum's host–parasite database. *Ecography* **39**, 391–393.

Dallas T and Poisot T (2018) Compositional turnover in host and parasite communities does not change network structure. *Ecography* **41**, 1534–1542.

Dallas T, Park AW and Drake JM (2016) Predictability of helminth parasite host range using information on geography, host traits and parasite community structure. *Parasitology* **144**, 1–6. doi: 10.1017/S0031182016001608

Dallas TA, Aguirre AA, Budischak S, Carlson C, Ezenwa V, Han B, Huang S and Stephens PR (2018) Gauging support for macroecological patterns in helminth parasites. *Global Ecology and Biogeography* **27**, 1437–1447.

Dallas T, Gehman A-LM, Aguirre AA, Budischak SA, Drake JM, Farrell MJ, Ghai R, Huang S and Morales-Castilla I (2019) Contrasting latitudinal gradients of body size in helminth parasites and their hosts. *Global Ecology and Biogeography* **28**, 804–813.

De'Ath G (2007) Boosted trees for ecological modeling and prediction. *Ecology* **88**, 243–251.

Dornelas M, Gotelli NJ, McGill B, Shimadzu H, Moyes F, Sievers C and Magurran AE (2014) Assemblage time series reveal biodiversity change but not systematic loss. *Science (New York, N.Y.)* **344**, 296–299.

Dornelas M, Gotelli NJ, Shimadzu H, Moyes F, Magurran AE and McGill BJ (2019) A balance of winners and losers in the Anthropocene. *Ecology Letters* **22**, 847–854.

Elith J, Leathwick JR and Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813.

Gibson DI, Bray RA and Harris EA (2005) *Host-parasite Database of the Natural History Museum, London*.

Graham CH, Storch D and Machac A (2018) Phylogenetic scale in ecology and evolution. *Global Ecology and Biogeography* **27**, 175–187.

Han BA, Schmidt JP, Bowden SE and Drake JM (2015) Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences* **112**, 7039–7044.

Jones KE, Bielby J, Cardillo M, Fritz SA, O'Dell J, Orme CDL, Safi K, Sechrest W, Boakes EH, Carbone C, Connolly C, Cutts MJ, Foster JK, Grenyer R, Habib M, Plaster CA, Price SA, Rigby EA, Rist J, Teacher A, Bininda-Emonds ORP, Gittleman JL, Mace GM, Purvis A and Michener WK (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**, 2648–2648.

Krasnov BR, Poulin R, Shenbrot GI, Mouillot D and Khokhlova IS (2004) Ectoparasitic "jacks-of-all-trades": relationship between abundance and host specificity in fleas (Siphonaptera) parasitic on small mammals. *The American Naturalist* **164**, 506–516.

Leggett HC, Buckling A, Long GH and Boots M (2013) Generalism and the evolution of parasite virulence. *Trends in Ecology & Evolution* **28**, 592–596.

Levin SA (1992) The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* **73**, 1943–1967.

Levine ND (1980) *Nematode Parasites of Domestic Animals and of Man*, 2nd Edn, Minneapolis, USA: Burgess Publishing Co.

Lindenfors P, Nunn CL, Jones KE, Cunningham AA, Sechrest W and Gittleman JL (2007) Parasite species richness in carnivores: effects of host body mass, latitude, geographical range and population density. *Global Ecology and Biogeography* **16**, 496–509.

Moorhouse DE and Tatchell RJ (1966) The feeding processes of the cattle-tick *Boophilus microplus* (Canestrini): a study in host-parasite relations: Part I. Attachment to the host. *Parasitology* **56**, 623–631.

Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL and Daszak P (2017) Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650.

Petrochenko V and Skrjabin K (1971) *Acanthocephala of Domestic and Wild Animals Vol. 1–2*. Jerusalem: Israel Program for Scientific Translations.

Pichler M, Boreux V, Klein A-M, Schleuning M and Hartig F (2020) Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution* **11**, 281–293.

Ridgeway G (2006) *gbm: Generalized Boosted Regression Models*. R package version 1.

Sears BF, Snyder PW and Rohr JR (2015) Host life history and host–parasite syntopy predict behavioural resistance and tolerance of parasites. *Journal of Animal Ecology* **84**, 625–636.

Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C and Balloux F (2020) The phylogenetic range of bacterial and viral pathogens of vertebrates. *Molecular Ecology* **29**, 3361–3379.

Sing T, Sander O, Beerenwinkel N and Lengauer T (2005) ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)* **21**, 3940–3941.

Streicker DG, Turmelle AS, Vonhof MJ, Kuzmin IV, McCracken GF and Rupprecht CE (2010) Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science (New York, N.Y.)* **329**, 676–679.

van Proosdij AS, Sosef MS, Wieringa JJ and Raes N (2016) Minimum required number of specimen records to develop accurate species distribution models. *Ecography* **39**, 542–552.

Washburne AD, Crowley DE, Becker DJ, Olival KJ, Taylor M, Munster VJ and Plowright RK (2018) Taxonomic patterns in the zoonotic potential of mammalian viruses. *PeerJ* **6**, e5979.

Washburne AD, Silverman JD, Morton JT, Becker DJ, Crowley D, Mukherjee S, David LA and Plowright RK (2019) Phylofactorization: a graph partitioning algorithm to identify phylogenetic scales of ecological data. *Ecological Monographs* **89**, e01353. doi: 10.1002/ecm.1353

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A and Group NPSDW (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions* **14**, 763–773.

Withenshaw SM, Devevey G, Pedersen AB and Fenton A (2016) Multihost Bartonella parasites display covert host specificity even when transmitted by generalist vectors. *Journal of Animal Ecology* **85**, 1442–1452.

Yamaguti S (1971) *Synopsis of Digenetic Trematodes of Vertebrates*. Vols I and II. Tokyo, Japan: Keigaku Publishing Co.