7 March 2012

Review for Nature Reviews Genetics (Study Designs series):

Molecular phylogenetics: principles and practice

Ziheng Yang [a,b] and Bruce Rannala [a,c]
[a] Center for Computational and Evolutionary Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[b] Department of Biology, University College London, London WC1E 6BT, United Kingdom
[c] Genome Center and Department of Evolution and Ecology, University of California, Davis, CA 95616, USA

Correspondence to: Ziheng Yang
Address: Department of Biology
University College London
Darwin Building
Gower Street
London WC1E 6BT
UK
Email: z.yang@ucl.ac.uk
Phone: +44 (20) 7679 4379
Fax: +44 (20) 7679 7096

Abstract

Phylogenies are important for addressing various biological questions such as relationships among species or genes, the origin and spread of viral infection and the demographic changes and migration patterns of species. The advancement of sequencing technologies has taken phylogenetic analysis to a new height. Phylogenies have permeated nearly every branch of biology, and the plethora of phylogenetic methods and software packages now available may seem daunting to an experimental biologist. Here, we review the major methods of phylogenetic analysis, including parsimony, distance, likelihood and Bayesian methods. We discuss their strengths and weaknesses and provide guidance as to their use.

Before the advent of DNA sequencing technologies, phylogenetic trees were used almost exclusively to describe relationships among species in SYSTEMATICS and TAXONOMY. Today, phylogenies are used in almost every branch of biology. Besides representing the relationships among species on the tree of life, phylogenies are used to describe relationships between paralogues in a gene family[1], histories of populations[2], the evolutionary and epidemiological dynamics of pathogens[3,4], the genealogical relationship of somatic cells during differentiation and cancer development[5], and the evolution of language[6]. More recently, molecular phylogenetics has become an indispensible tool for genome comparisons. It is used to classify metagenomic sequences[7]; to identify genes, regulatory elements, and noncoding RNAs in newly sequenced genomes[8-10]; to interpret modern and ancient individual genomes[11-13]; and to reconstruct ancestral genomes[14,15].

In other applications, the phylogeny itself may not be of direct interest, but must nevertheless be accounted for in the analysis. This 'tree thinking' has transformed many branches of biology. In population genetics, the development of the COALESCENT theory[16,17] and the widespread availability of gene sequences for multiple individuals from the same species have prompted the development of genealogy-based inference methods, which have revolutionized modern computational population genetics. Here, the GENE TREES describing the genealogy of sequences in a sample are highly uncertain. They are not of direct interest but nevertheless contain valuable information about parameters in the model. Tree thinking has also forged a deep synthesis of population genetics and phylogenetics, creating the emerging field of STATISTICAL PHYLOGEOGRAPHY. In SPECIES-TREE methods[18,2,19], the gene trees at individual loci may not be of direct interest and may be in conflict with the species tree. By averaging over the unobserved gene trees under the multi-species coalescent model[20], those methods infer the species tree despite uncertainty in the gene trees. In comparative analysis, inference of associations between traits (for example, testis size and sexual promiscuity) using the observed traits of modern species should consider the species phylogeny, to avoid misinterpreting historical contingencies as causal relationships[21]. In the inference of adaptive protein evolution, the phylogeny is used to trace the synonymous and nonsynonymous substitutions along branches to identify cases of accelerated amino acid change[22], even though the phylogeny is not of direct interest.

Nowadays, every biologist needs to know something about phylogenetic inference. However, to an experimental biologist unfamiliar with the field, the existence of many analytical methods and software packages may seem daunting. In this review, we describe the suite of current methodologies of phylogenetic inference using sequence data. We also discuss various statistical criteria that are useful for choosing suitable methods for a particular question and data type. Next-generation sequencing (NGS) technologies are generating huge datasets. In analysis of such datasets, reducing SYSTEMATIC ERRORS and increasing robustness to model violations are much more important than reducing RANDOM SAMPLING ERRORS. We discuss several issues in the analysis of large datasets, such as the impact of missing data and strategies of data partitioning. The literature of molecular phylogenetics is large and complex[23,24]; the aim of this review is to provide a starting point for exploring the methods further.

Phylogenetic tree reconstruction: basic concepts

A phylogeny is a tree containing nodes connected by branches. Each branch represents the persistence of a genetic lineage through time and each node represents a birth of a new lineage (see

BOX 1). If the tree represents the relationship among a group of species the nodes represent speciation events. In other contexts the interpretation may be different. For example, in a gene tree of sequences sampled from a population the nodes represent birth events of individuals ancestral to the sample, whereas in a tree of paralogous gene families the nodes might represent gene duplication events.

Phylogenetic trees are not observed directly and are instead inferred from sequence or other data. Phylogeny reconstruction methods are either distance-based or character-based. In distance matrix (DM) methods, the distance between every pair of sequences is calculated, and the resulting distance matrix is used for tree reconstruction. For instance, Neighbour Joining (NJ)[25] applies a CLUSTER ALGORITHM to the distance matrix to arrive at a fully resolved phylogeny. Character-based methods include maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI) methods. They compare all sequences in the alignment simultaneously, considering one character (a site in the alignment) at a time to calculate a score for each tree. The tree score is the minimum number of changes for MP, the log-likelihood value for ML, and the POSTERIOR PROBABILITY for BI. In theory, all possible trees should be compared to identify the tree with the best score. In practice such an exhaustive search is not feasible computationally except for very small datasets because of the huge number of possible trees. Instead, heuristic tree-search algorithms are used. They often generate a starting tree using a fast algorithm and then perform local rearrangements to attempt to improve the tree score. Heuristic tree search is not guaranteed to find the best tree under the criterion, but renders feasible the analysis of large datasets. To describe the data, DM, ML, and BI all make use of a substitution model and are therefore model-based, whereas MP does not have an explicit model and its assumptions are implicit.

Distance-matrix (DM) method

Distance calculation. Pairwise sequence distances are calculated assuming a MARKOV CHAIN model of nucleotide substitution. Several commonly used models are illustrated in **Figure 1**. The JC69 model[26] an equal rate of substitution between any two nucleotides whereas the K80 model[27] assumes different rates for TRANSITIONS and TRANSVERSIONS. Both models predict equal frequencies of the four nucleotides. The assumption of equal base frequencies is relaxed in the HKY85 model[28] and general-time-reversible (GTR) model[29,30]. Because of variation in local mutation rate and in selective constraint, different sites in a DNA or protein sequence often evolve at different rates. In distance calculation, such rate variation is accommodated by assuming a gamma distribution of rates for sites[31], leading to models such as JC69+Γ, HKY85+Γ, or GTR+Γ.

Distance matrix (DM) methods. Once the distances are calculated, the sequence alignment is no longer used in DM methods. Here we mention three such methods: least-squares (LS), minimum evolution (ME) and Neighbour Joining (NJ). The LS method[32, see also 33]) minimizes a measure of the differences between the calculated distances (d_{ij}) in the distance matrix and the expected distances $(\hat{d}_{ij})$ on the tree (that is, the sum of branch lengths on the tree linking the two species i and j):

$$Q = \sum_{i=1}^{s} \sum_{i=1}^{s} (\hat{d}_{ij} - d_{ij})^2 \ . \qquad (1)$$

This is the same least-squares method used in statistics for fitting a straight line $y = a + bx$ to a scatter-plot. Optimizing branch lengths (or $\hat{d}_{ij}$) leads to the score Q for the given tree, and the tree with the smallest score is the LS estimate of the true tree.

The minimum evolution (ME) method[34,35] uses the tree length (sum of branch lengths) instead of Q for tree selection, even though the branch lengths can still be estimated using the LS criterion. Under the ME criterion, shorter trees are more likely to be correct than longer trees.

The most widely used distance method is Neighbour Joining (NJ)[25]. This is a cluster algorithm and operates by starting with a star tree and successively choosing a pair of taxa to join (based on the taxon distances), until a fully-resolved tree is obtained. Taxa are chosen for joining to minimize an estimate of tree length (sum of branch lengths)[36]. The two joined taxa (e.g., species 1 and 2 in **FIG. 2**) are then represented by their ancestor (e.g., node y in **FIG. 2**), with the number of taxa connected to the root (node *x*) reduced by 1 (**FIG. 2**). The distance matrix is then updated with the joined taxa replacing the two original taxa. See ref.[36] for a discussion of the NJ updating formula. An efficient implementation of NJ is found in the program MEGA[37] (**Table 1**).

Strengths and weaknesses of distance methods. One advantage of distance methods (especially NJ) is their computational efficiency. The cluster algorithm is fast because it does not need to compare many trees under an optimality criterion (as do MP and ML). For this reason, NJ is useful for analysing large datasets that have low levels of sequence divergence. Note that it might be important to use a realistic substitution model to calculate the pairwise distances. Distance methods can perform poorly for very divergent sequences, because large distances involve large sampling errors and most distance methods (such as NJ) do not account for the high variances of large distance estimates. Distance methods are also sensitive to gaps in the sequence alignment[38].

Maximum parsimony

Parsimony tree score. The maximum parsimony (MP) method minimizes the number of changes on a phylogenetic tree by assigning character states to interior nodes on the tree. The *character (site) length* is the minimum number of changes required for that site while the *tree score* is the sum of character lengths over all sites. The MP tree is the tree that minimizes the tree score.

Some sites are not useful for tree comparison by parsimony. The *constant sites*, for which the same nucleotide occurs in all species, have character length 0 on any tree. *Singleton sites*, at which only one of the species has a distinct nucleotide while all others are the same, can also be ignored as the character length is always 1. The *parsimony-informative* sites are those at which at least two distinct characters are observed, each at least twice. For four species, only three site patterns are informative: *xxyy*, *xyxy*, and *xyyx*, where *x* and *y* are any two distinct nucleotides. There are three possible UNROOTED TREES for four species and which of them is the MP tree depends on which of the three site patterns occurs most often in the alignment.

An algorithm for finding the minimum number of changes on a binary tree (and for reconstructing the ancestral states to achieve the minimum) was developed by Fitch[39] and Hartigan[40]. PAUP[41], MEGA[37], and TNT[42] are commonly used parsimony programs.

Parsimony was originally developed for use in analysing discrete morphological characters. During the late 1970s it began to be applied to molecular data. A controversy arose concerning whether parsimony (without explicit assumptions) or likelihood (with an explicit evolutionary model) was a better method for phylogenetic analysis[23]. The controversy has subsided and the importance of model-based inference methods is broadly recognized. The use of parsimony is still common, not because it is believed to be assumption-free, but because it often produces reasonable results and is computationally efficient.

Strengths and weaknesses of parsimony. A strength of parsimony is its simplicity; it is easy to describe and understand, and is amenable to rigorous mathematical analysis. The simplicity also helps development of efficient computer algorithms.

A major weakness of parsimony is its lack of explicit assumptions, which makes it nearly impossible to incorporate any knowledge of the process of sequence evolution in tree reconstruction. The failure of parsimony to correct for multiple substitutions at the same site makes it suffer from a problem known as LONG-BRANCH ATTRACTION[43]. If the correct tree (T_1 in **FIG. 3**a) has two long external branches separated by a short internal branch, parsimony tends to infer the incorrect tree T_2 of **FIG 3b**, with the long branches grouped together. When the branch lengths in T_1 are extreme enough, the probability for site pattern *xxyy*, which supports the correct tree T_1, may be smaller than that for *xyxy*, which supports the incorrect tree T_2. Thus, the more sites are in the sequence, the more probable it is for the pattern *xxyy* to be observed at fewer sites than *xyxy* and the more certain the incorrect tree T_2 is chosen to be the MP tree. Parsimony thus converges to a wrong tree and is statistically inconsistent. Long-branch attraction has been demonstrated in many real and simulated datasets[44] and is due to the failure of parsimony to correct for multiple changes at the same site or to accommodate parallel changes on the two long branches. See refs. [45,24: pp. 192-6] for more discussions of the issue.

Note that model-based methods (distance, likelihood and Bayesian) also suffer from long-branch attraction if the assumed model is too simplistic and ignores among-site rate variation [46]. In reconstruction of deep phylogenies, long-branch attraction (as well as unequal nucleotide or amino acid frequencies among species) is an important source of systematic error[47,48] (**FIG. 3c, d**). In such analyses it is advisable to use realistic substitution models and likelihood or Bayesian methodologies. Dense taxon sampling to break long branches and removing fast-evolving proteins or sites can also be helpful.

Maximum likelihood

Basis of ML. Maximum likelihood (ML) was developed by R.A. Fisher in the 1920s as a statistical methodology for estimating unknown parameters in a model. The likelihood function is defined as the probability of the data given the parameters, but viewed as a function of the parameters with the data observed and fixed. It represents all information in the data about the parameters. The maximum likelihood estimates (MLEs) of parameters are the parameter values that maximize the likelihood. Most often the MLEs are found numerically using iterative optimization algorithms. The MLEs have nice asymptotic (large-sample) properties; they are unbiased, consistent (they approach the true values) and efficient (they have the smallest variance among unbiased estimates).

ML tree reconstruction. The first algorithm for ML analysis of DNA sequence data was developed by Felsenstein[49]. The method is now widely used due both to the increased computing power and software implementations and to the development of increasingly realistic models of sequence evolution. Note that two optimization steps are involved in ML tree estimation: optimization of branch lengths to calculate the tree score for each candidate tree and a search in the tree space for the ML tree. From a statistical point of view, the tree (topology) is a model instead of a parameter, whereas branch lengths on the given tree and substitution parameters are parameters in the model. ML tree inference is thus equivalent to comparing many statistical models, each having the same number of parameters. The nice asymptotic properties of MLEs mentioned above apply to parameter estimation when the true tree is given but not to ML tree[50,24].

Calculation of the likelihood on a given tree under a variety of substitution models is explained in refs.[23,24]. All substitution models used in distance calculation can be used here. Indeed, joint comparison of many sequences by likelihood makes it feasible to accommodate much more sophisticated models of sequence evolution. Most models used in molecular phylogenetics assume independent evolution of sites in the sequence so that the likelihood is a product of the probabilities for different sites. The probability at any particular site is an average over the unobserved character states at the ancestral nodes. Likelihood and parsimony analyses are similar in this respect, although parsimony uses only the optimal ancestral states while likelihood averages over all possible states.

Early ML implementations include PHYLIP[51], MOLPHY[52] and PAUP* 4.0 [41]. Modern implementations, such as PhyML[53], RAxML[54], and GARLI[55], are not only much faster computationally but also more effective in finding trees with high likelihood scores. The recent inclusion of ML in MEGA 5[37] has made the method more accessible to biologists who are not experienced computer users (**Table 1**).

Strengths and weaknesses of the ML method. One advantage of the ML method is that all its model assumptions are explicit, so that they can be evaluated and improved upon. The availability of a rich repertoire of sophisticated evolutionary models in the likelihood (and Bayesian) methods is one of their major advantages over MP. Modern inferences of deep phylogenies using conserved proteins rely almost exclusively on likelihood and Bayesian methods. For such inference, it is important for the model to accommodate variable amino acid substitution rates among sites[56] or even different amino acid frequencies among sites[57,58].

ML has a clear advantage over distance or parsimony methods if our aim is to understand the process of sequence evolution. The LIKELIHOOD RATIO TEST can be used to examine the fit of evolutionary models[59] and to test interesting biological hypotheses, such as the molecular clock[60,49] and Darwinian selection affecting protein evolution[61-63]. See references[64,65,24,22] for summaries of such tests in phylogenetics.

The main drawback of ML is that the likelihood calculation and, in particular, tree search under the likelihood criterion is computationally demanding. Another drawback is that the method has potentially poor statistical properties if the model is mis-specified. This is also true for Bayesian analysis (**Table 2**).

Bayesian methods

Basis of Bayesian inference. Bayesian inference (BI) is a general methodology of statistical inference. It differs from ML in that parameters in the model are considered random variables with statistical distributions whereas in ML they are (unknown) fixed constants. Before the analysis of the data, parameters are assigned a PRIOR DISTRIBUTION, which is combined with the data (likelihood) to generate the POSTERIOR DISTRIBUTION. All inferences concerning the parameters are then based on the posterior distribution. In the past two decades Bayesian inference has gained popularity thanks to advances in computational methods, especially MARKOV CHAIN MONTE CARLO (MCMC) algorithms.

Bayesian phylogenetics. Bayesian inference was introduced to molecular phylogenetics in the late 1990s[66-69]. The early methods assumed a molecular clock[60]. Development of more efficient MCMC algorithms[70] that eliminate the clock assumption (allowing independent branch lengths on unrooted trees) and the release of the program MrBayes[71] made the method very popular among molecular systematists. A more recent Bayesian implementation in the program BEAST[72] uses the so-called relaxed-clock models to infer rooted trees even though the model allows substitution rates to vary across lineages (**Table 1**).

Bayesian inference relies on Bayes's theorem, which states that

$$\mathrm{P}(T,\theta\,|\,D)=\frac{\mathrm{P}(T,\theta)\mathrm{P}(D\,|\,T,\theta)}{\mathrm{P}(D)},$$

where $\mathrm{P}(T, \theta)$ is the *prior probability* for tree T and parameter θ, $\mathrm{P}(D|T, \theta)$ is the likelihood or probability of the data given the tree and parameter, and $\mathrm{P}(T, \theta|D)$ is the *posterior probability*. The denominator $\mathrm{P}(D)$ is a *normalizing constant* as its role is to ensure that $\mathrm{P}(T, \theta|D)$ sums over the trees and integrates over the parameters to 1. The theorem states that *the posterior is proportional to the prior times the likelihood* or *the posterior information is the prior information plus the data information*.

In general the posterior probabilities of trees cannot be calculated directly. In particular, the normalizing constant $\mathrm{P}(D)$ involves high-dimensional integrals (over all possible parameter values θs) and summation over all possible trees. Instead, Bayesian phylogenetic inference relies on MCMC to generate a sample from the posterior distribution. This is illustrated in **Box** 2. See Chapter 5 of ref.[24] for an introduction.

Strengths and weaknesses of the BI method. Both likelihood and Bayesian methods use the likelihood function and thus share many statistical properties such as consistency and efficiency. However, ML and BI represent opposing philosophies of statistical inference. The same feature of BI may thus be viewed as either a strength or weakness depending on one's philosophy. See ref.[24] for a brief description of the controversy. Here we comment on two issues, of interpretability and of the prior.

First, Bayesian statistics is known to answer the biological questions directly and yields results that are easy to interpret: the posterior probability of a tree is simply the probability that the tree is correct, given the data and model. In contrast, concepts such as the confidence interval in a likelihood analysis have a contrived interpretation that eludes many users of statistics. In phylogenetics, it has not been possible to define a confidence interval for the tree. The widely used bootstrap method[73] (see **Box 3**) has been hard to interpret despite numerous efforts[74-77]. However, the odds are not entirely against ML. Posterior probabilities for trees and CLADES calculated from real datasets often appear too high[66,78-80]. In many analyses, nearly all nodes had posterior probabilities of ~100%. Posterior tree probabilities are also sensitive to model violations, and use of simplistic models may lead to inflated posterior probabilities[81].

Second, the prior probability allows one to incorporate *a priori* information about the trees or parameters. However, such information is rarely available, and specification of the prior is most often a burden on the user; almost all data analyses are conducted using the "default" priors in the computer program. High-dimensional priors are notoriously hard to specify, and an innocent-looking prior can have an undue and unexpected influence on the posterior. For example, it has recently been pointed out that the independent exponential prior on branch lengths used by MrBayes can induce a strongly informative and unreasonable prior on the tree length, producing unreasonably long trees in some datasets[82-84]. It is therefore important to conduct Bayesian robustness analysis to assess the impact of the prior on the posterior estimates.

Statistical assessments of phylogenetic methods

The aim of phylogenetic inference is to estimate the tree topology, and possibly also the branch lengths. A number of criteria have been used to judge tree-reconstruction methods.

Consistency. An estimation method is said to be consistent if the estimate converges to the true parameter value when the amount of data approaches infinity. A tree reconstruction method is consistent if the estimated tree converges to the true tree when the number of sites in the sequence grows. Model-based methods (DM, ML, and BI) are consistent if the assumed model is correct. Parsimony may be inconsistent under some model-tree combinations; Felsenstein's[43] demonstration of this has spurred much heated discussion.

Efficiency. In statistical estimation of a parameter, an unbiased estimate with a smaller variance is more efficient than one with a larger variance. In phylogenetics, efficiency may be measured by the probability of recovering the correct tree or subtree given the number of sites. This can be estimated by computer simulation. The complexity of tree reconstruction means that the asymptotic theory of MLEs does not apply. Nevertheless, computer simulations have generally found higher efficiency of ML than MP or NJ in recovering the correct tree[23].

Robustness. A method is robust if it gives correct answers even when its assumptions are violated. Obviously some assumptions matter more than others. With the rapid accumulation of sequence data, sampling errors in tree reconstruction are considerably reduced, so systematic errors or robustness of the method become more important.

Computational speed. The computational speed is easy to assess. NJ uses a cluster algorithm to arrive at a tree and is very fast. Methods that search for the best tree under a criterion, such as ME, MP, and ML, are slower. The computational speed of the Bayesian method depends on the length of the chain (MCMC), which is highly data-dependent. As calculation of the phylogenetic likelihood is very expensive, ML and BI are typically slower than MP. Nevertheless, considerable advancements in computational algorithms[53-55] have made likelihood-based methods feasible for analysis of large datasets. Algorithms taking advantage of new computers with multicore processors and GRAPHICAL PROCESSING UNITS (GPUs)[85,86] are pushing the boundary even further.

Phylogenomic analysis of large datasets

With the advent of new sequencing technologies and the completion of various genome projects, phylogenetics has entered the era of genome-scale datasets. Here we discuss a few issues related to the analysis of such large datasets.

Supertree and supermatrix approaches. Two approaches have been advocated for the phylogenetic analysis of hundreds or thousands of genes or proteins, especially when some loci are missing in some species.. The supertree approach analyses each gene separately and then uses heuristic algorithms to assemble the subtrees for individual genes into a supertree for all species[87]. The separate analysis is useful for studying the differences in the reconstructed subtrees or the prevalence of horizontal gene transfer. It is inefficient for estimating a common phylogeny underlying all genes. In the supermatrix approach, sequences for multiple genes are concatenated to generate a data supermatrix, in which missing data are replaced by question marks, and the supermatrix is then used for tree reconstruction[88]. Most supermatrix analyses ignore the differences in evolutionary dynamics among the genes. Note that a supermatrix analysis assuming different evolutionary models and different trees and branch lengths for the genes is equivalent to a separate or supertree analysis. When a common tree underlies all genes, the ideal approach should be a combined (supermatrix) analysis of all genes, using the likelihood to accommodate the among-gene heterogeneity in the evolutionary process[89-91]. Our comments below concern this combined approach.

Impact of missing data. Many genomic datasets are highly incomplete, with most cells in the species × gene matrix missing. While in theory the likelihood function (in the ML and Bayesian methods) can accommodate missing data properly[23: pp.255-6, 24: pp. 107-8], the impact of such large-scale missing data and alignment gaps is not well understood. Simulations suggest that ML and BI generally perform better than NJ or MP in dealing with missing data, with BI performing best[92-94].

The poor performance of NJ may be understood if one considers extreme cases where after removal of alignment gaps, the different pairwise distances are calculated from different sets of genes or sites, some of which are fast-evolving while others are slowly-evolving.

Importance of systematic errors. In analysis of very large datasets, virtually all bootstrap support values or Bayesian posterior probabilities are calculated to be 100%, even though the inferred phylogenies may be conflicting across genes or may depend on the method and model used[47]. Systematic biases are thus much more important than random sampling errors in such analyses, and methods robust to violations of model assumptions, even if they are less efficient, should be preferable.

Data partitioning strategies. The rationale for data partitioning is to group genes or sites with similar evolutionary characteristics into the same partition, so that all sites in the same partition are described using the same model, while different partitions use different models[89,90]. Partitioning too finely increases computation and may cause over-fitting, while partitioning too coarsely may lead to under-fitting or model violation. However the situation is complicated as the assumed model may deal with random variation among sites in substitution rate [31,56,61], in amino acid frequencies [57,58,95], or in the pattern of substitution[96]. Such *mixture models* use a statistical distribution to accommodate the among-site heterogeneity without data partitioning. Often the choice of partition and mixture models is a philosophical one: they correspond to fixed-effects and random-effects models in statistics, respectively. Current strategies for data partitioning include partitioning genes according to their relative rates[97], and separating the three codon positions in coding genes into different partitions[89]. The likelihood ratio test has also been used to decide whether two genes should be in the same or different partitions[98]. In sum, data partitioning is more art than science, and should rely on our knowledge of the biological system: for example, should the same phylogeny underlie all genes?

Perspectives

Here we discuss a few research areas that are currently the focus of much methodological development. The first is multiple sequence alignment. Many heuristic methods and programs for aligning sequences exist[99,100], and improved algorithms continue to appear[101,102]. Efforts have also been taken to infer alignment statistically under an explicit model of insertions and deletions[103,104], and to infer alignment and phylogeny jointly in a Bayesian framework[105,106]. An advantage of those model-based alignment methods is that they produce estimates of insertion and deletion rates. For the present those algorithms are based on simplistic insertion-deletion models and involve heavy computation so that they are not competitive, in either computational efficiency or alignment quality, against good heuristic algorithms. Nevertheless they are biologically appealing and improvements are very likely.

The second area is molecular clock estimation of divergence dates. Under the clock assumption, the distance between sequences increases linearly with the time of divergence, and if a particular divergence can be assigned an absolute geological age based on the fossil record, the substitution rate can be calculated and all divergences on the tree can be dated. Similar ideas can be used to estimate divergence times of viral strains when sample dates for viral sequences are available and act as calibrations. However, in practice, the molecular clock may be violated, especially for distantly-related species, and the fossil record can never provide unambiguous times of lineage divergence. In the past several years, advancements have been made in the Bayesian framework to deal with those issues. Since the pioneering work of Thorne and colleagues[107,108], models of evolutionary rate drift over time have been developed to relax the molecular clock[72,109]. Soft age bounds and flexible probability distributions have been implemented to accommodate uncertainties in fossil calibrations[72,110,111]. The fossil record (or the presence and absence of fossils in the rock layers) has also been analyzed statistically to generate calibration densities for molecular dating analysis[112,113].

The third area of exciting development, mentioned at the beginning of this review, is statistical phylogeography, at the interface of population genetics and phylogeography[20,114-116]. The availability of genomic data at both the species and population levels offers unprecedented opportunities for addressing interesting questions in evolutionary biology. Multi-locus sequence data can be used to estimate divergence times between closely-related species and the sizes of both extant and extinct populations[117,118], to infer population demographic changes and to estimate migration patterns and rates[119,120]. Such data can also be used to delimit species (that is, to infer one vs. two species, say)[121,122]. The last few years have seen the appearance of individual genomes and the rise of population genomics. Currently the data are mostly from the human and close relatives, but genomes from other species are being sequenced as well, such as the mastodon and mammoth[123] and the bacteria *Yersinia pestis* from Black Death victims[124]. Genomic sequence data from the human and the apes are used to infer the species divergence times and to test for possible hybridization during the human-chimpanzee separation[125-131]. Comparison of a few human individual genomes provides insights into the recent demographic history of our species[12,13], while sequencing of the Neandertal genome allows estimation of the Neandertal contribution to the genome of modern humans[11,132]. The size of the data and the complexity of the model pose great statistical and computational challenges. Again Bayesian MCMC, under the multi-species coalescent model[118], provides the natural framework for inference.

References

1. Maser, P. et al. Phylogenetic relationships within cation transporter families of Arabidopsis. *Plant Physiol* **126**, 1646-1667 (2001).
2. Edwards, S.V. Is a new and general theory of molecular systematics emerging? *Evolution* **63**, 1–19 (2009).
3. Marra, M.A. et al. The genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399-1404 (2003).
4. Grenfell, B.T. et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327-332 (2004).
5. Salipante, S.J. & Horwitz, M.S. Phylogenetic fate mapping. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5448-5453 (2006).
6. Gray, R.D., Drummond, A.J. & Greenhill, S.J. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* **323**, 479-483 (2009).
7. Brady, A. & Salzberg, S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nature Methods* **8**, 367 (2011).
8. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254 (2003).
9. Pedersen, J.S. et al. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2**, e33 (2006).
10. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482 (2011).
11. Green, R.E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710-722 (2010).
12. Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genet.* **43**, 1031-1034 (2011).
13. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496 (2011).
14. Paten, B. et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829-1843 (2008).
15. Ma, J. Reconstructing the history of large-scale genomic changes: biological questions and computational challenges. *J. Comput. Biol.* **18**, 879-893 (2011).
16. Kingman, J.F.C. On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27-43 (1982).
17. Kingman, J.F.C. The coalescent. *Stochastic Process Appl.* **13**, 235-248 (1982).
18. Edwards, S.V., Liu, L. & Pearl, D.K. High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5936-5941 (2007).
   Introduces a method for estimating the species tree despite the presence of conflicting gene trees.
19. Than, C. & Nakhleh, L. Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.* **5**, e1000501 (2009).
20. Rannala, B. & Yang, Z. Phylogenetic inference using whole genomes. *Ann. Rev. Genom. Hum. Genet.* **9**, 217-231 (2008).
21. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1-15 (1985).
   Introduces bootstrap approach to phylogenetic analysis. This is the most commonly used method for assessing sampling errors in estimated phylogenies.
22. Yang, Z. Adaptive molecular evolution. in *Handbook of Statistical Genetics* (eds. Balding, D., Bishop, M. & Cannings, C.) 377-406 (Wiley, New York, 2007).
23. Felsenstein, J. *Inferring Phylogenies*, (Sinauer Associates, Sunderland, Massachusetts, 2004).
24. Yang, Z. *Computational Molecular Evolution*, (Oxford University Press, Oxford, England, 2006).
25. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425 (1987).
26. Jukes, T.H. & Cantor, C.R. Evolution of protein molecules. in *Mammalian Protein Metabolism* (ed. Munro, H.N.) 21-123 (Academic Press, New York, 1969).
27. Kimura, M. A simple method for estimating evolutionary rate of base substitution through

comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120 (1980).

28. Hasegawa, M., Kishino, H. & Yano, T. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160-174 (1985).
29. Tavaré, S. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57-86 (1986).
30. Yang, Z. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105-111 (1994).
31. Yang, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**, 1396-1401 (1993).
32. Cavalli-Sforza, L.L. & Edwards, A.W.F. Phylogenetic analysis: models and estimation procedures. *Evolution* **21**, 550-570 (1967).
33. Fitch, W.M. & Margoliash, E. Construction of phylogenetic trees. *Science* **155**, 279-284 (1967).
34. Rzhetsky, A. & Nei, M. A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* **9**, 945-967 (1992).
35. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* **9**, 687-705 (2002).
36. Gascuel, O. & Steel, M. Neighbor-joining revealed. *Mol. Biol. Evol.* **23**, 1997-2000 (2006).
37. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731-2739 (2011).
38. Bruno, W.J., Socci, N.D. & Halpern, A.L. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* **17**, 189-197 (2000).
39. Fitch, W.M. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**, 406-416 (1971).
40. Hartigan, J.A. Minimum evolution fits to a given tree. *Biometrics* **29**, 53-65 (1973).
41. Swofford, D.L. PAUP*: Phylogenetic Analysis by Parsimony, Version 4. (Sinauer Associates, Sanderland, Massachusetts, 2000).
42. Goloboff, P.A., Farris, J.S. & Nixon, K.C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774-786 (2008).
43. Felsenstein, J. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401-410 (1978).
44. Huelsenbeck, J.P. Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Syst. Biol.* **47**, 519-537 (1998).
45. Swofford, D.L. et al. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* **50**, 525-539 (2001).
46. Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367-372 (1996).
47. Philippe, H. et al. Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature* **470**, 255-25 (2011).
48. Zhong, B. et al. Systematic error in seed plant phylogenomics. *Genome Biol. Evol.* **3**, 1340-1348 (2011).
49. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368-376 (1981).
    Introduces the pruning algorithm for likelihood calculation on a tree. This approach forms the basis for modern likelihood and Bayesian methods of phylogenetic analysis.
50. Yang, Z. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42**, 294-307 (1996).
51. Felsenstein, J. Phylip: Phylogenetic Inference Program, Version 3.6. *University of Washington*, Seattle (2005).
52. Adachi, J. & Hasegawa, M. MOLPHY Version 2.3: Programs for molecular phylogenetics based on maximum likelihood. *Computer Science Monographs* **28**, 1-150 (1996).
53. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704 (2003).
54. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).

55. Zwickl, D. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. *Ph.D. Thesis*, University of Texas at Austin (2006).
56. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306-314 (1994).
57. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095-1109 (2004).
58. Blanquart, S. & Lartillot, N. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **25**, 842-858 (2008).
59. Goldman, N. Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**, 182-198 (1993).
60. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. in *Evolving Genes and Proteins* (eds. Bryson, V. & Vogel, H.J.) 97-166 (Academic Press, New York, 1965).
61. Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936 (1998).
62. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568-573 (1998).
63. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908-917 (2002).
64. Huelsenbeck, J.P. & Rannala, B. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227-232 (1997).
65. Whelan, S., Liò, P. & Goldman, N. Molecular phylogenetics: state of the art methods for looking into the past. *Trends Genet.* **17**, 262-272 (2001).
66. Rannala, B. & Yang, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304-311 (1996).
67. Yang, Z. & Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Mol. Biol. Evol.* **14**, 717-724 (1997).
68. 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).
69. Li, S., Pearl, D. & Doss, H. Phylogenetic tree reconstruction using Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **95**, 493-508 (2000).
70. Larget, B. & Simon, D.L. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**, 750-759 (1999).
71. Huelsenbeck, J.P. & Ronquist, F. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).
72. Drummond, A.J., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Bio.* **4**, e88 (2006).
Introduces a Bayesian MCMC algorithm (the BEAST program) to estimate rooted trees under relaxed-clock models.
73. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791 (1985).
74. 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).
75. Efron, B., Halloran, E. & Holmes, S. Bootstrap confidence levels for phylogenetic trees [corrected and republished article originally printed in Proc. Natl. Acad. Sci. U.S.A. 1996 , **93**:7085-7090]. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 13429-13434 (1996).
76. Berry, V. & Gascuel, O. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Mol. Biol. Evol.* **13**, 999-1011 (1996).
77. Susko, E. First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. *Mol Biol Evol* **27**, 1621-1629 (2010).
78. Suzuki, Y., Glazko, G.V. & Nei, M. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 16138-16143 (2002).
79. Lewis, P.O., Holder, M.T. & Holsinger, K.E. Polytomies and Bayesian phylogenetic inference. *Syst. Biol.* **54**, 241-253 (2005).

80. Yang, Z. & Rannala, B. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* **54**, 455-470 (2005).
81. Huelsenbeck, J.P. & Rannala, B. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* **53**, 904-913 (2004).
82. Brown, J.M., Hedtke, S.M., Lemmon, A.R. & Lemmon, E.M. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. *Syst. Biol.* **59**, 145-161 (2010).
83. Rannala, B., Zhu, T. & Yang, Z. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. *Mol. Biol. Evol.* **29**, 325-335 (2012).
84. Zhang, C., Rannala, B. & Yang, Z. Robustness of compound Dirichlet priors for Bayesian inference of branch lengths. *Syst. Biol.*, 10 Feb 2012 (doi: 10.1093/sysbio/sys030) (2012).
85. Suchard, M. & Rambaut, A. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**, 1370-1376 (2009).
86. Zierke, S. & Bakos, J. FPGA acceleration of the phylogenetic likelihood function for Bayesian MCMC inference methods. *BMC Bioinform.* **11**, 184 (2010).
87. Bininda-Emonds, O.R.P. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, (Kluwer Academic, Dordrecht, the Netherlands, 2004).
88. de Queiroz, A. & Gatesy, J. The supermatrix approach to systematics. *Trends Ecol Evol* **22**, 34-41 (2007).
89. Yang, Z. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**, 587-596 (1996).
90. Shapiro, B., Rambaut, A. & Drummond, A.J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* **23**, 7-9 (2006).
91. Ren, F., Tanaka, H. & Yang, Z. A likelihood look at the supermatrix-supertree controversy. *Gene* **441**, 119-125 (2009).
92. Criscuolo, A., Berry, V., Douzery, E.J. & Gascuel, O. SDM: a fast distance-based approach for (super) tree building in phylogenomics. *Syst Biol* **55**, 740-755 (2006).
93. Wiens, J.J. & Moen, D.S. Missing data and the accuracy of Bayesian phylogenetics. *J. Syst. Evol.* **46**, 307-314 (2008).
94. Dwivedi, B. & Gadagkar, S. Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evol. Biol.* **9**, 1471-2148 (2009).
95. Rodrigue, N., Philippe, H. & Lartillot, N. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4629-4634 (2010).
96. 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).
97. Nishihara, H., Okada, N. & Hasegawa, M. Rooting the Eutherian tree - the power and pitfalls of phylogenomics. *Genome Biol* **8**, R199 (2007).
98. Leigh, J.W., Susko, E., Baumgartner, M. & Roger, A.J. Testing congruence in phylogenomic analysis. *Syst. Biol.* **57**, 104-115 (2008).
99. Higgins, D.G. & Sharp, P.M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237-244 (1988).
100. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).
101. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* **102**, 10557-10562 (2005).
102. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632-1635 (2008).
103. Thorne, J.L., Kishino, H. & Felsenstein, J. An evolutionary model for maximum likelihood alignment of DNA sequences [Erratum in J. Mol. Evol. 1992, 34:91]. *J. Mol. Evol.* **33**, 114-124 (1991).
104. Hein, J., Jensen, J.L. & Pedersen, C.N. Recursions for statistical multiple alignment. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14960-14965 (2003).
105. Redelings, B.D. & Suchard, M.A. Joint Bayesian estimation of alignment and phylogeny.

Syst. Biol. **54**, 401-418 (2005).

106. Lunter, G., Miklos, I., Drummond, A., Jensen, J.L. & Hein, J. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* **6**, 83 (2005).

107. Thorne, J.L., Kishino, H. & Painter, I.S. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647-1657 (1998).
Describes the first Bayesian MCMC method for dating species divergence using minimum and maximum bounds [NB to copyed 'bound' is a statistical term and should not be changed to 'boundary'] to incorporate fossil calibrations.

108. Kishino, H., Thorne, J.L. & Bruno, W.J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352-361 (2001).

109. Rannala, B. & Yang, Z. Inferring speciation times under an episodic molecular clock. *Syst. Biol.* **56**, 453-466 (2007).

110. Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* **23**, 212-226 (2006).

111. Inoue, J., Donoghue, P.C.H. & Yang, Z. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.* **59**, 74-89 (2010).

112. Tavaré, S., Marshall, C.R., Will, O., Soligos, C. & Martin, R.D. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416**, 726-729 (2002).

113. Wilkinson, R.D. et al. Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Syst. Biol.* **60**, 16-31 (2011).

114. Knowles, L.L. Statistical phylogeography. *Ann. Rev. Ecol. Syst.* **40**, 593-612 (2009).

115. Lemey, P., Rambaut, A., Drummond, A.J. & Suchard, M.A. Bayesian phylogeography finds its roots. *PLoS Comp. Biol.* **5**, e1000520 (2009).

116. Lemey, P., Rambaut, A., Welch, J.J. & Suchard, M.A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877-1885 (2010).

117. Takahata, N., Satta, Y. & Klein, J. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.* **48**, 198-221 (1995).

118. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645-1656 (2003).
Describes the multi-species coalescent model. This is the basis for carrying out comparative analyses of individual genomes, phylogeographic analyses, and species tree methods.

119. Drummond, A.J., Nicholls, G.K., Rodrigo, A.G. & Solomon, W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307-1320 (2002).

120. Hey, J. Isolation with migration models for more than two populations. *Mol. Biol. Evol.* **27**, 905-920 (2010).

121. Knowles, L.L. & Carstens, B.C. Delimiting species without monophyletic gene trees. *Syst. Biol.* **56**, 887-895 (2007).

122. Yang, Z. & Rannala, B. Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9264-9269 (2010).
Describes a Bayesian MCMC method for delimiting species using sequence data from multiple loci under the multi-species coalescent model.

123. Rohland, N. et al. Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biol.* **8**, e1000564 (2010).

124. Bos, K.I. et al. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506-510 (2011).

125. Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103-1108 (2006).

126. Innan, H. & Watanabe, H. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Mol. Biol. Evol.* **23**, 1040-1047 (2006).

127. Becquet, C. & Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* **17**, 1505-1519 (2007).

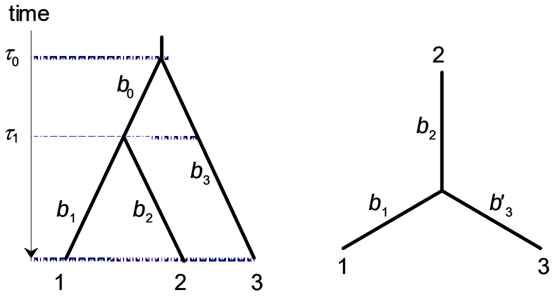
128. Hobolth, A., Christensen, O.F., Mailund, T. & Schierup, M.H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov

model. *PLoS Genet.* **3**, e7 (2007).
129. Burgess, R. & Yang, Z. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.* **25**, 1979-1994 (2008).
130. Becquet, C. & Przeworski, M. Learning about modes of speciation by computational approaches. *Evolution* **63**, 2547-2562 (2009).
131. Yang, Z. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genom. Biol. Evol.* **2**, 200-211 (2010).
132. Reich, D. et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060 (2010).
133. Sitnikova, T., Rzhetsky, A. & Nei, M. Interior-branch and bootstrap tests of phylogenetic trees. *Mol. Biol. Evol.* **12**, 319-333 (1995).
134. Zhong, B., Yonezawa, T., Zhong, Y. & Hasegawa, M. The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol. Biol. Evol.* **27**, 2855-2863 (2010).
135. Drummond, A.J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
136. Kosakovsky Pond, S.L., Frost, S.D.W. & Muse, S.V. HyPhy: hypothesis testing using phylogenies. *BioInformatics* **21**, 676-679 (2005).
137. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).
138. Lartillot, N. & Philippe, H. Computing Bayes factors using thermodynamic integration. *Syst. Biol.* **55**, 195-207 (2006).
139. Xie, W., Lewis, P.O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150-160 (2011).

Acknowledgements

We thank three referees for constructive comments, and Masami Hasegawa and Bojian Zhong for providing the seed-plant phylogenies of **Figure 3**. Z.Y. is supported by a Biotechnological and Biological Sciences Research Council Grant and a Royal Society/Wolfson Merit Award. B.R. is supported by a National Institutes of Health grant.

Box 1. TREE CONCEPTS



(a) Rooted tree (b) Unrooted tree

A phylogeny is a model of genealogical history in which the lengths of the branches are unknown parameters. For example, the phylogeny on the left is generated by two speciation events that occurred at time points τ_0 and τ_1. The branch lengths (b_0, b_1, b_2, b_3) are typically expressed in units of expected number of substitutions per site and measure the amount of evolution along the branches.

If the substitution rate is constant over time or among lineages, we say that the MOLECULAR CLOCK[60] holds. The tree will then be rooted and ultrametric, meaning that the distances from the tips of the tree to the root are all equal (e.g., $b_0 + b_1 = b_0 + b_2 = b_3$). A rooted tree for s species can then be represented by the ages of the $s - 1$ ancestral nodes and thus involves $s - 1$ branch-length parameters. The procedure of inferring rooted trees by assuming the molecular clock is called *molecular clock rooting*. For distantly related species, the clock hypothesis should not be assumed. Most phylogenetic analyses are therefore conducted without the assumption of the clock. If one allows every branch on the tree to have an independent evolutionary rate, commonly used models and methods are unable to identify the location of the root, so only unrooted trees are inferred. An unrooted tree for s species then has $2s - 3$ branch-length parameters. A commonly used strategy to 'root the tree' is to include in the analysis *outgroup species*, which are known to be more distantly related than the species we are interested in. While the inferred tree for all species is unrooted, the root is believed to be located along the branch leading to the outgroup, so that the tree for the *ingroup species* is rooted. This strategy is called *outgroup rooting*.

Box 2. Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo is a simulation algorithm in which one moves from one tree (or parameter value) to another and in the long run visits the trees (or parameters) in proportion to their posterior probabilities. The tree-parameter set (T, θ) constitutes the state of the algorithm. Here parameters θ may include the branch lengths of the tree and parameters in the evolutionary model such as the transition/transversion rate ratio. The following sketch illustrates the main features of MCMC algorithms.

- 1. (Initialization): Choose at random a starting tree and starting parameters (T, θ).
- 2. (Main loop)
    - 2a (Proposal to change T): Propose a new tree T^* by changing the current tree T. If T^* has higher posterior probability than the current tree, $\mathrm{P}(T^*, \theta|D) > \mathrm{P}(T, \theta|D)$, accept the new tree T^*. Otherwise accept T^* with probability $\frac{\mathrm{P}(T^*,\theta|D)}{\mathrm{P}(T,\theta|D)} = \frac{\mathrm{P}(T^*,\theta)\mathrm{P}(D|T^*,\theta)}{\mathrm{P}(T,\theta)\mathrm{P}(D|T,\theta)}$. If T^* is accepted, set $T = T^*$.
    - 2b (Proposal to change parameters θ): Propose new parameter values θ^* by changing the current θ. If $\mathrm{P}(T, \theta^*|D) > \mathrm{P}(T, \theta|D)$, accept the new θ^*. Otherwise accept θ^* with probability $\frac{\mathrm{P}(T,\theta^*|D)}{\mathrm{P}(T,\theta|D)} = \frac{\mathrm{P}(T,\theta^*)\mathrm{P}(D|T,\theta^*)}{\mathrm{P}(T,\theta)\mathrm{P}(D|T,\theta)}$. It the new θ^* is accepted, set $\theta = \theta^*$.
    - 2c (Sample from the chain): Print out (T, θ).

Note that first the algorithm does not need calculation of the normalizing constant $\mathrm{P}(D)$ as it cancels in the posterior ratios in proposal steps 2a and 2b. Second, in the long run a tree-parameter set (T_1, θ_1) will be visited more often by the algorithm than another set (T_2, θ_2) if its posterior probability is higher: $\mathrm{P}(T_1, \theta_1|D) > \mathrm{P}(T_2, \theta_2|D)$. Indeed the expected proportion of time that the algorithm spends in any tree T is exactly its posterior probability: $\mathrm{P}(T|D)$. Thus by counting the frequencies at which each tree is visited in the algorithm, we get an MCMC estimate of the posterior probabilities for the trees.

The sequence (chain) of values for (T, θ) generated by the algorithm has the property that given the current state (T, θ), the probabilities by which it moves to new states do not depend on past states. This memory-less property is known as the Markovian property, which states that *given the present, the future does not depend on the past*. The generated sequence is called a Markov chain. The algorithm is called Markov chain Monte Carlo because the Markov chain is generated by Monte Carlo simulation.

BOX 3. SAMPLING ERROR IN THE ESTIMATED TREE AND BOOTSTRAP

In traditional parameter estimation, we attach a confidence interval to indicate the uncertainty involved in the point estimate of the parameter. This has not been possible in molecular phylogenetics, as concepts such as the variance and confidence interval are not meaningful when applied to trees. For distance, parsimony, and likelihood methods, the most commonly used procedure to assess the confidence in a tree topology estimate is the bootstrap analysis[73]. In this approach, the sites in the sequence alignment are resampled with replacement as many times as the sequence length, generating a bootstrap pseudo-sample that is of the same size as the original dataset. Typically 100 or 1,000 bootstrap samples are generated in this way, and each one is analyzed in the same way as the original sequence alignment. An example using the maximum likelihood method is illustrated in the figure. The inferred trees from those bootstrap samples are then tabulated to calculate the bootstrap support values. For every CLADE in the estimated tree, its bootstrap support value is simply the proportion of bootstrap trees that include that clade[133,65,24]. The commonly used but less satisfactory approach is to use the bootstrap trees to generate a *majority-rule consensus tree*, which shows a clade if and only if it occurs in more than half of the bootstrap trees.

Figure legends

Figure 1. **Markov models of nucleotide substitution.**
The thickness of the lines indicates the substitution rates of the four nucleotides (T, C, A, G), and the sizes of the circles represent the nucleotide frequencies when the substitution process is in equilibrium. Note that both JC69 and K80 predict equal proportions of the four nucleotides.

Figure 2. **The neighbour joining (NJ) algorithm.**
The NJ algorithm is a divisive cluster algorithm. It starts from a star tree and joins two species or nodes, reducing the number of nodes at the root (node x) by one. The process is repeated until a fullyresolved tree is generated.

Figure 3. **Long-branch attraction in theory and in practice**.
Panels (**a**) and (**b**) show the 4-species case analyzed by Felsenstein[43]. If the correct tree (T_1 in **a**) has two long branches separated by a short internal branch, parsimony (as well as model-based methods such as likelihood and Bayesian methods under simplistic models) tends to recover a wrong tree (T_2 in **b**), in which the two long branches are grouped together. Panels (**c**) and (**d**) show a similar phenomenon in a real dataset, concerning the phylogeny of seed plants[134]. The Gnetales is a morphologically and ecologically diverse group of Gymnosperms including three genera (*Ephedra*, *Gnetum*, and *Welwitschia*), but its phylogenetic position has been controversial. Maximum likelihood analysis of 56 chloroplast proteins produced the GneCup tree (**d**), in which the Gnetales are grouped with Cupressophyta, apparently due to a long-branch attraction artefact. However, the Gnepine tree (**c**), in which the Gnetales joins the Pinaceae, was inferred by excluding the fastest-evolving 18 proteins as well as three proteins (psbC, rpl2 and rps7) that had experienced many parallel substitutions between the Cryptomeria branch and the branch ancestral to the Gnetales. The Gnepine tree (**c**) is also supported by two proteins from the nuclear genome and appears to be the correct tree. Branch lengths and bootstrap proportions are all calculated using RAxML. See ref. [134] for details.

Table 1. Functionalities of a few commonly-used phylogenetic programs

Name	Brief description	Link
Bayesian Evolutionary Analysis Sampling Trees (BEAST) [135]	A Bayesian MCMC program for inferring rooted trees under the clock or relaxed-clock models. It can be used to analyze nucleotide and amino acid sequences, as well as morphological data. A suite of programs, such as Tracer and FigTree, are also provided to diagnose, summarize and visualize results.	http://beast.bio.ed.ac.uk/
Genetic Algorithm for Rapid Likelihood Inference (GARLI) [55]	A program that uses genetic algorithms to search for ML trees. It includes the GTR+Γ model and special cases, and can analyze nucleotide, amino acid and codon sequences. A parallel version is also available.	http://code.google.com/p/garli/
Hypothesis testing using phylogenies (HYPHY)[136]	A ML program for fitting models of molecular evolution. It implements a high-level language that the user can use to specify models and set up likelihood ratio tests.	http://www.hyphy.org
Molecular Evolutionary Genetic Analysis (MEGA) [37]	A Windows-based program with a full graphical user interface that can be run under Mac OS X or Linux using Windows emulators. It includes distance, parsimony and likelihood methods of phylogeny reconstruction, although its strength lies in the distance methods. It incorporates the alignment program ClustalW, and can retrieve data from GenBank.	http://www.megasoftware.net
MRBAYES [71]	A Bayesian MCMC program for phylogenetic inference. It includes all the models of nucleotide, amino acid and codon substitution developed for likelihood analysis.	http://mrbayes.net
Phylogenetic Analysis by Maximum Likelihood (PAML) [137]	A collection of programs for estimating parameters and testing hypotheses using likelihood. It is mostly used for tests of positive selection, ancestral reconstruction, and molecular clock dating. It is not appropriate for tree searches.	http://abacus.gene.ucl.ac.uk/software/
Phylogenetic Analysis Using Parsimony *and other methods (PAUP* 4.0)	PAUP* 4.0 is still in beta version. It implements parsimony, distance and likelihood methods of phylogeny reconstruction.	http://www.sinauer.com/detail.php?id=8060
PHYLIP	A package of programs for phylogenetic inference by distance, parsimony, and likelihood methods.	http://evolution.gs.washington.edu/phylip.html

PHYML[53]	A fast program for searching for the ML trees using nucleotide or protein sequence data.	http://www.atgc-montpellier.fr/phyml/binaries.php
RAXML[54]	A fast program for searching for the ML trees under the GTR model using nucleotide or amino acid sequences. The parallel versions are particularly powerful.	http://wwwkramer.in.tum.de/exelixis/software.html
Tree analysis using New Technology (TNT)[42]	A fast parsimony program intended for very large data sets.	http://www.zmuc.dk/public/phylogeny/TNT

Note:— Essentially all programs can run on Windows, Mac OSX and Unix/Linux platforms. Except for PAUP*, which charges a nominal fee, all packages are free for download. See J. Felsenstein's comprehensive list of programs at http://evolution.genetics.washington.edu/phylip/software.html.
MCMC, Markov chain Monte Carlo; ML, Maximum likelihood; GTR (General Time Reversible) model.

Table 2. A summary of strengths and weaknesses of different tree-reconstruction methods

Strengths	Weaknesses
Parsimony methods	
1. Simplicity and intuitive appeal	1. Assumptions are implicit and poorly understood.
2. The only framework appropriate for some data (SINES, LINES, etc.)	2. Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution.
	2. Branch lengths are substantially underestimated when substitution rates are high.
	3. Maximum parsimony (MP) may suffer from long-branch attraction
Distance methods	
1. Fast computational speed	1. Most distance methods such as neighbour joining (NJ) do not consider variances of distance estimates.
2. Can be applied to any type of data as long as a genetic distance can be defined.	2. Distance calculation is problematic when sequences are divergent and involve many alignment gaps
3. Models for distance calculation can be chosen to fit data	3. Negative branch lengths are not meaningful.
Likelihood methods	
1. Can use complex substitution models to approach biological reality	1. Maximum likelihood (ML) iteration involves heavy computation.
2. Powerful framework for estimating parameters and testing hypotheses	2. The topology is not a parameter so that it is difficult to apply ML theory for its estimation. Bootstrap proportions are hard to interpret.
Bayesian methods	
1. Can use realistic substitution models, as in ML	1. Markov chain Monte Carlo (MCMC) involves heavy computation.
2. Prior probability allows the incorporation of information or expert knowledge	2. In large datasets, MCMC convergence and mixing problems can be hard to identify or rectify.
3. Posterior probabilities for trees and clades have easy interpretations	3. Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge.
	4. Posterior probabilities often appear too high.
	5. Model selection involves challenging computation [138,139].

Glossary terms

SYSTEMATICS. The inference of phylogenetic relationships among species and the use of such information to classify species.

TAXONOMY. The description, classification and naming of species.

COALESCENT. The process of joining ancestral lineages when the genealogical relationships of a random sample of sequences from a modern population are traced back.

GENE TREE. The phylogenetic or genealogical tree of sequences at a gene locus or genomic region.

STATISTICAL PHYLOGEOGRAPHY. The statistical analysis of population data from closely-related species to infer population parameters and processes such as population sizes, demography, migration patterns and rates.

SPECIES TREE. A phylogenetic tree for a set of species which underlies the gene trees at individual loci.

SYSTEMATIC ERROR. Error that is due to an incorrect model assumption. It is exacerbated when the data size increases.

RANDOM SAMPLING ERROR. Error or uncertainty in parameter estimates due to limited data.

CLUSTER ALGORITHM. An algorithm of assigning a set of individuals to groups (clusters) so that objects of the same cluster are more similar to each other than those from different clusters. Hierarchical cluster analysis can be *agglomerative* (starting with single elements and successively joining them into clusters) or *divisive* (starting with all objects and successively dividing them into partitions).

POSTERIOR PROBABILITY: See posterior distribution.

MARKOV CHAIN. A stochastic sequence (chain) of states having the property that given the current state, the probabilities for the next state do not depend on the past states.

TRANSITION. Substitution between the two pyrimidines (T ↔ C) or between the two purines (A ↔ G).

TRANSVERSION. Substitution between a pyrimidine and a purine (T, C ↔ A, G).

UNROOTED TREE. A phylogenetic tree for which the location of the root is unspecified.

LONG-BRANCH ATTRACTION. The phenomenon of inferring an incorrect tree with long branches grouped together by parsimony or by model-based methods under simplistic models.

MOLECULAR CLOCK. The hypothesis or observation that the evolutionary rate is constant over time or across lineages.

LIKELIHOOD RATIO TEST. A general hypothesis-testing method that uses the likelihood to compare two nested hypotheses, often using the χ^2 distribution to assess significance.

PRIOR DISTRIBUTION. The distribution assigned to parameters before the analysis of the data.

POSTERIOR DISTRIBUTION. The distribution of the parameters (or models) conditional on the data. It combines the information in the prior and in the data (likelihood).

MARKOV CHAIN MONTE CARLO (MCMC) ALGORITHM. A Monte Carlo simulation is a computer simulation of a biological process using random numbers. An MCMC algorithm is a Monte Carlo simulation algorithm that generates a sample from a target distribution (often a Bayesian posterior distribution).

GRAPHICAL PROCESSING UNITS (GPU). Specialized units traditionally used to manipulate output on a video display and recently explored for use in parallel computation.

CLADE: A group of species descended from a common ancestor.

Online summary

- The rapid accumulation of genome sequence data has made phylogenetics an indispensable tool to various branches of biology. However, it has also posed considerable statistical and computational challenges to data analysis.
- Distance, parsimony, likelihood, and Bayesian methods of phylogenetic analysis have different strengths and weaknesses. Although distance methods are good for large datasets of highly similar sequences, likelihood and Bayesian methods often have more power and are more robust, especially for inferring deep phylogenies.
- Assessing phylogenetic uncertainty remains a difficult statistical problem.
- Data partitioning may have an important influence on the phylogenetic analysis of genome-scale datesets.
- Systematic biases, such as long-branch attraction, may be more important than random sampling errors in the analysis of genomic-scale datasets.

Author biographies

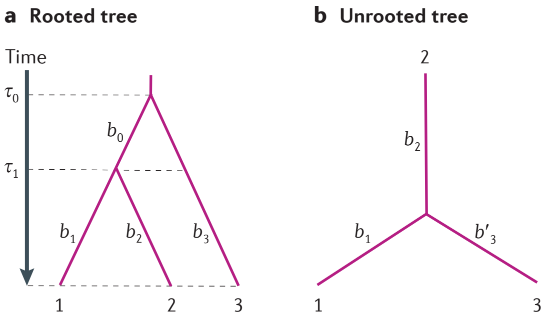
Ziheng Yang received his Ph.D. from Beijing Agricultural University in China. He currently holds the R.A. Fisher Chair of Statistical Genetics at University College London, UK. He is the Director of the Center for Computational and Evolutionary Biology at Institute of Zoology, Chinese Academy of Sciences in Beijing, China. His research focuses on the development of statistical methods and computer algorithms for phylogenetic and phylogeographic analyses of DNA sequence data.

Bruce Rannala received his Ph.D. from Yale University in Connecticut, USA. He is a Professor of Evolution and Ecology at the University of California, Davis, USA, and a guest professor at Institute of Zoology, Chinese Academy of Sciences in Beijing, China. His current research focuses on the estimation of recombination rates from genomic sequence data in relation to human disease mapping and on species delimitation using genomic data.

Online links
***Nature Reviews Genetics* article series on Study Designs:**
http://www.nature.com/nrg/series/studydesigns/index.html
A comprehensive list of phylogenetic programs maintained by Joe Felsenstein:
http://evolution.genetics.washington.edu/phylip/software.html

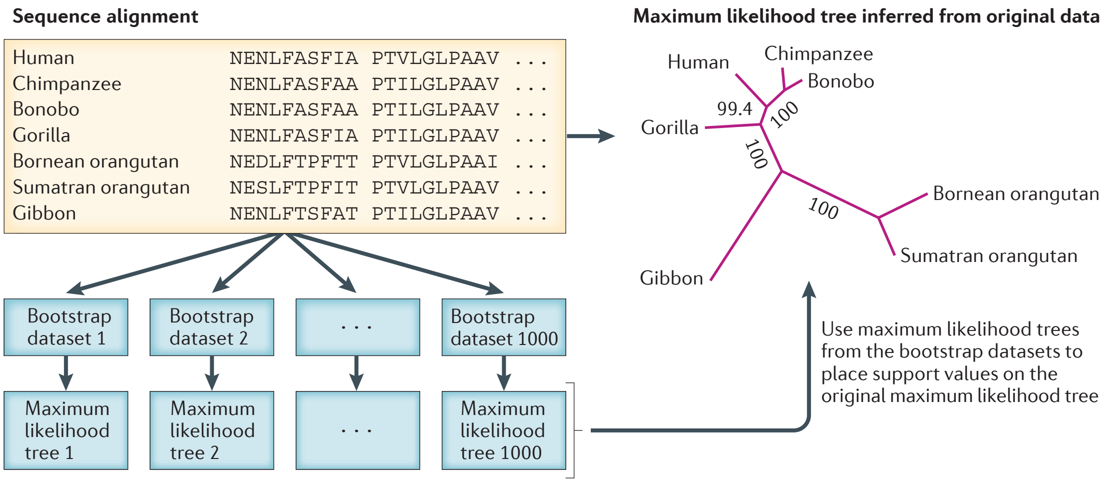
Subject categories: Evolutionary biology, Bioinformatics, Population genetics

Toc Blurb
Phylogenetic analysis is pervading every field of biological study. The authors review and assess the main methods of phylogenetic analysis — including parsimony, distance, likelihood and Bayesian methods — and provide guidance to selecting the most appropriate approach and software package.



Nature Reviews | Genetics

B1



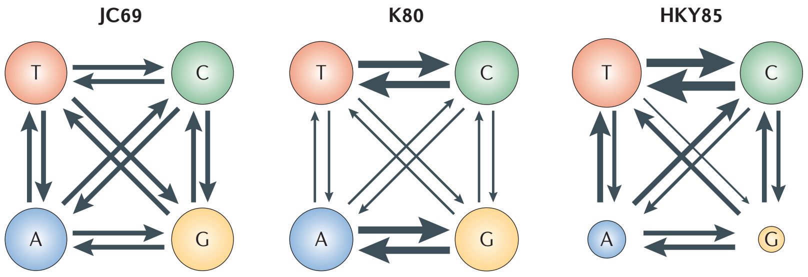
Sequence alignment

Human	NENLFASFIA PTVLGLPAAV	...
Chimpanzee	NENLFASFAA PTILGLPAAV	...
Bonobo	NENLFASFAA PTILGLPAAV	...
Gorilla	NENLFASFIA PTILGLPAAV	...
Bornean orangutan	NEDLFTPFTT PTVLGLPAAI	...
Sumatran orangutan	NESLFTPFIT PTVLGLPAAV	...
Gibbon	NENLFTSFAT PTILGLPAAV	...

Maximum likelihood tree inferred from original data

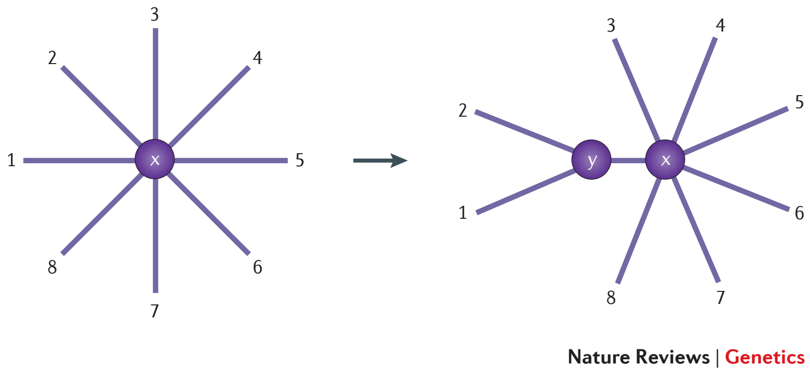
Use maximum likelihood trees from the bootstrap datasets to place support values on the original maximum likelihood tree

Nature Reviews | Genetics

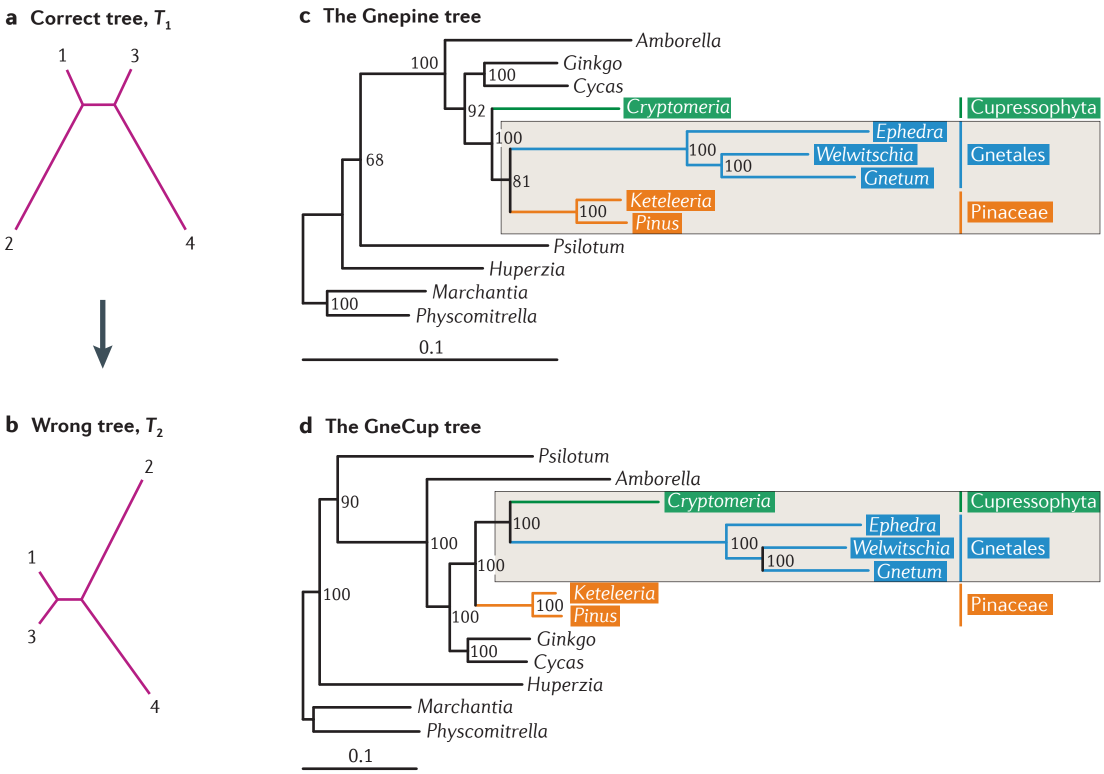


Nature Reviews | Genetics

F1



F2



Nature Reviews | Genetics