To What Extent Current Limits of Phylogenomics Can Be Overcome?

Paul Simion, Frédéric Delsuc, Herve Philippe

Chapter 2.1 To What Extent Current Limits of Phylogenomics Can Be Overcome?

Paul Simion
Laboratoire d'Ecologie et Génétique Evolutive (LEGE), URBE
University of Namur, Namur, Belgium
polo.simion@gmail.com

Frédéric Delsuc[1]
Institut des Sciences de l'Evolution de Montpellier (ISEM), CNRS, IRD, EPHE
Université de Montpellier, Montpellier, France

Hervé Philippe
Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321
Moulis, 09200, France
herve.philippe@sete.cnrs.fr

Abstract

Current phylogenomic methods are still a long way from implementing a realistic genome evolution model. An ideal approach would require a general joint analysis of genomic sequences, while including coding sequence annotation, protein evolution or gene transfer, among other mechanisms, to infer the complete evolutionary history of the studied genomes. Such an approach is computationally intractable and currently approximated by phylogenomic pipelines that implement a series of independent steps ranging from gene annotation to species tree inference or positive selection detection. Here we review the virtues and limits of current phylogenomic methods compared to what could be expected from an ideal method. We present five case studies to illustrate various issues and limits in current phylogenomic practices, while assessing their relative importance. We argue that data error is pervasive in modern datasets and models are still too simplistic compared to the complexity of biological and evolutionary processes. Importantly, joint analyses should be a research focus as the many steps of phylogenomic pipelines are not mutually independent. It is essential to recognize the hidden assumptions of the many types of analysis available to our community so as to circumvent model misspecifications and critically evaluate the relevance of their results. In conclusion, the quality of datasets should be enhanced via numerous, rigorous checkpoints, while also boosting the capability of models to handle biological complexity by the development of better models, particularly through joint analyses.

How to cite: Paul Simion, Frédéric Delsuc, and Hervé Philippe (2020). To What Extent Current Limits of Phylogenomics Can Be Overcome?. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 2.1, pp. 2.1:1–2.1:34. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE.

Supplement Material `https://github.com/psimion/SuppData_Simion_Chapter_2020_Limitations_Phylogenomics`

[1] FD was funded by the European Research Council via the ERC-2015-CoG-683257 ConvergeAnt project.

Phylogenetics in the genomic era.
Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 2.1; pp. 2.1:1–2.1:34

A book completely handled by researchers.
No publisher has been paid.

1 Evolutionary history modeling and inference

Reconstructing genome evolution

The ultimate goal of phylogenomics is to reconstruct the evolutionary history of species through their genomes. In theory, this involves reconstructing the genomes of all organisms that ultimately supplied a DNA fragment to extant organisms and all events that generated modifications in the genetic material. Such details obviously cannot all be determined, but it is likely that several major patterns of great interest could be inferred, as illustrated in Figure 1. First, the evolutionary history of species generated through speciation and hybridization should have left a clear majority signal in genomes. Second, the signal left by horizontal gene transfers from distant or sister species (or so-called "gene flow") should be discordant from the majority signal. Third, the extent of incomplete lineage sorting would inform us about ancestral population sizes and times between successive speciation events. Fourth, mutations that became fixed because they provided a selective advantage could be differentiated from the bulk of neutral or slightly deleterious mutations (e.g. through an unexpected synonymous to non-synonymous substitutions ratio). Mutations can, for instance, involve single point changes, insertions of a few random nucleotides or a long stretch of nucleotides (from a transposable element or through illegitimate recombination), deletions of a few nucleotides or a long fragment (even complete chromosomes), duplications (of genomes or some chromosomes) or rearrangements (e.g. chromosome translocation, fission or fusion). The order of magnitude signals generated by each of these events may vary from a single point mutation to genome duplication. Hence, some could likely be finely characterized (e.g. the timing of a genome duplication) while it might only be possible to describe others statistically.

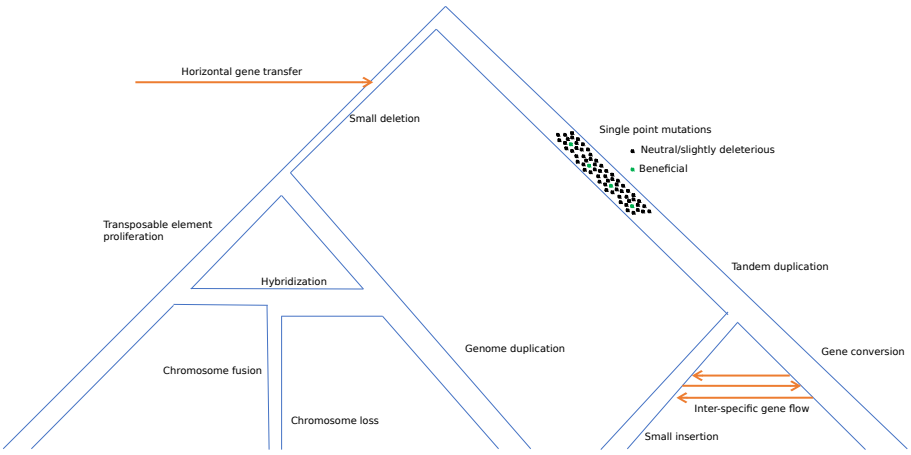


Figure 1 The historical processes shaping genome evolution. Schematic depiction of the main mutational processes that shape genomes. Their frequencies and impact on fitness are highly heterogeneous (e.g. from synonymous mutations to genome duplication or hybridization). These mutational processes are quite well known and relatively easy to model, whereas estimating the fitness of a given genome is much more difficult.

An ideal evolutionary model

The most natural way to infer this history from a series of genomes is to develop a genome evolution model and use standard statistical inference methods (in a Bayesian or maximum likelihood framework, see respectively Chapter 1.2 [Stamatakis and Kozlov 2020] and Chapter 1.4 [Lartillot 2020a]). The mechanisms depicted in Figure 1 have been the focus of massive in-depth studies for decades. Speciation is a pivotal theme in evolutionary biology, and DNA structure and change (including DNA repair) are crucial in molecular biology. All of this gives us an excellent idea of the most important mechanisms and how they work. So theoretically we have most of the knowledge required to develop a refined mechanistic model to reconstruct genome history. Naturally such a model can be designed in the mutation/selection framework (Bird, 1980), where the mutation process is independent of the DNA function, and a fitness function of the overall genome may be used to accept or reject a mutation. It is relatively easy to imagine how to model the mutational process inspired by simulators of genome evolution (Dalquen et al., 2012). For instance, single point mutations could be modelled with a general time reversible model (Tavare 1986; Chapter 1.1 [Pupko and Mayrose 2020]), and insertion/deletion with a hidden Markov model (Holmes and Bruno, 2001), while the mutational process would not necessarily be uniform across the genome (e.g. CpG hypermutability [Bird 1980] or proximity to the DNA minor groove [Pich et al. 2018]). Horizontal gene transfer should be considered as a mutation. Developing a fitness function is obviously much more difficult, but a function that only takes the major fitness components into account, i.e. non-coding RNAs and proteins, and their expression level, might be sufficient. A model similar to those used for gene annotation (see Chapter 4.1 [Necsulea 2020]) would enable prediction of non-coding RNA and protein sequences from genome sequences through the identification of transcription initiation sites and exon/intron structures. The fitness of these sequences could be estimated via a phenomenological approach (as in Yu and Thorne, 2006; Rodrigue et al., 2010). The expression level can be predicted based on promoter (nucleotide) and transcription factor (amino acid) sequences. Innovative solutions would certainly be required to be able to integrate all of these elementary fitness components into the fitness framework of a genome.

Ideal but beyond reach

Despite the attractiveness of such a theoretical model that could be used to infer major events which have occurred during genome evolution (see Figure 1), nobody has ever envisioned such a holistic approach. The reason may be that the approximations needed to compute genome fitness are so unrealistic that the extent of model violations would undoubtedly generate highly inconsistent results. But this is an unlikely explanation since, for instance, in phylogenetics the underlying maximum parsimony model, and to a lesser extent the Jukes-Cantor model are highly unrealistic, (e.g. based on the assumption that selective pressures are the same at every genome position). These models have nevertheless been and are still being frequently used. The most likely reason for not developing such a global genome evolution model is the tremendously high combinatorics. Sequence alignment and evolutionary tree inference independently constitute non-polynomial (NP) problems (see respectively Chapter 1.2 [Stamatakis and Kozlov 2020] and Chapter 2.2 [Ranwez and Chantret 2020]). As genomes are composed of millions of nucleotides, the number of possible ancestral genomes, evolutionary paths of organisms and DNA fragments is tremendous. Anyone who has ever tried to infer a phylogenetic tree from a relatively small dataset (e.g. 500 genes from 100 species) under the site-heterogeneous CAT-GTR model (PhyloBayes, Lartillot and

Philippe 2004) or a coalescent model (BEAST, Bouckaert et al. 2019) is aware of how far we are from making inferences with such a genome-scale model (see also Chapter 5.3 [Zhukova et al. 2020]).

Inferring the history of genomes therefore requires a divide and conquer approach combined with a clever choice of simplifying assumptions. The next section will roughly describe the main divisions that have been adopted by the phylogenomic research community.

2 The phylogenomic approach

Here we focus on species phylogeny inference using phylogenomics. We then exemplify the problems and advantages generated by the arbitrary division of a large-scale joint inference (see Figure 1) into several smaller elements (see Figure 2). Since less information (e.g. the evolutionary history of sequences is overlooked during alignment) is used at each step, errors may easily be made and their impact on subsequent steps is a concern. In contrast, working on a small-scale inference potentially allows us to use more complex models, hence reducing model violations.

2.1 A practical approximation

Divide and conquer

Essentially, reconstructing a given species phylogeny is a logical inference that uses both observable data (e.g. genomic sequences or morphological characters) and various premises (e.g. mutations are inherited through time, and transitions are more likely than transversions) to produce hypotheses regarding past evolutionary events. Unfortunately, the quantity and complexity of these premises currently hamper any practical holistic inference. With the aim of applying an approach capable of grasping the main aspects of the numerous evolutionary processes described in the previous section, while remaining practical, in phylogenomic analyses this integrated process is divided into independent blocks of computationally tractable units. These units form typical phylogenomic approaches, as illustrated in Figure 2, and they correspond to various common genomics procedures: (i) genome annotation, (ii) searching for homologous genes, (iii) defining orthologs, (iv) aligning homologous positions, (v) inferring species phylogeny (supermatrix or gene tree approaches), and (vi) reconciling single gene trees and the species phylogeny. Once the species tree is inferred, it is used in various methods to refine gene trees (duplication, loss, and horizontal gene transfer), infer the strength of selection, or reconstruct the gene order. Many of these procedures are detailed in other chapters of this book (see Chapters 1.4, 2.2, 2.4, 2.5, 3.2 and 4.5 [Lartillot 2020a; Ranwez and Chantret 2020; Fernández et al. 2020; Tannier et al. 2020; Boussau and Scornavacca 2020; Lowe and Rodrigue 2020]). The following description of conventional phylogenomic approaches is voluntarily brief and cursory and interested readers may find additional in-depth reviews on phylogenomics elsewhere (Delsuc et al., 2005; Philippe et al., 2005; Laumer, 2018).

Inferring orthologs

Phylogenomic pipelines are based on genomic data (e.g. often using coding sequences [CDSs] or ultra-conserved elements [UCEs]) and on transcriptomic data from multiple species. The choice of the molecular markers to be used is dependent on the biological question at hand, such as the evolutionary scale under investigation. These molecular datasets from multiple samples are then clustered into groups of homologous sequences. The criteria used

for homology prediction is pairwise sequence similarity estimated using BLAST (Altschul et al., 1990) or Smith–Waterman (Smith and Waterman, 1981) classical pairwise sequence alignment algorithms. Several algorithms have been proposed to make use of this pairwise similarity information to explicitly produce groups of homologous sequences, such as MCL (Enright et al., 2002), hcluster sg (Ruan et al., 2008) or SiLiX (Miele et al., 2011). Other tools use this pairwise similarity information to directly predict orthology relationships, such as Hieranoid (Kaduk and Sonnhammer, 2017), OrthoMCL (Li et al., 2003), OrthoFinder (Emms and Kelly, 2015), OMA (Altenhoff et al., 2019), Eggnog (Huerta-Cepas et al., 2019), UPhO (Ballesteros and Hormiga, 2016), Ortholog-Finder (Horiike et al., 2016) among others (see Chapter 2.4 [Fernández et al. 2020]), with variable complexity ranging from fairly straightforward (OrthoMCL) to complex procedures based on successive clustering loops and sequence alignments followed by gene tree inference and paralog splitting (Ortholog-Finder). Orthology prediction tools dovetail in many ways with the overall phylogenomic approach. They often rely on a variety of steps that include homology inference (i.e. similarity searches), pairwise species comparisons or species-overlap concepts, sequence alignment, gene genealogy inference, and even species tree inference or *a priori* knowledge of the species tree. Orthology prediction is thus often considered as a separate phylogenomic approach, and interested readers will find a more in-depth review of this topic in Chapter 2.4 (Fernández et al., 2020).

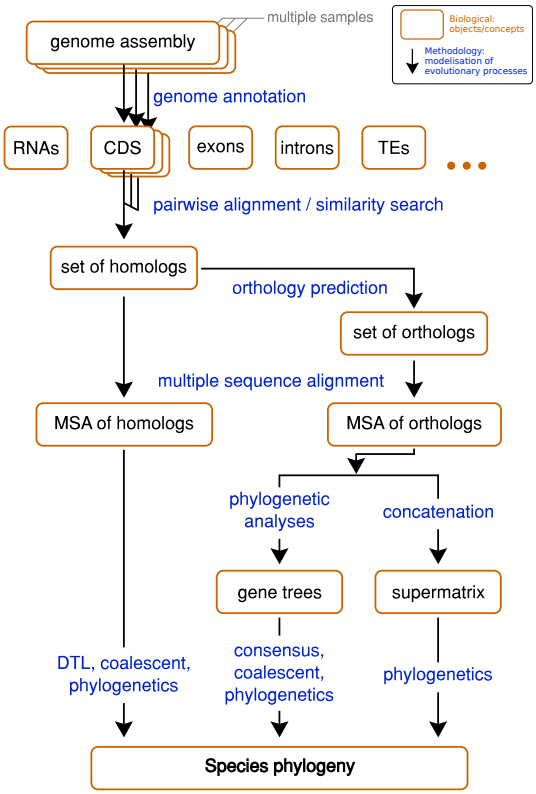


Figure 2 Common phylogenomic approaches. Schematic view of the series of practical analysis steps (in blue) of the phylogenomic approach.

Producing alignments

Multiple homologous or orthologous sequences are then jointly aligned using one of the many available sequence alignment software packages, such as MAFFT (Katoh and Standley, 2013), Clustal Omega (Sievers et al., 2011), MUSCLE (Edgar, 2004) or T-COFFEE (Notredame et al., 2000) (see also Chapter 2.2 [Ranwez and Chantret 2020]). Multiple sequence alignment software optimizes sequence alignment using various complex algorithms, but usually relative to a simplistic sequence evolution model (e.g. all substitutions and replacements are considered equiprobable) and the evolutionary history of the sequence is overlooked. Interestingly, one tool, i.e. MACSE v2 (Ranwez et al., 2018), models the codon structure of coding sequences during alignment. Using a more complex model unsurprisingly leads to better results and to some additional features, such as the ability to circumvent frameshifts that are often present in sequencing data (Ranwez et al. 2018; Chapter 2.3 [Ranwez and Delsuc 2020]).

Phylogenomic analyses

When using MSA of homologous sequences, directly concatenating them is impossible since multiple paralogous/xenologous sequences per species could be present. Instead, reconciliation methods are used to jointly analyse gene trees and species trees, notably by modeling gene duplication-transfer-loss (DTL) events. Available software includes Phyldog (Boussau et al., 2013) or ecceTERA (Jacox et al., 2016), among others (see Chapter 3.2 [Boussau and Scornavacca 2020]). Such analyses are complex but promising for the future of phylogenomics as they acknowledge the actual interdependency of two steps (i.e. jointly inferring gene and species trees), which are handled separately in the standard phylogenomic approach using orthologs only (see Figure 2). In addition, much more data can be used with these analyses than is possible with the reduced set of orthologs. Unfortunately, these methods are still very computationally-expensive to be widely used in a ML framework (Phyldog), although parsimony-based amalgamation methods such as ecceTERA could scale up with genomic data (ecceTERA).

Two main strategies are available to infer a species tree when using multiple sequence alignments (MSA) of orthologous sequences. Every alignment may be analysed independently to produce gene trees that may be incongruent because of incomplete lineage sorting (ILS), introgression or lateral gene transfer. This information may then be used to infer the underlying species tree, or otherwise every sequence per species may be concatenated in order to sum up their phylogenetic signals. There is ongoing debate on which strategy recovers the most accurate species trees (Springer and Gatesy, 2016; Edwards et al., 2016) and it is important to highlight three key arguments in this debate. First, taking ILS into account (see Chapter 3.4 [Bryant and Hahn 2020]) is impossible when using a concatenation approach which, despite the current use of more refined evolution models and more data, could never accurately solve a series of extremely fast speciation events given that it can be inconsistent under some evolutionary scenarios (Kubatko and Degnan, 2007). Second, single-gene tree reconstruction often yields little or no phylogenetic signal for difficult nodes (e.g. short internal branches) due to stochastic error. Third, only considering the species tree is not appropriate for subsequent evolutionary analyses (Hahn and Nakhleh, 2016). We believe that concatenation seems therefore more adequate to resolve ancient phylogenetic relationships or when the sampling is devoid of ultra-close speciation events, whereas the use of single gene trees is more appropriate for more recent speciation events, even when closely-spaced in time. Both methods rely on phylogenetic tree inference generally using

software based on ML, such as PHYML (Guindon et al., 2010), IQ-TREE (Nguyen et al., 2015), RAxML-NG (Kozlov et al. 2019; Chapter 1.3 [Kozlov and Stamatakis 2020]), or on Bayesian approaches such as MrBayes (Ronquist et al., 2012), BEAST (Bouckaert et al., 2019), and PhyloBayes (Lartillot et al. 2009; Chapter 1.5 [Lartillot 2020b]).

2.2 The costs of over-simplification and subdivision

While some model violations are often discussed concerning the phylogenetic inference step (e.g. ongoing debate on the development of the best sequence evolution models or on concatenation versus coalescence approaches), many other steps in the phylogenomic approach could also potentially lead to erroneous results. Below we discuss some of the problems encountered during the various practical steps in the phylogenomic approach when the ideal evolutionary model presented earlier is misspecified.

Information loss and implicit model violation

The subdivision of an ideal integrated model for reconstructing the evolutionary history of genomes into a series of independent blocks of computationally tractable units necessarily leads to the loss of potentially useful information, while forcing us to adopt an over-simplified model that cannot use the missing information. For instance, homology search through sequence similarity ignores the overall evolutionary history of the genomes being compared. Due to the loss of phylogenetic information, it implicitly makes the strong yet incorrect assumption that sequences were generated under a star-tree topology with equal branch lengths. However, the information that some species are closely related and that some others are fast-evolving is extremely useful for homology detection. The impacts of this model violation on the outcome are extremely hard to predict and study because of the substantial challenge of designing alternative non-star-tree models. Moreover, the model used to quantify similarity is extremely simplistic as it is solely based on an amino acid exchangeability matrix (e.g. JTT, or BLOSUM). It implicitly assumes that every position evolves at the same rate and that at most a single substitution has occurred at a given position, which are two obviously incorrect assumptions. This oversimplified model explains the poor sensitivity of the BLAST score (Koski and Golding, 2001). Interestingly, alongside the publication of the orthology inference tool Orthofinder (Emms and Kelly, 2015), the authors designed a blast score double-normalization. It normalises BLAST scores for alignment length and, more importantly, these pairwise scores are normalised across species according to their evolutionary distances, so it is striving to transform the scores as if the sequences had been generated under a star-tree topology. This interesting approach nevertheless cannot control saturation of the similarity score, which means that the correction will be much more accurate for closely related species than for divergent ones.

Genome annotation errors

As briefly introduced in Figure 2, genome annotation is one of the first steps of most phylogenomic pipelines. Annotating genes requires a set of complex methods that rely on knowledge regarding genetic code, intron structure, transcription and translation mechanisms or RNA-seq data (see Chapter 4.1 [Necsulea 2020]). Yet, it often assumes that genomes do not have any evolutionary history, again an obviously false assumption. A shortcut to input some evolutionary information is to compare predicted coding sequences with transcriptomes or proteomes from closely related species (Dunne and Kelly, 2017; Monnahan et al., 2019; Rey

et al., 2019). Unfortunately, current annotation methods do not model chromosome structure, protein folding or interaction with other genomic regions. These limitations lead to erroneous gene predictions that can ultimately mislead comparative genomic analyses (see examples in Section 3.1).

Sequence alignment model violations

Multiple sequence alignments are also hampered by model violations (see Chapter 2.2 [Ranwez and Chantret 2020]), i.e. some mutational processes are explicitly modelled while overlooking sequence function and protein structure as well as their chromosome-wise context, such as species-specific recombination hotspots or even lineage-specific evolutionary rates (i.e. heterotachy). Otherwise, when aligning multiple sequences, indels are implicitly considered as characters rather than historical events. The latter is a misspecification of an ideal evolutionary model, which is tackled by the dynamic homology concept. This issue has led various authors to develop methods for joint inference of sequence alignments and species trees (Fleissner et al., 2005; Redelings and Suchard, 2005; Herman et al., 2014; Wheeler et al., 2015). As expected, this interesting approach is computationally intensive, thus seriously limiting the dataset size and the complexity of the sequence evolution model that can be handled.

Unrealistic phylogenomic inference models

In contrast, phylogenomic analyses of aligned and concatenated sequences enable the use of more complex evolutionary models geared towards minimizing model violations. However, some potentially important aspects of genome evolution are still not taken into account by most phylogenomic inference methods, e.g. lineage-specific composition heterogeneity, site-specific substitution process heterogeneity, or heterogeneity of site-specific substitution process among lineages (i.e. heteropecilly, Roure and Philippe 2011). Note that even when some methods are available to model one aspect of genome evolutionary processes, e.g. modeling ILS, site-heterogeneity or DTL in reconciliation methods, it is seldom feasible to combine them, and if it were, the resulting implementation would surely be extremely time-consuming. For example, combining a CAT model with a GTR component, a Gamma component, amino acid compositional breakpoints along the tree and evolutionary rate breakpoints along the tree while allowing for gene transfer across lineages to analyse relationships between 300 complete genomes would clearly be beyond reach with current computation resources. Finally, knowledge on the genomic context of a sequence is still not used in the phylogenetic inference process.

Software errors

In addition to these errors — for which we know the origin albeit we do not know where they are in the dataset — there are unknown errors, i.e. errors in the implementation such that the script/software does not produce the intended results. These unknown errors are expected because limited funding and publish-or-perish pressure imply that an insufficient amount of time is generally devoted to quality control of both programs (Czech et al., 2017; Darriba et al., 2018) and pipelines (see Section 4.5).

All of the severe model violations described above, albeit unavoidable for computational tractability reasons, as well as information loss very likely generate errors at each phylogenomic pipeline step. Importantly, of all these errors accumulate along the pipelines, with a possible snowball effect. For instance, annotation errors alone will generate additional errors

in homology detection, which in turn will generate more errors in the alignment and finally in phylogenetic inference. Hereafter we briefly discuss the robustness of phylogenomics to these errors and how to reduce their impacts.

3 Relative robustness to pervasive errors

3.1 Types of error and methods to detect them

Theoretical limitations of the successive independent and simplistic steps of the phylogenomic approach inevitably lead to the production of errors. These can appear at all steps of a given pipeline and can propagate from step to step. Here we briefly describe various error types and some recent methods or tools that can detect them and thus reduce their impact on the phylogenomic approach as a whole. We have classified these errors into three arbitrary groups: i) observational errors during data acquisition and production, ii) errors during dataset assembly, and iii) errors during phylogenetic inference. These errors can be generated by experimental error (e.g. contamination by DNA from other species), stochastic error (e.g. insufficient coverage), and systematic error (i.e. due to model violations).

Observational errors

This type of error concerns data that are not what the user believes they are. These include contamination from organisms other than the target (e.g. bacteria, fungi, trypanosomes, viruses), cross-contamination between samples during sequencing data production, sequencing and assembly errors, fragmented transcriptomic contigs thought to be entire transcripts, gene exons thought to correspond to entire genes, gene introns thought to correspond to exons, amino acid sequences translated out of frame (i.e. frameshifts). Contamination in genomic data can partially be detected by Blobtools by combining coverage, GC content, and blast taxonomy (Laetsch and Blaxter, 2017), large scale similarity search with Conterminator (Steinegger and Salzberg, 2020) or by the consensus of various methods (Cornet et al., 2018). Contamination is not only present in the data generated during a given study, but also affects public databases: e.g. 5% of the publicly available cyanobacterial genomes turned out to be highly contaminated (Cornet et al., 2018) and a recent large-scale analysis of GenBank identified more than 2,000,000 contaminated sequences! (Steinegger and Salzberg, 2020). Cross-contamination affects both DNA and RNA data and is increasingly acknowledged as a pervasive issue (Ballenghien et al., 2017; Alié et al., 2018; Allio et al., 2020; Prous et al., 2020). It can be handled by the CroCo program which relies on coverage to detect the actual origin of a sequence in a set of samples (Simion et al., 2018). It has been shown that up to 30% of transcripts from a *de novo* assembled transcriptome could be cross contaminated (i.e. actually belong to another species, Simion et al. 2018) or up to 26% of ddRAD loci (Prous et al., 2020).

Assembly errors during *de novo* transcriptome assembly (e.g. fragmentation due to insufficient coverage) can be corrected by fusing non-overlapping fragmented transcripts based on a multi-species orthology context, as shown in Section 4.1, where 30.6% of the transcripts were fragmented (124,096 out of 405,055 transcripts analysed). Annotation errors are also pervasive, as illustrated by the fact that reannotation of the well-known model group *Drosophila* recently led to the discovery of 500 to 1,000 new genes per species (Yang et al., 2018). Recent studies have proposed tools to correct gene annotation based on comparisons with other species (Dunne and Kelly, 2017; Rey et al., 2019; Monnahan et al., 2019). Using a non-overlapping sequence criteria on gene annotation from tunicate genomes (see Section

4.1), we estimate that 5.6% of the predicted genes were split into, usually, two exons (3,178 out of the 56,694 genes analysed). A next step towards more holistic genome annotation approaches could involve joint annotation of several genomes at once, with tractable computations if the species tree is known. Current tools using such a comparative framework unfortunately only aim at correcting gene prediction *a posteriori* (Dunne and Kelly, 2017; Monnahan et al., 2019) or at improving transcriptomic assemblies that could then be used to help gene prediction (Rey et al., 2019). Finally, frameshifts (due to sequencing, assembly or annotation errors) produce very divergent sequence stretches. These can be corrected during sequence alignment by taking the codon structure of coding sequences into account using the MACSE v2 alignment tool (Ranwez et al., 2018). If already present, frameshifts can be detected and masked in a multi-sample alignment using two recent segment filtering tools, i.e. HMMcleaner (Di Franco et al., 2019) and Prequal (Whelan et al., 2018).

Orthology errors

Similarity searches, usually via BLAST (Altschul et al., 1990), are hampered by bias, where higher scores are given to long dissimilar alignments than to short highly similar ones. This bias, combined with the lack of precision of the method in detecting very dissimilar sequences, means that a poorly assembled or highly divergent transcriptome will likely yield poor homology results. Pairs of homologous sequences have different potential relationships: they can be orthologous (i.e. stemming from speciation), paralogous (i.e. stem from duplication) or xenologous (i.e. stemming from horizontal transfer) — see Chapter 2.4 (Fernández et al. 2020) for an in-depth review of this topic. When using orthology inference algorithms, sequence pairs can be erroneously categorised as a subtype due to several issues. For example, incomplete taxonomic sampling complicates the detection of xenologs Kuzniar et al. (2008), and inaccurate gene tree inference can hampers the classification of two sequences as orthologous (e.g. if only one of them evolved rapidly). While ongoing research is focused on improving orthology inference tools, it is also crucial to design a taxonomic sampling method tailored to the evolutionary scale at hand to ensure accurate inference of sequence relationships. Orthology sets can be further improved by *a posteriori* checking orthology set consistency (Simion et al., 2017). Lastly, several tools can eventually be used to analyse sequence alignments and gene trees in order to detect and limit the impact of orthology errors on the phylogenomic pipeline, e.g. Phylo-MCOA (de Vienne et al., 2012), treespex (Struck, 2014), Branch Length Correlation (BLC) methods (as in Simion et al. 2017), Treeshrink (Mai and Mirarab, 2018) or reconciliation methods (Dondi et al., 2016). See a comparison of some of these methods in Section 4.2.

Evaluating phylogenomic datasets

It should be noticed that phylogenomic pipelines could be highly diversified in terms of implementation, thus leading to highly diversified phylogenomic datasets. Unfortunately, these datasets are seldom compared with statistics other than simply the numbers of genes and species included. Looking beyond summary statistics in these phylogenomic datasets can reveal various levels of data quality (e.g. measured with the Robinson-Foulds distance between gene trees and species trees), with orders of magnitude difference in data quantity (see Figure 2 in Simion et al. 2017, and Figure S4D in Philippe et al. 2019). For instance, regarding the debated phylogenetic position of ctenophores, less than 30% of the gene tree bipartitions are congruent with the species tree in three datasets that support the ctenophora-first hypothesis (Dunn et al., 2008; Hejnol Andreas et al., 2009; Moroz et al., 2014), whereas more

than 60% are congruent in a dataset that supports the porifera-first hypothesis of Simion et al. (2017). Data quality governs the crucial phylogenetic signal-to-noise ratio upon which the accuracy of the inferred species tree strongly depends. When working on debated species phylogeny, we stress the need to carefully inspect the phylogenomic pipelines used and the respective virtues of the datasets they led to, as their signal-to-noise ratio might be pivotal in evaluating the reliability of a given phylogenomic result. See also Chapter 2.5 (Tannier et al. 2020) for an original approach to gene tree quality assessment based on ancestral genome reconstruction.

Phylogenetic inference errors

Many sequence evolution models are available for phylogenetic inference based on homologous sequence alignments. Early models were tailored for single-gene analyses but the datasets have increased in both their dimensions (i.e. more markers and taxa) and complexity. Model assumptions are now recognised as often being violated by complex datasets, thus prompting the need to also increase the model complexity. For example, introduction of the Gamma component in models discredited the assumption that all sites evolve at the same pace (Yang, 1994), and more recent site-heterogeneous CAT models refuted the assumption that all sites evolve under the same substitution process (Lartillot and Philippe 2004 and Chapter 1.4 [Lartillot 2020a]). Potential model misspecifications are still numerous and could result in erroneous topology inference (e.g. LBA artefact, compositional bias) and incorrect branch length estimation. Various methods and models have been developed to reduce these misspecifications, such as GHOST models implemented in IQ-TREE in order to model heterotachy (Crotty et al., 2019), CAT models to phenomenologically account for protein structure and function (Lartillot and Philippe, 2004), the PMSF approach recently implemented in the ML framework (Wang et al., 2018), compositional breakpoints (BP) to account for heterogeneity in the substitution process across lineages (Blanquart and Lartillot, 2006, 2008) or site-heterogeneous codon models (SelAC) to model stabilising selection (Beaulieu et al., 2019). Data recoding has a special role in current phylogenomics. It consists in grouping different character states into a single common character leading to alphabet reduction (e.g. the Dayhoff 6-state recoding scheme). It is used in order to reduce compositional bias and saturation in the data, thus enhancing the phylogenetic signal (Susko and Roger, 2007). The relative importance of signal loss in comparison with the reduction in compositional bias and saturation is still debatable and likely depends on the characteristics of the dataset under study. In our opinion, data recoding should be considered suitable for large supermatrices only and recoded datasets are still highly complex so that they still require to be analysed with complex models (e.g. CAT models, see Feuda et al. 2017). Finally, a recent study assessed the impact of modelling site-heterogeneity versus partition-wide heterotachy and convincingly concluded that modelling site-heterogeneity was more important than modelling partition-wide heterotachy (Wang et al., 2019).

So far we have discussed ways to limit errors stemming from systematic bias, but when the phylogenetic signal is weak, stochastic errors can occur even when using large phylogenomic datasets. This is particularly true for ancient and short internal branches. In fact, with such difficult relationships, the data quantity required is so high that even a large sampling of complete genomes would not be enough to resolve them (Philippe et al., 1994).

3.2 Consistent species phylogenies

A large quantity of signal

The phylogenomic approach is, despite its flaws, surprisingly robust, as most pipelines will lead to the recovery of a similar species tree topology. This can be explained by the sheer quantity of phylogenetic signal accumulated when thousands of molecular markers are combined. This is not surprising, as many parts of the tree of life have already been correctly inferred using comparatively small morphological character matrices or single gene phylogenies. Phylogenetic signal is additive, so the amount of signal increases with the data quantity. In fact, only additive errors can compete with phylogenetic signal by producing a non-phylogenetic signal, leading to the recovery of an erroneous tree. For an error to be additive it has to produce the same kind of bias repeatedly across markers and lineages. For example, genomic cross-contamination from sample A to sample B will repeat the same mislabelling of B sequences, and species B will eventually be attracted towards the phylogenetic position of species A (Laurin-Lemay et al., 2012). As another example, in a lineage that has a naturally high evolutionary rate, on average all markers will present more homoplasy with another fast-evolving lineage, and both long-branch lineages will attract each other (i.e. LBA artifact). Conversely, various randomly distributed errors will only produce non-additive signals (often called "noise') that will not severely distort the phylogenetic signal. Even if not correctly modelled, these errors will simply reduce the statistical power of phylogenomics (see Section 4.3). Overall, the phylogenomic approach produces a globally consistent species tree as long as the phylogenetic signal prevails over the systematic error.

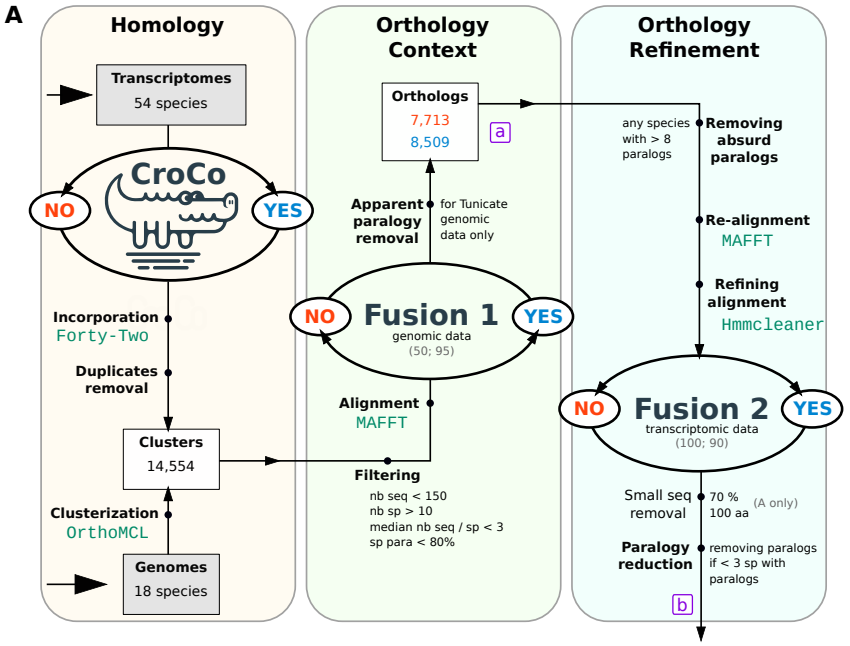
Few very difficult cases

Phylogenomic inconsistency only occurs in a few cases across the tree of life, all of which share the same characteristics and correspond to short internal branches. These branches bear a limited amount of phylogenetic signal so they are highly susceptible to errors, even random errors (e.g. see Section 4.3). Indeed, the signal-to-noise ratio for these branches is so low that any perturbation or noise will hamper signal extraction regardless of the intrinsic qualities of the model used. Branches are short when diversification has occurred rapidly through time, and the problem becomes more complex when the speciation event was too ancient, with progressive loss of the historical signal through multiple substitutions (i.e. saturation). Difficult relationships stemming from the first case triggered ILS modelling research (see Chapter 3.3 [Rannala et al. 2020]), while relationships derived from the second case underscore the need for better models to optimize the efficiency of extraction of the scant amount of remaining historical signal (see above and Chapter 1.4 [Lartillot 2020a]). As the phylogenomic approach is largely consistent across species phylogenies, except for short internal branches with low signal-to-noise ratio, it is not surprising that long-standing phylogenomic disputes are finally now focused on a few difficult relationships, i.e. the phylogenetic position of ctenophores, xenacoelomorphs, Stauromedusae, Laurasiatheria, the root of the placental tree or the early evolution of birds and eukaryotes.

4 Case-studies: Examples of current limits of phylogenomics

4.1 Correcting data errors in tunicates

Current phylogenomic practices involve similar handling of genomic and transcriptomic data from orthology inference to final sequence alignment, which violates the hypothetical ideal



B

	Orthologs at checkpoint a	Orthologs at checkpoint b	Min species recquired	Final Orthologs	congruence gene tree/ species tree	Max Missing% per site	Supermatrix length	Global Missing%	Total known residues
Standard Pipeline	**7,713**	**2,416**	5	2,024	mean = **63.3 %**	100	156,373	65.3	3,852,366
			10	1,770		80	95,037	46.9	3,582,646
			20	801		**50**	**54,878**	**34.9**	**2,536,964**
			30	**408** →		20	3,139	17.1	184,735
Corrected Pipeline	**8,509**	**6,732**	5	5,341	mean = **67.2 %**	100	1,860,030	57.5	56,067,810
			10	5,161		80	1,267,890	40.8	53,330,401
			20	3,993		**50**	**849,403**	**29.3**	**42,614,055**
			30	**3,039** →		20	210,163	16.0	12,534,696

Figure 3 Impact of data correction on tunicate phylogenomics. A) Detailed procedure for both standard and corrected pipelines. Three steps were only used in the "corrected" procedure: CroCo, Fusion 1 and Fusion 2. None of these steps were used in the "standard" pipeline. B) Summary statistics of the datasets produced from the two pipelines. The left side of the table corresponds to statistics obtained from the final orthologs highlighted in bold (i.e. 408 and 3,039). Conventional criteria for missing data filtering in supermatrices are highlighted in bold for both pipelines to ease comparison for readers.

holistic evolution model described at the beginning of this chapter. Indeed, a mix of genomic and transcriptomic data is often used in phylogenomic studies as transcriptomic data sequencing and assembly is both cheaper and faster than whole genomes. Transcriptomics is a cost-efficient way to enrich taxonomic sampling, which in turn helps to infer more accurate trees. Both types of sequencing data are potentially subject to contamination and cross-contamination. However, genomic and transcriptomic data differ in their nature. One the one hand, genomic data quality relies on accurate gene annotation, genes can occur in multiple copies and it might be expected that a genomic gene set is exhaustive for a given organism. On the other hand transcriptomic data quality mostly relies on the transcriptome assembly accuracy, whereas alternative splicing is often overlooked in favour of keeping the longest transcript, and transcriptomes are not exhaustive as rare transcripts are likely to be missed if the sequencing coverage is not deep enough. Regardless, both types of data are usually handled the same way. Data errors must be properly modelled in order to reduce discrepancies between data acquired by current protocols and real biological data. Here,

we simply tried to *a posteriori* correct some misspecifications, namely cross-contamination, split gene annotation and fragmentation of transcriptomic assemblies. We assessed the importance of these misspecifications by comparing datasets built with two different pipelines on the same mix of tunicate genomic and transcriptomic assemblies.

Both pipelines start with a set of putative orthogroups created with 18 genomes using OrthoMCL and then follow a series of filters and alignments until a final set of orthogroups is reached. The first pipeline is straightforward as it does not try to correct potential errors impacting genomic and transcriptomic data (no CroCo, no Fusion 1 and no fusion 2, see the *standard* pipeline in Figure 3A). The general procedure is designed as follows: transcriptomic data are incorporated into clusters of homologous sequences based on genomic data using Forty-Two (see https://bitbucket.org/dbaurain/42) and only alignments with reasonable size and diversity are kept. We then check tunicate species (only those based on genomic data) and remove all alignments in which at least one species presents two sequences (i.e. apparent paralogy). For each alignment, we then remove all sequences from species (based on transcriptomic data) that have too many paralogous sequences (i.e. more than eight copies per species). Short sequences that span less than 100 amino acids and have more than 70% missing data in a given alignment are discarded. Lastly, if less than three species present paralogs for a given alignment, their sequences are removed. The second pipeline adds three steps to the previous one (CroCo, Fusion 1 and Fusion 2, see the three blue "YES" of the *corrected* pipeline in Figure 3A). First, transcriptome assemblies from 54 species are cleaned from cross-contaminations using CroCo (Simion et al., 2018). Second, gene annotations of tunicate species (only those based on genomic data) are refined in the comparative context of an alignment by fusing together several sequences from the same species when they do not overlap. This is the "Fusion 1" step. Third, tunicate fragmented transcripts (based on transcriptomic data) are improved by also fusing same-species non-overlapping sequences in the alignments. This last "Fusion 2" step was only possible after an orthology context was reached by filtering out alignments containing paralogy for tunicate genomic data thus ensuring that we did not erroneous fuse non-overlaping paralogs.

The simultaneous use of the three corrections described above combined with paralogy filters led to a dramatic increase of roughly one order of magnitude in the size of the assembled dataset (see Figure 3B). The corrections improved the quantitative metrics measured here, i.e. gene number (408 to 3,039) and missing data (34.9 to 29.3%), thus increasing the number of known residues by 17-fold (2.5 M to 42.6 M). Gene alignments and supermatrices are available on the following website, containing the Supplementary Material of the current article: `https://github.com/psimion/SuppData_Simion_Chapter_2020_Limitations_Phylogenomics` . A recent study also reported a marked improvement in their dataset after cross-contamination removal with CroCo, with a gene number increase of 2,993 to 6,621 (Allio et al., 2020). This improvement was associated with an impact on the species tree topology and branch lengths, as expected given the findings of previous studies on the impact of the presence of cross-contamination in phylogenomics (Laurin-Lemay et al., 2012; Simion et al., 2018)). Importantly, these quantitative improvements were observed hand in hand with a qualitative improvement of the phylogenomic dataset. Indeed, the congruence between gene trees and species tree increased from 63.3% to 67.2% when using our three correcting steps (Figure 3B).

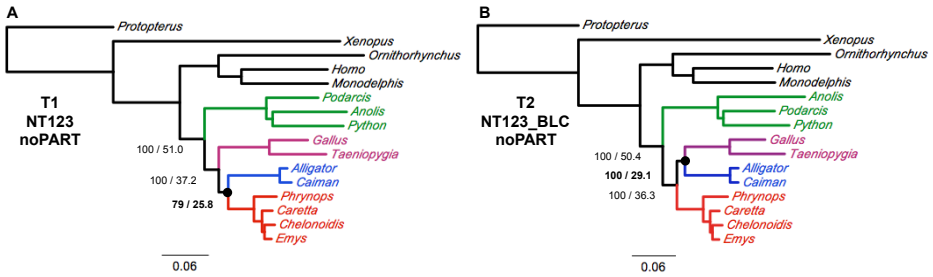
Why do cross-contamination removal and non-overlapping sequence fusion lead to a dataset that is an order of magnitude larger and of better quality? The answer lies in the existing interdependency between the different steps of the phylogenomic approach. We handled each step independently, but they share many assumptions that underlie an ideal genome evolu-

tion model, which means that a misspecification in one step will impact the next one. First, orthology inference tools are based on the assumption that sequences are correctly associated with an organism (i.e. no cross-contamination). Second, the accuracy of filtering clusters corresponding to 1-to-1 orthologs depends on the extent to which the genes are complete (i.e. correct gene annotation). In our example, gene annotation was *a posteriori* improved by being considered in the comparative context of a multi-species alignment. Third, high transcriptomic assembly quality (i.e. no fragmented transcripts) is essential when considering contigs from a transcriptomic assembly as biological transcripts. By carrying out simple *a posteriori* corrections to some known issues in the practical phylogenomic approach, we were able to better take into account some evolutionary and experimental processes that produced the genomic data under study. Orthology inference methods are more accurate if contaminants are removed and the number of apparent paralogs is reduced if split gene annotation and transcripts are fused back together. These simple corrections ultimately led to a vastly larger phylogenomic dataset.

Although useful, our simple corrections are still insufficient to build a truly genome-scale dataset. Indeed, a large percentage of the genes in the genomes and transcriptomes are still not present in the supermatrices. This might be due to incomplete sequencing and assembly of genomes and transcriptomes, to the limits of the orthology assignment method used (e.g. missing small and fast evolving genes) and/or to incomplete taxonomic sampling, which hampers accurate reconstruction of complex evolutionary histories (e.g. transfers, duplications, losses).

4.2 Cleaning outlier sequences and genes in turtle phylogenomics

The phylogenetic position of turtles within amniotes offers a great example of a longstanding question that has finally been answered through the resolving power of the phylogenomic approach. In 2012, two phylogenomic studies based respectively on transcriptomes (Chiari et al., 2012) and ultra-conserved DNA elements (Crawford Nicholas G. et al., 2012) independently found convincing support for positioning turtles as a sister group of archosaurs (birds and crocodiles), to the exclusion of lepidosaurs (lizards and snakes). This more derived position of turtles, recently confirmed by a larger scale phylogenomic analysis (Irisarri et al., 2017), implies that the anapsid condition of turtles (no temporal fenestration) is a derived state, whereas it was classically interpreted as the ancestral condition for amniotes. Chiari et al. (2012) built a phylogenomic dataset of 248 single copy nuclear genes for 16 vertebrate taxa, which was assembled according to best reciprocal hits obtained with BLAST based on the genomes and orthology annotations available at the time, along with newly generated transcriptomes. They showed that ML and Bayesian concatenations, and gene trees/species tree approaches performed under the best fitting nucleotide and amino acid substitution models unambiguously supported the classification of turtles as a sister group to birds and crocodiles (T2 in Figure 4B). However, the use of more simplistic nucleotide substitution models for both concatenation and gene trees/species tree reconstruction methods led to an alternative topology by artifactually grouping turtles and crocodiles (T1 in Figure 4A), likely because of third codon position saturation. This 248-gene dataset has since become an exemplary dataset in several studies aimed at testing phylogenetic reconstruction methods by comparing concatenation versus gene trees/species tree approaches (Bayzid et al., 2014; Mirarab et al., 2014, 2016; Simmons et al., 2016, 2019; Gatesy et al., 2019). Moreover, two recent studies have used this dataset as a core example to illustrate methods to detect outlier genes in phylogenomic analyses based respectively on Bayes factors (Brown and Thomson, 2017) and gene-wise likelihoods (Walker et al., 2018) between alternative



C

	AA	AA_BT	NT12	NT12_BT	NT12_TS	NT12_BLC	NT123	NT123_BT	NT123_TS	NT123_BLC
Genes	248	246	248	246	248	237	248	246	248	237
Sites	62,342	62,011	124,684	124,022	124,626	120,054	187,026	186,033	186,939	180,081
Variable sites	21,172 (33.96%)	21,034 (33.92%)	32,556 (26.11%)	32,352 (26.09%)	36,167 (29.02%)	36,282 (30.22%)	87,388 (46.73%)	86,868 (46.69%)	90,528 (48.43%)	89,311 (49.60%)
Missing	35.25%	35.26%	35.25%	35.26%	37.35%	35.81%	35.25%	35.26%	37.35%	35.81%
noPART BP (T1)	-	-	-	-	-	-	79	-	73	-
noPART BP (T2)	100	100	100	100	100	100	-	97	-	100
noPART CF_{gene} (T1)	-	-	-	-	-	-	25.8	-	28.3	-
noPART CF_{gene} (T2)	28.6	28.5	26.6	26.0	25.8	27.4	-	28.0	-	29.1
PART BP (T1)	-	-	-	-	-	-	-	-	-	-
PART BP (T2)	100	100	100	100	100	100	78	100	67	100
PART CF_{gene} (T1)	-	-	-	-	-	-	-	-	-	-
PART CF_{gene} (T2)	30.6	29.3	26.2	26.8	25.8	27.8	29.4	29.3	29.9	31.6

Figure 4 Assessing the effects of three filtering methods and sequence evolution models on the phylogenetic position of turtles. A) Maximum likelihood topology (T1) obtained with IQ-TREE under a single concatenated GTR+G model on the original nucleotide dataset of Chiari et al. (2012), including all codon positions. B) Maximum likelihood topology (T2) obtained with IQ-TREE under a single concatenated GTR+G model on the nucleotide dataset, including all codon positions filtered using the branch length correlation (BLC) method of Simion et al. (2017). Numbers at nodes indicate the standard ML bootstrap percentage/gene concordance factor, respectively. Bullets indicate nodes for which support values are reported in the table. Scale is based on the mean number of substitutions per site. C) Summary statistics and support values for the alternative T1 and T2 topologies obtained with different datasets (AA, NT12, and NT123) resulting from three following filtering methods: BT, TS, and BLC. The datasets were analysed under a single concatenated model (noPART) and a partitioned model by gene (PART). Abbreviations: BT: removal of the two paralogous genes identified by Brown and Thomson (2017); TS: TreeShrink filtering with default parameters (Mai and Mirarab, 2018); BLC: filtering with the branch length correlation method of Simion et al. (2017); AA: amino acid dataset; NT12: nucleotide dataset with saturated third codon positions removed; NT123: nucleotide dataset with all codon positions; noPART: single concatenated model (LG+G for amino acids and GTR+G for nucleotides); PART: partitioned model by gene (best models determined with ModelFinder); BP: ML bootstrap percentage; CFgene: gene concordance factor.

topologies. Brown and Thomson (2017) detected two genes as marked outliers within the 248 single-copy orthologous genes of Chiari et al. (2012) and showed that the corresponding alignments contained non-orthologous sequences, thus creating conflicting gene trees and resulting in the artefactual topology (T1) grouping of turtles and crocodiles when analysing the complete concatenated nucleotide dataset.

Protocols have been proposed to clean outlier sequences from alignments based on branch length analysis of the corresponding gene trees. Analyses of branch lengths between concatenation trees and those of individual gene trees have been used to exclude genes in phylogenomic analyses of metazoans (Simion et al., 2017), as well as to curate single-copy

orthologous gene alignments of the OrthoMaM database (Scornavacca et al., 2019), and to exclude outlier sequences when focusing on terminal branch lengths (Simion et al., 2017). This approach is here referred to as Branch Length Comparison (BLC). A similar method to detect outlier sequences that artificially inflate the diameter of individual gene trees was recently developed and implemented in the TreeShrink software package (Mai and Mirarab, 2018). Here we used the dataset of Chiari et al. (2012) to illustrate the impact of different cleaning methods on resolving phylogenetic conflicts regarding the position of turtles. First, we removed the two paralogs identified by Brown and Thomson (2017) of the original amino acid and nucleotide datasets (BT). Second, we used HMMCleaner to remove likely non-homologous sequence fragments from the original nucleotide datasets, removed residual sequences shorter than 50 nucleotides, inferred individual ML gene trees and concatenated trees under a GTR+G model using RAxML 8 (Stamatakis, 2014). Third, we applied TreeShrink with default parameters (TS, Mai and Mirarab 2018), and used the method of Simion et al. 2017 –here named BLC– to exclude outlier sequences having a terminal branch-length ratio >5 and gene alignments with an R2 Pearson correlation coefficient between all branch lengths in each gene tree and the corresponding supermatrix tree outside of the normal distribution (mean +/- 1.96 standard deviation). Finally, we performed phylogenetic reconstruction, along with gene and site concordance factors, on the resulting datasets using IQ-TREE (Minh et al., 2018) under a single concatenated LG+G or GTR+G model (noPART) and a gene partitioned model (PART), with model selection performed using ModelFinder (Kalyaanamoorthy et al., 2017) on amino acid datasets (AA), nucleotide datasets with only first and second codon positions (NT12), and nucleotide datasets with all codon positions (NT123).

The application of TreeShrink (TS) resulted in the removal of a total of 82 sequences in 76 gene alignments, whereas the BLC method removed 10 outlier sequences and 11 genes. Only BLC allowed automatic detection and exclusion of the two alignments containing paralogous sequences (ENSGALG00000008916 and ENSGALG00000011434). TS only excluded the *Monodelphis* sequence from ENSGALG00000008916. As previously shown by Brown and Thomson (2017), removing the two paralogous genes was enough to shift the topology inferred with the three codon positions (T1) to the highly supported amino acid topology (T2) in all cases (see Figure 4C). Automatic filtering with TreeShrink was inefficient as it resulted in supporting the artifactual T1 topology (BP = 73) and with an even higher gene concordance factor (CFgene = 28.3 vs. 25.8) than the original nucleotide dataset with all three codon positions included when analysed with a single concatenated model. In contrast, the BLC procedure retrieved the amino acid topology (T2) with strong support (BP = 100), even with a single concatenated model. As shown in Chiari et al. (2012), removing the saturated third codon positions worked for all filtering methods, as was also the case when using a gene partitioned model, which in all cases supported the T2 topology (see Figure 4C). However, in the latter case, the NT_123 and NT123_TS datasets only moderately supported the amino acid T2 topology (BP = 78 and 67, respectively), whereas all other methods and datasets provided strong support (BP = 100).

Finally, it is worth noting that the BLC filtering method generally resulted in higher gene- and site-concordance values compared to other filtering approaches (see Figure 4C), thus demonstrating that gene tree incongruence was reduced by efficient sequence filtering in individual gene alignments. Compared to Bayes factors and likelihood calculations, this method provides an automated and computationally efficient approach to decrease the impact of data error on phylogenomic inference. More generally, "gene incongruence" as detected by current methods does not seem to stem from biological processes. Instead, they

were mostly caused by orthology error and saturated positions, two typical methodological issues during phylogenomic data construction and analysis. Our experiment also highlights the importance of the signal-to-noise ratio. When low quality third codon positions were included, more noise than signal was incorporated because it is hard to extract signal from saturated data with the evolutionary models used here (note that codon models could behave differently as they explicitly account for genetic code structure). When only the first and second codon positions were retained, the absolute signal quantity was lower but the signal-to-noise ratio was higher as we discarded the noisy saturated data. Discarding data that would not be correctly modelled is one way of improving the accuracy of phylogenomic analyses.

4.3 Random contamination and short internal branches

Although it has been shown that cross-contamination (Laurin-Lemay et al., 2012; Simion et al., 2018) and contamination (Philippe et al., 2011b) can drastically distort phylogenomic trees, the level of contamination necessary to induce reconstruction errors as well as the nature of the errors are unknown. In this case study, we introduced various amounts of contaminants into a clean dataset and inferred phylogenomic trees to assess the impact of sequence contamination on phylogenomics. We used the eukaryotic reference alignments maintained in the Philippe lab from which we selected 33 molluscan species as well as 15 lophotrochozoan species as a close outgroup. For the contaminant sequences, we selected species from taxonomic groups observed to be frequent contaminant sources in real transcriptomic datasets (Philippe, unpublished observations): Amoebozoa, Apicomplexa, Arthropoda, Choanoflagellata, Ciliophora, Cryptophyta, Deuterostomia, Diplomonadida, Dinophycae, Fungi, Haptophyta, Heterolobosea, Kinetoplastida, Microsporidia, Nematoda, Platyhelminthes, Rotifera, Stramenopiles and Viridiplantae. Some chimaeras between closely related species were made to reduce the amount of missing data (for details, see Supplementary Material website). A total of 110 species were finally selected and the dataset was constructed by Philippe et al. (2011b). Briefly, ambiguously aligned positions were removed using Gblocks (Castresana, 2000) and a supermatrix was assembled using SCaFoS (Roure et al., 2007). Only 143 genes with less than 27 missing species were considered (see Supplementary Material website), yielding an alignment of 30,517 positions from 110 species with 13% missing data. From this alignment, we extracted an alignment of 48 uncontaminated data (the 33 molluscan and 15 lophotrochozoan species) with 25% missing data, which was used as reference.

For the sake of simplicity, the protocol is described for a contamination level of 5% of species and 5% of genes. The same protocol was repeated for all combinations between [5, 10, 25, 50] percent of genes and [5, 10, 25, 50] percent of species. Ten replicates were done for each contamination level. Briefly, 5% of genes (i.e. 7 genes) were randomly selected over the 143 proteins in the dataset. For each selected gene, 1 to 5% of the 48 species (i.e. 1 to 2 species) were randomly selected as contamination targets and we randomly drew a value between 1 and 5% to mimic the fact that in real alignments the contamination level varies greatly among species. Note that the target species were different for each gene. For each sequence to be contaminated, a species was randomly selected among the remaining 62 non lophotrochozoan species, and its sequence was used to replace the original target sequence. A supermatrix was then assembled using SCaFoS (Roure et al., 2007), yielding an alignment of the same size and level of completeness as the uncontaminated alignment (30,517 positions, 25% missing data).

The accuracy of the phylogenetic inferences performed in the presence of contamination

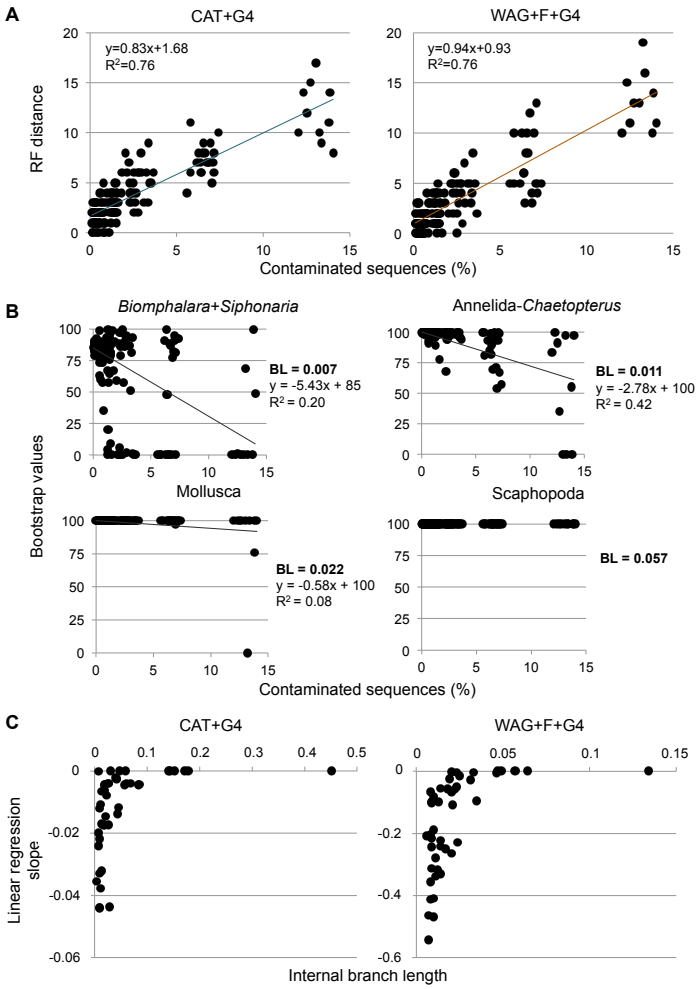


Figure 5 Effect of an increasing contamination level on the phylogenetic accuracy. A) The Robinson-Foulds distance was measured against the topology obtained from the uncontaminated dataset for all contaminated datasets. The inferences were conducted with the CAT+G4 and WAG+F+G4 models. The linear regression is plotted for each diagram. B) For four branches of different length, the bootstrap values of the inferences performed with the WAG+F+G4 model are plotted against the percentage of contaminated sequences. The linear regression is plotted for each node. BL is the internal branch length as estimated by the WAG+F+G4 model. C) The slope of the linear regression between the statistical support (as dependent variable) and the contamination level (as explanatory variable) is plotted against the internal branch length for all branches present in the topology obtained from the uncontaminated dataset with the CAT+G4 and WAG+F+G4 models.

was measured using the Robinson-Foulds distance (Robinson and Foulds, 1981) against the tree inferred from the dataset without contamination. As expected, the accuracy decreased with the amount of contaminated sequences, but no difference was detectable with respect to the sequence evolution model, with similar performances obtained with the CAT+G4 and WAG+F+G4 models (see Figure 5A).

As shown in Figure 5B in the case of the WAG+F+G4 model, the statistical support for a given node decreased with the contamination level. As expected, contamination had a greater impact for short branch lengths, while being negligible for medium and long

branches (compare slope and BL values on Figure 5B). Similar results were obtained with the CAT+G4 model (data not shown). To further validate this result, the linear regression slope (as in Figure 5B) was plotted against the branch length for all non-trivial bipartitions present in the trees inferred without contamination (see Figure 5C). For branches with a length greater than 0.1 (CAT) or 0.04 (WAG), the node was almost always recovered with maximal support and the slope was close to 0 (see Figure 5C). Otherwise the slope decreased with the branch shortness, indicating that inference was on average more sensitive to contamination with shorter branches. The dispersion of the slope values for a given branch length was likely due to the other parameters influencing phylogenetic inference (e.g. depth in the tree, number of taxa in the bipartitions, heterogeneity of evolutionary rates of species surrounding the branch). In summary, at realistic random contamination levels (<5%), only short branches will be negatively affected while most of the topology will remain unchanged. Note that this was due to the randomness of this experiment, thus producing noise. Had we simulated a contamination pattern consistently affecting the same taxa, thus introducing non-phylogenetic signal, even medium and long branches would have likely been affected. In conclusion, the main effect of a limited level of random contamination is a small reduction in the phylogenomics statistical power, so only frequent and biased contamination could explain incongruent phylogenomic trees.

4.4 Reappraisal of phylogenomic signal dissection methods

Recent years have brought forth a particular kind of phylogenomic analysis that aims at using single genes or single alignment sites to investigate phylogenetic relationships that are notoriously hard to resolve. We will here refer to these as "Constrained Topology Analyses" (CTA) as they include slightly different analyses schemes (e.g. Gene Genealogy Interrogation (GGI) [Arcila et al. 2017], ΔGLS and ΔSLS [Shen et al. 2017], Maximum Gene-Wise Edge (MGWE) [Walker et al. 2018], Bayes Factors [Brown and Thomson 2017]). These approaches use the general idea underlying all supertree methods that a "majority vote" from many small data subparts will help determine the best tree. They measure congruence and conflict of genes (or sites) relative to constrained tree topologies and then relies on a "majority vote" to determine which of these topologies is best supported by the data (see also Chapter 3.4 [Bryant and Hahn 2020]). CTA approaches were recently used to assess a variety of phylogenetic relationships, including the position of ctenophores (Arcila et al., 2017) (Shen et al., 2017), otophysans (Arcila et al., 2017), turtles (Brown and Thomson, 2017) (Walker et al., 2018), carnivorous Caryophyllales (Walker et al., 2018), or *Amborella* within angiosperms (Smith et al., 2015).

In this section, we reappraise the potential of CTA approaches to resolve notoriously difficult phylogenetic relationships while criticizing the scant amount of phylogenetic signal they rely on and highlighting the ever-important problem of model fit. CTA approaches are based on the assumption that every single-gene analysis yields enough phylogenetic signal to inform the difficult node under scrutiny. Since these nodes usually correspond to ancient events and/or short internal branches, it can be doubted that the findings of a single-gene analysis could reliably support a given topology versus alternative ones. Indeed, all recent CTA studies have reported a very high number of uninformative genes, and 82.8% to 97.3% of the genes did not reliably support either of the two main topologies under scrutiny (see Figure 6A). The data subparts are generally too small to resolve the problem at hand. This, in itself, is not problematic and is reminiscent of the era of single-gene phylogenetics where all phylogenetic analyses were interpreted in light of these limitations. In CTA, however, all of these single-gene analyses are often genuinely combined and the majority solution is con-

A

	Genes not preferring any of 2 main topologies	Total Genes	Genes supporting the best topology	Uninformative Genes	Informativeness Criteria
Smith et al. 2015	94,3%	5863	204	4024	Bootstrap > 50
Walker et al. 2017	82,8%	93	7	55	SH-like > 80
Arcila et al. 2017 (Metazoa)	96,2%	209	7	201	AU test p-value < 0,05
Arcila et al. 2017 (Eutheria)	97,3%	414	7	403	AU test p-value < 0,05
Arcila et al. 2017 (Neoavia)	93,8%	259	9	243	AU test p-value < 0,05
Shen et al. 2017	na	209	na	na	na

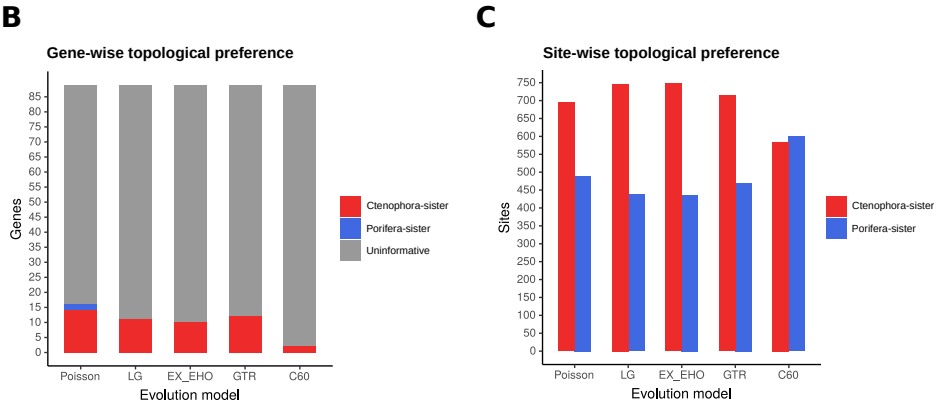


Figure 6 The limitations of CTA approaches. A) Summary statistics of recent CTA studies focused on hard to resolve phylogenetic nodes. Values in the analysis conducted by Smith et al. (2015) are means of the corresponding values for the three deep nodes under study (nodes 3, 4 and 6). B and C) Comparative reanalyses of a previous study (Shen et al., 2017) at gene and amino-acid site levels. B) Gene-wise information detected by CTA under several evolution models. Uninformative genes did not significantly reject the other topology tested according to the AU topology test (Shimodaira, 2002). C) Site-wise information detected by CTA under several evolution models when using the 5% strongest sites (as in the original study of Shen et al. 2017).

sidered to be the best. This is problematic for at least two reasons: (i) there is no guarantee that the majority signal is phylogenetic, as it could stem from an error source that would be additive in nature and therefore lead to a substantial systematic error (i.e. non-phylogenetic signal), and (ii) single gene inference is more complicated because less information is available to accurately infer the evolutionary model parameters. For example, if LBA artefacts affect many single-gene analyses, then the CTA results will erroneously consider that the majority LBA species tree is the correct topology. These issues are even more problematic when CTA is used at the scale of single site (see Shen et al. 2017). Using per-site likelihood differences between contradicting topologies is very similar to counting synapomorphies to build the best species tree: it would only work in the presence of low or negligible homoplasy, which is clearly not the case in large phylogenomic datasets. Historically, questions regarding dataset size and phylogenetic signal extraction triggered the emergence of "total-evidence", "multi-gene" and later phylogenomic approaches (Delsuc et al., 2005). This led to the resolution of many parts of the tree of life by maximising the amount of phylogenetic signal that could be extracted from the data. In this regard, recent CTA methods can be viewed as a return to a "low-evidence" approach as well as a violation of the ideal joint model described at the beginning of this chapter (see Figure 1).

CTA approaches can thus theoretically support an erroneous species tree, like conventional phylogenomic approaches. Since they are based on the sum of single-gene or single-site

signals, they are also impacted by the practical difficulties inherent to meeting phylogenetic objectives: using high quality data under a satisfactory evolutionary model in order to infer an accurate gene genealogy. The importance of data quality was stressed in Sections 4.1 and 4.2. We thus tested the impact of model choice on CTA approaches at gene and site scales by reanalysing a dataset from a recent study focused on the position of ctenophores within metazoans (i.e. the "D16_Opisthokonta" dataset of 89 genes, see Shen et al. 2017). This latter study originally did not check the significance of gene support for a given topology, so we used the same approach as another study instead, based on the approximately unbiased topology test (Arcila et al., 2017). Genes that do not significantly support any given topology are hereby called "uninformative". In addition, we checked the potential test rejection of *a priori* constrained topologies compared to the unconstrained, genuine, gene topology (i.e. *star tree*). This test has not been conducted in recent CTA studies, and our results shows that a large proportion of constrained topologies are significantly rejected when compared to the unconstrained topology, across evolution models: 98% (Poisson), 97% (LG), 98% (EX_EHO), 93% (GTR), and 93% (C60), see Supplementary Material website. This confirms that most single-gene datasets do not carry enough phylogenetic signal to recover a reasonable species tree, regardless of the evolution model used.

Our gene-wise reappraisal of CTA approaches using different sequence evolution models revealed the same trend as noted in previous studies: a very large amount of the genes are uninformative as they do not contain enough phylogenetic signal to favour any constrained topology (grey bars in Figure 6B). Moreover, improving the model complexity decreases the number of informative genes, suggesting that these genes might support a given topology under a simpler model because of model misspecifications. With the C60 site-heterogeneous model, virtually all genes are considered as uninformative regarding the two topologies under scrutiny (Figure 6B). The impact of the model choice is even greater when considering site-wise data. The absolute site-preference for the two topologies tested decreases with the model complexity, as for gene-wise analyses (see AU tests results provided in Supplementary Material website), again indicating that general site support for topologies might be overestimated with simple models. More importantly, whereas simple models show a greater number of sites seemingly supporting ctenophores as a sister-clade to other metazoans (as in Shen et al. 2017), using the more complex and better fitting C60 site-heterogeneous model leads to the opposite conclusion and favours sponges as a sister-clade to other metazoans (see Figure 6C). This highlights that CTA approaches can be highly impacted by model choice upon which the conclusions of studies are based on.

The critical reappraisal of CTA approaches presented here is not intended to discourage its potential use for phylogenomics. Dissecting conflicting signals in large heterogeneous phylogenomic datasets is both important and helpful. However, our results show that CTA still needs to be refined in the following areas: i) investigating the validity of using a constrained topology when the data significantly supports a different topology, ii) ensuring data quality, and iii) selecting the most adequate evolution model for each data subset. It is still doubtful that CTA could effectively resolve notoriously contentious relationships because it is hard for the models to extract enough phylogenetic signal from a small dataset (e.g. see Wang et al. 2019). Small datasets indeed contain a limited absolute amount of phylogenetic signal and the lack of data hampers complex models from accurately estimating parameter values. The main results of CTA studies to date are that outlier genes that generally channel phylogenomic analyses towards an erroneous species tree actually (Brown and Thomson, 2017) or likely (Walker et al., 2018) correspond to data errors (e.g. paralogs, contamination). In phylogenomics studies, CTA approaches might therefore be an efficient data quality

check tool for detecting outliers, rather than being an efficient phylogenetic signal extraction approach.

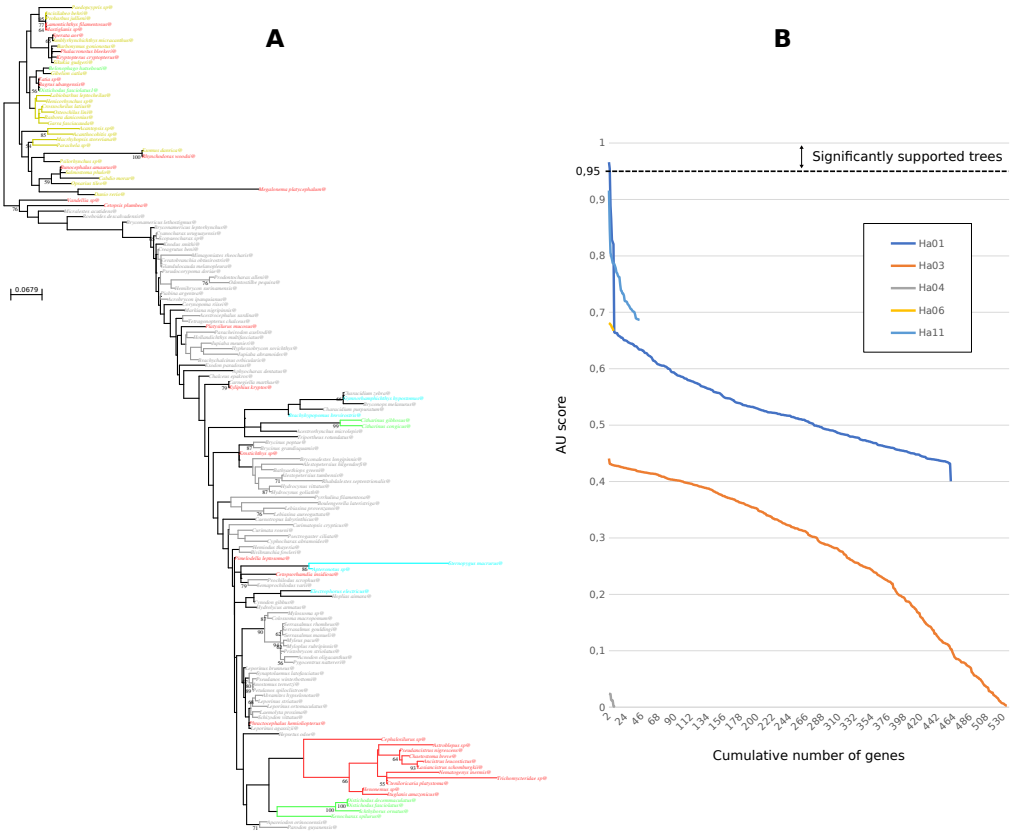


Figure 7 Example of cross-contaminations in the dataset of Arcila et al. (2017). The alignment of the ENSDARG00000061941_5_17739391_17739087 locus was analysed under the GTR+G model using RAxML. A) To illustrate cross-contamination simply, the five major monophyletic groups are shown in color (Cypriniformes in yellow, Gymnotiformes in blue, Siluriformes in red, Characoidei in grey and Citharinoidei in green). For instance, the sequence of the Characoidei, *Characidium zebra*, is identical to *Gymnorhamphichthys hypostomus*, a member of Gymnotiformes. It is likely that the correct sequence is that of *Characidium zebra* since it is closely related to *Characidium purpuratum*. B) Reanalysis of Arcila et al. (2017) dataset using IQ-TREE to perform the Approximately Unbiased (AU) topology test. Lines represent the cumulative number of genes (x axes) supporting each hypothesis with highest probability (rank 1) and their associated P-values (y axes) according to the AU topology test. Values above the dashed line indicate all rank 1 hypotheses that are significantly better than the alternatives ($P < 0.05$), whereas those below the dashed line are also rank 1 but without statistical significance.

4.5 Opening the phylogenomic black box

Phylogenomic pipelines and their various components are becoming increasingly complicated, and the volume of data to handle is exponentially increasing. Accordingly, it is tempting to consider the pipeline as a black box and to focus on the last step, i.e. alignment-based tree inference. However, the road from biological samples to the species tree is long and paved with multiple potential errors, as shown in this chapter. Placing too much trust in the phylogenomic black box is thus risky. For instance, Smith et al. (2011) stated that "Phy-

loBayes misidentified the data type of [their] matrix", without questioning their pipeline or looking at the data, despite the fact that six amino acids (E, F, I, L, P and Q) were entirely missing from their dataset. This led the authors to publish a corrigendum to their study stating that this error was due to a bug in their phylogenomic pipeline (Smith et al., 2013). Similarly, Finet et al. (2010) inferred that two well-established clades (Zygnematales and Coleochetales) were not monophyletic, although their study focused on a sister-group of land plants. Yet a close visual examination of their alignments and corresponding gene trees revealed numerous cross-contamination events that were responsible for these unexpected results (Laurin-Lemay et al., 2012).

It is thus time to open the phylogenomic black box, take a close look at the intermediate data and results, and check that they make sense in the light of current knowledge. We briefly illustrate this by dissecting a recent study on the radiation of Otophysi fish (Arcila et al., 2017). The relationships between four well-established clades (Gymnotiformes, Siluriformes, Characoidei and Citharinoidea) were recognized as difficult to resolve. The authors sequenced 1,051 genes for 225 species, using monophyletic Cypriniformes as outgroup. They used 45 different tree reconstruction methods based on concatenation and species tree approaches and observed that only 5 out of the 15 possible topologies were never supported. The two most frequently recovered topologies were retrieved in only 11 and 9 of the 45 results. In contrast, the monophyly of the 5 clades (Cypriniformes, Gymnotiformes, Siluriformes, Characoidei and Citharinoidea) was always recovered with maximal support. This result is in full agreement with previous knowledge: the length of internal branches is long for the monophyly of the 5 clades (plenty of phylogenetic signal) and short with regard to the relationships among the 5 clades (sparse phylogenetic signal). Instead of concluding that the data quantity was insufficient to resolve this radiation, the authors used the GGI approach in an attempt to find signal in single gene trees. The theoretical expectation is that single genes should have some signal in support of the monophyly of the 5 clades and virtually no signal for their inter-relationships. Indeed, if a single gene provides strong support for any inter-clade relationship, this gene likely did not follow the same historical path as the 1,050 other genes (e.g. contains paralogs). Surprisingly, Arcila et al. (2017) disregarded this theoretical expectation and instead decided "to gain additional insights [...] [to] constrain gene-tree space to a small number of relevant options (15 in this case; Figure1) [to] overcome gene tree estimation error." In other words, they constrained branches for which a single gene was expected to have some signal so as to find support for branches for which the single gene was not expected to bear any signal. Quite surprisingly, their approach seemed to succeed since they found 325 genes that strongly supported one topology (topology H_0 which was supported by seven of the 45 analyses described above) and 69 another one (topology H_a10 which was never supported by any of the 45 analyses). This raises two questions: (1) why was monophyly of the 5 clades constrained, and (2) why was there so much phylogenetic signal in 40% of the loci for the short internal branches connecting these 5 clades but none when concatenated genes were analysed? Indeed, it is striking that only one out of 23 concatenation approaches recovered a topology favoured by the gene-scale approach (i.e., H_0 and not H_a10); Instead, concatenations mostly supported H_a01, H_a03 and H_a04. Unfortunately, the answers are to be found in errors in the phylogenomic pipeline, not in any previously overlooked biological properties.

First, Arcila et al. (2017) recognised that: "it is possible that deep coalescences could result in particular gene histories that display non-monophyly of one or more subclades. This possibility, however, is highly unlikely because internal branches subtending subclades span 20–65 million years of evolution." But they did not foresee that the non-monophyly

of the 5 clades could be due to data error (i.e. cross-contamination events). A comparison of single gene and concatenation branch lengths using the BLC method (see previous section 4.2) highlighted numerous anomalous terminal branch lengths and led to some very poor correlation coefficients. For instance, the phylogeny based on the ENSDARG00000061941_5_17739391_17739087 locus revealed numerous cross-contamination events (see Figure 7A). Among other events, the Siluriformes *Tatia* and *Bagrus ubangensis* are identical to the Citharinoidei *Distichodus fasciolatus*1 (distant from other species of the *Distichodus* genus), and this cluster is deeply nested within the outgroup, i.e. Cypriniformes. Using a variety of approaches based mainly on branch length comparison and BLAST similarity, we estimated that about 12.3% of the sequences of this dataset stemmed from contamination events (20,000 out of the 162,555 sequences. For details, see Supplementary Material website). This indicates that the ENSDARG00000061941_5_17739391_17739087 locus is therefore not an exception. Our findings complement those of a later study from the same group (Betancur-R. et al., 2019), which showed that a dataset from a competing study was cross-contaminated but ironically did not check their own previous data. In summary, constraining the monophyly of the five clades in the Arcila et al. (2017)'s study was mainly necessitated by the presence of numerous incorrectly identified sequences.

Second, the presence of a very strong signal for deep relationships in 394 out of the 1,051 loci reported by Arcila et al. (2017) could solely be explained by software error, i.e. a bug in the RAxML version used, as confirmed by Alexis Stamatakis (personal communication). When we performed the computation with IQ-TREE (Nguyen et al., 2015), our results perfectly fit the theoretical expectations: only 2 out of the 1,051 loci significantly supported any of the 15 possible topologies (see Figure 7B). This observation is in agreement with results of Section 4.4. Moreover, contrary to Figure 2a in Arcila et al. (2017), where the best two topologies were H_0 and H_a10, our results suggested that the best two topologies were H_a01 and H_a03. The topology that was the most frequently recovered in the 45 experiments of Arcila et al. (see H_a01 in their Figure 1) displayed the highest average AU test p-value in our results. This was in full agreement with the theoretical expectation, as the accumulation of a very weak signal (phylogenetic or not) over a large number of genes appeared to be the dominant signal in the concatenated tree.

These examples demonstrated how easily errors can arise along the long road leading from biological samples to species trees and lead to erroneous conclusions (i.e. data error and software error in the case of Arcila et al. 2017). Not only a combination of automatic and manual quality control needs to be performed at each step along phylogenomic pipelines, but also all (intermediate) results should be evaluated in the light of theoretical expectations. For instance, a transcriptome assembly with a reasonable N50 but more than 100,000 contigs, or a very unexpected phylogenetic placement or branch length for a taxon should immediately trigger the opening of the phylogenomic black box. Indeed, such a transcriptome likely contains contaminants and/or fragmented sequences, and a surprisingly long branch taxon could be produced by incorrect underlying data.

5 Conclusions

Multiple types of error occur along the long road from organisms to the species phylogeny, all of which are ultimately due to model violations. They generally decrease the resolving power of phylogenomics (e.g. Figure 5), but occasionally increase it in favour of an erroneous solution (e.g. biased contamination or software bugs). Since very few studies have focused on tracking these errors, we have little idea on the extent of their impact on the species

phylogeny accuracy. We argue that it is essential to identify the most damaging errors (e.g. contamination, annotation errors, orthology errors, alignment errors, single gene tree errors, violation of sequence or gene evolution models) and devote more energy correcting them. Note that the extent of the damaging effect might differ markedly depending on the result of interest. For instance, an annotation error might have a limited impact on topology inference, but a huge impact on branch length and positive selection estimates (Di Franco et al., 2019).

Our current knowledge of biology and evolution could guide us in identifying relevant model violations (and the errors they introduced in the divide and conquer approach of the ideal model in Figure 1. For instance, ignoring ILS would likely only be important for very short internal branches, while over-simplified sequence evolution models could be adequate for minor phylogenetic issues. Posterior predictive checks could enhance studies on these intuitions by quantifying data aspects that are the most poorly explained by the model. For example, recent studies have stressed the importance of rampant discordance between gene trees (Hahn and Nakhleh, 2016), while it could be argued that rampant data error is an equally (or more) serious threat.

Two approaches are conventionally used to eschew these model violations. The first one consists in identifying and removing the most problematic data. For instance, cross-contaminants may be removed from transcriptomic data (Simion et al., 2018), poorly aligned regions (Di Franco et al., 2019) or genes/sites that seriously violate the model assumptions (Roure and Philippe, 2011). This approach is not well founded from a statistical standpoint because the data has to be analysed multiple times while removing data instead of developing an adequate model. However, it is computationally quite efficient and seems reasonnable when founded on solid external knowledge. We believe research in that direction should be pursued. The second approach involves the development of better models. Each sub-model may be improved along the phylogenomic pipeline independently, or sub-models could be combined to allow joint analysis (e.g. alignment and phylogeny). We feel that studies should be focused on sequence evolution models (see Chapter 1.4 [Lartillot 2020a]) and on joint inference of gene trees and species trees (as in Boussau et al. 2013). Yet we stress that software error is an emerging threat to the phylogenomic approach and that increasing the model complexity or implementing clever mathematical tricks to accelerate the computation will increase the software error risk. More resources need to be devoted to the development of high quality software.

Finally, in the current Anthropocene age, the question arises as to whether phylogenomics is a past or future science. Unfortunately, at a time when the environmental footprint of humanity should be drastically reduced [IPCC and IPBES reports], enhancing the accuracy of phylogenomics would require a sharp increase in the computational burden (more data and more complex joint models). As an illustrative example, the computational footprint of xenambulacrarian phylogenomics rose from 7T of CO2 in 2011 (Philippe et al., 2011a) to 260T in 2019 (Philippe et al., 2019), roughly equivalent to 137 round-trip flights between New York and Paris. This raises a legitimate question as to whether pursuing biodiversity science under current scientific practices is reasonable.

References

Alié, A., Hiebert, L. S., Simion, P., Scelzo, M., Prünster, M. M., Lotito, S., Delsuc, F., Douzery, E. J. P., Dantec, C., Lemaire, P., Darras, S., Kawamura, K., Brown, F. D., and Tiozzo, S. (2018). Convergent acquisition of nonembryonic development in styelid ascidians. *Molecular Biology and Evolution*, 35(7):1728–1743.

Allio, R., Scornavacca, C., Nabholz, B., Clamens, A.-L., Sperling, F. A., and Condamine, F. L. (2020). Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. *Systematic Biology*, 69(1):38–60.

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D., and Dessimoz, C. (2019). OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Research*, 29(7):1152–1163.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., Sabaj, M. H., Lundberg, J., Revell, L. J., and Betancur-R, R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(2):0020.

Ballenghien, M., Faivre, N., and Galtier, N. (2017). Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. *BMC Biology*, 15(1):25.

Ballesteros, J. A. and Hormiga, G. (2016). A new orthology assessment method for phylogenomic data: Unrooted phylogenetic orthology. *Molecular Biology and Evolution*, 33(8):2117–2134.

Bayzid, M. S., Hunt, T., and Warnow, T. (2014). Disk covering methods improve phylogenomic analyses. *BMC Genomics*, 15(6):S7.

Beaulieu, J. M., O'Meara, B. C., Zaretzki, R., Landerer, C., Chai, J., and Gilchrist, M. A. (2019). Population genetics based phylogenetics under stabilizing selection for an optimal amino acid sequence: A nested modeling approach. *Molecular Biology and Evolution*, 36(4):834–851.

Betancur-R., R., Arcila, D., Vari, R. P., Hughes, L. C., Oliveira, C., Sabaj, M. H., and Ortí, G. (2019). Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes. *Evolution*, 73(2):329–345.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504.

Blanquart, S. and Lartillot, N. (2006). A bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution*, 23(11):2058–2071.

Blanquart, S. and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution*, 25(5):842–858.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Plessis, L. d., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):e1006650.

Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.

Boussau, B., Szöllősi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330.

Brown, J. M. and Thomson, R. C. (2017). Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Systematic Biology*, 66(4):517–530.

Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552.

Chiari, Y., Cahais, V., Galtier, N., and Delsuc, F. (2012). Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (archosauria). *BMC Biology*, 10(1):65.

Cornet, L., Meunier, L., Vlierberghe, M. V., Léonard, R. R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E. J., Philippe, H., Wilmotte, A., and Baurain, D. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLOS ONE*, 13(7):e0200323.

Crawford Nicholas G., Faircloth Brant C., McCormack John E., Brumfield Robb T., Winker Kevin, and Glenn Travis C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, 8(5):783–786.

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermiin, L. S., and Haeseler, A. v. (2019). GHOST: Recovering historical signal from heterotachously-evolved sequence alignments. *bioRxiv*, page 174789.

Czech, L., Huerta-Cepas, J., and Stamatakis, A. (2017). A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular Biology and Evolution*, 34(6):1535–1542.

Dalquen, D. A., Anisimova, M., Gonnet, G. H., and Dessimoz, C. (2012). ALF—a simulation framework for genome evolution. *Molecular Biology and Evolution*, 29(4):1115–1123.

Darriba, D., Flouri, T., and Stamatakis, A. (2018). The state of software for evolutionary biology. *Molecular Biology and Evolution*, 35(5):1037–1046.

de Vienne, D. M., Ollier, S., and Aguileta, G. (2012). Phylo-MCOA: A fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Molecular Biology and Evolution*, 29(6):1587–1598.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361.

Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evolutionary Biology*, 19(1):21.

Dondi, R., El-Mabrouk, N., and Lafond, M. (2016). Correction of weighted orthology and paralogy relations - complexity and algorithmic results. In Frith, M. and Storm Pedersen, C. N., editors, *Algorithms in Bioinformatics*, Lecture Notes in Computer Science, pages 121–136. Springer International Publishing.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q.,

and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–749.

Dunne, M. P. and Kelly, S. (2017). OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations. *BMC Genomics*, 18(1):390.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1):157.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584.

Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.

Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., and Pisani, D. (2017). Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology*, 27(24):3864–3870.e4.

Finet, C., Timme, R. E., Delwiche, C. F., and Marlétaz, F. (2010). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, 20(24):2217–2222.

Fleissner, R., Metzler, D., and von Haeseler, A. (2005). Simultaneous statistical multiple alignment and phylogeny reconstruction. *Systematic Biology*, 54(4):548–561.

Gatesy, J., Sloan, D. B., Warren, J. M., Baker, R. H., Simmons, M. P., and Springer, M. S. (2019). Partitioned coalescence support reveals biases in species-tree methods and detects gene trees that determine phylogenomic conflicts. *Molecular Phylogenetics and Evolution*, 139:106539.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.

Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70(1):7–17.

Hejnol Andreas, Obst Matthias, Stamatakis Alexandros, Ott Michael, Rouse Greg W., Edgecombe Gregory D., Martinez Pedro, Baguñà Jaume, Bailly Xavier, Jondelius Ulf, Wiens Matthias, Müller Werner E. G., Seaver Elaine, Wheeler Ward C., Martindale Mark Q., Giribet Gonzalo, and Dunn Casey W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences*, 276(1677):4261–4270.

Herman, J. L., Challis, C. J., Novák, A., Hein, J., and Schmidler, S. C. (2014). Simultaneous bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Molecular Biology and Evolution*, 31(9):2251–2266.

Holmes, I. and Bruno, W. J. (2001). Evolutionary HMMs: a bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820.

Horiike, T., Minai, R., Miyata, D., Nakamura, Y., and Tateno, Y. (2016). Ortholog-finder: A tool for constructing an ortholog data set. *Genome Biology and Evolution*, 8(2):446–457.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47:D309–D314.

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.-Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., and Philippe, H. (2017). Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature ecology & evolution*, 1(9):1370–1378.

Jacox, E., Chauve, C., Szöllősi, G. J., Ponty, Y., and Scornavacca, C. (2016). ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058.

Kaduk, M. and Sonnhammer, E. (2017). Improved orthology inference with hieranoid 2. *Bioinformatics*, 33(8):1154–1159.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Koski, L. B. and Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, 52(6):540–542.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Kozlov, A. M. and Stamatakis, A. (2020). Using raxml-ng in practice. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.3, pages 1.3:1–1.3:25. No commercial publisher | Authors open access book.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

Kuzniar, A., [van Ham], R. C., Pongor, S., and Leunissen, J. A. (2008). The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539 – 551.

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

Lartillot, N. (2020a). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.

Lartillot, N. (2020b). Phylobayes: Bayesian phylogenetics using site-heterogeneous models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.5, pages 1.5:1–1.5:16. No commercial publisher | Authors open access book.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288.

Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Laumer, C. E. (2018). Inferring ancient relationships with genomic data: A commentary on current practices. *Integrative and Comparative Biology*, 58(4):623–639.

Laurin-Lemay, S., Brinkmann, H., and Philippe, H. (2012). Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology*, 22(15):R593–R594.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189.

Lowe, C. and Rodrigue, N. (2020). Detecting adaptation from multi-species protein-coding dna sequence alignments alignments. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.5, pages 4.5:1–4.5:18. No commercial publisher | Authors open access book.

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

Miele, V., Penel, S., and Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, 12:116.

Minh, B. Q., Hahn, M. W., and Lanfear, R. (2018). New methods to calculate concordance factors for phylogenomic datasets. *bioRxiv*, page 487801.

Mirarab, S., Bayzid, M. S., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65(3):366–380.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Monnahan, P. J., Michno, J.-M., O'Connor, C. H., Brohammer, A. B., Springer, N. M., McGaugh, S. E., and Hirsch, C. N. (2019). Using multiple reference genomes to identify and resolve annotation inconsistencies. *bioRxiv*, page 651984.

Moroz, L. L., Kocot, K. M., Citarella, M. R., Dosung, S., Norekian, T. P., Povolotskaya, I. S., Grigorenko, A. P., Dailey, C., Berezikov, E., Buckley, K. M., Ptitsyn, A., Reshetov, D., Mukherjee, K., Moroz, T. P., Bobkova, Y., Yu, F., Kapitonov, V. V., Jurka, J., Bobkov, Y. V., Swore, J. J., Girardo, D. O., Fodor, A., Gusev, F., Sanford, R., Bruders, R., Kittler, E., Mills, C. E., Rast, J. P., Derelle, R., Solovyev, V. V., Kondrashov, F. A., Swalla, B. J., Sweedler, J. V., Rogaev, E. I., Halanych, K. M., and Kohn, A. B. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, 510(7503):109–114.

Necsulea, A. (2020). Phylogenomics and genome annotation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.1, pages 4.1:1–4.1:26. No commercial publisher | Authors open access book.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment11edited by j. thornton. *Journal of Molecular Biology*, 302(1):205–217.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J., and Telford, M. J. (2011a). Acoelomorph flatworms are deuterostomes related to xenoturbella. *Nature*, 470(7333):255–258.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011b). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLOS Biology*, 9(3):e1000602.

Philippe, H., Chenuil, A., and Adoutte, A. (1994). Can the cambrian explosion be inferred through molecular phylogeny? *Development*, 1994:15–25.

Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005). Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, 36(1):541–562.

Philippe, H., Poustka, A. J., Chiodin, M., Hoff, K. J., Dessimoz, C., Tomiczek, B., Schiffer, P. H., Müller, S., Domman, D., Horn, M., Kuhl, H., Timmermann, B., Satoh, N., Hikosaka-Katayama, T., Nakano, H., Rowe, M. L., Elphick, M. R., Thomas-Chollier, M., Hankeln, T., Mertes, F., Wallberg, A., Rast, J. P., Copley, R. R., Martinez, P., and Telford, M. J. (2019). Mitigating anticipated effects of systematic errors supports sister-group relationship between xenacoelomorpha and ambulacraria. *Current Biology*, 29(11):1818–1826.e6.

Pich, O., Muiños, F., Sabarinathan, R., Reyes-Salazar, I., Gonzalez-Perez, A., and Lopez-Bigas, N. (2018). Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes. *Cell*, 175(4):1074–1087.e18.

Prous, M., Lee, K. M., and Mutanen, M. (2020). Cross-contamination and strong mitonuclear discordance in empria sawflies (hymenoptera, tenthredinidae) in the light of phylogenomic data. *Molecular Phylogenetics and Evolution*, 143:106670.

Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.

Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.

Ranwez, V. and Chantret, N. (2020). Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.2, pages 2.2:1–2.2:36. No commercial publisher | Authors open access book.

Ranwez, V. and Delsuc, F. (2020). Accurate alignment of (meta)barcoding datasets using macse. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.3, pages 2.3:1–2.3:31. No commercial publisher | Authors open access book.

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., and Delsuc, F. (2018). MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, 35(10):2582–2584.

Redelings, B. D. and Suchard, M. A. (2005). Joint bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3):401–418.

Rey, C., Veber, P., Boussau, B., and Sémon, M. (2019). CAARS: comparative assembly and annotation of RNA-seq data. *Bioinformatics*, 35(13):2199–2207.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10):4629–4634.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Roure, B. and Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evolutionary Biology*, 11(1):17.

Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evolutionary Biology*, 7(1):S2.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., Hériché, J.-K., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A. J., Ureta-Vidal, A., Bolund, L., Wang, J., and Durbin, R. (2008). TreeFam: 2008 update. *Nucleic Acids Research*, 36:D735–D740.

Scornavacca, C., Belkhir, K., Lopez, J., Dernat, R., Delsuc, F., Douzery, E. J. P., and Ranwez, V. (2019). OrthoMaM v10: Scaling-up orthologous coding sequence and exon alignments with more than one hundred mammalian genomes. *Molecular Biology and Evolution*, 36(4):861–862.

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5):0126.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7:539.

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J. C., Manuel, M., Philippe, H., and Telford, M. J. (2018). A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology*, 16(1):28.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., and Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, 27(7):958–967.

Simmons, M. P., Sloan, D. B., and Gatesy, J. (2016). The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Molecular Phylogenetics and Evolution*, 97:76–89.

Simmons, M. P., Sloan, D. B., Springer, M. S., and Gatesy, J. (2019). Gene-wise resampling outperforms site-wise resampling in phylogenetic coalescence analyses. *Molecular Phylogenetics and Evolution*, 131:80–92.

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, 15(1):150.

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C. S., Rouse, G. W., Giribet, G., and Dunn, C. W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, 480(7377):364–367.

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C. S., Rouse, G. W., Giribet, G., and Dunn, C. W. (2013). Corrigendum: Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, 493(7434):708–708.

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.

Steinegger, M. and Salzberg, S. L. (2020). Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Biorxiv preprint `https://doi.org/10.1101/2020.01.26.920173`.

Struck, T. H. (2014). TreSpEx—detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics*, 10:EBO.S14239.

Susko, E. and Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*, 24(9):2139–2150.

Tannier, E., Bazin, A., Davín, A. A., Guéguen, L., Bérard, S., and Chauve, C. (2020). Ancestral genome organization as a diagnosis tool for phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.5, pages 2.5:1–2.5:19. No commercial publisher | Authors open access book.

Tavare, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, pages 57–86.

Walker, J. F., Brown, J. W., and Smith, S. A. (2018). Analyzing contentious relationships and outlier genes in phylogenomics. *Systematic Biology*, 67(5):916–924.

Wang, H.-C., Minh, B. Q., Susko, E., and Roger, A. J. (2018). Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Systematic Biology*, 67(2):216–235.

Wang, H.-C., Susko, E., and Roger, A. J. (2019). The relative importance of modeling site pattern heterogeneity versus partition-wise heterotachy in phylogenomic inference. *Systematic Biology*, 68(6):1003–1019.

Wheeler, W. C., Lucaroni, N., Hong, L., Crowley, L. M., and Varón, A. (2015). POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. *Cladistics*, 31(2):189–196.

Whelan, S., Irisarri, I., and Burki, F. (2018). PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics*, 34(22):3929–3930.

Yang, H., Jaime, M., Polihronakis, M., Kanegawa, K., Markow, T., Kaneshiro, K., and Oliver, B. (2018). Reannotation of eight drosophila genomes. *bioRxiv*, page 350363.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.

Yu, J. and Thorne, J. L. (2006). Dependence among sites in RNA evolution. *Molecular Biology and Evolution*, 23(8):1525–1537.

Zhukova, A., Gascuel, O., Duchêne, S., Ayres, D. L., Lemey, P., and Baele, G. (2020). Efficiently analysing large viral data sets in computational phylogenomics. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.3, pages 5.3:1–5.3:43. No commercial publisher | Authors open access book.