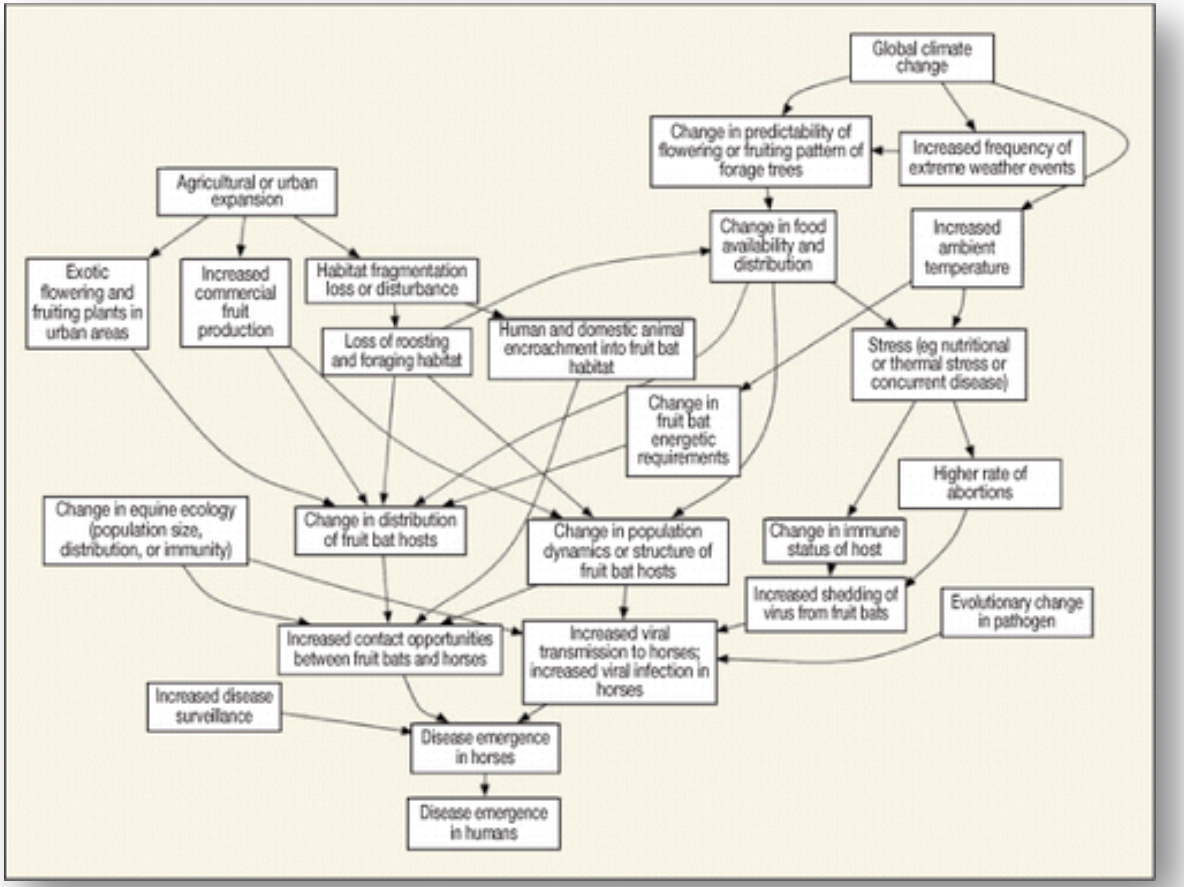
Multiple pathogens infect multiple hosts:

Inference for incidence, infection, & impact

Hypothesis

- Competing species can coexist if each is attacked when it becomes abundant
- Requires a different pathogen to regulate each host
- If **Janzen-Connell** effects maintain diversity through pathogens, then
  - Pathogens effects are **host-specific (N pathogens for N hosts)**
  - Strongest effect when host is abundant

Inference for EID



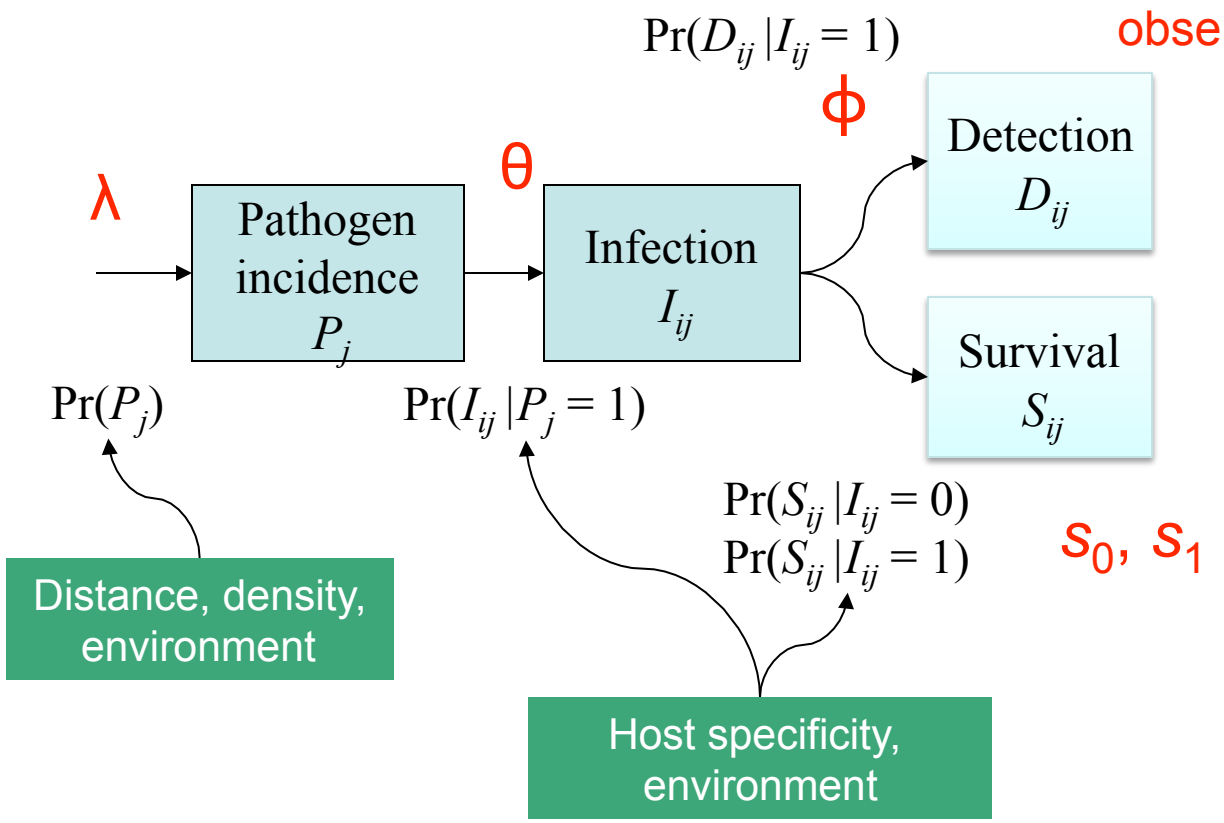
'ecological change and disease emergence are often mediated through complex processes that are not amenable to traditional causal inference'

Plowright et al. (2008) *Frontiers Ecol*

Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions

Janzen-Connell for one pathogen, one host



Clark and Hersh (2009) *Bayesian Analysis*

A classical analysis

- Observations of both pathogen and survival:

$$p_{D,S} = p\left(D,S\middle|P=1\right) = \left[\theta\phi s_1\right]^{y_{SD}} \left[\theta\phi(1-s_1)\right]^{y_{(1-S)D}} \times\left[(1-\theta)s_0 + \theta(1-\phi)s_1\right]^{y_{S(1-D)}} \times\left[(1-\theta)(1-s_0) + \theta(1-\phi)(1-s_1)\right]^{y_{(1-S)(1-D)}}$$

- Observations of survival only:

$$p_S = p\left(S\middle|P=1\right) = \left[(1-\theta)s_0 + \theta s_1\right]^{y_S} \left[1-(1-\theta)s_0 - \theta s_1\right]^{y_{1-S}}$$

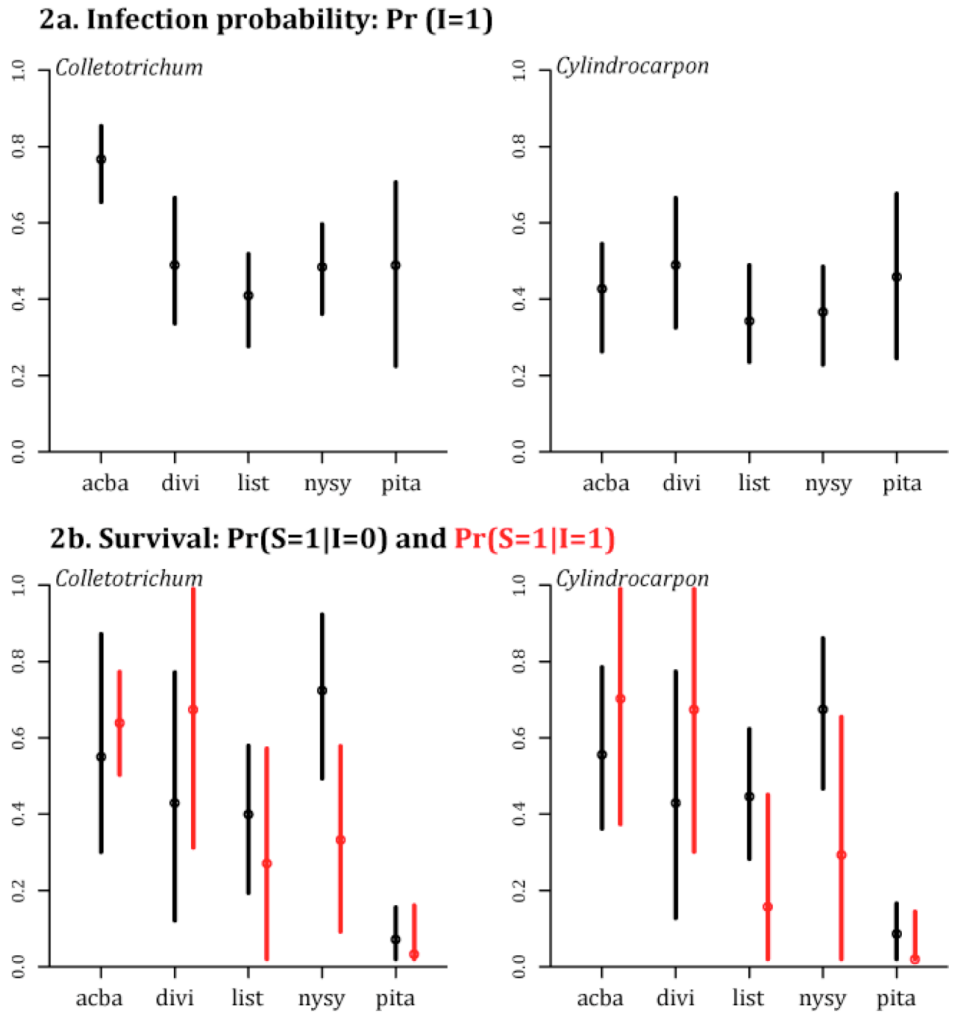
y_S - no. seedlings in two *S* classes

y_{SD} - no. seedlings in four (*S, D*) classes

$$multinom\left(\mathbf{y}_{D,S}\middle|n_{D,S},\mathbf{p}_{D,S}\right) binom\left(\mathbf{y}_S\middle|n_S,p_S\right) \propto \prod_{D,S} p_{D,S}^{y_{D,S}} \times p_S^{y_S}$$

A classical analysis

- Weak inference:
  - λθ not independently identifiable
  - Too much uncertainty
  - 80% of seedlings attached by > 1 pathogen
  - Effect of covariates?
  - Covariates vary among individuals



All hosts & pathogens must be modeled together

- Co-infection effects non-additive?
- All hosts provide information on incidence of all pathogens
- Environmental covariates affect pathogens, hosts, interactions

Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions

Application

- Fungal pathogens on seedling hosts
- Experimental plots across moisture, light gradients
  - **Culture and DNA sequence** pathogens on live and dead hosts
- **Infer** pathogen incidence, infection, survival impact
- Complication: **co-infection**

Data

- Weekly mortality, biannual growth
- Dead and live hosts sampled for pathogens
- Covariates measured
- Survival submodel: $Bernoulli\left(S_{hij} \middle| s_{hijL}\right)$

$$\text{logit}(s_{hijL}) = \mathbf{x}_{hijL}^{(s)} \mathbf{c}_{hL}$$

- Pathogen detection varies
  - DNA sequencing correct $D_{hijk}^{(s)}$
  - Cultures uncertain $p\left(D_{hijk}^{(c)} = 0 \middle| I_{hijk} = 1\right) > 0$

Model components

Incidence depends on soil moisture:

$$Bernoulli\left(P_{jk}\left|\lambda_{jk}\right.\right)$$

$$\text{logit}(\lambda_{jk}) = \mathbf{x}_{jk}^{(\lambda)}\mathbf{a}_k = a_k + a_{km}m_j$$

Infection of host *h* by *k*: $Bernoulli\left(I_{hijk}\left|\theta_{hk}\right.\right)$

Survival and detection:

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

Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions

Dimensionality of the interactions

$$\text{logit}(s_{hijL}) = c_{h0} + \sum_{L=1}^{15} c_{hL} I_{hijL} + c_{hm} m_j + c_{hl} l_j$$

$$= c_{h0} + c_{hL} + c_{hm} m_j + c_{hl} l_j$$

L - *K*-tuple of binary indicators in $\{0, 1\}^K$

For *K* pathogens on *H* hosts there are $H \times 2^K$ combinations

10 hosts & pathogens require $10 \times 2^{10} = 10{,}240$ models

Large model spaces increasingly common

- Complex systems (e.g., genomics, species and gene interactions)
- The dimensionality problem
  - *K* parasites on *H* hosts
- The multiplicity problem
  - Corrections for multiple comparisons (e.g., Bonferroni adjustment)

Traditional model selection criteria

- What they do: compare fit to the same data set
- What they do not do (well): model evaluation
- Why not to use them?
  - Fit to one data set is usually not a good criterion for building a model
- Scalability of AIC, BIC, DIC, …
  - 10 models ➔45 comparisons, 100 models ➔4950 comparisons
  - 10 hosts and pathogens ➔10×2^{10} = 10,240 models ➔ 52,423,680 comparisons
- MCMC:
  - standard M-H simulates posterior within a model, does not change dimension—each infection represents a different model

Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions

rjMcMC to evaluate high-dimensional model space

- θ_{hm} – vector of parameters for effects of each pathogen combination on survival of host h
- M_{hm} – model indicator, dimension of θ_{hm}
- Evaluate models of different dimension
- reversible jump Markov chain Monte Carlo

Posterior simulation

- Metropolis: random walk through posterior $p(\theta)$
  - Propose a parameter vector from symmetric $j(\,)$ $\quad \theta' \sim j(\theta)$
  - Accept with probability $\quad a = \dfrac{p(\theta')}{p(\theta)}$
- Metropolis-Hastings:
  - Propose from asymmetric $j(\,)$, accept with probability $a = \dfrac{p(\theta')}{p(\theta)} \times \dfrac{j(\theta|\theta')}{j(\theta'|\theta)}$
- Reversible Jump MCMC: random walk through posterior $p(\theta_m, M_m)$
  - Propose model & dimension variable $\quad M' \sim J(M) \;\; u \sim q(\theta_m, M_m)$
  - Set $\quad (\theta'_m, u') = G(M, M') \forall \left\{ G(M, M') = G^{-1}(M', M) \right\}$
  - Accept with probability $\quad a = \dfrac{p(\theta')}{p(\theta)} \times \dfrac{j(\theta|\theta')}{j(\theta'|\theta)} \times \dfrac{J(M|M')}{J(M'|M)} \times \dfrac{q(u')}{q(u)} \times \left| \dfrac{\partial G(\theta, M)}{\partial(\theta, M)} \right|$

Algorithm summary

- Algorithm
  - Propose a model
  - Select a dimension-matching variable
  - Evaluate parameter values from an invertible injection
  - Acceptance criterion
- Concerns:
  - MCMC won't mix: there are no 'local moves'
    - Metropolis can be optimized with arbitrarily small jumps
    - With RJMCMC we are changing dimensions
  - Interpretation of parameters changes with model
- This algorithm
  - Auxiliary variables and centering methods
  - Parameters have the same interpretation

Summary of inference goals

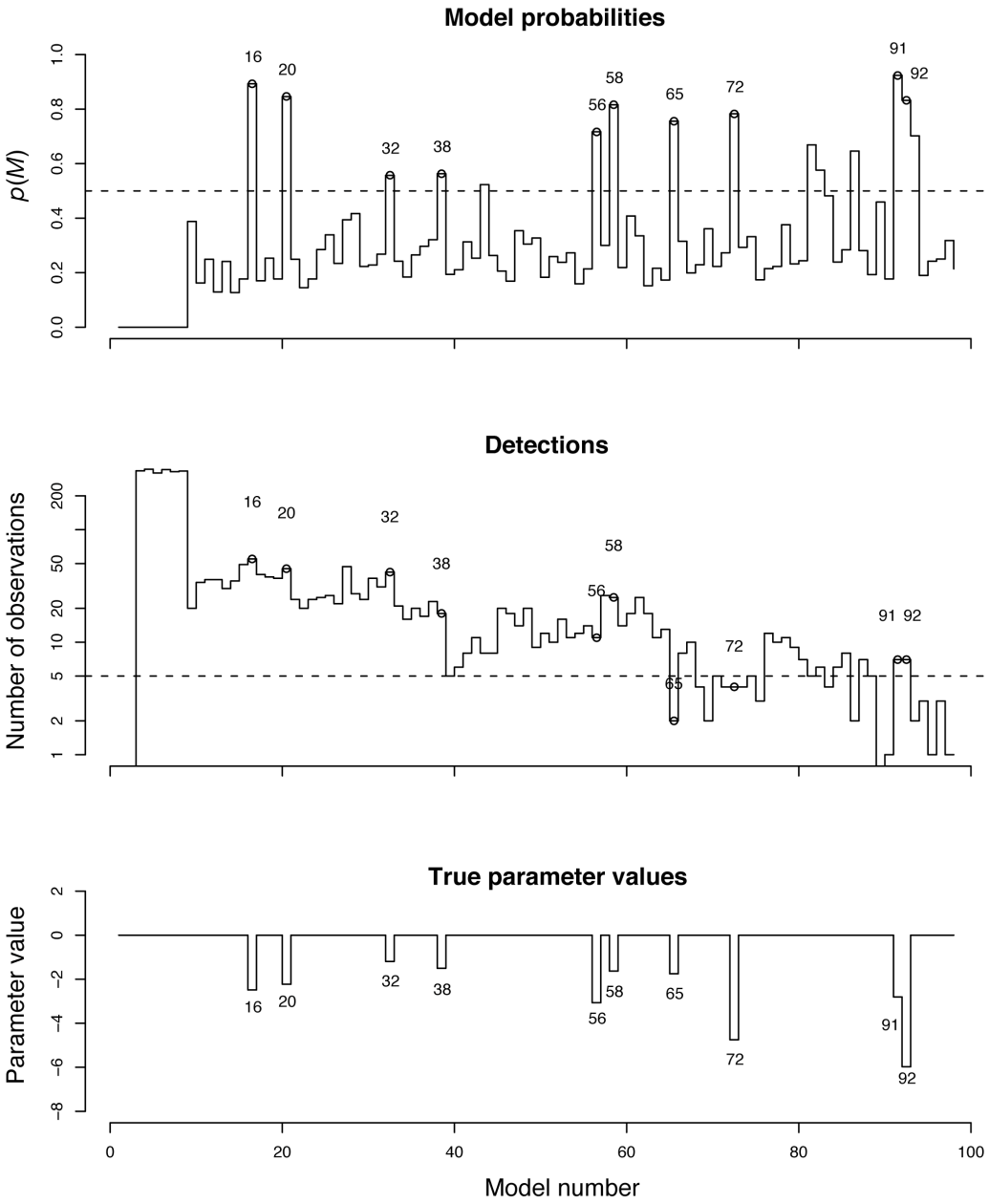
- Each host with each of 2^K pathogen combinations
  - Which are important?
  - Cannot test them all and compare, say, pairwise
- Can explore model and parameter spaces simultaneously, using RJMCMC
- Important relationships can be derived:
  - Pr(M)
  - Pr($\theta|M$)
  - Total Pr($S|P$)
  - Total Pr($S|E$)

Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions

Simulation for model evaluation

- Pathogen effect on survival
- Sample sizes like our experiment (H = 6, K = 4, n = 700)
- Correct models identified, false positives when few detections in data set



Clark and Hersh, *Bayesian Analysis* (2009)

RJMCMC chains from simulation

Chains (left) and posterior densities (right) for models having the 10 highest posterior probabilities.

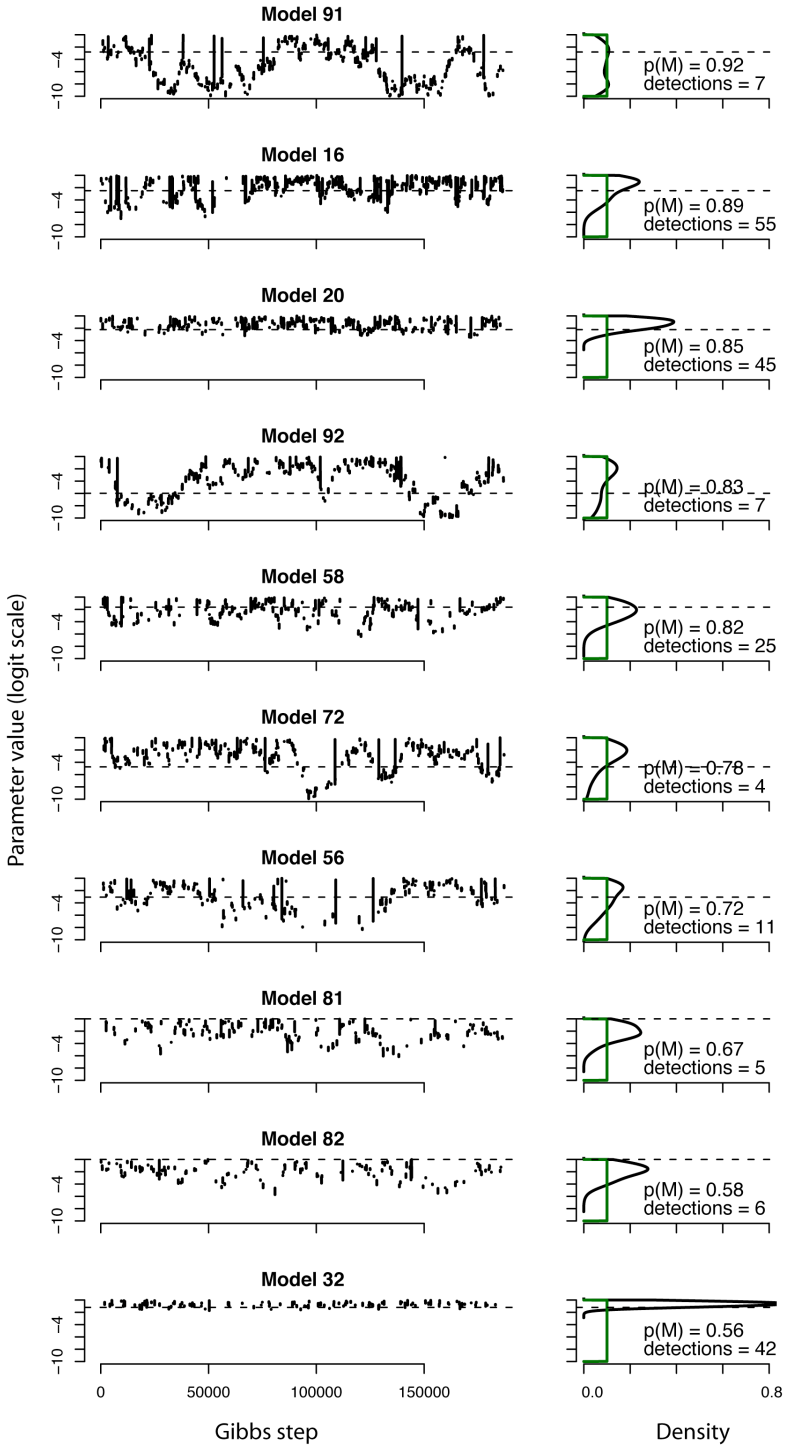
Chains are discontinuous as parameters are dropped and reinstated. Horizontal dashed lines are true values.

At right, prior densities are green, posterior densities black.

No. times the infection combination was detected is given at right.

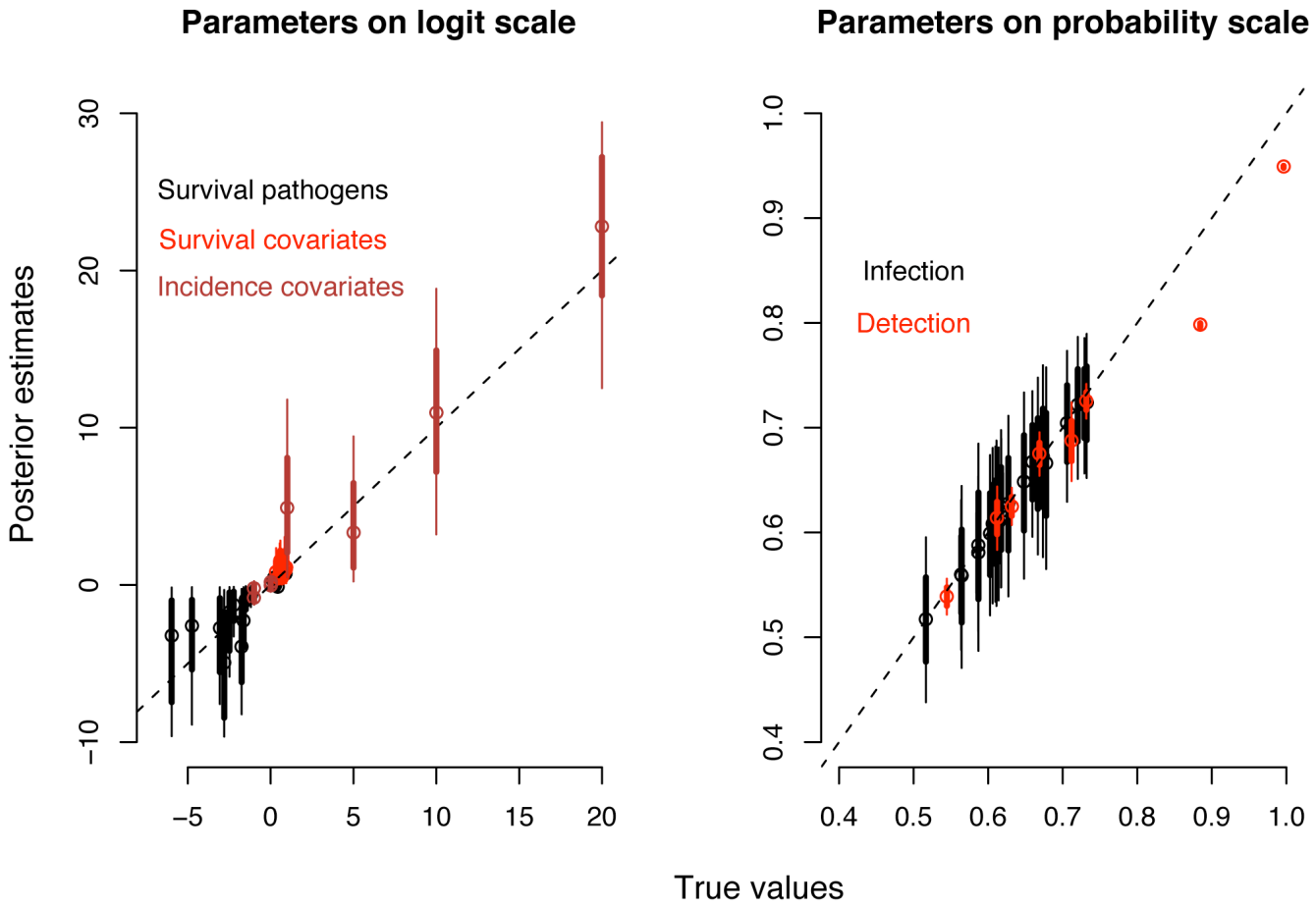
Models 81 and 82 are false positives.

Clark and Hersh, *Bayesian Analysis* (2009)



Simulation studies recover parameter values

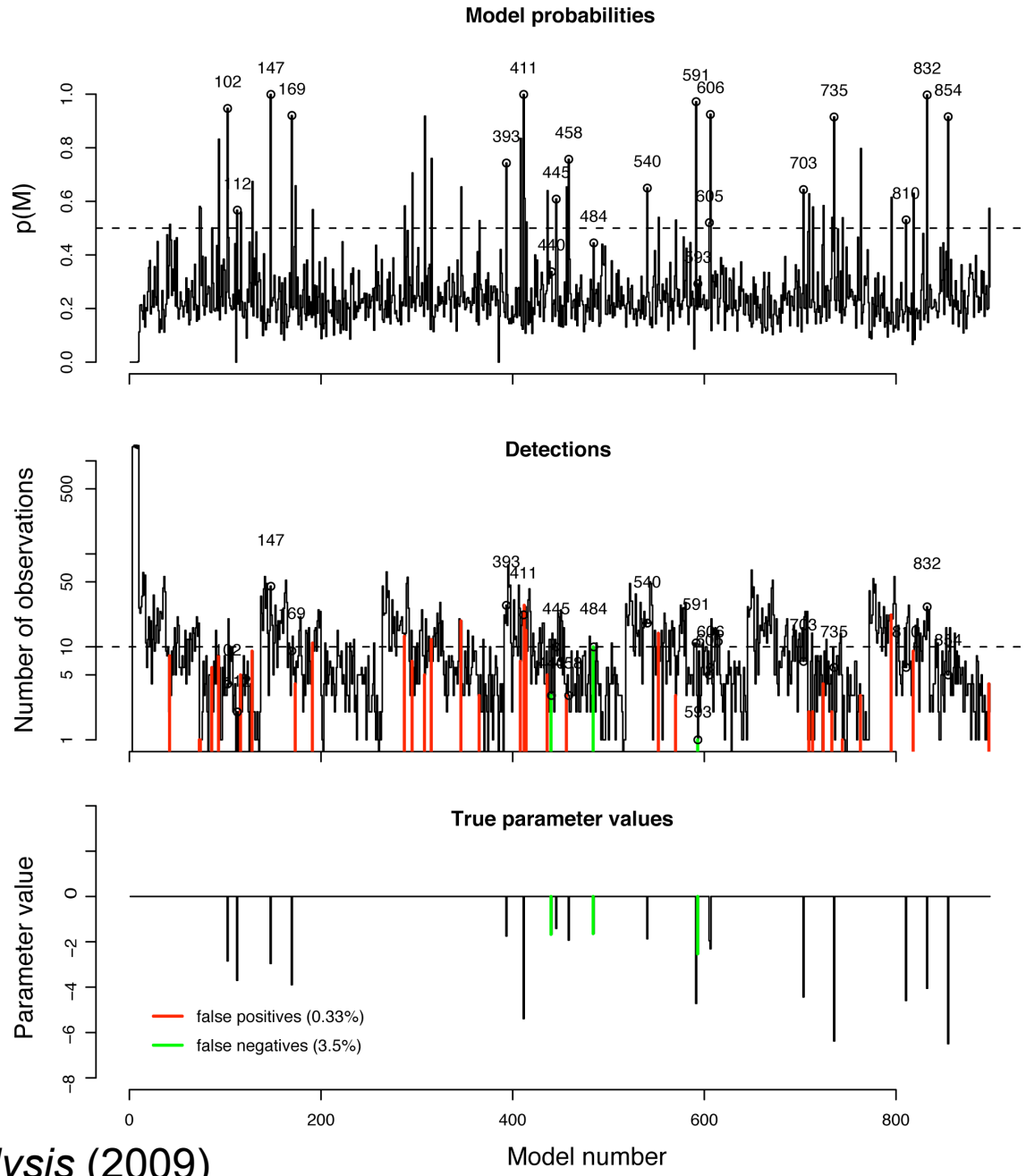
- 95% CIs include true values



Clark and Hersh, *Bayesian Analysis* (2009)

Simulation studies work with more spp

- 7 hosts, 7 pathogens = 896 models (20 correct)
- False positives when < 10 detections
- False negatives when effect is small

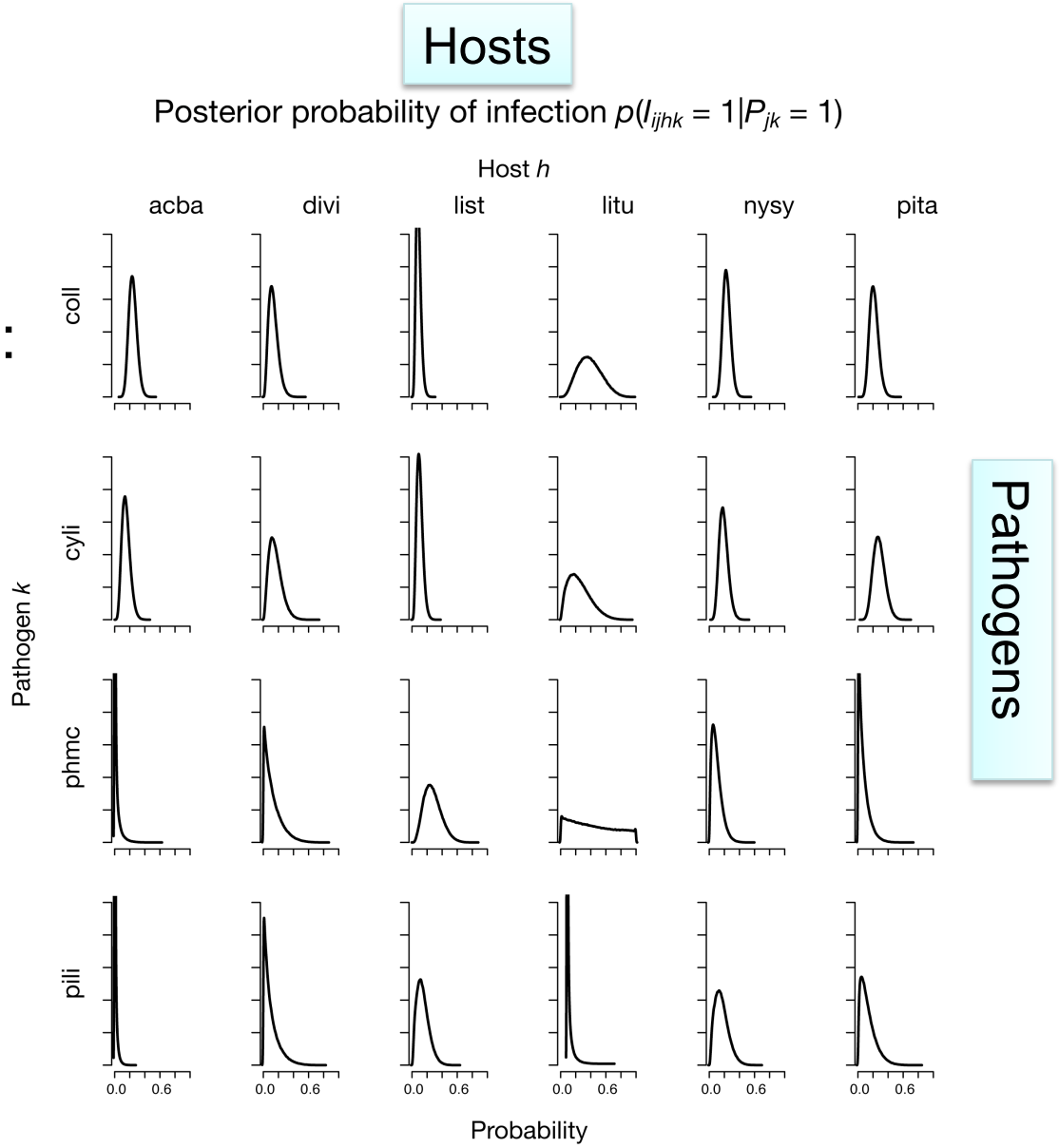


Clark and Hersh, *Bayesian Analysis* (2009)

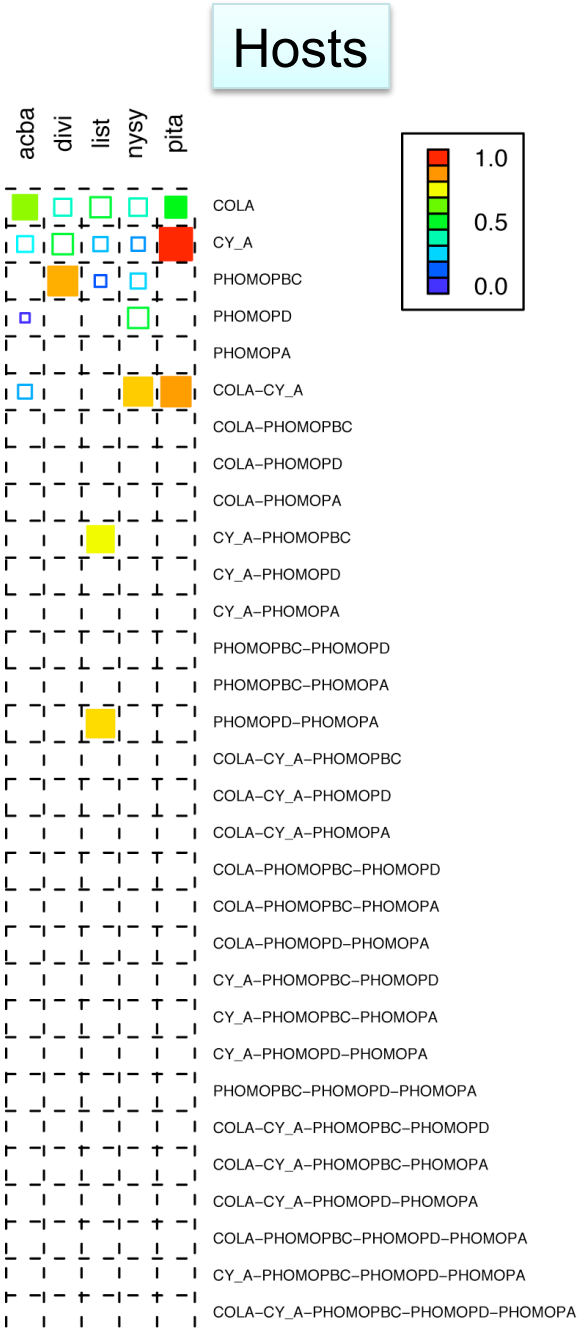
Outline

- A basic model
- An application
- The dimensionality problem
- RJMCMC
- Evaluation
- Finding the important interactions

Range of infection rates: host-pathogen combinations



Clark and Hersh, *Bayesian Analysis* (2009)



Differences in survival effect

- Posterior model probabilities for host-specific combinations of infection
- Different hosts susceptible to different **combinations of attack**

Clark and Hersh, *Bayesian Analysis* (2009)

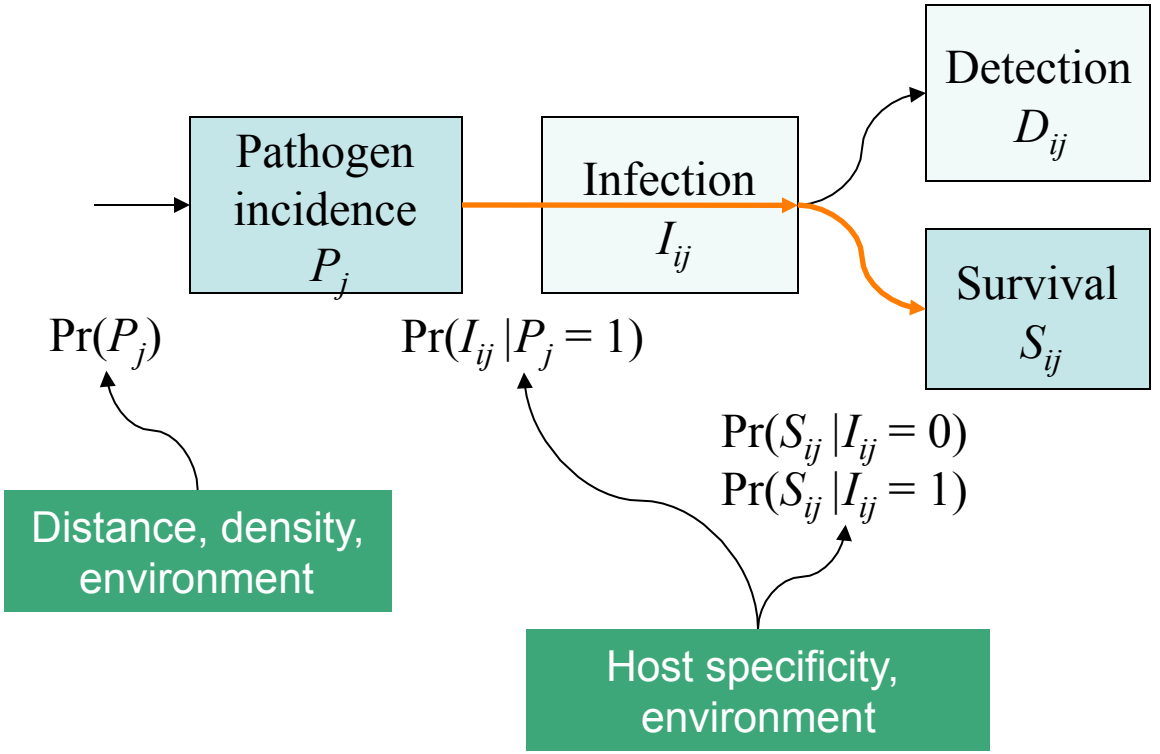
Predictive distributions

Survival given incidence marginalizes infection risk:

$$p\left(S_{hL}\middle|\mathbf{P}_L=1\right)=\sum_{I_{hL}=0,1}p\left(S_h\middle|I_{hL}\right)p\left(I_{hL}\middle|\mathbf{P}_L=1\right)$$

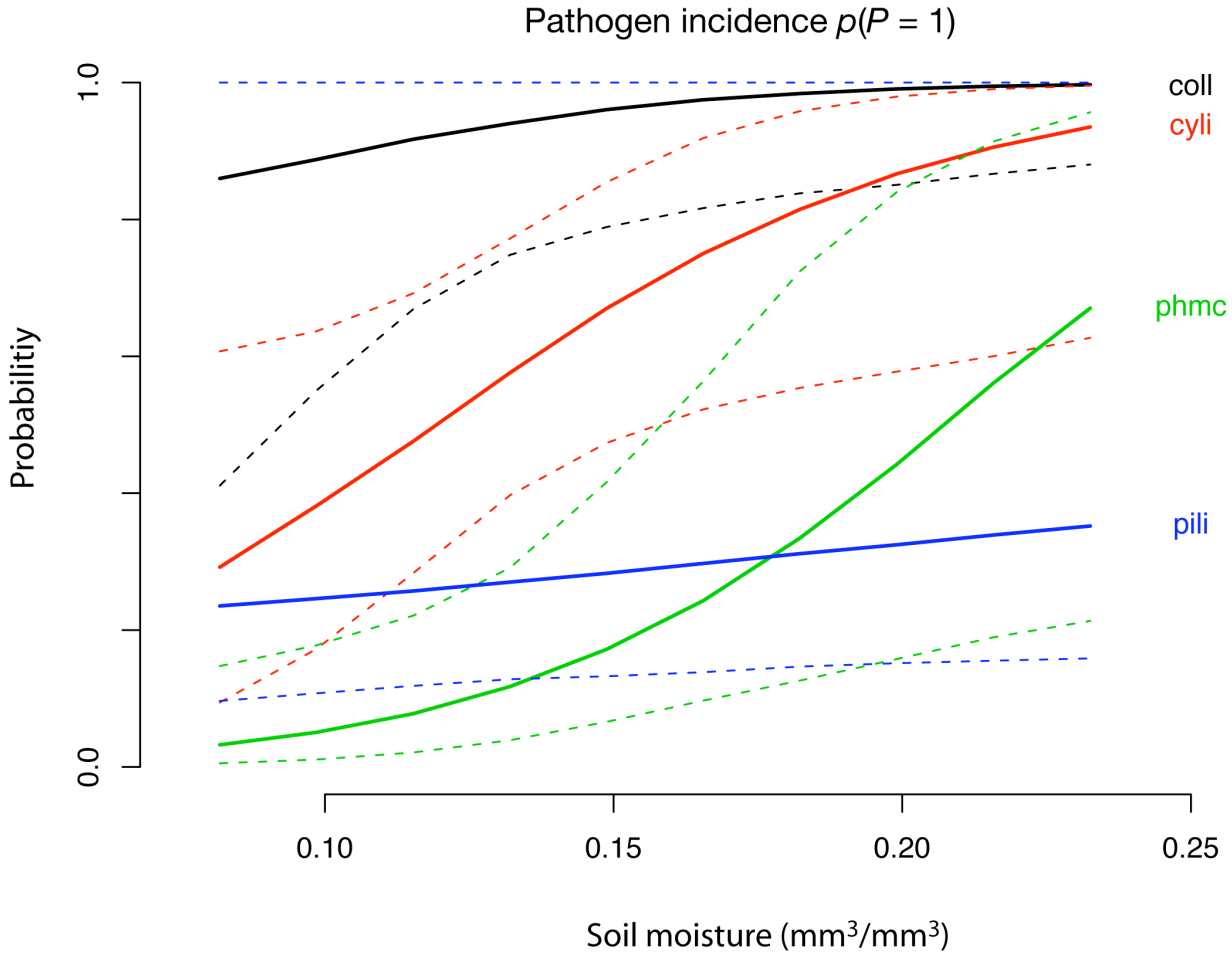
Survival given environment marginalizes incidence:

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



Environment at j affects incidence of pathogen k

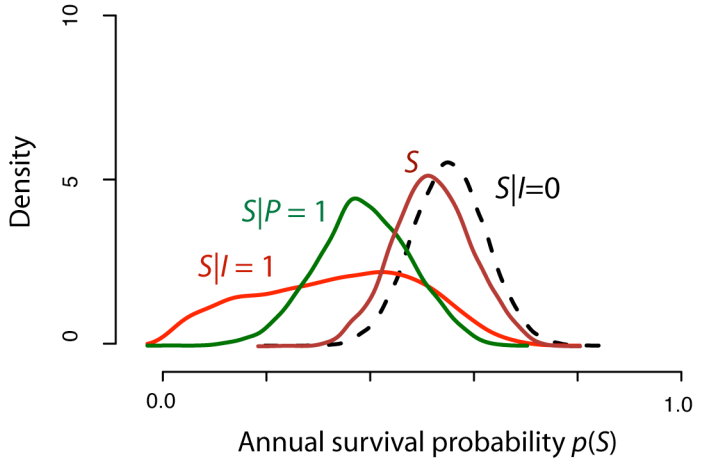
Pathogens do best where hosts do best



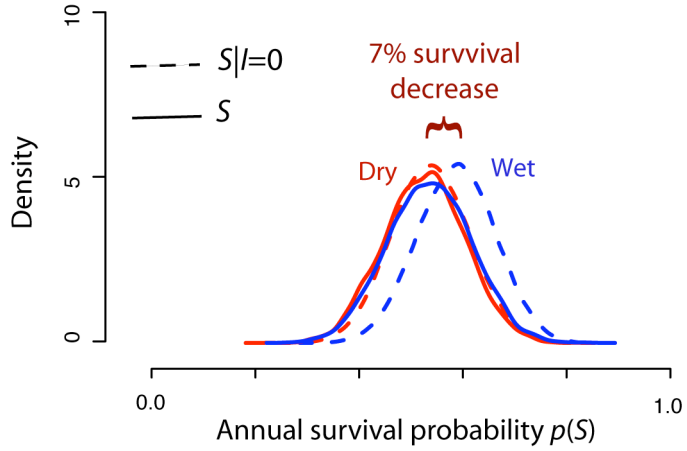
Clark and Hersh, *Bayesian Analysis* (2009)

Predictive density for annual survival rate at different scales

Effects depend on scale of information about risk factor



Interactions change the risk factor



Dry bad for host and pathogen

Clark and Hersh, Bayesian Analysis (2009)

Conclusions

- The complexity challenge
  - Reduce huge no. of potential interactions to those that matter
- Janzen Connell
  - The importance of interactions
    - Without them, no specificity
    - With them, specificity