A Practical Guide to Design and Assess a Phylogenomic Study

Jesus Lozano-Fernandez [1,2,*]

[1]Department of Genetics, Microbiology and Statistics, Biodiversity Research Institute (IRBio), University of Barcelona, Avd. Diagonal 643, 08028 Barcelona, Spain

[2]Institute of Evolutionary Biology (CSIC – Universitat Pompeu Fabra), Passeig marítim de la Barcelona 37-49, 08003 Barcelona, Spain

*Corresponding author: E-mail: jesus.lozano@ub.edu.

Accepted: 03 August 2022

Abstract

Over the last decade, molecular systematics has undergone a change of paradigm as high-throughput sequencing now makes it possible to reconstruct evolutionary relationships using genome-scale datasets. The advent of "big data" molecular phylogenetics provided a battery of new tools for biologists but simultaneously brought new methodological challenges. The increase in analytical complexity comes at the price of highly specific training in computational biology and molecular phylogenetics, resulting very often in a polarized accumulation of knowledge (technical on one side and biological on the other). Interpreting the robustness of genome-scale phylogenetic studies is not straightforward, particularly as new methodological developments have consistently shown that the general belief of "more genes, more robustness" often does not apply, and because there is a range of systematic errors that plague phylogenomic investigations. This is particularly problematic because phylogenomic studies are highly heterogeneous in their methodology, and best practices are often not clearly defined. The main aim of this article is to present what I consider as the ten most important points to take into consideration when planning a well-thought-out phylogenomic study and while evaluating the quality of published papers. The goal is to provide a practical step-by-step guide that can be easily followed by nonexperts and phylogenomic novices in order to assess the technical robustness of phylogenomic studies or improve the experimental design of a project.

Key words: systematics, systematic error, high-throughput sequencing, models of sequence evolution, phylogenetics, genomics.

Significance

The abundance of whole sequenced genomes, genomic fragments, and transcriptomes generated during the last decade has facilitated the interrogation of the interrelationships of all living beings at an unprecedented level. Nevertheless, big data methodologies are challenging to assess for systematists without expertise in molecular evolution and bioinformatics, a key aspect since the analysis of these massive amounts of information is subject to different kinds of errors that may strongly bias the inferred phylogeny. In this perspective, are discussed what I consider as the 10 most relevant points when planning a genome-scale phylogenetic project and assessing the quality and robustness of phylogenomic results.

Introduction

The establishment of evolutionary relationships by means of phylogenetic reconstruction is the basis to understand how species evolved and diversified. The output of these analyses is phylogenetic trees, which represent hypotheses of evolutionary relationships. Traditional approaches to establish phylogenetic relationships relied on comparing homologous morphological characters between organisms (Scotland et al. 2003). Thanks to the development and standardization of the polymerase chain reaction

Glossary

- **Complex models:** Models that handle the heterogeneity of substitution patterns by assuming that all loci or sites do not share the same substitution process. Therefore, they allow differentiated substitution processes at the gene or partition level (in the case of partition models) or site level (profile mixture models accommodating heterogeneity across sites).
- **Compositional heterogeneity:** Differences in nucleotide or amino acid composition of sequences between taxa, which may induce systematic error in phylogenetic inference. Taxa with randomly shared similarity in base composition due to convergent events may be artifactually clustered together.
- **Gene tree:** A phylogenetic tree resulting from the analysis of a single locus.
- **Gene duplication:** Event in which a gene is duplicated into multiple copies of itself. The duplicated genes may be retained in the genome and evolve independently.
- **Gene gain:** When a gene is only present in the branch leading to the last common ancestor of all the orthologs of that gene. Therefore, that gene lacks orthologs in any of the other clades present in that dataset and it originated at the onset of the lineage that contains it.
- **Gene loss:** When a gene has no homologs in a branch leading to a given clade, but homologs of that gene are present in their relatives.
- **Heterotachy:** Variations of the speed of evolutionary rates across time, sites and/or lineages.
- **Hidden paralogy:** Artifact affecting orthology inference in which paralogous genes are recovered as putative orthologs, affecting the phylogenetic reconstruction. It may be due to complex gene histories with multiple duplications and losses, or to incomplete datasets either caused by gene loss or partial sequencing of the genome/transcriptome.
- **Homologous sequences:** Sequences that originated from a common ancestor.
- **Horizontal gene transfer:** The nonvertical transfer of genetic material from a donor species to a receiver species.
- **Incomplete Lineage Sorting:** Phenomenon in which ancestral polymorphisms are retained and therefore coalesce deeper than speciation events. It causes discordance between the gene tree and species tree.
- **Long branch attraction (LBA):** Phylogenetic artifact in which rapidly evolving lineages are incorrectly inferred as closely related because they have undergone multiple molecular substitutions, and not because they are related by descent.
- **Model of sequence evolution (substitution models):** Models grounded in statistical theory that make use of explicit descriptions of the process of substitution in nucleotide or amino acid sequences. Profile mixture models are more fine-grained than global ones and better approximate the evolutionary process by including many parameters and being most computationally demanding.
- **Orthogroup:** Set of homologous sequences that are descended from a single ancestral sequence in the last common ancestor of all the taxa being considered.
- **Orthologous sequences:** Homologous sequences that diverged via speciation events.
- **Paralogous sequences:** Homologous sequences that diverged via duplication events.
- **Phylogenetic signal:** A measure of how much of the similarity between genetic sequences reflects common ancestry. A related concept is "phylogenetic noise", which describes the confounding signals in genetic sequences that cannot be used to reconstruct reliable phylogenies.
- **Saturation:** When there have been multiple mutations at the same site in a sequence alignment, and the apparent distances largely underestimate real genetic distances the alignment is said to be mutationally saturated. Nucleotide sequences saturate more rapidly than amino acids (four nucleotide bases versus 20 amino acids).
- **Sequencing depth:** The ratio of the total number of bases obtained by sequencing to the size of the genome or the average number of times each base is recorded in the genome.
- **Species tree:** A phylogenetic tree depicting the evolutionary relationship between a group of species.
- **Stochastic error:** Errors produced by insufficient amount of data.
- **Systematic error:** Errors during the phylogenetic inference mainly caused by incorrect model assumptions.

(commonly known as PCR), systematists started to interrogate molecular information to infer those relationships. Similarly to morphological characters, molecular data (in the form of nucleotide or amino acid sequences) can be used to build matrices of homologous positions (Hillis et al. 1996) and analyzed by phylogenetic methods. Starting in the 1980s, studies utilizing Sanger sequencing technology typically used a handful of molecular sequences to infer phylogenies. Although for some scientific questions using a few genes may be enough (e.g., biodiversity

inventories, barcoding studies, etc.), for others it is simply not sufficient, for instance when there are incongruent results among different studies and different loci (Gee 2003). Incongruence is common when studying speciation events that happened in a short amount of time. In the scenario of rapid speciation, the amount of phylogenetic signal (see Glossary) is often small, and therefore the internal branches are short, and hence phylogenies are more difficult to resolve and more data are needed (Philippe et al. 1994). Other examples in which few loci may not be enough to resolve relationships are those affecting deeper splits, in which multiple substitution may have occurred at the same position (i.e., homoplasy) (Rokas and Carroll 2006; Martin et al. 2016). In those instances, the results of using a few loci may not recover the "true" species relationships due errors, low information content, and to stochasticity in a small number of loci not mirroring the history of the species.

The advent of high throughput sequencing techniques in the last decade has aided in advancing the field of systematics by unlocking the access to massive amounts of sequence information (Metzker 2010). Scientists can now address phylogenetic hypotheses in depth and breadth by analyzing thousands of genes leveraged from genomic data, the so-called phylogenomic approach. Although the word "phylogenomics" was firstly coined in the context of gene function prediction using genome-scale data (Eisen 1998), shortly afterwards it was also applied to encompass phylogenetic inference using datasets of this magnitude (O'Brien and Stanyon 1999). Today, large amounts of sequence data from many living, and recently extinct, species are available in public repositories such as the National Center for Biotechnology Information (NCBI). Although the analysis of gene-rich datasets results in a drastic reduction of the random or sampling error, the reconstruction of the Tree of Life based on genome-scale data is not so straightforward. The analysis of big data comes at the price of an increasing methodological complexity, which hampers the critical appraisal of published pieces of work.

The inherent problems of inferring phylogenies using a few loci or using genome-scale data are substantially different. When relying on a small number of sequences, and therefore not many molecular characters, analyses might end up having low resolution or being poorly supported due to stochastic errors (see Glossary). Genome-scale datasets, instead, are less susceptible to stochastic or sampling error and often result in highly supported phylogenetic trees. However, other sources of error may strongly affect phylogenomic studies. On one hand, there are errors derived from the quality and appropriateness of data. On the other hand, there are problems stemming from the performance of the phylogenetic method, known as systematic error (see Glossary). This type of error is broadly derived by faulty assumptions on the analysis, such as when using models that do not properly describe the biological process of sequence evolution because certain model assumptions are violated (Philippe et al. 2011; Kapli et al. 2020; Simion et al. 2020). Systematic errors are consistently and repeatedly recovered unless the underlying biases are mitigated (Felsenstein 1978; Phillips et al. 2004), and the addition of more data can exacerbate their effect (Brown and Thomson 2017). An important issue concerns the correct modeling of the differences in substitution rates amongst nucleotide or amino acid sites (Lartillot and Philippe 2004) and across genes (Timmermans et al. 2016) and lineages (Foster 2004). Neither all sites in a gene evolve at the same pace nor all species do. These heterogeneities in evolutionary rates, combined with sequence saturation (see Glossary), in which hidden multiple substitutions that occurred at the same site of a sequence, can lead to phylogenetic reconstruction artifacts if not properly modelled. These errors might result in one of the most common artifacts in phylogenomics: the so-called long branch attraction (LBA) (see Glossary), a form of systematic error in which long branched taxa have a higher probability of clustering together artificially because of randomly shared similarity in base composition due to convergent or parallel changes. LBA is an artifact already present when dealing with a few molecular sequences, but the more molecular information we add, the more it can reinforce incorrect topologies. LBA, and systematic errors in general, can be addressed in various ways: denser taxon sampling (Zwickl and Hillis 2002; Heath et al. 2008), critically evaluating properties of the data, using more realistic models of sequence evolution, or selecting sets of the most reliable characters (Delsuc et al. 2005), among other things—as will be detailed below.

Other sources of error caused by true discrepancies between the history of specific genes or loci and the species phylogeny may also affect phylogenomic inference. The evolutionary dynamics of gene families is very complex, including gene duplications and losses, introgression or hybridization, or even cases of horizontally transferred genes (as opposed to vertically transmitted ones) (Maddison 1997), all of which often results in a discordant evolutionary histories of gene trees and species trees (Edwards 2009). Furthermore, the discordance between the gene and species tree might be caused by other processes such as incomplete lineage sorting (ILS) (see Glossary) due to retention of ancestral polymorphisms (Degnan and Rosenberg 2006; Edwards 2009), or the incorrect inclusion of paralogous sequences (see Glossary) as if these were orthologous in phylogenetic inference. For large datasets, there are two main model-based methods to infer the species tree using multiple sequence alignments of orthologous sequences (see Glossary). The first method, known as the supermatrix approach, is a "total-evidence" application that involves the concatenation of multiple orthologous sequences into a single alignment, assuming a commonly shared evolutionary history

(Rokas et al. 2003). This supermatrix is then analysed under a maximum likelihood (ML) or Bayesian framework using probabilistic methods that incorporate models of sequence evolution (Whelan et al. 2001). The second method involves using coalescent-based approaches (Rannala and Yang 2003). In a coalescent framework, genes may follow different histories with the species tree taking ILS into consideration (Liu et al. 2015).

As a result of these heterogeneous sources of gene tree discordance, and depending on how they are addressed in each study, phylogenomic studies using comparable datasets sometimes lead to contrasting results (Jeffroy et al. 2006; Gouy et al. 2015)—epitomized, among many others, by the ever-lasting conflict regarding the earliest splitting lineage in the Animal Tree of Life between the "ctenophora-sister" (Dunn et al. 2008; Ryan et al. 2013) and "sponge-sister" hypotheses (Pisani et al. 2015; Simion et al. 2017; Kapli and Telford 2020). Furthermore, there is no universal set of best practices in phylogenomics, which hampers the comparison of different studies. Although some recent reviews have addressed different aspects of phylogenomic inference (e.g., the accommodation of heterogeneous genomic signals (Bravo et al. 2019), the description of theory and main tools central to phylogenomics in insects (Young and Gillung 2020) or plants (McKain et al. 2018), or the exploration of major sources of error on the phylogenomic pipeline and strategies to mitigate them (Kapli et al. 2020; Simion et al. 2020)), these are usually highly technical and sometimes complex for the nonexpert in the field. In this article, I present and discuss what I consider the ten most important points that should be taken into consideration when assessing the reliability of the results of phylogenomic analyses or when setting up a new phylogenomic project aiming at inferring species phylogenies, with the goal of providing a comprehensive guidance to nonexperts of phylogenomics. Most of the suggestions are the reinterpretation of well-known phylogenetic practices but emphasizing their application into large-scale molecular datasets. Even though many of the examples used concern ancient divergences, the author ensured that they also apply to more recent splits. In addition to presenting an accessible synthesis of current phylogenomic practices, a flow diagram (fig. 1) is provided to help navigate the assessment of the methodology at each step of the analytical process.

1. Carefully Select the Taxa for Your Study (Including Outgroups)

This is an important point, as it should be addressed at the very beginning of the experimental design and has many downstream implications that can greatly influence the result of phylogenetic inference (Nabhan and Sarkar 2012). Early phylogenomic studies already highlighted that the number and choice of taxa included is as important as the number of molecular characters (Rokas et al. 2003), although in current datasets the latter routinely exceeds the number of taxa by orders of magnitude.

Assess the Diversity and Density of Your Taxon Sampling

Using a taxon sampling covering major lineages as much as possible generally improves the estimation of molecular rates and variation in base composition (Timmermans et al. 2016), except when including terminals with poor sequence data (that might lead to nonrandom distribution of missing data) or excessively unequal proportions of nucleotides or amino acids (see point 6 on ways of dealing with compositional heterogeneity). Furthermore, a carefully chosen taxon sampling can help improve the detection of hidden paralogy or horizontal gene transfer events (Kuzniar et al. 2008; Philippe et al. 2011) and in reducing problems associated with sequence saturation, which may lead to tree reconstruction errors. Saturation is driven by hidden multiple substitutions occurring at the same site, so models of molecular evolution can better resolve, or at least moderate, this problem when using a richer taxon set (Hendy and Penny 1989). The prevalent artifact of LBA can be ameliorated by adding more sequences and/or taxa that "break" these long branches (Graybeal 1998; Holton and Pisani 2010). The strategy of using a dense taxon set has been implemented successfully in phlyogenomic studies of eukaryotes. Using a rich taxon sampling, in conjunction with other corrective measures, led to the retrieval of the single-cell Microsporidia parasites as close relatives of Fungi, breaking the long branch that placed them on an artifactual position at the base of the eukaryotic tree (Foster et al. 2009). It is not possible to apply this strategy in all cases, though, as certain lineages are subtended by a long branch without any known living taxa that could break it.

Avoid Including Taxa with Highly Divergent Substitution Rate and Base Composition

In addition to including a high number of taxa, LBA can also be ameliorated by including sequences from slow-evolving taxa to eschew potential artifacts (Roure et al. 2013). This is particularly helpful when inferring the relationships of groups with ancient divergences, in which extinction processes may have erased informative intermediate states. Including these short-branch taxa helps increase the effective number of characters available for detecting hidden multiple substitutions (Hendy and Penny 1989), hence decreasing the amount of nonphylogenetic signal while preserving historical signal (Baurain et al. 2007; Philippe et al. 2011). A prime example of this practice resulted in the retrieval of the Ecdysozoa animal clade (Aguinaldo et al. 1997), now widely corroborated. This clade of moulting animals was retrieved after excluding long-branched

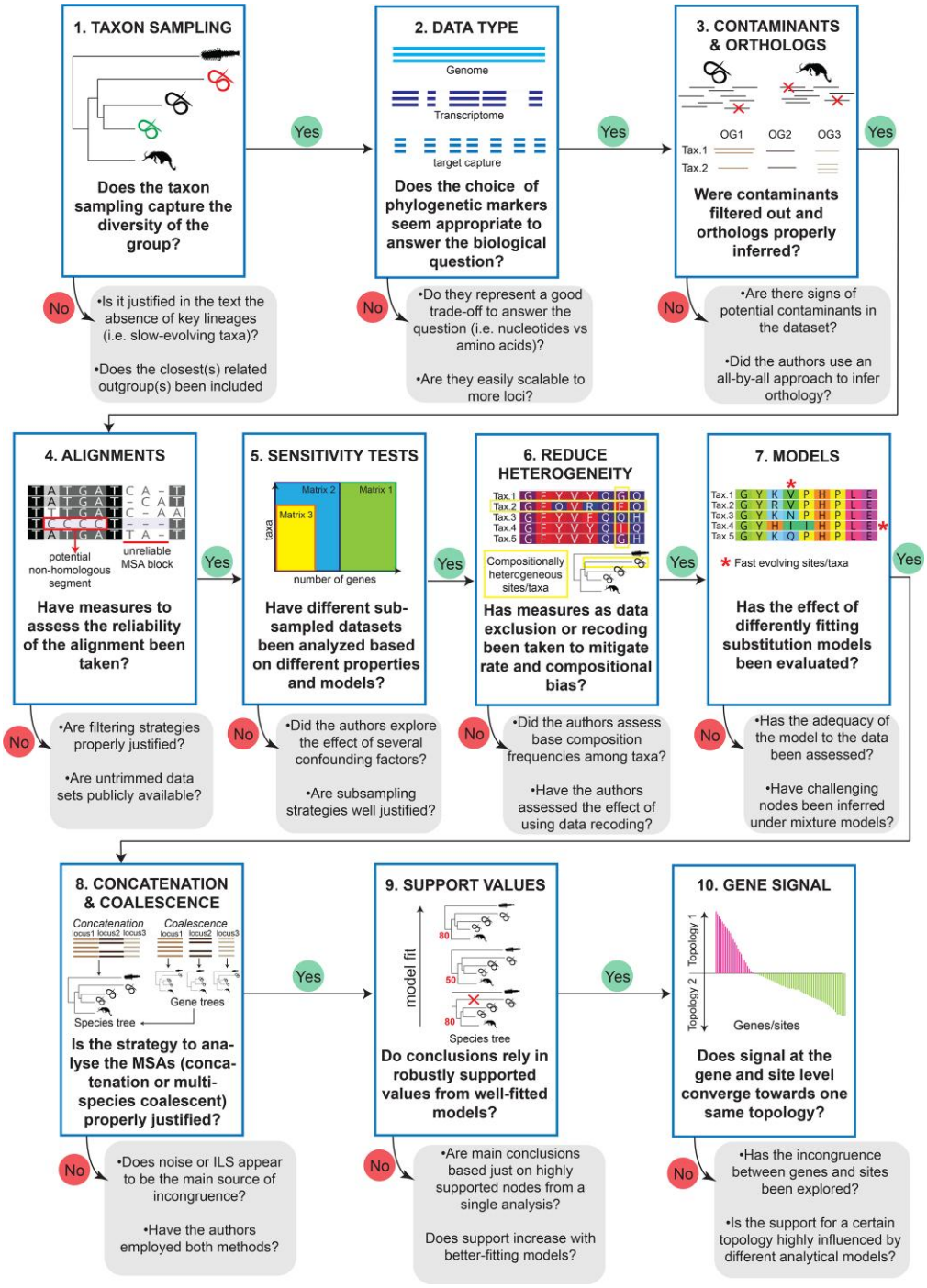

Fig. 1.—Flow diagram connecting 10 important points to take into consideration when planning and evaluating phylogenomic analyses. For each point, there are pictograms and yes/no questions to help navigating the review process and settinging up an experimental design. These questions are intended to assess globally the overall quality of each point. Negative replies take to more questions whose answers may lead to a better comprehension of the project and might serve as a starting point to improve the study. Silhouettes retrieved from Phylopic (phylopic.org).

nematodes from a single-gene dataset containing several animal phyla, keeping just slow evolving nematodes, and therefore correcting LBA artifacts that were affecting previous analyses. A second example is the inference of a sister-group relationship between the thermophilic bacterium *Thermus* and mesophilic *Deinococcus* recovered when a mesophilic relative of *Thermus* was included in the analyses (reviewed in Williams et al. 2021).

Table 1
Comparison of Sequencing Strategies in Terms of the Requirements They Place on the Analysed Samples and the Outputs They Produce

	Transcriptome Sequencing	Whole-Genome Sequencing	Genome Skimming	Target Capture
Technology				
Target molecule	RNA	DNA	DNA	DNA
Sequencing platform	Flexible	Preferably long-read technology	Flexible	Flexible
Samples				
Sample type	RNA has to be intact, i.e., fresh material or stored in specialised preserving solution	DNA has to be available (preferably) in sufficient quality	Can use samples in ethanol and museum specimens	Can use samples in ethanol and museum specimens
Sample amount	Enough RNA has to be available	Small specimens acceptable	Small specimens acceptable	Small specimens acceptable
Prior genomic resources required	No	No	No	Yes, to design probes
Recommended number of taxa	Flexible	Flexible	Flexible	More taxa are recommended
Genome size of taxa	Less relevant	Important	Important	Less relevant
Outputs				
Orthology inference	Confident	Confident	Less confident	Less confident
Ability to identify single-copy genes	Yes	Yes	Maybe	Maybe
Ability to analyse gene expression levels	Yes	No	No	No

Evaluate Your Outgroups

Phylogenies are rooted using taxa that are known to be outside the group under study. These taxa are known as outgroups, whereas the group under study is known as the ingroup. As outgroup sequences are often on long branches (commonly being more dissimilar to the ingroup sequences), long branched ingroup taxa tend to be attracted to the base of the tree (Holland et al. 2003; Bergsten 2005; Shavit et al. 2007; Dabert et al. 2010). Using the closest possible outgroup may reduce the impact of LBA on deep branches, as they will be less dissimilar to the ingroup (Philippe and Laurent 1998). This strategy has been suggested to counteract systematic biases in deep animal relationships (Philippe et al. 2011; Pisani et al. 2015; Simion et al. 2017). As a second example, the inclusion of the recently discovered Asgard archaea in phylogenomic analyses (Spang et al. 2015), a clade that emerged as the closest known prokaryotic relative of eukaryotes, has shifted the consensus toward two primary domains of life, Bacteria and Archaea, with eukaryotes nested within the latter instead of representing a third primary domain (Williams et al. 2020). Another strategy is to include several outgroups and experiment with subsets of them, to test if ingroup relationships change when certain outgroups are excluded (Shavit et al. 2007; Cox et al. 2008; Rota-Stabelli et al. 2011; Pisani et al. 2015).

2. Choose Your Destiny: Genomes, Transcriptomes, or Targeted Capture?

The choice of loci as phylogenetic markers is a crucial step in phylogenomic analyses (Reddy et al. 2017). Researchers have to decide 1) the molecule to sequence (DNA or RNA); 2) the sequencing strategy (transcriptomes, whole genomes, genome skimming, or target capture); and 3) the sequencing depth (table 1). Newly obtained sequences can be supplemented with genomic data available from public sequence databases such as the sequence read archive in NCBI. In the last decade, most sequences have been generated using short-read (50–300 nucleotides) sequencing platforms, such as Illumina. Emerging long-read sequencing platforms, such as SMRT Pacific Biosciences (PacBio) or Oxford Nanopore Technologies, are becoming more popular as they allow sequencing reads with median lengths over several thousand nucleotides (De Maio et al. 2019), making them more suitable for de novo assemblies. 22Short raw reads, are assembled to reconstruct contigs, with the most popular current softwares including SPAdes (Bankevich et al. 2012) and Velvet (Zerbino and Birney 2008) for genome assemblies, and Trinity (Grabherr et al. 2011) for transcriptome assemblies. Once assembled, the contigs can be used to extract molecular sequences and detect orthologs (see point 3).

Ideally, one would use complete and well-annotated genome assemblies containing all genes present in an organism (Petersen et al. 2017). Complete genomes reveal the full gene repertoire of an organism. Using fully sequenced genomes for orthology inference allows for a more confident assessment of gene histories by improving the detection of orthology and paralogy relationships, the estimation of gene gains and losses, and the detection of horizontal transfer events (see point 3). Complete genomes are the prevailing

source of sequence information in microbial phylogenomics, with an increasing number of genomes from uncultured taxa being reconstructed from metagenomic sequencing projects (Spang et al. 2015). However, whole genome data are still not frequently used in studies of eukaryotes due to their larger genome size and complexity. For example, most current phylogenomic studies on fungi or protists use draft genomes assemblies instead. Alternatively, the shallow sequencing of whole genomes at low-coverage, also known as genome skimming (Straub et al. 2012), may provide enough information to retrieve thousands of genes.

Phylogenomic analyses based on transcriptomic data—phylotranscriptomics—is one of the most common approaches, and transcriptomes have been the predominant source of large-scale phylogenomic studies focusing on deep divergences (Todd et al. 2016), even though it is currently changing towards using whole genomic data. Transcriptomes are routinely assembled from messenger RNA data in order to generate matrices of orthologous protein-coding genes. In the case of animals, RNA-seq datasets with more than 30 million reads seem to recover most of an organism's genes (Francis et al. 2013). Since most genes are expressed in almost all tissues, by sequencing the transcriptome from just a single tissue type we can obtain several thousand genes, as long as the sequencing depth (see Glossary) is adequate. Assembled transcriptomes can contain multiple isoforms for each gene, but for phylogenomic purposes just one of them is commonly retained, with the rest being removed. A common approach is to choose the longest isoform (Laumer et al. 2019) or the one with the highest read coverage as a representative of the gene (e.g., Trinity subcomponent Grabherr et al. 2011).

Amongst the ever-growing target enrichment methods, two major sources of phylogenetic markers include Hyb-Seq in plants and ultra-conserved elements (UCEs) in animals (Weitemier et al. 2014; Faircloth et al. 2012). Both methods rely on the construction of synthetic probes that hybridize with highly conserved genomic regions, which are sequenced along their flanking sequences—that are more variable and phylogenetically informative. One advantage of sequencing-capture techniques over transcriptomic methods is that the sample does not need to come from fresh tissue or having been fixed with RNA-preserving solutions, and therefore allows using samples fixed in ethanol, including museum-based specimens stored for long periods of time (McCormack et al. 2016). The use of target enrichment and genome skimming approaches may suffer analytical caveats mainly due to orthology inference (i.e., orthology inference may be challenging due bias in bait design and data processing can exacerbate hidden paralogy—see Glossary [Doolittle and Brown 1994; Rasmussen and Kellis 2012]).The type of phylogenomic markers selected for our study ultimately determines the subsequent analytical pipeline in three ways: 1) the possibility to analyse nucleotides, amino acids or both, 2) using exonic regions (portion of genes coding for amino acids), intronic regions (noncoding regions between exons in the genes), intergenic regions (DNA sequences between genes), or a combination of them, and 3) the extent of missing data that we may expect a priori in our dataset. All three conditions influence, to varying degrees, the accuracy of orthology inference (see point 3). The choice will be mainly influenced by the genomic organization of the group under study (i.e., genome size, proportion of genes and sequence conservation), the data type (i.e., falling in or outside of coding regions), and the divergences among the taxa of interest (i.e., whether we are aiming to resolve a shallow or deep phylogeny). The different types of molecular markers may be analysed at the nucleotide or amino acid level, with both kinds of molecules presenting different statistical properties (Huelsenbeck et al. 2008). Genome-based and transcriptomic datasets allow the analyses of protein-coding genes both at nucleotide and amino level. Amino acids are less susceptible to saturation than nucleotides (Philippe et al. 2011). Therefore, they are particularly useful to reduce the effects of saturation in molecular phylogenies (Whitfield and Lockhart 2007), which is common for ancient relationships (see point 6). Target enrichment techniques, like UCEs, are usually analysed at the nucleotide level because they can lie in nonprotein-coding parts of the genome. It has been shown that the regions that UCEs target can potentially be affected by GC biases, either by presenting high GC content or being heterogeneous amongst taxa, which can enhance topological conflict among gene trees (Bossert et al. 2017). Hence, it is advisable to assess whether the underlying incongruence may be caused by this kind of bias (Bossert et al. 2017).

When working with eukaryotes, the analysed data may belong either to coding (exonic) or noncoding (intronic) gene regions, each with their own characteristics. Exons commonly evolve under selection (Xing and Lee 2005), and thus are subject to many sources of heterogeneity (with selective pressures varying across sites, genes and over time). Intronic regions, instead, evolve more rapidly than coding sequences (Parenteau and Abou Elela 2019), and may quickly be subject to sequence saturation, causing difficulties in orthology assessment and during the alignment processes. In any case, protein sequences are under functional and structural selective constraints, and these are conserved over much longer periods than the individual codon choices.

As mentioned above, the robustness of orthology inference depends directly on the type of data. Using incomplete genomes or transcriptomes, as well as datasets containing misassemblies, can severely violate the underlying assumptions of orthology inference methods with existing heuristics (Yang and Smith 2014; Petersen et al. 2017). Under this scenario, distinguishing gene loss from missing data becomes even more difficult than when using

whole complete genomes, and paralogy and orthology cannot be accurately differentiated. As a consequence, hidden paralogy and inaccurate orthology and paralogy relationships are exacerbated (Rasmussen and Kellis 2012). Since loci recovery is determined by bait design in target enrichment techniques, and high levels of missing data are expected from genome skimming, these methods can be more prone to incorrect orthology inference than high-quality transcriptomes or completely sequenced genomes.

A common approach to assess the completeness and quality of the genomes or transcriptomes is to benchmark them against a clade-specific set of near-universal single-copy orthologs—BUSCO (Simão et al. 2015). In case of having multiple assemblies for a certain lineage, for example when there are multiple transcriptome projects for the same species, I recommend to select the one with highest the completeness—understood as having a higher percentage of complete BUSCO genes, ideally above 80%, or combine them. In some instances, as when maximizing taxon sampling, incomplete assemblies may still contain enough orthologous sequences to answer the question at hand, although maximizing completeness in the assemblies included in each dataset is strongly recommended.

3. Remove Contaminants and Make Sure Orthologs are Properly Inferred

Checking for Contamination

Common sources of unwanted nucleic acids are contaminants from other organisms, such as bacteria, parasites or endosymbionts, or cross-contamination between different samples processed together in the lab or in the same sequencing machine. These contaminants from undesired taxa can mislead the phylogenetic inference (Laurin-Lemay et al. 2012). A recently reported case of genome contamination, a wrongly purported massive acquisition of bacterial genes in tardigrades, warns about the perils of not eliminating contaminants before assembling the data (Koutsovoulos et al. 2016). One way of detecting foreign genetic material, and posteriorly removing them, is by mapping the sequences against public databases of known contaminants, such as certain bacteria or fungi (e.g., The NCBI Taxonomy database [Federhen 2012]). When sequences or scaffolds are almost identical to known contaminants, these nontarget DNA sequences can be filtered during the genome assembly stage with software packages such as BlobTools (Laetsch and Blaxter 2017). Softwares like CroCo (Simion et al. 2018), that rely on coverage, allow identifying and removing cross contaminants produced by biological samples from different species in assembled transcriptomes that have been processed or sequenced in parallel. For certain groups with fewer genomic resources, such as in protists, identifying contaminant sequences before assembling the data is harder. Nonetheless, it is possible to identify these contaminants, together with paralogous sequences, in later stages by visually inspecting gene trees for anomalous positions (expanded further in the following paragraph). Assessing the impact of contaminants in datasets of interest prior to further analysis is strongly recommended.

Orthology Inference

To reconstruct the species tree, it is a key to compare sequences that are orthologous (see Glossary). Orthologous sequences are derived from speciation processes and record the organismal phylogeny (Fitch 1970, 2000). Comparing sequences that are orthologous is crucial as their history reflects the species tree, and the inadvertent inclusion of paralogs—that derived from gene duplication (Ahrens et al. 2020)—may consequently distort phylogenetic reconstruction (Struck 2013; Siu-Ting et al. 2019). The correct inference of orthologs is not straightforward, because orthology and paralogy relationships are not transitive (except for one-to-one orthologs). For instance, consider one gene, gene *A*, that only has one copy in *Drosophila melanogaster*, but it got duplicated in *Apis mellifera* (we could then call them genes *A1* and *A2*). The *D. melanogaster* gene *A* is orthologous both to genes *A1* and *A2* in *A. mellifera*, however genes *A1* and *A2* are paralogs to each other since they originated through gene duplication. In this context, we move from pairwise comparisons to group orthology relationships (fig. 2). The main current softwares of homology inference aims to estimate the full set of orthologs and paralogs, known as orthogroups (see Glossary). Each orthogroup contains a set of sequences that are descended from a single one in the last common ancestor of all the species being considered (Emms and Kelly 2015). The inference of these orthogroups is nontrivial, and the methodologies vary a lot in their algorithms: graph-based vs. tree-based approaches, all-by-all vs. only analyzing a subset of taxa, clustering methodology, etc. The different methodologies have been covered and explained in several pieces of work (e.g., Altenhoff et al. 2016; Fernández et al. 2020) and hence we will just briefly touch upon them here. I recommend approaches that rely on doing an all-by-all comparison (in order to reduce biases associated with orthology inference) implemented in softwares such as OrthoFinder (Emms and Kelly 2015), and use those methods that had been favored in last benchmarking studies (e.g., Altenhoff et al. 2016; Emms and Kelly 2020).

Alternative pipelines to retrieve orthologs commonly use all versus all sequence similarity searches with BLAST (Tice et al. 2021), sometimes adding sequence data from genomes and transcriptomes to an existing set of multiple sequence alignments (i.e., Forty-Two—https://metacpan.org/release/Bio-MUST-Apps-FortyTwo as implemented in Irisarri et al. 2017). A recent study suggested that cases of undetected paralogy can remain in phylogenomic "single-copy orthogroup" datasets when using widely accepted all versus all BLAST

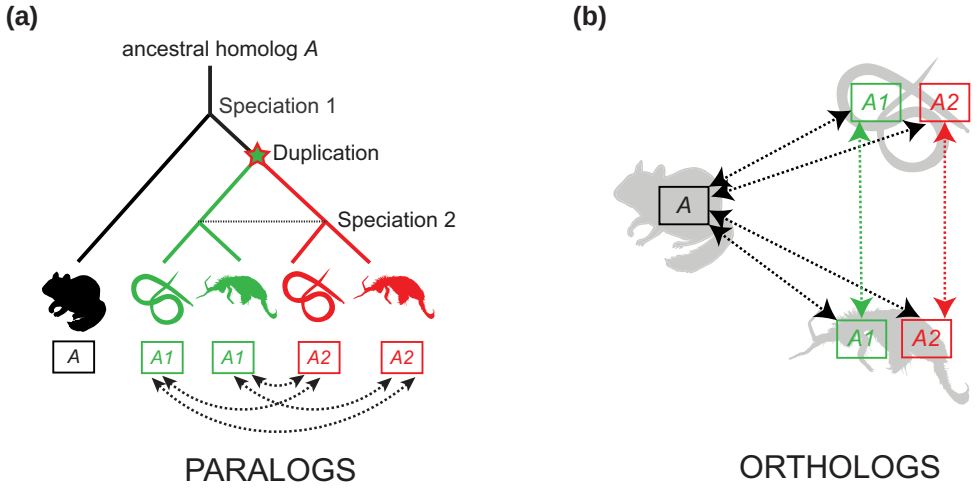

Fig. 2.—Evolutionary scenario of a gene family between three species, starting from an ancestral homolog (*A*) that is followed by two speciation events and one duplication occurring in the ancestor of two of the three species. (*a*) The single-copy gene (*A*) in the squirrel is orthologous to all other genes (*A1* and *A2*) and coalesce at the first speciation event. The green (*A1*) and red (*A2*) genes are orthologs between themselves, coalescing at the second speciation event, but paralogs between each other (*A1* vs. *A2*), coalescing at the duplication event. (*b*) Orthology graph between the three species just showing orthology relationships. The squirrel gene (*A*) forms one-to-many orthology with the other two species, being orthologous to more than one sequence on each of the species. The ancestrally duplicated genes, the green (*A1*) and red (*A2*), are just orthologs between those of the same color.

approaches followed by Markov Cluster Algorithm clustering and subsampling via automated tree pruning strategy (Tice et al. 2021). As datasets get larger, an increasingly common approach is to use a set of previously inferred orthologs as "seed" to create profile hidden Markov models (Eddy 1998) in order to find orthologous genes in the taxa that we want to add to the dataset (i.e.,: Orthograph—Petersen et al. 2017). Although this approach is fast, it only ensures finding homologs and not orthologs, particularly when datasets are incomplete (Petersen et al. 2017). As described above, orthology is nontransitive, and therefore this approach will not recover "true" orthogroups unless consecutive inference steps aiming at assessing paralogy and orthology are added to the pipeline.

Orthogroups usually contain both orthologous and paralogous genes. Finding single-copy genes in all species analysed is an exception, particularly when analyses are set up with many species (Emms and Kelly 2018). Certain tree-based methods, such as PhyloTreePruner (Kocot et al. 2013), PhyloPyPruner (Thalén 2018) or the pipeline implemented in Prasanna et al. (2020), allow to extract orthologous sequences based on the gene phylogenies of the orthogroups inferred with graph-based approaches. It is noted that there are ongoing efforts to evaluate the accuracy of several approaches to include paralogous loci and therefore circumvent the limitation of restricting analyses to strict orthologs when reconstructing the species tree (Smith and Hahn 2021; Yan et al. 2021).

Finally, sequences with unrealistic long branches, or nesting with an unrelated group, on individual gene trees may be suggestive of contamination, erroneous orthology assessment or poor sequence data. Methods to detect and filter those outlier sequences have been used to curate phylogenomic matrices of protists (Strassert et al. 2021; Irisarri et al. 2022), plants (Laurin-Lemay et al. 2012), fungi (Varga et al. 2019), metazoans (Simion et al. 2017) or mammals (Scornavacca et al. 2019), and there are softwares to detect sequences that lead to unrealistically long branch lengths (TreeShrink: Mai and Mirarab 2018). In order to allow reproducibility, the release by the authors of both the candidate set of homologs and orthologs for a particular locus is encouraged, as well as the curated final alignments (Salomaki et al. 2020).

4. Beware of the Multiple Sequence Alignment Step

Once the orthologs have been inferred, the sequences must be aligned to ensure evolutionary homology at the level of site—so each column of the multiple sequence alignment is homologous too. Errors in the alignment can introduce noise and bias in the phylogenetic reconstruction (Löytynoja and Goldman 2008). The most widely used alignment methods are progressive alignments that rely on heuristic searches, implemented in softwares such as MAFFT (Katoh and Standley 2013), MUSCLE (Edgar 2004), or T-coffee (Notredame et al. 2000). Progressive algorithms tend to produce compact alignments because they underestimate true insertion and deletion events (Löytynoja 2021). Nonetheless, newer algorithms implemented in MAFFT, such as L-INS-i and E-INS-i, better handle large gaps and large numbers of sequences, including divergent ones, so they are appropriate for genome-scale datasets. A second group of aligners are the

phylogeny-aware algorithms incorporated in programs such as PRANK (Löytynoja and Goldman 2008) or PAGAN (Löytynoja et al. 2012) that show improvements in modeling true insertion and deletion events and infer ancestral sequences. Although all alignment methods may struggle to align divergent sequences, phylogeny-aware algorithms might suffer with highly divergent ones because are sensitive to excessive levels of missing data (being confounded with gap patterns) and errors in the guiding trees, that are more likely for ancient divergences (Löytynoja 2021).

Multiple sequence alignments containing unreliable sections or erroneously aligned sites conflict with the genuine signal and may hamper phylogenetic reconstruction. Primary sequencing errors may lead to translated proteins with a disrupted reading frame or incorrect structural annotations of the coding regions. As heuristic approaches are used to assemble raw reads into longer contiguous regions or scaffolds, some assembly errors can generate chimeric sequences, which may affect the accuracy of the alignment. On the other side, biological changes in the sequences such as insertions, deletions, and translocations will also result in nonoptimal alignments. Even though alignment algorithms have considerably improved in the last decades, the alignment of columns outside highly conserved regions is still problematic. Hence, it is common to find alignment stretches of contiguous characters where homology is unreliable. Several filtering methods, referred as block-filtering tools (Di Franco et al. 2019), entirely remove sequences or columns formed by ambiguously aligned sites, such as those excessively variable (and potentially saturated) or with high levels of missing data (i.e., Gblocks: Talavera and Castresana 2007; trimAl: Capella-Gutiérrez et al. 2009; BMGE: Criscuolo and Gribaldo 2010), or alternatively, retain phylogenetically informative sites (ClipKIT: Steenwyk et al. 2020). Several studies found that character trimming increase the overall performance of the phylogenetic reconstruction by improving the phylogenetic signal-to-noise ratio and alleviating systematic artifacts caused by compositional heterogeneity (Talavera and Castresana 2007; Capella-Gutiérrez et al. 2009; Criscuolo and Gribaldo 2010). A recent comprehensive set of analyses comparing most filtering tools found that they all improve branch length estimation (Ranwez and Chantret 2020). Furthermore, filtering softwares masking nonhomologous segments, such as PREQUAL (Whelan et al. 2018) or HmmCleaner (Di Franco et al. 2019), can be more powerful than block-filtering tools because can detect errors in a set of unaligned or aligned homologous sequences and have a positive impact on the topology of gene trees (Ranwez and Chantret 2020). On the other hand, Tan et al. (2015) tested different block-filtering methods over single-gene phylogenies and concluded that those tools often lead to a decrease in accuracy, with loss of phylogenetic signal exceeding nonphylogenetic signal (noise). They recommended to avoid using current block-filtering tools that remove highly divergent sites, particularly with stringent options (removing more than 20% of the sites). More research is needed on this topic, particularly on the effect of trimming in the phylogenetic accuracy of large concatenated supermatrices, in which the presumable stochastic error of single-locus phylogenies is buffered by large amounts of loci (Philippe et al. 2017).

Some practical measures to select the most reliably aligned columns include performing alignments of the forward and reverse direction (Heads or Tails approach (Landan and Graur 2007)) or using different software methods and chose the columns of the alignment that show pairing consistency across these different treatments (e.g., the consensus alignment; Huerta-Cepas et al. 2010). Furthermore, to evaluate the robustness of the results to different filtering methods, I suggest performing a few sets of analyses involving unfiltered and filtered alignments. In order to ensure reproducibility and further investigations with published datasets, I recommend authors to release the original set of unaligned and aligned unfiltered alignments, in addition to the trimmed ones.

5. Subsample Loci Based on (Un)desirable Properties

Once orthologs have been identified and trimmed (if necessary), the next step would be to inspect the set of loci to understand their properties and assess how reliable they are for the species phylogeny. The individual history of genes contained in genomes is highly heterogeneous, and sometimes it differs significantly from the history underlying species diversification because of biological processes such as ILS or hybridization. In addition, different genes may be subject to biases, such as significant heterogeneity in base composition. This may result in varying degrees of systematic error, including instances in which the nonphylogenetic signal might overcome the genuine historical signal. As a consequence, the inference of the species tree may be dependent on which loci are analysed. As an example, Fernández et al. (2016) found that the resulting species tree phylogenies from analysing matrices with less than 250 genes differed from those using larger matrices from the same original dataset. Discarding loci that are deemed as potentially misleading is a common strategy that has been used to reduce heterogeneities in the data and improve the model fit, handle computational limitations, and testing the robustness of phylogenetic results (Mongiardino Koch 2021). A set of phylogenetically reliable candidate loci might be those with strong phylogenetic signal, low probability of being affected by systematic errors, and clock-like behavior, properties that can be roughly estimated using the branch lengths and support values of the gene trees (Doyle et al. 2015; Mongiardino Koch and Thompson 2021).

Possibly the most well-studied property that may affect phylogenomic reconstruction is missing data, which is the product of using partial or absent sequences in gene

alignments and results in site columns containing unknown or missing nucleotides/amino acids (known as gaps). There is no general agreement regarding the impact of high levels of missing data in phylogenomics. Some simulation studies concluded that their impact is small in large datasets, up to levels of missing data of 50% (Wiens and Morrill 2011), and empirical studies with similar levels have been proved to not bias the resulting topology at least in some lineages (e.g., Fernández et al. 2014, 2016). Others, instead, suggested that missing data are particularly problematic when it is not randomly distributed (Shavit Grievink et al. 2013; Xi et al. 2016) and could exacerbate systematic errors surpassing real historical signal (Roure et al. 2013). This may be because missing data reduce the number of homologous positions effectively available, re-creating an effective situation of poor taxon sampling, and hence make a dataset less informative for the calculations of the substitution models (see also points 1, 6, and 7). Since the impact of missing data is unclear, it is advisable to analyse several subsets of loci with increasing levels of missing data to evaluate the robustness of competing phylogenetic hypotheses, as well as exploring if missing data are randomly distributed in the dataset (i.e., assess if sequences from certain lineages are mostly present in loci with weak phylogenetic signal).

Several other properties have been explored in previous studies in an effort to optimize phylogenetic signal at the gene level (Shen, Salichos, et al. 2016; Smith et al. 2018; Dornburg et al. 2019; Mongiardino Koch 2021). This includes potential confounding factors such as rates of molecular evolution, compositional heterogeneity, phylogenetic informativeness or sequence saturation, among others. The rates of molecular evolution (Yang 1998; Klopfstein et al. 2017) can be measured either through tree-based methods (i.e., dividing the total gene tree length by the number of terminals) or by calculating the average percent pairwise identity between sequence pairs. In principle, slow evolving positions are less affected by saturation whereas fast evolving ones should contain more phylogenetic signal (Brinkmann and Philippe 1999). Yet there is no general consensus on whether it is better to subsample sites or genes characterised by high, intermediate, or slow evolutionary rates (Salichos and Rokas 2013; Betancur-R et al. 2014; Telford et al. 2014; Raymann et al. 2015; Klopfstein et al. 2017), although most prokaryote phylogenies or studies addressing deep divergences commonly excludes fast-evolving positions (Gerth et al. 2014; Strassert et al. 2019). Another desirable property is selecting loci with low levels of among-lineage compositional heterogeneity (Foster 2004; Nesnidal et al. 2010; Fernández et al. 2014; Gerth et al. 2014). Compositional bias may lead to artifactual relationships between taxa with similar base composition due to convergence, and is more prominent in nucleotides than in amino acids, as randomly shared composition within four possible nucleotides is more likely than in 20 amino acids (Hasegawa and Hashimoto 1993). Programs such as BaCoCa can measure the compositional heterogeneity by calculating the relative composition frequency variability in aligned sequence data (Kück and Struck 2014). Phylogenetic informativeness (Townsend 2007) is a method that uses site rate estimates to profile the suitability of particular sites or loci to correctly infer phylogenetic divergences that took place within a particular time scale, allowing to determine for which epoch is a locus most informative (Fong and Fujita 2011; Bellot et al. 2020). High levels of saturation have also been deemed as undesirable, as in saturated alignments the real genetic distances are underestimated because sites have undergone multiple substitutions, increasing the likelihood of being affected by statistical inconsistencies and LBA (Philippe et al. 2011; Nosenko et al. 2013). Therefore, minimization of saturation can be achieved by excluding saturated sites or loci, whose values are estimated as one minus the regression slope of patristic distances on *p*-distances (Mongiardino Koch 2021).

These different properties may impact phylogenetic inference in different ways, with some conditions being more likely to violate model assumptions (Philippe and Roure 2011). Therefore, it is strongly advised to build several matrices accounting for these different factors in order to assess the robustness of the results and the extent of systematic error. Loci can be filtered by tackling these factors separately, or alternatively, the correlation of those multiple properties can be assessed using principal component analysis and retain those loci that rank highest along one of the principal components axes that best explain the variance of the dataset (Struck et al. 2014; Mongiardino Koch 2021). Finally, alternative loci subsampling entailed the selection of those that were able to reconstruct well-known uncontested groups (Philippe et al. 2019; Kapli and Telford 2020) or that do support predefined hypotheses for a given question (Chen et al. 2015). Best practices for phylogenomics thus should include a series of sensitivity analyses, including building phylogenies using several matrices created from different subsampling strategies to deeply explore the robustness of our results. Analyses of large amounts of data, particularly in a Bayesian context, are hampered by problems with the mixing of chains and convergence (Laumer et al. 2019). I recommend subsampling analyses with smaller datasets that can be modeled adequately (Mulhair et al. 2021) and, based on our experience, use a minimum of 200 loci in order to have a good balance between computational burden and stochastic errors.

6. Reduce Heterogeneities in Data to Avoid Model Misspecifications

In addition to differences among genes, heterogeneity in genome-scale datasets also commonly come from the variance of rates and base composition across sites (Lartillot

and Philippe 2004) and taxa (Foster 2004). Hence, gene subsampling strategies are often not enough to overcome systematic biases caused by the improper modeling of these differences. Compositional heterogeneity is a common phenomenon because sites, genes, and organisms tend to contain an unequal proportion of nucleotides or amino acids (Foster 2004); for example, some cold-adapted fishes are known to contain a high portion of GC nucleotides compared with their tropical relatives (Zhang, Hu, et al. 2018). When the phylogenetic signal is weak, distantly related groups with similar base composition may cluster together in phylogenies due to convergence rather than descent from a common ancestor (Cummins and McInerney 2011). Therefore, compositional heterogeneity has been shown to induce systematic errors caused by model misspecifications and to reduce phylogenetic accuracy (Lartillot et al. 2007; Nesnidal et al. 2010). The influence of such biases can be mitigated by implementing strategies to reduce compositional heterogeneity in the data, using substitution models accounting for compositional heterogeneity, or a combination of both. A common approach is reducing the impact of nonphylogenetic signal by excluding the data that likely violates model assumptions (see point 5), for example by identifying and removing saturated or compositionally heterogeneous loci, sites, or taxa from the alignment (Nesnidal et al. 2010; Struck et al. 2014; Irisarri and Meyer 2016). Alternatively, making trees based on amino acid or nucleotide composition is an effective way of identifying sequences that group together due to a convergent process caused by shared base composition rather than shared history (Williams et al. 2021). Other examples of data filtering approaches used to disentangle phylogenetic signal from misleading effects include removing taxa falling on the longest branches (Rota-Stabelli et al. 2011; Struck et al. 2014) or minimize model violation by performing model adequacy tests and discarding those loci with worst absolute model fit (Prasanna et al. 2020).

An alternative solution to reduce compositional heterogeneity and substitution saturation is lumping together groups of nucleotides or amino acids that tend to have more frequent evolutionary changes within them than between them (Phillips and Penny 2003; Susko and Roger 2007). Recoding strategies sacrifice information at the expense of reducing homoplasy and/or large base compositional differences. In the case of nucleotides, RY recoding (transversion analysis) involves recoding nucleotides bases as either purines (R) or pyrimidines (Y) (Woese et al. 1991; Phillips and Penny 2003; Phillips et al. 2004; Gerth et al. 2014) and therefore only considering transversion events for phylogenetic reconstruction. Data recoding has also been implemented to reduce the amino acid data alphabets to less than 20 categories, with several schemes proposed based on similar substitution scores derived from empirical matrices, the most well-known being the six-state recoding Dayhoff 6 (Embley et al. 2003; Hrdy et al. 2004; Kosiol et al. 2004; Susko and Roger 2007). In recoded matrices, only amino acid changes between different categories, and not within categories, are considered substitutions. Four- and six-state recoding strategies have been used to test relationships across the tree of life. For example, recoding using four and six functional categories was used on a global eukaryotic phylogeny (Rodríguez-Ezpeleta et al. 2007) or to understand the evolutionary history of *Wolbachia* bacteria (Gerth et al. 2014). Dayhoff-6 recoding has been implemented in animal phylogenomic matrices (Rota-Stabelli et al. 2013; Feuda et al. 2017; Laumer et al. 2019), with Feuda et al. (2017) suggesting that this strategy may reduce potential artifacts due to differences in amino acid frequencies across species. These implementations have been contested by Hernandez and Ryan (2021), who concluded that the loss of information using six-state recodings outweighs its benefits in reducing saturation and compositional heterogeneity. Nonetheless, one recent study using simulated and empirical datasets suggests that amino acid recodings can significantly improve phylogenomic accuracy (Giacomelli et al. 2022), whereas a second one based on simulated datasets concluded that it can either increase or decrease the phylogenetic accuracy (Foster et al. 2022). Simulations derived from real datasets are also a powerful tool to detect sources of methodological errors. For example, based on how often the competing topologies concerning early animal evolution were recovered under different conditions, Kapli and Telford (2020) concluded that "ctenophora-sister" topology is driven by their unequal rates of evolution. I recommend using substitution models accounting for heterogeneity at the site level (see point 7). Furthermore, in cases when datasets are likely affected by saturation and compositional biases, analyzing recoded datasets is suggested and performing sensitivity analyses by removing heterogeneous sites or taxon to test the robustness of the inferred relationships.

7. Choose the Models That Best Fit the Data

ML and Bayesian inference are parametric approaches (i.e., model-based computations), that treat phylogenetic reconstruction as a statistical estimation. Both methods rely on explicit stochastic models of sequence evolution and on the evaluation of the likelihood function. In other words, model-based tree inference weights different substitutions differently based on a user-specified substitution model, and the rarity of substitutions is used to estimate the probability that the model generates the observed data. As such, these methods enjoy certain theoretical properties and offer approaches to evaluate the relative and absolute fit of the models to describe the evolutionary process (and thus characterize the impact of model violation). Accurate inference requires realistic substitution models accounting for

Table 2
Examples of Substitution Models Accounting for Different Aspects of the Heterogeneous Substitution Process Including Some of the Software Packages Where They Are Implemented. Models Can Be Combined to Account for Multiple Types of Heterogeneity. Computational Complexity Roughly Increases Further Down the Table

Type of Substitutional Heterogeneity	Substitution Models
No heterogeneity assumed (homogeneous models)	JC69
Replacement-rate heterogeneity	GTR, WAG, LG …
Across-sites rate heterogeneity	Gamma (G/Γ) model
Across-site compositional heterogeneity	CAT (PhyloBayes); UDM, C10–60 (IQ-TREE)
Heterotachy	GHOST (IQ-TREE)
Across-lineages compositional heterogeneity	Node discrete compositional heterogeneity model, correspondence and likelihood analysis

the heterogeneity in substitution patterns amongst lineages, genes, and sites (Philippe and Roure 2011). The process of nucleotide or amino acid evolution is complex and involves many heterogeneities, namely differences in nucleotide or amino acid compositions between species (compositional heterogeneity), differences in substitution rates between lineages or sites (rate heterogeneity), differences in substitution rates of sites through time (heterotachy; Lopez et al. 2002), and clade-specific substitution rates changing over time (heteropecily; Roure and Philippe 2011). Models that do not account for these complexities of genome evolution can lead to model violations and the recovery of incorrect trees. Although scores of substitution models have been developed (table 2), there is not yet a single model able to fit all these heterogeneities. Therefore, the choice of the model of sequence evolution has become a key decision in phylogenomic studies. To make matters more complicated, some models are available only in some specialised software.

There are two main approaches for choosing a model based on how well they describe the evolutionary process: relative vs. absolute model fit. The first approach—model comparison—consists of contrasting the relative fit of a range of models to the data (Posada and Buckley 2004). Models are ranked based on measures such as likelihood scores, Bayes factors, or information criteria, such as Bayesian information criterion or Akaike information criterion. These model-selection methods are implemented in widely used software packages such as ModelFinder or ModelTest-NG (i.e., Kalyaanamoorthy et al. 2017; Darriba et al. 2020). In the context of Bayesian inference, cross-validation has also been sometimes used as an alternative method (Lartillot et al. 2009). The model selected as the best from a set of candidates may not fit the data well simply because it has not been designed to model a specific aspect of the substitution process. The second approach to choose a model involves testing model adequacy as a measure of goodness-of-fit, giving an idea of how adequately the model describes important features of the data (Goldman 1993). Bayesian posterior predictive simulations provide a useful way of testing model adequacy (Bollback 2002; Brown 2014; Doyle et al. 2015), but they are not as widely implemented or used as model fit tests.

Some areas in a phylogeny are particularly recalcitrant to resolution. Major inconsistencies across phylogenomic studies usually occur at short internal branches or ancient diversification events (Delsuc et al. 2005). These branches are normally the product of a rapid diversification and bear a limited amount of phylogenetic signal that may be progressively lost through multiple substitutions (Rokas and Carroll 2006). They are challenging to resolve for any model, particularly those simpler ones that assume all sequences evolve under the same substitution process. These challenging nodes present a low signal-to-noise ratio and are susceptible to random or systematic errors. Although these short internodes could represent true divergence events between three or more lineages (hard polytomies—with methods available to test for polytomies [Chang et al. 2015]), better modeling commonly reveals they are the result of insufficient or biased information (soft polytomy). Several strategies using complex models (see Glossary) may help to better accommodate heterogeneous data and increase the phylogenetic signal by minimizing model violations in problematic nodes. The first strategy to improve model fit is partitioning the data (Yang 1996). In concatenated matrices, genes (or partitions) may evolve under different rates, so a different substitution model could be applied to each of them to improve accuracy (Lanfear et al. 2014). A recent study, though, suggested that partitioning data do not properly account for the substitution heterogeneity, particularly when site homogeneous models are used (Wang et al. 2019). The second strategy to handle differences in the evolution process between sites, genes, and lineages relies on the use of profile mixture models of evolution. These models account for heterogeneity at the site level (i.e., site-heterogeneous mixture models, assuming evolutionary processes vary widely, in particular the set of acceptable amino acids) and have been shown to fit empirical data significantly better than site homogeneous models in most cases (Philippe et al. 2011; Wang et al. 2019). Mixture models account for the differences in amino acid frequencies in the multiple sequence alignments, either as empirical compositional mixtures that have been previously estimated from large protein alignment collections (e.g., C10 to C60: Quang et al. 2008; universal distribution mixture [UDM] model: Schrempf et al. 2020), or as an infinite mixture

estimated directly from the data (CAT model: Lartillot and Philippe 2004). The main disadvantage of these types of models is that they require high computational times, which limits their use in the largest matrices. Recent methodological developments, such as the posterior mean site frequency (Wang et al. 2018) method, are promising in terms of improving the computational efficiency in the ML framework. Overall, there is usually a trade-off between the computational cost of an analysis and the size of the dataset that is being analyzed if we chose profile mixture models. A good compromise involves the analyses of large matrices under simpler substitution models, and the exploration of smaller matrices following subsampling strategies (see above) under mixture models. Crucially, if the two types of analyses do not agree, then a more extensive analysis of the fit of alternative models should be conducted, so as to establish which modeling strategy is more adequate for the case at hand.

8. To Concatenate or Not to Concatenate? This is the Question…

For large datasets, there are two main strategies to infer the species tree based on a set of orthologous multiple sequence alignments. In the first of them, each of the alignments are concatenated into a single alignment in order to sum up their genome-wide phylogenetic signal to get a global estimate of the species tree. Concatenation is inconsistent, understood as not converging towards the correct tree with more data, when there is a substantial amount of discordance between gene trees (Kubatko and Degnan 2007) because there are violations of the model of gene evolution (Di Franco et al. 2021). Among the most popular programs for analysing concatenated supermatrices in ML framework are: IQ-TREE (Nguyen et al. 2015), RaxML (Kozlov et al. 2019), PhyML (Guindon et al. 2010), FastTree (Price et al. 2010), and PAML (Yang 1997); and for Bayesian inference: PhyloBayes (Lartillot et al. 2009; Lartillot et al. 2013), RevBayes (Höhna et al. 2016), MrBayes (Huelsenbeck and Ronquist 2001), and BEAST (Suchard et al. 2018).

Alternatively, in a coalescence framework, such as the multispecies coalescent (MSC) model (Pamilo and Nei 1988), the topologies and branch length of the locus vary among loci due to the coalescent process in the ancestral populations (Rannala and Yang 2003). The latter method is robust to ILS, which can cause different parts of the genome to have different evolutionary histories and therefore results in heterogeneous gene trees with incongruent topologies compared with the species tree. Summary coalescent or two-step methods such as ASTRAL (Mirarab, Reaz, et al. 2014; Zhang, Rabiee, et al. 2018; Zhang et al. 2020), MP-EST (Liu et al. 2010), and SVDQuartets (Chifman and Kubatko 2014) rely on previously inferred gene trees to estimate the species tree, which are used as real observations without accounting for stochastic errors. Inaccurate or uninformative single gene reconstructions driven by the small size of loci might result in stochastic and/or systematic errors due to improper modeling (Richards et al. 2018). These biases might be ameliorated when using full likelihood implementations of the MSC (Shi and Yang 2018). Full likelihood methods or single-step ones such as BPP (Yang 2015; Flouri et al. 2018) or *BEAST (Heled and Drummond 2010), jointly infer gene and species trees, accounting for stochastic errors and deep coalescent events. However, current software implementations do not include mixture models (Flouri et al. 2018) that accommodate heterogeneity across sites or lineages. Even though simulated and empirical data suggests that full likelihood methods might be superior to summary coalescent ones (Shi and Yang 2018), the latter is the most widely used when analysing large datasets given their low computational demands. Coalescent approaches present other compound sets of problems, such assuming that there is no intralocus recombination. Recombination has been shown as common in protein-coding genes, with exons within genes having different stories (Scornavacca and Galtier 2017), despite this fact, MSC-based methods might not be severely impacted by the violation of this premise. A philosophically similar approach to full likelihood methods, but less computationally intensive, is the gene tree–species tree reconciliation method, that coestimate gene and species trees taking into account gene duplication, losses and horizontal gene transfers (Boussau et al. 2013; Szöllősi et al. 2013).

Therefore, concatenation and coalescence methods are both compromised by assumptions that are not completely satisfied (Gatesy and Springer 2014), and it is still controversial which strategy captures the more accurate tree (Edwards et al. 2016; Springer and Gatesy 2016). Incongruent topologies between concatenations and coalescent-based approaches are common among animals (Prum et al. 2015), fungi (Shen, Zhou, et al. 2016), and plants (Wickett et al. 2014). The presence of incongruence between both approaches represents a major challenge to infer robust species phylogenies using genome-scale data (Kubatko and Degnan 2007; Bravo et al. 2019; Shen et al. 2021). Hybrid approaches, such as "binning methods" have been developed to concatenate groups of loci to be used as input for coalescent methods (Mirarab, Bayzid, et al. 2014), but these strategies violate the implicit assumption of ILS-aware methods in which the history of a coalescent units is represented by a single tree (Gatesy and Springer 2014). Each biological question needs to account for the potential violations or misspecification of the assumptions of the different methods, and selecting the most appropriate method will heavily depend on the phylogenetic problem at hand (Bryant and Hahn 2020). For reconstructing deep branches, tree accuracy seems more impacted by high levels of homoplasy, heterogeneous rates, and lack of phylogenetic signal

at rapid radiations than levels of ILS, which has been shown to be a minor determinant of the phylogenetic conflict in a mammalian dataset (Scornavacca and Galtier 2017). Concatenation can therefore be the preferred strategy when noise is the dominant source of conflict in ancient phylogenetic relationships presenting distant speciation events (Bryant and Hahn 2020) or when gene tree estimation errors are high and ILS levels are low (Shen et al. 2021). ILS-aware methods, instead, would be more appropriate for more recent speciation events, when there have been successive splits within a small time interval, or when the source of genealogic tree discordance is presumed to be explained by ILS (Simion et al. 2020). When some of these assumptions to choose one method versus the other are unknown, which happens to be often the case, I recommend employing both methods and discussing the potential incongruences between the results of both phylogenetic approaches under the light of the different methodological assumptions.

9. High Support Values in Supermatrices Does Not Imply Accurate Trees

Measures of branch support are important for understanding the robustness of phylogenetic inference (Minh et al. 2020). The analyses of large phylogenetic datasets usually result in fully resolved trees with highly supported branches. However, these strongly supported topologies are not unequivocally correct. One approach to infer the confidence or support for each branch of the tree is done via performing bootstrap analyses (Felsenstein 1985). These analyses resample columns of the alignments, building trees with each new resampled set, and calculating how many times each clade is retrieved in the bootstrap replications. As calculating nonparametric bootstrap is time consuming, particularly for genome-scale datasets, faster and relatively unbiased support values are commonly used in phylogenomics, such as the ultrafast bootstrap approximation (Minh et al. 2013) or the Shimodaira-Hasegawa (SH)-like approximate likelihood ratio test (Guindon et al. 2010). In the Bayesian framework, branch support is measured by posterior probabilities (Huelsenbeck et al. 2001). In concatenated matrices, bootstrap values above 95% and posterior probability values above 0.9 are normally considered as strong support. These standard measures of statistical support, though, measure uncertainty in estimates given the data and a specific evolutionary model. Hence, they do not account for biases in the data or incorrect model assumptions (systematic errors). So if the model does not describe the properties of the data, an incorrect topology can receive high statistical support (Delsuc et al. 2005). It has been shown that in phylogenomic supermatrices one loci or a few sites can have a disproportionate amount of influence in recovering incorrect nodes with full support (Shen et al. 2017). I recommend evaluating these support measures in different submatrices under alternative inference approaches and models of evolution, from lesser to better fitting ones, which could inform on potential systematic errors (see point 5, 6 and 7 above on subsampling strategies and modeling). Finally, the trustworthiness of the phylogenomic results could be tested by checking if there is agreement across alternative sources of data, such as morphology, within the philosophical framework of consilience (McInerney et al. 2014).

10. Dissect Incongruence in the Phylogenetic Signal

The phylogenetic signal in phylogenomic data is heterogeneous, with sites and loci presenting evolutionary histories that often depart from the branching pattern of the species phylogeny (Bravo et al. 2019). Classic measures of branch support do not fully capture these heterogeneous histories. Alternative measures of support quantify the disagreement among loci and sites and can reveal sources of topological incongruence in phylogenomic matrices, being more informative in the context of genome-scale data. The calculation of site and gene concordance factors, for example, allows assessing the fraction of sites or loci within an alignment supporting a particular branch (Ané et al. 2007; Minh et al. 2020).

Another recently developed approach to understand how the phylogenetic signal is distributed is through constrained tree analyses (Simion et al. 2020). In these analyses, two or more competing topologies are constrained, and the difference in likelihood is measured (per gene or per site) among the set of alternative hypotheses, relying on a "majority vote" to determine which topology is best supported by the data. These constrained tree analyses, leveraging the information from single sites or loci, have been recently used to evaluate the likelihood of alternative topologies in recalcitrant nodes of multiple lineages and discriminate which ones better represent the species tree (Smith et al. 2015; Arcila et al. 2017; Shen et al. 2017). A similar measure of incongruence can be obtained in a coalescent framework via quantification of the quartet-based scores for alternative topologies (Gatesy et al. 2016; Shen et al. 2021).

Constrained tree analyses have revealed that the distribution of phylogenetic signal across individual locus and/or sites in large data matrices is unequal, and hence just a few sites may drive the resulting topology (Shen et al. 2017; Walker et al. 2018; Francis and Canfield 2020). These results suggest that dissecting the distribution of support for alternative topologies enables researchers to better explain the incongruence in phylogenomic analyses, quantify the distribution and strength of signal on contentious branches, and understand whether these topologies are robustly supported or not (Shen et al. 2017). An excessive gene or site support in a contentious node may be attributed to positive selection or other evolutionary processes

such as ILS, horizontal gene transfer, or hybridization. Alternatively, analytical errors such as insufficient taxon sampling or model misspecification (i.e., using an inadequate substitution model) can also give rise to loci with histories incongruent with the species tree (Shen et al. 2017). On the other hand, it has been argued that these analyses are also be affected by problems of model fit, because non-phylogenetic signal (such as high levels of homoplasy) can bias the recovered topology and models cannot accurately infer parameters using the limited amount of information contained in a single locus/site (Simion et al. 2020). Besides helping to clarify the nature of phylogenetic incongruence, analytical approaches that quantify signal at the level of sites and loci are useful for understanding the influence of substitution models by contrasting the distribution and strength of phylogenetic signal favoring a particular topology when different parameters are used. This type of analysis in phylogenomic studies allows to gauge the robustness of the phylogenetic signal and its dependence on the datasets and models used.

Closing Remarks

The aim of this article is to provide a simple guide for dissecting a phylogenomic study in order to assess its robustness, with the goal of bridging the gap between technical and biological knowledge and helping researchers without a formal background in bioinformatics to comprehensively evaluate the quality of genome-scale phylogenetic studies. I believe this is a key not only for efficient and fair manuscript review and grant evaluations but also for a correct biological interpretation of the results presented in the scientific literature. In the present era of massive amounts of data, an unknown portion of published phylogenomic results may be the product of inadvertent errors, nonoptimal steps in the pipeline, or methodological biases. The manuscript highlights that these discrepancies can be the result of contaminants, improper assembly or orthology inference, erroneous alignments, or model assumptions being breached (systematic error). I encourage readers to open the "phylogenomic black box" and understand the quality of genomic data, look at the intermediate steps of the analyses, and assess whether a satisfactory model of evolution has been used to correctly infer the phylogeny in several submatrices of carefully chosen loci or positions. To sum up, more is not always better, and the devil is in the detail.

Acknowledgment

I thank Omar Rota-Stabelli and three other anonymous reviewers, who revised and provided excellent suggestions over a previous version of the manuscript. Furthermore, I would like to thank Rosa Fernández, Erik Tihelka, Ya Yang, László Nagy, and Nicolás Mongiardino Koch, who provided fruitful discussion and critical evaluation on a final version of the manuscript and enriched it with useful comments on phylogenomics methods applied to a wide range of clades. J.L.-F. was supported by a Juan de la Cierva Incorporación fellowship (Ministerio de Ciencia e Innovación, IJC2018-035237-I).

Literature Cited

Aguinaldo AM et al. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 387(6632): 489–493.

Ahrens JB, Teufel AI, Siltberg-Liberles J. 2020. A phylogenetic rate parameter indicates different sequence divergence patterns in orthologs and paralogs. J Mol Evol. 88:720–730.

Altenhoff AM, et al. 2016. Standardized benchmarking in the quest for orthologs. Nat Methods. 13:425–430.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol Biol Evol. 24(2):412–426.

Arcila D, et al. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat Ecol Evol. 1(2):1–10.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5): 455–477.

Baurain D, Brinkmann H, Philippe H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol Biol Evol. 24(1):6–9.

Bellot S, Mitchell TC, Schaefer H. 2020. Phylogenetic informativeness analyses to clarify past diversification processes in Cucurbitaceae. Sci Rep. 10(1):1–3.

Bergsten J. 2005. A review of long-branch attraction. Cladistics. 21(2): 163–193.

Betancur-R R, Naylor GJ, Ortí G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst Biol. 63(2):257–262.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol. 19:1171–1180.

Bossert S, Murray EA, Blaimer BB, Danforth BN. 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol Phylogenet Evol. 111:149–157.

Boussau B et al. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23(2):323–330.

Bravo GA, et al. 2019. Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. PeerJ. 14(7):e6399.

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol. 16(6):817–825.

Brown JM. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst Biol. 63(3):334–338.

Brown JM, Thomson RC. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst Biol. 66(4):517–530.

Bryant D, Hahn MW. 2020. The concatenation question. In: Scornavacca C, et al., editors. Phylogenetics in the genomic era. No commercial publisher—Authors open access book. 3.4:1–3.4:23. Available from: https://hal.inria.fr/PGE/.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Chang Y, et al. 2015. Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants. Genome Biol Evol. 7(6):1590–1601.

Chen MY, Liang D, Zhang P. 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst Biol. 64(6):1104–1120.
Chifman J, Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30(23):3317–3324.
Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. PNAS. 105(51):20356–20361.
Criscuolo A, Gribaldo S. 2010. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 10(1):210.
Cummins CA, McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst Biol. 60(6):833–844.
Dabert M, Witalinski W, Kazmierski A, Olszanowski Z, Dabert J. 2010. Molecular phy-logeny of acariform mites (Acari, Arachnida): strong conflict between phylogenetic signal and long-branch attraction artifacts. Mol Phylogenet Evol. 56(1):222–241.
Darriba D, et al. 2020. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol. 37(1):291–294.
Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2(5):e68.
Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6(5):361–375.
De Maio N, et al. 2019. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 5(9):1–12.
Di Franco A, Baurain D, Glöckner G, Melkonian M, Philippe H. 2021. Lower statistical support with larger datasets: insights from the Ochrophyta radiation. bioRxiv. Available from: http://doi.org/10.1101/2021.01.14.426536.
Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol Biol. 19(1):21.
Doolittle WF, Brown JR. 1994. Tempo, mode, the progenote, and the universal root. PNAS. 91(15):6721–6728.
Dornburg A, Su Z, Townsend JP. 2019. Optimal rates for phylogenetic inference and experimental design in the era of genome-scale data sets. Syst Biol. 68(1):145–156.
Doyle VP, Young RE, Naylor GJ, Brown JM. 2015. Can we identify genes with increased phylogenetic reliability? Syst Biol. 264(5):824–837.
Dunn CW, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452(7188):745–749.
Eddy SR. 1998. Profile hidden Markov models. Bioinformatics (Oxford, England) 14(9):755–763.
Edgar R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.
Edwards SV, et al. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol Phylogenet Evol. 94:447–462.
Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8: 163–167.
Embley TM, et al. 2003. Hydrogenosomes, mitochondria and early eukaryotic evolution. IUBMB life. 55(7):387–395.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):1–4.
Emms DM, Kelly S. 2018. STAG: species tree inference from all genes. BioRxiv. 267914. Available from: http://doi.org/10.1101/267914.
Emms DM, Kelly S. 2020. Benchmarking orthogroup inference accuracy: revisiting orthobench. Genome Biol Evol. 12(12):2258–2266.
Faircloth BC, et al. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol. 61(5):717–726.
Federhen S. 2012. The NCBI taxonomy database. Nucleic Acids Res. 40(D1):D136–D143.
Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27(4):401–410.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.
Fernández R, Edgecombe GD, Giribet G. 2016. Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. Syst Biol. 65(5):871–889.
Fernández R, Gabaldon T, Dessimoz C. 2020. Orthology: definitions, prediction, and impact on species phylogeny inference. In: Scornavacca C, et al., editors. Phylogenetics in the genomic era. No commercial publisher—Authors open access book. 2.4:1–2.4:14. Available from: https://hal.inria.fr/PGE/.
Fernández R, Hormiga G, Giribet G. 2014. Phylogenomic analysis of spiders reveals nonmonophyly of orb weavers. Curr Biol. 24(15): 1772–1777.
Feuda R, et al. 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr Biol. 27-(24):3864–3870.
Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19(2):99–113.
Fitch WM. 2000. Homology: a personal view on some of the problems. Trends Genet. 16(5):227–231.
Flouri T, Jiao X, Rannala B, Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol Biol Evol. 35(10):2585–2593.
Fong JJ, Fujita MK. 2011. Evaluating phylogenetic informativeness and data-type usage for new protein-coding genes across Vertebrata. Mol Phylogenet Evol. 61(2):300–307.
Foster PG. 2004. Modeling compositional heterogeneity. Syst Biol. 53(3):485–495.
Foster PG, Cox CJ, Embley TM. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc B. 364(1527):2197–2207.
Foster PG, et al. 2022. Recoding amino acids to a reduced alphabet may increase or decrease phylogenetic accuracy. Syst Biol. syac042. doi:10.1093/sysbio/syac042.
Francis WR, et al. 2013. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics. 14(1):1–2.
Francis WR, Canfield DE. 2020. Very few sites can reshape the inferred phylogenetic tree. PeerJ. 8:e8865.
Gatesy J, et al. 2016. Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. Cladistics. 33(3):295–332.
Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol Phylogenet Evol. 80: 231–266.
Gee H. 2003. Ending incongruence. Nature. 425(6960):782–782.
Gerth M, Gansauge MT, Weigert A, Bleidorn C. 2014. Phylogenomic analyses uncover origin and spread of the Wolbachia pandemic. Nat Commun. 5(1):1–7.
Giacomelli M, Rossi ME, Lozano-Fernandez J, Feuda R, Pisani D. 2022. Resolving tricky nodes in the tree of life through amino acid recoding. bioRxiv. Available from: http://doi.org/10.1101/2022.02.24.479670.

Goldman N. 1993. Statistical tests of models of DNA substitution. J Mol Evol. 36(2):182–198.
Gouy R, Baurain D, Philippe H. 2015. Rooting the tree of life: the phylogenetic jury is still out. Philos Trans R Soc B. 370(1678):20140329.
Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652.
Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol. 47(1):9–17.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.
Hasegawa M, Hashimoto T. 1993. Ribosomal RNA trees misleading? Nature. 361(6407):23.
Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol. 46(3):239–257.
Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27(3):570–580.
Hendy MD, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool. 38:297–309.
Hernandez AM, Ryan JF. 2021. Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses. Syst Biol. 70(6):1200–1212.
Hillis D, Moritz C, Mable BK. 1996. Molecular systematics. 2nd ed. Sunderland (MA): Sinauer Associates.
Höhna S, et al. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol. 65(4):726–736.
Holland BR, Penny D, Hendy MD. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—a simulation study. Syst Biol. 52:229–238.
Holton TA, Pisani D. 2010. Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm. Genome Biol Evol. 2:310–324.
Hrdy I, et al. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432: 618–622.
Huelsenbeck JP, Joyce P, Lakner C, Ronquist F. 2008. Bayesian analysis of amino acid substitution models. Philos Trans R Soc B. 363(1512): 3941–3953.
Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17(8):754–755.
Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294(5550):2310–2314.
Huerta-Cepas J, et al. 2010. PhylomeDB v3. 0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. 39:D556–D560.
Irisarri I, et al. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol. 1(9):1370–1378.
Irisarri I, Meyer A. 2016. The identification of the closest living relative(s) of tetrapods: phylogenomic lessons for resolving short ancient internodes. Syst Biol. 65(6):1057–1075.
Irisarri I, Strassert JF, Burki F. 2022. Phylogenomic insights into the origin of primary plastids. Syst Biol. 71(1):105–120.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22(4):225–231.
Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14(6):587–589.
Kapli P, Telford MJ. 2020. Topology-dependent asymmetry in systematic errors affects phylogenetic placement of Ctenophora and Xenacoelomorpha. Sci Adv. 6(50):eabc5162.
Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. Nat Rev Genet. 18:1–7.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.
Klopfstein S, et al. 2017. More on the best evolutionary rate for phylogenetic analysis. Syst Biol. 66(5):769–785.
Kocot KM, Citarella MR, Moroz LL, Halanych KM. 2013. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol Bioinform. 9:429–435.
Kosiol C, Goldman N, Buttimore NH. 2004. A new criterion and method for amino acid classification. J Theor Biol. 228:97–106.
Koutsovoulos G, et al. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. PNAS. 113(18):5053–5058.
Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35(21):4453–4455.
Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 56(1):17–24.
Kück P, Struck TH. 2014. BaCoCa–A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Mol Phylogenet Evol. 70:94–98.
Kuzniar A, van Ham RC, Pongor S, Leunissen JA. 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet. 24(11):539–551.
Laetsch DR, Blaxter ML. 2017. BlobTools: interrogation of genome assemblies. F1000Research 6:1287.
Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24(6):1380–1383.
Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol. 14(1):1–4.
Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7(1):1–4.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25(17):2286–2288.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21(6):1095–1109.
Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 62(4):611–615.
Laumer CE, et al. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc Royal Soc B 286(1906):20190831.
Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Curr Biol. 22(15):R593–R594.
Liu L, Wu S, Yu L. 2015. Coalescent methods for estimating species trees from phylogenomic data. J Syst Evol. 53:380–390.
Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol. 10(1):302.
Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19(1):1–7.
Löytynoja A. 2021. Phylogeny-aware alignment with PRANK and PAGAN. In: Katoh K, editor. Multiple sequence alignment. Methods in molecular biology. Vol. 2231. New York (NY): Humana. p. 17–33.
Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320(5883):1632–1635.

Löytynoja A, Vilella AJ, Goldman N. 2012. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. Bioinformatics. 28(13):1684–1691.
Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46(3): 523–536.
Mai U, Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics. 19(5):23–40.
Martin WF, Weiss MC, Neukirchen S, Nelson-Sathi S, Sousa FL. 2016. Physiology, phylogeny, and LUCA. Microb Cell. 3(12):582.
McCormack JE, Tsai WL, Faircloth BC. 2016. Sequence capture of ultraconserved elements from bird museum specimens. Mol Ecol Resour. 16(5):1189–1203.
McInerney JO, O'connell MJ, Pisani D. 2014. The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nat Rev Microbiol. 12(6):449–455.
McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. 2018. Practical considerations for plant phylogenomics. Appl Plant Sci. 6(3):e1038.
Metzker ML. 2010. Sequencing technologies – the next generation. Nat Rev Genet. 11:31–46.
Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. Mol Biol Evol. 37(9): 2727–2733.
Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 30(5): 1188–1195.
Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346(6215):1250463.
Mirarab S, Reaz R, et al. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30-(17):i541–i548.
Mongiardino Koch N. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. Mol Biol Evol. 38(9):4025–4038.
Mongiardino Koch N, Thompson JR. 2021. A total-evidence dated phylogeny of Echinoidea combining phylogenomic and paleontological data. Syst Biol. 70(3):421–439.
Mulhair P, McCarthy CG, Siu Ting K, Creevey C, O'Connell MJ. 2021. Enriching for orthologs increases support for Xenacoelomorpha and Ambulacraria sister relationship. BioRxiv. Available from: http://doi.org/10.1101/2021.12.13.472462.
Nabhan AR, Sarkar IN. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Brief Bioinform. 13(1):122–134.
Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B. 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol Biol Evol. 27(9):2095–2104.
Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.
Nosenko T, et al. 2013. Deep metazoan phylogeny: when different genes tell different stories. Mol Phylogenet Evol. 67(1): 223–233.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302(1):205–217.
O'Brien SJ, Stanyon R. 1999. Phylogenomics: ancestral primate viewed. Nature 402:365–366.
Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol. 5(5):568–583.
Parenteau J, Abou Elela S. 2019. Introns: good day junk is bad day treasure. Trends Genet. 35(12):923–934.
Petersen M, et al. 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics. 18:111.
Philippe H, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9(3):e1000602.
Philippe H, et al. 2017. Pitfalls in supermatrix phylogenomics. Eur J Taxon. 283:1–25.
Philippe H, et al. 2019. Mitigating anticipated effects of systematic errors supports sister-group relationship between Xenacoelomorpha and Ambulacraria. Curr Biol. 29(11):1818–1826.
Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev. 8(6):616–623.
Philippe H, Roure B. 2011. Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biol. 9:91.
Philippe H, Chenuil A, Adoutte A. 1994. Can the Cambrian explosion be inferred through molecular phylogeny? Development 1994: 15–25.
Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 21(7):1455–1458.
Phillips MJ, Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol. 28: 171–185.
Pisani D, et al. 2015. Genomic data do not support comb jellies as the sister group to all other animals. PNAS 112:15402–15407.
Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol. 53(5): 793–808.
Prasanna AN, et al. 2020. Model choice, missing data, and taxon sampling impact phylogenomic inference of deep Basidiomycota relationships. Syst Biol. 69(1):17–37.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One. 5(3): e9490.
Prum RO, et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature. 526(7574): 569–573.
Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24: 2317–2323.
Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164(4):1645–1656.
Ranwez V, Chantret N. 2020. Strengths and limits of multiple sequence alignment and filtering methods. In: Scornavacca C, et al., editors. Phylogenetics in the genomic era. 2.2:1–2.2:36. Available from: https://hal.inria.fr/PGE/.
Rasmussen MD, Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22: 755–765.
Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the Archaea. PNAS 112-(21):6670–6675.
Reddy S, et al. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst Biol. 66(5):857–879.
Richards EJ, Brown JM, Barley AJ, Chong RA, Thomson RC. 2018. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological? Syst Biol. 67(5):847–860.
Rodríguez-Ezpeleta N, et al. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 56(3):389–399.
Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biol. 4(11): e352.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425(6960):798–804.
Rota-Stabelli O, et al. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc Royal Soc B. 278(1703):298–306.
Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2013. Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst Biol. 62(1):121–133.
Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol. 30(1):197–214.
Roure B, Philippe H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol Biol. 11(1):1–4.
Ryan JF, et al. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342:1242592.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497(7449):327–331.
Salomaki ED, Eme L, Brown MW, Kolisko M. 2020. Releasing uncurated datasets is essential for reproducible phylogenomics. Nat Ecol Evol. 4:1435–1437.
Schrempf D, Lartillot N, Szöllősi G. 2020. Scalable empirical mixture models that account for across-site compositional heterogeneity. Mol Biol Evol. 37(12):3616–3631.
Scornavacca C, et al. 2019. OrthoMaM v10: scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. Mol Biol Evol. 36(4):861–862.
Scornavacca C, Galtier N. 2017. Incomplete lineage sorting in mammalian phylogenomics. Syst Biol. 66(1):112–120.
Scotland RW, Olmstead RG, Bennett JR. 2003. Phylogeny reconstruction: the role of morphology. Syst Biol. 52(4):539–548.
Shavit L, Penny D, Hendy MD, Holland BR. 2007. The problem of rooting rapid radiations. Mol Biol Evol. 224(11):2400–2411.
Shavit Grievink L, Penny D, Holland BR. 2013. Missing data and influential sites: choice of sites for phylogenetic analysis can be as important as taxon sampling and model choice. Genome Biol Evol. 5:681–687.
Shen XX, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol. 1:0126.
Shen XX, Salichos L, Rokas A. 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol Evol. 8(8):2565–2580.
Shen XX, Steenwyk JL, Rokas A. 2021. Dissecting incongruence between concatenation-and quartet-based approaches in phylogenomic data. Syst Biol. 70(5):997–1014.
Shen XX, Zhou X, et al. 2016. Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. G3-Genes Genomes Genet. 6(12):3927–3939.
Shi CM, Yang Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Mol Biol Evol. 35(1):159–179.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Simion P, et al. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr Biol. 27(7):958–967.
Simion P, et al. 2018. A software tool 'CroCo'detects pervasive cross-species contamination in next generation sequencing data. BMC Biol. 16:28.
Simion P, Delsuc F, Philippe H. 2020. To what extent current limits of phylogenomics can be overcome? In: Scornavacca C, et al., editors. Phylogenetics in the genomic era. No commercial publisher—Authors open access book. 2.1:1–2.1:34. Available from: https://hal.inria.fr/PGE/.
Siu-Ting K, et al. 2019. Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. Mol Biol Evol. 36(6):1344–1356.
Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. PloS One 13(5):e0197433.
Smith ML, Hahn MW. 2021. New approaches for inferring phylogenies in the presence of paralogs. Trends Genet. 37(2):174–187.
Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol Biol. 15(1):1–5.
Spang A, et al. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521(7551):173–179.
Springer MS, Gatesy J. 2016. The gene tree delusion. Mol Phylogenet Evol. 94:1–33.
Steenwyk JL, Buida TJ III, Li Y, Shen XX, Rokas A. 2020. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 18(12):e3001007.
Strassert JF, Irisarri I, Williams TA, Burki F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat Commun. 12(1):1–3.
Strassert JF, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F. 2019. New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. Mol Biol Evol. 36(4):757–765.
Straub SC, et al. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot. 99(2):349–364.
Struck TH. 2013. The impact of paralogy on phylogenomic studies–a case study on annelid relationships. PloS One 8(5):e62892.
Struck TH, et al. 2014. Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. Mol Biol Evol. 31(7):1833–1849.
Suchard MA, et al. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4(1):vey016.
Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol. 24:2139–2150.
Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. Syst Biol. 62(6):901–912.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56(4):564–577.
Tan G, et al. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst Biol. 64(5):778–791.
Telford MJ, et al. 2014. Phylogenomic analysis of echinoderm class relationships supports Asterozoa. Proc Royal Soc B. 281(1786):20140479.
Thalén F. 2018. PhyloPyPruner: tree-based orthology inference for phylogenomics with new methods for identifying and excluding contamination. University of Tuebingen. Available from http://lup.lub.lu.se/student-papers/record/8963554.
Tice AK, et al. 2021. PhyloFisher: a phylogenomic package for resolving eukaryotic relationships. PLoS Biol. 19(8):e3001365.
Timmermans MJ, et al. 2016. Family-level sampling of mitochondrial genomes in Coleoptera: compositional heterogeneity and phylogenetics. Genome Biol Evol. 8(1):161–175.
Todd EV, Black MA, Gemmell NJ. 2016. The power and promise of RNA-seq in ecology and evolution. Mol Ecol. 25(6):1224–1241.

Townsend JP. 2007. Profiling phylogenetic informativeness. Syst Biol. 56(2):222–231.
Varga T, et al. 2019. Megaphylogeny resolves global patterns of mushroom evolution. Nat Ecol Evol. 3(4):668–678.
Walker JF, Brown JW, Smith SA. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. Syst Biol. 67(5):916–924.
Wang HC, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst Biol. 67(2):216–235.
Wang HC, Susko E, Roger AJ. 2019. The relative importance of modeling site pattern heterogeneity versus partition-wise heterotachy in phylogenomic inference. Syst Biol. 68(6):1003–1019.
Weitemier K, et al. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl Plant Sci. 2(9):1400042.
Whelan S, Irisarri I, Burki F. 2018. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. Bioinformatics 34(22):3929–3930.
Whelan S, Liò P, Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 17(5):262–272.
Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends Ecol Evol. 22(5):258–265.
Wickett NJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. PNAS 111(45): E4859–E4868.
Wiens JJ, Morrill MC. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst Biol. 60(5):719–731.
Williams TA, et al. 2021. Inferring the deep past from molecular data. Genome Biol Evol. 13(5):evab067.
Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. Nat Ecol Evol. 4(1):138–147.
Woese CR, Achenbach L, Rouviere P, Mandelco L. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst Appl Microbiol. 14:364–371.
Xi Z, Liu L, Davis CC. 2016. The impact of missing data on species tree estimation. Mol Biol Evol. 33(3):838–860.
Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. Genome Biol. 6(5):1–30.
Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L. 2021. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. Syst Biol. 71(2):367–381.
Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol. 42(5):587–596.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555–556.
Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst Biol. 47(1):125–133.
Yang Z. 2015. The BPP program for species tree estimation and species delimitation. Curr Zool. 61(5):854–865.
Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol Biol Evol. 31(11):3081–3092.
Young AD, Gillung JP. 2020. Phylogenomics—principles, opportunities and pitfalls of big-data phylogenetics. Syst Entomol. 45:225–247.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829.
Zhang D, Hu P, et al. 2018. GC bias lead to increased small amino acids and random coils of proteins in cold-water fishes. BMC Genomics 19(1):315.
Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:153.
Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. Mol Biol Evol. 37(11):3292–3307.
Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51(4):588–598.

Associate editor: Davide Pisani