LETTERS

Phylogenies reveal new interpretation of speciation and the Red Queen

Chris Venditti[1], Andrew Meade[1] & Mark Pagel[1,2]

The Red Queen[1] describes a view of nature in which species continually evolve but do not become better adapted. It is one of the more distinctive metaphors of evolutionary biology, but no test of its claim that speciation occurs at a constant rate[2] has ever been made against competing models that can predict virtually identical outcomes, nor has any mechanism been proposed that could cause the constant-rate phenomenon. Here we use 101 phylogenies of animal, plant and fungal taxa to test the constant-rate claim against four competing models. Phylogenetic branch lengths record the amount of time or evolutionary change between successive events of speciation. The models predict the distribution of these lengths by specifying how factors combine to bring about speciation, or by describing how rates of speciation vary throughout a tree. We find that the hypotheses that speciation follows the accumulation of many small events that act either multiplicatively or additively found support in 8% and none of the trees, respectively. A further 8% of trees hinted that the probability of speciation changes according to the amount of divergence from the ancestral species, and 6% suggested speciation rates vary among taxa. By comparison, 78% of the trees fit the simplest model in which new species emerge from single events, each rare but individually sufficient to cause speciation. This model predicts a constant rate of speciation, and provides a new interpretation of the Red Queen: the metaphor of species losing a race against a deteriorating environment is replaced by a view linking speciation to rare stochastic events that cause reproductive isolation. Attempts to understand species-radiations[3] or why some groups have more or fewer species should look to the size of the catalogue of potential causes of speciation shared by a group of closely related organisms rather than to how those causes combine.

Van Valen's original observations in support of the Red Queen were of the length of time a species persisted in the fossil record[1], and yielded the claim that individual species went extinct at a constant rate through time—longer lived species do not become better adapted. Van Valen suggested that extinction occurs when species lose a battle against a biotic environment that is constantly changing. Subsequent work showed that this model predicts not only a constant rate of extinction, but also a constant rate of speciation and evolution in a homogeneous group[2,4]. Van Valen did not test the constant-rate model against competing models, and apart from a suggestion that in some micro-fossil groups the probability of speciation decreases with age while the probability of extinction increases[5], nor have those who followed him. A burst of studies beginning in the 1990s, that estimated speciation and extinction rates from phylogenies[6], derived their expectations from the constant-rate model that underlies Red Queen predictions[7].

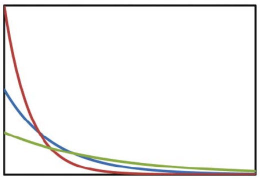
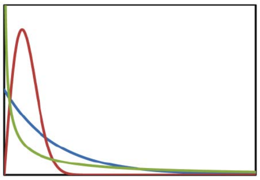
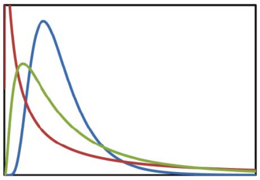
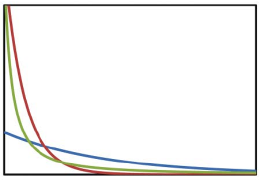
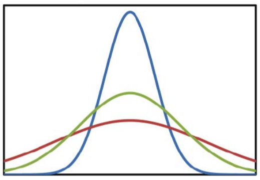
In addition to being used to study rates of speciation and extinction, the lengths of the branches of phylogenetic trees can reveal how the various factors that bring about speciation combine to do so, but we are not aware of any studies that have used phylogenies for this purpose. We studied the frequency distributions of these branch lengths in 101 phylogenies inferred from gene-sequence data, and selected for including a well-characterized and narrow taxonomic range of species. This reduces background differences in life histories, morphology and ecology that might affect rates of speciation. Our data sets include bumblebees, cats, turtles and roses (Supplementary Information). For each of the gene-sequence alignments, we inferred a Bayesian posterior probability sample of 750 phylogenetic trees using our phylogenetic mixture model[8] (Supplementary Information). We used uniform (0–10) priors on branch lengths to avoid biasing towards short or long branches, although exponential priors gave the same results. The mixture model improves on conventional single-rate-matrix models and on partitioned models, more accurately recovers branch lengths and reduces artefacts of phylogeny reconstruction[8,9]. Accurate reconstruction of branch lengths is crucial, as, for example, systematically underestimating the true lengths of long branches would bias the branch-length distribution away from long-tailed distributions. We excluded any data sets in which the inferred trees suffered from node-density artefacts[10,11].

We characterized the frequency distributions of the phylogenetic branches using statistical models that make differing assumptions about the expected amount of divergence or waiting times between successive speciation events. We suppose there are many potential causes of speciation, including environmental and behavioural changes, purely physical factors such as the uplifting of a mountain range that divides two populations, or genetic and genomic changes. If many independent factors combine additively to produce a speciation event, the distribution of branch lengths will conform to a normal probability density; if they combine multiplicatively, a log-normal density of lengths will arise. Suppose the factors are rare but large in number, where 'rare' means occurring at a rate less than the rate of speciation. Then their distribution over long periods spanning many speciation events will follow a Poisson density[12]. If these factors have the potential on their own to cause a speciation, the branch length distribution will follow an exponential density[12], that being the waiting time between successive events of a Poisson process. This is also the density that arises if there is a constant probability of speciation. A variant of the exponential model allows the multiple rare factors to affect species differently such that they have different constant rates[13] (hereafter the variable-rates model), as might be expected of a species radiation[3]. Another variation of the exponential—the Weibull density—can accommodate the probability of speciation changing according to the amount of divergence from the ancestral species. This model will fit the data if, for example, species are either more or less likely to speciate the older they get.

Table 1 (see also Supplementary Information) shows that with the exception of the normal, these statistical models can produce almost indistinguishable densities, but imply different modes of causation. For

[1]School of Biological Sciences, University of Reading, Reading, Berkshire, RG6 6BX, UK. [2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

Table 1 | Five biologically interpretable probability models of the distribution of branch lengths on a phylogeny

Model	Parameters	PDF
Exponential $p(x)=\frac{1}{\beta}e^{-x/\beta}$	β = scale	
Weibull $p(x)=\frac{\alpha}{\beta}x^{\alpha-1}e^{-(x/\beta)^{\alpha}}$	α = shape; β = scale: when $\alpha = 1$, Weibull = exponential	
Lognormal $p(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-[\ln x-\mu]^2/2\sigma^2}$	μ = mean of ln x; σ^2 = variance	
Variable rates $p(x)=\frac{\alpha\beta}{(1+\beta x)^{1+\alpha}}$	α = shape; β = scale (of the gamma distribution of the rates)	
Normal $p(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-[(x-\mu)/\sigma]^2/2}$	μ = mean; σ^2 = variance	

Scale parameters control the mean, shape parameters describe skew and mode. Four of the five models can produce similar shapes but have different biological interpretations (see text). Different coloured distributions correspond to different parameter values in each model. Colours distinguish shapes but do not correspond across models to equivalent parameter combinations. Models and parameters are further described in Supplementary Information. PDF, probability density function.

example, the Weibull density simplifies to the exponential density as a special case when $\alpha = 1$ and $\beta = 1$ (see Table 1 for definitions of these parameters). The variable-rates model is the convolution of multiple exponentials whose individual rates are assumed to follow a gamma probability distribution—in this model α and β describe the shape and scale, respectively[13]. If this gamma distribution is very narrow (small β), then the variable-rates model converges on an exponential.

For each posterior sample of 750 trees we normalized the branch lengths in each tree to be 1.0 (this does not affect the shape of the distribution), then studied the probability distribution of branch lengths by comparing the fits of the five models in Table 1. The length of each internal branch in a phylogeny records the expected amount of genetic divergence per site inferred to have occurred between two adjacent nodes of the tree, corresponding to events of inferred speciation minus any species missing from the tree owing to extinction or sampling[11]. We excluded the terminal branches because they do not record speciation events and so are not suitable for characterizing waiting intervals. We used genetic branch lengths in preference to branch lengths scaled to time because all temporal-scaling methods introduce uncertain nonlinear transformations (see Supplementary Information). Our preference is equivalent to assuming that any departures from a molecular clock in our data sets are random with respect to the underlying branch lengths.

Figure 1a reports the percentage of the data sets in which each of the five models provided the best overall description of the branch-length distributions. Each model's fit was evaluated using a reversible-jump Markov chain Monte Carlo (MCMC) criterion and by harmonic means (see Fig. 1 legend and Supplementary Information). For the reversible-jump approach this is the model with the highest posterior proportion in a data set and for the harmonic mean method we take the model with the largest score. The two methods are in close agreement. The reversible-jump method finds the exponential model provides the best fit in 78 ± 4.1% of the data sets, followed by the Weibull at 8 ± 2.7%, the lognormal at 8 ± 2.7% and the variable-rates model at 6 ± 2.4%. The normal distribution never provided the best fit. Despite fitting fewer parameters, only the exponential model's performance exceeds the prior expectation of fitting at least 20% of the trees. The harmonic mean approach returned similar results for the exponential (80 ± 4.0%) but gave more support to the variable-rates model (14 ± 3.4%), in preference to the lognormal (1 ± 1.2%), followed by the Weibull in 6 ± 2.4%, and again the normal distribution did not fit any of the data sets best. The lognormal's poorer performance under the harmonic mean largely reflects the disproportionate effects of a few trees in each of a handful of posterior samples that it fitted badly: the harmonic mean is known to be sensitive to extreme values and as such can be unstable[14,15]. For this reason we favour the reversible-jump approach and refer to its results for the remainder of the analyses, although none of the tests reported below differs qualitatively when the harmonic mean values are used.

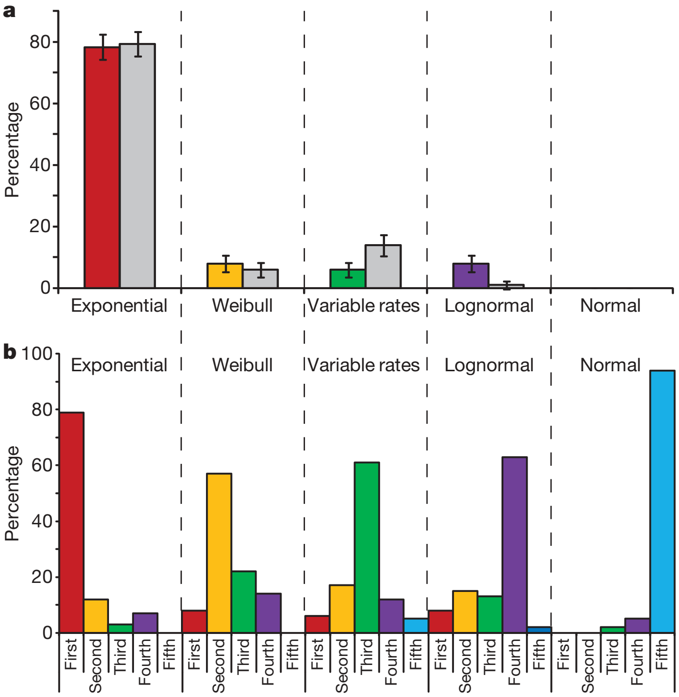


Figure 1 | Performance of the five models. We fitted the statistical models in two ways. In one, we implemented a Bayesian reversible-jump[27] Markov chain that moved among the five statistical models estimating their parameters while simultaneously jumping among trees in the posterior sample (Supplementary Information). Allowed to run for many iterations, the proportion of time the chain spends in each model measures its posterior probability of describing those data. In the second method, we fitted each model separately in its own Markov chain that estimated the parameters of the statistical model while moving among trees, recording the harmonic mean of the likelihoods (based on samples from chains that were run for 1.875×10^9 iterations) as an estimate of each model's marginal density[27]. In both approaches, we used Bayesian prior distributions chosen to favour the four two-parameter models over the one-parameter exponential (Supplementary Information). **a**, Percentage of data sets for which each model provided the best overall description of the branch-length distribution (models described in text). The coloured bars are the results from the reversible-jump procedure (see text), the grey bars record the results from the harmonic mean test. Error bars, standard error. **b**, The distribution of each model's finishing places (first to fifth) across the 101 data sets.

The two models that can most closely mimic the exponential—the Weibull and the variable-rates model—regularly come in second and third when each model's distribution of finishing places across the 101 data sets is recorded (Fig. 1b). These show that even though the lognormal model was in joint second place for the number of best fits, it does not in general provide a competitive description of the data, suggesting that the data sets it does fit best are unusual. When the Weibull scaling parameter equals 1 (see Table 1), this distribution simplifies to the exponential. For the $n = 22$ data sets in which the exponential was not the best fitting model, the Weibull scaling parameter is close to 1 (1.09 ± 0.25), indicating that there is no trend in those data sets either for longer-lived species to have a lower rate of speciation ($\alpha > 1$) or for short-lived species to have a higher rate ($\alpha < 1$).

We find no relationship between the size of the tree and the degree of posterior support for the exponential model ($r^2 = 0.03$, $P = 0.5327$). Neither did we find any association between Colless' index, I_c, of tree imbalance[16] and the degree of posterior support for the exponential model ($r^2 = 0.006$, $P = 0.4368$), suggesting that the popular Yule process[17] may not in general explain the shape and branch-length distributions in real trees. The posterior probability of a sample of trees being fitted best by the exponential model is not different in the subset of our data sets that sample greater than 50% of the extant clade ($F = 1.1713$, $P = 0.3143$). Similarly, there was no significant difference between the trees that span different taxonomic ranges; within genus, within family or within order ($F = 1.7642$, $P = 0.1767$), or between kingdom (plants, animals or fungi; $F = 0.8142$, $P = 0.4464$).

The observed cumulative density distributions of the branch lengths for trees best fitted by the exponential, the lognormal and the variable-rates models, respectively, are shown in Fig. 2a, b and c, along with the exact cumulative density of the statistical model that provides the best fit to those data. These show that data sets do differ in which model fits them best, and that simple models are able to characterize the branch-length distributions with precision. The steep rise in Fig. 2c is consistent with the suggestion that in some instances, rates of speciation are high initially and then taper off[3], but as we find this pattern in just 6% of the trees, it does not appear to be a general phenomenon in the growth of a phylogeny. Figure 2d plots the cumulative exponential density function, and the average of the cumulative densities for the 79 data sets best fitted by the exponential model, and separately for the 22 data sets fitted best by the other models. The data sets that are best described by the exponential provide a remarkably close fit to the true exponential density, showing that the simplest model of speciation describes the individual speciation rates of closely related species over millions of years.

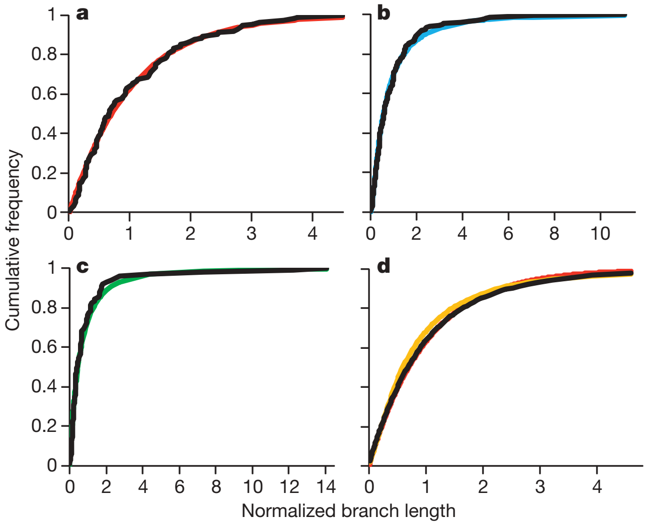


Figure 2 | Cumulative density distributions of branch lengths. Each panel plots the proportion of observed branch lengths expected to fall at or below a point on the *x* axis if the model is true. Coloured curves are the statistical cumulative distribution functions (CDFs). **a**, Cumulative density of branch lengths for a phylogeny of pit vipers[28] taken from the posterior sample (black) and the predicted exponential CDF (red); **b**, cumulative density of branch lengths for a phylogeny taken from the posterior sample of bumblebees[29] (black) and the predicted lognormal CDF (blue); **c**, cumulative density of branch lengths for a phylogeny taken from the posterior sample of *Symplocos* (a genus of flowering plants)[30] (black) and the expected variable-rates CDF (green). **d**, This plot compares the observed proportions for: (1) data sets that were best described by the exponential model (black) and (2) data sets best fitted by other models (yellow). The red line shows the predicted exponential density (this line is mostly masked by the black line).

Deviations from the molecular clock might be expected to produce excesses of very short or very long branches—corresponding to the clock speeding up or slowing down. We think that such deviations are unlikely, because the narrow taxonomic range of our species means they share life-history and metabolic factors that might influence the rate of evolution[18]. But even if they did occur they would tend to favour models other than the exponential, such as the variable rates. Missing species and extinctions could influence our results if not randomly distributed in the tree, but neither of these would bias the results towards the exponential model (Supplementary Information). We note that exponentially distributed branches in phylogenies provide support for the widespread use of exponential priors on branch lengths in MCMC models of phylogenetic inference.

Our results stringently test the idea that speciation occurs at a constant rate, and suggest a general mechanism by which shared rates of speciation will arise among independently evolving taxa. We derive the Red Queen prediction from a simple model that supposes that the causes of speciation are many and rare, not necessarily limited to biotic interactions, and each individually having the potential to cause a speciation event. It has been shown[12,19] in a different context that if these assumptions hold, an exponential distribution of divergences is expected between—in our case—successive bouts of speciation (Fig. 3). If the original Red Queen model had a 'whiff' of a species running out of breath from the accumulation of many detrimental biotic effects, and then being 'knocked off' by the next event, the interpretation we propose is different. Species do not so much 'run in place' as simply wait for the next sufficient cause of speciation to occur. Speciation is freed from the gradual tug of natural selection, there need not be an 'arms race' between the species and its environment, nor even any biotic effects. To the extent that this view is correct, the gradual genetic and other changes that normally accompany speciation[20] may often be consequential to the event that promotes the reproductive isolation, rather than causal themselves. Factors apart from biotic interactions that can cause speciation include polyploidy[21], altered sex determination mechanisms[22], chromosomal rearrangements[23], accumulation of genetic incompatibilities[24], sensory drive[25], hybridization[26] and the many physical factors included in the metaphor of mountain range uplift.

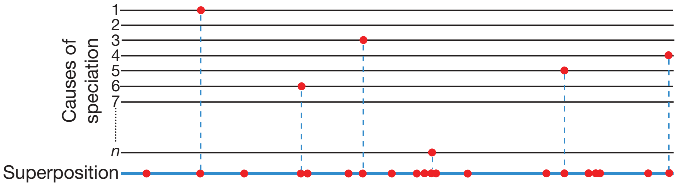


Figure 3 | Single rare-events model. We suppose there are many rare factors (n) each individually having the potential to cause a speciation event *per se* (red points indicate a speciation event). If the events are sufficiently rare and large in number (rare being defined as much less than the rate of speciation), the superposition of these events (bold blue horizontal line with red points) gives rise to a Poisson process over many speciation events, with the waiting time between successive events of the Poisson being exponentially distributed. The speciation events (red points) shown on the superposition line that do not connect to a red point above are presumed to be among the '$n-7$' causes not shown (illustration drawn after ref. 12).

This way of thinking about speciation also has implications for attempts to understand why some groups have so many more or fewer species than others. If speciation is driven by rare stochastic events, then it will be the number of different such events that sets the rate of speciation. This means that researchers seeking to develop explanatory theories of speciation should focus their attention on the size of the catalogue of sufficient causes (speciation factors) shared by a group of organisms, rather than on special driving forces or how these forces might combine.

Received 18 August; accepted 29 October 2009.
Published online 9 December 2009.

1. Van Valen, L. A new evolutionary law. *Evol. Theory* **1,** 1–30 (1973).
2. Stenseth, N. C. & Maynard Smith, J. Coevolution in ecosystems: Red Queen evolution or stasis? *Evolution* **38,** 870–880 (1984).
3. Rabosky, D. L. & Lovette, I. J. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution* **62,** 1866–1875 (2008).
4. Benton, M. J. in *Palaeobiology: A Synthesis* (eds Briggs, D. E. G. & Crowther, P. R.) 119–124 (Blackwells, 1990).
5. Pearson, P. N. Survivorship analysis of fossil taxa when real-time extinction rates vary: the Paleogene planktonic foraminifera. *Paleobiology* **18,** 115–131 (1992).
6. Nee, S., Mooers, A. O. & Harvey, P. H. Tempo and mode of evolution revealed from molecular phylogenies. *Proc. Natl Acad. Sci. USA* **89,** 8322–8326 (1992).
7. Nee, S. Birth-death models in macroevolution. *Annu. Rev. Ecol. Evol. Syst.* **37,** 1–17 (2006).
8. Pagel, M. & Meade, A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* **53,** 571–581 (2004).
9. Venditti, C., Meade, A. & Pagel, M. Phylogenetic mixture models can reduce node-density artifacts. *Syst. Biol.* **57,** 286–293 (2008).
10. Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Syst. Biol.* **55,** 637–643 (2006).
11. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314,** 119–121 (2006).
12. Gillespie, D. J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, 1991).
13. Pagel, M., Meade, A. & Scott, D. Assembly rules for protein networks derived from phylogenetic-statistical analysis of whole genomes. *BMC Evol. Biol.* **7,** S16 (2007).
14. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167,** 808–825 (2006).
15. Raftery, A. E. in *Markov Chain Monte Carlo in Practice* (eds Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.) 145–161 (Chapman & Hall, 1996).
16. Colless, D. H. Phylogenetics: the theory and practice of phylogenetic systematics II (book review). *Syst. Zool.* **31,** 100–104 (1982).
17. Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Phil. Trans. R. Soc. Lond. B.* **213,** 21–87 (1925).
18. Smith, S. A. & Donoghue, M. J. Rates of molecular evolution are linked to life history in flowering plants. *Science* **322,** 86–89 (2008).
19. Gillespie, J. H. The molecular clock may be an episodic clock. *Proc. Natl Acad. Sci. USA* **81,** 8009–8013 (1984).
20. Coyne, J. A. & Orr, H. A. *Speciation* 411–442 (Sinauer Associates, 2004).
21. Soltis, D. E., Soltis, P. S. & Tate, J. A. Advances in the study of polyploidy since plant speciation. *New Phytol.* **161,** 173–191 (2003).
22. Mitsainas, G. P., Rovatsos, M. T. H., Rizou, E. I. & Giagia-Athanasopoulou, E. Sex chromosome variability outlines the pathway to the chromosomal evolution in *Microtus thomasi* (Rodentia, Arvicolinae). *Biol. J. Linn. Soc.* **96,** 685–695 (2009).
23. Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence — accelerated evolution in rearranged chromosomes. *Science* **300,** 321–324 (2003).
24. Orr, H. A. & Turelli, M. The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. *Evolution* **55,** 1085–1094 (2001).
25. Seehausen, O. *et al.* Speciation through sensory drive in cichlid fish. *Nature* **455,** 620–626 (2008).
26. Mallet, J. Hybrid speciation. *Nature* **446,** 279–283 (2007).
27. Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biome* **82,** 711–732 (1995).
28. Castoe, T. A. & Parkinson, C. L. Bayesian mixed models and the phylogeny of pitvipers (Viperidae: Serpentes). *Mol. Phylogenet. Evol.* **39,** 91–110 (2006).
29. Cameron, S. A., Hines, H. M. & Williams, P. H. A comprehensive phylogeny of the bumble bees (*Bombus*). *Biol. J. Linn. Soc.* **91,** 161–188 (2007).
30. Fritsch, P. W., Cruz, B. C., Almeda, F., Wang, Y. & Shi, S. Phylogeny of Symplocos based on DNA sequences of the chloroplast trnC-trnD intergenic region. *Syst. Bot.* **31,** 181–192 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Centre for Advanced Computing and Emerging Technologies (ACET) at the University of Reading for the use of the ThamesBlue supercomputer. We thank M. Turelli for calling our attention to Gillespie's Poisson process model and for comments on earlier drafts of the paper. M. Steel pointed out the geometric distribution proof given in the Supplementary Information. This research was supported by grants to M.P. from the Natural Environment Research Council (NERC), UK, and the Leverhulme Trust.

Author Contributions C.V., A.M. and M.P. contributed to all aspects of this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.P. (m.pagel@reading.ac.uk).