Dated Bayesian phylogenetic analysis to infer parameters of epidemiological interest

Sam Hong
Rega Institute of Medical Research
KU Leuven - Belgium
samuel.hong@kuleuven.be

2023 CADDE Workshop - São Paulo, Brazil

Some applied questions addressed by viral genomic epidemiology

Tuesday	What virus is causing the outbreak?
Wednesday	What species are involved in virus transmission?
Thursday	When did the outbreak begin?
Thursday	How many introductions have there been?
Friday	Where did the outbreak begin?
Friday	How fast is the virus evolving?
Friday	How rapidly is the virus transmitting?
Friday	What factors drive an outbreak?

Ask questions, participate, network and join the Workshop's WhatsApp groups ☺

Workshop Materials and Detailed Agenda:

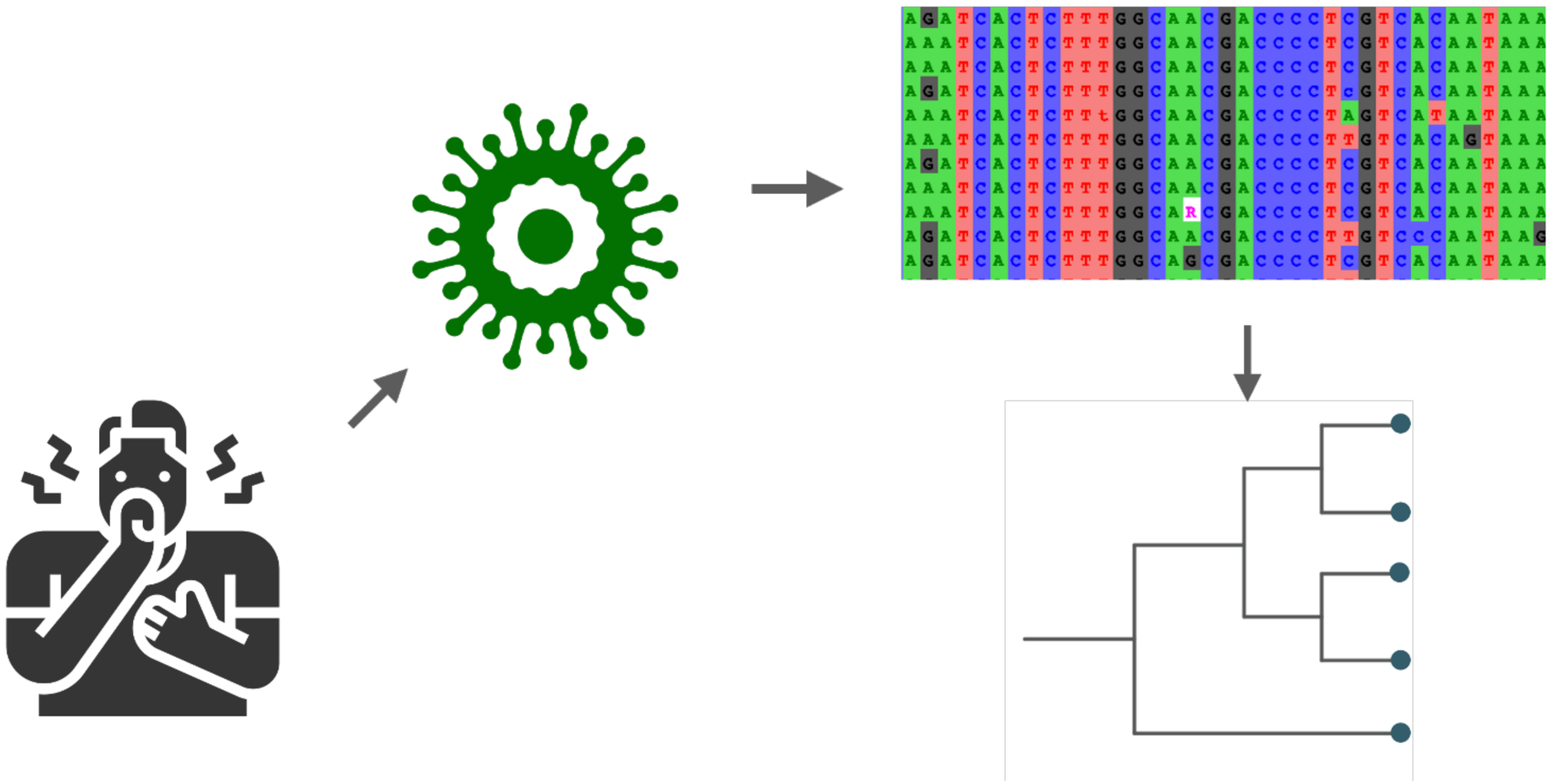
Some applied questions addressed by viral genomic epidemiology

Tuesday	What virus is causing the outbreak?
Wednesday	What species are involved in virus transmission?
Thursday	When did the outbreak begin?
Thursday	How many introductions have there been?
Friday	Where did the outbreak begin?
Friday	How fast is the virus evolving?
Friday	How rapidly is the virus transmitting?
Friday	What factors drive an outbreak?

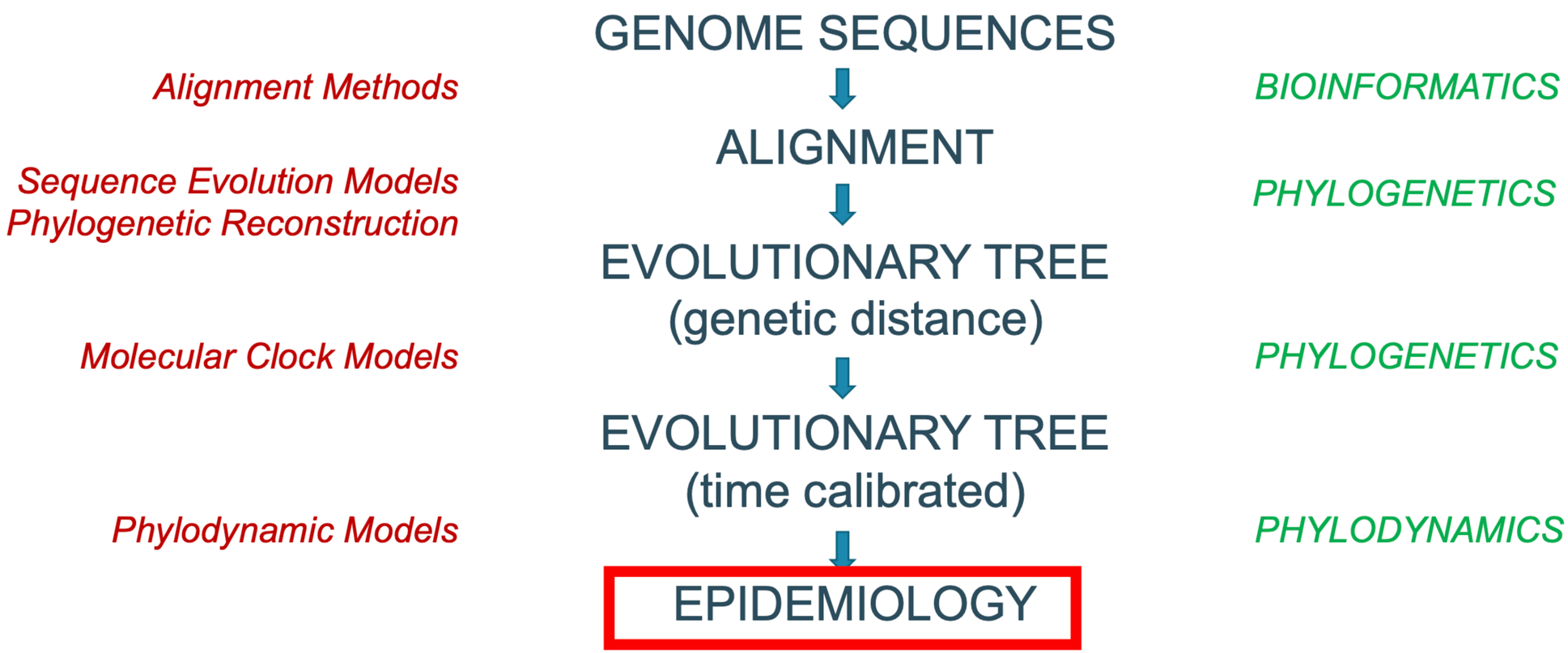
Ask questions, participate, network and join the Workshop's WhatsApp groups ☺

Workshop Materials and Detailed Agenda:

Genomic Epidemiology



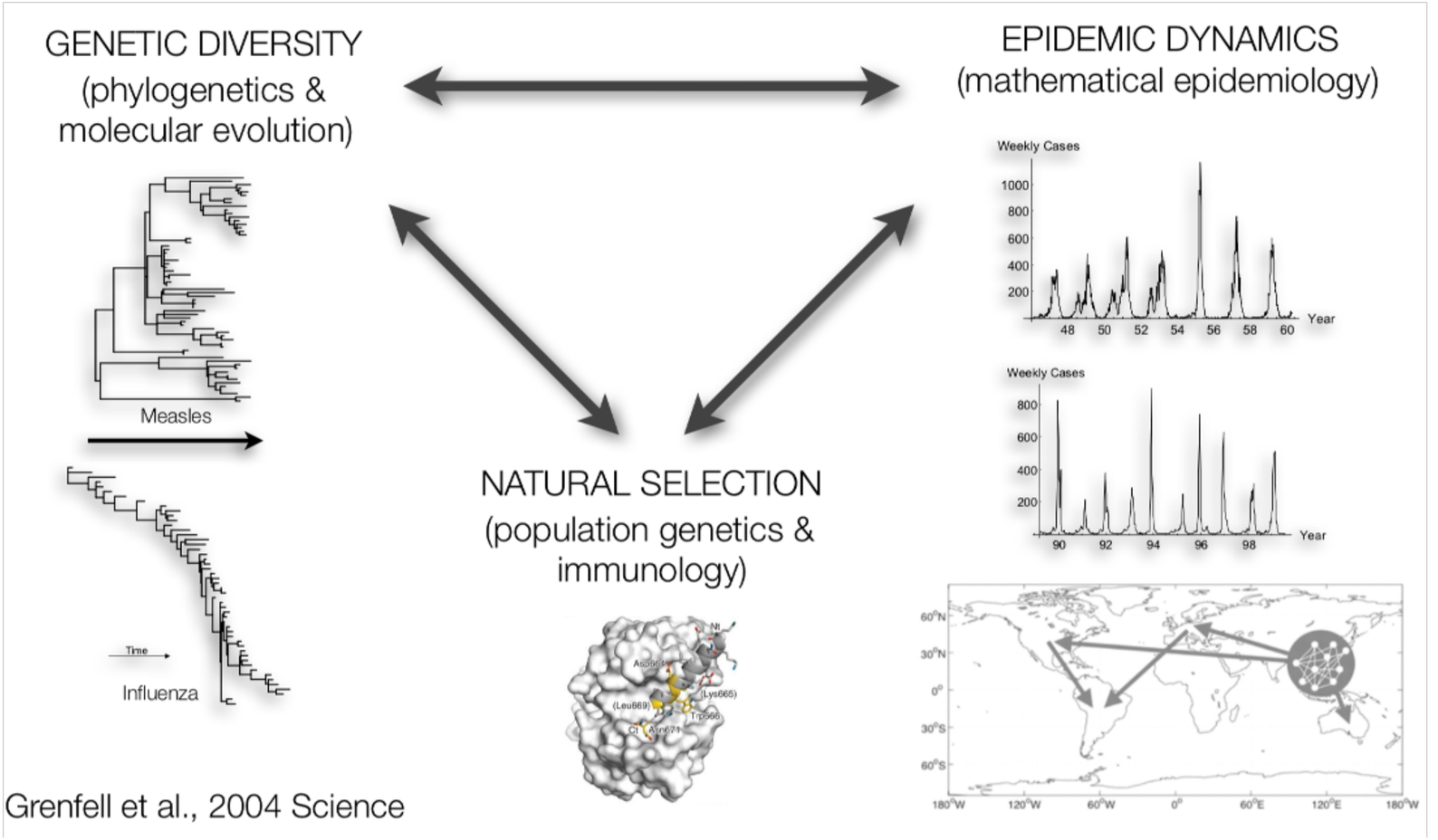
From sequence to epidemiology



Adapted from Philippe Lemey

Phylodynamics

In many infectious pathogens, evolutionary and epidemiological dynamics occur within the same time scale.



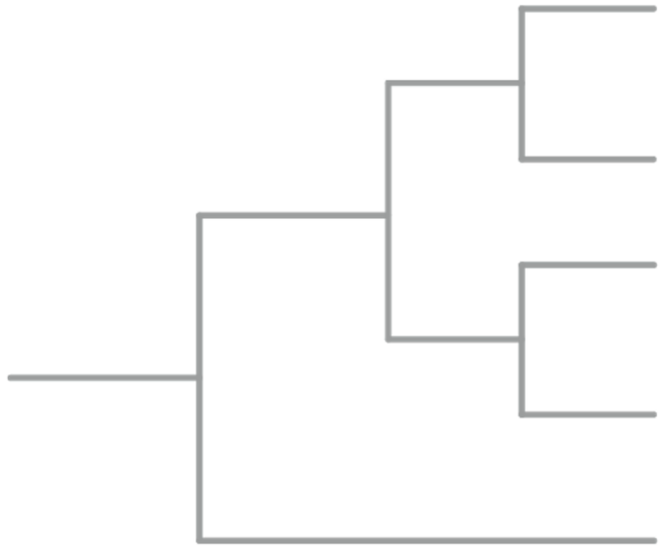
Grenfell et al., 2004 Science

Bayesian Phylodynamics

Maximum Likelihood

$$\arg\max_{\theta} L(\theta)$$

Likelihood: $L(\theta) \propto P(Data|\theta)$

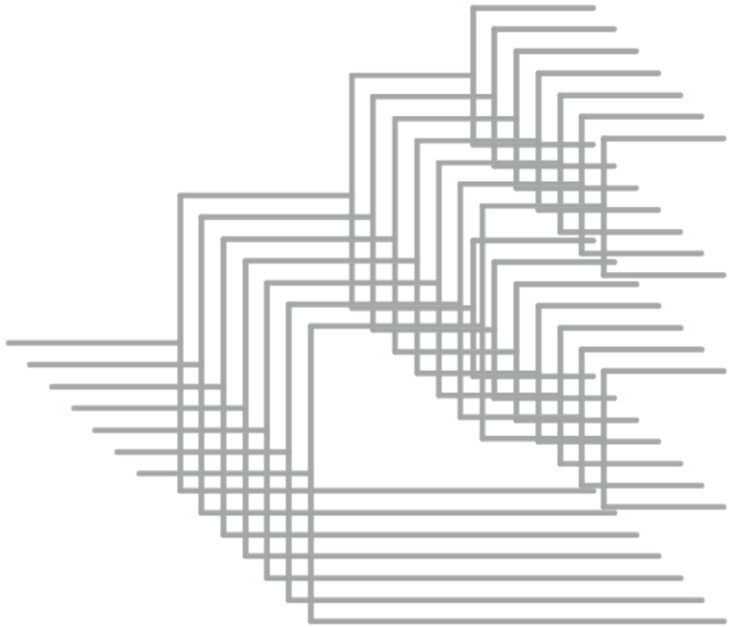


Bayesian

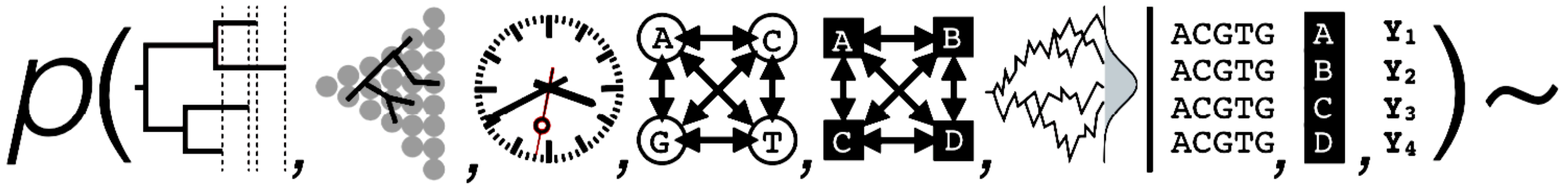
Posterior

Prior

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$



Bayesian phylodynamics with BEAST



Phylogeny

Population

Clock

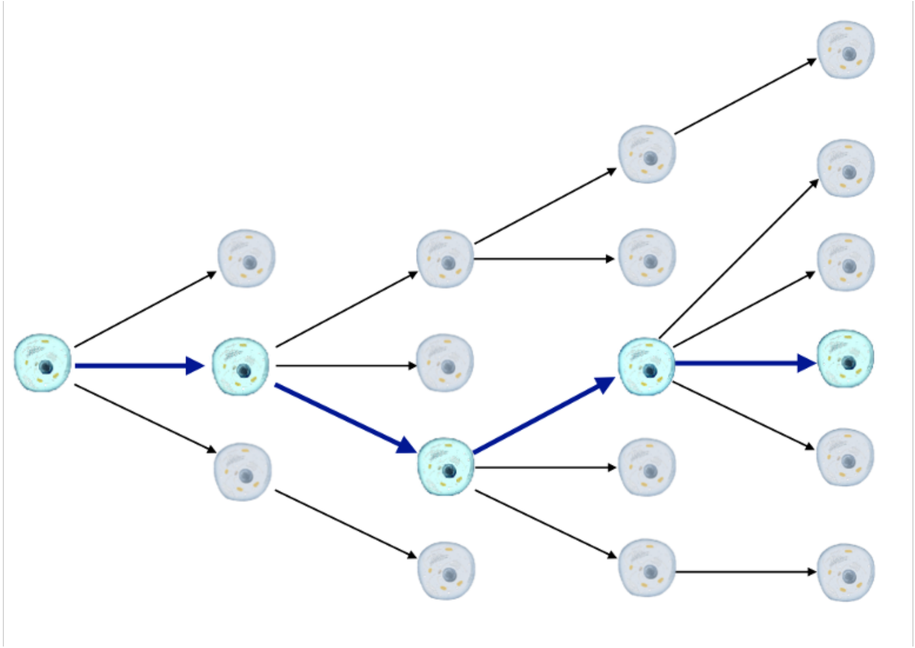
Substitution

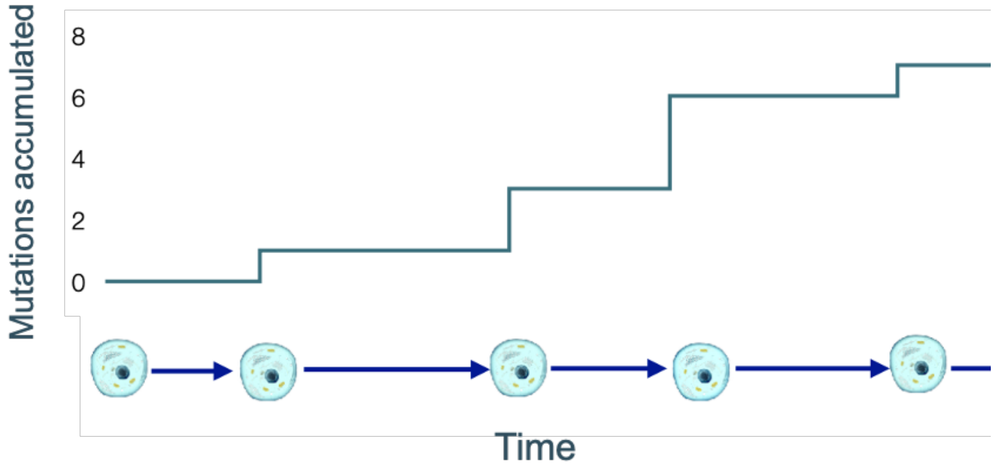
Traits

Sequence/Trait Data

Adapted from Philippe Lemey

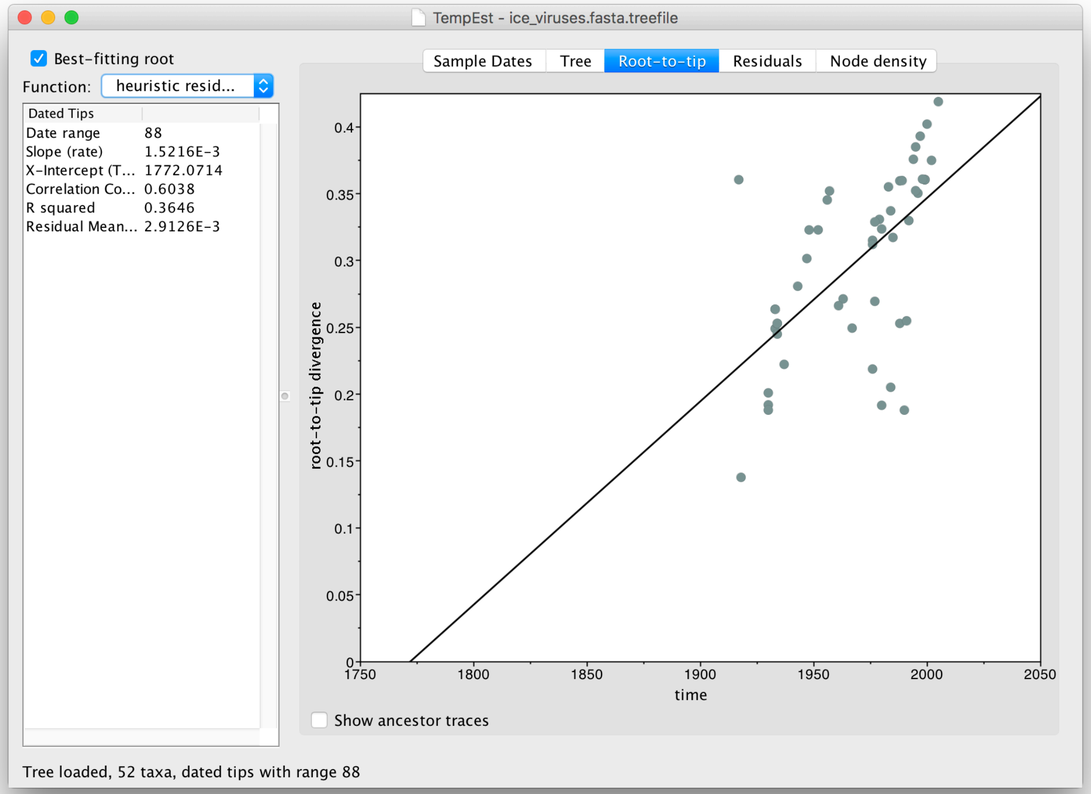
Learning from trees: molecular clocks



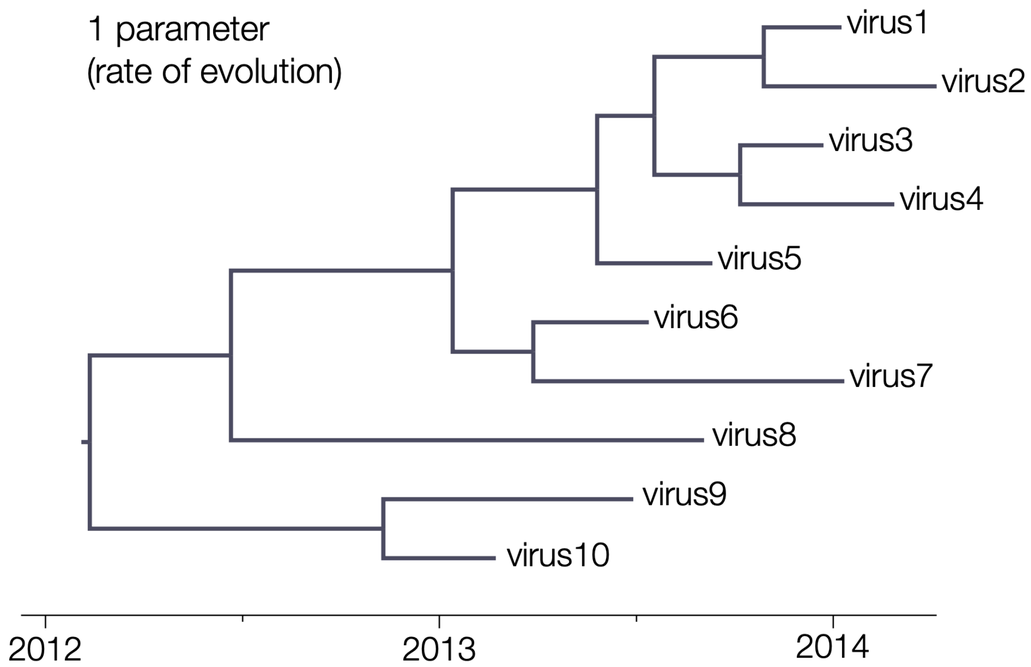


Adapted from Andrew Rambaut

Clock models

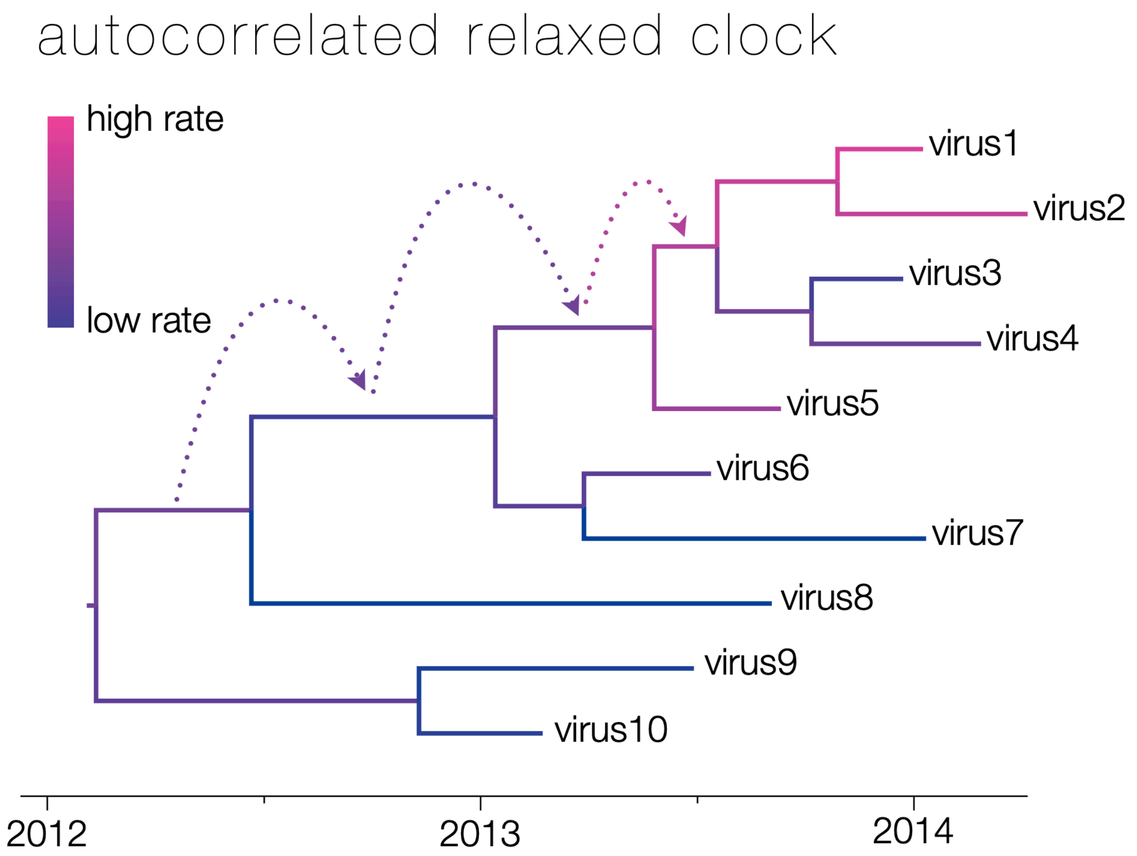


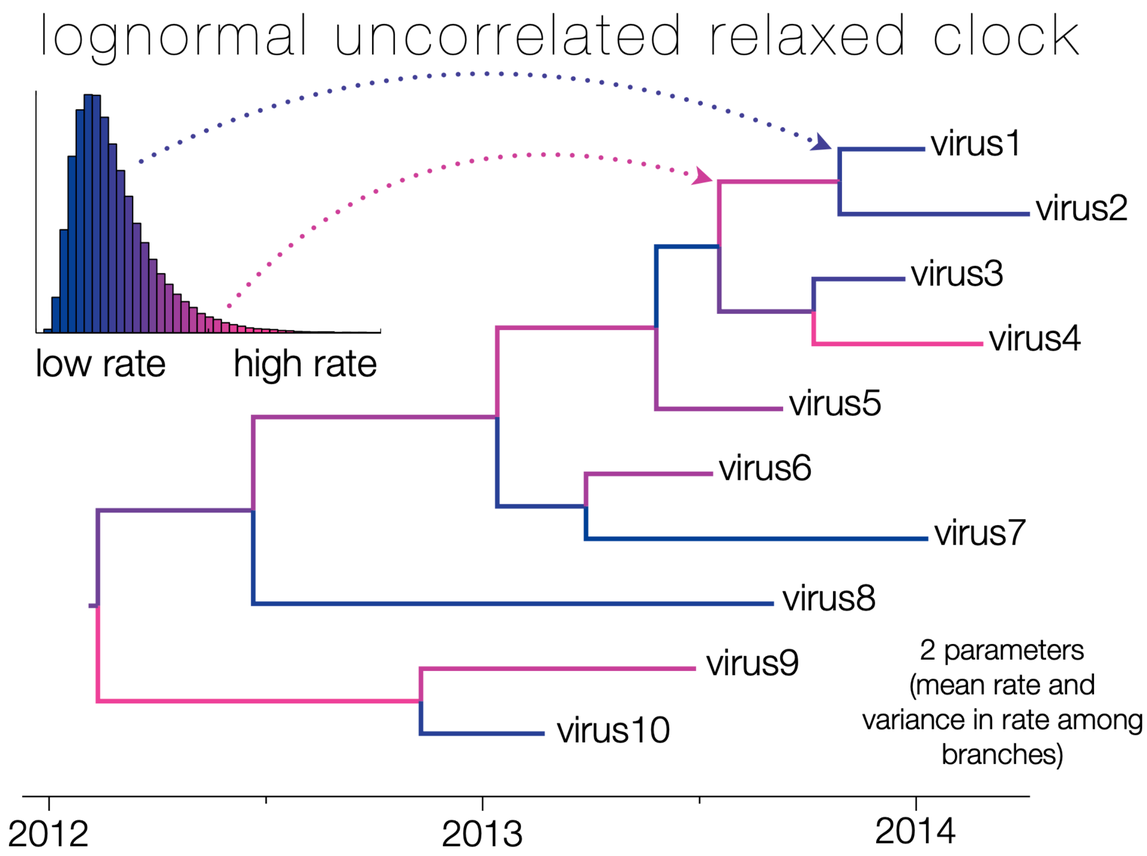
'strict' molecular clock



Adapted from Andrew Rambaut

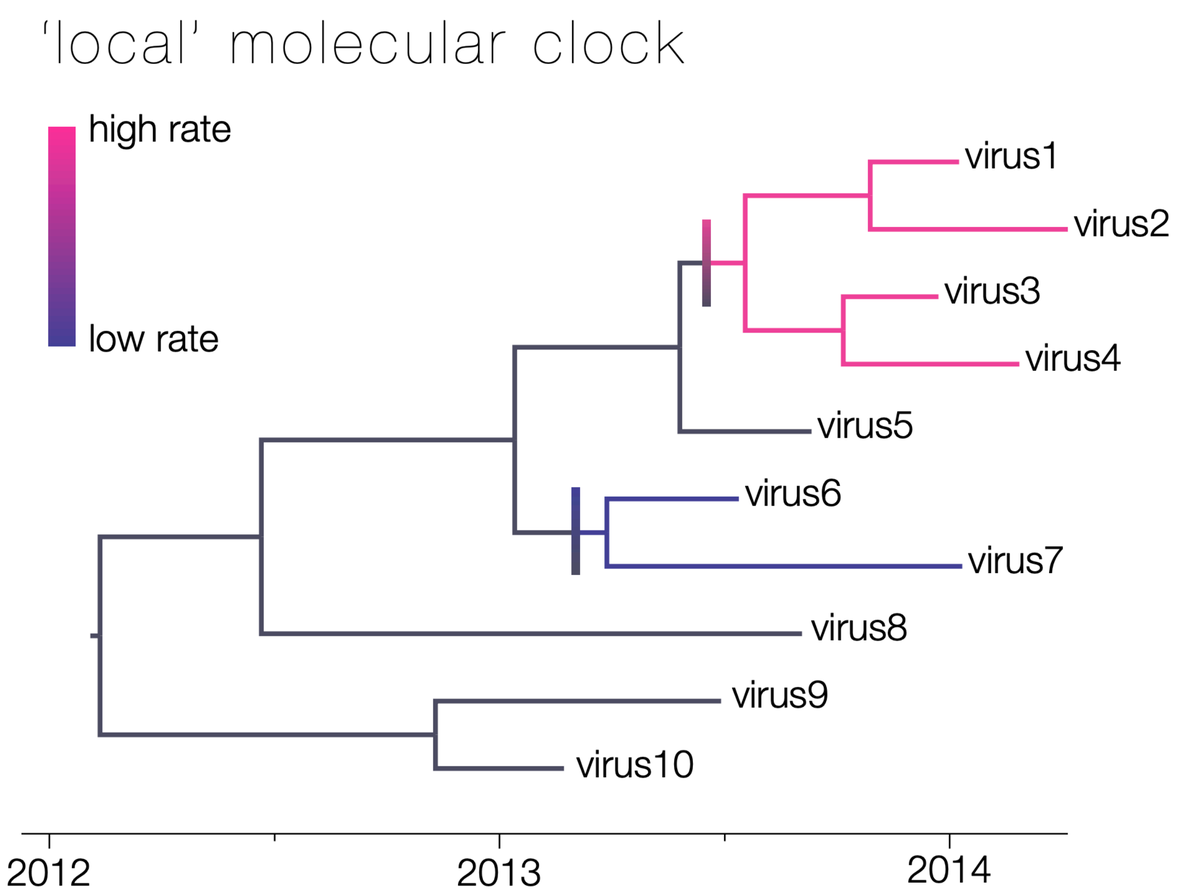
Relaxed Clocks

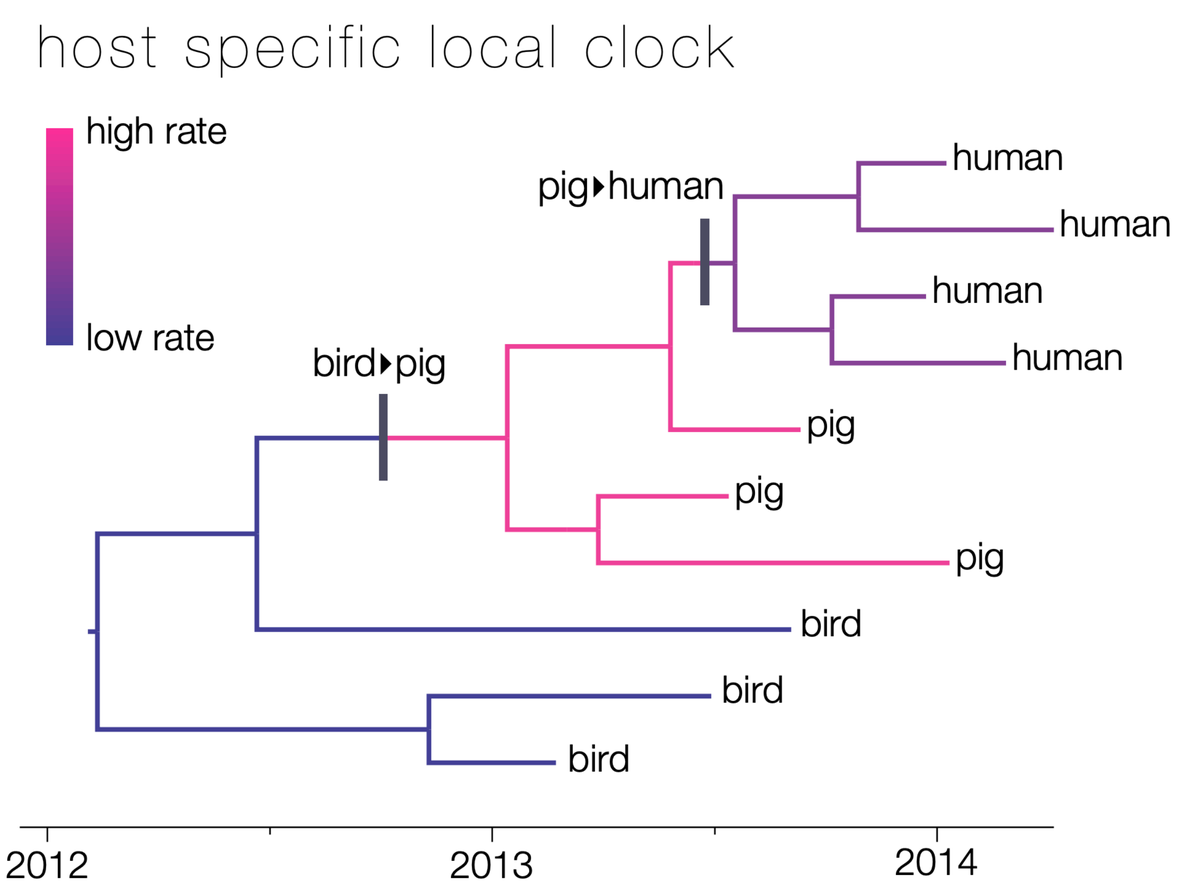




Adapted from Andrew Rambaut

Local clocks

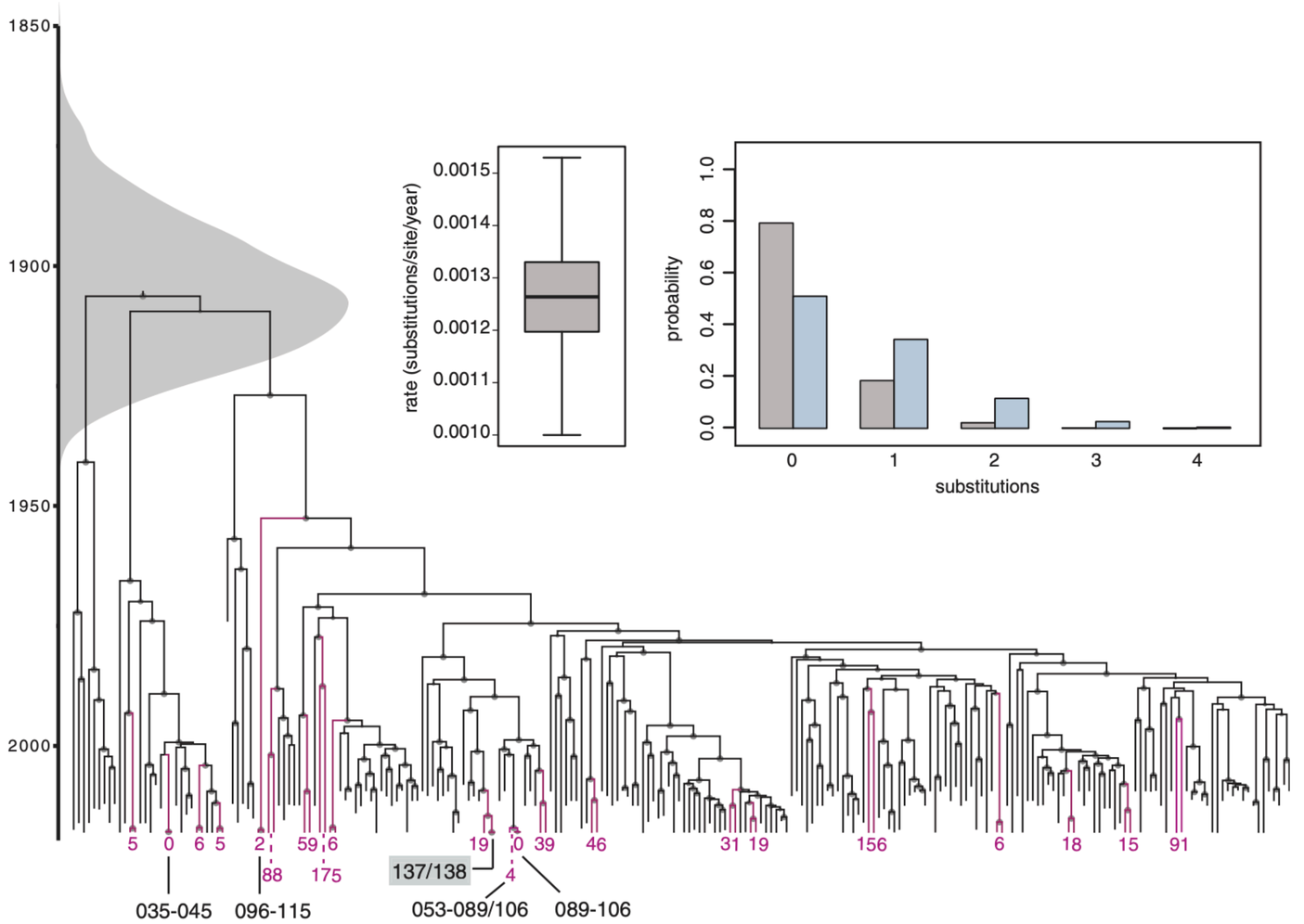




Adapted from Andrew Rambaut

Time calibration applications

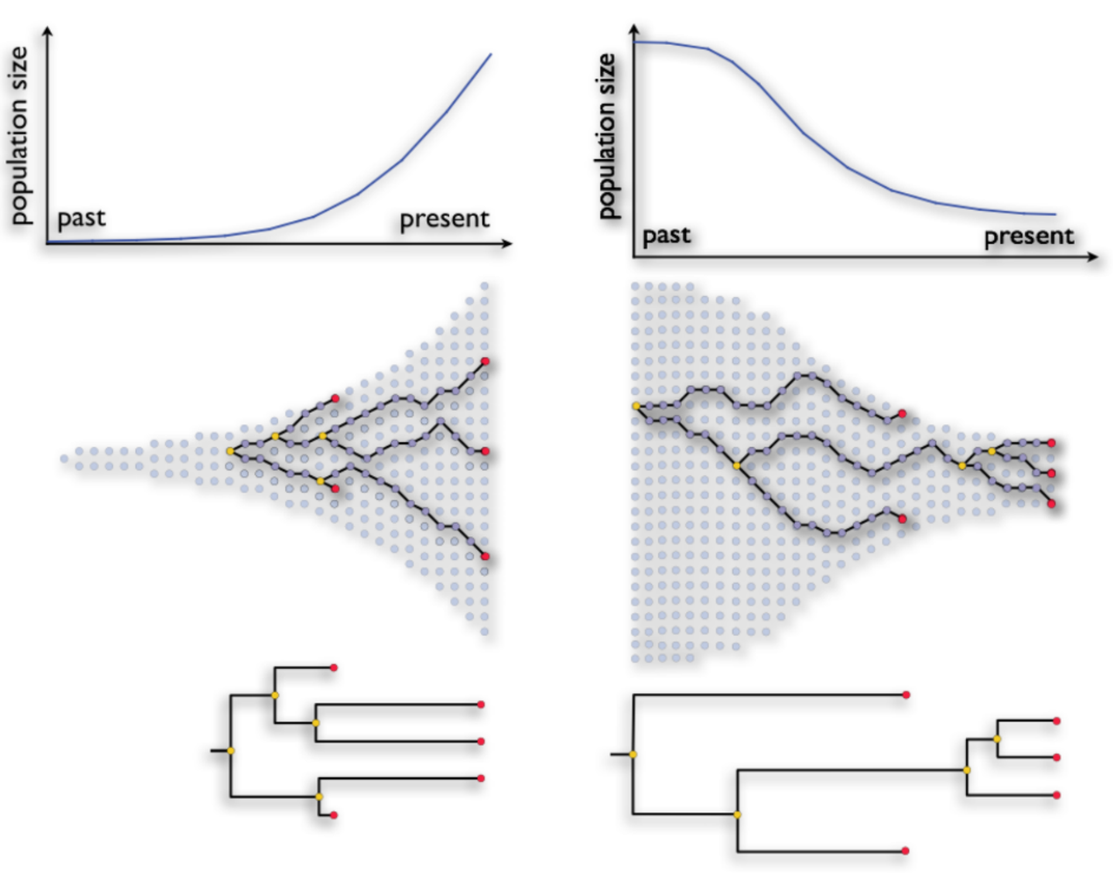
- Transforming branch lengths into units of time allow us to answer epidemiological questions such as :
  - When did the outbreak begin? (TMRCA)
  - How fast is the virus evolving? (Clock rate)
  - What is the mode of transmission?



Kafetzopoulou et al. (2019)

Learning from trees: Coalescent

- Coalescent theory: Statistical model of the probability of two samples to merge (coalesce) after a given time
- We can learn about changes in population size from the shape of a tree:
    - The larger the population size, the longer it will take for two lineages to coalesce
- Application: Calculate the exponential growth rate of an epidemic
    - How rapidly is the virus spreading? (Exponential growth)



Adapted from Volodymyr Minin

Coalescent: Exponential Growth

$$N_e(t) = N_0 e^{rt}$$

Growth rate

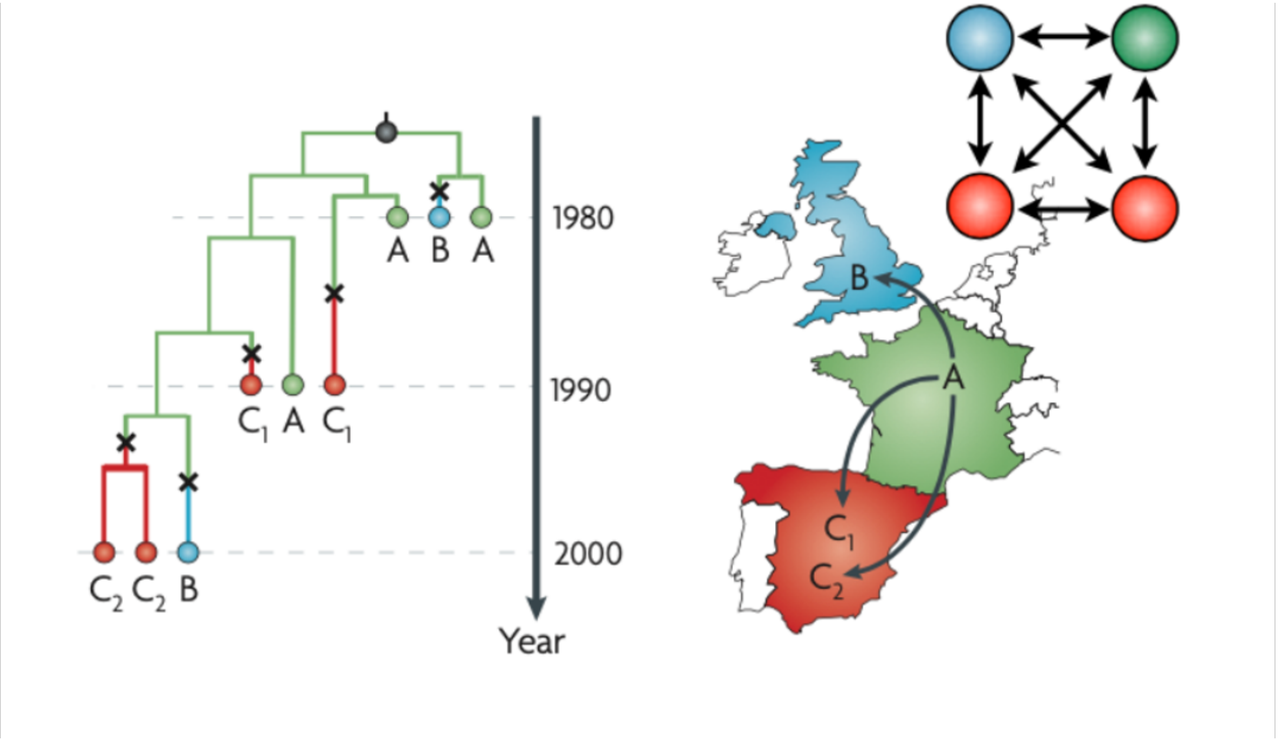
Effective population size

Initial population size

Doubling time: $N_e(t_{double}) = 2N_0 = N_0 e^{rt_{double}}$ → (Solve for t) $t_{double} = \frac{Log(2)}{\text{Growth rate}}$

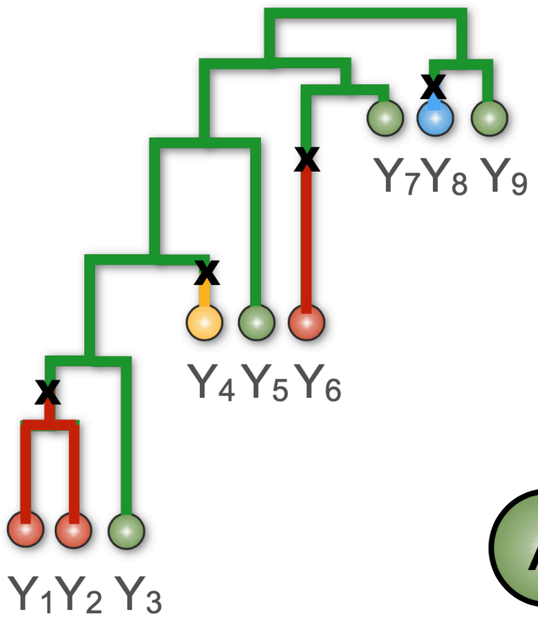
Learning from trees: Phylogeography

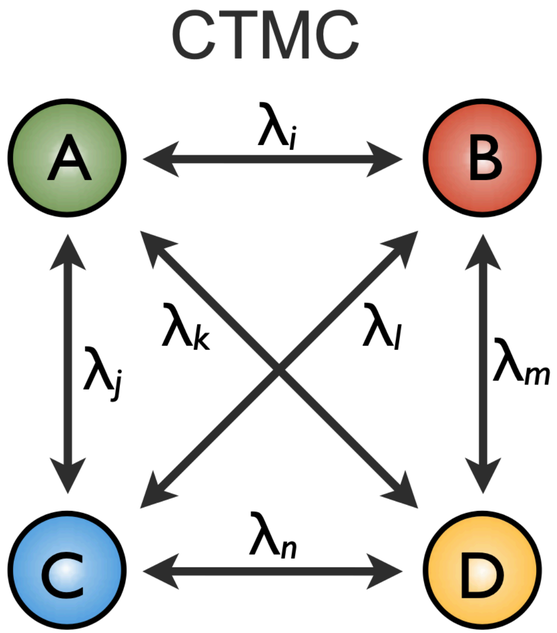
- Discrete trait analysis: model discrete locations as a trait coloring the tree
- Reconstruct probability of jumping from one location to another and expected number of jumps
- Answer questions such as:
  - How many introductions?
  - Where did the outbreak begin?
  - What factors drive spread?



Discrete phylogeography

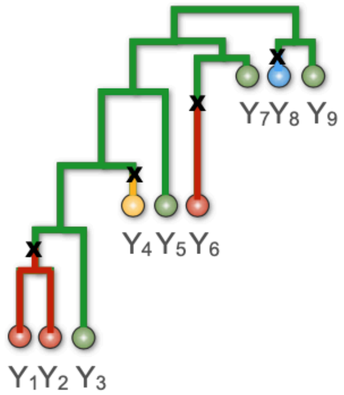
- Model spatial spread as a random process characterized by a rate matrix that describes the probability of transitioning to a location given the current location
    - Analogous to nucleotide substitution models
- Transition rates can be symmetric or asymmetric
- Use rates to reconstruct the location of internal nodes
    - How many introductions?
    - What is the origin?

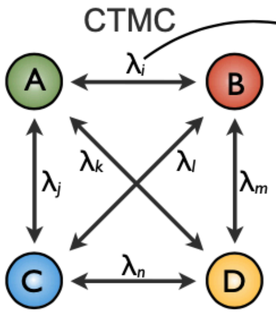




Adapted from Philippe Lemey

Factors that drive spread?

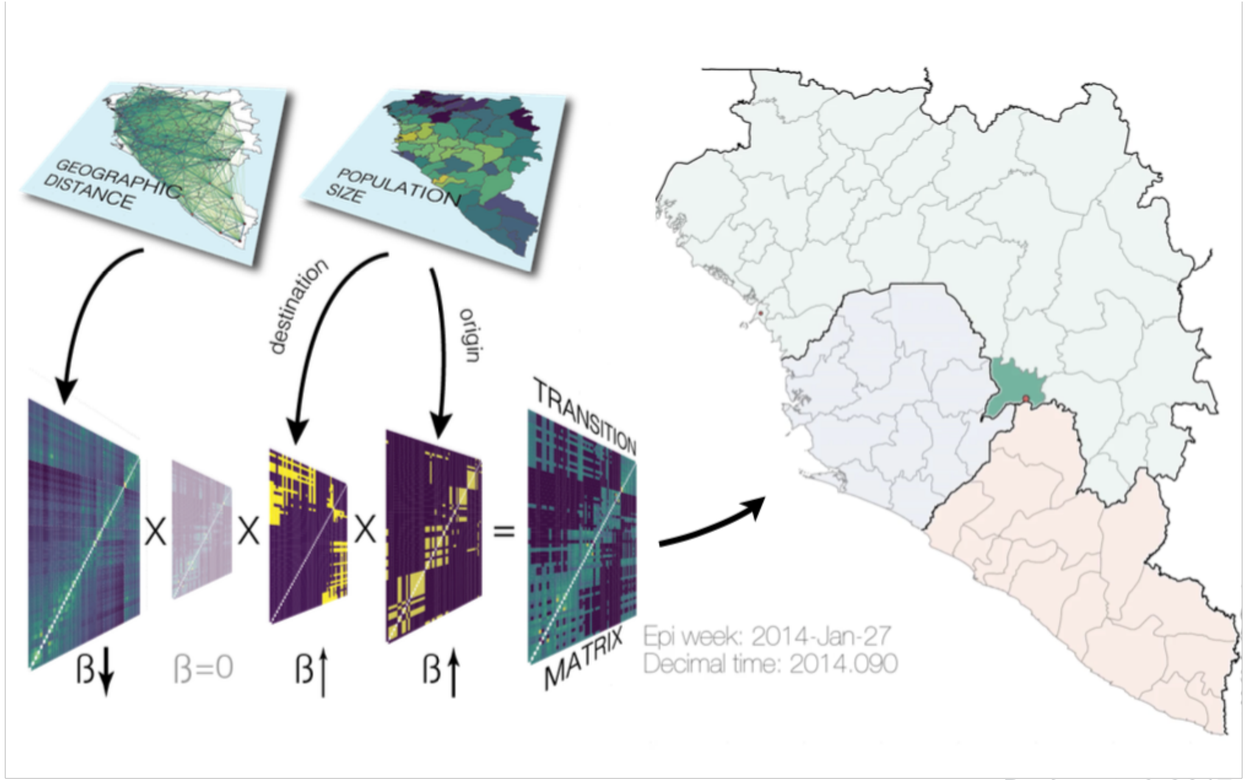




$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{i,j,1} + \beta_2 \delta_2 x_{i,j,2} + \ldots + \beta_P \delta_P x_{i,j,P}$$

Predictors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_P)$

$\mathbf{x}_p = (x_{1,2,p}, \ldots x_{K-1,K,p})$



Dudas et al. 2017

Adapted from Philippe Lemey

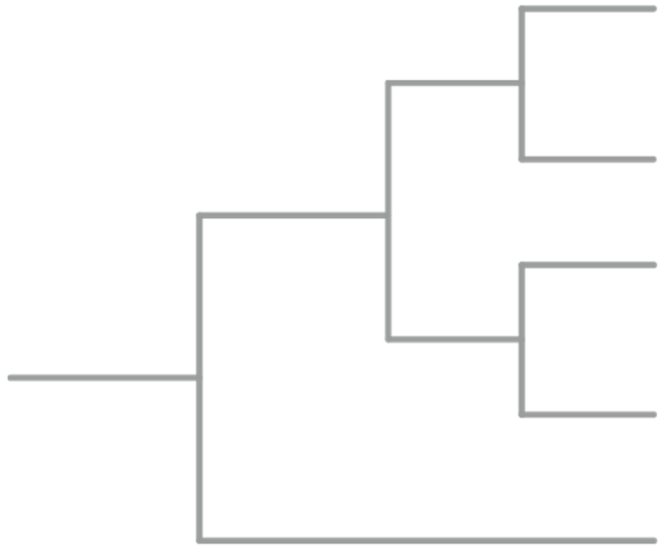
How does Bayesian Phylogenetic Inference work?

Bayesian Phylodynamics

Maximum Likelihood

$$\arg\max_{\theta} L(\theta)$$

Likelihood: $L(\theta) \propto P(Data|\theta)$

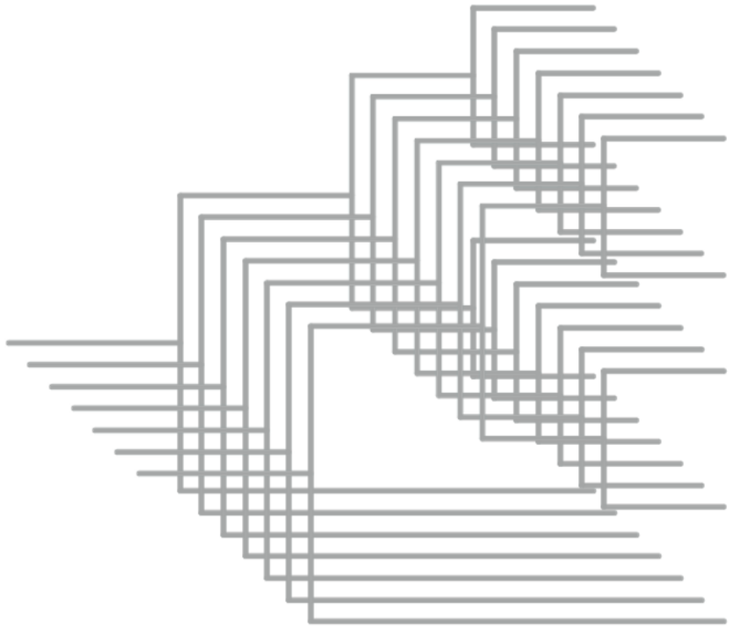


Bayesian

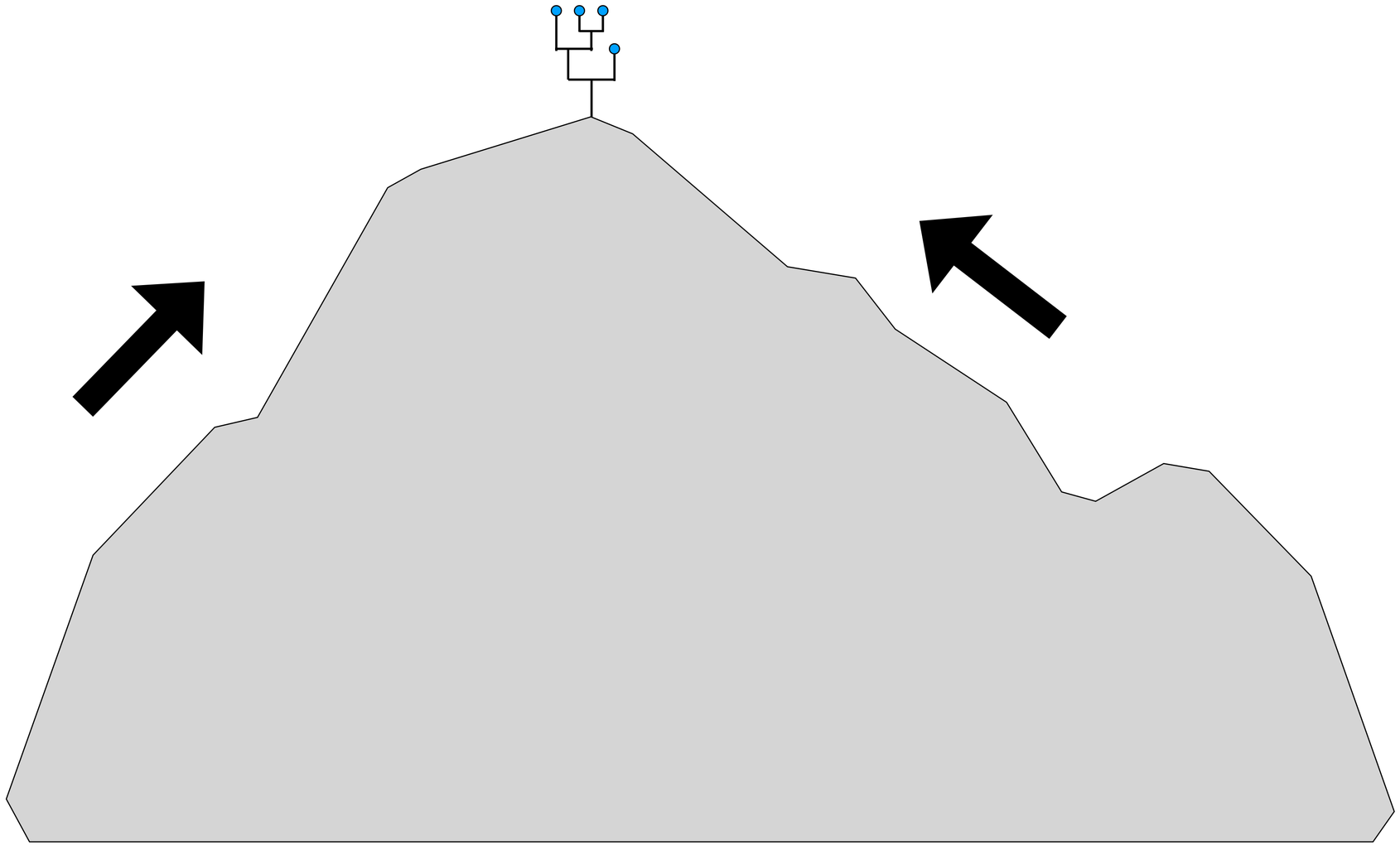
Posterior

Prior

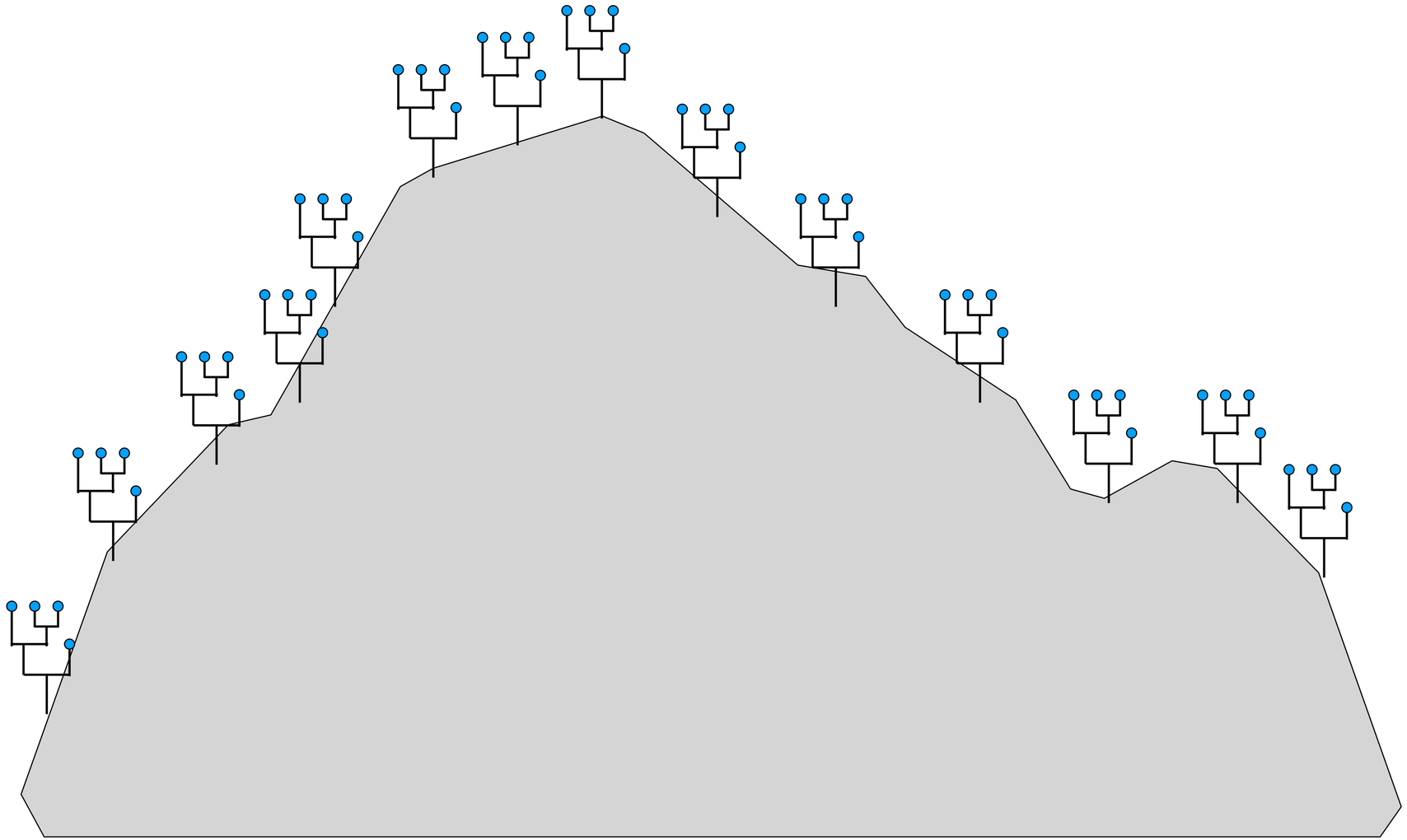
$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

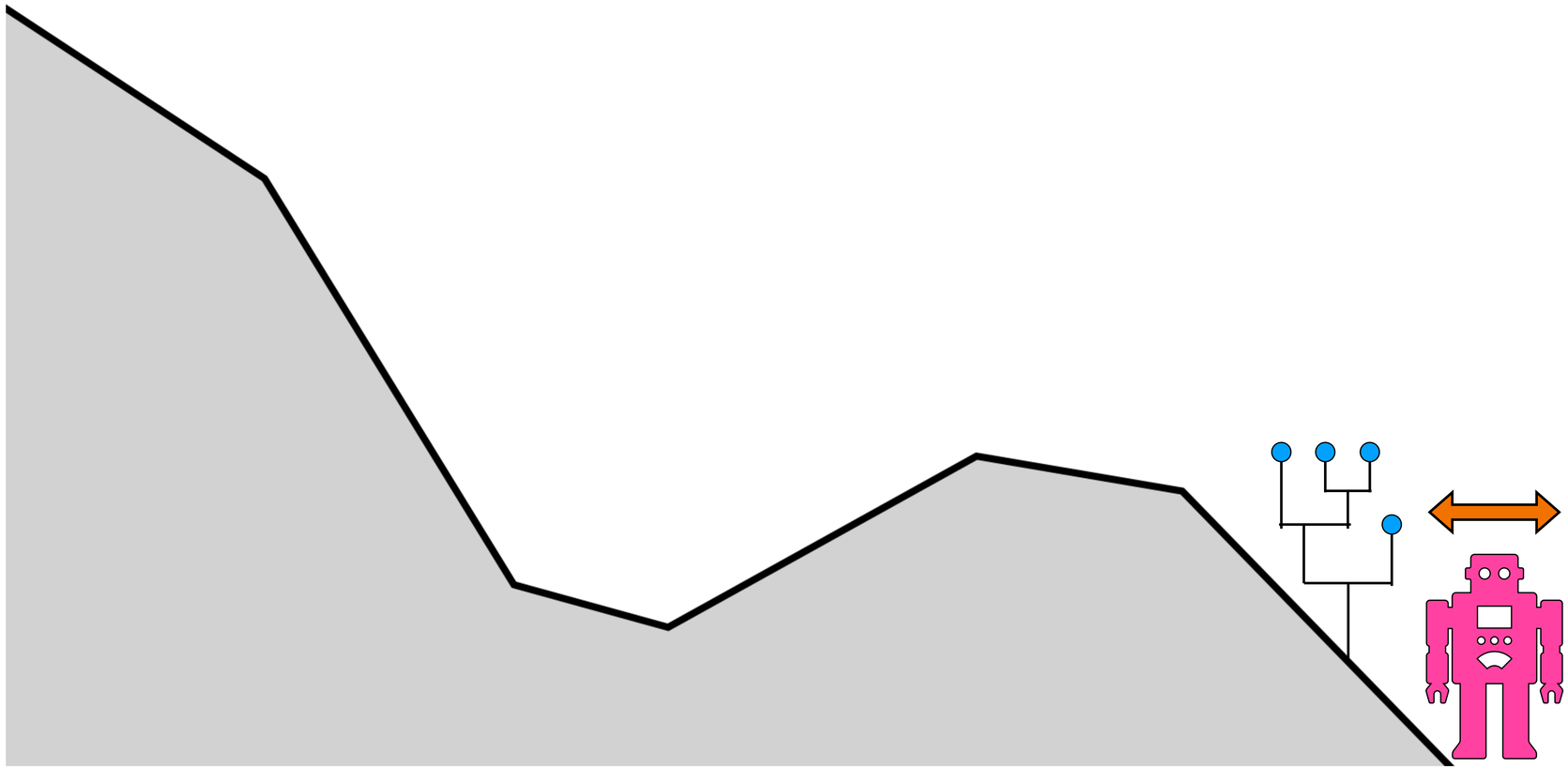


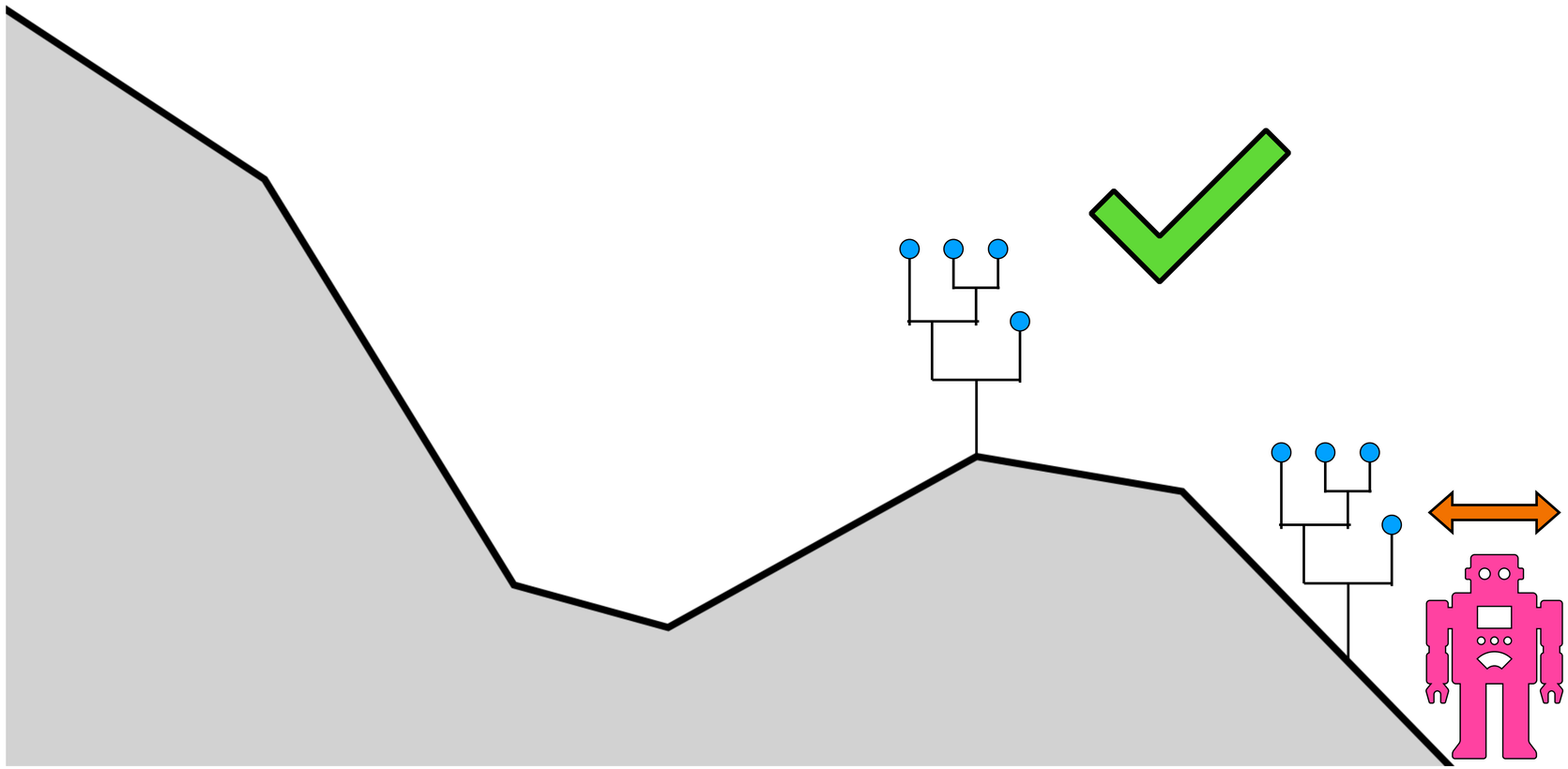
Maximum Likelihood

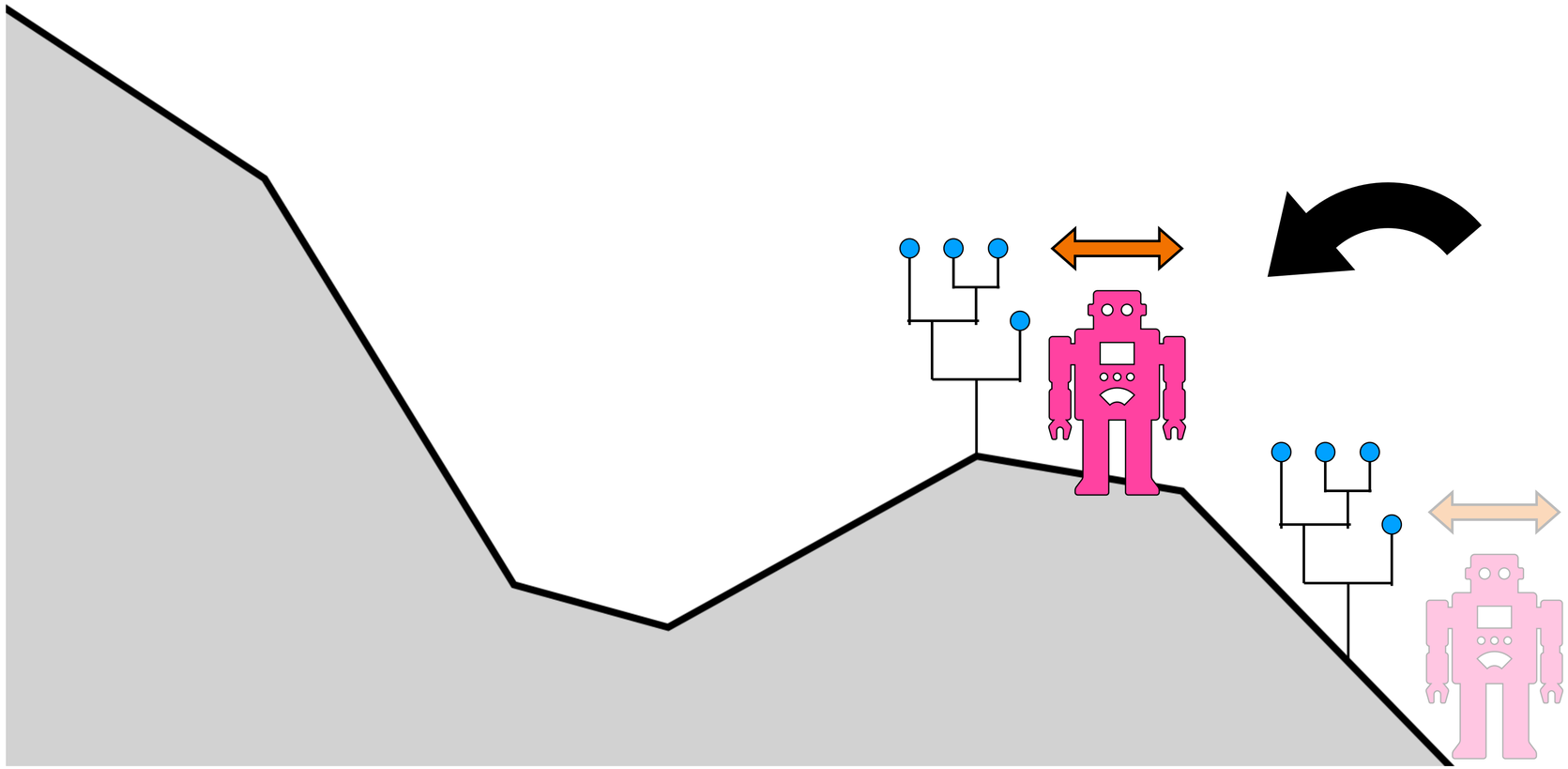


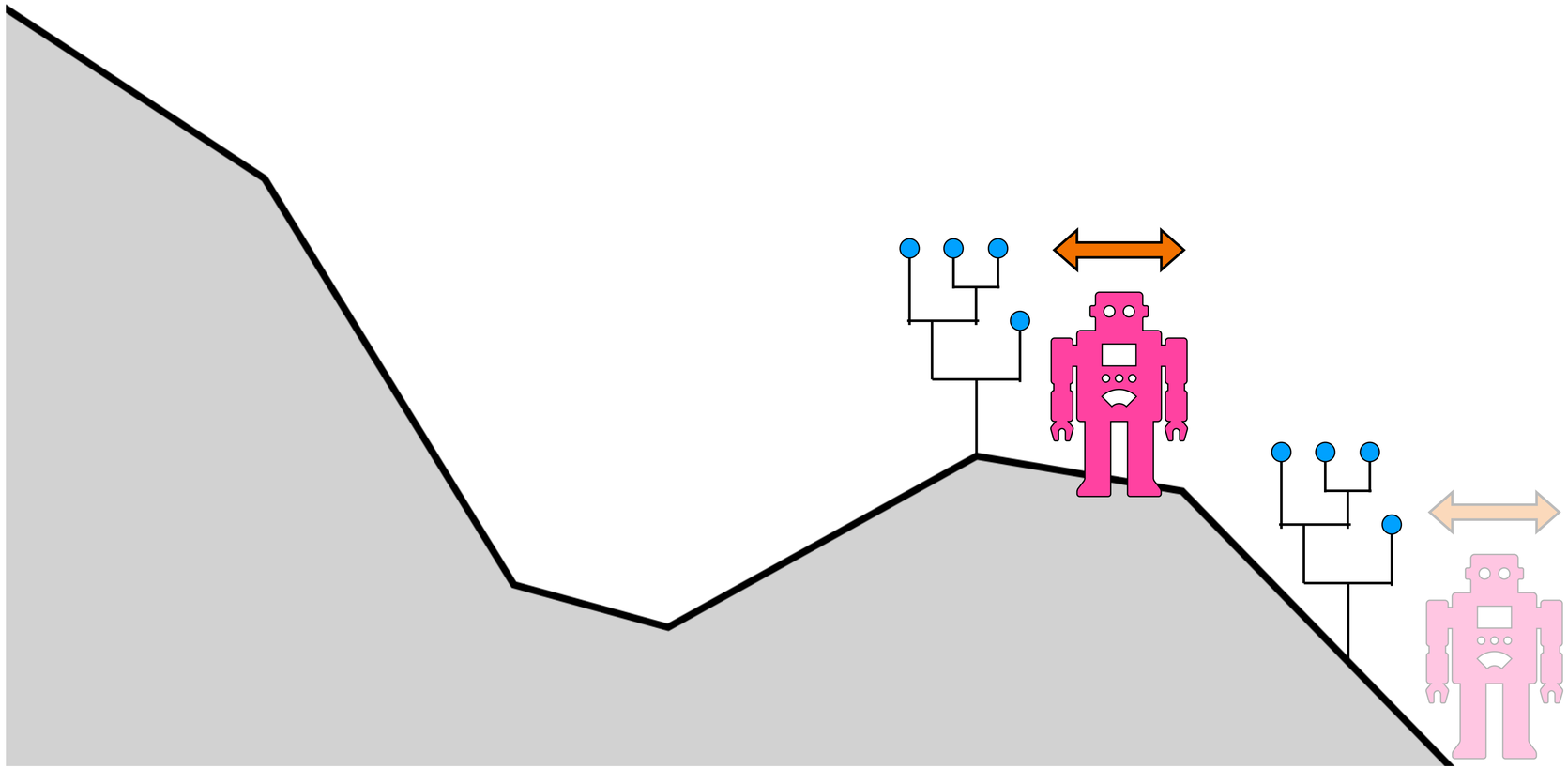
Bayesian

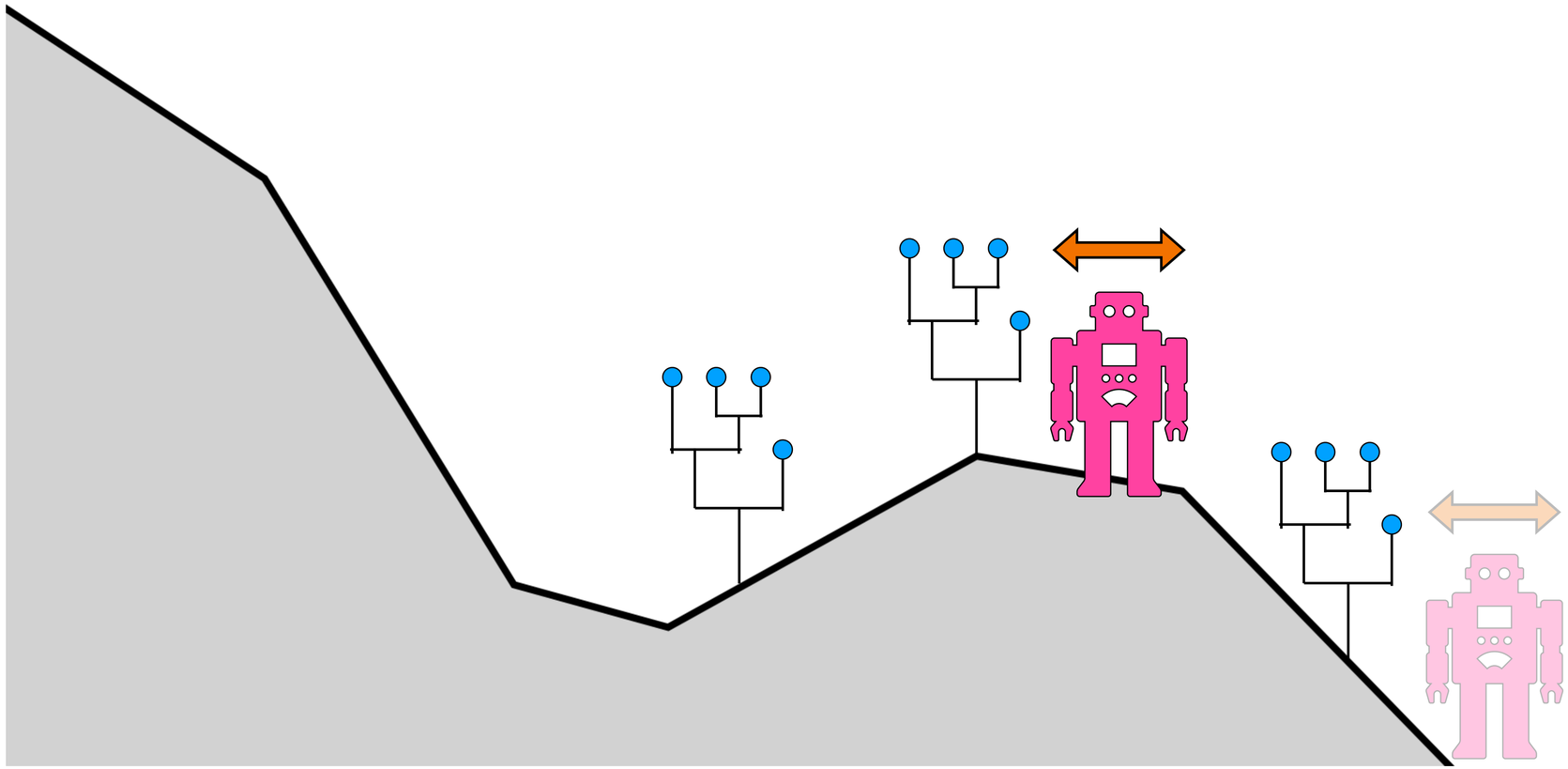


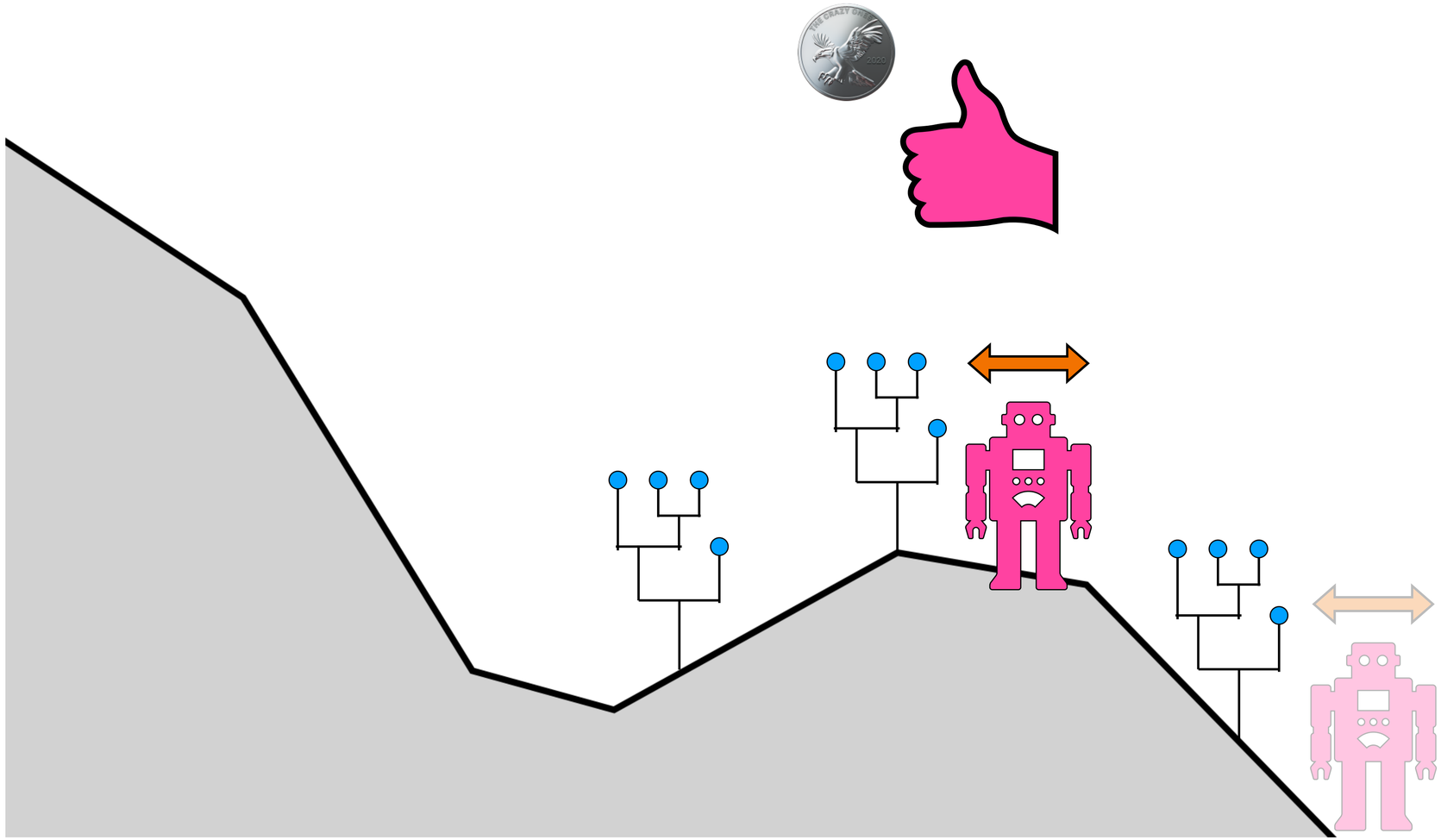


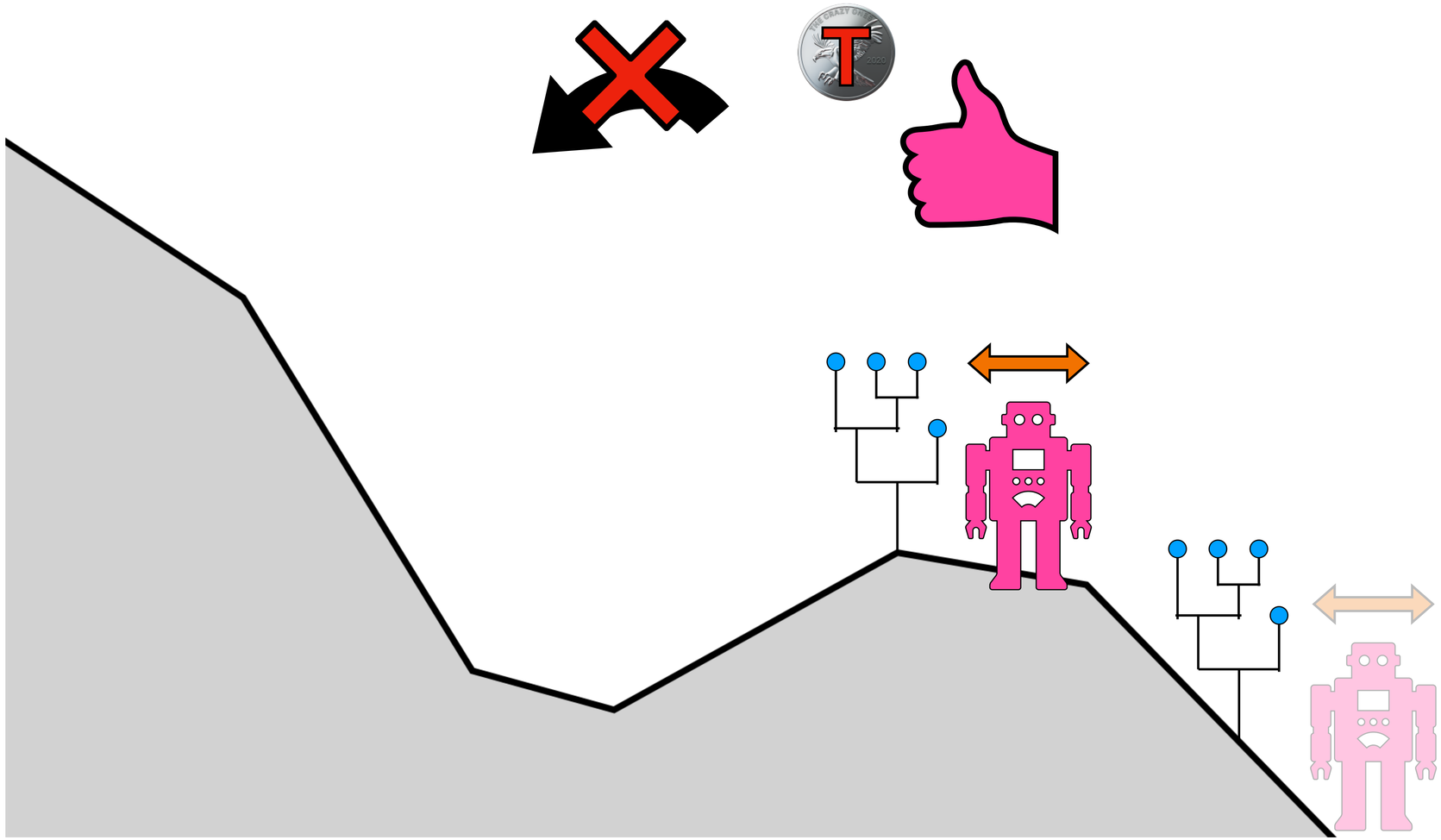


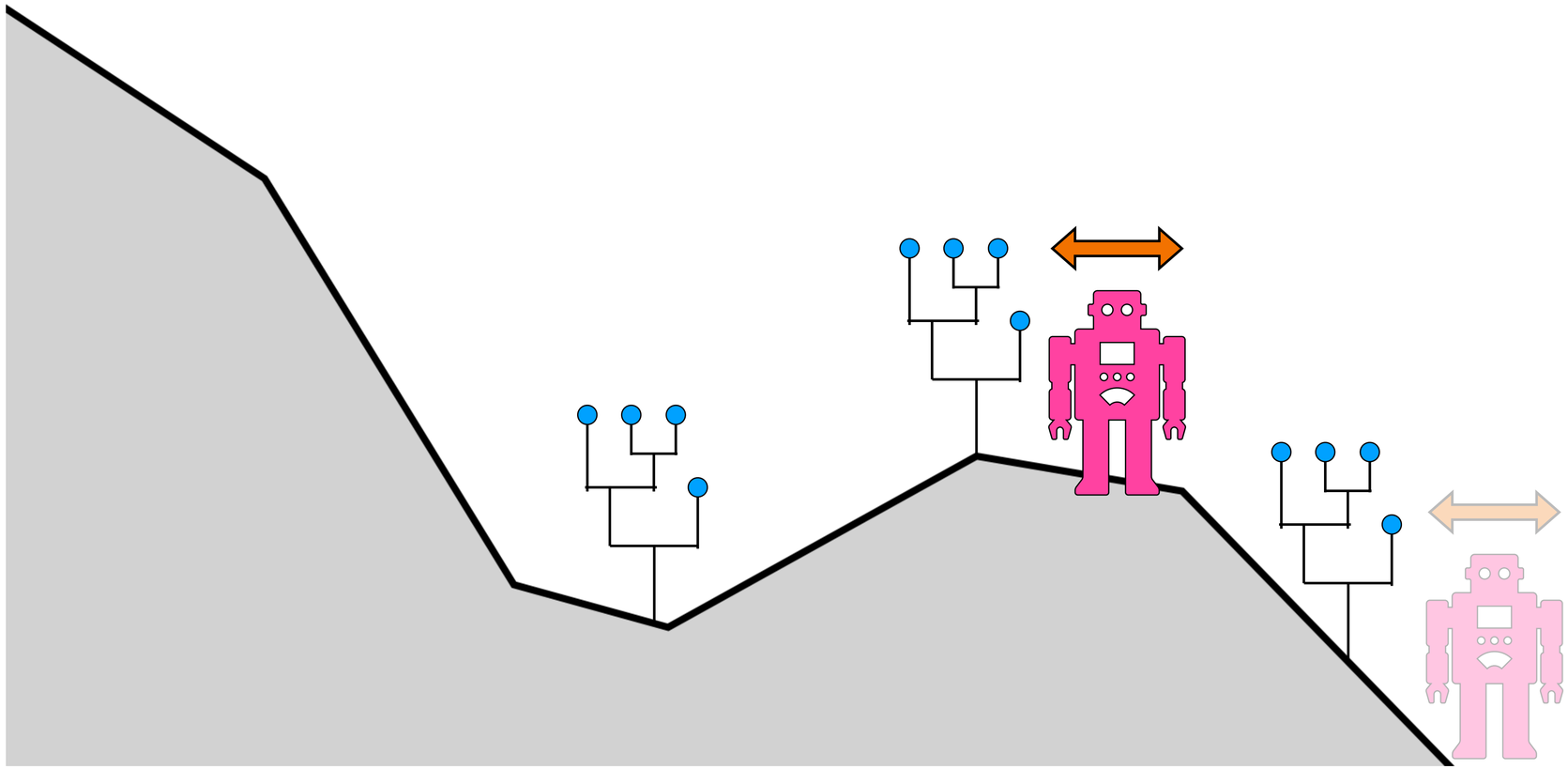


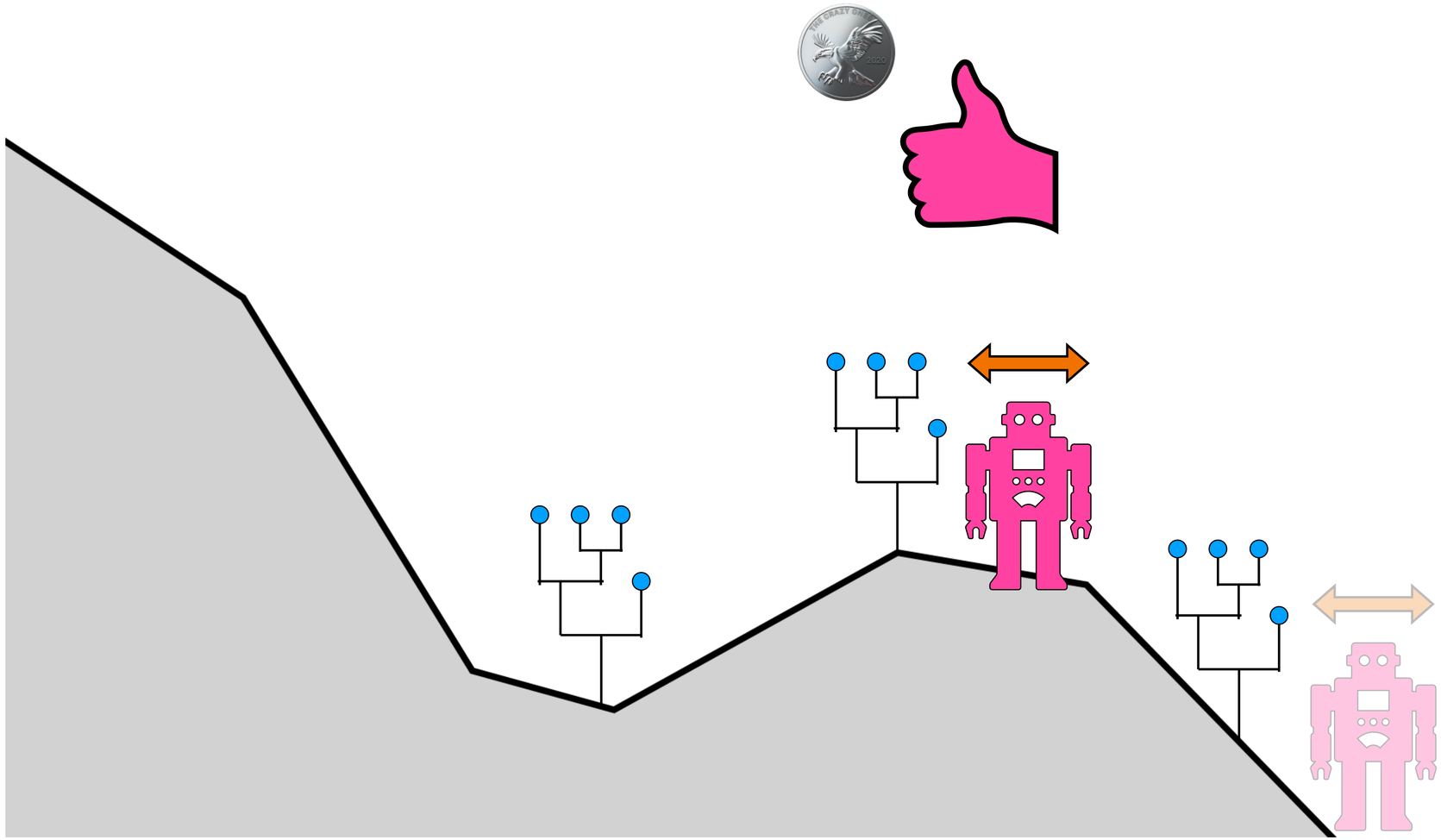


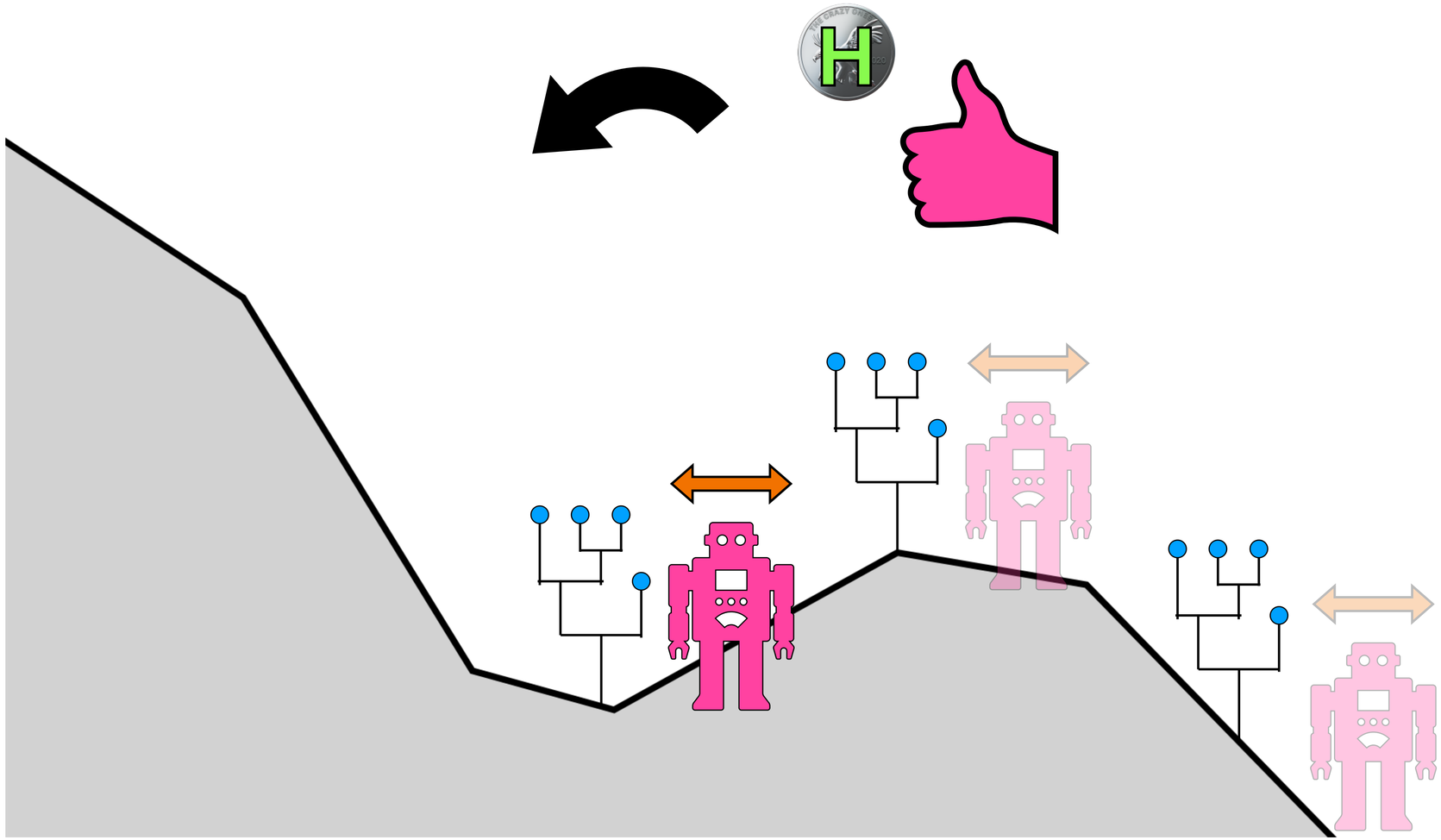


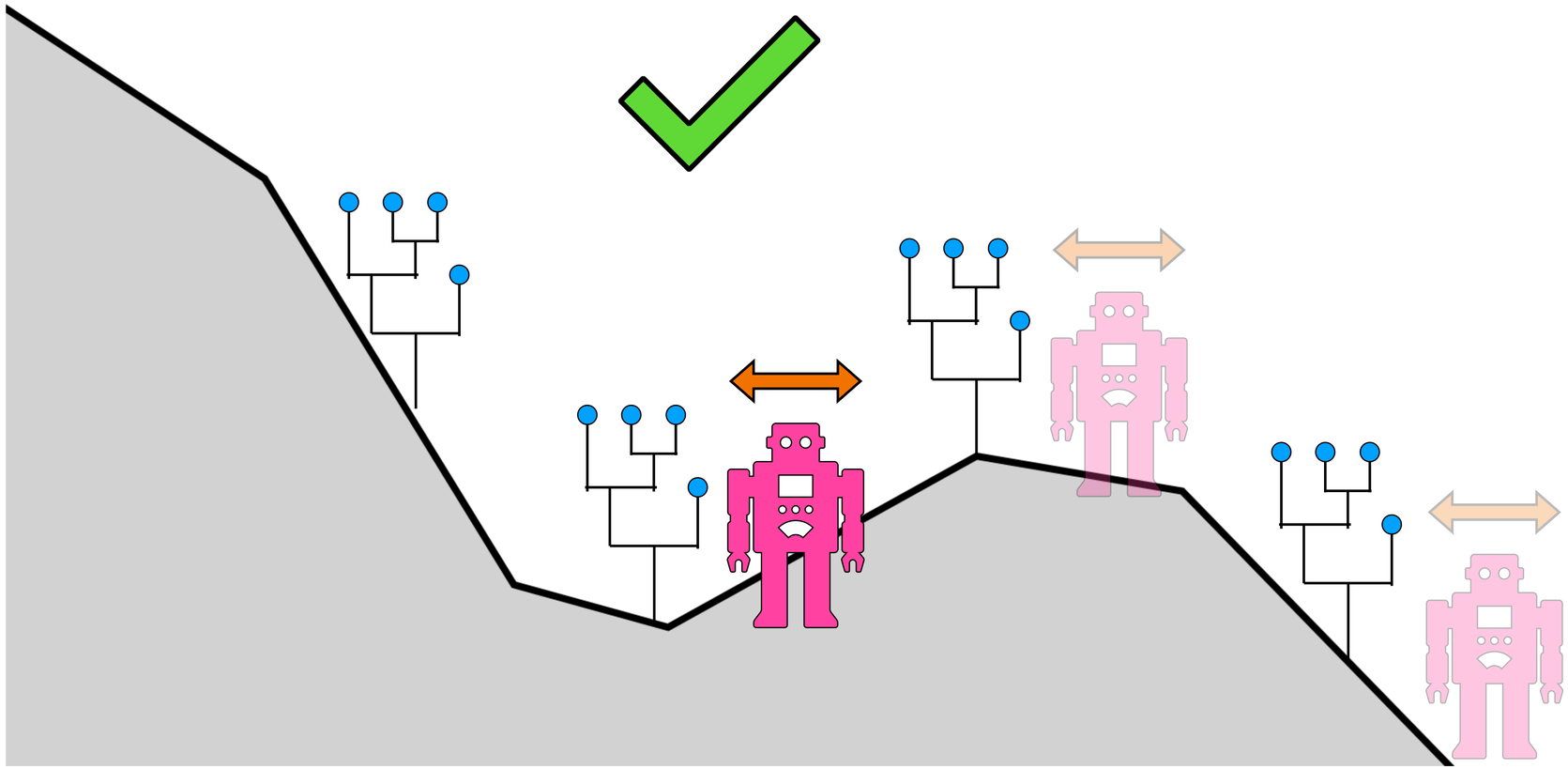


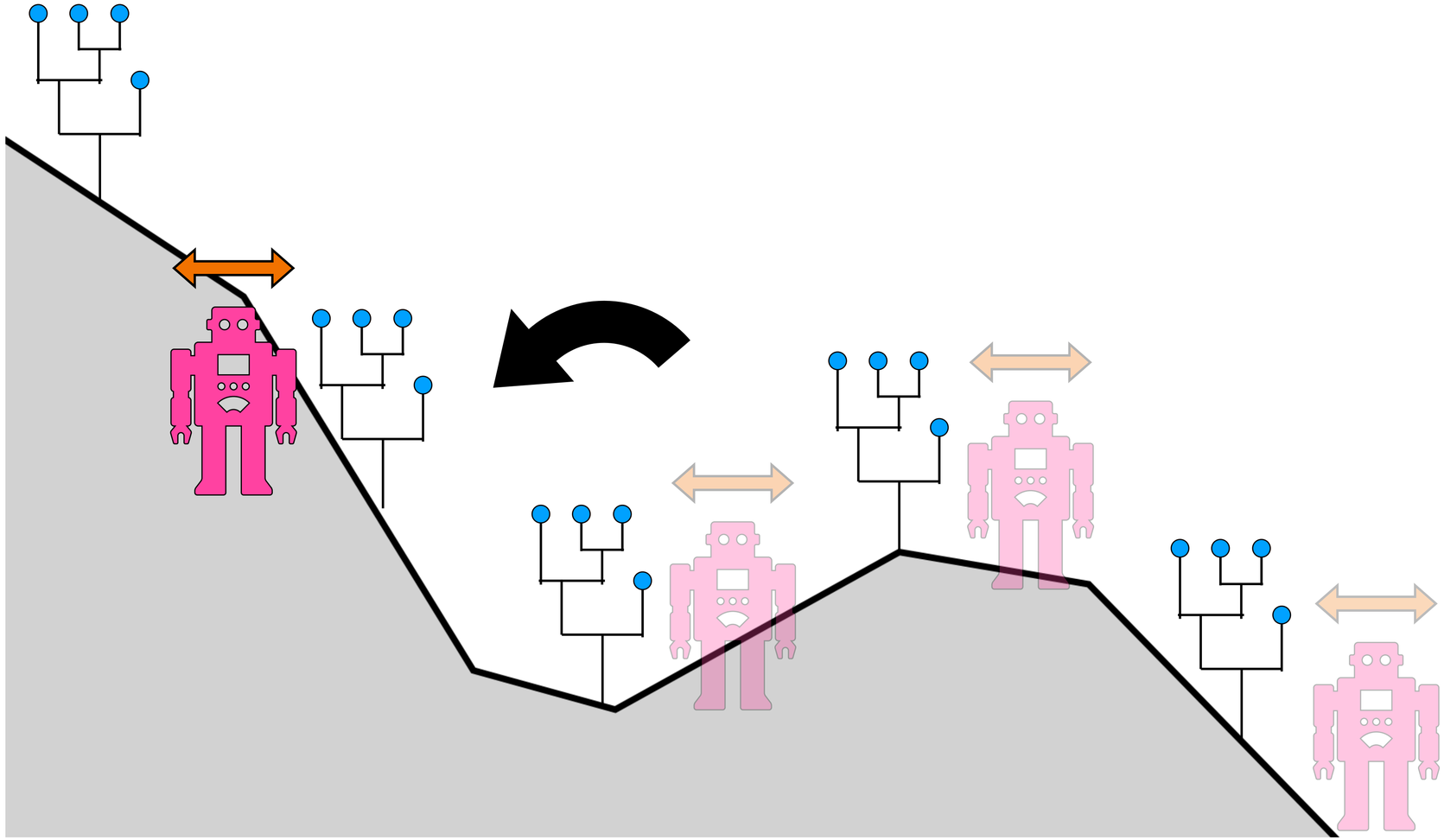

T

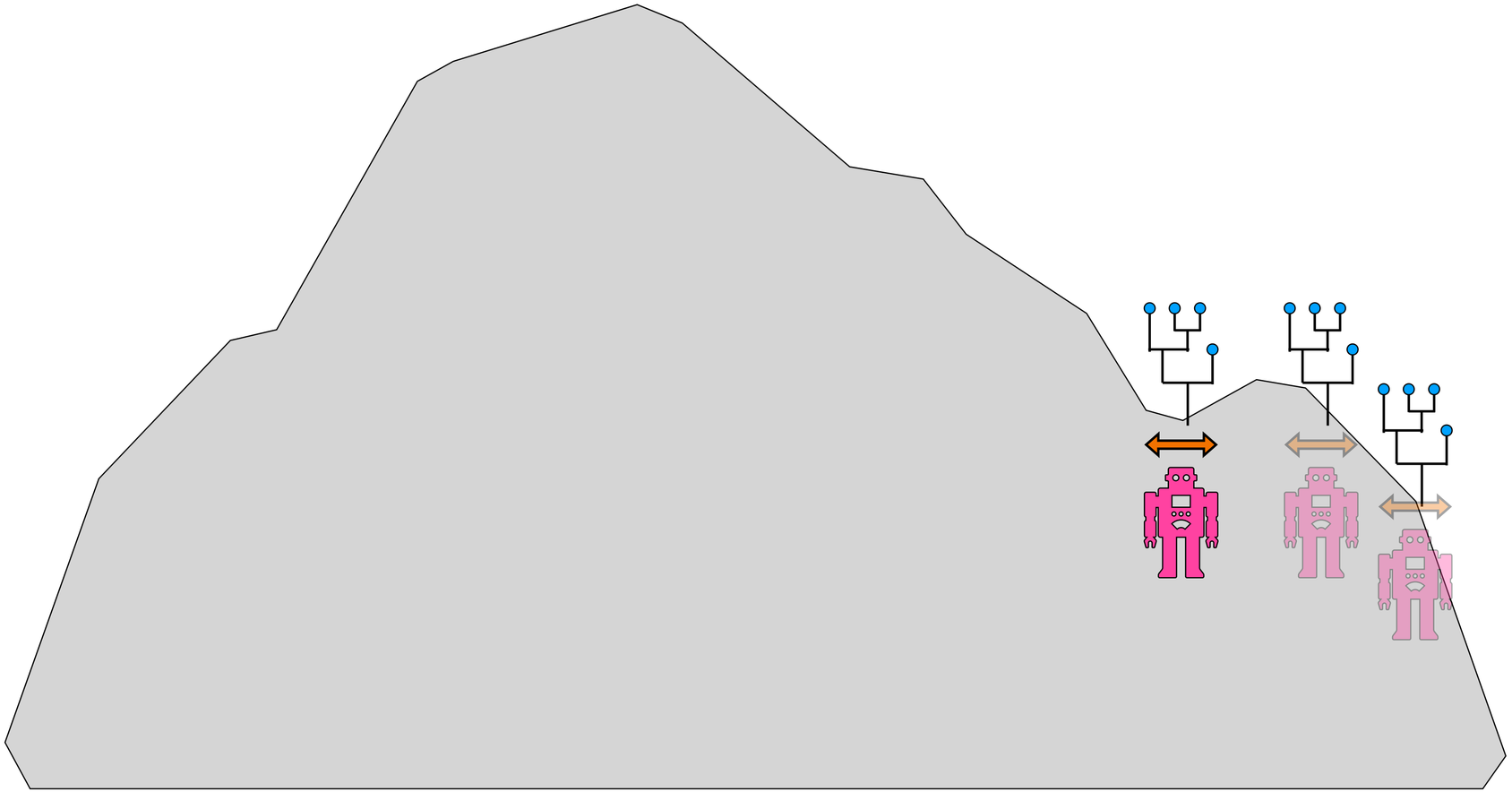


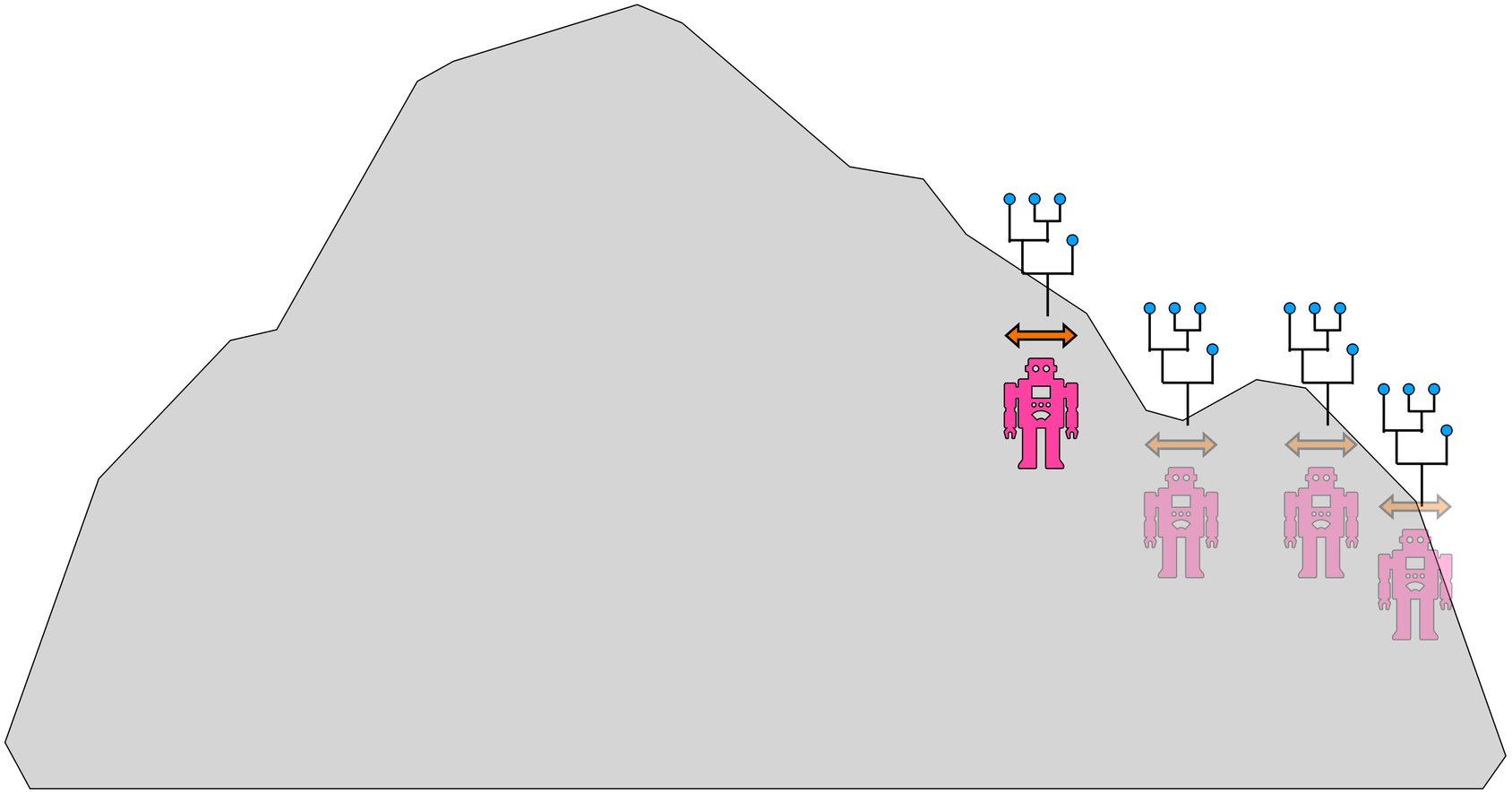


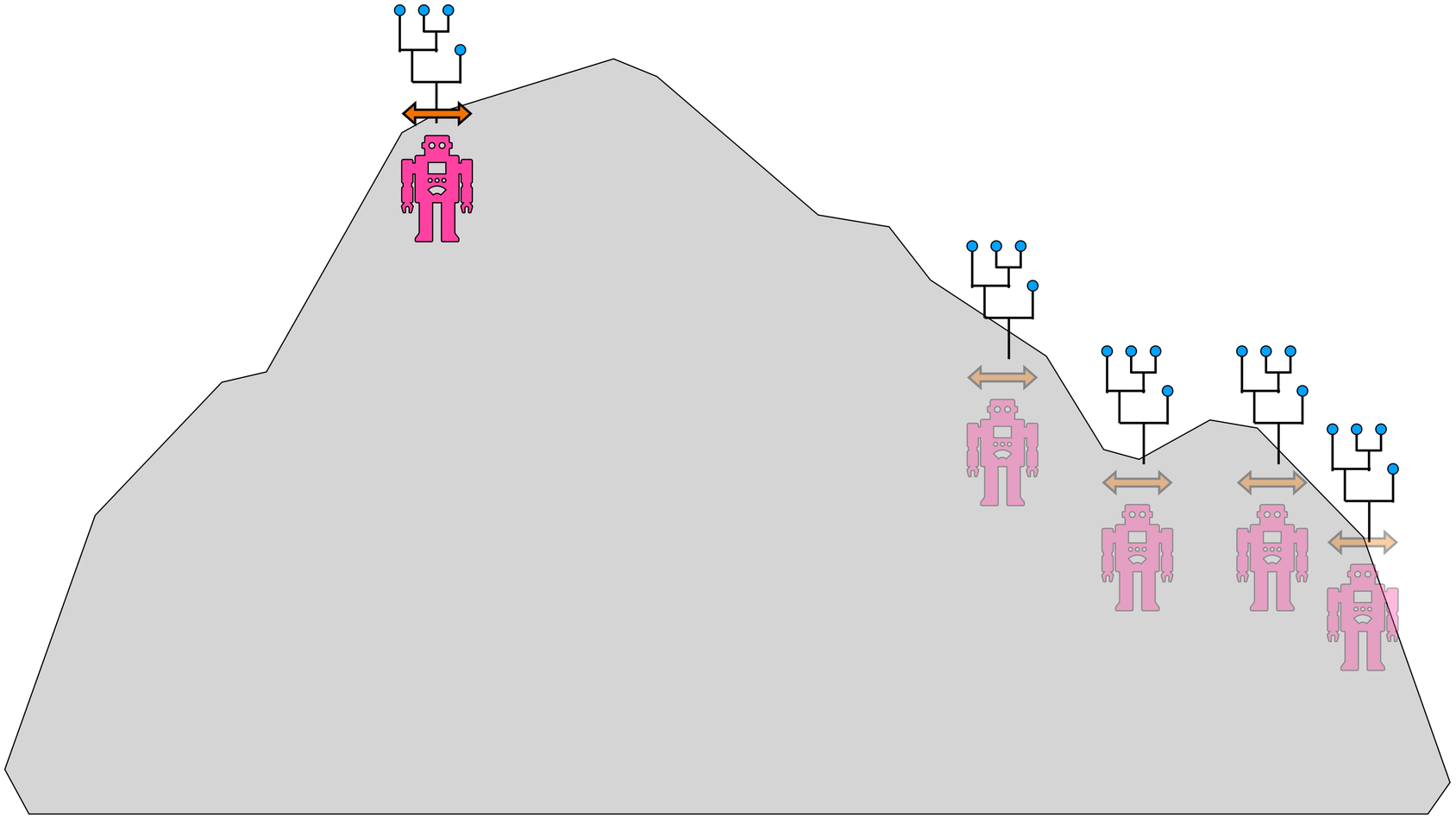


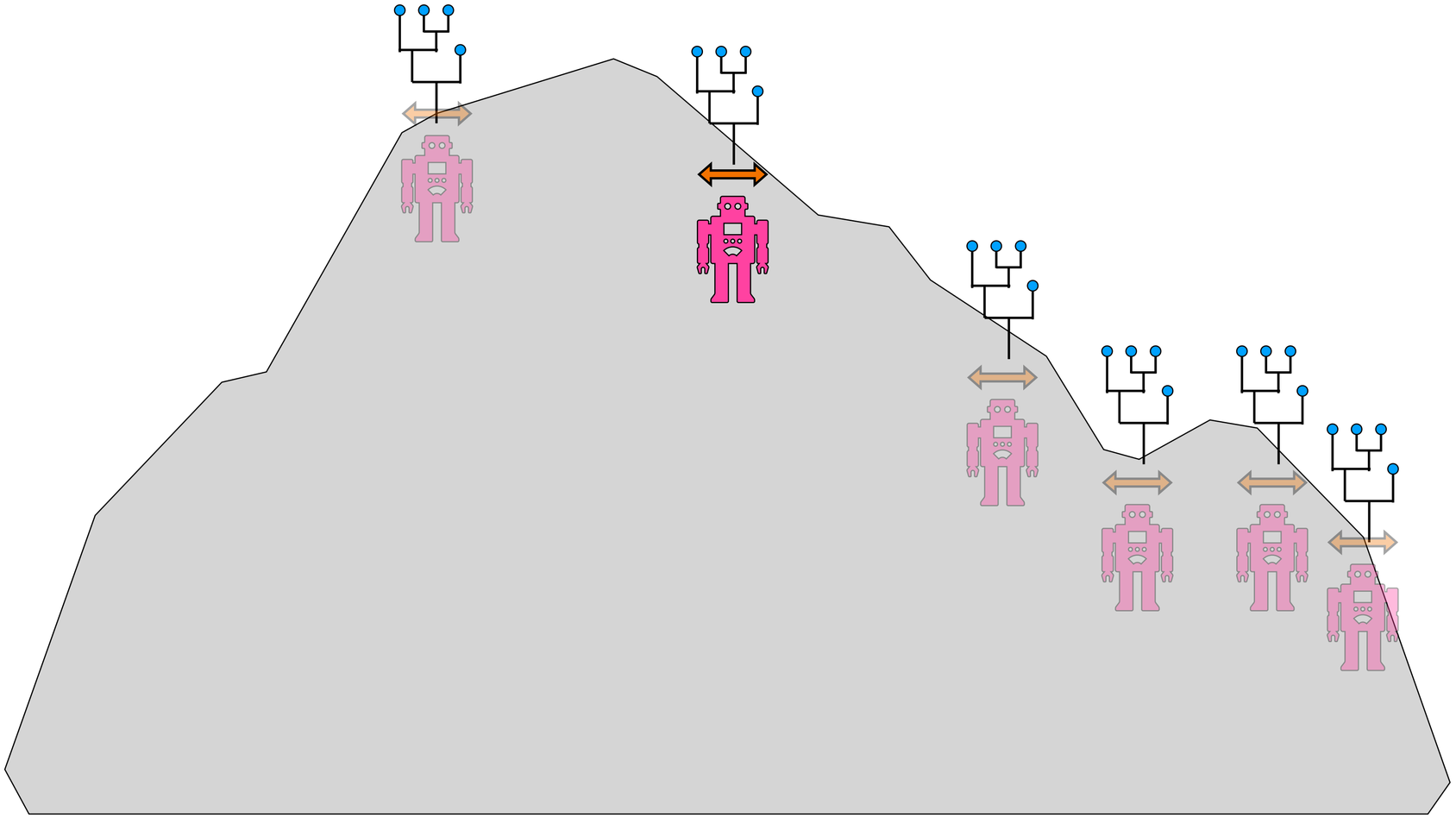


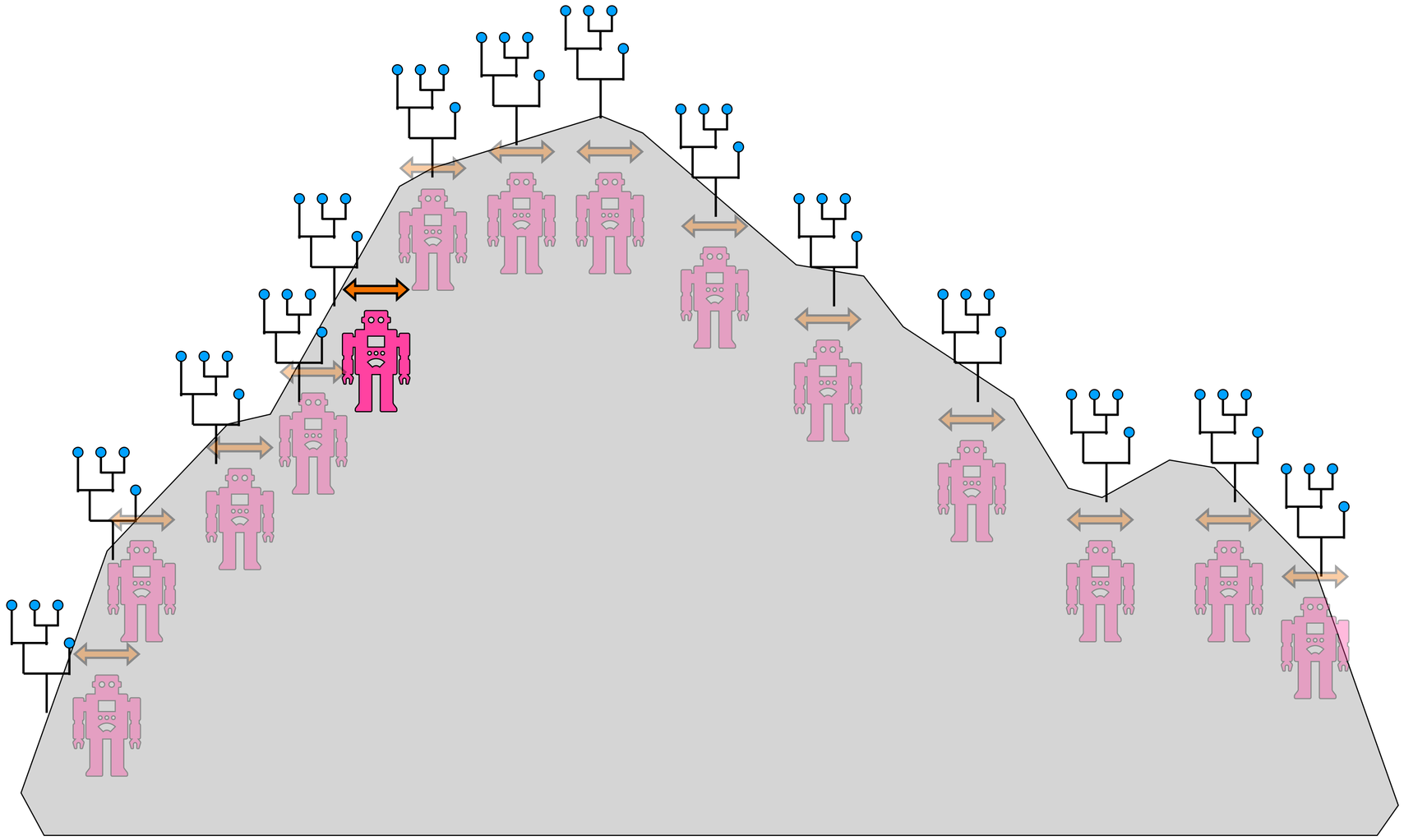






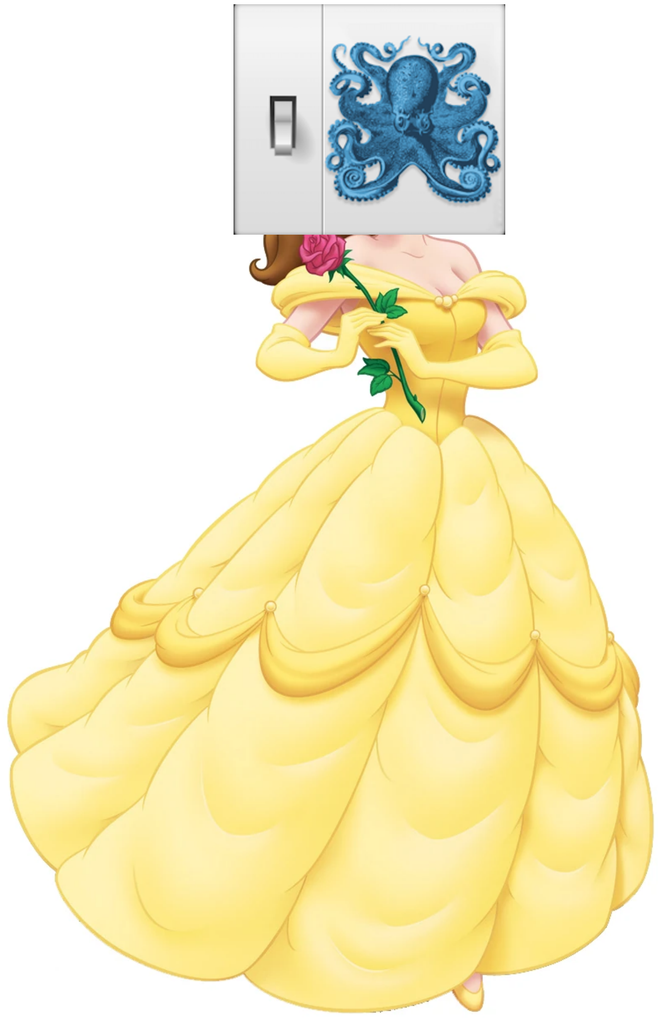






How do I infer trees with BEAST?

BEAUti & the BEAST



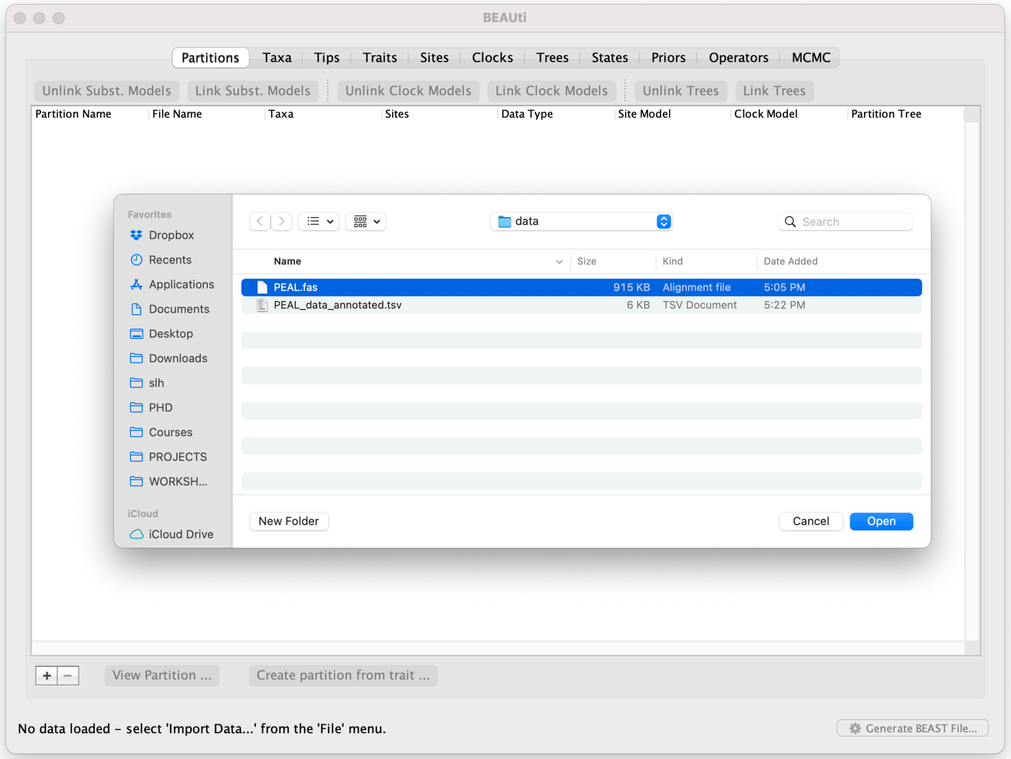
Model specification (BEAUti)

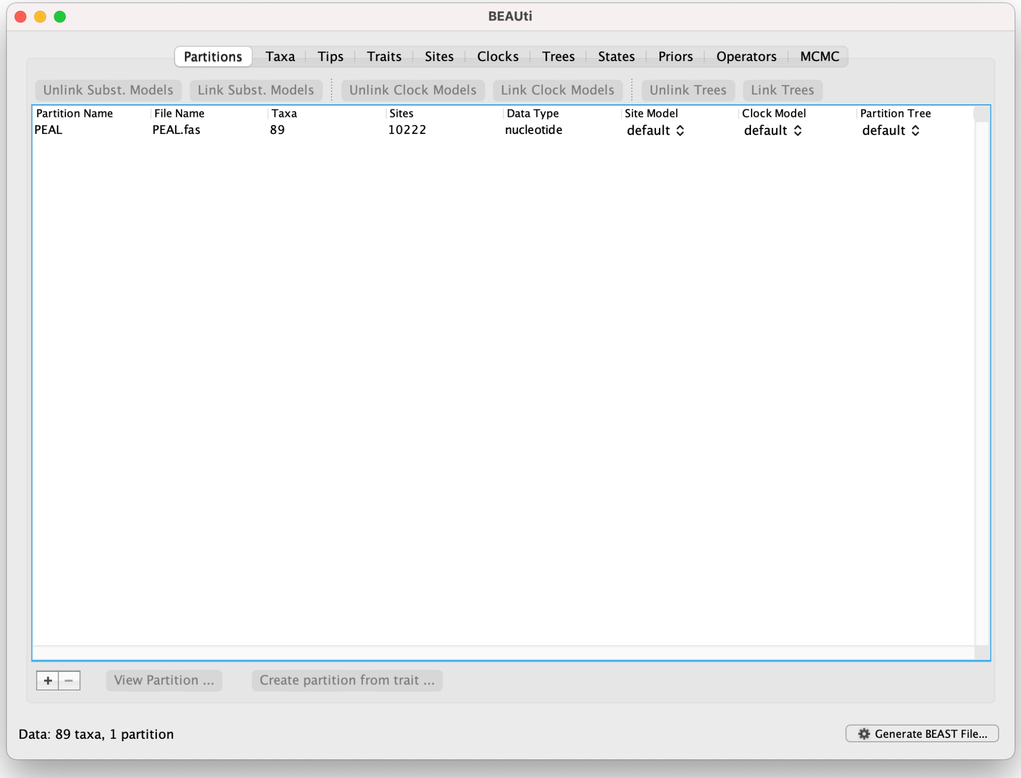
MCMC sampler (BEAST)

Setting up the analysis in BEAUti

- Go to <> and download the data required for the analysis
- The data consists of:
  - A fasta file with 89 sequences of YFV from the 2017-18 outbreak
  - A TSV file with metadata for the 89 sequences
- Open BEAUti by typing *beauti* in the terminal

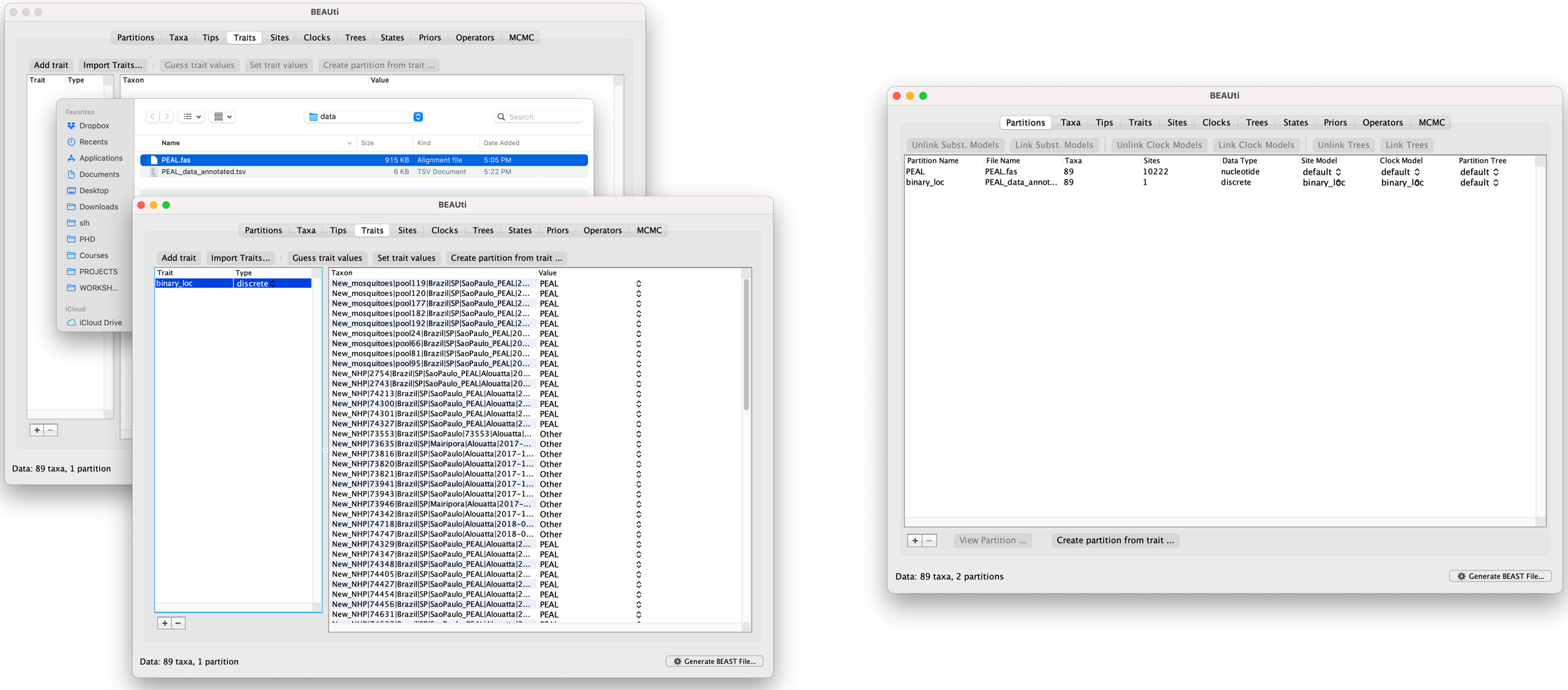
Loading sequence data





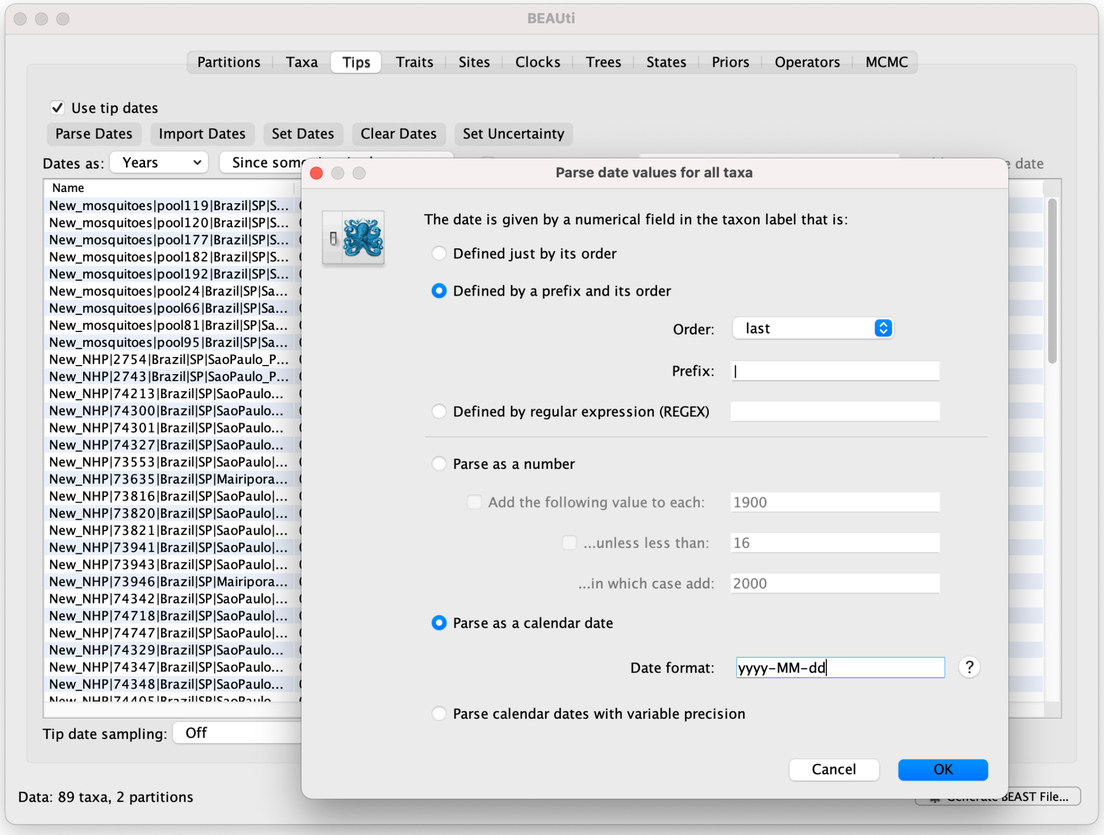
Click on File → Import data and load the fasta file (alternatively you can drag it to the window)

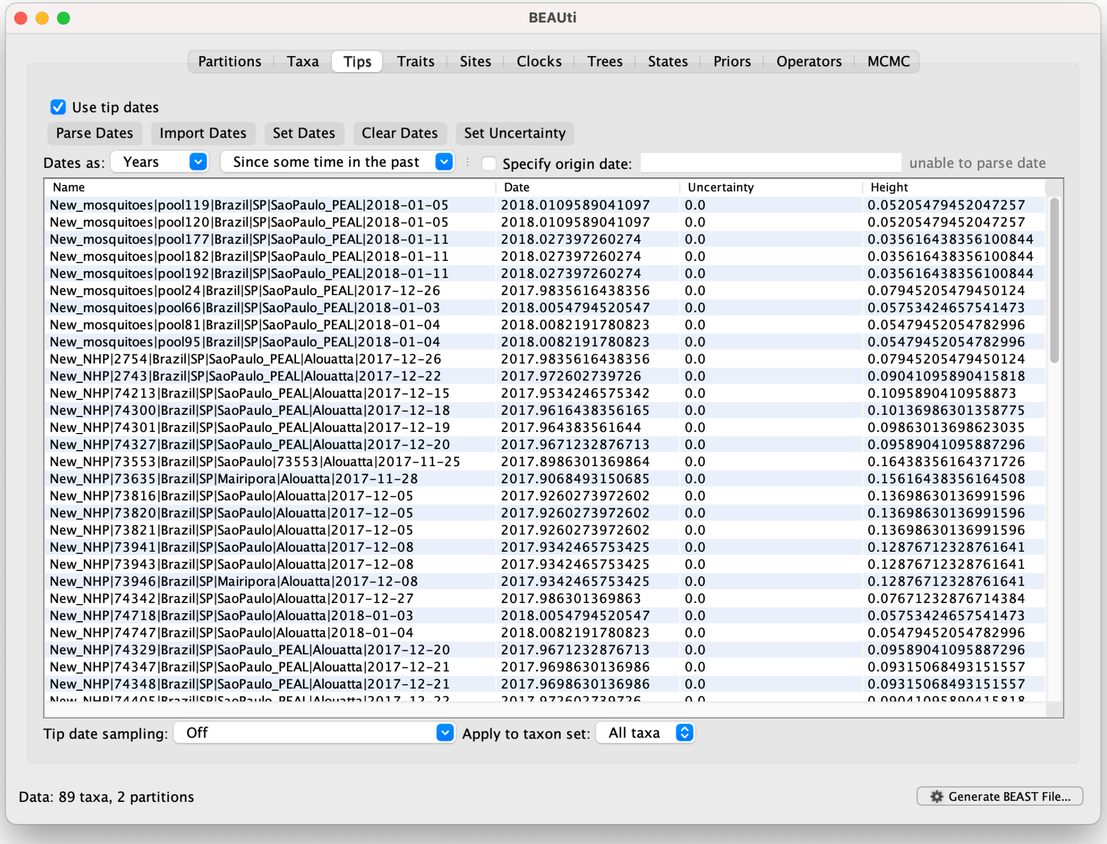
Loading location metadata



Click on *Traits* → *Import Traits* and load the metadata tsv. The click on *Create partition from trait*…

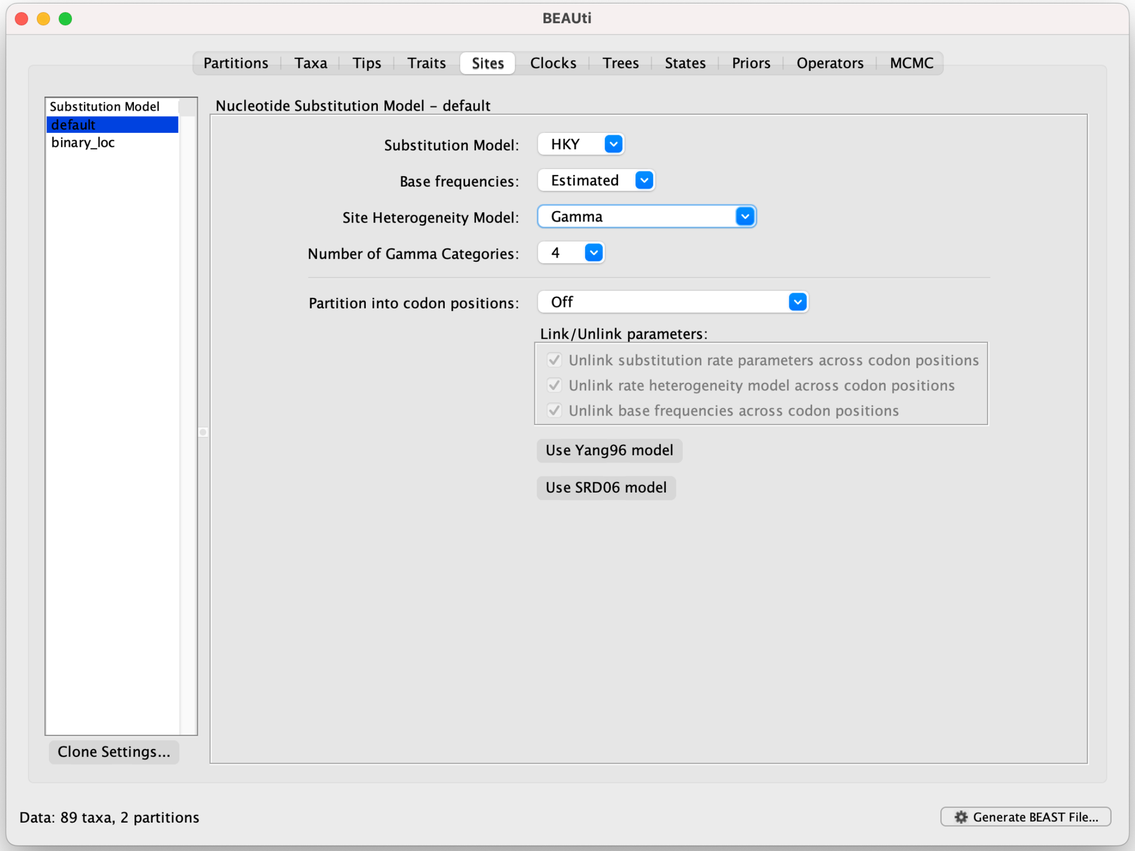
Setting up sequence dates

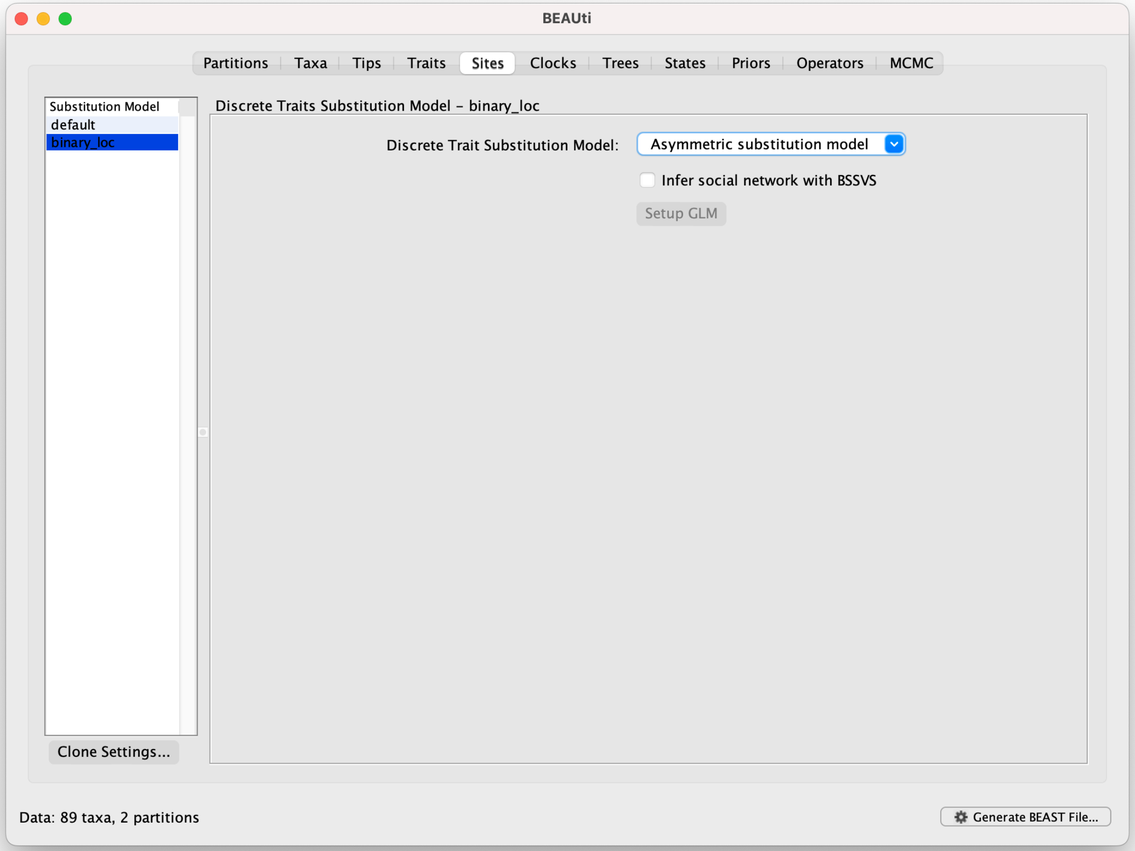




Click on *Tips* → *Parse Dates* and set up the dates by parsing them from the sequence names

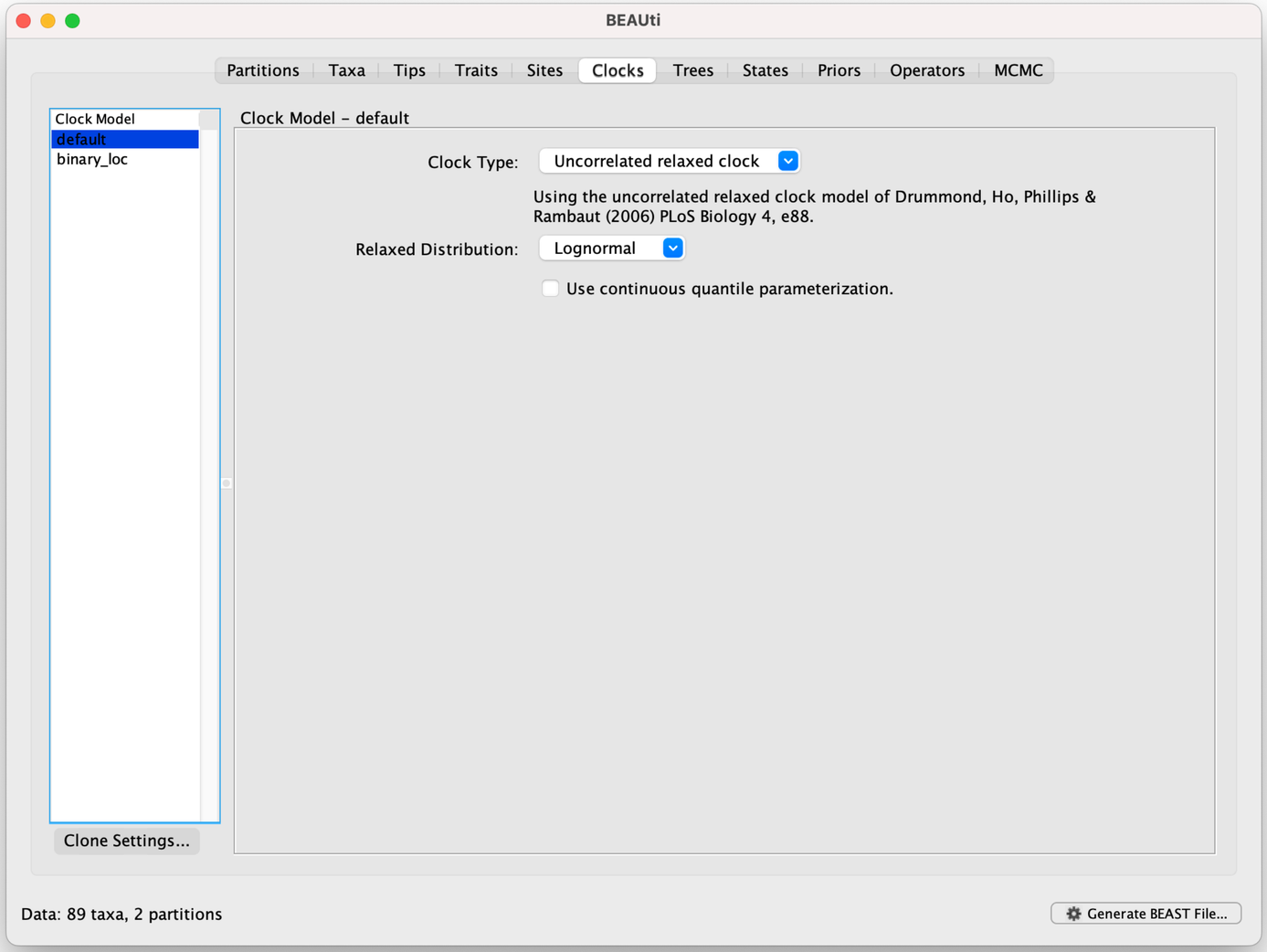
Setting up substitution models





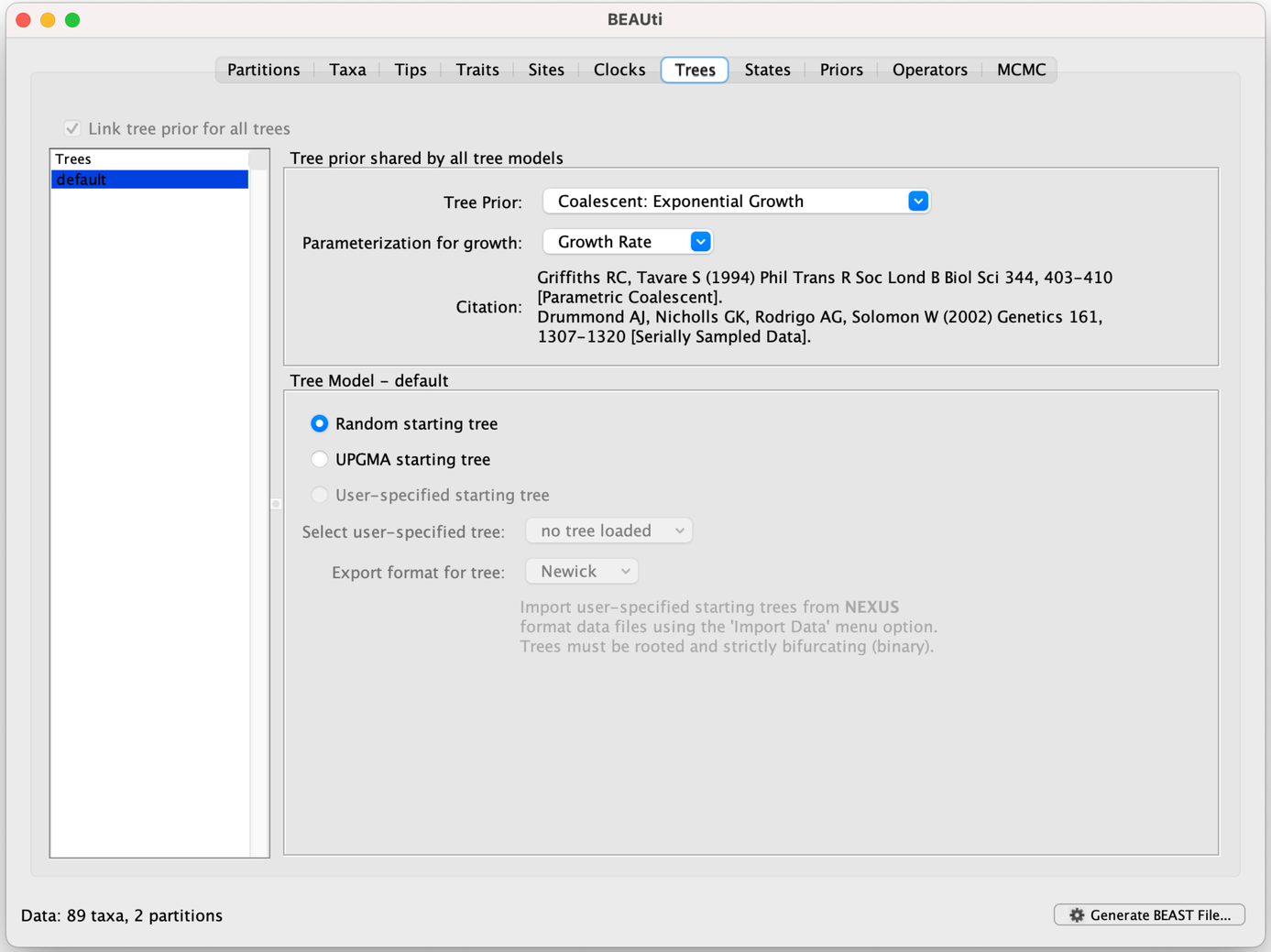
Click on the *Sites* tab and set up the nucleotide substitution model as *HKY with Gamma heterogeneity and 4 categories*. Set up an *Asymmetric substitution model* for the location trait

Setting up the clock model



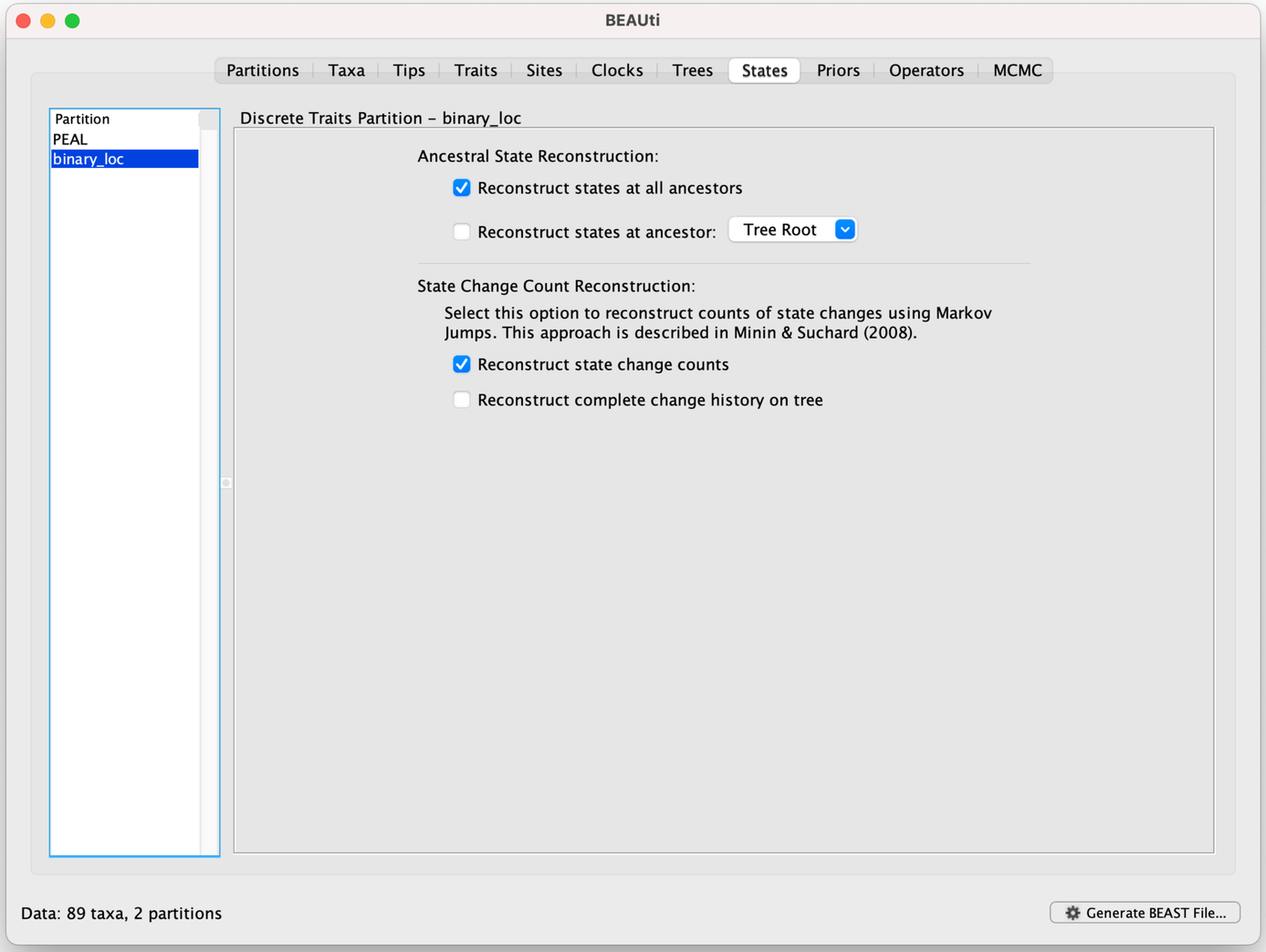
Click on the *Clocks* tab and set up the clock type as *Uncorrelated relaxed clock* with a *Lognormal* distribution. Leave the default *Strict clock* for the location partition.

Setting up the coalescent model



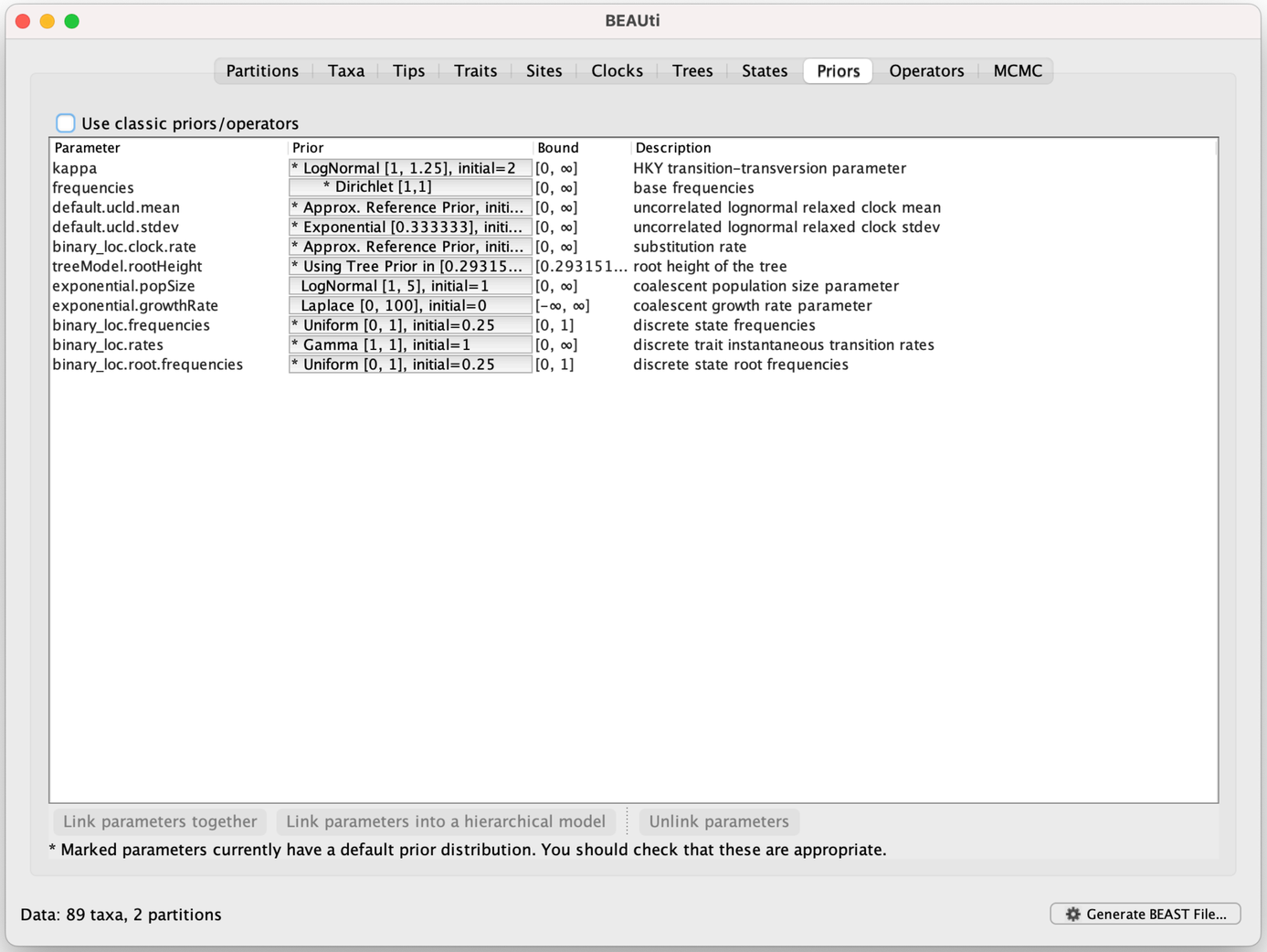
Click on the *Trees* tab and set up the tree prior as *Coalescent: Exponential Growth* with a *Growth Rate* parameterization.

Setting up the ancestral state reconstruction



Click on the *States* tab and on the discrete trait partition, click on *Reconstruct state change counts*

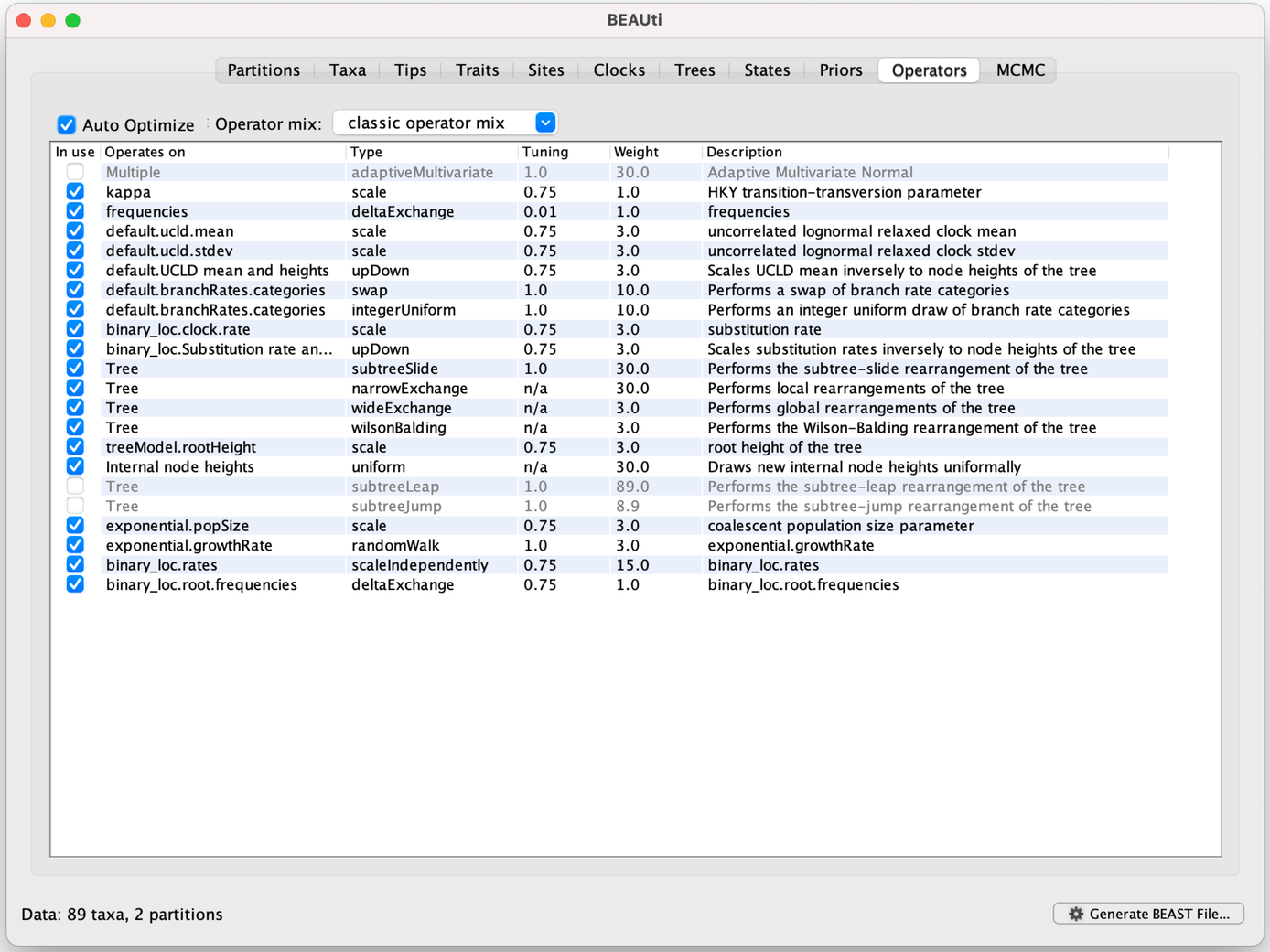
Setting up the priors



Click on the *Priors* tab and leave the default priors for all parameters except for

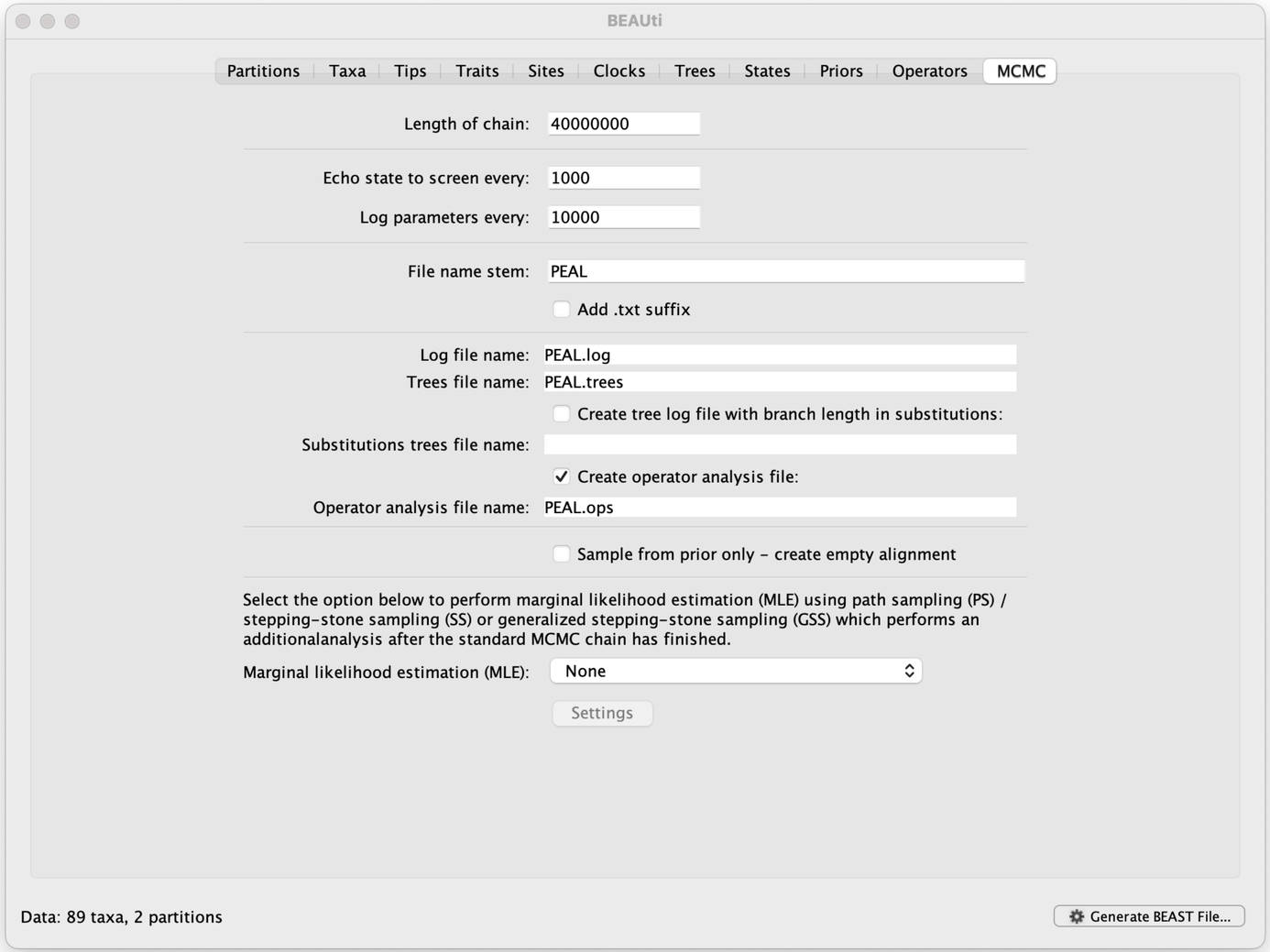
- *exponential.popSize:* LogNormal [1,5]
- *Exponential.growthRate:* Laplace [0,100]

Setting up the operators



On the *Operators* tab,
leave all default values.

Setting up the MCMC



Click on the *MCMC* tab and set the *Length of chain* to 40 million and *Log parameters every* 10,000 states. Generate an XML file for BEAST using the *Generate BEAST File* button on the bottom right.

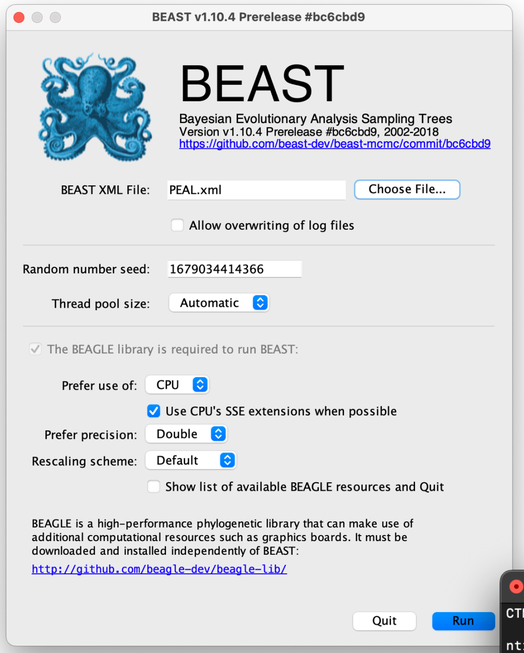
Setting up the Markov Jump counts

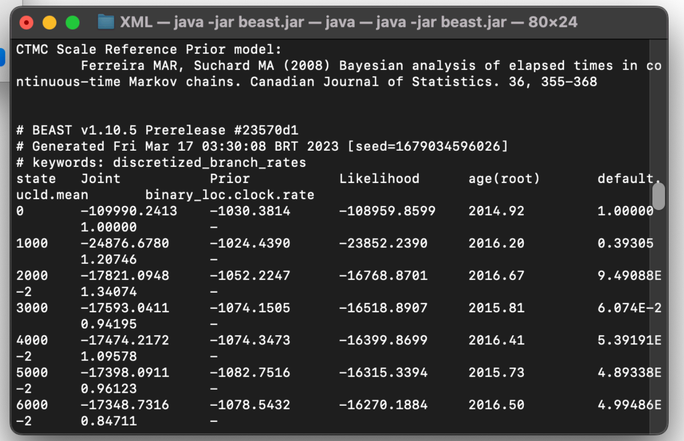
```
<!-- START Ancestral state
reconstruction                                          -->
<parameter id="binary_loc.count" value=" 0.0 1.0 1.0 0.0"/>
<parameter id="binary_loc.OthertoPEAL" value=" 0.0 1.0 0.0 0.0"/>
<parameter id="binary_loc.PEALtoOther" value=" 0.0 0.0 1.0 0.0"/>

<!-- END Ancestral state
reconstruction                                          -->
```

By default, *Reconstruct state change counts* in BEAUti will only create a parameter that logs all jumps from any two locations. To keep track of jumps into and out of Horto separately, open the generated XML file in a text editor and add the *OthertoPEAL* and *PEALtoOther* parameters below the parameter that logs all counts, in the Ancestral state reconstruction section of the *markovJumpsTreeLikelihood* block. The zeros and ones in the value field represent the transitions being counted represented with a flattened matrix with rows and columns for the different locations alphabetically ordered.

Running the analysis in BEAST

- Open BEAST by typing *beast* in the terminal
- Once BEAST is open, select the XML file for the analysis using the *Choose file* button.
- Click the *Run* button to start the analysis. Your terminal window should now be outputting the parameter values being explored by the MCMC sampler

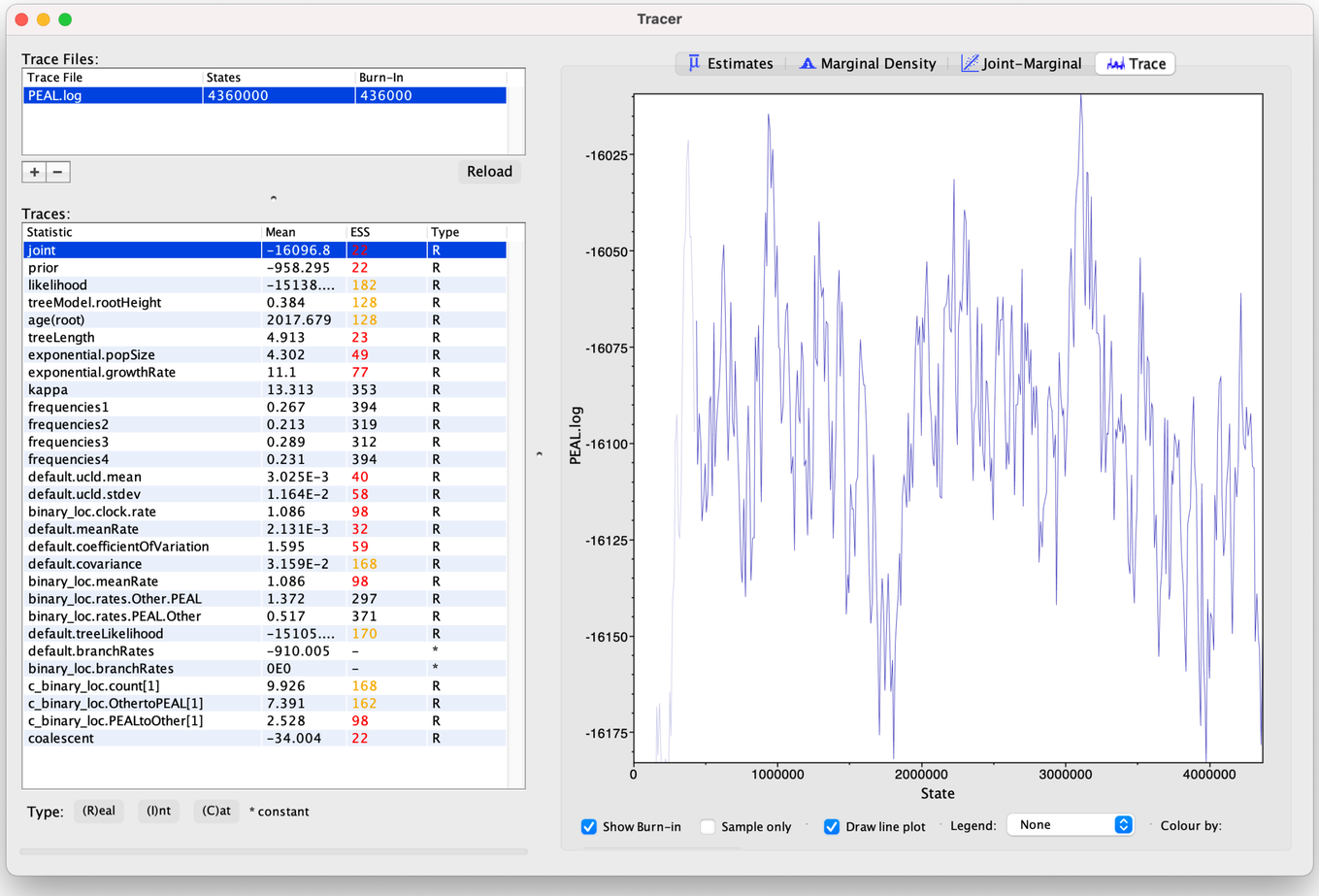




BEAST output files

- You should now see 3 output files being generated by BEAST:
  - A log file with all parameters being sampled in the analysis
  - A log file with the phylogeography rates being sampled
  - A trees file with all trees being sampled in the MCMC run
- We can use the Tracer to explore and visualize the output in the log files.

Tracer



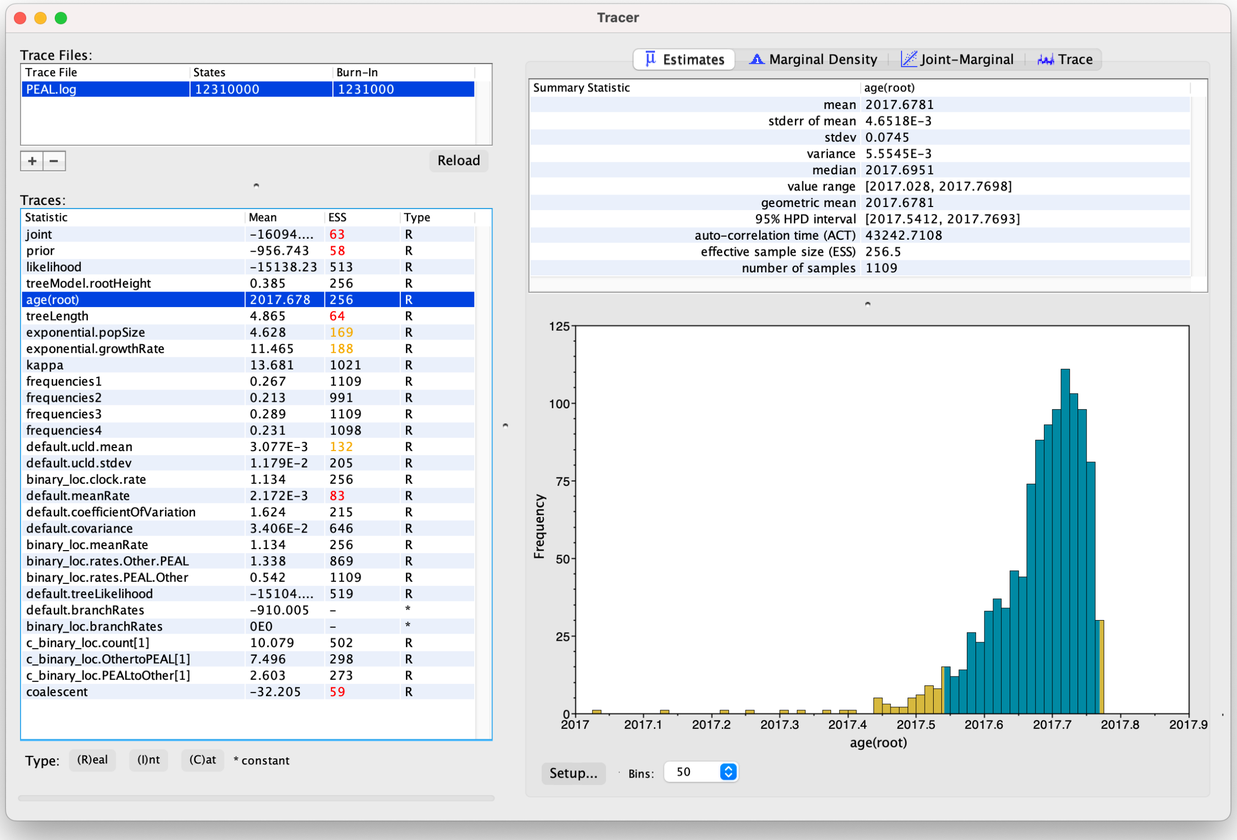
- Open Tracer and load the log file by dragging and dropping it into the *Trace Files* box
- Once loaded, you will see all parameters being logged on the *Traces* box on the left, and different visualizations on the tabs on the right.
- As a rule of thumb we want to keep running the analysis until all parameters in the log file have ESS values above 200.

Exploring the Tracer results

- The values shown in Tracer represent the estimates of the parameters in our model being obtained by sampling across all trees in our posterior distribution.

- Thus, the uncertainty in our estimates is given for free without having to do any additional analysis.

- We can check the *95% Highest Posterior Density (HPD)* range for our parameters in the *Estimates* tab on the right. We call this the *credible interval* of our parameter, and it states that there is a 95% probability that the parameter value falls within this range.



Back to the epidemiology

- When did the outbreak begin?
- How fast is the virus evolving on average?
- How many mutations would you expect after 1 week, given a genome size of 10kb?
- How many introductions to PEAL?
- How rapidly is the virus population growing?
- What is the doubling time of the virus?

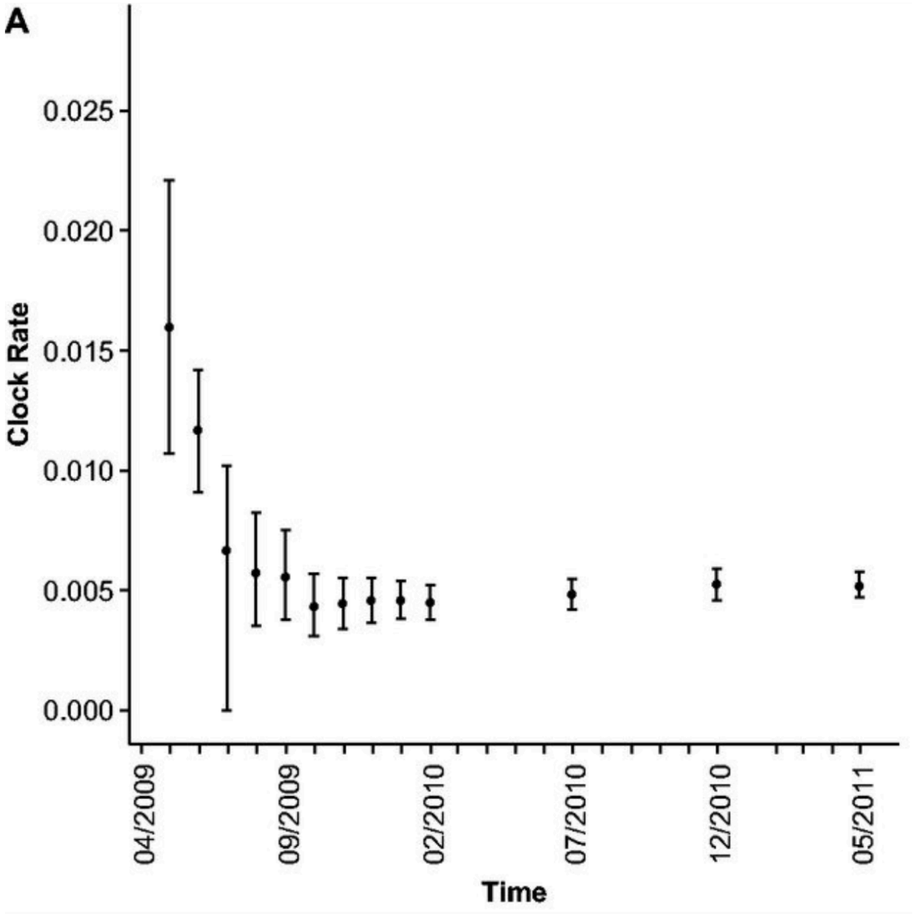
Back to the epidemiology

- When did the outbreak begin?
  - 2017.7 [2017.5 - 2017.8]
- How fast is the virus evolving on average?
  - 2.2E-3 [1.3E-3 - 3.2E-3]
- How many mutations would you expect after 1 week, given a genome size of 10kb?
  - 0.42 mutations [0.24 - 0.61]
- How many introductions to PEAL?
  - 7.45 [5 - 11]
- How rapidly is the virus population growing?
  - Growth rate 10.9 [2.3 - 20.2]
- What is the doubling time of the virus?
  - Doubling time 23 days [12 - 109]

A note on the evolutionary rate

- Notice that the mean rate we obtain from our analysis, 2.2E-3 [1.2E-3 to 3.2E-3] is much faster than the rate of ~9E-4 reported in previous studies.
- An important thing to consider is that all of our sequences come from a short and densely sampled period of time at the beginning of the outbreak. The short amount of time of rapid growth in the virus population means that many mutations we observe might actually be transient polymorphisms, deleterious mutations that are still present in the population because purifying selection hasn't had enough time to act.
- This phenomenon of inflated evolutionary rates at the beginning of an outbreak can also be seen in other pathogens such as with the case of the 2009 pandemic H1N1 influenza virus.

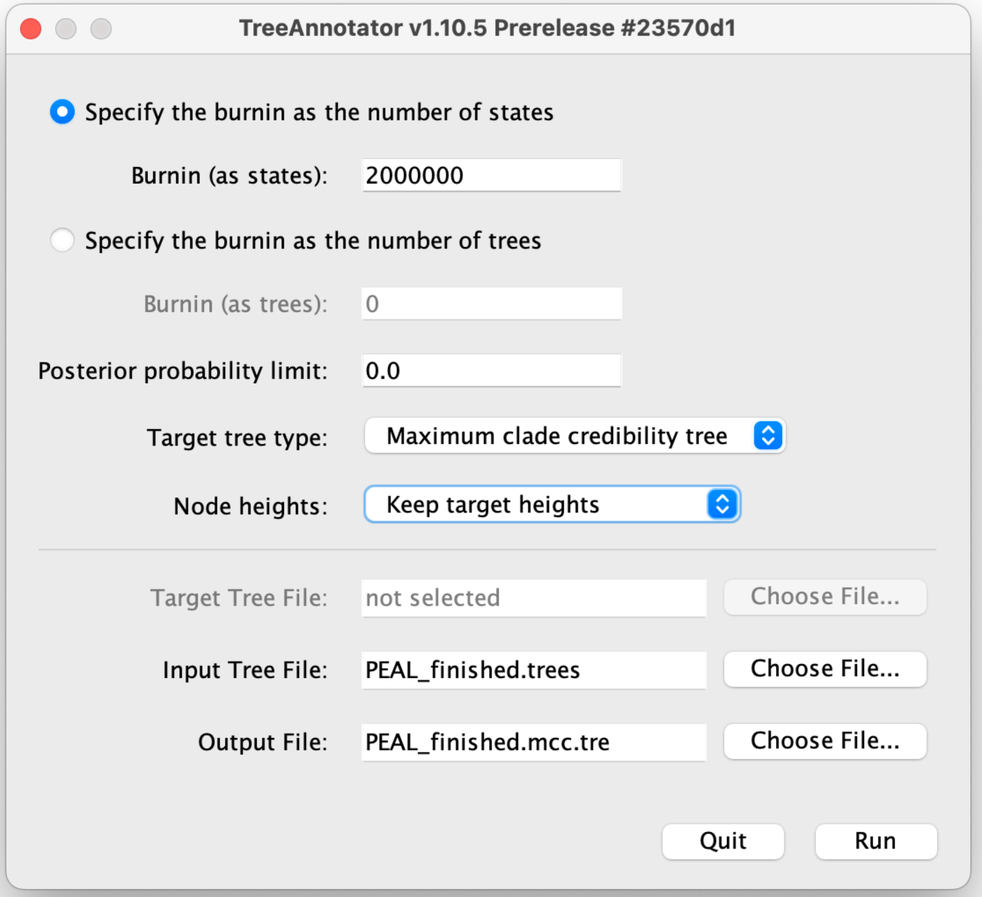
2009 pH1N1 clock rates



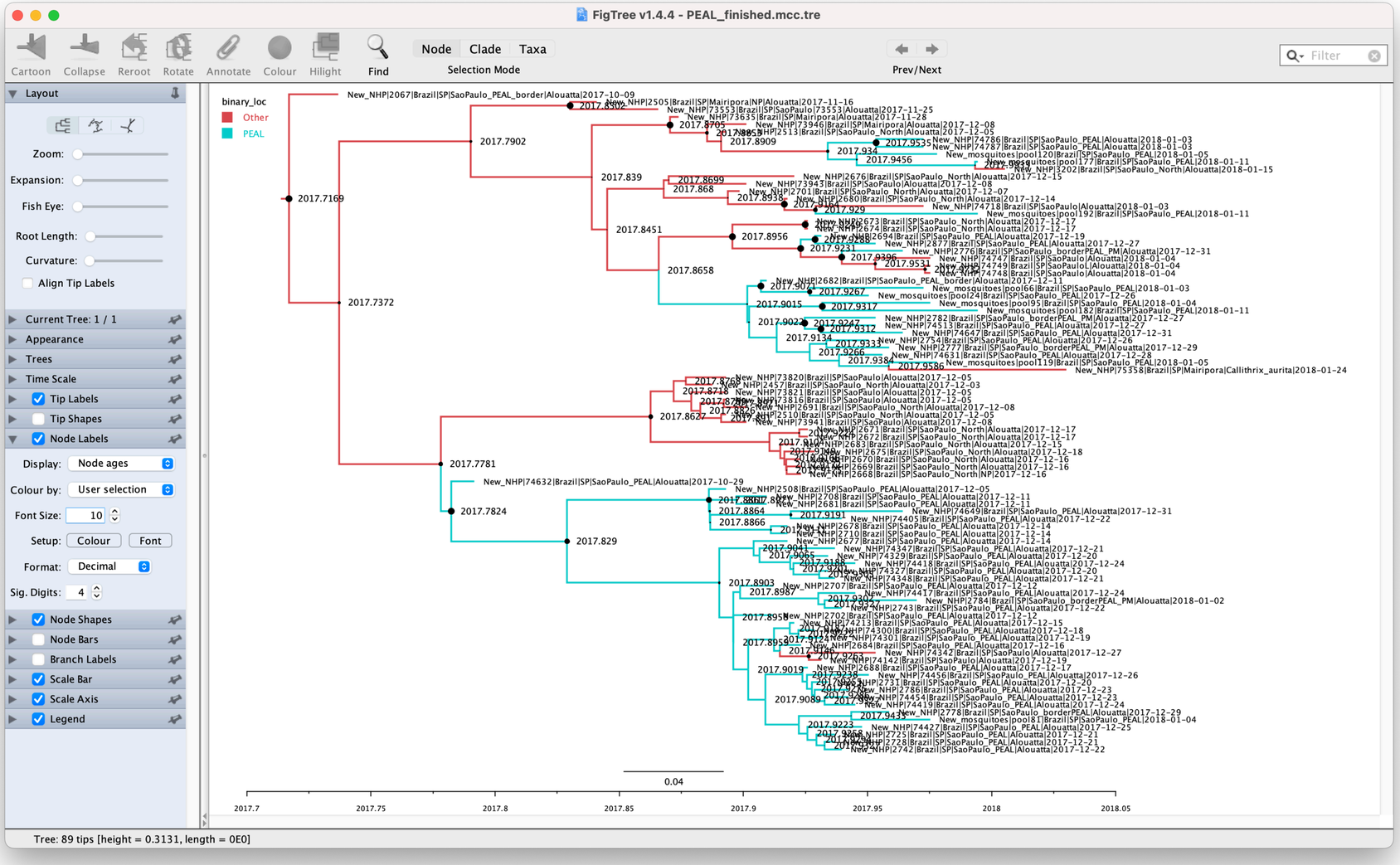
Meyer et al. 2009

Summarizing posterior trees

- We can summarize posterior trees using TreeAnnotator to obtain a Maximum Clade Credibility (MCC) tree
- For each tree in the posterior, calculate the fraction of times each clade appears in the collection of trees
- The MCC tree is the tree that maximizes the product of these fractions
- Posterior support values are annotated in the internal nodes denoting the probability of each clade



Visualizing the MCC tree



Finally, we can use FigTree to visualize the MCC tree with the annotations