Developing StarBEAST2 for faster species tree inference and accurate estimates of substitution rates

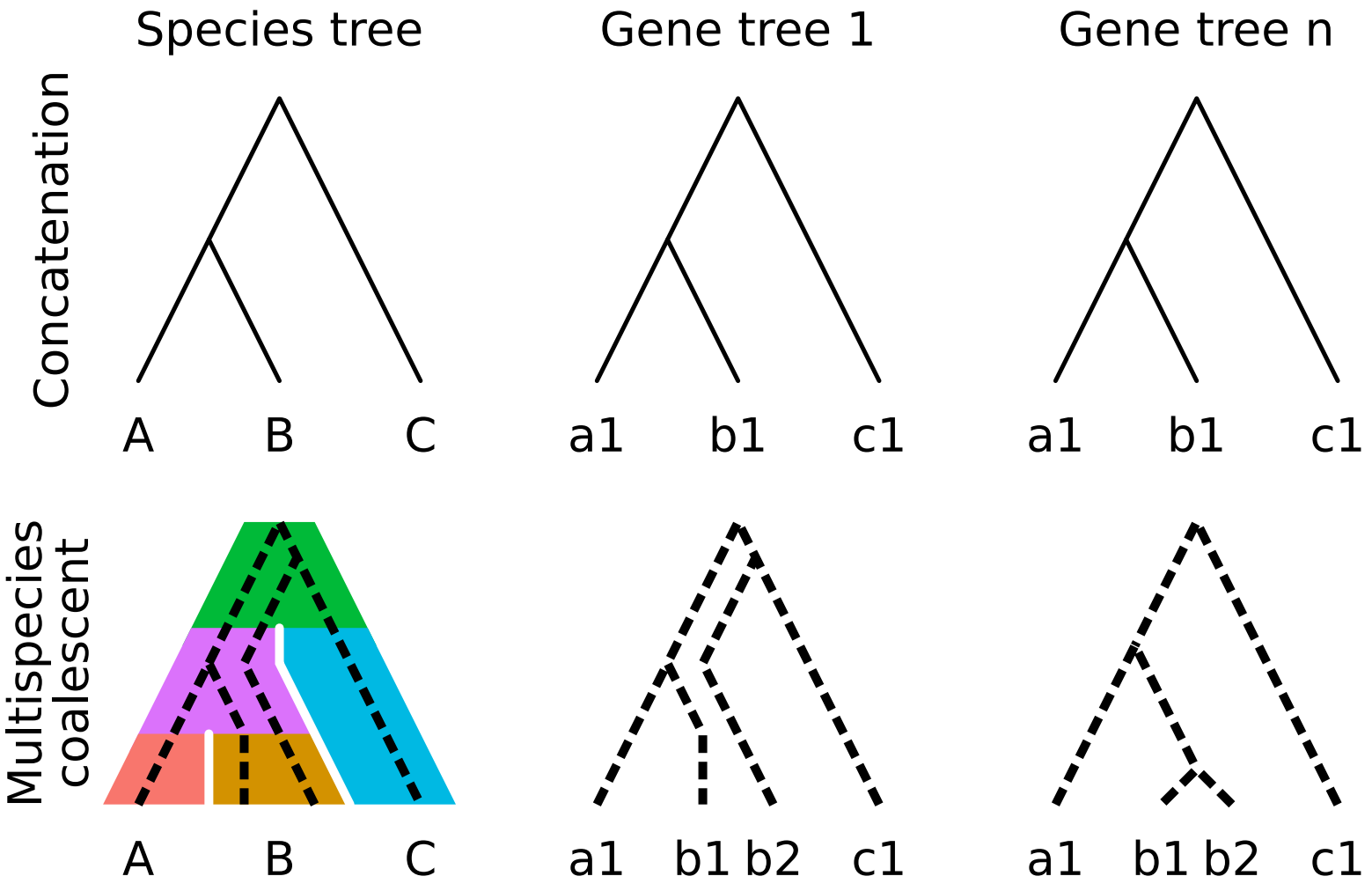
Huw A. Ogilvie[1,2]

[1]Research School of Biology
Australian National University

[2]Centre for Computational Evolution
University of Auckland

July 26, 2017

Models for inferring “species” trees

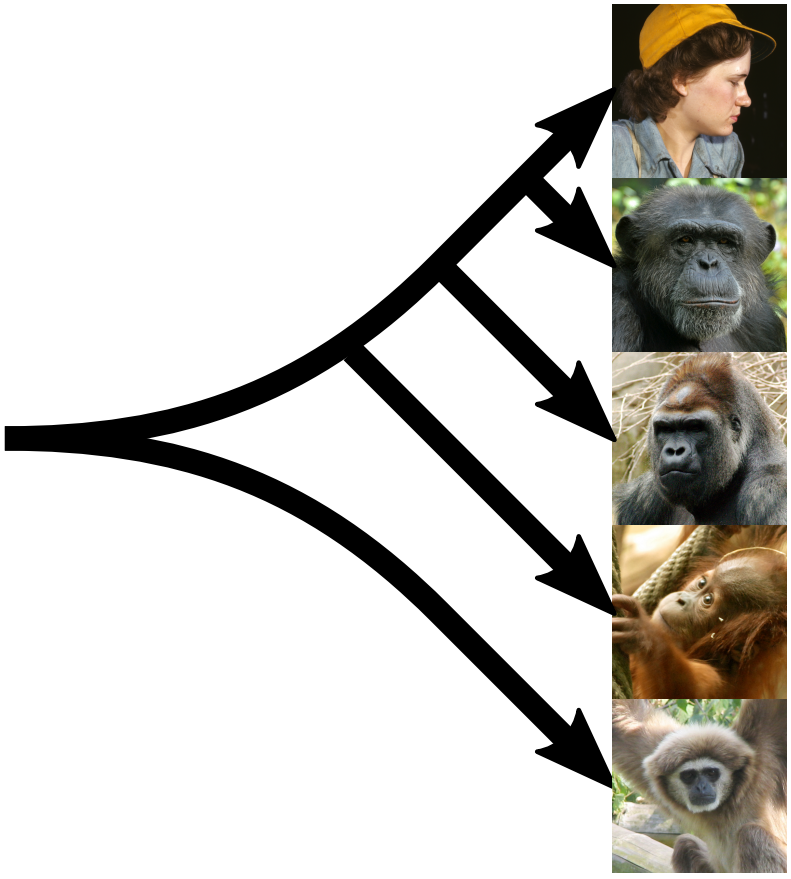


A note on the word “species”

In a multispecies coalescent (MSC) context, **species** means **independently evolving lineage**, or a genetically isolated population of some organism

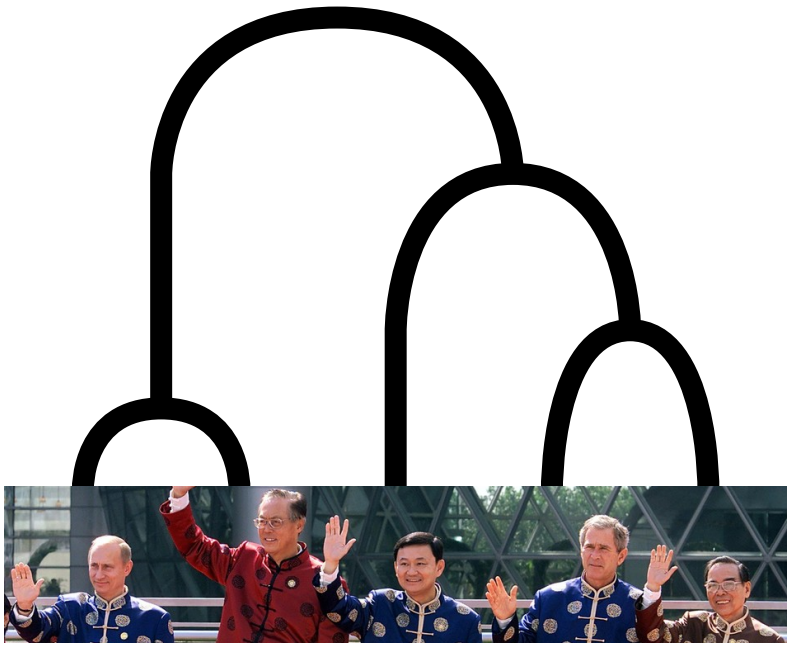
Introducing the multispecies coalescent

“Speciation” processes



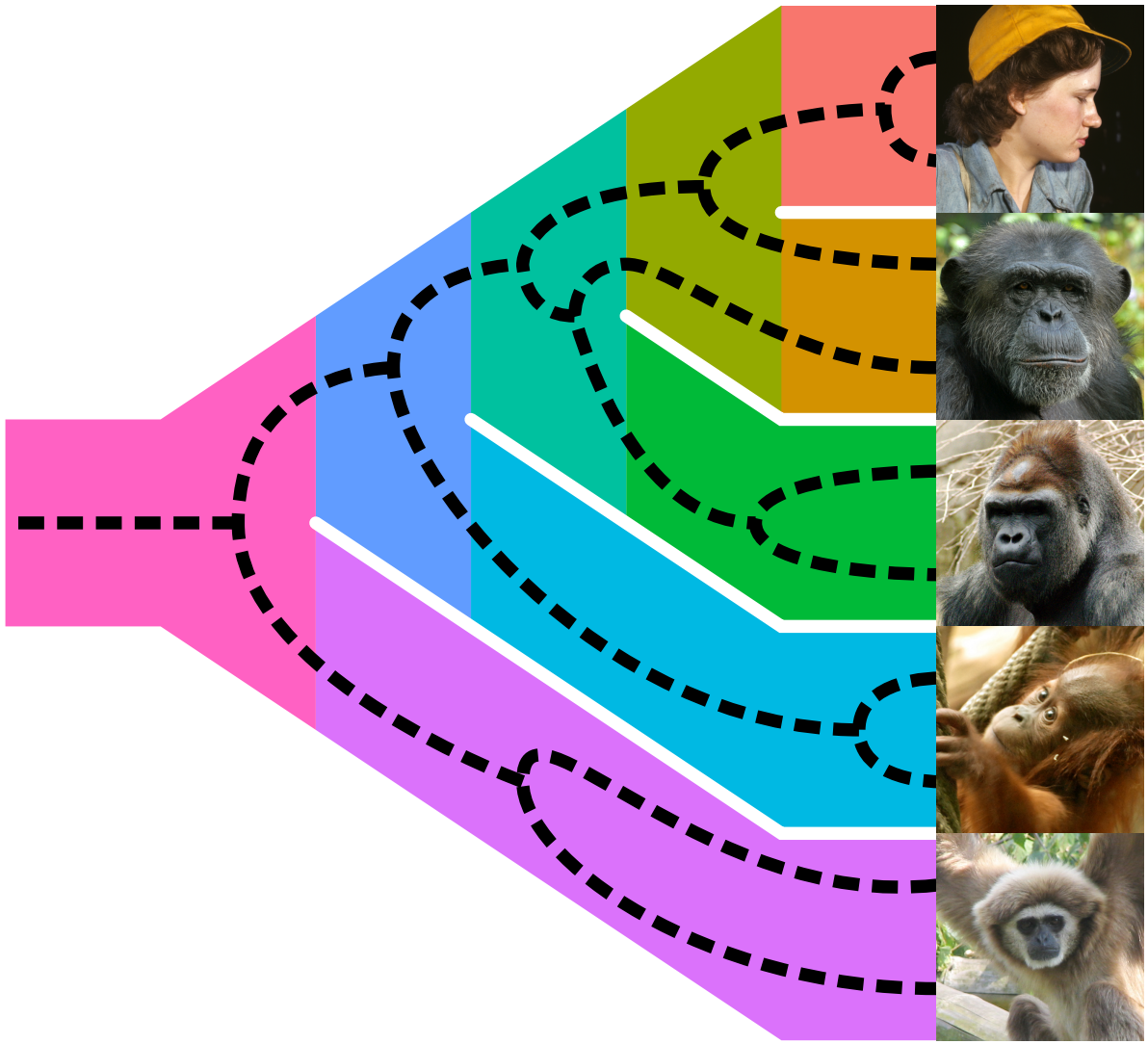
- Evolution of species (or populations)
- Birth-death process (speciation & extinction)
- Diversification rate $\text{Birth} - \text{Death} = \lambda$
- Extinction ratio $\text{Death} \div \text{Birth} = \nu$
- Forward in time

Coalescent processes

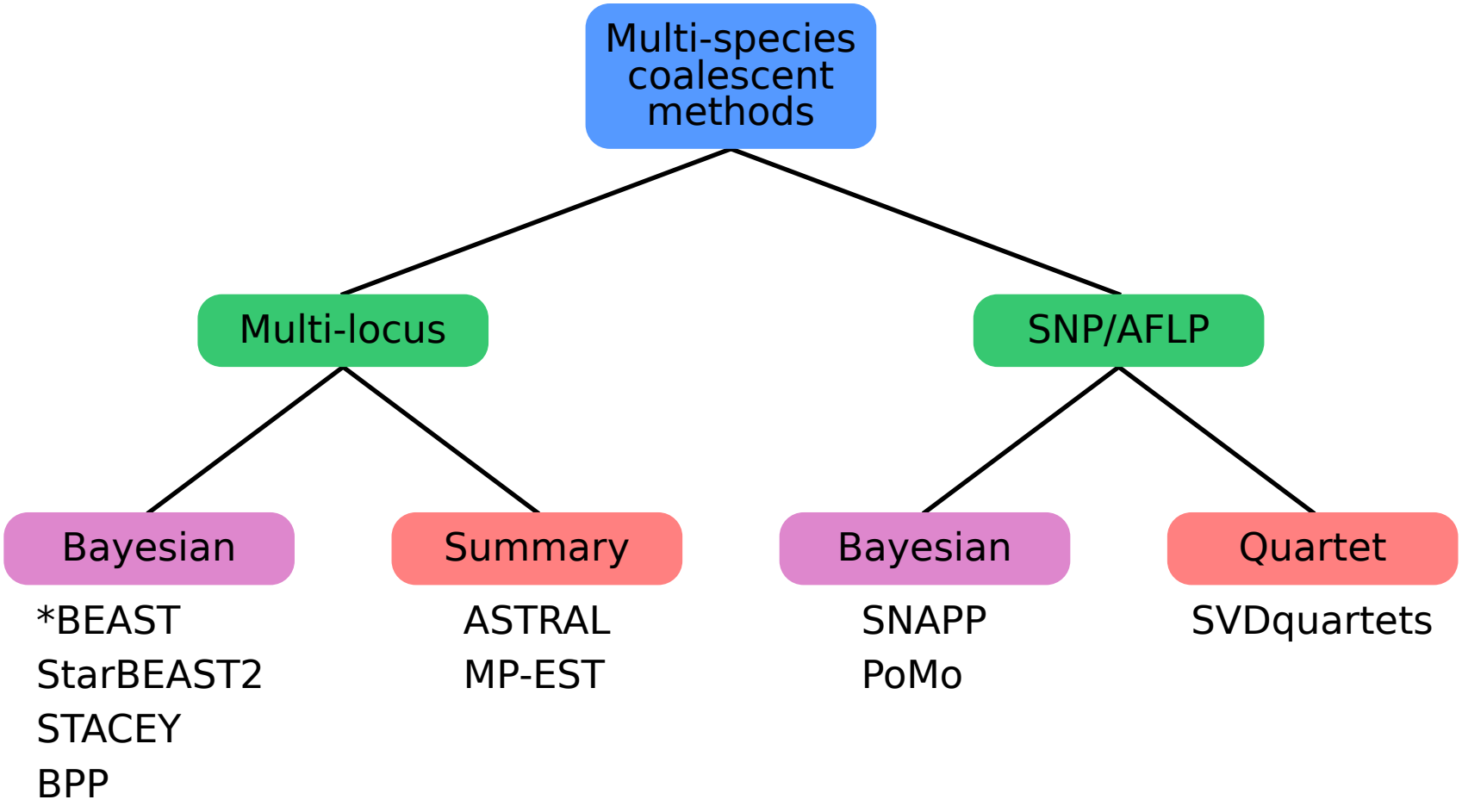


- Evolution of genes
- Dependent on effective population size N_e
- Constant, linear, exponential or stepwise (as in skyline plots)
- Backwards in time

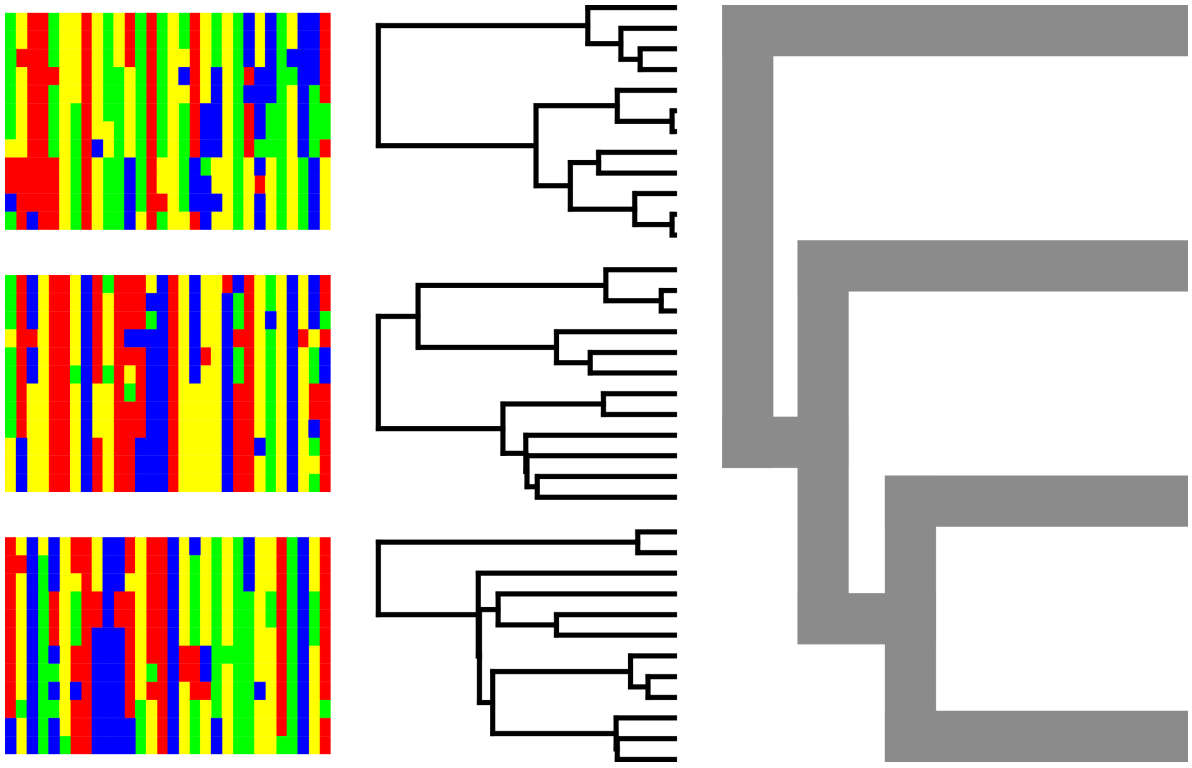
Putting them together



MSC methods


Multi-species coalescent methods
Multi-locus
SNP/AFLP
Bayesian
Summary
Bayesian
Quartet
*BEAST
StarBEAST2
STACEY
BPP
ASTRAL
MP-EST
SNAPP
PoMo
SVDquartets

Bayesian multi-locus MSC



$$\Pr(S|D) = \prod_{i=1}^{n} \Pr(d_i|g_i) \cdot \Pr(G|S) \cdot \Pr(S) \cdot \frac{1}{\Pr(D)} \qquad (1)$$

StarBEAST2

- Multilocus multispecies coalescent (MSC) inference
- Extends the MSC to estimate per-species clock rates
  - Uncorrelated and random local clocks implemented
- Roughly 13$\times$ faster than *BEAST
- Available as a package for the BEAST 2 platform

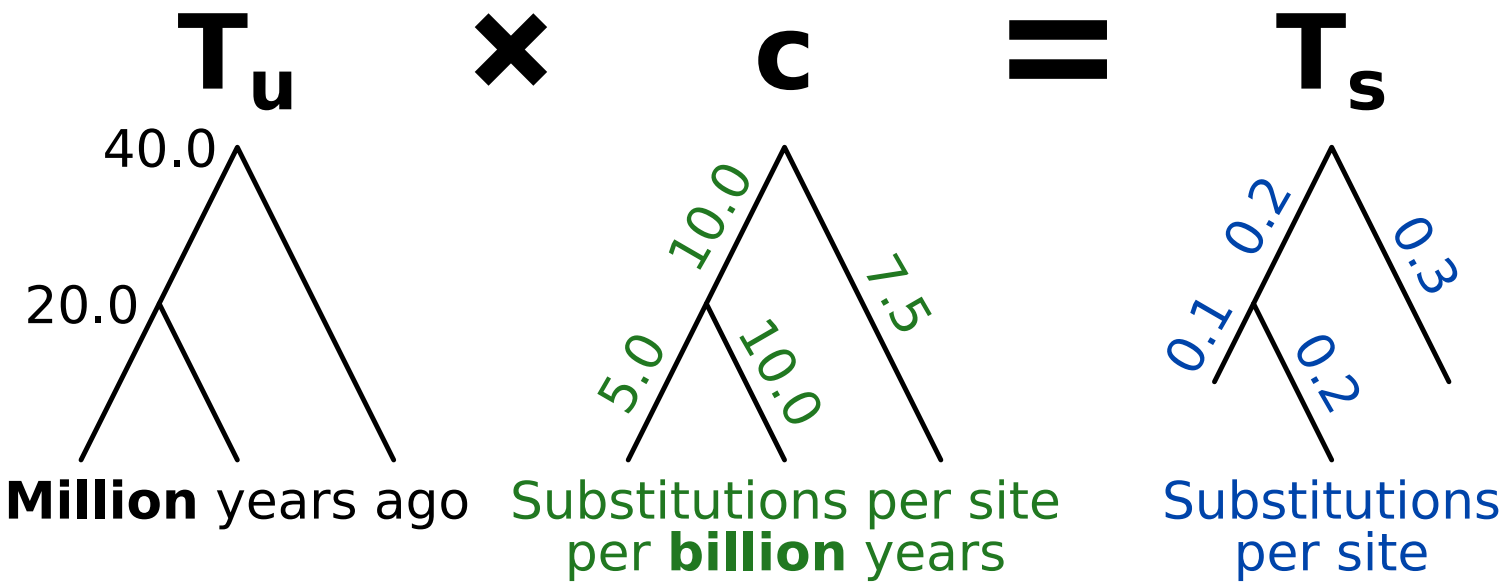
Ogilvie, Bouckaert and Drummond (2017)
Molecular Biology and Evolution

Estimating per-species molecular clock rates

Why clock rates are needed

Divergence times necessitate estimating a tree with branch lengths in units of time T_u

Phylogenetic likelihood $L(D|T_s)$ is the likelihood of multiple sequence alignments D given a tree with branch lengths in units of substitutions per site T_s

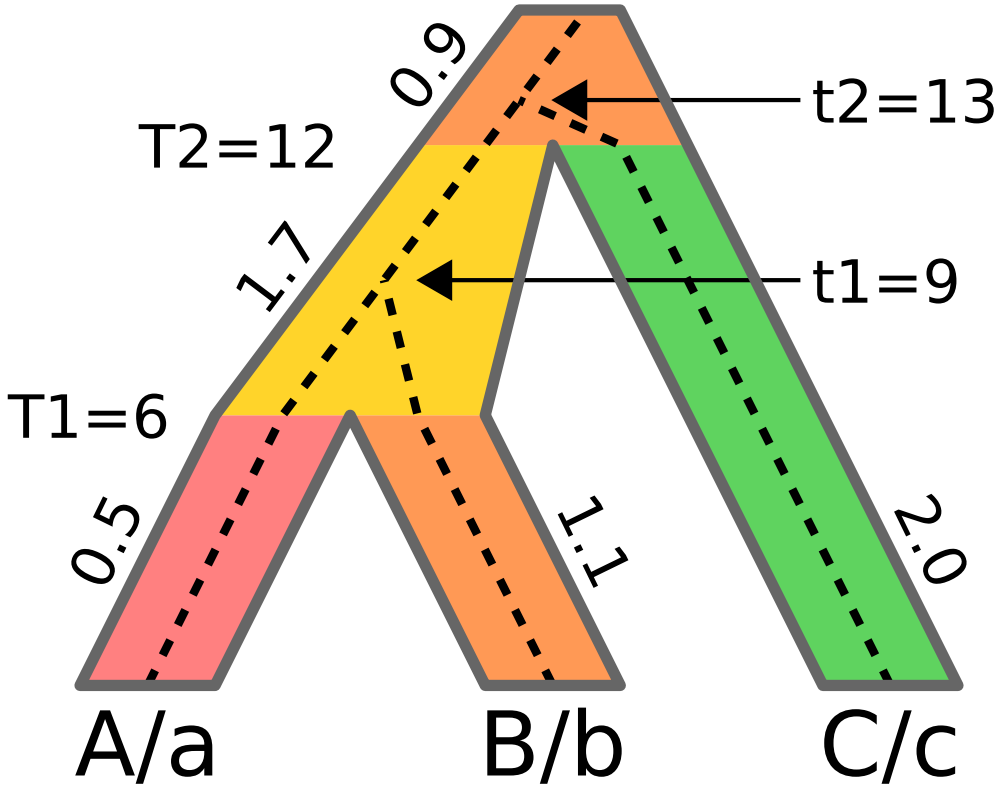


Per-species substitution rates

- Mutation rates vary between species, e.g. because of larger or smaller numbers of germline duplications per year
- Selection and drift pressures vary between species, e.g. because of varying effective population size (nearly neutral theory)

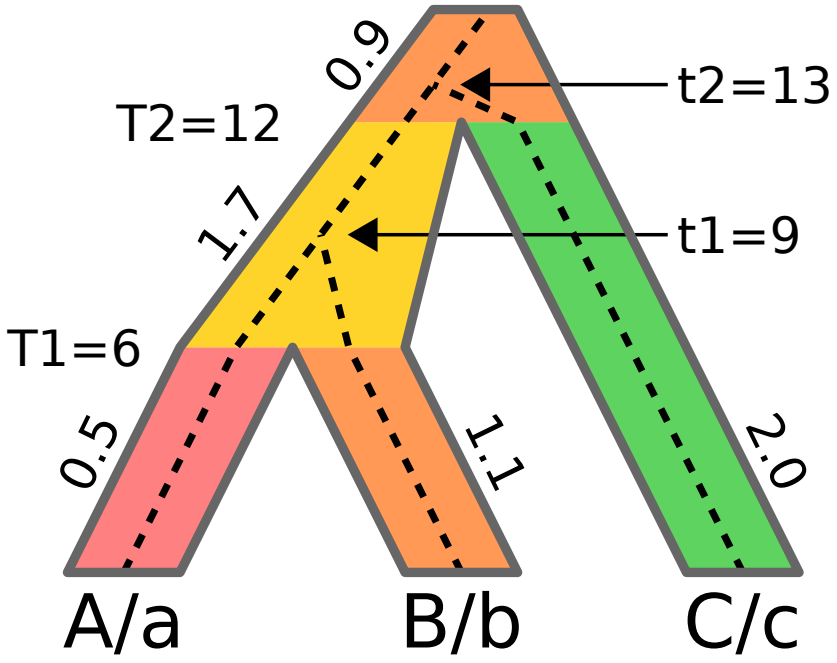
See Bromham (2009) “Why do species vary in their rate of molecular evolution?” Biology Letters

MSC clock rates I



$$rate_{branch} = rate_{locus} \sum_{species}^{n} \left(rate_{species} \times \frac{overlap_{species,branch}}{length_{branch}} \right) \quad (2)$$

MSC clock rates II

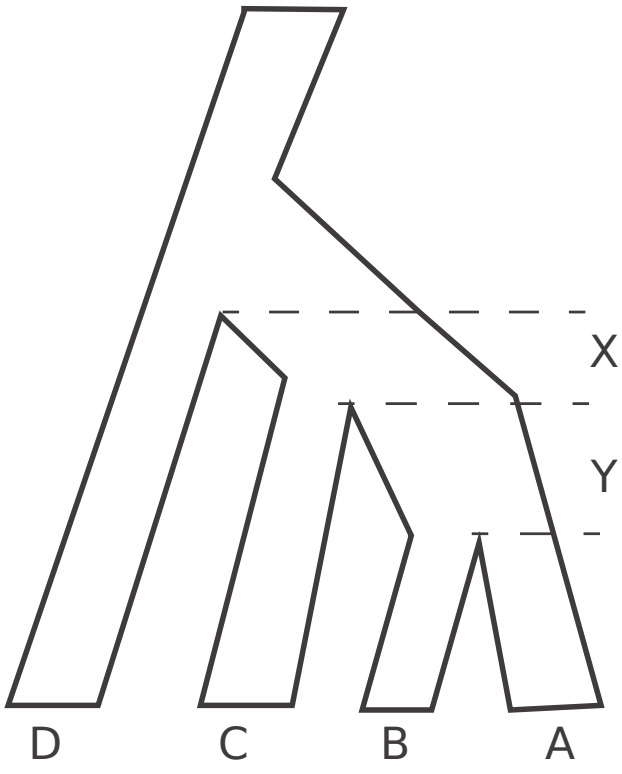


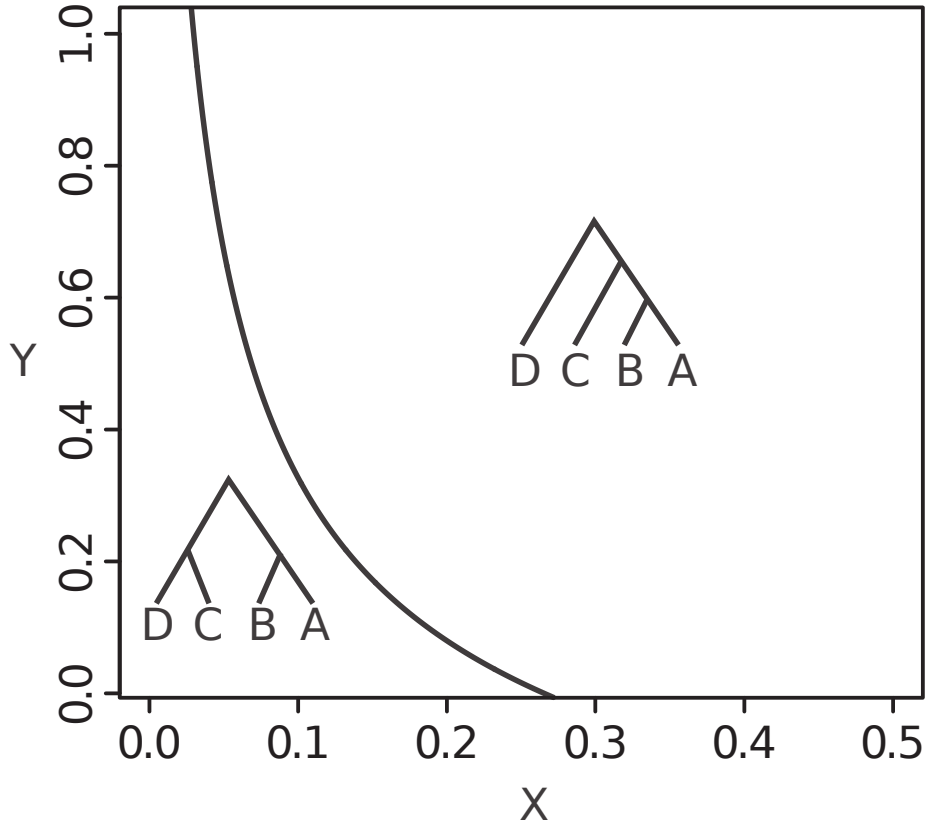
$$rate_a = 10^{-9}\left(0.5 \times \frac{6.0}{9.0} + 1.7 \times \frac{3.0}{9.0}\right) = 0.9 \times 10^{-9} \quad (3)$$

$$rate_b = 10^{-9}\left(1.1 \times \frac{6.0}{9.0} + 1.7 \times \frac{3.0}{9.0}\right) = 1.3 \times 10^{-9} \quad (4)$$

Comparing the accuracy of concatenation *vs* the MSC in theory

Topology accuracy: anomaly zone





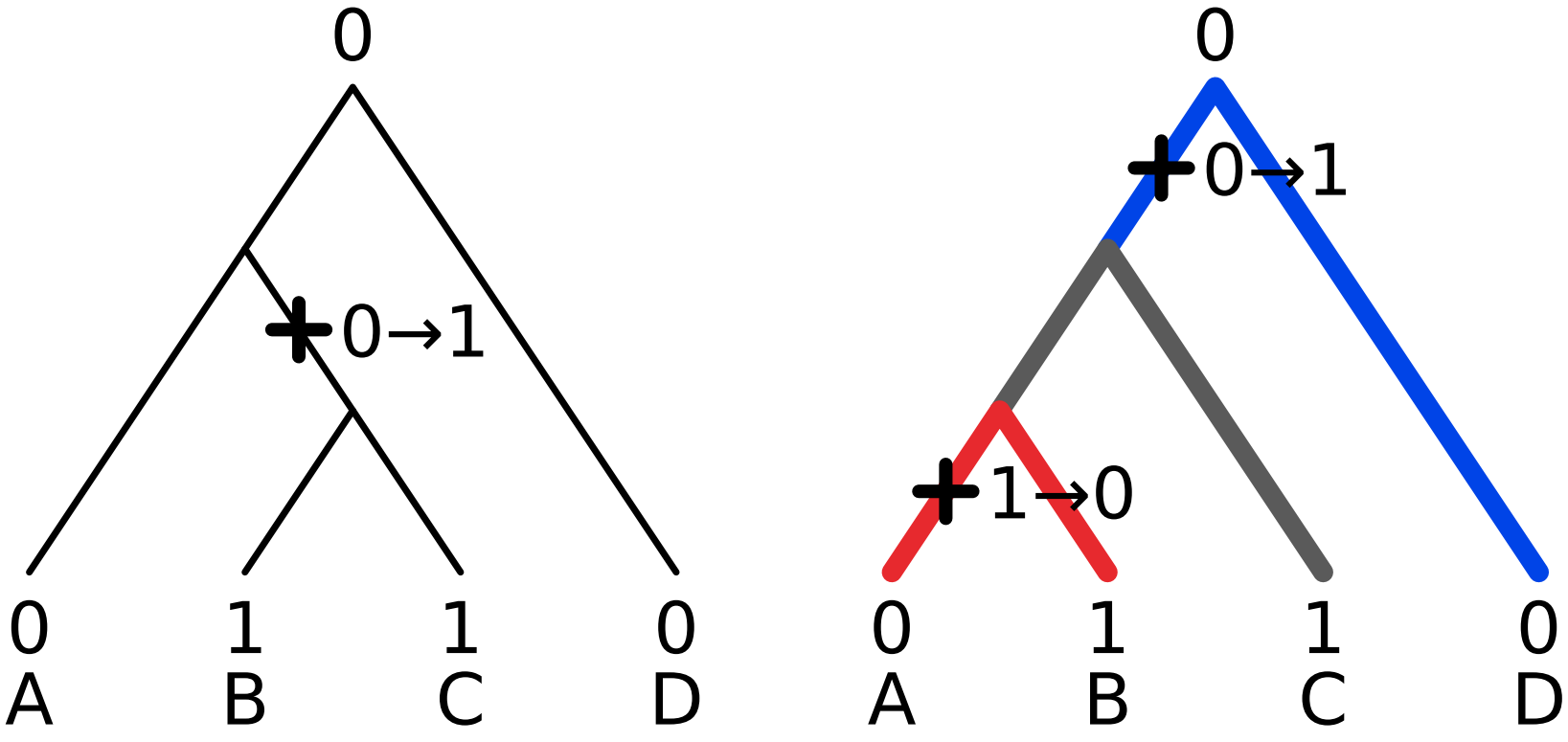
Linkem, Minin & Leaché (2016) Systematic Biology

Also: Mendes & Hahn (2017) “Why concatenation fails in the anomaly zone”

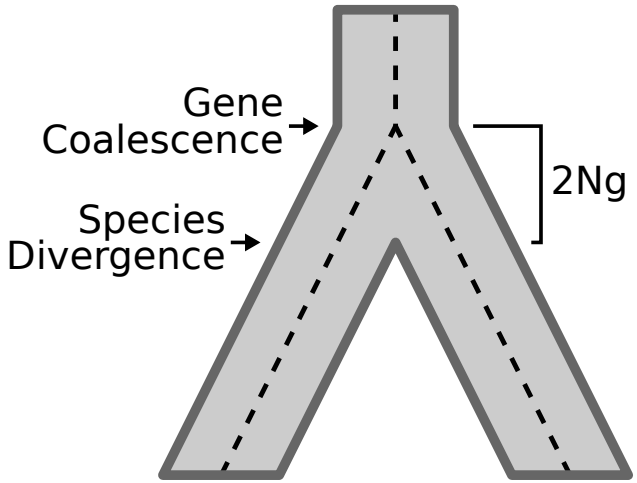
Clock rate accuracy: SPILS

“Substitutions Produced by Incomplete Lineage Sorting”

First reported by Mendes & Hahn (2016) in Systematic Biology



Divergence time accuracy: coalescence times

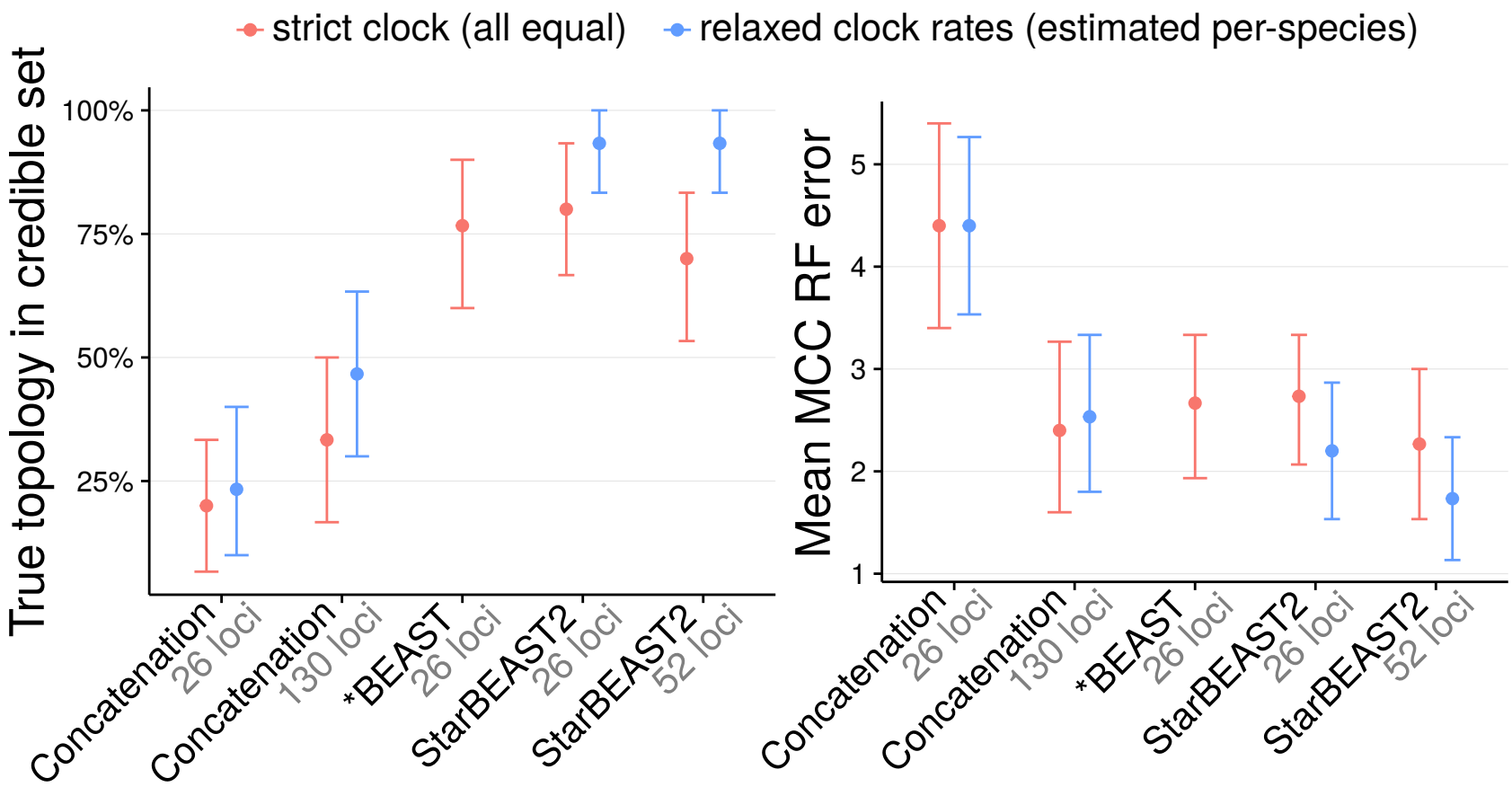


- Human-chip divergence time $\approx 4my$
- Ancestral effective population size $N \approx 50000$
- Ancestral generation time $g \approx 20y$
- $2Ng = 2 \cdot 50000 \cdot 20y = 2my$
- Estimated divergence time will be $4my + 2my = 6my$

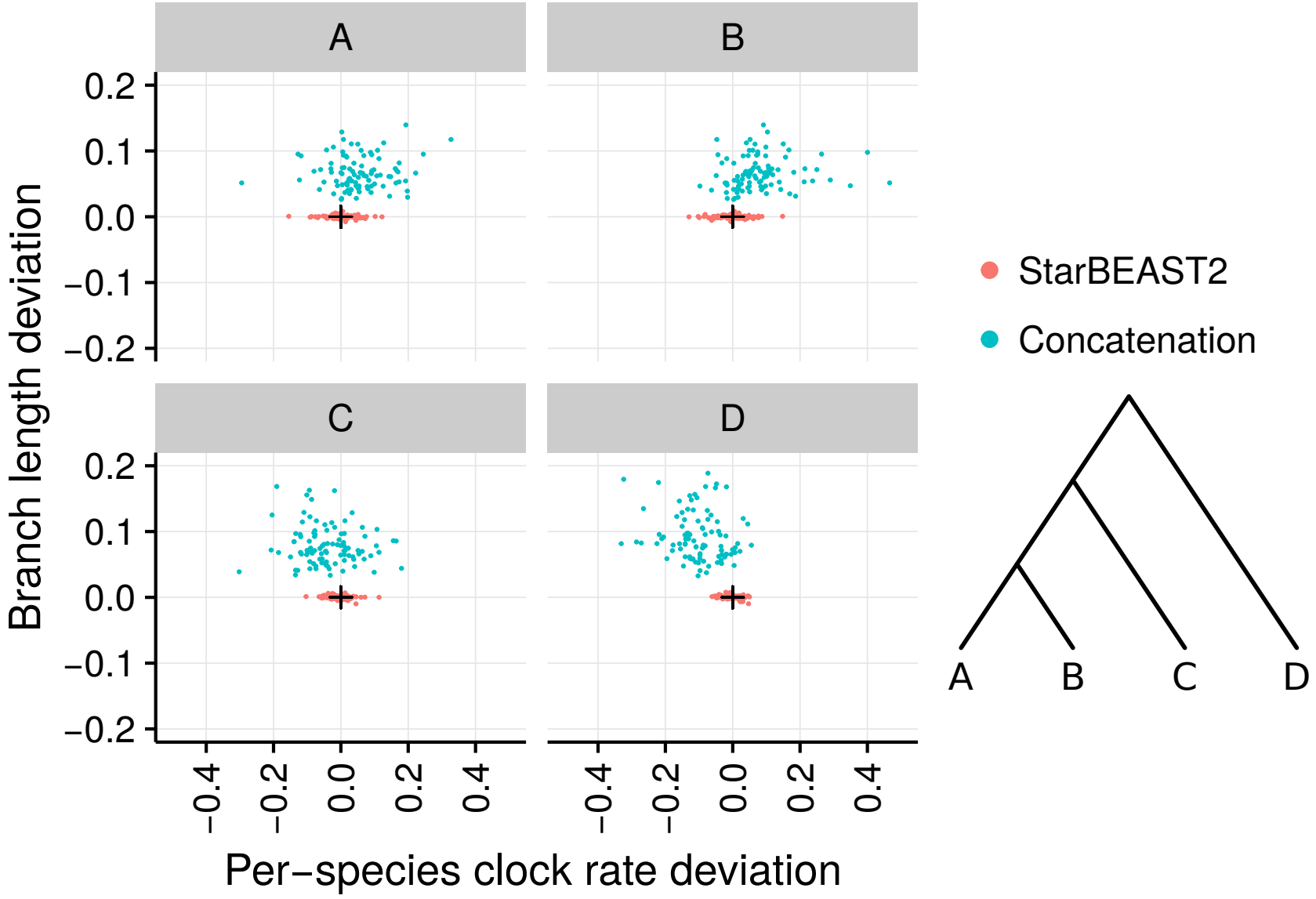
Hobolth *et al.* (2011) Genome Research

Comparing the accuracy of concatenation *vs* the MSC by simulation

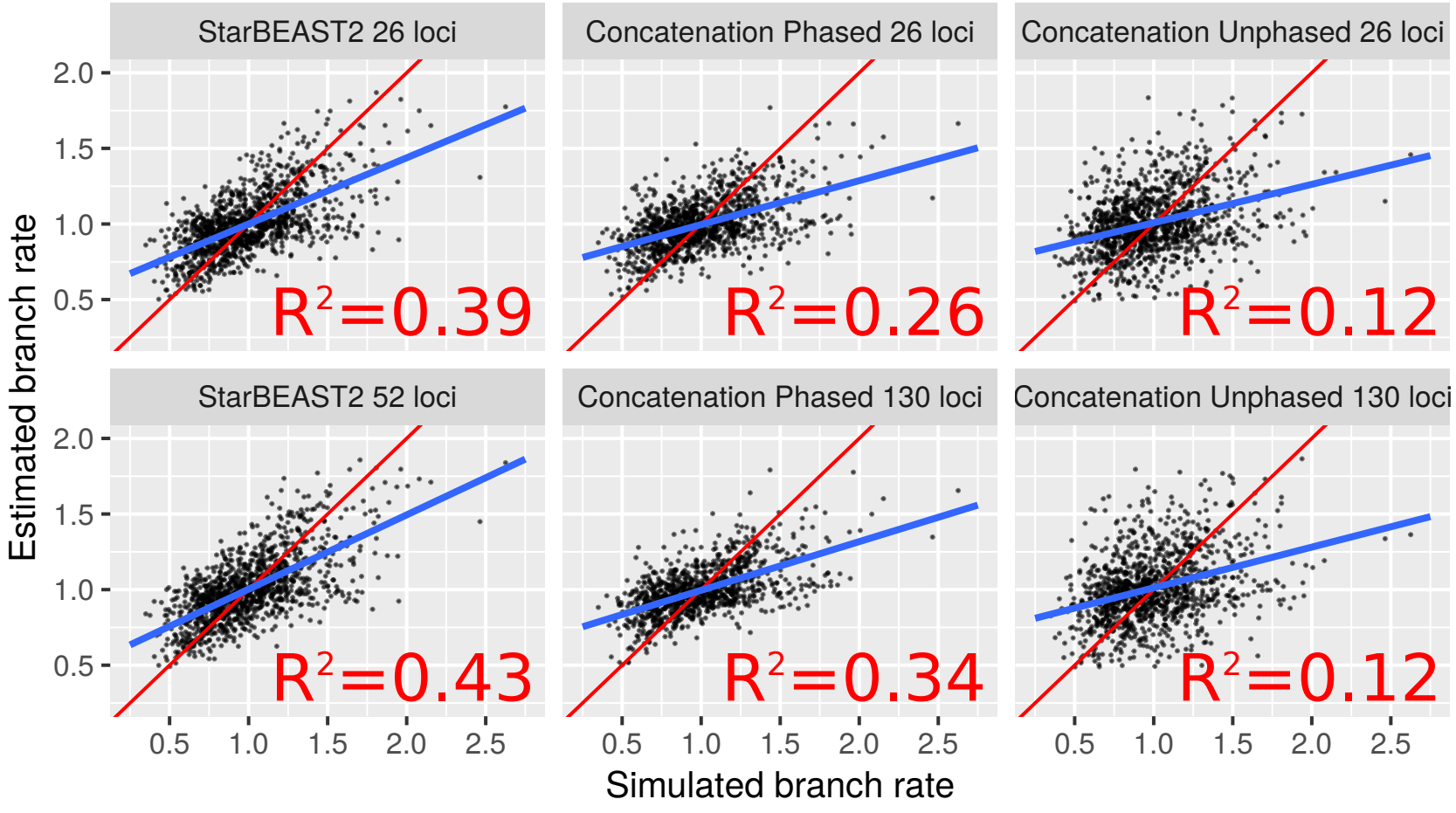
Topology accuracy: concatenation vs MSC



Rates and times: concatenation vs MSC



Rates II: concatenation vs MSC



Assorted whinging

"[*BEAST] analysis can be problematic for UCE data sets in that the large number of loci precludes use of many coalescent-based species-tree methods" — Streicher *et al.* (2015)

"We did not use [*BEAST] because [it] has shown poor performance with phylogenomic-scale data" — Pyron *et al.* (2014)

"[Using] *BEAST...is not possible given these data...because the number of loci is prohibitive" — Mandel *et al.* (2015)

Is it really better to get a wrong answer using all loci, than to get the right answer with a random subset?

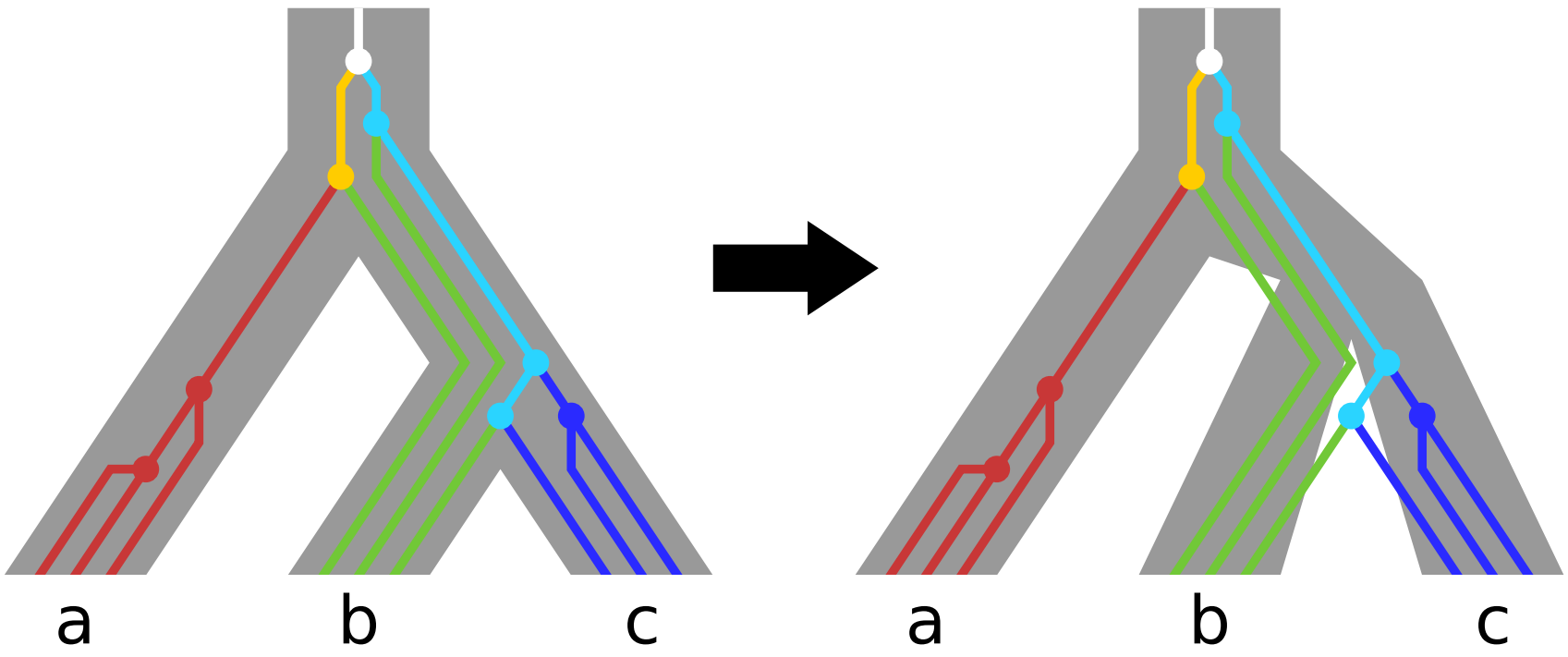
Many people ask this question

Why is StarBEAST2 faster than *BEAST?

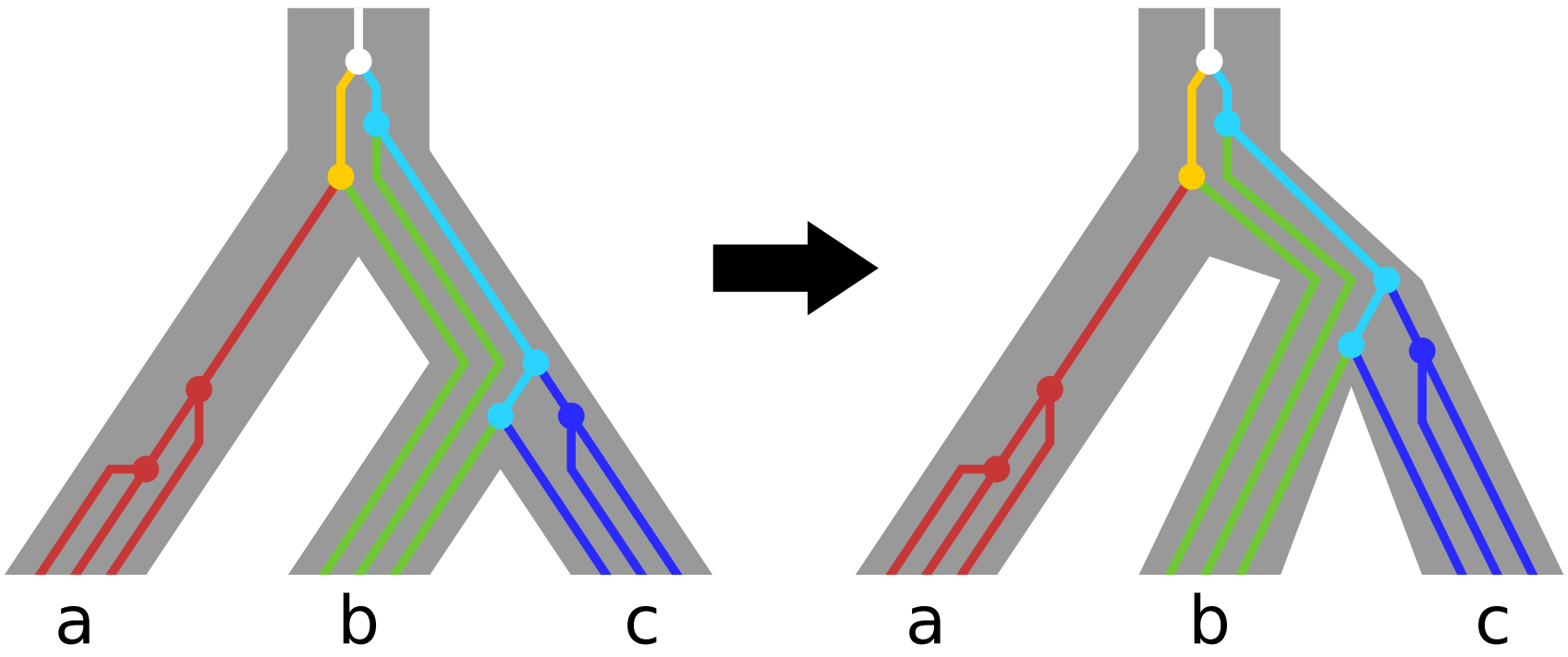
Improving mixing

- Analytical integration of population sizes
- Better tuning of default MCMC operator weights
- Novel MCMC operators to change node heights

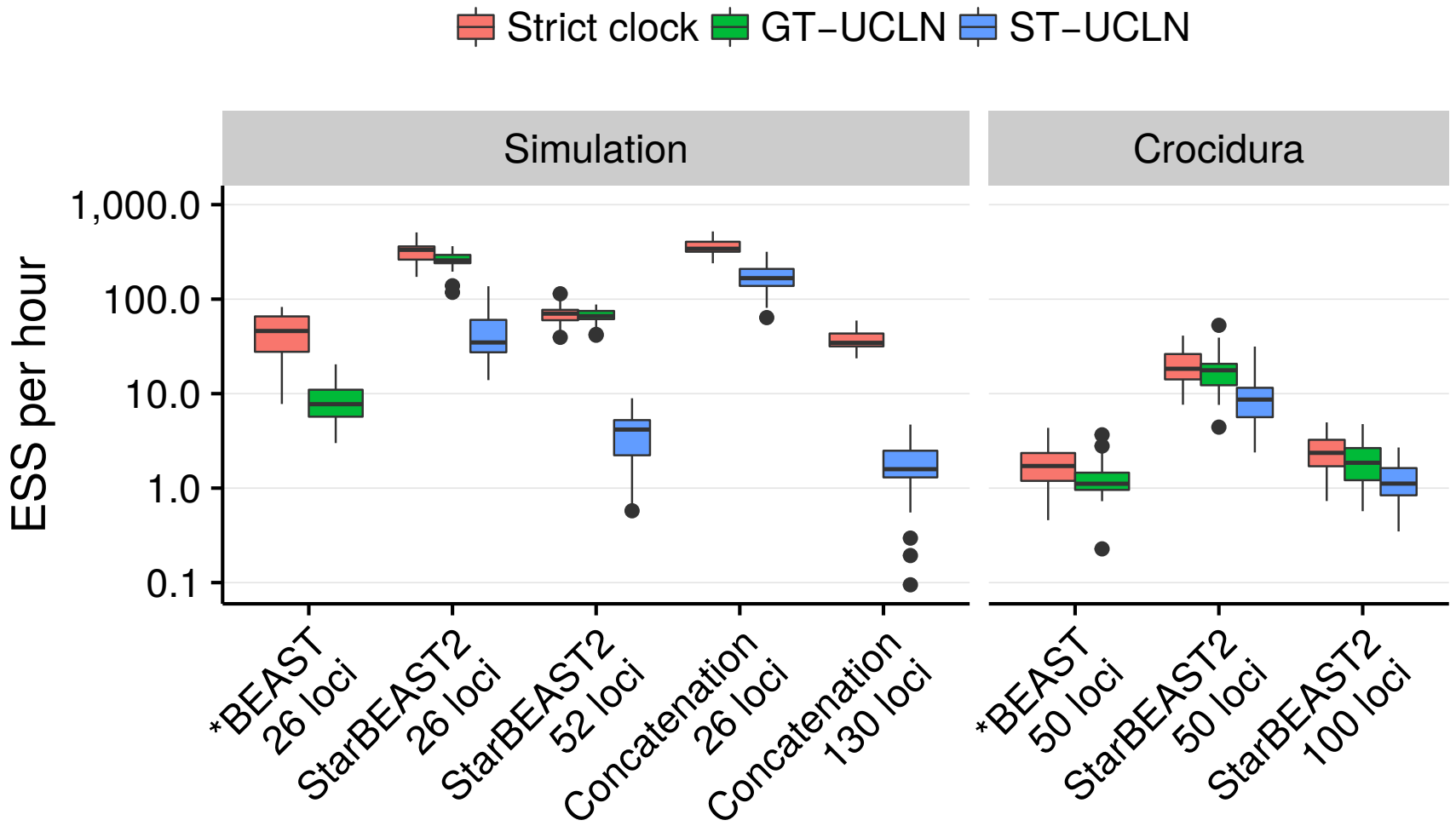
Naïve height changing proposals



Coordinated height changing proposals



StarBEAST2 performance



Next steps

- Apply to Eugongylus-group skinks, infer substitution rate shifts correlated with environmental changes
  — with Craig Moritz, Sally Potter and Jason Bragg
- Combine the MSC with morphological, fossilized birth death (FBD) and sampled ancestor (SA) models
  — with Alexei Drummond, Nick Matzke, Tanja Stadler and Tim Vaughan
  — Coming in the next version of StarBEAST2 (v14) available later this year

Acknowledgements

Auckland
Alexei Drummond
Tim Vaughan
Remco Bouckaert
Joseph Heled

ANU (Canberra)
Craig Moritz
Jason Bragg