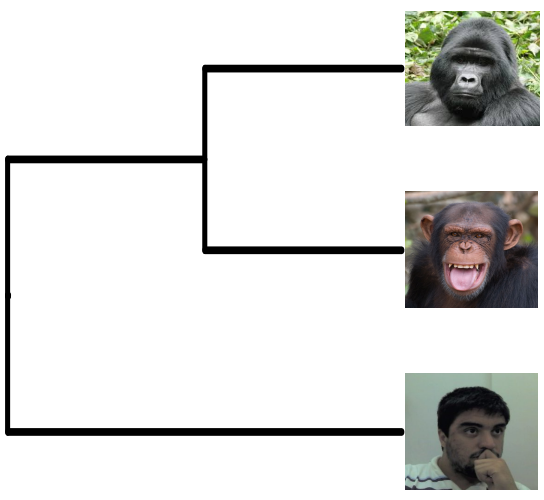
Simulation-based calibration for phylogenetics

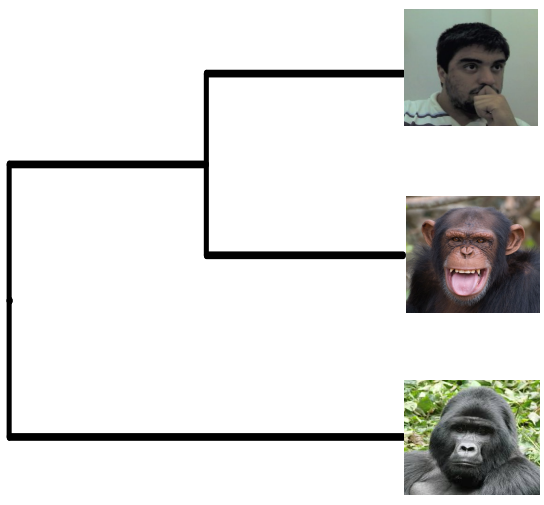
Challenges from a non-standard statistical problem

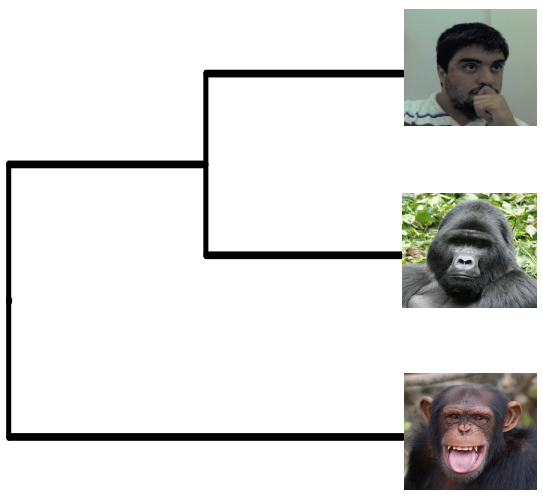
Luiz Max Carvalho & Remco Bouckaert (Auckland)

FGV EMAp

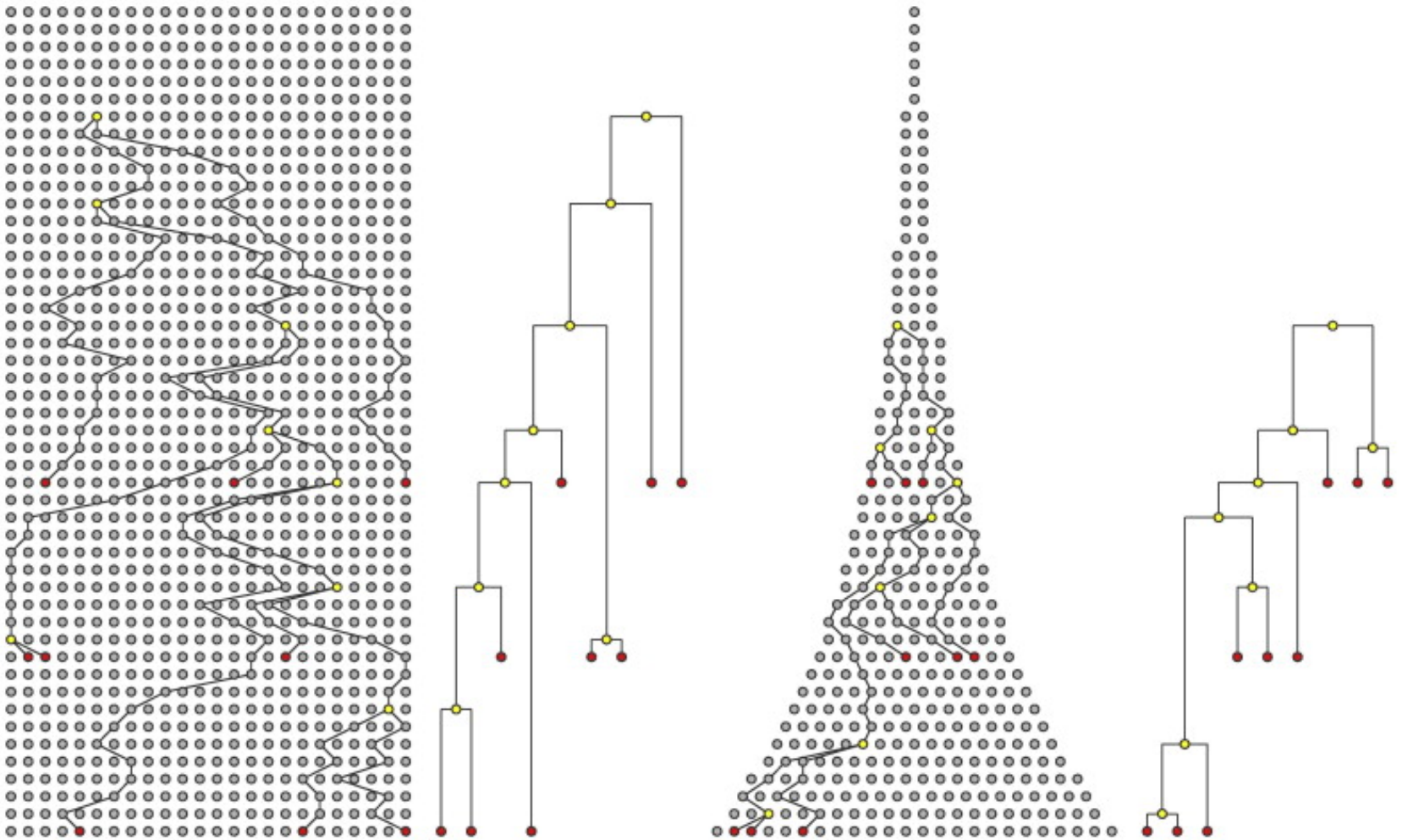
Trees are hypotheses







Trees and the coalescent



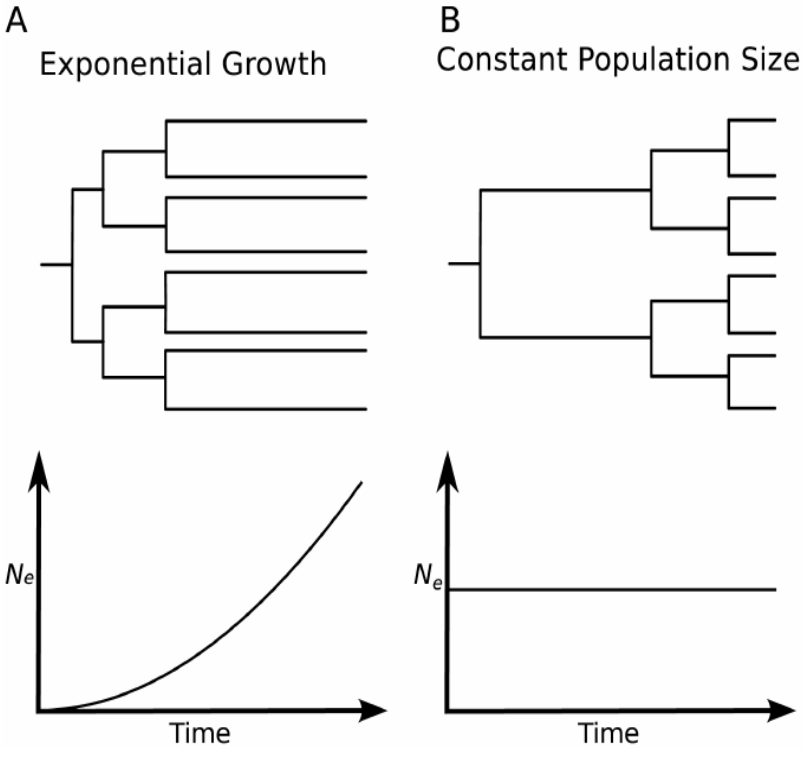
Motivation

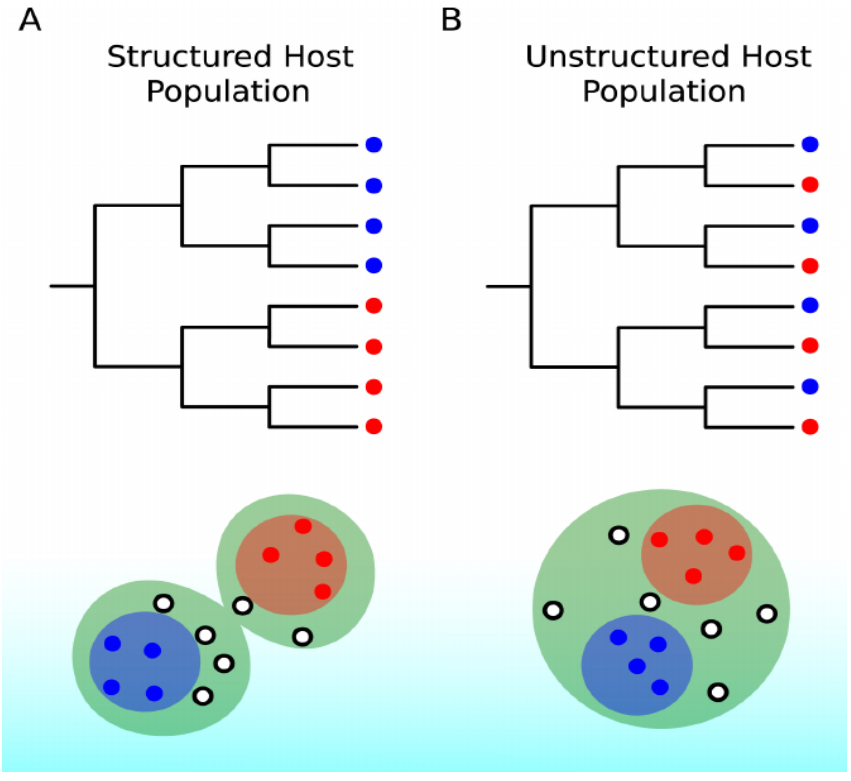
Phylodynamics of fast-evolving viruses

Inferring spatial and temporal dynamics from genomic data:

Phylogenies*!

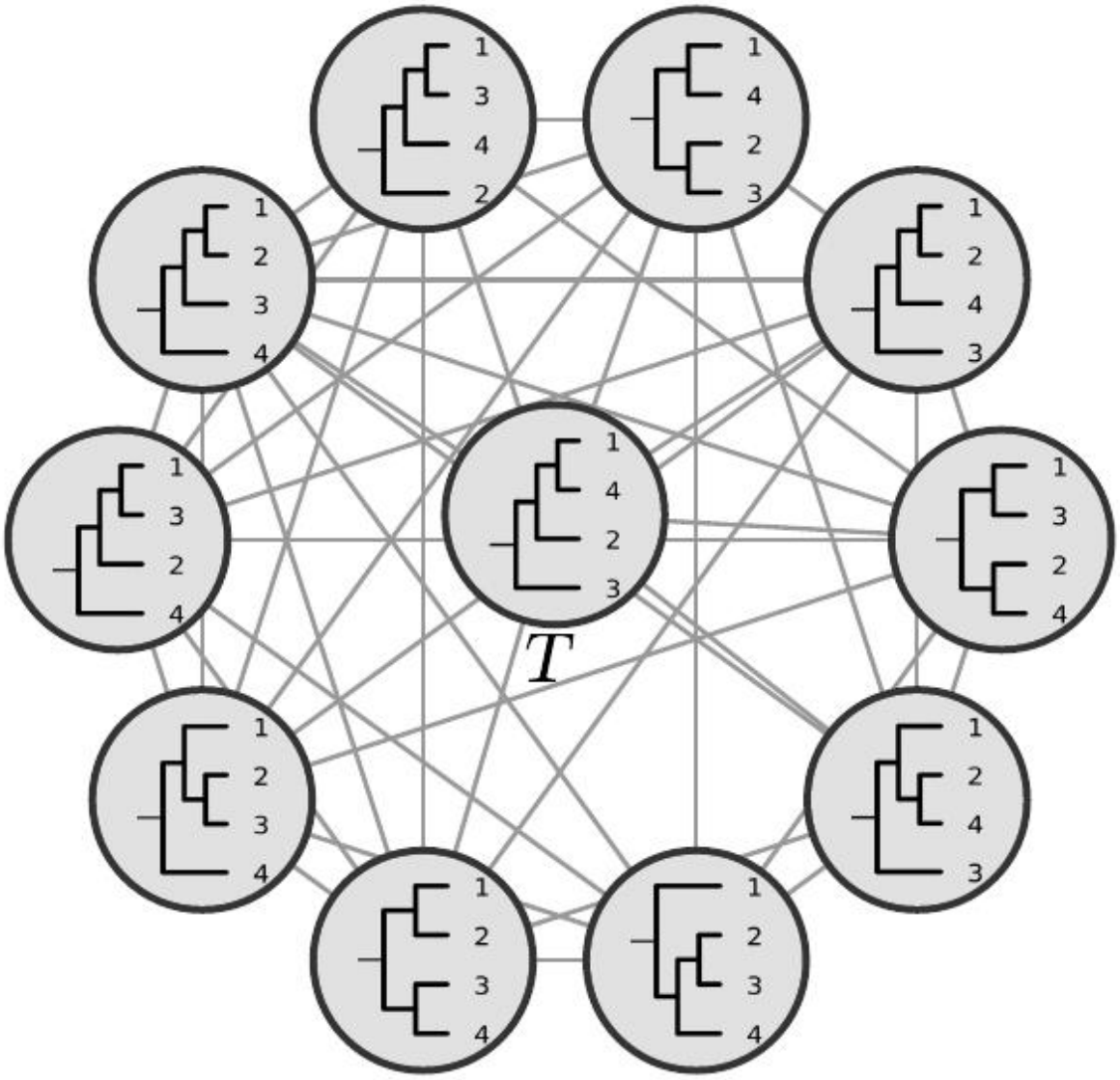
* plus complicated models





Discrete tree space: SPR graph

For curvature results, see Whidden & Matsen(2017).



Target

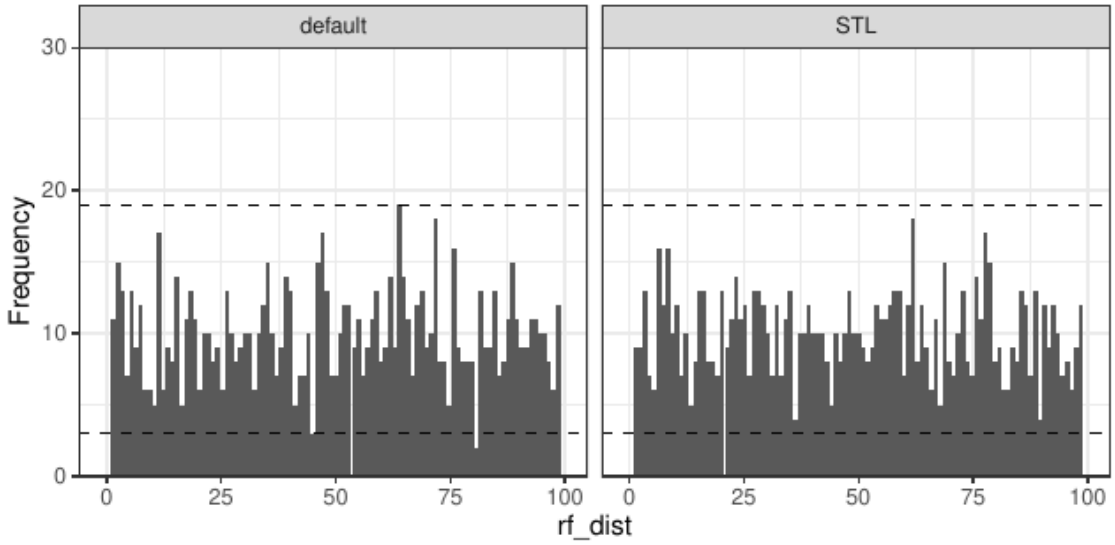
$$p(t, \boldsymbol{b}, \boldsymbol{\omega}|D) = \frac{f(D|t, \boldsymbol{b}, \boldsymbol{\omega})\pi(t, \boldsymbol{b}, \boldsymbol{\omega})}{\sum_{t_i \in \boldsymbol{T}_n} \int_{\boldsymbol{B}} \int_{\boldsymbol{\Omega}} f(D|t_i, \boldsymbol{b}_i, \boldsymbol{\omega})\pi(t_i, \boldsymbol{b}_i, \boldsymbol{\omega}) d\boldsymbol{\omega} d\boldsymbol{b}_i} \quad (1)$$

- D: observed sequence (DNA) data;
- $\boldsymbol{T}_n$: set of all binary ranked trees ($\mathbb{G}^{(2n-3)!!}$);
- $\boldsymbol{b}_k$: set of branch lengths of $t_k \in \boldsymbol{T}_n$ ($\mathbb{R}_+^{2n-2}$, kind of) ;
- $\boldsymbol{\omega}$: set of parameters of interest such as substitution model parameters, migration rates, heritability coefficients, etc.

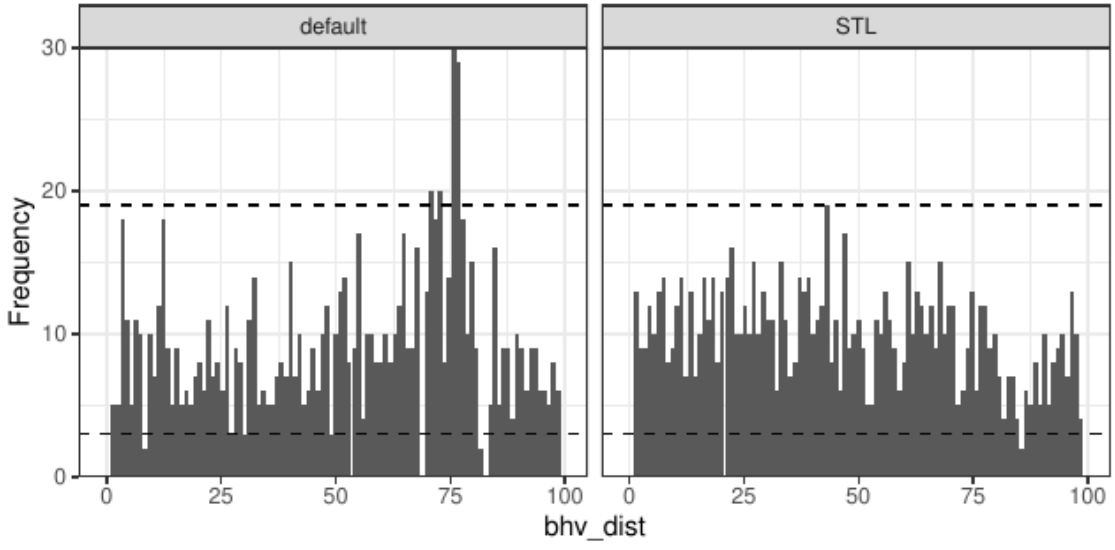
SBC for trees

0. Generate a reference tree from the prior $\bar{\tau}_0 \sim \pi_T(\tau|\boldsymbol{\gamma})$;
   for each iteration in 1:N, **do**:
1. Generate $\bar{\tau} \sim \pi_T(\tau|\boldsymbol{\gamma})$;
2. Compute the distance $\bar{\delta} = d_\sigma(\bar{\tau}, \bar{\tau}_0)$ according to the metric of choice;
3. Generate some (alignment) data $\tilde{y} \sim p(y|\bar{\tau}, \boldsymbol{\alpha})$;
4. Draw (approximately) $\boldsymbol{\tau}_s = \{\tau_s^{(1)}, \tau_s^{(2)}, \ldots, \tau_s^{(L)}\}$ from the posterior $\pi(\tau|\tilde{y})$;
5. Compute distances $\boldsymbol{\delta}_s = \{\delta_1, \delta_2, \ldots, \delta_L\}$ with $\delta_i = d_\sigma(\tau_s^{(i)}, \bar{\tau}_0)$;
6. Compute the rank $r(\boldsymbol{\delta}_s, \bar{\delta}) = \sum_{i=1}^{L} \mathbb{I}(\delta_i < \bar{\delta})$.

Some results: tree distances



(a) Robinson-Foulds, $\mathrm{RF}_0(\tau)$



(b) BHV, $\mathrm{BHV}_0(\tau)$

Some results: continuous parameters

Simulation Based Calibration

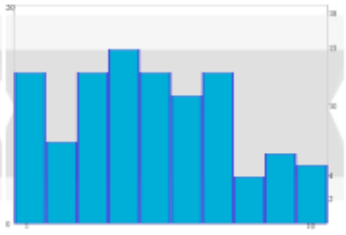
prior sample: ../../truth.log
posterior samples: combined.log
Use ranking for bins

Tree.height

Missed: 0

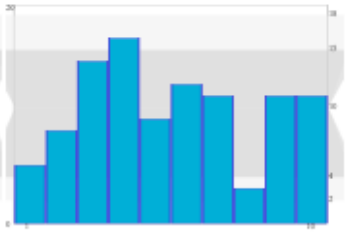
Tree.treeLength

Missed: 0



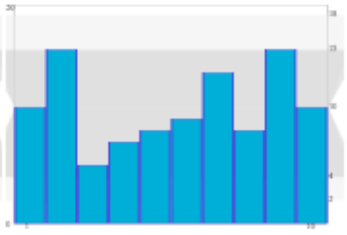
kappa

Missed: 0



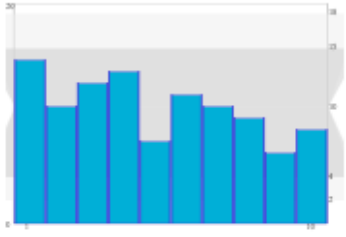
gammaShape

Missed: 0



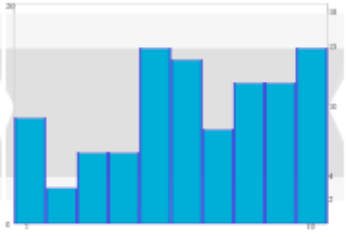
popSize

Missed: 0



CoalescentConstant

Missed: 0

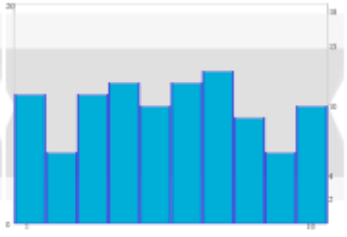


freqParameter.1

Missed: 1

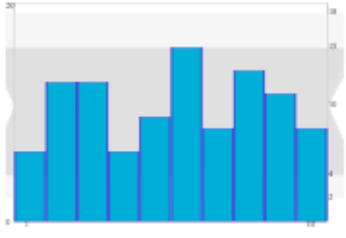
freqParameter.2

Missed: 0



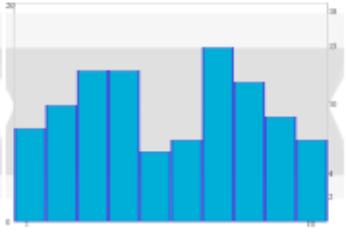
freqParameter.3

Missed: 0



freqParameter.4

Missed: 0



What makes phylogenetics hard?

Huge, non-standard parameter space

Difficult to represent (i.e., what are "ranks" for trees?)

Costly models

Runs take days, at a minimum. We **need** early stopping.

Mixture proposals

Contrast with Stan's "single" proposal.

THE END