Sources of Error and Incongruence in Phylogenomic Analyses

C. Bleidorn, *Phylogenomics*, DOI 10.1007/978-3-319-54064-1_9

- Phylogenomic analyses can be performed by analysing gene trees separately and using coalescent or supertree analyses or a concatenation of all genes (supermatrix approach).
- Several sources of systematic error may bias phylogenomic studies due to the violation of substitution model assumptions, including problems with compositional heterogeneity, among-lineage rate variation and heterotachy.
- Missing data is usually less problematic for phylogenomic studies, but taxon sampling can be critical.
- Data and taxa should be carefully selected for analysis; highly saturated genes as well as phylogenetically unstable (rogue) taxa should be avoided.
- Discordance of gene trees and species trees is not rare, and potential causes are incongruent lineage sorting, hybridization or horizontal gene transfer.
- Coalescent-based methods are able to reconstruct species tree inference when gene trees are incongruent due to incomplete lineage sorting.

9.1 Incongruence in Phylogenomic Analyses

During the end of the 1990s and the early 2000s, molecular phylogenetic analyses revolutionized phylogenetic systematics. Many results contributed to changing textbook knowledge about the evolutionary relationships of plant and animal systematics and enabled a new picture for the phylogeny of the entire group of eukaryotes (Donoghue and Doyle 2000; Halanych 2004; Adl et al. 2005). Many of these early analyses were based on a single or few genes, leaving many nodes – especially deep in time – unsupported or unresolved. Current practice of phylogenomic analyses can be broadly classified into two different approaches: supermatrix and gene tree-based analyses of hundreds or thousands of genes (Liu et al. 2015). In the case of supermatrix analyses, all gene alignments are concatenated into a single matrix, which is subsequently analysed using the chosen phylogenetic method. In the case of gene tree-based analyses, all genes are analysed separately, and in a second step, the resulting topologies are (subsequently or simultaneously) used to construct a supertree (Bininda-Emonds 2004) or a species tree based on coalescent theory (▶ see Sect. 9.4). Phylogenomic approaches are able to produce precise estimations of phylogeny; however, this does not mean the result reflects the true evolutionary history (Kumar et al. 2012), as several factors can mislead phylogenetic analyses even when a massive amount of data is available.

The era of phylogenomic analyses to resolve relationships among organisms was basically kick-started in 2003. By analysing 106 different genes to resolve the phylogeny of yeast, Rokas et al. (2003) found incongruence among them, sometimes strongly supporting competing hypotheses (◘ Fig. 9.1). Using a genome-scale approach, the incongruence disappeared when combining all of them. Moreover, it was shown that a concatenation of any 20 out of these 106 genes always recovered the best topology with bootstrap values of at least 95% for each node. Even though details of this study have been criticized to be unrealistic (Gatesy et al. 2007), it clearly supported the idea that phylogenomic approaches could end incongruence in phylogenetics (Gee 2003). Whereas genome-scale approaches for most groups of non-model organisms remained a pipe dream in 2003, the availability of next-generation sequencing (NGS) techniques allowed gathering huge datasets for basically every taxon of interest (Rokas and Abbot 2009).

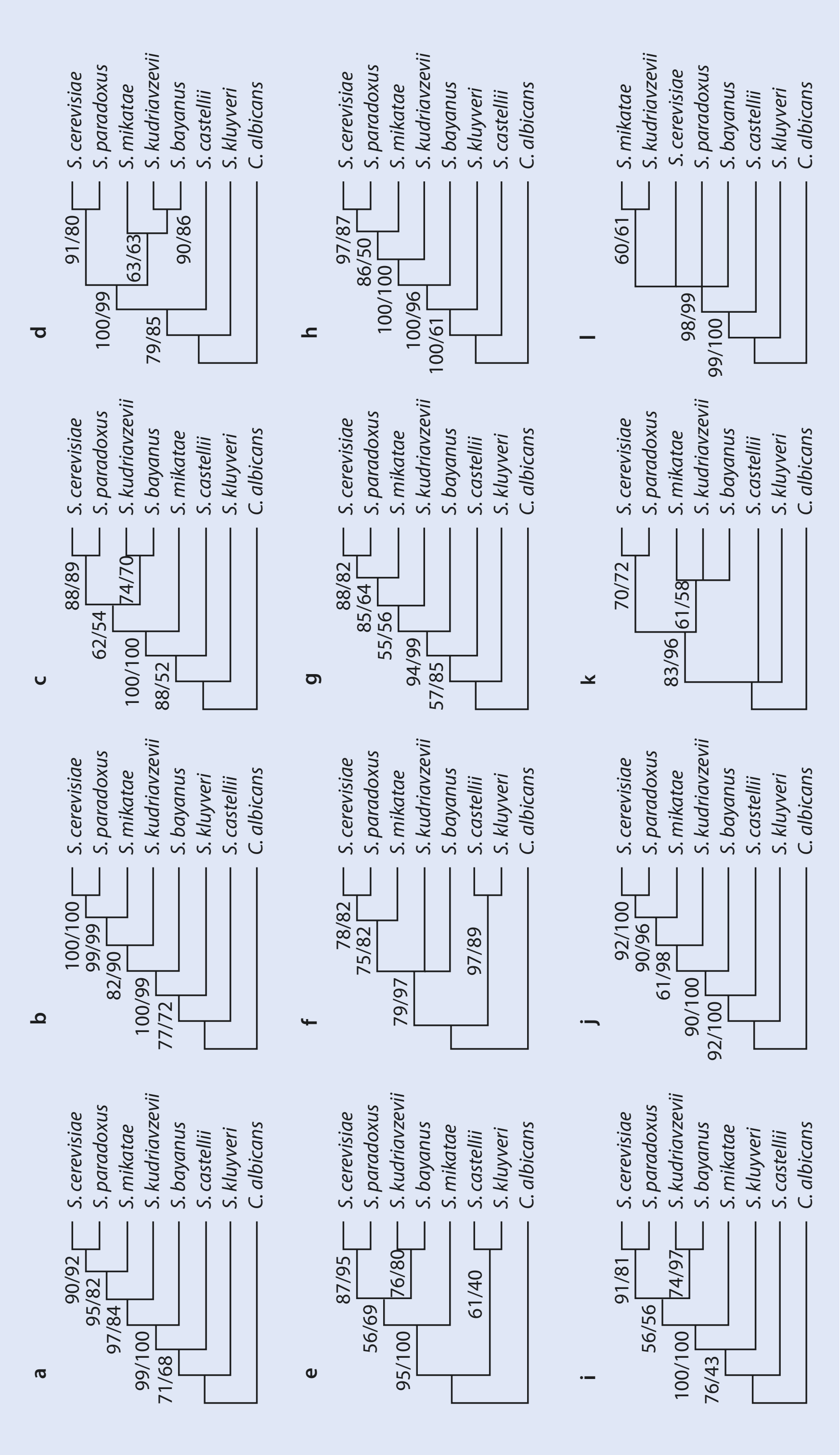


Fig. 9.1 **a–l** Incongruence among gene trees from a phylogenomic analysis of yeast relationships (Reprinted by permission from Macmillan Publishers Ltd.: [*Nature*] Rokas et al. (2003), copyright 2003)

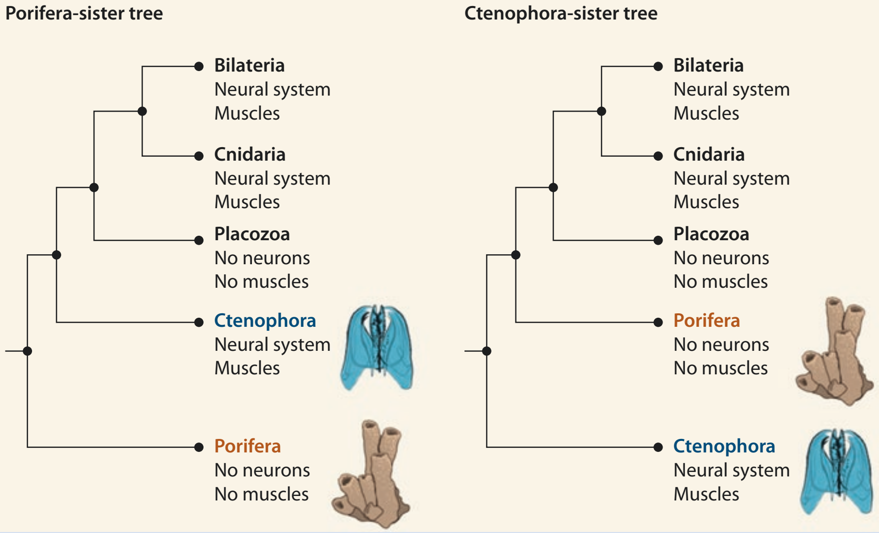
There are several reasons why trees inferred from single genes (i.e. gene trees) might differ with each other (Jeffroy et al. 2006). First, this might be a stochastic error associated with a lack of sufficient phylogenetic signal, which could be overcome by combining more (informative) genes. This approach assumes that combining more genes into a single data matrix should increase the phylogenetic signal-to-noise ratio compared to single genes (de Queiroz and Gatesy 2007). Second, the species tree will be different from a gene tree because of violation of the orthology assumption, incongruent lineage sorting or horizontal gene transfer. There are certain methods detecting such problems and dealing with them in phylogenomic datasets (▶ see Sect. 9.4). Third, systematic errors present in single genes might also lead to artefacts in the phylogenetic reconstruction (▶ see Sect. 9.2). Such systematic errors are usually due to the violation of assumptions of the underlying model for the analyses. Systematic errors can occur because the assumptions of the underlying model are violated, including (I) heterogeneity of the nucleotide/amino acid composition among lineages (compositional signal), (II) variation of the substitution rate among lineages (rate signal) and (III) variation in the substitution rate within nucleotide positions over time (heterotacheous signal). All these patterns are generally not accounted for by the evolutionary model and might negatively impact phylogenetic reconstruction.

Often, high statistical support (e.g. bootstrapping) is taken as a measure that the tree is correct. However, it is important to remember that these measures assess the stability of the obtained relationships to sampling error (White et al. 2007). Bootstrap analyses detect if datasets contain a pattern and how strong this is but are not able to decide whether or not this pattern represents genuine phylogenetic signal. Systematic error can negatively affect phylogenetic inference even with single genes, but it becomes stronger when multiple genes are combined into a supermatrix, simply because the addition of more (biased) genes will increase the support for a biased (wrong) result. As expressed by Jeffroy et al. (2006), phylogenomic analyses, rather than resolving the entire tree of life, might in fact be the beginning of incongruence (▶ see Infobox 9.1 for an example). Furthermore, combined datasets from hundreds of genes often contain large amounts of missing data (Roure et al. 2013), which could additionally influence the analysis (▶ see Sect. 9.3).

Infobox 9.1

Which Taxon Is the Sister Group of All Other Animals?

It was basically written in stone that sponges (Porifera) represent the sister taxon of all other animals, and it was rather discussed if sponges are monophyletic or if different sponge taxa branch off subsequently at the base of the animal tree (Sperling et al. 2007; Philippe et al. 2009). However, some phylogenomic analyses surprisingly started to find that the enigmatic Ctenophora (known as comb jellies or sea gooseberries) could represent the sister taxon of animals (Dunn et al. 2008; Moroz et al. 2014). This placement has important implications regarding how the evolution of several organ systems is understood (▣ Fig. 9.2) (Telford et al. 2016). Under the latter hypothesis, it has to be assumed either that the nervous system, muscles and epithelia evolved twice convergently or that all these characters were already present in the last common ancestor of animals and got lost in sponges. This controversy led to a heated debate about phylogenomics methodology and systematic error and how much trust can be put into phylogenomic analyses of very deep divergences. Proponents of the «Porifera-sister» scenario claimed that the result supporting the «Ctenophora-sister» hypothesis represents an LBA artefact, which might be introduced due to a poor fit of the used evolutionary models with the analysed data, as well as by the out-group choice (Pisani et al. 2015). In contrast, proponents of the «Ctenophora-sister» hypothesis analysed the sensitivity of phylogenomic analyses to model and gene choice (Whelan et al. 2015) and used an extensive taxon sampling. By analysing possible sources of systematic error, no biases affecting the position of Ctenophora as sister taxon to all other animals were found. Instead, some genes included in previous analyses supporting the «Porifera-sister» hypothesis were identified to introduce conflicting signal, thereby supporting a maybe wrong hypotheses of the placement of Ctenophora. This result is in line with a previous study by Nosenko et al. (2013), who by modifying gene and out-group taxon sampling were able to recover three different but well-supported phylogenies of non-bilaterian animals. This controversy remains still unresolved (Giribet 2016) and shifted to the question which models are suited to analyse datasets with massive substitutional heterogeneity and how to perform phylogenomic analyses for deep phylogenies (Whelan and Halanych 2016).



◘ **Fig. 9.2** Competing hypotheses regarding which taxon represents the sister group of all other animals and its evolutionary implications (Reprinted by permission from Macmillan Publishers Ltd.: [Nature] (Telford et al. 2016), copyright 2016)

9.2 Systematic Errors

The problem of systematic errors biasing phylogenetic analyses has been recognized early on by Felsenstein (1978). In this paper, he described conditions under which maximum parsimony (MP) inference is misled by the attraction of long branches in a tree irrespective of the true relationships (◘ Fig. 9.3). This phenomenon was termed «long edges attract» by Hendy and Penny (1989), and it is nowadays generally known as long-branch attraction (LBA). Despite maximum likelihood (ML) and Bayesian inference (BI) being more robust than MP to LBA (Philippe et al. 2005b), it was shown that probabilistic phylogenetic reconstruction methods could be also affected by LBA when the assumptions of the underlying model are violated by the data (Huelsenbeck 1995). Many simulation studies have shown that MP is the most sensitive method to the LBA artefact, whereas ML and BI are more robust (Philippe et al. 2005b). Even though LBA is often accounted for when phylogenetic analyses lead to unexpected results, a clear (statistically based) definition of the phenomenon is missing. Some authors defined LBA loosely as a condition where

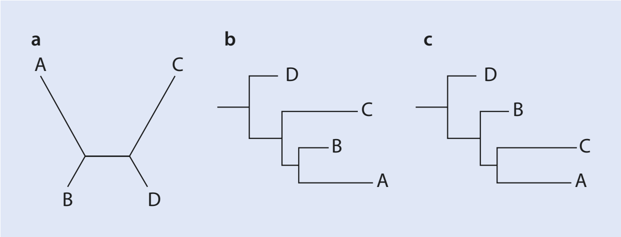


Fig. 9.3 **a** Unrooted four-taxon tree illustrating the classical example of long-branch attraction (LBA), with two long and two short branches, both unrelated. **b** A valid rooted tree of the unrooted topology shown in **a**. **c** Often analyses are misled by LBA, clustering together the long-branched terminals. This rooted topology is a typical artefact occurring in studies with tree A as the underlying true tree

analyses are biased due to a combination of short and long branches (Sanderson et al. 2000; Bergsten 2005), which basically translates to a bias due to variation of the substitution rate across lineages. Parks and Goldman (2014) systematically analysed the placement of long branches using simulation studies and found that also single long branches are difficult to place in a phylogeny, even when using ML. Interestingly they also found that there is no attraction between two long branches, even though they seem to be disproportionally often joined together. This observation has an impact on several approaches which were proposed to detect LBA in real datasets. For example, a common method was to remove one of the long branches from the analysis and to see if the placement of the other long branch remains consistent (Pol and Siddall 2001). However, as also the placement of single long branches is difficult, this might not be a good test. Other approaches to reduce LBA are the exclusion of terminals with very long branches (not an option when they are the taxon of interest) or the exclusion of fast-evolving genes or sites (Bergsten 2005; Pisani 2004; Rivera-Rivera and Montoya-Burgos 2016). Especially classifying all genes (or alignment sites) according to their evolutionary rate and successively removing them from the analysis starting with the fastest class will give a good overview if analyses are biased by the rate signal (Brinkmann et al. 2005). Finally, as LBA is basically a problem of model misspecification, the use of more sophisticated models is recommended. As such, it has been shown that site-heterogeneous CAT models are less affected by LBA due to their ability to better anticipate homoplasy in alignment site patterns (Lartillot et al. 2007), but also ML analyses with carefully selected partitions (and models for each partition) seem to be promising (Whelan and Halanych 2016). In summary, LBA is a very common yet not fully understood phenomenon, and the placement of long branches in phylogenetic analyses remains a difficult task.

Variation in the substitution rate across lineages (rate signal) can lead to the LBA phenomenon (Jeffroy et al. 2006), but this bias can often be handled by using models incorporating rate heterogeneity (Yang 1996). Additionally, the evolutionary rate of an alignment site can vary over time (heterotachy) (Lopez et al. 2002), and this process can also produce LBA (Lockhart and Steel 2005). A specific case of this phenomenon is known as the covarion hypothesis of molecular evolution, which states that substitutions at one alignment site may alter the substitution probability at other sites (Miyamoto and Fitch 1995). Kolaczkowski and Thornton (2004) used a clever simulation scheme to mimic another case of heterotachy. They simulated two sets of sequence alignments using the same topology, but under completely different models of DNA substitutions. By combining these two

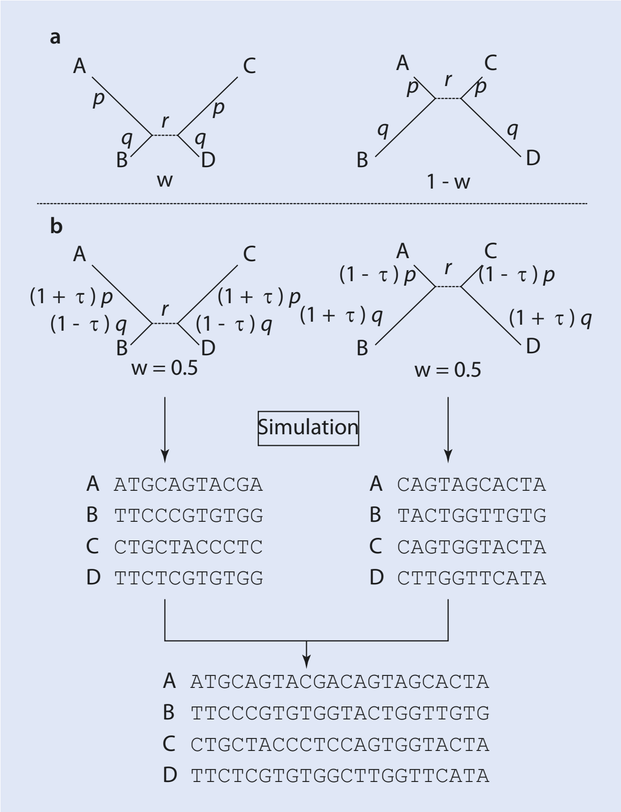


Fig. 9.4 Scheme for the simulation of different levels of heterotachy as used in Kolaczkowski and Thornton (2004). **a** Sequences are simulated under two different sets of branch lengths, including opposing sets of long (p) and short terminal branches. **b** Sequence alignments generated under this simulation scheme can be combined under different weights (w) to simulate different degrees of heterotachy (Figure reprinted from Philippe et al. (2005b))

datasets and giving different weights to the two data partitions, different levels of heterotachy were simulated (Fig. 9.4). Interestingly, these authors found that under higher levels of heterotachy, MP outperforms ML in recovering the correct tree. However, subsequent studies criticized this study for choosing very special and unrealistic parameters for their simulation, as well as for the way how ML analyses were conducted (Philippe et al. 2005b; Spencer et al. 2005). Instead, it could be shown that for realistic simulations of heterotacheous datasets, ML always outperforms MP and should be therefore the preferred method (Philippe et al. 2005b). This phenomenon of heterotachy has been demonstrated to be common in real datasets, where it affects phylogenetic reconstruction (Lopez et al. 2002; Whelan et al. 2011). Some statistical tests for the detection of heterotachy have been proposed (Wu and Susko 2011; Wang et al. 2011). Approaches specifically dealing with heterotachy are the CAT-BP model (Blanquart and Lartillot 2008), as well as a model allowing changing the rate heterogeneity as modelled by the gamma distribution along branches (Bouckaert and Lockhart 2015).

Another systematic error violating model assumptions is compositional bias, which describes significant differences in the nucleotide or amino acid composition across taxa. Most evolutionary models assume that the composition is homogenous across taxa. Several tests for compositional homogeneity are available, including frequency-dependent significance tests, matched-pairs tests or analyses based on Monte Carlo simulations of estimates of the standard deviation of the mean nucleotide or amino acid composition (Steel et al.

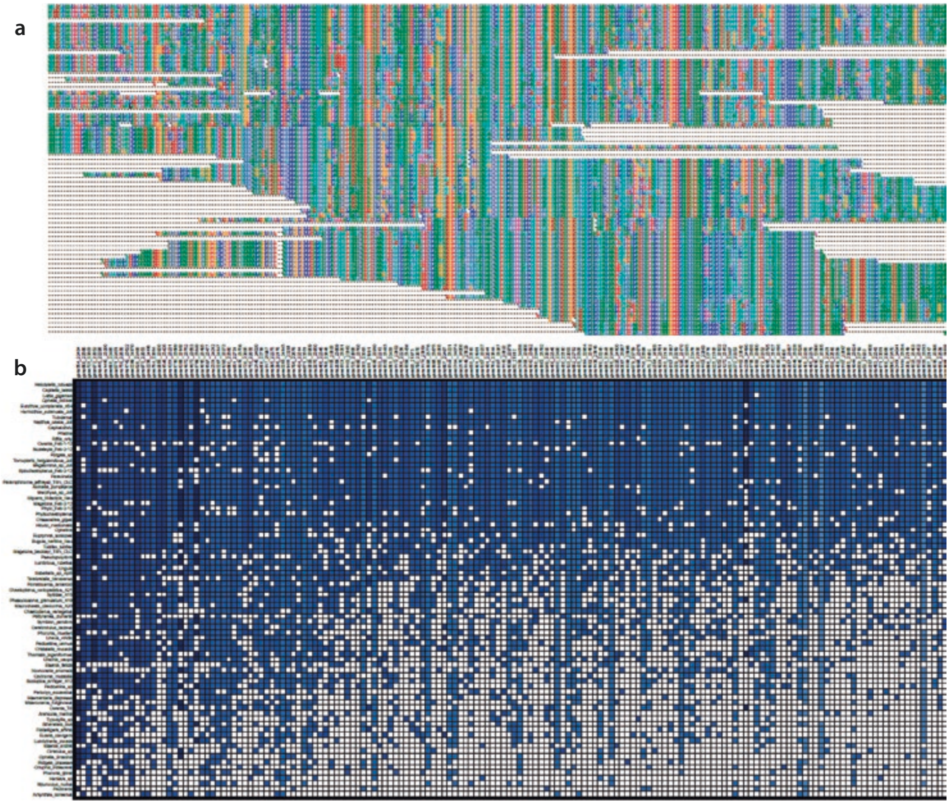
1993; Jermiin et al. 2004; Ababneh et al. 2006). With the software SEQVIS, it is possible to visualize compositional heterogeneity in nucleotide alignments (Ho et al. 2006).

A typical example of how compositional bias misleads phylogenetic analyses is that unrelated taxa with convergently evolved elevated GC content might group together, e.g. as demonstrated for drosophilids (Tarrío et al. 2001). Using simulation studies, Jermiin et al. (2004) found that the frequency of successful phylogenetic reconstruction is not only related to the difference in GC content (or base composition) but also to the length of internal branches. Analyses with short internal branches are more easily misled. Compositional bias is also related to rate variation, as especially fast-evolving sites are frequently compositionally biased (Rodríguez-Ezpeleta et al. 2007). Fittingly, third codon positions in protein-coding genes often have a stronger bias in composition, and their removal sometimes increases the accuracy of the phylogenetic analysis. One of the many negative effects of compositional heterogeneity can be the accumulation of convergencies. For example, transitions (replacement of a purine by a purine or pyrimidine by a pyrimidine) are usually more frequently observed than transversions (replacement of a purine by a pyrimidine or reverse), leading to coincident substitutions. It has been shown that recoding all nucleotides to R (purines, A and G) and Y (pyrimidines, C and T) reduces this misleading effect of compositional bias (Phillips and Penny 2003). Recoding can, for example, be conducted with the software BMGE (Criscuolo and Gribaldo 2010), which furthermore is able to identify and exclude characters which contribute to compositional biases based on a matched-pairs test of marginal symmetry. Finally, non-homogeneous nonstationary models that account for variations in the base composition can be used. The model of DNA sequence evolution by Galtier and Gouy (1998), which is implemented in PHYML (Boussau and Gouy 2006), allows varying equilibrium GC contents among lineages and estimation of five parameters: (I) ancestral GC content, (II) location of the root in its branch, (III) transition/transversion ratio, (IV) branch lengths and (V) equilibrium GC contents in each branch. Compositional bias was expected to be more frequent and also misleading on the nucleotide level, as only four different states exist and convergence is to be expected (Hasegawa and Hashimoto 1993; Foster and Hickey 1999). However, compositional bias on the protein level seems also to be frequent and thereby a problem for phylogenetic analyses as well (Lartillot and Philippe 2008; Nesnidal et al. 2010). Kück and Struck (2014) developed a package of scripts to analyse phylogenomic datasets (BACOCA), which can be used to investigate the compositional bias among amino acids. As with nucleotides, recoding of amino acids can reduce the compositional bias. The most commonly used recoding classifies amino acids according to six groups identified by Dayhoff et al. (1978), which tend to replace each other (Susko and Roger 2007). Furthermore, using the CAT-BP model for amino acid data allows lineage-specific compositional shifts across the phylogeny and thus deals with heterogeneous amino acid sequence compositions (Blanquart and Lartillot 2008).

9.3 Missing Data, Phylogenetic Information Content and Taxon Sampling

9.3.1 Missing Data

A typical way to compile a dataset for phylogenomic studies involves the generation of transcriptomes and subsequent selection of putative orthologs for the analyses. Ortholog sets often range from 100 to more than 1000 genes, and it is not unusual that not all genes are (completely) recovered for all taxa. As such, orthologs are often found incomplete using transcriptome sequencing (◘ Fig. 9.5a). In most cases, missing genes are due to the depth of the sequenced transcriptome or they are just not expressed in the sampled specimen (Roure et al. 2013). Moreover, many genes might have been lost for some taxa during evolution (◘ Fig. 9.5b). Percentages of missing data up to 80% have been reported for phylogenomic studies (Hejnol et al. 2009). The discussion if missing data should be reduced from phylogenetic analyses, e.g. excluding the most incomplete taxa and/or characters, has a long tradition in the literature (Wiens 2003; Wiens and Morrill 2011; Philippe et al. 2004; Wiens 1998). Initially, the question arose if incompletely sampled taxa should be included in phylogenetic analyses of one or few genes or in morphological character matrices. In the latter case, the discussion often centred on fossils, for which it was usually impossible to analyse all characters found in recent taxa. Later the discussion was expanded to genomic datasets, where often substantial amounts of data are missing. Even though some publications addressed missing data as problematic (Lemmon et al. 2009), most studies using real or simulated data could show that the inclusion of incomplete taxa is usually advantageous. One simple reason is that an improved taxon sampling



◘ **Fig. 9.5** Missing data in phylogenomic analyses. **a** Single gene alignment based on transcriptomic data often includes highly incomplete and partially nonoverlapping gene sequences. **b** The gene coverage (*columns*) is often highly uneven for taxa (*rows*) included in a phylogenomic study. *Blue squares* show presence of genes, *white squares* show absent genes. Matrix based on data from Weigert et al. (2014) constructed with MARE (Misof et al. 2013) helps to break long branches (Roure et al. 2013). By analysing a large dataset covering diverse eukaryotes, Philippe et al. (2004) could show that 25% of missing data in the original dataset did not negatively impact the analyses. Subsequent random deletion of 50% of the character matrix did not alter the outcome of the analysis, and even when analysing with up to 90% of missing data, similar trees could be obtained. Jiang et al. (2014) found that that adding incomplete data is in particular helpful for resolving poorly supported nodes and showed that missing data does not consistently bias branch lengths. Finally, Hovmöller et al. (2013) have shown that also species tree reconstruction methods relying on coalescent approaches (▶ see Sect. 9.4) are remarkably robust under the presence of up to 50% of missing data. However, if missing data is nonrandomly distributed over the matrix, it may bias analyses, leading to many trees (or subtrees) which are nearly indistinguishable by its likelihood value (Sanderson et al. 2010). A tool for the visualization of the completeness of the supermatrix (◘ Fig. 9.5b), as well as for the exclusion of incompletely sampled genes, is the software MARE (Misof et al. 2013). Using such an approach, differently covered data matrices can be constructed and analysed, and the sensitivity of phylogenomic analyses to missing data can be assessed (Weigert et al. 2014).

9.3.2 More Genes or More Taxa?

Taxon sampling has been profusely discussed in the phylogenetic literature prior to the genomic era. In particular, whether it was better centres the efforts in obtaining more data for a number of taxa or more taxa with relatively fewer data (Rokas and Carroll 2005; Mitchell et al. 2000). This discussion lost power with the (comparatively) cheap price of NGS technologies, which allows the recovery of large amounts of sequences for non-model taxa, and in most cases adding more data is not a bottleneck anymore. The first phylogenomic analyses often relied on a handful of model taxa where complete genomes were available. For example, focussing on animal relationships, these analyses seemed to support the so-called Coelomata hypothesis (arthropods + deuterostomes) and not the widely accepted Ecdysozoa hypothesis (arthropods + nematodes) (Philip et al. 2005). However, these results have been clearly demonstrated to be an artefact related to a limited taxon sampling (Philippe et al. 2005a). The discussion of experimental design has now shifted to which genes and which taxa to include in an analysis (Philippe et al. 2011).

9.3.3 Taxon Sampling

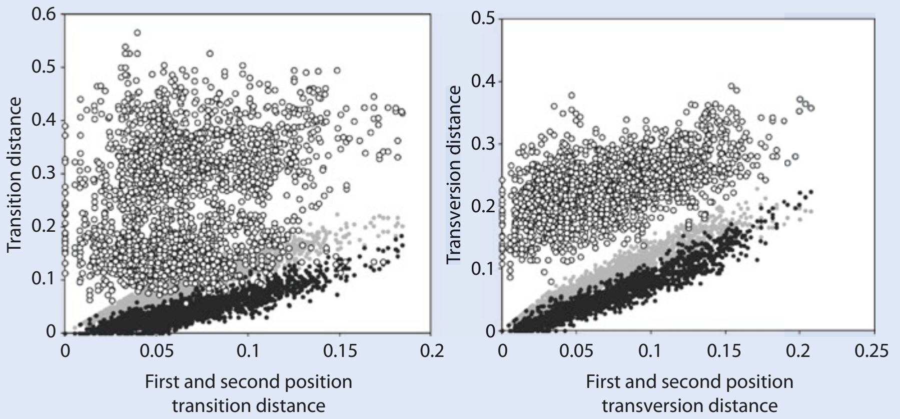
The importance of taxon sampling for phylogenetic analyses is widely acknowledged (Heath et al. 2008; Pollock et al. 2002; Zwickl and Hillis 2002), with only few studies coming to a different conclusion (Rosenberg and Kumar 2001). Rannala et al. (1998) demonstrated in a simulation study that a decrease in taxon sampling leads to an increase in the average branch length of terminals, which could make analyses more susceptible to LBA. This is in line with the finding that the estimation of rate heterogeneity is highly sensitive to taxon sampling (Sullivan et al. 1999). Moreover, estimation of branch lengths becomes also more challenging due to the so-called node density effect under a limited taxon sampling (Hugall and Lee 2007). This effect often leads to an underestimation of branch lengths in sparsely sampled tree regions, because less information is available to infer multiple substitutions, which could have been revealed under the presence of additional nodes. However, not all included taxa are equally helpful to improve phylogenetic analyses. Certain taxa, so-called rogue taxa, can show a phylogenetically unstable behaviour, characterized by widely different positions in tree topologies estimated from the same dataset (e.g. within bootstrap replicates) (Sanderson and Shaffer 2002). Often, but not always, rogue taxa are characterized by showing large amounts of missing data. Inclusion of such rogue taxa can have a negative impact on support values (especially when using bootstrap), but could also influence tree reconstruction in general (Mariadassou et al. 2012). In fact, Aberer et al. (2013) demonstrated that exclusion of rogue taxa increases the accuracy of phylogenetic analyses. These authors developed an algorithm for the identification and subsequent pruning of rogue taxa, implemented in the software ROGUENAROK. The idea behind the algorithm is to identify taxa, which exclusion results into an increase of support in bootstrap consensus trees. The measure of change in support is called relative bipartition information criterion (RBIC), which is the sum of all support values divided by the maximum support in a fully bifurcating tree of the original dataset. Taxa or combinations of taxa yielding the highest change in RBIC are excluded from the analysis. This analysis can be iteratively repeated until no significant change is observed. Alternatively, the leave stability index (LSI) has been used to identify rogue taxa. The LSI uses the occurrence of taxon triplets in trees from bootstrap analyses (Thorley and Wilkinson 1999). Three different possibilities for the relationship of three taxa (A, B, C) exist in a rooted, bifurcated tree: ((A, B), C), ((A, C), B) and ((B, C), A). The LSI is calculated as the difference of the relative frequency of the most common triplet and the second most common and is averaged over all triplets containing a certain taxon. LSI values of 1 or close to 1 indicate stable taxa, where values closer to 0 indicate instability. A LSI cut-off value can be defined for rogue taxa to be excluded from the analysis. Inference of the LSI is, for example, incorporated in the software PHYUTILITY (Smith and Dunn 2008). A third approach called multiple co-inertia analysis (MCOA) has been explored by de Vienne et al. (2012), which is based on the comparison of pairwise distances between species in all gene tree topologies to identify rogue taxa (described as outlier taxa in this publication).

9.3.4 Gene Sampling

Gene alignments can differ in their missing data, sequence saturation or phylogenetic information content. DNA and protein sequences are regarded as saturated, when sites have undergone multiple substitutions and the number of observed differences no longer reflects «true» evolutionary distances. Slight levels of saturation are corrected by the use of models of sequence evolution, but more saturated sequence alignments can mislead phylogenetic reconstruction. When analysing highly saturated sequences, phylogenetic inference can be driven by sequence composition to a large extent rather than true phylogeny (Xia et al. 2003). DNA sequences are normally more affected by saturation because only four different character states exist compared to the 20 states of amino acids (Philippe et al. 2011). However, saturation can also be problematic at the amino acid level (Van de Peer et al. 2002). A simple method to check for the presence of saturation in nucleotide sequences is by separately plotting the raw numbers of substitutions (p uncorrected distance) of transitions and transversions of all pairwise comparisons of taxa in an alignment against their genetic (usually ML-corrected) distance (Struck et al. 2008). For most protein-coding genes, transitions occur more frequently than transversions and thus are more likely saturated (◘ Fig. 9.6). Formalized measures of substitution saturation have been introduced by Xia et al. (2003), as implemented in the software DAMBE (Xia 2013), and Struck et al. (2008), as implemented in the BACOCA package of scripts (Kück and Struck 2014). Possible strategies to deal with saturated sequences are use of amino acids, exclusion of the saturated data or recoding (e.g. RY coding or the use of Dayhoff categories for amino acids).

It is important to remember the relationship between sequence saturation and sequence divergence: one gene might be saturated for old divergences but well suited to resolve young divergences, whereas a slower-evolving gene might not be saturated for old divergences but totally uninformative for young ones. The usefulness of a given gene for phylogenetic analyses can be estimated by its phylogenetic informativeness (PI) (Townsend 2007). Briefly summarized, PI estimates the probability that a character resolves a dated four-taxon alignment (more than four taxa can be analysed by providing a consistent topology). Thereby, PI provides an estimate of the amount of phylogenetic signal relative to noise across time (◘ Fig. 9.7). PI can be analysed using the software PHYDESIGN (López-Giráldez and Townsend 2011), which is available online, by providing an alignment, as well as an ultrametric tree as input. Some updates and modifications for the calculation of PI are available in the R package PHYLINFORMR (Dornburg et al. 2016). As an example on how to use PI, in ◘ Fig. 9.7, the utility of different classes of phylogenetic markers from percomorph fishes are compared (Gilbert et al. 2015).

A different approach to investigate and visualize phylogenetic information content is based on likelihood mapping (Strimmer and von Haeseler 1997). This method analyses possible four-taxon cases of a given dataset, called quartets. For every quartet, there are three possible fully resolved tree topologies, for which the posterior probability for each of the three possible topologies can be estimated using Bayes' theorem. The three



◘ **Fig. 9.6** Saturation at different codon positions. Uncorrected pairwise distances are plotted for pairs of taxa, separately for transitions (*left*) and transversions (*right*) and first (*grey*), second (*black*) and third (*white*) codon positions. For unsaturated sequences, the number of substitutions should increase linearly with time (e.g. transversions on first and second positions), whereas for saturated sequences, no increase in the number of substitutions is detected with increasing genetic distance (e.g. transitions on third codon positions) (Reprinted from Dávalos and Perkins (2008), with permission from Elsevier)

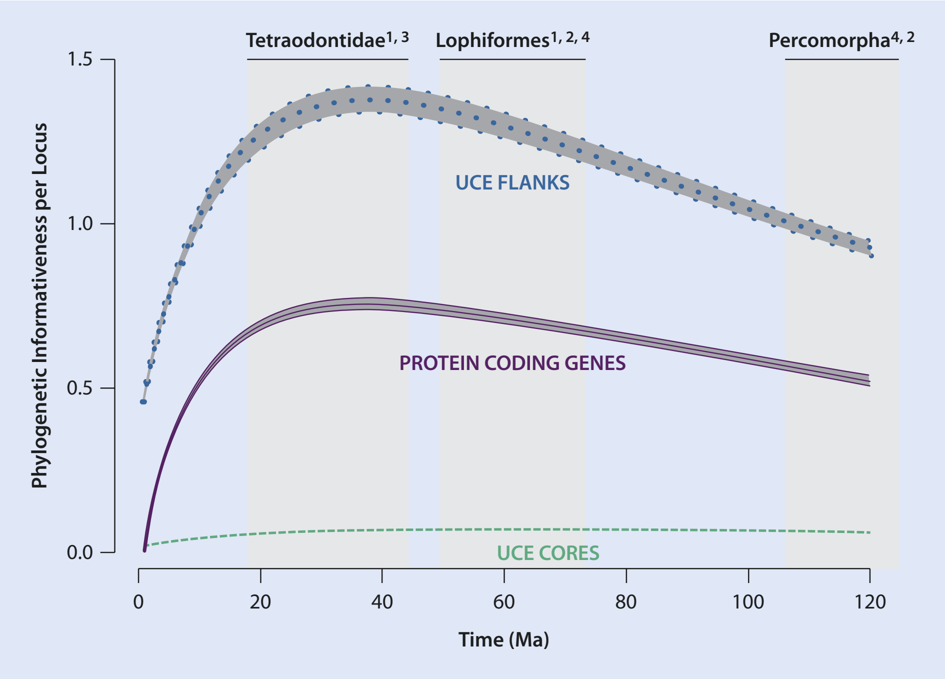


Fig. 9.7 Phylogenetic informativeness and its 95% confidence interval of three different classes of phylogenetic markers from percomorph fishes (UCE core regions, UCE flanking regions, protein-coding genes) plotted against time. Core regions of ultraconserved elements (UCEs) are basically uninformative, whereas flanking regions of UCE show a higher PI than protein-coding genes, with the highest resolution power for divergences between 20 and 40 million years old (Reprinted from (Gilbert et al. 2015), with permission from Elsevier)

posterior probabilities are then used as coordinates to locate a point within a triangular graph where each corner represents one topology. This calculation is repeated for all possible quartets, which are subsequently plotted in the triangle. In the case of an uninformative quartet (starlike evolution), all three probabilities are the same and the point is located in the middle of the triangle. If one tree topology is clearly supported with a probability close to 1, this would point to one of the corners of the triangle (according the supported topology). If two topologies gain similar probability, whereas one topology gets a probability close to 0, the point would be located at one edge of the triangle, between the corners representing the two supported topologies. By analysing all possible quartets of a dataset, the phylogenetic information content can be visualized. The more quartets can be located in one of the corners of the triangle, the higher is the information content of the dataset (Fig. 9.8). Likelihood mapping is implemented in the software TREE-PUZZLE (Schmidt et al. 2002).

Different strategies have been used to select sets of orthologous genes for phylogenetic analyses. Some authors recommend to only include highly informative genes in the analysis (Salichos and Rokas 2013), whereas others suggest that phylogenetic signal can be basically extracted from all ortholog alignments when combined in a supermatrix (Gatesy and Baker 2005). PI represents a possible way to select genes which are suitable for both, supermatrix and coalescent-based methods. Shen et al. (2016) systematically analysed the

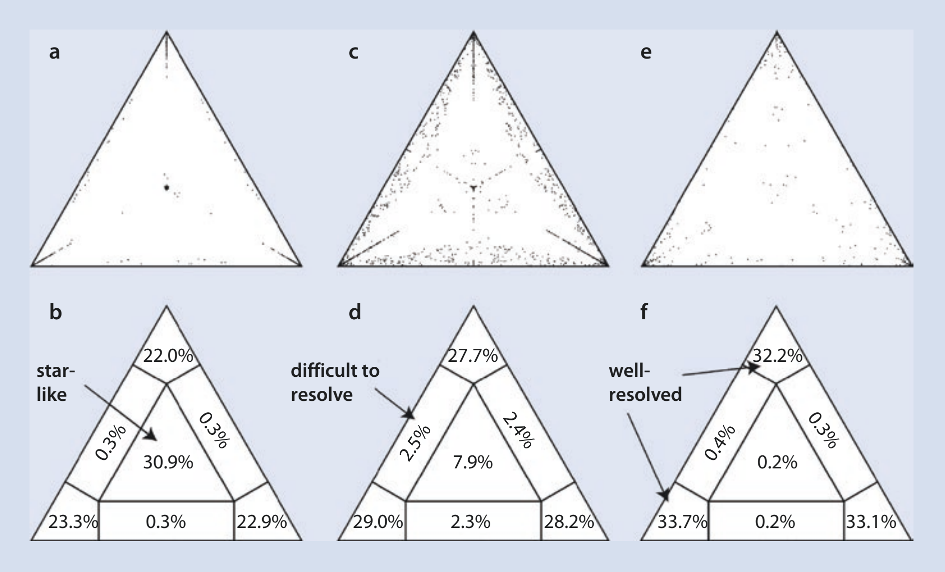


Fig. 9.8 Likelihood mapping using TREE-PUZZLE (Schmidt et al. 2002) for datasets with differences in phylogenetic information content. **a**, **b** In a dataset with low information content, a high percentage (30.9%) of quartets represent starlike evolution. **c**, **d** In this dataset 7.9% of the quartets represent starlike evolution, whereas 2.5% + 2.3% + 2.4% of quartets are in an area where it is difficult to distinguish between two of the three possible tree alternatives. **e**, **f** Most quartets (33.7% + 32.2%, 33.1%) are in well-resolved areas of the tree distribution, indicating high phylogenetic information content. **a**, **c**, **e** show distribution patterns of mapped quartets; **b**, **d**, **f** show occupancies (in percent) for seven areas of interest

association between sequence-based properties, gene function-based properties and gene tree-based properties with phylogenetic information content. The goal was to identify those properties which predict phylogenetic signal of a gene best. Even though most of the investigated properties correlate with each other, a set of properties with the highest relevance could be identified. Interestingly, the most important property to predict phylogenetic signal is gene alignment length, followed by number of parsimony-informative sites and variable sites. This result could be interpreted in favour of binning genes for coalescent analyses (see above), but also for the use of the supermatrix approach, which basically combines all alignments into a highly informative «supergene».

9.4 Incongruence Between Gene Trees and Species Trees

Gene trees may differ from the species tree simply by the stochastic sampling of alleles during speciation events (Degnan and Rosenberg 2009), a phenomenon known as incomplete lineage sorting or deep coalescence (Fig. 9.9). The term «hemiplasy» has been coined to describe incorrect inference of character-state evolution due to genetic polymorphisms which are retained across speciation events (Avise and Robinson 2008; Hahn and Nakhleh 2016). This term should reflect that in this case similarity does not reflect common ancestry, even though the considered character states are homologous (and apomorphic!).

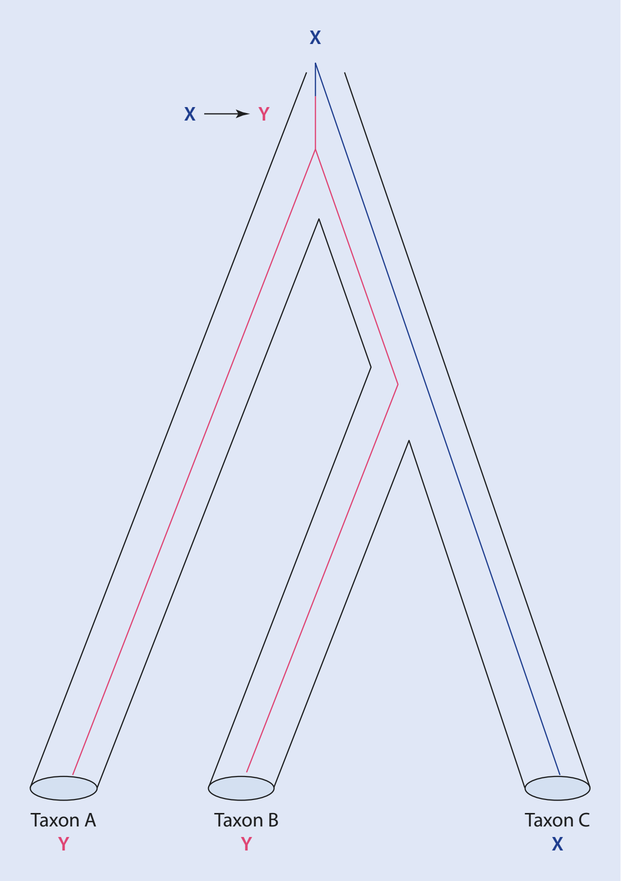


Fig. 9.9 Incomplete lineage sorting can lead to incongruence between gene trees and species trees. The gene tree is drawn in colour inside the species tree (*black*). The last common ancestor of taxa **a**–**c** had two paralogs of a gene (X and Y). Duplicates got lost before the split of the three species, but paralog sorting is incongruent with the species tree

It has been demonstrated that discordance between gene trees and species trees is common, especially in cases where speciation events happened in short time spans, i.e. separated by short branches (Degnan and Rosenberg 2006). A good example of incomplete lineage sorting is represented by the genome-scale analyses of the bird phylogeny, which includes a rapid radiation characterized by many short internal branches. For this phylogeny, not a single gene tree has been found to match the reconstructed species tree (Jarvis et al. 2014). Later on, lineage sorting has been shown to be frequent in the evolutionary history of birds, and a phylogenetic network was used to illustrate their complex history (Fig. 9.10) (Suh et al. 2015).

Several other evolutionary processes can lead to the disagreement between gene trees and species trees, including horizontal gene transfer (HGT), gene duplication and hybridization (Maddison 1997; Knowles and Kubatko 2010). HGT is a process where genes are transferred from one species to another across the phylogeny. Whereas HGT is rather rare in eukaryotes and therefore less problematic for phylogenetic reconstruction, it is common among prokaryotes (Ku and Martin 2016). Gene duplication complicates the inference of orthology (Philippe et al. 2011). Hybridization and introgression are biological processes by which the genetic material of two different species gives rise to hybrids and sometimes new species. Hybridization is most commonly found in plants, but also many examples have been described for animals (Mallet 2007).

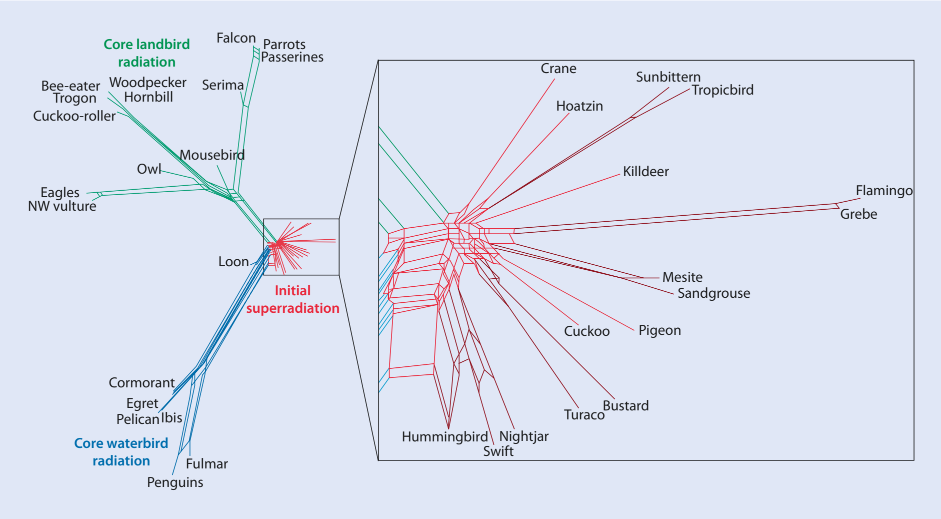


Fig. 9.10 Phylogenetic network analyses of rare genomic change markers reveal a strong discordance of markers, which can be explained by high levels of incomplete lineage sorting (Figure reprinted from Suh et al. (2015))

Several phylogenetic methods have been developed to detect and deal with incongruence of gene trees and species trees. In contrast to the supermatrix approach, where genes are concatenated into one single matrix, these methods are usually based on the separate reconstruction of gene trees, which are subsequently (or simultaneously) used to infer the species tree. Most species tree inference methods are rooted within the coalescence theory, a model which has been developed to follow the history of genes (or alleles) back in time. Coalescence models are commonly used in population genetics and are often based on the Wright-Fisher model of genetic drift, assuming nonoverlapping generations, neutral evolution and random joining of populations back in time (Degnan and Rosenberg 2009). The multispecies coalescent (MSC) is used to estimate the probability distribution of gene trees evolving along the branches of a species tree. Each branch of a species tree represents a single population, and lineages of genes entering these populations are traced back through time to a common ancestor at rates given by the model. The coalescence of different gene lineages of the gene trees finally provides the signal for the inference of the overlying species tree (Liu et al. 2015). The MSC has been implemented into ML approaches, e.g. STEM (Kubatko et al. 2009) or MP-EST (Liu et al. 2010), and a Bayesian framework, e.g. BEST (Liu 2008) or BEAST (Drummond et al. 2012). The performance of species tree inference methods is controversially discussed. Gatesy and Springer (2014) criticized that species tree inference is often misled by unreliable gene trees, especially when dealing with phylogenetic analyses at deep timescales. Similar to the idea that the phylogenetic signal-to-noise ratio gets improved by using concatenation of single gene alignments into a supermatrix, statistical binning of genes with a similar signal has been proposed to reduce gene tree estimation errors for species tree inference (Mirarab et al. 2014). Several simulation studies show a superior performance of species tree inference using a Bayesian framework in comparison with other methods, especially in the case when a high probability of gene tree discordance is simulated (Leaché and Rannala 2011). Interestingly, comparison of results from species tree inference and supermatrix methods for real datasets often show rather consistent results (Liu et al. 2015).

For the quantification of incongruence in phylogenomic datasets, Salichos and Rokas (2013) developed a measure called internode certainty (IC). Here, incongruence for a given internal node is measured by calculating the frequency of a bipartition found in the best tree in a given set of gene trees together with the occurrence of conflicting bipartition in these gene trees. Values close to 0 indicate the presence of strong conflict, whereas values close to 1 indicate the absence of conflictive signal. Summing overall ICs will give the tree certainty (TC). The calculation of IC and TC is implemented within the software RAxML (Stamatakis 2014; Kobert et al. 2016).

References

Ababneh F, Jermiin LS, Ma C, Robinson J (2006) Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225–1231

Aberer AJ, Krompass D, Stamatakis A (2013) Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst Biol 62:162–166

Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle GUY, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup Ø, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MFJR (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52:399–451

Avise JC, Robinson TJ (2008) Hemiplasy: a new term in the lexicon of phylogenetics. Syst Biol 57:503–507

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

Bininda-Emonds ORP (2004) The evolution of supertrees. Trends Ecol Evol 19:315–322

Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842–858

Bouckaert R, Lockhart P (2015) Capturing heterotachy through multi-gamma site models. bioRxiv. doi.org/10.1101/018101

Boussau B, Gouy M (2006) Efficient likelihood computations with nonreversible models of evolution. Syst Biol 55:756–768

Brinkmann H, van der Giezen M, Zhou Y, de Raucourt GP, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743–757

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 10:210

Dávalos LM, Perkins SL (2008) Saturation and base composition bias explain phylogenomic conflict in Plasmodium. Genomics 91:433–442

Dayhoff M, Schwarz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Dayhoff M (ed) Atlas of protein sequence and structure, vol 5, Suppl. 3. National Biomedical Research Foundation. Washington, DC, pp 345–352

de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. Trends Ecol Evol 22:34–41

de Vienne DM, Ollier S, Aguileta G (2012) Phylo-MCOA: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. Mol Biol Evol 29:1587–1598

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2:e68

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24:332–340

Donoghue MJ, Doyle JA (2000) Seed plant phylogeny: demise of the anthophyte hypothesis? Curr Biol 10:R106–R109

Dornburg A, Fisk JN, Tamagnan J, Townsend JP (2016) PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evol Biol 16:262

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale

MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–750

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol 48:284–290

Galtier N, Gouy M (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol Biol Evol 15:871–879

Gatesy J, Baker RH (2005) Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol 54:483–492

Gatesy J, DeSalle R, Wahlberg N (2007) How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst Biol 56:355–363

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) Evolution: ending incongruence. Nature 425:782–782

Gilbert PS, Chang J, Pan C, Sobel EM, Sinsheimer JS, Faircloth BC, Alfaro ME (2015) Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Mol Phylogenet Evol 92:140–146

Giribet G (2016) Genomics and the animal tree of life: conflicts and future prospects. Zool Scr 45:14–21

Hahn MW, Nakhleh L (2016) Irrational exuberance for resolved species trees. Evolution 70:7–17

Halanych KM (2004) The new view of animal phylogeny. Annu Rev Ecol Syst 35:229–256

Hasegawa M, Hashimoto T (1993) Ribosomal RNA trees misleading? Nature 361:23–23

Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–257

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, Wiens M, Müller WEG, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn CW (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc R Soc Lond B Biol Sci 276:4261–4270

Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Biol 38:297–309

Ho JWK, Adams CE, Lew JB, Matthews TJ, Ng CC, Shahabi-Sirjani A, Tan LH, Zhao Y, Easteal S, Wilson SR, Jermiin LS (2006) SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22:2162–2163

Hovmöller R, Lacey Knowles L, Kubatko LS (2013) Effects of missing data on species tree estimation under the coalescent. Mol Phylogenet Evol 69:1057–1062

Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. Syst Biol 44:17–48

Hugall AF, Lee MSY (2007) The likelihood node density effect and consequence for evolutionary studies of molecular rates. Evolution 61:2293–2307

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldón T, Capella-Gutiérrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MPC, Prosdocimi F, Samaniego JA, Velazquez AMV, Alfaro-Núñez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong Z, Zeng Y, Liu S, Li Z, Liu B, Wu K, Xiao J, Yinqi X, Zheng Q, Zhang Y, Yang H, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jønsson KA, Johnson W, Koepfli K-P, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, McCormack J, Burt D, Ellegren H, Alström P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MTP, Zhang G (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol 53:638–643

Jiang W, Chen S-Y, Wang H, Li D-Z, Wiens JJ (2014) Should genes with missing data be excluded from phylogenetic analyses? Mol Phylogenet Evol 80:308–318

Knowles LL, Kubatko LS (2010) Estimating species trees: an introduction to concepts and models. In: Knowles LL, Kubatko LS (eds) Estimating species trees: practical and theoretical aspects. Wiley-Balckwell, Hoboken, pp 1–14

Kobert K, Salichos L, Rokas A, Stamatakis A (2016) Computing the internode certainty and related measures from partial gene trees. Mol Biol Evol 33:1606–1617

Kolaczkowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984

Ku C, Martin WF (2016) A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. BMC Biol 14:89

Kubatko LS, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973

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

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K (2012) Statistics and Truth in Phylogenomics. Mol Biol Evol 29:457–472

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:S4

Lartillot N, Philippe H (2008) Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond Ser B Biol Sci 363:1463–1472

Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: a comparison of methods. Syst Biol 60:126–137

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Syst Biol 58:130–145

Liu L (2008) BEST: bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543

Liu L, Xi Z, Wu S, Davis CC, Edwards SV (2015) Estimating phylogenetic trees from genome-scale data. Ann N Y Acad Sci 1360:36–53

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

Lockhart P, Steel M (2005) A tale of two processes. Syst Biol 54:948–951

López-Giráldez F, Townsend JP (2011) PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol Biol 11:152

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

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536

Mallet J (2007) Hybrid speciation. Nature 446:279–283

Mariadassou M, Bar-Hen A, Kishino H (2012) Taxon influence index: assessing taxon-induced incongruities in phylogenetic inference. Syst Biol 61:337–345

Mirarab S, Bayzid MS, Boussau B, Warnow T (2014) Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346 1250463.

Misof B, Meyer B, von Reumont BM, Kück P, Misof K, Meusemann K (2013) Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics 14:348

Mitchell A, Mitter C, Regier JC (2000) More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of noctuoidea (Insecta: lepidoptera). Syst Biol 49:202–224

Miyamoto MM, Fitch WM (1995) Testing the covarion hypothesis of molecular evolution. Mol Biol Evol 12:503–513

Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov YV, Swore JJ, Girardo DO, Fodor A, Gusev F, Sanford R, Bruders R, Kittler E, Mills CE, Rast JP, Derelle R, Solovyev VV, Kondrashov FA, Swalla BJ, Sweedler JV, Rogaev EI, Halanych KM, Kohn AB (2014) The ctenophore genome and the evolutionary origins of neural systems. Nature 510: 109–114

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

Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WEG, Nickel M, Schierwater B, Vacelet J, Wiens M, Wörheide G (2013) Deep metazoan phylogeny: when different genes tell different stories. Mol Phylogenet Evol 67:223–233

Parks SL, Goldman N (2014) Maximum likelihood inference of small trees in the presence of long branches. Syst Biol 63:798–811

Philip GK, Creevey CJ, McInerney JO (2005) The opisthokonta and the ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the coelomata than ecdysozoa. Mol Biol Evol 22:1175–1184

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9:e1000602

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Quéinnec E, Da Silva C, Wincker P, Le Guyader H, Leys S, Jackson DJ, Schreiber F, Erpenbeck D, Morgenstern B, Wörheide G, Manuel M (2009) Phylogenomics revives traditional views on deep animal relationships. Curr Biol 19:706–712

Philippe H, Lartillot N, Brinkmann H (2005a) Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. Mol Biol Evol 22:1246–1253

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH, Casane D (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21:1740–1752

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005b) Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50

Phillips MJ, Penny D (2003) The root of the mammalian tree inferred from whole mitochondrial genomes. Mol Phylogenet Evol 28:171–185

Pisani D (2004) Identifying and removing fast-evolving sites using compatibility analysis: an example from the arthropoda. Syst Biol 53:978–989

Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, Lartillot N, Wörheide G (2015) Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci U S A 112:15402–15407

Pol D, Siddall ME (2001) Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. Cladistics 17:266–281

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM (2002) Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 51:664–671

Rannala B, Huelsenbeck JP, Yang Z, Nielsen R (1998) Taxon sampling and the accuracy of large phylogenies. Syst Biol 47:702–710

Rivera-Rivera CJ, Montoya-Burgos JI (2016) LS^3: a method for improving phylogenomic inferences when evolutionary rates are heterogeneous among taxa. Mol Biol Evol 33:1625–1634

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56:389–399

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. Trends Ecol Evol 24:192–200

Rokas A, Carroll SB (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337–1344

Rokas A, Williams B, King N, Caroll S (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804

Rosenberg MS, Kumar S (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci U S A 98:10751–10756

Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol 30:197–214

Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331

Sanderson MJ, McMahon MM, Steel M (2010) Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol Biol 10:155

Sanderson MJ, Shaffer HB (2002) Troubleshooting molecular phylogenetic analyses. Annu Rev Ecol Syst 33:49–72

Sanderson MJ, Wojciechowski MF, Hu J-M, Khan TS, Brady SG (2000) Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol Biol Evol 17:782–797

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Shen X-X, Salichos L, Rokas A (2016) A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol Evol 8:2565–2580

Smith SA, Dunn CW (2008) Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716

Spencer M, Susko E, Roger AJ (2005) Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161–1164

Sperling EA, Pisani D, Peterson KJ (2007) Poriferan paraphyly and its implications for Precambrian palaeobiology. Geol Soc Lond Spec Publ 286:355–368

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Steel MA, Lockhart PJ, Penny D (1993) Confidence in evolutionary trees from biological sequence data. Nature 364:440–442

Strimmer K, von Haeseler A (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A 94:6815–6819

Struck TH, Nesnidal MP, Purschke G, Halanych KM (2008) Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa). Mol Phylogenet Evol 48:628–645

Suh A, Smeds L, Ellegren H (2015) The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol 13:e1002224

Sullivan J, Swofford D, Naylor G (1999) The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol Biol Evol 16:1347

Susko E, Roger AJ (2007) On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol 24:2139–2150

Tarrío R, Rodríguez-Trelles F, Ayala FJ (2001) Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the drosophilidae. Mol Biol Evol 18:1464–1473

Telford MJ, Moroz LL, Halanych KM (2016) Evolution: a sisterly dispute. Nature 529:286–287

Thorley JL, Wilkinson M (1999) Testing the phylogenetic stability of early tetrapods. J Theor Biol 200:343–344

Townsend JP (2007) Profiling phylogenetic informativeness. Syst Biol 56:222–231

Van de Peer Y, Frickey T, Taylor JS, Meyer A (2002) Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene 295:205–211

Wang H-C, Susko E, Roger AJ (2011) Fast statistical tests for detecting heterotachy in protein evolution. Mol Biol Evol 28:2305–2315

Weigert A, Helm C, Meyer M, Nickel B, Arendt D, Hausdorf B, Santos SR, Halanych KM, Purschke G, Bleidorn C, Struck TH (2014) Illuminating the base of the annelid tree using transcriptomics. Mol Biol Evol 31:1391–1401

Whelan NV, Halanych KM (2016) Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst Biol 52:696–704

Whelan NV, Kocot KM, Moroz LL, Halanych KM (2015) Error, signal, and the placement of Ctenophora sister to all other animals. Proc Natl Acad Sci U S A 112:5773–5778

Whelan S, Blackburne BP, Spencer M (2011) Phylogenetic substitution models for detecting heterotachy during plastid evolution. Mol Biol Evol 28:449–458

White W, Hills S, Gaddam R, Holland B, Penny D (2007) Treeness triangles: visualizing the loss of phylogenetic signal. Mol Biol Evol 24:2029–2039

Wiens JJ (1998) Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst Biol 47:625–640

Wiens JJ (2003) Missing data, incomplete taxa, and phylogenetic accuracy. Syst Biol 52:528–538

Wiens JJ, Morrill MC (2011) Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst Biol 60:719–731

Wu J, Susko E (2011) A test for heterotachy using multiple pairs of sequences. Mol Biol Evol 28:1661–1673

Xia X (2013) DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30:1720–1728

Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phylogenet Evol 26:1–7

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372

Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51:588–598