Multi-locus multi-species coalescent phylogenetic inference

Huw A. Ogilvie

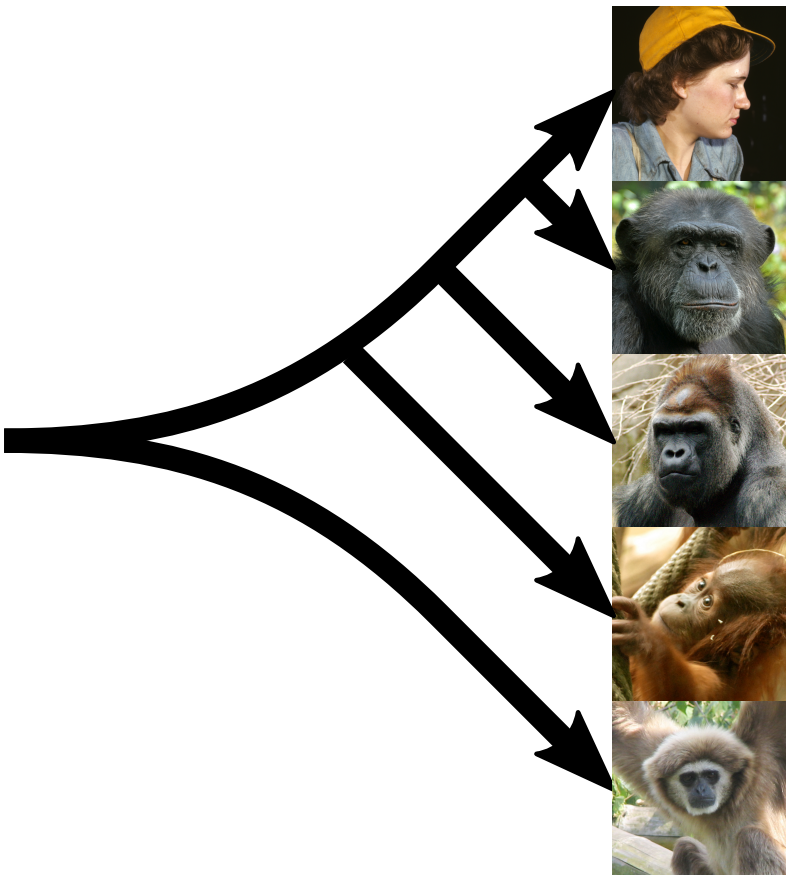
Research School of Biology, Australian National University

Centre for Computational Evolution, University of Auckland

February 2017

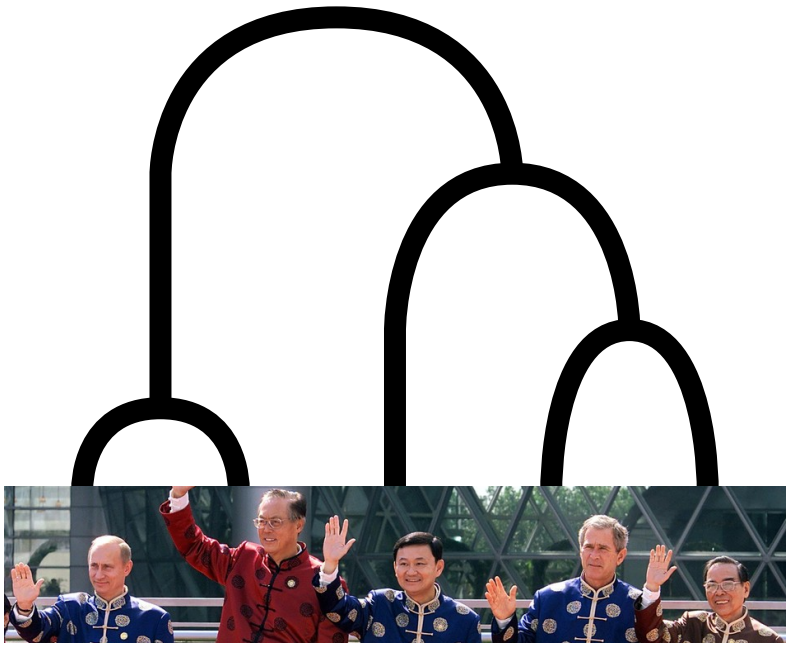
(Re-)introducing the multispecies coalescent

Speciation processes



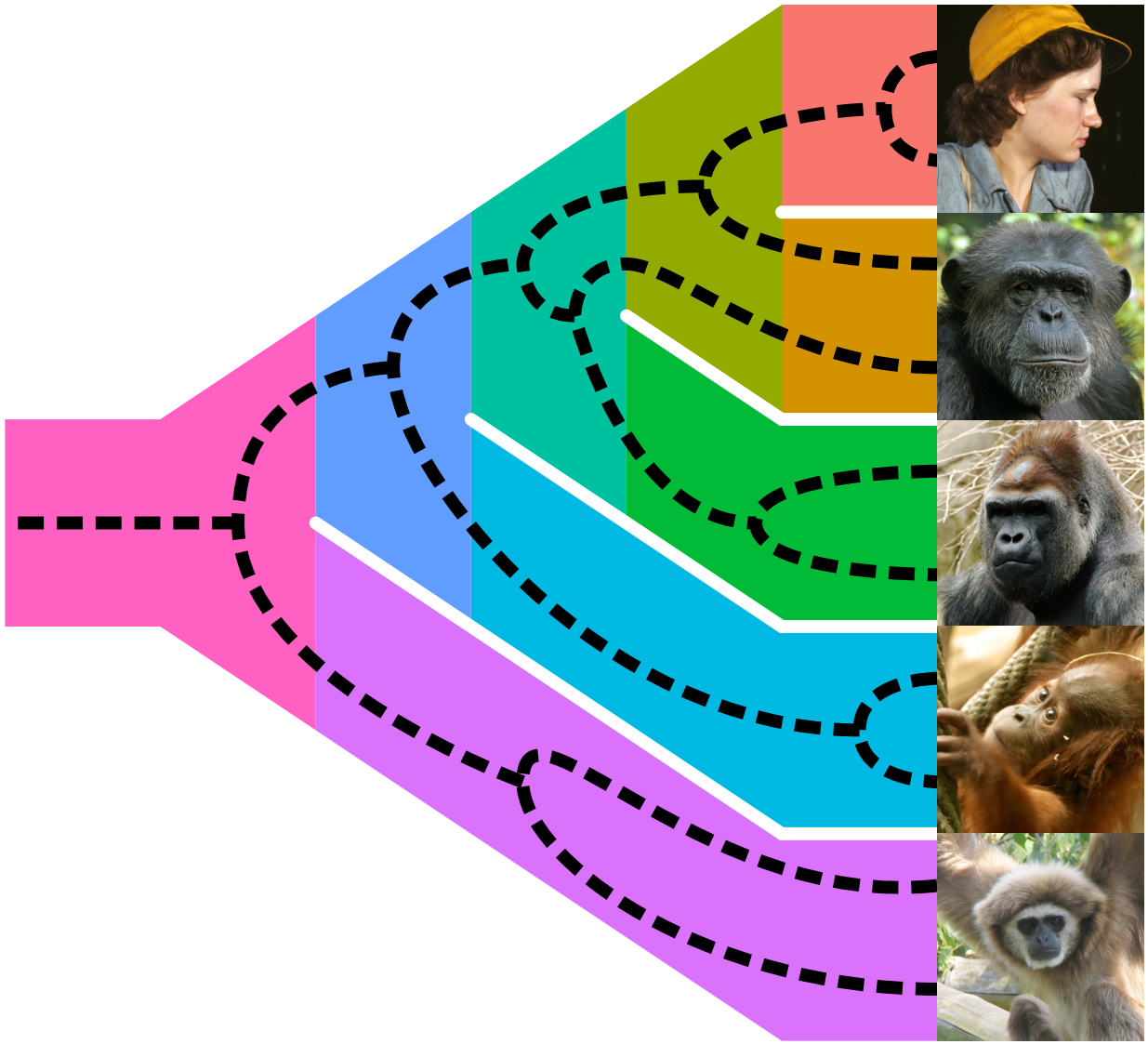
- Evolution of species
- Birth-death process (speciation & extinction)
- Diversification rate $\text{Birth} - \text{Death} = \lambda$
- Extinction ratio $\text{Death} \div \text{Birth} = \nu$
- Forward in time

Coalescent processes



- Evolution of genes
- Dependent on effective population size N_e
- Constant, linear, exponential or stepwise
- Backwards in time

Putting them together



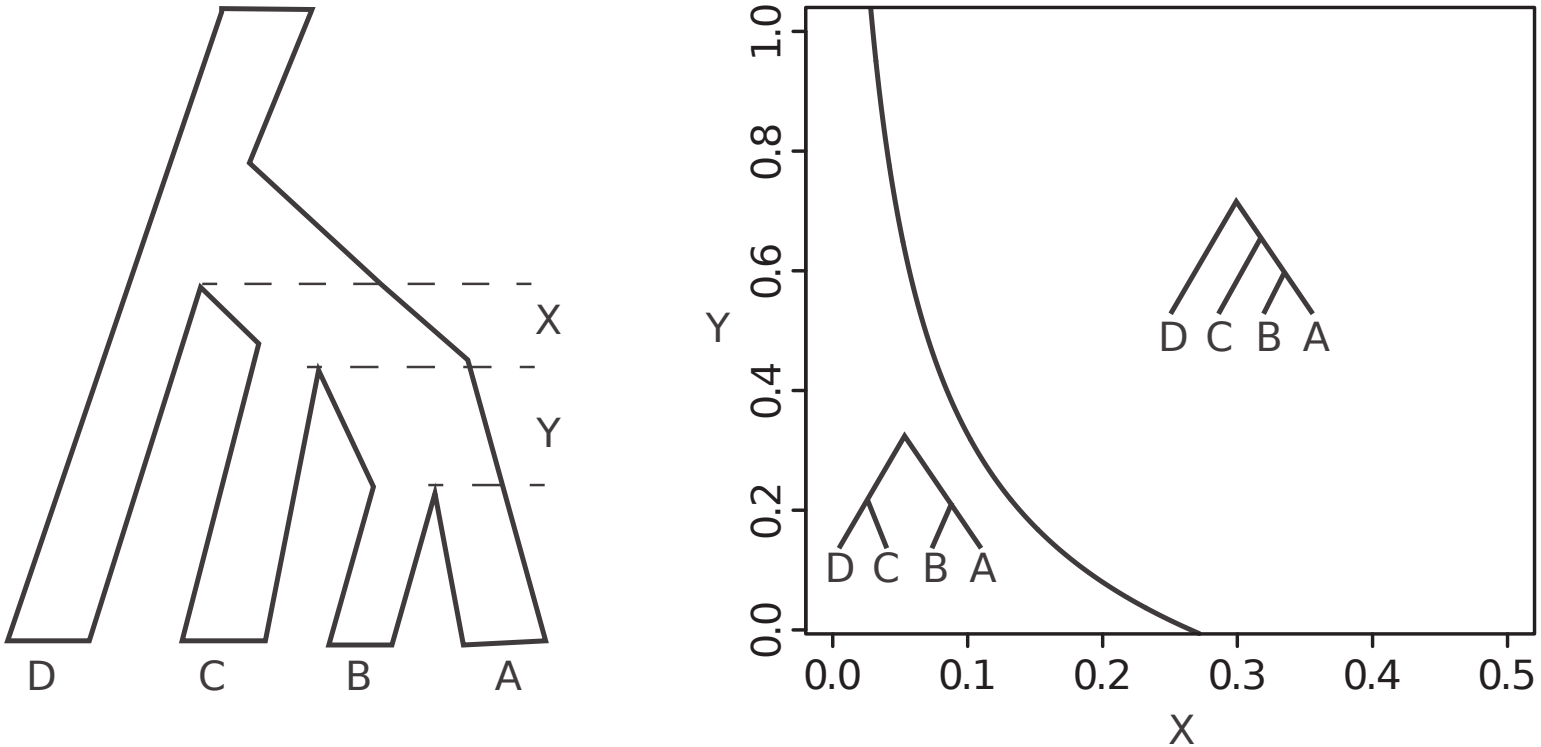
Why use the MSC

Why the MSC is important

or

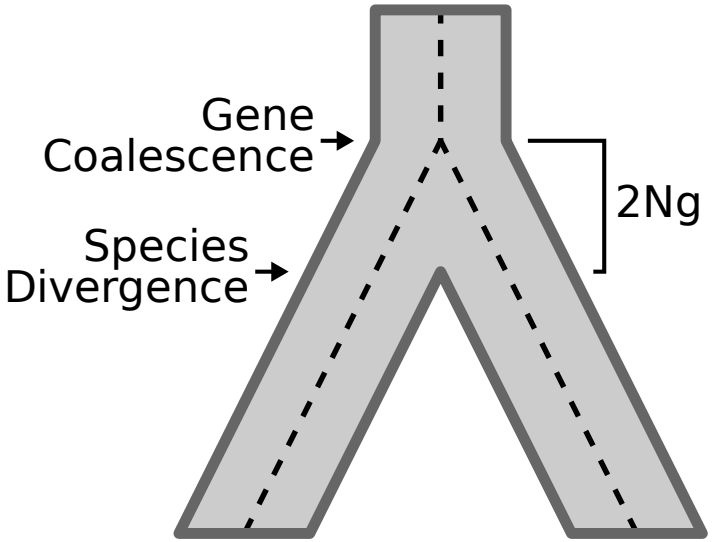
reasons to use MSC methods

Welcome to the anomaly zone



Linkem, Minin and Leaché (2016) Systematic Biology

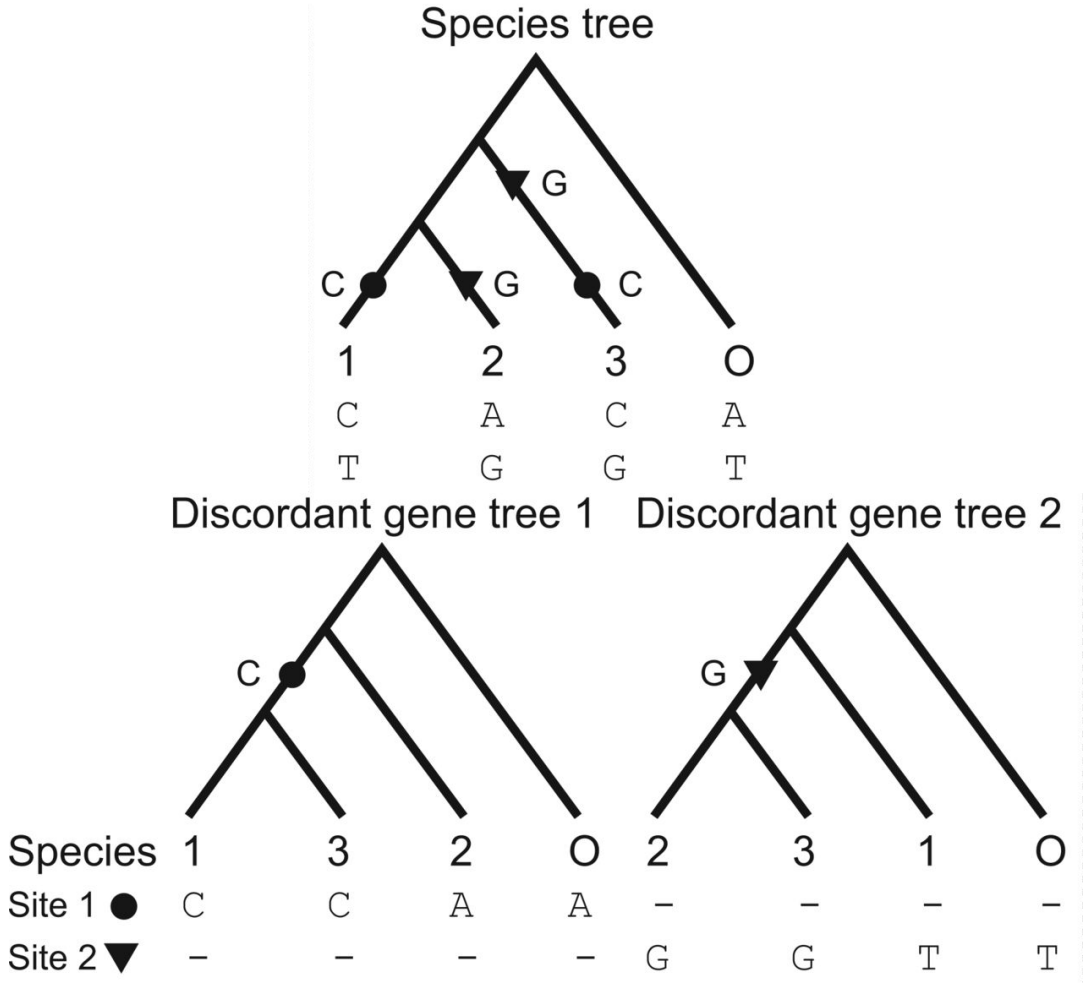
Impact on estimated divergence times



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

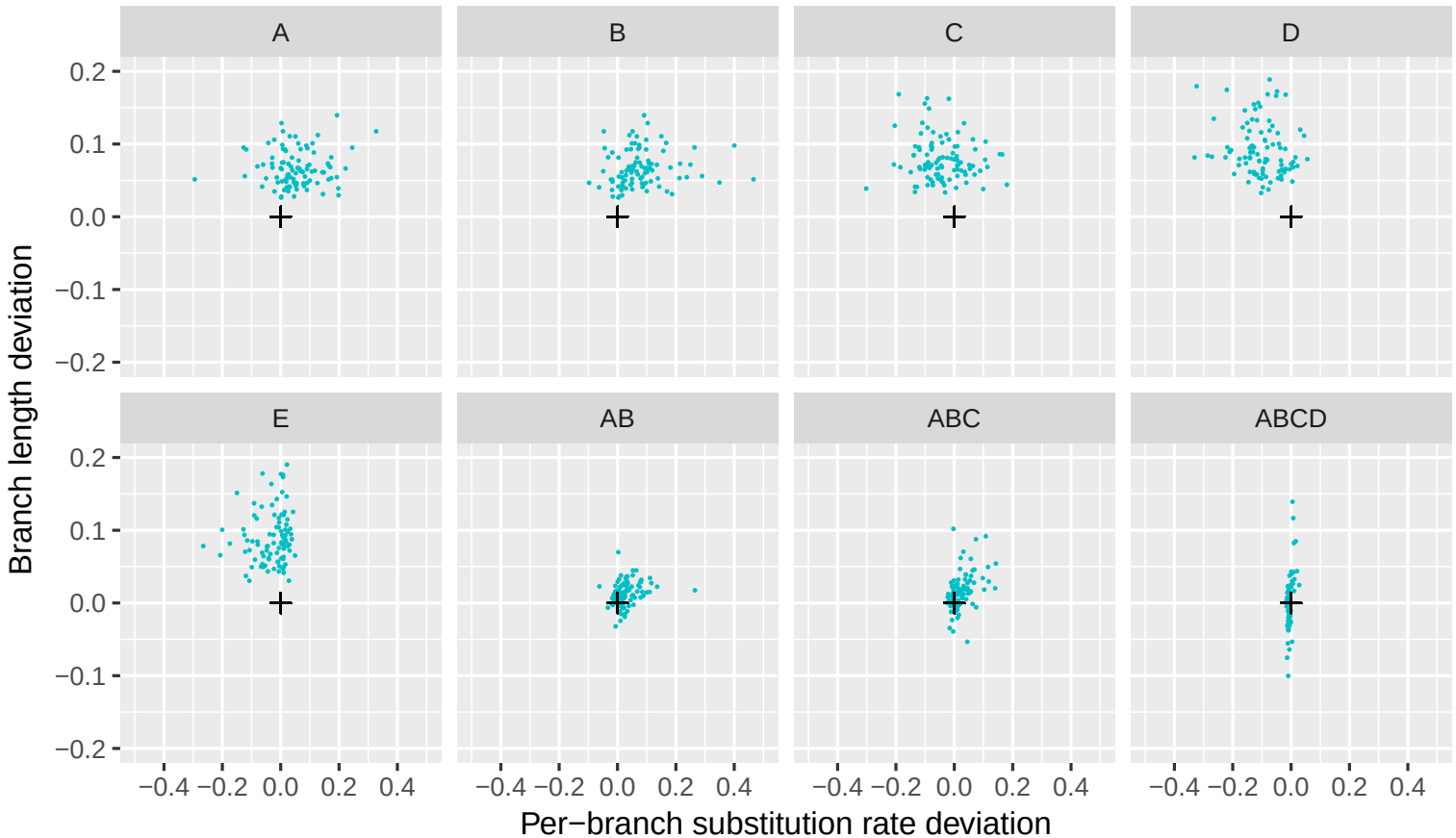
Hobolth *et al.* (2011) Genome Research

It gets worse



Mendes, Hahn & Hahn (2016) Molecular Biology and Evolution

The impact of SPILS

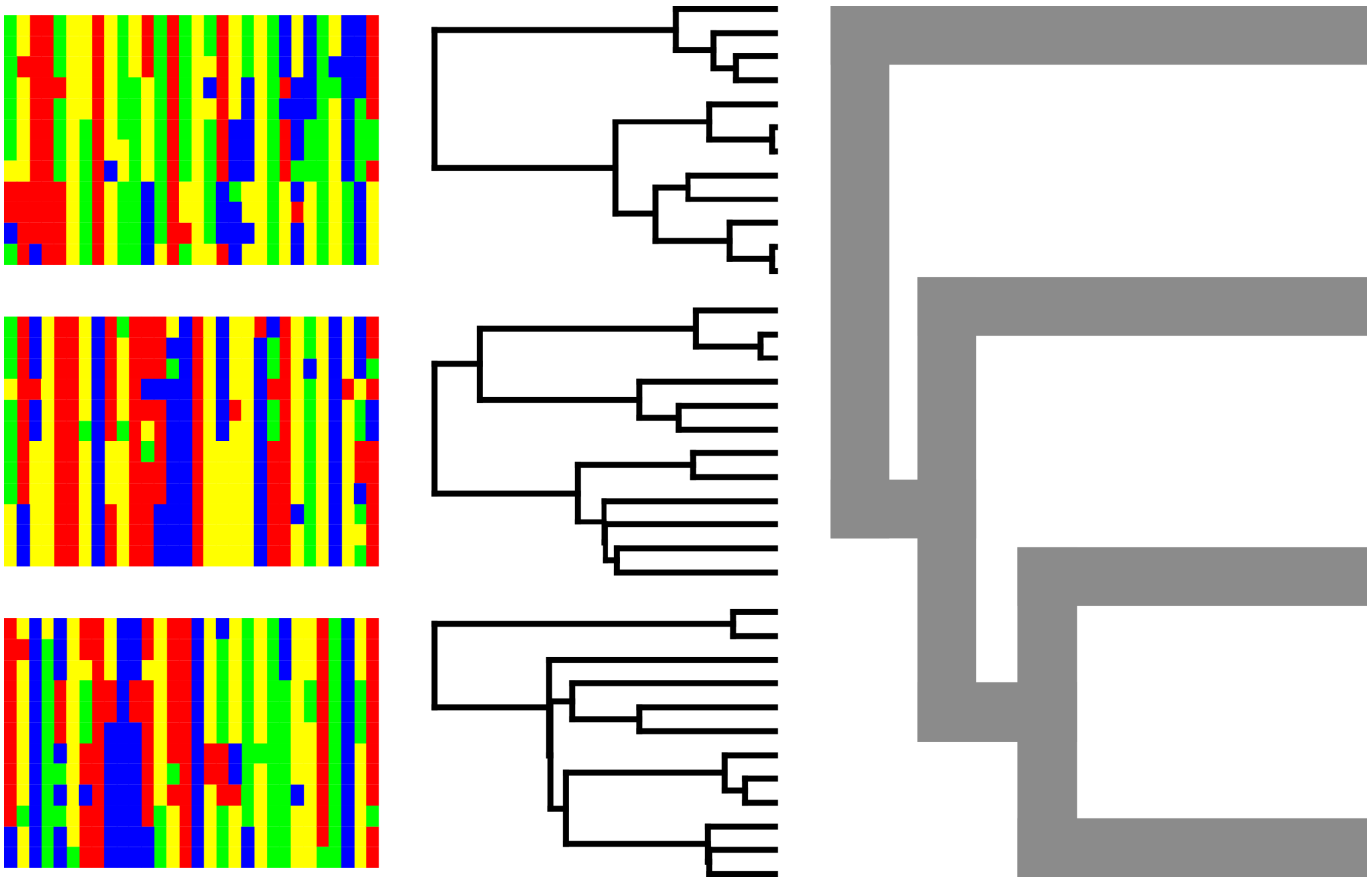


Ogilvie, Bouckaert & Drummond (2017, if the journal gods are kind)

Multi-locus MSC

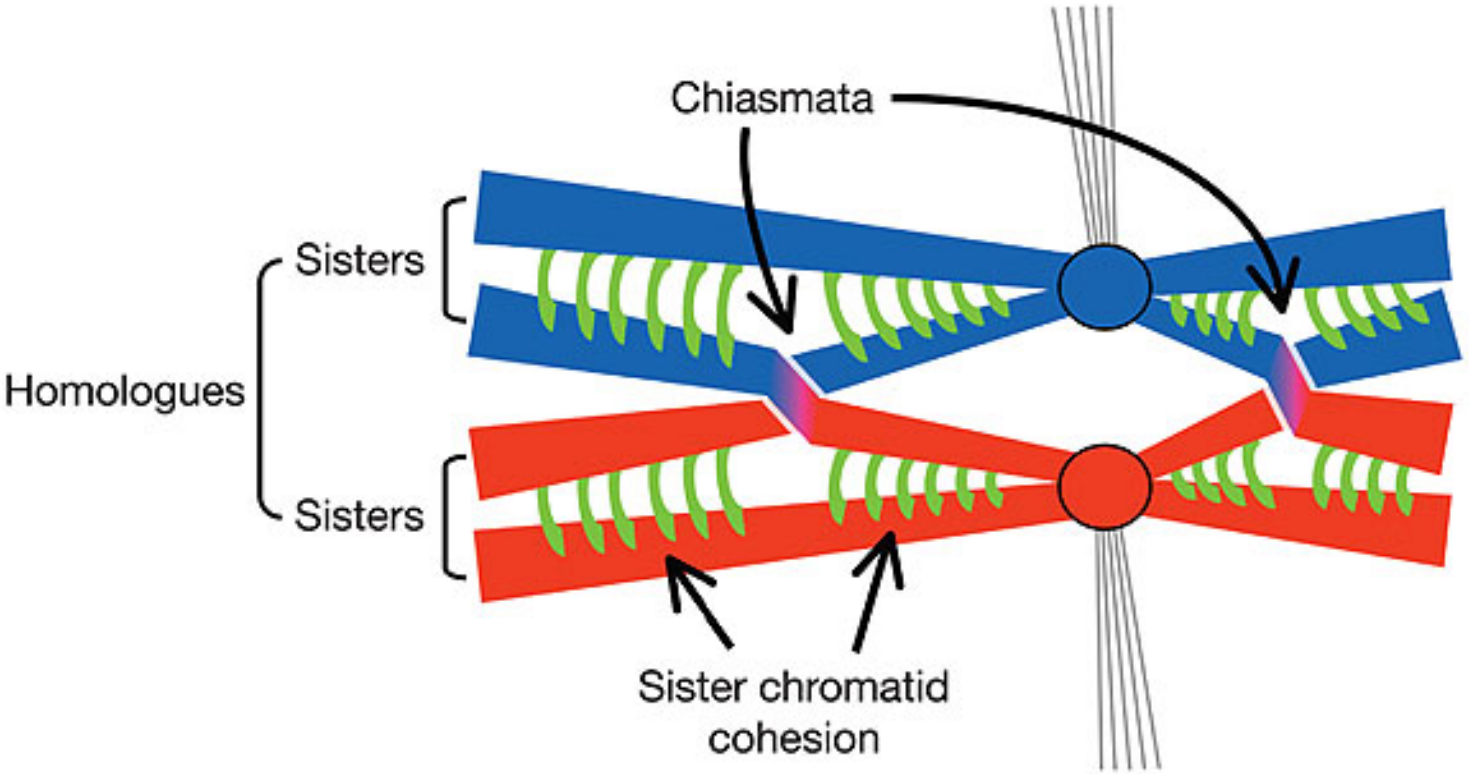
Estimating genes trees
and
species trees

Sequence alignments to species trees



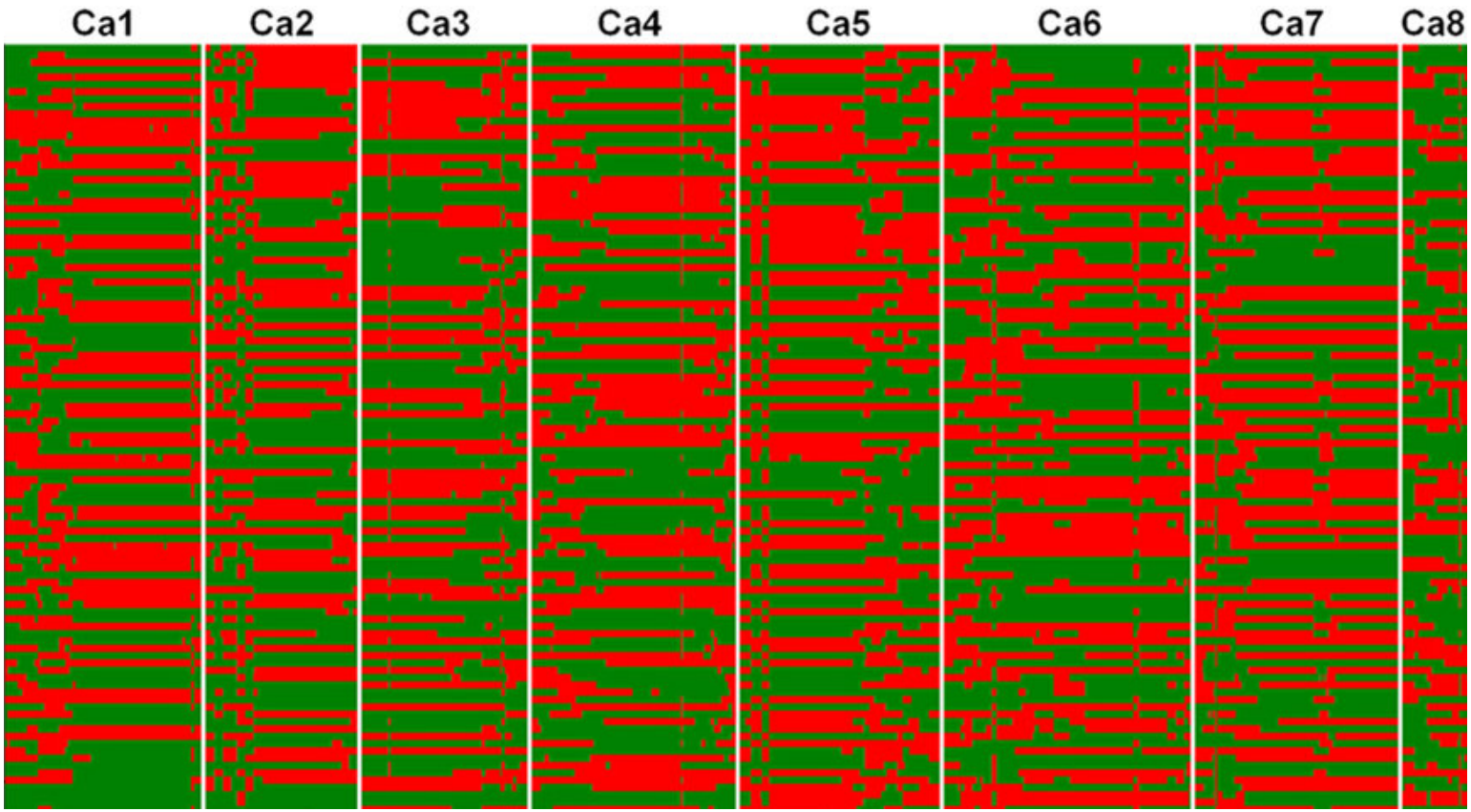
$$\Pr(S|D) = \frac{\Pr(D|G) \cdot \Pr(G|S) \cdot \Pr(S)}{\Pr(D)} \quad (1)$$

Recombination I



Neale & Keeney (2006) Nature

Recombination II

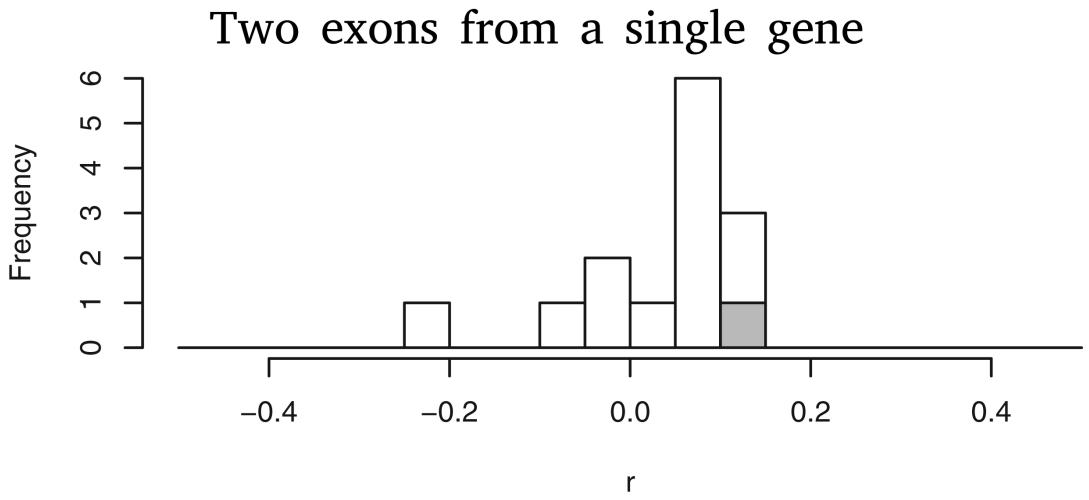


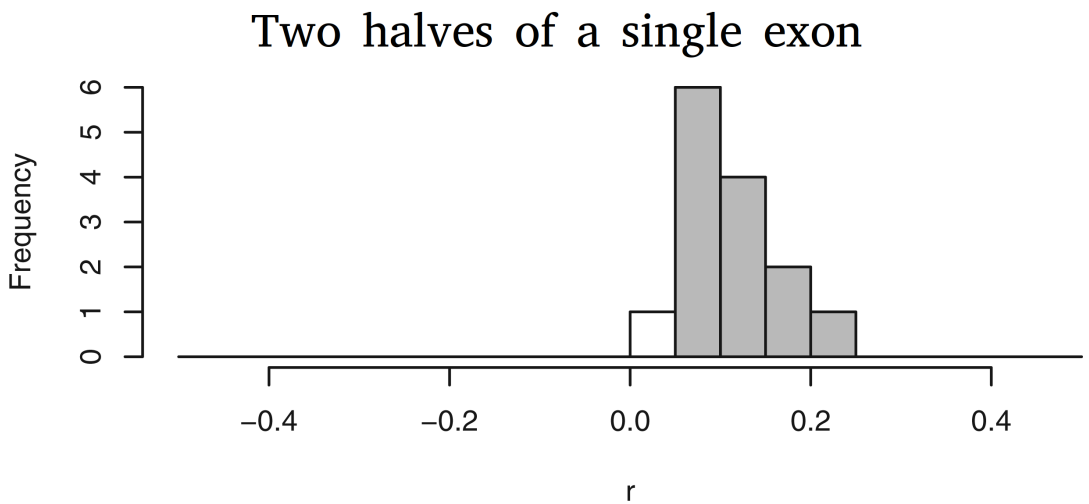
Kale *et al.* (2015) Scientific Reports

Multi-locus MSC assumptions

- Sites within a locus are linked
  - Sequences should be short enough that recombination is unlikely within a locus
  - Deeper trees $\rightarrow$ more recombination $\rightarrow$ shorter sequences
- No linkage between loci
  - Distances between loci should be large enough that recombination is common
  - Deeper trees $\rightarrow$ more recombination $\rightarrow$ shorter distances

Correlation of phylogenetic signal





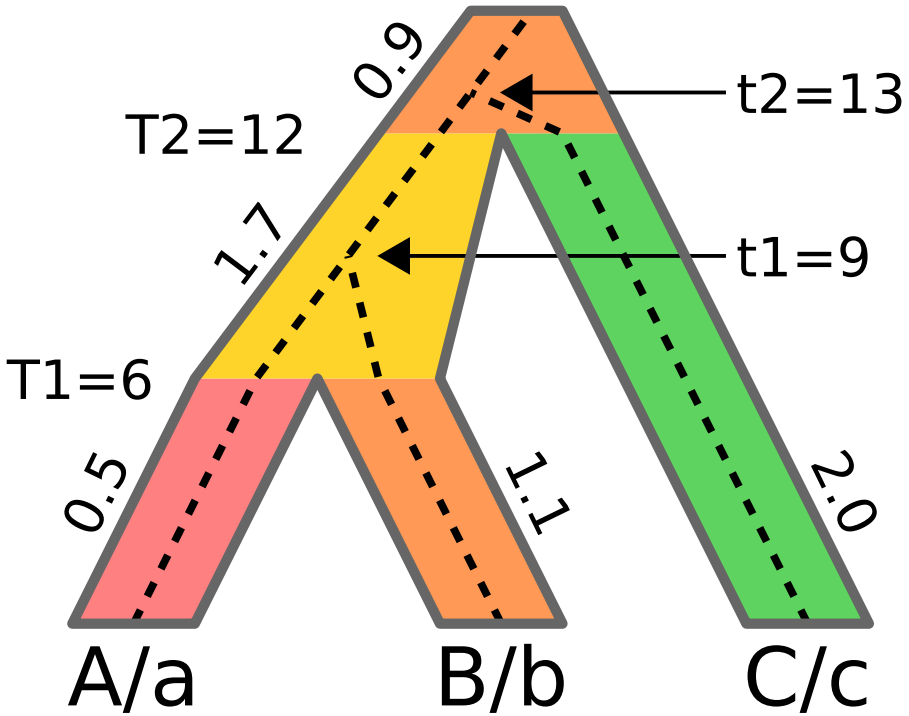
Scornavacca & Galtier (2017) Systematic Biology

Species tree clocks in StarBEAST2

Per-species substitution rates

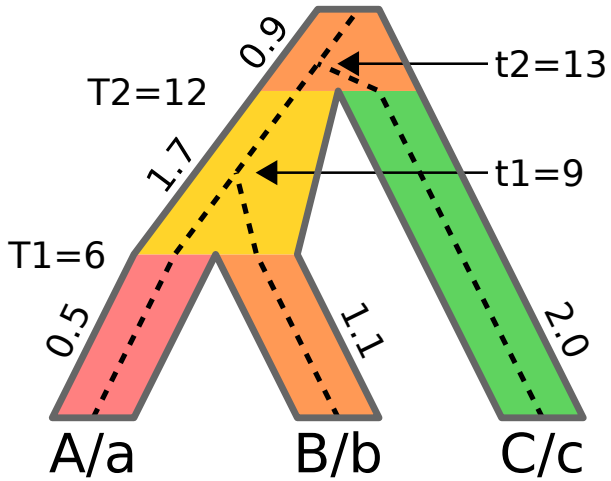
- Differences exist in substitution (clock) rates between loci
- Differences also exist in substitution rates between **species**
- But phylogenetic likelihood requires clock rates be applied to **gene** branches
- Species tree relaxed clocks extend the MSC to derive gene tree branch rates from species tree branch rates

Computing clock rates I



$$\text{rate}_{\text{branch}} = \text{rate}_{\text{locus}} \sum^{\text{species}} \left(\text{rate}_{\text{species}} \times \frac{\text{overlap}_{\text{species,branch}}}{\text{length}_{\text{branch}}} \right) \quad (2)$$

Computing clock rates II



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

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

$$\mathrm{rate_{ab}} = 10^{-9}\left(1.7 \times \frac{3.0}{4.0} + 0.9 \times \frac{1.0}{4.0}\right) = 1.5 \times 10^{-9} \qquad (5)$$

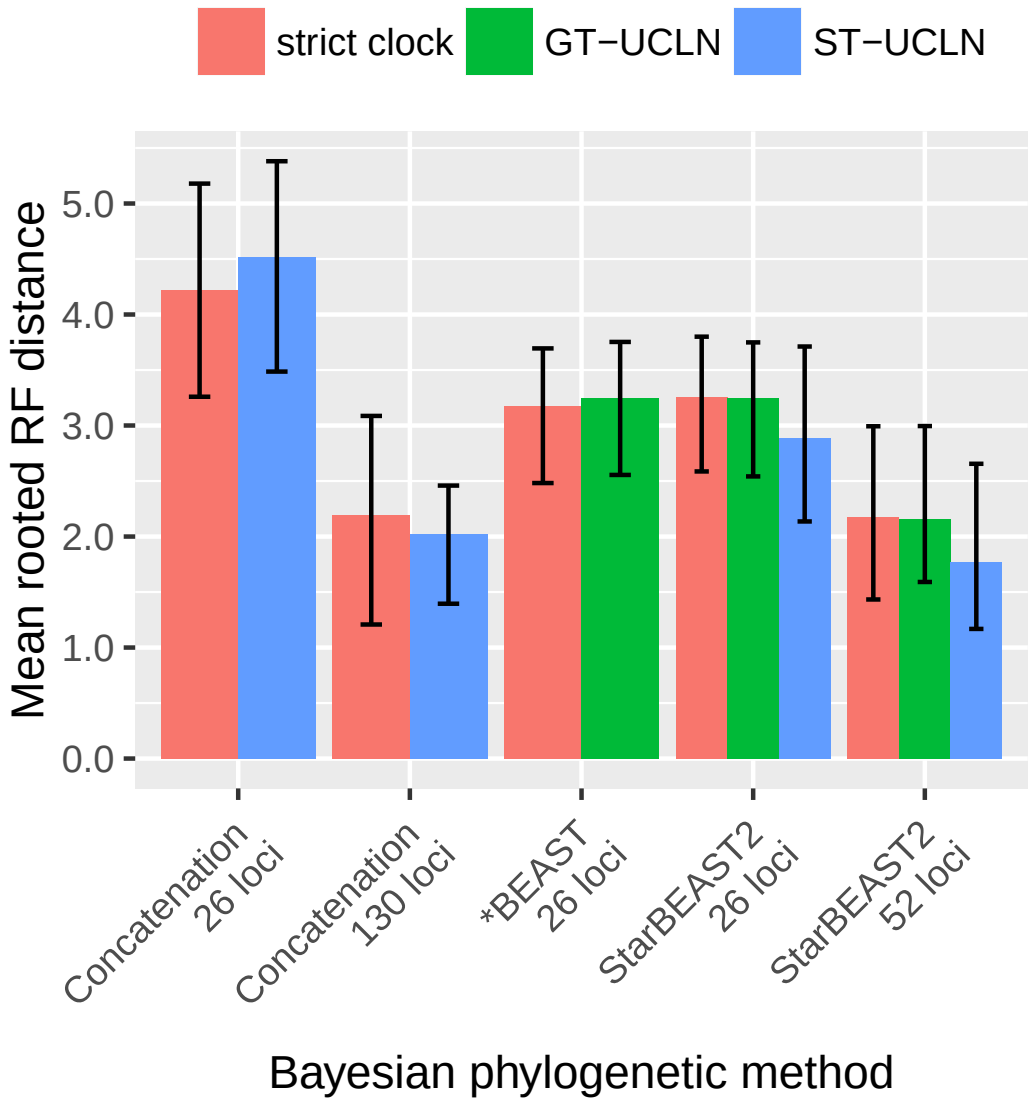
MSC accuracy

The power
of the
multi-locus MSC

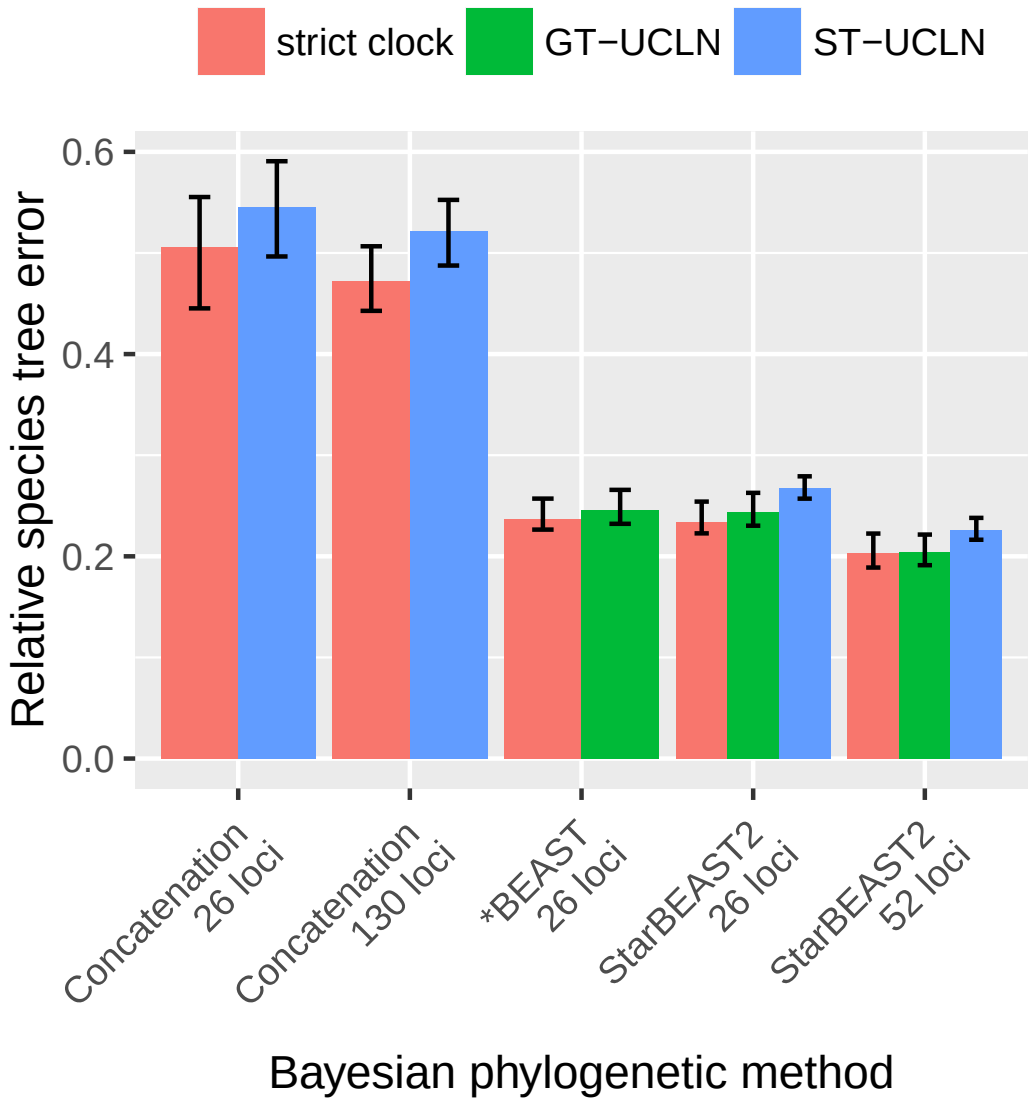
Simulation details

- 21 species, two phased sequences per species
- 600nt long sequence alignments
- Substitution rate $\approx 10^{-3}$ per million years
- $N_e \approx 2$ million assuming annual generation time
- Species tree height ≈ 34 million years
- "StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates" on bioRχiv

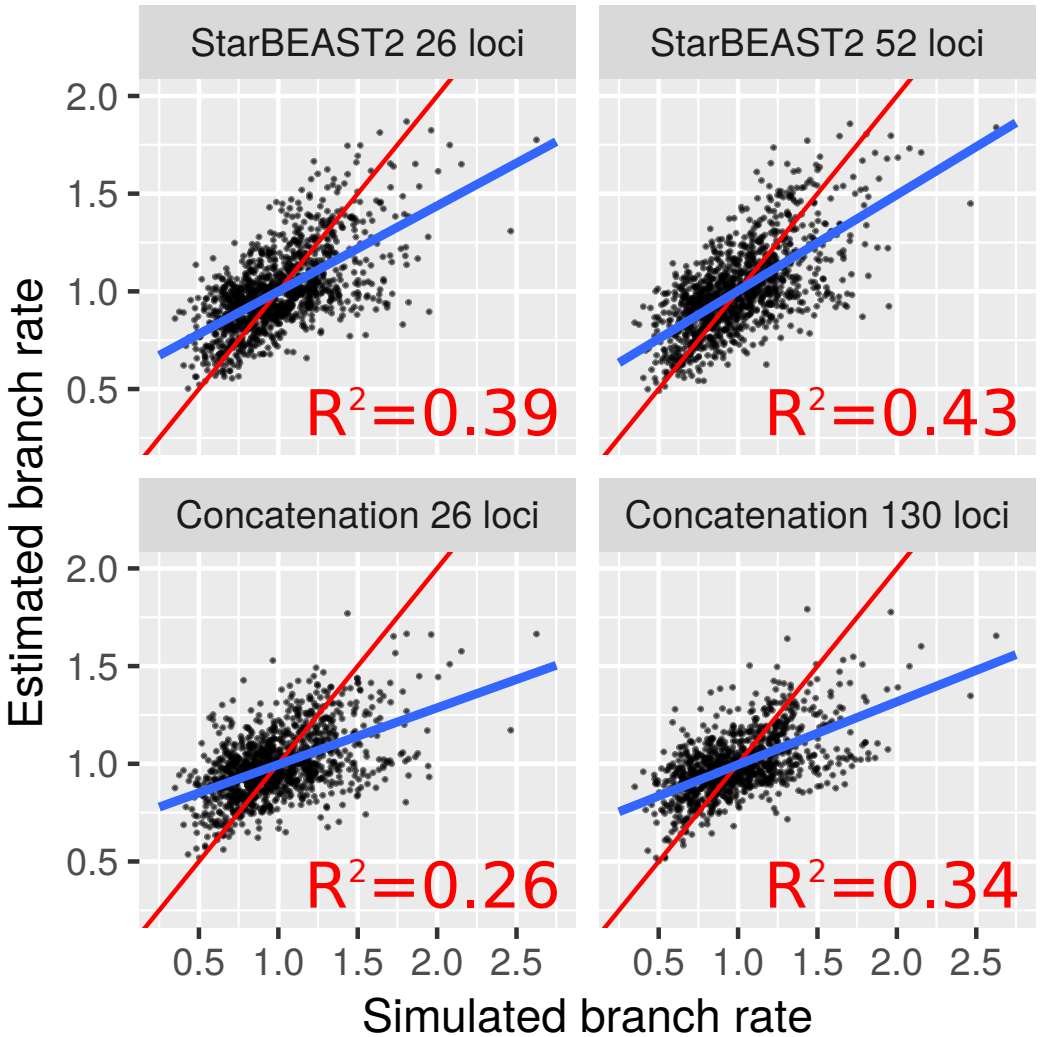
Topological accuracy



Branch length accuracy



Substitution rate accuracy

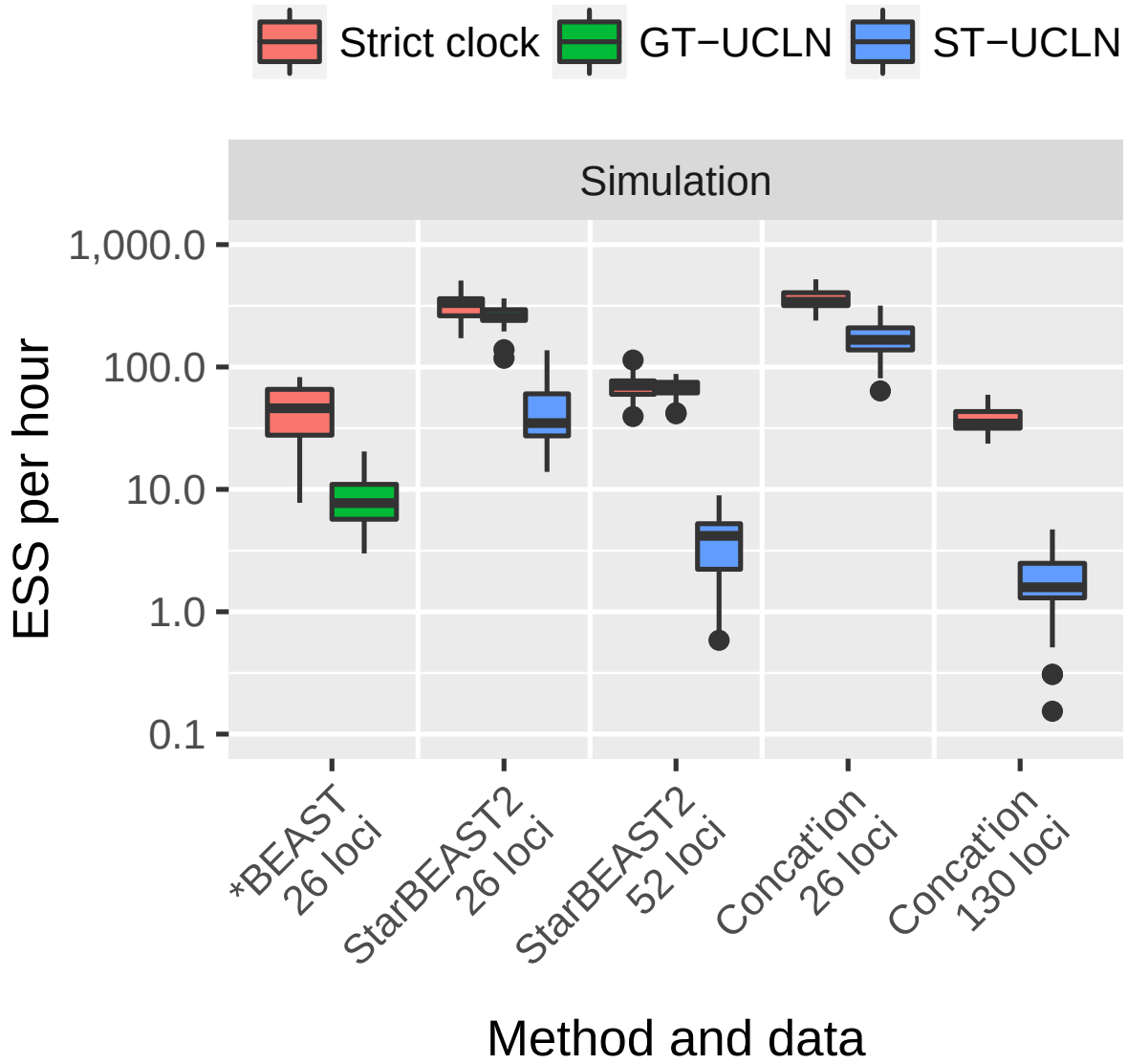


Implementations of the multi-locus MSC

StarBEAST2

- Successor to *BEAST, a method originally developed by Joseph Heled and Alexei Drummond
- 10 to 30 $\times$ faster than *BEAST — depending on the data and model
- Can integrate out population sizes analytically — slightly faster if population sizes are nuisance parameters
- The first method to implement species tree relaxed clocks

StarBEAST2 performance



Other methods

BPP Simultaneous species delimitation and species tree inference. Limited to Jukes-Cantor, strict clocks and gamma priors.

RevBayes A challenger approaches!

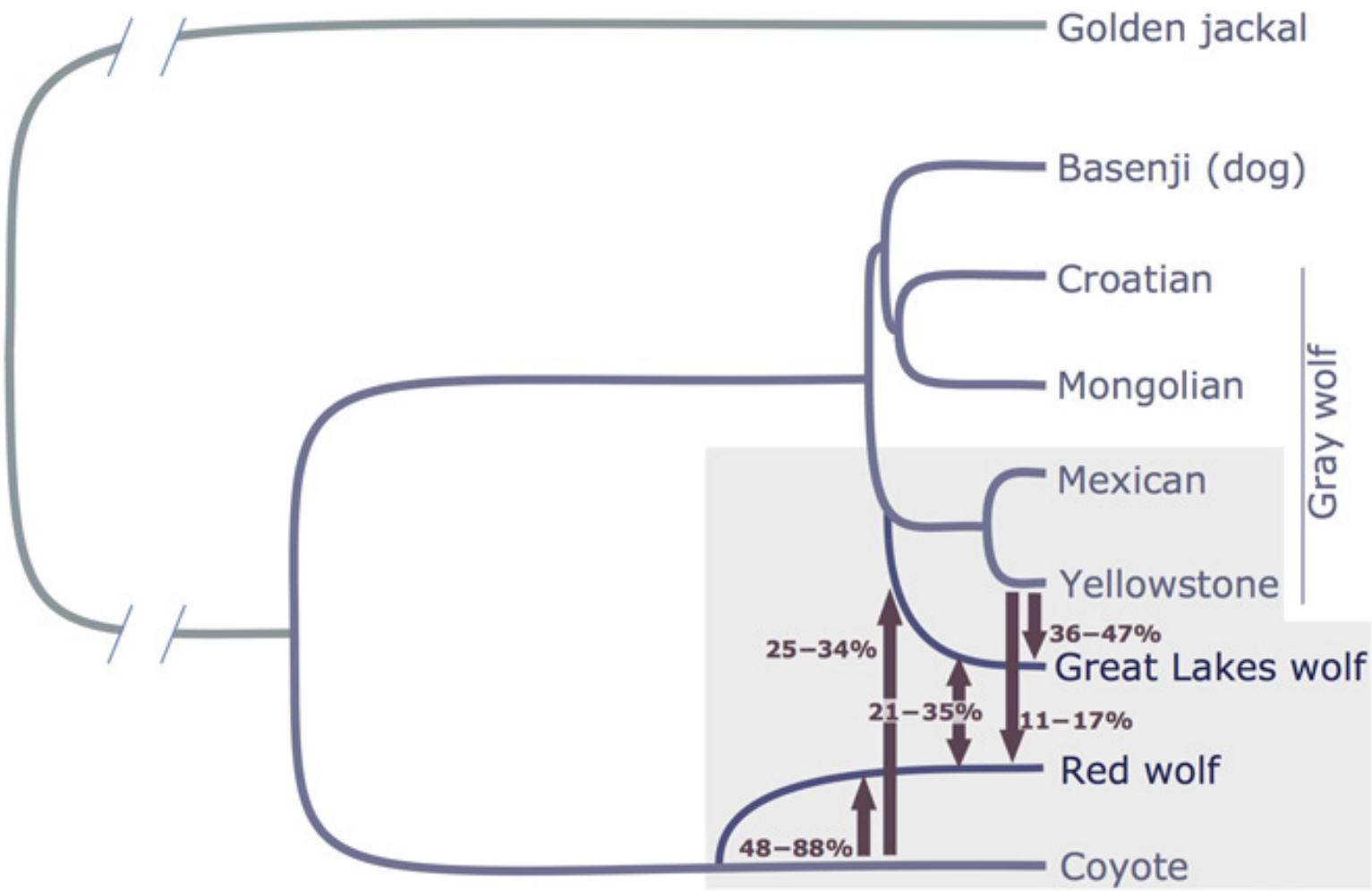
ASTRAL Can estimate topology but not branch lengths (except in coalescent units which are useless)

MP-EST Like ASTRAL but does a terrible job at estimating topology (what's the point???)

Multi-species-network coalescent

Because biology likes to play tricks on biologists

North American canids



vonHoldt *et al.* (2016) Science Advances

MSNC implementations

Multi-locus MSNC: joint estimation of gene trees, species networks and inheritance probabilities

PhyloNet	Limited range of substitution models and prior distributions implemented
BEAST2 Species Networks	Chi Zhang, Alexei Drummond, Tanja Stadler and me

Acknowledgements

Auckland
Alexei Drummond
Tim Vaughan
Remco Bouckaert
Joseph Heled

Canberra
Craig Moritz
Jason Bragg

Basel
Tanja Stadler
Chi Zhang

Photo credits:

Human, Howard R. Hollem; public domain U.S. Government work
Chimpanzee, Clément Bardot; CC-BY-SA
Gorilla, Kabir Bakie; CC-BY-SA
Orangutan, "Winkelbohrer"; CC-BY-SA
Gibbon, Kabir Bakie; CC-BY-SA
APEC group photo, Stephen Jaffe/Agence France-Presse/Getty Images