Bayesian Analysis (2009) **4**, Number 2, pp. 1–30

# Inference in incidence, infection, and impact: Co-infection of multiple hosts by multiple pathogens

James S. Clark* and Michelle H. Hersh†

**Abstract.** A large literature concerns the epidemiology of single pathogens on single hosts. Yet in some environmental applications, such as fungal pathogens of forest tree seedlings, the "one host-one pathogen" paradigm may not be applicable. Multiple potential pathogens are often found in a single individual and/or multiple hosts share the same pathogens. Understanding diversity requires techniques to infer how multiple pathogens might regulate multiple hosts and to predict how impacts might vary with the environment. Here we present a hierarchical framework for the case where there is detection information based on multiple sources (cultures, gene sequencing, and survival observations), and the inference problem includes not only parameters that describe environmental influences on pathogen incidence, infection, and host survival, but also on latent states themselves–pathogen incidence at a site and infection statuses of hosts. Due to the large size of the model space, we develop a reversible jump Markov chain Monte Carlo approach to select models, estimate posterior distributions, and predict environmental influences on host survival. We demonstrate with application to a data set involving fungal pathogens on tree hosts, where data include host survival and fungal detection using cultures and DNA sequencing.

**Keywords:** DNA sequence data, forest dynamics, Janzen Connell hypothesis, reversible jump MCMC, species diversity, variable selection

## 1 Introduction

One of the most efficacious mechanisms for maintaining plant diversity could be the so-called 'rare-species advantage' or 'Janzen-Connell (JC) effect' (Janzen 1970, Connell 1971). Nature supports large numbers of species that compete for a small number of resources (e.g., trees compete predominantly for light, water, and several nutrients). Models used to explore how the high diversity is maintained suggest few options, most predicting extinction of all but a few species, with coexistence predicted only if there are precise combinations of traits (Tilman 1988), but these combinations are not observed in nature (Clark et al. 2007). The JC effect is an exception, predicting that coexistence of multiple species results from host-specific natural enemies, such as insect herbivores, seed predators, and plant pathogens (Gillett 1962, Janzen 1970, Connell 1971). Under this model, seedlings are more likely to survive when dispersed far from parent trees

*Nicholas School of the Environment, Department of Biology, Duke University, Durham, NC, mailto:jimclark@duke.edu

†University Program in Ecology, Department of Biology, Duke University, Durham, NC, mailto:michelle.hersh@duke.edu

 

due to escape from host-specific enemies concentrated near conspecific adults. Patterns of plant demography consistent with some aspect of JC have been found repeatedly in both tropical (e.g. Augspurger 1984, Clark and Clark 1984, Wills et al. 1997, Webb and Peart 1999, Harms et al. 2000, Peters 2003) and temperate forests (Streng et al. 1989, Jones et al. 1994, HilleRisLambers et al. 2002, HilleRisLambers and Clark 2003), but in all but a few cases the mechanism has not been directly tested.

Microbial plant pathogens, such as fungi, bacteria, and oomycetes (fungi-like protists) are frequently discussed as the most probable drivers of JC effects (Wright 2002, Gilbert 2002), in part because the microbial community is often abundant and diverse (Torsvik and Ovreas 2002, O'Brien et al. 2005), a requisite for JC to be effective. The bulk of evidence supporting pathogen-driven JC effects has highlighted examples of single host-pathogen combinations (e.g. Augspurger 1983, Packer and Clay 2000, Bell et al. 2006) or, in some cases, multiple hosts with an unknown number of pathogens (Augspurger 1984, Augspurger and Kelly 1984). However, there is mounting evidence that plants are simultaneously infected by multiple microbial species, including many that could act as pathogens (Gallery et al. 2007, Morris et al. 2007), and that these effects may be non-additive (Bradley et al. 2008). JC has not been tested in a way that allows for i) the fact that each of the competing plant species could host multiple pathogens, and ii) the fact that diversity could be enhanced by this mechanism only if there is a unique pathogen or pathogen combination regulating each host. If natural enemies are host-specific and thus disproportionately regulate their hosts only when those hosts are abundant, then as many host species as there are unique limiting pathogens or combinations thereof might coexist. An increase in pathogen infection with higher host plant density is a well-documented phenomenon in natural systems (Burdon and Chilvers 1982). Agricultural monocultures could be taken as an extreme example of density dependent host regulation, where pathogen and insect outbreaks become much more likely when a large area is occupied by a single host plant.

Until recently, testing the hypothesis of JC regulation has been challenging due to lack of adequate data on the pathogens that might regulate natural vegetation. Implicating any one pathogen as the causal agent of disease is challenging given that multiple species of fungi and oomycetes (alone or possibly in combination) can cause similar disease symptoms in seedlings (Agrios 2005). Now that data on fungi and oomycetes residing in plant tissue are becoming increasingly available, the problem becomes one of complexity: when there are observations from a large number of host species affected by a large suite of pathogens, how do we infer their combined effects? We note that for $H$ host and $K$ pathogen species, the number of host-pathogen combinations is $H \cdot 2^K$ which exceeds $10^4$ for ten species of each. In this paper we provide a general approach to this problem, showing how a hierarchical model of incidence, infection, and survival can be implemented in a variable selection context allowing for inference on the JC effect for a full suite of hosts and their fungal pathogens.

Models of plant disease infection and impact are the subjects of a huge literature (e.g., Madden et al. 2007), but largely limited to single pathogens that affect single hosts, sometimes involving an intermediate vector for the pathogen. From the perspective of a single host and a single pathogen, we could begin with a simple model of incidence $P_j$

at location $j$, infection $I_{ij}$ of individual $i$, and detection of infection $D_{ij}$. Ecologists have increasingly turned to a combination of culture-based and molecular methods to detect fungal infection (Arnold et al. 2007, Peay et al. 2008), the former characterized by relatively high uncertainty, the latter by high cost. However, observations of infection or lack thereof can be relatively uninformative by any technique. Consider a simple causal diagram of the problem (Fig. 1). Failure to detect infection could mean that the pathogen was not present at site $j$, $P_j = 0$, that the pathogen was present, but individual $ij$ was not infected $I_{ij} = 0|\ P_j = 1$, or that individual $ij$ was infected, but infection was not detected, $D_{ij} = 0|I_{ij} = 1$. Observables include not only detection $D_{ij}$, but also survival $S_{ij}$. The effect of a pathogen on host survival $p\,(S_{ij}\,|P_j = 1\,)$ depends on both the probability of infection and the probability of survival given infection or not. Of particular interest is the probability of infection given that the pathogen is present, $p(I_{ij} = 1|P_{ij} = 1)$. As described thus far, $P_j$ and $I_{ij}$ are not independently identifiable. Clearly, we require inference not only on the parameters that link each of these events, but also on the latent states in the model (incidence and infection).

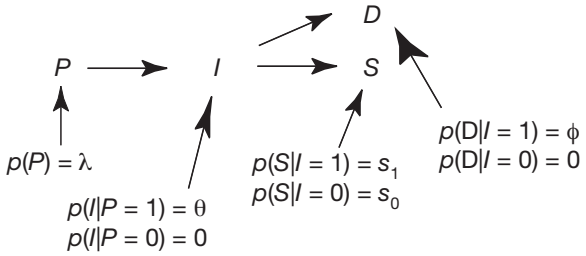


Figure 1: Graph of incidence ($P$), infection ($I$), survival ($S$), and detection ($D$) for a single pathogen and a single host. Only $D$ and $S$ are observed.

In fact the problem is substantially more complex than Figure 1, because all hosts and pathogens must be modeled together. Pathogens must be modeled together, because host survival should depend on the full pathogen load–the JC effect could as readily apply to combinations of pathogens as it could to individual pathogens. Host species must be modeled together, because incidence of a given pathogen marginally depends on all of the host species it might infect–incidence must be inferred, and all host species contribute information on incidence. The full model then includes incidence $P_{jk}$ of pathogen $k$ at $j$, infection $I_{hijk}$ of individual $i$ of host species $h$ by pathogen $k$ at $j$, and host survival $S_{hijL}$ given the full pathogen load $L$, ranging from uninfected to co-infection by all pathogens included in the study. An efficacious JC effect would mean that there is a high probability that host species $h$ would be regulated by pathogen $k$, $p\,(S_{hk} = 0\,|P_k = 1\,)\,\forall\,(k \in L)$ for one and only one $h$-host/$k$-pathogen combination. Any other result would diminish the effect, meaning i) that some host species go unregulated (there exists an $h$ for which $p\,(S_{hk} = 1\,|P_k = 1\,)$ is close to 1 for all $k$ that commonly occur at all $j$) and thus could dominate to the detriment of others or ii) that pathogens tend to regulate all hosts, never allowing a rare species to benefit disproportionately from regulation of its abundant competitors (any $L$ for which $p\,(S_{hL} = 1\,|P_k = 1\,)$ is close to

0 for all $h$ does not contribute to a rare species advantage). We desire a probability statement to this effect, a prediction problem involving

$$p\left(S_h \,|E\right) = \sum_{P_L=0,1} \sum_{I_{hL}=0,1} p\left(S_h \,|I_{hL}, E\right) p\left(I_{hL} \,|P_L\right) p\left(P_L \,|E\right) \tag{1}$$

or the probability that host species $h$ survives marginalized over environmental effects $E$ on incidence, infection by combination $I_{hL}$, and its effects on host survival.

In Section 2 we describe a model that accommodates these relationships for multiple fungal taxa that may or may not regulate multiple hosts. Computation follows in Section 3, including reversible jump MCMC (Green 1995, Dellaportas and Forster 1999) as a basis for evaluating which of the potentially many host-pathogen combinations influence survival. We further discuss the prediction problem, which addresses the core issue of how survival of each host is affected by the existence of each host pathogen combination, recognizing context dependence of both hosts and pathogens. In Section 4 we illustrate the model with simulated data. In Section 5 we apply it to a data set involving fungal cultures and DNA sequence data on host tree species planted in experimental plots in the Duke Forest, NC.

# 2 Model

## 2.1 Notation and model relationships

Consider a landscape with locations $j = 1, \ldots, J$, each supporting individuals $i = 1, \ldots, n_{hj}$ of host species $h = 1, \ldots, H$ infected with pathogens having index $k = 1,\ldots, K$. There are $n = \sum_{h,j} n_{hj}$ total seedlings. To indicate infection, detection, or survival for an individual with a specific pathogen we use the subscript $k$. A population will consist of some individuals that are uninfected and others that are infected by one or more pathogens. We wish to infer the pathogen load and what it can tell us about incidence of all pathogens, probability of infection if the pathogen is present, and the effect of infection on survival.

To indicate the full pathogen load we use the subscript $L$, which is a $K$-tuple of binary indicators, taking values in the discrete space $\{0, 1\}^K$. In this application (Section 5) we consider $K = 4$ pathogens, including the two found in the largest number of hosts and two at intermediate levels of abundance that, based on previous knowledge, potentially have some pathogenic activity (Hersh et al. in prep). For any given host, there are 16 pathogen combinations; no infection is the indicator $L = 0$ for the quadruple (0,0,0,0); the indicators $L = 1, \ldots, 4$ represent, respectively, the four singletons (1,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1); $L = 5$ indicates (1,1,0,0) and so forth. The survival model for species $h$ hosting pathogen load $L$ is designated $M_{hL}$, with $M_{h0}$ indicating uninfected. There are $H \times 2^K = 96$ models in $\{M_{hL}\colon\ h = 1,\ldots, 6,\ L = 0,\ldots, 15\}$.

Detection data include cultures $D_{hijk}$, which can have false negatives (failure to detect

$k$), $p\,(D_{hijk}=0\,|I_{hijk}=1\,)>0$, but not false positives $p\,(D_{hijk}=1\,|I_{hijk}=0\,)=0$, and DNA sequence data, which are obtained directly from the cultures themselves–thus, they are taken to be correct. Culture data are available for most sampled individuals; nuclear ribosomal DNA was sequenced from subset of cultures for further identification. Because there is observation error, we need to infer infection status for any individual $ij$ for which detection data are unavailable and for those where a pathogen $k$ was not detected using either method. We further need to infer pathogen incidence $P_{jk}$ for any plot $j$ where no individual is detected with pathogen $k$. Finally, each infection combination represents a different model, so we need to compare them all, integrate them, and assess effects of each of the combinations for each host. The submodels that follow apply to the state variables and edges of the graph in Figure 1.

The probability that pathogen $k$ occurs at site $j$ is modeled as a logit

$$\begin{array}{l} Bernoulli\,(P_{jk}\,|\lambda_{jk}\,) \\ \text{logit}(\lambda_{jk}) = x_{jk}^{(\lambda)}\mathbf{a}_k = a_k + a_{km}m_j, \end{array} \tag{2}$$

where $x_{jk}^{(\lambda)}=(1,m_j)$ is a design (row) vector for intercept and soil moisture $m_j$, and $\mathbf{a}_k$ is the corresponding parameter vector. We assume *a priori* that pathogens occur independently of one another (conditionally), with dependence structure coming through covariates and hosts.

Infection of individual $hij$ by pathogen $k$ is modeled as

$$Bernoulli\,(I_{hijk}\,|\theta_{hk}\,)\,. \tag{3}$$

Note there is a different infection risk $\theta_{hk}$ for each host-pathogen combination $hk$. Infection of a single host by multiple pathogens is also taken to be (*a priori*) conditionally independent, i.e., $p\,(I_{hijL})=\prod_{k\in L}\theta_{hk}$. Survival depends on the full pathogen load

$$\begin{array}{l} Bernoulli\,(S_{hij}\,|s_{hijL}\,) \\ \text{logit}(s_{hijL}) = x_{hijL}^{(s)}\mathbf{c}_{hL}. \end{array} \tag{4}$$

where the design vector includes an intercept and the effects of light and soil moisture. If infection does not affect survival, we have model $M_{h0}$,

$$\text{logit}(s_{hij0}) = c_h + c_{hm}m_j + c_{hl}l_j, \tag{5}$$

where $m_j$ and $l_j$ are soil moisture and light, respectively, covariates known to affect plant survival. If infection by a single pathogen affects survival, we have a coefficient for host/pathogen combination $hk$,

$$\text{logit}(s_{hijk}) = c_h + c_{hk}I_{hijk} + c_{hm}m_j + c_{hl}l_j. \tag{6}$$

If $hij$ is not infected by $k$, then $I_{hijk}=0$. In other words a design matrix for this model would include a column containing infection status (0, 1) for each individual, $\{I_{hijk}\}$.

Thus far, the survival submodel seems general–infection by $k$ changes the logit by quantity $c_{hk}$. If $k$ is indeed pathogenic for $h$, then $c_{hk} < 0$. Once we move beyond a single pathogen, there is no prior knowledge to guide model building. For example, the parsimonious model when co-infected by the first two pathogens $L$ = (1,1,0,0) might seem to be

$$\text{logit}(s_{hijL}) = c_h + c_{h1}I_{hij1} + c_{h2}I_{hij2} + c_{h12}I_{hij1}I_{hij2} + c_{hm}m_j + c_{hl}l_j. \tag{7}$$

But we have no reason to believe that the effects are additive or that the interaction term should be of this simplistic form. For example, if already infected by a pathogen with large effects on survival, co-infection by a second pathogen might not have additional impact. Conversely, if two pathogens attack a host in different ways, their combined impact could far exceed those of individual infections. And the complexity compounds as we move to co-infection by combinations of three and more pathogens. It would be hard to extend a model like eqn (7) in a way that could accommodate the many types of potential effects from the $2^K$ combinations. We would inevitably resort to assumptions that individual parameters have similar impact when combined with different infection combinations. For example, the parameters $c_{h1}$ and $c_{h12}$ could appear in a model that would contribute to the response in eqn (7) for individuals some of which are co-infected by a third pathogen, and others not. We have no reason to believe that parameters should be combined in this way.

Our alternative approach of assigning a distinct parameter to each pathogen combination $L$ suggests the model

$$\begin{aligned} \text{logit}(s_{hijL}) &= c_{h0} + \sum_{L=1}^{15} c_{hL}I_{hijL} + c_{hm}m_j + c_{hl}l_j \\ &= c_{h0} + c_{hL} + c_{hm}m_j + c_{hl}l_j. \end{aligned} \tag{8}$$

The second line corresponds to individuals infected by combination $L$. For uninfected individuals this becomes $c_{h0} + c_{hm}m_j + c_{hl}l_j$. Model $M_{h0}$ has three parameters, and all other models $M_{hL}$ have four parameters. This structure has the advantage that it does not rely on arbitrary ways that we might represent a pathogen's effect when it occurs in different combinations. Moreover, this approach makes it easy to assess whether combinations have stronger or weaker effects than single infections, by simply comparing parameter estimates for each unique $L$. Because models differ in dimension, we use a reversible jump algorithm to compare models and to derive model averaged parameter estimates (Section 3).

The model for detection is

$$binom\left(D_{hijk}\,|N_{hijk}, \phi_k\right), \tag{9}$$

where $N_{hijk}$ is the number of cultures for individual $hijk$ with detection probabilities $\phi_k$ potentially differing among pathogens, but not depending on the host in which the

pathogen occurs. The likelihood conditioned on infection for an individual host plant is eqn (4) or (if detection data are present), the multinomial,

$$p\left(S_{hij}, D_{hijL} \,|I_{hijL}\right) = s_{hijL}^{S_{hij}} \left(1-s_{hijL}\right)^{1-S_{hij}} \prod_{k\in L} \left(\phi_k\right)^{D_{hijk}} \left(1-\phi_k\right)^{(N_{hijk}-D_{hijk})I_{hijk}} \quad (10)$$

the product involving detection being included when there are assays.

## 2.2 Prior distributions

For purposes of transparency, prior distributions for some parameters are truncated. For the most part, we felt confident in excluding certain parameter ranges (e.g., positive or negative values), and we believed that these ranges would be acceptable by other ecologists. For most parameters we did not have strong opinions on central tendency of prior distributions or their weights relative to the likelihood. The use of normal prior distributions with large variances, truncated at clearly defensible values facilitates sensitivity analysis.

For incidence parameters (2) we used the prior

$$N_2\left(\left[\begin{array}{c} a_k \\ a_{km} \end{array}\right] \middle| \left[\begin{array}{c} a_k^{(p)} \\ a_{km}^{(p)} \end{array}\right], 1000\times I_2\right) \mathbf{1}\left(\left[\begin{array}{c} -30 \\ 0 \end{array}\right] < \left[\begin{array}{c} a_k \\ a_{km} \end{array}\right] < \left[\begin{array}{c} 0 \\ 20 \end{array}\right]\right) \forall k, \quad (11)$$

where $\mathbf{1}$ is the indicator function, reflecting the fact that soil moisture is known to have a non-negative direct effect on fungal growth. The large variance makes this prior distribution weak. For infection parameters (eqn 3), we used the conjugate prior $\theta_{hk} \sim beta\left(1,1\right)$. For survival (eqn 8), we used

$$\begin{aligned} c_{hL} = \left[c_{h0}, c_{hL}, c_{hm}, c_{hl}\right]^T & \\ \sim N\left(\left[0, -1, c_{hm}^{(p)}, c_{hl}^{(p)}\right]^T, \left[10^5, 10^5, 100/n, 100/n\right] I_4\right) & \quad (12) \\ \times \mathbf{1}\left(\left[-5,-5,0,0\right]^T < c_{hL} < \left[5,0,5,5\right]^T\right) & \end{aligned}$$

This prior reflects the fact that effects of soil moisture and light are known to have positive effects on seedling survival. The values for $\left(c_{hm}^{(p)}, c_{hl}^{(p)}\right)$ are given in Table ??, based on relative differences in survival from previous studies (Ibañez et al. 2007). Infection may or may not be pathogenic. The non-negative prior means that these

fungi cannot be mutualists. However, effects of infection on survival may be zero, a possibility that enters through the reversible jump implementation (Section 3). For detection, we used a conjugate beta truncated at values below and above which we did not expect detection rates for fall. These bounds were based on observations of multiple trials, $\phi_k \sim beta\,(1,1)\,\mathbf{1}\,(g_{1k} < \phi_k < g_{2k})$. The length-$K$ bounding vectors are $g_1 = [0.3, 0.4, 0.3, 0.4]^T$ and $g_2 = [0.95, 0.85, 0.9, 0.8]^T$.

| Host | Light $c_{hm}^{(p)}$ | Soil moisture $c_{hl}^{(p)}$ |
|---|---|---|
| acba | 0.8 | 1.5 |
| divi | 0.5 | 0.4 |
| list | 1.6 | 2.0 |
| litu | 3.0 | 3.5 |
| nysy | 0.5 | 0.4 |
| pita | 4.0 | 2.0 |

Table 1: Prior parameter values for survival model (eqn 11).

# 3 Computation

## 3.1 Posterior simulation

Gibbs sampling was used to simulate the posterior. We begin here with a description for a given model $M_{hL}$ followed by a description of the reversible jump algorithm in Section 3.2.

The probability that pathogen $k$ occurs at location $j$ conditionally depends on soil moisture and infection (2). If any seedling is known or imputed to be infected at location $j$ by $k$, then we impute $P_{jk} = 1$. If no seedlings are imputed to be infected by $k$ at location $j$, then

$$p\,(P_{jk}\,|I_{jk} = 0\,) = \frac{p\,(I_{jk} = 0\,|P_{jk}\,)\,p\,(P_{jk})}{\sum_{P=0,1} p\,(I_{jk} = 0\,|P_{jk}\,)\,p\,(P_{jk})} = \frac{\lambda_{jk}\prod_h (1-\theta_{hk})^{n_{hj}}}{1-\lambda_{jk}+\lambda_{jk}\prod_h (1-\theta_{hk})^{n_{hj}}}$$

where $n_{hj}$ is the number of host plants of species $h$ at site $j$, and $I_{jk} = 1 - \prod_{h,i}(1 - I_{hijk}) = 0$ indicates that all host plants at $j$ are imputed to be not infected by pathogen $k$. We draw from a Bernoulli with this probability for each $(j,k)$ for which $I_{jk} = 0$. Note that imputation of $P_{jk}$ is constrained not only by detections on all host individuals (of all species) at $j$, but also by informative prior distributions concerning soil moisture effects (eqn 12).

We need only impute infection status for host-pathogen combinations for which the pathogen is currently imputed to be present at $j$, because $p\,(I_{hijk}\,|P_{jk} = 0\,) = 0$, and for

which there is no detection. We have the conditional probability for pathogen load $L$, ignoring for the moment other subscripts,

$$p\left(I_L \left| D_L = 0, S, P_L = 1\right.\right) = \frac{p\left(D_L = 0, S \left| I_L\right.\right) p\left(I_L\right)}{\sum_{\{L\}} p\left(D_L = 0, S \left| I_L\right.\right) p\left(I_L\right)},$$

where $P_L = \prod_{k \in L} P_k \prod_{k \notin L} \left(1 - P_k\right) = 1$ if pathogens in $L$ are present. The normalizer in the denominator is complex, because the probability $s_{ijkL}$ must be computed for all individuals and for all $L$ combinations. We avoid these calculations with a Metropolis step, which allows us to simply compare current and proposed infection statuses. The infection of all *hij* by each $k$ is proposed with probability 0.5. Each individual now has a current and a proposed $L$. Of course all known host/pathogen infections–those occurring on sites $j$ where pathogen $k$ is currently imputed to be absent and any positive detections of $k$ on *hij*–are set to their known values. Conditional on incidence and infection we have

$$\begin{aligned}
&p\left(I_{hijL} \left| S_{hij,}\right.\right. \left.\left(P_{jk} = 1, D_{hijk} = 0\right) \forall k \in L\right) \\
&\quad \propto p\left(S_{hij}, D_{hijL} = 0, \left| I_{hijL}\right.\right) p\left(I_{hijL} \left| P_L = 1\right.\right) \\
&\quad = s_{hijL}^{S_{hij}} \left(1 - s_{hijL}\right)^{1 - S_{hij}} \prod_{k \in L} \left(1 - \phi_k\right)^{N_{hijk} I_{hijk}} \\
&\quad \times \prod_k \theta_{hk}^{I_{hijk}} \left(1 - \theta_{hk}\right)^{1 - I_{hijk}}.
\end{aligned}$$

This quantity is computed for current and proposed infections.

For incidence parameters $\mathbf{a}$, values were proposal from a normal truncated at the bounds given in eqn (12) and centered on the current value. Acceptance was based on the ratio of likelihoods with proposed and current values in eqn (2). For infection, we sampled from

$$\theta_{hk} \sim beta\left(1 + \sum_{ij} I_{hijk}, 1 + \sum_{ij} \left(1 - I_{hijk}\right) P_{jk}\right).$$

For detection, we sampled from

$$\phi_k \sim beta\left(1 + \sum_{hij} D_{hijk}, 1 + \sum_{hij} \left(N_{hijk} - D_{hijk}\right) I_{hijk}\right) \mathbf{1}\left(g_{1k} < \phi_k < g_{2k}\right).$$

For survival, parameters are updated as part of the reversible jump algorithm (Section 3.2).

## 3.2 Reversible jump variable selection

We are interested in the joint distribution of survival effects for each model-parameter combination$(c_M, M)$. We implement a reversible jump algorithm (Green 1995, Brooks et al. 2003) to determine model probabilities and posterior distributions conditional on specific models and averaged over models. The model where none of the pathogens affect survival is indicated by $M_{h0}$ (eqn 8, second line) and models where at least one combination of pathogens affects survival is indicated by $M_{hL}$: $L$ =1,. . . , 15 (eqn 8, third line). The parameter vectors for two such models are

$$M_{h0} : c_0 = (c_{h0}, c_m, c_l)$$
$$M_{hL} : c_L = (c_{h0}, c_{hL}, c_m, c_l)$$

having dimensions $d(M_{h0}) = 3$ and $d(M_{hL}) = 4$, respectively.

Although our application includes covariates, we recognize that this may not always be the case and first address the simpler problem without covariates. In the absence of covariates we have

$$M_{h0} : c_0 = (c_{h0})$$
$$M_{hL} : c_L = (c_{h0}, c_{hL})$$

Let $w_{hL} = n_h^{-1} \sum_{ij} I_{ijhL}$ be the fraction of individuals of host $h$ co-infected with the combination of pathogens $L$. By expressing the model in terms of a unique combination of pathogens for each individual $ij$ we have $I_{hijL} I_{hijL\prime} = 0, \forall (L \neq L\prime)$, which accounts for the sparseness of $X^{(s)}_{hijL}$, the design matrix for which eqn (4) represents a single row, and simplifies results that follow. Let $c_M$ be the parameter vector associated with model $M$, where we do not wish to distinguish between the reduced model $M_0$ and one of the enlarged models $M_L$. The target density for the survival component of the model is

$$p\left(c_M, M \,|\{S_{hij}\}\right) \propto \prod_{ij} Bernoulli\left(S_{hij} \,|c_M\right) p\left(c_M \,|M\right) p\left(M\right)), \tag{13}$$

where $p(\mathbf{c}_M)$ is given by (eqn 11). For a move between models$(c_M, M) \rightarrow (c_{M\prime}, M\prime)$ we have the acceptance criterion $A$= min(1, $a$), where

$$a = \frac{p\left(c_{M\prime}, M\prime \,|\{S_{hij}\}\right)}{p\left(c_M, M \,|\{S_{hij}\}\right)} \times \frac{J_{M\prime} q\left(u \,|u\prime\right)}{J_M q\left(u\prime \,|u\right)} \times \left| \frac{\partial G_{M,M\prime}\left(c_M, u\right)}{\partial\left(c_M, u\right)} \right|, \tag{14}$$

$J_M$ is the probability of drawing model $M$, $u \sim q() = N(0, U)$ is the density for a dimension-matching parameter $u$, and $G$ is the injection that maps parameters from model $M$ to $M\prime$ (Green 1995). In the application that follows (Section 4) the proposal probability for model $M$ is $J_M = 1/m_h$. $m_h$ is the number of models available for host $h$, which includes all combinations in $\{L\}$ for which there is currently at least

one (imputed) infected individual–any combination $L$ including infections not currently imputed to occur is not included in the sample of models. We now discuss a function that maps the current parameter vector and $u$ onto $M\prime$. We assume *a priori* that all models have equal probability.

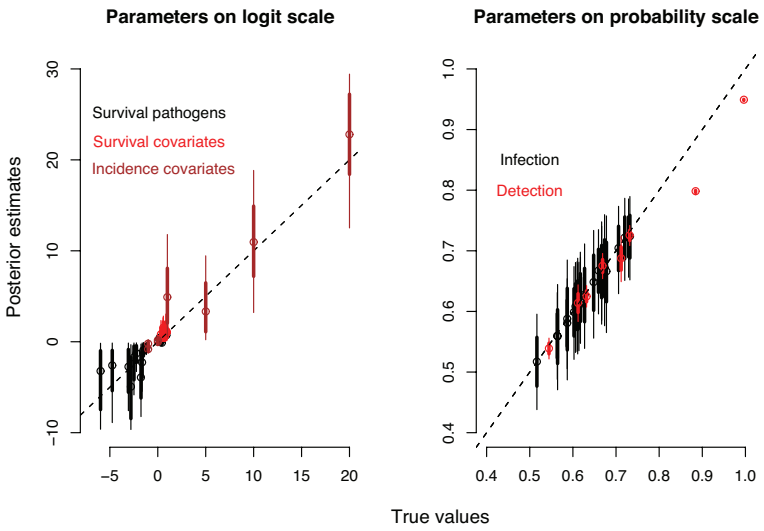


Figure 2: Posterior estimates for $c_{hk}$ ('survival pathogens'), $\{c_{hm} \times \bar{m}\}$ and $\{c_{hl} \times \bar{l}\}$ ('survival covariates'), $\mathbf{a}_k$ ('incidence covariates'), $\{\theta_{hk}\}$ ('infection'), and $\{\phi_k\}$ ('detection') plotted against true values used to simulate data. To place them on the common logit scale, soil moisture and light parameters are scaled by mean soil moisture $\bar{m}$ and mean light $\bar{l}$. Posteriors are represented by median (dot), 68% (thick line), and 95% (thin line).

The basic mapping for the transition $(c_M, M) \rightarrow (c_{M\prime}, M\prime)$ is

$$c_{M\prime} = G_{MM\prime} = g_{MM\prime} c_M, \tag{15}$$

where $g_{MM\prime} = \frac{\partial G(c_M)}{\partial (c_M)}$ is the Jacobian, $c_M$ is the current parameter vector, and $c_{M\prime}$ is the new parameter vector. There are three possible changes in dimension. First, the proposed model $M\prime$ may be of the same dimension as $M$. If so, there is no $u$. To optimize acceptance rates, we equate likelihood support, setting $X_{M\prime} c_{M\prime} = X_M c_M$ and solving to obtain

$$g_{MM\prime} = \left(X_{M\prime}^T X_{M\prime}\right)^{-1} X_{M\prime}^T X_M = \begin{bmatrix} 1 & \frac{w_{hM}}{1-w_{hM\prime}} \\ 0 & -\frac{w_{hM}}{1-w_{hM\prime}} \end{bmatrix}.$$

The determinant is $|g_{MM\prime}| = \frac{w_{hM}}{1-w_{hM\prime}}$. One might propose from, say, $N_2\left(c_M, 0.2 I_2\right)$ and accept with probability

$$a = \frac{p\left(M\prime, c_{M\prime} \left| \{S_{hij}\} \right.\right)}{p\left(M, c_M \left| \{S_{hij}\} \right.\right)} \times \frac{w_{hM}}{1 - w_{hM\prime}}.$$

The reverse move has

$$g_{M\prime M} = g_{MM\prime}^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & -\frac{(1-w_{hM\prime})}{w_{hM}} \end{bmatrix}$$

and determinant $|g_{M\prime M}| = \frac{1-w_{hM\prime}}{w_{hM}}$. If the chosen models are the same, then $g_{MM\prime}$ is the 2 by 2 identity matrix, the determinant is 1, and we have the simple Metropolis acceptance probability,

$$a = \frac{p\left(c_{M\prime}, M\prime \left| \{S_{hij}\} \right.\right)}{p\left(c_M, M \left| \{S_{hij}\} \right.\right)}.$$

For increases or decreases in dimension an injection can be defined in terms of weights $w_{hM}$. We do not pursue this further, but turn instead to the application that includes covariates.

Consider now the models in eqn (8). For the case where current and proposed are the same we use simple Metropolis. If the same dimension, we equate likelihood support with $G_{MM\prime} = g_{MM\prime} c_M$, where $g_{MM\prime} = \left(X_{M\prime}^T X_{M\prime}\right)^{-1} X_{M\prime}^T X_M$. If the current model is $M = (M_0, c_0)$, and a proposed model is$M\prime = (M_L, c_L)\, \forall L > 0$ we have an increase in dimension and draw a new vector $(c_h, u)$ based on a draw from a truncated normal having mean given by the current value $N_4\left((c_h, u\prime), UI_4\right) 1\left(u < 0\right)$, where $u\prime$ is an auxiliary variable, taken to be most recent value for the parameter $c_{hL\prime}$, i.e., when it was last included in the model (Brooks et al. 2003). The truncation at zero reflects the prior that these could be pathogens and can reduce survival, but not increase it. Means for the other parameters are taken at the current values. The determinant of the Jacobian is 1 and the acceptance probability is

$$a = \frac{p\left(c_{M\prime}, M\prime \left| \{S_{hij}\} \right.\right)}{p\left(c_M, M \left| \{S_{hij}\} \right.\right) N\left(u \left| u\prime, U \right.\right)}.$$

If $U$ is small, moves to the smaller of the two models will be rare, due to the density in the denominator (Brooks et al. 2003).

If the present model is $M = (c_L, M_L)$ and the proposed model is $M\prime = (c_0, M_0)$, then the proposal represents a decrease in dimension. Then $u$ is deterministically set equal to $c_{hL}$. The determinant of the Jacobian is still 1, and acceptance is

$$a = \frac{p\left(c_{M\prime}, M\prime \left| \{S_{hij}\} \right.\right)}{p\left(c_M, M \left| \{S_{hij}\} \right.\right)} \times N\left(u \left| u\prime, U \right.\right).$$

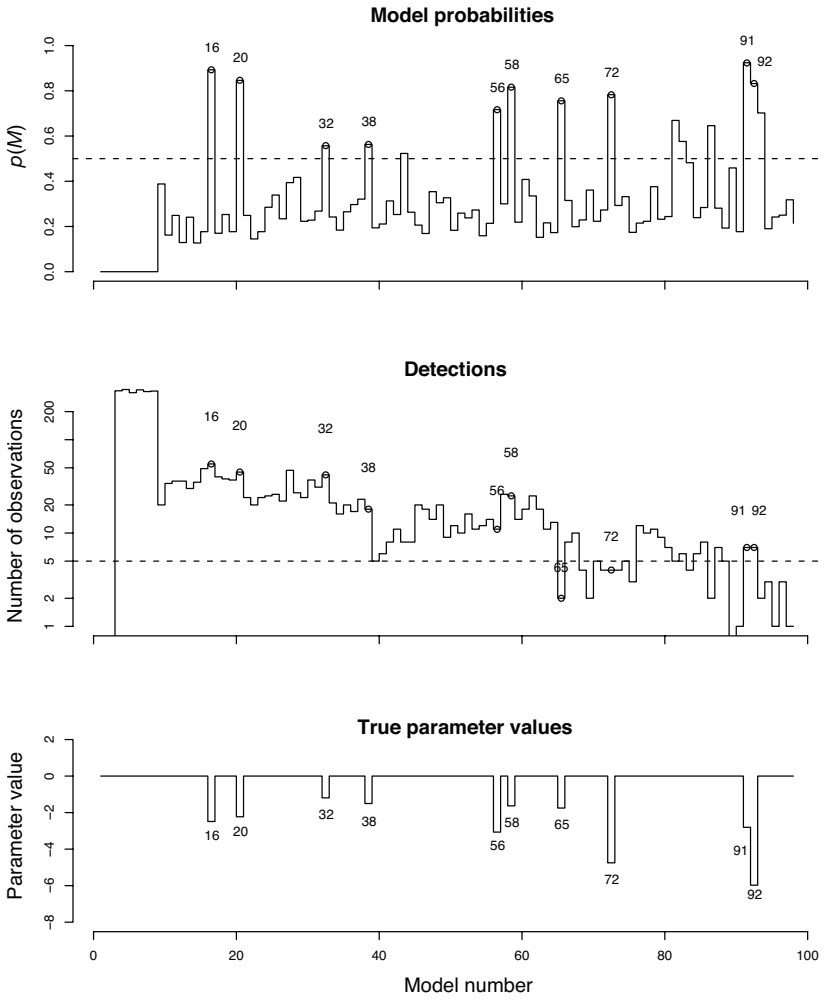


Figure 3: Model probabilities relative to the reduced model (upper), numbers of detections for each host-pathogen combination (middle), and 'true' parameter values (lower) in simulated data from Figs. 2, 3. Detections decline for higher model numbers, because fewer host individuals are infected with increasing numbers of pathogens.

The posterior model probabilities are compared with reference to the reduced model for the corresponding host $h$. An efficient scheme is to simply tally the fraction of times in which the larger model is selected over the reduced model. This approach assures that model $M$ is currently available (we only propose models for which there are currently imputed infections) and provides the probability with reference to $M_{h0}$. We call this $p(M_{hL})$.

## 3.3 Pathogen-effects prediction

We desire the probability of survival of hosts $\{h\}$ for each pathogen load $L$, given that all pathogens in $L$ are present, as compared to that for individuals in the absence of pathogens, $p\left(S_{h0}=1\,|P_K=0\right)= s_{h0}$, where $P_K = 1-\prod_{k=1}^{K}\left(1-P_k\right)=0$ indicates that no pathogens are present. The probability of survival with pathogens in $L$ being present at the site can be expressed as

$$p\left(S_{hL}=1\,|P_L=1\right) = p\left(S_{hL}\,|I_{h0}\right)p\left(I_{h0}\,|P_L=1\right) + p\left(S_{hL}\,|I_{hL}\right)p\left(I_{hL}\,|P_L=1\right)$$

$$= s_{h0}\prod_{k\in L}\left(1-\theta_{hk}\right) + s_{hL}\prod_{k\in L}\theta_{hk}, \tag{16}$$

where $I_{h0}$ indicates uninfected by pathogens in $L$, $I_{hL}$ indicates infection by all pathogens in $L$, and $P_L = 1$ indicates that only pathogens in $L$ are present. Of course, we could include terms for survival given infection by subsets of $L$, in which case we would have the combined effects of all subsets of $L$. Here we are interested in isolating the effects of the combination $L$ vs 'not $L$'. These posterior distributions are taken at the mean values for light and soil moisture of the data set.

We further predict how covariates mediate effects of individual pathogens. Pathogen incidence depends on soil moisture, and host survival depends on both soil moisture and light.

$$p\left(S_h\,|m,l\right) = \sum_{P_k=0,1}\sum_{I_k=0,1} p\left(S_h\,|I_k,m,l\right)p\left(I_k\,|P_k\right)p\left(P_k\,|m\right)$$

$$= s_{h0}(m,l)\left(1-\lambda_k\left(m\right)\right) + s_{h0}(m,l)\left(1-\theta_{hk}\right)\lambda_k\left(m\right) + s_{hk}(m,l)\theta_{hk}\lambda_k\left(m\right). \tag{17}$$

The three terms are, respectively, pathogen absent, incidence but not infection, and infection. For predictive distributions we approximate integrals over the posterior with draws from Gibbs chains.

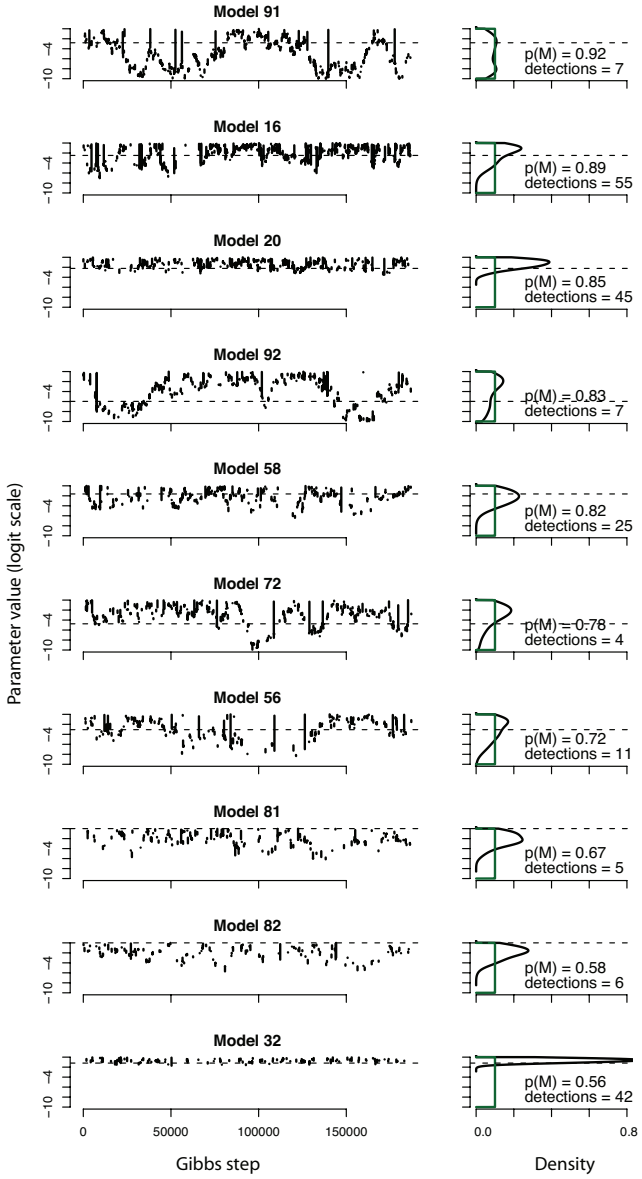


Figure 4: Example Gibbs chains (left) and posterior densities (right) for survival parameters in models having the 10 highest posterior probabilities $p(M)$, from the simulation example in Figure 3. Chains are discontinuous as parameters are dropped and reinstated in the full model. Horizontal dashed lines indicate true values. At right, prior densities are shown in green, posterior densities in black. The number of times the infection combination was detected in data is given at right. Models 81 and 82 are false positives.

# 4 Simulation studies

We used simulation studies to evaluate the model. The simulation involved these steps:

1. *Observations*: To match the application that follows, assume $H = 6$ host species, and $K = 4$ pathogens for a total of $H \times 2^K = 96$ possible infection combinations. We show an illustration with $J = 100$ plots and $n = 2000$ individuals, enough to provide a range of representations for each infection combination and be comparable to our application.

2. *Parameters*: Specify parameter values for $\mathbf{a}_k$ (incidence, eqn. 2), $\{\phi_k\}$ (detection, eqn 9), $\theta_{hk}$ (infection, eqn. 3), and $\mathbf{c}_{hL}$ (survival, eqn 4). For incidence, we used a range of positive values for effects of soil moisture. For detection, we drew from $\phi_k \sim unif(0.5, 1)$, and for infection we drew from $\theta_{hk} \sim unif\,(0.2, 0.8)$. For survival, we assumed that the 'true' model consists of a subset of the 96 total models, drawn at random uniformly from the 96. All models include a host-specific intercept $c_h \sim unif\,(0, 1)$, and the effects of light and soil moisture, spanning the range of values used as prior distributions. The survival parameters are drawn from $c_{hk} \sim N\,(-2, 4)\,I\,(-10 < c_{hk} < -1)$. Distributions reflect assumptions that light and soil moisture have a positive effect and pathogens have no effect or a negative effect on survival.

3. *Pathogen incidence* (eqn 2): Draw a random length-$J$ covariate vector from a uniform distribution for soil moisture values and concatenate with a column of ones to produce the design matrix $\mathbf{X}^{(\lambda)}$. Calculate $\lambda_{jk}$ and simulate 'true' incidence as $P_{jk} \sim Bernoulli\,(\lambda_{jk})$.

4. *Infection* (eqn 3): Draw 'true' infections from $I_{hijk} \sim Bernoulli\,(\theta_{hk}) \times P_{jk}$. For each individual there is a $K$ tuple of zeros and ones indicating infection by each pathogen.

5. *Detection* (eqn 9): Comparable to the application that follows, assume five cultures per individual for each of detection types, i.e., $N_{hijk} = 5$ and simulate detection data, that for cultures being $D_{hijk} \sim binom\,(N_{hijk}, \phi_k) \times I_{hijk}$. Also comparable to the application that follows we assumed that 30% of individual cultures were sequenced.

6. *Survival* (eqn 4): Assemble the design matrix for survival from soil moisture (step 3) and $J$ light levels, the latter simulated as a vector of uniformly distributed random variables. The design vector $x^{(s)}_{hijL}$ contains light and soil moisture values and ones in columns corresponding to the host-specific intercept and the subset of 96 columns corresponding to the infection combination for individual *hij*. All other elements are zeros. Calculate $\text{logit}(s_{hijL}) = x^{(s)}_{hijL}\mathbf{c}_{hL}$, where $\mathbf{c}_{hL}$ are zeros except for those that are contained in the 'true' model, and draw $S_{hij} \sim Bernoulli\,(s_{hijL})$.

Some general guidelines emerged from simulation studies, illustrated with an example in Figures 2 through 4. First, simulations confirm that the Gibbs sampler converges and

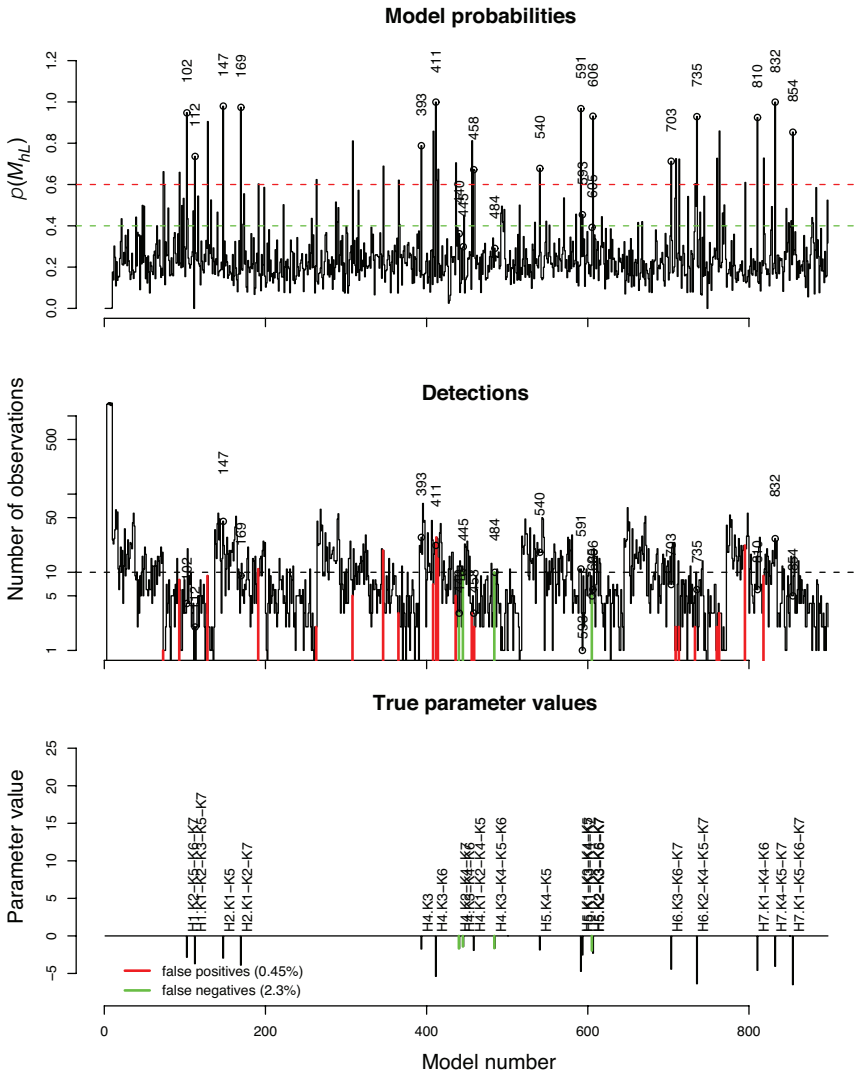


Figure 5: As in Figure 3, but with $H = K = 7$, or 896 host-pathogen combinations.

can recover parameter values. All models are visited by the Gibbs sampler, and chains for model probabilities and parameter estimates converge within $10^3$ to $10^4$ iterations. The accuracy and precision varies across the different parameter estimates in part due to the fact that some host-pathogen combinations can be abundant while others are rare. Despite weak prior densities, parameter values are recovered, the example in Figure 2 being typical. The pathogen effects on survival are of particular interest and are explored in further detail in Figures 3 and 4.

With the exception of cases where detections are rare and/or effects are weak, correct models are typically recovered. The posterior model probabilities from this example (upper panel of Figure 3) show the tendency for some false positives to occur for host-pathogen combinations that are not well represented in the data–$p(M_{hL})$ values $> 0.5$ can occur for the wrong models when detections fall to the range of 5 to 10 observations (middle panel of Figure 3). However, posterior densities for these false positives have mass centered near zero (models 81 and 82 are false positives, shown in Figure 4). Models 91 and 92 represent a different situation, being correctly identified, despite few detections (Fig. 3). Rather than being concentrated near zero, as in the case of false positives, the posterior simply recovers the prior (upper panel in Figure 4). Thus, we apparently have a distinction between false positives and correct models, when sample size is low.

False negatives can occur because effects are small, in which case we might not call them 'false', or because risk is high for other reasons, such that the additional risk from a deleterious pathogen would not be detected. A unit change on the logit scale has larger effect on probability when the logit is near zero and probability near 0.5. False negatives do not occur in the example shown in Figure 3.

The model performs well for larger numbers of combinations, provided sample sizes are large enough to include adequate representation. The model space is enlarged by an order of magnitude in Figure 5, with $H = K = 7$, or 896 combination, a sample size of $n = 10{,}000$, and 20 'correct' models. To omit the region of 'maximum uncertainty', taken to be 0.4 to 0.6, we determined false positives to be above this region and false negatives below. We still correctly identify those where detections exceed 5 to 10– false positives (0.45% in this example) are dominated by combinations that are rarely observed. Additionally, false negatives occur if both the detections are few (middle panel) and the effect is small (lower panel). Experiments of this size are feasible and relevant. It is worth noting that even when many combinations have detections in the range of five, the error rate is quite low (lower panel).

A provisional rule of thumb appears to be that, if there are few detections, a model with high posterior probability, but density concentrated near zero, might be viewed as a false positive. On the other hand, a broad posterior density for the effect of a model assigned high probability might signal need for larger sample size. Finally, models assigned posterior probability $> 0.5$ with numbers of detections exceeding 10 observations are likely correct.

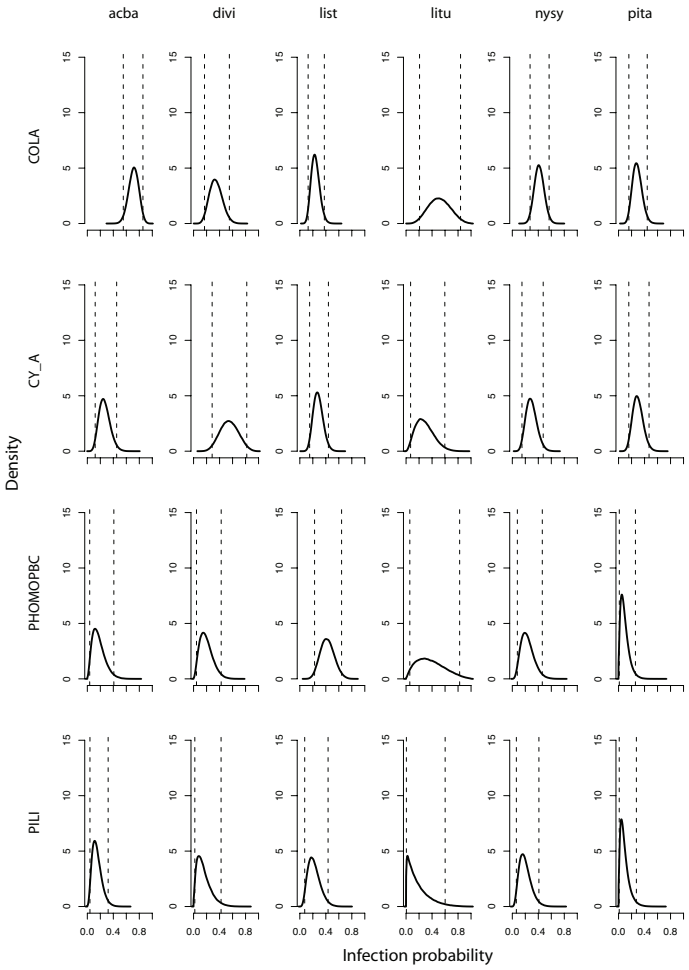


Figure 6: Posterior densities of infection risk $\theta_{hk}$ for hosts $h$ (columns) and pathogens $k$ (rows).

# 5 Application

We used this model to determine if fungi inhabiting seedlings in the Duke Forest could be differentially affecting tree seedling survival, and whether these impacts on survival are consistent with the Janzen-Connell hypothesis, i.e., there are specific pathogen combinations regulating different hosts. The data set consisted of $H = 6$ host plants and $K = 4$ potentially pathogenic fungal species. There are $J = 60$ plots from the Duke Forest, Chapel Hill, NC, where planted seedlings were assayed by culture and DNA sequencing for potentially pathogenic fungi (Table 2). Hosts are referenced by codes acba (*Acer barbatum*), divi (*Diospyros virginiana*), list (*Liquidambar styraciflua*), litu (*Liriodendron tulipifera*), nysy (*Nyssa sylvatica*), and pita (*Pinus taeda*) and fungi by codes COLA (*Colletotrichum acutatum*), CY_A (*Cylindrocarpon destructans*), PHOMOPBC (*Phomopsis* sp. BC), and PILI (*Pilidiella* sp. A). *C. acutatum* and *C. destructans* are known pathogens of multiple plant hosts, including trees (Peres et al. 2005, Dahm and Strzelczyk 1987); they were also the two most commonly isolated fungi in this study (Hersh et al. *in prep*). *Phomopsis* sp. BC and *Pilidiella* sp. A, though not yet identified to the species level, are members of genera that contain multiple known tree pathogens (Rossman et al. 2007). *Phomopsis* sp. BC and *Pilidiella* sp. A were of intermediate abundance (Hersh et al. *in prep*). These fungi are also capable of infecting some hosts asymptomatically and may have minimal effects on survival; thus we refer to them as "potential pathogens" given that pathogenicity likely depends on both host identity and environmental conditions.

There are light and soil moisture measurements for each plot $\{(l_j,\ m_j)\colon\ j = 1,\ldots,\ J\}$, which are uncorrelated ($r$ = -0.065). Seedlings were followed for one year. Those that died, along with a subset of survivors, were assayed for potential pathogens by isolating cultures on alkaline water agar (AWA) and Penicillin Rifampicin Amlicillin Pentachloronitrobenzene (PARP) media. Cultures were then transferred to Potato Dextrose Agar (PDA) or Corn Meal Agar (CMA), respectively, to allow for further growth. PDA cultures were then initially scored by macroscopic morphology (cultures growing on CMA do not exhibit the same level of morphological variation). A subset of cultures was further identified by DNA sequencing of the internal transcribed spacer (ITS) region of ribosomal DNA using primers ITS1F (Gardes and Bruns 1993) and ITS4 (White et al. 1990), recognizing cost considerations. Morphological identifications are less costly, but subject to greater error (e.g. Stefani and Berube 2006). Molecular identifications, obtained for a subset of cultures, are more reliable. Thus, individual-level observations consist of survival and zero, one, or two types of detection data. With the exception of the combination of host *L. tulipifera* and fungus *Pilidiella* sp. A, all individual host/fungus combinations were directly detected in the data (Table 2), but only some of the combinations that involve multiple fungi on a single host were detected. Moreover, only one *L. tulipifera* individual survived (Table 2). Given simulation results (Section 5), we anticipate that detections could limit inference for many of the host-fungus combinations in our data set.

The Gibbs sampler was initialized with infection status $I_{hijk} = 1$ if $k$ was detected on $h$ and with a Bernoulli draw and probability 0.5 otherwise, and with $P_{jk} = 1$ if any

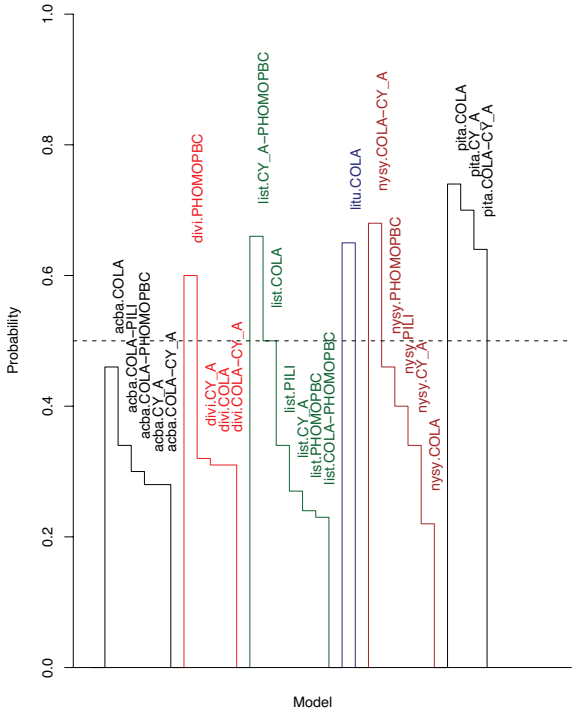


Figure 7: Posterior model probabilities ranked in order by host, then probability. Models for which there were $< 2$ detections are not shown.

| | Host | | | | | |
|---|---|---|---|---|---|---|
| | acba | divi | list | litu | nysy | pita |
| Survived | 81 | 25 | 52 | 1 | 62 | 8 |
| Total | 131 | 47 | 150 | 29 | 116 | 153 |
| Detections | | | | | | |
| PDA-COLA | 28 | 5 | 9 | 4 | 14 | 9 |
| PDA-CYA | 5 | 7 | 6 | 2 | 8 | 8 |
| PDA-PHOMOPBC | 0 | 1 | 9 | 1 | 2 | 1 |
| PDA-PILI | 2 | 1 | 4 | 0 | 3 | 1 |
| ITS-COLA | 22 | 6 | 5 | 2 | 11 | 4 |
| ITS-CYA | 7 | 6 | 8 | 2 | 8 | 9 |
| ITS-PHOMOPBC | 2 | 1 | 6 | 1 | 4 | 0 |
| ITS-PILI | 2 | 0 | 3 | 0 | 4 | 0 |
| Posterior inclusion probabilities for survival | | | | | | |
| COLA | 0.34 | 0.45 | 0.39 | 0.62 | 0.42 | 0.64 |
| CYA | 0.34 | 0.37 | 0.45 | 0.60 | 0.37 | 0.65 |
| PHOMOPBC | 0.32 | 0.52 | 0.42 | 0.61 | 0.43 | 0.58 |
| PILI | 0.34 | 0.56 | 0.37 | 0.56 | 0.45 | 0.60 |

Table 2: Number of survived and total seedlings (above), number of detections (middle), and model averaged estimates that a pathogen is included in the survival model (below).

host were found to be infected by $k$ at $j$ and randomly otherwise. All parameters were initialized with draws from the prior distributions. We report results from a single long chain following $10^6$ burnin iterations. Repeated simulation from different initializations confirmed convergence. All models including combinations for which there were no detections were visited by the RJMCMC algorithm.

Estimates of infection rate $\theta_{hk} = p(I_{hijk}|P_{jk} = 1)$ varied substantially, with only modest tendency for different hosts to experience highest infection rates by different fungi. *A. barbatum*, *N. sylvatica* and *P. taeda* were most often infected by *C. acutatum* (Figure 6). *C. acutatum* and *C. destructans* were estimated to be the most common infections. There is large uncertainty for infection of *L. tulipifera,* the host tree with the smallest sample size and almost no surviving individuals. *P. taeda* tends not to be infected by *Phomopsis* sp. BC and *Pilidiella* sp. A. *A. barbatum* also has low risk of infection by *Pilidiella* sp. A.

Given infection, survival did show evidence that different hosts are regulated by different pathogens. Figure 7 shows posterior model probabilities, omitting combinations with few detections. In light of simulations, values below 0.4 provide substantial evidence for weak or no effect on host survival in this data set. In simulation, false negatives are rare and limited to cases where there are few detections or effects too small to have impact. False positives as high as 0.6 are also rare, but cannot be ruled out given that numbers of detections in our data set is marginal. With the exception of *Acer barbatum*, each

host showed posterior model probabilities > 0.5 for a different infection combination. However, posterior densities suggest that there is not strong evidence for the strength of the effects, with many posterior densities not differing substantially from the prior (Fig. 8).

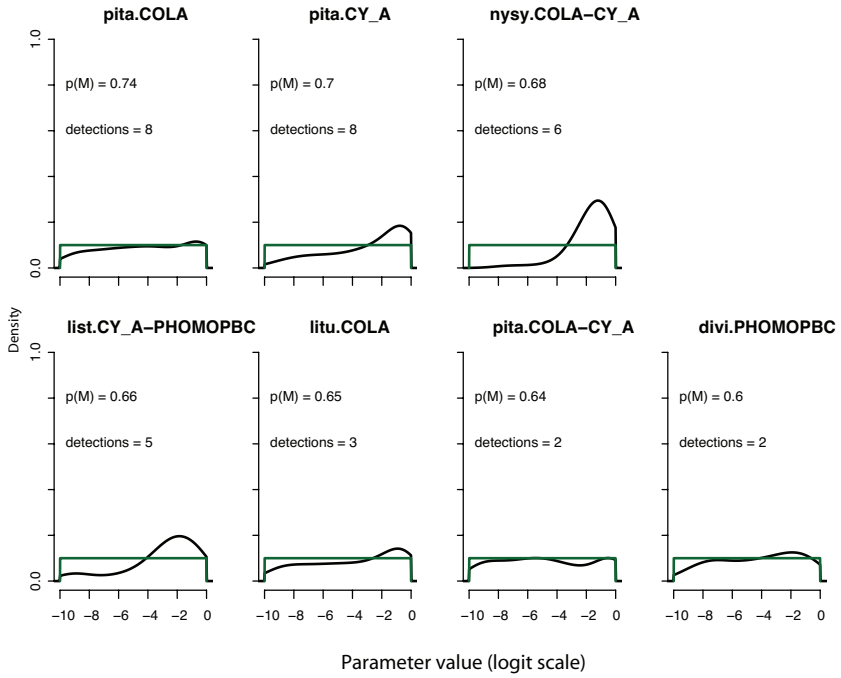


Figure 8: Posterior (black) and prior (green) densities for $c_{mL}$ for parameters in models with highest posterior probability and adequate sample size.

The posterior inclusion probability (Scott and Berger 2009) of pathogen $k$ on host $h$ is calculated as the fraction of times that any of the models containing pathogen $k$ is chosen relative to proposed, a calculation that is conditional on infection (Table 2). These estimates suggest that all four potential pathogens occur in combinations affecting survival of *L. tulipifera* and *P. taeda* at probabilities > 0.5. None exceed 0.5 for *A. barbatum*. The contrast between the probabilities for individual pathogens compared with their appearance in combinations (Fig. 7) highlights the importance of considering the full pathogen load.

Due to modest effects on survival parameter estimates, the integrated effects of infection on survival were also modest. Predictions from eqn 17 (Fig. 9) were lower than for uninfected cases for *D. virginiana, L. styraciflua*, and *N. sylvatica* (dashed lines in Figure 9). Broad overlap with the uninfected densities (dashed lines) suggests a weak effect on survival. However, limited differences on an annual basis could accumulate over years, or magnify during years of extreme soil moisture levels, beyond the range observed in this study. Predictive distributions for specific light-soil moisture combinations (eqn

17, densities not shown) did not show large effects of covariates for this study.

# 6 Discussion

With the increased accessibility of molecular tools for identification of potential microbial pathogens and continued interest in the role of disease in shaping plant communities, we expect data sets like ours to become increasingly available, with observations consisting of survival, detection of infections, and, sometimes, environmental covariates. Because the diversity of fungi in plant tissue can be high (Vandenkoornhuyse et al. 2002) and many fungi commonly isolated from plant tissue are capable of multiple lifestyles (Moricca and Ragazzi 2008), even preliminary information on how fungi affect host survival are largely unavailable. There is little prior guidance for variable selection, because pathogenic potential is likely dependent on environmental conditions and potentially contingent on the combination of infections to which a host plant is exposed. Due to the many ways in which combinations of potential pathogens and context-dependent host resistance could influence survival, there can be advantages to the modeling and computation approach taken here, where each combination of potential pathogens is represented by a specific model, and it may or may not affect survival in ways that differ from other combinations. The computational approach mixes over what could be a large number of combinations, extracts those supported by the evidence, and allows for direct comparison of their effects. Because we can readily integrate over the combined impacts of covariates on incidence and infection, we can directly address a critical question for rare-species advantage: How are our predictions of host survival across environmental gradients modified when analyses incorporate pathogens both singly and in combination? (e.g., eqns 13, 14). This capacity to extract from a large model space specific combinations that might have impact could facilitate analysis of environmental effects on incidence, infection conditioned on incidence, survival conditioned the full pathogen load, and covariates that could affect not only efficacy of the pathogen, but also host capacity to survive infection.

The initial results of this analysis demonstrate that the effects of these fungi on plant survival are not additive. In some cases, such as *Pinus taeda,* the effects of infection by either *Colletotrichum acutatum* or *Cylindrocarpon destructans* are roughly equivalent, and in fact also equivalent to their combined effects. In other cases, impacts on survival only become apparent when seedlings are infected in combination. In this case, *L. styraciflua* is not negatively affected by *Phomopsis* sp. BC or *Cylindrocarpon destructans,* alone, but is when these fungi are found in combination. These preliminary results suggest that some fungi, though isolated from multiple hosts, may not have equivalent impacts on all potential hosts (for example, *Phomopsis* sp. BC only impacts the survival of *D. virginiana* and *L. styraciflua*) while others have minimal impacts on all hosts (*Pilidiella* sp. A). Conversely, some hosts (*A. barbatum*) seem to not be strongly affected by any of these fungi, singly or in combination. The incorporation of additional fungal species isolated from this site (Hersh et al. *in prep*) will shed further light on whether unique limiting pathogens or combinations thereof might exist for each host species, and if the impacts of the fungi identified in this system are consistent with

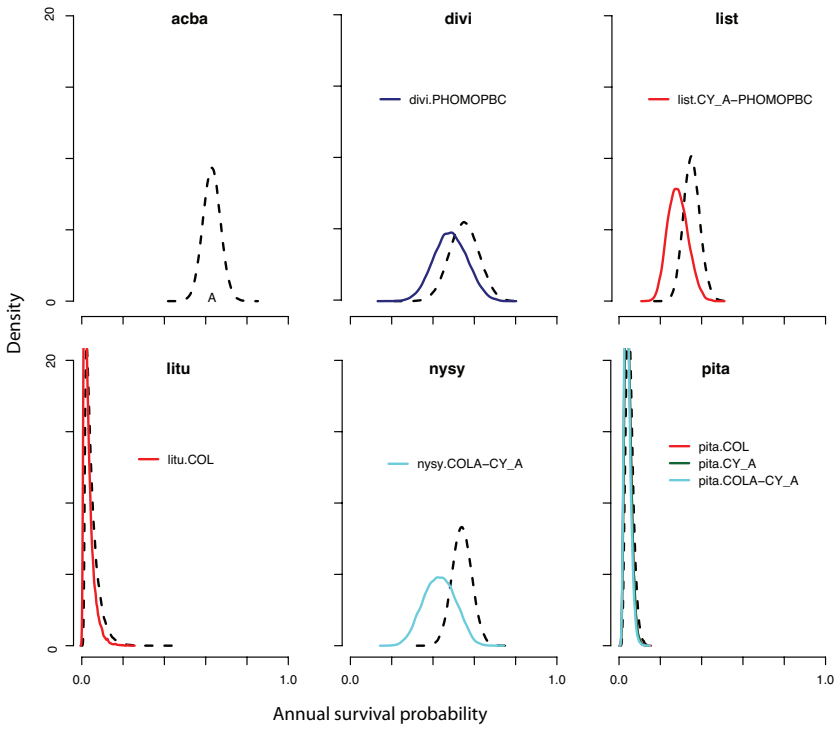


Figure 9: Posterior probability of survival accounting for the probability that fungal combination $L$ is present at a location, probability of infection by all taxa in $L$, and effects on survival.

the predictions of the Janzen-Connell hypothesis. Given that combinations of infections are common in nature, but many models only allow for a single infection at a time, this model provides novel insight into how multiple infections may affect survival.

## References

Agrios, G. 2005. Plant Pathology. Fifth edition. Elsevier Academic Press, San Diego.

Arnold, A.E., D.A. Henk, R.L. Eells, F. Lutzoni, and R. Vilgalys. 2007. Diversity and phylogenetic affinities of foliar fungal endophytes in loblolly pine inferred by culturing and environmental PCR. Mycologia 99: 185-206.

Augspurger, C. K. 1983. Seed dispersal of the tropical tree, *Platypodium elegans*, and the escape of its seedlings from fungal pathogens. Journal of Ecology 71:759-771.

Augspurger, C. K. 1984. Seedling survival of tropical tree species - interactions of dispersal distance, light-gaps, and pathogens. Ecology 65:1705-1712.

Augspurger, C. K. and C. K. Kelly. 1984. Pathogen mortality of tropical tree seedlings - experimental studies of the effects of dispersal distance, seedling density, and light conditions. Oecologia 61:211-217.

Bell, T., R. P. Freckleton, and O. T. Lewis. 2006. Plant pathogens drive density-dependent seedling mortality in a tropical tree. Ecology Letters 9:569-574.

Bradley, D. J., G. S. Gilbert, and J. B. H. Martiny. 2008. Pathogens promote plant diversity through a compensatory response. Ecology Letters 11:461-469.

Brooks, S.P., P. Giudici, and G.O. Roberts. 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. J. R. Statist. Soc. B 65: 3-55.

Burdon, J. J. and G. A. Chilvers. 1982. Host density as A factor In plant-disease ecology. Annual Review of Phytopathology **20**:143-166.

Clark, D. A. and D. B. Clark. 1984. Spacing dynamics of a tropical rain-forest tree - evaluation of the Janzen-Connell model. American Naturalist **124**:769-788.

Clark, J.S., M. Dietze, P. Agarwal, S. Chakraborty, I. Ibañez, S. LaDeau, and M. Wolosin. 2007. Resolving the biodiversity debate. Ecology Letters **10**: 647–662

Connell, J. H. 1971. On the role of natural enemies in preventing competitive exclusion in some marine animals and in rain forest trees. Pages 298-312 in P. J. Boer and G. R. Graadwell, editors. Dynamics of Numbers in Populations. Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands.

Dahm, H. and E. Strzelczyk. 1987. Cellulolytic and pectolytic activity of *Cylindrocarpon destructans* (Zins) Scholt isolates pathogenic and nonpathogenic to fir (*Abies alba* Mill) and pine (*Pinus sylvestris* L). Journal of Phytopathology **118**:76-83.

Dellaportas, P. and J.J. Forster. Markov chain Monte Carlo model determination for

hierarchical and graphical log-linear models. Biometrika 86: 615-633.

Freckleton, R. P., and O. T. Lewis. 2006. Pathogens, density dependence and the coexistence of tropical trees. Proceedings Of The Royal Society B-Biological Sciences 273:2909-2916.

Gallery, R. E., J. W. Dalling, and A. E. Arnold. 2007. Diversity, host affinity, and distribution of seed-infecting fungi: A case study with *Cecropia*. Ecology 88:582-588.

Gardes, M. and T. D. Bruns. 1993. ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. Molecular Ecology **2**:113-118.

Gilbert, G. S. 2002. Evolutionary ecology of plant diseases in natural ecosystems. Annual Review of Phytopathology **40**:13-43.

Gillett, J. B. 1962. Pest pressure, an underestimated factor in evolution. Systematic Association Publication **4**:37-46.

Green, P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711-732.

Harms, K. E., S. J. Wright, O. Calderon, A. Hernandez, and E. A. Herre. 2000. Pervasive density-dependent recruitment enhances seedling diversity in a tropical forest. Nature **404**:493-495.

HilleRisLambers, J. and J. S. Clark. 2003. Effects of dispersal, shrubs, and density-dependent mortality on seed and seedling distributions in temperate forests. Canadian Journal Of Forest Research **33**:783-795.

HilleRisLambers, J., J. S. Clark, and B. Beckage. 2002. Density-dependent mortality and the latitudinal gradient in species diversity. Nature **417**:732-735.

Ibañez, I., J.S. Clark, S. LaDeau, and J. Hille Ris Lambers 2007. Exploiting temporal variability to understand tree recruitment response to climate change, Ecological Monographs, 77:163-177.

Janzen, D. H. 1970. Herbivores and the number of tree species in tropical forests. American Naturalist 104:501-527.

Jones, R. H., R. R. Sharitz, P. M. Dixon, D. S. Segal, and R. L. Schneider. 1994. Woody plant regeneration in four floodplain forests Ecological Monographs **64**:345-367.

Madden, L. V., G. Hughes, and F. van den Bosch. 2007. The Study of Plant Disease Epidemics. APS Press, St. Paul, MN.

Moricca, S. and A. Ragazzi. 2008. Fungal endophytes in Mediterranean oak forests: A lesson from *Discula quercina*. Phytopathology **98**:380-386.

Morris, W. F., R. A. Hufbauer, A. A. Agrawal, J. D. Bever, V. A. Borowicz, G. S. Gilbert, J. L. Maron, C. E. Mitchell, I. M. Parker, A. G. Power, M. E. Torchin, and D. P. Vazquez. 2007. Direct and interactive effects of enemies and mutualists on plant performance: A meta-analysis. Ecology 88:1021-1029.

O'Brien, H. E., J. L. Parrent, J. A. Jackson, J. M. Moncalvo, and R. Vilgalys. 2005. Fungal community analysis by large-scale sequencing of environmental samples. Applied and Environmental Microbiology **71**:5544-5550.

Packer, A., and K. Clay. 2000. Soil pathogens and spatial patterns of seedling mortality in a temperate tree. Nature 404:278-281.

Peay, K. G., P. G. Kennedy, and T. D. Bruns. 2008. Fungal community ecology: A hybrid beast with a molecular master. Bioscience **58**:799-810.

Peres, N. A., L. W. Timmer, J. E. Adaskaveg, and J. C. Correll. 2005. Lifestyles of *Colletotrichum acutatum*. Plant Disease 89:784-796.

Peters, H. A. 2003. Neighbour-regulated mortality: the influence of positive and negative density dependence on tree populations in species-rich tropical forests. Ecology Letters **6**:757-765.

Rossman AY, Farr DF, Castlebury LA. 2007. A review of the phylogeny and biology of the Diaporthales. Mycoscience **48**: 135–144.

Scott, J.G. and J.O. Berger. 2009. Bayes and empirical Bayes multiplicity adjustment in the variable-selection problem, in review.

Stefani, F. O. P. and J. Berube. 2006. Biodiversity of foliar fungal endophytes in white spruce (*Picea glauca*) from southern Quebec. Canadian Journal Of Botany **84**:777-790.

Streng, D. R., J. S. Glitzenstein, and P. A. Harcombe. 1989. Woody seedling dynamics in an East Texas floodplain forest. Ecological Monographs **59**:177-204.

Torsvik, V. and L. Ovreas. 2002. Microbial diversity and function in soil: from genes to ecosystems. Current Opinion in Microbiology **5**:240-245.

Vandenkoornhuyse, P., S. L. Baldauf, C. Leyval, J. Straczek, and J. P. W. Young. 2002. Extensive fungal diversity in plant roots. Science **295**:2051-2051.

Webb, C. O. and D. R. Peart. 1999. Seedling density dependence promotes coexistence of Bornean rain forest trees. Ecology **80**:2006-2017.

White, T. J., T. D. Bruns, S. Lee, and J. W. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. . Pages 315-322 *in* M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J. White, editors. PCR Protocols: A Guide to Methods and Applications. Academic Press, New York.

Wills, C., R. Condit, R. B. Foster, and S. P. Hubbell. 1997. Strong density- and diversity-related effects help to maintain tree species diversity in a neotropical forest. Proceedings of the National Academy of Sciences of the United States of America **94**:1252-1257.

Wright, S. J. 2002. Plant diversity in tropical forests: a review of mechanisms of species coexistence. Oecologia **130**:1-14.

**Acknowledgments**

We thank two anonymous reviewers for helpful comments. This research was supported by NSF grants DDDAS 0540347, DEB 0425465, IDEA-0308498, and a NSF DDIG to Hersh. Sarah Rorick and Emily White assisted with data collection.