Title: What controls the range of hosts a fish parasite infects?

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

Affiliations:

1. University of Georgia, Odum School of Ecology, 140 E. Green Street, Athens GA, 30602.
2. Corresponding author: `tdallas@uga.edu`

Abstract

Keywords

Take-home messages

1. It is possible to predict parasite niche breadth using either parasite community similarity. This means that freshwater fish parasites are not random assemblages, but that parasites with similar niches infect similar host species. It is possible that parasite community information condenses host trait variation, (perhaps) evolutionary history, and geographic location (to some extent).
2. Predictive accuracy does not vary as a function of host specificity (though I only consider parasites with 20 or more occurrence records, but these could all be on the same host species).
3. It may be possible to predict parasite spillover from invasive hosts to native host communities, or to predict biotic resistance of a community to invasion.

Selling points

1. Similar parasites infect similar hosts. This means that the host represents a patch, patches vary in quality, and host-parasite interactions are not neutral. If there were neutral, geographic variables would have been king, right?

Introduction

Host-parasite relationships are complex, intimate (non-neutral) interactions with lots of impacts

Most of parasite population and community ecology is about looking for patterns, and lots of folks don't find them, and have called host-parasite interactions neutral, or random Parasite communities are sometimes conserved across host species, such that the presence of one parasite may increase the likelihood of finding another parasite species.

Being able to predict parasite occurrence is pretty important, for a number of reasons (species invasions/biotic resistance, spillover to human hosts, etc.

Previous work and knowledge gap One of the largest factors holding predictive models of parasite distributions among potential hosts is the relative paucity of data (but see). However, this barrier is being overcome both by scientists Strona and Lafferty [2012], Nunn and Altizer [2005] and museums (Gibson et al. [2005]). Studies utilizing these large datasets have largely asked questions about parasite co-occurrence patterns Strona et al. [2013], or the factors influencing parasite sharing Braga et al. [2014], Dallas and Presley [2014]. These studies largely examine parasite community composition, and determine the influence of host traits or phylogenetic relationships on parasite community composition. However, almost no studies have attempted to predict what host species a parasite will infect. Previously, Strona et al. [2013] developed a framework to predict parasite co-occurrence likelihood given host community, and habitat variables. This approach is predicated on the idea that information on hosts and geography are what determine the likelihood of a parasite infecting a given host.

Thesis (what I did, what I found) Here, we test this idea by quantifying the predictive capability of models trained on host traits, geographic variables, or parasite community variables. To do this, we examined a large dataset on interactions between

freshwater fish and their parasite communities Strona and Lafferty [2012]. The number and identity of host species that a given parasite could infect may be constrained by geographic location, host trait variables (i.e. patch quality), or the existing parasite community of the given host species (i.e. parasite community structure). We trained boosed regression models on each of these three variable types, predicting the potential distribution 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 on a set of hosts. This suggests that either parasite community composition contains information about host susceptibility to infection, and potentially information on the relevant ranges of host and parasite, serving as a measure of geographic location. Taken together, this suggests that parasite community composition can determine parasite occurrence in a host community. This has important applications to the study of invasive species, as it may be possible to predict what parasites will spillover to other hosts in the case of a non-native host introduction, or the biotic resistance of the native host community, as the native parasite community may be capable of infecting the non-native host.

Methods

Data and processing We use an existing global database of fish-parasite associations (hereafter referred to as FishPest; citepstrona2013) consisting of over 38000 helminth parasite records spanning a large diversity of parasites (Acanthocephala, Cestoda, Monogenea, Nematoda, Trematoda). Many of these entries represent isolated occurrences of parasites on hosts. In order to allow for cross-validation and accurate prediction, we constrained our ananlyses to parasites with a minimum of 20 host records. In other words, we only examined parasites that had been recorded more than 20 times, but these occurrences could be on fewer than 20 host species. The inclusion of duplicate occurrences was only permitted if the parasite was recorded on a host in a different geographic location, based on latitude and longitude values. This resulted in a total of 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 Strona and Lafferty [2012], Strona et al. [2013], supplemented by variables obtained from FishBase Froese and Pauly [2010]. However, there were still some missing predictor variable values. Missing predictor variable values were imputed using the imputation procedure in the `randomForest` *R* package (Liaw and Wiener [2002]). Details of host trait and geographic variable determination are provided in Strona et al. [2013]. Parasite community similarity was considered as a the first five principal components from a principal components analysis on the host-parasite matrix. This matrix contains information on all parasites infecting all host species, except with the parasite species of interest removed. Thus, the principal components represent a measure of parasite community similarity among host species without any information about host range of the parasite species under consideration. In addition, we included parasite species richness of a host as a predictor variable.

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) Elith et al. [2008], Dallas and Drake [2014]. 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.

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 used to train the boosted regression tree model, and a test set, the remaining 30% of the data used to test the predictive accuracy of the trained model.

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. 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 (Anderson 2001). 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).

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. 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.

Results

Importance of host traits, geographic variables, and parasite community similarity All models performed better than our null predictions (null model AUC = 0.50). With varying degrees of accuracy, models were able to predict parasite occur-

rence for the 238 parasite species examined using host traits ($A\bar{U}C = 0.66$), geographic variables ($A\bar{U}C = 0.79$), and parasite community similarity ($A\bar{U}C = 0.88$). The full model containing all variables was able to successfully predict parasite occurrences in the hold-out test dataset with high accuracy ($A\bar{U}C = 0.90$). The relative contribution (RC) values for each separate model, and the model trained on all available data are provided in Figure 2.

Details about full model (RC values, etc.) However, the importance of variable group (host traits, geographic variables, or parasite community variables) to predictive power differed among parasite species in the full model (Figure 3). Despite the apparent variation in the relative contributions of each class of variables, the error bars on the relative contribution values are small (Figure 2). Further, variables related to parasite community similarity had the highest relative contribution values in the full model, summing to more than half of the relative contribution values, and constituting four of the top five explanatory variables.

Was predictive ability influence by parasite specificity, parasite type, or...? It is possible that predictive power could be a function of parasite taxonomic affiliation, or host specificity. However, models trained on each of the variable classes, including the full model, did not differ in predictive power among parasite taxonomic groups (Figure S1). Further, predictive accuracy was unrelated to parasite specificity, where specificity is defined as the number of unique host species that a parasite species was observed to infect (Figure S2). This may be caused by our *a priori* condition for the minimum number of parasite occurrences ($n = 20$), set in an effort to provide the model enough data to train and cross validate.

Discussion

State most important findings

what does this mean? Tie into concepts of invasion and biotic resistance. Also emphasize that this suggests parasites are likely not random assemblages, but can be predicted with limited accuracy based on host traits, and with much greater accuracy using only data on parasite community composition.

Acknowledgements

References

Mariana P Braga, Emanuel Razzolini, and Walter A Boeger. Drivers of parasite sharing among neotropical freshwater fishes. *Journal of Animal Ecology*, 2014.

Tad Dallas and John M Drake. Relative importance of environmental, geographic, and spatial variables on zooplankton metacommunities. *Ecosphere*, 5(9):art104, 2014.

Tad Dallas and Steven J Presley. Relative importance of host environment, transmission potential and host phylogeny to the structure of parasite metacommunities. *Oikos*, 123(7):866–874, 2014.

Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

Rainer Froese and Daniel Pauly. Fishbase, 2010.

DI Gibson, RA Bray, and EA Harris. 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*, 2005.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

174 Charles L Nunn and Sonia M Altizer. The global mammal parasite database: an online
175 resource for infectious disease records in wild primates. *Evolutionary Anthropology:*
176 *Issues, News, and Reviews*, 14(1):1–2, 2005.

177 Giovanni Strona and Kevin D Lafferty. FishPEST: an innovative software suite for fish
178 parasitologists. *Trends in parasitology*, 28(4):123, 2012.

179 Giovanni Strona, Maria Lourdes D Palomares, Nicolas Bailly, Paolo Galli, and Kevin D
180 Lafferty. Host range, host ecology, and distribution of more than 11 800 fish parasite
181 species: Ecological archives e094-045. *Ecology*, 94(2):544–544, 2013.

Tables

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

Variable	Units	Description	Range
Max length	cm	Maximum fish species length	1 – 2000
Trophic level	–	1 + mean trophic level of food items	2 – 5
Age at maturity	years	Age at sexual maturity	0.1 – 34
Life span	years	Estimated maximum age	0 – 145
Growth rate	$years^{-1}$	Rate to approach asymptotic length	0.02 – 9.87
Marine	–	Is host found in marine habitat?	0, 1
Freshwater	–	Is host found in freshwater habitat?	0, 1
Brackish	–	Is host found in brackish habitat?	0, 1
Geographic region	–	Biogegraphic region	–
Area of occupancy	No. 1x1 ° cells	Global host distribution	1 – 1610
Latitude	max - min degrees	Latitudinal distribution	1 – 148 ????
Longitude	max - min degrees	Longitudinal distribution	1 – 359 ?????

Figures

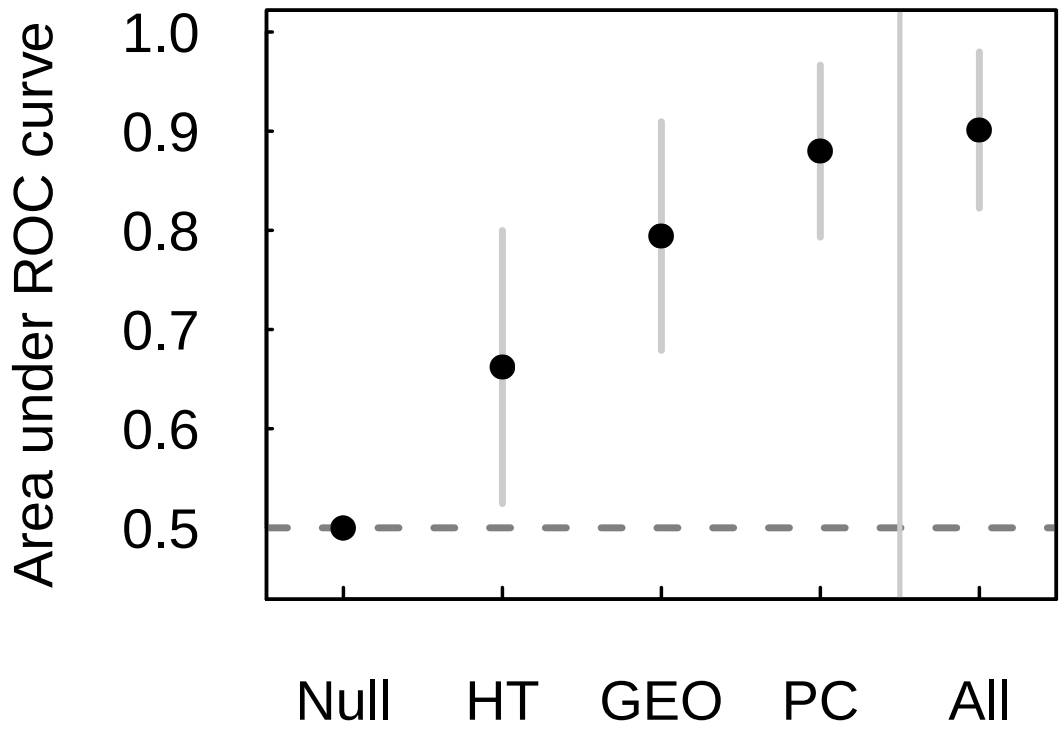

Figure 1: Accuracy (Area under Receiver operator characteristic (ROC) curves) for our predictive models incorporating host traits ('HT'), our null model ('Null') that maintained interaction number (number of occurrence records), but assigned occurrences equiprobably among potential interactors.

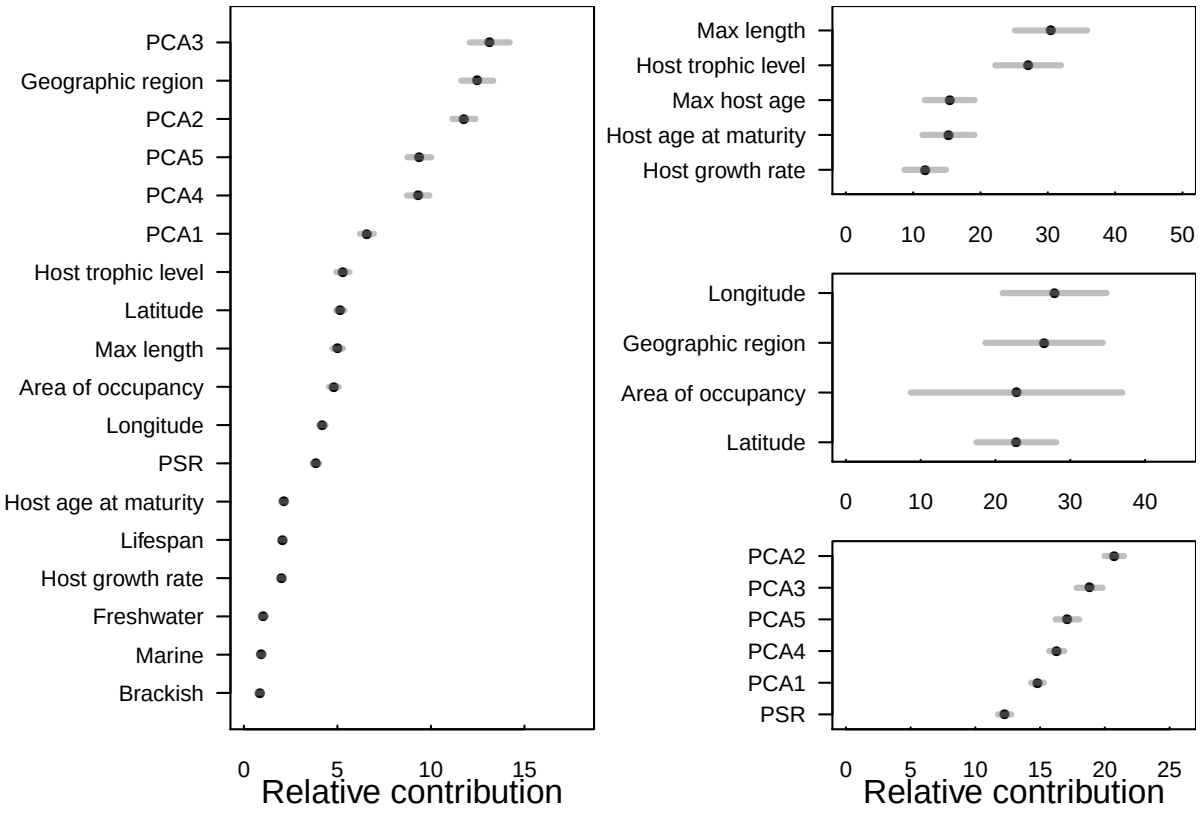

Figure 2: The average relative contribution values from the boosted regression tree models trained on all available data (left), host trait data (top right), geographic variables (middle right), and parasite community similarity (bottom right). Variables named "PCA" are principal components axes, and "PSR" refers to parasite species richness. Other variable definitions and units are available in Table 1.

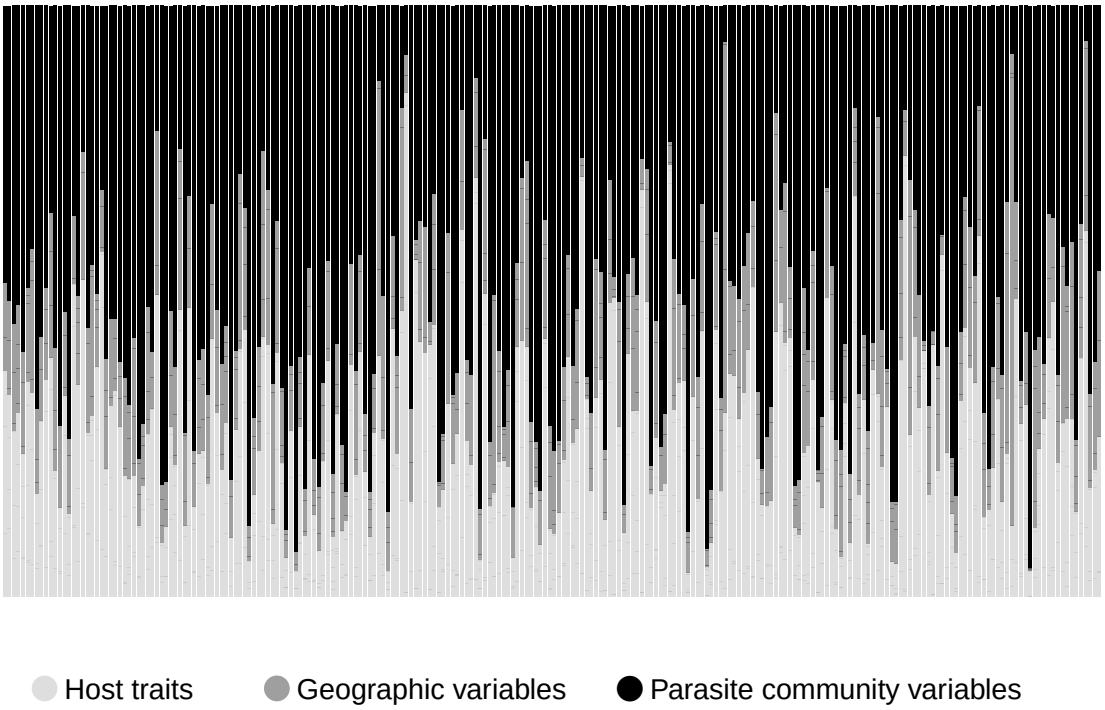

Figure 3: Relative contribution values for variables of one of three classes; host traits, geographic variables, or parasite community variables. Each column represents a model trained on occurrence data for a single parasite species.

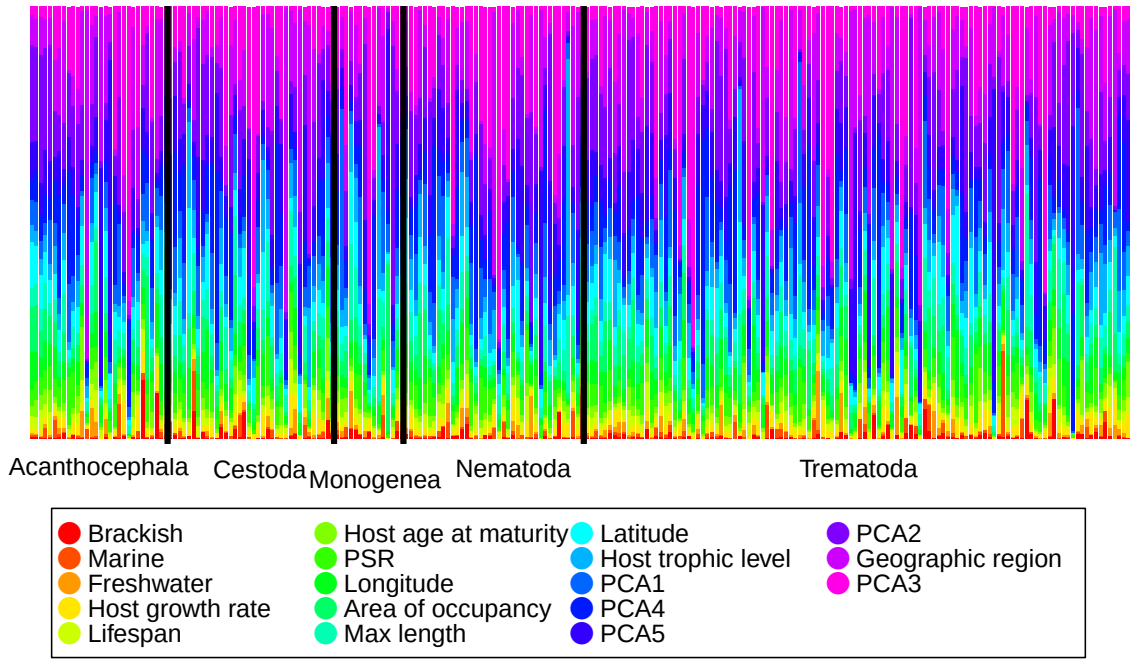

Figure 4: Variation in variable importance as a function of parasite type. Each column represents a model trained on occurrence data for a single parasite species. Variables are sorted by their mean relative importance values across parasite species, but these values vary among parasite species.

Supplementary Material

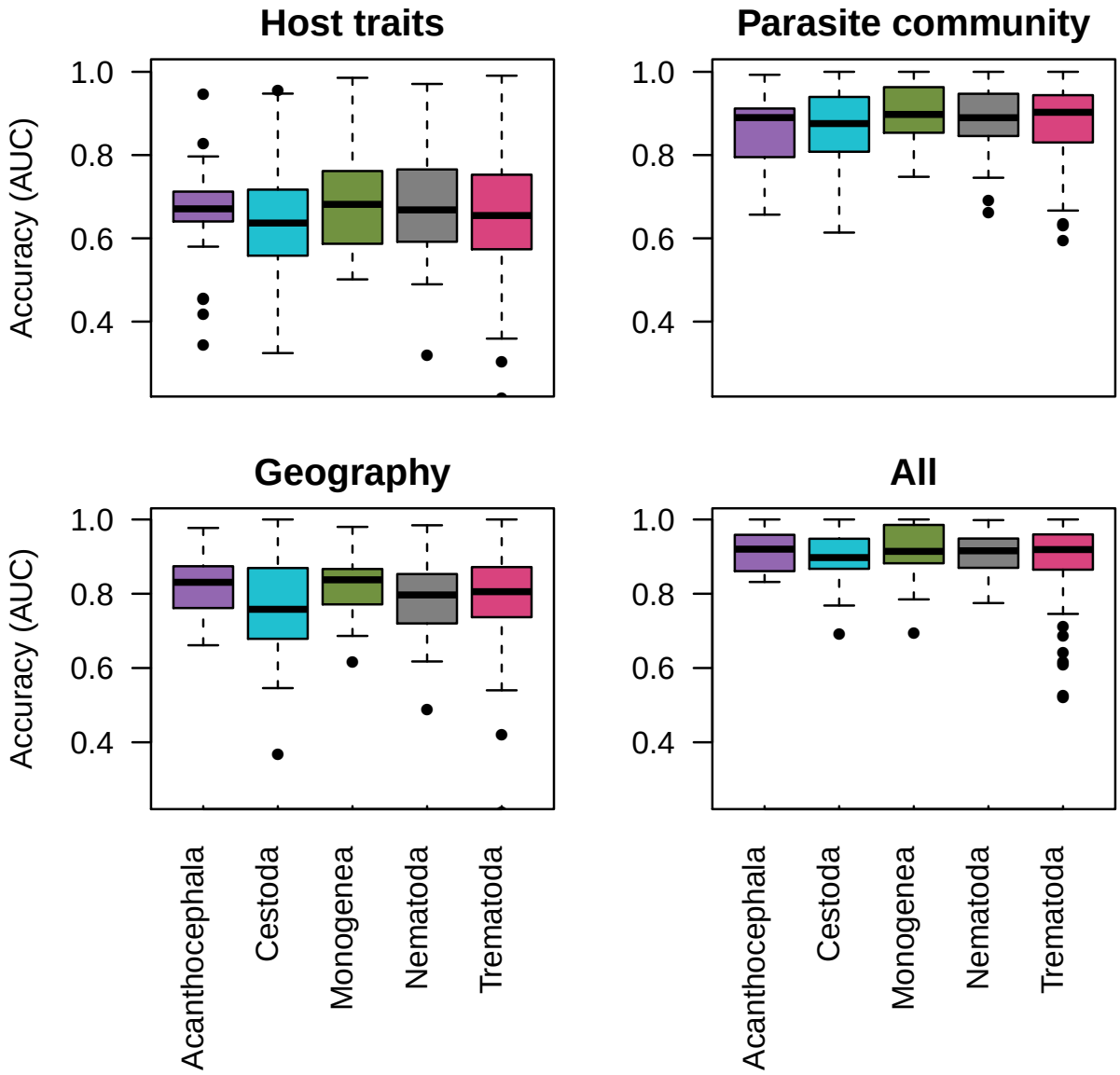

Figure 5: Accuracy (Area under Receiver operator characteristic (ROC) curves) for boosted regression models trained using host traits (top left), parasite community similarity (top right), geographic variables (bottom left), and all available data (bottom right) as a function of parasite type (x-axis).

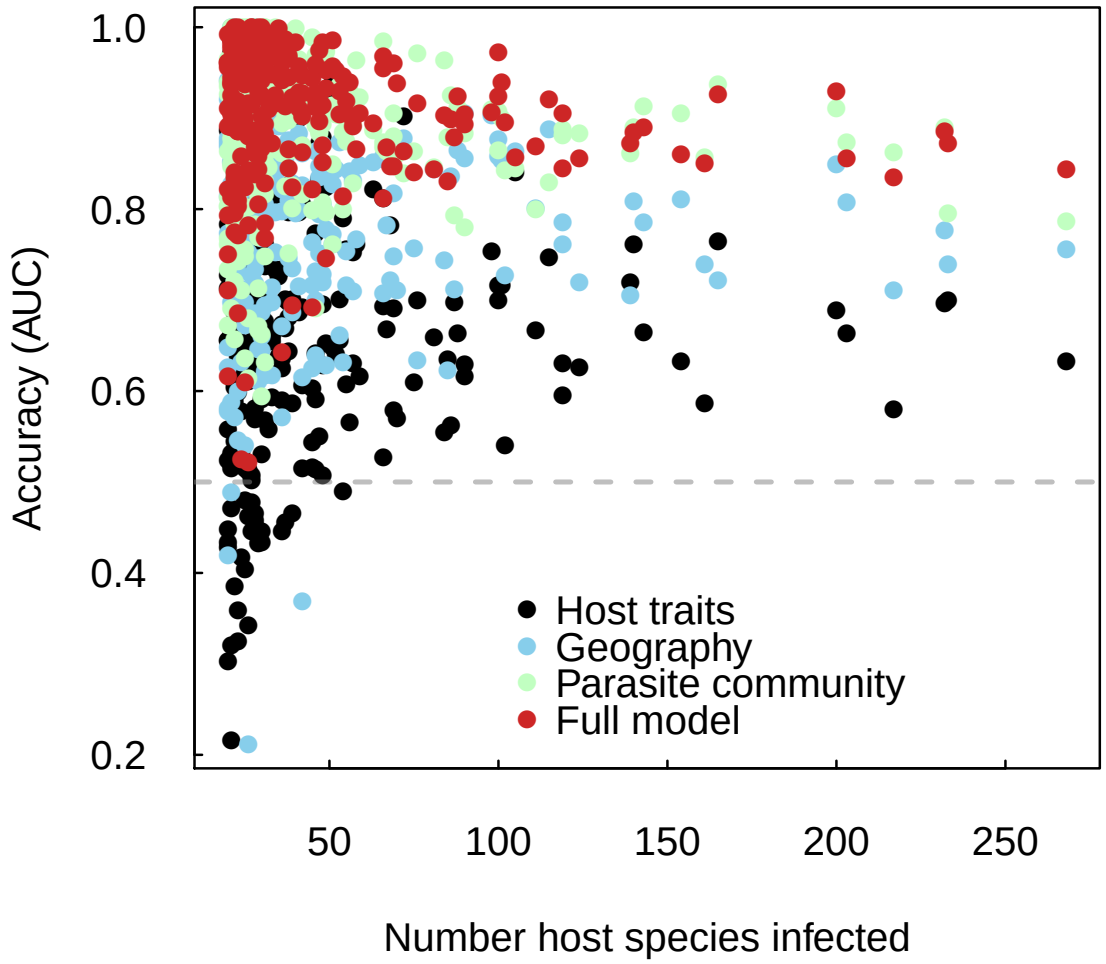

Figure 6: Accuracy (Area under Receiver operator characteristic (ROC) curves) for boosted regression models trained using host traits, parasite community similarity, geographic variables, and all available data as a function of the number of hosts the parasite species was collected on (x-axis). This host range estimate was uncorrelated to accuracy, though some models trained on host traits (black dots) performed poorly when parasite species were specific to a smaller number of host species.