Monophyletic concordance between species trees and gene genealogies with multiple mergers

Bjarki Eldon and James Degnan

Phylomania 2010

University of Tasmania

November 4-5, 2010

Low offspring number models

Kingman (1982) introduced the n-coalescent from an exchangeable Cannings offspring model; let ν_i denote the number of offspring of individual i

$$\mathbb{E}[\nu_1^k] < \infty \quad \text{as} \quad N \to \infty; \quad k \geq 1$$

Möhle and Sagitov (2001) characterised coalescent processes based on the timescale c_N

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

Conditions for convergence to Kingman's coalescent

Wright-Fisher and Moran models are exchangeable Cannings models with

$$\lim_{N\to\infty} \frac{\mathbb{E}[\nu_1(\nu_1-1)(\nu_1-2)]}{N^2 c_N} = 0$$

implying $c_N \to 0$ and convergence to Kingman's coalescent.

High variance in offspring distribution

Ecology, reproductive biology, and genetics of a diverse group of marine organisms suggest many offspring contributed by few individuals (Beckenbach 94; Hedgecock 94)

Direct genotyping of parents and offspring provides evidence of large families in Pacific oyster (Boudry etal 2002) and Lion-Paw Scallop (Petersen etal 2008)

Cod, oysters, mussels, barnacles, sea stars, plants ?

Evidence for large offspring distribution

- broadcast spawning and external fertilization
- high initial mortality
- very large population sizes
- low genetic diversity
- large number of singleton genetic variants

Λ coalescent allows multiple mergers

Donnelly and Kurtz (1999), Pitman (1999), and Sagitov (1999) independently introduce a multiple merger coalescent; Λ-coalescent with coalescence rate

$$\lambda_{b,k} = \binom{b}{k} \int_0^1 x^k (1-x)^{b-k} x^{-2} \Lambda(dx)$$

Kingman's coalescent is obtained if $\Lambda = \delta_0$
For simultaneous multiple merger coalescent processes, see Schweinsberg (2000) and Möhle and Sagitov (2001).

Schweinsberg's heavy-tail model

Schweinsberg (2003)
Each individual produces a random number X_i of potential offspring; $C > 0$ and $a > 0$ and constant population size N

$$\mathbb{P}[X_i \geq k] \sim C/k^a$$

and

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

From the pool of potential offspring, sample without replacement to form the new generation

Coalescent process depends on a

Coalescent timescale in units of $c_N \sim N^{a-1}$ if $1 < a < 2$

case	coalescent	coalescence rate
$a \geq 2$	Kingman coalescent	$\binom{b}{2}$
$1 \leq a < 2$	$\Lambda \sim Beta(2-a, a)$	$\binom{b}{k} \frac{B(k-a, b-k+a)}{B(2-a, a)}$
$0 < a < 1$	Ξ-coalescent	

A modified Moran model

Eldon and Wakeley (2006)
A modified Moran model, in which the offspring number U is random rather than fixed at one as in the usual Moran model

$$\mathbb{P}[U = u] = (1 - \varepsilon_N)\delta_2 + \varepsilon_N \delta_{[\psi N]}$$

and

$$\varepsilon_N \sim 1/N^\gamma, \quad \gamma > 0$$

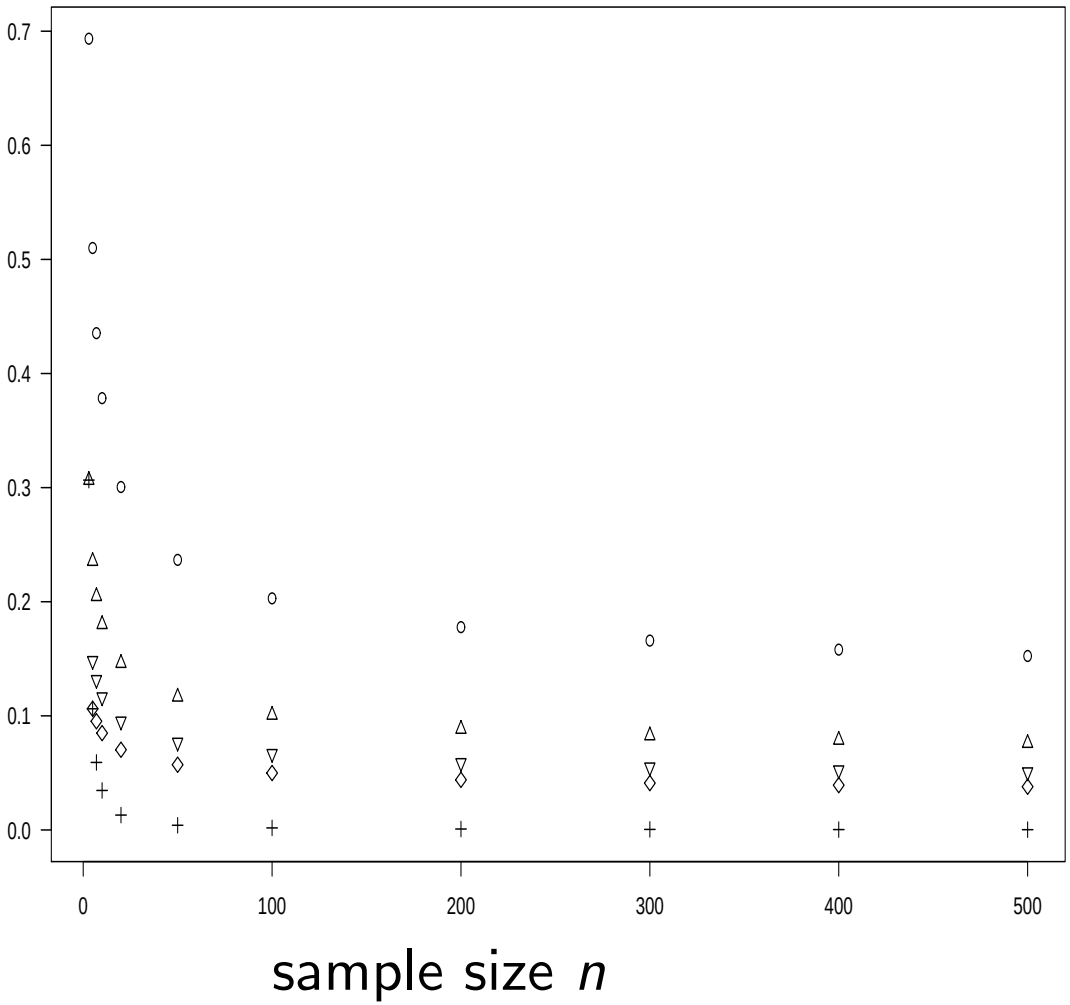
Coalescent process depends on γ

Coalescent timescale is $N_\gamma = \min\left(N^\gamma, N^2\right), \gamma > 0$

case	coalescence rate	timescale
$\gamma > 2$	$\binom{n}{2}$	N^2
$\gamma = 2$	$\binom{b}{k}\left(\delta_2 + \psi^k(1-\psi)^{b-k}\right)$	N^2
$\gamma < 2$	$\binom{b}{k}\psi^k(1-\psi)^{b-k}$	N^γ, $1 < \gamma < 2$

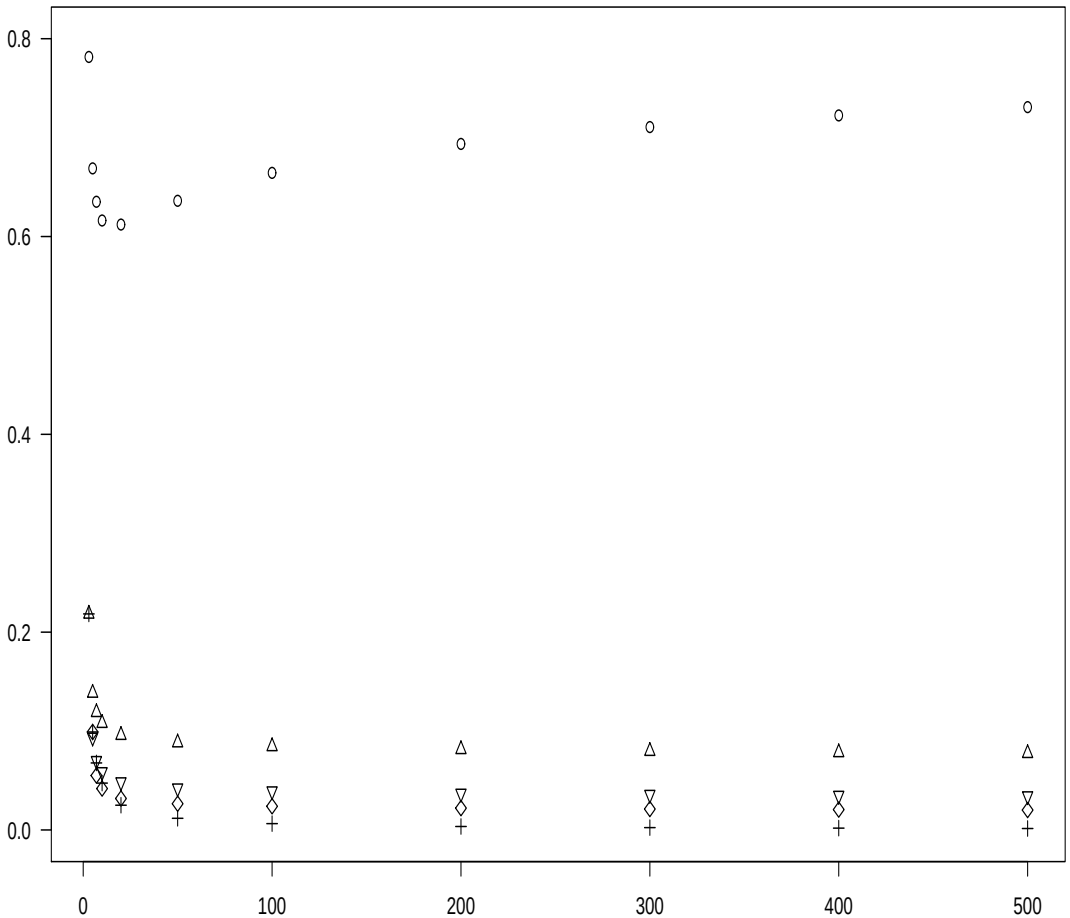
Ratios of coalescence times for $\Lambda = K + \Lambda_\psi$

$\circ : \overline{R_1}; \quad \triangle : \overline{R_2}; \quad \bigtriangledown : \overline{R_3}; \quad \diamond : \overline{R_4}; \quad + : \overline{R_{n-1}}$



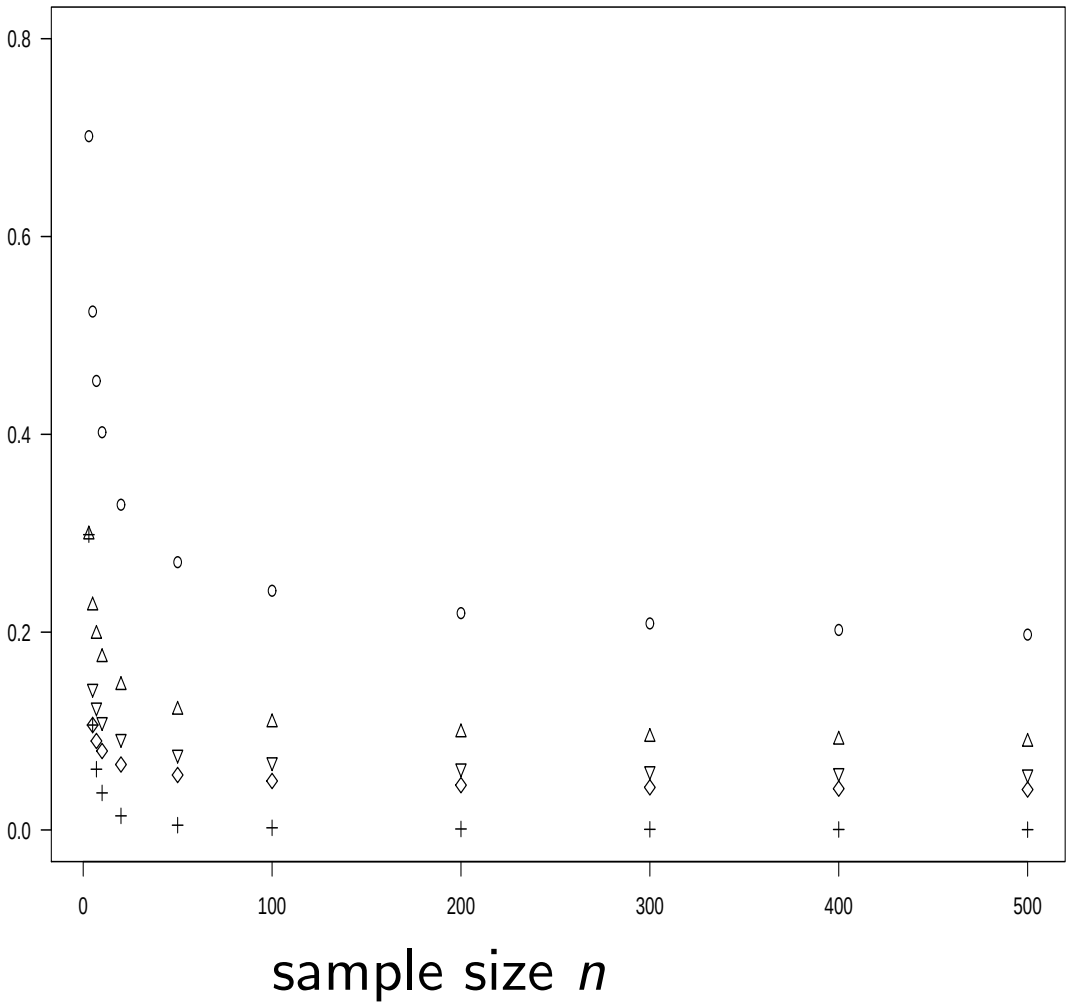
Ratios of coalescence times for $\Lambda = Beta(0.9, 1.1)$

$\circ : \overline{R_1}; \quad \triangle : \overline{R_2}; \quad \bigtriangledown : \overline{R_3}; \quad \diamond : \overline{R_4}; \quad + : \overline{R_{n-1}}$

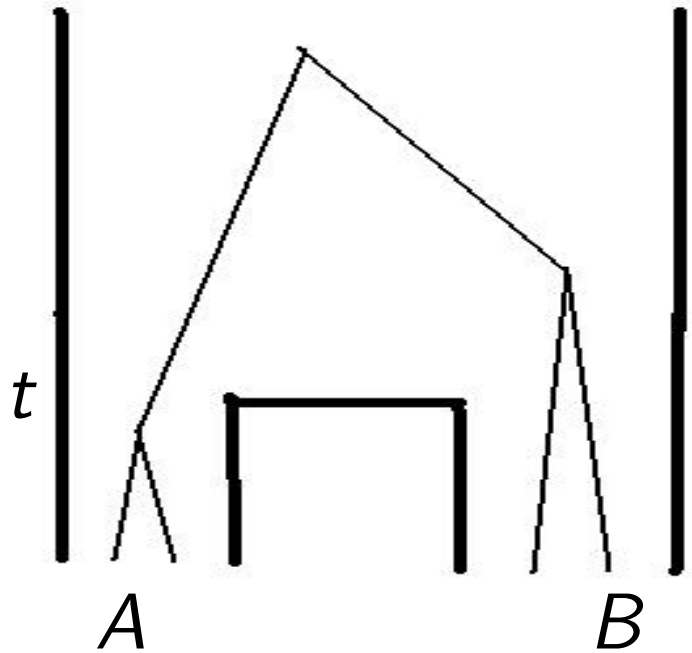


Ratios of coalescence times for $\Lambda = Beta(0.1, 1.9)$

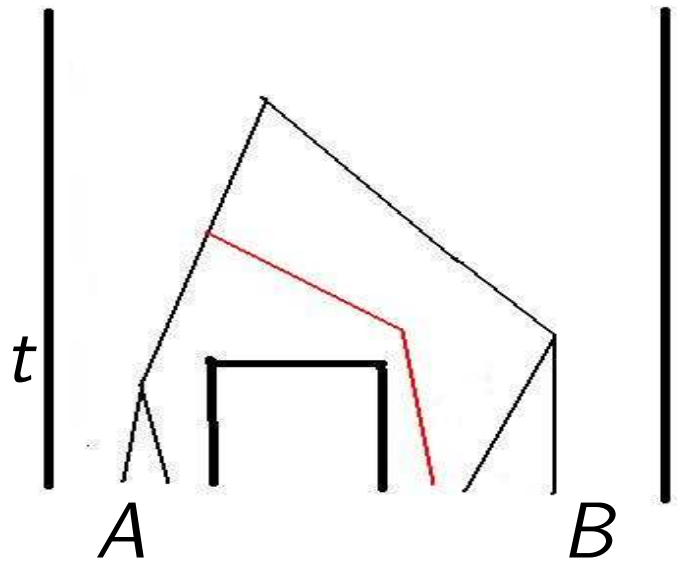
$\circ : \overline{R_1}; \quad \triangle : \overline{R_2}; \quad \bigtriangledown : \overline{R_3}; \quad \diamond : \overline{R_4}; \quad + : \overline{R_{n-1}}$



Monophyletic concordance for Λ coalescents



Not monophyletic concordance



General form for $\mathbb{P}[MC]$ for two species

$$\mathbb{P}[MC] = \sum_{m_A, m_B} \mathbb{P}[MC; m_A, m_B]\mathbb{P}[m_A, m_B]$$

with

$$\mathbb{P}[n_A, n_B] = G_{n_A, m_A}(t) G_{n_B, m_B}(t)$$

and

$$\mathbb{P}[MC; m_A, m_B] = \sum_{k=2}^{m_A+m_B} \beta_{m_A+m_B, k} \left(\mathbb{P}[MC; m_A - k + 1, m_B] \binom{m_A}{k} \right.$$

$$\left. + \mathbb{P}[MC; m_A, m_B - k + 1] \binom{m_B}{k} \right) / \binom{m_A + m_B}{k}$$

Computing $G_{i,j}(t)$

$G_{i,j}(t)$ is the probability of j lines at time t when starting from i lines at time zero within one population

A vector c of ordered mergers associated with Kingman's coalescent is simply $\{2, 2, \ldots, 2\}$

By way of example, starting from 10 lines, say, a coalescence sequence could be $\{3, 2, 5, 3\}$ in a Λ coalescent.

Conditioning on the *embedded chain*, or the order of mergers

Transition probabilities

$$\beta_{i,j} = \begin{cases} \dfrac{q_{i,j}}{\sum_{k \neq i} q_{i,k}} & \text{if } i \neq j \\ \\ 0 & \text{otherwise} \end{cases}$$

The rate matrix $\mathbf{Q}_A$ of $(A_t; t \geq 0)$ is

$$q_{j,i} = \binom{j}{j-i+1} \int_0^1 x^{j-i-1}(1-x)^{i-1}\Lambda(dx)$$

$$q_{j,j} = -\sum_{i=1}^{j-1} q_{j,i}, \quad 2 \leq j \leq n$$

$$q_{j,i} = 0, \quad \text{otherwise}$$

Using eigenvectors and eigenvalues of $\mathbf{Q}_A$

Eigenvalues of $\mathbf{Q}_A$ are $\alpha(k) = q_{k,k}$

Left eigenvector $\mathbf{l}^{(k)} = \left(l_1^{(k)}, \ldots, l_n^{(k)}\right)$

Right eigenvector $\mathbf{r}^{(k)} = \left(r_1^{(k)}, \ldots, r_n^{(k)}\right)$

Obtained by recursions

$$l_j^{(k)} = \frac{q_{j+1,j} l_{j+1}^{(k)} + \cdots + q_{k,j} l_k^{(k)}}{q_{k,k} - q_{j,j}}, \quad 1 \leq j < k$$

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

The spectral decomposition of $\mathbf{Q}_A$ yields the transition probabilites

$$G_{i,j}(t) \equiv \mathbb{P}[A_t = j | A_0 = i]$$

as

$$G_{i,j}(t) = \sum_{k=j}^{i} e^{-\alpha(k)t} r_i^{(k)} l_j^{(k)}$$

Transition probabilities $G_{i,j}$ for $i = 3$

$$G_{3,2}(t) = \frac{q_{3,2}}{q_{3,2} + q_{3,3}} \mathbb{P}[T_3 \leq t,\, T_3 + T_2 > t]$$

$$G_{3,1}(t) = \frac{q_{3,2}}{q_{3,2} + q_{3,3}} \mathbb{P}[T_3 + T_2 \leq t] + \frac{q_{3,3}}{q_{3,2} + q_{3,3}} \mathbb{P}[T_3 \leq t]$$

$$G_{3,3}(t) = \mathbb{P}[T_3 > t]$$

and

$$G_{3,1}(t) + G_{3,2}(t) + G_{3,3}(t) = 1$$

An example with Λ_ψ

Process with infinitesimal parameters

$$q_{ij} = \binom{i}{j} \psi^{i-j+1} (1-\psi)^{j-1}$$

For $i = 3$ we obtain, with $\alpha(k) \equiv \sum_{k=i-1}^{1} q_{ik}$

$$G_{3,2}(t) = \frac{3}{2}\left(e^{-\alpha(2)t} - e^{-\alpha(3)t}\right)$$

$$G_{3,1}(t) = 1 - \frac{3}{2}e^{-\alpha(2)t} + \frac{1}{2}e^{-\alpha(3)t}$$

$$G_{3,3}(t) = e^{-\alpha(3)t}$$

In general,

$$G_{i,j}(t) = \sum_{c \in C_{i,j}} g_c(t), \quad 1 \leq j < i$$

in which c is a *coalescence sequence*; or a particular order of mergers in going from i to j sequences.
Number of possible sequences is

$$|C_{i,j}| = 2^{i-j-1}$$

$$g_c(t) = \begin{cases} p(c)\mathbb{P}[T(c) \leq t, T(c) + T_j > t] & \text{if } j > 1 \\ p(c)\mathbb{P}[T(c) \leq t] & \text{if } j = 1 \\ \mathbb{P}[T_i > t] & \text{if } j = i \end{cases}$$

in which

$$\mathbb{P}[T(c) \leq t, T(c) + T_j > t] = e^{-\alpha(j)t} \sum_{k=1}^{I} \frac{\gamma_k}{\beta(i_k, j)} \left(1 - e^{-\beta(i_k,j)t}\right)$$

with $\beta(i_k, j) \equiv \alpha(i_k) - \alpha(j)$;
and

$$\mathbb{P}[T(c) \leq t] = \sum_{k=1}^{I} \gamma'_k \left(1 - e^{-\alpha(i_k)t}\right)$$

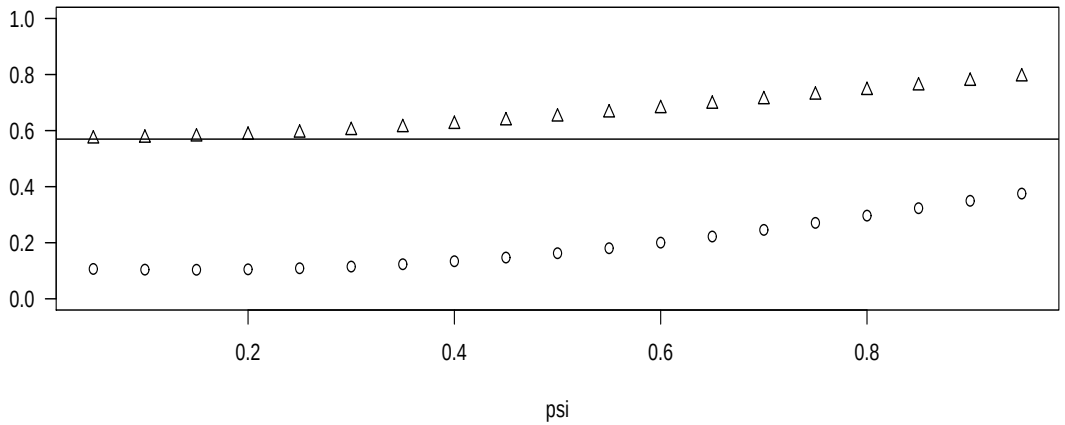
Example: two species

The probability $\mathbb{P}[MC]$ of *monophyletic concordance* for two lines from each of two species, with $\alpha_X(k) = \sum_{1 \le k \le i-1} q_{ik}$ (for species X)

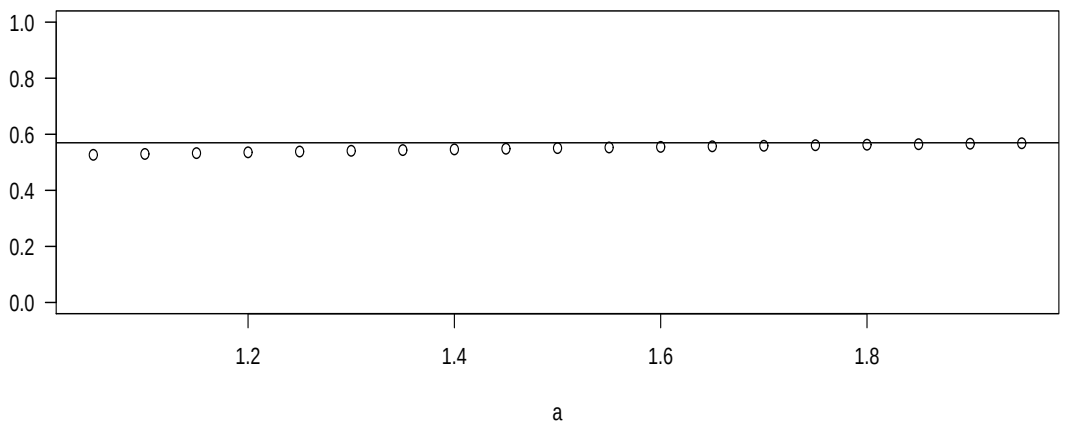
$$\begin{aligned}
\mathbb{P}[MC] = \quad & (1 - e^{-\alpha_A(2)t})(1 - e^{-\alpha_B(2)t}) \\
+ \quad & e^{-\alpha_A(2)t}(1 - e^{-\alpha_B(2)t})\beta_{3,2}/3 \\
+ \quad & (1 - e^{-\alpha_A(2)t})e^{-\alpha_B(2)t}\beta_{3,2}/3 \\
+ \quad & e^{-\alpha_A(2)t}e^{-\alpha_B(2)t}\beta_{4,2}\beta_{3,2}/9
\end{aligned}$$

Two species and two lines each

$\circ : \Lambda_\psi; \quad \triangle : K + \Lambda_\psi$

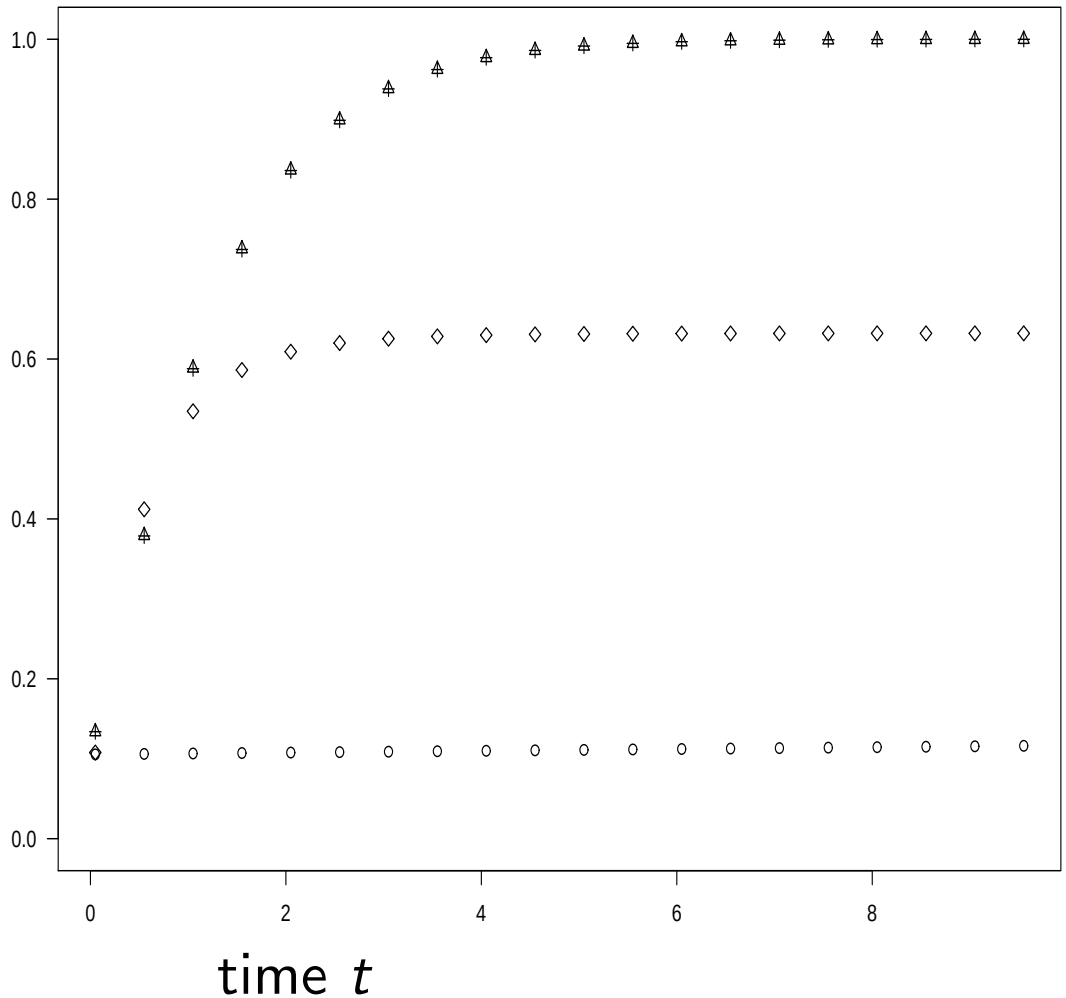


$\circ : Beta(2 - a, a)$



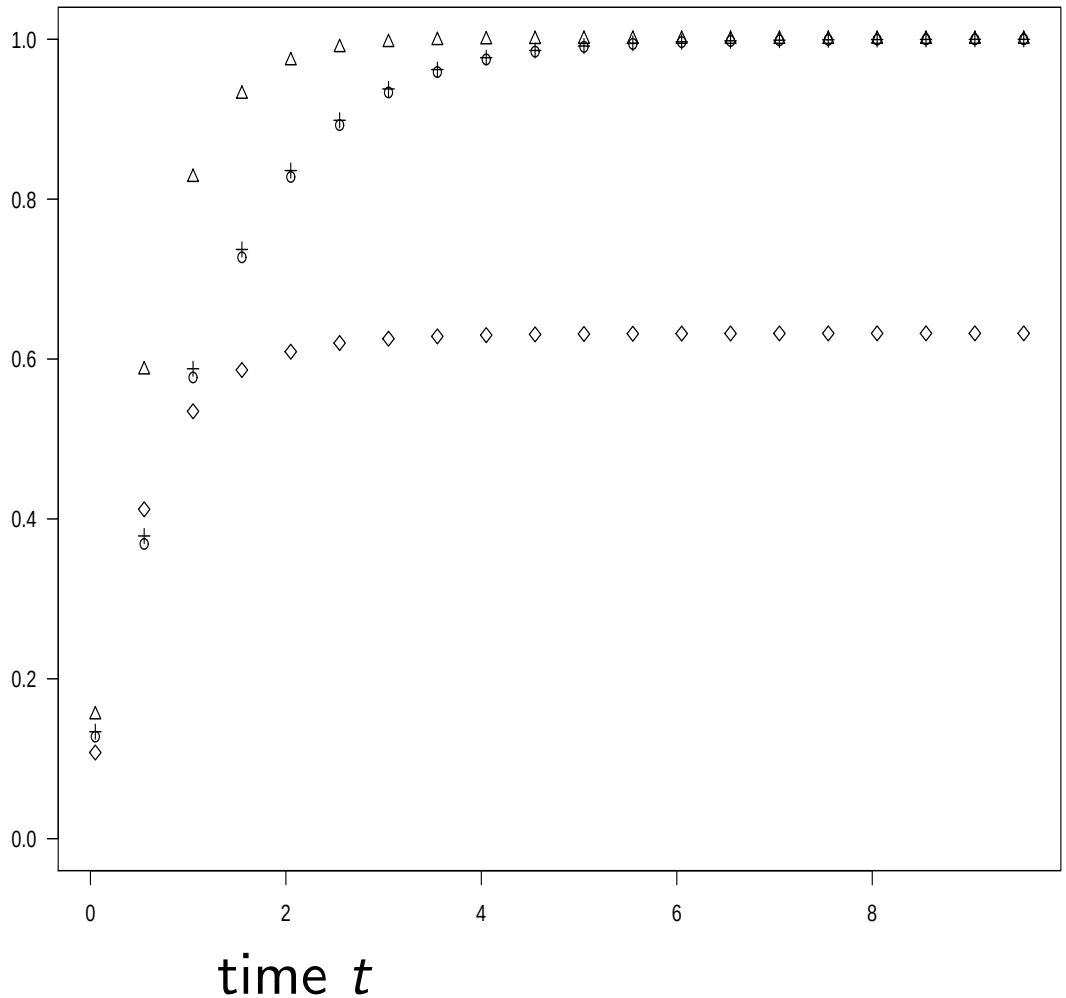
Two species and two lines each

$\circ : \Lambda_{0.05}; \quad \triangle : K + \Lambda_{0.05}; \quad \diamond : Beta(0.95, 1.05); \quad + : K$



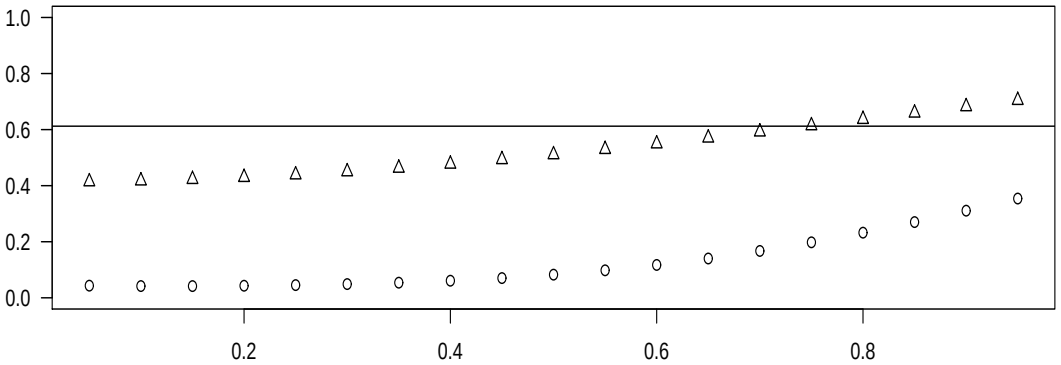
Two species and two lines each

$\circ : \Lambda_{0.99}; \quad \triangle : K + \Lambda_{0.99}; \quad \diamond : Beta(0.05, 1.95); \quad + : K$

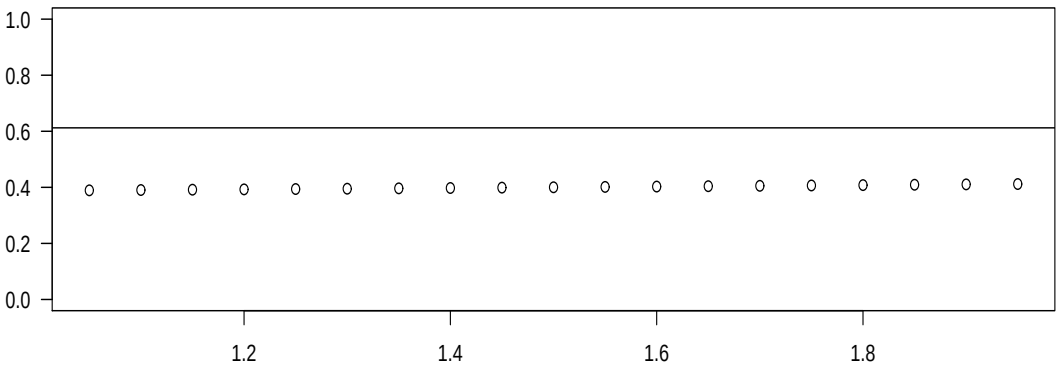


Two species and three lines each

$\circ : \Lambda_\psi; \quad \triangle : K + \Lambda_\psi$

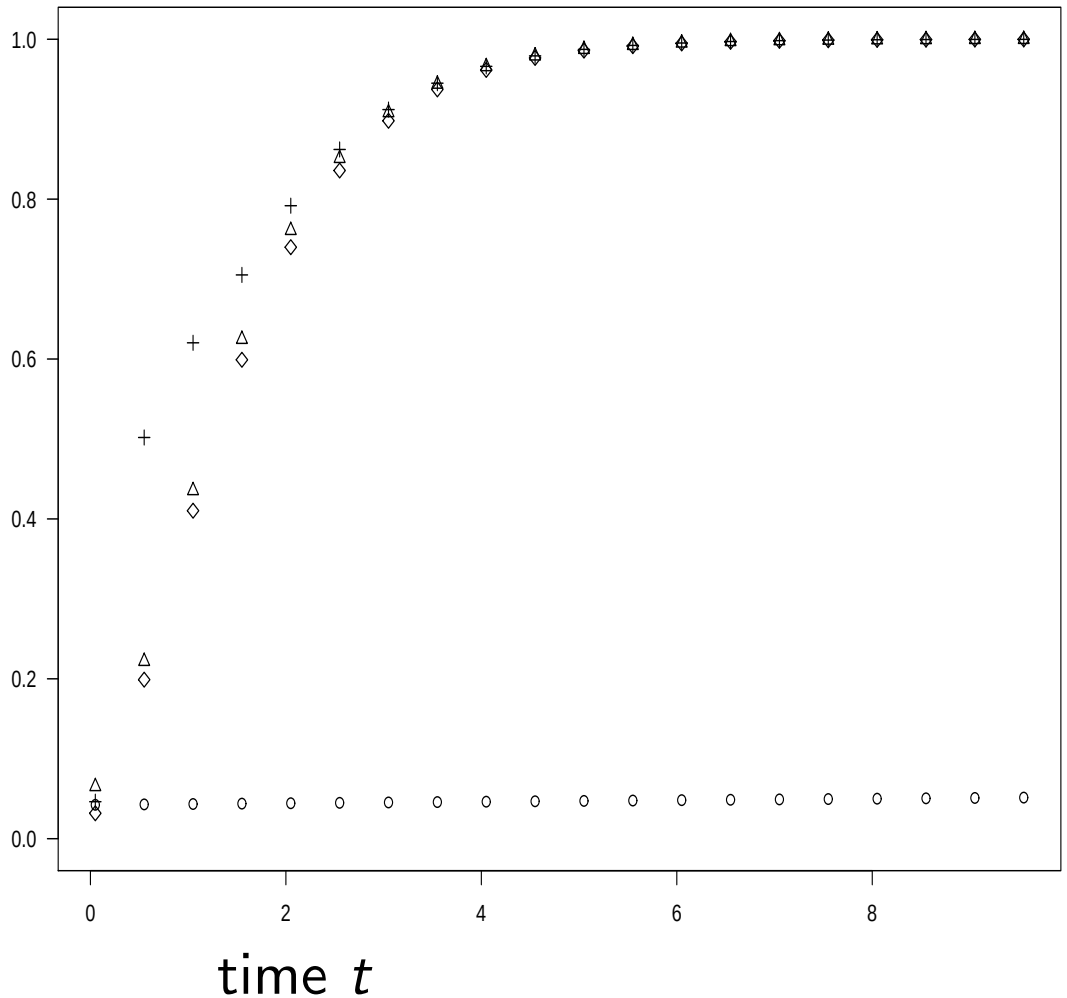


$\circ : Beta(2 - a, a)$



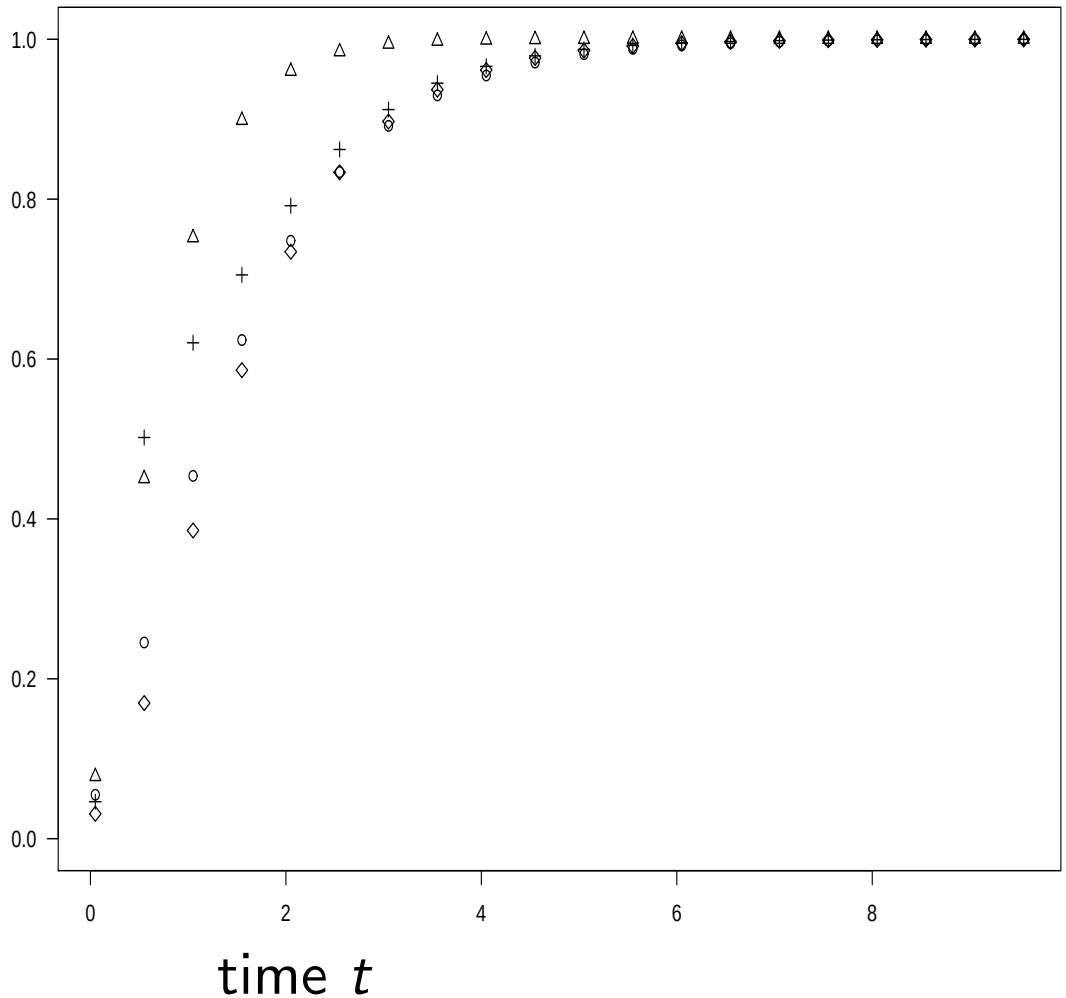
Two species and three lines each

$\circ : \Lambda_{0.05}; \quad \triangle : K + \Lambda_{0.05}; \quad \diamond : Beta(0.95, 1.05)$



Two species and three lines each

$\circ : \Lambda_{0.95}; \quad \triangle : K + \Lambda_{0.95}; \quad \diamond : Beta(0.05, 1.95)$



Recursive approach for s species

Let $\tilde{n} = n_1 + \cdots + n_s$ in which n_i denotes the number of ancestral lines for species i in a population; and let $n = (n_1, \ldots, n_s)$

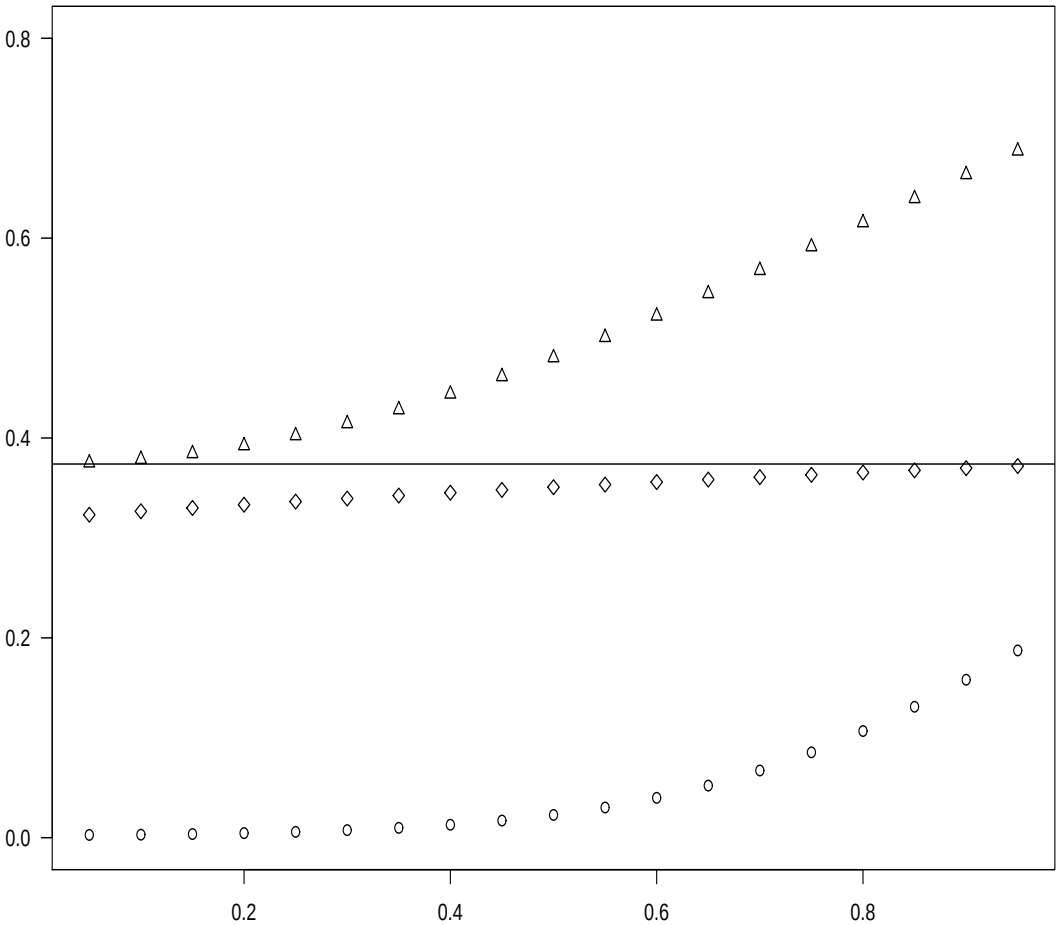
$$\mathbb{P}[MC; n] = \sum_{k=2}^{\tilde{n}} \beta_{\tilde{n},k} \sum_{r=1}^{s} \mathbb{P}[MC; m] \binom{n_r}{k} / \binom{\tilde{n}}{k}$$

in which $m = (n_1, n_2, \ldots, n_{r-1}, n_r - k + 1, n_{r+1}, \ldots, n_s)$ and

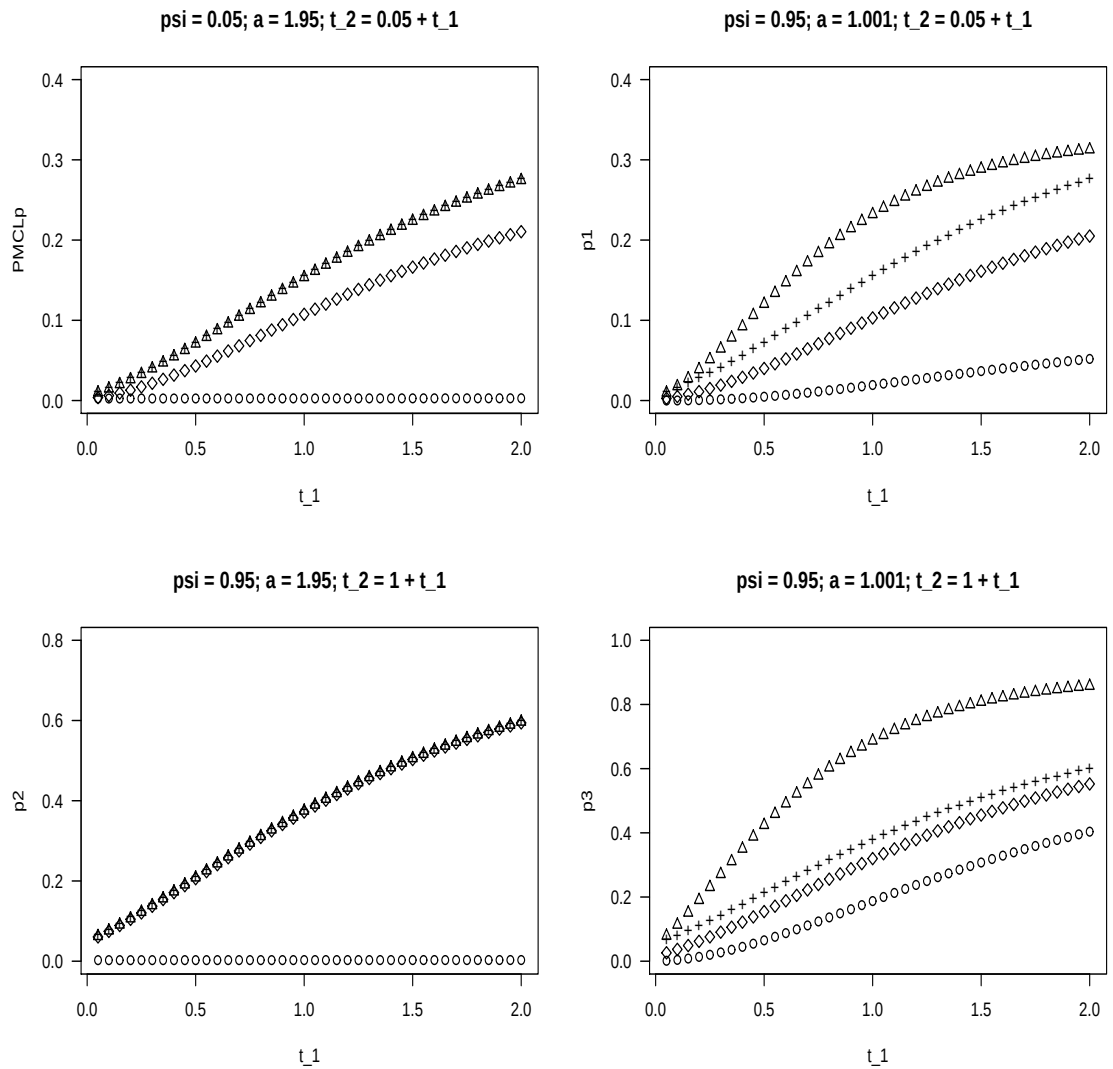
$$\mathbb{P}[MC; (0, 0, \ldots, 0, 1)] = \mathbb{P}[MC; (0, 0, \ldots, 0, 1, 1)] = 1$$

Three species and two lines each ($t_1 = 1, t_2 = 2$)

$\circ : \Lambda_{0.05}; \quad \triangle : K + \Lambda_{0.05}; \quad \diamond : Beta(0.95, 1.05)$



Three species and two lines each ($+ : K$)



$\circ : \Lambda_{0.05}$; $\triangle : K + \Lambda_{0.05}$; $\diamond : Beta(0.95, 1.05)$

Conclusions

- Probability of monophyletic concordance depends on parameters of multiple merger coalescent processes
- Presence of multiple mergers complicates computations
- Scaling time appropriately is important

Acknowledgments

EPSRC and Marsden Fund for funding