

Bayesian Model Comparison with MIGRATE

Peter Beerli, Scientific Computing, Florida State University

Twitter: @peterbeerli

Inference of parameters

Model of prime interest:

- ◆ Geographic structure, colonization, recurrent gene flow, past population splitting, ...

Inference of parameters

Model of prime interest:

- ◆ Geographic structure, colonization, recurrent gene flow, past population splitting, ...

But our data is usually **not a detailed historical record**, so we depend on genetic data. This is problematic because we only see differences in the sequences thus we need some more models.

Inference of parameters

Model of prime interest:

- ◆ Geographic structure, colonization, recurrent gene flow, past population splitting, ...

But our data is usually not a detailed historical record, so we depend on genetic data. This is problematic because we only see differences in the sequences thus we need some more models.

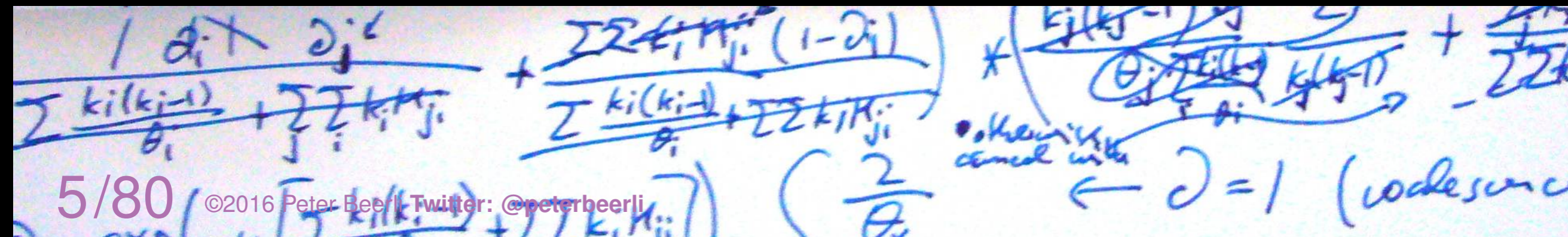
Nuisances (we are not really interested in estimating these)

- ◆ Mutation model, genealogies of individuals

The nitty gritty detail

infer the posterior probability of parameters of a population model

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} = \frac{P(\theta) \int_G P(G|\theta)P(D|G, \mu)dG}{\int_{\theta} P(\theta) \int_G P(G|\theta)P(D|G, \mu)dGd\theta}$$



The nitty gritty detail

- infer the posterior probability of parameters of a population model, usually using Markov Chain Monte Carlo
- report the posteriors and highlight some differences of the parameter, **done!?**

Handwritten mathematical derivations for a Markov Chain Monte Carlo algorithm, showing log-likelihood and log-posterior expressions with various terms and annotations.

Log-likelihood: $\log L(\theta) = \sum_i \log \left(\frac{k_i!}{k_i! k_{i-1}!} \right) + \sum_i \sum_j k_i \log \pi_{ji}$

Log-posterior: $\log \pi(\theta) = \sum_i \log \left(\frac{k_i!}{k_i! k_{i-1}!} \right) + \sum_i \sum_j k_i \log \pi_{ji} + \sum_i \log \left(\frac{1}{\theta_i} \right) + \sum_i \log \left(\frac{1}{\theta_j} \right)$

Annotations: "other terms cancel with", " $\theta = 1$ (code source)"

The nitty gritty detail

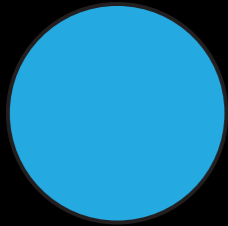
infer the posterior probability of parameters of a population model, usually using Markov Chain Monte Carlo

report the posteriors and highlight some differences of the parameter, done!

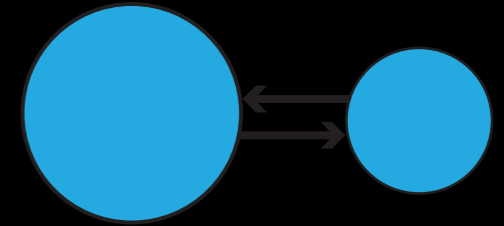
We can do better than that and statistically compare different models.

Handwritten mathematical derivations for a multinomial distribution. The top part shows the log-likelihood function $\ln L(\theta) = \sum_i k_i \ln \theta_i + \sum_{i,j} k_{ij} \ln \theta_j$. Below it, the derivative with respect to θ_i is shown as $\frac{\partial \ln L}{\partial \theta_i} = \frac{k_i}{\theta_i} - \frac{\sum_j k_{ij}}{\theta_j}$. The derivative with respect to θ_j is $\frac{\partial \ln L}{\partial \theta_j} = \frac{\sum_i k_{ij}}{\theta_j} - 1$. A note indicates that the derivative is zero at the maximum likelihood estimate, leading to $\theta_j = 1$ (code source).

Structured vs non-structured populations



A single population allows free interbreeding of all individuals, mutation accumulate approximately by $N \times \mu$ where N is the population size, and μ is the mutation rate per generation. Highly variable populations persist longer and can resist catastrophes better.

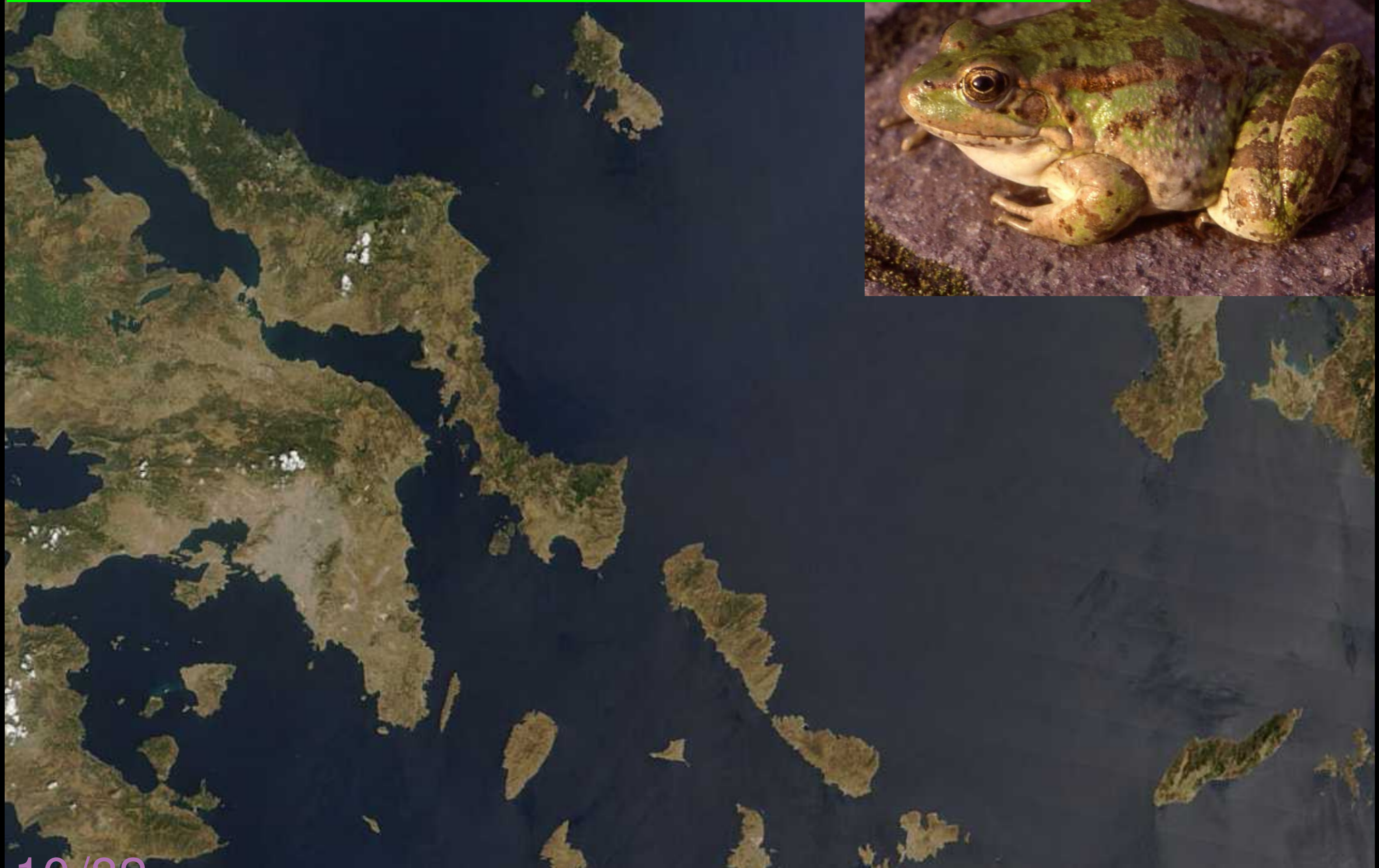


A structured population restricts interbreeding to the subpopulations. Variability in a subpopulation is gained about $N_{\text{subpop}} \times (m + \mu)$ where m is the immigration rate per generation. With very high immigration rates the structured population behaves like a single population. If N_{subpop} is small the risk of extinction is high, but such systems are often more resistant to extinction by a parasite/virus/bacteria because the transmission of these is slowed down compared to a single population.

Location versus Population



Location versus Population



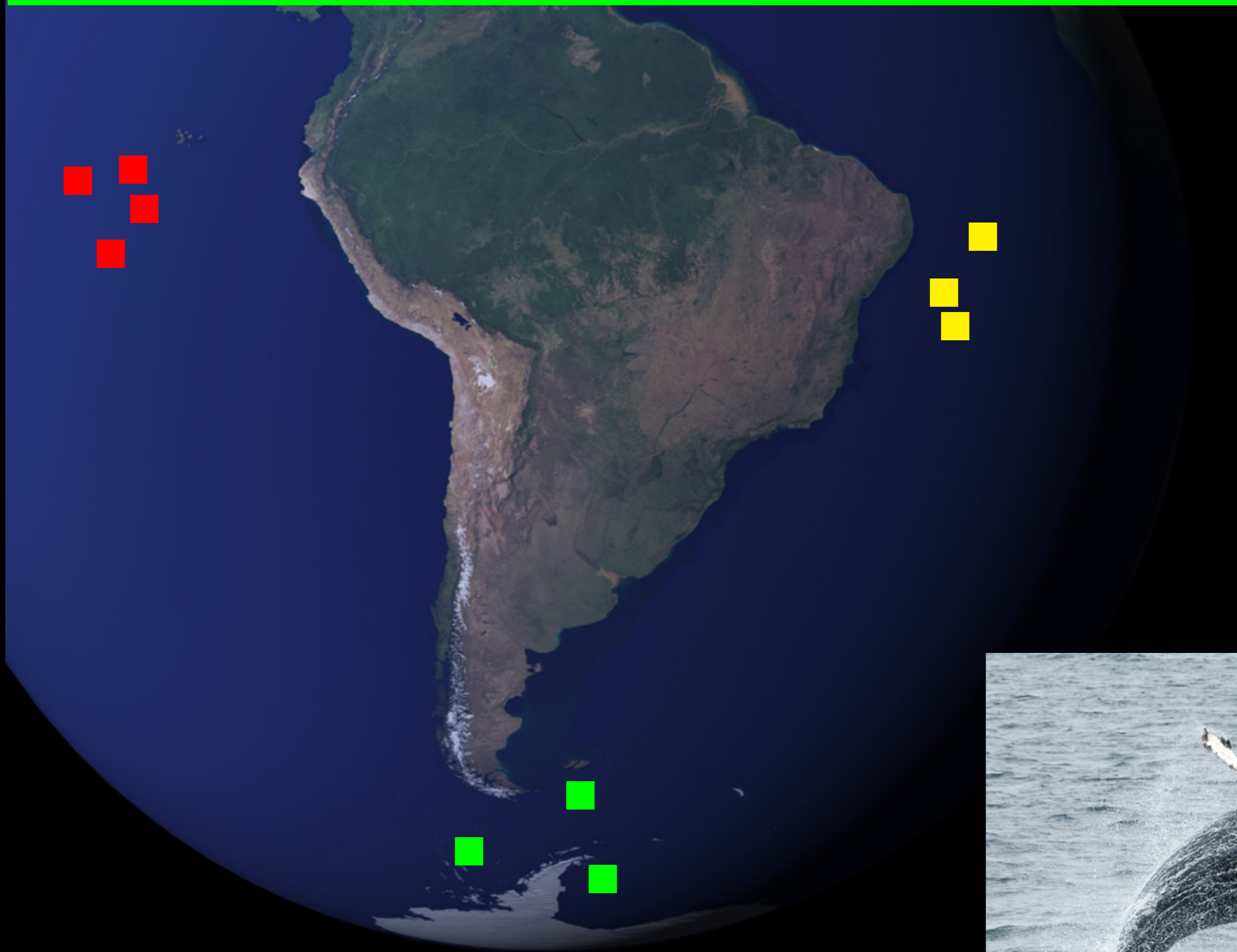
Location \approx Population



Location versus Population

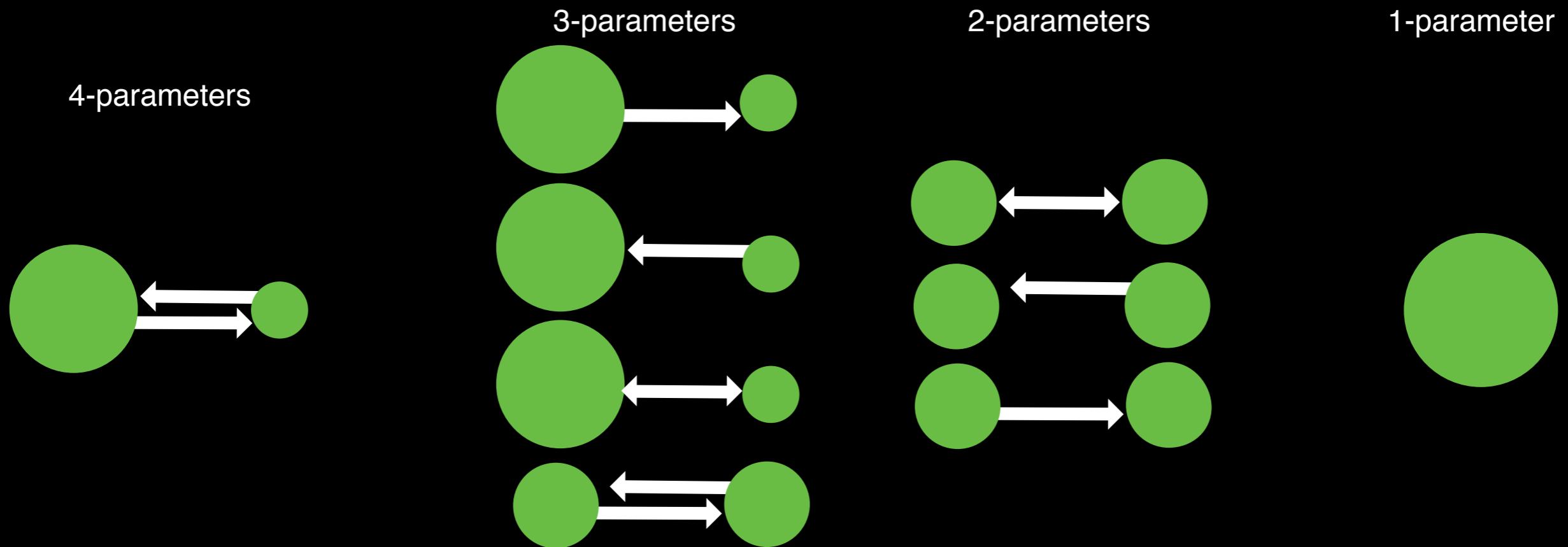


Location $\stackrel{?}{=}$ Population



Model comparison

All simple “two-population” population models that can be use in my software MIGRATE to estimate population parameters using Bayesian inference based on the population genetic framework of the coalescence theory.



Model comparison

With a criterium such as likelihood we can compare nested models. In phylogenetics, we commonly use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different from each other, or which mutation model provides the best answers among the tested models.

Kass and Raftery (1995) popularized the [Bayes Factor](#) as a Bayesian alternative to the LRT.

Bayesian Odds Ratios



*Knew that we ventured on such dangerous seas
That if we wrought out life 'twas ten to one*
William Shakespeare (Henry IV). [1597]

Bayesian Odds Ratios

Using Bayes' theorem:

$$p(M_1|X) = \frac{p(M_1)p(X|M_1)}{p(X)}$$



we can express support of one model over another as a ratio:

$$\frac{p(M_1|X)}{p(M_2|X)} = \frac{\frac{p(M_1)p(X|M_1)}{p(X)}}{\frac{p(M_2)p(X|M_2)}{p(X)}}$$

$$\text{Posterior Odds} \quad \frac{p(M_1|X)}{p(M_2|X)} = \text{Prior Odds} \quad \frac{p(M_1)}{p(M_2)} \times \text{Bayes Factor} \quad \frac{p(X|M_1)}{p(X|M_2)}$$

Bayes factor

We can use the **posterior odds ratio** or equivalently the **Bayes factors** for model comparison:

$$\text{BF} = \frac{p(X|M_1)}{p(X|M_2)} \quad \text{LBF} = 2 \ln \text{BF} = 2 \ln \left(\frac{p(X|M_1)}{p(X|M_2)} \right)$$

The magnitude of BF gives us evidence how different the models are

$$\text{LBF} = 2 \ln \text{BF} = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

IMPORTANT: recognize that $p(X|M_i)$ is equivalent to $p(X)$ in the denominator in the standard Bayesian posterior. This is the **marginal likelihood** integrated over the whole parameter space.

Marginal likelihood calculation

In MCMC application it is often complicated to calculate marginal likelihoods. Several approaches were put forward, of which the easiest, the [harmonic mean estimator](#), has turned out to be [unreliable](#) and sometimes [wrong](#).

Several other methods give accurate marginal likelihoods:

- ◆ Thermodynamic integration [MIGRATE uses this]
- ◆ Stepping-stone integration
- ◆ Inflated Density Ratio

A simple example

We want to establish a direction of geneflow between n populations.

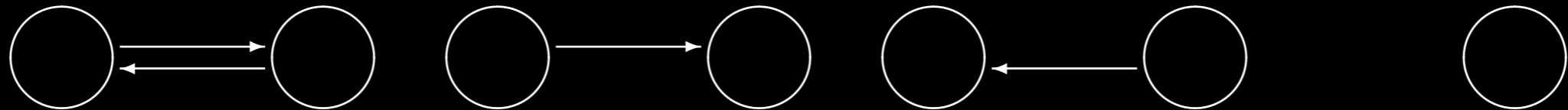
A simple example

We want to establish a direction of geneflow between 2 populations.

A simple example

We want to establish a direction of geneflow between 2 populations.

We generate 4 hypotheses



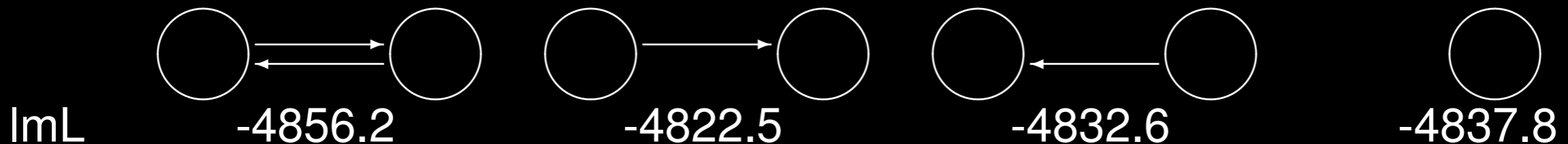
We collect data from individuals in the two populations

Analyze the data in MIGRATE

A simple example

Recipe: starting with the finished dish

Log Marginal likelihoods [ImL] of the 4 hypotheses:

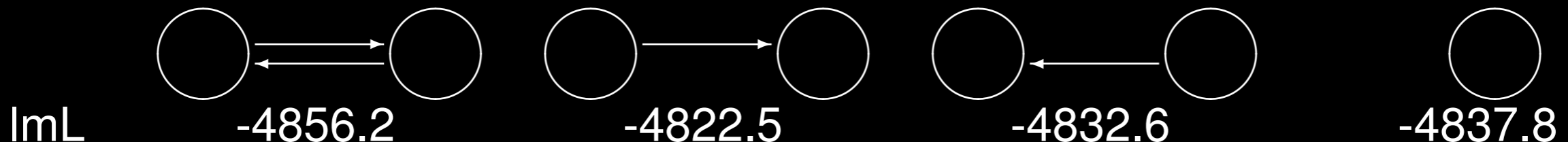


Data was simulated using the second model (2) from the left.

A simple example

Recipe: starting with the finished dish

Log Marginal likelihoods [lmL] of the 4 hypotheses:



The best model (highest lmL) is the model second from left (model 2).

We can calculate the log Bayes factor for two leftmost models as

$$LBF_{12} = 2(lmL_1 - lmL_2) = 2(-4856.2 - -4822.5) = -67.4$$

The value suggests that we should strongly prefer model 2 over model 1.

Data was simulated using the second model from the left (model 2).



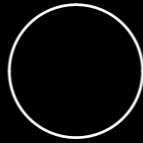
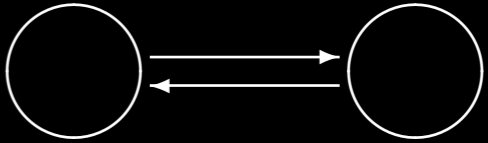
A simple example

Recipe:

1. Pick the hypothesis with largest number of parameters
2. Set priors and run parameters (use heated chains) so that you are comfortable with the result (converged, etc)
3. Record the log marginal likelihood from the output.
4. Pick next hypothesis, adjust migration model, and run and record the log marginal likelihood.
5. Repeat (4) until all log marginal likelihoods are calculated
6. Compare the log marginal likelihoods, for example order the hypothesis accordingly, or calculate the model probability

A simple example

Ordered models

				
lmL	-4822.5	-4832.6	-4837.8	-4856.2
P(model)	0.99	0.01	0.0	0.0

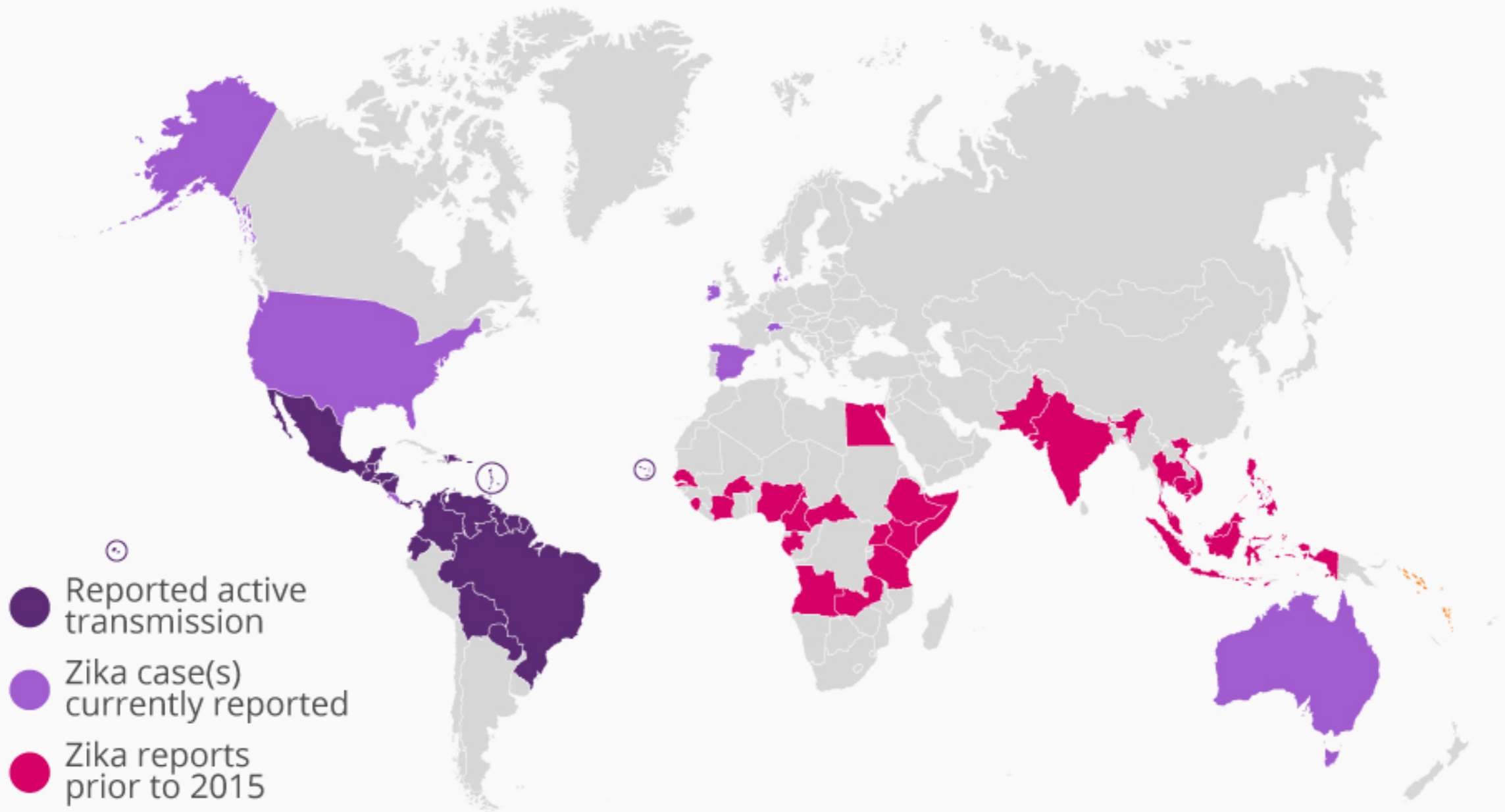
Model probability (Burnham and Anderson 2002) calculation:

$$P(M_i) = \frac{\exp(lmL_i)}{\sum_j \exp(lmL_j)} = \frac{mL_i}{\sum_j mL_j}$$

Splitting populations

The Spread Of The Zika Virus

Countries and territories with active Zika virus transmission* and reported cases



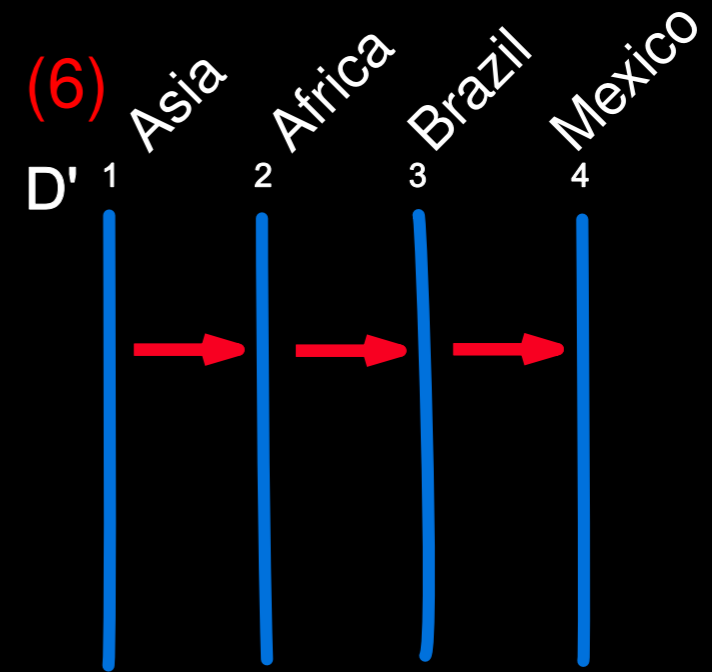
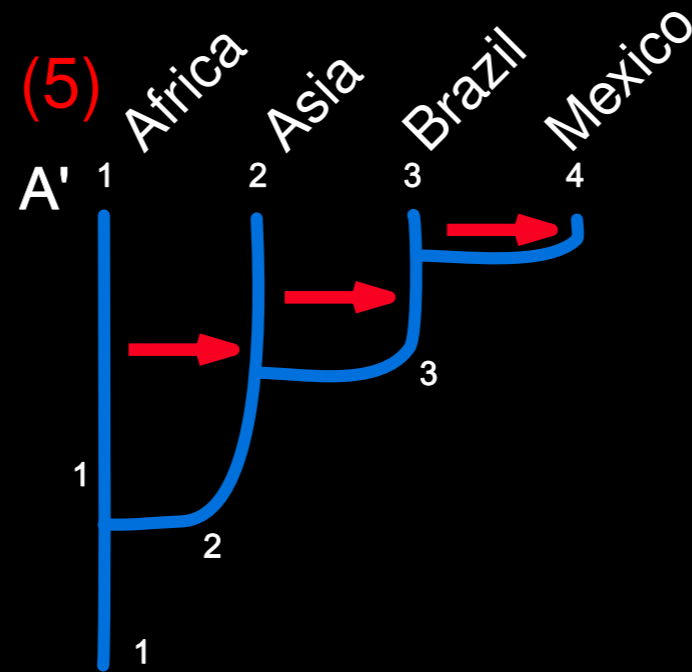
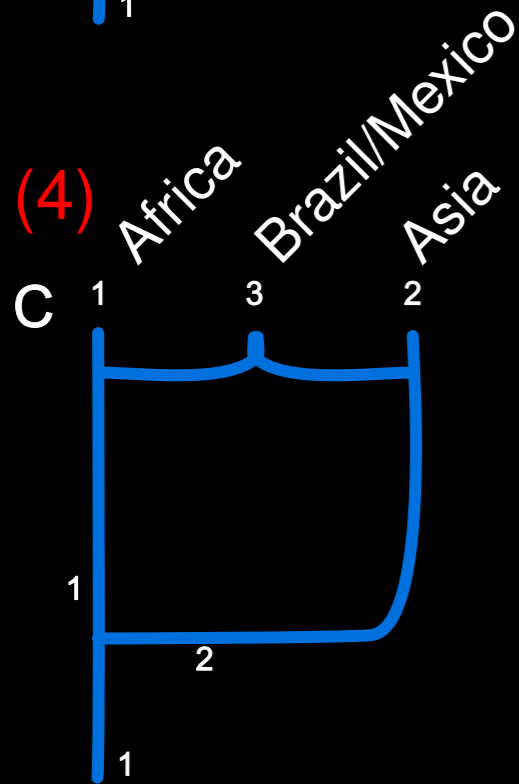
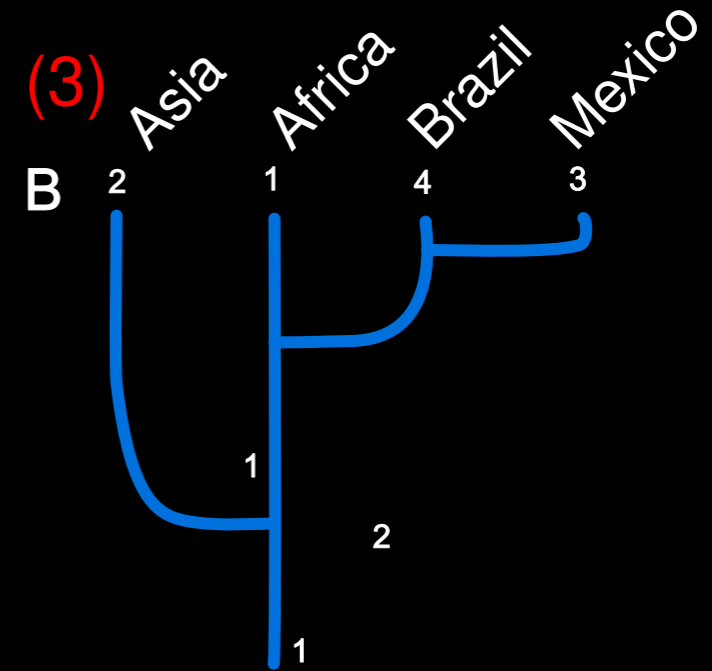
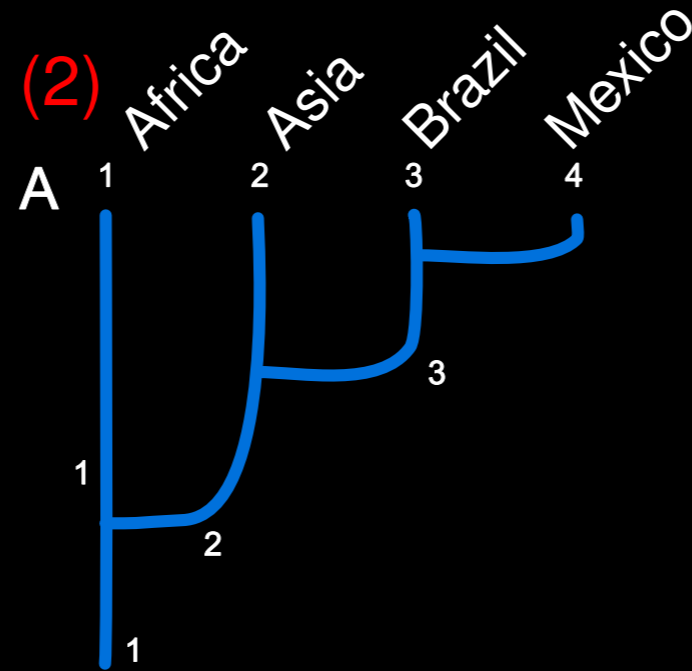
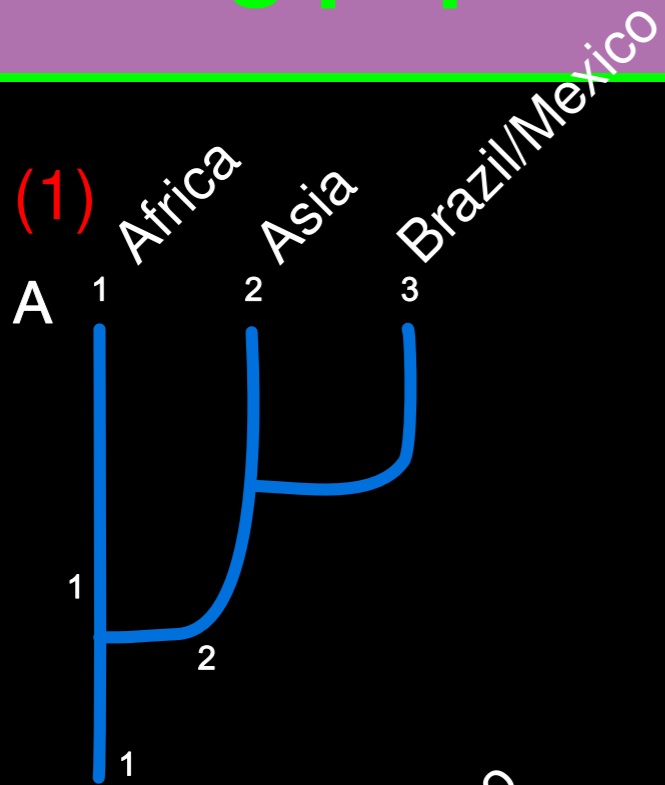
Source: Centers for Disease Control and Prevention

*As of February 2016

©2016 Peter Beerli Twitter: @peterbeerli

Splitting populations

Best model order: Zika

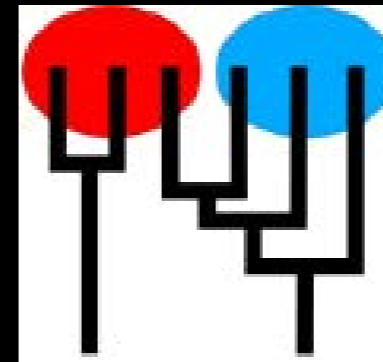
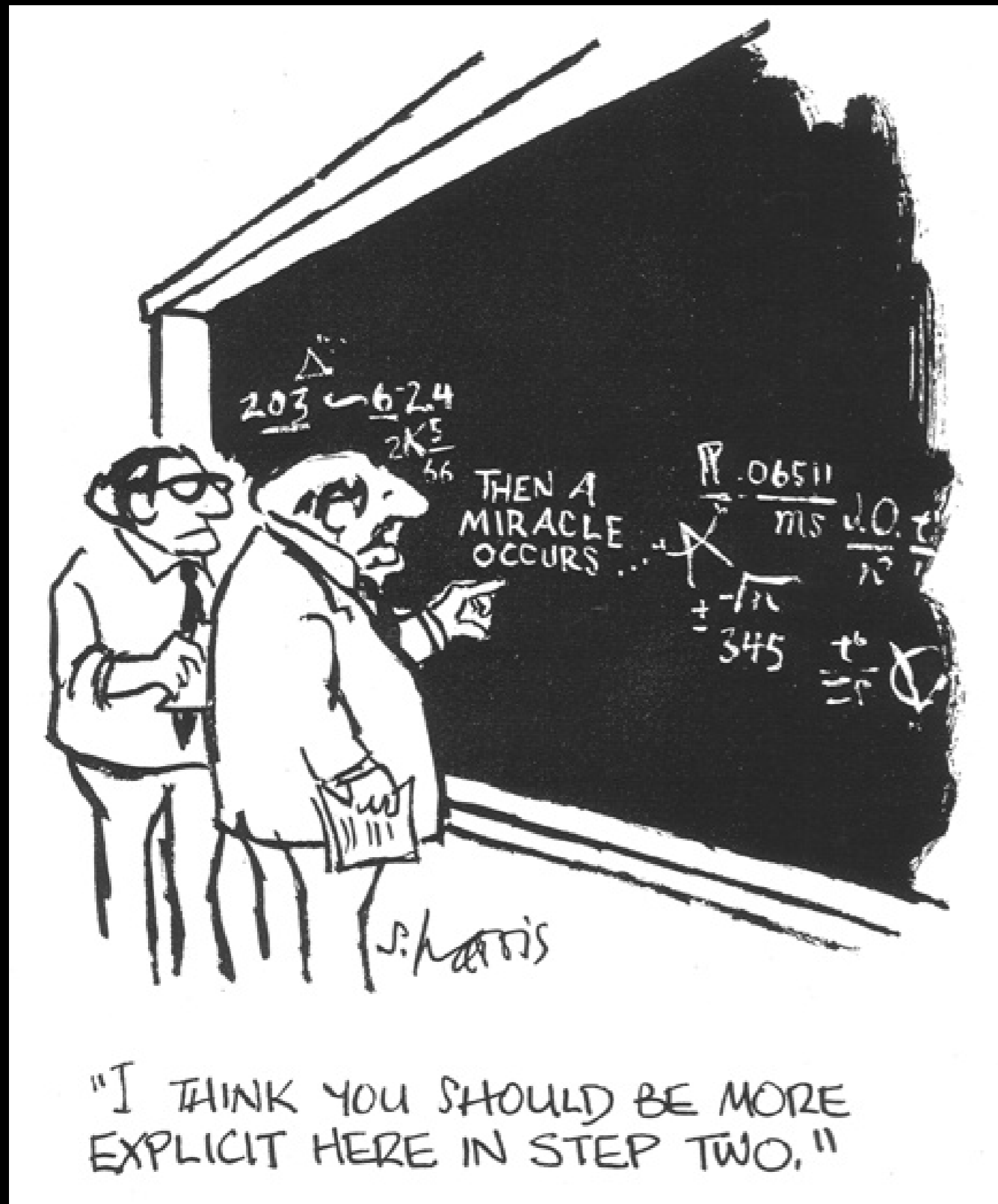


Summary

- ◆ Bayesian model selection using marginal likelihoods allows comparison of non-nested models.
- ◆ Complex biogeographic models can be compared easily.
- ◆ Data partitioning models can be compared and partition model specification affect the magnitude of parameters such as effective population size size.
- ◆ MIGRATE can run in parallel, therefore we can analyze large numbers of loci ($\gg 100$) in decent time and also compare models.

Questions?

Twitter: @peterbeerli



MIGRATE website:
<http://popgen.sc.fsu.edu>

