Predictability of helminth parasite host range using information on geography, host traits and parasite community structure

Tad Dallas 1,2,\*, Andrew Park 1,3 and John M. Drake 1,3

1 Odum School of Ecology, University of Georgia, Athens GA, 30602

2 Department of Environmental Sciences and Policy, University of California, Davis, CA 95616

3 Center for the Ecology of Infectious Diseases, University of Georgia, Athens, GA, 30602

∗ Corresponding author; e-mail: tdallas@uga.edu

‡ ORCIDs: Dallas, 0000-0003-3328-9958; Park, 0000-0003-4080-7274; Drake, 0000-0003-4646-1235.

## SUPPLEMENTARY MATERIALS

### HOSTS FOUND IN MULTIPLE GEOGRAPHIC REALMS

All host covariates had a single value for each host species except for host biogeographic realm. This is because biogeographic region is a characteristic of the location of the host-parasite interaction, while other measures are host-level measures of geographic range or mean trait values. To accommodate for this, we encoded biogeographic region as a series of seven binary dummy variables. This resulted in a single unique row entry for each host species. The main text reports the results of permutation tests to obtain relative contribution (Rc) values, in which the binary geographic variables (e.g., Paleartic) are independently permuted, and then all seven Rc values from each biogeographic region are summed.

### DOES PARASITE TAXONOMY OR HOST SPECIFICITY INFLUENCE PREDICTIVE ACCURACY?

We investigated if model predictive accuracy was biased by either parasite specificity (number of hosts), or type (taxonomic group). We found no evidence that parasite taxonomic group influenced predictive accuracy for models trained on host traits, parasite community structure, geographic variables, or the full model (Figure 1). Many helminth parasites, especially trematodes, have complex life cycles. While the current manuscript dealt only with fish, and is therefore unlikely to be confounded by parasite stage-structured infection of multiple host species, this could be an issue in predicting the permissive host set for other parasite species. Parasite specificity, defined as the number of unique hosts a single parasite species infects, only slightly influenced model predictive accuracy (Figure 2). Predictive accuracy became more variable with decreasing parasite specificity, but the mean accuracy remained unchanged over the range of parasite specificity values.

### DOES THE NUMBER OF BACKGROUND POINTS INFLUENCE MODEL ACCURACY?

In the main text, we sample background points proportional to the number of positive occurrence points, selecting a value of 10x the number of occurrence points as a rule of thumb. Here, we explore the sensitivity of our models to this scaling factor by training models on a sample of 50 parasites with at least 40 occurrence records. We examined model accuracy, measured as AUC, for models trained with background data consisting of 5, 10, 20, or 40 times the number of occurrence points. When the background point size would be larger than the available data (which occurred rarely), we simply used all the data as background data. We trained models using all of the covariates available (the “full model”). We found no evidence that model accuracy, defined here as AUC, differed as a function of the size of the background data (Figure 3). This may relate to our quantification of accuracy as AUC, which focuses on discrimination, or the ability to rank order prospective hosts correctly.

### DETAILS ON THE FORMATION OF PARASITE COMMUNITY VARIABLES

Parasite community covariates were created by calculating principal component axes on the binary occurrence matrix of hosts (rows) and parasites (columns). Only the top five PCA axes were included in our analyses, as additional axes did not explain much variation. In fact, each axis after the first five explained approximately 2% of the remaining variation (Figure 4).

SUPPLEMENTAL FIGURES

Figure 1: Model accuracy (area under ROC curves) for boosted regression models trained using host traits (top left), parasite community variables (bottom left), geographic variables (top right), host taxonomic variables (bottom right), and all available data (center panel) as a function of parasite type (*x*-axis). Parasite taxa did not influence model accuracy.



Figure 2: Model accuracy (area under ROC curves) for boosted regression models trained using host traits, host taxonomy, parasite community composition, geographic variables, and all available data as a function of the number of hosts infected by a given parasite (*x*-axis). The number of host species infected did not influence model accuracy, though variation did exist as parasite species became specialized to a smaller number of host species. Lines correspond to fitted splines (spar = 0.5, to restrict possibility of oversmoothing).



Figure 3: Model accuracy (area under ROC curves) did not change as a result of the size of the background data used to train the model. The blue point indicates the value used in the main text for background point sampling. Dotted lines are ± 1 SD away from the mean.



Figure 4: Scree plot of the first ten principal components calculated on the binary host-parasite network. Numbers on top of plotted bars correspond to the cumulative variance explained. PCA axes indicated by blue bars were included as covariates in the main text, as the remaining axes (red bars) did not explain much variation.

