Phylogenetic tree building in the genomic age

Paschalia Kapli[1], *Ziheng Yang*[1], *Maximilian J Telford*[1*]

[1] **Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK**

***e-mail: m.telford@ucl.ac.uk**

Abstract

Knowing phylogenetic relationships among species is fundamental for many studies in biology. An accurate phylogenetic tree underpins our understanding of the major transitions in evolution such as the emergence of new body plans or metabolism and is key to inferring the origin of new genes, detecting molecular adaptation, understanding morphological character evolution, and reconstructing demographic changes in recently diverged species. While data are ever more plentiful and powerful analysis methods are available, there remain many challenges to reliable tree building. Here we discuss the major steps of phylogenetic analysis, including identification of orthologous genes or proteins, multiple sequence alignment and choice of substitution models and inference methodologies. Understanding different sources of errors and strategies to mitigate them is essential for assembling an accurate tree of life.

Introduction

Knowledge of how living (and extinct) species are related to one another underpins much of evolutionary biology. Knowing the relationships between species is an important goal in its own right and underlies our system of phylogenetic classification. The tree of life is also the essential framework for studying the origins of novel phenotypes and the processes that underlie biological evolution[1,2]. Mapping heritable character states (phenotypic or genotypic) onto a tree is the basis of different evolutionary analyses: it allows us, for example, to make inferences about character homology and also give insights into character loss and convergent evolution. Homologous characters of two taxa were, by definition, present in their common ancestor allowing us to infer the characteristics of these ancestors and more generally, character mapping allows us to follow the changing character states across a tree to reconstruct the historical path of evolution. Trees (and molecular data) also underpin methods for fitting a timescale to the evolutionary process and trees underlie the comparative method used to establish trends in the processes of evolution[2].

Reconstructing the relationships across all life, while prefigured in attempts at classification as long ago as Aristotle and Linneaus, is an endeavour that began seriously in the 19th century with Darwinism. While trees were initially based to a great extent on morphological characters, biological molecules — nucleic acids and proteins — provide a far more powerful and plentiful source of information for reconstructing trees [3]. Since DNA sequencing was developed and sequence data were first used for phylogenetics, our understanding of the tree of life has changed radically and huge progress has been made towards Darwin's dream of "very fairly true genealogical trees of each great kingdom of nature"[4].

For almost two decades, molecular phylogenies depended on data from one or a few genes, typically generated using PCR amplification and Sanger sequencing [5,6]. The development of new sequencing technologies has resulted in huge datasets containing numbers of genes that have increased by orders of magnitude[7]. The ease and low cost of genome and transcriptome sequencing have also meant that the number of taxa that can be considered is expanding massively, as manifest in recent proposals to sequence the genomes of all species on earth[8]. The data for reconstructing the tree of life are increasingly available but accurate tree reconstruction is not always straightforward.

In this Review we describe the major steps in the phylogenetic pipeline (Figure 1) involving hundreds or thousands of genes (the so-called phylogenomic approach). For every step, we outline the various methodological choices and several corresponding trade-offs between model sophistication and computational demands. We begin with the identification of orthologous genes (i.e. genes whose relationships will reliably reflect species relationships) from sets of genome or transcriptome sequences. We then discuss how to align orthologs from different species, to account for insertions and deletions, and strategies for trimming unreliably aligned regions. Finally, we discuss in detail the choice of inference methods and substitution models and consider potential errors as well as approaches to identify and avoid or mitigate them.

Generating databases of orthologous genes

The first years of molecular phylogenetics were dominated by studies using a small set of universal orthologous genes including the small and large subunit ribosomal RNAs (SSU and LSU rRNAs)[9] and (for eukaryote phylogenies) the mitochondrial genome[10]. The widespread use of rRNAs stemmed from the ease of PCR amplification using universal primers (unlike protein coding genes for which degenerate primers are required), the fact that orthology was clear amongst these universal genes, and the existence of a large database of these sequences.

The advances in high-throughput sequencing technologies of recent years mean that gene sequence data are abundant in sequence databases and new data are cheaply and easily produced. The challenges for data collection we are now faced with are: to ensure the data are free from contaminants; to identify orthologous genes that will reflect species relationships; and, ideally, to select those genes that are less prone to biases that may result in inaccurate trees[11].

Data compilation and preparation. Initial gene sequence data can be derived either from gene predictions based on genome sequences (even from draft-quality genomes) or from transcriptomes generated by sequencing libraries derived from mRNA[12] (Figure 1, A). An important part of this step is to identify and eliminate contamination (either by bacteria, commensals, parasites or gut contents or by cross contamination post DNA extraction)[13,14] (Figure 1, B). We start the description of the phylogenomic pipeline assuming the availability of gene sequences from each of the organisms of interest. Our ultimate aim is to produce an accurate tree of species relationships.

Orthology predictions. Two genes are homologous if they are inherited from an ancestral gene (Figure 2). Orthology is a special type of homology in which genes in different species have diverged from each other due to speciation[15,16]. As a result, orthologous genes recapitulate the relationships among the species they derive from (Figure 2, B). Other forms of homology include paralogy, in which genes from two species are derived from gene duplications deeper in time than the common ancestor of the two species (Figure 2, C), and xenology, where a gene in one species derives from a distantly related species through horizontal gene transfer (HGT). Paralogy and xenology do not reflect the relationships among species (Figure 2, C). Determining orthologous genes, therefore, is an essential step for reconstructing species phylogenies[16,17] (Figure 1, C). Gene duplications and losses in different lineages are common and may lead to paralogous relationships even among single-copy genes, posing a challenge to orthology identification.

Approaches for de novo identification of orthologs fall into two main categories: tree based and graph-based[18]. Tree-based orthology inference identifies orthologs by aligning homologous sequences and reconstructing a tree to find those which are most plausibly related by speciation rather than by duplication or HGT[19–21]. These methods are conceptually closest to the definition of orthology, but they are computationally expensive as they require both alignment and phylogenetic inference of entire gene families that often comprise hundreds of sequences. Deeper divergences pose a greater challenge to gene-tree inference as the phylogenetic signal erodes (i.e. multiple mutations accumulate resulting in homoplasy) and the risks systematic error increases (discussed in more detail in the later sections)[22,23]. Gene family relationships may be further obscured if other processes causing gene-tree discordance are not accounted for such as incomplete lineage sorting, horizontal gene transfer, hybridization, introgression and non-allelic gene conversion[22,24]. Particular groups of organisms are characterised by high occurrence of some of these processes, e.g., horizontal gene transfer in bacteria and hybridization, genome duplication and polyploidy in plants, which makes them more likely to suffer from orthology prediction errors.

Graph-based orthology inference methods[25–29] rely on the assumption that a gene in one species should be more similar to its ortholog than to any other gene in a second species and vice versa[30,31]. This concept of orthology gave rise to the most popular graph-based approach, the "bidirectional best hits" method[31] and several subsequent alternatives[25,32,33] . All such methods are based on all against all pairwise sequence comparisons mostly performed using Basic Local Alignment Search Tool (BLAST) for defining sequence similarity[34]. Graph-based approaches are not immune to the problems described for the tree-based ones, but they have the advantage of being computationally efficient and scaling well

with large datasets[27].

Given the complexity of sequence and gene evolution, *de novo* orthology prediction is bound to be approximate. It is encouraging that phylogenomic studies based on these procedures yield consistent and accurate phylogenies. Orthology prediction errors can, however, be a source of incongruence in challenging phylogenetic problems[27]. An alternative to *de novo* prediction is to use a set of reference orthologs and to identify their co-orthologs in newly sequenced species. Several dedicated databases offer orthologous sequences suitable for this cause, some spanning all domains of life (e.g., OrthoDB[29], OMA[35]) and others focussed on specific groups of organisms such as plants (Plaza[36]) and mammals (OrthoMam[37]). Several pipelines are available for automating this procedure (e.g., [38]) and there are two advantages in following this strategy, first it is computationally cheaper than *de novo* inference, and, second, it may alleviate errors associated with incomplete gene sampling. This is particularly relevant when using transcriptomic data, which usually contain only a subset of the genes. Incompleteness of data combined with differential gene loss may increase misidentification of paralogs for orthologs in *de novo* prediction. Using reference orthologous groups based on good quality genome data minimises this risk by ensuring the completeness of the gene repertoire for the group of interest[39].

Given that the identification of orthologs and the inference of a species phylogeny are intertwined, the hypothesis of orthology may also be tested at the same time as the species phylogeny. In particular, multi-copy gene data can be used simultaneously to estimate the species and gene-family evolution[40,41] Several methods have been described in a full Bayesian framework[41] as well as heuristic alternatives (e.g., [42–44]). Comparative assessment of the performance of these methods shows promising results[45].

A final consideration for orthology prediction is the genetic fragment that is used as unit. It is typical to use genes (entire or partial) as a means of identifying orthologous parts of a species' genome. However, most genes consist of multiple domains and, through time, their order and number may change (REF). In this context, it has been suggested that domains may be more suitable units for orthology and consequently for phylogenetic inference[46,47].

Alignment and trimming

Sequence alignment. Due to insertions and deletions (indels), genes and proteins typically differ in length between species, and, even in genes of identical length today, residues at the same location in a gene need not necessarily be homologous. Identifying homologous residues across genes entails aligning the genes, through the addition of gaps within the sequences[48] so that, in the final multiple sequence alignment (MSA), the residues in each column of the alignment should have descended from the same ancestral residue (Figure 1 D). Accurate alignment is fundamental in the inference of evolutionary relationships but, for genes in which indels have been frequent, it is a challenging task. When aligning protein-coding DNA sequences, the nucleotides naturally evolve as codon triplets rather than as single nucleotides. This property, as well as the fact that amino acid sequences change less rapidly than the corresponding nucleotides, means initial alignment at the protein rather than DNA level is usually appropriate. The codon triplets can then be aligned according to their corresponding amino acids[49–51] .

Alignment methods can be classified into three main categories. The most commonly used methods adopt the progressive approach, including Muscle[52], Clustal[53] and Mafft[54]. These methods first make a rough estimate of how similar each pair of sequences is and use this information to produce an approximate guide tree of relationships between sequences. They then build up the alignment by first aligning the most similar pair of sequences and progressively adding more distantly related sequences, according to the guide tree, to this fixed alignment.

Second are the consistency-based methods, including T-Coffee[55], ProbCons[56] and some versions of Mafft[54]. Initially, these estimate all pairwise alignments and, for each sequence pair, keep a record of alternative high-scoring solutions. Subsequently, they attempt to identify the overall alignment that maximizes the consistency among all pairs. Consistency-based methods are slower but overall more

accurate than progressive methods[57].

Finally, the most computationally expensive are the statistical or evolution-based methods such as BaliPhy[58] and StatAlign[59]. These assume an explicit evolutionary model of insertions and deletions[60] and jointly infer, in a Bayesian framework, both the alignment and the tree relating the sequences[58,59,61]. The statistical approach is the most methodologically sound, however, with large data sets it may become computationally demanding. In such cases, compromising with well performing heuristics such as PRANK[62] and Mafft. For deeper divergences in particular, versions of Mafft ("Mafft - **E-INS-i**" and "Mafft - **L-INS-i**") which accommodate the possibility of long internal or terminal gaps, respectively[54],) may be practical alternatives[50,63].

Filtering aligned putative orthologs. Any orthology identification procedure may falsely identify contaminants, paralogs or xenologs as orthologs. Such errors may have an effect on the accuracy of phylogenetic inference, for example, by yielding longer branches, biased model parameters or even changes to tree topology. To minimize this source of error, phylogenomic projects typically follow methods that aim to identify outlier sequences, often employing BLAST-based sequence comparisons[34] to test compatibility of closest neighbours with phylogenetic expectations[64,65](Figure 1 E). A true insect ortholog, for example, is expected to show higher similarity to homologs from bilaterian phyla than to those from non-bilaterians, and if such an assumption is not met, the sequence can be removed from the dataset. These protocols can be efficient in data sanitizing, but they typically require some knowledge of the phylogenetic relationships of the taxa involved.

Several tools are available that either automate such BLAST-based procedures (e.g. [64,66]) or use alternative approaches for outlier detection (e.g., PhyloMCoa is based on multiple co-inertia analysis[67]). Tools aiming to identify and eliminate sequences with characteristics that may be associated with systematic error[66,68] or low phylogenetic information also exist[65]. Finally, to enrich ortholog groups that might have been produced by too-stringent orthology prediction (i.e. leading to many false negatives), it is possible to use reference-based orthology prediction pipelines[38,64] under more relaxed criteria.

Alignment trimming. Alignment quality naturally decreases with increasing sequence divergence[69] . Because alignment errors may impact subsequent phylogenetic analyses[69,70] , it is common to filter ambiguously aligned regions (Figure 1 F). Filtering can be based on *ad hoc* criteria regarding alignment quality such as gappyness and sequence similarity[71–73] or by retaining only the alignment positions that are robust to changes to alignment parameters[74]. Reports on the impact of alignment trimming on the quality of downstream phylogenetic analysis vary[75,76], and hence trimming should be used cautiously.

Phylogenetic inference methods

Classification of phylogenetic inference methods.
Given a set of aligned and trimmed orthologous genes, there are two approaches to deriving a species tree. First, each of the gene alignments can be analysed independently to provide an estimate of the tree and the different trees can then be integrated to produce an estimate of the species tree. This is known as the super-tree approach. Second, the aligned genes can be concatenated into a supermatrix, which is analysed to produce a global estimate of the species tree. While we discuss methods for the reconciliation of multiple gene trees in the context genealogical heterogeneity across genes (below), the supermatrix method (Figure 1 G) is most commonly used and is the main focus of this Review.

Phylogeny reconstruction methods fall into two categories: distance-based and character-based. Distance methods involve calculating a genetic distance between every pair of species (based on comparison of their aligned sequences) and using the resulting distance matrix iteratively to construct a tree. The most popular distance method is the Neighbor Joining algorithm[77] (NJ). Because NJ does not search (according to a certain criterion) for the optimal tree in the huge space of all possible trees it is computationally very efficient. There are several implementations of the NJ method or variants[78] as well as versions capable of producing phylogenies of several thousands of samples[79,80]. Distance methods tend to perform poorly for distantly related species, however, because large distances are hard to estimate, and distance methods exacerbate this problem by summing up the branch lengths on the path between

species on the phylogeny when defining the pairwise distance.

Character based phylogenetic inference methods.
Character-based methods include maximum parsimony, maximum likelihood (ML) and Bayesian inference (BI)[81–83]. The maximum parsimony method calculates the minimum number of nucleotide or amino acid changes that are required to explain the data using each possible tree topology[84,85]. The tree topology with the smallest number of changes is known as the most parsimonious tree and is the estimate of the species phylogeny. For large datasets, exhaustive comparison of all possible trees is impossible (for 10 species there are 8.2×10^{21} possible rooted trees), and various heuristic tree searching approaches are typically used. Parsimony is attractive because of its mathematical simplicity and computational efficiency. Nevertheless, the method involves apparently unrealistic, implicit assumptions about the evolutionary process[86]. The lack of an explicitly stated model in the method makes it hard to incorporate well-known features of the process of sequence evolution, such as different rates between character states (e.g., different rates for transitions and transversions) and different rates among sites (e.g., higher rates at the third codon position than at the first and second). Parsimony is known to be more prone than likelihood methods to systematic errors including long branch attraction[87] (see below). The method is nevertheless useful for data types for which it is difficult to devise appropriate models of character evolution such as rare-event characters based on genome rearrangements or unique morphological characters.

In contrast to parsimony, both ML and BI methods are based on an explicitly stated model of sequence evolution and on the likelihood function. Under a statistical model parametrised by unknown parameter θ, the likelihood $L(\theta)$ is the probability of the observed data viewed as a function of θ. Here θ may include the parameters of the substitution model and the branch lengths on the tree. In phylogenetics, almost all models assume that different sites or columns in the alignment are independent; the likelihood is then the product of the probability of observing the data at the different sites. The likelihood contains all the information in the data concerning the unknown parameter under the model[88]. In other words, a parameter value that makes the observed data look highly likely to occur is expected to be closer to the truth than a parameter value that makes the data look nearly impossible. The ML estimate of the parameter is the parameter value that maximizes the likelihood. The ML method of tree estimation was introduced by Felsenstein[89] and has been implemented in programs such as PAML[90] PhyML[91], RAxML-NG[92], IQ-Tree[93] and FastTree[94](Table1). For each tree topology, the substitution parameters and branch lengths are optimized to maximize the likelihood, and the tree topology that achieves the highest likelihood is the ML tree.

The Bayesian method also relies on an explicitly stated model and on the likelihood function. It differs from ML in that it uses statistical distributions to quantify uncertainties in the parameters. Before the data are observed, the *prior* distribution is used to describe our prior information concerning the species tree and model parameters. After the data have been collected and analysed, the *posterior* distribution does the same thing. The posterior is the prior multiplied by the likelihood, rescaled so that it becomes a proper distribution. The posterior thus captures all information relevant for the parameters from the data and is an update of the prior.

The Bayesian method was introduced into molecular phylogenetics in the 1990s[95–97] and has been implemented in programs such as MrBayes[98], RevBayes[99], BEAST1[100],BEAST2[101] , and PhyloBayes[102,103] (Table 1). Computation in Bayesian phylogenetics is achieved using the Markov chain Monte Carlo (MCMC) algorithm, which is a computer simulation algorithm that generates a sample of the tree topologies and parameters from their posterior. In practical terms, the frequency with which the algorithm visits a given tree topology is an estimate of the posterior probability for that tree. The maximum posterior probability tree (or the MAP tree) is our best estimate of the true tree[95]. The 95% credible set of trees includes the most probable trees with the total posterior probability $\geq$95%; the credible set has the interpretation that the set includes the true tree with probability 95%, given the data and model[95,104]
.

A serious drawback of likelihood-based methods, including both ML and BI, is the heavy computational

demand, and they may take many thousands of CPU hours to run. This is particularly true of MCMC algorithms. Formulation of the likelihood function requires explicit specification of model assumptions concerning sequence evolution. This was considered by some as a disadvantage (because all models are wrong). However, it means that the assumed model can be tested, its impact on the analysis can be assessed and the model can be improved by incorporating important features of the evolutionary process. Indeed, most modern developments in statistical phylogenetics have been achieved in the likelihood framework[83,105].

Confidence in clades using the Bootstrap.
The NJ tree, parsimony tree or ML tree may be considered a point estimate of the true phylogeny from the respective methods. It is desirable to attach a measure of confidence in the point estimate as the confidence interval on a conventional parameter does. The most commonly used method for this purpose is bootstrapping, introduced to phylogenetics by Felsenstein[106]. This generates a number of bootstrap pseudo-datasets (say, 100), of the same size as the original dataset formed by resampling, with replacement of alignment sites. The pseudo-datasets are then analysed in the same way as the original dataset. The bootstrap support for a tree is the frequency at which that tree is inferred among the pseudo-datasets. The bootstrap is often used to attach support values for clades (as opposed to the whole tree): the support for a clade is the frequency at which the clade is recovered following phylogenetic tree reconstruction based on the bootstrap datasets. Unlike the bootstrap in other applications of statistics, the phylogenetic bootstrap does not have well-accepted or straightforward interpretations[107].

The bootstrap is applied to assess confidence in estimated trees for the distance, parsimony and maximum likelihood methods. For Bayesian methods, the posterior probabilities for trees and clades provide the natural measure of confidence so that bootstrap is unnecessary.

In analyses of phylogenomic datasets, a common observation is that bootstrap and posterior support values are very high (near 100%) whether the relationships are correct or not. This is particularly obvious for Bayesian posterior probabilities[108]. In phylogenomic-scale datasets random errors become unimportant, and such strong support for incorrect relationships typically derives from systematic errors.

We now review the most common and important sources of error in phylogenetic analysis of deep phylogenies. The reader may consult Felsenstein (2004)[81] and Yang[83] for more detailed discussions.

Accommodating phylogenetic errors

There are two kinds of errors in phylogenetic inference. Random errors are due to the dataset having a finite size (i.e., a limited number of sites in the alignment), whereas systematic errors are due to the violation of the model assumptions in the method[11] . In general, systematic errors arise when phylogenies are inferred under a simple homogeneous-process model of sequence evolution (assuming homogeneous rates of evolution between character states, among sites or genes, and across taxa or time) when in reality the process is heterogeneous. The explosive accumulation of sequence data in recent years means that random errors in phylogenetic analysis have been greatly reduced, but systematic errors actually increase with longer alignments.

Heterogeneity of rates across taxa and long branch attraction. Long Branch Attraction (LBA) is perhaps the best-known systematic error affecting phylogenetic reconstruction. At the root of LBA errors are unequal rates of evolution in different lineages; the resulting variance in the expected amount of change per lineage is represented by long branches (highly divergent sequences) and short branches (less divergent sequences) on a tree[87] LBA manifests itself as the incorrect grouping of long but, in reality, distantly related branches on the tree (Figure 3). Two unrelated long branches can experience occasional identical substitutions. Parsimony methods will reconstruct these convergences as a homologous shared character inherited from a common ancestor. Likelihood methods (ML and BI) are more robust to LBA errors than is parsimony, as they are branch length aware and hence take into account the increased possibility of convergence on two long branches. ML and BI can nevertheless suffer from LBA if the assumed substitution model is incorrect or too simplistic[109] such as wrongly assuming a

homogeneous rate of change across sites.

LBA may be hard to identify in empirical datasets. Its symptoms include two or more rapidly evolving lineages grouping together or a long-branch taxon joining a distant outgroup. It is then important to assess the robustness of such relationships to changes of the substitution model.

Several *ad hoc* strategies have been suggested to alleviate potential LBA artifacts, including exclusion of problematic species with very high evolutionary rates[6], removal of genes or gene regions with very high rates (which also tend to have poor alignment quality) and the addition of species which serve to break up long branches on the tree[110–112]. More recently, measures of branch length heterogeneity[66,68] have been used to identify genes that appeared less rate heterogeneous and which were therefore assumed to be less susceptible to LBA. In a similar spirit are methods for identifying and removing substantially longer branches from individual genes-trees thereby reducing rate heterogeneity[67,68,113].

Heterogeneity of nucleotide or amino acid compositions across taxa (Compositional Bias). Most phylogenetic inference models assume that the substitution process has been stationary throughout the history of the species under study and that all species therefore share the same frequencies of the 4 nucleotides or 20 amino acids. This assumption of compositional homogeneity is often violated in analysis of distantly related species, and an obvious example is when distantly related taxa have independently evolved Adenine/Thymine-rich genomes. In such a case, assumption of the homogeneous model will tend to artifactually group species with similar base compositions[114].

The optimal approach to dealing with compositional bias is to relax the assumption of compositional homogeneity by allowing the character state frequency parameters to drift across the phylogeny[115–117]. Such models involve a set of frequency parameters for every branch on the tree and resulting in a large number of parameters, with a high computational cost.

A more practical approach to circumventing this problem is to identify and remove from the analysis genes or taxa that show compositional bias[118]. There are several measures of compositional deviation that are available (e.g. in the software packages p4[116], IQtree and PhyloBayes). However, removing genes or taxa will not be possible if the most biased taxa are of central interest or if the majority of genes fails the homogeneity tests.

A final approach that has been proposed is to aggregate character states[119]. The four nucleotides can, for example, be recoded into pyrimidines (A and G) and purines (C and T) which removes any AT bias. Similarly, the 20 amino acids have been recoded into a reduced set, grouped according to their interexchangeability as represented in a substitution matrix[120]. The recoding naturally leads to information loss, which on its own may lead to topological changes. However, it can be informative to examine how the placement of compositionally divergent taxa changes when the data are recoded.

Heterogeneity of rates across sites. Different parts of the genome evolve at different rates. Collagens change more quickly than histones; introns change more quickly than exons; third positions in a codon change more quickly than first and second; and some amino acids within a protein are under strong stabilizing selection while others are free to vary. Ultimately, assuming a constant rate among sites of a gene is unrealistic. Assuming a single (average) rate results in a systematic underestimation of the likelihood of change at sites with higher rates[121]. As we have seen, underestimating the likelihood of change (and hence the probability of convergent evolution) tends to exacerbate long branch attraction. To accommodate this among-site rate variation, Yang (1993,1994)[121,122] proposed to model rates of sites as a random variable following a gamma distribution (Figure 4A). The resulting model is represented by a suffix '+Γ' or '+G' and can be combined with any nucleotide or amino acid substitution model (e.g., "JC69+Γ", "GTR+Γ", "LG+Γ"). This strategy for accounting for rate heterogeneity among sites is implemented in all phylogenetic inference and model-selection tools. Alternative models to accommodate among-site rate variation include the free-rates model (which assumes a few discrete rate classes)[123,124] and the gamma-mixture model (which assumes a mixture of two gamma distributions)[125]. In addition to heterogeneities across sites in an alignment, substitution rate and processes can also vary over time perhaps reflecting structural and functional changes in the proteins in different taxa[126]. As a consequence, the substitution rate and pattern at a given site may differ substantially among the lineages

of a phylogeny (Figure 5). This phenomenon is called "heterotachy"[127,128] and current methods for dealing with it are only computationally feasible for tree searching on very small datasets or for comparisons of single trees for larger data sets[129].

Heterogeneity of substitution patterns across sites — partition and mixture models. Different rates for different types of substitutions are easily accommodated in the Markov models used in phylogenetics. For example, transitions and transversions can be assigned distinct rates, with two parameters used[130]. The General Time Reversible (GTR) model assumes all nucleotides occur at different frequencies (i.e., three free model parameters) and change to one another at different rates (i.e. six exchangeability parameters).

For the 20 amino acids the GTR model will involve 209 parameters (19 frequencies and 190 exchangeabilities). This model is parameter-rich but can be fitted to moderately-sized datasets [Cite Yang et al. 1998]. However it is computationally expensive to estimate so many parameters during tree searching. Instead, empirical amino acid models derived from analysis of hundreds or thousands of protein sequences are more often used, including Dayhoff[131], JTT[132], WAG[133] and LG[134]. Empirical models have also been calculated based on specific subsets of proteins (e.g., viral[135], chloroplast[136] , and mitochondrial[137]). Different genes will fit different models best.

The common practice in a phylogenomic study has been to concatenate all genes into a super gene from which a single tree is inferred. Nevertheless, genes may differ in the rate and process of evolution. Such differences between genes may be accommodated by partition models that construct partitions with distinct parameters, such that sites in the same partition share evolutionary features and parameters whereas different partitions have distinct parameters[95]. Partition models provide a way of reducing errors from model misspecification by accounting for large-scale heterogeneity in rates and substitution patterns.

In a dataset of hundreds of genes and with dozens of models to choose from, it is not simple to assign models to genes or to construct a partitioning strategy. Automated model selection methods typically assume a fixed tree topology and try to maximize the likelihood of the data by altering the substitution models per gene (e.g.[93,138,139]). Some tools combine the process of model selection with the evaluation of alternative partition schemes, in which case genes that fit the same model are merged into one larger partition. For large datasets, the combined task of partition selection and model optimization is computationally intensive. Phylogenetic inference using empirical data under different substitution schemes may, however, result in differences in topology, branch lengths and statistical support[140,141]. Simulations show that optimised partitioning schemes are similar to partitioning based on biological common sense (e.g. by gene or by codon) and that both approaches are substantially better than unpartitioned data[141,142].

Mixture models. Mixture models may also accommodate among-site heterogeneity in substitution rates and patterns (Figure 4A and B). In a mixture model, instead of assigning each site to a specific partition, the model averages over all possible assignments of a site to the site classes. The gamma model of variable rates among sites discussed above is a typical mixture model. When biological knowledge is available to assign sites to well defined partitions (e.g., to assign sites of a gene to the three codon positions), it is natural to use partition models. When such knowledge is lacking, mixture models offer a flexible alternative.

In the analysis of protein data, different parts of a protein may have very different substitution rates, as well as having preferences for different amino acids dictated by local selective constraints. A one-size-fits-all empirical substitution matrix or even a partitioning approach is unlikely to capture these subtleties in the process of evolution. A mixture model may then be natural for accommodating the among-site heterogeneity in the rate and mode of amino acid substitution. A mixture model involves far more computation than a partition model, because without the knowledge of which component each site is from, one has to average over all components in the likelihood calculation (Figure 6).

Mixture models can be used to account for site heterogeneity in both the rate and pattern of substitution.

The model may assume multiple substitution matrices[143–145] or multiple sets of amino acid frequencies[146,147]. Profile models use multiple components that differ in the frequencies of the 20 amino acids, while assuming a single set of exchangeability rates among them[146,148–150]. The C10-C60[146] empirical models, for example, include empirically estimated amino acid frequencies from known protein sequences. These models are implemented both in a Bayesian[102] and a maximum likelihood framework[120,146,150] . The 'CAT' (categories) model, implemented in PhyloBayes[102], is the broadest generalization of the profile models. The CAT model treats the mixture components as free parameters and estimates the amino acid frequencies as well as the mixing proportions from the data (Figure 4B). Importantly, the CAT model and other mixture models appear to be much less prone to underestimating branch lengths and more robust against LBA artefacts in analyses of distantly related species than site-homogeneous models[151] .

Genealogical heterogeneity across genes

Concatenating all genes into a supermatrix and inferring a single tree assumes that one single gene tree underlies all genes and that it corresponds to the species tree. However, due to multiple biological processes — such as polymorphism in ancestral species, gene duplication and loss, and horizontal gene transfer — different genes or proteins may have different histories or gene trees[119,120].

Ancestral polymorphism means that orthologous genes from different species may not coalesce as soon as they reach the common ancestral species when we trace their history backwards in time; as a result, the genes may not track the species phylogeny and may have a different tree topology from the species tree (Figure 7). The phenomenon is variously termed incomplete lineage sorting (ILS), deep coalescence or gene-tree species-tree incongruence. Incongruence is more likely to occur if the interior branches of the species tree are short and if the ancestral species had large population sizes. Phylogenetic relationships represented by long interior branches in the species tree will most likely be resolved confidently even if the analytical method ignores ILS. However, for species that arose through a radiative speciation process (which generates short interior branches in the species tree), ILS may pose serious challenges to species tree estimation[152].

The framework for accommodating ILS is the multispecies coalescent (MSC[153,154]), an extension of the single-population coalescent[155] to the case of multiple species. Under the MSC model, the gene trees (topologies and branch lengths) vary among genes or genomic regions due to the coalescent process in the ancestral species: they have a statistical distribution specified by the species tree and by parameters such as the species divergence times and population sizes[156]. Thus, the MSC process is a natural consequence of reproduction and genetic drift. The simple MSC model has been extended to incorporate cross-species gene flow, leading to models such as MSC with migration (the isolation-with-migration or IM model[157–159] and MSC with introgression (the MSci or multispecies network coalescent or MSNC model)[160–162]. See [156,163,164] for recent reviews.

There are two major classes of species tree methods that incorporate the MSC model. The summary or two-step methods use phylogenetic programs to infer the gene trees for individual loci, and then use the estimated gene trees as data to construct the species tree. Popular two-step programs include ASTRAL[165] and MP-EST[166]. These methods are computationally efficient and can analyse thousands of genes but may suffer from errors in reconstructed gene trees.

In contrast, the full likelihood methods calculate the likelihood of the sequence alignments and therefore accommodate the uncertainties in the gene trees. Commonly used programs implementing the MSC model include *BEAST[167,168] and BP&P[169,170] . Both are MCMC algorithms[171] and involve heavy computation, although algorithmic improvements have made it possible to analyse datasets of 10,000 loci[162,172,173].

Analyses of both simulated and empirical data suggest that full likelihood methods are superior to the approximate coalescent methods and to concatenation[172–174]. A number of coalescent-based methods have been applied and evaluated in relatively shallow divergences, but the effectiveness of these methods in reconstructing deep phylogenies is poorly understood. However, the root cause of ILS is the short

internal branches in the species tree, rather than the shallowness of the nodes: deep phylogenies are just as affected by ILS as shallow phylogenies[152] . We expect that the next few years will see much effort in evaluating and overcoming the impact of ILS in deep parts of the tree of life.

Conclusions and perspectives

We have discussed a phylogenomic pipeline for accurate tree building, from careful data compilation including ortholog identification and contamination avoidance, via multiple sequence alignment, to selection of tree reconstruction methods and substitution models to avoid systematic errors in phylogenetic reconstruction. For challenging phylogenies — in particular deep phylogenies involving distant species — the choice of likelihood-based methods and the selection of adequate models to accommodate heterogeneities in the process of molecular evolution across sites, taxa and time (Figure 3, 4 and 5) appears to be as important as the generation of the underlying data. Here we discuss a few areas in phylogenetic research that may see progress in the next few years.

One approach that has so far received little attention is the development of computationally tractable models for accommodating heterogeneity across clades. Besides compositional bias, amino acid exchangeabilities have also been reported to vary across the tree of life[175,176]. The strategy adopted to address this issue so far has been to remove data (taxa or genes) or to attempt to reduce other related problems such as among-site heterogeneity. Nevertheless, directly modelling tree heterogeneity should provide more accurate tree estimates.

Species radiations and the resulting short branches in the species phylogeny are responsible for many of the challenges in resolving the tree of life. This is particularly true for species radiations in deep time (examples within the animal kingdom include the divergences of mammals and birds, and the spirally cleaving phyla within the Lophotrochozoa). With deep radiations, the problem of ILS is exacerbated by the erosion of phylogenetic signal resulting from substitutional saturation on the terminal branches. The performance of the MSC methods in deep divergences, when the molecular clock is seriously violated, needs careful study. Recent work shows that existing approximate methods may be vulnerable to LBA artifacts[177], and research is needed to evaluate the performance of coalescent approaches under relaxed-clock models in inference of deep divergences.

Phylogenomic datasets pose enormous computational burdens, in particular, when complex models (for example, heterogeneous or MSC models) are used to reduce systematic errors. Great progress has been made in computational phylogenetics by algorithms for speeding up[178] and parallelizing[179] the likelihood calculation, and the implementation of software making use of modern multi-processor multi-core computer architecture[92,102,180–182]. Additional improvements have been achieved in the mixing efficiency of MCMC sampling methods in Bayesian inference[183–185]. Yet there appears to be much room for further improvement in the computational efficiency of those algorithms. Such advances will enhance the biological realism of phylogenomic models and will improve overall phylogenetic accuracy.

Figure legends

Figure 1 | **Phylogenomic Pipeline**. **A** | The starting material is a set of gene sequences (typically translated protein sequences) predicted from a genome sequence or derived from transcriptome sequencing. **B** | Contamination from commensals/symbionts, parasites, gut contents in animals, environmental sources, or experimental errors, especially in multiplexed transcriptome sequencing, must be identified and removed. Contaminants can be identified and excluded on the basis of GC content of sequences, read coverage and taxonomy of sequence similarity matches. **C** | All-against-all comparisons (BLAST or similar) are used to identify sequences that are homologous between all species of interest. Clustering algorithms are used to identify putative orthologous genes whose relationships should reflect the species phylogeny. **D** | The sequences of putative orthologs are aligned to generate a multiple sequence alignment (MSA). **E** | The MSA can be analysed to produce an initial phylogenetic tree for the putative orthologs, which can be used to identify paralogs, contaminants and other

problematic sequences indicated by unusually long branches. **F** | The MSA is typically filtered to remove regions of unreliable alignment. **G** | The orthologs are concatenated to produce a super-matrix, which is analysed to infer the species phylogeny. Different models (or independently estimated substitution parameters under the same general model) may be used for different partitions of the MSA.

Figure 2 | **Distinguishing orthologous and paralogous relationships between genes.** **A** | A gene has duplicated in a common ancestor of the species of interest. Two paralogous copies (red and grey) now evolve independently and each is inherited by descendent taxa following speciation events. **B** | Each of the duplicated genes (red or grey) have orthologous relationships amongst themselves such that reconstructing the relationships using just red or just grey orthologs will result in a tree that reflects the species relationships. **C** | Red and grey copies are related by duplication so that a tree based on a mixture of red and grey genes will not reflect the correct branch lengths of the species tree (left, the asterisk denotes the part of the branch length that corresponds to the time among the duplication and the speciation events) and can also result in an incorrect species tree topology (right).

Figure 3 | **Heterogeneous rates across lineages and long branch attraction.** Heterogenous rates of substitution across lineages, if not accommodated by the model, may result in a Long Branch Attraction (LBA) artefact. **A** | The upper tree is the true tree relating four species, in which substitution rates are heterogeneous among taxa, with long branches reflecting rapid changes along the lineages. The lower tree shows the effect of interpreting the occasional convergent changes arising in the long branches as shared characters indicating a close relationship between the long branches. This erroneous tree is inferred using Maximum Parsimony (MP), whereas branch-length-aware likelihood methods such as Maximum Likelihood (ML) and Bayesian Inference (BI) are less prone to this error. **B** | We used the true tree of panel A to simulate 1000 replicate datasets (sequence alignments) of increasing length (with 50-10,000 sites) under the Jukes-Cantor model. We analysed each replicate dataset using ML (blue) and MP (orange) and recorded whether the correct (solid lines) or LBA tree (dotted lines) was recovered. For ML, small data sets show errors due to small sample size (stochastic errors) which decrease with larger samples. For MP the systematic errors caused by LBA become larger with increasing sample size.

Figure 4 | **Heterogeneous substitution rates and patterns across sites.** **A** | When different sites have different substitution rates (in different colours in the multiple sequence alignment (MSA) at the bottom), more mutations (black circles) are accumulated at fast evolving sites, resulting in more homoplasy (convergent substitutions independently acquired by unrelated taxa, black stars). A model that assumes homogenous rates across sites will lead to underestimation of the amount of change at the fast evolving sites and underestimation of the likelihood of convergence. Such systematic errors can lead to the erroneous Long Branch Attraction (LBA) tree, whereas assuming the heterogeneous model incorporating variable rates across sites recovers the true tree. **B** | When different sites in the protein prefer different amino acids (e.g., with hydrophobic or hydrophilic amino acids shown in different colours in the MSA and corresponding amino acid frequency bar charts), rates of change within each composition category (e.g. amongst hydrophobic amino acids) are higher than the average rate across the whole alignment. The homogeneous model ignoring among-site composition heterogeneity tends to underestimate the amount of change expected for sites with restricted compositions and to underestimate the likelihood of convergence, resulting in the erroneous LBA tree. The heterogeneous model incorporating among-site composition variation recovers the true tree.

Figure 5 | **Heterogeneities across time or lineages.** The first half of the sites in species A and C evolve faster than those in species B and D but the opposite is true for the second half. Such a heterogeneous substitution process is called "heterotachy"[144]□ When

heterotachy is ignored, the tree shown on the right will be inferred, with A and C erroneously joined together.

Figure 6 | **Homogeneous, partition & mixture models**. All the three types of models assume that sites evolve independently so that the likelihood (or the probability of all data) is the product of the probabilities for different sites, i.e., $L = \prod_{i=1}^{s} p_i$. **A** | The site-homogeneous model assumes the same substitution rate and process for all the sites in the alignment. The probability of each site p_i is calculated under the shared model. **B** | In a partitioned model, each site is assigned to a partition, with sites in the same partition evolving according to the same model whereas different partitions have different models or mode parameters. Therefore, the probability of observing each site i, is calculated under the model it is assigned to. **C** | In a mixture model the sites in the alignment are a mixture of m classes, but we do not know a priori which site class each site is from. The probability of observing data at a site i is then an average over the m site classes $p_i = \sum_{k=1}^{m} w_k \, P(X_i M_k)$, where w_k is the proportion for site class k.

Figure 7 **| Gene-tree Species-tree Incongruence** Ancestral polymorphism and deep coalescence may cause the most common gene tree to have a different topology from the species tree. The species tree is then said to be in the anomaly zone and the gene trees are called anomalous gene trees[123]□ Anomaly zones do not exist for three species but exist for four or more species. Here the four species A, B, C, and D arose through rapid speciation events, with the speciation times τ_{AB}, τ_{ABC} and τ_{ABCD} nearly equal. When we sample one sequence from each species and trace the genealogical history of the four sequences a, b, c, and d, there will be little chance for coalescence events to happen in the ancestral species AB or ABC because of the very short time interval. All four sequences are very likely to trace back to the common ancestor ABCD, in which coalescent events occur in random order. As a result, each sequence of coalescent events will occur with the same probability. Each sequence of coalescent events generates a unique labelled history or ranked gene tree, which is a rooted tree with internal nodes ordered by age. There are 18 possible ranked gene trees and each has probability 1/18. Thus the unbalanced gene tree G_1, which matches the species tree, has probability 1/18 as it represents one sequence of coalescent events (a and b joining first, then their ancestor joining with c, then their ancestor joining with d). The balanced gene tree ((ab),(cd)), which does not match the species tree, has probability 2/18, as it represents two distinct sequences of coalescent events or two labelled histories: G_2 in which $t_{ab} < t_{cd}$ and G_3 in which $t_{ab} > t_{cd}$. When the speciation times (τ_{AB}, τ_{ABC} and τ_{ABCD}) are close but not exactly the same, the mismatching balanced gene tree may still have a probability greater than for the matching unbalanced gene tree, even if not twice as large. For such speciation times, the species tree of (A) is in the anomaly zone. In the anomaly zone, the majority-vote method of species tree estimation, which uses the most common gene tree as the species tree estimate, is inconsistent and will approach a wrong species tree when more and more loci or gene trees are used. The problem of the anomaly zone (and in general the problem of incomplete lineage sorting) is more serious the shorter the internal branches in the species tree are or the larger the ancestral populations are.

Table 1 | **Features of different orthology prediction and sequence alignment programs (with references and links to software)**

Table 2 | **Features of different tree reconstruction programs (with references and links to software)**

Glossary

HOMOLOGY – HOMOLOGS

Features, including morphological characters and gene loci, inherited from a common ancestor e.g. a gene in two species originating from a single ancestral gene.

ORTHOLOGY – ORTHOLOGS

Homologous sequences that have diverged due to speciation events

PARALOGY – PARALOGS

Homologous sequences that have diverged due to duplication events so that both copies have descended side by side during the history of an organism.

XENOLOGS

Homologous sequences originating from lateral gene transfer.

TOPOLOGY

The branching pattern of a phylogenetic tree indicating relationships between taxa.

CLADE

A group of taxa on a tree that includes their most recent common ancestor and all its descendants, also known as a monophyletic group.

ALIGNMENT

Insertion of gaps in homologous sequences so that nucleotides or amino acids in the same column are homologous.

SUBSTITUTION MODEL

Continuous time Markov Chain probabilistic models that describe changes between nucleotides or amino acids over evolutionary time.

RATE VARIATION ACROSS SITES

The phenomenon where different sites of a gene sequence evolve at different rates.

RATE HETEROGENEITY ACROSS TAXA

The phenomenon where different taxa evolve at different rates.

HOMOGENEOUS MODEL

A model that assumes the same substitution rate or process across alignment sites, taxa and time.

MIXTURE MODEL

A model that assumes different substitution rates or processes across sites the alignment.

PROFILE MIXTURE MODEL

A model that assumes multiple sets of state frequencies for sites (e.g. CAT, C10-C60).

STOCHASTIC ERROR

Error due to the finite length of sequences in alignment.

SYSTEMATIC ERROR

Errors due to incorrect model assumptions.

LONG-BRANCH ATTRACTION

The phenomenon of inferring an incorrect tree in which taxa with long branches are grouped together.

COMPOSITIONAL HETEROGENEITY

Heterogeneity in nucleotide or amino-acid frequencies across lineages of a phylogeny.

SPECIES TREE

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

GENE TREES

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

INCOMPLETE LINEAGE SORTING

Discordance of gene-trees from the species tree due to ancestral polymorphism.

COALESCENT

The process of lineage joining when one traces the history of a sample of sequences backwards in time.

GENETIC DRIFT

The process of random changes in allele frequencies over generations due to the stochastic nature of reproduction.

References

1. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**, 361–375 (2005).

2. Telford, M. J. & Budd, G. E. The place of phylogeny and cladistics in Evo-Devo research. *Int. J. Dev. Biol.* **47**, 479–490 (2003).

3. Fitch, W. M. & Margoliash, E. Construction of phylogenetic trees. *Science (80-.).* **155**, 279–284 (1967).

4. Darwin, C. R. Darwin Correspondence Project, 'Letter no. 2143'.

5. Field, K. G. *et al.* Molecular phylogeny of the animal kingdom. *Science (80-.).* **239**, 748–753 (1988).

6. Aguinaldo, A. M. A. *et al.* Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387, 489–493 (1997).

Classic paper on LBA that shows the benefit of excluding long branch taxa.

7. Telford, M. J., Budd, G. E. & Philippe, H. Phylogenomic insights into animal evolution. *Curr. Biol.* **25**, R876–R887 (2015).

8. Lewin, H. A. *et al.* Earth BioGenome project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).

9. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5088–5090 (1977).

10. Kocher, T. D. *et al.* Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6196–6200 (1989).

11. Philippe, H. & Telford, M. J. Large-scale sequencing and the new animal phylogeny. *Trends Ecol. Evol.* **21**, 614–620 (2006).

12. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinforma.* **65**, e57 (2019).

13. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).

14. Simion, P. *et al.* A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biol.* **16, 28 (2018).

Identifies cross contamination between multiplexed sequence samples as frequent occurrence and provides the means to detect this source of error.

15. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19, 99–113 (1970).

Original paper defining different forms of homology.

16. Kristensen, D. M., Wolf, Y. I., Mushegian, A. R. & Koonin, E. V. Computational methods for gene orthology inference. *Brief. Bioinform.* **12**, 379–391 (2011).

17. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).

18. Trachana, K. *et al.* Orthology prediction methods: A quality assessment using curated protein families. *BioEssays* **33**, 769–780 (2011).

19. Li, H. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006).

20. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M. & Gabaldón, T. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **42**, D897–D902 (2014).

21. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377–D386 (2013).

22. Glover, N. *et al.* Advances and Applications in the Quest for Orthologs. *Mol. Biol. Evol.* **36**, 2157–2164 (2019).

23. Boeckmann, B. *et al.* Quest for Orthologs Entails Quest for Tree of Life: In Search of the Gene Stream. *Genome Biol. Evol.* **7**, 1988–1999 (2015).

24. Harpak, A., Lan, X., Gao, Z. & Pritchard, J. K. Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 12779–12784 (2017).

25. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

26. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

27. Altenhoff, A. M. *et al.* OMA standalone: Orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **29**, 1152–1163 (2019).

28. Kaduk, M., Riegler, C., Lemp, O. & Sonnhammer, E. L. L. HieranoiDB: A database of orthologs inferred by Hieranoid. *Nucleic Acids Res.* **45**, D687–D690 (2017).

29. Kriventseva, E. V. *et al.* OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).

30. Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 10268–10273 (1996).

31. Overbeek, R., Fonstein, M., D'Souza, M., Push, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2896–2901 (1999).

32. Wall, D. P., Fraser, H. B. & Hirsh, A. E. Detecting putative orthologs. *Bioinformatics* **19**, 1710–1711 (2003).

33. Dessimoz, C., Boeckmann, B., Roth, A. C. J. & Gonnet, G. H. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.* **34**, 3309–3316 (2006).

34. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

35. Altenhoff, A. M. *et al.* The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.* **46**, D477–D485 (2018).

36. Van Bel, M. *et al.* PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.* **46**, D1190–D1196 (2018).

37. Scornavacca, C. *et al.* OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes. *Mol. Biol. Evol.* **36**, 861–862 (2019).

38. Petersen, M. *et al.* Orthograph: A versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics* **18**, (2017).

39. Kuzniar, A., van Ham, R. C. H. J., Pongor, S. & Leunissen, J. A. M. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* **24**, 539–551 (2008).

40. Szöllősi, G. J., Tannier, E., Daubin, V. & Boussau, B. The inference of gene trees with species trees. *Syst. Biol.* **64**, e42–e62 (2015).

41. Boussau, B. *et al.* Genome-scale coestimation of species and gene trees. *Genome Res.* (2013) doi:10.1101/gr.141978.112.

42. Wehe, A., Bansal, M. S., Burleigh, J. G. & Eulenstein, O. DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* **24**, 1540–1541 (2008).

43. Bansal, M. S., Burleigh, J. G. & Eulenstein, O. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. *BMC Bioinformatics* **11**, S42 (2010).

44. Chaudhary, R., Burleigh, J. G. & Fernández-Baca, D. Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol. Biol.* **28**, 8 (2013).

45. Chaudhary, R., Boussau, B., Burleigh, J. G. & Fernández-Baca, D. Assessing approaches for inferring species trees from multi-copy genes. *Syst. Biol.* **64**, 325–339 (2015).

46. Scornavacca, C. & Galtier, N. Incomplete lineage sorting in mammalian phylogenomics. *Syst. Biol.* **66**, 112–120 (2017).

47. Sonnhammer, E. L. L. *et al.* Big data and other challenges in the quest for orthologs. *Bioinformatics* **30**, 2993–2998 (2014).

48. Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244 (1988).

49. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).

50. Dessimoz, C. & Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.* **11**, R37 (2010).

51. Hall, B. G. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol. Biol. Evol.* **22**, 792–802 (2005).

52. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

53. Sievers, F. & Higgins, D. G. Clustal Omega. *Curr. Protoc. Bioinforma.* **48**, 3–13 (2014).

54. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537**, 39–64 (2009).

55. Notredame, C., Higgins, D. G. & Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).

56. Do, C. B., Mahabhashyam, M. S. P., Brudno, M. & Batzoglou, S. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330–340 (2005).

57. Chatzou, M. *et al.* Multiple sequence alignment modeling: Methods and applications. *Brief. Bioinform.* **17**, 1009–1023 (2016).

58. Suchard, M. A. & Redelings, B. D. BAli-Phy: Simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics* **22**, 2047–2048 (2006).

59. Novák, Á., Miklós, I., Lyngsø, R. & Hein, J. StatAlign: An extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics* **24**, 2403–2404 (2008).

60. Thorne, J. L., Kishino, H. & Felsenstein, J. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114–124 (1991).

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

62. Löytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (80-.).* **320**, 1632–1635 (2008).

63. Vialle, R. A., Tamuri, A. U. & Goldman, N. Alignment modulates ancestral sequence reconstruction accuracy. *Mol. Biol. Evol.* **35**, 1783–1797 (2018).

64. Simion, P. *et al.* A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **27**, 958–967 (2017).

65. Philippe, H. *et al.* Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. *Curr. Biol.* **29**, 1818–1826 (2019).

66. Struck, T. H. Trespex-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol. Bioinforma.* **10**, 51–67 (2014).

67. De Vienne, D. M., Ollier, S. & Aguileta, G. Phylo-MCOA: A fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol. Biol. Evol.* **29**, 1587–1598 (2012).

68. Mai, U. & Mirarab, S. TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* **19**, 272 (2018).

69. Ogden, T. H. & Rosenberg, M. S. Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.* **55**, 314–328 (2006).

70. Fletcher, W. & Yang, Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267 (2010).

71. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).

72. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

73. Misof, B., Katharina, M., Misof, B. & Katharina, M. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst. Biol.* **58**, 21–34 (2009).

74. Moretti, S. *et al.* The M-Coffee web server: A meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucleic Acids Res.* **35**, W645–W648 (2007).

75. Tan, G. *et al.* Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst. Biol.* **64**, 778–791 (2015).

76. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).

77. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

78. Gascuel, O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).

79. Saitou, N. *Introduction to Evolutionary Genomics*. (Springer, 2018). doi:10.1007/978-3-319-92642-1.

80. Wheeler, T. J. Large-scale Neighbor-Joining with NINJA. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and*

Lecture Notes in Bioinformatics) 375–389 (2009). doi:10.1007/978-3-642-04241-6_31.

81. Felsenstein, J. *Inferring phylogenies*. (Sinauer associates Sunderland, MA, 2004).

82. Yang, Z. & Rannala, B. Molecular phylogenetics: Principles and practice. *Nat. Rev. Genet.* **13**, 303–314 (2012).

83. Yang, Z. *Molecular Evolution A Statistical Approach*. (Oxford University Press, 2014).

84. Fitch, W. M. Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Biol.* **20**, 406–416 (1971).

85. Hartigan, J. A. Minimum Mutation Fits to a Given Tree. *Biometrics* (1973) doi:10.2307/2529676.

86. Felsenstein, J. Parsimony in systematics: biological and statistical issues. *Annu. Rev. Ecol. Syst.* **14**, 313–333 (1983).

87. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.* **27, 401–410 (1978).

Clear explanation and demonstration of the effects of Long Branch Attraction.

88. Stuart, A., Ord, K. & Arnold, S. *Kendall's Advanced Theory of Statistics. The Statistician* (Arnold, 1999). doi:10.2307/2348968.

89. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).

90. Yang, Z. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).

91. Guindon, S. *et al.* PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

92. Kozlov, A. M. *et al.* RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz305.

93. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

94. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, 3 (2010).

95. Rannala, B. & Yang, Z. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* **43, 304–311 (1996).

Introduces Bayesian methods to phylogenetics.

96. Li, S., Pearl, D. K. & Doss, H. Phylogenetic tree construction using Markov chain monte carlo. *J. Am. Stat. Assoc.* **95**, 493–508 (2000).

97. Mau, B. & Newton, M. A. Phylogenetic Inference for binary data on dendograms using markov chain monte carlo. *J. Comput. Graph. Stat.* **6**, 122–131 (1997).

98. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).

99. Höhna, S. *et al.* RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **25**, 726–736 (2016).

100. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, (2018).

101. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, 1–28 (2019).

102. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. Phylobayes mpi: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).

103. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25, 2286–2288 (2009).

Implementation of the CAT (Categories) model that accommodates site heterogenous evolution in a Bayesian framework.

104. 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).

105. Chen, M.-H., Kuo, L. & Lewis, P. *Bayesian phylogenetics: methods, algorithms, and applications*. (2014).

106. Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution (N. Y).* **39**, 783 (1985).

107. Susko, E. Bootstrap support is not first-order correct. *Syst. Biol.* **58**, 211–223 (2009).

108. Yang, Z. & Zhu, T. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1854–1859 (2018).

109. Huelsenbeck, J. P. Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48 (1995).

110. Baurain, D., Brinkmann, H. & Philippe, H. Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors? *Mol. Biol. Evol.* **24**, 6–9 (2007).

111. Rodréguez-Ezpeleta, N. *et al.* Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* **56**, 389–399 (2007).

112. Brinkmann, H., Van Der Giezen, M., Zhou, Y., De Raucourt, G. P. & Philippe, H. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.* **54**, 743–757 (2005).

113. Rivera-Rivera, C. J. & Montoya-Burgos, J. I. LS3: A method for improving phylogenomic inferences when evolutionary rates are heterogeneous among taxa. *Mol. Biol. Evol.* **33**, 1625–1634 (2016).

114. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11**, 605–612 (1994).

115. 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).

116. Foster, P. G. Modeling compositional heterogeneity. *Syst. Biol.* **53, 485–495 (2004).

Method to detect compositional heterogeneity in sequence alignments.

117. Blanquart, S. & Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23**, 2058–2071 (2006).

118. Nesnidal, M. P., Helmkampf, M., Bruchhaus, I. & Hausdorf, B. Compositional heterogeneity and phylogenomic inference of metazoan relationships. *Mol. Biol. Evol.* **27**, 2095–2104 (2010).

119. Phillips, M. J. & Penny, D. The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol. Phylogenet. Evol.* **28**, 171–185 (2003).

120. Susko, E., Lincker, L. & Roger, A. J. Accelerated estimation of frequency classes in site-heterogeneous profile mixture models. *Mol. Biol. Evol.* **35**, 1266–1283 (2018).

121. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).

122. Yang, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10, 1396–1401 (1993).

Introduces the Gamma distribution to model rate heterogeneity across sites.

123. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005 (1995).

124. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

125. Mayrose, I., Friedman, N. & Pupko, T. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics* **21**, 151–158 (2005).

126. Fitch, W. M. & Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–593 (1970).

127. Philippe, H. & Lopez, P. On the conservation of protein sequences in evolution. *Trends Biochem. Sci.* **26**, 414–416 (2001).

128. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19, 1–7 (2002).

Introduces the process of heterotachy and effects on tree reconstruction.

129. Zhou, Y., Rodrigue, N., Lartillot, N. & Philippe, H. Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evol. Biol.* **7**, (2007).

130. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

131. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. in *Atlas of protein sequence and structure* 345–352 (1978).

132. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275–282 (1992).

133. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).

134. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).

135. Dang, C. C., Le, Q. S., Gascuel, O. & Le, V. S. FLU, an amino acid substitution model for influenza proteins. *BMC Evol. Biol.* **10**, (2010).

136. Adachi, J., Waddell, P. J., Martin, W. & Hasegawa, M. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.* **50**, 348–358 (2000).

137. Rota-Stabelli, O., Yang, Z. & Telford, M. J. MtZoa: A general mitochondrial amino acid substitutions model for animal evolutionary studies. *Mol. Phylogenet. Evol.* **52**, 268–272 (2009).

138. Darriba, D. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* (2019).

139. Morel, B., Kozlov, A. M. & Stamatakis, A. ParGenes: A tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* **35**, 1771–1773 (2019).

140. Hoff, M., Orf, S., Riehm, B., Darriba, D. & Stamatakis, A. Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics* **17**, 143 (2016).

141. Kainer, D. & Lanfear, R. The effects of partitioning on phylogenetic inference. *Mol. Biol. Evol.* **32**, 1611–1627 (2015).

142. Darriba, D. & Posada, D. The impact of partitioning on phylogenomic accuracy. *bioRxiv* 023978 (2015).

143. Goldman, N., Thorne, J. L. & Jones, D. T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458 (1998).

144. Le, S. Q., Dang, C. C. & Gascuel, O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **29**, 2921–2936 (2012).

145. Le, S. Q. & Gascuel, O. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst. Biol.* **59**, 277–278 (2010).

146. Le, S. Q., Lartillot, N. & Gascuel, O. Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 3965–3976 (2008).

147. Wang, H. C., Li, K., Susko, E. & Roger, A. J. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evol. Biol.* **8**, 331 (2008).

148. Halpern, A. L. & Bruno, W. J. Evolutionary distances for protein-coding sequences: Modeling site- specific residue frequencies. *Mol. Biol. Evol.* **15**, 910–917 (1998).

149. 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).

Introduces the CAT model to accommodate site heterogeneity.

150. Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67, 216–235 (2018).

Approximate site heterogenous models for Maximum Likelihood framework applicable to large data sets.

151. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7**, S4 (2007).

152. Edwards, S. V. Is a new and general theory of molecular systematics emerging? *Evolution (N. Y).* **63**, 1–19 (2009).

153. 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).

**154. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* vol. 24 332–340 (2009).

Introduces the discordance between gene trees and species trees due to incomplete lineage sorting.

155. Kingman, J. F. C. The coalescent. *Stoch. Process. their Appl.* **13**, 235–248 (1982).

156. Xu, B. & Yang, Z. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* **204**, 1353–1368 (2016).

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

158. Hey, J. *et al.* Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.* **35**, 2805–2818 (2018).

159. Dalquen, D. A., Zhu, T. & Yang, A. Z. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst. Biol.* **66**, 379–398 (2017).

160. Wen, D. & Nakhleh, L. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst. Biol.* **67**, 439–457 (2018).

161. Zhang, C., Ogilvie, H. A., Drummond, A. J. & Stadler, T. Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* **35**, 504–517 (2018).

162. Flouri, T., Jiao, X., Rannala, B. & Yang, Z. A Bayesian Implementation of the Multispecies Coalescent Model with Introgression for Phylogenomic Analysis. *Mol. Biol. Evol.* (2019) doi:10.1093/molbev/msz296.

163. Kubatko, L. The mutlispecies coalescent. in *Handbook of Statistical Genomics* (eds. Balding, D., Moltke, I. & Marioni, J.) 219–245 (Wiley, 2019).

164. Rannala, B., Edwards, S., Leaché, A. D. & Yang, Z. The multi-species coalescent model and species tree inference. in *Phylogenetics in the Genomic Era* (eds. Galtier, N., Delsuc, F. & Scornavacca, C.) (2020).

165. Mirarab, S. *et al.* ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).

166. Liu, L., Yu, L. & Edwards, S. V. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* **10**, 302 (2010).

167. Ogilvie, H. A., Bouckaert, R. R. & Drummond, A. J. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.* **34**, 2101–2114 (2017).

168. Heled, J. & Drummond, A. J. Bayesian Inference of Species Trees from Multilocus Data. *Mol. Biol. Evol.* **27**, 570–580 (2010).

169. Yang, Z. & Rannala, B. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.* **31**, 3125–3135 (2014).

170. Flouri, T., Jiao, X., Rannala, B. & Yang, Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.* **35**, 2585–2593 (2018).

171. Nascimento, F. F., Reis, M. Dos & Yang, Z. A biologist's guide to Bayesian phylogenetic analysis. *Nat. Ecol. Evol.* **1**, 1446–1454 (2017).

172. Thawornwattana, Y., Dalquen, D. & Yang, Z. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the Anopheles gambiae species complex. *Mol. Biol. Evol.* **35**, 2512–2527 (2018).

173. Shi, C. M. & Yang, Z. Coalescent-Based Analyses of Genomic Sequence Data Provide a Robust Resolution of Phylogenetic Relationships among Major Groups of Gibbons. *Mol. Biol. Evol.* **35**, 159–179 (2018).

174. Mirarab, S., Bayzid, M. S. & Warnow, T. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.* **65**, 366–380 (2016).

175. Morgan, C. C. *et al.* Heterogeneous models place the root of the placental mammal phylogeny. *Mol. Biol. Evol.* **30**, 2145–2156 (2013).

176. Zhengting, Z. & Jianzhi, Z. Amino acid exchangeabilities vary across the tree of life. *Sci. Adv.* **5**, eaax3124 (2019).

177. Roch, S., Nute, M. & Warnow, T. Long-Branch Attraction in Species Tree Estimation: Inconsistency of Partitioned Likelihood and Topology-Based Summary Methods. *Syst. Biol.* **68**, 281–297 (2019).

178. Kobert, K., Stamatakis, A. & Flouri, T. Efficient detection of repeating sites to accelerate phylogenetic likelihood calculations. *Syst. Biol.* **66, 205–217 (2017).

Important speed up of Likelihood calculation.

179. Kobert, K., Flouri, T., Aberer, A. & Stamatakis, A. The divisible load balance problem and its application to phylogenetic inference. in *Algorithms in Bioinformatics. WABI 2014. Lecture Notes in Computer Science.* (eds. Brown, D. & Morgenstern, B.) 204–216 (Springer, 2014). doi:10.1007/978-3-662-44753-6_16.

180. Aberer, A. J., Kobert, K. & Stamatakis, A. ExaBayes: Massively parallel bayesian tree inference for the whole-genome era. *Mol. Biol. Evol.* **31**, 2553–2556 (2014).

181. Flouri, T. *et al.* The phylogenetic likelihood library. *Syst. Biol.* **64**, 356–362 (2015).

182. Ayres, D. L. *et al.* BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019).

183. Rannala, B. & Yang, Z. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.* **66**, 823–842 (2017).

184. Höhna, S. & Drummond, A. J. Guided tree topology proposals for Bayesian phylogenetic inference. *Syst. Biol.* **61**, 1–11 (2012).

185. Baele, G., Lemey, P., Rambaut, A. & Suchard, M. A. Adaptive MCMC in Bayesian phylogenetics: An application to analyzing partitioned data in BEAST. *Bioinformatics* **33**, 1798–1805 (2017).

Acknowledgements

The authors would like to thank members of the Z. Yang and the M. Telford lab as well as three anonymous reviewers for valuable feedback on previous versions of the manuscript. The writing of this review was supported by BBSRC grant reference BB/R016240/1.

Author Information.

Affiliations

Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment,

University College London, London WC1E 6BT, UK

Contributions

K.P., Z.Y. and M.J.T. contributed to all aspects of the article.

Corresponding author

Correspondence to MJT.

Ethics declarations.

Competing interests

The authors declare no competing interests.