Supplementary material

Likelihoods. A cladogram with n tips contains the set of splits

$$(m, i) = (\text{ size of the parent clade, size of the smaller daughter clade})$$

defined at each internal node (there are $n-1$ such nodes). Under the beta splitting model, the probability of a tree is

$$\mathrm{P}(\text{tree}; \beta) \propto \prod_{\text{internal node}} p(i|m; \beta)$$

with $p(i|m; \beta)$ given by Equation 4 in the text. Maximum likelihood estimation of β was performed using a combination of golden section search and successive parabolic interpolation as implemented in the R function `optimize`. See also (Brent, 1973. Algorithms for Minimization without Derivatives. Englewood Cliffs N.J.: Prentice-Hall).

The Beta-binomial (BB) model. The biased-speciation model (Kirkpatrick and Slatkin, 1993) assumes that the two sister branches that result from a speciation event differ in their further speciation rates. In our model one sister branch inherits the rate $p\lambda$ and the other inherits the rate $(1-p)\lambda$, where λ is the ancestral rate. In addition the probability coefficient p is random, and is sampled at each speciation event from the Beta distribution $\mathrm{Beta}(\alpha+1, \alpha+1)$, $\alpha > -1$.

The BB split distribution. The biased-speciation process starts with two branches and $\lambda = 1$. Assume that the left branch has speciation rate $(1-p)$ and the right branch has speciation rate p. For $n = 4$ taxa, the three possible splits at the root are $(1, 3)$, $(2, 2)$, $(3, 1)$. It is routine to check that they are given the probabilities p^2, $p(1-p)+(1-p)p$, $(1-p)^2$, i.e. a bin(2,p) distribution. With n species, we obtain that the number of species in the left sister clade minus one is distributed as $\mathrm{bin}(n-2, p)$.

Because the process may also start with the symmetric situation in which the left branch speciation rate is equal to p, the size of the left sister clade is actually given the same distribution as the random variable

$$L_n = 1 + \mathrm{bin}(n-2, p)X + \mathrm{bin}(n-2, 1-p)(1-X)$$

where X is an independent (coin flipping) Bernoulli random variable. This description of L_n is strictly equivalent to Equation 6 in the text. For i in $1,\ldots,n-1$, we obtain

$$\mathrm{P}(L_n = i) = \frac{1}{2}\left(\begin{array}{c} n-2 \\ i-1 \end{array}\right)\left(p^{i-1}(1-p)^{n-1-i} + (1-p)^{i-1}p^{n-1-i}\right).$$

Now p is also a random variable. Integrating over all p's leads to the split distribution of the Beta-binomial model

$$p_{\mathrm{BB}}(i|n) = \left(\begin{array}{c} n-2 \\ i-1 \end{array}\right)\int_0^1 p^{i-1}(1-p)^{n-2-(i-1)}\frac{p^\alpha(1-p)^\alpha}{\mathrm{B}(\alpha+1,\alpha+1)}dp$$

where $\mathrm{B}(a,b)$ denotes the Beta function (Abramowitz and Stegun, 1970). This leads to

$$p_{\mathrm{BB}}(i|n) = \frac{\mathrm{B}(i+\alpha, n-i+\alpha)}{(n-1)\mathrm{B}(\alpha+1,\alpha+1)\mathrm{B}(i,n-i)}, \quad 1 \le i \le n-1.$$

which is strictly equivalent to the Beta-binomial (BB) split distribution

$$p_{\mathrm{BB}}(i|n) = \frac{1}{b_n(\alpha)}\frac{\Gamma(i+\alpha)\Gamma(n-i+\alpha)}{\Gamma(i)\Gamma(n-i)}, \quad 1 \le i \le n-1,$$

Changing i into $i-1$, this distribution corresponds to the Beta-binomial distribution as it is called in the statistical literature (Johnson N. L., S. Kotz, and A. W. Klemp. 1993. Univariate Discrete Distributions. Wiley & Sons, New York.). For $\alpha = 0$, we have $p(i|n) = 1/(n-1)$ which corresponds to the ERM model. For $\alpha \to -1^+$, We have $p(i|n) \to 0$ for $2 \le i \le n-2$, and $p(i|n) \to 1/2$ for $i = 1$ or $i = n-1$. Hence the limiting process corresponds to the comb tree. However there is no value in the BB model that corresponds exactly to the PDA or the AB models.

Connection between the BB and Beta-splitting models. The Beta-splitting and BB model families both work on the basis of a binomial split distribution with random parameter p sampled from a beta density $f(p)$. For $\beta \in (-1,\infty)$ the Beta-splitting model has the Beta$(\beta+1,\beta+1)$ distribution. The AB model can be viewed as the limit case where $\beta = -1$. In the BB model, α plays a role similar to β.

The difference between the two model families comes from the conditional split distribution given the probability parameter p. Aldous' choice comes from the usual binomial bin(n,p) distribution. But this distribution has the

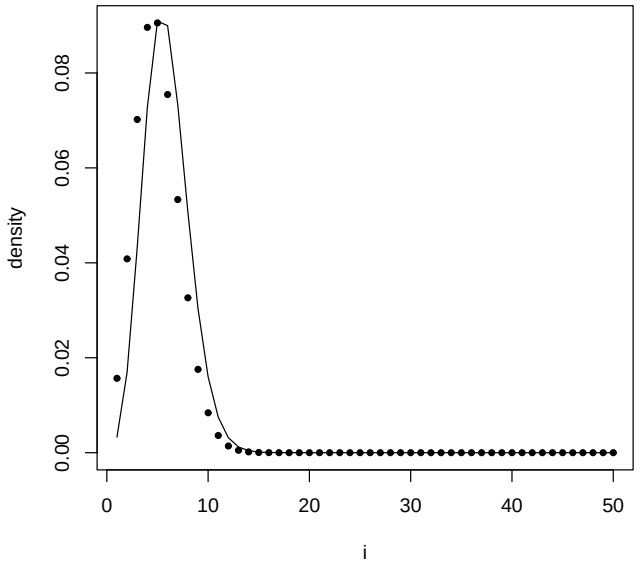
drawback of giving positive probabilities to 0 and n. To avoid this problem, sampling is iterated until these two extreme values are not produced. This leads to the conditional binomial distribution described in (Aldous, 1996, eq.(4))

$$p(i|n) = \left(\begin{array}{c} n \\ i \end{array} \right) \frac{p^i(1-p)^{n-i}}{1-p^n-(1-p)^n}, \quad 1 \leq i \leq n-1. \quad \text{(cond. BS)}$$

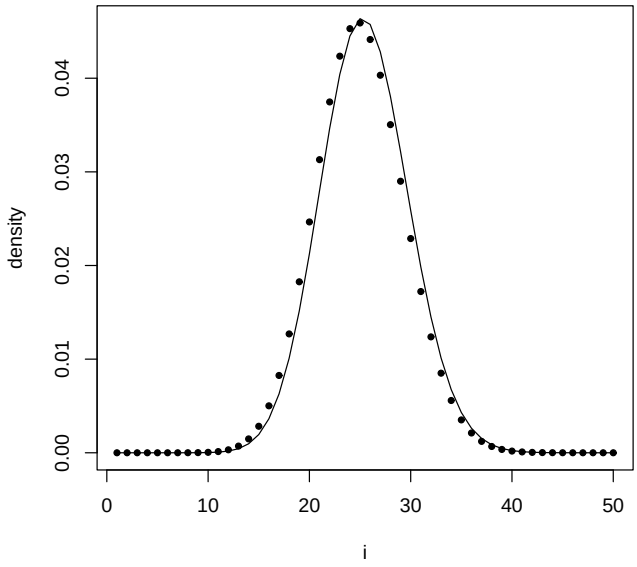
The BB model assumes a simpler strategy for transforming the binomial distribution into a distribution over the set $\{1, \ldots, n-1\}$

$$p(.|n) \equiv 1 + \text{bin}(n-2, p) \quad \text{(cond. BB)}$$

Taking label switching into account and averaging over p leads to the same formulas as in Equations 4 and 8 in the text. The connection between the BS and BB models can be seen through the distributions with fixed p's given in Equations (cond. BS) and (cond. BB). In fact the density curves give strong evidence that the trajectory of the BB model lies very close to the Beta-splitting in the space of probability models on $\{1, \ldots, n-1\}$ (see Fig. SM1 below). In addition for fixed numbers of taxa, there are optimal α's which minimize the distance between the BB and the AB model ($\beta = -1$). The optimal α depends on n as shown in Fig. SM2 where we used the mutual information distance between probability distributions (the same was obtained with the χ^2 and the total variation distances). The minimum distance parameters provide accurate approximations for the AB model split distribution for each n (see the result for $n = 100$ in Fig. SM2B). However finding α values that provide the best approximation for tree distributions (and not just split distributions with fixed number of taxa) requires solving a difficult combinatorial problem, because distances are now between probability measures in tree space. We just noticed that for values around $\alpha \approx -0.58$, shape statistics computed in the BB model showed non-significant departures from the AB model (Fig. SM3).

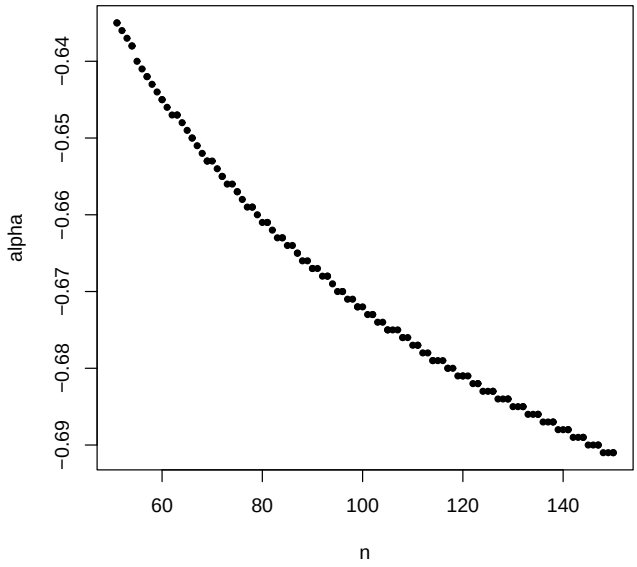


(A)

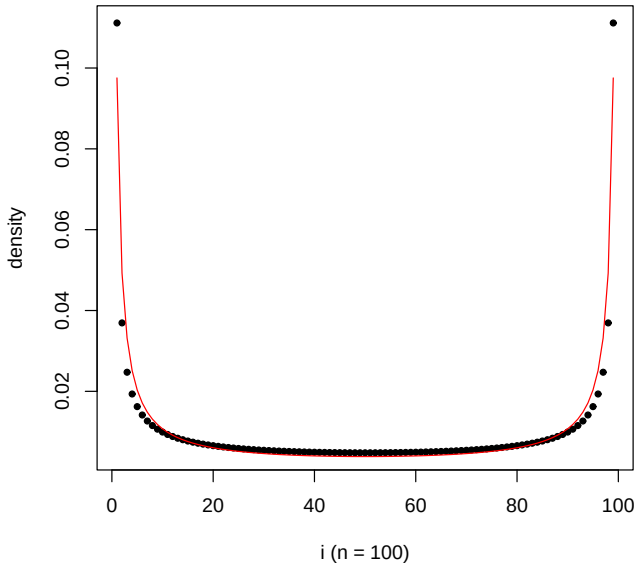


(B)

FIGURE SM1 Conditional distributions for the BB (dots) and Beta splitting (solid lines) for $n = 100$ given (A) $p = 0.05$ and (B) $p = 0.25$. These curve show that the two distributions are very close to each other although the match is not absolute (half of the distribution shown).



(A)



(B)

FIGURE SM2 (A) Values of α that provide the best approximation of the AB model by the BB model (mutual information distance). (B) $n = 100$. Split distribution for the AB model (solid line) and the BB model at $\alpha = -0.67$ (dots).

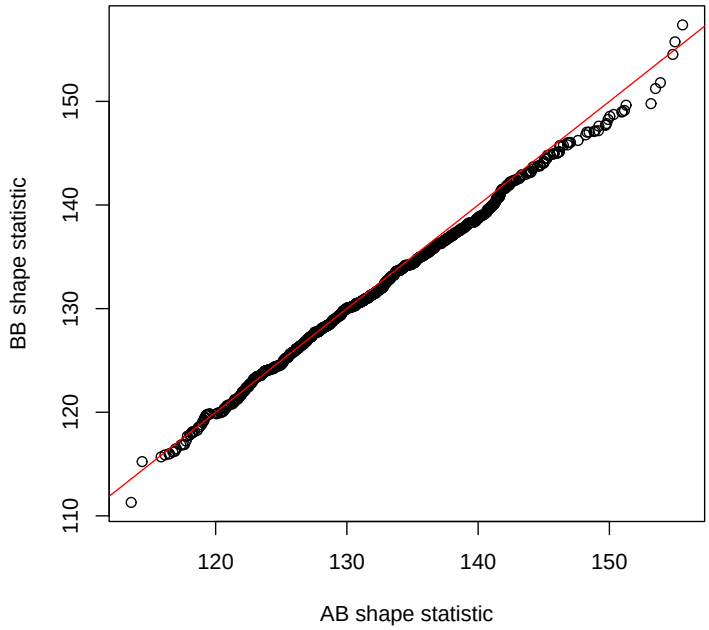
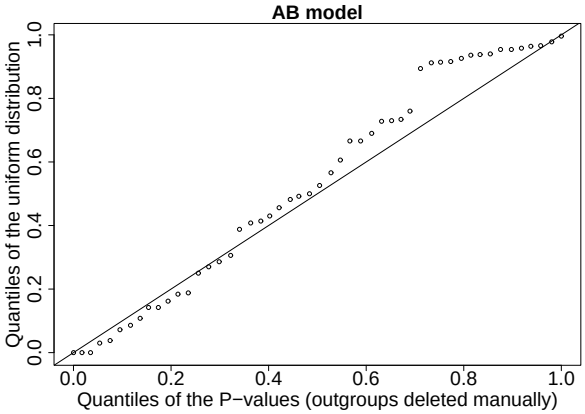
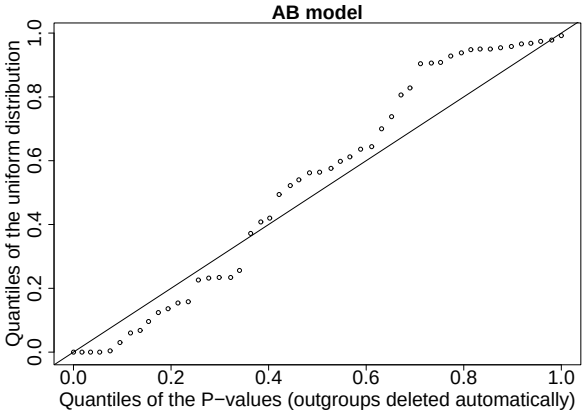


FIGURE SM3 Evidence for similar distributions of tree shape statistics with n tips under the AB and BB models. A quantile-quantile plot for $\alpha = -0.58$ and $n = 100$.

(A)



(B)



(C)

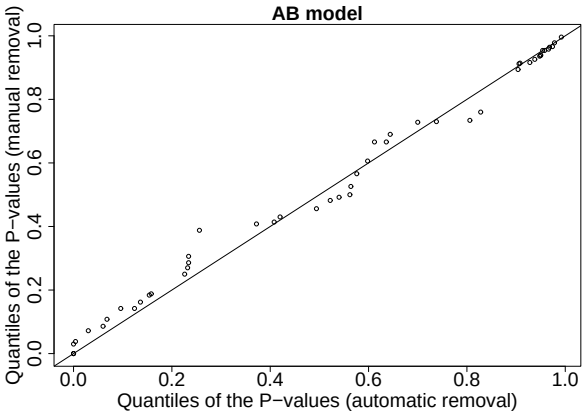


FIGURE SM4 Bias of the automatic outgroup removal method assessed from 50 arbitrary trees from TreeBASE. (A) Goodness-of-fit of the AB model after removing the true outgroups. (B) Goodness-of-fit of the AB model after automatic removal. (C) Comparison of shape statistics for the manually and automatically removed outgroup species. The TreeBASE entries of the 50 used trees are available upon request to the authors.