When can we infer mechanism from parasite aggregation? A constraint-based approach to disease ecology

Mark Q. Wilber[1,*]

Pieter T. J. Johnson[2]

Cheryl J. Briggs[1]

June 10, 2016

1. University of California, Santa Barbara, Santa Barbara, CA, 93106;

2. University of Colorado, Boulder, CO 80309;

$*$ Corresponding author; e-mail: mark.wilber@lifesci.ucsb.edu

Keywords: macroparasites, feasible sets, maximum entropy, trematodes, amphibians, negative binomial, geometric, host-heterogeneity, parasite-induced host mortality

Abstract

A few hosts have many parasites while many hosts have a few parasites - this axiom of macroparasite aggregation is so pervasive it is considered a general law in disease ecology, with important implications for the dynamics of host-parasite systems. Because of these dynamical implications, a significant amount of work has explored both the various mechanisms leading to parasite aggregation patterns and how to infer mechanism from these patterns. However, as many disease mechanisms can produce similar aggregation patterns, it is not clear whether aggregation itself provides any additional information about mechanism. Here we apply a "constraint-based" approach developed in macroecology that allows us to explore whether parasite aggregation contains any additional information beyond what is provided by mean parasite load. We tested two constraint-based null models, both of which were constrained on the total number of parasites P and hosts H found in a sample, on 842 observed amphibian host-trematode distributions. We found that 85-88% of the observed distributions did not significantly deviate from the constraint-based models, suggesting that many host-parasite distributions contain little biological information beyond what is already contained in the number of sampled parasites P and hosts H. However, for host-parasite distributions that were not sufficiently explained by P and H, we show that extending the constraint-based null models can identify the potential role of known aggregating mechanisms such as host-heterogeneity and disaggregating mechanisms such as parasite-induced host mortality in constraining host-parasite distributions. A constraint-based approach to host-parasite ecology can provide robust, low parameter null models for parasite aggregation to help determine when host-parasite distributions contain additional biological information beyond that contained in mean parasite load.

Introduction

Disease ecology has traditionally emphasized mechanistic descriptions of infection patterns (Anderson and May 1978, Duerr et al. 2003, Poulin 2007). One particular pattern observed in macroparasites, such as parasitic helminths and arthropods that do not directly reproduce within their host (Anderson and May 1979), is that many hosts in a population tend to have few parasites and a few hosts tend to have many. In statistical parlance this means that parasites tend to be aggregated within their hosts. This pattern is so ubiquitous in parasites that is has been called one of the few general laws in disease ecology (Poulin 2007).

Canonical models of host-macroparasite dynamics have illustrated that a balance between parasite pathogenicity (α) and parasite aggregation plays an important role in the ability of a parasite to regulate a host population (Anderson and May 1978, Tompkins et al. 2002). If parasites are highly aggregated and highly pathogenic, they cannot regulate a host population because parasites will be concentrated in a few hosts who will eventually experience parasite-induced mortality, extirpating the parasites from the host population. Moreover, models also show that when parasites are nearly uniformly distributed among hosts and highly pathogenic they cannot stably regulate a host population (Anderson and May 1978). In general, the stability of a host-parasite system and the regulation of a host population by parasites requires some level of parasite aggregation and that parasite pathogenicity is not too high. Because of the importance of parasite aggregation, much empirical and theoretical work has sought to understand both the mechanisms that can lead to aggregation in host-macroparasite systems (Anderson and Gordon 1982, Wilson et al. 2002, Raffel et al. 2011) and how to infer the dominant mechanisms structuring a host-parasite system from observed aggregation patterns (Duerr et al. 2003, Grear and Hudson 2011, Wilber et al. 2016).

Traditionally, studies of macroparasite aggregation have relied on a process-based approach where various aggregating and disaggregating mechanisms are sequentially incorporated into unaggregated null models until observed levels of aggregation are obtained (Anderson and Gor-

don 1982, Isham 1995, Chan and Isham 1998, Pugliese et al. 1998, Rosà and Pugliese 2002, Rosà et al. 2003, Grear and Hudson 2011, Fowler and Hollingsworth 2016). While the process-based approach has usefully illuminated various aggregating and disaggregating mechanisms in host-parasite systems (summarized in Wilson et al. 2002), it suffers from the "many-to-one problem" inherent in much of ecology (Frank 2014): there are many process-based models that can result in similar levels of parasite aggregation making it difficult to identify the specific processes leading to aggregation from patterns alone. When lab or field experiments are not a viable option to identify mechanism, it would be useful to have some criteria to identify when observed patterns of parasite aggregation may provide some information about the mechanisms influencing a host-parasite system or when all of the information is provided in the mean parasite load.

Recently developed constraint-based models used in macroecology provide such a criteria. These models are different from the process-based approach in that they attempt to predict the most-likely form of a population- or community-level distribution using only a known set of statistical constraints (Harte 2011, Locey and White 2013, Newman et al. 2014, Xiao et al. 2015b). The constraint-based approach does not propose that biological mechanisms are not acting in a system; it contends that many different combinations of these mechanisms lead to similar patterns of aggregation with predictable statistical properties (Frank 2009, McGill and Nekola 2010, Frank 2014). This is important because these models can then be used as robust null models (i.e. models that do not trivially fail) to identify when a given observed distribution contains biological information beyond that given by the constraints used to predict the distribution (Locey and White 2013, Harte and Newman 2014). These constraint-based approaches have had much success as robust null models in empirical free-living populations and communities (White et al. 2012, Locey and White 2013, Newman et al. 2014, Xiao et al. 2015a, Harte et al. 2015) and we argue that they can also be useful in addressing mechanistic questions about parasite aggregation in disease ecology.

For example, any observed host-parasite distribution is constrained by the total number of parasites P and the total number of hosts H in the sample. Given this, there are only a finite

number of shapes that this sampled host-parasite distribution can take (i.e. the feasible set of the host-parasite distribution, Locey and White 2013). If the shape of this observed host-parasite distribution is similar to the most likely distribution within this feasible set, then making inferences about the biological mechanisms leading to the shape of this distribution is difficult as the observed distribution is simply the most-likely distribution of all possible distributions (Haegeman and Loreau 2009). However, if the observed host-parasite distribution is not similar to the most-likely distribution, then one has reason to infer that biological mechanism may be further constraining the possible shape of the host-parasite distribution.

This has important implications for disease ecology where one often wants to understand something about the mechanisms affecting a host-parasite system from the level of aggregation observed (Anderson and Gordon 1982). Having some robust criteria for when a sampled host-parasite distribution shows "unusual" aggregation can help identify host-parasite systems where particular aggregating or disaggregating mechanisms are disproportionately constraining the distribution beyond the inherent (but biologically important) constraints imposed by P and H. This is potentially a much more powerful criteria for determining what is "unusual" aggregation in a host-parasite sample than the often-used Poisson distribution, which is widely acknowledged as a straw-man null hypothesis and trivially rejected in most all studies (Shaw and Dobson 1995, Shaw et al. 1998, Wilson et al. 2002).

This study has two goals. First, we use a dataset consisting of 22 unique amphibian host-trematode parasite pairings with over 8000 amphibians sampled at 205 sites over 5 years to test whether constraint-based models used in free-living systems also provide robust null models for host-parasite distributions. Second, we explore how, upon failing to describe host-parasite distributions, these constraint-based models can be extended to account for known aggregating and disaggregating mechanisms in host-parasite systems. We find that the shape of host-macroparasite distributions are generally well-predicted by the constraint-based approach, suggesting that there often may be little to infer about biological mechanism from host-parasite distributions beyond what is provided by P and H.

Methods

Constraint-based null models for parasite aggregation

We examine two different constraint-based null models that have been recently used in the macroecological literature. The first model is based on the concept of feasible sets (Haegeman and Loreau 2009, Locey and White 2013, Locey and McGlinn 2013) and begins by identifying two primary state variables of the system: here the total number of hosts H and the total number of parasites P. The feasible set is defined as all possible configurations in which P unlabeled parasites can be divided amongst H unlabeled hosts (see Appendix A; Locey and McGlinn 2013). A configuration $\mathbf{x}$ is one realized distribution resulting from distributing P parasites among H hosts. The central tendency of this feasible set is the most likely configuration given the constraints on the system (Appendix Figure 1). A common assumption is that all configurations are equally likely (Locey and McGlinn 2013, Locey and White 2013), but by weighting configurations differently one can obtain very different feasible sets (Xiao et al. 2015a).

The second constraint-based model is conceptually related to the feasible set approach and is derived using the principle of maximum entropy (Haegeman and Loreau 2009). The maximum entropy approach finds the distribution $p(\mathbf{x})$, where $\mathbf{x}$ is a configuration of P parasites amongst H hosts, that maximizes the entropy equation $H = -\sum_{\mathbf{x}} p(\mathbf{x}) \ln \left(p(\mathbf{x}) / p_0(\mathbf{x}) \right)$ given a set of constraints on $p(\mathbf{x})$, where $p(\mathbf{x})$ is the probability of observing a host-parasite configuration $\mathbf{x}$ and $p_0(\mathbf{x})$ is a prior weight on the configuration $\mathbf{x}$ (Frank 2009, Haegeman and Etienne 2010). Finding the maximum entropy distribution is equivalent to finding the most uniform distribution given a prior distribution and a set of constraints. For example, one may assume that each possible configuration is equally likely (i.e. $p_0(\mathbf{x})$ is uniform) and $\mathbf{x}$ is constrained such the total number of parasites in any configuration is P. By assuming that individual parasites are unlabeled, one could then derive the constraint-based model $p(\mathbf{x})$ that contains the least amount of information given these constraints (Jaynes 1982, Haegeman and Etienne 2010). We discuss the relationship between the feasible set and maximum entropy constraint-based models in Appendix A.

Without any prior reason to choose otherwise, we assumed that all configurations of parasites across hosts were equally likely for both the maximum entropy approach and the feasible set approach. For each constraint-based approach, we constrained the resulting distribution to have P parasites and H hosts, leading to a mean parasite load of P/H. It is well-known that mean parasite abundance is a highly significant predictor of the variance of a host-parasite distribution (Poulin 2013, Johnson and Hoverman 2014, Lagrue et al. 2015), and the constraint-based approach generalizes this result by asking whether mean parasite load provides enough information to predict the entire host-parasite distribution.

We computed the predicted feasible set host-parasite distribution with these constraints by randomly sampling from the full feasible set using the algorithms provided by Locey and McGlinn (2013). We drew 1000 samples from the full feasible set and used the center of this feasible set as the predicted host-parasite distribution. The predicted maximum entropy distribution with constraints P and H is given analytically as (Haegeman and Etienne 2010)

$$f(x|P,H) = \binom{P-x+H-2}{P-x} / \binom{P+H-1}{P} \tag{1}$$

where $f(x|P,H)$ is probability of observing a host with x parasites under the maximum entropy model constrained on P and H. This is equivalent to a finite negative binomial distribution with $k = 1$ (Zillio and He 2010). In the context of more commonly used distributions in disease ecology, these constraint-based models have one less parameter than a negative binomial model, which is a very flexible distribution that often fits host-parasite distributions very well (Shaw et al. 1998). We stress that the goal of this study is not to ask whether these distributions do better or worse than a negative binomial in predicting a host-parasite distribution, but whether host-parasite distributions tend to contain information beyond what is given by P and H.

Empirical data and constraint-based null model analysis

To test whether empirical host-parasite distributions contained information beyond that given by the constraint-based models, we used an extensive dataset of all macroparasites found in 8099 amphibian hosts across 205 ponds (sites) in the East Bay region of California (Alameda, Contra Costa and Santa Clara counties) from 2009-2014. In this field study, we sampled recently metamorphosed amphibians, as these provide a reliable and standardized indicator of infections acquired during aquatic development. To measure parasite abundance, we collected at least 10 of each host species as they approached metamorphosis, and performed a systematic examination of all major tissues and organs for parasites (Hartson et al. 2011). The sampled amphibians consisted of *Pseudacris regilla* (Pacific Tree Frog, $n = 4431$), *Anaxyrus boreas* (Western Toad, $n = 1309$), *Lithobates catesbeianus* (American Bullfrog, $n = 410$), *Taricha torosa* (California Newt, $n = 1568$), and *Taricha granulosa* (Rough-skinned Newt, $n = 381$).

We focused the following analyses on the five most common macroparasites in the system in terms of both prevalence and abundance. These were the larval trematodes *Ribeiroia ondatrae* (RION), *Echinostoma* sp. (ECSP), *Alaria* sp. (ALAR), *Cephalogonimus* sp. (CEPH), and *Manodistomum* sp. (MANO). All of these trematodes have complex life cycles in which their first intermediate hosts are pulmonate snails, their second intermediate host can be amphibians, snails or fish, and their definitive hosts are water-associated vertebrates (reptiles, amphibians, birds, or mammals) (Johnson and McKenzie 2008).

We determined whether the distributions of parasites across hosts deviated from the predictions of the constraint-based models for each combination of host species and parasite species at each site during each year. We included a year-by-site-by-host-by-parasite distribution only if it had at least 10 parasites and 10 hosts. Given this criterion, we were able to fit the constraint-based models to 842 host-parasite distributions. As expected, 837 of these distributions were aggregated with a variance to mean ratio greater than one (Figure 1). For each of these distributions, we extracted the total number of individuals of a given amphibian species (H) and parasites of a given trematode species (P) and calculated the corresponding rank abundance distribution (RAD) for

the constraint-based models. The RAD gives the predicted parasite abundances from a given distribution for H hosts and assigns a rank of 1 to the host with highest abundance and a rank of H to the host with the lowest abundance (Harte 2011, White et al. 2012).

We measured whether an observed host-parasite distribution deviated significantly from the central tendency of a constraint-based model using two criteria. The first criterion was plotting the observed RAD (obs_i) versus the predicted RAD (pred_i) and calculating the R^2 value based on a fit to the 1:1 line using the equation (White et al. 2012, Xiao et al. 2015b)

$$R^2 = 1 - \frac{\sum_i (\ln(\text{obs}_i + 1) - \ln(\text{pred}_i + 1)^2}{\sum_i (\ln(\text{obs}_i + 1) - \overline{\ln(\text{obs}_i + 1)})^2}$$

where i is the rank ($i = 1, \ldots, H$) of each observed or predicted host in a distribution. If the observed RAD exactly followed the constraint-based model then the R^2 value from the 1:1 line would be unity. If the model was a poor fit, the R^2 value will be less than unity and possibly negative if the 1:1 line was a worse fit than assuming that each host had a parasite abundance equal to mean of the observed distribution (White et al. 2012). We calculated R^2 values for each distribution independently as well as for all distributions combined.

The second criterion we used was an Anderson-Darling test to determine whether the observed RAD was significantly different than the constraint-based RAD for a given host-parasite distribution (Engmann and Cousineau 2011). The null hypothesis of this test is that the two distributions are the same. To account for the discrete nature of the data, we used a bootstrapped Anderson-Darling test with 10,000 bootstrapped samples (`kSamples` R package; Scholz and Zhu 2015). We considered two distributions to be significantly different if the p-value of the Anderson-Darling test was less than 0.1. We chose this conservative cutoff value to make it easier to reject the constraint-based models and thus harder for us to conclude that the constraint-based models were adequately describing the observed host-parasite distributions.

Extending the constraint-based null models

When an observed host-parasite distribution deviates from the central tendency of a constraint-based null model, this provides evidence that additional constraints/mechanisms beyond just P and H are disproportionately affecting the system (Harte and Newman 2014). We developed two ways to extend the constraint-based null models to detect whether classic aggregating and disaggregating mechanisms may be affecting host-parasite distributions beyond P and H.

Accounting for disaggregating mechanisms

Disaggregating mechanisms such as parasite-induced host mortality can play an important role in structuring empirically observed host-parasite distributions (Anderson and Gordon 1982). The parasite *Ribeiroia ondatrae* is known to have a strong, intensity-dependent effect on the survival of some amphibian hosts where increased parasite intensity leads to increased limb-malformations and decreased survival (Johnson 1999). This means that hosts with large parasite burdens are removed from the system, making the parasite distribution more uniform. Therefore, *Ribeiroia*-induced host mortality may interact with P and H to further constrain the shape of host-*Ribeiroia* distributions.

We included *Ribeiroia*-induced host mortality as an additional constraint in the feasible set and maximum entropy models using laboratory-derived survival curves that describe how *Ribeiroia* intensity affects amphibian host survival probability (Johnson et al. 2012). We focused on the amphibian species *Pseudacris regilla* because *Ribeiroia*-induced mortality and malformations in this species have been documented in the field and in the lab (Johnson 1999, Johnson and McKenzie 2008), an intensity-dependent *P. regilla*-*Ribeiroia* survival curve has been experimentally derived in the lab (Johnson 1999), and there were a large number of *P. regilla*-*Ribeiroia* distributions in the dataset on which to test the extended models ($n = 133$). We assumed that the intensity-dependent survival curve, which specifies the probability of an amphibian host surviving from larva to recent metamorph with some observed parasite load, followed a logistic function and

estimated the parameters of this function from the laboratory data given in Johnson (1999) (see Appendix B for more information).

We then used this laboratory-derived survival curve to further constrain the feasible set and maximum entropy predictions by assigning each possible configuration of P parasites amongst H hosts a likelihood based on the estimated survival function (Appendix B). For each constraint-based model, potential configurations were then weighted by this likelihood such that configurations with small likelihoods (e.g. ones that contained hosts with high parasite loads) were less likely to be observed than configurations with large likelihoods. Using this weighting scheme, we sampled from models that were constrained on P, H and *Ribeiroia*-induced host mortality using a Metropolis-Hastings algorithm (see Appendix B for a full description of the algorithm used). Once we obtained estimates of the mortality-constrained feasible set and maximum entropy models, we compared the resulting predictions to the observed *P. regilla-Ribeiroia* distributions using the methods described in *Empirical data and null model analysis*. As we did not perform any additional model fitting to derive the mortality-constrained model, it is not statistically inevitable that the central tendencies of the mortality-constrained models will provide a better representation of the data. Therefore, an improvement in agreement between model and data is strong evidence that *Ribeiroia*-induced mortality is constraining the distribution beyond P and H.

Accounting for aggregating mechanisms

Host heterogeneity, whether it be in susceptibility, parasite encounter rates, behavior or other factors, is an important mechanism leading to aggregation in host-parasite systems (Cornell 2010, Raffel et al. 2011). We accounted for this aggregating mechanism by extending the constraint-based models to include empirically observed levels of host heterogeneity. In particular, we explored discrete host heterogeneity where we assumed that overaggregation relative to the predicted model was a result of mixing discrete groups of hosts (Grafen and Woolhouse 1993, Wilson et al. 2002). This approach is conceptually and practically distinct from the standard practice of

fitting a negative binomial distribution to overaggregated host-parasite distributions. If the goal of an analysis is to obtain the best possible fit to an observed host-parasite distribution then it is well-known that fitting a negative binomial distribution provides an excellent model of overaggregated host-parasite distributions (Shaw et al. 1998, Calabrese et al. 2011). However, if the goal of an analysis is to determine whether a host-parasite distribution contains any information beyond what is contained in P and H, fitting a negative binomial model does not provide immediate insight into what constitutes unusual aggregation or the potential host attributes leading to this overaggregation (but see Alonso and Pascual 2006, Fowler and Hollingsworth 2016, for various mechanistic interpretations of the negative binomial k parameter). Extending a constraint-based model to include discrete host-heterogeneity, as is done here, can help generate more specific hypotheses as to the relative importance of different levels of host-heterogeneity in structuring a host-parasite distribution.

To incorporate discrete host-heterogeneity, we used 5 observed host attributes by which we could bin hosts into groups of heterogeneity. The first attribute was host body size (i.e. snout-vent length), which is a well-known attribute affecting parasite exposure and aggregation (Grutter and Poulin 1998, Poulin 2013). The other 4 host attributes were the parasite intensities of other larval trematodes. For example, if we were examining the host-parasite distribution of the parasite *Echinostoma* in the host *P. regilla*, the host attributes that we considered were the individual body-sizes of *P. regilla* and the intensities of *Ribeiroia*, *Alaria*, *Cephalogonimus*, and *Manodistomum* in *P. regilla* individuals. Co-infection can potentially increase aggregation by increasing heterogeneity in host susceptibility to the focal parasite (Cattadori et al. 2008), but can also decrease aggregation by increasing intra-host parasite negative density dependence (Pacala and Dobson 1988). Here we consider co-infection as a mechanism leading to increased aggregation.

Using these 5 host attributes, we used regression trees in which the response variable was the focal parasite abundance and the predictor variables were body size and the larval trematode intensities of a given host, excluding the focal parasite. Separate regression trees were run for each of the 842 host-parasite distributions. For a given host-parasite distribution, we found

the best regression tree with 2-5 of groups of host heterogeneity and calculated the relative importance of each predictor variable based on how much they reduced the sum of squared error compared to the other predictor variables (Fig. 2). We restricted each group to have at least 2 hosts and within each of these j groups, we determined the total number parasites P_j and the total number of hosts H_j. We chose to use a regression tree approach to analyze parasite abundance because it makes fewer assumptions than commonly used generalized linear models and it could be easily combined with the constraint-based feasible set model that does not have an analytically defined likelihood. Similar to a generalized linear model, the regression tree approach explores how various predictor variables affect mean parasite load (James et al. 2013), which is consistent with the constraint-based assumption that much of the information about the host-parasite distribution is contained in P and H.

To generate a constraint-based model RAD from the results of the regression tree, the RADs for each group j were computed with P_j and H_j and the predicted RAD was given by the concatenation for these j vectors (Fig. 2). This predicted mixture RAD could then be analyzed using the various methods described above. Finally, we also employed a randomization test to ensure that any improvement in model fit after including host heterogeneity was due to the host attributes considered, rather than just the act of grouping itself (described in Appendix C). In summary, while P and H alone may sometimes not sufficiently constrain an observed host-parasite distribution, this approach is testing whether allowing P and H to vary as a function of host covariates (i.e. heterogeneity) can account for deviations from the constraint-based null models. All analyses were performed in R and Python and the code to replicate the analysis can be found at `https://github.com/mqwilber/feasible_parasites`.

Results

Constraint-based null models

Overall, the majority of the host-parasite distributions that we considered did not significantly deviate from the constraint-based null models. 88% of the 842 year-by-site-by-host-by-parasite distributions were not significantly different from a feasible set model and 85% were not significantly different than a maximum entropy model. The median R^2 for all empirical distributions compared to the feasible set and maximum entropy models was 0.78 and 0.76, respectively.

Examining the models with regard to host-by-parasite combinations, typically over 70% of the empirical distributions for a given host-by-parasite combination did not significantly deviate from a constraint-based model, with some notable exceptions for the host *Lithobates catesbeianus* and the parasites *Alaria* sp. and *Cephalogonimus* sp. (Appendix Fig. 2-4). The median R^2 for the constraint-based models tended to be close to 80% for the various host-by-parasite combinations (Fig. 3).

How do host-parasite distributions deviate from constraint-based null models?

When a given host-parasite distribution significantly deviated from a constraint-based null models, they deviated being both overaggregated and underaggregated relative to the constraint-based model (89 and 120 distributions more aggregated and 13 and 7 distributions were less aggregated than predicted by feasible sets and maximum entropy, respectively). *Ribeiroia*, a parasite known to cause intensity-dependent host mortality (Johnson et al. 2012), accounted for 11 of the 13 significantly underaggregated distributions relative to the constraint-based model. Considering the empirical distributions that were overaggregated relative to the constraint-based models, *Alaria* had the highest proportion that were overaggregated (39%), followed by *Cephalogonimus* (26%), *Manodistomum* (25%), *Echinostoma* (13%) and *Ribeiroia* (4%).

Accounting for disaggregating and aggregating mechanisms

Parasite-induced host mortality

Including independently-estimated *Ribeiroia*-induced parasite mortality into the constraint-based models improved the overall fit of the models to *Pseudacris regilla-Ribeiroia* distributions (bootstrapped 95% confidence interval for the difference in overall R^2 between the mortality constraint-based model and the null constraint-based model from 1000 resamples: feasible set model, [0.019, 0.033]; maximum entropy model, [0.018, 0.035]; Fig 4A-D). This improvement in fit can be visualized by observing the tightening of the points to the 1:1 line when *Ribeiroia*-induced mortality was included in the model (Fig. 4A-D).

Host-heterogeneity

There were 124 unique host-parasite distribution that were overaggregated relative to one of the constraint-based models. Considering only these overaggregated distributions, we used the regression tree analysis described above to test whether further constraining P and H based on known host attributes improved the fit of the constraint-based models to the observed host-parasite distributions. The mixture models built from the regression tree analysis improved the fit of the constraint-based models to the empirical data beyond what would be expected by the inevitable increase in fit by simply grouping hosts (overall R^2 greater than the 95% interval from randomly permuting hosts into groups; Fig. 5). The improvement in fit can be visualized in Figure 4 by noting how the data points compress to the 1:1 line as more groups of heterogeneity are included. In particular, the overaggregation relative to a given model became less pronounced (though not absent) as host heterogeneity was included. Note that this increase in model fit was not achieved by minimizing or maximizing any criteria about how well the mixture model fit the observed host-parasite distribution. Finally, the largest improvement in the fit of the constraint-based models was when body-size heterogeneity, rather than co-infection, was included as an additional constraint on P and H (Fig. 6A-D).

Discussion

The shape of a sampled host-parasite distribution is necessarily constrained by the total number of hosts H and the total number of parasites P found in that sample (Haegeman and Etienne 2010, Locey and White 2013). While there are indisputably biological mechanisms leading to the shape of this distribution (Wilson et al. 2002), there may be little biological information in this distribution beyond what is contained in P and H. Here we use an extensive dataset of 22 host-parasite combinations and 842 empirical host-parasite distributions to show that aggregated host-parasite distributions tend to be consistent with the most likely distribution given P and H, suggesting that analyzing the shape of a host-parasite distribution may be less important than understanding the mechanisms leading to the observed mean parasite load.

This finding has three important implications for disease ecology. First, there is a rich history in parasitology of using the shape of host-parasite distributions in combination with dynamic models and statistical techniques to infer which mechanisms may be affecting a given host-parasite system (Crofton 1971, Anderson and Gordon 1982, Grear and Hudson 2011). While these approaches are in no way inappropriate, our results show that the instances in which the shape of a host-parasite distribution contains more information beyond what is contained in P and H may be more rare than previously thought. This result is consistent with other findings showing that log mean parasite load describes up to 88% of the variation in the log variance of parasite load, leaving only 13% of the variation to be described by biological mechanisms acting on something other than the mean (Shaw and Dobson 1995, Poulin 2013). Our results take this a step further by using constraint-based models to explicitly predict the entire host-parasite distribution given P and H. We find that, similar to Poulin (2013), much of the variability in the entire host-parasite distribution (not just the variance) is well predicted by mean parasite load and how many hosts are present in the sample. As a next step, explicitly considering whether higher order moments of the observed host-parasite distributions systematically deviate from constraint-based predictions could shed additional light on whether host-parasite distributions

contain much information beyond P and H.

The second implication is that the success of constraint-based models in disease ecology will allow them to be adopted as robust null models against which empirical host-parasite distributions can be compared. Constraint-based models are being increasingly used as robust null models in community ecology to determine when ecological mechanism may be disproportionately affecting the shape of population- and community-level distributions (Locey and White 2013, Newman et al. 2014, Xiao et al. 2015b;a). In disease ecology, by using a constraint-based model to predict parasite aggregation given P and H, we can determine when a host-parasite system is showing unusual levels of aggregation (i.e. aggregation that is higher and lower than the level predicted by the constraint-based model) to help direct modeling and experimental efforts.

Third, the general success of constraint-based models in describing host-parasite distributions has important implications for understanding the dynamics of host-macroparasite systems. Most macroparasite models explicitly model the state variables H and P (Anderson and May 1978, Dobson and Hudson 1992) and examine, in addition to other biological factors, how either fixed (Anderson and May 1978) or dynamic aggregation (Kretzschmar and Alder 1993, Rosà et al. 2003) influences host and parasite dynamics. Constraint-based models in turn predict that aggregation is largely determined by exactly these state variables. Therefore, a constraint-based approach to parasite ecology can be directly linked back to a more familiar mechanistic framework by examining the implications of constraint-based, aggregation predictions on dynamics of the total number of hosts and parasites in a system.

For example, assuming that parasite aggregation follows a feasible set would allow the aggregation parameter k to change over time as this aggregation measure tracks the total number of hosts and parasites in a given system (Appendix Figure 5). While there is no closed form solution for how k depends on P and H in a feasible set, it could be easily approximated via simulation and the resulting model dynamics examined. Linking constraint-based models for describing aggregation to dynamic equations for the state variables of a system has often been alluded to

in macroecology (Supp et al. 2012, White et al. 2012), but has been difficult to implement (Harte 2011). The rich empirical and theoretical understanding of biological factors affecting the total number of hosts and the total number of parasites in a system (Kretzschmar and Alder 1993, Hudson et al. 1992, Dobson and Hudson 1992) makes disease ecology an ideal field in which to make this connection.

In addition to providing a robust null model and a unique opportunity to link dynamic, mechanistic models with a constraint-based approach, the constraint-based models can also be extended beyond null models to test the importance of potential aggregating and disaggregating mechanisms affecting host-parasite distributions. In this study, we extended the constraint-based models to include independently estimated relationships between parasite load and amphibian survival and found that accounting for the well-described negative effect of *Ribeiroia* on *P. regilla* (Johnson 1999, Johnson et al. 2012) improved the fit of the constraint-based model to empirical host-parasite distributions. While this improvement in model fit was not drastic as P and H already accounted for 87% of the variation in the distributions, it was achieved using a survival curve estimated from an independent dataset (Johnson 1999), providing strong evidence that parasite-induced mortality is influencing *P. regilla-Ribeiroia* distributions beyond just changes to P and H.

Moreover, we also found that extending constraint-based models to include host-heterogeneity helped explain observed overaggregation in parasite distributions. Including host heterogeneity by allowing P and H to depend on host attributes such as host body size and co-infection with other trematode parasites accounted for much of the overaggregation observed in these host-parasite distributions. In particular, we found that host body size was generally a more important constraint on the host-parasite distribution than a host's level of co-infection with other trematodes. This result is consistent with previous studies which have shown the importance of host age/body size heterogeneity for increasing parasite aggregation due to changes in host immunity and/or exposure to parasites with host age/body size (Pugliese et al. 1998, Poulin 2013). Moreover, while previous work has shown that co-infection can act as a type of host heterogene-

ity and increase parasite aggregation (Cattadori et al. 2008), this same work has also shown that host characteristics such as age/body size, sex, and breeding status can often be more important factors affecting parasite aggregation and host-parasite dynamics than co-infection. We acknowledge that this analysis simplifies the complex effects of co-infection as both an aggregating and disaggregating mechanism in host-parasite systems (Wilson et al. 2002, Cattadori et al. 2008) and future work should look to extend constraint-based models to more robustly account for the possible effects of co-infection on parasite aggregation.

Constraint-based models have seen little use in disease ecology despite their success in describing the population and community-level distributions of free-living individuals (Conlisk et al. 2007, White et al. 2012, Newman et al. 2014, Xiao et al. 2015a). This study shows that many host-parasite distribution are well-predicted by constraint-based models, indicating that they can be used as robust null models for understanding when the shape of a host-parasite distribution may contain information about the dominant mechanisms acting in a system beyond the information that is contained in the total number of hosts and the total number of parasites. Moreover, we show that these models can be extended to explore dominant mechanisms structuring host-parasite systems such as host-heterogeneity or parasite-induced host mortality. We are not advocating that the constraint-based approach should replace the process-based approach that has been so successful in disease ecology. Rather, the constraint-based approach is another tool in the disease ecologist's belt that can increase our understanding about when parasite aggregation is telling us something novel about our system and when we should acknowledge the statistical inevitability that sometimes host-parasite distributions simply look how they must look given their mean.

Acknowledgments

We would like to thank the numerous members of the Johnson Lab at University of Colorado, Boulder who collected and processed the thousands of samples that comprise this dataset. We

would also like to thank Bill Murdoch and Roger Nisbet for helpful comments on earlier drafts of this manuscript. The National Institutes of Health (USA) Grant 1R01GM109499 from the Ecology of Infectious Disease program provided support for this work. M.W. was supported by a National Science Foundation, USA, Graduate Research Fellowship (Grant No. DGE 1144085) and the University of California Regents (USA).

References

Alonso, D. and M. Pascual, 2006. Comment on "A keystone mutualism drives pattern in a power function". *Science* **313**:1739; author reply 1739.

Anderson, R. M. and D. M. Gordon, 1982. Processes influencing the distribution of parasite numbers within host populations with special emphasis on parasite-induced host mortalities. *Parasitology* **85**:373–398.

Anderson, R. M. and R. M. May, 1978. Regulation and stability of host-parasite interactions: I. Regulatory processes. *Journal of Animal Ecology* **47**:219–247.

Anderson, R. M. and R. M. May, 1979. Population biology of infectious diseases: Part I. *Nature* **280**:361 – 367.

Calabrese, J. M., J. L. Brunner, and R. S. Ostfeld, 2011. Partitioning the aggregation of parasites on hosts into intrinsic and extrinsic components via an extended Poisson-gamma mixture model. *PloS one* **6**:e29215.

Cattadori, I. M., B. Boag, and P. J. Hudson, 2008. Parasite co-infection and interaction as drivers of host heterogeneity. *International Journal for Parasitology* **38**:371–380.

Chan, M. S. and V. S. Isham, 1998. A stochastic model of schistosomiasis immuno-epidemiology. *Mathematical Biosciences* **151**:179–198.

Conlisk, E., M. Bloxham, J. Conlisk, B. J. Enquist, and J. Harte, 2007. A new class of models of spatial distribution. *Ecological Monographs* **77**:269–284.

Cornell, S. J., 2010. Modelling stochastic transmission processes in helminth infections. In *Modelling Parasite Transmission and Control*, chapter 5, pages 66–78.

Crofton, H. D., 1971. A quantitative approach to parasitism. *Parasitology* **62**:179–193.

Dobson, A. P. and P. J. Hudson, 1992. Regulation and stability of a free-living host-parasite system: *Trichostrongylus tenuis* in red grouse. II. Population models. *Journal of Animal Ecology* **61**:487–498.

Duerr, H. P., K. Dietz, and M. Eichner, 2003. On the interpretation of age–intensity profiles and dispersion patterns in parasitological surveys. *Parasitology* **126**:87–101.

Engmann, S. and D. Cousineau, 2011. Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. *Journal of Applied Quantitative Methods* **6**:1–17.

Fowler, A. C. and T. D. Hollingsworth, 2016. The Dynamics of Ascaris lumbricoides Infections. *Bulletin of Mathematical Biology* pages 1–19.

Frank, S. A., 2009. The common patterns of nature. *Journal of Evolutionary Biology* **22**:1563–1585.

Frank, S. A., 2014. Generative models versus underlying symmetries to explain biological pattern. *Journal of Evolutionary Biology* **27**:1172–1178.

Grafen, A. and M. E. J. Woolhouse, 1993. Does the negative binomial distribution add up? *Parasitology Today* **9**:475–477.

Grear, D. A. and P. Hudson, 2011. The dynamics of macroparasite host-self-infection: a study of the patterns and processes of pinworm (Oxyuridae) aggregation. *Parasitology* **138**:619–27.

Grutter, A. and R. Poulin, 1998. Intraspecific and interspecific relationships between host size and the abundance of parasitic larval gnathiid isopods on coral reef fishes. *Marine Ecology Progress Series* **164**:263–271.

Haegeman, B. and R. S. Etienne, 2010. Entropy maximization and the spatial distribution of species. *The American Naturalist* **175**:E74–90.

Haegeman, B. and M. Loreau, 2009. Trivial and non-trivial applications of entropy maximization in ecology: A reply to Shipley. *Oikos* **118**:1270–1278.

Harte, J., 2011. Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics. Oxford University Press, Oxford, United Kingdom.

Harte, J. and E. A. Newman, 2014. Maximum information entropy: a foundation for ecological theory. *Trends in Ecology & Evolution* **29**:384–389.

Harte, J., A. Rominger, and W. Zhang, 2015. Integrating macroecological metrics and community taxonomic structure. *Ecology Letters* **18**:1068–1077.

Hartson, R. B., S. A. Orlofske, V. E. Melin, R. T. Dillon, and P. T. J. Johnson, 2011. Land use and wetland spatial position jointly determine amphibian parasite communities. *EcoHealth* **8**:485–500.

Hudson, P. J., D. Newborn, and A. P. Dobson, 1992. Regulation and stability of a free-living host-parasite system: *Trichostrongylus tenuis* in red grouse. 1. Monitoring and parasite reduction experiments. *Journal of Animal Ecology* **61**:477–486.

Isham, V., 1995. Stochastic models of host-macroparasite interaction. *The Annals of Applied Probability* **5**:720–740.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013. Introduction to Statistical Learning with Applications in R. Springer, New York, USA.

Jaynes, E., 1982. On the rationale of maximum-entropy methods. *Proceedings of the IEEE* **70**:939–952.

Johnson, P. T., 1999. The effect of trematode infection on amphibian limb development and survivorship. *Science* **284**:802–804.

Johnson, P. T. J. and J. T. Hoverman, 2014. Heterogeneous hosts: how variation in host size, behaviour and immunity affects parasite aggregation. *The Journal of Animal Ecology* pages 1–10.

Johnson, P. T. J. and V. J. McKenzie, 2008. Effects of Environmental Change on Helminth Infections in Amphibians: Exploring the Emergence of *Ribeiroia* and *Echinostoma* Infections in North America. In *The Biology of Echinostomes*, chapter 11, pages 249–280.

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

Kretzschmar, M. and F. R. Alder, 1993. Aggregated distributions in models for patchy populations. *Theoretical Population Biology* **43**:1–30.

Lagrue, C., R. Poulin, and J. E. Cohen, 2015. Parasitism alters three power laws of scaling in a metazoan community: Taylor's law, density-mass allometry, and variance-mass allometry. *Proceedings of the National Academy of Sciences* **112**:1791–1796.

Locey, K. J. and D. J. McGlinn, 2013. Efficient algorithms for sampling feasible sets of macroecological patterns. *PeerJ* pages 1–23.

Locey, K. J. and E. P. White, 2013. How species richness and total abundance constrain the distribution of abundance. *Ecology Letters* **16**:1177–85.

McGill, B. J. and J. C. Nekola, 2010. Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. *Oikos* **119**:591–603.

Newman, E. N., M. E. Harte, N. Lowell, M. Wilber, and J. Harte, 2014. Empirical tests of within- and across species energetics in a diverse plant community. *Ecology* **95**:2815–2825.

Pacala, S. W. and A. P. Dobson, 1988. The relation between the number of parasites/host and host age: population dynamic causes and maximum likelihood estimation. *Parasitology* **96**:197–210.

Poulin, R., 2007. Are there general laws in parasite ecology? *Parasitology* **134**:763–76.

Poulin, R., 2013. Explaining variability in parasite aggregation levels among host samples. *Parasitology* **140**:541–6.

Pugliese, A., R. Rosà, and M. L. Damaggio, 1998. Analysis of model for macroparasitic infection with variable aggregation and clumped infections. *Journal of Mathematical Biology* **36**:419–47.

Raffel, T. R., J. O. Lloyd-Smith, S. K. Sessions, P. J. Hudson, and J. R. Rohr, 2011. Does the early frog catch the worm? Disentangling potential drivers of a parasite age–intensity relationship in tadpoles. *Oecologia* **165**:1031–1042.

Rosà, R. and A. Pugliese, 2002. Aggregation, stability, and oscillations in different models for host-macroparasite interactions. *Theoretical Population Biology* **61**:319–34.

Rosà, R., A. Pugliese, A. Villani, and A. Rizzoli, 2003. Individual-based vs. deterministic models for macroparasites: host cycles and extinction. *Theoretical Population Biology* **63**:295–307.

Scholz, F. and A. Zhu, 2015. kSamples: K-sample rank tests and their combinations.

Shaw, D. J. and A. P. Dobson, 1995. Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. *Parasitology* **111**:111–133.

Shaw, D. J., B. T. Grenfell, and A. P. Dobson, 1998. Patterns of macroparasite aggregation in wildlife host populations. *Parasitology* **117**:597–610.

Supp, S. R., X. Xiao, K. M. Ernest, and E. P. White, 2012. An experimental test of the response of macroecological patterns to altered species interactions. *Ecology* **93**:2505–2511.

Tompkins, D. M., A. P. Dobson, P. Arneberg, M. Begon, I. M. Cattadori, J. V. Greenman, J. A. P. Heesterbeek, P. J. Hudson, D. Newborn, A. Pugliese, A. P. Rizzoli, R. Rosa, F. Rosso, and K. Wilson, 2002. Parasites and host population dynamics. In P. J. Hudson, A. Rizzoli, B. T. Grenfell, H. Heessterbeck, and A. P. Dobson, editors, *The Ecology of Wildlife Diseases*, chapter 3, pages 45–62. Oxford University Press, Oxford.

White, E. P., K. M. Thibault, and X. Xiao, 2012. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology* **93**:1772–8.

Wilber, M. Q., S. B. Weinstein, and C. J. Briggs, 2016. Detecting and quantifying parasite-induced host mortality from intensity data: Method comparisons and limitations. *International Journal for Parasitology* **46**:59–66.

Wilson, K., O. N. Bjoernstad, A. P. Dobson, S. Merler, G. Poglayen, A. F. Read, and A. Skorping, 2002. Heterogeneities in macroparasite infections: patterns and processes. In P. J. Hudson, A. Rizzoli, B. Grenfell, H. Heesterbeek, and A. Dobson, editors, *The Ecology of Wildlife Diseases*, chapter 2, pages 6–44. Oxford University Press, Oxford.

Xiao, X., K. J. Locey, and E. P. White, 2015a. A process-independent explanation for the general form of taylor's law. *The American Naturalist* **186**:E000–E000.

Xiao, X., D. J. McGlinn, and E. P. White, 2015b. A strong test of the Maximum Entropy Theory of Ecology. *The American Naturalist* **185**:E70–80.

Zillio, T. and F. He, 2010. Modeling spatial aggregation of finite populations. *Ecology* **91**:3698–3706.

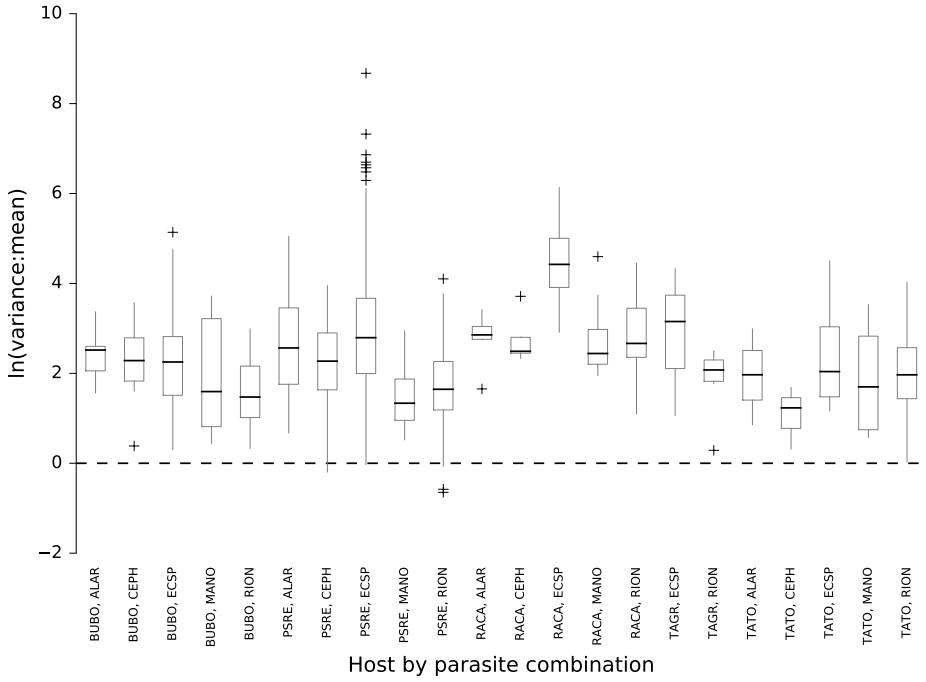

Figure 1: Boxplots of the log variance to mean ratio for each host by parasite combination for all 842 host-parasite distributions used in this analysis. The black dashed line indicates where the log variance to mean ratio is zero, consistent with an unaggregated Poisson distribution. As expected, nearly all of the host-parasite distributions have a log variance to mean ratio greater than 0, indicating an aggregated distribution.

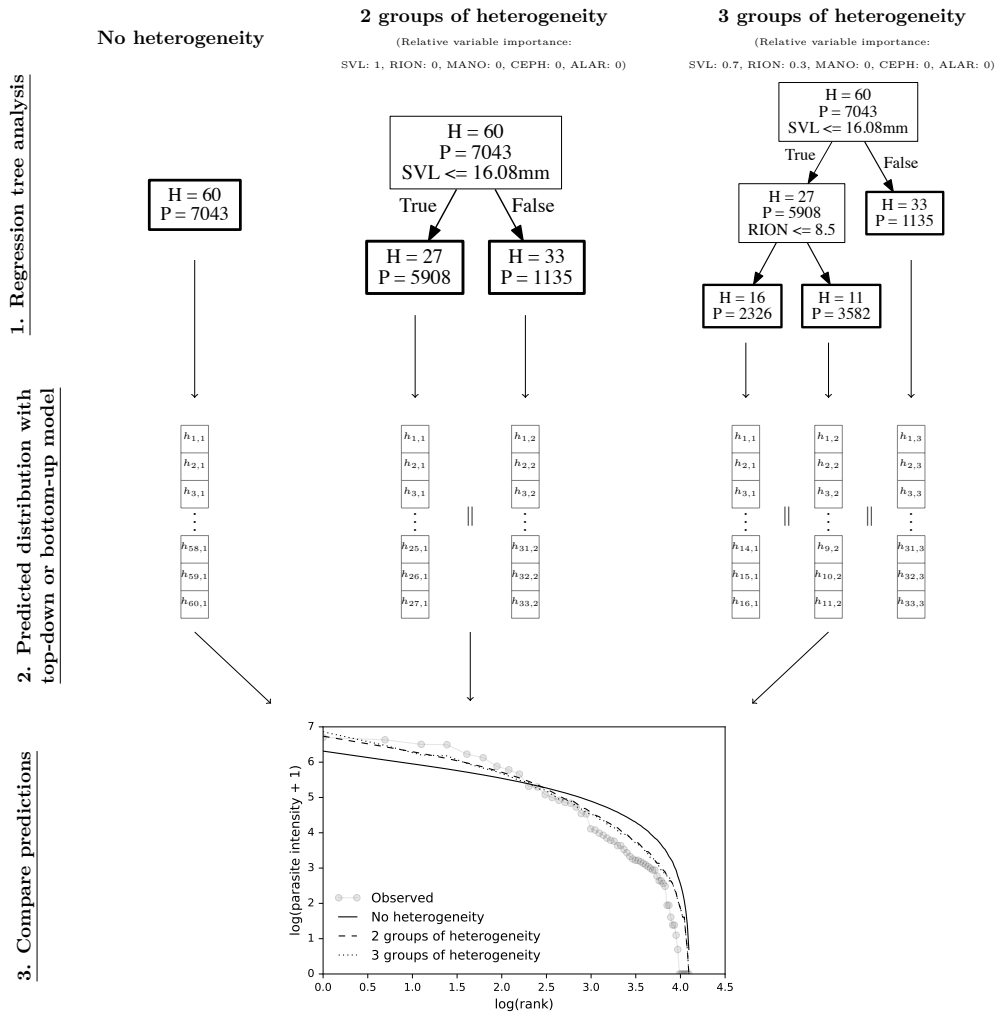

Figure 2: A diagram showing how host-heterogeneity can be incorporated into constraint-based models. **Step 1:** Consider, for example, a distribution for the parasite *Echinostoma* sp. in the host *Pseudacris regilla* with $H = 60$ hosts and $P = 7043$ parasites. When no host heterogeneity is included, the central tendency of the constraint-based model can be computed directly from H and P as described in the main text. To include groups of heterogeneity, a regression tree analysis is performed in which the response variable is *Echinostoma* abundance and the predictor variables are *Pseudacris regilla* body size (snout-vent length, SVL) and the abundance of *Ribeiroia ondatrae* (RION), *Alaria* sp. (ALAR), *Cephalogonimus* sp. (CEPH), and *Manodistomum* sp. (MANO) in a particular host. In the example above, the regression tree analysis shows that the "best" way to make two groups of heterogeneity given the predictor variables is to split the 60 *P. regilla* individuals into those with SVL $\leq$ 16.08 mm and those with SVL $>$ 16.08 mm. To make three groups of heterogeneity, *P. regilla* individuals with SVL $\leq$ 16.08 are again split into individuals with RION abundance $\leq$ 8.5. For each of these regression trees, we can determine the relative importance of each variable in building the regression tree by how much they decrease the sum of squared error compared to the other predictors. **Step 2:** We can then compute central tendency of the constraint-based model for each of these groups of heterogeneity (the bold boxes above) using the total number of hosts and parasites in each heterogeneity group. Each heterogeneity group has its own rank abundance distribution with $h_{j,i}$ being the jth ranked host with some number of parasites in the ith heterogeneity group. Concatenating ($\|$) these RADs together and re-ordering the resulting vector gives the predicted constraint-based model after allowing for P and H to vary with host heterogeneity. **Step 3:** These predicted distributions can then be compared to the observed host-parasite distribution.

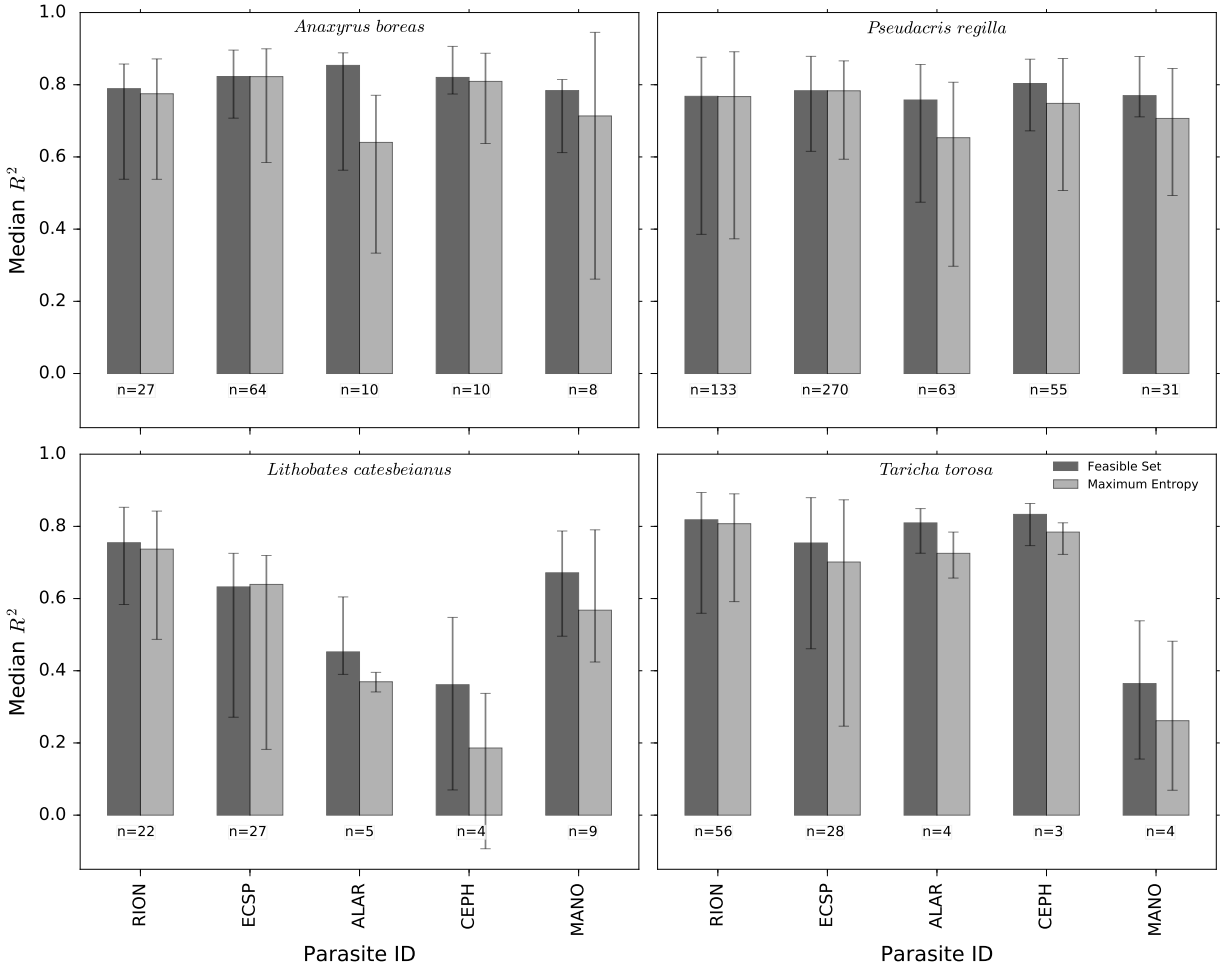

Figure 3: The height of each bar gives the median R^2 about the 1:1 line comparing the observed and predicted rank abundance distributions for various host-parasite combinations. The error bars give the first and third quartiles of the distribution of R^2 values for each distribution of a species host-parasite distribution. The number of distributions for each host-parasite combination that were used to compute this median R^2 are shown in the figure. The x-axis gives the 5 trematode parasites examined in this analysis: *Ribeiroia ondatrae* (RION), *Echinostoma* sp. (ECSP), *Alaria* sp. (ALAR), *Cephalogonimus* sp. (CEPH), and *Manodistomum* sp. (MANO). *Taricha granulosa* is not shown in this plot as it was never infected with ALAR, CEPH or MANO.

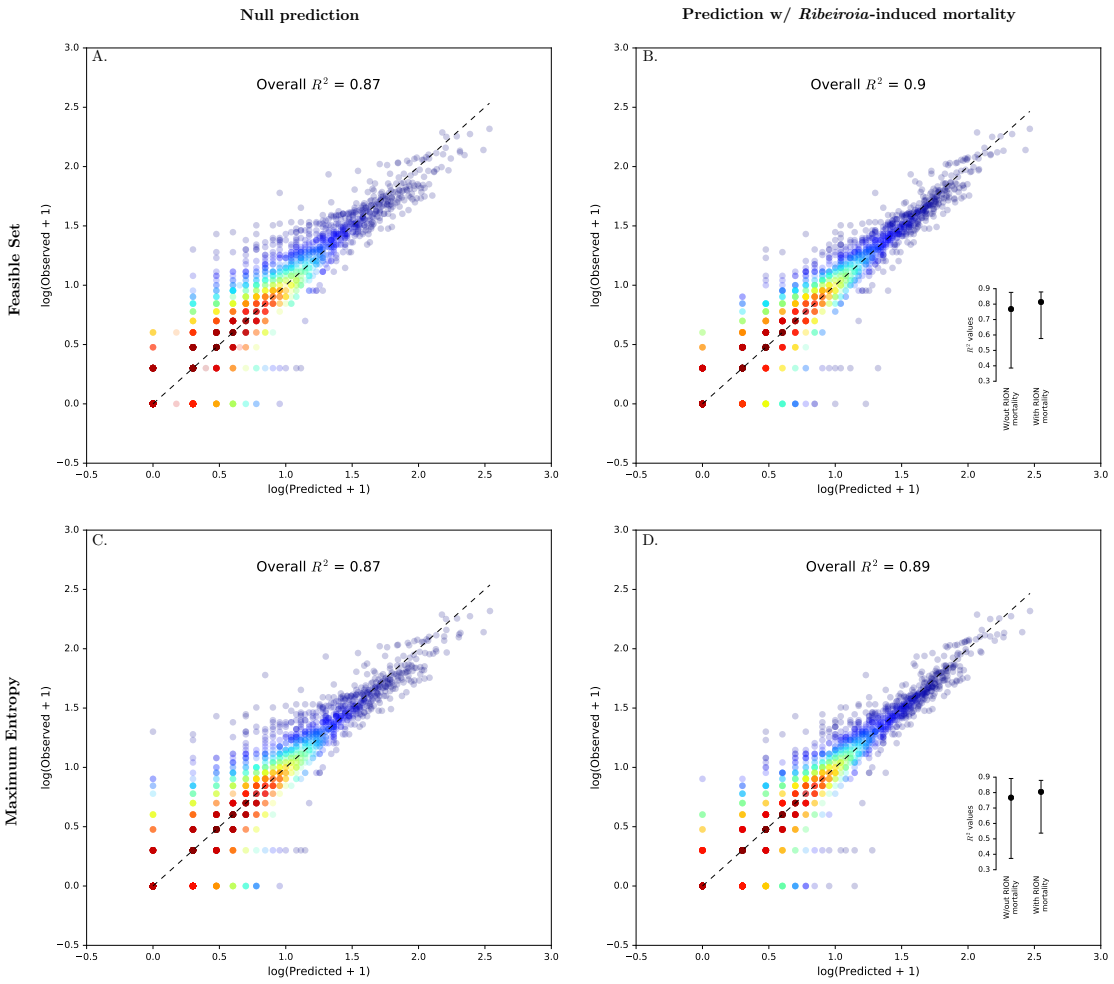

Figure 4: The figure shows the effect of including empirically-estimated *Ribeiroia*-induced *Pseudacris regilla* mortality (Johnson 1999) into the constraint-based models (feasible set and maximum entropy). The first column in this plot (**A.**, **C.**) compares 133 observed rank abundance distributions (RAD) of *Ribeiroia-P. regilla* with the RADs predicted by the constraint-based models before they were constrained on parasite-induced host mortality. The second column (**B.**, **D.**) compares the observed and predicted RADs after they were constrained on parasite-induced host mortality. The overall R^2 gives the fit of these distributions to the 1:1 line (black, dashed line). Each point represents a single host with a given predicted and observed parasite abundance. "Hotter" colors indicate a higher density of points in the region than "cooler" colors. The inset plots in B. and D. show the first through third quartiles of the R^2s from each of the 133 distributions without and with *Ribeiroia*-induced mortality. Including empirically-estimated *Ribeiroia*-induced mortality improves the fit of both constraint-based models to the *P. regilla-Ribeiroia* distributions.

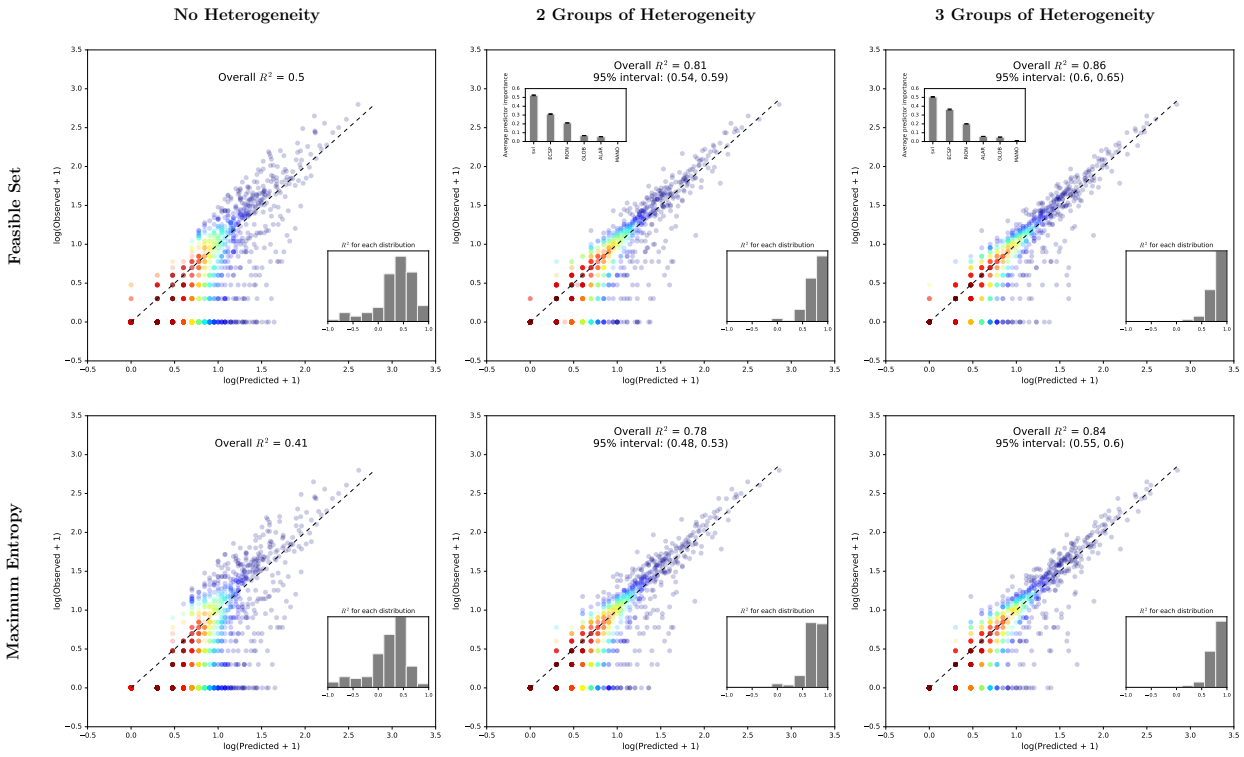

Figure 5: The figure shows the effect of discrete heterogeneity on the host-parasite distributions that were overaggregated relative to the constraint-based models (all hosts and parasites shown together). The first column in this plot shows the predicted rank abundance distributions (RAD) compared to the observed RADs when no host heterogeneity was included in the model for the two constraint-based models (feasible set and maximum entropy). The overall R^2 gives the fit of these distributions to the 1:1 line (black, dashed line) where an R^2 of 1 indicates a perfect fit to the 1:1 line and an R^2 of less than 0 indicates that 1:1 line fit is even worse than assuming that each host has a parasite abundance equal to the mean of the observed distribution. Each point represents a single host with a given predicted and observed parasite abundance. "Hotter" colors indicate a higher density of points in the region than "cooler" colors. The histogram in the lower right hand side gives the distribution of R^2 values for each particular host-parasite distribution (124 distributions). The second and third columns in this plot show the effect of adding 2 and 3 groups of host heterogeneity, respectively, on the predicted host-parasite distributions based on the results from a regression tree analysis on known host attributes in the dataset. The plots in the upper left hand corner show the mean importance of a given host attribute in structuring the regression tree for all the 124 overaggregated host-parasite distributions. The predictor variables were body-size (svl), *Echinostoma* sp. (ECSP), *Ribeiroia* (RION), *Cephalogonimus* (CEPH), *Alaria* (ALAR), and *Manodistomum* (MANO). The predictor importance was the same for all models within a heterogeneity group and are therefore only displayed once for each group. Finally, the 95% interval displayed in the plot gives the 95% quantiles of overall R^2 values based on randomly permuting parasites into the groups predicted by the regression tree analysis. If the overall R^2 is greater than the interval, it shows that the increase in R^2 from the regression tree is a result of the predictors used in the regression tree analysis, rather than just grouping itself.

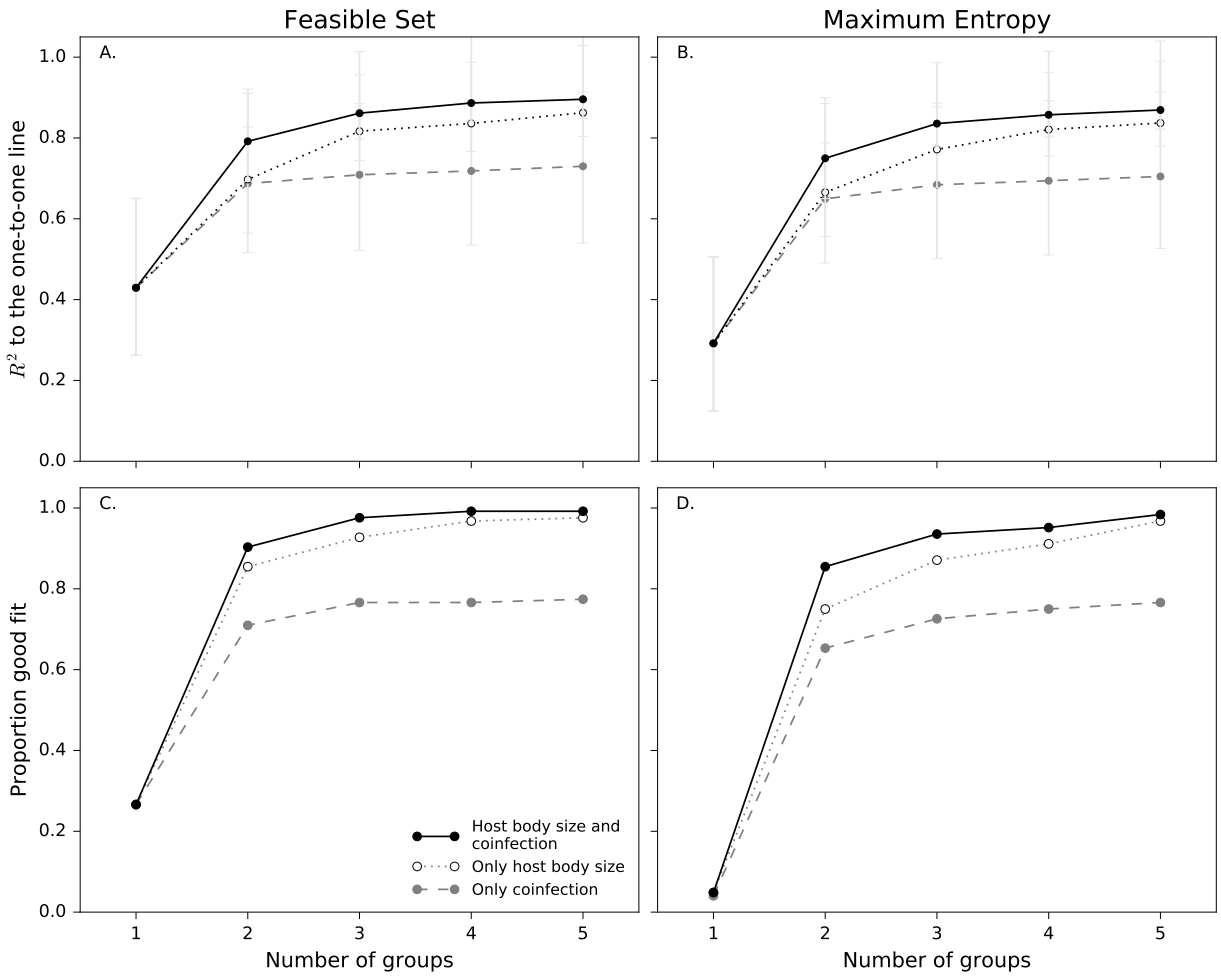

Figure 6: The effect of adding groups of adding heterogeneity to two constraint-based models (feasible sets and maximum entropy). **A.** and **B.** The black points and solid line gives the resulting median R^2 of the 124 overaggregated distributions when host heterogeneity was assumed to be predicted by host body size and/or co-infection with other larval trematodes. The white points and dotted line gives the median R^2 when host heterogeneity was assumed to be only predicted by host body size (i.e. co-infection was not included). The gray points and dashed line gives the median R^2 when host heterogeneity was assumed to be only predicted by co-infection with other larval trematodes (i.e. host body size was not included). The error bars give the first and third quartiles of the distribution of R^2 for each group number. **C.** and **D.** Same as above, but these plots give the proportion of distributions that were statistically indistinguishable from a constraint-based model based on an Anderson-Darling test.