Bayesian time-tree priors

Alexei Drummond, a.drummond@auckland.ac.nz
Dr Joseph Heled, University of Auckland
Denise Kuehnert, University of Auckland
Dr Walter Xie, University of Auckland
Dr Tanja Stadler, ETH Zurich

Isaac Newton Institute, 22 June 2011

BEAST - MCMC software for time-trees

Drummond and Rambaut, 2007

BEAST focuses on **time-trees** (phylochronologies); both species trees and gene trees

Currently useful for

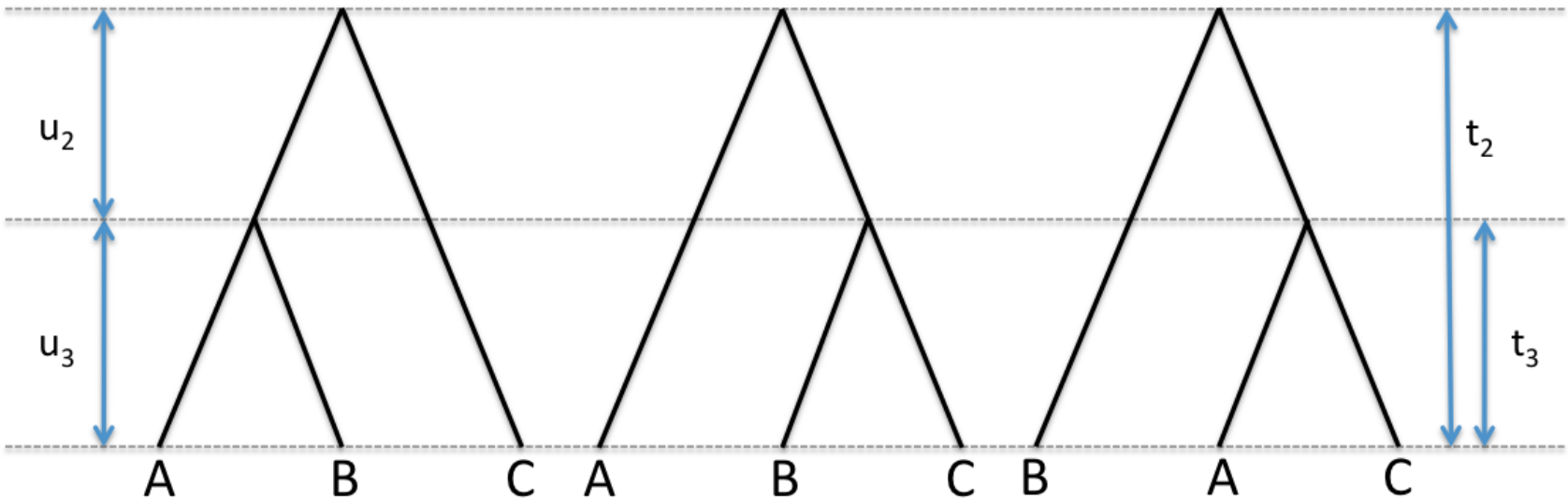
- Divergence time dating
- Estimating phylogenies under relaxed clock models
- Single population coalescent reconstruction
- Estimation of rates from viruses or ancient DNA
- Co-estimation of species trees and gene trees
- Simple models for statistical phylogeography

Working on

- More tree priors, relaxed clock models, substitution models
- Transdimensional model averaging
- More efficient tree sampling techniques

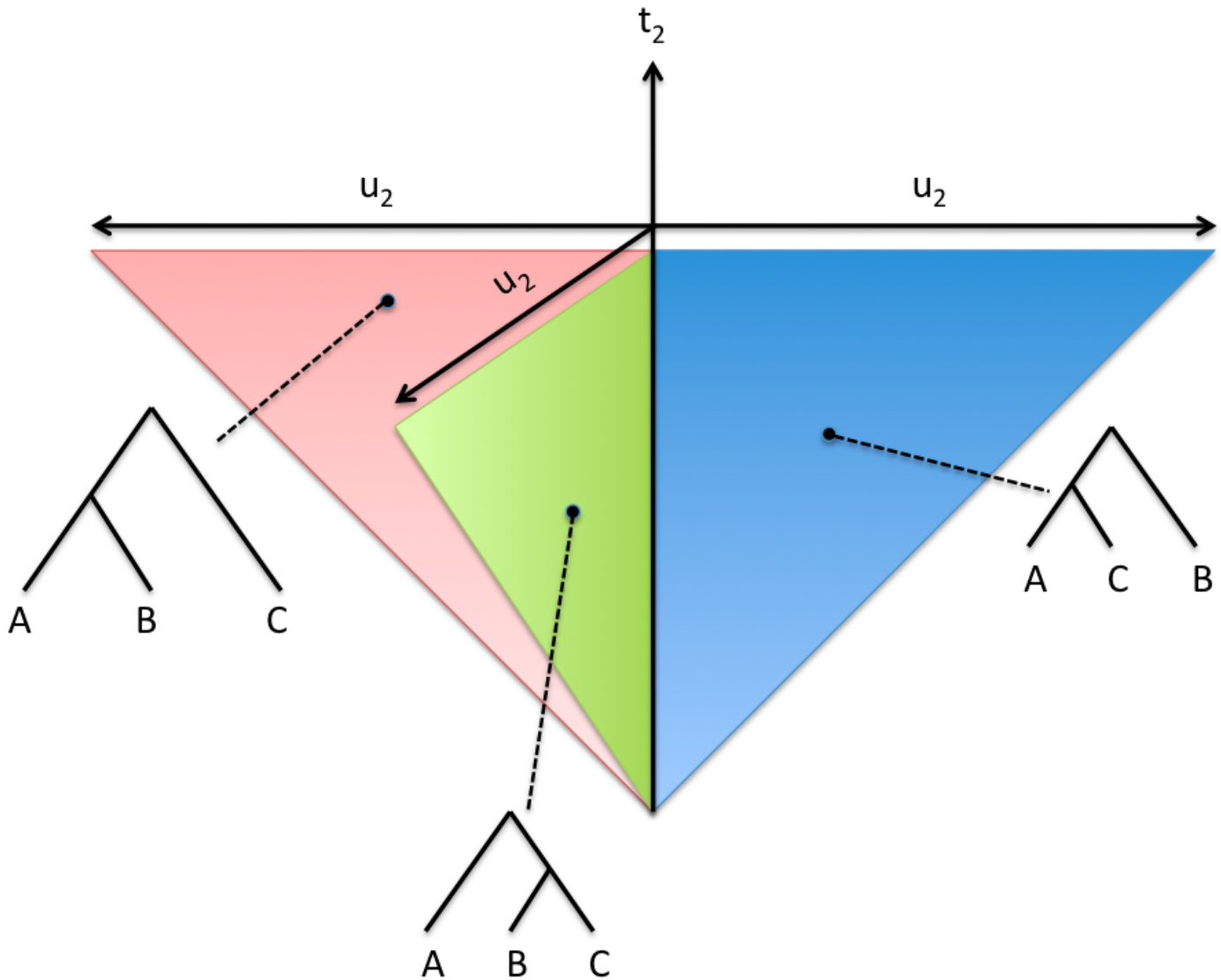
The tip-labeled time-tree

A tip-labeled time-tree is described by a *tip-labeled ranked topology* of size k and *coalescent times*, $\mathbf{u} = \{u_2, \ldots, u_k\}$.

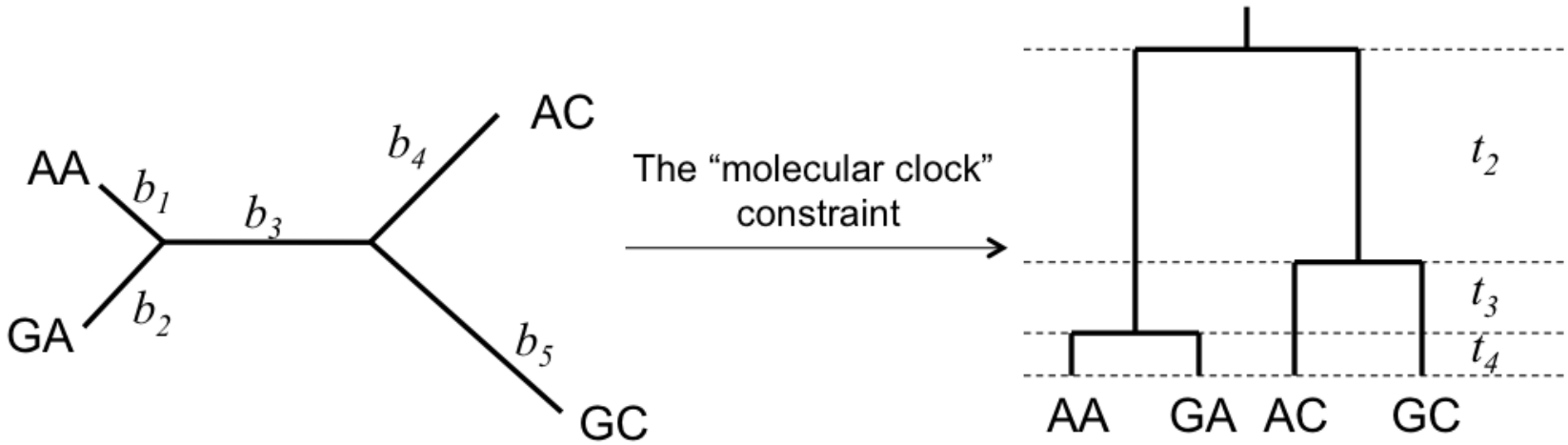


These time-trees of size 3 can be interpreted as describing the possible alternative evolutionary histories or (uniparental) ancestries of the three individuals represented by the labeled tips.

The space of tip-labeled time-trees of size 3



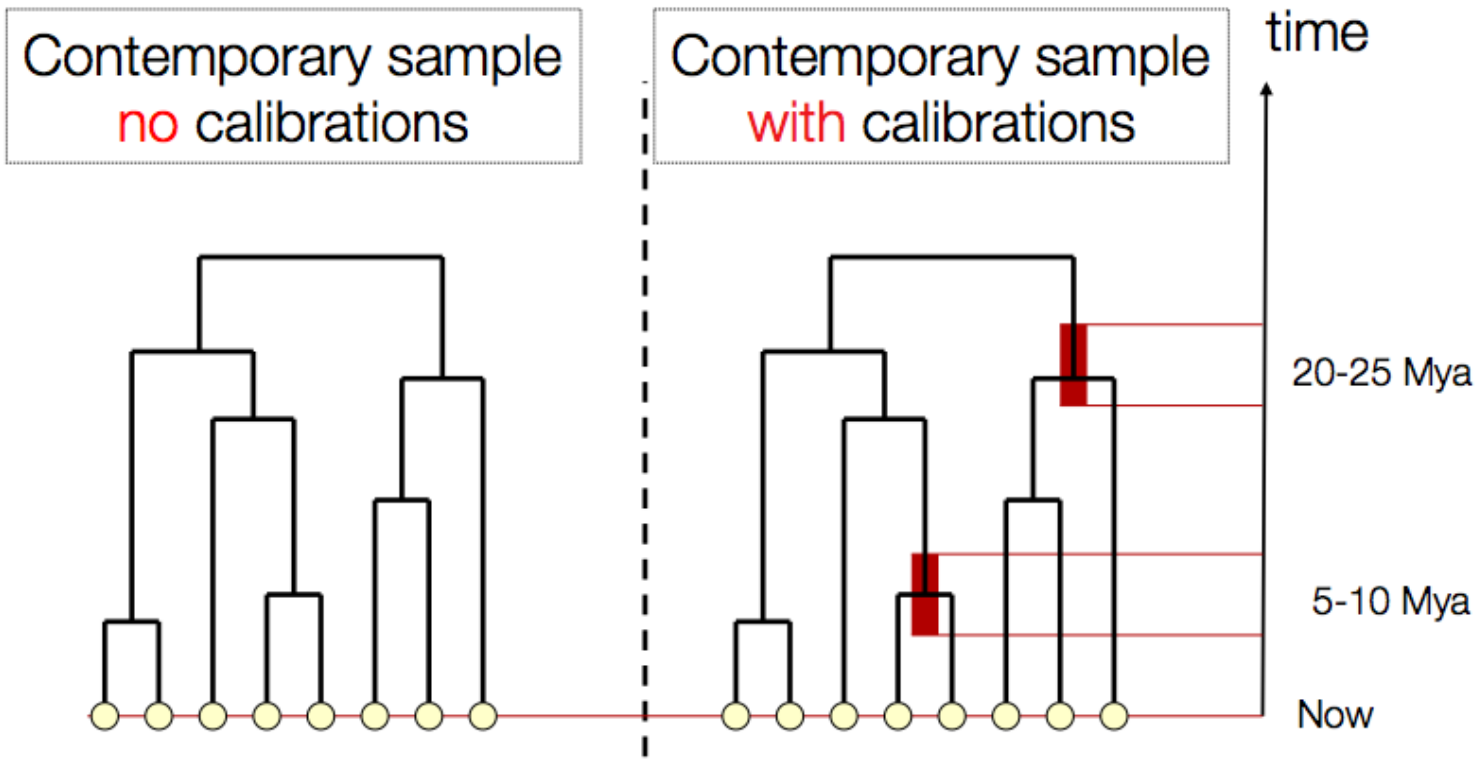
The molecular clock constraint



$$h(g, Q|D) \propto Pr\{D|\vec{\mu}g, Q\} f_G(g) f_Q(Q)$$

The joint posterior probability of the **rooted** time-tree (g) and the substitution matrix (Q) are estimated using Markov chain Monte Carlo (Drummond *et al*, 2002; 2006)

Absolute time via calibrations



Let $\rho_G(g)$ be "calibrated" $f_G(g)$ and allow the rate(s), $\vec{\mu}$, to be estimated:

$$p(\vec{\mu}, g, Q|D) \propto Pr\{D|\vec{\mu}g, Q\}\rho_G(g)f_Q(Q)f_M(\vec{\mu})$$

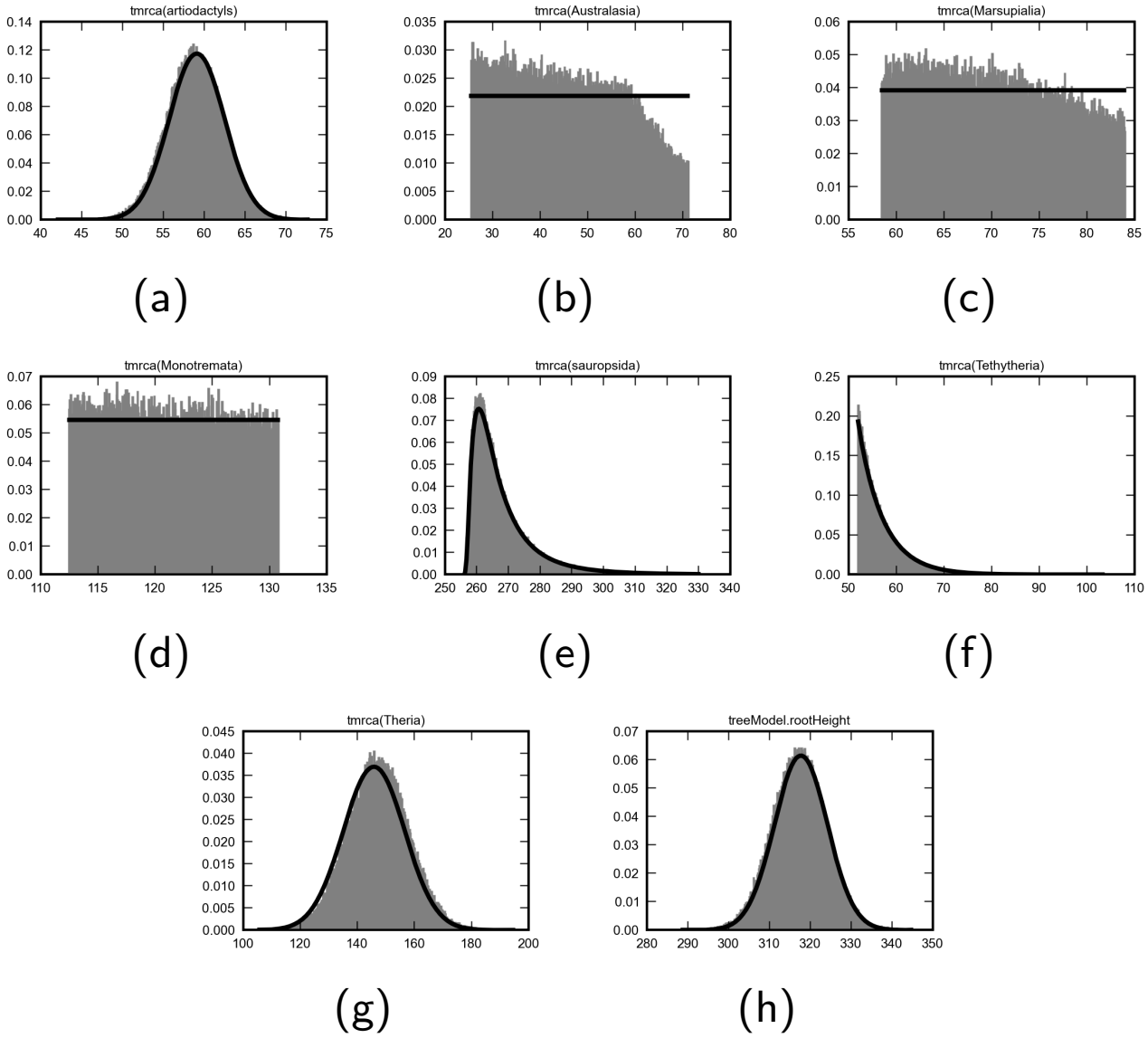


Figure: A simple construction of calibrated tree prior: $\rho_G(g) \propto f_G(g) \times \prod_{i=1}^{k} f_i(s_i)$. Where $f_i()$ is the univariate "calibration density" for the divergence time of the i'th calibrated node in the tree.

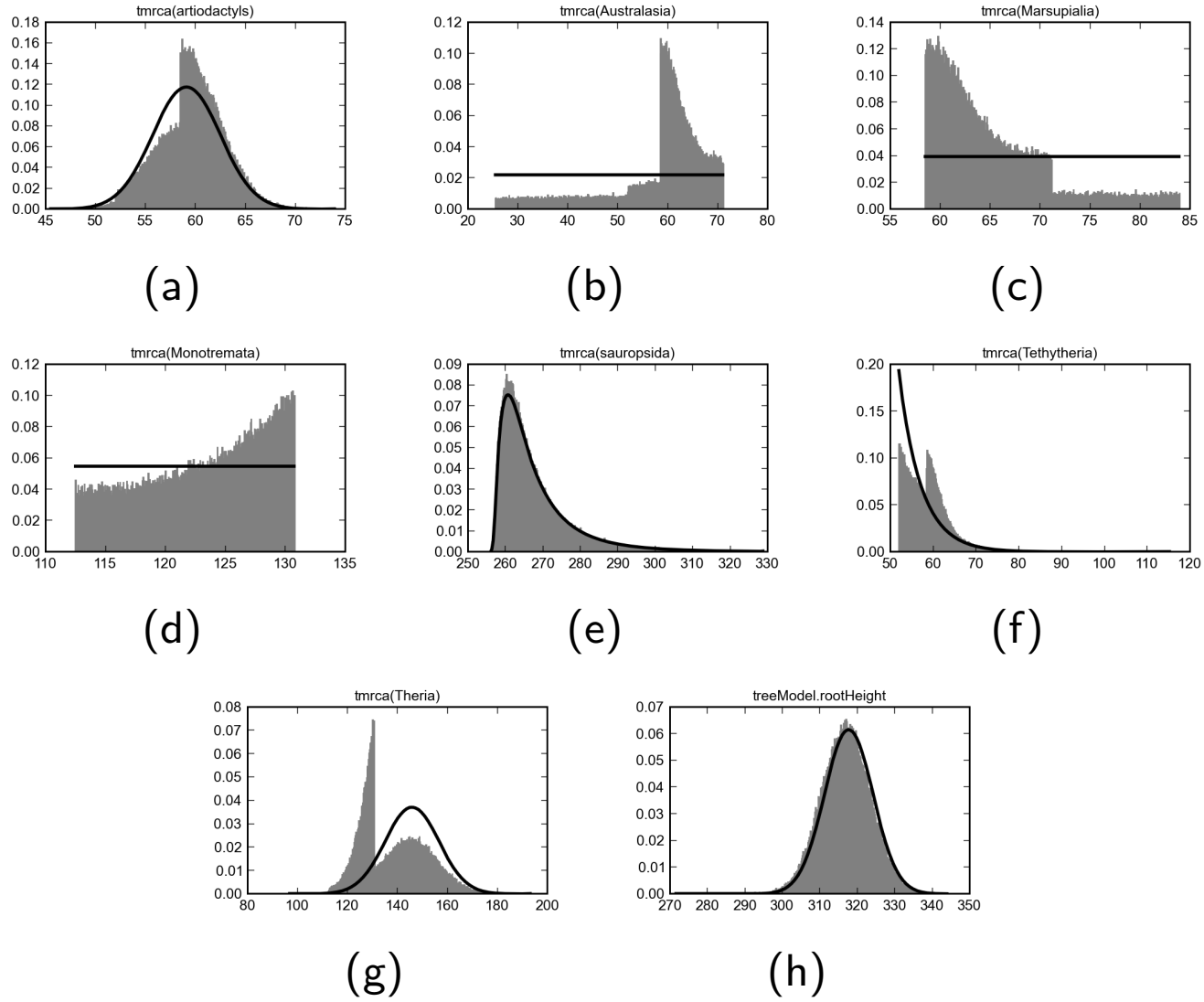


Figure: The marginal prior distributions that result from BEAST (gray) versus calibration densities (black) specified for the calibrated nodes from [**?**]. The marginal prior distributions were obtained from a MCMC run using the prior only. The calibration densities are as defined by the authors.

Calibration prior, main idea

Let $\tau(g)$ be TMRCA for calibrated taxa on tree $g \in G$.
Consider $\rho_{\mathrm{G}}(g)$, a candidate for a calibrated tree prior on the space of trees and $\rho_{\mathrm{T}}(\cdot)$, the desired marginal prior on $\tau(g)$. The following properties are desired:

(I) The marginal density on the calibrated node is equal to the calibration density:

$$\rho_{\mathrm{T}}(x) = \int_{\substack{g \in G \\ \tau(g)=x}} \rho_{\mathrm{G}}(g) = \int_{g \in G} 1(\tau(g) = x)\rho_{\mathrm{G}}(g)\mathrm{d}g \quad (1)$$

(II) When restricted to a subset of trees with equal calibrated node height, the density is proportional to f_G density:

$$\tau(g_1) = \tau(g_2) \quad \Longrightarrow \quad \frac{\rho_{\mathrm{G}}(g_1)}{\rho_{\mathrm{G}}(g_2)} = \frac{f_{\mathrm{G}}(g_1)}{f_{\mathrm{G}}(g_2)}. \quad (2)$$

Calibration prior, main idea

$$\rho_{\mathrm{G}}(g) = f_{\mathrm{G}}(g)\,\frac{\rho_{\mathrm{T}}(\tau(g))}{f_{\mathrm{T}}(\tau(g))}, \tag{3}$$

where $f_{\mathrm{T}}(\cdot)$ is the marginal distribution of τ under f_{G}. We call this the *conditional-construction*. Informally, equation 3 can be written as

$$\text{new-joint-prior} = \text{old-joint-prior} \times \frac{\text{new-marginal}}{\text{old-marginal}}.$$

This replaces the current calibrated tree prior in BEAST which was just proportional to $f_{\mathrm{G}}(g) \times \rho_{\mathrm{T}}(\tau(g))$.

Calibration prior, main idea

Heled and Drummond, 2011, in press

Our small contribution was computing $f_{\mathrm{T}}(\cdot)$ for a single monophyletic divergence time τ when f_{G} is the Yule prior. The marginal has a nice closed-form:

$$f_{\mathrm{T}}(x) = \begin{cases} \frac{1}{2}(n_c^3 - n_c)\lambda e^{-3\lambda x}(1 - e^{-\lambda x})^{n_c-2} & \text{if } n_c < n, \\ n_c(n_c - 1)\lambda e^{-2\lambda x}(1 - e^{-\lambda x})^{n_c-2} & \text{if } n_c = n. \end{cases} \quad (4)$$

where n_c is number of taxa under the clade and n is the size of the whole tree.

We also computed $f_{\mathrm{T}}(\cdot)$ when the monophyly restriction is relaxed and we are currently working on the case when more than one clade is calibrated (seems hard).

Multiple calibrations: root + nested clade

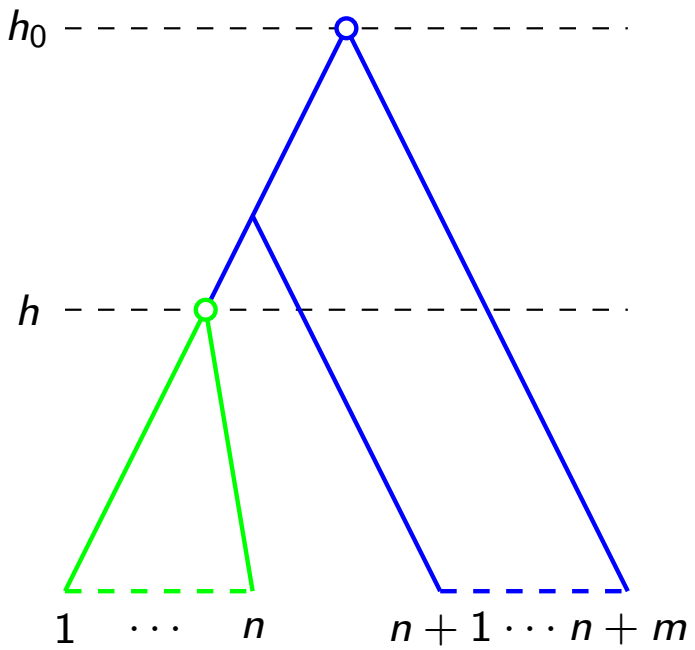


Figure: Monophyletic clade of size n and root with $n+m$ taxa.

$$
\begin{aligned}
f(h_0, h|n, m) = \quad & (n-1)n(n+1)\lambda^2 e^{-\lambda(3h+4h_0)} \\
& (1-e^{-\lambda h})^{n-2}(1-e^{-\lambda h_0})^{m-3} \\
& (e^{2\lambda(h_0+h)} + 2(m-1)e^{\lambda(2h_0+h)} - \\
& 2me^{\lambda(h_0+2h)} - m(m-1)e^{\lambda h_0+\lambda h} + \\
& \frac{(m-1)(m-2)}{2} e^{2\lambda h_0} + \frac{m(m+1)}{2} e^{2\lambda h})
\end{aligned}
\tag{5}
$$

Multiple calibrations: two nested clades

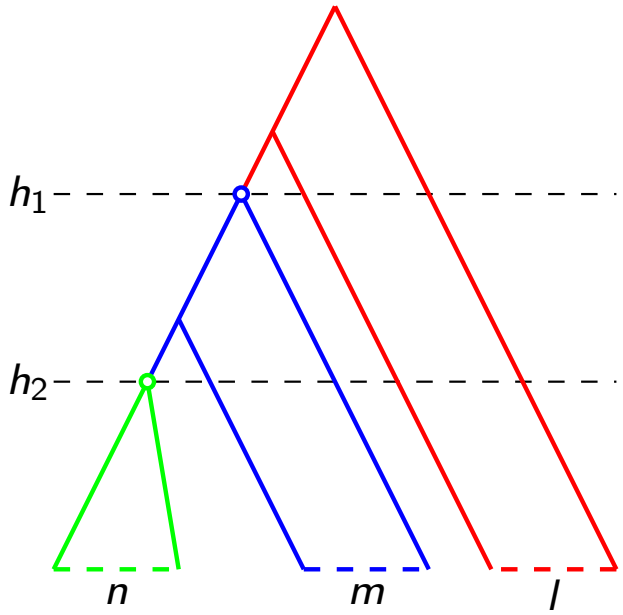


Figure: Two nested monophyletic clades of size n and $n+m$ taxa in a $n+m+l$ taxa tree ($l > 0$).

$$
\begin{aligned}
f(h_1, h_2 | n, m) = \quad & {}^1/_2 (n-1)n(n+1)(n+1+m)\lambda^2 e^{-\lambda(3h_2+5h_1)} \\
& (1-e^{-\lambda h_2})^{n-2}(1-e^{-\lambda h_1})^{m-3} \\
& (e^{2\lambda(h_2+h_1)} - 2me^{\lambda(2h_2+h_1)} + \\
& 2(m-1)e^{\lambda(h_2+2h_1)} - m(m-1)e^{\lambda(h_2+h_1)} + \\
& \frac{m(m+1)}{2} e^{2\lambda h_2} + \frac{(m-1)(m-2)}{2} e^{2\lambda h_1})
\end{aligned}
\tag{6}
$$

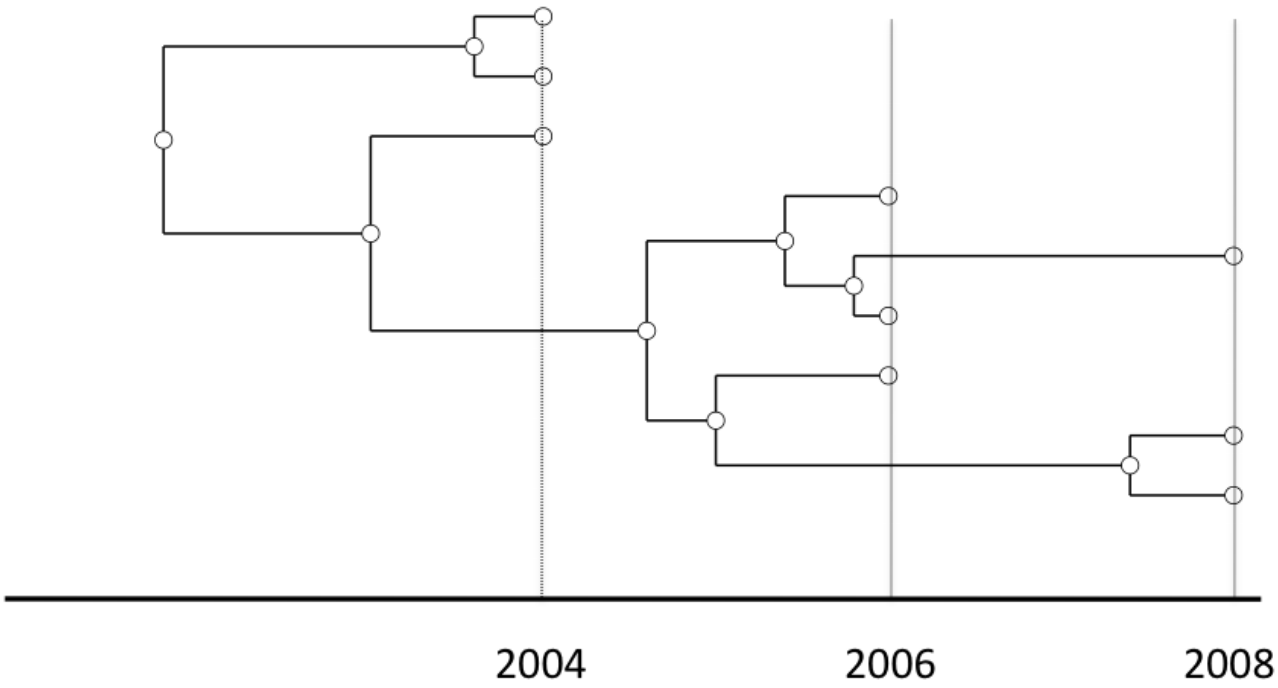
Questions

- What is the marginal distribution of the t_{MRCA} of set of taxa of size k in a random Yule tree of size $n \geq k$? With or without monophyly constraint. **[solved]**
- What is the marginal distribution of the t_{MRCA} of set of taxa of size k in a random Birth-death tree of size $n \geq k$? With or without monophyly constraint.
  - For Birth-death serial sampling model?
  - For coalescent models?
- What is the joint marginal distribution of the t_{MRCA}'s of two or more sets of (potentially nested) sets of taxa (e.g. $k_1[k_2, k_3[k_4]], k_5$), under Yule prior **[solved for a few special cases]**
  - For Birth-death model?
  - For Birth-death serial sampling?
  - For coalescent models?

Evolution is happening right now!

Rodrigo and Felsenstein, 1999; Drummond *et al*, 2002

Many pathogens, such as HIV, Hepatitis C and Influenza A, evolve very rapidly, so that samples of the virus population from different times directly reveal evolutionary change.



In fact it becomes possible to **calibrate** the tree and thus place the tree on a time scale - by constraining the tips to known sampling times

Phylodynamics

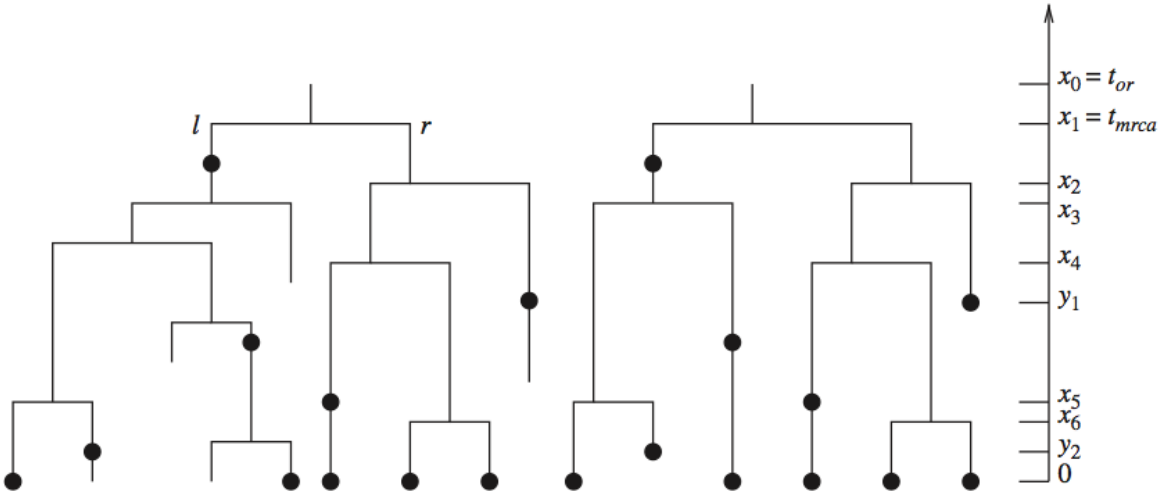
- The intersection of **phylogenetics** and **mathematical epidemiology**
- Includes estimation of epidemiological parameters from phylogenetic data
- In a Bayesian setting, this has the familiar flavor of a hierarchical tree prior
- The hyperparameters of the tree prior become dynamical parameters of the epidemiological model
- The most common approach is to leverage coalescent theory, by using coalescent machinery augmented with deterministic models of effective population size parametrized by R_0 or its epidemiological constituents (net infection rate *et cetera*).

Coalescent versus Birth-death-sampling tree prior

- Coalescent theory assumes that the sample is **small** compared to the background population.
- In cases where a large proportion of the infecteds are sampled, the small-sample approximation on which coalescent theory is based no longer applies.
- The **Swiss HIV cohort database** is estimated to contain approximately 75% of all cases of HIV infection in residents of Switzerland.
- Stadler (2010) developed an extension of the Birth death process to model transmission rate λ, removal rate μ and sampling rate ψ in cases where a viral phylogeny relating a substantial fraction of the infected individuals is available.

Birth-death-serial-sampled (BDSS) tree prior

Stadler, 2010



The per-lineage dynamics are captured by a simple set of rate equations:

$$I \xrightarrow{\lambda} 2I \qquad I \xrightarrow{\mu} R \qquad I \xrightarrow{r\psi} U \tag{7}$$

R_0 is the expected number of secondary infections per infected individual:

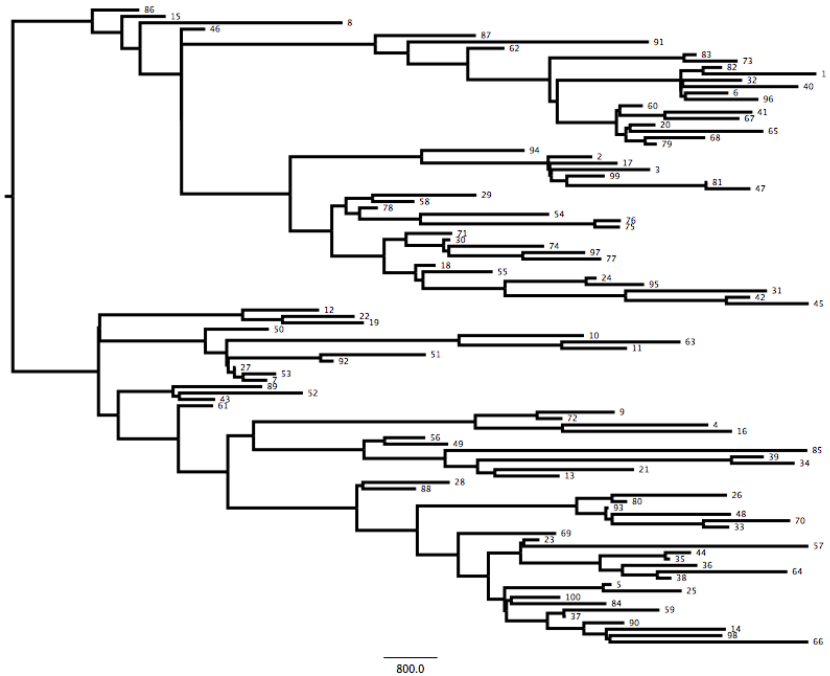
$$R_0 = \frac{\lambda}{\mu + r\psi} \tag{8}$$

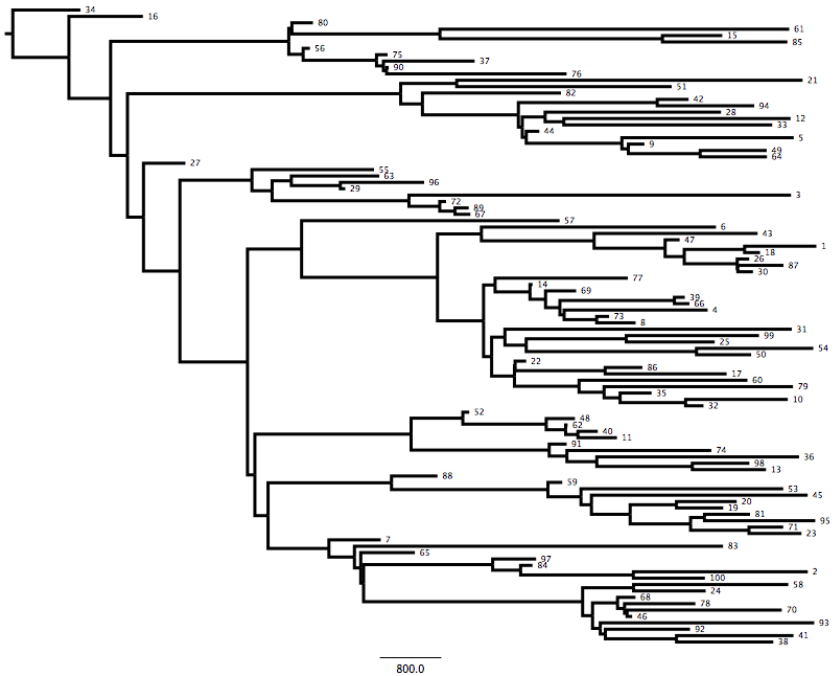
Where r is the probability that sampling removes the lineage from infectious category.

Simulation study for BDSS tree prior

- Implement BDSS likelihood (Stadler, 2010) in BEAST 1.6
- Fix $\lambda = 2$, $\mu = 1$, $\psi = 0.5$ and $r = 1$ and use MCMC to sample 100-tip trees from the BDSS prior [implies $R_0 = 4/3$]
  - This requires sampling the terminal node ages as well as internal node ages.
- Select 100 trees at random from posterior and, for each, attempt to re-estimate R_0 by two alternative parametric tree priors:
  - BDSS
  - Exponential growth coalescent tree prior

Simulated Birth-Death-Sampling trees





Parameter estimation with BDSS

Table: The measure of accuracy of re-estimating λ, μ_r, and ψ when $r = 1$.

true value	$r = 1$	original
$\lambda = 2$	mean of median	1.672845
	relative bias	−0.1635773
	interval width	4.369588
	95% HPD accuracy	100%
$\mu_r = 0.5$	mean of median	0.4510442
	relative bias	−0.09791154
	interval width	0.9122245
	95% HPD accuracy	100%
$\psi = 0.5$	mean of median	0.6180152
	relative bias	0.2360303
	interval width	0.8335251
	95% HPD accuracy	100%

Parameter estimation with BDSS

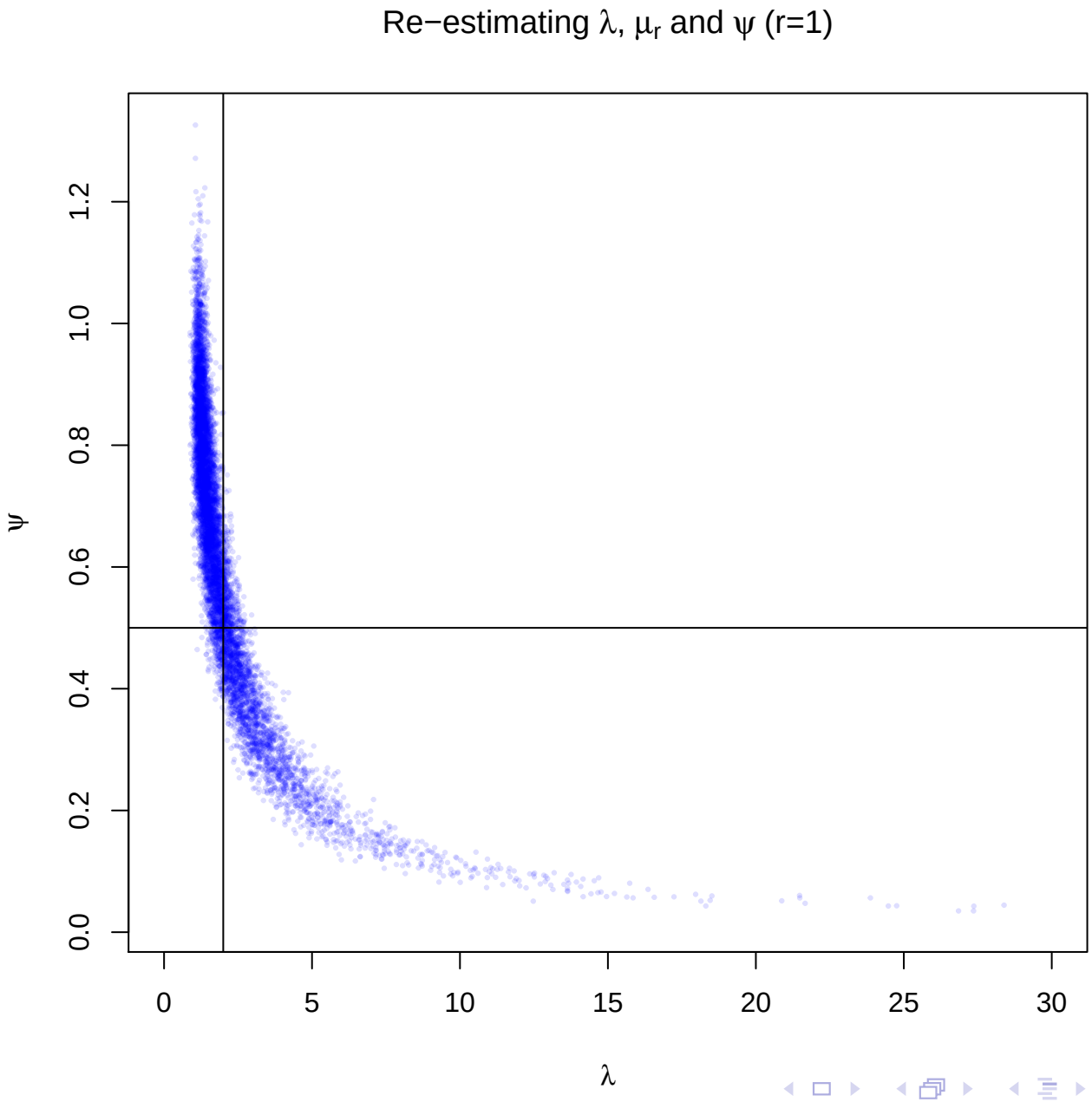
Table: The measure of accuracy of re-estimating λ, μ_r, and ψ when $r = 1$.

true value	$r = 1$	original
$R_0 = 1\frac{1}{3}$	mean of median	1.179817
	relative bias	-0.1151375
	mean of 95%HPD within	0.6028624
	95% HPD accuracy	94%
$\overline{\lambda\mu} = 1$	mean of median	0.9095235
	relative bias	-0.09047647
	95%HPD interval width	0.9159325
	95% HPD accuracy	100%
$Contour(\overline{\lambda\mu}, \psi)$	95% HPD accuracy	95%

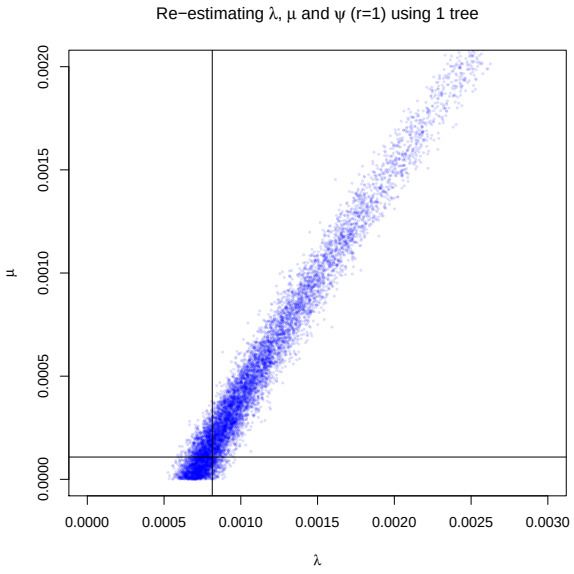
Table: The measure of accuracy of re-estimating λ, μ_r, and ψ when $r = 1$.

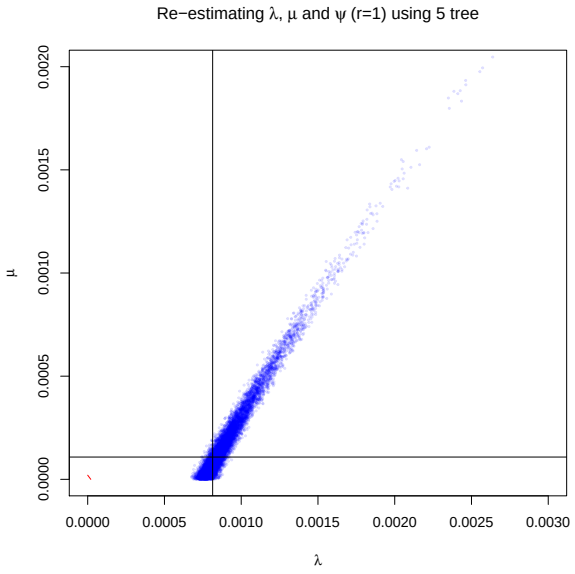
true value	$r = 1$	1 tree	2 trees	5 trees
$\lambda = 8.14 \times 10^{-4}$	mean of median	0.001085686	0.0009568504	0.0008405419
	relative bias	−0.3337661	0.1754918	0.03260674
	HPD interval width	0.002852294	0.001412868	0.0005498708
	95% HPD accuracy	100%	100%	100%
$\mu = 1.08 \times 10^{-4}$	mean of median	0.000409305	0.0002606213	0.0001247519
	relative bias	2.789861	1.413160	0.1551097
	HPD interval width	0.0029275	0.001495289	0.000602317
	95% HPD accuracy	100%	100%	100%
$\psi = 2.82 \times 10^{-4}$	mean of median	0.0002149549	0.0002406566	0.0002690238
	relative error	0.2379791	0.1516403	0.06201653
	relative bias	−0.2377486	−0.1466078	−0.04601496
	HPD interval width	0.0003096437	0.0002512993	0.0001617272
	95% HPD accuracy	91%	97%	99%
$R_0 = 2.087$	mean of median	1.674307	1.851440	2.079501
	relative error	0.1986405	0.1208391	0.0499755
	relative bias	−0.1977448	−0.1128703	−0.003593053
	HPD interval width	1.640055	1.516478	1.197511
	95% HPD accuracy	93%	99%	100%
$\overline{\lambda\mu} = 7.06 \times 10^{-4}$	mean of median	0.0006664413	0.0006867128	0.0007078903
	relative error	0.08665282	0.05635478	0.03325597
	relative bias	−0.05603218	−0.02731894	0.002677537
	HPD interval width	0.000406993	0.0003025207	0.0001876417
	95% HPD accuracy	98%	100%	100%
$Contour(\overline{\lambda\mu}, \psi)$	95% HPD accuracy	100%	100%	100%

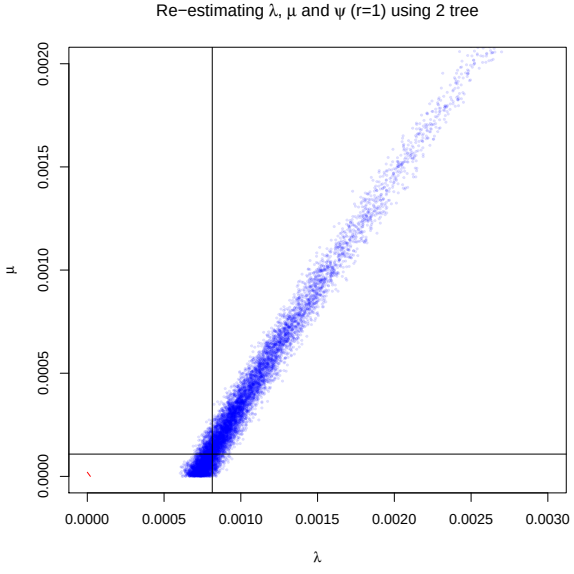
Joint estimate of λ and ψ from a single 100 tip tree

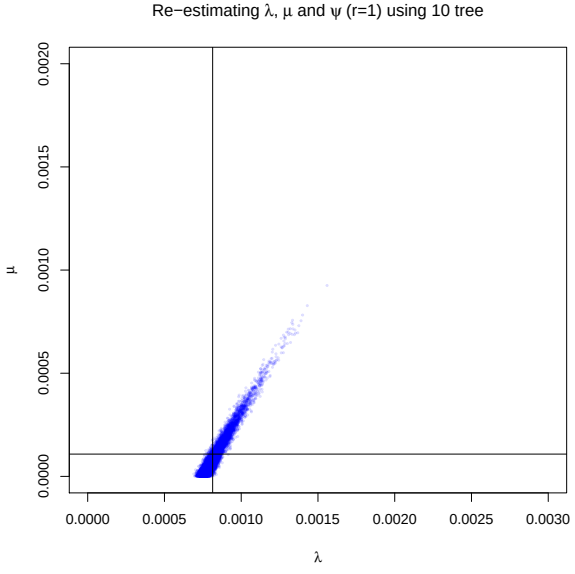


Joint estimate of λ and μ from 1, 2, 5, 10 trees

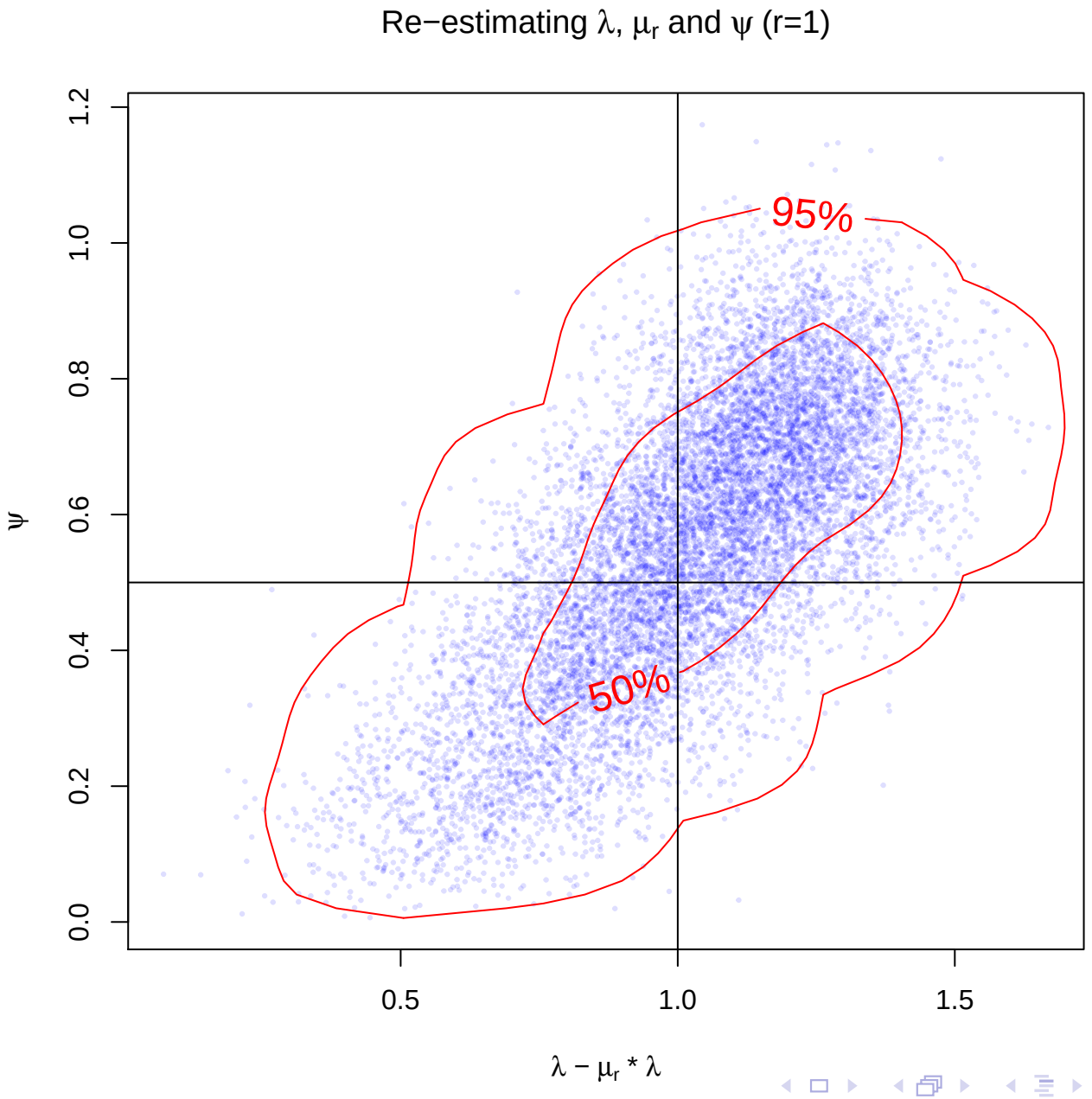








Joint estimate of $\lambda - \mu$ and ψ from a single 100 tip tree



Connecting coalescent growth rates and epidemic models

There is a simple relationship between R_0 and growth rate g at the start of the epidemic:

$$R_0 = 1 + \frac{g}{d} \tag{9}$$

where d is total death rate (Wallinga & Lipsitch, 2007). Taking:

$$d = \mu + r\psi \tag{10}$$

$$g = \lambda - d \tag{11}$$

it is easy to show this R_0 is the same as for BDSS model, so coalescent-estimated g is also an estimate of $\lambda - \mu - r\psi$.

However the proportion of infected individuals sampled by the BDSS in their lifetime is $\frac{r\psi}{\mu + r\psi} = \frac{0.5}{1+0.5} = 1/3$, a clear violation of the small sample condition that coalescent inference relies on.

Can we still estimate g accurately with exponential coalescent?

Estimating growth rate based on coalescent approach

Table: The measure of accuracy of estimating growth rate g in exponential growth tree prior, where true value $g = \lambda - \mu - r\psi = 4.24 \times 10^{-4}$

BDSS	1 tree	2 trees	5 trees
mean of median	0.0004488872	0.0004460723	0.0004396722
relative error	0.1705658	0.1316335	0.07757073
relative bias	0.05869633	0.05205729	0.03696277
HPD interval width	0.0003617696	0.0002531587	0.0001581470
95% HPD accuracy	95%	96%	93%
Coalescent	1 tree	2 trees	5 trees
mean of median	0.0004845768	0.0004319822	0.0004147897
relative error	0.2701248	0.1972525	0.1244552
relative bias	0.1428698	0.01882604	−0.02172247
HPD interval width	0.0001942935	0.0001265572	7.699674×10^{-5}
95% HPD accuracy	48%	46%	46%

Conclusions

- Using relaxed molecular clocks allows reconciliation of non-clock-like sequence alignment data with time-trees, for which rich parameter stochastic models are available.
- Calculating the joint distribution of a set of (possibly nested) t_{MRCA} under random trees from common stochastic tree models would be very useful for constructing **calibrated Bayesian time tree priors.**
- Explicitly modeling birth, death and sampling permits a richer epidemic inference from phylogenetic trees than Kingman's coalescent, but parameter estimates are highly correlated.