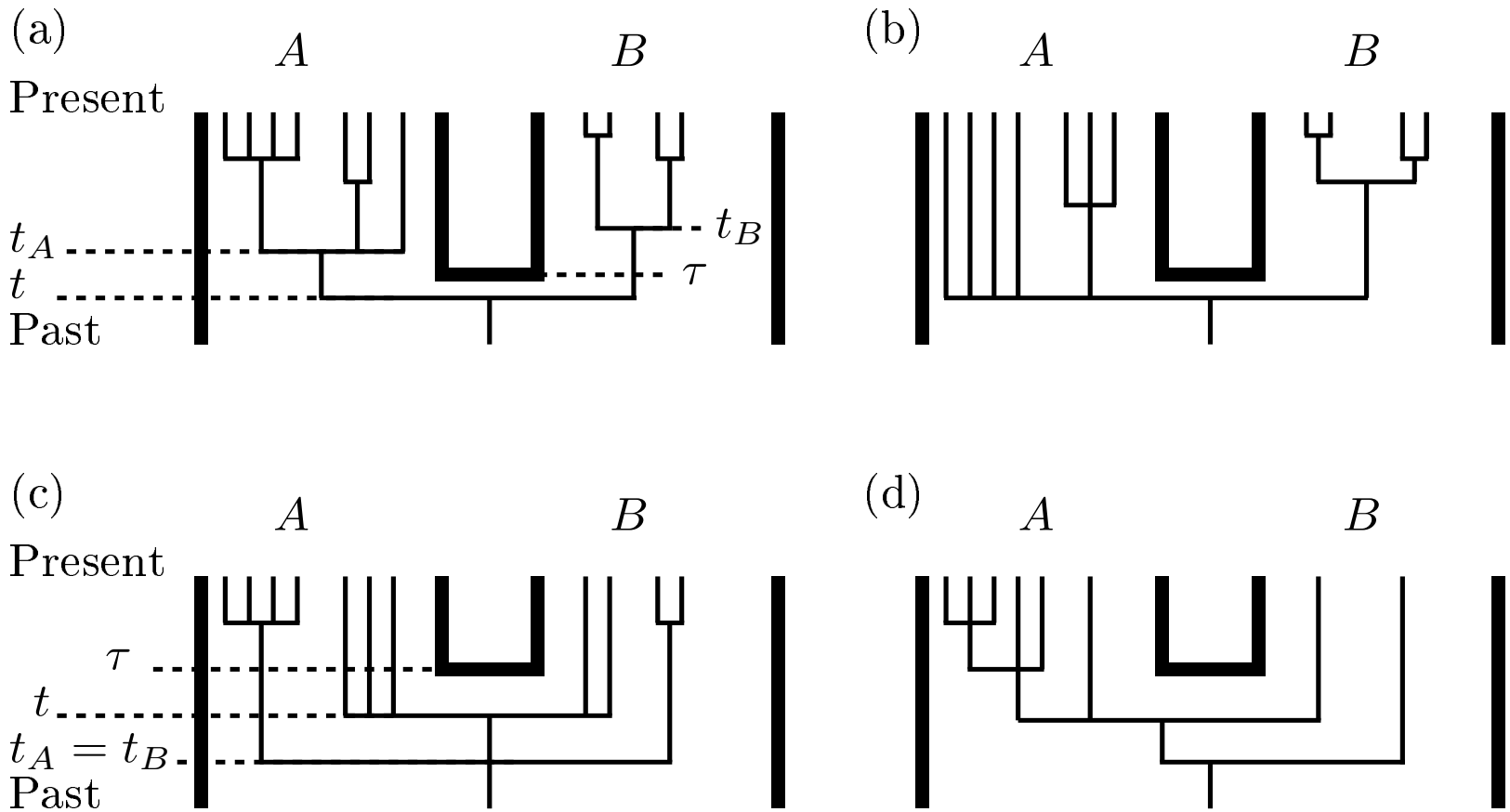
Concordance between species trees and gene genealogies with asynchronous multiple mergers

Bjarki Eldon[1] and James Degnan[2]
[1]University of Oxford, Oxford, UK
[2]University of Canterbury, Christchurch, New Zealand

Gene genealogies in two species A and B



Kingman coalescent

Kingman (1982) coalescent only allows two ancestral lineages to coalesce each time

$$q_{\xi\to\eta} = \begin{cases} 1 & \xi \prec \eta \\ -\binom{|\xi|}{2} & \xi = \eta \\ 0 & \text{otherwise} \end{cases}$$

Λ coalescent

Donnelly and Kurtz (1999), Sagitov (1999), and Pitman (1999) study Λ coalescent which allows asynchronous multiple mergers of ancestral lineages

$$q_{\xi,\eta} = \begin{cases} \int_0^1 x^{k-2}(1-x)^{|\xi|-k} d\Lambda(x) & \xi \prec \eta \\ & 2 \leq k \leq |xi| \\ -\sum_{\xi \prec \eta} q_{\xi,\eta} & \xi = \eta \\ \\ 0 & \text{otherwise} \end{cases}$$

Heavy-tail population model

Schweinsberg (2003)
Each individual contributes X_i potential offspring with tail probabilities, C and α positive constants,

$$\lim_{k \to \infty} Ck^{\alpha} \mathbb{P}[X_i \geq k] = 1$$

Can sample N offspring from pool of potential offspring since

$$\mathbb{E}[X_i] > 1$$

Beta$(2-\alpha, \alpha)$ coalescent

If $1 < \alpha < 2$ obtain a Λ coalescent, $B(\cdot, \cdot)$ is beta function,

$$q_{\xi,\eta} = \begin{cases} \dfrac{B(k-\alpha, |\xi| - k + \alpha)}{B(2-\alpha, \alpha)} & \xi \prec \eta \\ & 2 \leq k \leq |xi| \\ -\sum_{\xi \prec \eta} q_{\xi,\eta} & \xi = \eta \\ 0 & \text{otherwise} \end{cases}$$

Point mass model

Eldon and Wakeley (2006)
One parent each timestep contributes a random number U of offspring to replace those who perished

$$\mathbb{P}[U = u] = \left(1 - N^{-\gamma}\right) \delta_{u,1} + N^{-\gamma} \delta_{u,\lfloor \psi N \rfloor}$$

Gives coalescence rates, if $0 < \gamma < 2$,

$$\lambda_{b,k} = \binom{b}{k} \psi^k \left(1 - \psi\right)^{b-k}$$

Spectral expansion of the rate matrix

Let $(A_t)_{t\geq 0}$ denote the Markov chain counting the number of ancestral lineages in the coalescent.
Need to compute $\mathbb{P}[A_t = j | A_0 = i]$

$$\mathbb{P}[A_t = j | A_0 = i] = \sum_{k=j}^{i} e^{-t\lambda_k} r_i^{(k)} \ell_j^{(k)}$$

The right and left eigenvectors $r^{(k)}$ and $\ell^{(k)}$, respectively, are straightforward to obtain for the Kingman coalescent (Tavaré 1984)

Sprectal expansion for Λ coalescent

In case of a Λ coalescent, $r^{(k)}$ and $\ell^{(k)}$ can be obtained by recursion

$$\ell_j^{(k)} = \frac{q_{j+1,j}\ell_{j+1}^{(k)} + \cdots + q_{k,j}\ell_k^{(k)}}{q_k - q_j}, \quad 1 \le j < k$$

$\ell_j^{(k)} = 0$ if $j > k$; and

$$r_j^{(k)} = \frac{q_{j,k}r_k^{(k)} + \cdots + q_{j,j-1}r_{j-1}^{(k)}}{q_k - q_j}, \quad 1 < k < j \le n,$$

$r_j^{(k)} = 0$ if $j < k$.

Conditioning on the embedded chain not practical ...

Another method is conditioning on the paths of A_t
Conditional on path a of A_t, $T(a)$ sum of indep. Exponentials

$$g_{i,j}(t,a) = \begin{cases} \mathbb{P}[T(a) \leq t, T(a) + T_j > t] & 2 \leq j < i \\ \mathbb{P}[T(a) \leq t] & j = 1 \\ e^{-q_i t} & j = i \end{cases}$$

and

$$g_{i,j}(t) = \sum_a g_{i,j}(t,a)p(a)$$

Not practical, since 2^{i-j-1} possible paths from i to $j < i$

... unless the most probable paths can be identified

		fractiles			
π	i	50%	75%	90%	$c^\star$
1.01	5	0.324	0.449	0.643	$(2, \ldots, 2)$
	20	$6.7 \cdot 10^{-4}$	0.0013	0.0024	$(19, 2)$
1.2	5	0.194	0.306	0.533	$(2, \ldots, 2)$
	20	$1.4 \cdot 10^{-3}$	$3.3 \cdot 10^{-3}$	$7.4 \cdot 10^{-3}$	$(18, 2, 2)$
1.5	5	0.083	0.153	0.417	$(2, \ldots, 2)$
	20	$1.8 \cdot 10^{-5}$	$7.5 \cdot 10^{-5}$	$2.9 \cdot 10^{-4}$	$(2, \ldots, 2)$
0.01	5	0.002	0.008	0.307	$(2, \ldots, 2)$
	20	$1.2 \cdot 10^{-15}$	$4.3 \cdot 10^{-13}$	$7.9 \cdot 10^{-11}$	$(2, \ldots, 2)$
0.2	5	0.062	0.209	0.506	$(2, \ldots, 2)$
	20	$5.1 \cdot 10^{-4}$	0.003	0.016	$(5, 4, 4, 3, 2, \ldots, 2)$
0.5	5	0.409	0.576	0.767	$(3, 2, 2)$
	20	$1.2 \cdot 10^{-13}$	$2.1 \cdot 10^{-10}$	$1.2 \cdot 10^{-7}$	$(11, 6, 3, 2, 2)$

Computing probability P of reciprocal monophyly ...

Define T as the time when lineages from populations A *and* B first coalesce
Define $T_X \equiv \inf\{t : A_t = 1\}$ for population $X \in \{A, B\}$
The probability P of reciprocal monophyly is given by

$$P = \mathbb{P}[T > T_A, T > T_B]$$

... recursively

The probability P is computed recursively

$$P = \sum_{m_A=1}^{n_A} \sum_{m_B=1}^{n_B} P(m_A, m_B) g_{n_A,m_A,\pi_A}(\tau) g_{n_B,m_B,\pi_B}(\tau)$$

where τ is the time of divergence, and, with $m = m_A + m_B$,

$$P(m_A, m_B) = \sum_{k=2}^{\max(m_A,m_B)} \left(\frac{\binom{m_A}{k}}{\binom{m}{k}} p(m,k) P(m_A - k + 1, m_B) + \frac{\binom{m_B}{k}}{\binom{m}{k}} p(m,k) P(m_A, m_B - k + 1) \right)$$

with $P(1,1) = 1$, and $p(m,k)$ is probability of k-merger among $m = m_A + m_B$ lineages

Paraphyly and polyphyly

Probabilities of paraphyly and polyphyly can be obtained similarly. Define $P_A = \mathbb{P}[T > T_A]$. The probability P_B^* of paraphyly of B with respect to A is

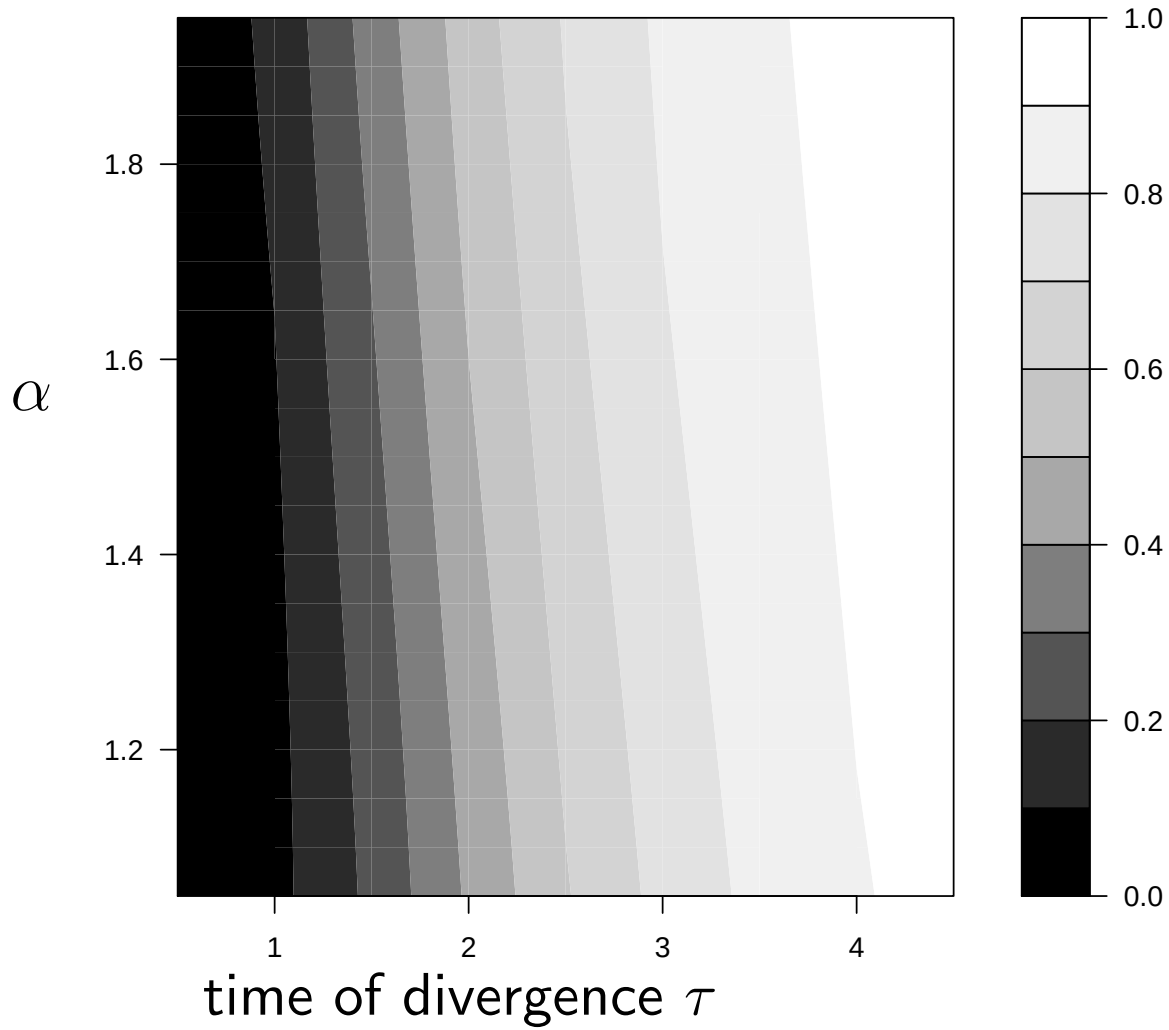
$$P_B^* = \mathbb{P}[T > T_A, T \leq T_B] = \mathbb{P}[T > T_A] - \mathbb{P}[T > T_A, T > T_B]$$

Polyphyly is the event $\{T \leq T_A\} \cap \{T \leq T_B\}$ which occurs with probability $P^\star$ given by

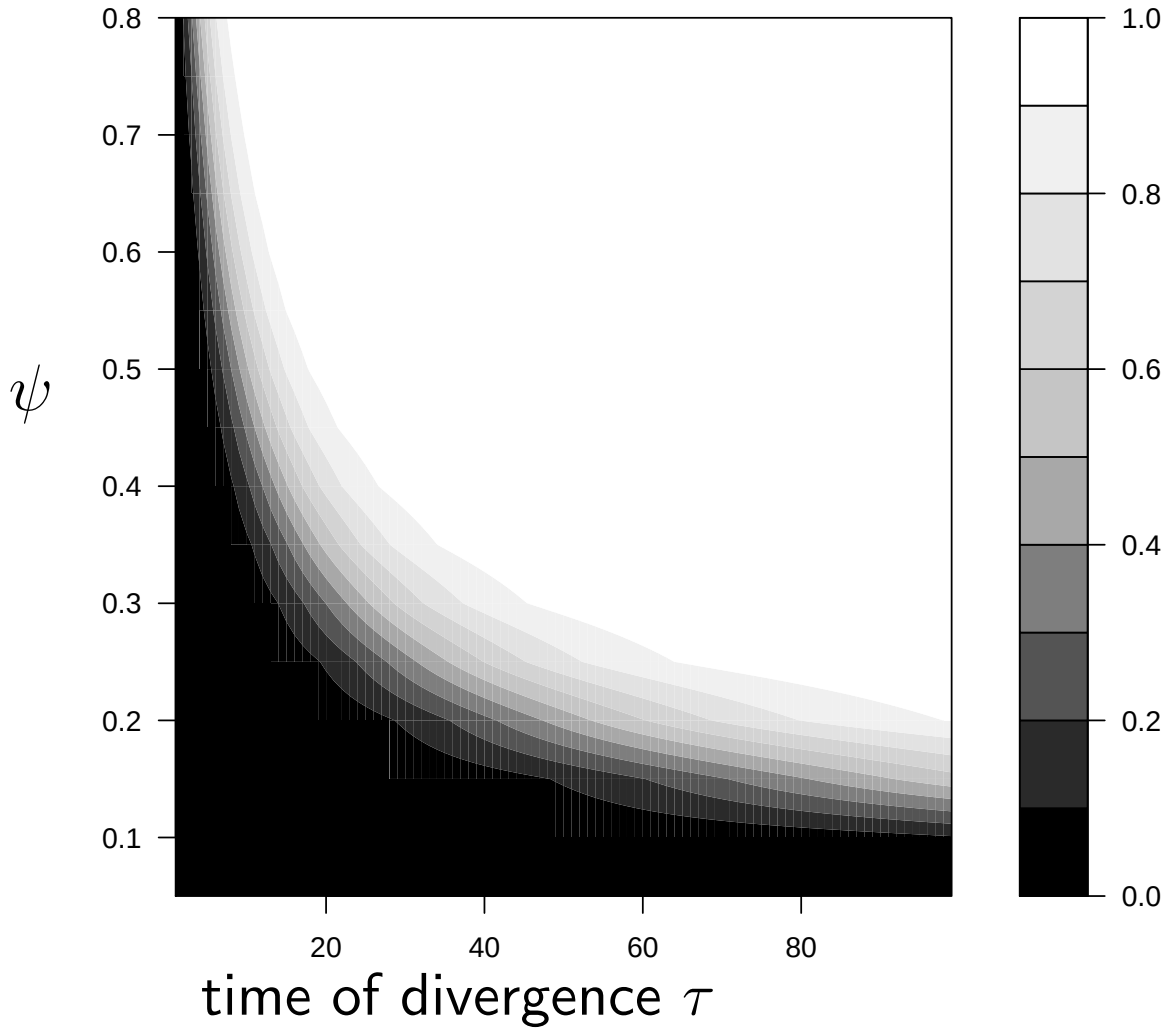
$$P^\star = \mathbb{P}[T \leq T_A, T \leq T_B] = 1 - P_A - P_B + P.$$

The probabilities P_A and P_B can be obtained recursively analogously to P

Probability P of monophyly as a function of τ and α

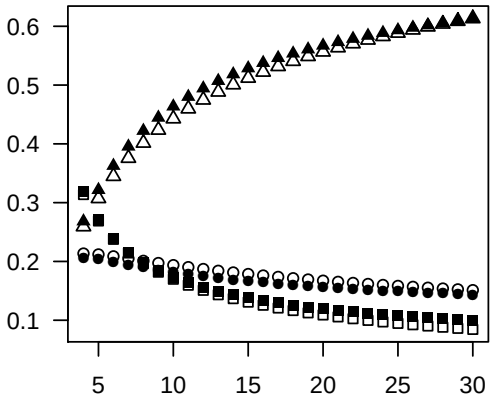


Probability P of monophyly as a function of τ and ψ

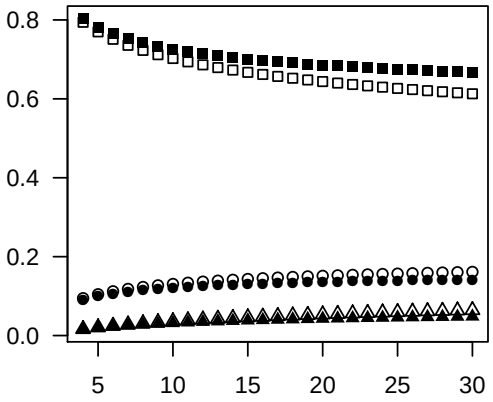


Monophyly (□, ■), Paraphyly (○, ●), Polyphyly (△, ▲)

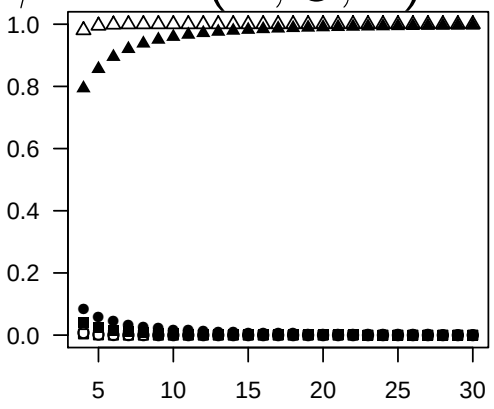
$\alpha = 1.05$ (□, ○, △)
$\alpha = 1.5$ (■, ●, ▲)



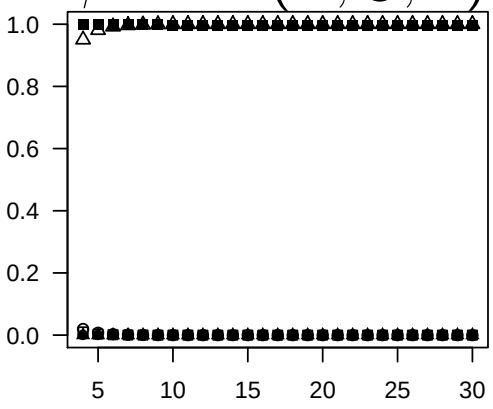
$\alpha = 1.05$ (□, ○, △)
$\alpha = 1.5$ (■, ●, ▲)



$\psi = 0.05$ (□, ○, △)
$\psi = 0.5$ (■, ●, ▲)

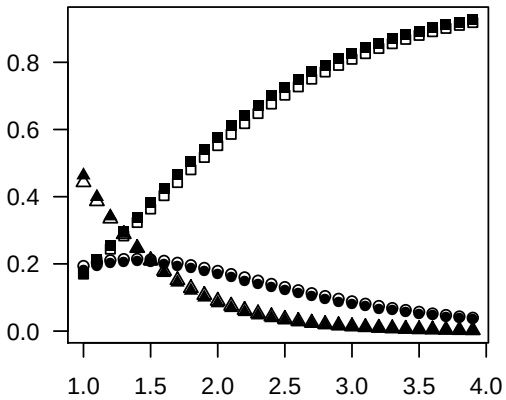


$\psi = 0.05$ (□, ○, △)
$\psi = 0.5$ (■, ●, ▲)

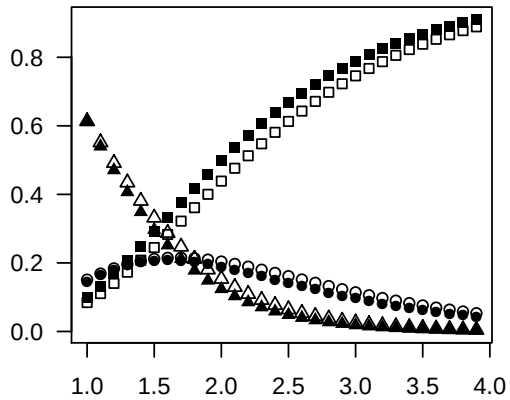


Monophyly (□, ■), Paraphyly (○, ●), Polyphyly (△, ▲)

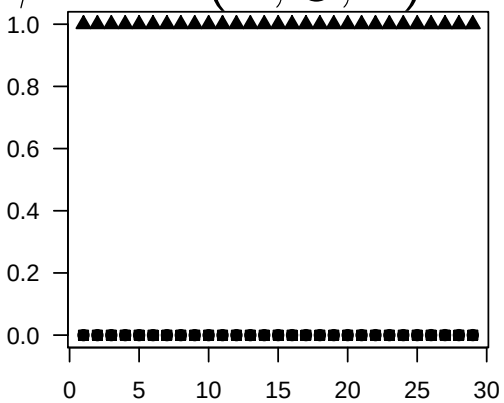
$\alpha = 1.05$ (□, ○, △)
$\alpha = 1.5$ (■, ●, ▲)



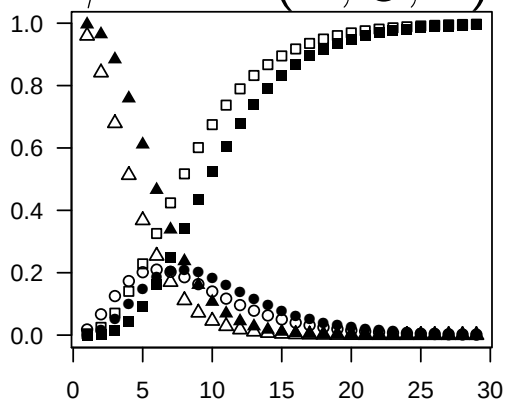
$\alpha = 1.05$ (□, ○, △)
$\alpha = 1.5$ (■, ●, ▲)



$\psi = 0.05$ (□, ○, △)
$\psi = 0.5$ (■, ●, ▲)



$\psi = 0.05$ (□, ○, △)
$\psi = 0.5$ (■, ●, ▲)



time of divergence τ

Unit of time of divergence

Different coalescent processes have different timescales
Unit of time is c_N given by

$$c_N = \frac{\mathbb{E}[\nu_1(\nu_1 - 1)]}{N - 1}$$

For the Beta$(2 - \alpha, \alpha)$ coalescent, $c_N^{-1} N^{\alpha - 1}$, $1 < \alpha < 2$
For the ψ-coalescent, $c_N^{-1} \sim N^{\gamma}$ with $1 < \gamma < 2$
For the Kingman coalescent $c_N \sim N$ (WF), or N^2 (Moran)

Conclusions

- The effects of coalescent parameters on probabilities on monophyly, paraphyly, and polyphyly depend on the coalescent process and if the population is ancestral or derived
- When different populations have different coalescent processes running on different timescales, scaling time of divergence becomes a key issue in terms of inference

Acknowledgments

- The organisers: Vincent Moulton, Mike Steel, Tandy Warnow
- Newton Institute
- EPSRC, Royal Society, Marsden fund