ETH zürich

Taming the BEAST

Bayesian evolutionary analysis by sampling trees

ACACACCTACAGACTTACAGACCC
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT
ACAGACTTTCAGACTTTCAGACCC
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT

Computational Evolution

Louis du Plessis and Carsten Magnus

D-BSSE
Department of Biosystems
Science and Engineering

ETH zürich

Taming the BEAST

Bayesian evolutionary analysis by sampling trees

ACACACCTACAGACTTACAGACCC
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT
ACAGACTTTCAGACTTTCAGACCC
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT

1. The problem

Louis du Plessis and Carsten Magnus

Computational Evolution

D-BSSE
Department of Biosystems Science and Engineering

ETH zürich

Taming the BEAST

Bayesian evolutionary analysis by sampling trees

ACACACCTACAGACTTACAGACCC
TCACACCTACACACCCACAGACTT
TCAGACTTTCAGACTTTCAGACCC
ACAGACTTTCAGACTTTCAGACCC
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT

1. The problem
2. What goes into a BEAST model?

Louis du Plessis and Carsten Magnus

Computational Evolution

D-BSSE
Department of Biosystems Science and Engineering

ETH zürich

Taming the BEAST

Bayesian evolutionary analysis by sampling trees

1. The problem
2. What goes into a BEAST model?
3. MCMC inference

Computational Evolution

Louis du Plessis and Carsten Magnus

D-BSSE
Department of Biosystems
Science and Engineering

ETH zürich

Taming the BEAST

Bayesian evolutionary analysis by sampling trees

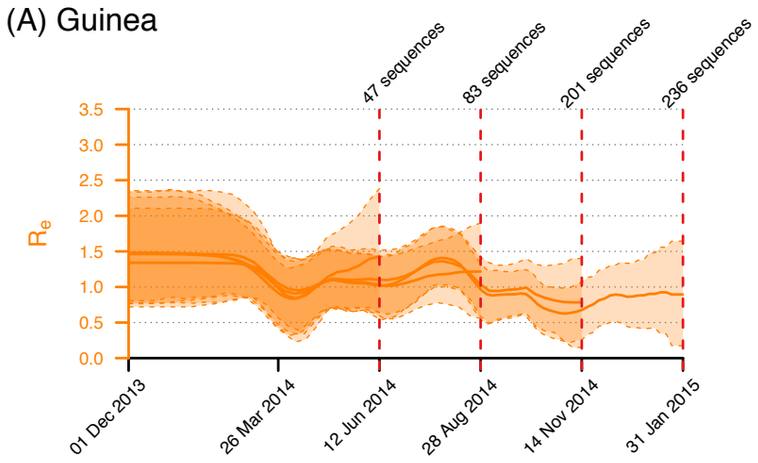
1. The problem
2. What goes into a BEAST model?
3. MCMC inference
4. Bayesian evolutionary analysis by sampling trees

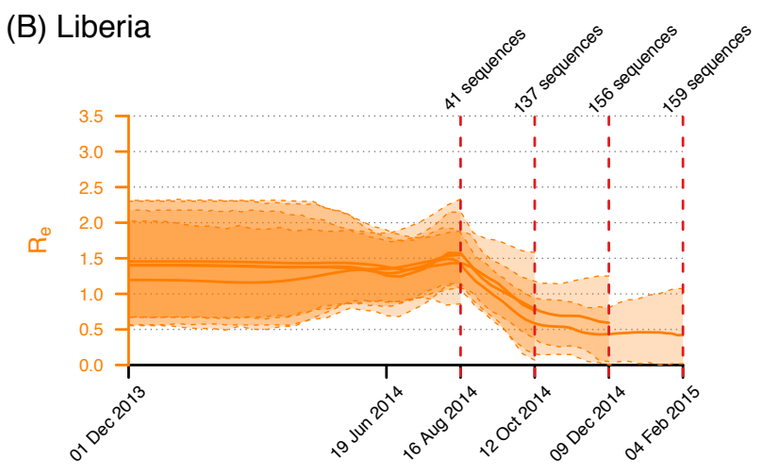
Computational Evolution

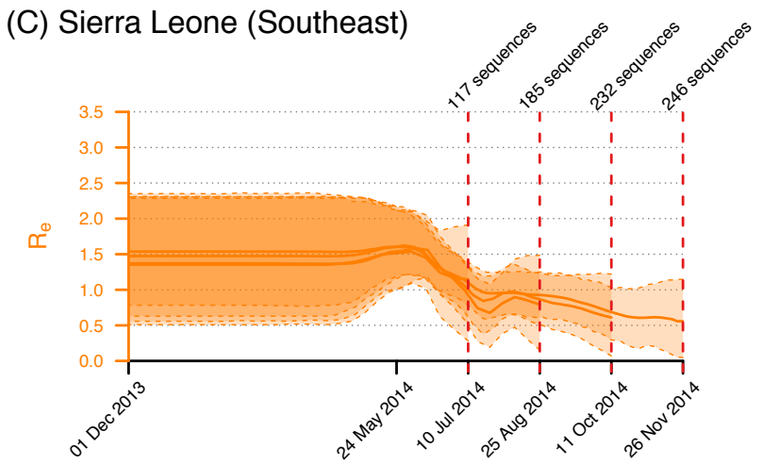
Louis du Plessis and Carsten Magnus

D-BSSE
Department of Biosystems Science and Engineering

Myself







Real-time phylodynamics

- Viral data sampled over the course of an outbreak

Model biases and limitations

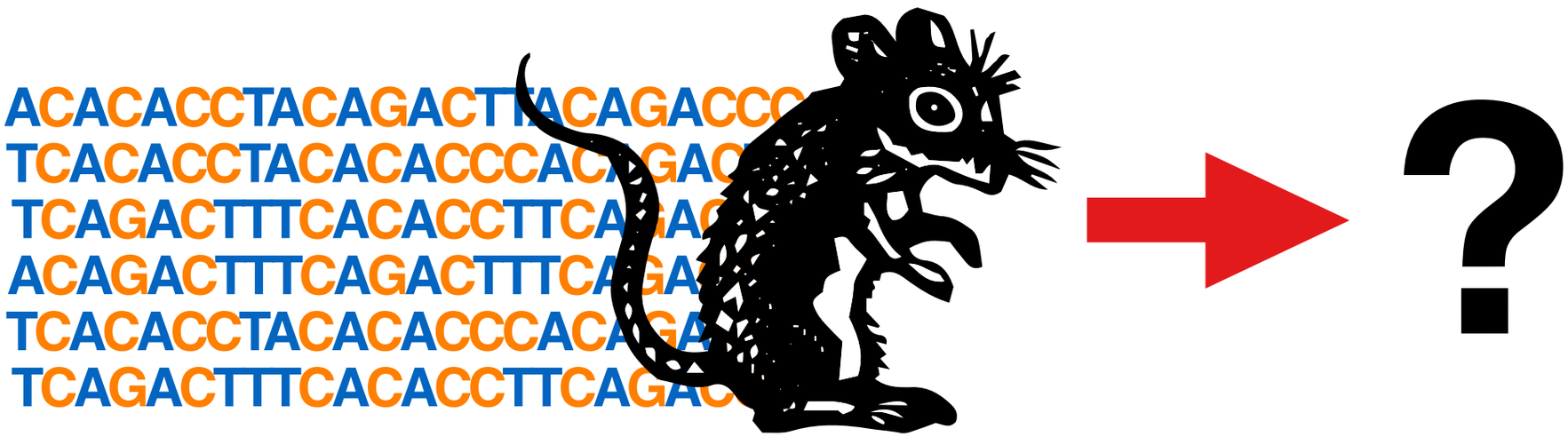
- Simulated data

Everyone else

- Evolution of HIV-1, tuberculosis, leprosy, Influenza etc.
- Evolution of drug resistance
- Host-parasite co-evolution
- Disease spread by migration/travel
- Species delimitation
- Bird taxonomy
- Plant diversity and dispersal
- Adaptive radiations and diversification rates
- Time calibration using fossils
- Ancient DNA
- Influence of environmental factors on phylogenetic diversity
- Total evidence dating

…etc

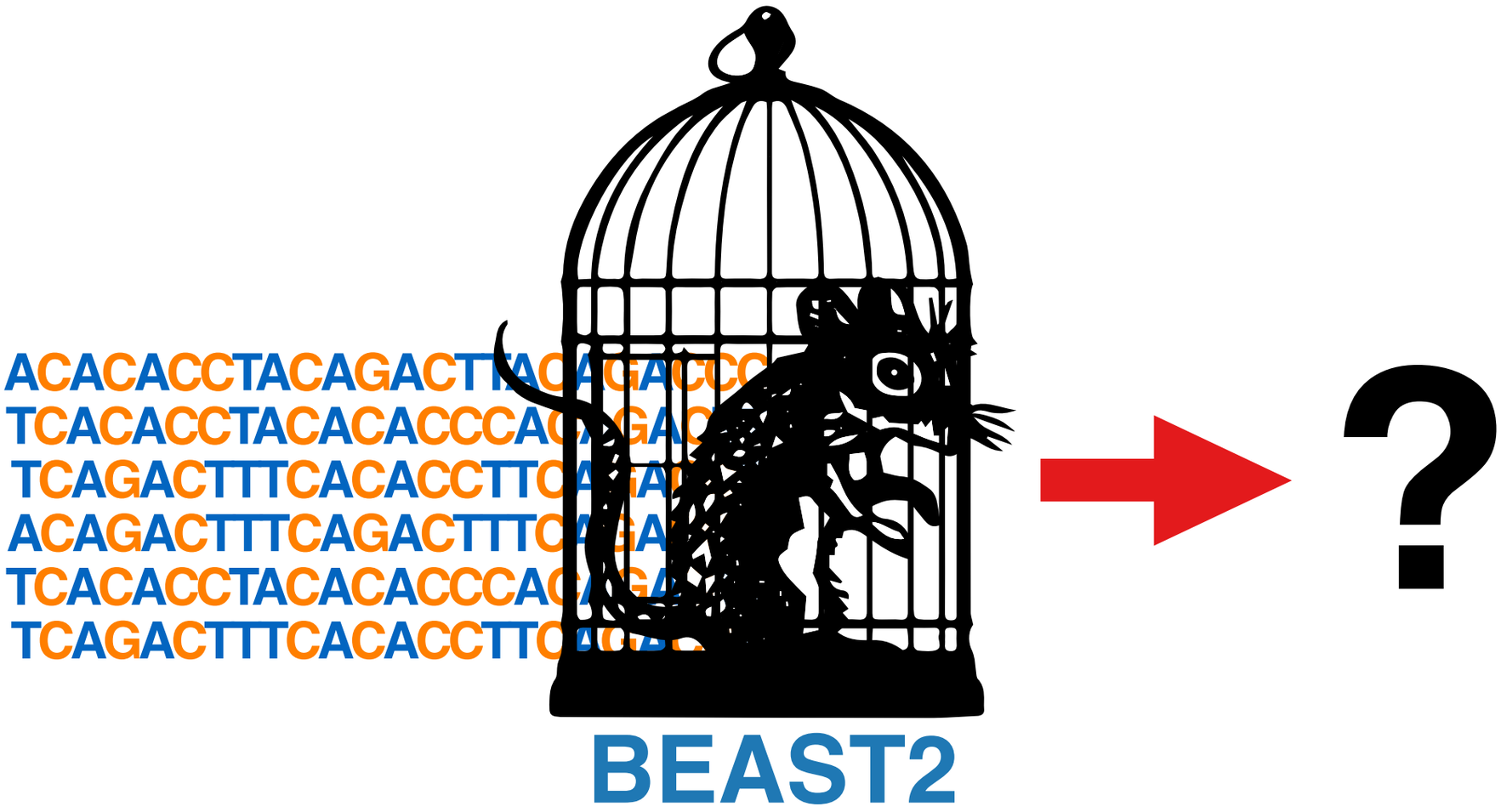
We all have one thing in common…



BEAST2

All of us use genomic sequencing data
to answer questions in **BEAST**

We all have one thing in common…



All of us use genomic sequencing data
to answer questions in **BEAST**

Bayesian inference recap

$$P(\textbf{model} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{model})P(\textbf{model})}{P(\textbf{data})}$$

Bayesian inference recap

Likelihood

$$P(\textbf{model} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{model})P(\textbf{model})}{P(\textbf{data})}$$

Posterior

Prior

Bayesian inference recap

Likelihood

$$P(\textbf{model} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{model})P(\textbf{model})}{P(\textbf{data})}$$

Posterior

Marginal likelihood of the data

Prior

Bayesian inference recap

Data

- One or more alignments of sequencing data
- Sampled at one or many time points
- Timescale of days to millions of years
- Realisation of a stochastic process

P(el)

Posterior

Marginal likelihood of the data

Prior

Bayesian inference recap

Data

- One or more alignments of sequencing data
- Sampled at one or many time points
- Timescale of days to millions of years
- Realisation of a stochastic process

P(el)

Model

- Description of the process that generated the data
- Parameters are random variables
- We may only be interested in some of the parameters
- Still need a prior for all the model parameters!

What goes into a BEAST model?

genealogy

demographic model

site model

molecular clock model

What goes into a BEAST model?

genealogy

demographic model

site model

molecular clock model

What goes into a BEAST model?

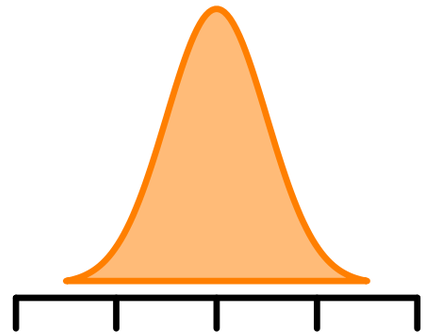
genealogy

demographic model

site model

molecular clock model

BEAST2



What goes into a BEAST model?

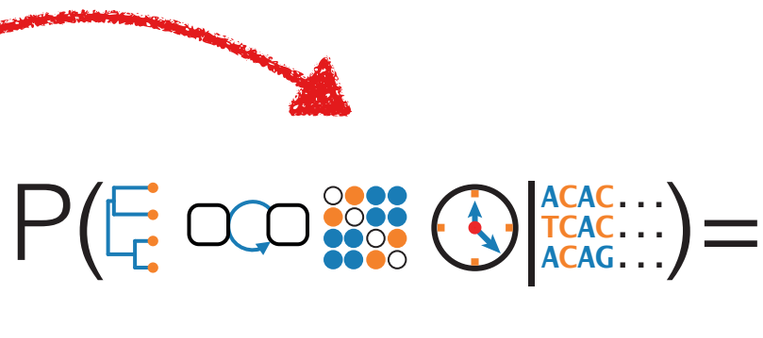
genealogy

demographic model

site model

molecular clock model

BEAST2



Estimate posterior distributions!

Demographic model

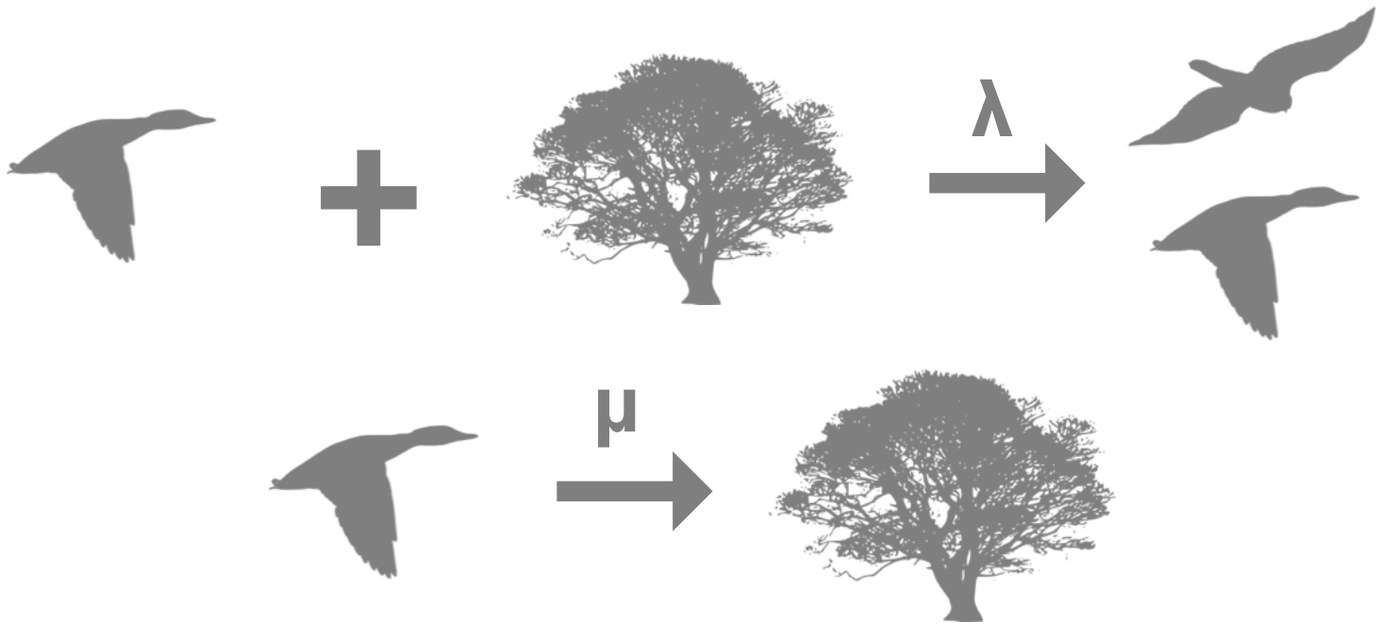
- Description of the population dynamics
- How does the population grow over time?
  - Infectious diseases - transmission/recovery
  - Species - speciation/extinction
- Migration events
- Different classes of individuals
- Host/vector dynamics

…etc

Demographic model

- Description of the population dynamics
- How does the population grow over time?
  - Infectious diseases - transmission/recovery
  - Species - speciation/extinction
- Migration events
- Different classes of individuals
- Host/vector dynamics

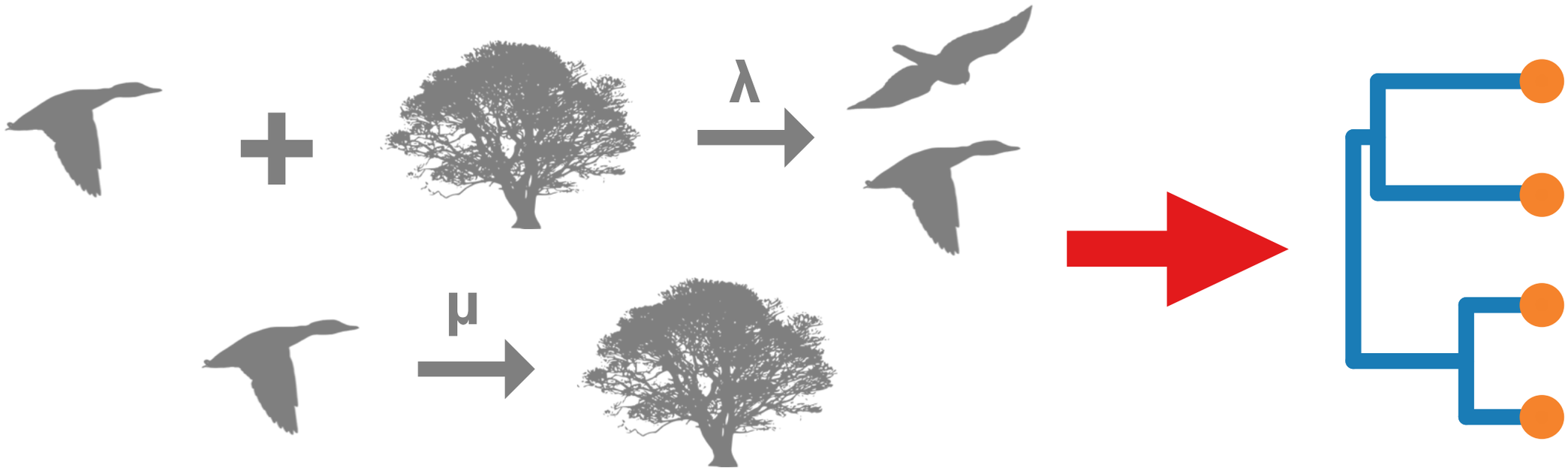
…etc



Demographic model

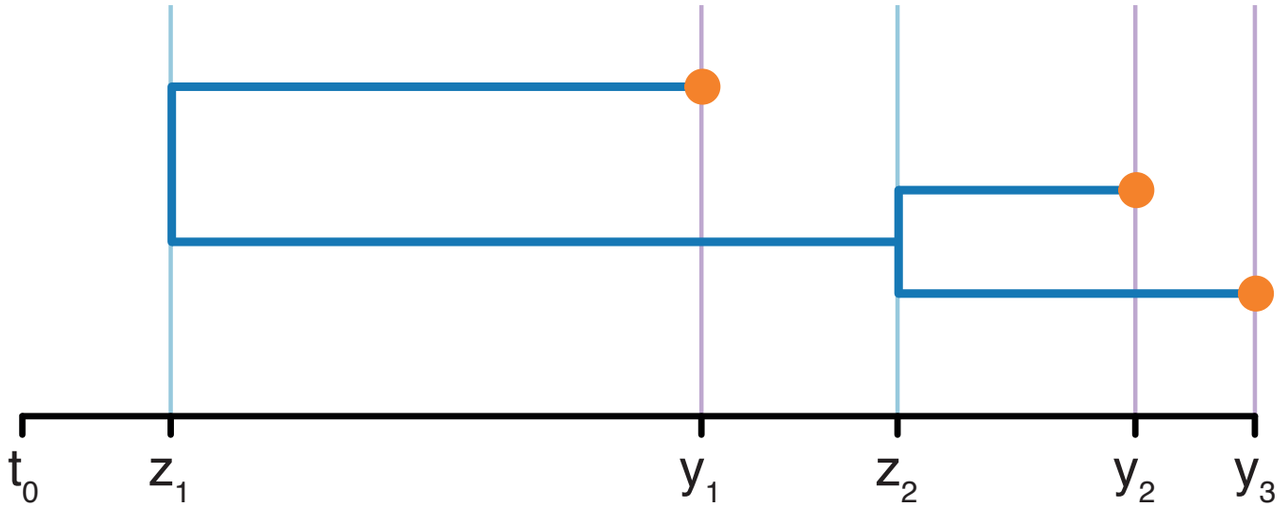
- Description of the population dynamics
- How does the population grow over time?
  - Infectious diseases - transmission/recovery
  - Species - speciation/extinction
- Migration events
- Different classes of individuals
- Host/vector dynamics

…etc



Types of demographic models

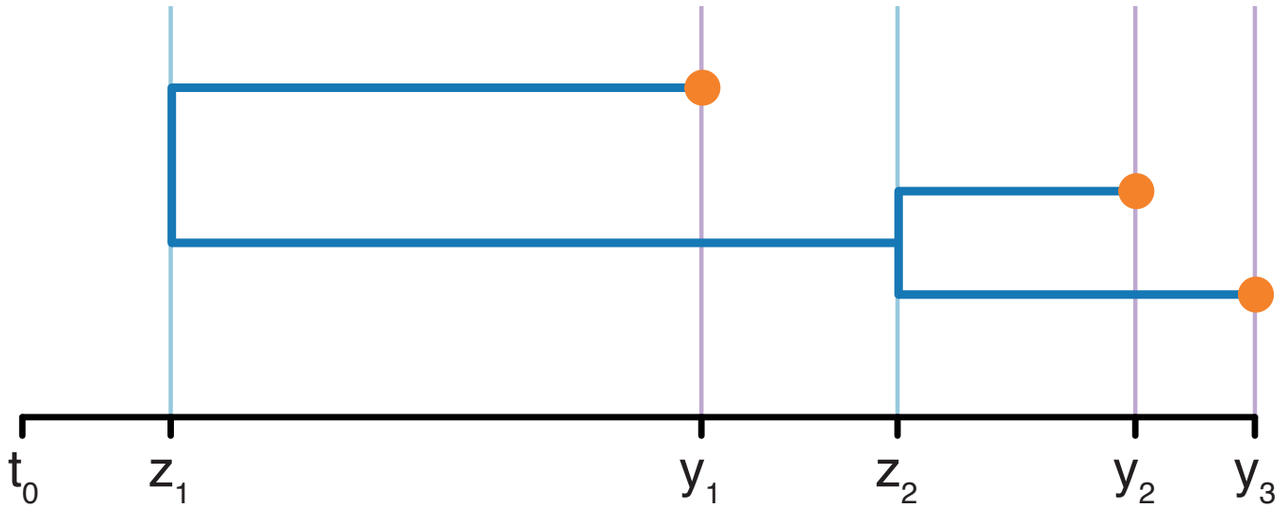
Birth-death →



← Coalescent

Types of demographic models

Birth-death →



← **Coalescent**

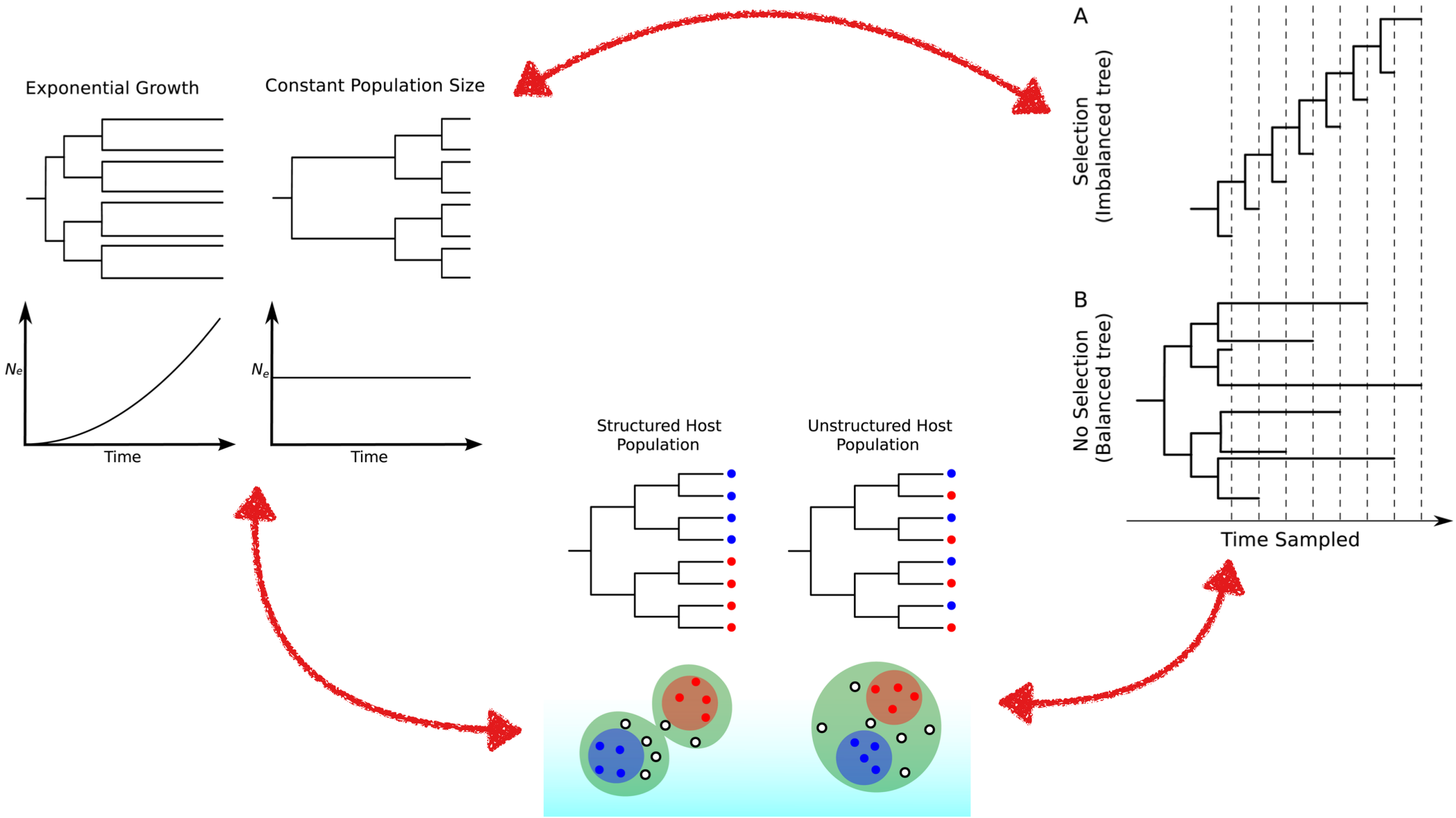
Birth-death models

- Speciation/infection rate
- Extinction/recovery rate
- Explicitly models sampling

Coalescent

- Effective population size
- Conditions on sampling a small random sample

Different population dynamics generate different trees

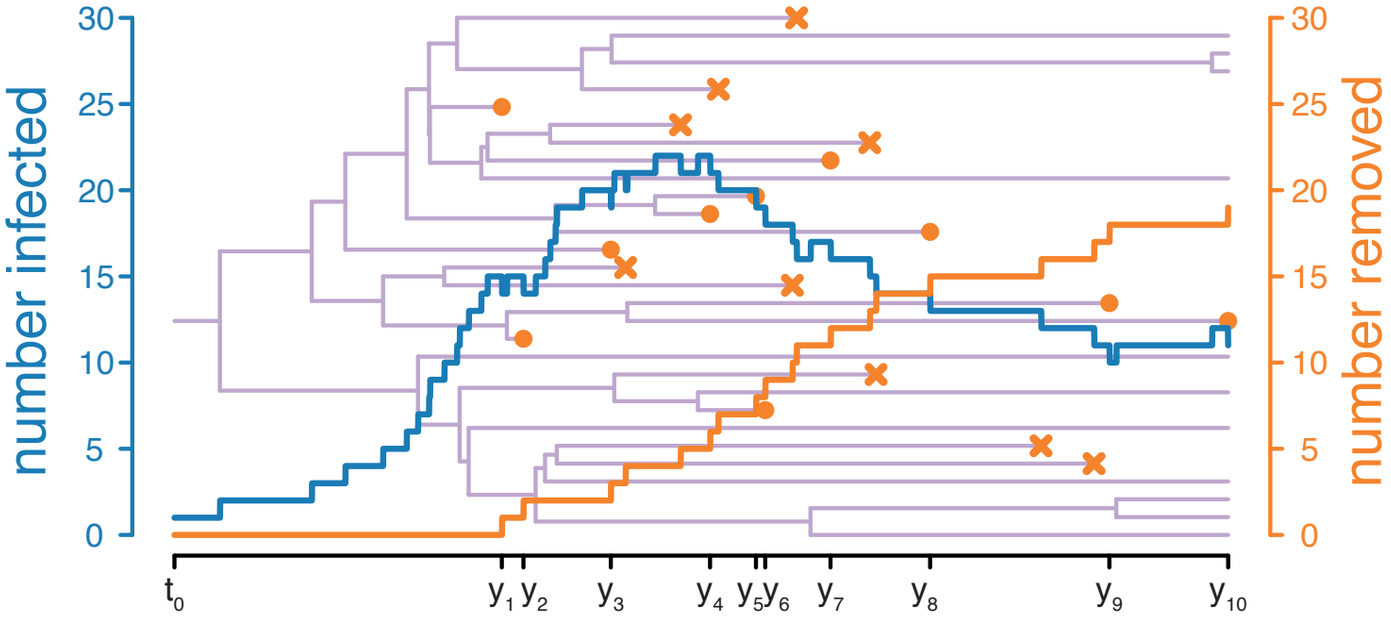


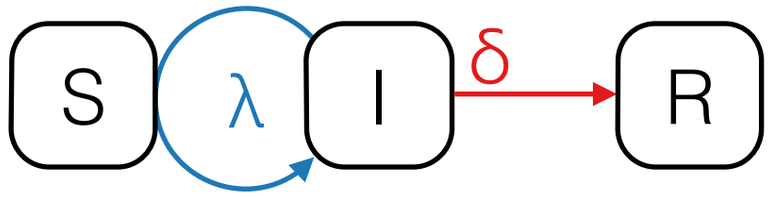
The genealogy

- What are the ancestral relationships between the sequences in our dataset?
- Only the relationships between the **sampled** sequences!

The genealogy

- What are the ancestral relationships between the sequences in our dataset?
- Only the relationships between the **sampled** sequences!

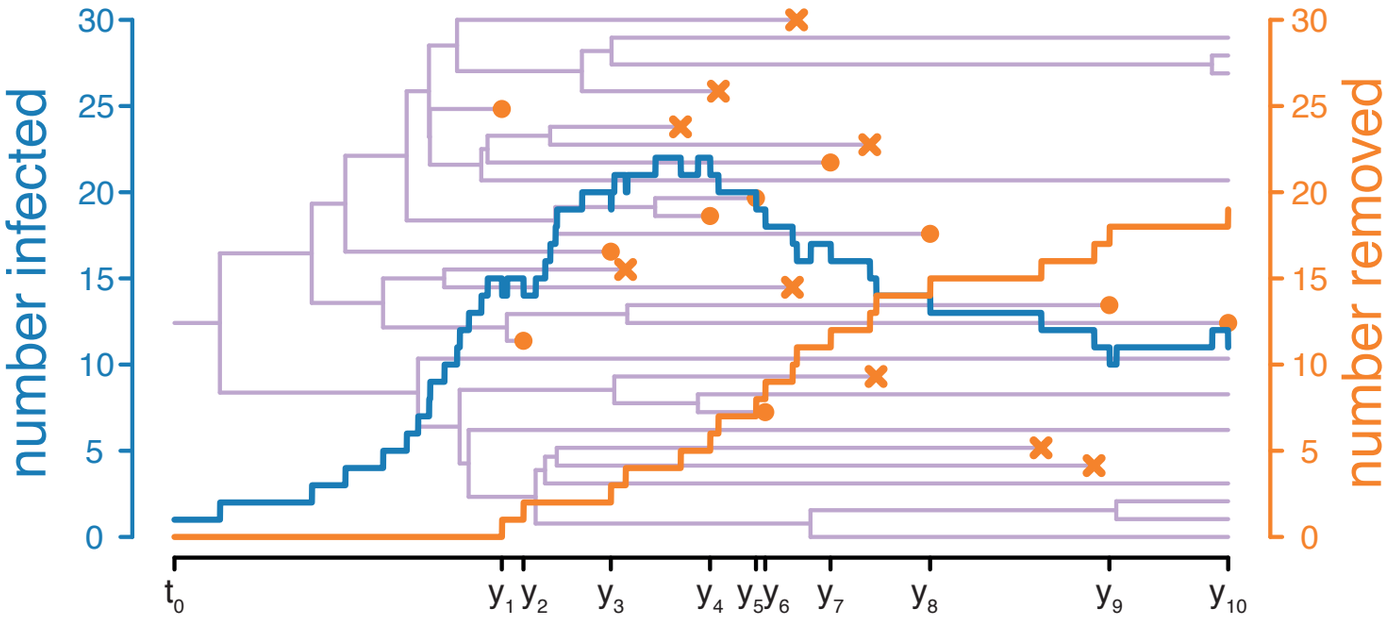


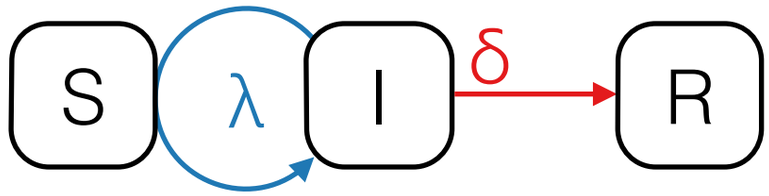


Full tree generated by an SIR model

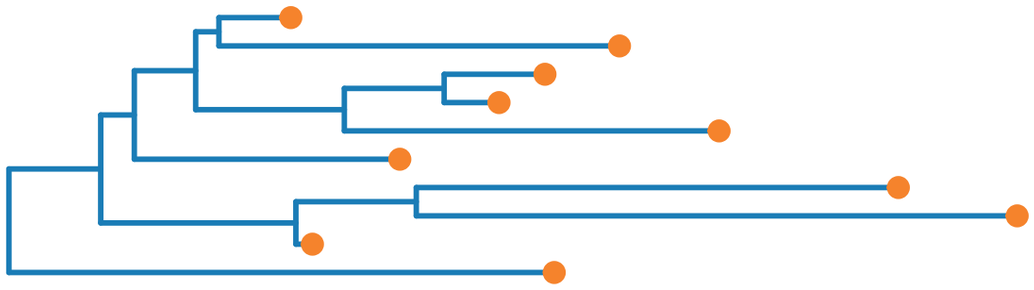
The genealogy

- What are the ancestral relationships between the sequences in our dataset?
- Only the relationships between the **sampled** sequences!



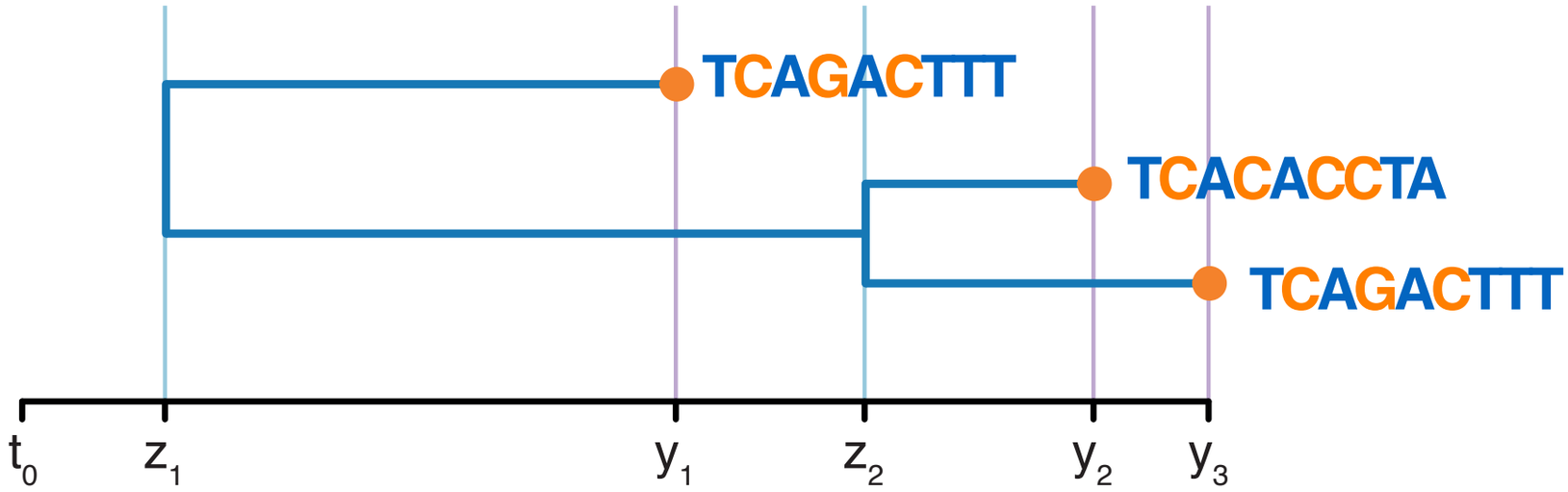


Full tree generated by an SIR model

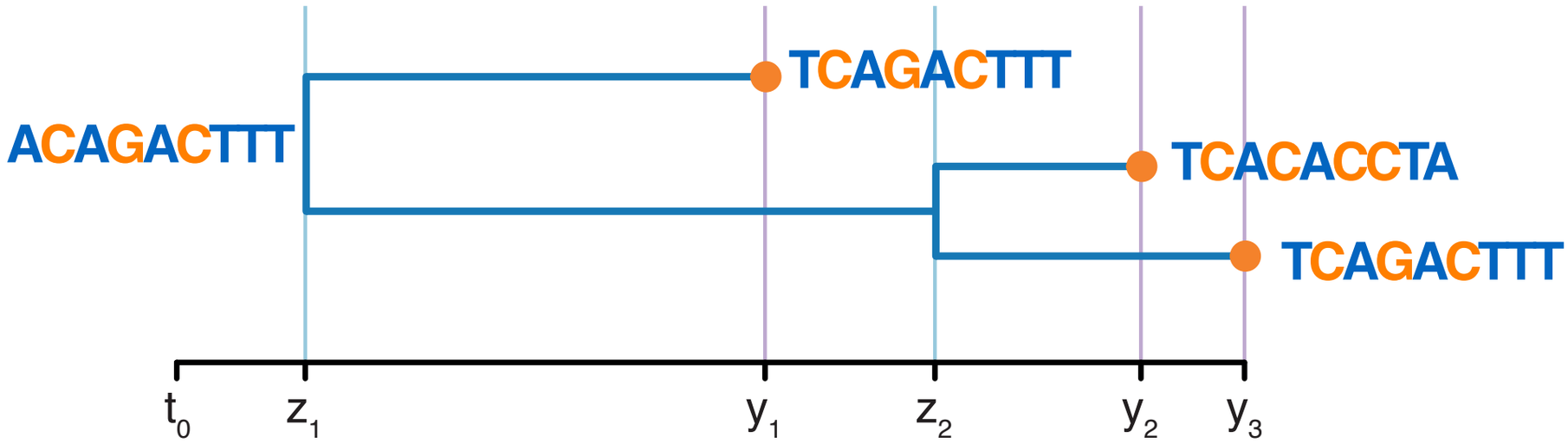


Sampled tree

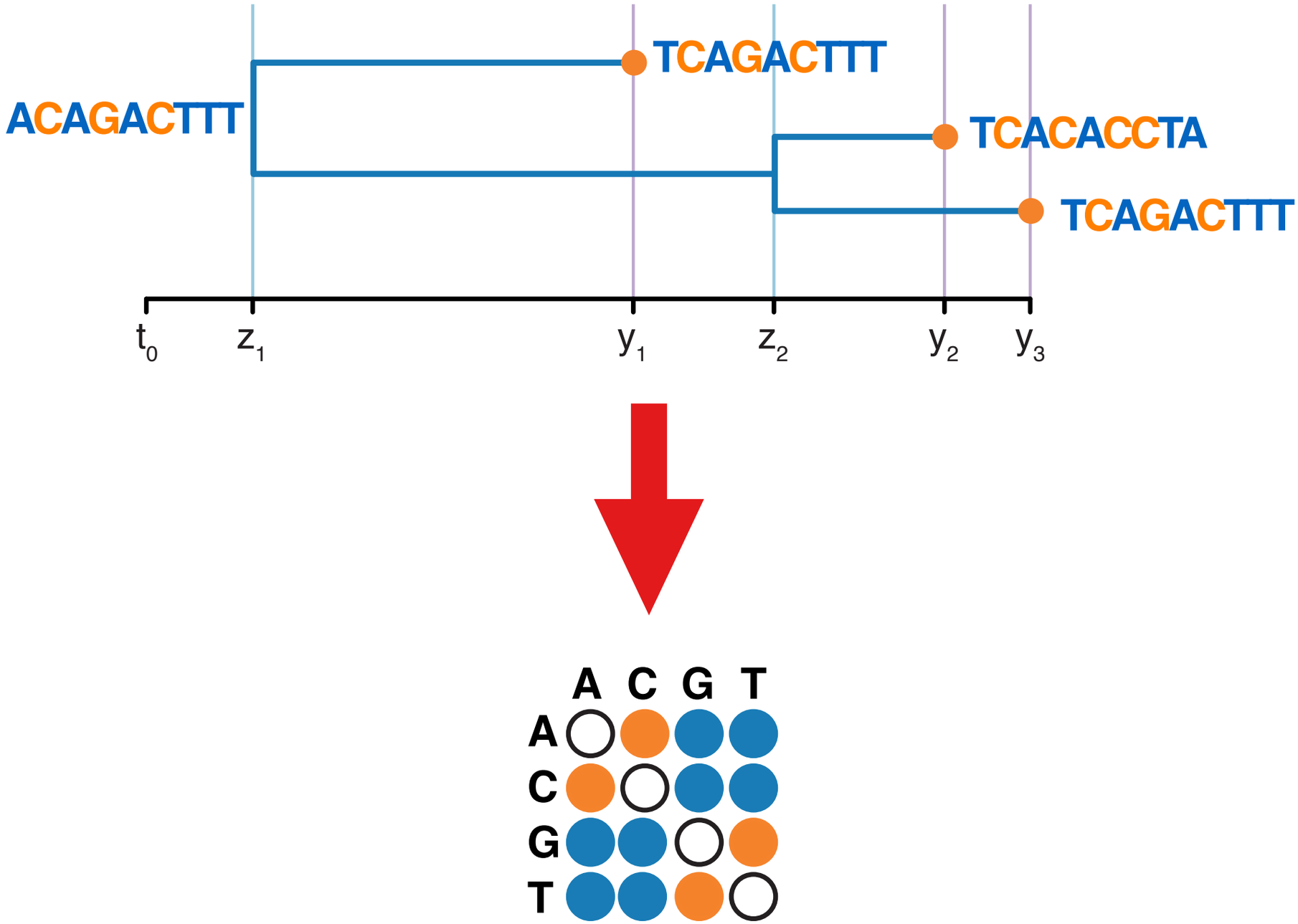
The genealogy



The genealogy



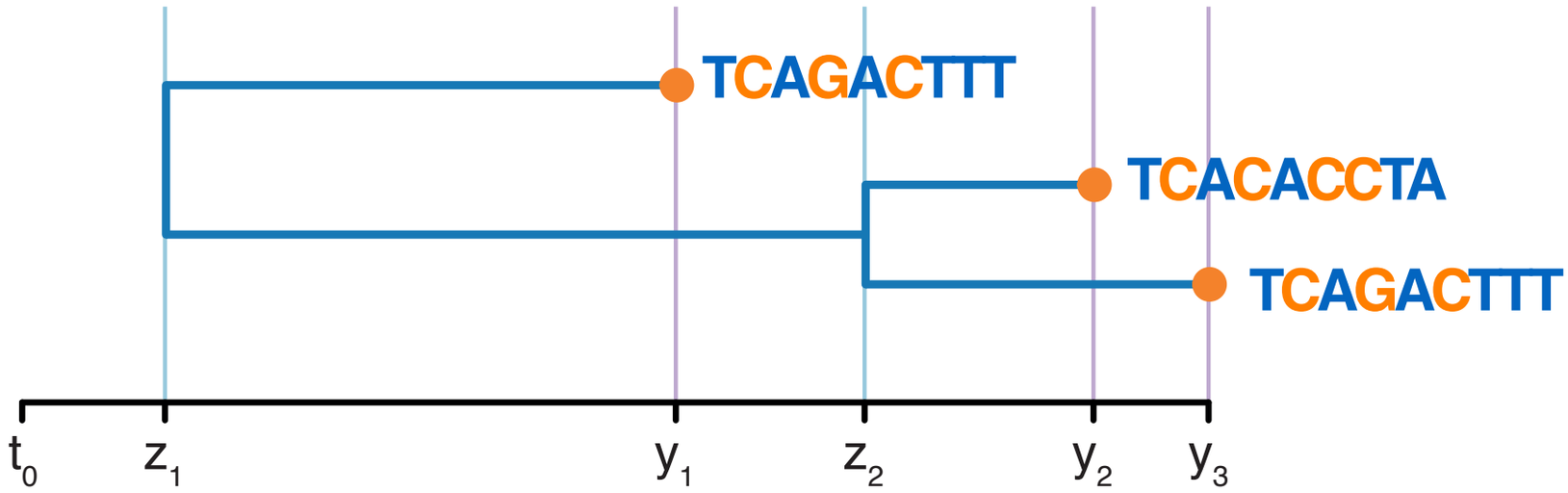
The genealogy



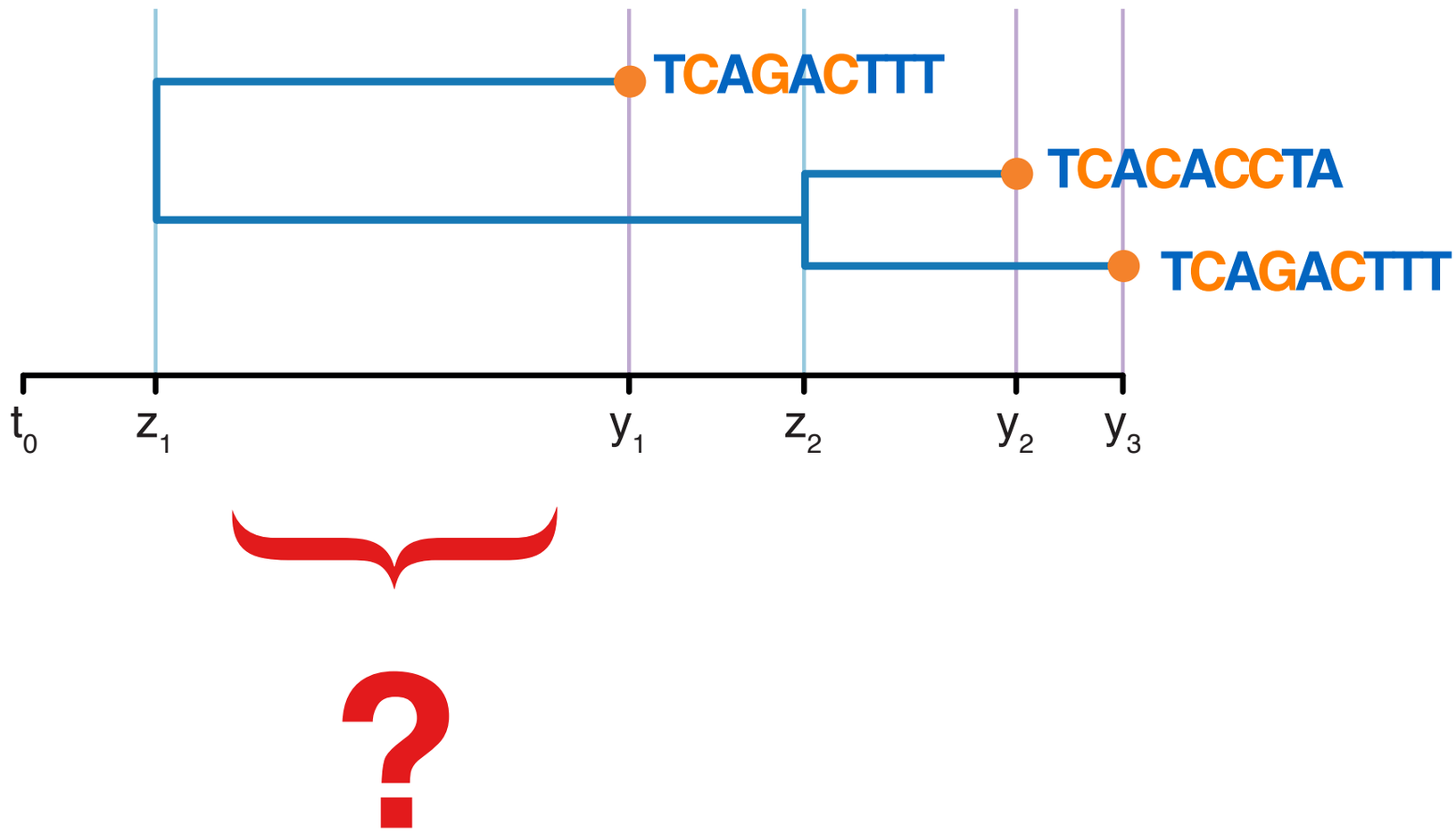
Site model

- How does one sequence change into another?
- Basic substitution model
  - Relative rate of substitution from one nucleotide to another
  - Or amino acids
  - Or codons
- Site variation
- Invariant sites
- Separate site model for each data partition
  - Multiple genes/loci
  - Model 1st, 2nd, 3rd codon positions differently

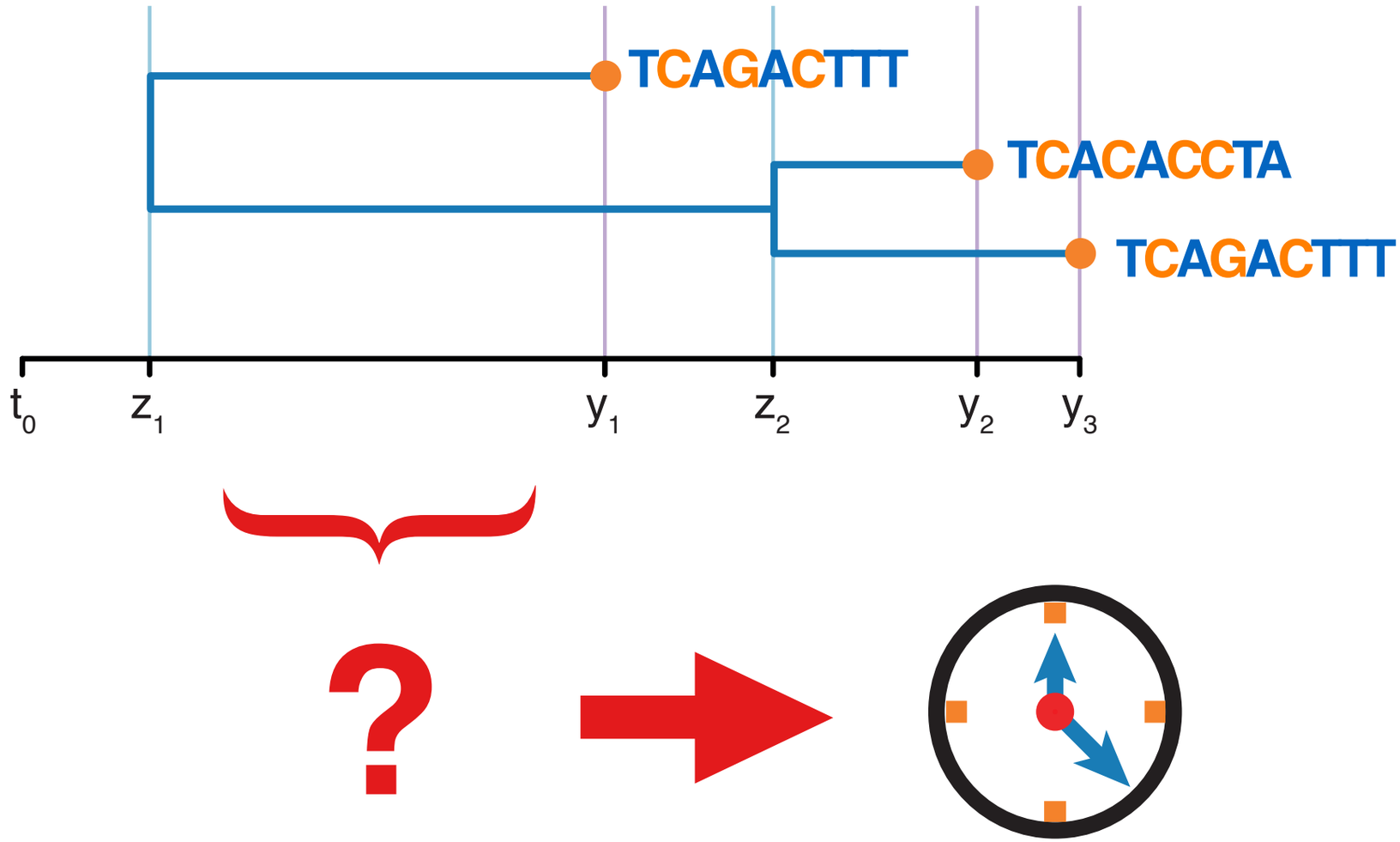
Site model

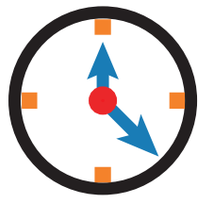


Site model



Site model





Molecular clock

- How long does it take for substitutions to appear?
- Scales branch lengths to calendar time
- Different branches may have different clock rates

Putting it all together

genetic sequences

genealogy

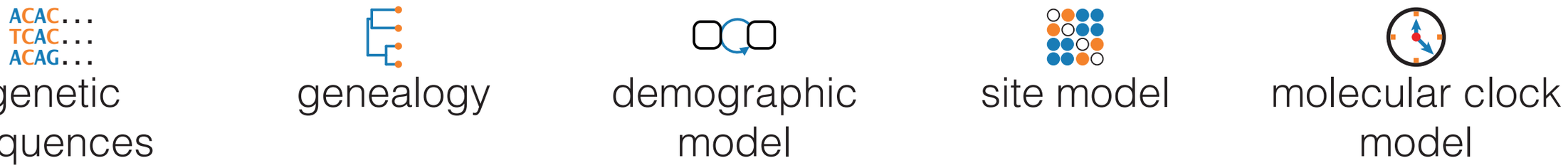
demographic model

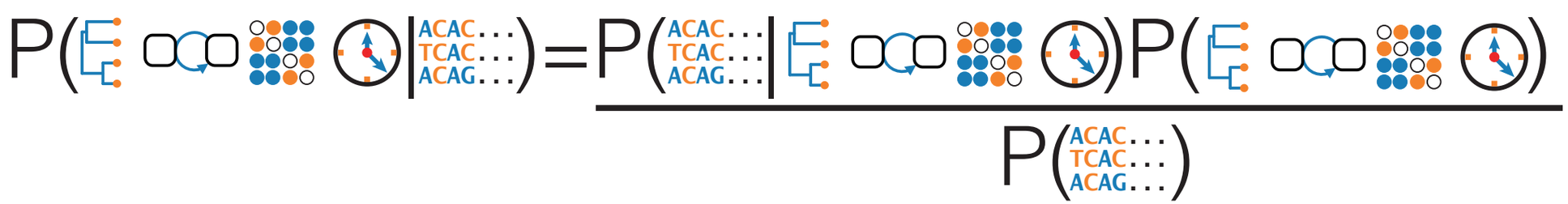
site model

molecular clock model

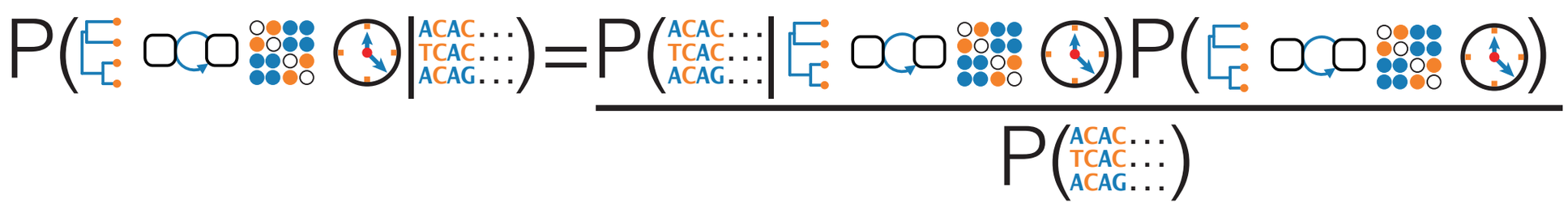
$$P(\textbf{model} \mid \textbf{data}) = \frac{P(\textbf{data} \mid \textbf{model})P(\textbf{model})}{P(\textbf{data})}$$

Putting it all together





Putting it all together



Assume independence

Posterior distribution in BEAST2

genetic sequences

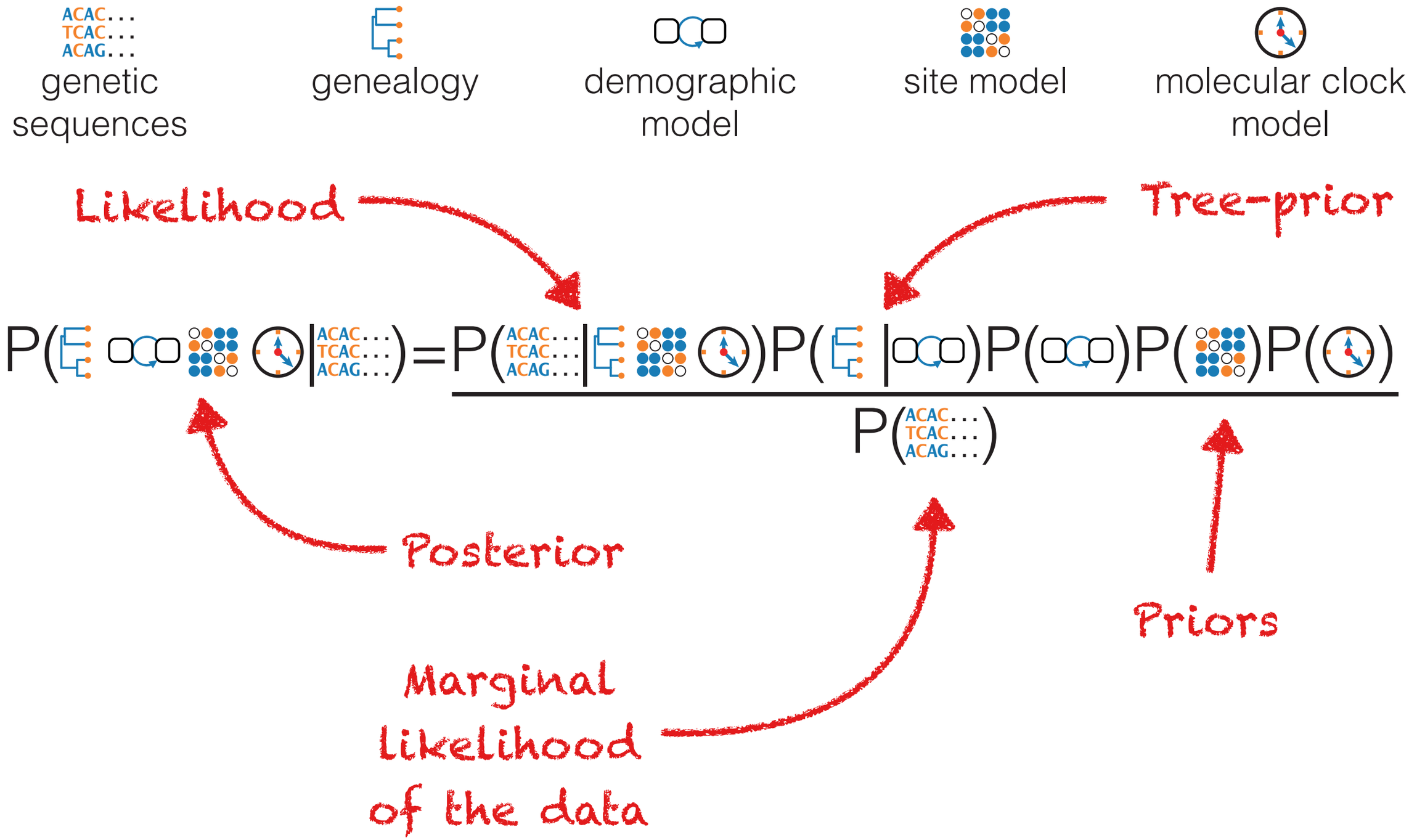

genealogy


demographic model


site model


molecular clock model

Posterior distribution in BEAST2



Posterio

ACAC...
TCAC...
ACAG...
genetic sequences

Likelih

P(

ecular clock model

e-prior

)P(○)

rs

MCMC inference

- We want to calculate the posterior
- But we cannot easily calculate the marginal likelihood
  → use MCMC!

Markov-chain

- Stochastic process
- Jumps between different states
- Memoryless

Monte Carlo algorithm

- Randomized algorithm
- Deterministic runtime
  (it **will** finish)
- Output may **not** be correct
  (with some small probability)

MCMC inference

- MCMC performs a random walk on the posterior, preferentially sampling high-density areas
- Only need to compare which posterior density is higher
- So we only need the ratio of posteriors and the marginal likelihoods cancel out

MCMC inference

- MCMC performs a random walk on the posterior, preferentially sampling high-density areas
- Only need to compare which posterior density is higher
- So we only need the ratio of posteriors and the marginal likelihoods cancel out

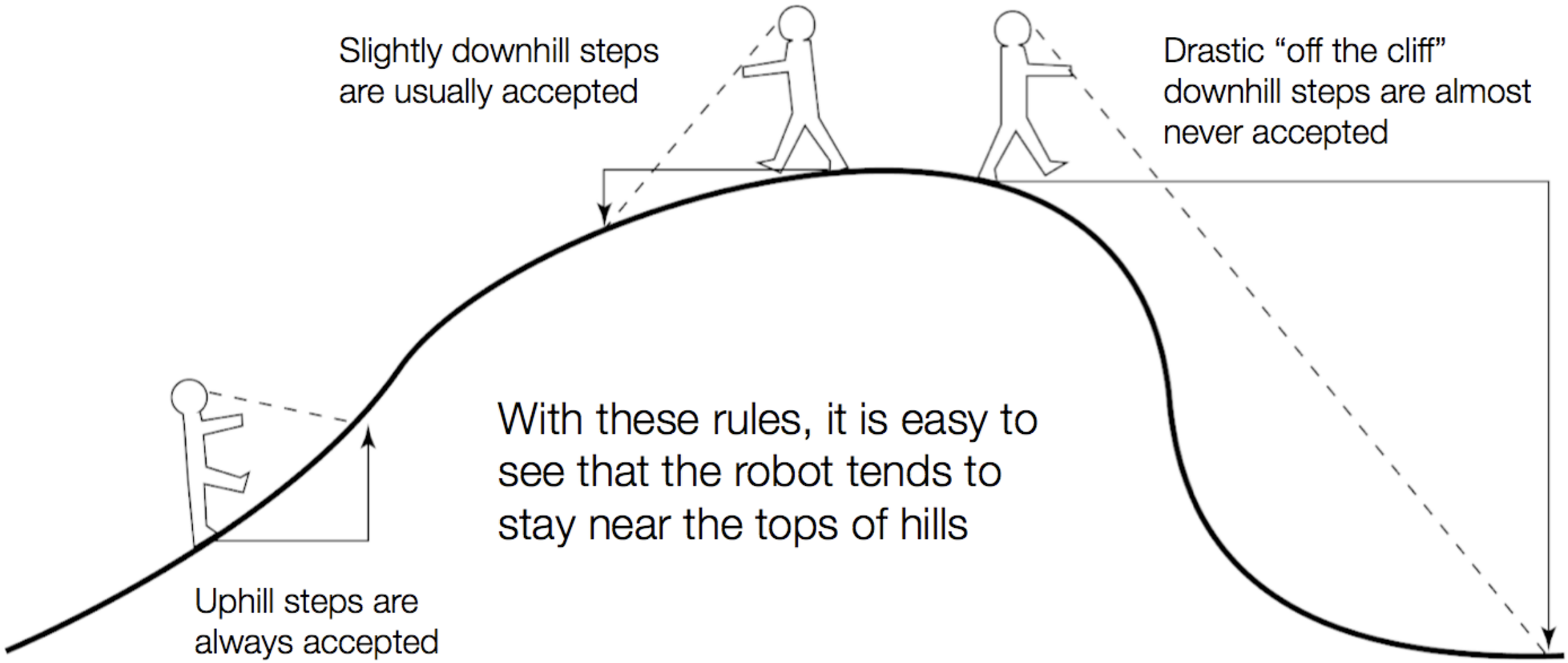
$$\frac{P(\mathbf{model_1} \mid \mathbf{data})}{P(\mathbf{model_2} \mid \mathbf{data})} = \frac{\frac{P(\mathbf{data} \mid \mathbf{model_1})P(\mathbf{model_1})}{P(\mathbf{data})}}{\frac{P(\mathbf{data} \mid \mathbf{model_2})P(\mathbf{model_2})}{P(\mathbf{data})}}$$

MCMC inference

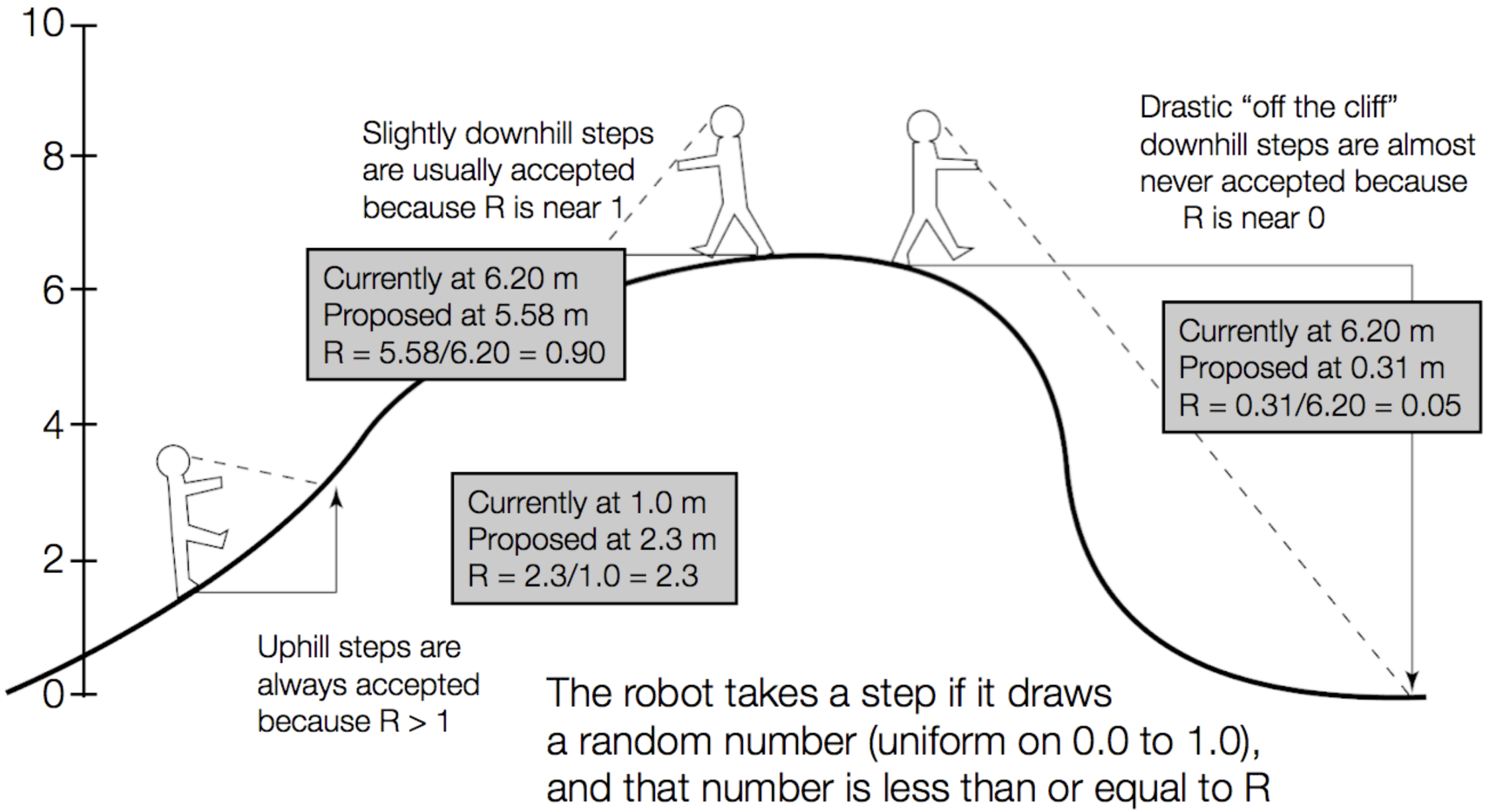
- MCMC performs a random walk on the posterior, preferentially sampling high-density areas
- Only need to compare which posterior density is higher
- So we only need the ratio of posteriors and the marginal likelihoods cancel out

$$\frac{P(\mathbf{model_1} \mid \mathbf{data})}{P(\mathbf{model_2} \mid \mathbf{data})} = \frac{\frac{P(\mathbf{data} \mid \mathbf{model_1})P(\mathbf{model_1})}{\cancel{P(\mathbf{data})}}}{\frac{P(\mathbf{data} \mid \mathbf{model_2})P(\mathbf{model_2})}{\cancel{P(\mathbf{data})}}}$$

MCMC robot (courtesy of Paul Lewis)

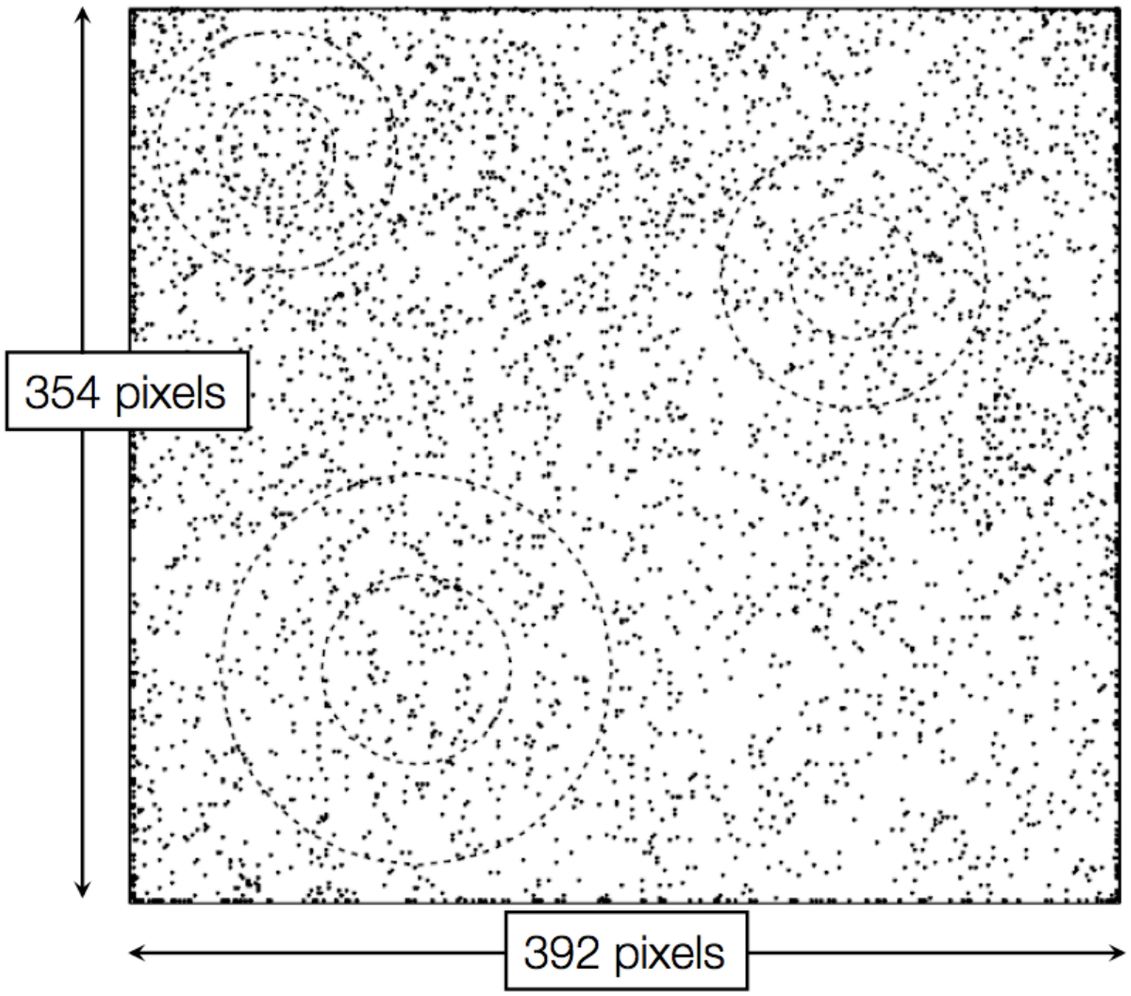


MCMC robot (courtesy of Paul Lewis)



(R is the ratio between the posterior densities)

Pure random walk (courtesy of Paul Lewis)



Random walk

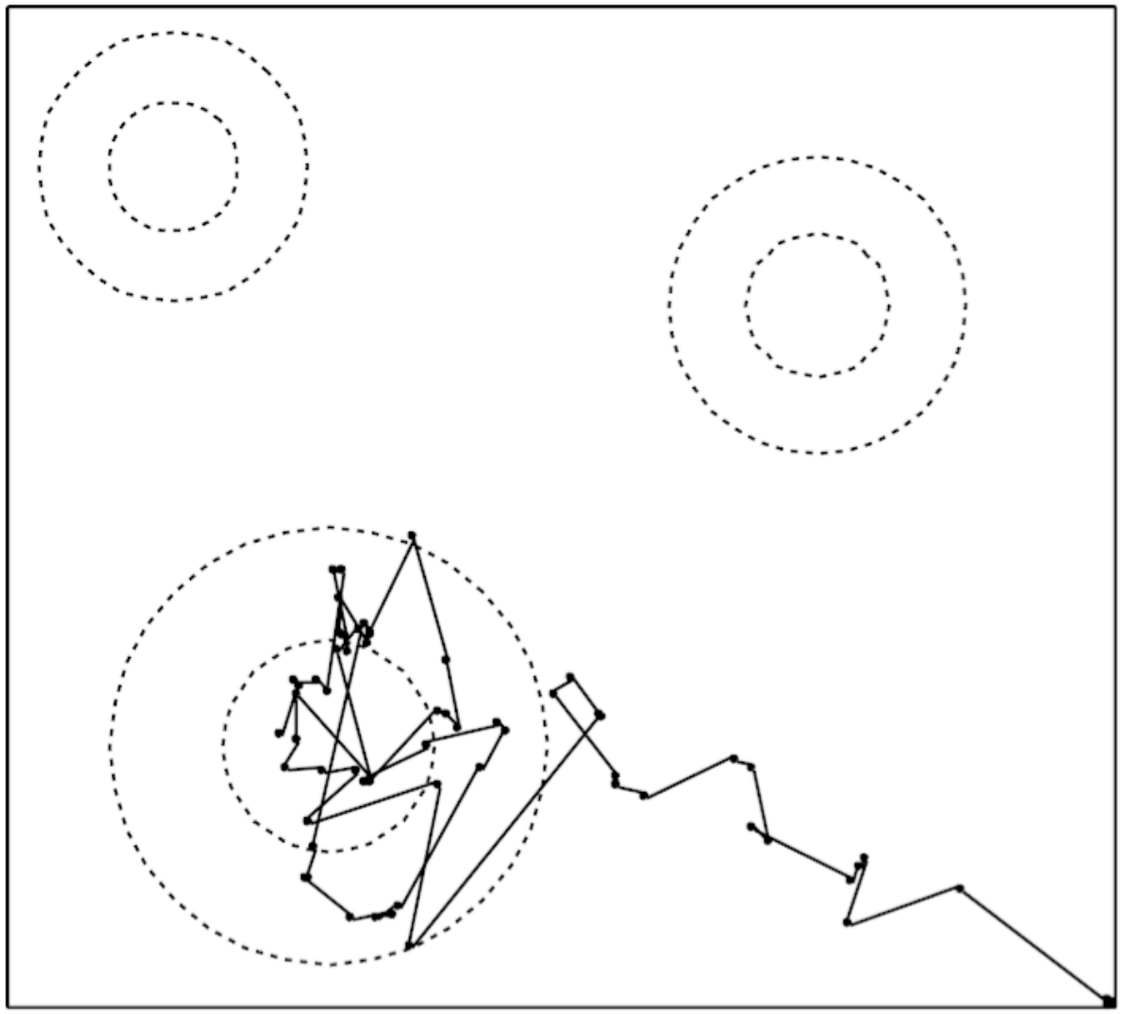
- Random direction
- Gamma distributed step size
- Reflection at edges

Target distribution

- Equal mixture of 3 bivariate normal hills
- Inners contours: 50%
- Outer contours: 95%

5000 steps by the robot

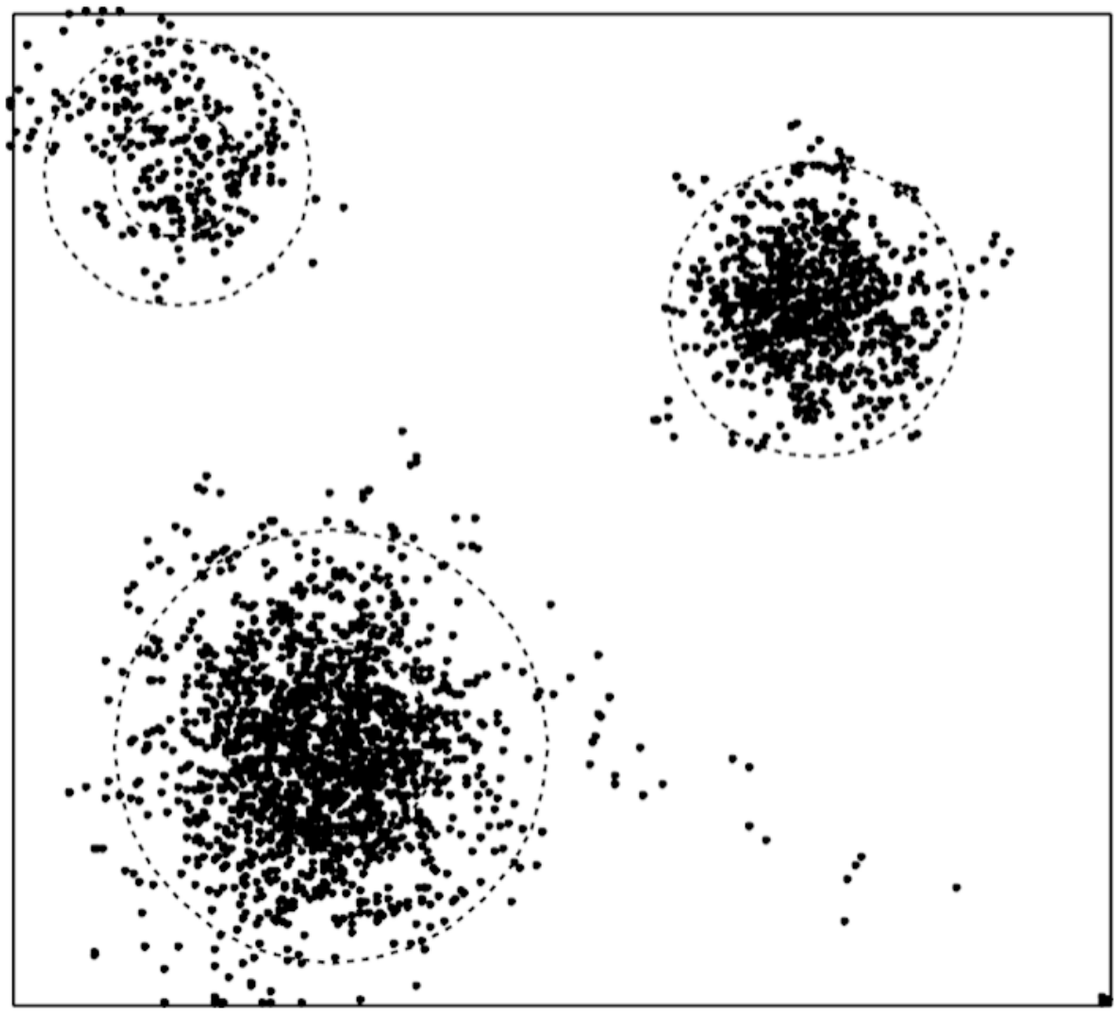
Burn in (courtesy of Paul Lewis)



- Using MCMC rules
- Quickly finds one of the 3 hills
- First few steps are not representative of the distribution

100 steps by the robot

MCMC approximation (courtesy of Paul Lewis)



5000 steps by the robot

How good is the approximation?

- 51.2% of points inside 50% contours
- 93.6% of points inside 95% contours

The more steps, the better the accuracy

Summary: MCMC inference

Target distribution

- This is the posterior in BEAST2: $P($  $)$

Proposal distribution

- Used to decide where to step to next
- The choice only affects the efficiency of the algorithm
- Metropolis algorithm: symmetric proposals
- Metropolis-Hastings: asymmetric proposals

Summary: MCMC inference

Target distribution

- This is the posterior in BEAST2: $P($  $)$

Proposal distribution

- Used to decide where to step to next
- The choice only affects the efficiency of the algorithm
- Metropolis algorithm: symmetric proposals
- Metropolis-Hastings: asymmetric proposals

In BEAST and BEAST2 operators are used to propose the next step

Marginal distributions

- We only have the joint posterior: 

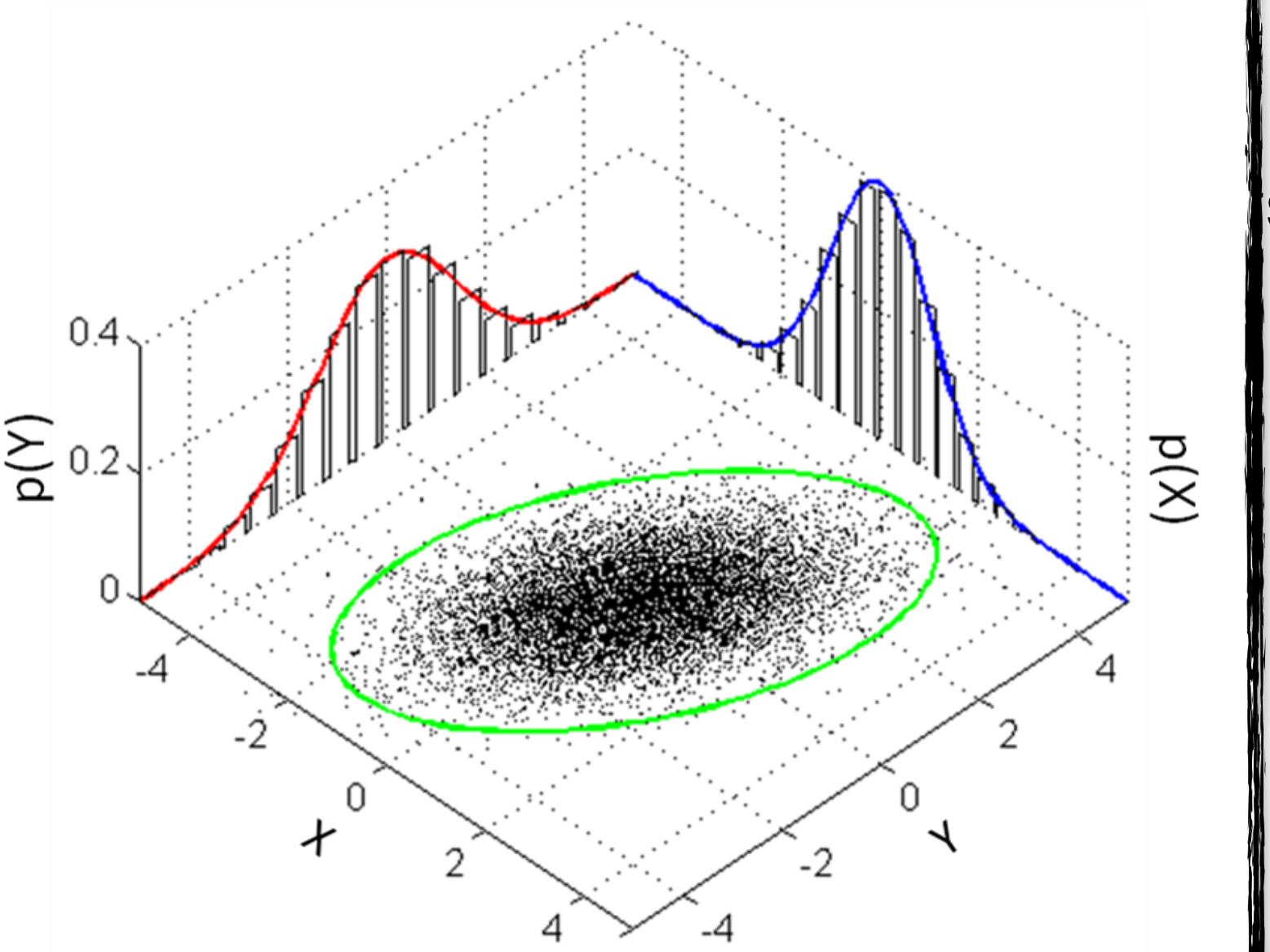
- But we want distributions for each of the parameters we are interested in → marginalize

$$P(\phi) = \int_{\theta} P(\phi|\theta)P(\theta)d\theta$$

Margin

- We on ::)
- But we s
  we are

In practice



One tree to rule them all?

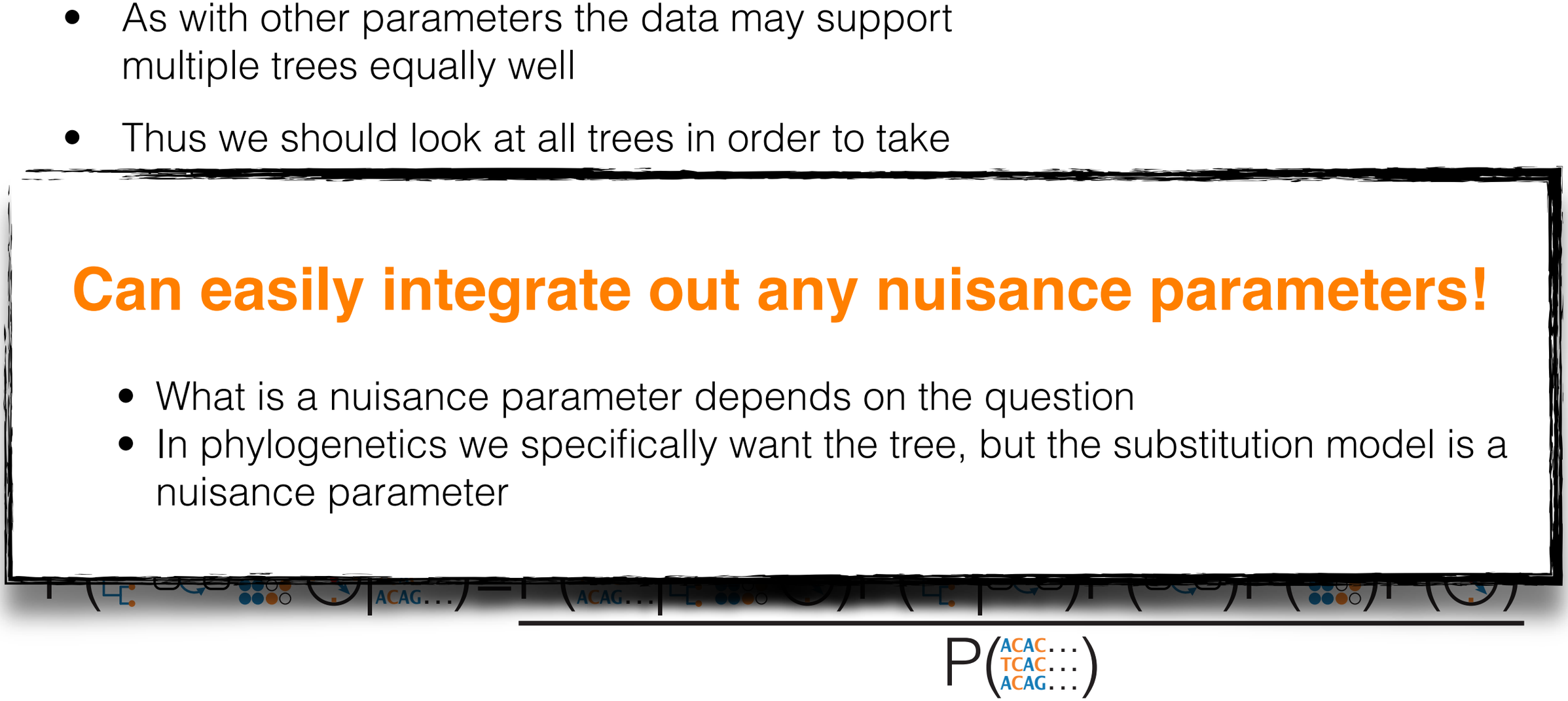
- As with other parameters the data may support multiple trees equally well
- Thus we should look at all trees in order to take phylogenetic uncertainty into account
- If we are only interested in the epidemiological parameters we can integrate out the tree (but this is very difficult in an ML framework)
- Naturally integrate out the tree in an MCMC framework by inferring the distribution of trees

One tree to rule them all?

- As with other parameters the data may support multiple trees equally well
- Thus we should look at all trees in order to take

Can easily integrate out any nuisance parameters!

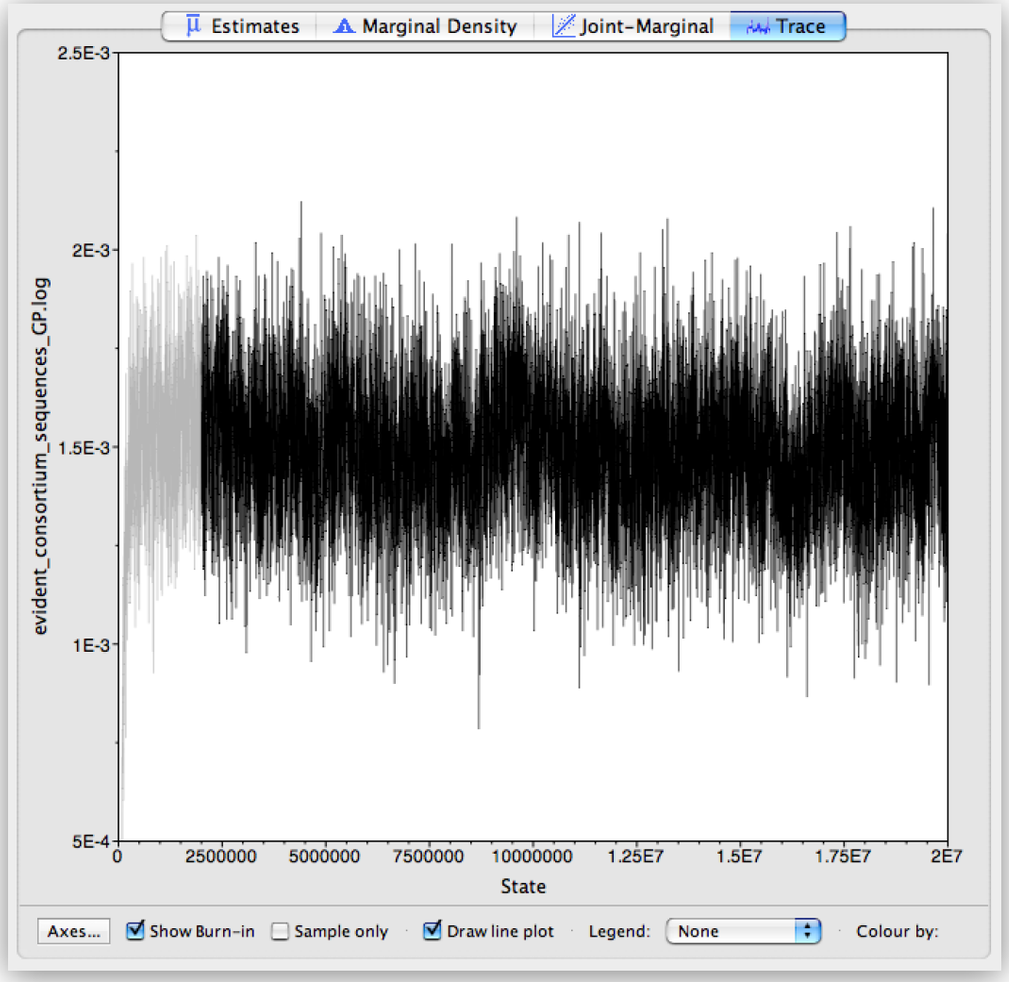
- What is a nuisance parameter depends on the question
- In phylogenetics we specifically want the tree, but the substitution model is a nuisance parameter



What we hope will happen

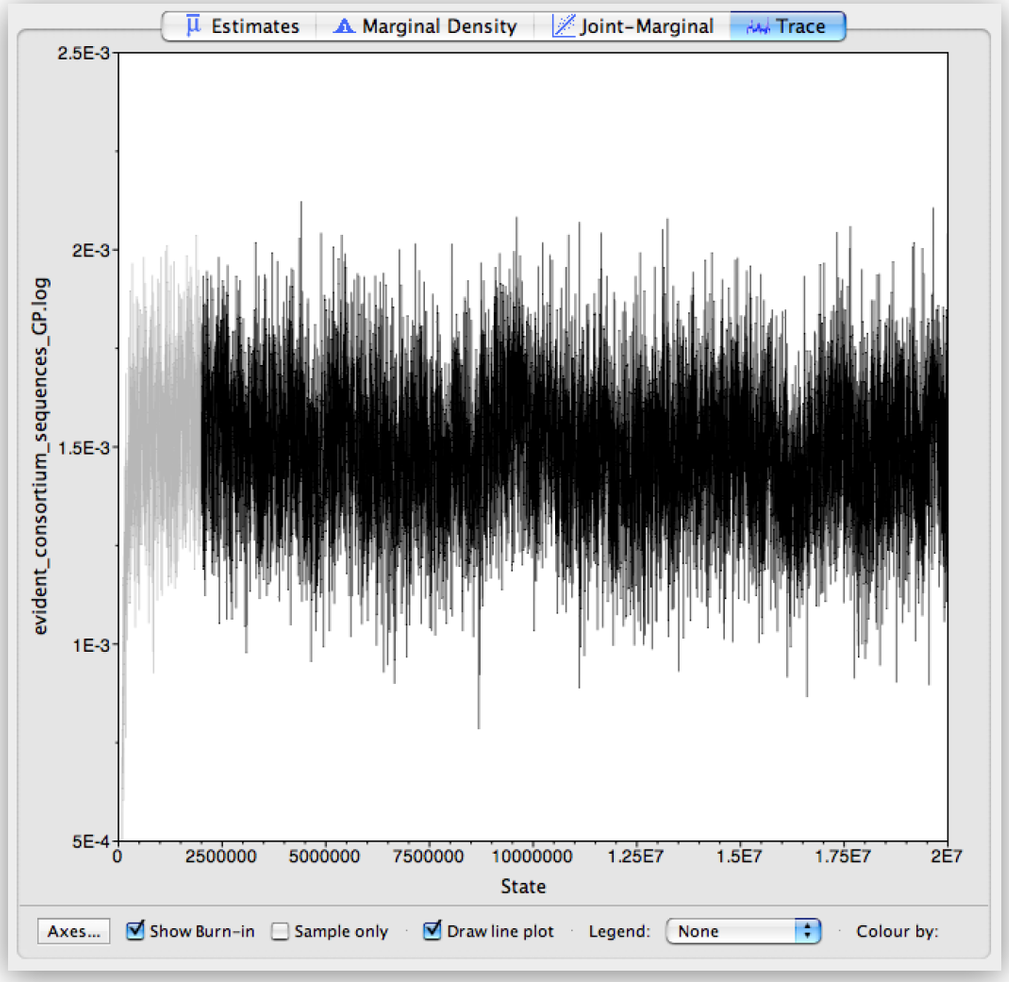
What we hope will happen

- The MCMC algorithm samples efficiently from high density areas of the posterior distribution



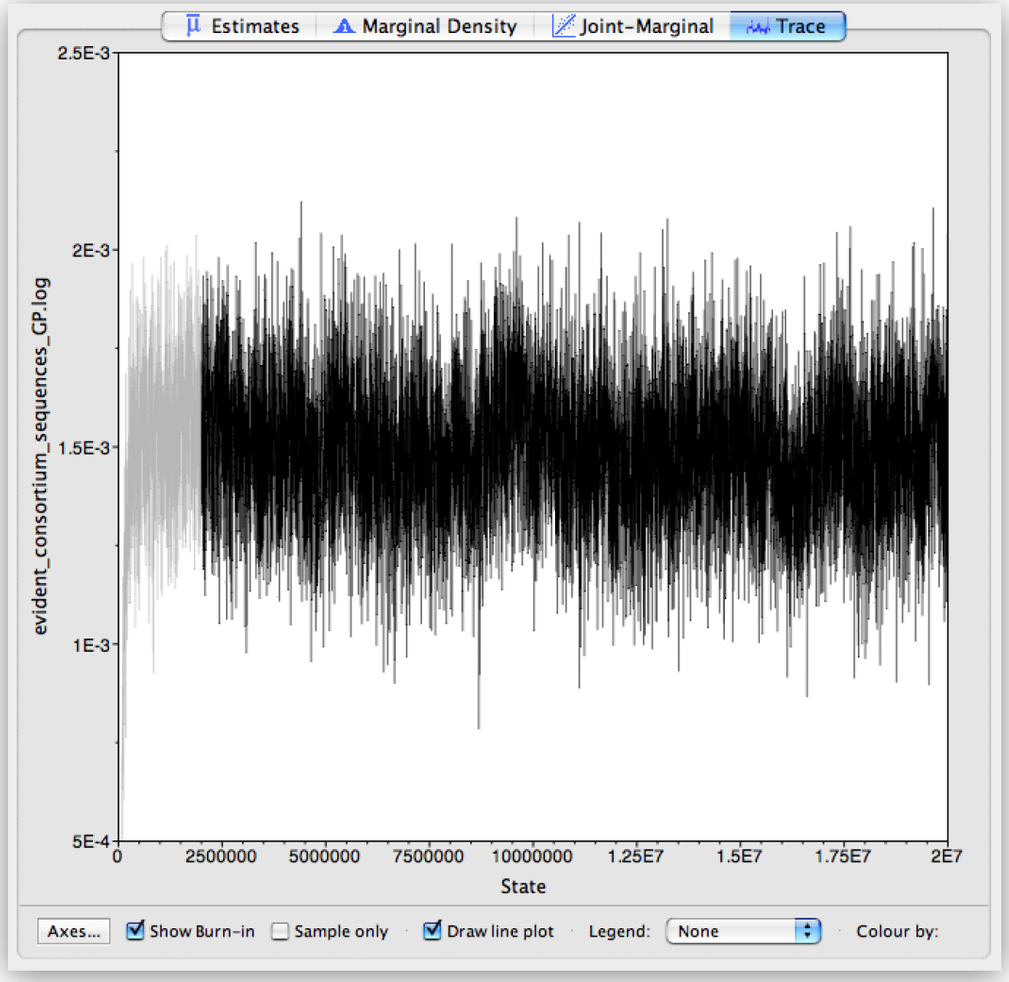
What we hope will happen

- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a **good** approximation of the posterior distribution in **finite** time



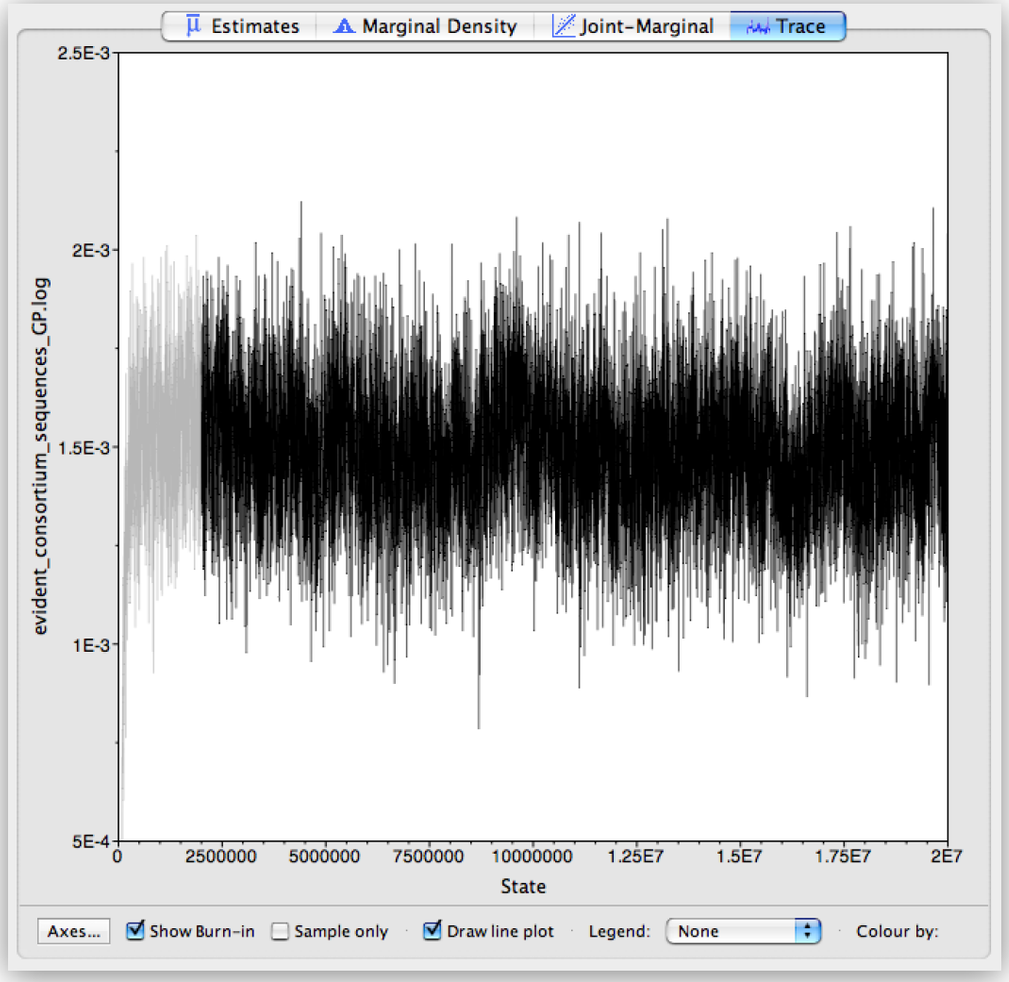
What we hope will happen

- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a **good** approximation of the posterior distribution in **finite** time
- Everything is awesome!



What we hope will happen

- The MCMC algorithm samples efficiently from high density areas of the posterior distribution
- We end up with a **good** approximation of the posterior distribution in **finite** time
- Everything is awesome!

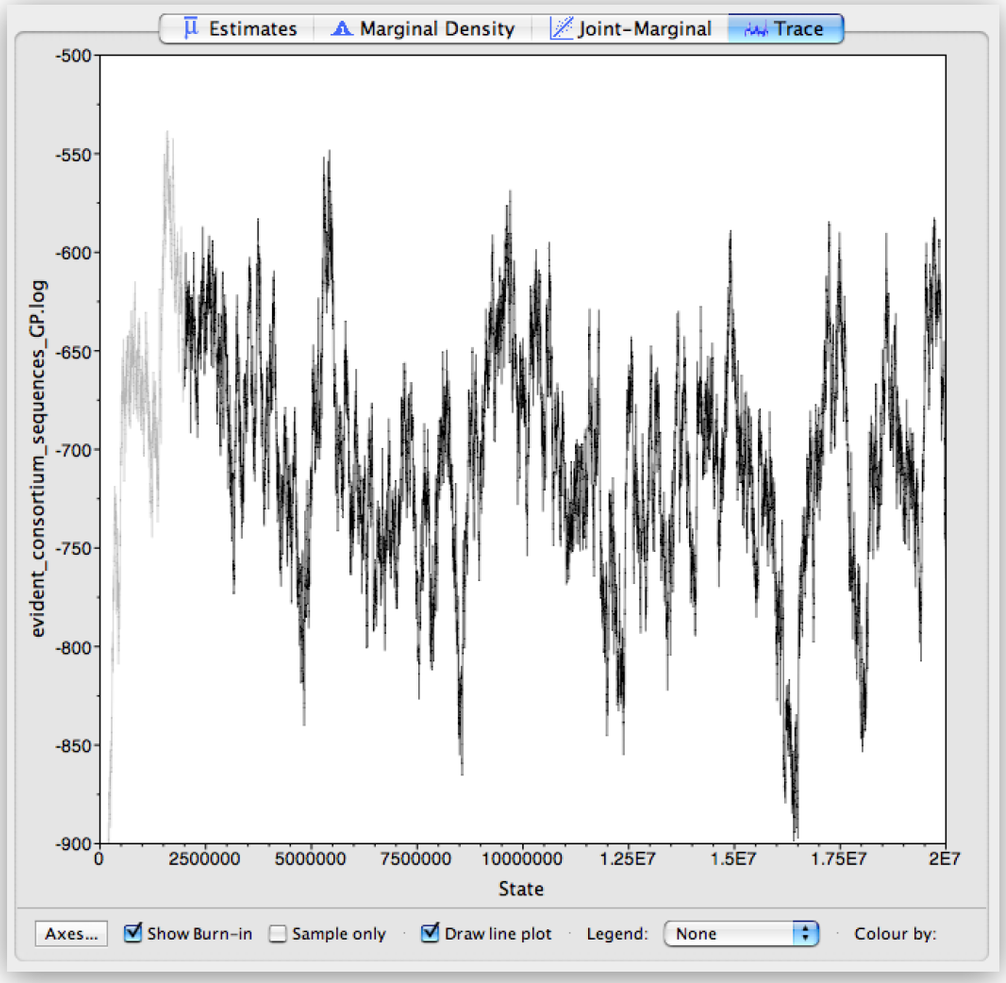


Mixing well! 😄

What often happens in practice…

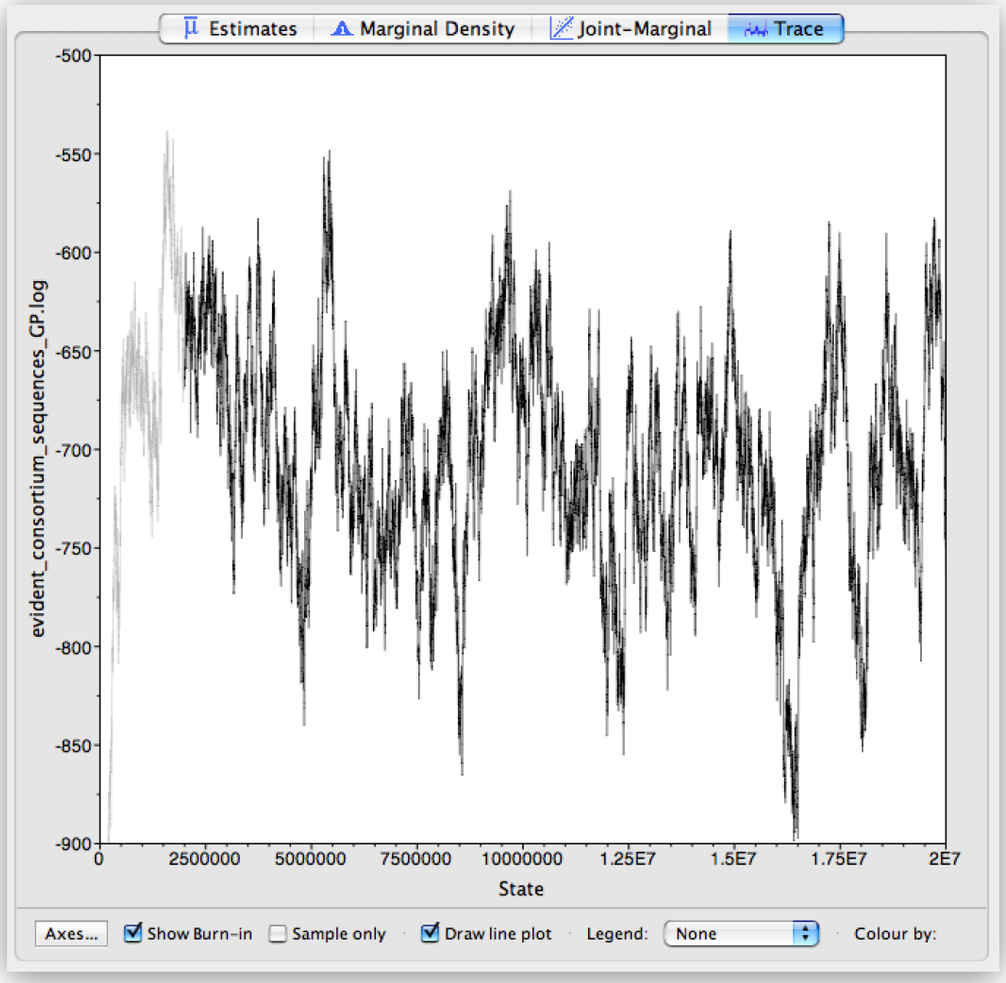
What often happens in practice…

- Not so much…



What often happens in practice…

- Not so much…



Not mixing! 😭

What oft

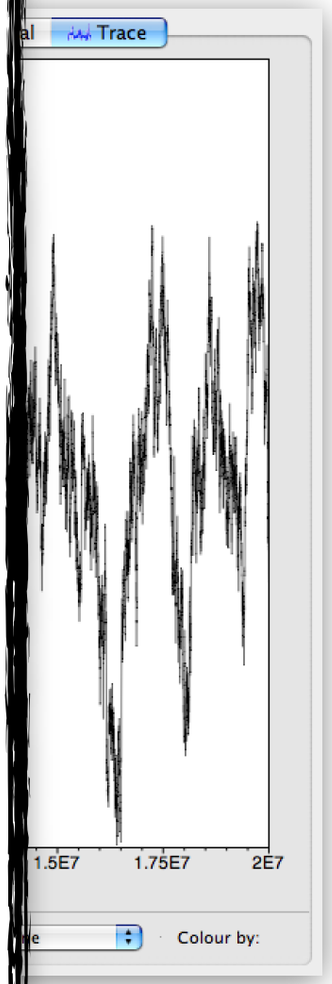
- Not so m

What went wrong?

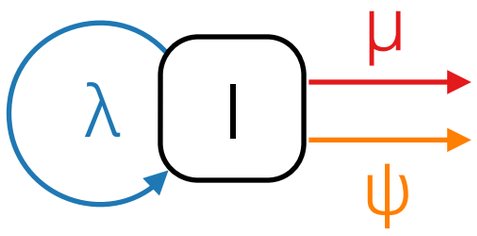
- Recall that MCMC is a Monte Carlo algorithm — it is not **guaranteed** to find the correct solution in finite time!

How can we make it work better?

- Tweak the operators to make good proposals
  (increase operator efficiency)
- Fine tune specifics about the MCMC run
  (sampling frequency, length etc.)
- Poor model choice?
- Poor choice of parameterization?

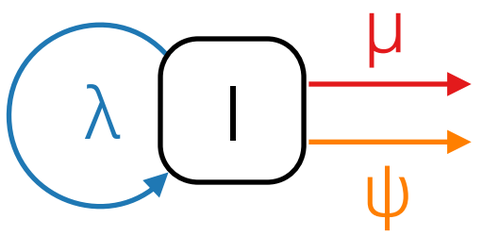


Model Parameterization: Birth-death process



- λ — Birth-rate (speciation / transmission)
- μ — Death-rate (extinction / death)
- ψ — Sampling rate

Model Parameterization: Birth-death process



- λ — Birth-rate (speciation / transmission)
- μ — Death-rate (extinction / death)
- ψ — Sampling rate

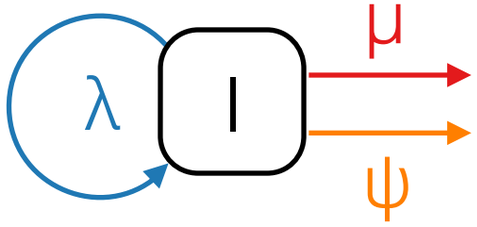
$\lambda \in [0,\infty)$

$\mu \in [0,\infty)$

$\psi \in [0,\infty)$

Priors?

Model Parameterization: Birth-death process



- λ — Birth-rate (speciation / transmission)
- μ — Death-rate (extinction / death)
- ψ — Sampling rate

$\lambda \in [0,\infty)$

$\mu \in [0,\infty)$

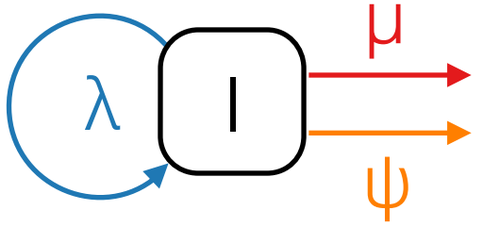
$\psi \in [0,\infty)$

Priors?

Infectious diseases

- $R = \lambda/(\mu+\psi)$ — Reproduction number
  (is it spreading or not?)
- $\delta = \mu + \psi$ — Becoming noninfectious rate
  ($1/\delta$ = infectious period)
- $p = \psi/(\mu + \psi)$ — Sampling proportion

Model Parameterization: Birth-death process



- λ — Birth-rate (speciation / transmission)
- μ — Death-rate (extinction / death)
- ψ — Sampling rate

$\lambda \in [0,\infty)$

$\mu \in [0,\infty)$

$\psi \in [0,\infty)$

Priors?

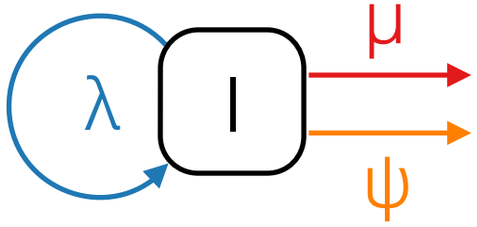
Infectious diseases

- $R = \lambda/(\mu+\psi)$ — Reproduction number
  (is it spreading or not?)
- $\delta = \mu + \psi$ — Becoming noninfectious rate
  ($1/\delta$ = infectious period)
- $p = \psi/(\mu + \psi)$ — Sampling proportion

$R \in [0,\infty)$ $\delta \in [0,\infty)$ $p \in [0,1]$

More natural priors!

Model Parameterization: Birth-death process



- λ — Birth-rate (speciation / transmission)
- μ — Death-rate (extinction / death)
- ψ — Sampling rate

$\lambda \in [0,\infty)$

$\mu \in [0,\infty)$

$\psi \in [0,\infty)$

Priors?

Infectious diseases

- $R = \lambda/(\mu+\psi)$ — Reproduction number (is it spreading or not?)
- $\delta = \mu + \psi$ — Becoming noninfectious rate ($1/\delta$ = infectious period)
- $p = \psi/(\mu + \psi)$ — Sampling proportion

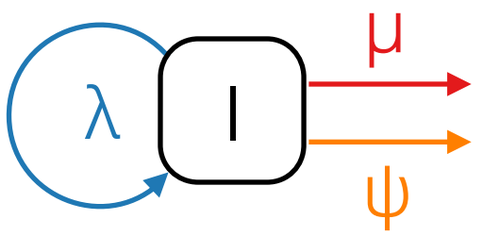
$R \in [0,\infty)$ $\delta \in [0,\infty)$ $p \in [0,1]$

More natural priors!

Speciation

- $\lambda - \mu$ — Growth rate
- μ/λ — Relative death rate
- $p = \psi/(\mu + \psi)$ — Sampling proportion

Model Parameterization: Birth-death process



- λ — Birth-rate (speciation / transmission)
- μ — Death-rate (extinction / death)
- ψ — Sampling rate

$\lambda \in [0,\infty)$

$\mu \in [0,\infty)$

$\psi \in [0,\infty)$

Priors?

Infectious diseases

- $R = \lambda/(\mu+\psi)$ — Reproduction number (is it spreading or not?)
- $\delta = \mu + \psi$ — Becoming noninfectious rate ($1/\delta$ = infectious period)
- $p = \psi/(\mu + \psi)$ — Sampling proportion

$R \in [0,\infty)$ $\delta \in [0,\infty)$ $p \in [0,1]$

More natural priors!

Speciation

- $\lambda - \mu$ — Growth rate
- μ/λ — Relative death rate
- $p = \psi/(\mu + \psi)$ — Sampling proportion

$(\lambda - \mu) \in [0,\infty)$ $\mu/\lambda \in [0,1)$ $p \in [0,1]$

Smaller parameter space!

Summary

- Model consists of a demographic model, a site model and a molecular clock model

- Combine priors for model parameters with the likelihood to get the posterior

- Using MCMC we can efficiently approximate the joint posterior of the model

- By marginalising we get posterior distributions of parameters of interest while integrating out uncertainty in nuisance parameters

- Need a good set of operators to propose new steps

- Efficient inference may need some fine-tuning!
  (or it may not converge to the posterior distribution in a reasonable time)