

# Bayesian estimation of time-trees:

A journey through a strange land

---

Luiz Max Carvalho

Institute of Evolutionary Biology

Maxwell Institute seminar series 2017



THE UNIVERSITY  
*of* EDINBURGH

# Acknowledgements



Andrew Rambaut  
UoE



Marc Suchard  
UCLA



Guy Baele  
KU Leuven

# Plan for today

## Problem

What are trees and why are interested in them?

## Parameter space

What does the space we are trying to sample look like?

## MCMC in tree space

A journey through a strange land

## Preliminary results and perspectives

Performance analyses and open problems.

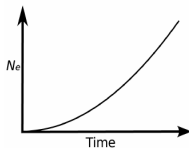
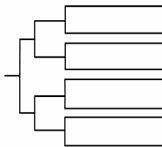
## Phylogenetics of fast-evolving viruses

Inferring spatial and temporal dynamics from genomic data:

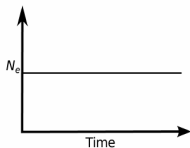
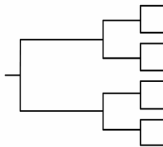
### Phylogenies\*!

\* plus complicated models

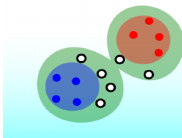
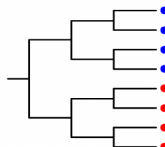
A  
Exponential Growth



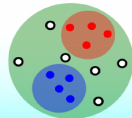
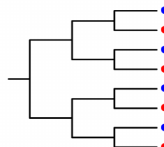
B  
Constant Population Size



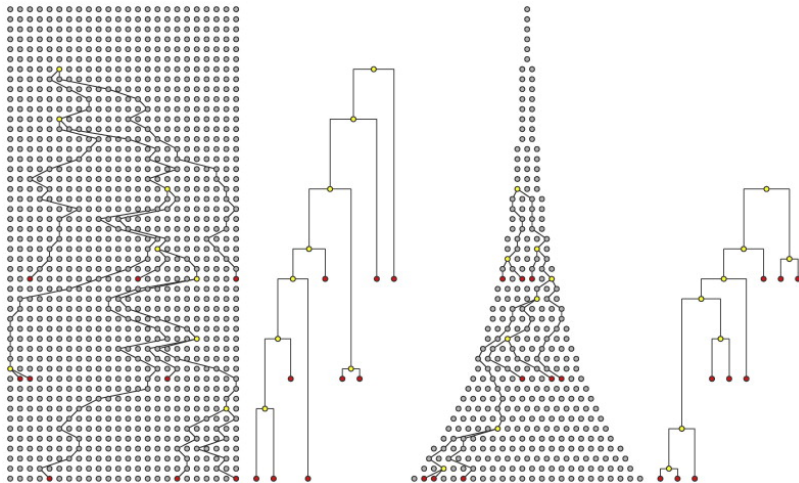
A  
Structured Host Population



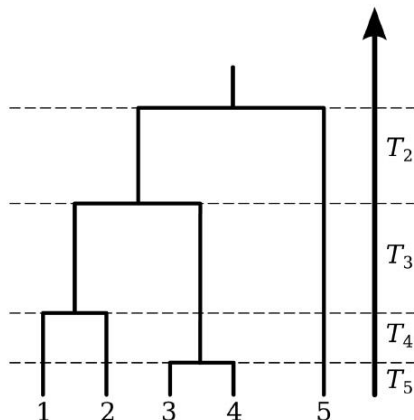
B  
Unstructured Host Population



# Trees and the coalescent



# Central object: time-calibrated trees



Let  $T_n$  denote the time for  $n$  lineages to *coalesce*, i.e., merge into one ancestral lineage, in a population of size  $N_e$ . Then:

$$\Pr(T_n = t) = \lambda_n e^{-\lambda_n t}$$

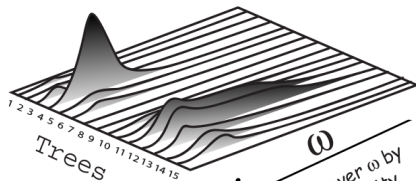
$$\lambda_n = \binom{n}{2} \frac{1}{N_e} = \binom{n}{2} \frac{1}{N_e \tau}$$

where  $N_e$  is the effective population size and  $\tau$  is the generation time. Let  $T_{\text{mrca}}$  denote the age of the most recent common ancestor:

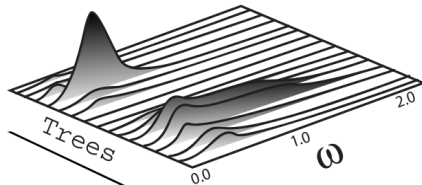
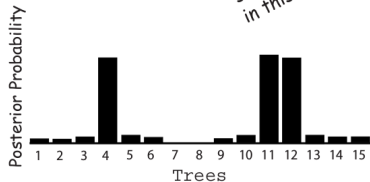
$$\begin{aligned} \mathbb{E}[T_{\text{mrca}}] &= \mathbb{E}[T_n] + \mathbb{E}[T_{n-1}] + \dots + \mathbb{E}[T_2] \\ &= 1/\lambda_n + 1/\lambda_{n-1} + \dots + 1/\lambda_2 \\ &= 2N_e \left(1 - \frac{1}{n}\right) \end{aligned}$$

Figure: Figure 4 from Volz et al. (2013).

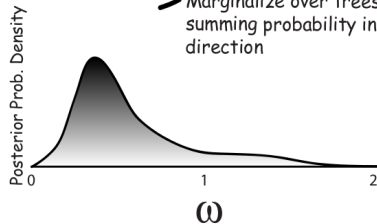
# Bayesian phylogenetics



← Marginalize over  $\omega$  by summing probability in this direction



→ Marginalize over trees by summing probability in this direction

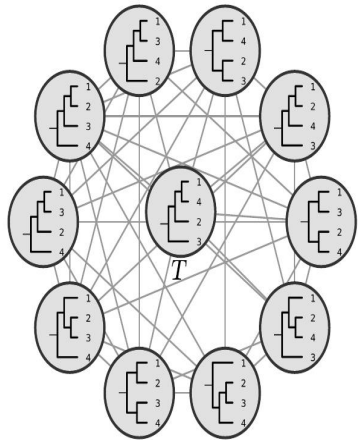
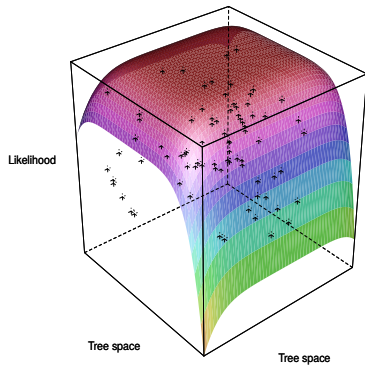


$$p(t, \mathbf{b}, \boldsymbol{\theta} | D) = \frac{f(D|t, \mathbf{b}, \boldsymbol{\theta})\pi(t, \mathbf{b}, \boldsymbol{\theta})}{\sum_{t_i \in T_n} \int_B \int_{\Theta} f(D|t_i, \mathbf{b}_i, \boldsymbol{\theta})\pi(t_i, \mathbf{b}_i, \boldsymbol{\theta})d\boldsymbol{\theta}d\mathbf{b}_i} \quad (1)$$

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

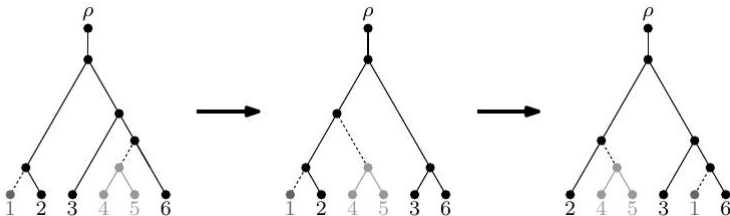


# Tree space: a strange land

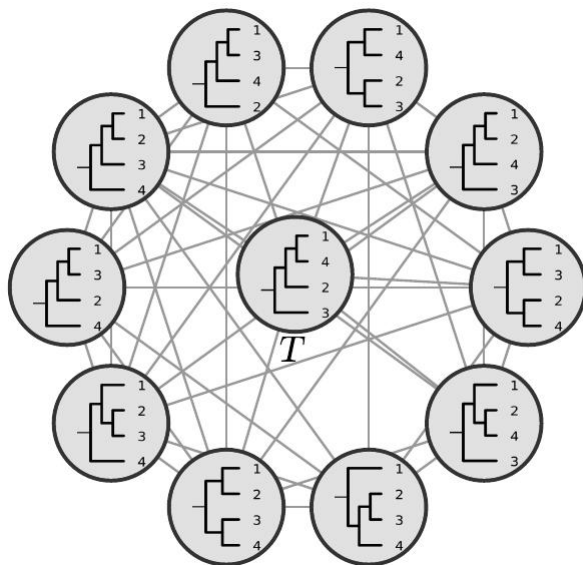


# Discrete tree space: tree surgery

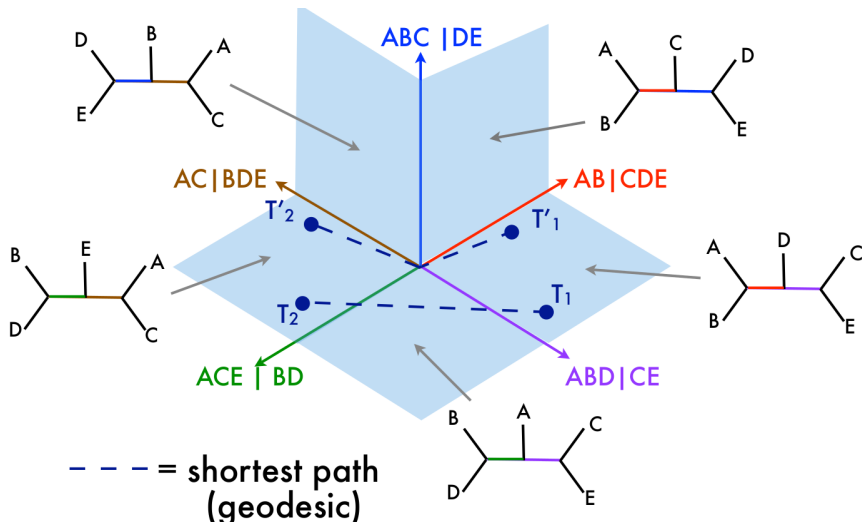
Subtree prune-and-regraft (SPR):



# Discrete tree space: SPR graph

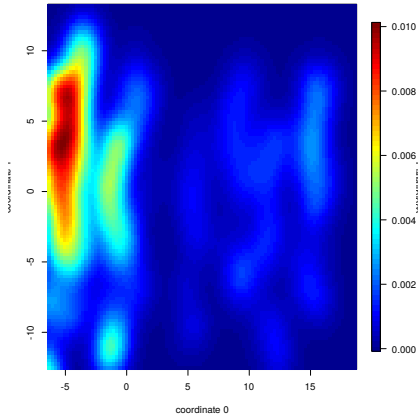


# Continuous tree space: BHV

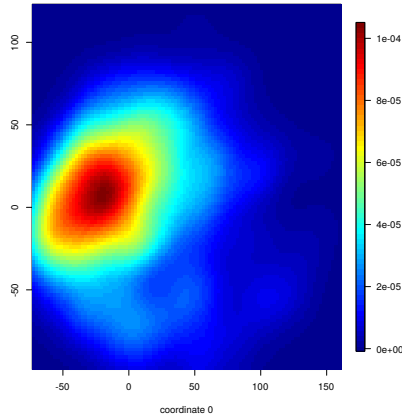


# Multi-dimensional scaling

Topology



Topology + branches



- ⊙ Non-standard, **huge** parameter space;
- ⊙ No canonical representation
- ⊙ Tip (leaf) heights impose constraints.

Open problems:

- Random walks on the SPR graph (and others);
- Useful representation for time-trees;

# Metropolis-Hastings for trees

General MH setup.

Let  $\tau = (t, \mathbf{b})$  denote a tree with topology  $t$  and branch lengths  $\mathbf{b}$ . For two trees  $\tau$  and  $\tau'$ , denote the transition kernel by  $q_\gamma(\tau|\tau') := Pr(\tau' \rightarrow \tau|\gamma)$ .

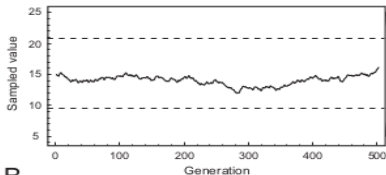
Accepting with probability

$$A_\gamma(\tau|\tau') = \min \left( 1, \frac{p(\tau', \boldsymbol{\theta}|D)q_\gamma(\tau|\tau')}{p(\tau, \boldsymbol{\theta}|D)q_\gamma(\tau'|\tau)} \right)$$

leads to the desired target.

# Exploring parameter space: **mixing**

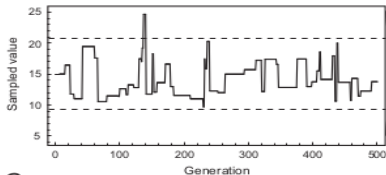
A



Target distribution

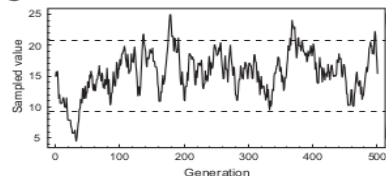
Too modest proposals  
Acceptance rate too high  
Poor mixing

B



Too bold proposals  
Acceptance rate too low  
Poor mixing

C

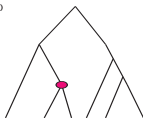


Moderately bold proposals  
Acceptance rate intermediate  
Good mixing



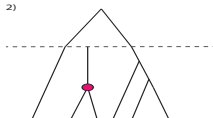
# Height-constrained kernels: SubTreeLeap

1)



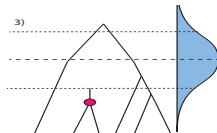
Pick a node

2)



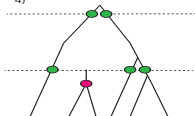
Disconnect its parent

3)



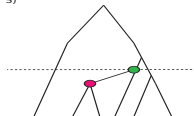
Draw a new height from a normal centred on old height of parent. Also consider the symmetrical height above or below the old height.

4)



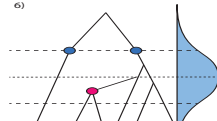
Pick uniformly from branches subtending that height and the symmetrical height above or below (in this case 5).

5)



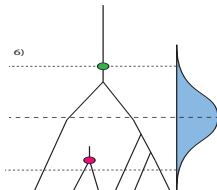
Attach parent to the chosen location.

6)



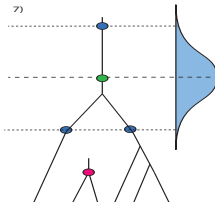
Hastings ratio: ratio of reverse probability (1 / number of reverse locations, i.e., 1/2) to forwards probability (i.e., 1/5).  
Hastings ratio = 5 / 2

6)



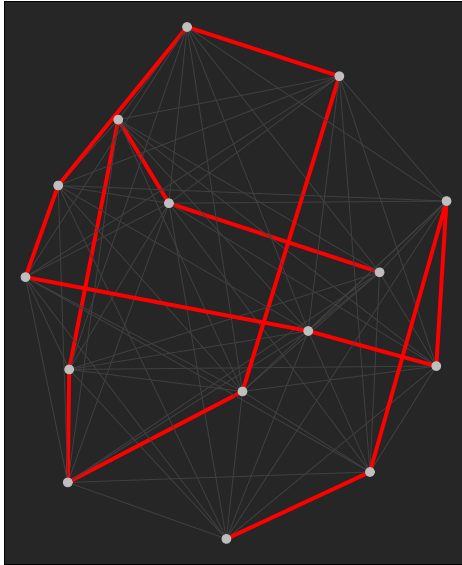
There is always at least 1 target location (above the root).

7)



In this case the HR would be 1/3

# SubTreeLeap is Hamiltonian\* on $T_n$



# Quantifying exploration

- ⊙ MDS;
- ⊙ Clade – aka subtree – frequencies;
- ⊙ Clade switching;
- ⊙ Effective sample size (ESS) of continuous parameters.

# Clade “space”

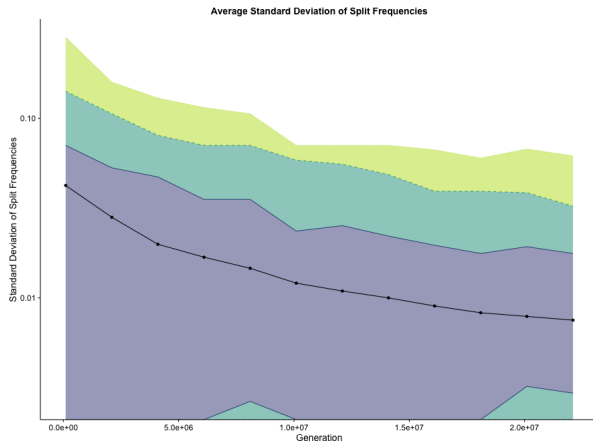
A clade  $c$  is any collection of leaves  $s_1, s_2, \dots, s_n$  such that they share a common ancestor in the tree. For  $n$  taxa (leaves) there are  $A(n) = 2^{n-1} - 1$  possible clades.

Let  $\mathbf{X}_i = \{X^{(1)}, X^{(2)}, \dots, X^{(n)}\} \in [0, 1]^n$  be a collection of samples from a Markov chain such that  $X_i^{(j)} = 1$  if clade  $i$  was sampled in the  $j$ -th iteration and 0 otherwise. Also, for  $s_i = \sum_k X_i^{(k)}$  we call  $f_i^c = s_i/n$  the *frequency* of clade  $i$ .

# Clade frequencies – deviation

$$\delta := \max_{1 \leq i \leq A(n)} \frac{|f_i^c - r_i^c|}{r_i^c},$$

where  $f^c$  and  $r^c$  are the observed and true clade frequencies.



# Clade switching

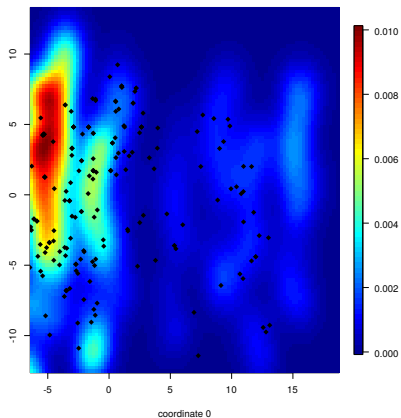
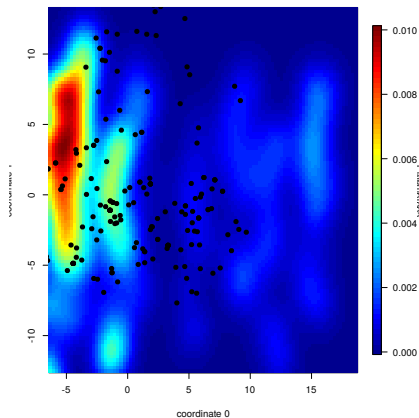
Let  $m_i = \min(n - s_i, s_i)$ , it can be shown that the maximum number of transitions that can be observed from  $X_i$  is either  $J_i = 2m_i$ .

Let  $\delta_i = \Delta(X_i)$ , where  $\Delta(\cdot)$  a function that counts the number of state transitions in  $X_i$ . Then  $\sigma_i = \delta_i / J_i \in [0, 1]$  is a score that measures the relative efficiency of sampling by comparing how many transitions happened compared to the theoretical maximum.

# Traversing tree space – Topology

Default kernels

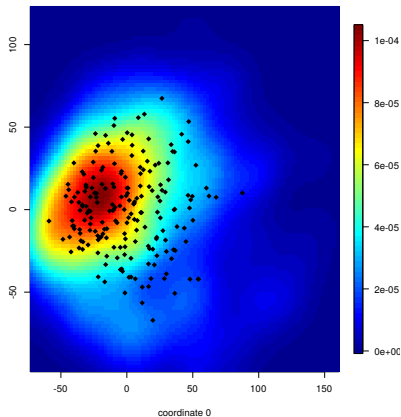
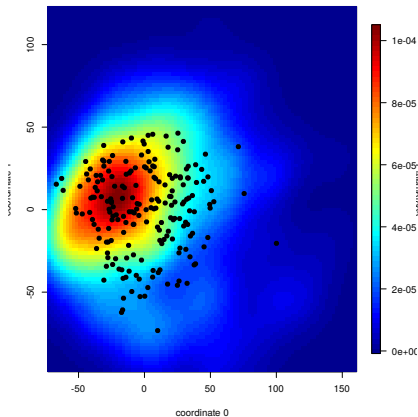
STL



# Traversing tree space – Topology + branch lengths

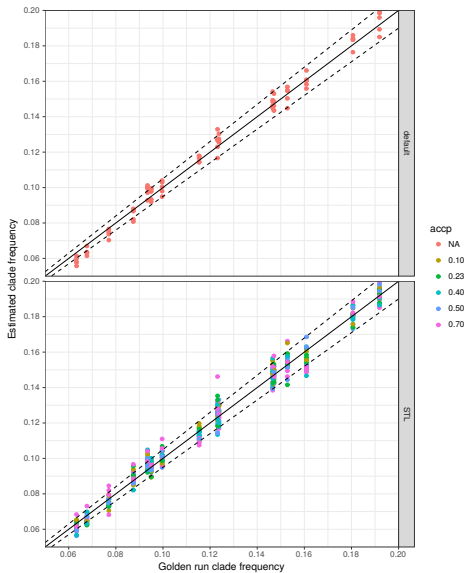
Default kernels

STL

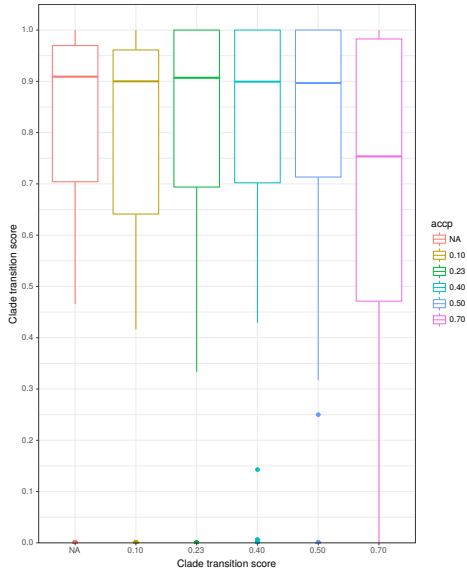




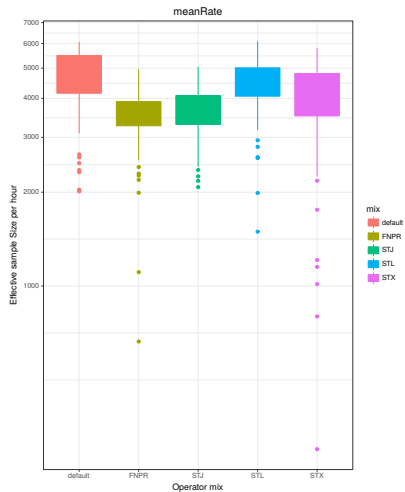
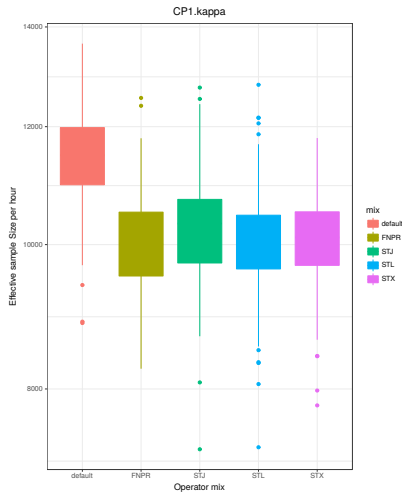
# Clade frequencies – example results



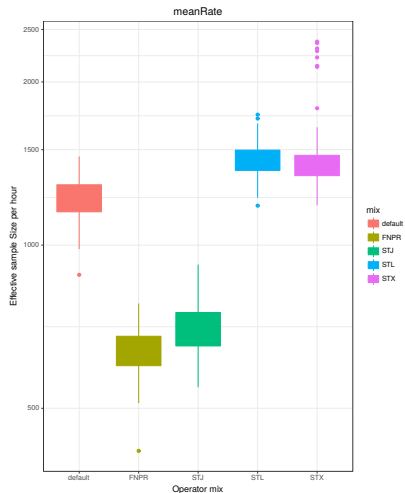
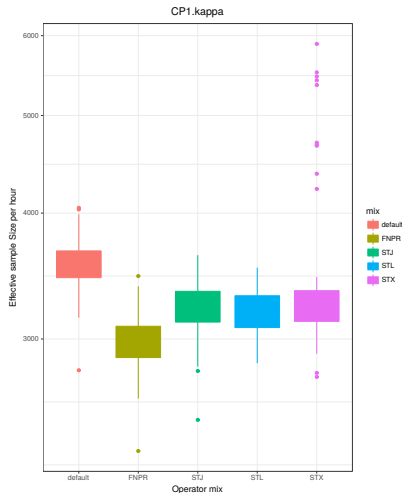
# Clade switching – example results



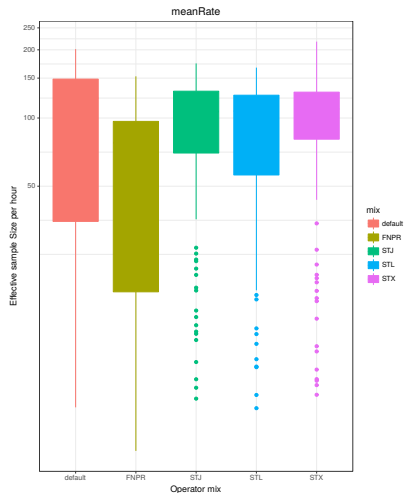
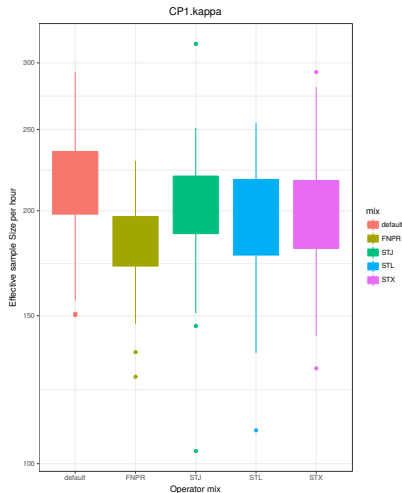
# Dengue 4 env (17 taxa, 1485 sites)



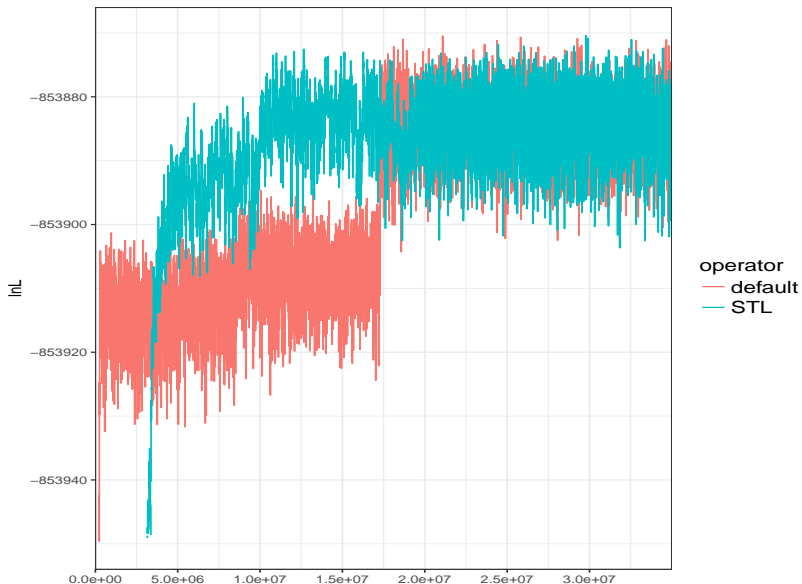
# RSVA G protein (35 taxa, 629 sites)



# YFV *prM/E* gene (71 taxa, 654 sites)



# Metazoans (contemporaneous, 55 taxa, 30257 AA sites)



SubTreeLeap seems to

- ⊙ explore topology space more thoroughly (and mix better in clade space);
- ⊙ facilitate sampling other parameters of interest **conditional** on the tree;

Open problems:

- Optimal scaling: what's the optimal acceptance probability?
- Can we construct even more efficient proposals? How to exploit structure?

Searching trees is **hard**

Complex, discrete and **HUGE** parameter space

---

<sup>1</sup>this talk is available [online](#)



Searching trees is **hard**

Complex, discrete and **HUGE** parameter space

Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

---

<sup>1</sup>this talk is available [online](#)

## Searching trees is **hard**

Complex, discrete and **HUGE** parameter space

## Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

## Adaptive moves are more efficient

Avoid wasting computing power

---

<sup>1</sup>this talk is available [online](#)

Searching trees is **hard**

Complex, discrete and **HUGE** parameter space

Height-preserving tree rearrangements are **good**

Use the extra information provided by the tip dates

Adaptive moves are more efficient

Avoid wasting computing power

Much more work is needed

We should prepare for an era of plenty

---

<sup>1</sup>this talk is available [online](#)

THE  
END