nature reviews genetics

https://doi.org/10.1038/s41576-023-00620-x

Review article

Check for updates

Incongruence in the phylogenomics era

Jacob L. Steenwyk [1,2,3], Yuanning Li [4], Xiaofan Zhou [5], Xing-Xing Shen [6] & Antonis Rokas [2,3,7]

Abstract

Genome-scale data and the development of novel statistical phylogenetic approaches have greatly aided the reconstruction of a broad sketch of the tree of life and resolved many of its branches. However, incongruence — the inference of conflicting evolutionary histories — remains pervasive in phylogenomic data, hampering our ability to reconstruct and interpret the tree of life. Biological factors, such as incomplete lineage sorting, horizontal gene transfer, hybridization, introgression, recombination and convergent molecular evolution, can lead to gene phylogenies that differ from the species tree. In addition, analytical factors, including stochastic, systematic and treatment errors, can drive incongruence. Here, we review these factors, discuss methodological advances to identify and handle incongruence, and highlight avenues for future research.

Sections

[1]Howards Hughes Medical Institute and the Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA. [3]Vanderbilt Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN, USA. [4]Institute of Marine Science and Technology, Shandong University, Qingdao, China. [5]Guangdong Laboratory for Lingnan Modern Agriculture, Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou, China. [6]Key Laboratory of Biology of Crop Pathogens and Insects of Zhejiang Province, Institute of Insect Sciences, Zhejiang University, Hangzhou, China. [7]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany. e-mail: antonis.rokas@vanderbilt.edu

Introduction

"The stream of heredity makes phylogeny; in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream."
George Gaylord Simpson[1]

Phylogenetics aims to reconstruct the evolutionary histories of organisms, genes, traits or other biological features by examining the distribution of inherited characters in descendant lineages and tracing them back in time to identify ones that shared a common ancestor. 'Trees' inferred from phylogenetic analyses of biological features represent the best-supported hypotheses of their evolutionary histories, that is, the statistically most probable path rather than the ground truth. Phylogenetic approaches that use genome-scale data, or phylogenomics, have become the gold standard for understanding the evolution of lineages in the tree of life, a prerequisite for understanding the evolution of biological features[2–5]. Defined initially as predicting gene function from phylogenies of homologous genes[6], phylogenomics was later expanded to include phylogenetic inference using genome-scale data[7]. Phylogenomics has revolutionized systematic biology, resolving numerous branches of the tree of life that were previously contentious and increasing our confidence in many others[8–14].

Despite these successes, phylogenomic studies can sometimes support conflicting tree topologies[15,16], which suggests that certain branches of the tree of life are challenging to resolve, even with genome-scale data. Some of these branches concern relationships that are key to our understanding of the most exciting episodes in evolution (for one example, see Box 1), hindering our ability to resolve the tree of life. Incongruence is an umbrella term that describes the inference of conflicting tree topologies. This phenomenon can be observed at all time scales, from very ancient (hundreds of millions to billions of years old) to very recent (tens of thousands to millions of years old), and at all levels of genomic organization, from whole chromosomes to individual sites in a multiple sequence alignment (Fig. 1). The primary drivers of incongruence are biological processes that cause the genealogies of DNA sequences to differ from the genealogy of their species (for example, hybridization or horizontal gene transfer events)[2,5] and analytical shortcomings that lead to errors in inference (for instance, erroneous orthologue detection or poor model fit)[17]. Dissecting the contribution of biological and analytical drivers of incongruence can improve phylogenetic inference and deepen our understanding of phylogenesis and the evolutionary process.

Now, more than two decades after the dawn of phylogenomics, our understanding of the factors contributing to incongruence has matured. Concomitant development of methods and software that aid in identifying and accounting for incongruence in phylogenomic analyses has improved accuracy in inference. This Review synthesizes the biological and analytical factors that drive incongruence, discusses methodological advances to identify and handle incongruence, and highlights avenues for future research.

Biological factors

Several processes influence the evolution of genomic regions; these biological factors can cause a gene tree (which shows the evolutionary relationships between sequences of a single gene) to differ from the species tree (which shows the evolutionary relationships between different species) and contribute to incongruence (Fig. 2). Note that the term 'gene tree' is often used as shorthand for any locus (for example, a protein-coding or a non-coding region) in the genome; for simplicity, we follow this convention.

Incomplete lineage sorting

Incomplete lineage sorting can occur during the speciation process when alleles in a population fail to coalesce due to retention and random sorting of ancestral polymorphisms, causing, at times, alleles to first coalesce with alleles from more distantly related species (Fig. 2). It is common across sexually reproducing organisms because allelic polymorphisms often persist across multiple speciation events[18–20]. Incomplete lineage sorting does not always result in gene trees that are incongruent with the species phylogeny but, when it does, it is referred to as hemiplasy[21] (Table 1). Hemiplasy is particularly prevalent when populations are large and the time interval between speciation events is short[22] and can affect a substantial fraction of the genome. Examination of the evolutionary history among 500-bp windows of the human, chimpanzee, bonobo, gorilla and orangutan genomes revealed that ~37% of the human genome exhibits hemiplasy, and the evolutionary histories of these loci conflict with the species tree topology[18] (Fig. 2).

By modelling the underlying probability distribution of gene trees within a species tree, the multispecies coalescent model provides a framework that incorporates incomplete lineage sorting in phylogenomic inference[23]. One approach for evaluating whether hemiplasy explains gene tree–species tree incongruence is by simulating trees under the multispecies coalescent model and comparing levels of observed and expected gene tree incongruence[24]. If the observed incongruence is equal to the expected incongruence under the model, then hemiplasy is the major contributor; if not, other analytical or biological factors are likely (also) at play.

Other approaches, such as the one implemented by the BEAST software, use Bayesian statistics to co-estimate gene trees and species phylogenies in the presence of incomplete lineage sorting[25,26] (Table 2). These fully coalescent methods are computationally expensive, hindering their use for large phylogenomic data matrices. To reduce computational costs, summary coalescent-based methods implemented in various software packages, including STAR, MP-EST, ASTRAL, ASTER and ASTEROID[27–31] (Table 2), infer the species tree from pre-inferred single-gene trees in phylogenomic data matrices but at the cost of increased error rates in gene and species tree inference, especially for ancient divergences (see the section Analytical factors). Thus, although hemiplasy can contribute to the incongruence of both ancient and recent divergences, it is much more likely to be detectable in the latter.

Horizontal gene transfer

Genomic regions that experienced horizontal gene transfer also have histories that deviate from the species tree (Fig. 2 and Table 1). For example, eukaryotic acquisition of bacterial loci leads to gene phylogenies where eukaryotic sequences are nested within clades of bacterial sequences[5,32]. The contribution of horizontal gene transfer to incongruence is asymmetric across the tree of life; horizontal gene transfer is very common in Bacteria and Archaea and is a notable driver of genome evolution in these lineages[33,34]. Horizontal gene transfer in eukaryotes is less common, although evidence of its importance in eukaryotic genome evolution is increasing[35].

For lineages with low levels of horizontal gene transfer, incongruence stemming from horizontal gene transfer can be ameliorated by removing genes with signatures of transfer from the phylogenomic data matrix[36]. Horizontally transferred genes can be identified using phylogeny-based methods such as topology tests (implemented

in major programmes such as RAxML and IQ-TREE 2) that evaluate whether the gene tree topology indicative of horizontal gene transfer is significantly better than topologies that do not invoke transfer[37]. Horizontally transferred loci can also be detected by sequence composition-based methods, where notable changes in the GC content or codon usage bias of one or more loci relative to the rest of the genome are used to identify signatures of horizontal transfer[38], or using sequence similarity-based methods to detect foreign sequences such as alien index[39]. Sequence composition-based and similarity-based methods are faster, can be implemented across entire genomes, and are primarily suitable for recent events where the acquired sequence has not substantially diverged from the donor sequence; by contrast, phylogeny-based methods are generally more accurate, especially for ancient episodes of transfer, but slower and typically used to test horizontal transfer for one or a few loci.

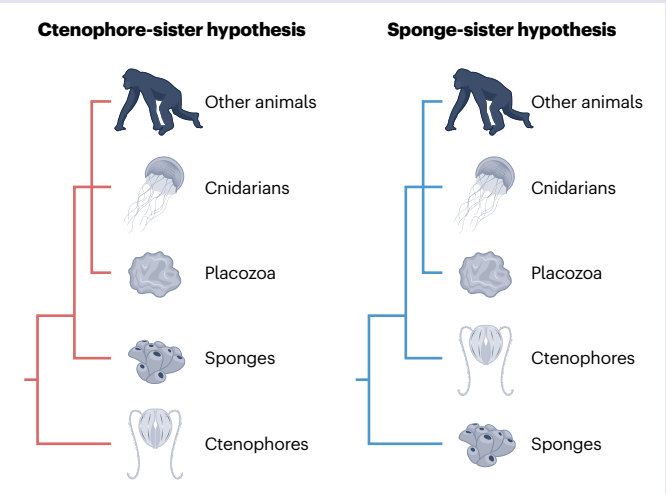
An alternative approach is to infer the species phylogeny through a probabilistic model of sequence evolution that explicitly models horizontal gene transfer as one of the processes that lead to gene tree–species tree incongruence[40,41], using programmes such as SpeciesRax[42]. Horizontal transfer can occur between both closely related species as well as between distantly related ones. However, irrespective of the method used, inference of gene transfer — and amelioration of its effects on incongruence — among distantly related species is much easier than among close relatives. This is because horizontal

Box 1

Rooting the animal tree

Few branches in the tree of life are as intensely debated as the root of animal phylogeny. The two leading hypotheses debate whether sponges[15,95,212–214] or comb jellies (ctenophores)[12,16,67,198,215,216] are the sister group to a clade of all other animals. These two hypotheses have come to be known as the sponge-sister and ctenophore-sister hypotheses, respectively (see the figure). Resolution of the root of the animal tree has a bearing on our understanding of how animal cell types and tissues evolved[217]. Sponges lack muscles and a nervous system and are sometimes thought of as morphologically 'simpler' animals compared to ctenophores, which have both[218–220]. Which hypothesis is correct also has implications for whether ctenophore nervous systems are structurally and genetically homologous to those of bilaterian animals[221,222], with some arguing that the ctenophore nervous system evolved independently[223].

Numerous biological and analytical factors contribute to this challenging phylogenetic problem. Much of the controversy has centred around whether site-homogeneous (with gene partitioning) or site-heterogeneous models of sequence evolution are most appropriate for reconstructing the animal phylogeny[198,224]. These models are largely employed to combat long-branch attraction, an artefact central to the debate because ctenophores have a long branch leading up to the lineage[225]. Site-heterogeneous models with many categories tend to support the sponge-sister hypothesis[15,198], whereas site-heterogeneous models with fewer categories and site-homogeneous models tend to support the ctenophore-sister hypothesis[198]. Some simulation analyses suggest that site-heterogeneous models underperform site-homogeneous models with gene partitioning[226] and others suggest the opposite[225]. Aimed at reducing saturation and compositional biases, data matrix recoding analyses supported the sponge-sister hypothesis[153,212]; however, some of these analyses[212] failed to recover well-established monophyletic clades, such as Chordata, suggesting that analyses of non-recoded data were more accurate[227]. Poor taxon sampling has also long impacted this phylogenetic question, but new genomes and transcriptomes have recently been made available for key lineages — sponges, ctenophores, cnidarians and placozoans[15,16,153]. Outgroup choice has also been important to the debate: the sponge-sister hypothesis is most frequently supported when choanoflagellates are chosen as the outgroup, whereas



the ctenophore-sister hypothesis is supported when a broader sampling of single-celled relatives of animals (Holozoa) and fungi (Opisthokonta) is used[198].

Several other factors, such as orthologue inference errors and multiple sequence alignment errors, are likely at play. The possibility that additional biological factors, such as hybridization or incomplete lineage sorting, also contribute cannot be excluded; however, detecting the effect of multiple analytical and biological factors in such an ancient divergence is challenging. Resolving the root of the animal tree may require extensive amounts of new (high-quality) data such as expanded taxon sampling of sponge, ctenophore and choanoflagellate genomes[217]. Similarly, other lines of evidence, such as investigations of synteny conservation using chromosome-level genome assemblies[228] — an independent line of evidence that does not have the same pitfalls as sequence data analyses — may shed light on the root of the animal tree. Interestingly, a very recent study identified seven genomic regions displaying synteny conservation across animals except for ctenophores but no regions with conserved synteny across animals except for sponges, providing support for the ctenophore-sister hypothesis[229].

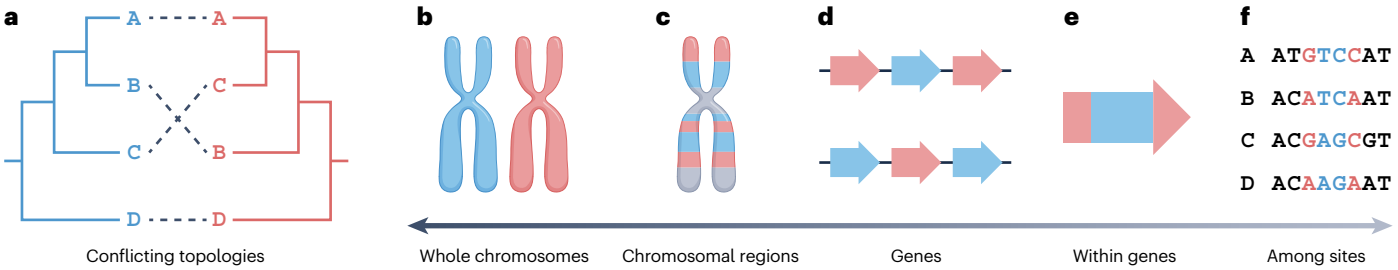


Fig. 1 | Incongruence at different levels of genomic organization. The topology shown in blue supports a sister group relationship of taxa A and B, whereas the red topology supports a sister group relationship of taxa A and C (part **a**). The inference of such conflicting topologies defines incongruence. Incongruence can occur at different levels in the genome, such as among whole chromosomes (for example, analyses of one chromosome support the blue topology but analyses of another support the red topology) (part **b**), regions of a chromosome (grey regions represent lack of homology) (part **c**), genes (or loci) (part **d**), within a gene or locus (for example, different domains support different topologies) (part **e**) and among sites in a multiple sequence alignment (part **f**). Note that incongruence is also prevalent in other types of data (for example, behavioural or morphological traits) and can occur at all evolutionary depths.

gene transfer between closely related species is much more difficult to distinguish from other evolutionary processes such as differential gene duplication and loss.

Hybridization, introgression and recombination

The exchange of genetic material between species during hybridization or introgression introduces alleles with evolutionary histories that deviate from the history of species, leading to gene tree–species tree incongruence[43,44] (Fig. 2). When the hybrid species has the same ploidy as the parental species, hybridization can be detected through phylogeny-based and sequence read-mapping methods. In phylogeny-based methods, phylogenomic data matrices containing loci from the hybrid and both parental species are expected to show nearly equal support (using measures such as internode certainty, gene support frequency and concordance factors; see the section Detecting incongruence) for two distinct topologies because roughly one-half of the hybrid genome derives from each parent[45]. Similarly, in sequence read-mapping methods, such as the one implemented in sppIDer[46] (Table 2), half of the sequence reads of the hybrid are expected to map to one parental species and the other half to the other parental species. Hybrid species that differ in their ploidy from the parental species (for example, allodiploid hybrids) can also be detected using the above methods but their gene number is also expected to be the sum of the genes in the parental species[47]. Approaches that ameliorate the contribution of hybridization to incongruence include first separating the hybrid genome into parental subgenomes prior to phylogenomic inference[47] and using probabilistic models that explicitly incorporate hybridization as one of the processes contributing to incongruence[48].

Introgression can also affect large genomic regions that can be several megabases in size or greater and lead to incongruence but it is potentially more challenging to detect because the percentage and distribution of introgressed regions can vary. Methods for introgression detection typically aim to identify allele patterns across species that significantly deviate from a null model in which these patterns are governed only by incomplete lineage sorting (and no introgression). These include the D-statistic (also known as the ABBA-BABA test), which is designed to detect gene flow between two taxa in a four-taxon phylogeny[49]; D_{FOIL}, which expands the D-statistic for the five-taxon case[50]; and D_3 and the branch-length test, which use the signal of pairwise divergence[51] – wherein gene trees that support introgression have shorter branch lengths[52] – for introgression detection (Table 2). Removing loci with signatures of introgression or directly modelling the process can ameliorate incongruence stemming from introgression[44]. For example, inclusion of introgressed regions (detected using the D-statistic) in a phylogenomic data set of passerine birds led to an incorrect species phylogeny; more accurate inference of the species phylogeny required careful examination not only of the topologies of individual loci but also of some of their properties such as recombination frequency and nucleotide diversity[43].

Recombination, a frequent phenomenon in diverse lineages, including prokaryotes and viruses, can also give rise to mosaic sequences and incongruence. In these instances, incongruence depends on the fraction of recombinant sites and how closely related the taxa are[53]. Sequences with evidence of recombination can be detected using PhyPack or RDP[54,55] and removed from the data matrix before inference. Accurate inference of all three processes is inversely proportional to the ages of the events; therefore, evaluating their contribution to incongruence in ancient divergences is challenging.

Natural selection

Natural selection generally leads to the divergence of sequences; however, selection for the same or similar traits in distantly related taxa can result in convergent molecular evolution[56] (Table 1). Thus, trees that contain sequences that experienced convergent evolution may erroneously suggest that these sequences are closely related, reflecting the shared influence of selection rather than common ancestry (Fig. 2). For example, phylogenetic analysis of the gene *prestin*, which encodes a transport protein present on the membrane of cochlear outer hair cells, groups sequences from echolocating organisms, such as bats and whales, together. This grouping occurs because the bat and whale sequences of *prestin* have experienced convergent molecular evolution as bats and whales are not sister lineages[57]. One method for detecting convergent sequence evolution is reconstructing ancestral sequences and identifying convergent amino acid substitutions in independent branches of the species phylogeny, if known[58]. Ancestral sequence reconstruction can be done with diverse software, including IQ-TREE[59], FireProt^{ASR}[60] and PhyloBot[61] (Table 2). Cases of convergent molecular evolution that affect one or a few genes are best handled by removing such genes, if they are known, from the data matrix before tree inference but their contribution to incongruence is generally expected to be small.

Convergent molecular evolution can also be observed in phylogenomic analyses of entire genomes or proteomes. For example,

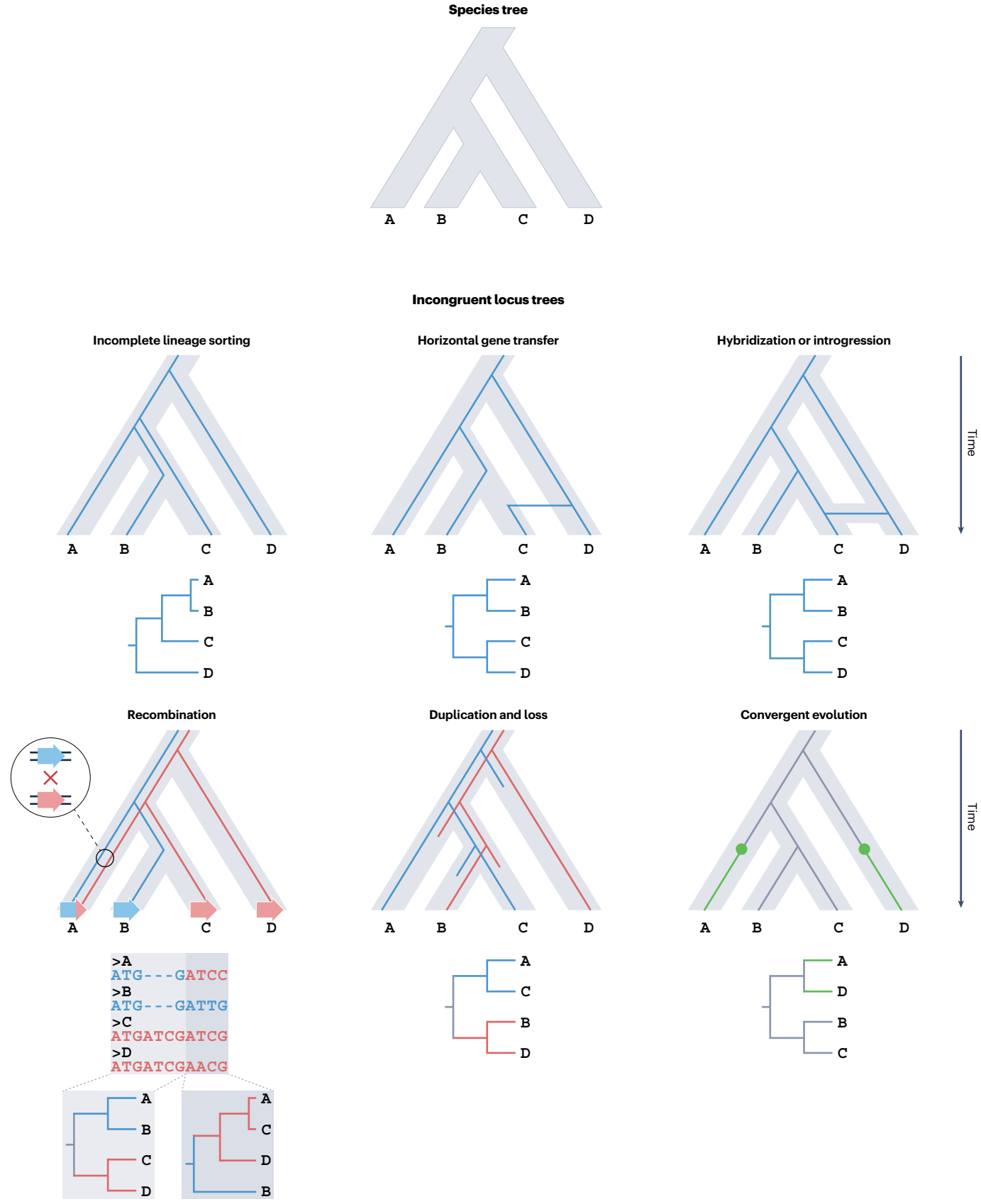

Species tree
A
B
C
D
Incongruent locus trees
Incomplete lineage sorting
Horizontal gene transfer
Hybridization or introgression
Time
Recombination
Duplication and loss
Convergent evolution
Time
>A
ATG---GATCC
>B
ATG---GATTG
>C
ATGATCGATCG
>D
ATGATCGAACG

Fig. 2 | Major biological factors that contribute to incongruence. The true species phylogeny is depicted in grey (top). Six exemplary processes that contribute to incongruence are depicted (bottom). For each example, the true species history is depicted in grey and the evolutionary history of a locus is depicted as lines within the species history. Incomplete lineage sorting can lead to gene trees that differ from the species phylogeny due to variation in the sorting of ancestral polymorphisms. Horizontal gene transfer, hybridization and introgression can also lead to gene phylogenies that differ from the species tree due to non-vertical evolution. Recombination can result in loci with chimeric evolutionary histories. Here, two distinct loci depicted in red and blue recombine on the branch leading to species A, resulting in one part of the gene that tracks with the evolutionary history of the red locus and another that tracks with the evolutionary history of the blue locus. Ancestral gene duplication followed by asymmetric patterns of paralogue loss can lead to hidden paralogy. Here, gene duplication occurred in the last common ancestor of all four taxa, giving rise to blue and red paralogues. Loss of the red paralogue in taxa A and C and loss of the blue paralogue in taxa B and D results in a one-to-one paralogous group of genes whose evolutionary history differs from that of the species. Independently evolved traits in different phylogenetic lineages can be associated with convergent molecular evolution (green), meaning that loci independently underwent similar nucleotide or amino acid substitutions in their sequences. One or a combination of these factors can contribute to loci incongruent with species histories.

convergent amino acid usage, such as the convergence observed in high-salt adapted Methanonatronarchaeia and Haloarchaea towards similarly acidified amino acid compositions in their proteomes, can obfuscate phylogenomic inference[62]. In such cases, incongruence can be reduced through exclusion of affected sites, character recoding (see the section Character recoding) and the use of models that explicitly account for compositional heterogeneity. For example, recent analyses on the evolutionary origins of mitochondrial genomes, a case of incongruence where compositional biases are at play[63], using a model that accommodates both across-site and across-branch compositional heterogeneity supported mitochondria as the sister lineage to Alphaproteobacteria[64].

Analytical factors

The content of phylogenomic data sets and choices in how these data sets are constructed and analysed can also contribute to incongruence (Fig. 3). Incongruence due to stochastic errors stems from statistical uncertainty when too few molecular markers or taxa are analysed. Incongruence from systematic errors stems from incorrect or inadequate assumptions in analysis such as substitution model misspecifications, a lack of realistic models or erroneous orthologue detection. Finally, choices in experimental design or treatment of phylogenomic data are an emerging category of error that can also lead to incongruence, sometimes exacerbating or leading to additional stochastic and/or systematic errors; we term these treatment errors.

Stochastic errors

Taxon sampling. Taxon sampling has a critical role in species tree inference and incongruence (Fig. 3a) because the number and taxonomic distribution of the sampled taxa influence numerous downstream analyses such as predicting orthologous groups of genes and the estimation of substitution model parameters (Table 1). Generally, including more taxa improves tree inference but can lead to speed versus accuracy trade-offs (see the section Treatment errors). In some cases, incongruence can guide the sampling of additional taxa. For example, the placement of the family *Ascoideaceae*, represented by a single taxon, was unstable in early phylogenomic studies of Saccharomycotina yeasts[65–67], but the inclusion of three additional taxa from *Ascoideaceae* stabilized its placement[68]. Similarly, the inclusion of additional taxa that diverged near the base of the land plant phylogeny increased the stability of phylogenetic inference[69–71]. However, taxon pruning, such as removing rogue taxa whose placement is unstable across a set of trees (for example, across a set of gene trees), may also improve congruence and accuracy in some cases[72,73]. Comprehensive taxon sampling may not always be possible, for example, for ancient lineages that contain one or a few closely related extant species such as coelacanths and lungfish[74]. However, studies of ancient DNA can shed light on phylogenetic relationships in cases where extant taxon sampling is difficult or impossible[75,76].

Locus sampling. How much sampling of sequence data is required depends on the specific evolutionary history of the lineage examined and how ancient or recent it is, on the information content of the loci used to reconstruct it, and on the evolutionary history of the loci[9,77,78] (see the section Biological factors). Thus, incongruence stemming from limited sampling of sequence data can affect the resolution of ancient and recent divergences[79,80] but can generally be improved with additional sampling of molecular markers (Table 1). Additional molecular markers can be obtained using programmes that identify single-copy orthologues from multi-copy gene families, for example, OrthoSNAP or DISCO[81,82] (Table 2). However, there is a limit imposed by the sequence divergence of the genomes examined, such that the resolution of relationships of genome sequences that contain relatively few informative sites and/or many taxa, such as the SARS-CoV-2 whole-genome alignments[80], will be challenging from sequence data alone. Additionally, data sets that contain short sequences (for example, gene fragments or short genes) often contain insufficient numbers of sites for robust gene tree inference when using summary-based coalescence methods and can contribute to incongruence[83] (Fig. 3a) but, at times, this limitation can be overcome by collapsing poorly supported branches before species tree inference[29].

Molecular markers included in phylogenomic data matrices typically exhibit partial or incomplete taxon coverage. This can increase statistical uncertainty, leading to identical support for multiple topologies, referred to as tree terraces[84,85]. For example, in a 3-locus, 298-taxon data matrix from grasses (Poaceae), with taxon coverage of 66%, the optimal tree is on a terrace with 61.2 million other equally supported topologies[84]. Tree terraces can be addressed through increased taxon coverage across molecular markers and locus sampling. For example, analysis of a 129-locus, 117-taxon data matrix of arthropods, with a coverage density similar to that of the data set of grasses (65%), yielded a single optimal tree[84,86]. The Gentrius function in IQ-TREE can help identify and characterize phylogenetic terraces[59] (Table 2).

Partial taxon coverage can stem from genuine differences in the gene content of organisms or from missing data (for example, from incomplete genome assemblies or errors in gene annotation). To reduce the negative effects of partial taxon coverage on inference, phylogenomic studies typically implement a taxon occupancy threshold of 50% or higher per locus[68,87]. However, different taxon occupancy thresholds may be optimal for different clades. For example, among Lori and Lorikeet birds, a taxon occupancy threshold of 70% was necessary to ameliorate the impact of missing data[88].

Systematic errors

Orthologue inference. Phylogenomic analyses often rely on single-copy orthologous genes, but errors in orthology inference, such as hidden orthology, can lead to incongruence. The over-splitting of orthologous groups of genes can stem from sequence length biases among orthologues because both BLAST bit scores and expectation values have a length dependency: longer sequences that contain many hundreds or thousands of base pairs have higher maximum bit scores and lower expectation values. Thus, variation in sequence length within an orthologous group of genes can lead to the exclusion of shorter sequences[89] (Fig. 3a and Table 1). Hidden orthology can also stem from a failure to detect rapidly evolving orthologues, an issue exacerbated across large evolutionary distances[90], resulting in artefactual inferences of lineage-specific genes. Hidden orthologues can be detected using 'bridging' methods such as Leapfrog, an algorithm for the identification of instances of reciprocal best BLAST hits in two different orthologous groups of genes[91] (Table 1). Probabilistic modelling approaches, such as profile Hidden Markov Models implemented in HMMER that leverage site-specific parameterization of conservation (or lack thereof) from multiple sequence alignments, are more sensitive in detecting rapidly evolving orthologues[92] and reduce the risk of hidden orthology (Table 2). Improved taxon sampling (for example, inclusion of under-represented lineages) in multiple sequence alignments used to construct profile Hidden Markov Models, such as those implemented in TIAMMAt, can further improve the sensitivity of sequence similarity searches[93] (Table 2).

Another systematic error source is the asymmetry in rates of gene duplication and loss between species, which can result in hidden paralogy. At shallow evolutionary depths (that is, when comparing species that diverged less than a hundred million years ago), hidden paralogy can be detected by examining synteny, for example, examining the synteny of six yeast species that underwent differential patterns of gene loss since a shared whole-genome duplication event revealed that ~10% of inferred single-copy orthologues were hidden paralogues[94]. Detecting hidden paralogy instances deeper in time (that is, when comparing species that diverged hundreds of millions or billions of years ago) is more challenging because synteny is likely not conserved. In such cases, hidden paralogues can potentially be detected by searching for gene trees where well-known clades are not monophyletic[95,96]. Alternatively, because hidden paralogues can be quite divergent from the rest of the sequences in an orthogroup, they can also be identified by examining gene trees for taxa that have unexpectedly long terminal branches using software such as TreeShrink[97], PhyloFisher[98] and PhyKIT[96] (Table 2). Inparalogues, especially species-specific ones, can easily be handled by retaining one of the two sequences as implemented in PhyloTreePruner[99] and OrthoSNAP[81].

Errors in orthologue inference can also stem from contaminated sequences in genome assemblies, a key concern in metagenome-assembled genomes. The degree of contamination (and completeness) of a given genome can be evaluated with the CheckM[100] and miComplete[101] programmes, and contaminant sequences can be removed prior to inference.

Table 1 | Drivers of incongruence

Driver of incongruence	Factor	Refs.
Incomplete lineage sorting	Biological	20,24,194
Horizontal gene transfer	Biological	36,40,195,196
Hybridization or introgression and recombination	Biological	43,44
Natural selection	Biological	57,58
Sampling (taxon and locus)	Analytical, stochastic error	83,197,198
Insufficient number of genes or divergent sites	Analytical, stochastic error	2,9,11,80
Erroneous orthologue detection	Analytical, systematic error	87,89,95,97,199
Model misspecification	Analytical, systematic error	8,125,126,200
Multiple sequence alignment errors	Analytical, treatment error	143,144
Excessive trimming	Analytical, treatment error	148,149
Inappropriate character recoding	Analytical, treatment error	154,155

Modelling substitutions. Traditional substitution models are site-homogeneous models, which use one reversible substitution matrix and the same nucleotide or amino acid frequencies for all sites in a data matrix. Early nucleotide models assumed equal substitution rates and base frequencies[102]. Later models incorporated biologically informed parameters such as accounting for differences in the rates of transitions and transversions or base frequencies[103,104]. The most parameter-rich model among reversible models for nucleotide sequences is the generalized time-reversible model, which uses unequal substitution rates and base frequencies[105]. Nucleotide substitution models that relax the assumptions of reversibility (that is, the rate at which a particular nucleotide, say A, changes to another one, say G, is not the same as the rate of a G changing to an A), stationarity (nucleotide frequencies do not change over time) and independence (changes at each site in the alignment are independent of changes at other sites) also exist, but they are computationally expensive and not typically used in phylogenomic studies[106].

In contrast to these mechanistic substitution models for nucleotide sequences, substitution models for amino acid sequences are often inferred from empirical multiple sequence alignments. For example, the amino acid exchange probabilities in the mtMAM substitution model were estimated empirically by examining the rates of amino acid substitutions across the mitochondrial proteomes of 20 mammals[107]; other substitution models, such as WAG[108] and LG[109], are derived by estimating substitution rates from larger, more diverse data bases of amino acid sequence alignments like Pfam.

Determining the best-fitting nucleotide and amino acid substitution models is often done using likelihood ratio tests and Akaike or Bayesian information criteria[110]. The latter outperform likelihood ratio tests but also have their shortcomings, which can result in the wrong model being favoured[111]. Of note, model fit does not always predict phylogenetic tree accuracy, and models of variable fit can sometimes result in consistent phylogenetic trees[112]. For example, the generalized time-reversible model is often the best-fitting nucleotide reversible model; however, the large number of estimated parameters in this model may need to be revised for specific analyses[113]. In general, the modelling of substitutions is more challenging in ancient divergences than in more recent ones because the variation of mutational processes and evolutionary rates is typically greater in analyses of distantly related taxa. Another avenue of modelling sequence evolution is through direct experimental measurement — mutagenesis, functional selection and deep sequencing.

Table 2 | Tools to investigate incongruence in large genomic data sets

Software or method	Utility category	Utility details	Refs.
Bag of little bootstraps	Bipartition support metric	Median bagging of bootstrap support assessed using few little samples and a small subset of sites is a rapid method to infer bootstrap trees and provides similar patterns of support compared to traditional bootstrapping procedures	201
Gene and site concordance factors	Bipartition support metric	Bipartition support that details how many 'decisive' genes or sites support a given bipartition in a reference tree	168
Internode or tree certainty	Bipartition support metric	Identifies bipartitions in a reference phylogeny that also have a well-supported alternative topology	172–175
UFBoot2	Bipartition support metric	Ultrafast bootstrap approximations that are robust to model violation	202
IQ-TREE 2, FireProt^{ASR}, PhyloBot	Convergent sequence evolution	Software for inferring ancestral sequences across nodes of a phylogeny; these pieces of software can be used to detect convergent sequence evolution	59–61
RERconverge	Convergent sequence evolution	Identifies genes in phylogenomic data matrices with signatures of convergent relative evolutionary rates in lineages with similar phenotypes	203
ClipKIT	Data processing and analysis	Multiple sequence alignment trimming wherein informative sites are retained rather than removing highly divergent sites	149
Concaterpillar	Data processing and analysis	Identifies congruent loci in a phylogenomic data matrix	204
ConJak	Data processing and analysis	Identifies sequence outliers compared to the central mean of a phylogenomic data matrix	205
ConWin	Data processing and analysis	Tests for within-protein incongruence using a sliding window approach	205
PhyKIT	Data processing and analysis	Broadly applicable phylogenomic toolkit for data processing and analysis such as examining information content biases, gene–gene coevolution and polytomy testing	96
PhyloFisher	Data processing and analysis	Collection of scripts for data set building and trimming phylogenomic data sets; also features a data base of eukaryotic orthologues	98
RogueNaRok	Data processing and analysis	Identification of rogue taxa in a phylogenomic data set	72
Root Digger	Data processing and analysis	Uses a non-reversible Markov model to calculate the likelihood of the root position in a tree	140
TreeShrink, PhyloFisher and PhyKIT	Data processing and analysis	Identifies spurious orthologues from unexpectedly long terminal branches	96–98
abSENSE	Homology and/or orthology detection	Calculates the probability that homologue detection may fail	90
BLAST	Homology and/or orthology detection	Searches for similar sequences by using measures of local similarity	206
Leapfrog	Homology and/or orthology detection	Combines over split orthologues using reciprocal best BLAST hits	91
OrthoFinder	Homology and/or orthology detection	Infers groups of orthologous genes	89
OrthoSNAP and DISCO	Homology and/or orthology detection	Decompose multi-copy gene families into subgroups of single-copy orthologous genes	81,82
Profile hidden Markov models	Homology and/or orthology detection	Probabilistic inference method that accounts for position-specific variation in sequences	92
TIAMMAt	Homology and/or orthology detection	Increases sensitivity of sequence similarity searches by incorporating under-represented lineages in profile Hidden Markov Models	93
ASTRAL and PhyKIT	Hypothesis testing	Both pieces of software enable researchers to conduct polytomy testing at a specific bipartition in a phylogeny	29,96
Gene-wise and site-wise log-likelihood scores; gene-wise quartet scores	Hypothesis testing	Allows researchers to examine gene-wise and site-wise support between two topologies using maximum likelihood; gene-wise support can also be examined using quartet scores	67,158
D-statistic (also known as the ABBA-BABA test), D_{FOIL}, D_3 and the branch-length test	Introgression detection	Diverse methods that detect introgression events using sequence or phylogenetic information	44, 49–51
NetRAX	Phylogenetic network inference	Maximum likelihood inference of phylogenetic networks when incomplete lineage sorting is not a factor	180
PhyloNet	Tree inference	Maximum parsimony, maximum likelihood and Bayesian inference of phylogenetic networks from locus tree estimates	179

Table 2 (continued) | Tools to investigate incongruence in large genomic data sets

Software or method	Utility category	Utility details	Refs.
SplitsTree	Phylogenetic network inference	Splits graph inference using multiple sequence alignments, distance matrices or sets of trees	177
GHOST	Substitution models	Edge-unlinked mixture model consisting of several site classes with separate sets of model parameters and edge lengths on the same tree topology	8
QMaker	Substitution models	Estimates general time-reversible protein matrices, which describe rates of substitutions between amino acids, from multiple sequence alignments	200
Asteroid	Tree inference	Supertree method for species tree inference that is robust to missing data	31
ASTRAL, ASTRAL-PRO and ASTER	Tree inference	Quartet-based supertree method that accounts for partial gene trees, paralogs and gene tree uncertainty	29,30, 207
BEAST	Tree inference	Bayesian approach for phylogenetic tree inference and divergence time estimation	26
BPP	Tree inference	Full-likelihood implementation of the multispecies coalescent	25
IQ-TREE 2	Tree inference	Maximum likelihood tree inference method that uses hill-climbing and stochastic perturbation to search tree space; moreover, the Gentrius function can help identify and characterize phylogenetic terraces	59
MP-EST	Tree inference	Maximum pseudo-likelihood approach for species tree inference	28
PhyloBayes MPI	Tree inference	Bayesian tree inference method that incorporates finite and infinite mixture models to account for site variation	208
RAxML-NG	Tree inference	Maximum likelihood tree inference method that uses a greedy tree search algorithm to explore tree space	209
STAR	Tree inference	Inference of species trees using average ranks of coalescences	210
SpeciesRax	Tree inference	Maximum likelihood species tree inference method that explicitly accounts for incomplete lineage sorting, gene duplication, gene loss and horizontal gene transfer	42
SVDQuartets	Tree inference	Inference of relationships using quartets and the coalescent model	211

These experimentally derived models have substantially improved fit compared to those with few or hundreds of parameters[114].

Partitioning concatenated data matrices — that is, applying different site-homogeneous substitution models to distinct molecular markers or portions of an alignment — can account for heterogeneity in substitutions among sites and lead to more accurate estimates of phylogeny[115]. Supermatrices can be partitioned by biological features (for example, genes or codon positions) or be algorithmically defined[116]. An alternative to partitioning is site-heterogeneous models, wherein nucleotide or amino acid equilibrium frequencies differ across sites of a multiple sequence alignment. Site-heterogeneous models fit data better than site-homogeneous models and are thought to be superior at ameliorating long-branch attraction artefacts[117,118]. Consequently, site-heterogeneous models have risen in popularity and helped resolve the placement of several anciently diverged lineages[119,120] but are also the focal point of controversies such as the rooting of the animal tree (Box 1). In other cases, using site-heterogeneous models has shed light on the evolutionary relationships among the three domains of life, supporting the hypothesis that eukaryotes originated from within Archaea (the two-domain hypothesis)[121].

Substitution model misspecification can bias topology estimation, contributing to incongruence[17,122–124] (Fig. 3c and Table 1). One well-known source of incongruence that stems from model misspecification is long-branch attraction[125,126]. Long-branch attraction is common in phylogenomic data matrices containing taxa that greatly vary in their evolutionary rates or lineages undergoing accelerated evolutionary rates as observed in bacterial endosymbionts[127] and parasitic fungi[128]. Outgroup taxa may also introduce long branches, increasing the potential for long-branch attraction artefacts (see the section Rooting strategy). In addition to using site-heterogeneous models[125], long-branch attraction artefacts can sometimes be ameliorated by including taxa whose placements break long branches[129,130] (see the section Taxon sampling). Notably, long-branch attraction can also occur when models are correctly specified and be exacerbated when partitioning phylogenomic data sets[126]. Other approaches attempt to better approximate true processes of sequence evolution. For example, heterotachy, which is not accounted for by either site-homogeneous or site-heterogeneous models[131], can decrease phylogenetic accuracy due to long-branch attraction artefacts[126,132]. The GHOST (general heterogeneous evolution on a single topology) model of sequence evolution can partly account for heterotachy by incorporating features of mixed substitution and mixed branch-length models. The GHOST model has helped resolve some phylogenetic controversies such as the placement of turtles[8].

Rooting strategy. Rooting strategies have been debated for a long time, especially in the context of outgroup taxa driving long-branch attraction artefacts[133]. The recent controversy surrounding the root of animal phylogeny has highlighted the relevance of these debates (Box 1). Although there is no consensus on selecting outgroup taxa[134], it is broadly accepted that thorough sampling of representatives of diverse lineages improves phylogenetic inference[135].

Other methods aim to infer the root of a phylogenetic tree without using outgroup taxa. These include the use of paralogues as implemented in the software STRIDE[136–138]; non-reversible Markov

models as the one implemented in the software Root Digger[139,140]; relaxed molecular clock models as implemented in BEAST[141]; the minimal ancestor deviation method, which is also based on molecular clocks[142]; and modelling dynamics of gene family evolution[41]. For example, modelling genome duplication, horizontal gene transfer and gene loss helped root the archaeal tree of life, placing it between Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaea (known as DPANN) and other Archaea[41].

Treatment errors

Multiple sequence alignment. Errors in multiple sequence alignment can result in inaccurate phylogenetic inferences and incongruence[143,144]. Alignment errors can stem from errors in orthologue inference (from either hidden paralogy or hidden orthology) but can also occur when truly orthologous sequences are aligned. Such errors are particularly common when sequences in the alignment exhibit high levels of divergence[145] (Fig. 3b). Approaches to remedy errors in multiple sequence alignments include alignment trimming (see the section Alignment trimming), probabilistic modelling to identify clusters of homologous characters and dividing the alignment accordingly (as implemented in Divvier)[146], or masking putative errors in multiple sequence alignments using two-dimensional outlier detection methods (as implemented in TAPER)[147].

Alignment trimming. Although trimming of sites during multiple sequence alignment is a widespread practice to reduce alignment errors, it can also reduce the accuracy of phylogenetic inference, increase statistical uncertainty and lead to incongruence (Fig. 3b and Table 1). Generally, more aggressive alignment trimming that removes larger numbers of sites increases errors in single-gene tree inferences[148]. For example, entropy-based trimming, which removes divergent sites, or multiple rounds of trimming, which often remove more than 20% of sites in an alignment, can significantly worsen phylogenetic inferences of tree topology, support and branch-length estimation[148,149]. Recently developed approaches that focus on retaining phylogenetically informative sites, such as ClipKIT (Table 2), can be equally accurate and are more time-saving than no-trimming approaches[149].

Character recoding. Saturation by multiple substitutions and compositional biases can lead to inaccurate phylogenetic inferences and contribute to incongruence. Recoding nucleotides or amino acids into fewer character states can combat these issues[150–153] (Fig. 3b). However, the benefit of combating compositional heterogeneity and substitutional saturation can be outweighed by the loss of information from reducing the number of character states during recoding and increase statistical uncertainty, especially among shorter alignments[154,155]. Thus, recoding can also increase, rather than ameliorate, error. Appropriate ways forward include adequately assessing how recoding affects compositional heterogeneity or implementing alternative recoding schemes; for example, in amino acid sequence alignments, a greater number of recoding states outperformed the most frequently implemented six-state recoding strategies[154]. Notably, errors in multiple sequence alignment, excessive trimming and inappropriate character recoding all contribute to erosion of the phylogenetic signal.

Concatenation versus coalescence. Phylogenomic data matrices can be analysed as a single supermatrix (an approach known as concatenation) or each gene alignment can be analysed separately under the multispecies coalescent framework (an approach known as coalescence). The two approaches sometimes yield different tree topologies, contributing to incongruence[68,87]. Determining which approach is more appropriate for a phylogenomic data set is difficult. For example, using simulated multi-locus data, concatenation slightly outperformed a fully coalescent-based approach (wherein gene trees and species trees are coestimated), whereas using coalescent independent sites, both approaches performed comparably[156]. However, an extensive evaluation of coalescent-based and concatenation-based approaches when different biological and analytical factors are at play is lacking, hindering our knowledge of best practices. Moreover, there can be differences in the performance of fully and summary coalescent-based methods (wherein gene trees are first estimated and then the species tree is estimated by summarizing the collection of gene trees). Summary coalescent-based methods are more vulnerable to errors in gene tree inference than fully coalescent-based methods but newer implementations of summary coalescent-based methods take gene tree uncertainty into account[30]. Analyses with both fully and summary coalescent-based methods can be improved through targeted data filtering such as removing loci with low phylogenetic informativeness[157]. Loci that are inconsistent between concatenation-based and coalescence-based methods can also be pruned from data matrices[158].

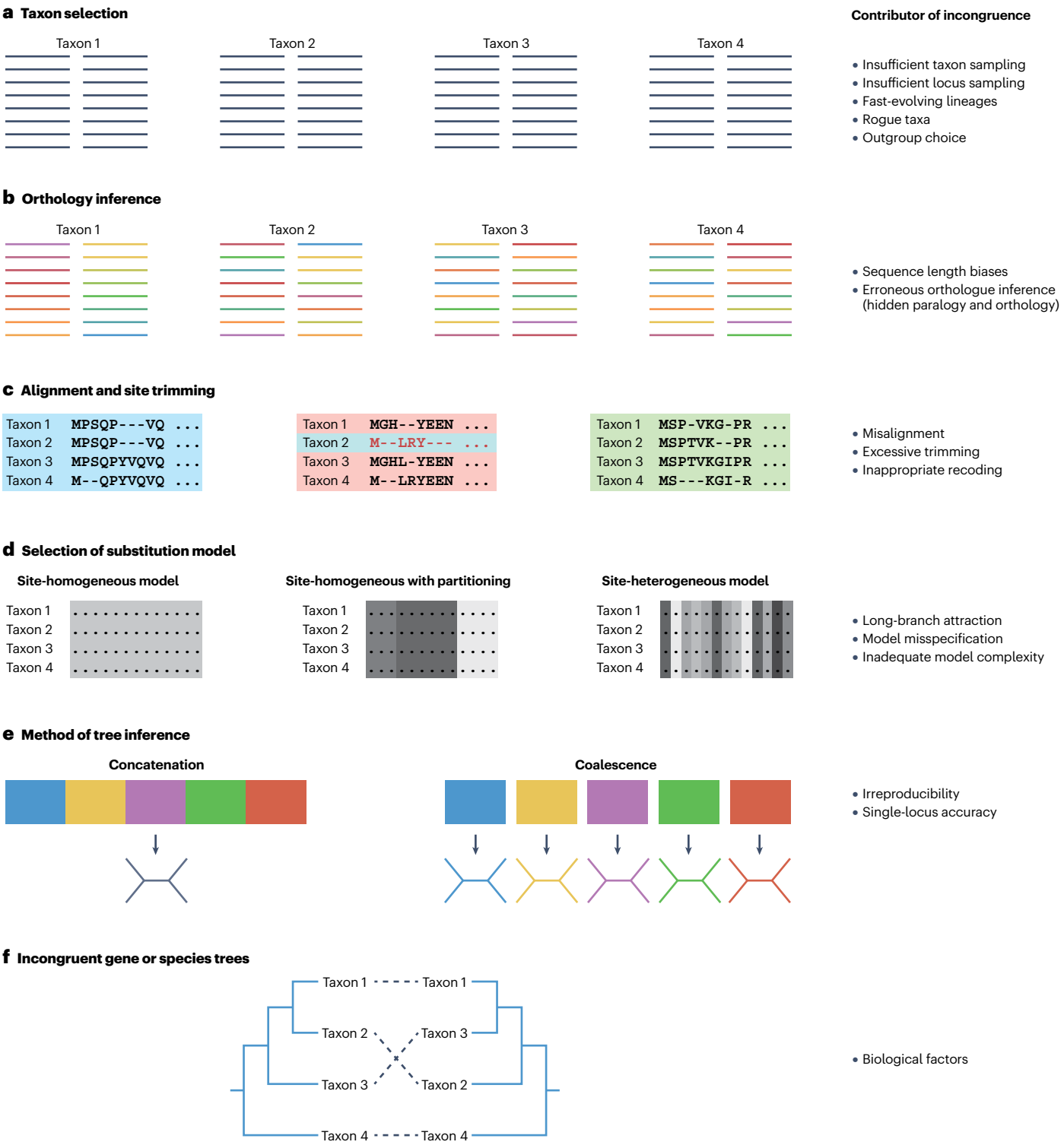
Irreproducibility. A tenet of scientific inquiry is reproducibility. Phylogenetic irreproducibility contributes to incongruence and can be caused by increasing the number of threads (because threads can be initialized in different orders between runs); errors in floating point arithmetic, such as rounding errors, and numerical overflow and underflow (the storing of a value greater than or smaller than the maximum and minimum supported value, respectively); and differences in software compilers that result in binaries with slightly different orders of operations[159,160]. Genes with a low phylogenetic signal (few parsimony-informative sites) are particularly susceptible to irreproducibility; this means that summary coalescent-based methods, which typically rely on accurately inferred gene tree topologies, can be particularly susceptible[160]. Some problems of irreproducibility and issues plaguing bioinformatic software can be remedied through rigorous software development practices such as extensive testing and continuous integration pipelines[149,159]. Studies that further our understanding of the accuracy and information content of multiple sequence alignments may facilitate predicting genes with a greater phylogenetic signal[77,161–163].

Detecting incongruence

Because multiple biological and analytical factors, often initially unknown, can contribute to incongruence, several methods examine the presence and magnitude of incongruence per se in phylogenomic data sets without assuming the presence of specific underlying biological or analytical factors.

Measures of branch support

Traditional approaches, such as non-parametric bootstrapping[164] and Bayesian posterior probabilities, are frequently used to examine bipartition support in a phylogeny; low branch support values can be indicative of incongruence. Other branch support methods include approximate likelihood ratio tests and the Shimodaira–Hasegawa approximate likelihood ratio test[165]. The transfer bootstrap



expectation method — an approach based on traditional bootstrapping but that measures the presence of branches among bootstrap trees as a gradual 'transfer' distance rather than as a binary presence or absence — is more accurate for assessing support among deep branches in data sets with large numbers of taxa[166]. The usefulness of many of these measures in concatenation analyses of phylogenomic data sets is rather low because they almost invariably yield absolute support values, even if there is substantial incongruence between sites or loci[79]. However, these measures are highly informative when using summary coalescent-based methods to remove loci with low amounts of phylogenetic signal[167].

Fig. 3 | Analytical factors can contribute to incongruence at every step in a phylogenomic workflow. a, Taxon sampling, including sampling of taxa from fast-evolving lineages or of rogue taxa, can affect all downstream analyses in phylogenomic studies. **b**, During orthology inference, biases (for example, sequence length biases) and analytical errors (for example, erroneous orthology inferences) can contribute to incongruence. Each colour corresponds to a unique orthologue present in each of the four taxa. **c**, Misalignment and excessive trimming of individual groups of orthologous genes can further decrease the accuracy of phylogenetic inferences and contribute to incongruence. An example of erroneous orthologue inclusion is depicted using red font. **d**, Selection of substitution model can influence phylogenetic inferences. Some common approaches include site-homogeneous models, site-homogeneous models with partitioning or site-heterogeneous models. Sources of error include model misspecification, inadequate model complexity and long-branch attraction artefacts. **e**, The method of tree inference, for example, concatenation (left) or coalescence (right), can be susceptible to multiple additional sources of error, including irreproducibility and poor accuracy of trees inferred from single genes (single-locus accuracy). **f**, Even if all analytical factors have been adequately addressed and there is no incongruence due to them, the resulting locus or species trees may still be exhibiting incongruence due to the action of diverse biological factors (Fig. 2).

Gene support frequencies and concordance factors

Gene support frequencies measure the frequency of recovering an individual branch in a set of gene trees from a phylogenomic data matrix[96,168]. Branches with low gene support frequencies are likely to be incongruent. Concordance factors were initially defined as the proportion of the genome that supports a given branch in the species tree[169,170] and can be measured using BUCKy, a Bayesian approach that estimates the joint probability distribution of genes and their phylogenies (or a gene-to-tree map) genome-wide[169,171]. Recently, concordance factors were redefined as equivalent to gene support frequencies[168], which can be calculated using IQ-TREE and PhyKIT[59,96] (Table 2).

Internode certainty

Internode certainty is an information theory-based approach that considers the relative prevalence of a branch and the second most common conflicting branch in a set of trees; internode certainty-all considers the relative prevalence of a branch relative to all alternative conflicting branches in a set of trees[172–175]. Internode certainty measures can help identify branches with substantial conflict, which can then be examined further for underlying causes contributing to incongruence. Internode certainty measures are distinct in that the prevalence of conflicting alternative branches is accounted for, thereby providing a measure of the degree of conflict for every branch in a phylogenomic tree. Internode certainty can be calculated using the software QuartetScores[174] (Table 2).

Phylogenetic networks

Evolutionary relationships among organisms are often depicted as bifurcating trees, which may not always be appropriate. As discussed earlier, many genomes bear the hallmarks of biological factors that make the histories of genes and genomes deviate from strict vertical inheritance. By relaxing the assumption of a strictly bifurcating topology, reconstruction of the histories of loci from such lineages as phylogenetic networks enables the description and visualization of incongruence. The underlying data and theory used to infer a phylogenetic network can differ[176]; for example, split networks depict all possible splits in a set of phylogenies[177], whereas reticulate networks depict putative evolutionary events such as hybridization[178]. Software for inferring phylogenetic networks includes SplitsTree[177], PhyloNet[179] and NetRAX[180] (Table 2).

Incongruence search protocols

In addition to the above methods, several protocols have been used to search for incongruence in phylogenomic data sets. These include repeated subsampling of smaller subsets of loci with a robust phylogenetic signal and re-inference of the species phylogeny[162], gene genealogy interrogation[181], examination of phylogenetic signal[67] and quartet sampling[182].

Polytomies

Several clades in the tree of life, such as cichlids and finches, have experienced elevated rates of speciation, giving rise to evolutionary radiation. Such clades have often been influenced by multiple biological (for example, introgression or lineage sorting) and analytical (for example, long-branch attraction for ancient radiations) factors, making phylogenomic inference particularly challenging. They often present as polytomies, a node where more than two descendant lineages stem from an ancestral one. Polytomies can be detected by identifying cases of equal support for multiple distinct topologies in sets of single-gene trees[96,183]. Support can be measured using gene trees or the quartets of taxa present in these gene trees using ASTRAL[29], PhyKIT[96] and IQ-TREE[59] (Table 2).

Future directions

Our knowledge of the tree of life and of the evolution of traits and genomes has been transformed by phylogenomics, but incongruence continues to cloud our understanding of some of its branches. We discussed biological and analytical factors contributing to incongruence, methods for its detection, and approaches that have helped improve the accuracy of phylogenomic inference. In this final section, we identify avenues ripe for research and discovery.

Which factors matter and when?

Although the effects of multiple factors on specific instances of incongruence have been investigated[32,157,160], a unified framework to assess the contribution of multiple biological and analytical factors to a given case of incongruence is lacking. The evolutionary depth of each case of incongruence further complicates assessment of the relative importance of any factor because our ability to detect their effects varies across time scales. For example, incomplete lineage sorting and hybridization likely contribute to incongruence of ancient and recent relationships but are typically detectable only in studies of recently diverged lineages. By contrast, it is typically much easier to detect horizontal gene transfer between distantly related taxa than between closely related ones. We also know that errors in orthologue inference or multiple sequence alignment are greater contributors to incongruence when studying ancient divergences than recent ones[90,184]. However, for a given case of incongruence in deep time, simultaneously evaluating the relative contribution of incongruence stemming from multiple biological and analytical factors is challenging (Box 1). A related issue is identifiability, that is, ascribing an observed conflict to certain factors and ruling out others. For example, ancient horizontal gene transfer is often difficult to distinguish from gene

duplication followed by extensive gene loss; attributing incongruence to one specific factor is challenging and often depends on a priori knowledge regarding which process is more likely. Developing methods and computational pipelines that enable simultaneous evaluation of potential contributing factors will be key to fully understanding the drivers of incongruence.

Data and data sets of ever higher quality

Data quality is paramount to phylogenomic inference. As sequencing technologies and other downstream processes, such as methods for genome assembly and gene annotation, improve, so does the field of phylogenomics. Higher quality and more complete genomes, coupled with increased sampling of organisms from taxa under-represented in genomic data bases, will help to reduce the impact of hidden paralogy and orthology in phylogenomic data sets. Denser data sets will also help increase confidence in inferences of the underlying analytical or biological drivers of incongruence; for example, confidence in inferring hybridization as a potential driver of incongruence may be weak in a data set of 100 molecular markers but strong in a 5,000-marker data set.

Mitigating errors in data set construction

Errors that contribute to incongruence can be introduced at all stages of phylogenomic analyses, including data matrix construction. Some errors may stem from certain strategies employed in a phylogenomic pipeline, such as multiple sequence alignment and trimming, being suitable for some, but not all, genes. Some features that may influence the efficacy of alignment and trimming strategies may be the taxa sampled and their evolutionary breadth, although numerous other technical contributors of incongruence may be at play. The development of pipelines to reproducibly handle phylogenomic data matrix construction will greatly facilitate comparative analyses of analytical drivers of incongruence across studies.

The forest grows: how can tree space be efficiently examined?

As genomic data increase, phylogenomic studies sampling several hundreds to thousands of organisms are becoming commonplace. One challenge with inferring phylogenies from such taxon-rich data sets is that tree space is vast, making computation challenging. For example, the numbers of possible unrooted trees for 3, 5, 7 and 9 taxa are 1, 15, 945 and 135,135, respectively. As tree space grows, the likelihood of finding the non-optimal tree increases, leading to speed-accuracy trade-offs and incongruence. However, efficiently searching tree space is key to finding an optimal tree; phylogenetic inference programmes that yield the highest likelihood scores on phylogenomic data matrices are the ones that perform the most extensive explorations of tree space and require the longest runtimes[185]. Moreover, gene-rich data sets present their own challenges such as optimizing tree parameters. It is possible that the phylogenetic signal in whole genomes will prove insufficient for resolving phylogenies of all known species in each major lineage. Developing algorithms, including those that leverage the power of machine learning[163,186–188], that can heuristically explore tree space in a reasonable amount of time or evaluate the degree of difficulty in the inference task will be critical for resolving the tree of life.

Phylogenomics and green computing

End-to-end phylogenomic analysis requires substantial computational resources and large amounts of energy. As the planet grapples with the consequences of global climate change, we must work to minimize the environmental toll of phylogenomic analyses[189]. We can reduce the carbon footprint of phylogenomics through judicious use of computing infrastructure, careful experimental design and software choice.

Glossary

Convergent molecular evolution
Independent evolution of similar or identical molecular changes (for example, gene deletions, nucleotide substitutions, gene order rearrangements) in organisms from different lineages that exhibit similar adaptations.

Evolutionary radiation
The occurrence of an elevated rate of speciation events in a narrow window of evolutionary time.

Heterotachy
The phenomenon of changes in the evolutionary rate of a nucleotide or amino acid sequence through time.

Hidden orthology
Undetected orthologous relationships of genes.

Hidden paralogy
Orthologous groups of genes that contain orthologues and paralogues (inparalogues and outparalogues) stemming from asymmetric patterns of duplication and loss.

Horizontal gene transfer
Also known as lateral gene transfer. The transfer of genetic material between organisms of the same or different species through non-reproductive means.

Hybridization
The interbreeding of two distinct species or lineages.

Inparalogues
Lineage-specific or species-specific paralogues wherein the duplication event occurred after divergence from a reference common ancestor.

Introgression
The interbreeding of two distinct species or lineages, followed by backcrossing with one of the parental species.

Long-branch attraction
The inaccurate inference of taxa with high evolutionary rates (giving rise to long branches in their phylogenetic trees) as closely related.

Model of sequence evolution
Also known as the substitution model. Markov models that describe rates of nucleotide or amino acid substitutions in a locus during evolution.

Partial or incomplete taxon coverage
The lack of sequences (either because they are genuinely absent or because they were not collected) from particular taxa in a group of orthologous genes.

Phylogenetic irreproducibility
Lack of reproducibility of a tree topology between two replicate tree inferences using the same software parameters (for example, same model of sequence evolution or starting seed).

Phylogenetic networks
Graphs of evolutionary relationships that, in addition to depicting the splitting of lineages, also depict the merging of lineages (due to events such as hybridization and convergent molecular evolution or due to different gene tree topologies).

Taxon sampling
Which and how many taxa are selected for a phylogenetic analysis.

For example, evaluating substitution model fit using fast and robust software such as ModelTest-NG[190] and jModelTest[191] can result in a 90% reduction in energy use, resulting in 10% less greenhouse gas emissions[192]. Similarly, choosing faster programmes in quantifiably difficult-to-analyse data sets does not alter the quality of inference but can save energy, according to a recent preprint[193].

Conclusions

Phylogenomics has revolutionized evolutionary biology by providing a clearer picture of the tree of life and a more accurate reconstruction of the evolution of biological features. The study of genome-scale data from diverse organisms has deepened our understanding of the various biological factors that cause the histories of genomic regions to differ from those of their species, spurring the development of models that consider them in inference. At the same time, it has improved knowledge of the analytical factors that contribute to errors in inference and the development of models and protocols for their amelioration. These advances, together with ongoing work that tackles some of the greatest challenges in the field, will continue to improve our mapping and understanding of, as Simpson eloquently put it, "*The stream of heredity* [that] *makes phylogeny*"[1].

Published online: 27 June 2023

References

1. Simpson, G. G. *The Principles of Classification and a Classification of Mammals* Vol. 85 (American Museum of Natural History, 1945).
2. Jarvis, E. D. et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
3. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
4. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
5. Li, Y. et al. HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* **185**, 2975–2987.e10 (2022).
6. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**, 163–167 (1998).
7. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **6**, 361–375 (2005).
8. Crotty, S. M. et al. GHOST: recovering historical signal from heterotachously evolved sequence alignments. *Syst. Biol.* **69**, 249–264 (2020).
9. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804 (2003).
10. Kawahara, A. Y. et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl Acad. Sci. USA* **116**, 22657–22663 (2019).
11. Misof, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
12. Dunn, C. W. et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749 (2008).
13. Bond, J. E. et al. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for Orb web evolution. *Curr. Biol.* **24**, 1765–1771 (2014).
14. Li, Y. et al. A genome-scale phylogeny of the kingdom Fungi. *Curr. Biol.* **31**, 1653–1665.e5 (2021).
15. 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).
16. Whelan, N. V. et al. Ctenophore relationships and their placement as the sister group to all other animals. *Nat. Ecol. Evol.* **1**, 1737–1746 (2017).
17. Lemmon, A. R. & Moriarty, E. C. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* **53**, 265–277 (2004).
18. Mao, Y. et al. A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**, 77–81 (2021).
19. Meleshko, O. et al. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Mol. Biol. Evol.* **38**, 2750–2766 (2021).
20. Feng, S. et al. Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell* **185**, 1646–1660.e18 (2022).
21. Avise, J. C. & Robinson, T. J. Hemiplasy: a new term in the lexicon of phylogenetics. *Syst. Biol.* **57**, 503–507 (2008).
22. Maddison, W. P. & Knowles, L. L. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* **55**, 21–30 (2006).
23. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).
24. Song, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl Acad. Sci. USA* **109**, 14942–14947 (2012).
25. 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).
26. Bouckaert, R. et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
27. Liu, L., Yu, L., Kubatko, L., Pearl, D. K. & Edwards, S. V. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* **53**, 320–328 (2009).
28. 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).
29. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153 (2018).
30. Zhang, C. & Mirarab, S. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Mol. Biol. Evol.* **39**, msac215 (2022).
This study describes the latest version of the state-of-the-art software for phylogenomic inference using summary-based coalescence methods. By incorporating weighting schemes that reduce the contribution of weakly supported gene trees and/or of trees with long branch lengths.
31. Morel, B., Williams, T. A. & Stamatakis, A. Asteroid: a new algorithm to infer species trees from gene trees under high proportions of missing data. *Bioinformatics* **39**, btac832 (2023).
32. Kominek, J. et al. Eukaryotic acquisition of a bacterial operon. *Cell* **176**, 1356–1366.e10 (2019).
33. Arnold, B. J., Huang, I.-T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* **20**, 206–218 (2022).
34. Gophna, U. & Altman-Price, N. Horizontal gene transfer in Archaea — from mechanisms to genome evolution. *Annu. Rev. Microbiol.* **76**, 481–502 (2022).
35. Van Etten, J. & Bhattacharya, D. Horizontal gene transfer in eukaryotes: not if, but how much? *Trends Genet.* **36**, 915–925 (2020).
36. Lapierre, P., Lasek-Nesselquist, E. & Gogarten, J. P. The impact of HGT on phylogenomic reconstruction methods. *Brief. Bioinform.* **15**, 79–90 (2014).
37. Wisecaver, J. H. & Rokas, A. Fungal metabolic gene clusters: caravans traveling across genomes and environments. *Front. Microbiol.* **6**, 161 (2015).
38. Sevillya, G., Adato, O. & Snir, S. Detecting horizontal gene transfer: a probabilistic approach. *BMC Genomics* **21**, 106 (2020).
39. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in Bdelloid rotifers. *Science* **320**, 1210–1213 (2008).
40. Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E. & Daubin, V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl Acad. Sci. USA* **109**, 17513–17518 (2012).
This study uses a statistical model of genome evolution that considers gene duplications, gene losses and horizontal gene transfers in phylogenetic reconstruction, demonstrating that incongruence stemming from these processes can inform inferences of evolutionary history.
41. Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl Acad. Sci. USA* **114**, E4602–E4611 (2017).
42. Morel, B. et al. SpeciesRax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *Mol. Biol. Evol.* **39**, msab365 (2022).
43. Zhang, D. et al. Most genomic loci misrepresent the phylogeny of an avian radiation because of ancient gene flow. *Syst. Biol.* **70**, 961–975 (2021).
44. Hibbins, M. S. & Hahn, M. W. Phylogenomic approaches to detecting and characterizing introgression. *Genetics* **220**, iyab173 (2022).
45. Sang, T. & Zhong, Y. Testing hybridization hypotheses based on incongruent gene trees. *Syst. Biol.* **49**, 422–434 (2000).
46. Langdon, Q. K., Peris, D., Kyle, B. & Hittinger, C. T. sppIDer: a species identification tool to investigate hybrid genomes with high-throughput sequencing. *Mol. Biol. Evol.* **35**, 2835–2849 (2018).
47. Steenwyk, J. L. et al. Pathogenic allodiploid hybrids of *Aspergillus* fungi. *Curr. Biol.* **30**, 2495–2507.e7 (2020).
48. Yu, Y., Dong, J., Liu, K. J. & Nakhleh, L. Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl Acad. Sci. USA* **111**, 16448–16453 (2014).
49. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
50. Pease, J. B. & Hahn, M. W. Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* **64**, 651–662 (2015).
This work describes a method for detecting incomplete lineage sorting and introgression in the five-taxon case, enabling identification of the taxa involved and the direction of introgression.
51. Hahn, M. W. & Hibbins, M. S. A three-sample test for introgression. *Mol. Biol. Evol.* **36**, 2878–2882 (2019).
52. Suvorov, A. et al. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Curr. Biol.* **32**, 111–123.e5 (2022).
53. Posada, D. & Crandall, K. A. The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.* **54**, 396–402 (2002).
54. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
55. Martin, D. P. et al. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **7**, veaa087 (2021).

56. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: new insights and directions. *Phil. Trans. R. Soc. B* **374**, 20190102 (2019).
57. Li, Y., Liu, Z., Shi, P. & Zhang, J. The hearing gene Prestin unites echolocating bats and whales. *Curr. Biol.* **20**, R55–R56 (2010).
Striking example of convergent molecular evolution in *Prestin*, a gene that encodes a protein involved in echolocation. Even though echolocating bats and whales are not sister lineages, bat and whale sequences of *Prestin* group these lineages together, demonstrating how convergent evolution can contribute to incongruence.
58. Castoe, T. A. et al. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl Acad. Sci. USA* **106**, 8986–8991 (2009).
59. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
60. Musil, M. et al. FireProt^{ASR}: a web server for fully automated ancestral sequence reconstruction. *Brief. Bioinform.* **22**, bbaa337 (2021).
61. Hanson-Smith, V. & Johnson, A. PhyloBot: a web portal for automated phylogenetics, ancestral sequence reconstruction, and exploration of mutational trajectories. *PLoS Comput. Biol.* **12**, e1004976 (2016).
62. Martijn, J. et al. Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat. Commun.* **11**, 5490 (2020).
63. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
64. Muñoz-Gómez, S. A. et al. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* **6**, 253–262 (2022).
This article describes a novel model of protein evolution that considers compositional heterogeneity both across sites of a data matrix and across branches of a phylogeny. This model is likely better than site-homogeneous or site-heterogenous models in cases where compositional heterogeneity varies across time and across the phylogeny such as the thorny question of the origin of mitochondria.
65. Riley, R. et al. Comparative genomics of biotechnologically important yeasts. *Proc. Natl Acad. Sci. USA* **113**, 9882–9887 (2016).
66. Shen, X.-X. et al. Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. *G3* **6**, 3927–3939 (2016).
67. Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* **1**, 0126 (2017).
This article describes a novel approach to visualize single-gene and single-site support for conflicting phylogenetic hypotheses. Application of this approach on phylogenomic data from different instances of incongruence reveals that a few, or even single, genes or sites in very large phylogenomic data matrices can drive incongruence.
68. Shen, X.-X. et al. Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533–1545.e20 (2018).
69. Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., Ruhfel, B. R. & Soltis, D. E. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* **105**, 291–301 (2018).
70. Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl Acad. Sci. USA* **111**, E4859–E4868 (2014).
71. Cheng, S. et al. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* **179**, 1057–1067.e14 (2019).
72. Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* **62**, 162–166 (2013).
73. Struck, T. H. TreSpEx — detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol. Bioinform. Online* **10**, EBO.S14239 (2014).
74. Amemiya, C. T. et al. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).
75. Liu, S. et al. Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell* **184**, 4874–4885.e16 (2021).
76. Perri, A. R. et al. Dire wolves were the last of an ancient New World canid lineage. *Nature* **591**, 87–91 (2021).
77. Townsend, J. P. Profiling phylogenetic informativeness. *Syst. Biol.* **56**, 222–231 (2007).
78. Patel, S., Kimball, R. T. & Braun, E. L. Error in phylogenetic estimation for bushes in the tree of life. *J. Phylogenet. Evol. Biol.* **01**, 1000110 (2013).
79. Rokas, A. & Carroll, S. B. Bushes in the tree of life. *PLoS Biol.* **4**, e352 (2006).
80. Pipes, L., Wang, H., Huelsenbeck, J. P. & Nielsen, R. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *Mol. Biol. Evol.* **38**, 1537–1543 (2021).
This article shows that statistical support for the rooting of the SAR-CoV-2 phylogeny is weak, suggesting that there is a limit in our power to resolve certain phylogenetic branches.
81. Steenwyk, J. L. et al. OrthoSNAP: a tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees. *PLoS Biol.* **20**, e3001827 (2022).
82. Willson, J., Roddur, M. S., Liu, B., Zaharias, P. & Warnow, T. DISCO: species tree inference using multicopy gene family tree decomposition. *Syst. Biol.* **71**, 610–629 (2022).
83. Springer, M. S. & Gatesy, J. The gene tree delusion. *Mol. Phylogenet. Evol.* **94**, 1–33 (2016).
84. Sanderson, M. J., McMahon, M. M. & Steel, M. Terraces in phylogenetic tree space. *Science* **333**, 448–450 (2011).
85. Xi, Z. et al. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation *Malpighiales*. *Proc. Natl Acad. Sci. USA* **109**, 17519–17524 (2012).
86. Sanderson, M. J., McMahon, M. M., Stamatakis, A., Zwickl, D. J. & Steel, M. Impacts of terraces on phylogenetic inference. *Syst. Biol.* **64**, 709–726 (2015).
87. Steenwyk, J. L., Shen, X.-X., Lind, A. L., Goldman, G. H. & Rokas, A. A robust phylogenomic time tree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*. *mBio* **10**, e00925-19 (2019).
88. Smith, B. T., Mauck, W. M., Benz, B. W. & Andersen, M. J. Uneven missing data skew phylogenomic relationships within the lories and lorikeets. *Genome Biol. Evol.* **12**, 1131–1147 (2020).
89. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
This article describes OrthoFinder, a state-of-the-art software for the identification of groups of orthologous genes that considers incomplete lineage sorting and gene duplication and loss, improving the accuracy of ortholog inference.
90. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**, e3000862 (2020).
91. Martín-Durán, J. M., Ryan, J. F., Vellutini, B. C., Pang, K. & Hejnol, A. Increased taxon sampling reveals thousands of hidden orthologs in flatworms. *Genome Res.* **27**, 1263–1272 (2017).
92. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
93. Tassia, M. G., David, K. T., Townsend, J. P. & Halanych, K. M. TIAMMAt: leveraging biodiversity to revise protein domain models, evidence from innate immunity. *Mol. Biol. Evol.* **38**, 5806–5818 (2021).
94. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**, 341–345 (2006).
95. Philippe, H. et al. Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19**, 706–712 (2009).
96. Steenwyk, J. L. et al. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics* **37**, 2325–2331 (2021).
97. Mai, U. & Mirarab, S. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genom.* **19**, 272 (2018).
98. Tice, A. K. et al. PhyloFisher: a phylogenomic package for resolving eukaryotic relationships. *PLoS Biol.* **19**, e3001365 (2021).
99. Kocot, K. M., Citarella, M. R., Moroz, L. L. & Halanych, K. M. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinform. Online* **9**, EBO.S12813 (2013).
100. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
101. Hugoson, E., Lam, W. T. & Guy, L. miComplete: weighted quality evaluation of assembled microbial genomes. *Bioinformatics* **36**, 936–937 (2020).
102. Jukes, T. H. & Cantor, C. R. In *Mammalian Protein Metabolism* 1st edn, Vol. III (ed. Munro, H. N.) Ch. 24 (Academic Press, 1969).
103. 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).
104. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
105. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86 (1986).
106. Arenas, M. Trends in substitution models of molecular evolution. *Front. Genet.* **6**, 319 (2015).
107. Yang, Z., Nielsen, R. & Hasegawa, M. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**, 1600–1611 (1998).
108. 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).
109. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
110. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772–772 (2012).
111. Susko, E. & Roger, A. J. On the use of information criteria for model selection in phylogenetics. *Mol. Biol. Evol.* **37**, 549–562 (2020).
112. Spielman, S. J. Relative model fit does not predict topological accuracy in single-gene protein phylogenetics. *Mol. Biol. Evol.* **37**, 2110–2123 (2020).
113. Abadi, S., Azouri, D., Pupko, T. & Mayrose, I. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat. Commun.* **10**, 934 (2019).
114. Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* **31**, 1956–1978 (2014).
Through systematic mutagenesis, functional selection and sequencing experiments, this study experimentally determines a substitution model for a viral protein. This parameter-free model is a much better fit than models with hundreds of parameters, highlighting the potential of high-throughput experimental strategies in improving the accuracy of phylogenetic inference.
115. Kainer, D. & Lanfear, R. The effects of partitioning on phylogenetic inference. *Mol. Biol. Evol.* **32**, 1611–1627 (2015).
116. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **34**, 772–773 (2016).

117. 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).
This landmark study introduces site-heterogeneous models of sequence evolution. By considering compositional heterogeneity across sites, these models can better ameliorate the impact of long-branch attraction artefacts.
118. Si Quang, L., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
119. Stairs, C. W. et al. Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Curr. Biol.* **31**, 5605–5612.e5 (2021).
120. Galindo, L. J., López-García, P., Torruella, G., Karpov, S. & Moreira, D. Phylogenomics of a new fungal phylum reveals multiple waves of reductive evolution across Holomycota. *Nat. Commun.* **12**, 4973 (2021).
121. Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* **4**, 138–147 (2019).
122. Minin, V., Abdo, Z., Joyce, P. & Sullivan, J. Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* **52**, 674–683 (2003).
123. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13**, 303–314 (2012).
124. Sullivan, J. & Swofford, D. L. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mamm. Evol.* **4**, 77–86 (1997).
125. 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).
126. Susko, E. & Roger, A. J. Long branch attraction biases in phylogenetics. *Syst. Biol.* **70**, 838–843 (2021).
127. Husník, F., Chrudimský, T. & Hypša, V. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches. *BMC Biol.* **9**, 87 (2011).
128. Capella-Gutiérrez, S., Marcet-Houben, M. & Gabaldón, T. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biol.* **10**, 47 (2012).
129. Graybeal, A. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47**, 9–17 (1998).
130. Hillis, D. M. Inferring complex phytogenies. *Nature* **383**, 130–131 (1996).
131. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**, 1–7 (2002).
132. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* **5**, 50 (2005).
133. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).
134. Geuten, K., Massingham, T., Darius, P., Smets, E. & Goldman, N. Experimental design criteria in phylogenetics: where to add taxa. *Syst. Biol.* **56**, 609–622 (2007).
135. Pollock, D. D., Zwickl, D. J., McGuire, J. A. & Hillis, D. M. Increased taxon sampling is advantageous for phylogenetic inference. *Syst. Biol.* **51**, 664–671 (2002).
136. Brady, S. G., Litman, J. R. & Danforth, B. N. Rooting phylogenies using gene duplications: an empirical example from the bees (Apoidea). *Mol. Phylogenet. Evol.* **60**, 295–304 (2011).
137. Mathews, S., Clements, M. D. & Beilstein, M. A. A duplicate gene rooting of seed plants and the phylogenetic position of flowering plants. *Phil. Trans. R. Soc. B* **365**, 383–395 (2010).
138. Emms, D. M. & Kelly, S. STRIDE: species tree root inference from gene duplication events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017).
139. Naser-Khdour, S., Quang Minh, B. & Lanfear, R. Assessing confidence in root placement on phylogenies: an empirical study using nonreversible models for mammals. *Syst. Biol.* **71**, 959–972 (2022).
140. Bettisworth, B. & Stamatakis, A. Root Digger: a root placement program for phylogenetic trees. *BMC Bioinformatics* **22**, 225 (2021).
141. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
142. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat. Ecol. Evol.* **1**, 0193 (2017).
143. Ashkenazy, H., Sela, I., Levy, K. E., Landan, G. & Pupko, T. Multiple sequence alignment averaging improves phylogeny reconstruction. *Syst. Biol.* **68**, 117–130 (2019).
144. Li-San, W. et al. The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 1108–1119 (2011).
145. Landan, G. & Graur, D. Characterization of pairwise and multiple sequence alignment errors. *Gene* **441**, 141–147 (2009).
146. Ali, R. H., Bogusz, M. & Whelan, S. Identifying clusters of high confidence homologies in multiple sequence alignments. *Mol. Biol. Evol.* **36**, 2340–2351 (2019).
147. Zhang, C., Zhao, Y., Braun, E. L. & Mirarab, S. TAPER: pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods Ecol. Evol.* **12**, 2145–2158 (2021).
148. 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).
Upending conventional wisdom, this study convincingly demonstrates that trimming typically reduces the accuracy of phylogenetic inference and contributes to incongruence.
149. Steenwyk, J. L., Buida, T. J., Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
This article describes a novel and more accurate approach to multiple sequence alignment trimming, where phylogenetically informative sites, which are more easily defined than phylogenetically uninformative sites, are retained and other sites are removed.
150. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).
151. Blanquart, S. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23**, 2058–2071 (2006).
152. Phillips, M. J., Delsuc, F. & Penny, D. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21**, 1455–1458 (2004).
153. Laumer, C. E. et al. Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. *eLife* **7**, e36278 (2018).
154. Hernandez, A. M. & Ryan, J. F. Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses. *Syst. Biol.* **70**, 1200–1212 (2021).
155. Foster, P. G. et al. Recoding amino acids to a reduced alphabet may increase or decrease phylogenetic accuracy. *Syst. Biol.* https://doi.org/10.1093/sysbio/syac042 (2022).
156. Wascher, M. & Kubatko, L. Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. *Syst. Biol.* **70**, 33–48 (2021).
157. Alda, F. et al. Resolving deep nodes in an ancient radiation of neotropical fishes in the presence of conflicting signals from incomplete lineage sorting. *Syst. Biol.* **68**, 573–593 (2019).
158. Shen, X.-X., Steenwyk, J. L. & Rokas, A. Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. *Syst. Biol.* **70**, 997–1014 (2021).
159. Darriba, D., Flouri, T. & Stamatakis, A. The state of software for evolutionary biology. *Mol. Biol. Evol.* **35**, 1037–1046 (2018).
160. Shen, X.-X., Li, Y., Hittinger, C. T., Chen, X. & Rokas, A. An investigation of irreproducibility in maximum likelihood phylogenetic inference. *Nat. Commun.* **11**, 6096 (2020).
This study reports that a considerable fraction of single gene phylogenies inferred from phylogenomic data matrices is irreproducible, leading to a novel source of incongruence in phylogenomic studies.
161. Shen, X.-X., Salichos, L. & Rokas, A. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. *Genome Biol. Evol.* **8**, 2565–2580 (2016).
162. Mongiardino Koch, N. Phylogenomic subsampling and the search for phylogenetically reliable loci. *Mol. Biol. Evol.* **38**, 4025–4038 (2021).
163. Haag, J., Höhler, D., Bettisworth, B. & Stamatakis, A. From easy to hopeless — predicting the difficulty of phylogenetic analyses. *Mol. Biol. Evol.* **39**, msac254 (2022).
164. Hillis, D. M. & Bull, J. J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–192 (1993).
165. Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).
166. Lemoine, F. et al. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).
167. Molloy, E. K. & Warnow, T. To include or not to include: the impact of gene filtering on species tree estimation methods. *Syst. Biol.* **67**, 285–303 (2018).
168. Minh, B. Q., Hahn, M. W. & Lanfear, R. New methods to calculate concordance factors for phylogenomic datasets. *Mol. Biol. Evol.* **37**, 2727–2733 (2020).
This article reports the development of methods to calculate the degree to which sites or genes support a particular branch of a phylogeny, also known as concordance factors, and their implementation in the IQ-TREE software. Concordance factors are very useful in identifying the presence of incongruence among a set of trees.
169. Ane, C., Larget, B., Baum, D. A., Smith, S. D. & Rokas, A. Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* **24**, 412–426 (2006).
170. Baum, D. A. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon* **56**, 417–426 (2007).
171. Larget, B. R., Kotha, S. K., Dewey, C. N. & Ané, C. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* **26**, 2910–2911 (2010).
172. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).
173. Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. Computing the internode certainty and related measures from partial gene trees. *Mol. Biol. Evol.* **33**, 1606–1617 (2016).
174. Zhou, X. et al. Quartet-based computations of internode certainty provide robust measures of phylogenetic incongruence. *Syst. Biol.* **69**, 308–324 (2020).
This article reports the development of internode certainty measures for phylogenomic data matrices with partial taxon coverage. By explicitly quantifying the level of incongruence of a given internal branch among a set of phylogenetic trees, internode certainty measures are a key tool for diagnosing the presence of incongruence in phylogenomic studies.
175. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).
176. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
177. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).

178. Huson, D. H., Klöpper, T., Lockhart, P. J. & Steel, M. A. Reconstruction of reticulate networks from gene trees. In *Proc. 9th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2005* (eds Miyano, S. et al.) 233–249 (Springer, Berlin, 2005).
179. Wen, D., Yu, Y., Zhu, J. & Nakhleh, L. Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* **67**, 735–740 (2018).
180. Lutteropp, S., Scornavacca, C., Kozlov, A. M., Morel, B. & Stamatakis, A. NetRAX: accurate and fast maximum likelihood phylogenetic network inference. *Bioinformatics* **38**, 3725–3733 (2022).
181. Arcila, D. et al. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat. Ecol. Evol.* **1**, 0020 (2017).
182. Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E. & Smith, S. A. Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. *Am. J. Bot.* **105**, 385–403 (2018).
183. Sayyari, E. & Mirarab, S. Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* **9**, 132 (2018).
184. Ogden, T. H. & Rosenberg, M. S. Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.* **55**, 314–328 (2006).
185. Zhou, X., Shen, X.-X., Hittinger, C. T. & Rokas, A. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol. Biol. Evol.* **35**, 486–503 (2018).
186. Suvorov, A., Hochuli, J. & Schrider, D. R. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Syst. Biol.* **69**, 221–233 (2020).
187. Azouri, D., Abadi, S., Mansour, Y., Mayrose, I. & Pupko, T. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nat. Commun.* **12**, 1983 (2021).
188. Rosenzweig, B. K., Hahn, M. W. & Kern, A. Accurate detection of incomplete lineage sorting via supervised machine learning. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.09.515828 (2022).
189. Grealey, J. et al. The carbon footprint of bioinformatics. *Mol. Biol. Evol.* **39**, msac034 (2022).
This article examines the environmental impact and carbon footprint of bioinformatic analyses, including phylogenetics, offering numerous suggestions for greener computing.
190. Darriba, D. et al. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* **37**, 291–294 (2020).
191. Posada, D. jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* **25**, 1253–1256 (2008).
192. Kumar, S. Embracing green computing in molecular phylogenetics. *Mol. Biol. Evol.* **39**, msac043 (2022).
193. Höhler, D., Haag, J., Kozlov, A. M. & Stamatakis, A. A representative performance assessment of maximum likelihood based phylogenetic inference tools. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.31.514545 (2022).
194. Scornavacca, C. & Galtier, N. Incomplete lineage sorting in mammalian phylogenomics. *Syst. Biol.* **66**, 112–120 (2016).
195. Galtier, N. A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst. Biol.* **56**, 633–642 (2007).
196. Stolzer, M. et al. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**, i409–i415 (2012).
197. Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief. Bioinform.* **13**, 122–134 (2012).
198. Li, Y., Shen, X.-X., Evans, B., Dunn, C. W. & Rokas, A. Rooting the animal tree of life. *Mol. Biol. Evol.* **38**, 4322–4333 (2021).
A systematic and in-depth examination of the evidence in favour of the sponge-sister and ctenophore-sister hypotheses concerning the rooting of the animal tree of life.
199. Cheon, S., Zhang, J. & Park, C. Is phylotranscriptomics as reliable as phylogenomics? *Mol. Biol. Evol.* **37**, 3672–3683 (2020).
200. Minh, B. Q., Dang, C. C., Vinh, L. S. & Lanfear, R. QMaker: fast and accurate method to estimate empirical models of protein evolution. *Syst. Biol.* **70**, 1046–1060 (2021).
201. Sharma, S. & Kumar, S. Fast and accurate bootstrap confidence limits on genome-scale phylogenies using little bootstraps. *Nat. Comput. Sci.* **1**, 573–577 (2021).
202. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
203. Kowalczyk, A. et al. RERconverge: an R package for associating evolutionary rates with convergent traits. *Bioinformatics* **35**, 4815–4817 (2019).
204. Leigh, J. W., Susko, E., Baumgartner, M. & Roger, A. J. Testing congruence in phylogenomic analysis. *Syst. Biol.* **57**, 104–115 (2008).
205. Al Jewari, C. & Baldauf, S. L. Conflict over the Eukaryote root resides in strong outliers, mosaics and missing data sensitivity of site-specific (CAT) mixture models. *Syst. Biol.* **72**, 1–16 (2023).
206. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
207. Zhang, C., Scornavacca, C., Molloy, E. K. & Mirarab, S. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* **37**, 3292–3307 (2020).
208. 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).
209. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
210. Liu, L., Yu, L., Pearl, D. K. & Edwards, S. V. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* **58**, 468–477 (2009).
211. Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317–3324 (2014).
212. Redmond, A. K. & McLysaght, A. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nat. Commun.* **12**, 1783 (2021).
213. Pisani, D. et al. Genomic data do not support comb jellies as the sister group to all other animals. *Proc. Natl Acad. Sci. USA* **112**, 15402–15407 (2015).
214. Feuda, R. et al. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Curr. Biol.* **27**, 3864–3870.e4 (2017).
215. Ryan, J. F. et al. The genome of the Ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **342**, 1242592 (2013).
216. Moroz, L. L. et al. The ctenophore genome and the evolutionary origins of neural systems. *Nature* **510**, 109–114 (2014).
217. King, N. & Rokas, A. Embracing uncertainty in reconstructing early animal evolution. *Curr. Biol.* **27**, R1081–R1088 (2017).
218. Dunn, C. W., Leys, S. P. & Haddock, S. H. D. The hidden biology of sponges and ctenophores. *Trends Ecol. Evol.* **30**, 282–291 (2015).
219. Nielsen, C. Early animal evolution: a morphologist's view. *R. Soc. Open Sci.* **6**, 190638 (2019).
220. Burkhardt, P. et al. Syncytial nerve net in a ctenophore adds insights on the evolution of nervous systems. *Science* **380**, 293–297 (2023).
221. Liebeskind, B. J., Hillis, D. M., Zakon, H. H. & Hofmann, H. A. Complex homology and the evolution of nervous systems. *Trends Ecol. Evol.* **31**, 127–135 (2016).
222. Sachkova, M. Y. et al. Neuropeptide repertoire and 3D anatomy of the ctenophore nervous system. *Curr. Biol.* **31**, 5274–5285.e6 (2021).
223. Burkhardt, P. Ctenophores and the evolutionary origin(s) of neurons. *Trends Neurosci.* **45**, 878–880 (2022).
224. Baños, H., Susko, E. & Roger, A. J. Is over-parameterization a problem for profile mixture models? Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.18.481053 (2022).
225. Kapli, P. & Telford, M. J. Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. *Sci. Adv.* **6**, eabc5162 (2020).
226. Whelan, N. V. & Halanych, K. M. Who let the CAT out of the Bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Syst. Biol.* **66**, 232–255 (2017).
227. Whelan, N. V. & Halanych, K. M. Available data do not rule out Ctenophora as the sister group to all other Metazoa. *Nat. Commun.* **14**, 711 (2023).
228. Parey, E. et al. Genome structures resolve the early diversification of teleost fishes. *Science* **379**, 572–575 (2023).
This study uses conservation of genome structure or synteny as an independent source of phylogenomic data. In combination with phylogenomic sequence data, these rare genomic changes resolve controversial relationships in early fish evolution.
229. Schultz, D. T. et al. Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).

Acknowledgements

J.L.S. and A.R. were funded by the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study Program. Research in A.R.'s lab is supported by grants from the National Science Foundation (DEB-2110404), the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R01 AI153356), and the Burroughs Wellcome Fund. A.R. acknowledges support from a Klaus Tschira Guest Professorship from the Heidelberg Institute for Theoretical Studies and from a Visiting Research Fellowship from Merton College of the University of Oxford. X.X.S. was supported by the National Key R&D Program of China (2022YFD1401600). Y.L. was supported by Shandong University Outstanding Youth Fund (62420082260514). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

All authors researched the literature, contributed substantially to discussions of the content, and reviewed and/or edited the manuscript before submission. J.L.S. and A.R. wrote the article.

Competing interests

J.L.S. is a scientific adviser for WittGen Biotechnologies and an adviser for ForensisGroup. A.R. is a scientific consultant for LifeMine Therapeutics. The other authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Genetics* thanks Thijs J.G. Ettema, who co-reviewed with Daniel Tamarit, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.