Incongruence in the phylogenomics era

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

[1] Howards Hughes Medical Institute and the Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[2] Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA

[3] Vanderbilt Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA

[4] Institute of Marine Science and Technology, Shandong University, Qingdao 266237, 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 510642, China

[6] Key Laboratory of Biology of Crop Pathogens and Insects of Zhejiang Province, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China

[7] Heidelberg Institute for Theoretical Studies, Heidelberg 69118, Germany

[†]e-mail: antonis.rokas@vanderbilt.edu

Abstract | Genome-scale amounts of 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. We synthesize the biological and analytical factors that drive incongruence, discuss methodological advances to diagnose and handle incongruence, and identify avenues for future research. The study of incongruence has enabled a deeper understanding of phylogenesis and improved our ability to reconstruct and interpret the tree of life.

"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, 1945[1]

Introduction

Phylogenetics aims to reconstruct the evolutionary histories of organisms, genes, traits, and other biological features. Trees inferred from phylogenetic analyses of biological features represent the best-supported hypotheses of their evolutionary histories, not the ground truth. Phylogenetic approaches that use genome-scale amounts of 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]. Phylogenomics revolutionized systematic biology, resolving numerous branches of the tree of life that were previously contentious and increasing our confidence in many others[6–12].

Despite these successes, different phylogenomic studies can sometimes support conflicting tree topologies[13,14], suggesting that certain branches of the tree of life are challenging to resolve, even with genome-scale data. Some of these branches concern relationships key to our understanding of evolution's most exciting episodes (see Box 1 for one example) and hinder 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 levels of genomic organization, from whole chromosomes to individual sites (Fig. 1). The primary drivers of incongruence are biological processes that cause the histories of DNA sequences to differ from the histories of their species—hybridization or horizontal gene transfer events, for example[2,5]—and analytical shortcomings that lead to errors in inference—erroneous ortholog detection or poor model fit, for instance[15]. 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, roughly two decades after the dawn of phylogenomics, the field's understanding of the factors contributing to incongruence has matured. Concomitant development of methods and software that aid in diagnosing and accounting for incongruence in phylogenomic analyses has improved accuracy in inference. This review synthesizes the factors that drive incongruence, methodological advances to diagnose and handle incongruence, and highlights avenues for future research.

Biological factors

Several evolutionary processes influence the evolutionary histories of genomic regions while others erase these histories; these biological factors cause the histories of genomic regions to deviate from the history of the species and contribute to incongruence (Fig. 2).

Incomplete lineage sorting

INCOMPLETE LINEAGE SORTING is common across sexually reproducing organisms[16–18]. 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[19] (Fig. 2, Table 1). Hemiplasy is particularly prevalent when populations are large and the time interval between speciation events is short[20], and can affect a substantial fraction of the genome. Examination of the evolutionary history of 500 base pair windows from 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[16] (Fig. 2).

By modeling 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[21]. 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[22]. If the observed incongruence is equal to the expected incongruence under the model, then hemiplasy is the major contributor to incongruence; 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 coestimate gene trees and species phylogenies in the presence of incomplete lineage sorting[23,24]

(Table 2). These full 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[25–29] (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 *Analytical factors* section). Thus, while hemiplasy may contribute to 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 OR LATERAL GENE TRANSFER also have histories that deviate from the species tree (Fig. 2, Table 1). For example, eukaryotic acquisition of bacterial loci leads to gene phylogenies where eukaryotic sequences are nested within clades of bacterial sequences[5,30]. 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 significant driver of genome evolution in these lineages[31,32]. Horizontal gene transfer in eukaryotes is less common, but evidence of its importance in eukaryotic genome evolution is increasing[33].

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[34]. Horizontally transferred genes can be identified using phylogeny-based methods, such as topology tests (implemented in major programs, 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[35]. Horizontally transferred loci can also be detected by sequence composition-based methods wherein 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[36] or using sequence similarity-based methods to detect foreign sequences, such as alien index[37]. Sequence composition- and similarity-based methods are faster, can be implemented across entire genomes, and are primarily suitable for recent events, whereas phylogeny-based methods are generally more accurate 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 genome evolution that explicitly models horizontal gene transfer as one of the processes that lead to gene tree-species tree incongruence[38,39], using programs such as SpeciesRax[40]. 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 transfers – and amelioration of its effects on incongruence – among distantly related species is much easier than among close relatives.

Hybridization, Introgression, and Recombination

The exchange of genetic material between species during HYBRIDIZATION or INTROGRESSION introduces alleles with evolutionary histories that deviate from the species' history, leading to locus tree-species tree incongruence[41,42]. 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 equal support (using measures such as internode certainty and concordance factors; see *In search for incongruence* section) for two distinct topologies because half of the hybrid's genome comes from one parent and half from the other[43]. Similarly, in sequence read-mapping methods, such as the one implemented in sppIDer[44] (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 (e.g., 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[45]. Approaches that ameliorate the contribution of hybridization to incongruence include to first separate the hybrid genome into parental subgenomes prior to phylogenomic inference[46] and using probabilistic models that explicitly incorporate hybridization as one of the processes contributing to incongruence[47].

Introgression can also impact large genomic regions 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) designed to detect gene flow between two taxa in a four-taxon phylogeny, D_{FOIL}, which expands the *D*-statistic for the five-taxon case, D_3, and the branch-length test that use the signal of pairwise divergence—wherein gene trees that support introgression have shorter branch lengths [48]—for

introgression detection[42,49,50] (Table 2). Removing loci with signatures of introgression or directly modeling the process can ameliorate incongruence stemming from introgression[42]. For example, inclusion of introgressed regions (detected using the *D-statistic*) in a phylogenomic dataset of passerine birds led to an incorrect species phylogeny; inference of the true 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[41].

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 [51]. Sequences with evidence of recombination can be detected using PhyPack or RDP[52,53] and removed from the data matrix before inference. Accurate inference of all three processes is inversely proportional to the age(s) of the event(s), such that evaluating whether they are contributing 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[54] (Table 1). Thus, gene trees of genes that have experienced convergent evolution may erroneously infer that they are closely related, reflecting the shared influence of selection rather than common ancestry (Fig. 2). Phylogenetic analysis of the gene *prestin*, which encodes a transport protein present on the membrane of cochlear outer hair cells, shows that sequences from echolocating organisms, such as bats and whales, group together because they have experienced convergent molecular evolution even though bats and whales are not sister lineages[55]. 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[56]. Ancestral sequence reconstruction can be done with diverse software including IQ-TREE[57], FireProt[ASR58], and PhyloBot[59] (Table 2). Cases of convergent molecular evolution that affect one or a few genes are best handled by removing those genes from the data matrix prior to inference.

Convergent molecular evolution can also be observed in phylogenomic analyses of entire genomes or proteomes. For example, convergent amino acid usage—such as the convergence observed in high-salt adapted Methanonatronarchaeia and Haloarchaea toward similarly acidified amino acid

compositions in their proteomes—can obfuscate phylogenomic inference[60]. In such cases, incongruence can be reduced either through exclusion or recoding (see *Characterter recoding* section) of affected sites or through the use of models that explicitly account for compositional heterogeneity. For example, resolving the evolutionary origins of mitochondrial genomes, a case of incongruence where compositional biases are at play[61], recent analyses using a model that accomodates both across-site and across-branch compositional heterogeneity supported mitochondria as the sister lineage to Alphaproteobacteria[62].

Analytical factors

The content of phylogenomic datasets and choices in how these datasets are constructed and analyzed can also contribute to incongruence. These stochastic, systematic, and treatment errors are collectively called analytical factors (Fig. 3). Incongruence due to stochastic errors stems from statistical uncertainty when too few molecular markers or taxa are analyzed. Incongruence from systematic errors stems from incorrect or inadequate assumptions in analysis—such as substitution model misspecifications or the lack of realistic models and erroneous ortholog detection. Finally, choices in experimental design or treatment of phylogenomic data are an emerging category of error, sometimes exacerbating or leading to additional stochastic and / or systematic errors; they can also lead to incongruence. We term these treatment errors.

Stochastic errors

Taxon Sampling. TAXON SAMPLING plays 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 *Treatment errors* section). 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[63–65], but the inclusion of three additional taxa from Ascoideaceae stabilized its placement[66]. Similarly, the inclusion of additional taxa that diverged near the base of the land plant phylogeny increased the stability of phylogenetic inference[67–69]. However, taxon pruning—such as removing ROGUE TAXA—may also improve congruence and accuracy in some cases[70,71]. Comprehensive taxon sampling may not always be possible, such as for ancient lineages that contain one or a few closely related extant species, such as coelacanths and lungfish[72]. However, studies of

ancient DNA can shed light on phylogenetic relationships in cases where extant taxon sampling is difficult or impossible[73,74].

Locus sampling. How much sampling of sequence data is required is dependent 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 (see the previous section on *biological factors*)[7,75,76]. Thus, incongruence stemming from limited sampling of sequence data can affect the resolution of ancient and recent divergences[77,78], but can generally be ameliorated with additional sampling of molecular markers (Table 1). Additional molecular markers can be sampled using programs that can identify single-copy orthologs from gene families, for example, OrthoSNAP or DISCO[79,80] (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—will be challenging from sequence data alone[78]. Additionally, datasets that contain short sequences (e.g., 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[81] (Fig. 3a), but these can be overcome by collapsing poorly supported branches before species tree inference[82].

Molecular markers included in phylogenomic data matrices typically exhibit PARTIAL TAXON COVERAGE. This can increase statistical uncertainty, leading to identical support for multiple topologies, referred to as tree terraces[83,84]. For example, in a three-locus, 298-taxon data matrix from grasses with taxon coverage of 66%, the optimal tree is on a terrace with 61.2 million other equally supported topologies[83]. Tree terraces can be addressed through increased taxon coverage across molecular markers and locus sampling. Case in point, analysis of a 129-locus, 117-taxon data matrix of arthropods with a coverage density similar to that of the dataset of grasses, 65%, yielded a single optimal tree[83,85]. The gentrius function in IQ-TREE can help identify and characterize phylogenetic terraces[86] (Table 2).

Systematic errors

Ortholog 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 orthologs because both

253 BLAST bit scores and expectation values have a length dependency such that longer sequences can
254 have higher maximum bit scores and lower expectation values; thus, variation in sequence length
255 within an orthologous group of genes can lead to exclusion of shorter sequences[87] (Fig. 3a, Table 1).
256 Hidden orthology can also stem from detection failure of rapidly evolving orthologs, an issue
257 exacerbated across large evolutionary distances[88], resulting in artifactual inferences of lineage-
258 specific genes. Hidden orthologs can be detected using "bridging" methods such as Leapfrog, an
259 algorithm for identifying instances of reciprocal best BLAST hits in two different orthologous groups
260 of genes[89] (Table 1). Probabilistic modeling approaches, such as profile Hidden Markov Models
261 implemented in HMMER that leverage site-specific parameterization of conservation (or lack thereof)
262 from multiple sequence alignments are more sensitive in detecting rapidly evolving orthologs[90] and
263 reduce the risk of hidden orthology (Table 2). Improved taxon sampling (e.g., inclusion of under-
264 represented lineages) in multiple sequence alignments used to construct profile Hidden Markov
265 Models, such as those implemented in TIAMMAt, can further improve the sensitivity of sequence
266 similarity searches[91] (Table 2).
267
268 Another systematic error source is the asymmetry in rates of gene duplication and loss between
269 species, which can result in HIDDEN PARALOGY. At shallow evolutionary depths, hidden paralogy can
270 be detected by examining synteny. For example, examining the synteny of six yeast species that
271 underwent differential patterns of gene loss since a shared whole-genome duplication event revealed
272 that ~10% of inferred single-copy orthologs were hidden paralogs[92]. Detecting hidden paralogy
273 instances in deep time is more challenging because synteny is likely not conserved. In such cases,
274 hidden paralogs can potentially be detected by searching for gene trees where well-known clades
275 are not monophyletic[93,94]. Alternatively, because hidden paralogs can be quite divergent from the
276 rest of the sequences in an orthogroup, they can also be identified by examining gene trees for taxa
277 that have unexpectedly long terminal branches using software such as TreeShrink, PhyloFisher, and
278 PhyKIT[94–97] (Table 2). INPARALOGS, especially species-specific ones, can easily be handled by retaining
279 one of the two sequences, as implemented in PhyloTreePruner and OrthoSNAP[98,99].
280
281 Errors in ortholog inference can also stem from contaminated sequences in genome assemblies, a
282 key concern in metagenome-assembled genomes. The degree of contamination (and completeness)
283 of a given genome can be evaluated with the CheckM and miComplete programs[61,100] and
284 contaminant sequences can be removed prior to inference.
285

Modeling substitutions. Traditional substitution models are site-homogeneous models, which use one reversible substitution matrix and the same nucleotide / amino acid frequencies for all sites in a data matrix. Early nucleotide models assumed equal substitution rates and base frequencies[101] but later models incorporated biologically informed parameters, such as accounting for differences in the rates of transitions and transversions or base frequencies[102,103]. The most parameter-rich model among reversible models for nucleotide sequences is the generalized time-reversible model, which uses unequal substitution rates and unequal base frequencies[104]. Nucleotide substitution models that relax the assumptions of reversibility (i.e., 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[105].

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[106]; other substitution models—such as WAG and LG—are derived by estimating substitution rates from larger, more diverse databases of amino acid sequence alignments like Pfam[107,108].

Determining the best-fitting nucleotide and amino acid substitution models is often done using likelihood ratio tests and Akaike or Bayesian information criteria[109]. The latter outperform likelihood ratio tests but also have their shortcomings resulting, at times, in the wrong model being favored[110]. Of note, model fit does not always predict phylogenetic tree accuracy, and models of variable fit can sometimes result in consistent phylogenetic trees[111]. 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[112]. In general, the modeling 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 modeling sequence evolution is through direct experimental measurement—mutagenesis, functional selection, and deep sequencing. These experimentally

derived models have substantially improved fit compared to those with few or hundreds of parameters[113].

Partitioning concatenated data matrices—i.e., 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[114]. Supermatrices can be partitioned by biological features (e.g., genes or codon positions) or be algorithmically defined[115]. 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 artifacts[116,117]. Consequently, site-heterogeneous models have risen in popularity and helped resolve the placement of several anciently diverged lineages[118,119], but are also the focal point of controversies such as the rooting the animal tree (Box 1). In other cases, using site-heterogeneous models has shed light on the evolutionary relationships among life's three domains, supporting the hypothesis that eukaryotes originated from within Archaea (the two-domain hypothesis)[120].

Substitution model misspecification can bias topology estimation, contributing to incongruence[15,121–123] (Fig. 3c, Table 1). One well-known source of incongruence that stems from model misspecification is long-branch attraction[124,125]. 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[126] and parasitic fungi[127]. Outgroup taxa may also introduce long branches, increasing the potential for long-branch attraction artifacts (see next section). In addition to using site-heterogeneous models[124], long-branch attraction artifacts can sometimes also be ameliorated by including taxa whose placements break long branches[128,129] (see also *Taxon sampling* section). Notably, long-branch attraction can also occur when models are correctly specified and be exacerbated when partitioning phylogenomic datasets[125].

Other approaches attempt to approximate true processes of sequence evolution better. For example, HETEROTACHY, which is not accounted for by either site-homogeneous or heterogeneous models[130], can decrease phylogenetic accuracy due to long-branch attraction artifacts[125,131]. The General Heterogeneous evolution On a Single Topology (or GHOST) model of sequence evolution can account for heterotachy, in part, 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[6].

Rooting strategy. Rooting strategies have been debated for a long time, especially in the context of outgroup taxa driving long-branch attraction artifacts[132]. 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[133], it is broadly accepted that thorough sampling of representatives of diverse lineages improves phylogenetic inference[134].

Other methods aim to infer the root of a phylogenetic tree without using outgroup taxa. These include the use of paralogs such as implemented in the software STRIDE[135–137], nonreversible Markov models such as the one implemented in the software Root Digger[138,139], relaxed molecular clock models as implemented in BEAST[140], the minimal ancestor deviation method that is also molecular clock-based[141], and modeling dynamics of gene family evolution[39]. For example, modeling 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[39].

Treatment errors

Multiple sequence alignment. Errors in multiple sequence alignment can result in inaccurate phylogenetic inferences and incongruence[142,143]. Alignment errors can stem from errors in ortholog 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[144] (Fig. 3b). Approaches to remedy errors in multiple sequence alignments include alignment trimming (see next section), probabilistic modeling to identify clusters of homologous characters and dividing the alignment accordingly (as implemented in Divvier[145]) or masking putative errors in multiple sequence alignments using two-dimensional outlier detection methods (as implemented in TAPER[146]).

Alignment trimming. Although trimming of sites during multiple sequence alignment is a widespread practice for reducing errors in multiple sequence alignment, it can also reduce the accuracy of phylogenetic inference, increase statistical uncertainty, and lead to incongruence (Fig. 3b, Table 1).

Generally, more aggressive alignment trimming that removes larger numbers of sites increases errors in single gene tree inferences[147]. 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[147,148]. Recently developed approaches that focus on retaining phylogenetically informative sites—such as ClipKIT (Table 2)—are equally accurate and more time-saving than no-trimming approaches[148].

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[149–152] (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[153,154]. Thus, recoding can also increase, rather than ameliorate, error. Appropriate ways forward include adequately assessing how recoding impacts 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[153]. Notably, errors in multiple sequence alignment, excessive trimming, and inappropriate character recoding all contribute to erosion of phylogenetic signal.

Concatenation vs. coalescence. Phylogenomic data matrices can be analyzed as a single supermatrix, an approach known as concatenation, or each gene alignment can be analyzed separately under the multispecies coalescent framework, an approach known as coalescence. The two approaches sometimes yield different tree topologies, contributing to incongruence[66,155]. Determining which approach is more appropriate for a phylogenomic dataset is difficult. For example, using simulated multilocus data, concatenation slightly outperformed a full coalescent-based approach (wherein gene trees and species trees are coestimated), whereas using coalescent independent sites, both approaches performed comparably[156]. Moreover, there can be differences in the performance of full 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, but newer implementations of summary coalescent-based methods take gene tree uncertainty into account[28]. Analyses with both full and

summary coalescent-based methods can be improved through targetted data filtering, such as removing loci with low phylogenetic informativeness[157]. Loci that are inconsistent between concatenation- 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 over- and under-flows (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 low phylogenetic signal (i.e., 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[148,159]. Studies that further our understanding of the accuracy and information content of multiple sequence alignments may facilitate predicting genes with greater phylogenetic signal[75,161–163].

Detecting incongruence

Because several biological and analytical factors, often initially unknown, can contribute to incongruence, several methods examine the presence and magnitude of incongruence *per se* in phylogenomic datasets without assuming the presence of a specific underlying biological or analytical factor(s).

Measures of branch support. Traditional approaches, such as nonparametric 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 a binary presence/absence—is more accurate for assessing support among deep branches in datasets with large numbers of taxa[166]. The usefulness of many of these measures in concatenation analyses of phylogenomic datasets is rather low because they almost invariably yield absolute support values, even if there is substantial incongruence between sites or loci[77]. However, these measures are highly informative when using summary coalescent-based methods to remove loci with low amounts of phylogenetic signal[167].

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[94,168]. 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[57,172] (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[173–176]. Internode certainty measures can help identify branches with substantial conflict, which can be then further examined 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[177] (Table 2).

Phylogenetic networks. Evolutionary relationships among organisms are often depicted as bifurcating trees, but this 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[178]—for example, split networks depict all possible splits in a set of phylogenies[179]; reticulate networks depict putative evolutionary events, such as hybridizations[180]. Software for inferring phylogenetic networks include SplitsTree[181], PhyloNet[182], and NetRAX[183] (Table 2).

Incongruence search protocols. In addition to the above methods, several protocols have been used to search for incongruence in phylogenomic datasets. These include repeated subsampling of smaller subsets of loci with robust phylogenetic signal and re-inference of the species phylogeny[**162**], gene genealogy interrogation[**184**], examination of phylogenetic signal[**185**], and quartet sampling[**186**].

Polytomies. Several clades in the tree of life, such as cichlids and finches, have experienced elevated rates of speciation giving rise to EVOLUTIONARY RADIATIONS. Such clades have often been influenced by multiple biological (e.g., introgression, lineage sorting) and analytical (e.g., long branch attraction for ancient radiations) factors, making phylogenomic inference particularly challenging and often present as a POLYTOMIES. Polytomies can be detected by identifying cases of equal support for multiple distinct topologies in sets of single gene trees[**94,187**]. Support can be measured using gene trees or the quartets of taxa present in these gene trees using ASTRAL[**82**], PhyKIT[**172**], and IQ-TREE[**57**] (Table 2).

Future Directions

Our knowledge of the tree of life, and 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 identified 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[31,157,160], a general framework for assessing 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 assessing any factor's relative importance because our ability to detect their effects varies across time scales. For example, incomplete lineage sorting and hybridization are biological factors that likely contribute to incongruence of ancient and recent relationships but are typically detectable only in studies of recently diverged lineages. In 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 ortholog inference or multiple sequence alignment are greater contributors to incongruence when studying ancient divergences than recent ones[188,189]. 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 (see also Box 1). A related issue is identifiability, that is figuring out why the observed conflict should be ascribed to certain factors and not others. For example, ancient horizontal gene transfer is often difficult to distinguish from gene duplication followed by extensive gene loss; attributing incongruence to one factor and ruling out another is challenging and often depends on *a priori* knowledge on which process is more likely. Developing methods and computational pipelines that enable simultaneous evaluation of potential contributing factors will be key for fully understanding the drivers of incongruence.

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

As the amount of genomic data increases, phylogenomic studies sampling several hundreds to thousands of organisms are becoming commonplace. One challenge with inferring phylogenies from such taxon-rich datasets is that tree space is vast, making computation challenging. For example, the numbers of possible unrooted trees for three, five, seven, and nine taxa are one, 15, 945, and 135,135, respectively. As tree space grows, the likelihood of finding the nonoptimal tree increases, leading to speed-accuracy trade-offs and incongruence. Efficiently searching tree space, however, is key to finding an optimal tree; phylogenetic inference programs 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[190]. Moreover, gene-rich datasets 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[163,191–193], 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.

Data and datasets 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 underrepresented in genomic databases, will help reduce the impact of hidden paralogy and orthology in phylogenomic datasets. Denser datasets 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 dataset of 100 molecular markers but strong in a 5,000-marker dataset.

Mitigating errors in dataset construction

Errors can be introduced at all stages of phylogenomic analyses, including data matrix construction, and contribute to incongruence. 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 for reproducibly handling phylogenomic data matrix construction will greatly facilitate comparative analyses of analytical drivers of incongruence across studies.

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[194]. We can reduce the carbon footprint of phylogenomics through judicious use of computing infrastructure, careful experimental design, and software choice. For example, evaluating substitution model fit using fast and robust software like ModelTest-NG[195] and jModelTest[196] can result in a 90% reduction in energy use, resulting in 10% less greenhouse gas emissions[197]. Similarly, choosing faster programs in quantifiably difficult-to-analyze datasets does not alter the quality of inference but can save energy[198].

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.

Competing interests

J.L.S. is a scientific consultant for Latch AI Inc. J.L.S. is a scientific advisor for WittGen Biotechnologies. J.L.S. is an advisor for ForensisGroup Inc. A.R. is a scientific consultant for LifeMine Therapeutics, Inc.

Table 1 | Drivers of incongruence.

Driver of incongruence	Factor	Literature about topic
Sampling, taxon and locus	Analytical, Stochastic error	81,199,200
Insufficient number of genes or divergent sites	Analytical, Stochastic error	2,7,9,78
Erroneous ortholog detection	Analytical, Systematic error	93,96,201–203
Model misspecification	Analytical, Systematic error	6,124,125,204
Multiple sequence alignment errors	Analytical, Treatment error	142,143
Excessive trimming	Analytical, Treatment error	147,148
Inappropriate character recoding	Analytical, Treatment error	205,206
Incomplete lineage sorting	Biological	18,22,207
Horizontal gene transfer	Biological	34,208–210
Hybridization / Introgression and Recombination	Biological	41,42
Natural selection	Biological	55,56

593 **Table 2** | Tools for investigating incongruence large genomic data sets.

Software/Method	Utility category	Utility details	Reference
Bag of little bootstraps	Bipartition support metric	Median bagging of bootstrap support assessed using few little samples and small subset of sites is a rapid method to infer bootstrap trees and provides similar patterns of support compared to traditional bootstrapping procedures	211
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 certainty / Tree certainty	Bipartition support metric	Identifies bipartitions in a reference phylogeny that also have a well-supported alternative topology	173–175
UFBoot2	Bipartition support metric	Ultrafast bootstrap approximations that are robust to model violation	212
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.	57–59
RERconverge	Convergent sequence evolution	Identifies genes in phylogenomic data matrices with signatures of convergent relative evolutionary rates in lineages with similar phenotypes	213
ClipKIT	Data processing and analysis	Multiple sequence alignment trimming wherein informative sites are retained rather than removing highly divergent sites	148
Concaterpillar	Data processing and analysis	Identifies congruent loci in a phylogenomic data matrix	214
ConJak	Data processing and analysis	Identifies sequence outliers compared to the central mean of a phylogenomic data matrix	215

ConWin	Data processing and analysis	Tests for within protein incongruence using a sliding window approach	215
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	216
PhyloFisher	Data processing and analysis	Collection of scripts for dataset building and trimming phylogenomic data sets. Also features a database of eukaryotic orthologs	97
RogueNaRok	Data processing and analysis	Identification of rogue taxa in a phylogenomic dataset	70
Root Digger	Data processing and analysis	Uses a non-reversible Markov model to calculate the likelihood of the root position in a tree	217
TreeShrink, PhyloFisher, and PhyKIT	Data processing and analysis	Identifies spurious orthologs from unexpectedly long terminal branches	96,216,218
abSENSE	Homology/ortholog detection	Calculates probability that homolog detection may fail	88
BLAST	Homology/ortholog detection	Searches for similar sequences by using measures of local similarity	219
Leapfrog	Homology/ortholog detection	Combines over split orthologs using reciprocal best BLAST hits	89
OrthoFinder	Homology/ortholog detection	Infers groups of orthologous genes	201
OrthoSNAP and DISCO	Homology/ortholog detection	Decompose multi-copy gene families into subgroups of single-copy orthologous genes	99,220
Profile Hidden Markov Models	Homology/ortholog detection	Probabilistic inference method that accounts for position-specific variation in sequences	90
TIAMMAt	Homology/ortholog detection	Increases sensitivity of sequence similarity searches by incorporating underrepresented lineages in profile Hidden Markov Models	91

ASTRAL and PhyKIT	Hypothesis testing	Both pieces of software enable researchers to conduct polytomy testing at a specific bipartition in a phylogeny	27,216
Gene- and site-wise log likelihood scores; gene-wise quartet scores	Hypothesis testing	Allows researchers to examine gene- and site-wise support between two topologies using maximum likelihood; gene-wise support can also be examined using quartet scores	158,221
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	42,49,50
NetRAX	Phylogenetic network inference	Maximum likelihood inference of phylogenetic networks when incomplete lineage sorting is not a factor	183
PhyloNet	Tree inference	Maximum parsimony, maximum likelihood, and Bayesian inference of phylogenetic networks from locus tree estimates	222
SplitsTree	Phylogenetic network inference	Splits graph inference using multiple sequence alignments, distance matrices, or sets of trees	181
General Heterogeneous evolution On a Single Topology model	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	6
QMaker	Substitution models	Estimates general time-reversible protein matrices—which describe rates of substitutions between amino acids—from multiple sequence alignments	204
Asteroid	Tree inference	Supertree method for species tree inference that is robust to missing data	223

ASTRAL, ASTRAL-PRO and ASTER	Tree inference	Quartet-based supertree method that accounts for partial gene trees, paralogs, and gene tree uncertainty	27,224,225
BEAST	Tree inference	Bayesian approach for phylogenetic tree inference and divergence time estimation	226
BPP	Tree inference	Full-likelihood implementation of the multispecies coalescent	227
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	86
MP-EST	Tree inference	Maximum pseudo-likelihood approach for species tree inference	228
PhyloBayes MPI	Tree inference	Bayesian tree inference method that incorporates finite and infinite mixture models to account for site variation	229
RAxML-NG	Tree inference	Maximum likelihood tree inference method that uses a greedy tree search algorithm to explore tree space	230
STAR	Tree inference	Inference of species trees using average ranks of coalescences	231
SVDQuartets	Tree inference	Inference of relationships using quartets and the coalescent model	232

Figure legends

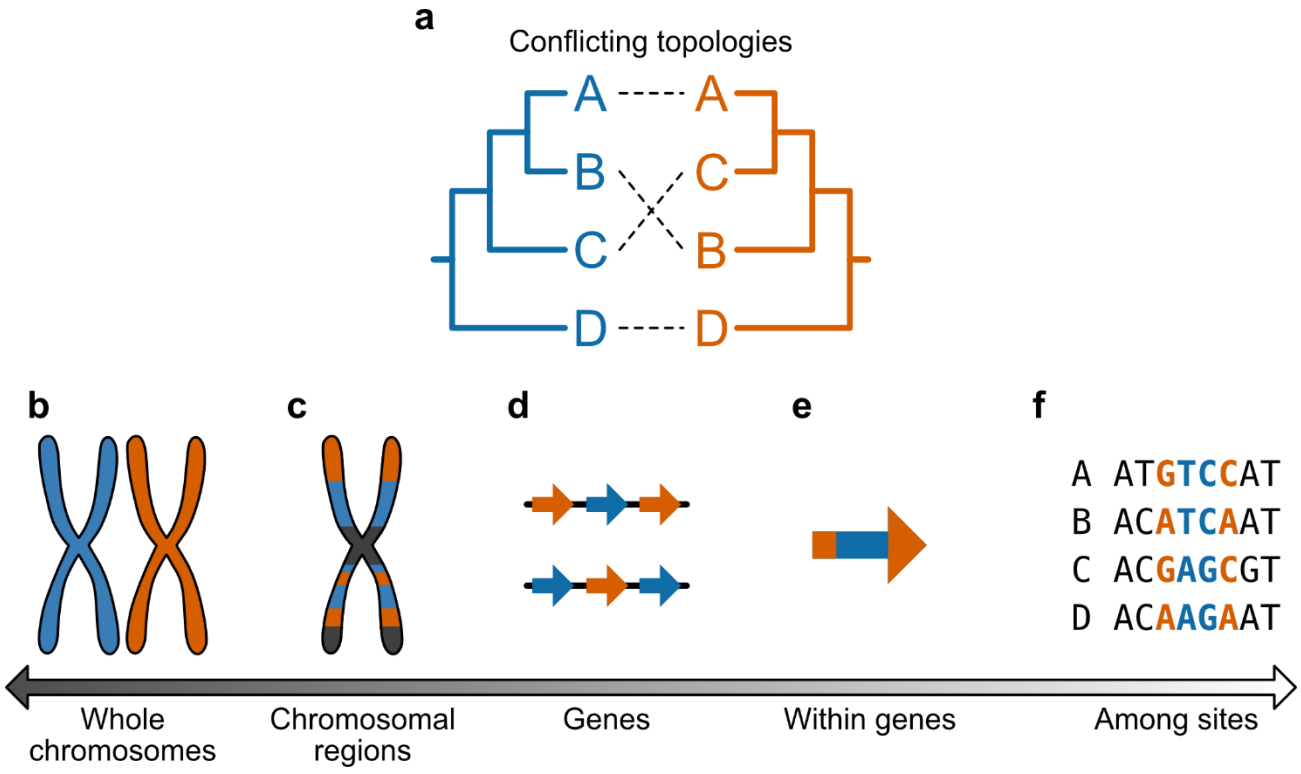


Figure 1 | Incongruence at different levels of genomic organization. a | The topology shown in blue supports a sister group relationship of taxa A and B, whereas the orange topology supports a sister group relationship of taxa A and C. The inference of such conflicting topologies defines incongruence.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

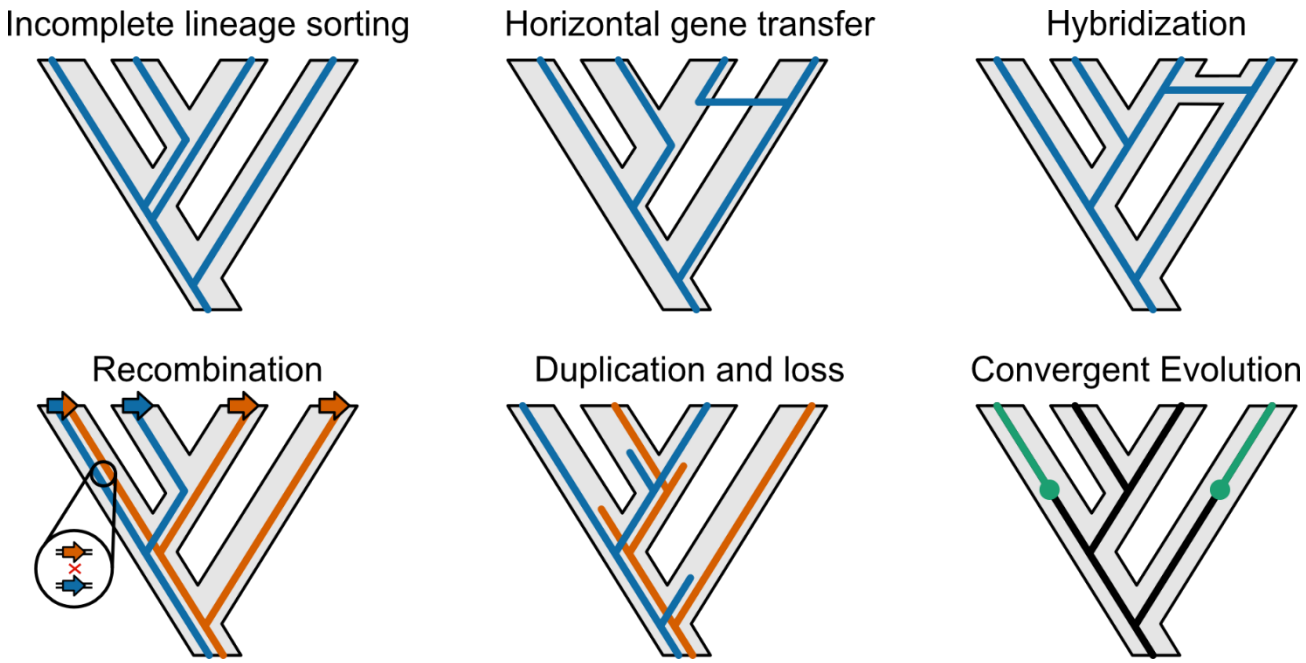


Figure 2 | Major biological factors that contribute to incongruence. *Incomplete lineage sorting* can lead to to gene trees that differ from the species phylogeny due to variation in the sorting of ancestral polymorphisms. *Horizontal gene transfer*, *hybridization*, and *introgression* can all lead to gene phylogenies that differ from the species tree. *Recombination* can result in loci with chimeric evolutionary histories. *Duplication and loss* can lead to hidden paralogy. Independently evolved traits in different phylogenetic lineages can be associated with *convergent molecular evolution* (green), contributing to incongruence.

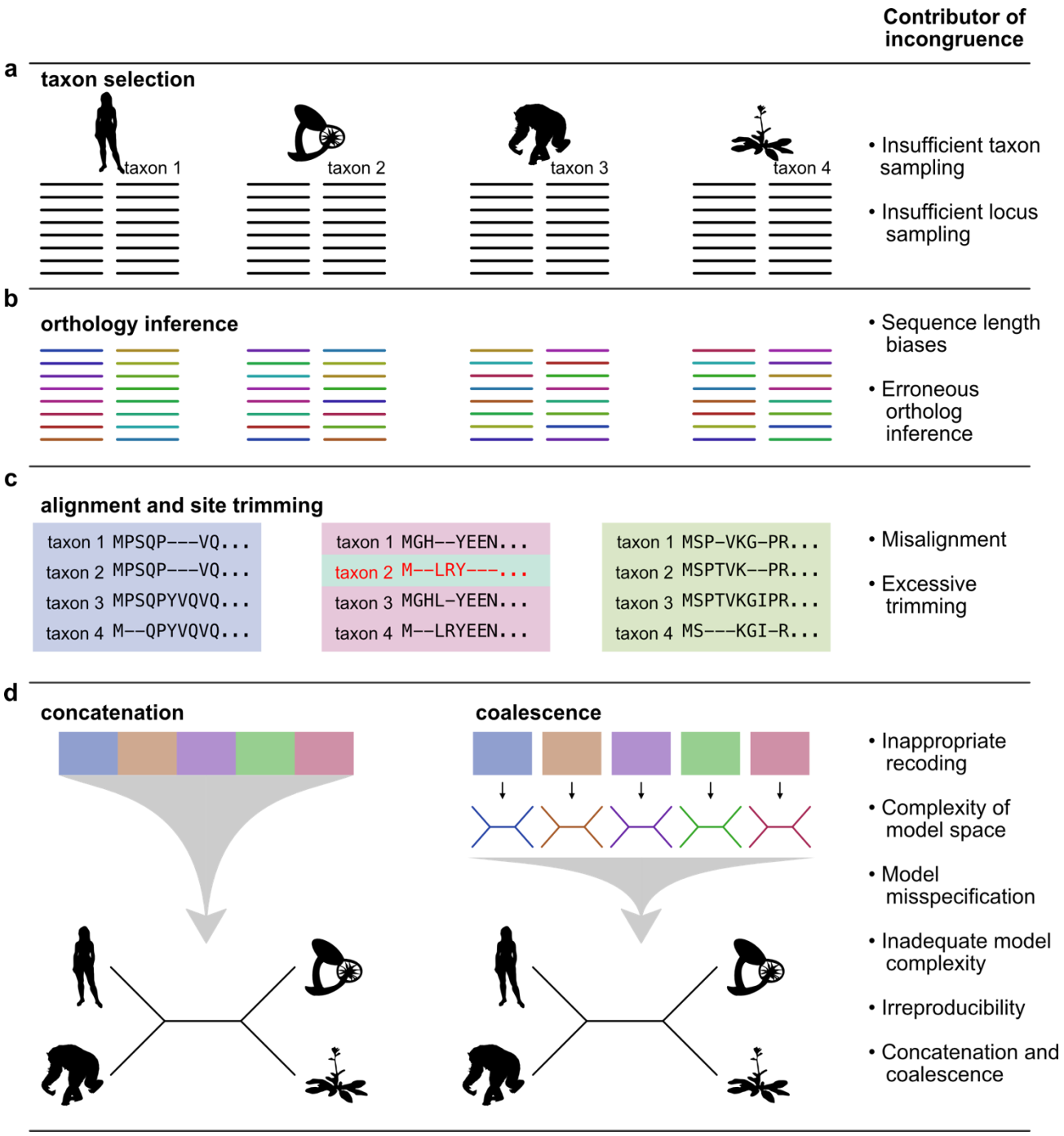
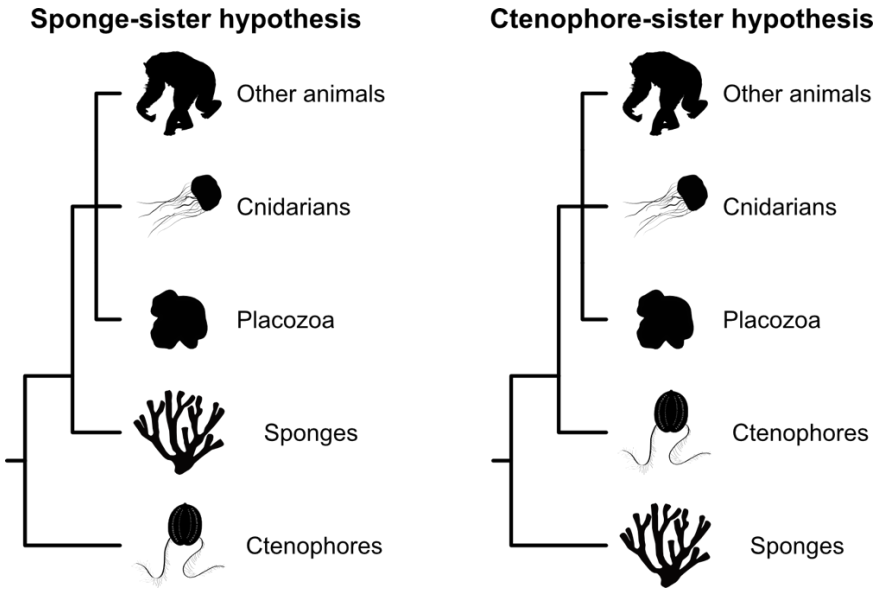


Figure 3 | **Analytical factors can contribute to incongruence at every step in a phylogenomic workflow**. **a** | Taxon sampling can impact all downstream analyses in phylogenomic studies. **b** | During orthology inference, biases (e.g., sequence length biases) and analytical errors (e.g., erroneous orthology inferences) can contribute to incongruence. Each color corresponds to a unique ortholog 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. An example of erroneous ortholog inclusion is depicted using red font. **d** |

624 Species tree inference via concatenation (left) or coalescence (right) is susceptible to multiple
625 additional sources of error—complexity of model space, model misspecification, and inadequate
626 model complexity, to name just a few.
627

Box 1 | **Rooting the animal tree**.



Few branches in the tree of life are as intensely debated as the root of the animal phylogeny. The two leading hypotheses debate whether sponges[93,233–236] or comb jellies (ctenophores)[10,14,65,200,237,238] 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 figure). Resolution of the root of the animal tree bears on our understanding of how animal cell types and tissues evolved[239]. Sponges lack muscles and a nervous system and are thought of as morphologically "simpler" animals compared to ctenophores, which have both[240,241]. Which hypothesis is correct also has implications for whether ctenophore nervous systems are structurally and genetically homologous to those of bilaterian animals[242,243], with some arguing that the ctenophore nervous system evolved independently[244].

Numerous biological and analytical factors contribute to this challenging phylogenetic problem. Much of the controversy has centered around whether site-homogeneous (with gene partitioning) or site-heterogeneous models of sequence evolution are most appropriate for reconstructing the animal phylogeny[200,245]. These models are largely employed to combat long-branch attraction, an artifact central to the debate because ctenophores have a long branch leading up to the lineage[246]. Site-heterogeneous models with many categories tend to support the sponge-sister hypothesis[13,247], whereas site-heterogeneous models with fewer categories and site-homogeneous models tend to support the ctenophore-sister hypothesis[247]. Some simulation analyses suggest that site-heterogeneous models underperform site-homogeneous models with gene partitioning[248] and others

suggest the opposite[246]. Aimed at reducing saturation and compositional biases, data matrix recoding analyses supported the sponge-sister hypothesis[152,249]; however, some of these analyses[249] failed to recover well-established monophyletic clades, such as Chordata, suggesting that analyses of non-recoded data were more accurate[250]. 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[13,14,152]. 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[200].

Several other factors, such as ortholog 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 contributed 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 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**[239]**. Similarly, other lines of evidence, such as investigations of synteny conservation using chromosome-level genome assemblies**[251]**, 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.

Glossary

CONVERGENT MOLECULAR EVOLUTION

Independent evolution of similar or identical molecular changes (e.g., 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 orthologs and paralogs (inparalogs and outparalogs) stemming from asymmetric patterns of duplication and loss

HORIZONTAL OR LATERAL GENE TRANSFER

The transfer of genetic material from one organism to another by mechanisms other than sexual reproduction

HYBRIDIZATION

The interbreeding of two distinct species or lineages

INCOMPLETE LINEAGE SORTING

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 more distantly related alleles

INPARALOG

Lineage- or species-specific paralogs 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 OR SUBSTITUTION

Models that describe rates of nucleotide or amino acid substitutions in a locus during evolution

OHNOLOGS

Paralogs that stem from a whole genome duplication event

OUTPARALOGS

Paralogs wherein the duplication event occurred before divergence from a reference common ancestor

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)

PHYLOGENOMICS

Defined initially as predicting gene function from phylogenies of homologous genes [252], the term was later expanded also to include phylogenetic inference using genome-scale amounts of data [253]

POLYTOMY

The node where more than two desendant lineages stem from an ancestral one

TAXON SAMPLING

Which and how many taxa are selected for a phylogenetic analysis

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

743 Lack of reproducibility of a tree topology between two replicate tree inferences using the same software
744 parameters (e.g., same model of sequence evolution, starting seed, etc.)
745
746 ROGUE TAXA
747 Taxa whose placement is unstable across a set of trees (e.g., across a set of gene trees)
748
749 STOCHASTIC ERROR
750 Error that occurs due to limited sampling and/or statistical uncertainty; can be eliminated by increasing the
751 amount of data
752
753 SYSTEMATIC ERROR
754 Error that occurs due to incorrect assumptions (e.g., model misspecification); it leads to bias in inference and
755 certainty in an incorrect result increases as larger amounts of data are used
756
757 TREATMENT ERROR
758 Error that stems from incorrect handling of data; depending on the source, it can result in stochastic or
759 systematic error
760
761
762
763

References

1. Simpson, G. G., 1902-. *The principles of classification and a classification of mammals*. ([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. *Nature Biotechnology* **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. Crotty, S. M. *et al.* GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. *Systematic Biology* (2019) doi:10.1093/sysbio/syz051.
7. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804 (2003).
8. Kawahara, A. Y. *et al.* Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences* **116**, 22657–22663 (2019).
9. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
10. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749 (2008).
11. Bond, J. E. *et al.* Phylogenomics Resolves a Spider Backbone Phylogeny and Rejects a Prevailing Paradigm for Orb Web Evolution. *Current Biology* **24**, 1765–1771 (2014).

12. Li, Y. *et al.* A genome-scale phylogeny of the kingdom Fungi. *Current Biology* **31**, 1653-1665.e5 (2021).
13. Simion, P. *et al.* A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology* **27**, 958–967 (2017).
14. Whelan, N. V. *et al.* Ctenophore relationships and their placement as the sister group to all other animals. *Nature Ecology & Evolution* **1**, 1737–1746 (2017).
15. Lemmon, A. R. & Moriarty, E. C. The Importance of Proper Model Assumption in Bayesian Phylogenetics. *Systematic Biology* **53**, 265–277 (2004).
16. Mao, Y. *et al.* A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**, 77–81 (2021).
17. 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).
18. Feng, S. *et al.* Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell* **185**, 1646-1660.e18 (2022).
19. Avise, J. C. & Robinson, T. J. Hemiplasy: A New Term in the Lexicon of Phylogenetics. *Systematic Biology* **57**, 503–507 (2008).
20. Maddison, W. & Knowles, L. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* **55**, 21–30 (2006).
21. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24**, 332–340 (2009).
22. Song, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences* **109**, 14942–14947 (2012).

23. 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).
24. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019).
25. 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).
26. 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).
27. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
28. Zhang, C. & Mirarab, S. *Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees*. http://biorxiv.org/lookup/doi/10.1101/2022.02.19.481132 (2022) doi:10.1101/2022.02.19.481132.
29. Morel, B., Williams, T. A. & Stamatakis, A. *Asteroid: a new minimum balanced evolution supertree algorithm robust to missing data*. http://biorxiv.org/lookup/doi/10.1101/2022.07.22.501101 (2022) doi:10.1101/2022.07.22.501101.
30. Kominek, J. *et al.* Eukaryotic Acquisition of a Bacterial Operon. *Cell* **176**, 1356-1366.e10 (2019).
31. Arnold, B. J., Huang, I.-T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* **20**, 206–218 (2022).

32. Gophna, U. & Altman-Price, N. Horizontal Gene Transfer in Archaea—From Mechanisms to Genome Evolution. *Annu. Rev. Microbiol.* **76**, 481–502 (2022).
33. Van Etten, J. & Bhattacharya, D. Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends in Genetics* **36**, 915–925 (2020).
34. Lapierre, P., Lasek-Nesselquist, E. & Gogarten, J. P. The impact of HGT on phylogenomic reconstruction methods. *Briefings in Bioinformatics* **15**, 79–90 (2014).
35. Wisecaver, J. H. & Rokas, A. Fungal metabolic gene clustersâ€"caravans traveling across genomes and environments. *Front. Microbiol.* **6**, (2015).
36. Sevillya, G., Adato, O. & Snir, S. Detecting horizontal gene transfer: a probabilistic approach. *BMC Genomics* **21**, 106 (2020).
37. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive Horizontal Gene Transfer in Bdelloid Rotifers. *Science* **320**, 1210–1213 (2008).
38. 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. U.S.A.* **109**, 17513–17518 (2012).
39. Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. U.S.A.* **114**, (2017).
40. Morel, B. *et al.* SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Molecular Biology and Evolution* **39**, msab365 (2022).
41. Zhang, D. *et al.* Most Genomic Loci Misrepresent the Phylogeny of an Avian Radiation Because of Ancient Gene Flow. *Systematic Biology* **70**, 961–975 (2021).
42. Hibbins, M. S. & Hahn, M. W. Phylogenomic approaches to detecting and characterizing introgression. *Genetics* **220**, iyab173 (2022).

43. Sang, T. & Zhong, Y. Testing Hybridization Hypotheses Based on Incongruent Gene Trees. *Systematic Biology* **49**, 422–434 (2000).

44. Langdon, Q. K., Peris, D., Kyle, B. & Hittinger, C. T. sppIDer: A Species Identification Tool to Investigate Hybrid Genomes with High-Throughput Sequencing. *Molecular Biology and Evolution* (2018) doi:10.1093/molbev/msy166.

45. Steenwyk, J. L. *et al.* Pathogenic Allodiploid Hybrids of Aspergillus Fungi. *Current Biology* **30**, 2495-2507.e7 (2020).

46. Steenwyk, J. L. *et al.* Pathogenic Allodiploid Hybrids of Aspergillus Fungi. *Current Biology* **30**, 2495-2507.e7 (2020).

47. Yu, Y., Dong, J., Liu, K. J. & Nakhleh, L. Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16448–16453 (2014).

48. Suvorov, A. *et al.* Widespread introgression across a phylogeny of 155 Drosophila genomes. *Current Biology* **32**, 111-123.e5 (2022).

49. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution* **28**, 2239–2252 (2011).

50. Pease, J. B. & Hahn, M. W. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology* **64**, 651–662 (2015).

51. Posada, D. & Crandall, K. A. The Effect of Recombination on the Accuracy of Phylogeny Estimation. *J Mol Evol* **54**, 396–402 (2002).

52. Bruen, T. C., Philippe, H. & Bryant, D. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics* **172**, 2665–2681 (2006).

53. Martin, D. P. *et al.* RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evolution* **7**, veaa087 (2021).
54. Sackton, T. B. & Clark, N. Convergent evolution in the genomics era: new insights and directions. *Phil. Trans. R. Soc. B* **374**, 20190102 (2019).
55. Li, Y., Liu, Z., Shi, P. & Zhang, J. The hearing gene Prestin unites echolocating bats and whales. *Current Biology* **20**, R55–R56 (2010).
56. Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences* **106**, 8986–8991 (2009).
57. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
58. Musil, M. *et al.* FireProtASR: A Web Server for Fully Automated Ancestral Sequence Reconstruction. *Briefings in Bioinformatics* **22**, bbaa337 (2021).
59. 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).
60. Martijn, J. *et al.* Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat Commun* **11**, 5490 (2020).
61. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
62. 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).

63. Riley, R. *et al.* Comparative genomics of biotechnologically important yeasts. *Proceedings of the National Academy of Sciences* **113**, 9882–9887 (2016).
64. Shen, X.-X. *et al.* Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3: Genes|Genomes|Genetics* **6**, 3927–3939 (2016).
65. Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**, 0126 (2017).
66. Shen, X.-X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**, 1533-1545.e20 (2018).
67. 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. *American Journal of Botany* **105**, 291–301 (2018).
68. Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111**, E4859–E4868 (2014).
69. Cheng, S. *et al.* Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell* **179**, 1057-1067.e14 (2019).
70. Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Systematic Biology* **62**, 162–166 (2013).
71. Struck, T. H. TreSpEx–-Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. *Evolutionary Bioinformatics* **10**, EBO.S14239 (2014).

72. Amemiya, C. T. *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).
73. Liu, S. *et al.* Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell* **184**, 4874-4885.e16 (2021).
74. Perri, A. R. *et al.* Dire wolves were the last of an ancient New World canid lineage. *Nature* **591**, 87–91 (2021).
75. Townsend, J. P. Profiling Phylogenetic Informativeness. *Systematic Biology* **56**, 222–231 (2007).
76. Patel, S. Error in Phylogenetic Estimation for Bushes in the Tree of Life. *Phylogenetics Evolutionary* **01**, (2013).
77. Rokas, A. & Carroll, S. B. Bushes in the Tree of Life. *PLoS Biology* **4**, e352 (2006).
78. Pipes, L., Wang, H., Huelsenbeck, J. P. & Nielsen, R. Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny. *Molecular Biology and Evolution* **38**, 1537–1543 (2021).
79. 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).
80. Willson, J., Roddur, M. S., Liu, B., Zaharias, P. & Warnow, T. DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. *Systematic Biology* (2021) doi:10.1093/sysbio/syab070.
81. Springer, M. S. & Gatesy, J. The gene tree delusion. *Molecular Phylogenetics and Evolution* **94**, 1–33 (2016).
82. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).

83. Sanderson, M. J., McMahon, M. M. & Steel, M. Terraces in Phylogenetic Tree Space. *Science* **333**, 448–450 (2011).
84. Xi, Z. *et al.* Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17519–17524 (2012).
85. 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).
86. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).
87. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 157 (2015).
88. 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).
89. 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).
90. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7**, e1002195 (2011).
91. Tassia, M. G., David, K. T., Townsend, J. P. & Halanych, K. M. TIAMMAt: Leveraging Biodiversity to Revise Protein Domain Models, Evidence from Innate Immunity. *Molecular Biology and Evolution* **38**, 5806–5818 (2021).

92. 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).
93. Philippe, H. *et al.* Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Current Biology* **19**, 706–712 (2009).
94. Steenwyk, J. L. *et al.* PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics (Oxford, England)* (2021) doi:10.1093/bioinformatics/btab096.
95. Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic “supergroups”. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3859–3864 (2009).
96. Mai, U. & Mirarab, S. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* **19**, 272 (2018).
97. Tice, A. K. *et al.* PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. *PLOS Biology* **19**, e3001365 (2021).
98. 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).
99. 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).
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. Jukes, T. H. & Cantor, C. R. Evolution of Protein Molecules. in *Mammalian Protein Metabolism* 21–132 (Elsevier, 1969). doi:10.1016/B978-1-4832-3211-9.50009-7.
102. 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).
103. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376 (1981).
104. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17**, 57–86 (1986).
105. Arenas, M. Trends in substitution models of molecular evolution. *Front. Genet.* **6**, (2015).
106. Yang, Z., Nielsen, R. & Hasegawa, M. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* **15**, 1600–1611 (1998).
107. Whelan, S. & Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution* **18**, 691–699 (2001).
108. Le, S. Q. & Gascuel, O. An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**, 1307–1320 (2008).
109. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772–772 (2012).
110. Susko, E. & Roger, A. J. On the Use of Information Criteria for Model Selection in Phylogenetics. *Molecular Biology and Evolution* **37**, 549–562 (2020).
111. Spielman, S. J. Relative Model Fit Does Not Predict Topological Accuracy in Single-Gene Protein Phylogenetics. *Molecular Biology and Evolution* **37**, 2110–2123 (2020).

112. Abadi, S., Azouri, D., Pupko, T. & Mayrose, I. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun* **10**, 934 (2019).
113. Bloom, J. D. An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit. *Molecular Biology and Evolution* **31**, 1956–1978 (2014).
114. Kainer, D. & Lanfear, R. The Effects of Partitioning on Phylogenetic Inference. *Molecular Biology and Evolution* **32**, 1611–1627 (2015).
115. 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* msw260 (2016) doi:10.1093/molbev/msw260.
116. Lartillot, N. & Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution* **21**, 1095–1109 (2004).
117. Si Quang, L., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
118. Stairs, C. W. *et al.* Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Current Biology* **31**, 5605-5612.e5 (2021).
119. 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. *Nature Communications* **12**, 4973 (2021).
120. 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).

121. Minin, V., Abdo, Z., Joyce, P. & Sullivan, J. Performance-Based Selection of Likelihood Models for Phylogeny Estimation. *Systematic Biology* **52**, 674–683 (2003).
122. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**, 303–314 (2012).
123. Sullivan, J. & Swofford, D. L. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution* (1997) doi:10.1023/A:1027314112438.
124. 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).
125. Susko, E. & Roger, A. J. Long Branch Attraction Biases in Phylogenetics. *Systematic Biology* **70**, 838–843 (2021).
126. 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).
127. Capella-Gutiérrez, S., Marcet-Houben, M. & Gabaldón, T. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biol* **10**, 47 (2012).
128. Graybeal, A. Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem? *Systematic Biology* **47**, 9–17 (1998).
129. Hillis, D. M. Inferring complex phytogenies. *Nature* **383**, 130–131 (1996).
130. Lopez, P., Casane, D. & Philippe, H. Heterotachy, an Important Process of Protein Evolution. *Molecular Biology and Evolution* **19**, 1–7 (2002).

131. Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. & Delsuc, F. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* **5**, 50 (2005).
132. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).
133. Geuten, K., Massingham, T., Darius, P., Smets, E. & Goldman, N. Experimental Design Criteria in Phylogenetics: Where to Add Taxa. *Systematic Biology* **56**, 609–622 (2007).
134. Pollock, D. D., Zwickl, D. J., McGuire, J. A. & Hillis, D. M. Increased Taxon Sampling Is Advantageous for Phylogenetic Inference. *Systematic Biology* **51**, 664–671 (2002).
135. Brady, S. G., Litman, J. R. & Danforth, B. N. Rooting phylogenies using gene duplications: An empirical example from the bees (Apoidea). *Molecular Phylogenetics and Evolution* **60**, 295–304 (2011).
136. 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).
137. Emms, D. M. & Kelly, S. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution* **34**, 3267–3278 (2017).
138. Naser-Khdour, S., Quang Minh, B. & Lanfear, R. Assessing Confidence in Root Placement on Phylogenies: An Empirical Study Using Nonreversible Models for Mammals. *Systematic Biology* **71**, 959–972 (2022).
139. Bettisworth, B. & Stamatakis, A. Root Digger: a root placement program for phylogenetic trees. *BMC Bioinformatics* **22**, 225 (2021).
140. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol* **4**, e88 (2006).

141. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* **1**, 0193 (2017).
142. Ashkenazy, H., Sela, I., Levy Karin, E., Landan, G. & Pupko, T. Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction. *Systematic Biology* **68**, 117–130 (2019).
143. Li-San Wang *et al.* The Impact of Multiple Protein Sequence Alignment on Phylogenetic Estimation. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **8**, 1108–1119 (2011).
144. Landan, G. & Graur, D. Characterization of pairwise and multiple sequence alignment errors. *Gene* **441**, 141–147 (2009).
145. Ali, R. H., Bogusz, M. & Whelan, S. Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. *Molecular Biology and Evolution* **36**, 2340–2351 (2019).
146. 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).
147. Tan, G. *et al.* Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Systematic Biology* **64**, 778–791 (2015).
148. 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 Biology* **18**, e3001007 (2020).
149. Susko, E. & Roger, A. J. On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Molecular Biology and Evolution* **24**, 2139–2150 (2007).

150. Blanquart, S. A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution. *Molecular Biology and Evolution* **23**, 2058–2071 (2006).
151. Phillips, M. J., Delsuc, F. & Penny, D. Genome-Scale Phylogeny and the Detection of Systematic Biases. *Molecular Biology and Evolution* **21**, 1455–1458 (2004).
152. Laumer, C. E. *et al.* Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. *eLife* **7**, e36278 (2018).
153. 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. *Systematic Biology* (2021) doi:10.1093/sysbio/syab027.
154. Foster, P. G. *et al.* Recoding amino acids to a reduced alphabet may increase or decrease phylogenetic accuracy. *Systematic Biology* syac042 (2022) doi:10.1093/sysbio/syac042.
155. 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**, (2019).
156. Wascher, M. & Kubatko, L. Consistency of SVDQuartets and Maximum Likelihood for Coalescent-Based Species Tree Estimation. *Systematic Biology* **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. *Systematic Biology* **68**, 573–593 (2019).
158. Shen, X.-X., Steenwyk, J. L. & Rokas, A. Dissecting Incongruence between Concatenation- and Quartet-Based Approaches in Phylogenomic Data. *Systematic Biology* **70**, 997–1014 (2021).

159. Darriba, D., Flouri, T. & Stamatakis, A. The State of Software for Evolutionary Biology. *Molecular Biology and Evolution* **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. *Nature Communications* **11**, 6096 (2020).
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 Biology and Evolution* **8**, 2565–2580 (2016).
162. Mongiardino Koch, N. Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci. *Molecular Biology and Evolution* (2021) doi:10.1093/molbev/msab151.
163. Haag, J., Höhler, D., Bettisworth, B. & Stamatakis, A. *From Easy to Hopeless - Predicting the Difficulty of Phylogenetic Analyses*. http://biorxiv.org/lookup/doi/10.1101/2022.06.20.496790 (2022) doi:10.1101/2022.06.20.496790.
164. Hillis, D. M. & Bull, J. J. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology* **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. *Systematic Biology* **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. *Systematic Biology* **67**, 285–303 (2018).

168. Minh, B. Q., Hahn, M. W. & Lanfear, R. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Molecular Biology and Evolution* **37**, 2727–2733 (2020).
169. Ane, C., Larget, B., Baum, D. A., Smith, S. D. & Rokas, A. Bayesian Estimation of Concordance among Gene Trees. *Molecular Biology and Evolution* **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. Steenwyk, J. L. *et al.* PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics* **37**, 2325–2331 (2021).
173. Salichos, L. & Rokas, A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).
174. Kobert, K., Salichos, L., Rokas, A. & Stamatakis, A. Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Molecular Biology and Evolution* **33**, 1606–1617 (2016).
175. Zhou, X. *et al.* Quartet-Based Computations of Internode Certainty Provide Robust Measures of Phylogenetic Incongruence. *Systematic Biology* **69**, 308–324 (2020).
176. Salichos, L., Stamatakis, A. & Rokas, A. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. *Molecular Biology and Evolution* **31**, 1261–1271 (2014).

177. Zhou, X. *et al.* Quartet-Based Computations of Internode Certainty Provide Robust Measures of Phylogenetic Incongruence. *Systematic Biology* **69**, 308–324 (2020).
178. Huson, D. H. & Bryant, D. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* **23**, 254–267 (2006).
179. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics (Oxford, England)* **14**, 68–73 (1998).
180. Huson, D. H., Klöpper, T., Lockhart, P. J. & Steel, M. A. Reconstruction of Reticulate Networks from Gene Trees. in *Research in Computational Molecular Biology* (eds. Miyano, S. et al.) 233–249 (Springer Berlin Heidelberg, 2005).
181. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
182. Wen, D., Yu, Y., Zhu, J. & Nakhleh, L. Inferring Phylogenetic Networks Using PhyloNet. *Systematic Biology* **67**, 735–740 (2018).
183. 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).
184. Arcila, D. *et al.* Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat Ecol Evol* **1**, 0020 (2017).
185. 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).
186. 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).

187. Sayyari, E. & Mirarab, S. Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. *Genes* **9**, (2018).
188. Ogden, T. H. & Rosenberg, M. S. Multiple Sequence Alignment Accuracy and Phylogenetic Inference. *Systematic Biology* **55**, 314–328 (2006).
189. 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).
190. Zhou, X., Shen, X.-X., Hittinger, C. T. & Rokas, A. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Molecular Biology and Evolution* **35**, 486–503 (2018).
191. Suvorov, A., Hochuli, J. & Schrider, D. R. Accurate Inference of Tree Topologies from Multiple Sequence Alignments Using Deep Learning. *Systematic Biology* **69**, 221–233 (2020).
192. Azouri, D., Abadi, S., Mansour, Y., Mayrose, I. & Pupko, T. Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nature Communications* **12**, 1983 (2021).
193. Rosenzweig, B. K., Hahn, M. W. & Kern, A. *Accurate Detection of Incomplete Lineage Sorting via Supervised Machine Learning*. http://biorxiv.org/lookup/doi/10.1101/2022.11.09.515828 (2022) doi:10.1101/2022.11.09.515828.
194. Grealey, J. *et al.* The Carbon Footprint of Bioinformatics. *Molecular Biology and Evolution* **39**, msac034 (2022).
195. Darriba, D. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution* **37**, 291–294 (2020).

196. Posada, D. jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution* **25**, 1253–1256 (2008).
197. Kumar, S. Embracing Green Computing in Molecular Phylogenetics. *Molecular Biology and Evolution* **39**, msac043 (2022).
198. Höhler, D., Haag, J., Kozlov, A. M. & Stamatakis, A. *A representative Performance Assessment of Maximum Likelihood based Phylogenetic Inference Tools*. http://biorxiv.org/lookup/doi/10.1101/2022.10.31.514545 (2022) doi:10.1101/2022.10.31.514545.
199. Nabhan, A. R. & Sarkar, I. N. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics* **13**, 122–134 (2012).
200. Li, Y., Shen, X.-X., Evans, B., Dunn, C. W. & Rokas, A. Rooting the Animal Tree of Life. *Molecular Biology and Evolution* **38**, 4322–4333 (2021).
201. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
202. 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).
203. Cheon, S., Zhang, J. & Park, C. Is Phylotranscriptomics as Reliable as Phylogenomics? *Mol Biol Evol* **37**, 3672–3683 (2020).
204. Minh, B. Q., Dang, C. C., Vinh, L. S. & Lanfear, R. QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Systematic Biology* **70**, 1046–1060 (2021).

205. 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. *Systematic Biology* **70**, 1200–1212 (2021).
206. Giacomelli, M., Rossi, M. E., Lozano-Fernandez, J., Feuda, R. & Pisani, D. Resolving tricky nodes in the tree of life through amino acid recoding. *iScience* **25**, 105594 (2022).
207. Scornavacca, C. & Galtier, N. Incomplete Lineage Sorting in Mammalian Phylogenomics. *Syst Biol* syw082 (2016) doi:10.1093/sysbio/syw082.
208. Galtier, N. A Model of Horizontal Gene Transfer and the Bacterial Phylogeny Problem. *Systematic Biology* **56**, 633–642 (2007).
209. Stolzer, M. *et al.* Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**, i409–i415 (2012).
210. 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. U.S.A.* **109**, 17513–17518 (2012).
211. Sharma, S. & Kumar, S. Fast and accurate bootstrap confidence limits on genome-scale phylogenies using little bootstraps. *Nat Comput Sci* **1**, 573–577 (2021).
212. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**, 518–522 (2018).
213. Kowalczyk, A. *et al.* RERconverge: an R package for associating evolutionary rates with convergent traits. *Bioinformatics* **35**, 4815–4817 (2019).
214. Leigh, J. W., Susko, E., Baumgartner, M. & Roger, A. J. Testing Congruence in Phylogenomic Analysis. *Systematic Biology* **57**, 104–115 (2008).

215. 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. *Systematic Biology* syac029 (2022) doi:10.1093/sysbio/syac029.
216. Steenwyk, J. L. *et al.* PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics (Oxford, England)* (2021) doi:10.1093/bioinformatics/btab096.
217. Bettisworth, B. & Stamatakis, A. Root Digger: a root placement program for phylogenetic trees. *BMC Bioinformatics* **22**, 225 (2021).
218. Tice, A. K. *et al.* PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. *PLoS Biol* **19**, e3001365 (2021).
219. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, (1990).
220. Willson, J., Roddur, M. S., Liu, B., Zaharias, P. & Warnow, T. DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. *Systematic Biology* **71**, 610–629 (2022).
221. Shen, X.-X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**, 0126 (2017).
222. Wen, D., Yu, Y., Zhu, J. & Nakhleh, L. Inferring Phylogenetic Networks Using PhyloNet. *Systematic Biology* **67**, 735–740 (2018).
223. Morel, B., Williams, T. A. & Stamatakis, A. *Asteroid: a new minimum balanced evolution supertree algorithm robust to missing data*. http://biorxiv.org/lookup/doi/10.1101/2022.07.22.501101 (2022) doi:10.1101/2022.07.22.501101.

224. Zhang, C. & Mirarab, S. *Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees*. http://biorxiv.org/lookup/doi/10.1101/2022.02.19.481132 (2022) doi:10.1101/2022.02.19.481132.

225. Zhang, C., Scornavacca, C., Molloy, E. K. & Mirarab, S. ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Molecular Biology and Evolution* **37**, 3292–3307 (2020).

226. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214 (2007).

227. Flouri, T., Jiao, X., Rannala, B. & Yang, Z. Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution* **35**, 2585–2593 (2018).

228. 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).

229. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology* **62**, 611–615 (2013).

230. 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).

231. Liu, L., Yu, L., Pearl, D. K. & Edwards, S. V. Estimating Species Phylogenies Using Coalescence Times among Sequences. *Systematic Biology* **58**, 468–477 (2009).

232. Chifman, J. & Kubatko, L. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics* **30**, 3317–3324 (2014).

233. Simion, P. *et al.* A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology* **27**, 958–967 (2017).
234. Redmond, A. K. & McLysaght, A. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nature Communications* **12**, 1783 (2021).
235. Pisani, D. *et al.* Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the National Academy of Sciences* **112**, 15402–15407 (2015).
236. Feuda, R. *et al.* Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology* **27**, 3864-3870.e4 (2017).
237. Ryan, J. F. *et al.* The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. *Science* **342**, (2013).
238. Moroz, L. L. *et al.* The ctenophore genome and the evolutionary origins of neural systems. *Nature* **510**, 109–114 (2014).
239. King, N. & Rokas, A. Embracing Uncertainty in Reconstructing Early Animal Evolution. *Current Biology* **27**, R1081–R1088 (2017).
240. Dunn, C. W., Leys, S. P. & Haddock, S. H. D. The hidden biology of sponges and ctenophores. *Trends in Ecology & Evolution* **30**, 282–291 (2015).
241. Nielsen, C. Early animal evolution: a morphologist's view. *Royal Society Open Science* **6**, 190638 (2019).
242. Liebeskind, B. J., Hillis, D. M., Zakon, H. H. & Hofmann, H. A. Complex Homology and the Evolution of Nervous Systems. *Trends in Ecology & Evolution* **31**, 127–135 (2016).
243. Sachkova, M. Y. *et al.* Neuropeptide repertoire and 3D anatomy of the ctenophore nervous system. *Current Biology* **31**, 5274-5285.e6 (2021).

244. Burkhardt, P. Ctenophores and the evolutionary origin(s) of neurons. *Trends in Neurosciences* **45**, 878–880 (2022).
245. Baños, H., Susko, E. & Roger, A. J. Are profile mixture models over-parameterized? *bioRxiv* 2022.02.18.481053 (2022) doi:10.1101/2022.02.18.481053.
246. Kapli, P. & Telford, M. J. Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. *Sci. Adv.* **6**, eabc5162 (2020).
247. Li, Y., Shen, X.-X., Evans, B., Dunn, C. W. & Rokas, A. Rooting the Animal Tree of Life. *Molecular Biology and Evolution* **38**, 4322–4333 (2021).
248. Whelan, N. V. & Halanych, K. M. Who Let the CAT Out of the Bag? Accurately Dealing with Substitutional Heterogeneity in Phylogenomic Analyses. *Syst Biol* syw084 (2016) doi:10.1093/sysbio/syw084.
249. 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).
250. 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).
251. Parey, E. *et al.* Genome structures resolve the early diversification of teleost fishes. *Science* **379**, 572–575 (2023).
252. Eisen, J. A. Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Res.* **8**, 163–167 (1998).
253. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**, 361–375 (2005).