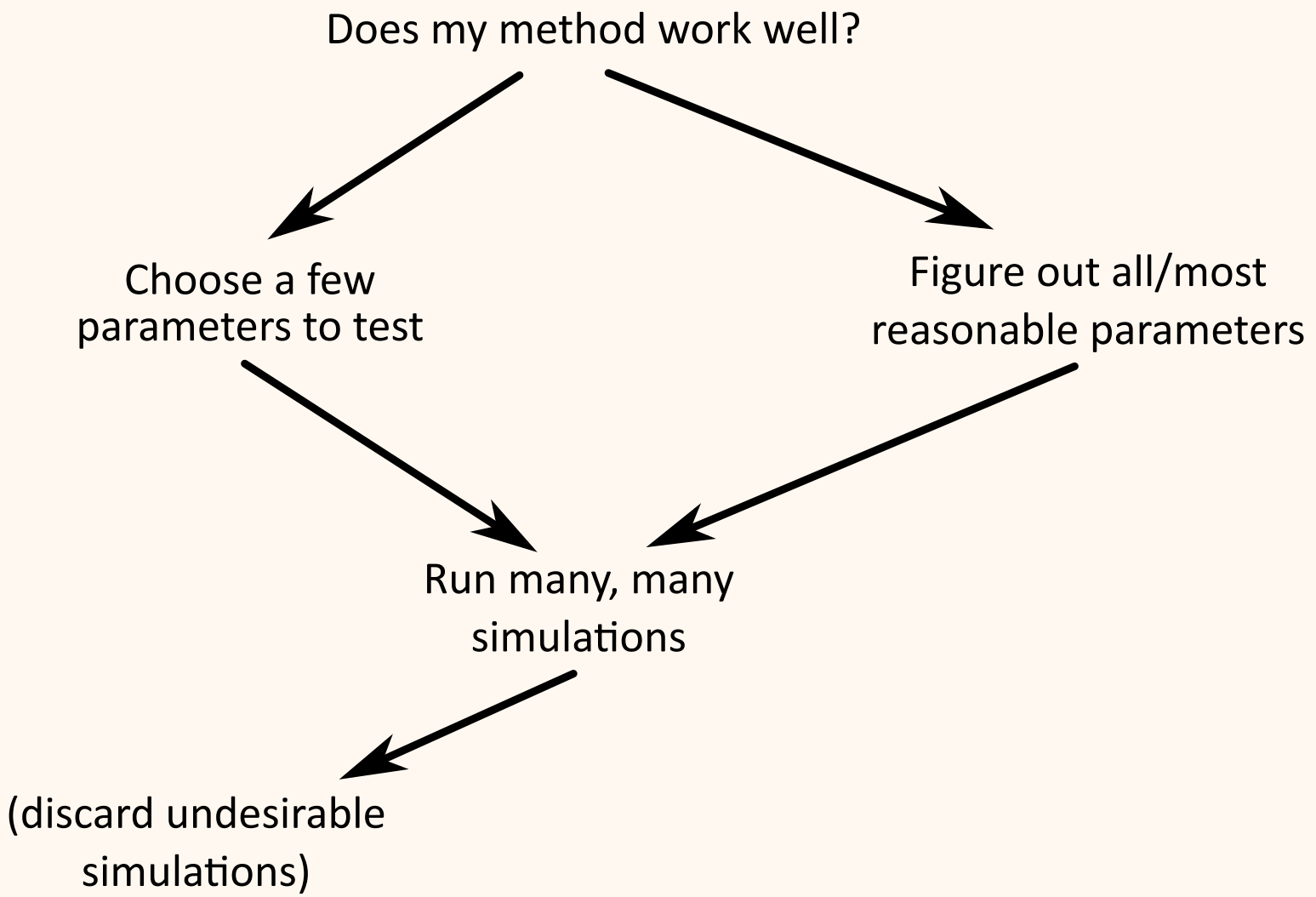
Simulation-based testing in an approximate Bayesian framework

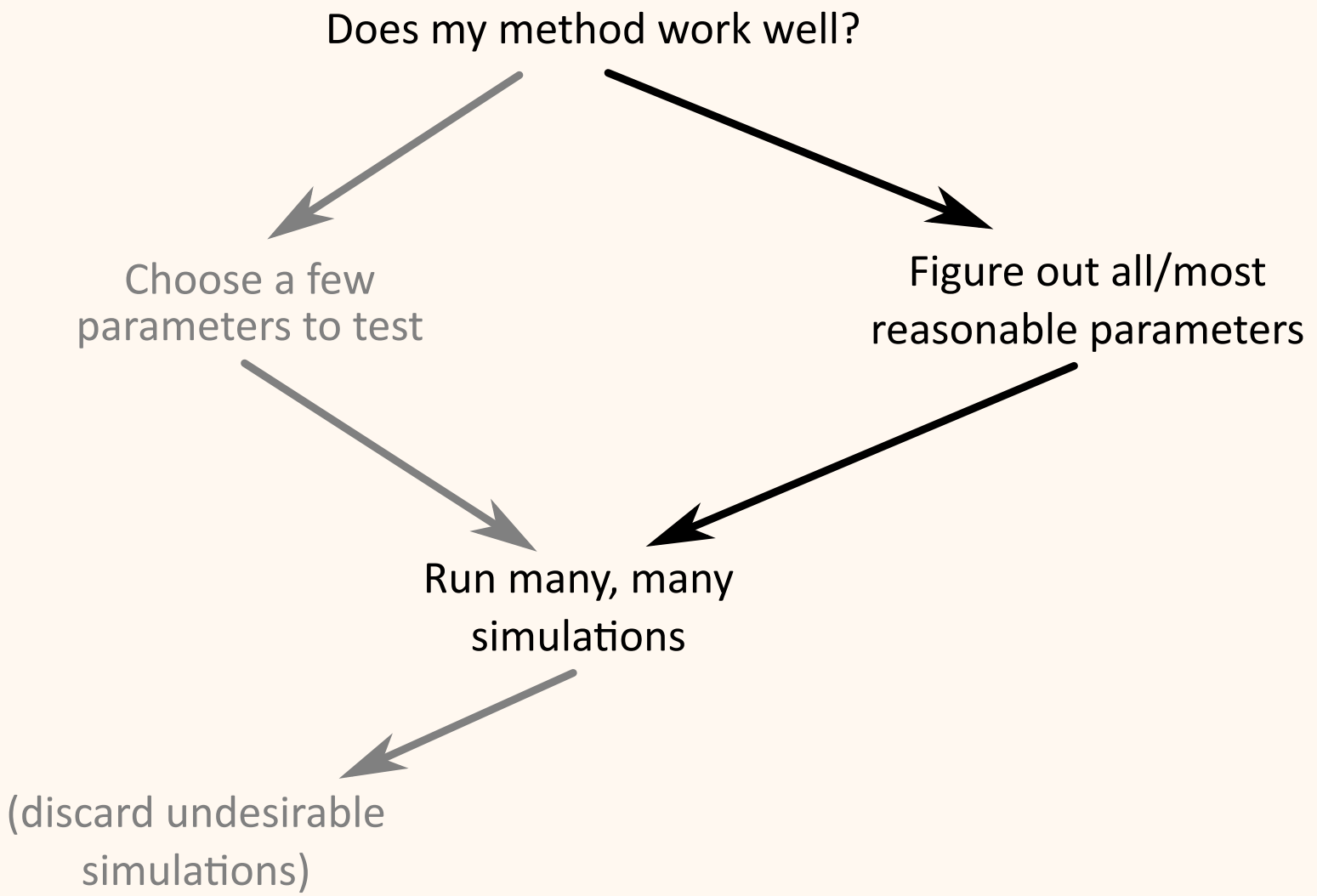
Jessica W. Leigh and David Bryant

5 November 2010

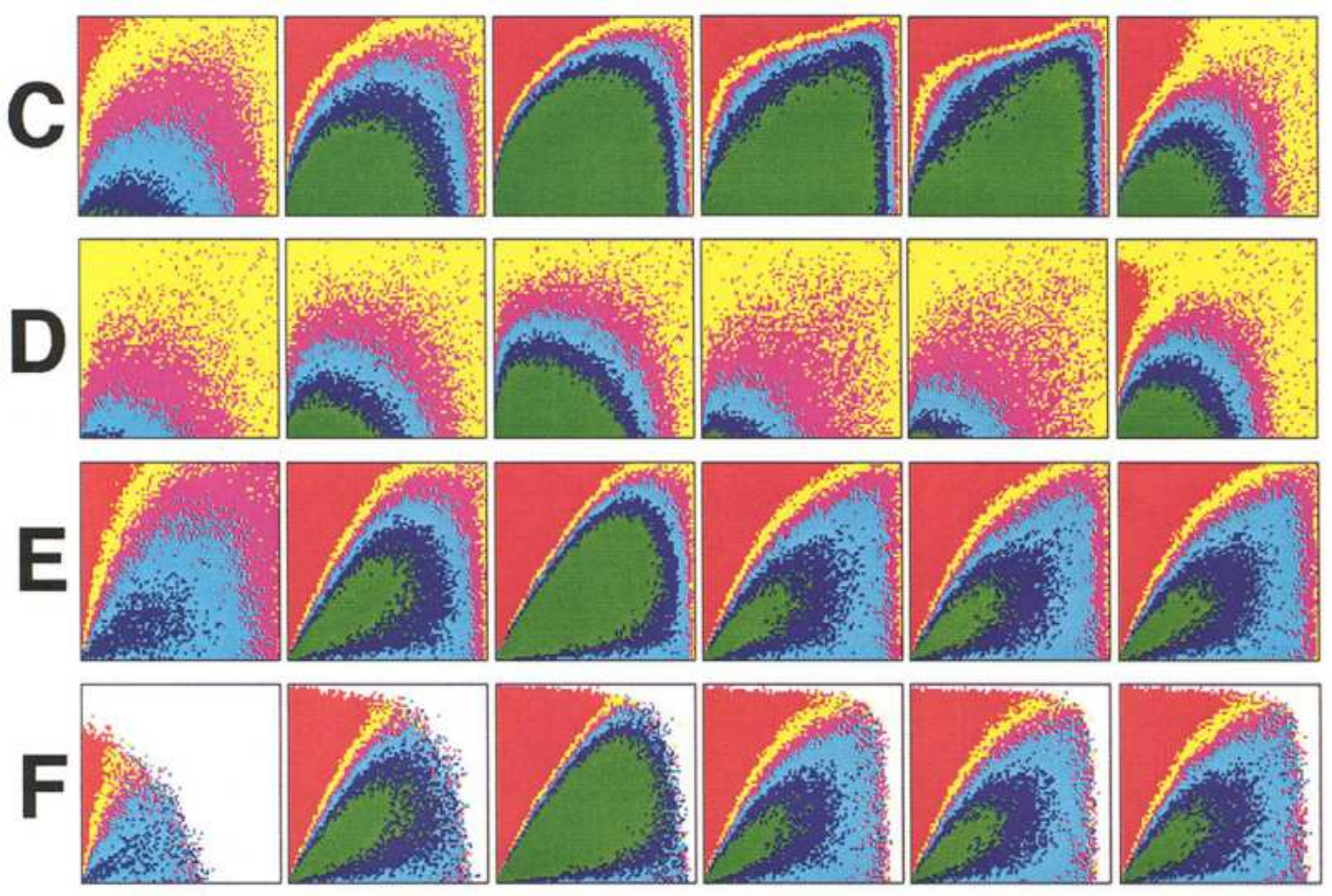
Simulation-Based Test Methodology



Simulation-Based Test Methodology

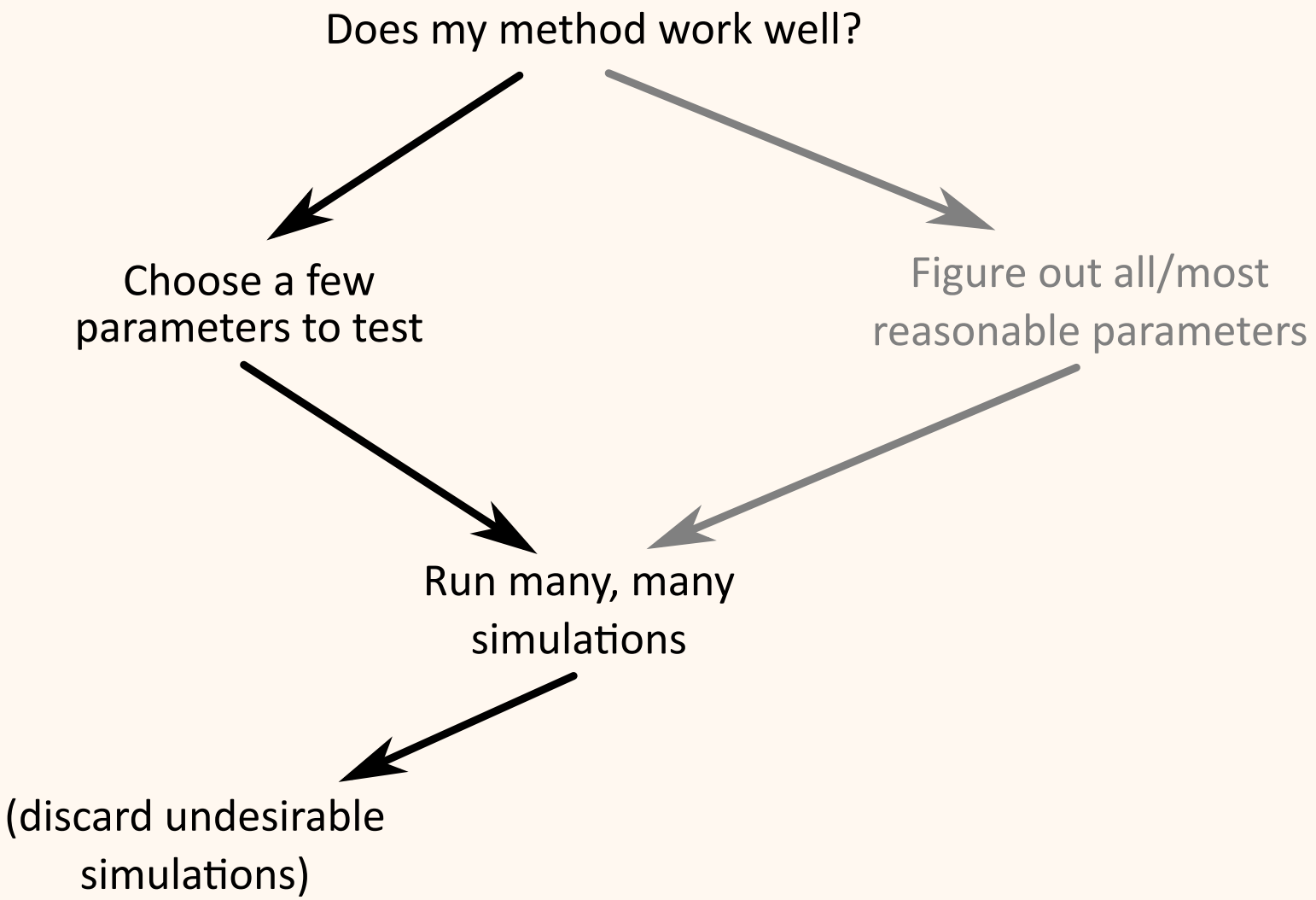


Example: Success of Phylogenetic Methods

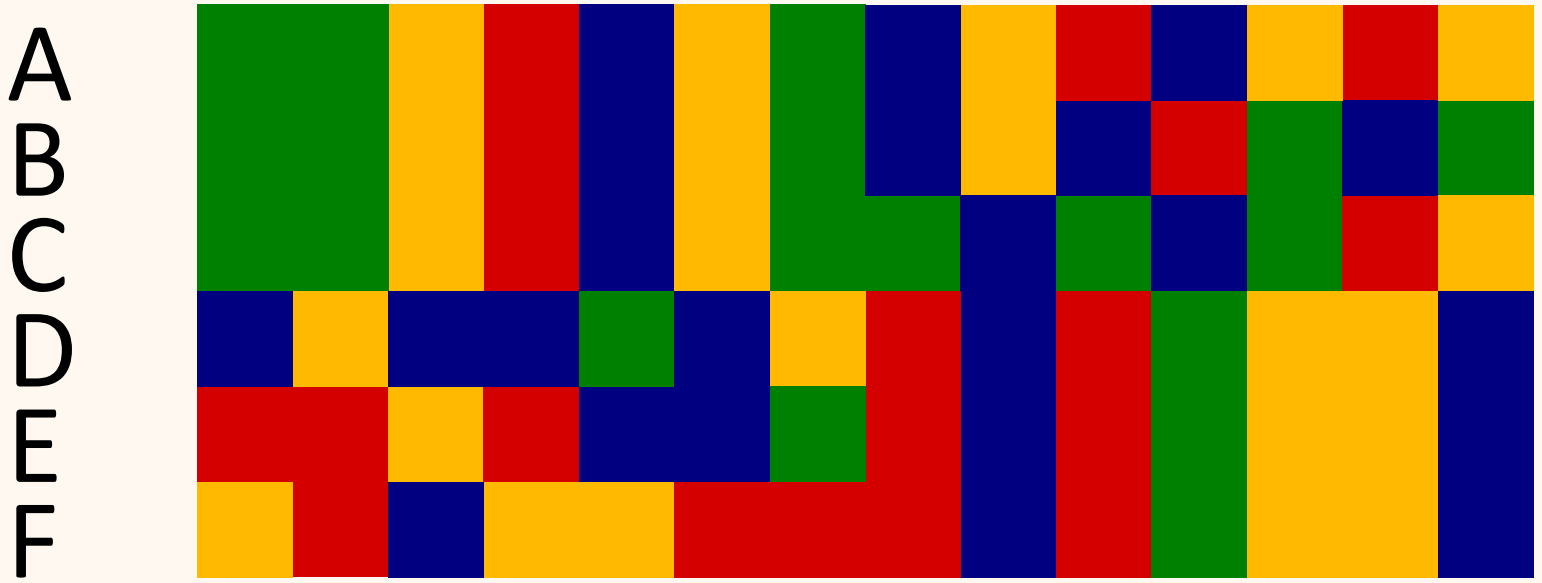


Huelsenbeck and Hillis, *Syst Biol* 1993

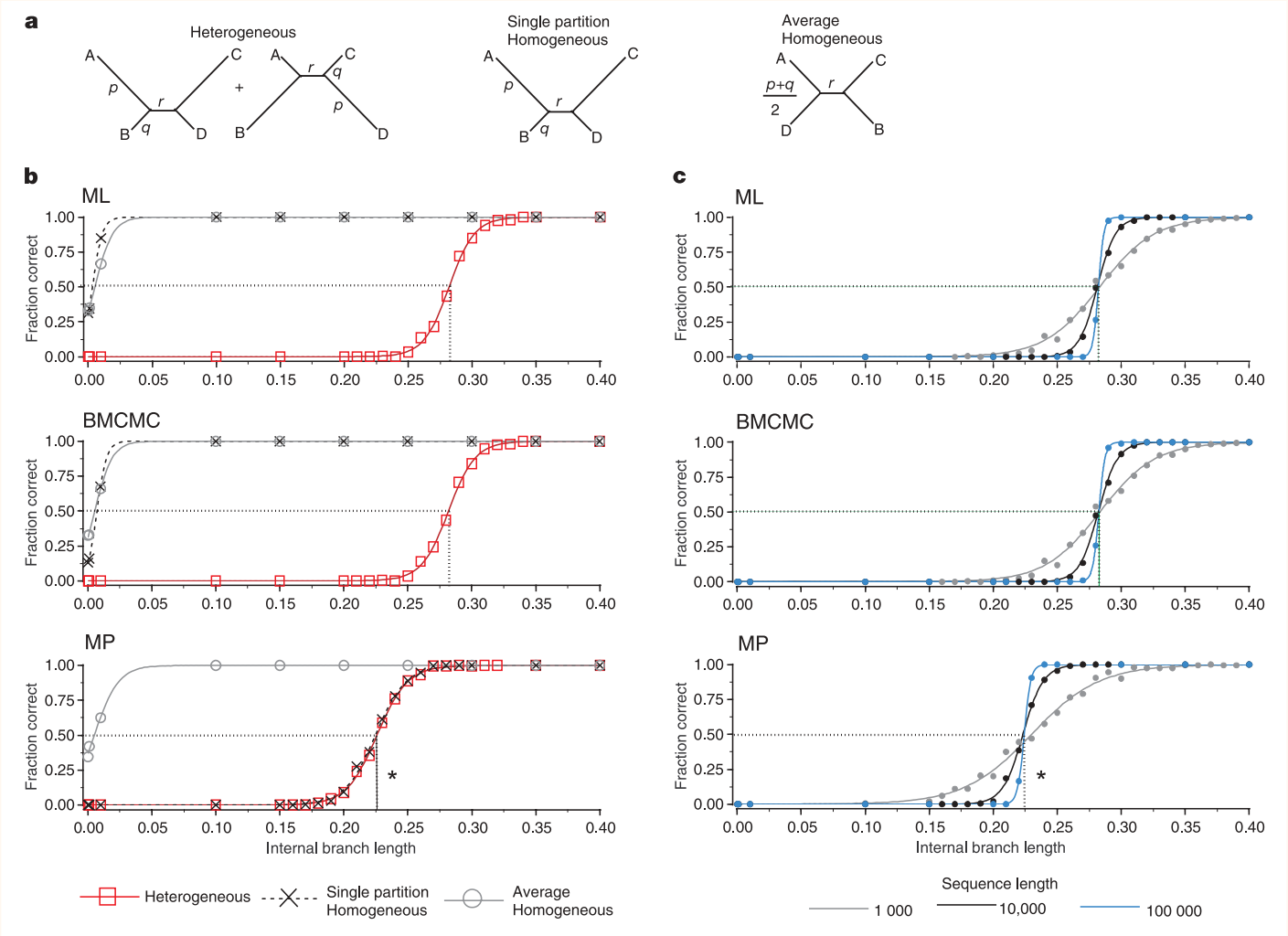
Simulation-Based Test Methodology



Example: Heterotachy

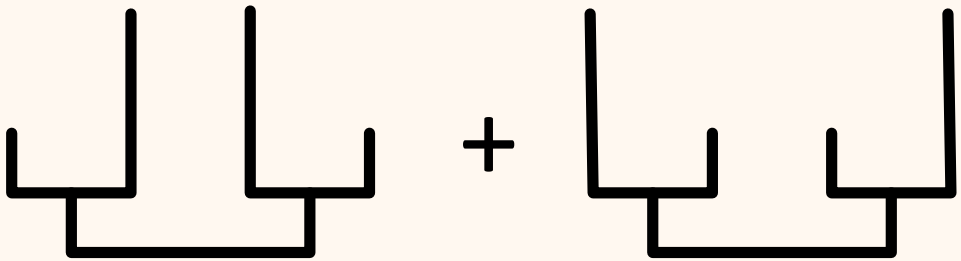


Example: Heterotachy

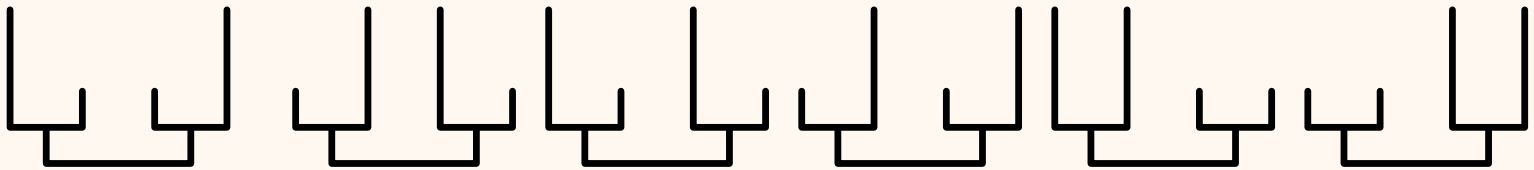


Kolaczkowski and Thornton, *Nature* 2004

Ideology Wars

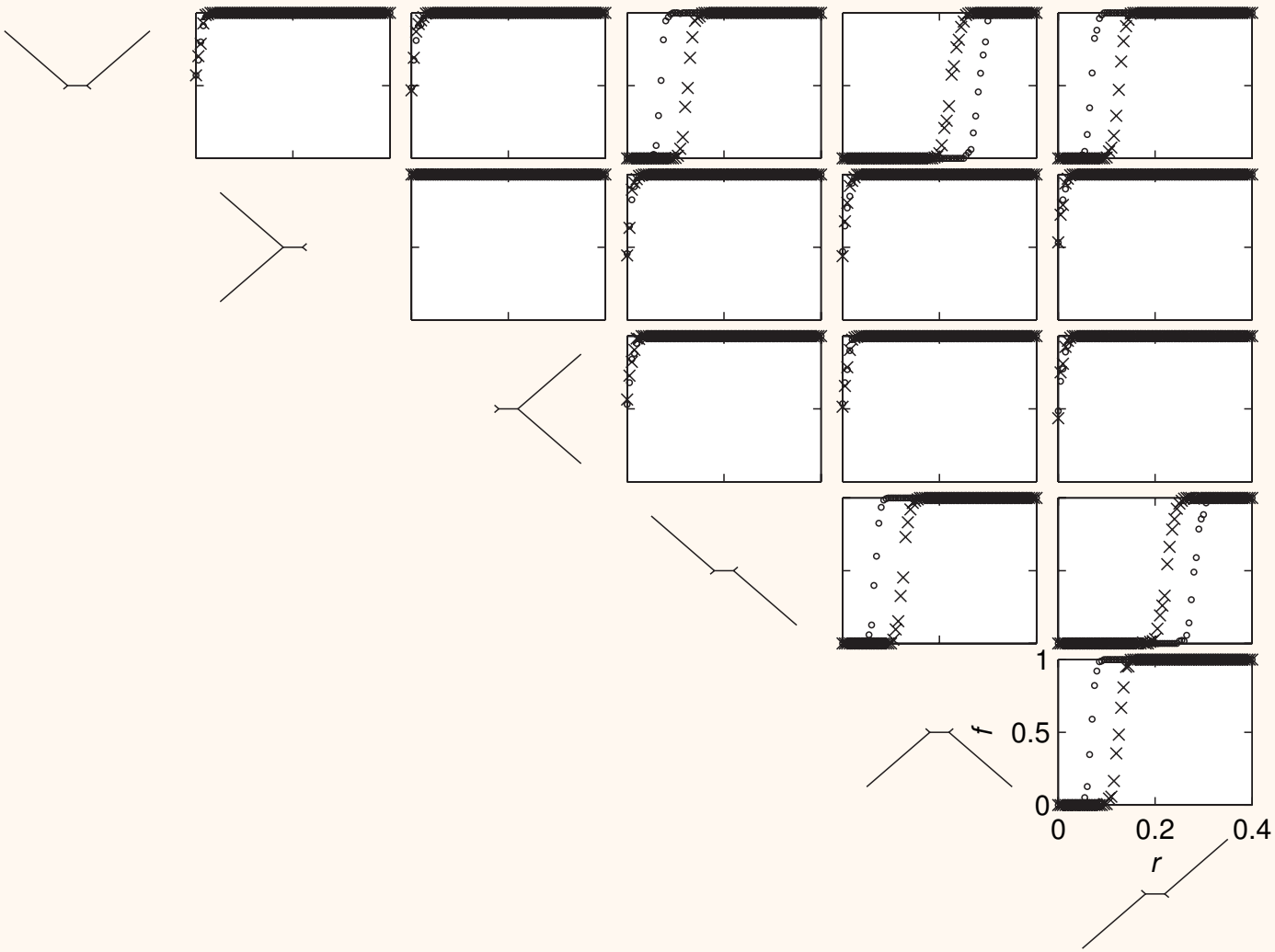


Ideology Wars



(15 combinations)

Example: Heterotachy (revisited)



Spencer, Susko, Roger, *Mol Biol Evol* 2004

Is There a Better Way?

- Simulation-based method assessment is inefficient: grid search requires too many different combinations of values for relevant parameters
- Not very rigorous if only a few select parameter values are tested
- Potentially dishonest
- **We can do better!**

Is There a Better Way?

- Simulation-based method assessment is inefficient: grid search requires too many different combinations of values for relevant parameters
- Not very rigorous if only a few select parameter values are tested
- Potentially dishonest
- **We can do better!**
- Needed: a method to explore parameters where the test does well and where it does poorly
- **MCMC can do this**

Recipe: MCMC-Based Simulation Test

- Let $\phi(X)$ denote a specific question addressing the performance of the method using simulated data X
  - Does one method outperform another?
  - Does a method produce a false positive?
- Sample from the probability distribution of parameter θ given a "true" answer to the question we asked ($P(\theta|\phi(X) = 1)$)

Markov chain Monte Carlo

- Suppose we wish to sample from some distribution $\pi(X)$
- Generate a Markov chain $X_1, X_2, \ldots, X_k, \ldots$ by repeatedly accepting or rejecting states drawn from a proposal distribution
- The chain is set up such that its stationary distribution is the distribution of interest
- Moves satisfy the detailed balance condition: $f(X_i, X_{i+1})\,\pi(X_i) = f(X_{i+1}, X_i)\,\pi(X_{i+1})$, where $f(X_i, X_{i+1})$ is the probability of moving from state X_i to X_{i+1}

MCMC: Metropolis-Hastings Algorithm

- Uses likelihoods to accept or reject moves, samples from the distribution $P(\theta|\mathcal{D})$
- Let $q(\theta_{i+1}|\theta_i)$ be the probability of proposing state θ_{i+1} given the current state θ_i and let $\pi(\theta_i) = P(\mathcal{D}|\theta_i)$
- Consider the k^{th} iteration of the chain:

$\theta_{k+1} \sim q(\cdot|\theta_k)$
$\alpha \leftarrow \min\left\{1, \frac{\pi(\theta_{k+1})}{\pi(\theta_k)} \frac{q(\theta_k|\theta_{k+1})}{q(\theta_{k+1}|\theta_k)}\right\}$
$u \sim U(0,1)$
if $u > a$ **then**
    $\theta_{k+1} \leftarrow \theta_k$
end if

MCMC: Exact Approximate Bayesian Simulation Framework

- Recall: $\phi(X)$ is a question that can be asked using data X and $P(\theta|\phi(X)=1)$ is the distribution of interest

Our algorithm

$\theta_{k+1} \sim q(\cdot|\theta_k)$
$X \leftarrow$ simulate using θ_{k+1}
if $\phi(X)=0$ **then**
    $\theta_{k+1} \leftarrow \theta_k$
end if

MCMC: Exact Approximate Bayesian Simulation Framework

- Recall: $\phi(X)$ is a question that can be asked using data X and $P(\theta|\phi(X)=1)$ is the distribution of interest
- . . . and it satisfies the detailed balance condition

Our algorithm

$\theta_{k+1} \sim q(\cdot|\theta_k)$
$X \leftarrow$ simulate using θ_{k+1}
if $\phi(X) = 0$ **then**
    $\theta_{k+1} \leftarrow \theta_k$
end if

MCMC: Exact Approximate Bayesian Simulation Framework

- Recall: $\phi(X)$ is a question that can be asked using data X and $P(\theta|\phi(X)=1)$ is the distribution of interest
- . . . and it satisfies the detailed balance condition
- An application of Approximate Bayesian Computation (Marjoram et al, PNAS 2003) that samples exactly from the distribution of interest

Our algorithm

$\theta_{k+1} \sim q(\cdot|\theta_k)$
$X \leftarrow$ simulate using θ_{k+1}
if $\phi(X) = 0$ **then**
    $\theta_{k+1} \leftarrow \theta_k$
end if

ABC

$\theta_{k+1} \sim q(\cdot|\theta_k)$
$X^* \leftarrow$ simulate using θ_{k+1}
if $\rho(X, X^*) > \varepsilon$ **then**
    $\theta_{k+1} \leftarrow \theta_k$
end if

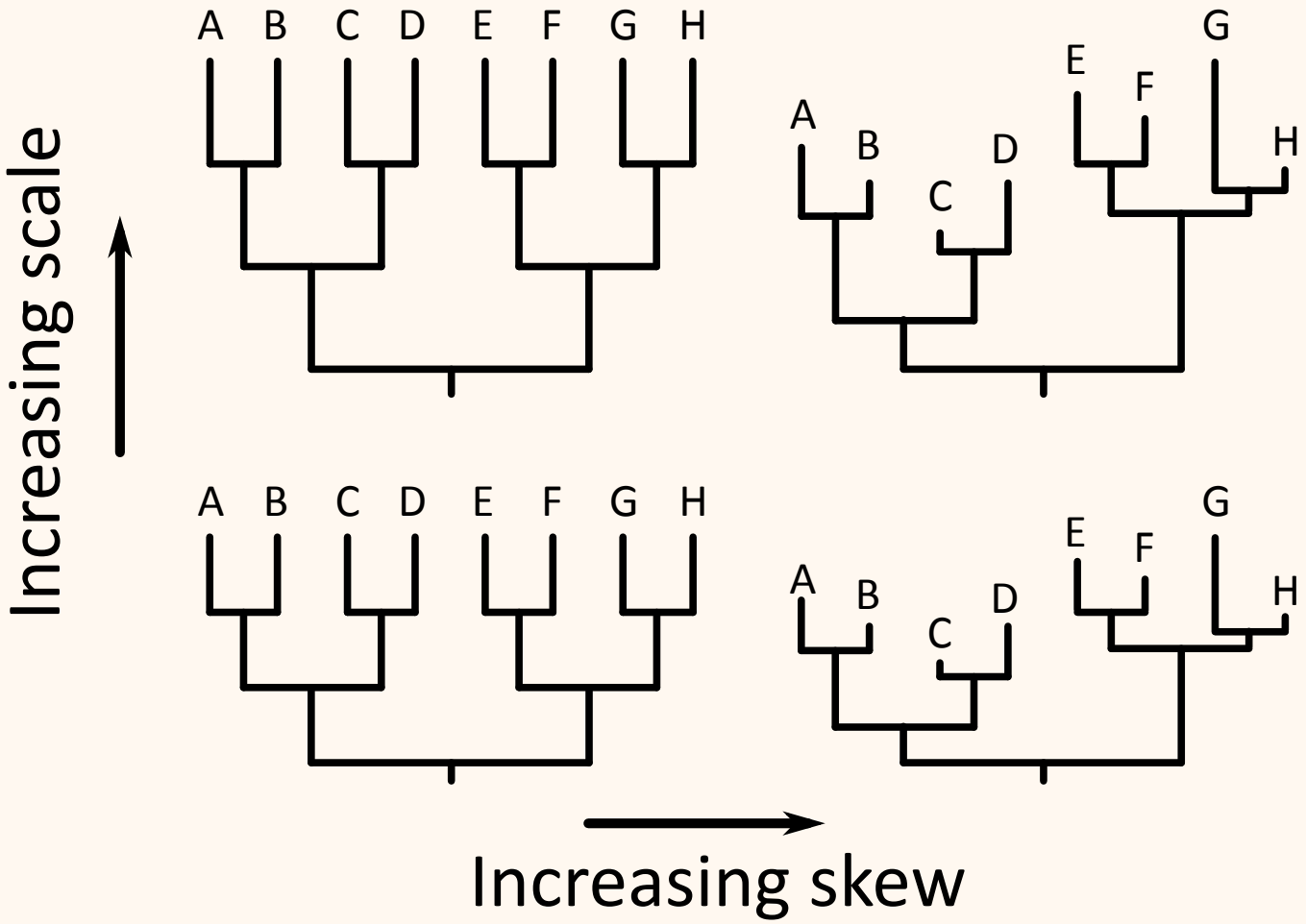
Example: UPGMA vs. NJ

- UPGMA and Neighbour-Joining are methods that produce phylogenetic trees given a matrix of pairwise distances between biological sequences representing the tips of a true tree
- Neighbour-Joining (Saitou and Nei, MBE 1987) remains a popular phylogenetic inference method and has been cited over 22,000 times (according to Google Scholar)
- Earned Masatoshi Nei an award presented by Emperor Akihito who stated that he himself had used NJ!
- UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is average linkage hierarchical clustering applied to phylogenetic data; it is generally no longer used for phylogenetic analysis because it is very sensitive to variation in evolutionary rate across lineages

Example: UPGMA vs. NJ (cont'd)

- Let T be a true phylogenetic tree, and $\hat{T}^X_{UPGMA}$ and $\hat{T}^X_{NJ}$ be trees inferred from dataset X by UPGMA and NJ, respectively
- Let $\theta = (s, \gamma)$ be a pair of parameters describing edge length scale (tree height) and skewness (non-clocklikeness)
- At each iteration, a new value for either s or γ is proposed, and a sequence alignment X is simulated from T with edge lengths described by θ
- If $\hat{T}^X_{UPGMA}$ is at least as close to T as $\hat{T}^X_{NJ}$ the new value is accepted

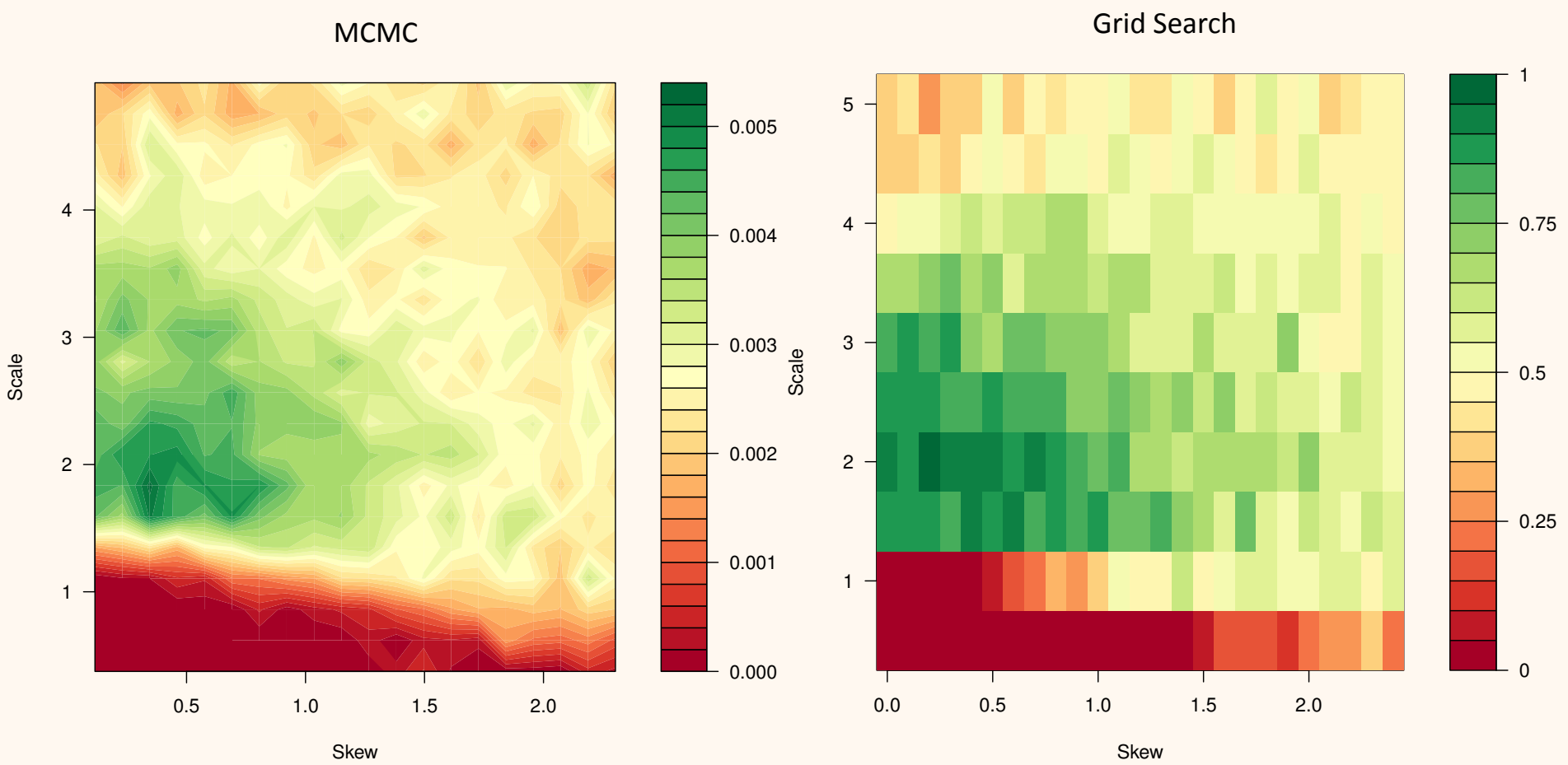
UPGMA vs. NJ: Skew and Scale Explained



Example: UPGMA vs. NJ (cont'd)

- Let T be a true phylogenetic tree, and $\hat{T}^X_{UPGMA}$ and $\hat{T}^X_{NJ}$ be trees inferred from dataset X by UPGMA and NJ, respectively
- Let $\theta = (s, \gamma)$ be a pair of parameters describing edge length scale (tree height) and skewness (non-clocklikeness)
- At each iteration, a new value for either s or γ is proposed, and a sequence alignment X is simulated from T with edge lengths described by θ
- If $\hat{T}^X_{UPGMA}$ is at least as close to T as $\hat{T}^X_{NJ}$ the new value is accepted

UPGMA vs. NJ


MCMC
Scale
Skew
0.5
1.0
1.5
2.0
1
2
3
4
0.005
0.004
0.003
0.002
0.001
0.000
Grid Search
Scale
Skew
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
1
0.75
0.5
0.25
0

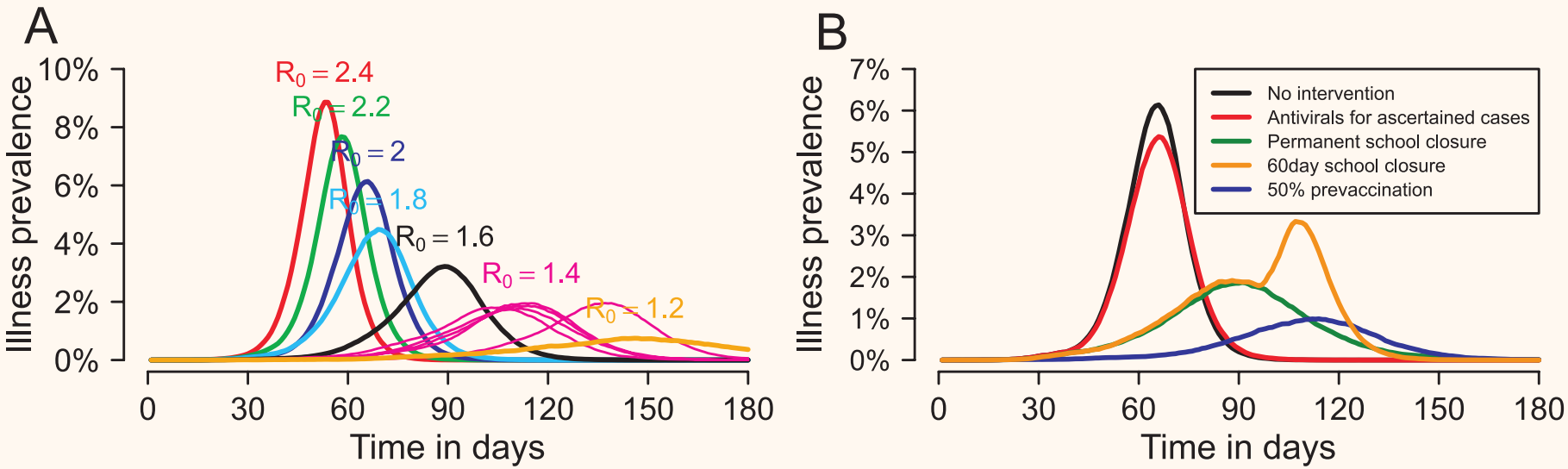
FluTE Simulator

- An influenza outbreak simulator run in half-day intervals
- Uses census data to simulate individuals, with contact probabilities based on on age and type of relationship (tuned to produce results similar to historical epidemics)

	preschool child	child	young adult	adult	older adult
Family, infectious is child	0.8	0.8	0.35	0.35	0.35
Family, infectious is adult	0.25	0.25	0.4	0.4	0.4
Household cluster, infectious is child	0.08	0.08	0.035	0.035	0.035
Household cluster, infectious is adult	0.025	0.025	0.04	0.04	0.04
Neighborhood	0.0000435	0.0001305	0.000348	0.000348	0.000696
Community	0.0000109	0.0000326	0.000087	0.000087	0.000174
Workplace			0.05	0.05	
Playgroup	0.28				
Daycare	0.12				
Elementary school		0.0348			
Middle school		0.03			
High school		0.0252			

Chao, Halloran et al, *PLoS Comp Biol* 2010

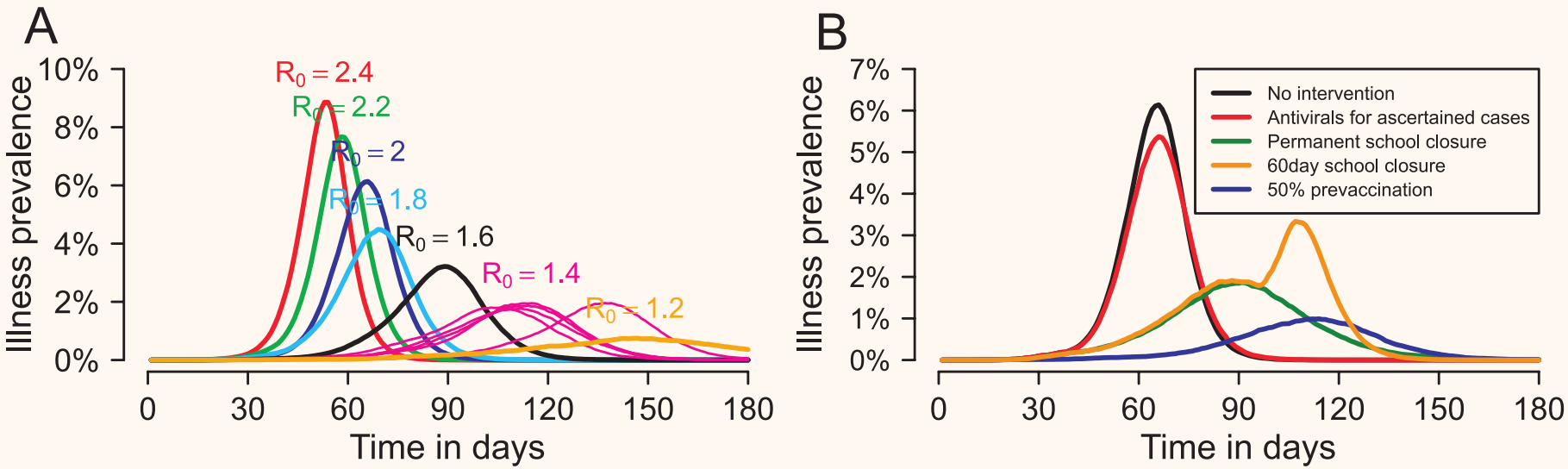
The FluTE Influenza Simulator



Chao, Halloran et al, *PLoS Comp Biol* 2010

- Various parameters, including **basic reproductive number ($\mathcal{R}_0$)**, and **prevaccinated fraction** of the population

The FluTE Influenza Simulator



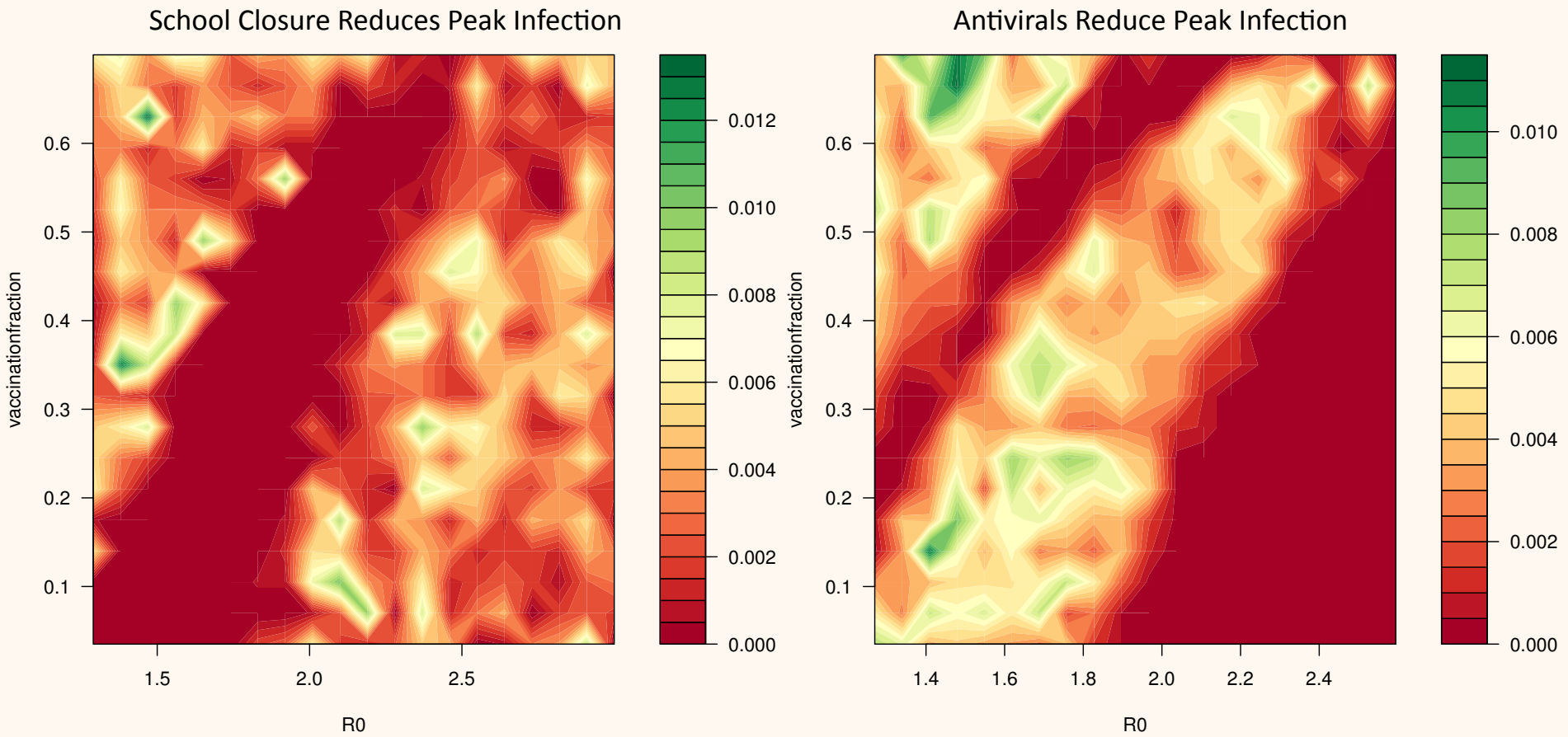
Chao, Halloran et al, *PLoS Comp Biol* 2010

- Various parameters, including **basic reproductive number ($\mathcal{R}_0$)**, and **prevaccinated fraction** of the population

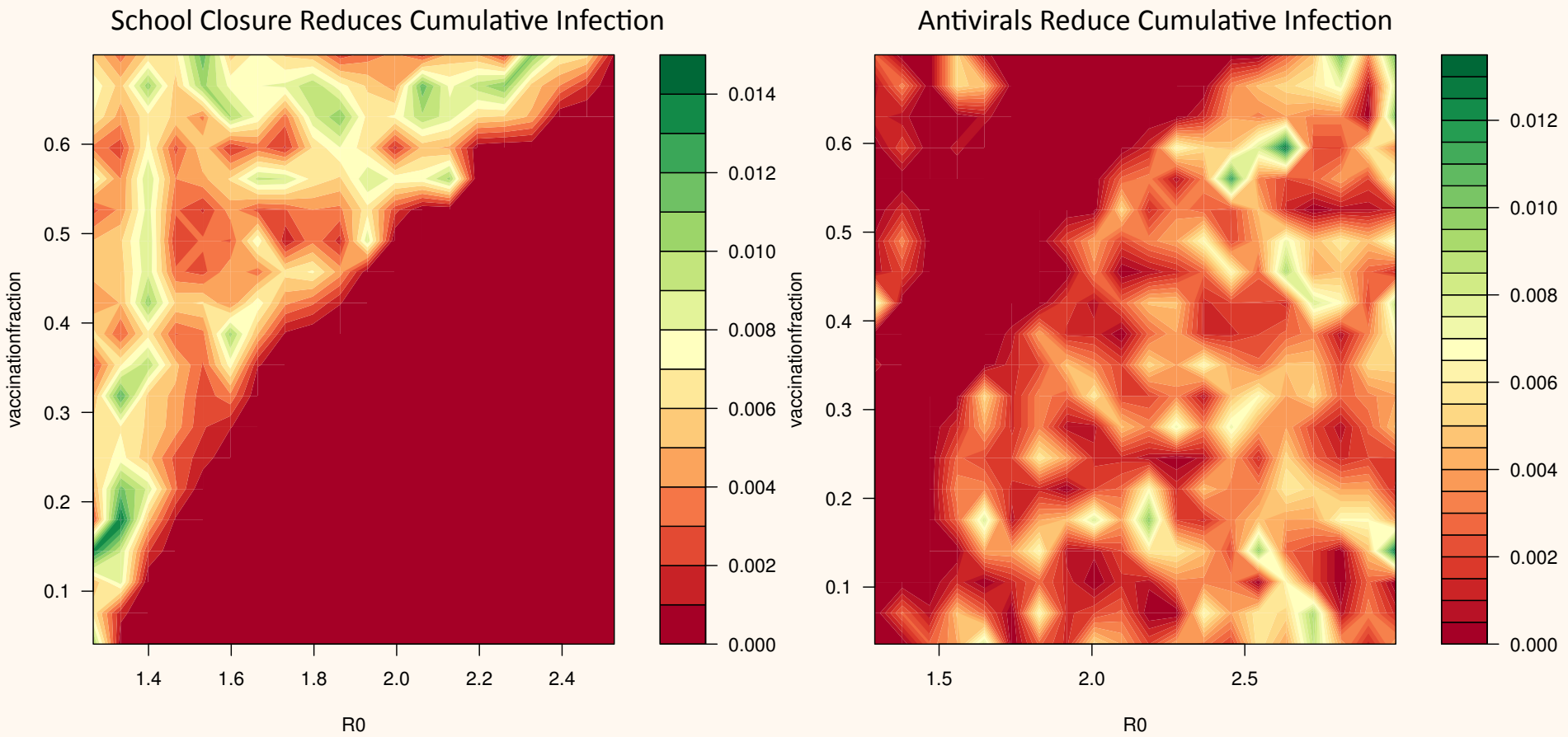
School Closure and Influenza

- School closure **might** help prevent epidemics because children have very high contact probability within a school
- In reality, if communities tend to organise social groups of children that mimic schools,
- School closure can be expensive in terms of parental absence from work
- Published simulation studies suggest that school closure might reduce the peak number of infected individuals and delay epidemics
  - Delay could be useful: often matched vaccines are unavailable at the onset of a pandemic
- A different question: given that school closure is effective, what is the distribution of $\mathcal{R}_0$ and prevaccinated fraction?

FluTE MCMC Results



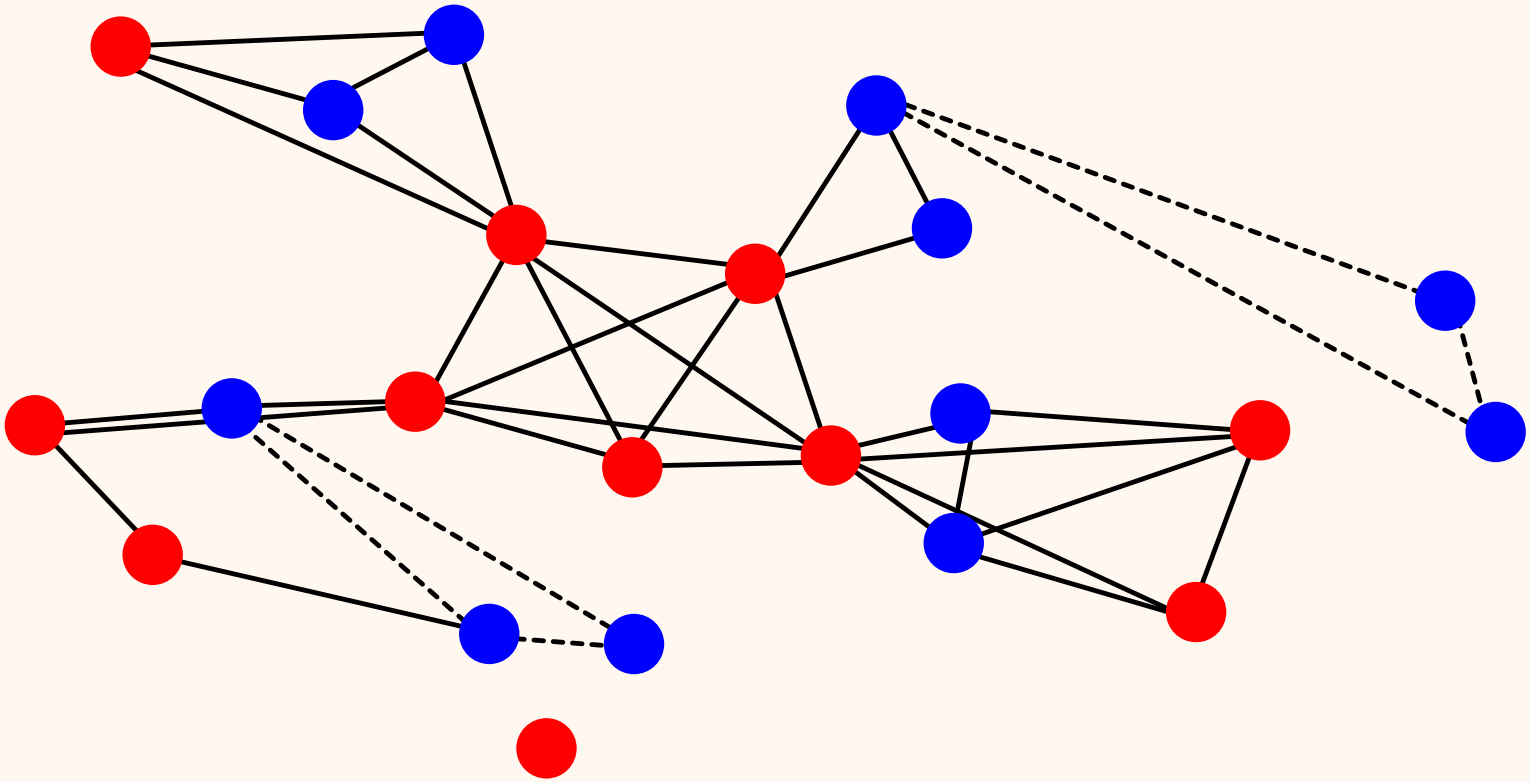
FluTE MCMC Results (Part 2)



FluTE MCMC Discussion

- For combinations of high $\mathcal{R}_0$ and low vaccination, school closure reduced the **peak** but not the **cumulative** infection level
- School closure reduced the cumulative infection level only for combinations of low $\mathcal{R}_0$ and high vaccination

FluTE MCMC Discussion



Conclusions

- MCMC can indeed be used to sample the parameter space where methods succeed (or fail)!
  - Result: probability distribution of parameter space, given success (or failure) of a method
  - (Or really the parameter distribution where you get a given answer to any true/false question that can be addressed by simulation)
- **Rigorous**, **objective**, and **efficient**
- Stupidly easy to implement (you can too!)