19 November 2015
Target: 2000-3000 words
Word count: **3052**

AWF Edwards and the origin of Bayesian phylogenetics

Ziheng Yang

Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

Summary

In the early 1960s, Anthony Edwards and Luca Cavalli-Sforza made an effort to apply R.A. Fisher's maximum likelihood (ML) method to estimate genealogical trees of human populations using gene frequency data. They used the Yule branching process to describe the probabilities of the trees and branching times and the Brownian motion process to model the drift of gene frequencies (after a suitable transformation) over time along the branches. They experienced considerable difficulties, including "singularities" in the likelihood surface, mainly because a distinction between parameters and random variables was not clearly made. In the process they invented the distance (additive-tree) and parsimony (minimum-evolution) methods, both of which they viewed as heuristic approximations to ML. The statistical nature of the inference problem was not clarified until Edwards [1], which pointed out that the trees should be estimated from their conditional distribution given the genetic data, rather than from the "likelihood function". In modern terminology, this is the Bayesian approach to phylogeny estimation: the Yule process specifies a prior on trees, while the conditional distribution of the trees given the data is the posterior. This article discusses the connections of the remarkable paper of Edwards [1] to modern Bayesian phylogenetics, and briefly comments on some modelling decisions Edwards made then that still concern us today in modern Bayesian phylogenetics. The reader I have in mind is familiar with modern phylogenetic methods but may not have read Edwards [1], which is published in a statistics journal.

The model and the statistical problem of phylogeny estimation. The data considered by Edwards and Cavalli-Sforza [2,3] consist of gene frequencies of common blood groups from different human populations. Edwards treated different human populations while I focus on different species here. The data-generating model consists of two components. A Yule branching process, with a constant per-lineage rate λ of splitting, is used to describe the probability distribution, $f(F, \boldsymbol{t}|\lambda, n)$, of the phylogeny ($F$) and the branching times: $\boldsymbol{t} = \{t_1 = 1, t_2, \ldots, t_{n-1}\}$ (fig. 1). The Yule process assigns uniform probabilities to the labelled histories. The term labelled history, due to Edwards [1], refers to a rooted tree topology with internal nodes ordered by time. For example, the rooted tree topology ((*a*,

b), (*c*, *d*)) corresponds to two distinct labelled histories depending on whether the age of the *a*-*b* ancestor is older or younger than the age of the *c*-*d* ancestor. We note that other models of cladogenesis, such as the coalescent process [4] and the constant-rate birth-death process [5], all generate labelled histories with equal probabilities. In using the Yule process to describe the process of species formation, Edwards fixed the first branching event (the root of the tree) at time $t_1 = 1$, and conditioned on the number of species at the present time to be n. He derived the joint density for the tree form (F) and branching times ($\boldsymbol{t}$) as

$$f(F, \boldsymbol{t} | \lambda) = \frac{2^{n-1}\lambda^{n-1}\exp\{-\lambda\sum_{i=2}^{n-1} t_i\}}{(n-1)n!(1-\mathrm{e}^{-\lambda})^{n-2}}. \quad (1)$$

The second component of the model is the Brownian motion process used to describe the evolution of the continuous characters over time. Note that such Brownian models are now widely used in phylogenetic and phylogeographic analyses of morphological measurements from different species [6,7]. The Brownian motion or random walk in one dimension gives the location of the particle time t later, given that it is at location x_0 at time 0, as a Gaussian variable, $x_t \sim \mathrm{N}(x_0, t\sigma^2)$, where the parameter σ^2 controls how fast the particle drifts and represents the evolutionary rate for the continuous character. Gene frequencies at multiple blood group loci are treated as a p-dimensional Brownian motion. The different dimensions (different variables) are moving independently, with the same σ^2. However, the measurements for the same character observed in the modern species (which are the tips of the tree) are correlated because they may have shared some branches. For example, $\mathrm{cov}(x_{7k}, x_{9k}) = (t_1 - t_3)\sigma^2$, where $(t_1 - t_3)$ is the time shared by the two paths from the root to the tips 7 and 9 (fig. 1). Thus given σ^2, the tree form (labelled history F) and branching times ($\boldsymbol{t}$), and the state at the root (x_{1k}), each character k observed in a modern species is normally distributed with mean x_{1k} and variance $t_1\sigma^2$. The data or the measurements of the k characters among all modern species have the probability density

$$f(\boldsymbol{\xi}|F, \boldsymbol{t}, \sigma^2, \boldsymbol{x}_1) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{T}|^{p/2}}\exp\left\{\sum_{k=1}^{p} -\tfrac{1}{2}(\boldsymbol{\xi}_k - x_{1k}\mathbf{1})'\boldsymbol{T}^{-1}(\boldsymbol{\xi}_k - x_{1k}\mathbf{1})\right\}, \quad (2)$$

where $\mathbf{1}$ is a $p \times 1$ vector with all elements to be 1 and $\boldsymbol{T}$ is the variance-covariance matrix given by the tree. This is a n-variate normal density, and is nowadays known as the phylogenetic likelihood. This is equation (8) of Edwards.

The true parameters in the model are the Yule branching rate λ, the Brownian parameter σ^2, and the initial state $\boldsymbol{x}_1$. Given those it should be possible to simulate the process. The other unknowns, including the tree F, the branching times $\boldsymbol{t}$, and the character states at the interior nodes x_{ik}, $2 \le i \le n - 1$, $1 \le k \le p$, are random variables. If one simulates the process using the true parameters, those random variables will have different realized values among simulated replicates. Edwards [1] pointed out that the true parameters should be estimated by ML, with the likelihood calculated by integrating

over the random variables, that is, by summing over the trees (F) and integrating over the branching times ($\boldsymbol{t}$) as well as the ancestral states. The likelihood function is thus

$$L(\lambda, \sigma^2, \boldsymbol{x}_1) = f(\boldsymbol{\xi}|\lambda, \sigma^2, \boldsymbol{x}_1) = \sum_F \int_{\boldsymbol{t}} f(F, \boldsymbol{t} \mid \lambda)\, f(\boldsymbol{\xi} \mid F, \boldsymbol{t}, \sigma^2, \boldsymbol{x}_1)\, \mathrm{d}\boldsymbol{t} \,. \quad (3)$$

Here the notation is heuristic, and the sum is over all possible tree forms and the integral is over the (n – 2) branching times within each tree. The computation was deemed impossible and the model was not analyzed. Nevertheless, Edwards pointed out that random variables, F and $\boldsymbol{t}$, should be estimated from their conditional distribution given the data, with the true parameters replaced by their MLEs.

The singularity on the "likelihood" surface. Early attempts by Edwards and Cavalli-Sforza [2,8] treated the branching times ($\boldsymbol{t}$) and ancestral states ($\boldsymbol{x}$) as parameters. The "likelihood" function was defined as the product of the multivariate normal densities for character changes along the branches. For any branch $i \to j$, with branch length $t_i - t_j$, the density is

$$\left[2\pi(t_j - t_i)\sigma^2\right]^{-\frac{p}{2}} \cdot \exp\left\{-\frac{1}{2(t_j - t_i)\sigma^2} \sum_k (x_{jk} - x_{ik})^2\right\}, \quad (4)$$

given by the Brownian motion model. As pointed out by Edwards [1], this "likelihood" increases without bound if $x_{ik} = x_{jk}$ for all k and if $t_i \to t_j$: there are lines of singularity in the likelihood surface when there is no change along a branch for any of the p characters and when the branch disappears. Here the mistake was to treat the ancestral states and branching times as parameters, when they are in fact random variables with statistical distributions under the model.

Some remarks

Application of the Yule process. While the Yule process of pure birth is implemented in some Bayesian programs such as BEAST [9], it is more common to use the birth-death process, which is more general by allowing species extinctions and species sampling [10,11]. Often the distribution of times is obtained by conditioning on the number of lineages at the present time and on the age of the root, as in Edwards [1], but variations exist. For example, Thompson [12] suggests fixing $\sigma^2 = 1$, instead of fixing $t_1 = 1$, and she treats n as data (as n may be informative about the birth rate λ), rather than conditioning on n tips at the present time. Gomberg [13], in an unfinished report, prefers not conditioning on the present time, although the details are not so clear. The differences among those variants are not well understood.

Ancestral states at the root. The characters states at the interior nodes of the tree are known as ancestral states. On a species phylogeny, they represent the states of the characters in the extinct common ancestors of modern species. Edwards [1] and also Thompson [12, p.119] treated the states at the root ($\boldsymbol{x}_1$) as parameters. This has the problem of introducing many parameters in the model, such that the number of parameters increases without bound when the number of characters increases. For discrete characters such as the nucleotides (T, C, A, and G) in DNA sequences, it is customary to use continuous-time Markov chains to describe transitions between character states and to assume that the

process has been stationary for exceptions see [14]. Then the root states have a distribution given by the stationary distribution of the Markov chain. For continuous characters, the Brownian motion does not have a stationary distribution. Felsenstein [15] discussed an algorithm to estimate the ancestral states, eliminating $\boldsymbol{x}_1$. Statistical justifications for this procedure were discussed by Thompson [12, p.119].

The maximum likelihood method of phylogeny estimation. The Yule process component of the model was dropped when Felsenstein [15] revisited the problem of phylogeny reconstruction using continuous characters. Thus the tree (F) and times ($\boldsymbol{t}$) do not have statistical distributions anymore and become true parameters, which are estimated by maximizing the likelihood function, which averages over the ancestral states. Similarly, Thompson [12, p. 60], in extending the method of Edwards [1], dropped the Yule process. These are the early ML implementations of the Brownian-motion model for continuous characters.

Origin of Bayesian phylogenetics

Felsenstein [16, p.291] has included a discussion of early applications of Bayesian or Bayesian-like ideas to phylogenetics. Perhaps the most relevant is the calculation of posterior probabilities for trees by Kishino and Hasegawa [17,] see also [18], who calculated the likelihood by optimizing the branch lengths for each tree, while a fully Bayesian approach should average over the branch lengths or branching times. The modern approach to Bayesian inference of molecular phylogenies was introduced by three groups working independently in the 1990s: Bob Mau and Michael Newton in Wisconsin [19], a research student in Ohio State University, Shuying Li [20], and Bruce Rannala and me in Berkeley [10,21]. All those works integrate over the branching times through a prior to calculate the posterior probabilities of trees. The first two groups are statisticians, applying Bayesian MCMC algorithms to phylogeny estimation. Note that the 1990s was the time when Bayesian MCMC algorithms were introduced into various branches of sciences, even though they were developed a few decades earlier [22,23].

In our case, we owe the motivation entirely to Edwards [1]. After finishing my PhD in Beijing in 1992, I went to Cambridge to work with Adrian Friday and Nick Goldman. We occasionally saw Edwards, but I believe he was working on his book on Venn diagrams, rather than phylogenetics. Adrian, Nick and I were developing Markov models of sequence evolution for use in the ML method, and we had much discussion about whether the tree should be treated as a discrete parameter or a statistical model. I have provided detailed argument elsewhere that the distinction is not a semantic one see, e.g., [24,25, pp. 159-163]. For example, it is not so clear how to decide whether a log likelihood difference between two trees is due to chance. We have the rule of thumb [26, p.202] that an improvement of 2 log-likelihood units (or 1.92 if we use the asymptotic χ_1^2) is good enough for including one additional parameter, but we lack such a "calibration" when two trees are compared. The use of bootstrap to evaluate the significance of trees has met with difficulties in interpretation [27-29]. I was also concerned that the ML tree topology does not have the large-sample efficiency of the ML

estimate of a conventional parameter see, e.g., [24]. Those concerns motivated my work with Bruce, when both of us were postdocs in Monty Slatkin's phylogenetics laboratory in Berkeley. We were curious to see what the alternative statistical methodology, the Bayesian, might offer, given the difficulties with the ML. We decided to try Edwards's [1] prescription, but with two changes. First, we worked with DNA sequence data, using a continuous-time Markov chain model instead of the Brownian motion model for continuous characters, with summation over the ancestral states achieved using Felsenstein's [30,31] pruning algorithm. Second we used the birth-death process (instead of the Yule) to specify a prior on the trees and times but this was an easy replacement. We used numerical integration to integrate over the times, so that the method is applicable to small trees only. This effort led to Rannala and Yang [10].

This is one of the first Bayesian molecular phylogenetic analyses, and the results are interesting. We applied our program to two datasets of four or five ape species (human, common chimpanzee, pygmy chimpanzee, gorilla, and orangutan). We estimated the birth rate λ, death rate μ, and the transition/transversion rate ratio parameter κ by ML, and used those to calculate the posterior probabilities for trees, as stipulated by Edwards [1]. The maximum *a posteriori* (MAP) trees are reasonable in both datasets, grouping the human with the chimpanzees, but the posterior probability in one dataset, at 0.9999, is uncomfortably high. This dataset, of 11 mitochondrial tRNA genes (739bp) and published by Horai *et al.* [32], is fairly small, and the human-chimpanzee-gorilla relationship was a hard phylogenetic problem at the time. Spuriously high posterior probabilities for trees continue to trouble us today, especially as datasets for phylogenetic analysis are getting increasingly large [33-35]. The problem appears to have to do with the asymptotic behaviour of Bayesian model selection when applied to opposing and nearly equally wrong models.

Phylogeny estimation and statistical inference

I suppose my characterization of Edwards [1] as the first effort to apply Bayesian statistics (rather than ML) to phylogeny estimation might not be looked upon by Edwards himself as reasonable. Indeed in his Discussion, Edwards [1, p. 104] has this to say about the Bayesian approach:

> *and a detailed study of the present problem has, if anything, strengthened my conviction that a Bayesian approach in this instance would be a gross over-simplification. I am not prepared to give the true parameters prior probability distributions because I can see no model which would justify them. We may also note that the adoption of a Bayesian approach would not automatically resolve the dilemma of how to summarize a posterior distribution in a great many variables in terms of a few descriptive parameters, for maximizing the posterior probability would lead to the singularities, …*

A few words of clarification may thus be called for, related to the changing usage of the term "Bayesian" and the philosophy of statistical inference. The word "Bayesian" was apparently coined by R.A. Fisher in 1950 [36,37], to refer derogatorily to the method of *inverse probability*, the inference method that uses the Bayes theorem to derive probability distributions for parameters. The word

"inverse" refers to the fact that the probability is here defined backwards from the data to the parameter or hypothesis, or from effects to causes. Given the probability of heads for a fair coin, $\theta = \frac{1}{2}$, say, the probability of $x = 2$ heads in $n = 4$ coin tosses is $\frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x} = \frac{3}{8}$. This has a simple frequency interpretation: if we do many experiments, each of which involves tossing the coin 4 times, then in $\frac{3}{8}$ of the experiments, we will see 2 heads. Now suppose we have observed $x = 2$ heads in $n = 4$ tosses of a coin, what is the probability distribution of θ? To a non-Bayesian, this question is not meaningful. The approach of assigning a prior on θ based on subjective beliefs or without a physical model is the inverse method, and Edwards's formulation is not such a method.

In the tree problem, the Yule (or birth-death) process is a biologically plausible even if simple model. The true parameters are the Yule branching rate λ and the Brownian drift parameter σ^2, which should be estimated by maximizing the likelihood function, while the tree is a random variable and its realized value should be estimated from the conditional probability given the data. This is a standard likelihood method for estimating realized values of random variables in the model, nowadays known as empirical Bayes. Edwards would thus consider his method to be a likelihood method (although not maximum likelihood) of phylogeny estimation.

Nevertheless, modern use of the term "Bayesian" does not have the derogatory tone, and use of a physical/biological process is a common approach to specifying a prior. Of course, one may argue that the full or hierarchical Bayesian approach would assign priors on parameters λ and σ^2, rather than using their MLEs, but in this article, I have not made an effort to distinguish the empirical Bayes and the full (hierarchical) Bayes, or whether the prior is based on a biological model or chosen for convenience.

In passing, it may be noted that the challenge of summarizing the posterior distribution of phylogenetic trees still exists today. However, the singularity in the posterior mentioned by Edwards does not exist because the posterior for a tree is calculated by integrating over the ancestral states and branching times.

Modern times

The early studies of Rananla and Yang [10,21], Mau and Newton [19], and Li *et al.* [20] assumed the molecular clock (rate constancy over time), which is often violated in comparisons of distant species. Bayesian phylogenetics really took off with the development of the program MrBayes [38,39], which adapted branch-swapping algorithms such as nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR) [40] into MCMC proposal algorithms to move between trees. The clock constraint was relaxed, enabling phylogenetic inference to be conducted under more realistic models. A more recent program, BEAST, infers rooted trees under the clock and relaxed-clock models [9], while PhyloBayes implements sophisticated nonstationary models to deal with substitution heterogeneity among lineages that may be important for deep phylogenies [41].

Nowadays those Bayesian programs are standard tools in molecular phylogenetics, together with fast likelihood programs such as RAxML [42] and PhyML [43].

A brief introduction to Bayesian phylogenetics is provided in Yang [44]. More extensive recent reviews include Zwickl and Holder [45] and Yang [25: Chapters 8 and 9]. An edited book by Chen et al. [46] summarizes current research topics in the field. Phylogenetics may well be the largest application area of Bayesian statistics. It provides a rich testing ground for advanced Monte Carlo computational algorithms. Jerzy Neyman [47] was certainly right to identify molecular phylogenetics as "a source of novel statistical problems".

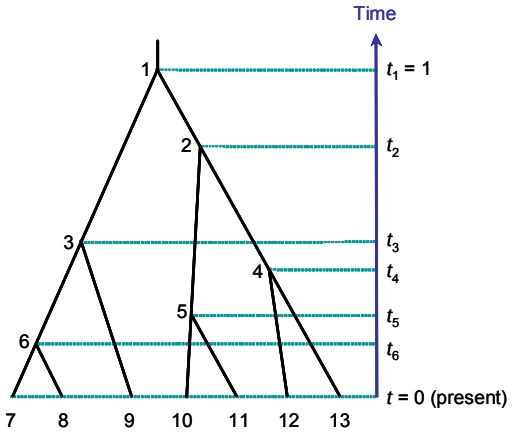

Figure 1. A phylogeny for seven ($n = 7$) species or populations used to illustrate the inference problem considered by Edwards [1]. The tips (final particles) are numbered n, $n + 1$, …, $2n - 1$. The interior nodes are numbered 1, 2, …, $n - 1$. They represent the branching events and are ordered by time: $t_1 > t_2 > \ldots > t_{n-1}$. The time machine runs backwards, so that the present time is $t = 0$ while the age of the root (the origin at the first split) is fixed at $t_1 = 1$. The data are observed measurements in p characters from the n modern species: $\boldsymbol{\xi} = \{\xi_{ik}\}$, where ξ_{ik} ($i = n, \ldots, 2n - 1$; $k = 1, \ldots, p$) is the measurement from species i in character k. Edwards considered the tree form (labelled history F), the times of non-root internal nodes, $\boldsymbol{t} = \{t_1 = 1, t_2, \ldots, t_{n-1}\}$, as well as the ancestral character states $\boldsymbol{x} = \{x_{ik}\}$, $i = 1, .., n - 1$; $k = 1, \ldots, p$, as quantities of interest.

1. Edwards, A. W. F. Estimation of the branch points of a branching diffusion process (with discussion). *J. R. Statist. Soc. B.* **32**, 155-174 (1970).
2. Edwards, A. W. F. & Cavalli-Sforza, L. L. Reconstruction of evolutionary trees. *Phenetic and Phylogenetic Classifications, Systematics Assoc. Publ.* **6**, 67-76 (1964).
3. Cavalli-Sforza, L. L. & Edwards, A. W. F. Estimation procedures for evolutionary branching processes. *Bull. Int. Statist. Inst.* **21**, 803-808 (1966).
4. Kingman, J. F. C. The coalescent. *Stochastic Process Appl.* **13**, 235-248 (1982).
5. Kendall, D. G. On the generalized birth-and-death process. *Ann. Math. Stat.* **19**, 1-15 (1948).
6. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.* **28**, 729-744 (2011).
7. Solis-Lemus, C., Knowles, L. L. & Ane, C. Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution* **69**, 492-507 (2015).
8. Cavalli-Sforza, L. L. & Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Evolution* **21**, 550-570 (1967).
9. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
10. Rannala, B. & Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304-311 (1996).
11. Stadler, T. Sampling-through-time in birth-death trees. *J. Theor. Biol.* **267**, 396-404 (2010).
12. Thompson, E. A. *Human Evolutionary Trees* (Cambridge University Press, Cambridge, England, 1975).
13. Gomberg, D. "Bayesian" postdiction in an evolution process. *Unfinished* (1966).
14. Yang, Z. & Roberts, D. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.* **12**, 451-458 (1995).
15. Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**, 471-492 (1973).
16. Felsenstein, J. *Inferring Phylogenies* (Sinauer Associates, Sunderland, Massachusetts, 2004).
17. Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* **29**, 170-179 (1989).
18. Smouse, P. E. & Li, W.-H. Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. *Evolution* **41**, 1162-1176 (1987).
19. Mau, B. & Newton, M. A. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Computat. Graph. Stat.* **6**, 122-131 (1997).
20. Li, S., Pearl, D. & Doss, H. Phylogenetic tree reconstruction using Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **95**, 493-508 (2000).
21. Yang, Z. & Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Mol. Biol. Evol.* **14**, 717-724 (1997).
22. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Physi.* **21**, 1087-1092 (1953).
23. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their application. *Biometrika* **57**, 97-109 (1970).
24. Yang, Z. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* **14**, 105-108 (1997).
25. Yang, Z. *Molecular Evolution: A Statistical Approach* (Oxford University Press, Oxford, England, 2014).
26. Edwards, A. W. F. *Likelihood* (Cambridge University Press, Cambridge, 1972).
27. Hillis, D. M. & Bull, J. J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182-192 (1993).
28. Felsenstein, J. & Kishino, H. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.* **42**, 193-200 (1993).
29. Zharkikh, A. & Li, W.-H. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. four taxa with a molecular clock. *Mol. Biol. Evol.* **9**,

1119-1147 (1992).
30. Felsenstein, J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22**, 240-249 (1973).
31. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376 (1981).
32. Horai, S. et al. Man's place in Hominoidea revealed by mitochondrial DNA genealogy [Erratum J Mol Evol 1993; 37:89]. *J. Mol. Evol.* **35**, 32-43 (1992).
33. Lewis, P. O., Holder, M. T. & Holsinger, K. E. Polytomies and Bayesian phylogenetic inference. *Syst. Biol.* **54**, 241-253 (2005).
34. Yang, Z. & Rannala, B. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* **54**, 455-470 (2005).
35. Yang, Z. Fair-balance paradox, star-tree paradox and Bayesian phylogenetics. *Mol. Biol. Evol.* **24**, 1639-1655 (2007).
36. Edwards, A. W. F. Comment on Bellhouse, David R. 'The Reverend Thomas Bayes FRS: A Biography to Celebrate the Tercentenary of his Birth'. *Stat. Sci.* **19**, 34-37 (2004).
37. Fienberg, S. E. When did Bayesian inference become "Bayesian"? *Bayesian Analysis* **1**, 1-40 (2006).
38. Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).
39. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539-542 (2012).
40. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* (eds. Hillis, D. M., Moritz, C. & Mable, B. K.) 407-514 (Sinauer Associates, Sunderland, Massachusetts, 1996).
41. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286-8 (2009).
42. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
43. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).
44. Yang, Z. in *Encyclopedia of Evolutionary Biology* (ed. Kliman) (Elsevier, 2015).
45. Zwickl, D. J. & Holder, M. T. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. *Syst. Biol.* **53**, 877-888 (2004).
46. Chen, M.-H., Kuo, L. & Lewis, P. *Bayesian Phylogenetics: Methods, Algorithms, and Applications* (Chapman & Hall/CRC, 2014).
47. Neyman, J. in *Statistical decision theory and related topics* (eds. Gupta, S. S. & Yackel, J.) 1-27 (Academic Press, New York, 1971).