What controls the range of hosts a fish parasite infects?

Tad Dallas[1,2], Andrew Park [1], and John M. Drake[1]

1. University of Georgia, Odum School of Ecology, 140 E. Green Street, Athens GA, 30602.

2. Corresponding author: tdallas@uga.edu

Abstract

Host-parasite interactions are complex interactions capable of being influenced by a multitude of factors. Predicting what hosts a given parasite can infect is a central goal in parasite ecology. Here, we develop predictive models capable of accurately determining which potential hosts a given parasite infects. We do this using a large database of freshwater fish-parasite interactions (FishPest). For each of 238 parasite species that were recorded at least 20 times in the database, we trained boosted regression tree models on host trait variables, geographic variables, and parasite community variables in order to determine both the predictive capability of our models on hold-out data, and the relative influence of our three different variable classes on predictive accuracy. We found that models trained on parasite community variables had high predictive accuracy ($\overline{AUC}$ = 0.89) relative to models trained on geographic variables ($\overline{AUC}$ = 0.79), or host traits ($\overline{AUC}$ = 0.66). Taken together, our findings suggest that the parasite communities of host species contain valuable information on the likelihood of infection by a novel parasite, which has implications for predicting how introduced parasites will integrate themselves into natural communities.

Keywords

FishPEST, species distribution model, boosted regression tree, parasite niche

Introduction

Parasites are ubiquitous in nature, and are incredibly diverse in their life histories, transmission modes, and degree of host specificity [1]. The question of what determines which hosts a parasite infects is a central question to disease ecology. On one hand, host-parasite relationships may be considered as complex interactions determined by environment [2], geography [3], co-evolutionary history [4], or trait matching between host and parasite [5]. On the other, host-parasite interactions may be considered as random [6], or neutral interactions, such that predicting which hosts a parasite will infect is either impossible, or determined simply based on host abundance [7]. The degree to which host-parasite interactions are environmentally constrained, and therefore predictable, is unclear. Previous efforts to characterize parasite communities have largely focused on parasite richness [8, 9, 10, 11] instead of parasite community composition. Further, studies examining parasite community composition have focused efforts on topological measures of host-parasite networks [12, 7, 4, 13] or examined distance decay relationships in parasite community dissimilarity [14, 2, 15]. The ability to 1) discern if host-parasite interactions are simply neutral processes, and 2) predict the identity of the subset of hosts able to be infected by a particular parasite is a large knowledge gap in the study of parasite ecology. From an applied perspective, parasite host range prediction could be useful for forecasting potential parasite spillover events to novel hosts [16], including humans [17]. More generally, understanding the factors that determine which hosts a parasite could infect would allow for the generation of testable hypotheses concerning parasite generalism/specialism, and the role of host functional diversity and community composition on the distribution of parasites.

One of the largest factors holding predictive models of parasite distributions among

potential hosts is the relative paucity of data (but see [18]). However, this barrier is being overcome both by scientists [18, 19] and museums ([20]). Studies utilizing these large datasets have largely asked questions about parasite co-occurrence patterns [21], or the factors influencing parasite sharing [22, 23]. These studies largely examine parasite community composition, and determine the influence of host traits or phylogenetic relationships on parasite community composition. However, nearly no studies have attempted to predict which hosts a given parasite species will infect (see [24]), despite the importance of this question to public health, and host-parasite network structure. Specifically, efforts predicting parasite spillover to humans may be essential for mitigation of zoonotic diseases. Further, the ability to predict which hosts an introduced parasite will infect in the native community can guide management decisions, and effectively predict change in host-parasite network structure.

Previous efforts to predict parasite species host ranges have been hampered by the use of deprecated niche modeling algorithms, and conceptual differences in study goals [18]. Specifically, [18] developed a framework, specifically `PaNic` [24], that estimates parasite niche boundaries as defined by the host traits of known hosts, and outputs a list of potential hosts given user-imposed constraints (e.g., host family, geographic location). This could be useful to identifying unsampled hosts that may contain a given parasite. However, our goal is conceptually different, in that we aimed to develop predictive, cross-validated models that would allow for the determination of the relative importance of host traits, geographic variables, and other parasites that infect a given host species. The core difference lies in the assumption of a well sampled host community. [24] asks the question "given known parasite occurrences in a set of species, what other species might we expect to be parasitized", while we ask "what variables determine host community composition for a given parasite?". The predictive models we develop may be used to predict probability of parasite occurrence given new data, which is not far removed from the goals of [24], and determine the relative importance of variables in determining parasite occurrence probabilities among a set of potential host species. This work addresses a central

gap in our understanding of host specificity in parasites; what determines which hosts get infected by a given parasite species, and, more generally, is the host community of a given parasite predictable?

We address this knowledge gap by using a large database on freshwater fish prasites [21], in order to develop predictive parasite species distribution models for a number of parasite species ($n = 238$). In doing so, we can address the predictive capability of models trained on different variable classes. Host traits may influence the likelihood of parasite occurrence, but are likely not the only determinants of a parasite's host range. Specifically, apart from measures of host quality (host traits) the number and identity of host species that a given parasite could infect may be constrained by geographic location or the existing parasite community of the given host species (i.e. parasite community structure). To address the relative importance of these variable classes, we trained boosted regression models on each of these three variable classes, and compared the accuracy obtained from each model in predicting the potential distributions of 238 parasite species. Parasite species distributions were most constrained by the existing parasite community, as this model allowed for highly accurate prediction of parasite occurrence likelihood. Models trained on host traits had poor predictive capabilities, and models trained on geographic variables had intermediate predictive accuracy. Taken together, our findings suggest that predicting the host distribution of a given parasite species requires information on the parasite communities of the potential hosts, and not necessarily any information on host traits.

Methods

Data and processing We use an existing global database of fish-parasite associations (hereafter referred to as FishPest; [21]) consisting of over 38,000 helminth parasite records spanning a large diversity of parasites (Acanthocephala, Cestoda, Monogenea, Nematoda, Trematoda). We defined an occurrence of host-parasite relationship as a geographically

unique record reporting a parasite species infecting a particular host species. In order to allow for cross-validation and accurate prediction, we constrained our ananlyses to parasites with a minimum of 20 host records, which limited our analysse to 238 parasite species. Our response variable was parasite occurrence (binary), and was predicted using three classes of variables, representing host life history traits, geographic location, and parasite community similarity (Table 1).

Values of predictor variables were obtained largely through the FishPest database [18, 21], supplemented by variables obtained from FishBase [25]. However, there were still some missing predictor variable values. Missing predictor variable values were imputed based on proximity to a random forest using the `rfImpute` function in the `randomForest` `R` package ([26]). Details of host trait and geographic variable determination are provided in [21] and Table 1. Parasite community variables were formed by performing a principal components analysis on the binary host-parasite matrix, which serves to compress the data to a series of vectors where each host has a single value which represents parasite community similarity in one dimension. In order to remove information about focal parasite occurrence, the parasite being modeled was removed from the host-parasite matrix before ordination. Thus, the principal components represent a measure of parasite community structure among host species without any information about host range of the parasite species under consideration. We used the first five principal components as a measure of parasite community structure, which explained 28% of the cumulative variance on average. In addition, we included parasite species richness of a host as a predictor variable, as this may reflect the susceptibility of host species to parasitism.

The absence of a recorded interaction between host and parasite does not mean that the parasite does not infect that host. Borrowing from the idea behind Maximum Entropy modeling, we sampled the data to obtain background interactions, which we define here as a set of possible interactions between host and parasites. This background set was not composed of the entire dataset, but rather a sample of five times the number of positive occurrence records for a given parasite. These data were subset into a training

set (70% of the data), and a test set (30% of the data).

Predictive model formulation Here, we used boosted regression trees to predict parasite occurrence among potential host species for each of our 238 parasite species. Regression tree analysis is an extremely powerful tool for prediction and feature selection, bypassing many of the issues of simple regression models (e.g. multicollinearity, nonlinear relationships) [27, 28]. Boosting refers to the process of creating a large number of regression trees, and weighting them by their predictive power to extract general weak rules, which are then combined to enhance predictive ability. The optimal number of trees was determined using the out-of-bag (`OOB`) estimation procedure, with the upper limit set to 50000. Other pertinent parameters include the learning rate ($l = 0.001$), which controls the degree each new tree contributes to the overall model, and interaction depth ($id = 4$), which allows for up to four-way interactions among predictor variables.

From the final boosted regression tree models, we are able to extract variable relative contribution (RC) measures, which provide information about the importance of each variable to the final model predictions [29]. Relative contribution values for each predictor variable was determined by permuting each predictor variable and quantifying the reduction in model performance, a method that is free of classical assumptions about normality and equal variance [30]. Relative contribution estimates were then based on the number of times a given predictor variable was selected for splitting, weighted by the degree the split improves model performance, and scaled between 0 (no contribution) to 100 (maximum contribution).

Model performance was assessed using receiver operating characteristic curves, which relate true positive and false positive (type I error) rates graphically. The area between the curve generated by true and false positives and the 1:1 line from the origin gives a measure of predictive accuracy. It is possible that predictive models could overfit, predicting nearly no parasite occurrences, since parasite presences are a fraction of the

number of background data points. To account for this, all models were compared to a random null model, which randomized occurrence values in the test dataset, but kept them constrained to the total number of occurrences.

Results

Importance of host traits, geographic variables, and parasite community similarity All models performed better than our null predictions (null model $\overline{AUC} = 0.50$). With varying degrees of accuracy, models were able to predict parasite occurrence for the 238 parasite species examined using host traits ($\overline{AUC} = 0.66$), geographic variables ($\overline{AUC} = 0.79$), and parasite community similarity ($\overline{AUC} = 0.88$). The full model containing all variables was able to successfully predict parasite occurrences in the hold-out test dataset with high accuracy ($\overline{AUC} = 0.89$), only marginally more accurate than the model trained with only parasite community similarity variables (Figure 2). The relative contribution (RC) values for each separate model, and the model trained on all available data are provided in Figure 1.

In the full model, variables of different classes were allowed to have different relative contributions, which allows for the determination of variables driving the predictive accuracy of the full model. For instance, relative contribution values were largest for the parasite community similarity values obtained from the principal components analysis on the host-parasite network with the parasite species of interest removed (Figure 1), with the five PCA vectors comprising around 52% of the relative contribution values, and four of the top five predictive variables. On the other side of the predictive spectrum, host trait variables, specifically host age at maturity, lifespan, and growth rate, contributed very little to model performance.

Was predictive ability influence by parasite ecology? The relative importance of variable classes, or the general predictive power of the trained model, may differ as a

function of parasite taxonomic group or host specificity. We tested for variation in predictive power among parasite taxonomic group (Acanthocephalans, Cestodes, Monogeneans, Nematodes, and Trematodes) and as a function of the number of host specificity. Here, we defined parasite host specificity as the number of hosts a parasite infects. We failed to detect evidence that parasite taxonomic group (Figure S1) influenced predictive power in any of our trained models. We did, however, observe an effect of host specificity (Figure S2), as predictive accuracy became more variable as host specificity increased (i.e. the number of hosts a given parasite infected became smaller). Despite this variability, the the mean predictive accuracy over a range of host specificity values remained constant (Figure S2).

Discussion

We provide evidence that host traits are not as important to determining which hosts a given parasite will infect relative to geographic location, or parasite community similarity. This suggests two things. First, the current paradigm that attempts to define a parasite's niche based on qualities of the host [31, 1], such as host phylogenetic distance [32] or host life history traits [33], may need to be reconsidered. Second, it suggests that the parasite community infecting a given host species contains information that can either predict or preclude the occurrence of a novel parasite species on that host. Further, model predictive accuracy did not vary strongly as a function of parasite type or specificity, suggesting that our findings may be broadly applicable to parasites of different transmission modes, life histories, and degrees of specificity. Taken together, our analyses suggest that parasite communities of freshwater fish are not simply random assemblages, but are predictable with high accuracy given only information on coinfecting parasites. Further, our findings have implications to host community invasibility by novel parasites, parasite spillover, and host-parasite network structure.

We found that host traits generally poorly predicted parasite occurrences, a striking

finding given that a parasite's niche is often defined using information on host life history and phylogeny [18, 34]. It is possible that host traits are important in structuring a parasite niche, especially host traits related to the ability of a parasite to infect a given host species (e.g. immune defenses, diet breadth, or geographic range) [35]. This would be the case if the host traits measured here were not the traits that most constrain parasite occurrences, and if geographic and parasite community variables were highly correlated with unmeasured host traits. However, host traits examined here should have captured at least some information related to likelihood of parasite infection. Specifically, host population growth rate and host age at maturity are likely related to immune defenses [36], as host trophic level is related to exposure [37]. Geographic variables predicted parasite occurrences more accurately, but the importance of individual geographic variables varied greatly among parasite species (Figure 1), suggesting either that parasite species are responding to different geographic variables, or that a methodological aspect of the analysis causes no single variable to dominate predictions. This could occur if geographic variables were highly correlated, or if interactions among variables were very important (i.e. the interaction between latitude or longitude with geographic region).

Parasites may be introduced with the addition of non-native host species to communities. The successful integration of the non-native host may be either enhanced if parasites of the non-native host "spillover" to the resident host community, or reduced by the sharing of parasites from the resident host community to the potential invader (the so-called biotic resistance hypothesis; [38]). The ability to predict the likelihood of both avenues of parasite sharing could allow for the prediction of invasion probability based on the parasite community of host species present. Based on our work, predicting what resident host species are likely to become infected by a novel parasite requires only information on the parasite communities of the resident host species. This means that pre-invasion prediction of parasite spillover may be possible.

Acknowledgements

The Macroecology of Infectious Disease Research Coordination Network (funded by NSF DEB 131223) provided useful discussions and support for this work. TD, AP, and JMD were supported by the Odum School of Ecology at the University of Georgia.

References

[1] Poulin, R., 2011 *Evolutionary ecology of parasites*. Princeton university press.

[2] Locke, S. A., McLaughlin, J. D. & Marcogliese, D. J., 2013 Predicting the similarity of parasite communities in freshwater fishes using the phylogeny, ecology and proximity of hosts. *Oikos* **122**, 73–83.

[3] Nieberding, C. M., Durette-Desset, M.-C., Vanderpoorten, A., Casanova, J. C., Ribas, A., Deffontaine, V., Feliu, C., Morand, S., Libois, R. & Michaux, J. R., 2008 Geography and host biogeography matter for understanding the phylogeography of a parasite. *Molecular phylogenetics and evolution* **47**, 538–554.

[4] Krasnov, B. R., Fortuna, M. A., Mouillot, D., Khokhlova, I. S., Shenbrot, G. I. & Poulin, R., 2012 Phylogenetic signal in module composition and species connectivity in compartmentalized host-parasite networks. *The American Naturalist* **179**, 501–511.

[5] Rohr, R. P., Naisbit, R. E., Mazza, C. & Bersier, L.-F., 2013 Matching-centrality decomposition and the forecasting of new links in networks. *arXiv preprint arXiv:1310.4633* .

[6] Kennedy, C., 2009 The ecology of parasites of freshwater fishes: the search for patterns. *Parasitology* **136**, 1653–1662.

[7] Canard, E., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D. & Gravel, D., 2014 Empirical evaluation of neutral interactions in host-parasite networks. *The American Naturalist* **183**, 468–479.

[8] Arneberg, P., 2002 Host population density and body mass as determinants of species richness in parasite communities: comparative analyses of directly transmitted nematodes of mammals. *Ecography* **25**, 88–94.

[9] Nunn, C. L., Altizer, S., Jones, K. E. & Sechrest, W., 2003 Comparative tests of parasite species richness in primates. *The American Naturalist* **162**, 597–614.

[10] Ezenwa, V. O., Price, S. A., Altizer, S., Vitone, N. D. & Cook, K. C., 2006 Host traits and parasite species richness in even and odd-toed hoofed mammals, artiodactyla and perissodactyla. *Oikos* **115**, 526–536.

[11] Poulin, R. & Rohde, K., 1997 Comparing the richness of metazoan ectoparasite communities of marine fishes: controlling for host phylogeny. *Oecologia* **110**, 278–283.

[12] Guégan, J.-F. & Hugueny, B., 1994 A nested parasite species subset pattern in tropical fish: host as major determinant of parasite infracommunity structure. *Oecologia* **100**, 184–189.

[13] Poulin, R., 2010 Network analysis shining light on parasite ecology and diversity. *Trends in Parasitology* **26**, 492–498.

[14] Locke, S. A., Levy, M. S., Marcogliese, D. J., Ackerman, S. & McLaughlin, J. D., 2012 The decay of parasite community similarity in ring-billed gulls larus delawarensis and other hosts. *Ecography* **35**, 530–538.

[15] Poulin, R., 2003 The decay of similarity with geographical distance in parasite communities of vertebrate hosts. *Journal of Biogeography* **30**, 1609–1615.

[16] Colautti, R. I., Ricciardi, A., Grigorovich, I. A. & MacIsaac, H. J., 2004 Is invasion success explained by the enemy release hypothesis? *Ecology letters* **7**, 721–733.

[17] Daszak, P., Cunningham, A. A. & Hyatt, A. D., 2000 Emerging infectious diseases of wildlife–threats to biodiversity and human health. *Science* **287**, 443–449.

[18] Strona, G. & Lafferty, K. D., 2012 FishPEST: an innovative software suite for fish parasitologists. *Trends in parasitology* **28**, 123.

[19] Nunn, C. L. & Altizer, S. M., 2005 The global mammal parasite database: an online resource for infectious disease records in wild primates. *Evolutionary Anthropology: Issues, News, and Reviews* **14**, 1–2.

[20] Gibson, D., Bray, R. & Harris, E., 2005 Host-parasite database of the natural history museum, london. *URL http://www. nhm. ac. uk/research-curation/scientific-resources/taxonomy-systematics/host-parasites/database/index. jsp* .

[21] Strona, G., Palomares, M. L. D., Bailly, N., Galli, P. & Lafferty, K. D., 2013 Host range, host ecology, and distribution of more than 11,800 fish parasite species: Ecological archives e094-045. *Ecology* **94**, 544–544.

[22] Braga, M. P., Razzolini, E. & Boeger, W. A., 2014 Drivers of parasite sharing among neotropical freshwater fishes. *Journal of Animal Ecology* .

[23] Dallas, T. & Presley, S. J., 2014 Relative importance of host environment, transmission potential and host phylogeny to the structure of parasite metacommunities. *Oikos* **123**, 866–874.

[24] Strona, G. & Lafferty, K. D., 2012 How to catch a parasite: parasite niche modeler (panic) meets fishbase. *Ecography* **35**, 481–486.

[25] Froese, R. & Pauly, D., 2010. Fishbase.

[26] Liaw, A. & Wiener, M., 2002 Classification and regression by randomforest. *R News* **2**, 18–22.

[27] Elith, J., Leathwick, J. R. & Hastie, T., 2008 A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813.

[28] Dallas, T. & Drake, J. M., 2014 Relative importance of environmental, geographic, and spatial variables on zooplankton metacommunities. *Ecosphere* **5**, art104.

[29] Breiman, L., 2001 Random forests. *Machine learning* **45**, 5–32.

[30] Anderson, M. J., 2001 Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian journal of fisheries and aquatic sciences* **58**, 626–639.

[31] Bush, A. O., 2001 Communities of parasites. In *Parasitism: the diversity and ecology of animal parasites*, pp. 405–433. Cambridge University Press.

[32] Adamson, M. & Caira, J., 1994 Evolutionary factors influencing the nature of parasite specificity. *Parasitology* **109**, S85–S95.

[33] Sasal, P., Trouvé, S., Müller-Graf, C. & Morand, S., 1999 Specificity and host predictability: a comparative analysis among monogenean parasites of fish. *Journal of Animal Ecology* **68**, 437–444.

[34] Rohde, K., 1993 Niche restriction in parasites: Proximate and ultimate causes. *Parasitology* **109**, S69–84.

[35] Johnson, P. T., Rohr, J. R., Hoverman, J. T., Kellermanns, E., Bowerman, J. & Lunde, K. B., 2012 Living fast and dying of infection: host life history drives interspecific variation in infection and disease risk. *Ecology Letters* **15**, 235–242.

[36] Zuk, M. & Stoehr, A. M., 2002 Immune defense and host life history. *The American Naturalist* **160**, S9–S22.

[37] Price, P. W., 1990 Host populations as resources defining parasite community organization. In *Parasite communities: patterns and processes*, pp. 21–40. Springer.

[38] Britton, J. R., 2013 Introduced parasites in food webs: new species, shifting structures? *Trends in Ecology & Evolution* **28**, 93–99.

Tables

Table 1: Description and units of variables used to predict parasite occurrences.

Variable	Units	Description	Range
Age at maturity	years	Age at sexual maturity	0.1 – 34
Growth rate	$years^{-1}$	Rate to approach asymptotic length	0.02 – 9.87
Life span	years	Estimated maximum age	0 – 145
Max length	cm	Maximum fish species length	1 – 2000
Trophic level	–	1 + mean trophic level of food	2 – 5
Area of occupancy	No. 1x1 ° cells	Global host distribution	1 – 1610
Geographic region	–	Biogegraphic region	–
Latitude	|max - min| degrees	Latitudinal distribution	1 – 148
Longitude	|max - min| degrees	Longitudinal distribution	1 – 359
Parasite species richness	#	No. parasite species of host species	0 – 89
Principal components	–	PCA axes of host-parasite network	-11.7 – 9.8

Table 2: Accuracy, measured as the Area under Receiver operating characteristic (ROC) curves, for our predictive boosted regression tree models trained on n variables relating to host traits, geographic variables, parasite community information, and the full model. These trained models were compared to a null model that maintained interaction number (number of occurrence records), but assigned occurrences equiprobably among potential interactors.

Variable class	n	A$\overline{U}$C	SE
Null model	-	0.50	$7e^{-5}$
Host traits	5	0.66	0.009
Geography	4	0.79	0.007
Parasite community	6	0.88	0.006
Full model	15	0.89	0.005

Figures

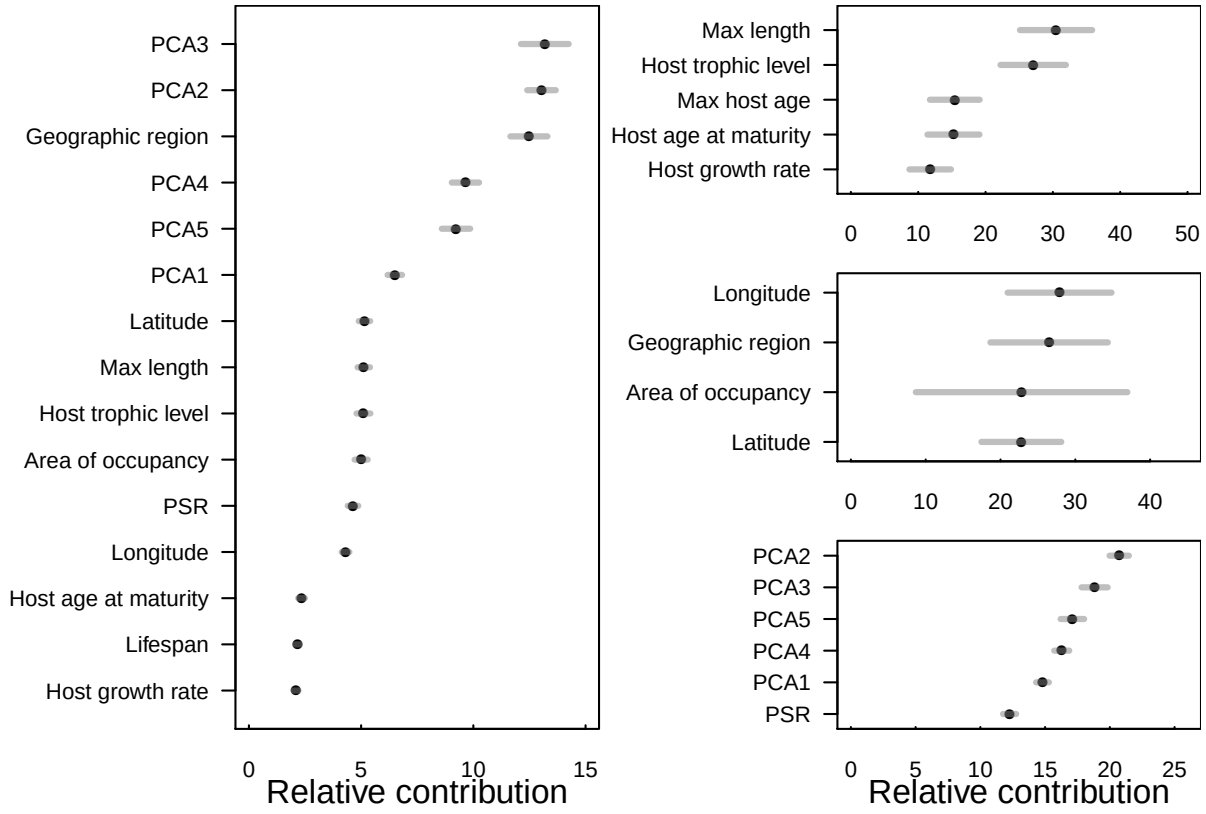

Figure 1: The average relative contribution values from the boosted regression tree models trained on all available data (left), host trait data (top right), geographic variables (middle right), and parasite community similarity (bottom right). Variables named "PCA" are principal components axes, and "PSR" refers to parasite species richness. Other variable definitions and units are available in Table 1.