# Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics

**Gustavo A Bravo** [Corresp., 1], **Alexandre Antonelli** [1, 2, 3, 4], **Christine D Bacon** [2, 3], **Krzysztof Bartoszek** [5], **Mozes P K Blom** [6], **Stella Huynh** [7], **Graham Jones** [3], **L. Lacey Knowles** [8], **Sangeet Lamichhaney** [1], **Thomas Marcussen** [9], **Hélène Morlon** [10], **Luay K Nakhleh** [11], **Bengt Oxelman** [2, 3], **Bernard Pfeil** [3], **Alexander Schliep** [12], **Niklas Wahlberg** [13], **Fernanda P Werneck** [14], **John Wiedenhoeft** [12, 15], **Sandi Willows-Munro** [16], **Scott V Edwards** [1, 17]

[1] Department of Organismic Biology and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States
[2] Gothenburg Global Biodiversity Centre, Göteborg, Sweden
[3] Department of Biological and Environmental Sciences, University of Gothenburg, Göteborg, Sweden
[4] Gothenburg Botanical Garden, Göteborg, Sweden
[5] Department of Computer and Information Science, Linköping University, Linköping, Sweden
[6] Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden
[7] Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland
[8] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA
[9] Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway
[10] Institut de Biologie, Ecole Normale Supérieure de Paris, Paris, France
[11] Department of Computer Science, Rice University, Houston, TX, United States
[12] Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Göteborg, Sweden
[13] Department of Biology, Lund University, Lund, Sweden
[14] Coordenação de Biodiversidade, Programa de Coleções Científicas Biológicas, Instituto Nacional de Pesquisa da Amazônia, Manaus, AM, Brazil
[15] Department of Computer Science, Rutgers University, Piscataway, NJ, USA
[16] School of Life Sciences, University of Kwazulu-Natal, Pietermaritzburg, South Africa
[17] Gothenburg Centre for Advanced Studies in Science and Technology, Chalmers University of Technology and University of Gothenburg, Göteborg, Sweden

Corresponding Author: Gustavo A Bravo
Email address: gustavo_bravo@fas.harvard.edu

Building the Tree of Life (ToL) is a major challenge of modern biology, requiring advances in cyberinfrastructure, data collection, theory, and more. Here, we argue that phylogenomics stands to benefit by embracing the many heterogeneous genomic signals emerging from the first decade of large-scale phylogenetic analysis spawned by high-throughput sequencing (HTS). Such signals include those most commonly encountered in phylogenomic datasets, such as incomplete lineage sorting, but also those reticulate processes emerging with greater frequency, such as recombination and introgression. Here we focus specifically on how phylogenetic methods can accommodate the heterogeneity incurred by such population genetic processes; we do not discuss phylogenetic methods that ignore such processes, such as concatenation or supermatrix approaches or supertrees. We suggest that methods of data acquisition and the types of markers used in phylogenomics will remain restricted until *a posteriori* methods of marker choice are made possible with routine whole-genome sequencing of taxa of interest. We

discuss limitations and potential extensions of a model supporting innovation in phylogenomics today, the multispecies coalescent model (MSC). Macroevolutionary models that use phylogenies, such as character mapping, often ignore the heterogeneity on which building phylogenies increasingly rely and suggest that assimilating such heterogeneity is an important goal moving forward. Finally, we argue that an integrative cyberinfrastructure linking all steps of the process of building the ToL, from specimen acquisition in the field to publication and tracking of phylogenomic data, as well as a culture that values contributors at each step, are essential for progress.

Article submission to *PeerJ*
Manuscript category: Literature Review Articles
Collection: "*Endless forms: Advances in evolutionary analyses of biodiversity*"

Article title: **Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics**

Gustavo A. Bravo*[1], Alexandre Antonelli[1,2,3,4], Christine D. Bacon[2,3], Krzysztof Bartoszek[5], Mozes P. K. Blom[6], Stella Huynh[7], Graham Jones[3], L. Lacey Knowles[8], Sangeet Lamichhaney[1], Thomas Marcussen[9], Hélène Morlon[10], Luay K. Nakhleh[11], Bengt Oxelman[2,3], Bernard Pfeil[3], Alexander Schliep[12], Niklas Wahlberg[13], Fernanda P. Werneck[14], John Wiedenhoeft[12,15], Sandi Willows-Munro[16], Scott V. Edwards[1,17]

[1] Harvard University, Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, 26 Oxford St., Cambridge, MA 02138, USA.
[2] Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Göteborg, Sweden.
[3] University of Gothenburg, Department of Biological and Environmental Sciences, Box 461, 405 30 Göteborg, Sweden.
[4] Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, SE-41319 Göteborg, Sweden.
[5] Linköping University, Department of Computer and Information Science, SE-581 83 Linköping, Sweden.
[6] Swedish Museum of Natural History, Department of Bioinformatics and Genetics, Box 50007, 104 05 Stockholm, Sweden.
[7] Université de Neuchâtel, Institut de Biologie, Rue Emile-Argand 11, Bureau D322, Neuchâtel, Switzerland.
[8] University of Michigan, Department of Ecology and Evolutionary Biology, Ann Arbor, MI 48109, USA.
[9] University of Oslo, Centre for Ecological and Evolutionary Synthesis, P.O. Box 1066 Blindem, NO-0316, Oslo, Norway.
[10] Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, Ecole Normale Supérieure, Paris Sciences et Lettres (PSL) Research University, F-75005 Paris, France.
[11] Rice University, Department of Computer Science, 6100 Main St. MS 132, Houston, TX 77005, USA.
[12] Chalmers University of Technology and University of Gothenburg, Department of Computer Science and Engineering, Computing Science Division, SE-412 96, Göteborg, Sweden.
[13] Lund University, Department of Biology, Sölvegatan 37, SE-223 62 Lund, Sweden.
[14] Instituto Nacional de Pesquisa da Amazônia, Coordenação de Biodiversidade, Programa de Coleções Científicas Biológicas, 69060-000 Manaus, AM, Brazil.
[15] Rutgers University, Department of Computer Science, 110 Frelinghuysen Rd. Piscataway, NJ 08854, USA.
[16] University of Kwazulu-Natal, School of Life Sciences, Pietermaritzburg, South Africa.
[17] Chalmers University of Technology and University of Gothenburg, Gothenburg Centre for Advanced Studies in Science and Technology, SE-412 96, Göteborg, Sweden.

*Corresponding author: Gustavo A. Bravo, E-mail: gustavo_bravo@fas.harvard.edu

**Abstract**

Building the Tree of Life (ToL) is a major challenge of modern biology, requiring advances in cyberinfrastructure, data collection, theory, and more. Here, we argue that phylogenomics stands to benefit by embracing the many heterogeneous genomic signals emerging from the first decade of large-scale phylogenetic analysis spawned by high-throughput sequencing (HTS). Such signals include those most commonly encountered in phylogenomic datasets, such as incomplete lineage sorting, but also those reticulate processes emerging with greater frequency, such as recombination and introgression. Here we focus specifically on how phylogenetic methods can accommodate the heterogeneity incurred by such population genetic processes; we do not discuss phylogenetic methods that ignore such processes, such as concatenation or supermatrix approaches or supertrees. We suggest that methods of data acquisition and the types of markers used in phylogenomics will remain restricted until *a posteriori* methods of marker choice are made possible with routine whole-genome sequencing of taxa of interest. We discuss limitations and potential extensions of a model supporting innovation in phylogenomics today, the multispecies coalescent model (MSC). Macroevolutionary models that use phylogenies, such as character mapping, often ignore the heterogeneity on which building phylogenies increasingly rely and suggest that assimilating such heterogeneity is an important goal moving forward. Finally, we argue that an integrative cyberinfrastructure linking all steps of the process of building the ToL, from specimen acquisition in the field to publication and tracking of phylogenomic data, as well as a culture that values contributors at each step, are essential for progress.

## 1. Introduction

Phylogenomics has been greatly enriched with the introduction of high-throughput sequencing (HTS) and increased breadth of phylogenomic sampling, which have allowed researchers interested in the ToL to scale up in several dimensions and placing both fields squarely in the era of 'big data'. Additionally, conceptual advances and improvements of statistical models used to analyze these data are helping bridge what some have perceived as a gap between phylogenetics and phylogeography (e.g., Felsenstein, 1988; Huson, 2006; Edwards et al., 2016a). However, as datasets become larger, researchers are inevitably faced with a plethora of heterogeneous signals across different genomic regions that often depart from a dichotomously-branching phylogeny (Kunin, Goldovsky & Darzentas, 2005; Jeffroy et al., 2006; Mallet, Besansky & Hahn, 2015). These signals cover an increasingly large array of biological processes at the level of genes and genomes, as well as individual organisms and populations, including processes such as recombination, hybridization, gene flow, and polyploidization; they can be thought of as conflicting, but in truth, they are simply a record of the singular history that we commonly refer to as the ToL (King & Rokas, 2017). One of the grand challenges of evolutionary biology is deciphering this history, whether at the level of genes, populations, species, or genomes, however, currently phylogenomicists have not yet determined how to fully exploit the diverse signals in molecular data.

In this perspective piece, our goal is to highlight opportunities for the field of phylogenomics to overcome conceptual and practical challenges as we navigate our way through the era of big data. We argue that conceptually and analytically embracing heterogeneity generated by population-level processes and associated with different histories across the genome will lead to increased insight into the ToL and its underlying processes. Because major theoretical advances in phylogenomics and population genetics resulting from the advent of genome-scale data have been reviewed elsewhere (e.g., Delsuc, Brinkmann & Philippe, 2005; Degnan & Rosenberg, 2009; Ellegren & Galtier, 2016; Payseur & Rieseberg, 2016; Hahn, 2018), here we focus on advances and challenges to phylogenomics specifically brought about by population genetic processes, which inevitably leads us to focus on the major conceptual framework dealing with such processes, the multispecies coalescent model (MSC). Likewise, advances in other

methods such as supertrees (reviewed by Warnow, 2018), supermatrices (de Queiroz & Gatesy, 2007 and Philippe et al., 2017), and partitioned analyses (e.g., Lanfear et al., 2014; Kainier & Lanfear, 2015) are not central to the objectives of this perspective piece and further details on those topics can be found elsewhere.

A key concept introduced by the scaling up from phylogeography to phylogenomics is the continuum of processes and analytical methods, – the so-called phylogeography-phylogenetics continuum (Edwards et al., 2016a). We argue here that bridging this continuum is critical for advancing phylogenetics. This can be done by either developing phylogenomic approaches that acknowledge and explicitly account for phylogeographic processes, or by determining the regions of parameter space (e.g., branch lengths in tree, level of gene flow) if any, where such within-species processes are not relevant. For example, the choice of markers in a given phylogenomics project is currently guided more by convenience and cost than by evaluating the biological properties and phylogenomic signals in those data; but comparisons of signals across various types of markers (e.g., transcriptomes, noncoding regions) reveal that marker choice is a critical step toward shedding light on the history of populations and unraveling potential processes underlying such history (Rokas et al., 2003; Cutter, 2013; Jarvis et al., 2014; Reddy et al., 2017). On the analysis side, we are in desperate need of methods that can handle the increasingly large data sets being produced by empiricists, but at the same time there is a desire to include increasingly diverse sources of signal in estimates of divergence times, biogeographic history, and models of diversification (Delsuc, Brinkmann & Philippe, 2005; Jeffroy et al., 2006; Kumar et al., 2012). Finding the balance between breadth, depth, and computational feasibility in project design and statistical analysis is crucial for the field today.

Additionally, we discuss ways in which data archiving and integration can benefit access to phylogenomics data and the contributions of phylogenomics to society. Are the priorities that society places on the many disciplines feeding into scientific efforts toward the ToL – fieldwork, museum collections, databases – appropriate for this grand mission? Although we cannot possibly answer all of

these questions within the scope of this perspective, we hope to at least spur discussion on the wide range of field, laboratory, conceptual, and societal issues that allow phylogenomics to move forward.

We first describe the types of genomic data that researchers can generate to perform phylogenomic analyses and how those are more or less suitable for phylogenomic and phylogeographic analyses. We then discuss key concepts around the MSC and highlight the need to expand this model beyond its current limitations. We then discuss how the interplay between phylogenomics and macroevolution might strengthen our understanding of diversity patterns and offer suggestions as to how the community can overcome limitations posed by current methods and models in both fields. Finally, we discuss desired practices that, as a community, phylogenomicists, museum scientists, field biologists, bioinformaticians, and other scientists can embrace toward the goal of assembling the ToL.

## 2. Survey methodology

During the 'Origins of Biodiversity Workshop' organized during May 15–19, 2017 by Chalmers University of Technology and the University of Gothenburg, Sweden, under the auspices of the Gothenburg Centre for Advanced Studies (GoCAS), we gathered scholars and students from several countries and scientific backgrounds to discuss future perspectives in the fields of phylogenomics and phylogeography. We spent one week sharing our recent experiences in these fields and outlining the topics presented here and continued remotely to complete this review. Our goal is not to provide a complete overview of phylogenomic and phylogeographic research, but rather present a number of conceptual and practical aspects that we feel are essential to keep the momentum that these fields have gained during recent times.

## 3. Data generation and data types in phylogenomics

### 3.1. The need for large-scale phylogenomic data

147 One of the fundamental challenges in evolutionary biology is to estimate a ToL for all species. The
148 potential impact of such large phylogenies is reflected in their publication in the highest impact journals,
149 but also in their broad contribution, which extends beyond big data, to methodological innovations, and
150 downstream understanding of macroevolutionary processes (e.g., coalescent methods of species tree
151 inference; accounting for hybridization and unsampled species or localities in datasets; understanding
152 community or genome evolution through large-scale phylogenetics). Hence, the phylogenomics
153 community is now placing a high priority on very large-scale trees, whether in terms of number of taxa,
154 number of genes, or both. The current need for large phylogenies and the high priority placed on them by
155 high-impact journals can also result in short-cuts, wherein extant species lacking any molecular data are
156 often placed in trees based on current taxonomy (Jetz et al., 2012; Zanne et al., 2014; Faurby & Svenning,
157 2015), which can result in conflicts with more robust estimates based on actual data (Brown, Wang &
158 Smith, 2017). At the same time, however, hypothesis-testing in areas such as macroevolution,
159 macroecology, biodiversity, and systematics require these large-scale trees, even as they present
160 challenges being built on high quality data. The phylogenetic knowledge on which we lay a foundation
161 for downstream analyses must be robust, and therefore it is essential that the input phylogenetic
162 hypotheses themselves are robust (Pyron, 2015). Indeed, the current bottlenecks in large-scale
163 phylogenomic data do not appear to be the sequencing, but rather the compilation of high quality, well-
164 curated genomic resources and the availability of adequate software and methods that can fuel
165 phylogenomics for the next century (e.g., Global Genome Initiative, www.mnh.si.edu/ggi/).
166

167 **3.2. Data quality**

168 Genome-scale data in the form of multiple alignments and other homology statements are the foundation
169 of phylogenomics. A major challenge is the difficulty of comprehensive quality checks of data, given that
170 HTS datasets are so large. As researchers collect datasets consisting of thousands of alignments across
171 scores of species and data quality is a serious concern that is left for detection and handling primarily by
172 computer algorithms. In addition to inherent systematic errors in the data (Kocot et al., 2017), several

examples of errors in phylogenomic data sets have been reported in the literature, including the use of unintended paralogous sequences in alignments (e.g., Struck, 2013); mistaking the genome sequence of one species for another (Philippe et al., 2011); and inclusion of genome sequence from parasites into the genome of the host (Kumar et al., 2013). However, the incidence of smaller errors in alignments that are not easily discerned from natural allelic variation, such as base miscalls or misplaced indels, are probably much more widespread than has been reported in the literature. Combined with the sensitivity of some phylogenomic datasets to individual loci or Single Nucleotide Polymorphisms (SNP) within loci (Shen, Hittinger & Rokas, 2017), such errors could have damaging consequences for phylogenomic studies, for the inference of topology and branch lengths of both gene phylogenies and species phylogenies (Marcussen et al., 2014; Bleidorn, 2017). Furthermore, as phylogenomic datasets increase in size, it is likely that the accumulation of errors due to the use of different sequencing chemistries and sequencing depths (Quail et al., 2012; Goodwin, McPherson & McCombie, 2016) will ultimately influence phylogenetic inference. We predict that the impact of these errors will largely depend on the sampling breadth and taxonomic scale of each study, and whether the species phylogeny is a tree or a network.

Sequencing high-quality samples from well-archived voucher specimens is a good first step to increase reproducibility and alleviate issues related to sample identity (Peterson et al., 2007; Pleijel et al., 2008; Chakrabarty, 2010; Turney et al., 2015; Troudet et al., 2018). For individual phylogenomic studies, wholesale manual inspection of every locus is unsustainable (Irisarri et al. 2017), but spot checks of a subset of the data (e.g., 5-10% of the alignments) is a recommended best practice (Phillipe et al., 2011) that is beginning to be encouraged in peer review and in published papers (Montague et al., 2014; Liu et al., 2017). Such checking is important not only for new data generated by a given study, but also for data downloaded from public repositories such as NCBI and Orthomam, which are well known to contain errors (Wesche, Gaffney & Keightley, 2004; Ranwez et al., 2007; Douzery et al., 2014). Because several databases do not include the raw sequence data it is often impossible to evaluate whether oddities may derive from poor sequencing. Robust pipelines for flagging poorly aligned sites or non-homologous sequences, based on existing tools or novel scripts such as Gblocks (Castresana, 2000; Talavera &

199 Castresana, 2007) or TrimAl (Capella-Gutiérrez, Silla-Martínez & Gabaldon, 2009) are gradually being
200 put into practice (Marcussen et al., 2014; He et al., 2016; Irisarri et al., 2017).
201 Coding regions, whether derived from transcriptomes or whole-genome data, are particularly
202 amenable to spot checking of alignments and to filtering out of low-quality data with bioinformatic
203 pipelines (e.g., Dunn et al., 2013; Blom, 2015). Coding regions have the advantage of allowing amino
204 acids to guide alignments, which is particularly useful for highly divergent sequences. Stop codons can
205 help flag errors or genuine pseudogenes. Examining gene tree topologies is also widely used to detect
206 paralogs in phylogenomic data (e.g., Betancur, Naylor & Ortí, 2014). Examining gene trees for aberrantly
207 long branch lengths can also reveal misalignments (e.g., He et al., 2016); sensitivity analyses of methods
208 for indirectly detecting errors in alignments are sorely needed.
209
210 **3.3. Data generation and marker development**
211 *3.3.1. Genome reduction methods*
212 A growing number of genome reduction methods are now providing empiricists with the means to
213 generate genomic subsets suitable for phylogenetic and phylogeographic inference (reviewed by
214 McCormack et al., 2013; Leaché & Oaks 2017; Lemmon & Lemmon, 2013). For phylogenomics, most
215 prominently featured are sequence-capture, focusing on highly conserved regions (e.g., Faircloth et al.,
216 2012; Lemmon, Emme & Lemmon, 2012; reviewed by Jones & Good, 2016) and transcriptomes (e.g.,
217 Misof et al., 2014; Cohen et al., 2016; Fernández et al., 2014; Park et al., 2015; Simion et al., 2017;
218 Irisarri et al., 2017), but phylogenomic trees have also been constructed based on restriction-digest
219 methods that focus on single nucleotide polymorphisms or SNPs (Leaché, Chavez & Jones, 2015; Harvey
220 et al., 2016) and analysis of transposable elements (e.g., Suh, Smeds & Ellegren, 2015). This diversity of
221 marker types for phylogenetics should be celebrated, but each marker type brings with it a list of pros and
222 cons. For example, many questions in the higher level phylogenetics of animals and plants have so far
223 relied almost exclusively on transcriptome data. However, the uncritical use of transcriptomes in
224 phylogenetics is not without caveats. At high taxonomic levels, coding regions can exhibit extreme levels

of among-taxon base composition, sometimes resulting in strong violations of phylogenetic models (Romiguier et al., 2016; Romiguier & Roux, 2017). Coding regions can exhibit reduced levels of incomplete lineage sorting (ILS) compared to noncoding regions (Scally et al., 2012). Such reduced ILS could in fact be helpful in building complex phylogenies with rapid radiations (Edwards, 2009a), but it will certainly distort estimated branch lengths when coalescent methods, which assume neutrality, are used. In addition to yielding abundant sequence variations and SNPs, transcriptome data also yields information on the levels of expression of various genes, and in which tissue-specific genes are expressed. However, using these aspects of transcriptome data are less likely to be of use to phylogeneticists, precisely because specific genes are often tissue-specific and because expression data can exhibit high levels of variation among individuals, populations and species in space and time (Todd, Black & Gemmell, 2016). Such variation will certainly pose limitations for long term phylogenomic endeavors that will likely combine data collected originally for different purposes.

Although non-coding portions of the genome have been largely neglected in phylogenomics because they are difficult to align and analyze, we are now making progress in understanding their sequence evolution and how it might be informative for comparative purposes (Ulitsky, 2016; Edwards, Cloutier & Baker, 2017). For instance, studies on enhancer and promoter evolution in mammals have shown that despite low levels of sequence conservation, there is conservation of regulatory function and 3D structure across species that carries information for comparative purposes (Villar et al., 2015; Berthelot et al., 2018). The development of methods to infer and interpret the evolutionary history and phylogenetic signal of non-coding elements and 3D genome structure is a critical priority.

SNPs have been advocated by some authors for higher level phylogenetics (Leaché & Oaks, 2017), but the available methods for analyzing such data are still extremely limited. For example, concatenation and two coalescent methods (SNAPP and SVD quartets: Bryant et al., 2012; Chifman & Kubatko, 2014) have recently been highlighted as the main methods available for phylogenomic analysis of SNPs (Leaché & Oaks, 2017). But each of these methods has its shortcomings. It is likely that concatenation of SNPs will be misleading for many of the same reasons that concatenating genes can be

misleading, due to different gene histories (see section 'Concepts and models in phylogenomics' for further details; Kubatko and Degnan 2007). SNAPP, a coalescent method suitable for analysis of SNPs (Bryant et al., 2012), works well only on relatively small data sets, and it is unclear how well SVD quartets performs on some data sets (Shi & Yang, 2018). Although SNPs do provide a helpful route around the often-violated assumption in coalescent models of no recombination within loci (Bryant et al., 2012), and are widely seen as excellent markers for phylogeography and population genetics, it remains to be seen how powerful they are at higher phylogenetic levels.

Despite the diversity of marker types for phylogenomics, it remains unclear whether features specific to each marker type can ultimately result in phylogenomic datasets that can strongly mislead. For example, incongruence in the phylogeny of modern birds developed by Jarvis et al. (2014; 48 whole genomes) and Prum et al. (2015; 259 anchored phylogenomics loci, 198 species) has recently been attributed to differences in marker type rather than number of taxa (Reddy et al., 2017). Whereas Jarvis et al. (2014) used primarily noncoding loci because they observed gross incongruence when using coding regions, the loci used by Prum et al. (2015), although nominally focused on broadly "anchored" conserved regions, came primarily from coding regions. Thus, at least one marker type is likely inappropriate when applied across modern birds (ca. 100 Ma). These data type effects can stem from multiple sources. Selection on exons might lead to localized differences in effective population size across the genome, as previous studies have highlighted issues with base composition heterogeneity within exons across taxa (Figuet et al., 2015; Scally et al., 2012). On the other hand, alignment quality of introns and ultraconserved elements (UCEs) can sometimes be less than desired (Edwards, Cloutier & Baker, 2017). Clearly marker effects can potentially have substantial consequences on species tree estimates and need to be further evaluated and compared side-by-side by the phylogenetic community (Shen, Hittinger & Rokas, 2017).

*3.3.2. A priori versus a posteriori selection of loci for phylogenomics*

In an ideal world, phylogeneticists would have whole and fully annotated genomes of all taxa available, allowing them to select loci for phylogenomics based on the relative merits of different loci. This *a posteriori* (after data generation) selection would be carried out after assessing desired phylogenetic and biological properties of a wide array of markers for which the data are already in hand. *A posteriori* selection of loci for phylogenomics is clearly a long-term goal that will yield greater choice and justification for specific loci. Today, the loci for phylogenomics are selected *a priori* (i.e., before data generation) based primarily on cost and ease of collection and alignment, disregarding potentially useful regions of the genome. Thus, an attractive aspect of whole-genome sequencing (WGS) for phylogenomics is to have the opportunity to select markers *a posteriori* once genomes are in hand (e.g., Edwards, Cloutier & Baker 2017; Fig. 1). WGS is less prone to sequencing biases and also allows for further expansion into different research fields and questions based on the same initial data (Lelieveld et al., 2015). In contrast, *a priori* marker selection often limits the kinds of questions and methods that researchers can apply and represents a real constraint for phylogenomics and other disciplines.

An important constraint for using WGS for downstream phylogenomic analyses is genome quality. Obtaining high-coverage well-assembled and thoroughly annotated genomes is still very expensive and time-consuming, and even low coverage genomes are still outside reach for large portions of the scientific community. However, even low coverage genomes, which should be cautiously used for maker selection due to potential problems caused by poor annotation and coverage, can sometimes yield a modest number of markers for phylogenomics, and in the short term might even yield data sets allowing a broader diversity of markers for analysis. Although we are fully aware of its constraints, we are particularly excited about the potential that we see in routinely using WGS to produce phylogenomic data sets.

*3.3.3. More taxa versus more loci*

The question of whether to add more genes or more taxa was a dominant theme in phylogenetics in the 1990s and early 2000s (e.g., Hillis, 1996; Kim, 1996), and remains a persistent question in phylogenomics

today. After much debate in the literature (e.g., Hillis, 1996; Graybeal, 1998; Hillis, 1998; Poe, 1998; Mitchell, Mitter & Regier, 2000), the initial consensus view from the Sanger sequencing era of phylogenetics, is that adding more taxa generally improves phylogenetic analysis more so than more markers (e.g., Hillis, 1996; Graybeal, 1998; Poe, 1998). However, phylogenomics is adding a new twist to this consensus, both from the standpoint of data acquisition and from theory (e.g., Rokas, 2005; Nabhan & Sarkar, 2012; Xi et al., 2012; Patel, Kimball & Braun, 2013). Amassing large data sets, both in terms of more taxa and more loci, is still a guiding principle of phylogenomics. But with the ability now to bring together many different types of markers in a single analysis, and to analyze them in ways that were not previously available, the "more taxa vs. more genes" debate is becoming more nuanced (Nabhan & Sarkar, 2012). For example, recent work shows that this debate can be highly context-specific and model-dependent (e.g., Baurain, Brinkmann & Philippe, 2006; Dell Ampio et al., 2013; Edwards et al., 2016b). Also, some phylogenetic methods, such as coalescent methods, appear to be more robust to limited taxon sampling than traditional methods like concatenation (Song et al., 2012; Liu, Xi & Davis, 2015). Some researchers favor "horizontal" data matrices, wherein the number of loci far exceeds the number of taxa, whereas other researchers favor "vertical" matrices, where many taxa are analyzed at just a few (1-5) loci. Whereas the PCR era of phylogenetics was often dominated by vertical matrices, HTS is allowing data matrices to become more horizontal (Fig. 2). It is important to note that as these more horizontal data become more prevalent, they increase the amount of missing data and aligning problems that can contribute to misleading or low phylogenetic resolution. At least in a coalescent framework, scaling up in both dimensions will be crucial for improved phylogenies and the number of loci required to resolve a given phylogenetic problem is often a function of the coalescent branch lengths in the phylogenetic tree being resolved, with longer branches requiring fewer loci (Edwards et al., 2007; Huang et al., 2010). At deeper time scales, increasing the number of loci have not proven particularly useful to resolve problematic areas in the ToL (e.g.sister group to animals; King & Rokas, 2017). Despite methodological complications in the accurate estimation of population genetic parameters and computational limitations for MCMC convergence at deep times in coalescent-based analyses, recalcitrant nodes likely represent

328 true complexities in the diversification history of these groups and not necessarily reflect failures of
329 coalescent-based phylogenetics (Lanier & Knowles, 2015).
330 To study how researchers have resolved challenges of balancing numbers of taxa versus numbers
331 of loci, we quantified trends in phylogenomic data set size and structure over the past 13 years, drawing
332 data from 164 data sets across diverse taxa (Supplementary Table S1). We found that, whereas the
333 number of species per paper has not increased significantly over time (Fig. 2A), there were significant
334 increases with time in number of loci (Fig. 2B), total length of sequence analyzed (Fig. 2C), as well as
335 total data set size, as measured by the product of species times locus number (Fig. 2E) or species times
336 total alignment length (Fig. 2F). These results mirror similar trends evaluated for the size of data sets in
337 phylogeography (Garrick et al., 2015). Surprisingly, we found no evidence for a tradeoff between the
338 number of species investigated and the number of loci analyzed (Fig. 2D); perhaps HTS data sets have
339 plateaued somewhat in terms of number of loci, whereas the number of species analyzed is more a
340 function of the questions being asked and the clade being investigated. Regardless, we suspect that, in
341 general, the number of loci and total alignment lengths in phylogenomic data sets are likely a function of
342 resources and sequencing effort. The era of whole genome sequencing in phylogenomics is still dawning,
343 given that most studies thus far have used targeted approaches for sampling loci (Supplementary Table
344 S1). We suspect that once whole genome sequencing on a clade-wide basis become routine (e.g., Genome
345 10K Community of Scientists, 2009; Grigoriev et al., 2014; Cheng et al., 2018), we will witness yet
346 another jump in the sizes of phylogenomic data sets.
347
348 *3.3.4. Filtering heterogeneous phylogenomic data sets*
349 Recent studies show that the addition of more loci and more taxa can result in higher levels of gene-tree
350 discordance (e.g., Smith et al., 2015; Shen, Hittinger & Rokas, 2017). This is not unexpected - as the
351 number of taxa and loci increase, the greater the likelihood that the dataset will capture the heterogeneous
352 evolutionary history (e.g., incomplete lineage sorting [ILS], lateral gene transfer [LGT], hybridization,
353 gene duplication and loss [GDL]; See Table 1 for a definition of these concepts), misidentification of

orthologs from paralogs, and patterns of molecular evolution (e.g., noise/lack of signal in the sequences, and nonstationarity in base composition) that can contribute to gene tree discord. At the same time, the variance in gene tree topologies could also have been caused by errors in gene tree estimation. Such observations have been used to argue that the accuracy of gene tree inference should be maximized or at least evaluated, but it is not clear what criteria should be used to filter sets of gene trees. For example, filters can be based on rates of molecular evolution (Klopfstein, Massingham & Goldman, 2017), levels of phylogenetic informativeness (Fong et al., 2012), or on the cause of gene-tree discord itself, if known (Huang et al., 2010). Chen, Liang & Zhang (2015) found that selecting genes whose trees contained a well-known uncontested and long branch in a given species phylogeny (long enough so as not to incur substantial ILS) was a better way to improve phylogenomic signal than selecting genes based on characteristics of sequence evolution. All of these methods are excellent suggestions and should be explored further. Still, the effects of such culling on the distribution of gene trees, and whether it could distort the distribution so that it departs from models like the multispecies coalescent, are unknown, and potentially of concern (but see Huang et al., 2017). Aside from the use of some explicit methods for detection of outlier genes (e.g., de Vienne, Ollier & Aguileta, 2012), rogue taxa (e.g., Aberer, Krompas & Stamatakis, 2013), outlier long branches (e.g., Mai & Mirarab, 2018), and tree space visualization (e.g., Huang et al., 2016; Jombart et al., 2017), an obvious way to alleviate potential effects of gene tree outliers is a more balanced taxon sampling (Hedtke, Townsend & Hillis, 2006). Nonetheless, we need further studies on the effects of different types of phylogenomic filters on the properties of large-scale phylogenomic datasets.

### 3.4. High throughput sequencing opens possibilities for new information and marker types

#### *3.4.1. Heterozygosity and Intra-Individual Site Polymorphisms*

Organisms with multiple ploidy levels are intra-individual polymorphisms and heterozygosity is, of course, common in diploid and polyploid organisms. However, due to issues such as lack of sufficient

read coverage and connectivity, confident identification of such polymorphism continues to be challenging (Garrick, Sunnnucks & Dyer, 2010; Lischer, Excoffier & Heckel, 2014; Schrempff et al., 2016) and many data sets do not permit statistical approaches, such as PHASE (Stephens, Smith & Donnelly, 2001) to robustly determine haplotypes of different alleles (Garrick, Sunnnucks & Dyer, 2010). Consequently, in phylogenetics, heterozygosity and intra-specific polymorphic sites are often accommodated using UIPAC ambiguity codes or ignored entirely or by randomly selecting alleles (Iqbal et al., 2012). In fact, most "one sequence per individual/species" phylogenomic data sets consists of haplotypes that might not occur in nature because many methods, including de novo assemblies of genomes, yield single haplotypes consisting of consensus or other haplotype summaries from diploid organisms. The fact that HTS produces several reads of the same region allows for the identification of heterozygosity and intra-specific polymorphic sites represents an untapped opportunity to incorporate intra-individual variation in our phylogenetic estimates (Lischer, Excoffier & Heckel, 2014; Schrempff et al., 2016; Andermann et al., 2018). Recent models have been proposed to improve calling and sorting such polymorphisms (De Maio, Schlötterer & Kosiol, 2013; Lischer, Excoffier & Heckel, 2014; Potts et al., 2014; Schrempff et al., 2016) and, although results of different studies vary (Kubatko, Gibbs & Bloomquist, 2011; Lischer et al. 2014), estimation of individual, naturally occurring haplotypes has been shown to improve phylogenomic reconstructions based on genome-scale data (Andermann et al., 2018).

*3.4.2. Rare genomic changes*

As noted above, molecular phylogenetics has primarily used alignments of sequence-level data for phylogenetic inference. This bias is perhaps driven by the notion that genome evolution occurs by aggregating small changes, such as point substitutions, over time, but more likely it is due to the challenges of characterizing rare genomic changes, such as indels, transpositions, inversions, and other large-scale genomic events (Rokas & Holland, 2000; Boore 2006; Bleidorn 2017). This emphasis on sequence data has produced a vast ecosystem of algorithms tailored to analyze such data, but most phylogeneticists would agree that rare genomic changes would be a welcome addition to the toolkit of

406 phylogenomics, because they are generally regarded as highly informative markers, providing strong
407 evidence of homology and monophyly (Boore 2006; Rogozin et al., 2008). With the increased availability
408 and affordability of WGS, our view of genome plasticity has changed drastically in recent years and we
409 are now capable of exploring other genomic features beyond the signals encapsulated in DNA or amino
410 acid sequences (e.g., Ryan et al., 2013). The question then arises of how to identify and utilize these rare
411 genomic markers, as well and assess their phylogenetic informativeness (Rokas & Holland, 2000; King &
412 Rokas, 2017). Genome-level characters will likely have different evolutionary properties than sequence-
413 based markers, suggesting that one of the biggest challenges we face for incorporating genomic changes
414 into phylogenetic analyses is to find informative evolutionary models and tools suited for these kinds of
415 data and assess how congruent or discordant they are with respect to other markers (e.g., Rota-Stabelli et
416 al., 2011). This will not only shed light on how phylogenetically informative different genomic changes
417 are, but also will broaden our understanding of the evolutionary intricacies across different genomic
418 regions (Rokas, 2011; Leigh et al., 2011).
419
420 *3.4.3. Gene order and synteny*
421 Computational algorithms to use gene order and rearrangements as markers in phylogenetics (Tang et al.,
422 2004; Ghiurcuta & Moret, 2014; Kowada et al., 2016) were spurred in part by the seminal paper by
423 Boore, Daehler & Brown (1999) using mitochondrial gene rearrangements to understand the phylogeny of
424 arthropods. Initially, algorithms for making use of gene order and synteny were applied primarily to
425 microbial genomes, but recent efforts have extended such methods to the analysis of eukaryotes as well
426 (see Lin et al., 2013). Gene order and synteny appear most promising at high phylogenetic levels,
427 although we still do not know how informative gene order will be at many levels, such as within
428 mammals. Chromosomal rearrangements appear highly dynamic in some groups, such as mammals, and
429 further study of their use in phylogenomics is warranted (Murphy et al., 2005).
430
431 *3.4.4. Indels and transpositions*

Indels and transpositions are two types of molecular characters that are underutilized in phylogenomics, the former perhaps because standard methods of analysis often treat indels as missing data and the latter because they are technically challenging to collect without whole genome data. Indels have been used sporadically in phylogenomics and several have argued for their utility and informativeness, given appropriate analytical tools (Jarvis et al., 2014; Ashkenazy et al., 2014; Roncal et al., 2016). Murphy et al. (2007) used indels in protein-coding regions to bolster estimates of mammalian phylogeny and found that the Atlantogenata hypothesis was supported after scrutinizing proteome-wide indels for spurious alignments and orthology. The Avian Phylogenomics Project found that indels had less homoplasy than SNPs and, despite showing high levels of ILS, was largely congruent with other markers across the avian tree. Transposable elements arguably are even more highly favored by phylogenomics researchers, but are much more difficult to isolate and analyze and have been used principally across various studies in mammals and birds (Kaiser et al., 2007; Churakov et al., 2010; Kriegs et al., 2010; Suh et al., 2011; Baker et al., 2014; Cloutier et al. 2018a). Whereas they are generally considered to have a low rate of homoplasy, most researchers agree that they can in some circumstances exhibit insertional homoplasy. Moreover, no marker is immune to the challenges of ILS, and transposable elements and indels are no exception (Matzke et al., 2012; Suh, Smeds & Ellegren, 2015). Still, the exceptional resolution afforded by some studies employing transposable elements is exciting, and we expect this marker type to increase in use as whole genomes are collected with higher frequency.

*3.4.5. Copy Number Variations (CNV)*

The 1000 Genomes Project estimates that in humans about 20 million base pairs are affected by structural variants, including copy number variations (CNV) and large deletions (1000 Genomes Project Consortium, 2015), suggesting that these types of mutations encompass a higher fraction of the genome than do SNPs in humans. A CNV is a DNA segment of at least one kilobase (kb) that varies in copy number compared with a reference genome (Redon et al., 2006). CNVs appear as deletions, insertions, duplications, and complex multi-site variants (Fredman et al., 2004). Such a profusion of CNVs across

458 human genomes has proven useful in tracking population structure (Sjödin & Jakobsson, 2012), but still
459 remains underappreciated in phylogenetics.
460 Newly available methods allow inference of CNV at high resolution with great accuracy
461 (Wiedenhoeft, Brugel & Schliep, 2016). The frequency with which CNVs occur in animal and plant
462 populations raises the question of how informative they would be at higher phylogenetic levels, and
463 whether they would incur unwanted homoplasy that would obscure homology and phylogenetic
464 relationships. For example, some CNVs evolve so quickly that they can be used with success at the sub-
465 individual level, for example, in tracking clonal evolution of cancer cells using CNV-specific
466 phylogenetic methods (e.g., Schwartz & Schäffer, 2017; Liang, Liao & Zhu, 2017; Ricketts et al., 2018;
467 Urrutia et al., 2018). Moreover, their interspecific variation has been shown to correlate with the
468 phylogeny of some groups, such as the highly pathogenic and rapidly-evolving barley powdery mildews
469 (Frantzeskakis et al., 2018). However, such fast evolution may mean that these markers might be less
470 useful at higher levels of biological organization. Additionally, the adaptive nature of CNVs may or may
471 not facilitate clear phylogenetic signals. For example, a study in *Arabidopsis thaliana* (DeBolt, 2010)
472 showed that adaptation to novel environments, or to varying temperatures, is associated with mutations in
473 CNVs. If CNVs are to become a useful tool in phylogenomics or phylogeography, we must understand
474 their microevolutionary properties in greater detail. For example, the pattern of evolution of CNVs,
475 wherein deletions of genetic material may not easily revert, resulting in a type of Dollo evolution, might
476 help clarify the overall structure of the models applied to them (Rogozin et al., 2006; Gusfield, 2015).
477
478 **3.5. Recent advances in the generation of high throughput sequence data and their impact on the**
479 **reconstruction of the Tree of Life**
480 As sequencing technology rapidly moves forward (reviewed by Goodwin, McPherson & McCombie,
481 2016), our ability to accurately identify the aforementioned marker types increases considerably. For
482 instance, the low per-base error rate of short-read sequencing technologies, such as the Illumina HiSeq X
483 Ten and NovaSeq (Illumina, San Diego, CA), allow for a significant reduction in the cost of sequencing

484 which can result in data for more taxa at a higher coverage. This is certainly beneficial for the accurate
485 identification of SNPs, heterozygosity, and intra-individual site polymorphisms (Goodwin, McPherson &
486 McCombie, 2016) and their use in a phylogenomic context. Moreover, single molecule real-time
487 sequencing technologies, such as Pacific Biosciences (Pacific Biosciences of California, Menlo Park, CA)
488 and Oxford Nanopore (Oxford Nanopore Technologies, ONT) produce reads that exceed 10 kb (Rhoads
489 & Au, 2015; Lu, Giordano & Ning, 2016). The advent of these technologies has led to improved and
490 more efficient assembly methods that allow accurate identification of structural changes (e.g., Khost,
491 Eickbush & Larracuente, 2017; Merker et al., 2018). When combined with data resulting from short-read
492 sequencing, they represent a powerful tool to correctly identify and use a wide array of genomic markers
493 for numerous purposes. These technical advances, which include portable devices that can be carried into
494 the field (i.e., Oxford Nanopore; Johnson et al., 2017), will certainly yield an increase in the genomic loci
495 and taxa available for genomic and phylogenomic studies.
496
497 **4. Concepts and models in phylogenomics**
498 For decades, phylogenetics has struggled with how best to translate evolutionary changes in DNA
499 sequences and other characters into phylogenies, and genomic data are no exception to this trend.
500 Phylogenomics is still in a developing stage of formulating models that effectively represent the
501 underlying mechanisms for genome-scale variation while remaining efficient and within reasonable
502 analytical and bioinformatic capacities. The current focus on models and evolutionary forces generating
503 the patterns that we recover as branching and reticulation events in our phylogenetic reconstructions is a
504 healthy one and can be extended to other important topics in phylogenomics, such as species delimitation,
505 character mapping, and trait evolution (e.g., Yang and Rannala 2014). All of these areas are developing
506 rapidly and are in need of updated models and bioinformatics applications to cope with the heterogeneity
507 brought by genome-scale data.
508
509 **4.1. The multispecies coalescent model**

510 One of the key practical advances in molecular phylogenetics has been the incorporation of gene tree
511 stochasticity into the inference of species phylogenies, via the multispecies coalescent model (MSC:
512 Rannala & Yang, 2003; Liu & Pearl, 2007; Heled & Drummond, 2010). The MSC allows gene trees to be
513 inferred with their own histories, including coalescent-appropriate branching models, but contained
514 within independent yet connected lineages within a species phylogeny, with speciation-appropriate
515 branching models (Degnan & Rosenberg, 2009). The main conceptual advance has been to understand
516 and separately manage the variation at different levels of biological organization – an advance that began
517 years ago (Doyle 1992; Maddison, 1997; Pamilo and Nei 1988), but has only recently been widely
518 embraced and put into practice (Edwards, 2009b). Given its ability to accommodate heterogeneous
519 histories across loci scattered throughout the genome, the MSC lays at the core of the conceptual
520 framework to deal with genome-scale data (e.g., Rannala & Yang, 2008; Liu et al., 2015). In the few
521 instances in which model comparison and fit has been evaluated (Liu & Pearl 2007; Edwards et al. 2007),
522 the MSC vastly outperforms concatenation. This of course does not mean that the MSC is the correct, or
523 even an adequate, model for phylogenomic data (Reid et al., 2014). Despite concerns regarding some of
524 its implementations when dealing with genomic data (e.g., Springer & Gatesy, 2016), the MSC is a
525 powerful theoretical model for phylogenomics and there is room for refinement and improvement for its
526 applications (e.g., Edwards et al., 2016b, Xu & Yang, 2016).
527
528 **4.2. Bypassing full likelihood models by relying on summaries of the coalescent process**
529 Given the huge computational resources required for modelling all the complexities of evolutionary
530 processes in a statistical framework, there is interest in methods that will accommodate genome-scale data
531 for large numbers of species within a coalescent framework. The utility of such methods cannot be
532 overstated: the rapid rise of large-scale genomic data sets has clearly outstripped theoretical and
533 computational methods required to analyze them. For example, although progress is being made
534 regarding scalability of full Bayesian methods of species phylogeny inference (e.g., Ogilvie, Bouckaert &
535 Drummond, 2017), they are still unable to accommodate large phylogenomic datasets, which often consist

of hundreds of species for thousands of loci (Supplementary Table S1). A common approach to speeding up species phylogeny inference consists of 'two-step' methods, wherein gene trees are estimated first and separately from the species phylogeny; then, using various summaries of the coalescent process for collections of gene trees, a species phylogeny is estimated. Many useful methods for estimating species phylogenies in this way have been proposed (see Marcussen et al., 2014; Liu, Wu & Yu, 2015; Mirarab & Warnow, 2015; Mirarab, Bayzid & Warnow, 2016), taking advantage of summary statistics of the coalescent process, such as the average ranks of pairs of species in the collection of gene trees (e.g., STAR: Liu et al., 2009; ASTRAL-II: Mirarab & Warnow, 2015) or the distribution of gene trees containing triplets of species (e.g., MP-EST; Liu, Yu & Edwards, 2010). Some of these two-step methods, while approximate, nonetheless allow for statistical testing in a likelihood framework. For example, MP-EST can evaluate the (pseudo)likelihood of two proposed species phylogenies given a collection of gene trees and the difference in likelihood can be used to evaluate two proposed species phylogenies against each other. However, such statistical approaches have rarely been used thus far, and bootstrapping or approximate posterior probabilities on branches are by far the most common statistics applied to species phylogenies (Sayyari & Mirarab, 2016). Speeding up the estimation process using two-step methods can be effective, but it can also accumulate errors or misallocate sources of variance which cannot be corrected at later stages (Xu & Yang, 2016). If gene trees are biased or uninformative, then downstream analyses for species phylogeny estimation or species delimitation may similarly be compromised (e.g., Olave, Sola & Knowles, 2014). For example, MP-EST can sometimes perform poorly when PhyML (Guindon et al., 2010) is used to build low-information gene trees because PhyML may produce biased gene trees when the alignments contain very similar sequences (Xi, Liu & Davis, 2015). This may account for the lower performance of MP-EST compared to ASTRAL in some simulation conditions, because ASTRAL resolves input polytomies and zero-length branches in gene trees more appropriately. This difference between MP-EST and ASTRAL is eliminated when RAxML (Stamatakis, 2014) is used to build gene trees (Xi, Liu & Davis, 2015).

562 **4.3. Beyond the multispecies coalescent model**

563 *4.3.1. Reticulation at multiple levels challenges the standard multispecies coalescent model*

564 The phylogenetic processes of branching and reticulation can operate at several levels of organization,
565 including within genes, within genomes, and within populations or species (Figs. 3, 4 and 5). For
566 example, recombination can cause reticulations within genes, allopolyploidization can cause reticulations
567 at the level of whole genomes, and introgression and hybridization can cause reticulations at the level of
568 populations. These levels are nested so that branching processes (and in part reticulations) acting at a
569 higher level will cause correlated branching patterns at lower levels. At the same time, reticulations at
570 lower levels, such as recombination acting within genes, will cause inference problems at higher levels,
571 such as estimating population histories. Crucially, however, it is only recombination that will break one
572 key element driving many recent models of phylogenetics and population histories, namely dichotomous
573 gene trees. Reticulations at levels of organization higher than the genome, such as the fusing of
574 populations, as well as gene duplication, will still yield collections of dichotomous gene trees, even if the
575 higher-level history is reticulated. Ultimately, the additive effects of these reticulate processes result in
576 our observed phylogenetic reconstructions, and we expect all of these scenarios to produce bifurcating,
577 dichotomous gene trees. From a modelling point of view, another key distinction is whether at the species
578 level, we still have a phylogeny that is tree-like, or whether a network is needed. The process whereby
579 two populations jointly produce a third requires a network to model properly. Allopolyploidy is another
580 situation requiring a network. There are several statistical methods for inferring homoploid networks (Yu
581 et al., 2014; Solís-Lemus & Ané, 2016; Wen, Yu & Nakhleh, 2016; Wen & Nakhleh, 2018), species
582 histories under allopolyploidy (Jones, Sagitov & Oxelman, 2013), and some two-step methods such as
583 PADRE (Huber et al., 2006; Lott et al., 2009). In general, dealing with multiple simultaneous violations
584 of the MSC, such as introgression and allopolyploidy, remains challenging (Degnan, 2018). It is likely
585 that the history of many radiations involves parts of the genome with a dichotomous history and parts that
586 exhibit reticulation, demanding methods that accommodate both scenarios. Alternatively, rather than
587 trying to accommodate multiple processes in our methods for phylogenetic inference, we might instead

focus our attention on subsets of loci that would not violate the MSC (e.g., Knowles et al., 2018). In cases where processes other than incomplete lineage sorting contribute to gene tree discord (i.e., the distribution of trees is statistically inconsistent with expectations under the MSC; see Smith et al., 2015), loci consistent with the MSC model can be identified (e.g., separated from loci with horizontal gene transfer) using CLASSIPHY (Huang et al., 2017). It has also been suggested that in order to distinguish violations of the MSC, the MSC can be used as a null model to be compared with increasingly complex models that would invoke processes such as hybridization and recombination using networks (Degnan, 2018). To follow this promising approach, further research must be conducted to not only model specific processes, but also distinguish them.

Models accommodating dichotomous divergence with gene flow are somewhat limited. For example, in IMa2 (Hey & Nielsen, 2004; Hey & Nielsen, 2007; Hey, 2010) the species phylogeny must be known and fairly small; in the method of Dalquen, Zhu & Yang (2017), both the species phylogeny and gene trees are restricted to three tips. Looking forward, it may be useful to deal with two sub-problems: The first is estimating the species phylogeny despite migration, for example by identifying which loci are interfering with the species phylogeny inference or causing reticulations in the form of gene flow. The second sub-problem is to incorporate a gradual speciation process (Fig. 6), where gene flow after speciation slowly declines, perhaps according to some simple function like an exponential. Such a model would capture what is thought to be a more common speciation process than the instantaneous process modelled by the MSC (Jones, 2018).

In some cases, it is possible to model one situation with either a species network or a tree with gene flow. Long (1991) discussed two models of admixture: Intermixture and gene flow (Fig. 7). The phylogenetics community has mainly focused on methods for inference under the intermixture model (e.g., the multispecies network coalescent; Yu et al., 2014, whereas the population genetics community has focused more on models including gene flow (e.g., IM, admixture graphs, G-PhoCS, Phrapl). While some initial work to test inference based on one of these models on data generated by the other has recently appeared (Wen & Nakhleh, 2018; Solís-Lemus et al., 2017; Blischak et al., 2018; Zhang et al.,

614 2018), much more work is needed to bring together these two lines of work. Simulations and comparisons
615 of observed and expected summary statistics, such as the site-frequency spectrum (Excoffier et al., 2013),
616 have proven especially useful in distinguishing such scenarios (Fig. 7).
617 Reticulation in the form of gene flow or introgression is probably the most difficult violation of
618 the MSC to address (but see Hibbins & Hahn, 2018 for a model that estimates the timing and direction of
619 introgression based on the multispecies network coalescent), in part because the number of potential trees
620 accommodating a reticulating network is even higher than the already high number of trees for a given
621 number of taxa. There is at least one issue where reticulation presents an opportunity as well as a
622 challenge. Any kind of gene flow/hybridization means that there is the possibility of inferring the
623 existence of extinct species, because extinct species contribute novel alleles that exceed the coalescence
624 time of most alleles in the focal species under study (Hammer et al., 2011). Well-known examples are the
625 documented presence of Neanderthal genes in most human genomes due to introgression (e.g., Meyer et
626 al., 2012) and the presence of genomes derived from now-extinct diploids in extant allopolyploids (i.e.
627 meso-allopolyploids; e.g., Mandáková et al., 2010; Marcussen et al., 2015). Some current models can
628 explain the data as containing genetic information from extinct species, but they do not model the full
629 species phylogeny: such a generalized approach seems a promising avenue to explore.
630
631 *4.3.2. Polyploidy and the challenges of analyzing gene duplication and loss*
632 The MSC model describes well allelic lineages and the mutations they accumulate (Fig. 3; Degnan &
633 Rosenberg, 2009; Liu, Xi & Davis, 2015). The simple MSC model is challenging to apply to evolutionary
634 events in which the evolving entities (genes or paralogs) duplicate and occasionally go extinct during the
635 evolutionary history of the populations/species and thus cannot be sampled in contemporary population or
636 species. Estimating the existence and number of these “ghost” lineages remains challenging. For example,
637 how can we detect duplication events if one of the duplicated loci is lost in descendant lineages? In the
638 case of polyploidy, two (or more) genomes having separate evolutionary histories end up together in a

single individual. What consequences for evolutionary history do genomic conflicts and dosage variation in gene expression impose? Polyploidy also raises technical issues, such as whether or not homoeologous sequences are recovered in standard genomic surveys.

The complication that gene duplication and loss (GDL) brings to the inference of species phylogenies has long been recognized (Fitch, 1970). It is therefore surprising that practical solutions to the problem of GDL are almost non-existent, with empirical examples usually based on *ad hoc* methods and deductions. Ancient duplications where most additional copies are retained in descendent species can be fairly easy to diagnose based on phylogeny (Oxelman et al., 2004; Pfeil et al., 2004). However, resolving duplications becomes more difficult when copy number changes quickly (Ashfield et al., 2012), or when duplications are recent and copy loss is complete or nearly so, thus returning the locus to a single-copy state (Ramadugu et al., 2013). In the latter case, the phylogenetic pattern can mimic that of ILS and become indistinguishable from it (Sousa et al., 2017), generally leaving no trace at all of the loss.

Why is GDL so challenging to implement in theory? The topological and coalescent-time similarities between ILS and GDL complicates extending the MSC to include both processes, unless copy number exceeds one in at least some samples (Fig. 4). Assuming that allelic and homoeologous variation is not confused with the copy number of independently duplicated genes, at the very least, duplicated genes could be handled as independent loci with missing data for some samples with MSC inference. When copy loss is complete, or when the duplication is so recent so as to conflate allelic versus copy variation, these GDL loci have little effect on species phylogeny inference and divergence times, especially if the algorithms used employ averages over coalescence times or other parameters across many gene trees (Liu et al., 2009; Sousa et al., 2017). At high proportions, though, they may cause serious issues for phylogenetic reconstruction, because the unexpected positions of gene duplications in a species phylogeny, coupled with random copy loss, means that no specific pattern is expected among the affected gene trees (Fig. 4). This scenario contrasts with the retention of ancestral polymorphisms, where we know that branches in short species phylogenies (in coalescent units) are the cause (Rosenberg & Nordborg, 2002). Thus, we expect deeply coalescing lineages to occur in specific parts of a species phylogeny with a

665 limited number of topological outcomes and branch lengths limited by effective population size, which is
666 not the case for duplicated genes. A recent approach to identifying genes that are single copy, but have
667 nonetheless been affected by GDL, was made using the genomic location of the loci (Sousa et al., 2017),
668 and could prove useful for distinguishing GDL and ILS.
669
670 *4.3.3. Recombination*
671 All existing methods for coalescent estimation of species trees and networks make two important
672 assumptions, namely that (1) there is free recombination between loci, and (2) there is no recombination
673 within a locus. These two assumptions address a key concept distinguishing MSC models from
674 concatenation or supermatrix models: it is the conditional independence of loci, mediated by
675 recombination between loci, and not the ability to address ILS or discordance among genes *per se*.
676 Moving forward, three important questions to address are: (1) How robust are methods to the presence of
677 recombination within loci and/or to the violation of independence among loci? (2) How should we model
678 recombination within the species phylogeny inference framework? and (3) How do we detect it and
679 differentiate recombination-free loci?
680 Researchers have started to examine the first question and found a detectable effect of
681 recombination only under extreme levels of ILS and gene tree heterogeneity (e.g., Lanier & Knowles,
682 2012). However, more analyses and studies are still needed to explore a wider range of factors and
683 parameters that could affect phylogenetic inference when the assumption of recombination-free loci is
684 violated (e.g., Li et al., 2018). For answering the second question, one approach involves combining the
685 multispecies coalescent with hidden Markov models (e.g., Hobolth et al., 2007). These methods suffer
686 from the "state explosion problem", where individual states are needed for the different coalescent
687 histories, and they increase rapidly with the number of taxa in the dataset, making them infeasible except
688 for very small (~4 taxa) datasets. New methods that scale to larger datasets are needed if such approaches
689 are to be useful in practice. A different direction is to devise novel methods for inferring species

690 phylogeny while assuming that the genealogies of the individual loci could take the form of an ancestral
691 recombination graph (ARG: Griffiths & Marjoram, 1996; Siepel, 2009; Rasmussen et al., 2014).
692 Extending these approaches to address recombination would require the development of new
693 models that significantly extend the multispecies coalescent to account for ARGs within the branches of a
694 species phylogeny. For two-step species tree methods, this entails developing new methods that infer
695 ARGs for the individual loci and methods that infer species phylogenies from collections of ARGs. For
696 single-step Bayesian methods, novel developments are needed to sample species phylogenies, locus-
697 specific ARGs, and their related parameters. It will also be important to better understand the conditions
698 under which ignoring recombination will still yield reasonable estimates of phylogeny. Extending the
699 theory to accommodate ARGs may be of intrinsic interest, but if the parameter space in which
700 recombination is relevant is very small, then practitioners may be able to ignore recombination.
701
702 **4.4. Species concepts and delimitation**
703 Coalescent methods have played an important role in the development and critical evaluation of species
704 delimitation methods because they provide hypotheses for species boundaries based on genetic data and
705 be integrated with phenotypic data (e.g., Solís-Lemus et al., 2015). Irrespective of traditional species
706 concepts, it is essential that the entities at the tips of the species tree do not violate the assumptions of the
707 MSC, wherein species are defined mathematically (e.g., Rannala & Yang, 2003, Degnan & Rosenberg,
708 2009): the branches of the species tree constitute species or populations that do not exchange genes.
709 However, the MSC model also carries strict assumptions about the divergence process if the delimited
710 units are to be interpreted as species. Specifically, it is important to emphasize that in the “standard” MSC
711 model, these species represent populations that, immediately after divergence, no longer experience gene
712 flow. Therefore, the species of the MSC model do not necessarily correspond with species as a taxonomic
713 rank, defined by traditional species concepts (Heled & Drummond, 2010): “MSC” species could simply
714 be populations by other criteria, so long as they have ceased to exchange genes, even for a short period of
715 time. In other words, a species tree built under the MSC might then be interpreted as a depiction of the

history of the barriers to gene flow among diverging structured populations and this may be particularly true when there is dense spatial and genomic sampling across individuals (Sukumaran & Knowles, 2017). Therefore, in those species tree methods requiring *a priori* assignments of individuals to species, such assignments may strongly influence the inferred species phylogeny, in the same way that hybridization will have serious consequences on an estimated species phylogeny (Leaché et al., 2014).

Recently, several MSC-based methods that have the ability to simultaneously perform species delimitation and estimate the species phylogenies have been developed and implemented (e.g., Yang & Rannala, 2014; Jones, Aydin, & Oxelman, 2015; Jones, 2017). These methods seem to consistently recover the correct number of "MSC species" given the assumptions of the model. However, it is probable that the assumption of no gene flow between the descendant populations is often violated and that most reproductive isolation processes are gradual or episodic rather than sudden and permanent (e.g., Rosindell et al., 2010). There is thus need for methods that perform simultaneous species phylogeny estimation and assignment of individuals to species while considering the limitations of the MSC (Jones, Aydin, & Oxelman, 2015).

If one prefers a species concept that affirms that most recently diverged populations are necessarily reproductively isolated, current methods will overestimate the number of species as defined by traditional species concepts and will likely reveal instead intraspecific population structure (Sukumaran & Knowles, 2017). Toprak et al. (2016) used DISSECT (Jones, Aydin & Oxelman, 2015) but also employed checks as to the integrity of various hypotheses of species boundaries suggested by the data. From a computational point of view, any species delimitation method will need an operational definition of species. Therefore, a possible development of MSC-based species delimitation methods could be allowing migration and assuming that speciation is complete when a certain proportion of the migrations is reached or when the migration rate is sufficiently low. However, this solution will not be suited for the protracted speciation model because other kinds of information besides the movement of genes will still be needed to identify when a clade becomes reproductively isolated. Possibly the best way to avoid confusion is to restrict the word "species" to taxonomy and base it on multiple sources of

742 information which are synthesized in an integrative fashion (Dayrat, 2005; Will, Mishler & Wheeler,
743 2005; Bacon et al., 2012; Solís-Lemus, Knowles & Ané, 2015), and refer to the reproductively isolated
744 units of MSC analysis as “MSC units” or “MSC taxa”.
745
746 **5. Models at the intersection of phylogenomics and macroevolution**
747 At the intersection of phylogenomics and macroevolution, phylogeneticists aim at shedding light on how
748 patterns of organismal diversity have been generated and maintained through time. The former focuses
749 primarily on building phylogenies, whereas the latter uses them to study the tempo and mode of
750 diversification over time. In many important respects, these two sub-disciplines have remained distinct
751 and non-communicative. On the one hand, phylogenomics and phylogeography have not exhaustively
752 aimed to address the type of questions - related to diversification and trait evolution - that macroevolution
753 focuses on. On the other hand, macroevolution ignores many kinds of complexities inherent to the
754 phylogeny building process that phylogenomics has recently begun to address.
755 Macroevolutionary models focus on long-term processes, in terms of both species richness and
756 phenotypic diversity. They rely on two types of models: birth-death models of diversification aimed at
757 understanding how and why speciation and extinction rates vary through time and across lineages (Hey
758 1992; Nee, Mooers & Harvey, 1992; see Stadler, 2013 and Morlon, 2014 for review) and models of trait
759 evolution aimed at understanding the mode and tempo of phenotypic evolution (Felsenstein, 1973; see
760 Pennell & Harmon, 2013 and Manceau, Lambert & Morlon, 2017 for reviews). These models are
761 typically constructed at the level of species, ignoring the populations or individuals that constitute these
762 species (but see Manceau, Lambert & Morlon, 2015 and Rosindell, Harmon & Etienne, 2015 for
763 exceptions). As a consequence, microevolutionary processes such as coalescence have informed
764 phylogenetic methods for building phylogenies more so than have macroevolutionary methods that use
765 them. For example, the most widely used phylogenetic dating methods generally do not acknowledge the
766 critical distinction between speciation times, which are usually of primary interest, and coalescence times,
767 which are often assumed to represent speciation times but in fact represent events older than the

768 divergence of the species concerned (Edwards & Beerli, 2000; dos Reis, Donoghue & Yang, 2016;
769 Angelis & dos Reis 2015). In addition, macroevolutionary models are fit to species phylogenies
770 (diversification models) or a combination of species phylogenies and phenotypic data (trait evolution
771 models), most often assuming that evolution is best represented by a species tree, not a network (but see
772 Jhwueng & O'Meara, 2015; Bastide et al., 2018; Solís-Lemus et al., 2017 for models of trait evolution on
773 networks), and that the species phylogeny is known. Nearly all models that use phylogenies to study
774 character evolution assume a single underlying species phylogeny on which characters evolve. But it has
775 become evident recently that different characters in principle have different phylogenies, for the same
776 reason that genes themselves might have different phylogenies (Hahn & Nahkleh, 2016). Analyzing
777 incongruences between character evolution inferred from the species tree versus from gene trees that are
778 more directly linked to the character under study would provide a refined understanding of character
779 evolution. Recent work on the phylogeny of quantitative characters may be helpful in this endeavor
780 (Felsenstein, 2012).

781 Developing research projects that integrate the heterogeneity inherent in phylogenomics and
782 macroevolution will bring important new insights into the evolutionary process. For example, developing
783 diversification and phenotypic evolution models to be fit to networks rather than dichotomous trees will
784 allow estimates of rates of hybrid speciation and phenotypic evolution as well as a better understanding of
785 factors influencing such rates (Bastide et al., 2018). Embracing genetic heterogeneity and the
786 incongruence between gene trees and species phylogenies when applying macroevolutionary models
787 could help us to better understand how speciation proceeds, and also to analyze the coupling between
788 genetic and phenotypic evolution (e.g., is phenotypic convergence coupled or not with genetic
789 convergence in relevant genes?). Developing macroevolutionary models accounting for within-species
790 heterogeneity linked to biogeography could help us understand how biogeographic structuring influences
791 speciation, extinction, and phenotypic evolution.

792 More generally, evolutionary biologists have yet to thoroughly explore the type of new questions
793 that we are going to be able to address if we are given genomic data at the tips of all species from a

794 phylogeny. Such data could allow us to gain an integrative understanding of three fundamental aspects of
795 evolution: evolution at the molecular level, at the phenotypic level, and at the clade level, as well as the
796 links among them. Are rates of evolution at these three levels correlated? If so, how? Do features of
797 genomes or of genome evolution, such as quantity of transposable elements, substitution rates, number of
798 gene duplications, influence rates of diversification and phenotypic evolution? Clearly, we are only at the
799 beginning of exploring these new possibilities.
800
801 **5.1. Mapping trait evolution on heterogeneous genomic datasets**
802 Mapping the genomic basis of phenotypic traits is a major trend in evolutionary biology today (Elmer &
803 Meyer, 2011; Hoban et al., 2016). Such mapping can be conducted in the context of populations of a
804 single species or, increasingly, via comparisons of species on a phylogeny (e.g., Hiller et al., 2012;
805 Marcovitz, Jia & Bejerano, 2016). Phylogenetic genome-wide association studies ("PhyloGWAS")
806 methods identify genomic features in coding or non-coding DNA that exhibit unusual patterns of
807 evolution on branches concerned with repeated evolution of phenotypes, thereby drawing connections
808 between the genomic and phenotypic levels (Pease et al., 2016). Such phylogenomic mapping usually
809 assumes a single phylogeny, the species phylogeny, as a framework for analysis, and therefore ignores
810 genomic heterogeneity. To make phyloGWAS mapping most efficient it might be more appropriate to use
811 the local topology in the genome for inference and estimation of ancestral states. Estimating genotype-
812 phenotype associations solely on the species phylogeny might yield misleading results regarding the
813 origin and evolution of phenotypic traits (Hahn & Nakleh, 2016; Mendes, Hahn & Hahn, 2016; Guerrero
814 & Hahn, 2018). Heterogeneity across gene histories has been traditionally considered as "biological
815 noise" when using comparative genomics to map traits, but of course such heterogeneity is the focus of
816 gene mapping efforts at lower taxonomic levels. A recently proposed application of the MSC for
817 quantitative traits, accounting for genealogical heterogeneity improves downstream estimates of mean
818 trait values, phylogenetic signal, and evolutionary rates of traits (Mendes et al., 2018). Furthermore,
819 genome-wide or gene-specific selective sweeps associated with the evolution of a particular phenotypic

820 trait are a major source of genetic heterogeneity among closely related populations or species, and can be
821 captured using outlier statistics, such as Fst or Dxy (Pease et al., 2016). Such selective sweep mapping of
822 genes with large phenotypic effect can now be accomplished with high resolution and precision in
823 genomically poorly studied organisms (Lamichhaney et al., 2015). Apart from providing valuable
824 knowledge on the genetic basis of trait diversification, such data are providing increasing support to the
825 fact that cases of genetic heterogeneity can be profitably used in the effort to understand and resolve
826 evolutionary history, rather than considering it "biological noise." Such thinking needs to be incorporated
827 into comparative genomics more frequently (e.g., Mendes et al., 2018).
828
829 **5.2. Tree-free methods of character evolution**
830 We have seen that incorporating phylogenetic heterogeneity is a challenge for macroevolutionary models
831 of character evolution. At the other end of the spectrum are a class of methods (so called "tree-free
832 methods") that attempt to draw inferences and principles about trait evolution without assuming a
833 particular phylogeny. The common situation when analyzing character or trait data correlated by a
834 phylogeny is to assume a stochastic process for the trait, commonly a variation of the Brownian motion
835 (BM; Felsenstein, 1985) or Ornstein-Uhlenbeck (OU; Hansen, 1997) processes. Then, using the estimated
836 phylogeny and measured trait data for each species, the parameters of various evolutionary processes –
837 trait variation, patterns and rates of change, etc. - are estimated, often using maximum-likelihood or
838 Bayesian approaches (see Pennell & Harmon, 2013 and Manceau, Lambert & Morlon, 2017 for reviews).
839 However, given the various logistical and technical challenges of inferring robust phylogenies, exploring
840 tree-free methods might represent a useful mechanism for guiding the study of character evolution for
841 certain groups.
842 Tree-free comparative methods work by integrating over the space of trees (under a given
843 branching process model). For example, under a pure birth model and with enough tip measurements, the
844 optimum value of the OU process can be estimated as the sample average (Bartoszek & Sagitov, 2015a).
845 Similar results have now been derived for other models of tree growth that include extinction (Adamczak

and Miłoś, 2014; 2015; Ané, Ho & Roch, 2017). Similarly, the rate of adaptation under the OU process, often modeled as the stationary variance – the ratio of the squared "rate of evolution" (sigma parameter in the OU model) and twice the "rate of adaptation" (the alpha parameter) can be estimated as the sample variance (Bartoszek & Sagitov, 2015a). Teasing sigma and alpha apart, however, requires a tree. The key parameter of the BM model, the rate of evolution, is similarly estimable directly from the trait sample (Bartoszek & Sagitov, 2015b; Crawford & Suchard, 2013), whereas the root state cannot be consistently estimated without a tree (Ané, 2008; Sagitov & Bartoszek, 2012). In addition to providing tree-free estimators of some model parameters, the studies mentioned above also derived Central Limit Theorems that allow computing confidence intervals around these point estimates as well as the sample sizes needed to obtain reliable estimates.

**5.3. Extinct and unsampled species**

A notable case when phylogenomics and macroevolution do meet is in the treatment of extinct or unsampled species in phylogenetic reconstruction and dating. Despite the avalanche of genomic data for an increasing number of species, we still lack sequence data for most species, making it difficult to place them in a phylogeny. Some researchers (e.g., Jetz et al., 2012; Tonini et al., 2016) have used a combination of consensus trees and current taxonomic classifications to impute the phylogenetic relationships of unsampled species. In this case, polytomies are often resolved by using distributions of branching times obtained from macroevolutionary birth-death models (Kuhn, Mooers & Thomas, 2011). While we appreciate the value of these approaches given the real logistic difficulties researchers face as they attempt to obtain samples from around the globe and that methods are now easily applicable, such approaches have generated well-founded concerns about biases in our inferences (Davies et al., 2012; Rabosky, 2015) and the extent of these biases remain largely unknown (Pennell, FitzJohn, & Cornwell, 2016). Nonetheless, there are methods that can objectively alleviate issues arising from using incomplete phylogenies in downstream phylogenetic comparative analyses. For example, recent results using conditioned birth-death processes (e.g., Gernhard, 2008a; b; Sagitov & Bartoszek, 2012) show that under

872 constant rate processes the size of the clade contributes information on the height of the tree and also on
873 the coalescence times. Such results can be used to improve the calibration and node dating of the
874 phylogeny when some species are not sampled. One would expect that ignoring the non-sequenced
875 species would incur a bias resulting in shorter tree heights, because less time is usually required to
876 generate fewer tips. Conditioned branching process models can help alleviate this bias. Also,
877 macroevolutionary birth-death models are used as branching process priors in Bayesian molecular dating.
878 The availability of likelihood expressions for incompletely sampled phylogenies (Stadler 2009; Stadler &
879 Steel 2012; Morlon, Parsons & Plotkin, 2011) thus allow to date phylogenies while accounting for the
880 fact that we have observed only a certain fraction of unsampled species.
881 An important source of genetic data for extinct and unsampled species comes from ancient DNA.
882 Current methods now allow sequencing ancestral DNA and incorporating it into phylogenomic,
883 phylogeographic, and population genetic analyses (reviewed by Leonardi et al., 2017), and even obtaining
884 whole-genomes from various samples dating back up to a few thousand years (e.g., Lynch et al., 2015;
885 Cloutier et al., 2018b). The paleontological and paleobiological records represent another source of
886 extinct species for which genomic data is out of reach. Having the opportunity of integrating the temporal
887 and phylogenetic information of extinct data provided by fossils into phylogenomic estimates of extant
888 diversity is a crucial task toward a complete ToL (reviewed by Hunt & Slater, 2016). Besides the
889 possibility of placing fossil data into phylogenetic trees, a growing field of development is that of time-
890 scaling phylogenies using fossil data. Traditionally, node-dating methods have used fossils to bound the
891 minimum age of nodes at the base of extant clades in which fossils likely belong based on morphological
892 comparisons (reviewed by Ksepka et al., 2011) or estimated fossil ages have been used to post-hoc-scale
893 branch lengths (Sanderson, 1992; Smith & O'Meara, 2012; Bapst, 2014). However, there is an increasing
894 interest in simultaneously performing phylogenetic inference and time-scaling phylogenies (Pyron, 2011;
895 Ronquist et al., 2012; Heath, Huelsenbeck & Stadler, 2014). These methods, known as tip-dating, have
896 the advantage of accommodating phylogenetic uncertainty implemented in a Bayesian model-based
897 framework (Hunt & Slater, 2016) and have been implemented in several groups spanning a wide temporal

scale (e.g., Pyron, 2011; Pyron & Burbrink, 2012; Ronquist et al., 2012; Heath, Huelsenbeck & Stadler, 2014; Slater, 2015; Larson-Johnson, 2016; Matzke & Wright, 2016). Lastly, these phylogenies containing both extinct and extant diversity serve as framework to conduct robust diversification and trait evolution studies (Hunt, 2013; Pennell & Harmon, 2013; Pyron, 2015).

## 6. Building, updating and sustaining the Tree of Life

### 6.1. Scalability challenges

Inferring the phylogeny of all living organisms represents a different challenge than inferring the relationships of just a few terminals; often the scale at which new methods are developed and tested is on this latter scale. For instance, for eukaryotes alone, recent conservative estimates indicate that there are ~8.7 million species on Earth and only 9-14% of them have been formally described (Mora et al., 2011). Furthermore, out of 2.6 million taxa currently represented in the Open Tree of Life (https://tree.opentreeoflife.org; Hinchliff et al., 2015), only ~55,000 were gathered from hard-data phylogenies, whereas phylogenetic affinities of the rest were inferred from current taxonomic classifications (McTavish et al., 2015; 2017). These observations suggest that the vast majority of taxa on Earth still await formal taxonomic description and placement in the ToL (Mora et al., 2011; McTavish et al., 2017). As mentioned above, one common challenge that phylogeneticists encounter is the difficulty in accessing samples from rare, endangered, or extinct taxa, particularly in countries where collecting and exporting is difficult. Recent genomic techniques now allow successful results in obtaining valuable ancient DNA data from museum specimens (e.g., Staats et al., 2013; Hykin, Bi & McGuire, 2015; McCormack, Tsai & Faircloth, 2016; McCormack et al., 2017; Ruane & Austin, 2017), and here, we advocate for routine use of these resources to improve taxon sampling, enhance research in phylogenomics and phylogeography, and increase awareness and usefulness of natural history collections.

Despite the great increase in the generation of genomic data across organisms, we are often limited to fast and efficient but simpler, less realistic phylogenetic methods and assumptions, such as IQ-Tree (Nguyen et al., 2015), PhyloBayes (Lartillot et al., 2013), and ExaML (Kozlov, Aberer &

924 Stamatakis, 2015), to deal with large, heterogeneous datasets. The main downside of these options is that
925 they are not a full coalescent framework and thus far rely on data concatenation. Fully coalescent
926 methods, such as the popular phylogenetic software program *BEAST (Heled & Drummond, 2010), are
927 not capable of dealing with more than a few hundred taxa and some dozen loci at a time for a common
928 analysis, and only recently the release of StarBeast2 allows for the use of thousands of loci for tens of
929 taxa (Ogilvie, Bouckaert & Drummond, 2017). To tackle this problem, we encourage the continuing
930 development of methods that are fully scalable and ideally only increase analytical time linearly rather
931 than exponentially with the number of taxa and loci. Phylogenetic methods should also be fully
932 parallelizable (in order to run natively in computer clusters) and contain checkpoints, i.e., be able to
933 resume the analyses from the latest logged file in case an analysis crashes or the user wishes to evaluate
934 partial results. Another point of possible improvement is in dealing with new sequences to be added to a
935 previously large dataset: should the analysis start from scratch, or could there be substantial time gains by
936 letting those sequences find their placement in the phylogeny 'on the fly' (e.g., Siu-Ting et al., 2014).
937 Large scale phylogenies should ideally be based on the best (or most comprehensive) available
938 datasets in terms of taxonomic and molecular sampling and be constructed from the data itself. However,
939 even supermatrix inference conducted under a single analysis can add bias on tree heights and
940 coalescence times when performed across unbalanced sampled clades (a very common case for species-
941 rich clades or understudied taxa), and therefore affect downstream analyses that rely on these parameters
942 (e.g., biogeography, trait evolution, diversification rates). Computing optimally populated datasets that
943 combine the largest number of taxa and loci simultaneously is a complex mathematical problem, but
944 recent approaches (e.g., SUPERSMART – Antonelli et al., 2017; PyPHLAWD – Smith & Brown, 2018)
945 attempt to overcome it objectively, such as applying the knapsack problem to phylogenetics by packing
946 the optimal choice of species and suitable alignments into a minimally sparse supermatrix.
947
948 **6.2. Community initiatives**

949 Building the ToL is a grand challenge in molecular phylogenetics, and one that cannot be accomplished
950 by a single person or institution's efforts. Several initiatives have been developed in recent years to
951 coordinate efforts and provide the research community with synthetic information. A prominent project is
952 the Open Tree of Life (https://tree.opentreeoflife.org/; Hinchliff et al., 2015). This project provides a
953 synthesis of previously published phylogenies merged through supertree and other grafting methods. One
954 issue faced by the initiative is that it relies on authors uploading their phylogenetic trees to open data
955 repositories, such as Dryad Data Repository (http://datadryad.org/pages/organization; Vision, 2010) or
956 TreeBase (Sanderson et al., 1994; Piel et al., 2009), which at least until recently only occurred in about
957 17% of cases (Drew, 2013). Substantial curatorial efforts are also critical to facilitate reusability of
958 deposited trees (McTavish et al., 2015). A different approach was taken by Antonelli et al. (2017), who
959 developed a framework for continuously inferring time-calibrated large phylogenies from raw sequence
960 data deposited in GenBank (Clark et al., 2016) in a multi-step method. Similarly, various tools have been
961 developed to make information contained in the ToL available for the general public (e.g., Rosindell &
962 Harmon, 2012; Harmon et al., 2013).
963
964 **6.3. Mapping the Tree of Life**
965 While progress has been made in mapping species distributions at the large scale aiming for improved
966 conservation practices (e.g., the Map of Life collaborative project; https://mol.org/), most initiatives do
967 not map the tips of phylogenetic trees directly onto the geographic space, and therefore are limited by
968 current taxonomic knowledge. As spatial variation in biodiversity results from interactions between
969 evolutionary history and environmental factors, explicit connections between the tips of the ToL and
970 geographic ranges will greatly improve biogeographic inferences (Quintero et al., 2015) and our
971 understanding of biodiversity patterns and future trends. Advances in mapping the ToL through earth
972 history using genomic-based phylogenetic inferences over broad scales and explicit spatial models (e.g.,
973 geophylogenies and continuous diffusion models: Kidd, 2010) depend directly on locality data that should
974 be made available in raw and ready-to-use formats. Data sharing policies for associated data, such as

975 geographic coordinates and voucher information, is not well established among journals. We argue that
976 editorial boards should try as best as possible to establish data policies that value and encourage the
977 deposit of geographic data associated to vouchered specimens and other associated information available
978 for future reference.
979
980 **6.4. Best practices for building the Tree of Life**
981 *6.4.1. Data must be well curated in databases and publicly available*
982 As we are now in the era of big data in biological sciences, adequate reproducibility must be a
983 fundamental endeavor of biodiversity research. Therefore, data publication in open access repositories
984 represents a powerful tool that not only ensures long-term storage and public availability for future
985 research, but also serves as a vehicle for clarifying intellectual rights and scientific merits (Costello &
986 Wieczorek, 2014). Biocuration, the activity of organizing, representing, and making biological
987 information accessible to both biologists and bioinformaticians, has now become an important
988 consideration in building, updating, and sustaining the ToL (McTavish et al, 2017). The exponential
989 growth in the amount of genomic scale data and the increased dependence on the availability of each
990 other's' data to answer complex biological questions means that there is a need for improved data
991 management, analysis, and accessibility. GenBank has been the main open access repository for annotated
992 collections of publicly available molecular data. Although the data stored in this database usually lists
993 information such as organism of origin and publication details, the utility of molecular data in this
994 database to answer multiple biological questions, such as biogeographic patterns of biodiversity, is often
995 hampered by lack of associated information such as collection locality or (Scotch et al., 2011; Gratton et
996 al., 2017) attachment to a specific voucher specimen. Moreover, recent surveys have shown that fewer
997 than 20% of phylogenetic studies provide access to phylogenetic data (i.e., alignments and phylogenies)
998 and when they do, critical biological information such as complete taxon names is missing (Drew et al.,
999 2013; Magee, May & Moore, 2014).

1000 We propose two urgent actions to advance this key field. First, authors should be encouraged to
1001 submit molecular data that is linked to voucher specimens deposited in recognized scientific collection
1002 and museums. Second, authors, journals, and curators should encourage all molecular data submitted to
1003 include information such as collection locality and details of voucher specimens. In this regard, other
1004 global initiatives such as the International Barcode of Life Project (iBOL; http://www.ibolproject.org)
1005 have had great success linking molecular data with morphological and distributional data. When all the
1006 data produced or published are curated to high standards and made accessible as soon as available,
1007 biological research will be able to process massive amounts of complex data much more quickly.
1008 Submitting sequence and tree data during publication is now routine. However, making available
1009 all analytical methods such as software and code used to process and analyze data is less widely
1010 employed by the phylogenetic community. Facilities such as TreeBase, Dryad Digital Repository, and
1011 Github (https://github.com/) provide a platform for the curated storage of the data and bioinformatic
1012 pipelines underlying the scientific literature (see McTavish et al., 2015; 2017). Authors and journals
1013 should require all published research to include links to raw data, processed data, and all analytical
1014 methods used to produce the results presented. In general, we advocate for following best practices of
1015 data management and publication to ensure the quality and utility of phylogenomic data and their
1016 associated biological information (Stoltzfus et al., 2012; Drew et al., 2013; Costello & Wieczorek, 2014).
1017 In putting together Figure 2, for example, we found that basic information on a given phylogenomic
1018 study, such as the number of species or sequences analyzed, or the total number of base pairs in an
1019 alignment, were often not reported or difficult to recover; including such information in easy-to-access
1020 tables prior to article acceptance would greatly facilitate meta-analyses and syntheses as the number of
1021 studies grows (Supplementary Table S1).
1022
1023 *6.4.2. The need of adequate curation of analytical tools*
1024 In the same way that data must be adequately stored and curated, analytical tools must be available for
1025 future use and should guarantee proper reproducibility (Wilson et al., 2014; Darriba, Flouri & Stamatakis,

2018). One of the reasons behind the dramatic increase in the number of phylogeographic and phylogenetic studies during the last 20 years is the proliferation of software and bioinformatic tools to process and analyze these data. Thanks to these new methods, it is now possible to implement a wide array of theoretical models that sustain the fields of phylogenomics and phylogeography. As stated above, genome-wide data have notoriously increased the necessity to expand our analytical models, ultimately leading to a stronger demand for computing resources (Darriba, Flouri & Stamatakis, 2018). Given their key role in phylogenomic research, it is advisable that software development, documentation, and availability follow the best possible practices (e.g., Leprevost et al., 2014; Wilson et al., 2014; Guang et al., 2016; Darriba, Flouri & Stamatakis, 2018). Having both data and analytical tools adequately stored and accessible to the public not only will ensure high reproducibility of previous studies, but, more importantly, will facilitate continuing the construction of the ToL (McTavish et al., 2017).

*6.4.3. All contributions toward building the Tree of Life must be properly recognized*

Some current publishing practices in the scientific community may unintentionally represent hurdles toward the ultimate end of collecting and disseminating phylogenetic data on which to build a ToL. For instance, the increasing need in many countries and communities for publishing high-impact papers understandably often discourages researchers from releasing their data until their studies are complete and have passed the peer-review process. This is partially explained by the heavy emphasis of top journals on unusually novel and flashy findings as compared to those studies that represent more modest, but just as critical, advances in the understanding of the phylogenetic relationships of the groups. Similarly, this urge to publish high-impact papers often impedes adequate long-term studies that could potentially generate a wider variety of basic data. With the cultural emphasis on impact and numbers and rates of publication, in practice there is often a penalty for long-term studies. Our current climate often values novel results produced in the short term. Consequently, as a community, we must reach an equilibrium between short- and long-term scientific production in a way that values both, encouraging high impact studies bringing

radical reorganizations of the ToL, without hurting lower impact research and the ongoing search for innovation.

Moreover, because building the ToL is a slow and daunting task, it is important that, as a scientific community, all contributors to the process receive proper recognition for their contributions, thereby keeping motivation high and retaining our best talent. Unfortunately, some contributors, both institutions and roles within them, receive less recognition in this grand task than others. For example, field biologists that obtain basic natural history information and specimens used for building the ToL (Suarez & Tsutsui, 2004), and the natural history museums that house those specimens, are often not recognized sufficiently. As a community, we have been following a trend in which, perhaps inadvertently, we do not value as much the production of basic biological and natural history data. This can certainly be recognized in our national funding practices, which often do not support basic taxonomic or natural history fieldwork at the expense of flashier end-uses of biological specimens. Specimens are the foundation of most phylogenomic and phylogeographic studies, and we should find standard mechanisms not only to acknowledge, but also to encourage the production of these data in an integrative framework. It is time to strengthen those initiatives aimed at recognizing scientific production beyond citations of peer-reviewed literature (e.g., ORCID; https://orcid.org) by giving also credit to the production and impact of basic biology datasets and collected specimens. Providing credit for depositing and generating data by tracking, for example, number of access and downloads or number of studies using genetic data associated to specimens, could represent a formal recognition of the importance of producing and sharing basic biological data could help bridge the gap between naturalists, taxonomists, empiricists, and mathematicians invested on the study of life history.

It will be exciting to have objective estimates that allow tracking the direct and indirect impact of how these data and samples are being used. We are confident that such initiatives will highlight the importance of continuing field- and museum-based research in various fields of biological research (Buerki & Baker, 2016). Furthermore, such cultural shifts will undoubtedly encourage discerning young

minds to embrace basic biological research in their academic endeavors, rather than embracing more lucrative and societally appreciated applied fields.

## 7. Conclusions

In this perspective, we have attempted to cover ground in the vast arena of issues facing modern phylogenomics today. In Table 2, we summarize some of the most pressing challenges that the field of phylogenomics is experiencing as we use coalescent-based methods toward building the ToL. We have seen how genome-scale phylogenomics, currently on a strong footing as a result of the multispecies coalescent model, is increasingly improved by models that recognize reticulate processes, such as recombination and introgression. In contrast, macroevolutionary models that use phylogenies have yet to embrace the heterogeneity that currently drives many theoretical innovations in phylogenetic reconstruction itself. We have emphasized the need for the phylogenomics community to embrace high standards of data quality, curation and accessibility in its long-term pursuit of the ToL. Such a grand mission requires value and recognition placed not only on the end products of the process, such as publications and trees, but also on the natural history specimens on which phylogenies are based and which are cared for by the community of natural history museums. Building the ToL will require contributions from all sectors of biological and related sciences – from field biology to theory and everything in between – and robust cyberinfrastructures to integrate these diverse and increasingly massive data streams.

## ACKNOWLEDGEMENTS

This paper is a product of the ‘Origin of Biodiversity Workshop’ organized by Chalmers University of Technology and the University of Gothenburg, under the auspices of the Gothenburg Centre for Advanced Studies (GoCAS). We are particularly grateful to the GoCAS organizers and facilitators, in particular Karin Hårding, Mattias Marklund, Bernt Wennberg, Sandra Johansson, and Lotta Fernström. We thank Johnathan Clark, Alison Cloutier, Phil Grayson, Kathrin Näpflin, Flavia Termignoni, Jonathan

Schmitt, Simon Sin, João Tonini, and Pengcheng Wang for help compiling Supplementary Table 1. Thomas Couvreur, Tobias Andermann, Prosanta Chakrabarty, Alex Pyron, Claire Morgan, Chris Creevey, and one anonymous reviewer provided useful comments that improved the contents of this manuscript.

**REFERENCES**

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526: 68–74.

Aberer, A. J., Krompas, D., Stamatakis, A. 2013. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. Systematic Biology 62: 162–166.

Adamczak, R., Miloś, P. 2014. U-statistics of Ornstein-Uhlenbeck branching particle system. Journal of Theoretical Probability 27: 1071–1111.

Adamczak, R., Miloś, P. 2015. CLT for Ornstein-Uhlenbeck branching particle system. Electronic Journal of Probability 20: 1–35.

Ané, C. 2008. Analysis of comparative data with hierarchical autocorrelation. Annals of Applied Statistics 2: 1078–1102.

Ané, C., Ho, L. S. T., Roch, S. 2017. Phase transition on the convergence rate of parameter estimation under an Ornstein-Uhlenbeck diffusion on a tree. Journal of Mathematical Biology 74: 355–385.

Andermann, T., Fernandes, A. M., Olsson, U., Topel, M., Pfeil, B., Oxelman, B., Aleixo, A., Faircloth, B. C., Antonelli, A. 2018. Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements. Systematic Biology syy039, https://doi.org/10.1093/sysbio/syy039

Angelis, K., dos Reis, M. 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. Current Zoology 61:874–885.

Antonelli, A., Hettling, H., Condamine, F. L., Vos, K., Nielsson, R. H., Sanderson, J., Sauquet, H., Scharn, R., Silvestro, D., Töpel, M., Bacon, C.D., Oxelman, B., Vos, R. A. 2017. Towards a Self-Updating Platform for Estimating Rates of Speciation and Migration, Ages, and Relationships of Taxa. Systematic Biology 66:152–166.

1128 Ashfield, T., Egan, A. N., Pfeil, B. E., Chen, N. W. G., Podicheti, R., Ratnaparkhe, M. B., Ameline-
1129 Torregrosa, C., Denny, R., Cannon, S., Doyle, J. J., Geffroy, V., Roe, B. A., Saghai-Maroof, M. A.,
1130 Young, N. D., Innes, R. W. 2012. Evolution of a complex disease resistance gene cluster in diploid
1131 Phaseolus and tetraploid Glycine. Plant Physiology 159: 336–354.
1132 Ashkenazy, H., Cohen, O., Pupko, T., Huchon, D. 2014. Indel Reliability in Indel-Based Phylogenetic
1133 Inference. Genome Biology and Evolution 6: 3199–3209.
1134 Bacon, C. D., McKenna, M. J., Simmons, M. P., Wagner, W. L. 2012. Evaluating multiple criteria for
1135 species delimitation: an empirical example using Hawaiian palms (Arecaceae: *Pritchardia*). BMC
1136 Evolutionary Biology 2012: 12–23.
1137 Baker, A. J., Haddrath, O., McPherson, J. D., Cloutier, A. 2014. Genomic Support for a Moa-Tinamou
1138 Clade and Adaptive Morphological Convergence in Flightless Ratites. Molecular Biology and
1139 Evolution 31: 1686–1696.
1140 Bapst, D. W. 2014. Assessing the effect of time-scaling methods on phylogeny-based analyses in the
1141 fossil record. Paleobiology 40: 331–351.
1142 Bartoszek, K., Sagitov, S. 2015a. Phylogenetic confidence intervals for the optimal trait value. Journal of
1143 Applied Probability 52: 1115–1132.
1144 Bartoszek, K., Sagitov, S. 2015b. A consistent estimator of the evolutionary rate. Journal of Theoretical
1145 Biology 371: 69–78.
1146 Bastide, P., Solís-Lemus, C., Kriebel, R., Sparks, K. W., Ané, C. 2018. Phylogenetic Comparative
1147 Methods on Phylogenetic Networks with Reticulations. Systematic Biology 67: 800–820.
1148 Baurain, D., Brinkmann, H., Philippe, H. 2006. Lack of Resolution in the Animal Phylogeny: Closely
1149 Spaced Cladogeneses or Undetected Systematic Errors? Molecular Biology and Evolution 24: 6–9.
1150 Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T., Flicek, P. 2018. Complexity and conservation of
1151 regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nature
1152 Ecology and Evolution 2: 152–163.
1153 Betancur, R., Naylor, G. J. P., Ortí, G. 2014. Conserved genes, sampling error, and phylogenomic

1154 inference. Systematic Biology 63: 257–262.
1155 Bleidorn, C. 2017. Sources of Error and Incongruence in Phylogenomic Analyses. In: Phylogenomics.
1156 Cham: Springer International Publishing, 173–193.
1157 Blom, M. P. K. 2015. EAPhy: A Flexible Tool for High-throughput Quality Filtering of Exon-alignments
1158 and Data Processing for Phylogenetic Methods. PLoS ToL.
1159 Boore, J. L. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology
1160 and Evolution 21:439–446.
1161 Boore, J. L., Daehler, L. L., Brown, W. M. Complete sequence, gene arrangement, and genetic code of
1162 mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). 1999. Molecular
1163 Biology and Evolution 16: 410–418.
1164 Brown, J. W., Wang, N., Smith, S. A. 2017. The development of scientific consensus: Analyzing conflict
1165 and concordance among avian phylogenies. Molecular Phylogenetics and Evolution 116: 69–77.
1166 Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., RoyChoudhury, A. 2012. Inferring species
1167 trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis.
1168 Molecular Biology and Evolution 29: 1917–1932.
1169 Buerki, S., Baker, W. J. 2016. Collections-based research in the genomics era. Biological Journal of the
1170 Linnean Society 117: 5–10.
1171 Burbrink, F. T., Pyron, R. A. 2011. The Impact of Gene-Tree/Species-Tree Discordance on
1172 Diversification-Rate Estimation. Evolution 65: 1851–1861.
1173 Capella-Gutierrez, S., Silla-Martinez, J. M., Gabaldon, T. 2009. trimAl: a tool for automated alignment
1174 trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.
1175 Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic
1176 analysis. Molecular Biology and Evolution 17: 540–552.
1177 Chakrabarty, P. 2010. Genetypes: a concept to help integrate molecular phylogenetics and taxonomy.
1178 Zootaxa 2632: 67–68.
1179 Chen, M. Y., Liang, D., Zhang, P. 2015. Selecting Question-Specific Genes to Reduce Incongruence in

1180 Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. Systematic Biology 64:
1181 1104–1120.
1182 Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P-M., Li, F-W.,
1183 Melkonian, B., Mavrodiev, E. V., Fu, W. S. Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S.,
1184 Liu, X., Xu, X., Wong, G. K-S. 2018. 10KP: A phylodiverse genome sequencing plan. GigaScience
1185 7: 1–9.
1186 Chifman, J., Kubatko, L. 2014. Quartet inference from SNP data under the coalescent model.
1187 Bioinformatics 30: 3317–3324.
1188 Churakov, G., Sadasivuni, M. K, Rosenbloom, K. R., Huchon, D., Brosius, J., Schmitz, J. 2010. Rodent
1189 Evolution: Back to the Root. Molecular Biology and Evolution 27: 1315–1326.
1190 Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E.W. 2016. GenBank. Nucleic Acids
1191 Research 44: D67–D72.
1192 Cloutier, A., Sackton, T. B., Grayson, P., Clamp, M., Baker, A. J., Edwards, S. V. 2018a. Whole-genome
1193 analyses resolve the phylogeny of flightless birds (Palaeognathae) in the presence of an empirical
1194 anomaly zone. bioRxiv https://doi.org/10.1101/262949.
1195 Cloutier, A., Sackton, T. B., Grayson, P., Edwards, S. V., Baker, A. J. 2018b. First nuclear genome
1196 assembly of an extinct moa species, the little bush moa (*Anomalopteryx didiformis*). bioRxiv doi:
1197 https://doi.org/10.1101/262816.
1198 Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., Mick, E., Sorek, R. 2016.
1199 Comparative transcriptomics across the prokaryotic Tree of Life. Nucleic Acids Research 44: W46–
1200 W53.
1201 Costello, M. J., Wieczorek, J. 2014. Best practice for biodiversity data management and publication.
1202 Biological Conservation 173: 68–73.
1203 Crawford, F. W., Suchard, M. A., 2013. Diversity, disparity, and evolutionary rate estimation for
1204 unresolved Yule trees. Systematic Biology 62: 439–455.
1205 Cutter, A. D. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes

1206 and evolutionary theory. Molecular Phylogenetics and Evolution 69: 1172–1185.
1207 Dalquen, D. A., Zhu, T., Yang, Z. 2017. Maximum Likelihood Implementation of an Isolation-with-
1208 Migration Model for Three Species. Systematic Biology 66: 379–398.
1209 Darriba, D., Flouri, T., Stamatakis, A. 2018. The state of software for evolutionary biology. Molecular
1210 Biology and Evolution 35: 1037–1046.
1211 Davies, T. J., Kraft, N. J. B., Salamin, N., Wolkovich, E. M. 2012. Incompletely resolved phylogenetic
1212 trees inflate estimates of phylogenetic conservatism. Ecology 93: 242–247.
1213 Dayrat, B. 2005. Towards integrative taxonomy. Biological Journal of the Linnean Society 85: 407–415.
1214 de Queiroz, A., Gatesy, J. 2007. The supermatrix approach to systematics. Trends in Ecology and
1215 Evolution 22: 34–41.
1216 de Vienne, D. M., Ollier, S., Aguileta, G. 2012. Phylo-MCOA: a fast and efficient method to detect
1217 outlier genes and species in phylogenomics using multiple co-inertia analysis. Molecular Biology and
1218 Evolution 29: 1587–1598.
1219 De Maio, N., Schlötterer, C., Kosiol, C. 2013. Linking Great Apes Genome Evolution across Time Scales
1220 Using Polymorphism-Aware Phylogenetic Models. Molecular Biology and Evolution 30: 2249–
1221 2262.
1222 DeBolt, S. 2010. Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate
1223 Family Generational Scales. Genome Biology and Evolution 2: 441–453.
1224 Degnan, J. H. 2018. Modeling Hybridization Under the Network Multispecies Coalescent. Systematic
1225 Biology 67: 786–799.
1226 Degnan, J. H., Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the
1227 multispecies coalescent. Trends in Ecology & Evolution 24: 332–340.
1228 Dell Ampio, E., Meusemann, K., Szucsich, N. U., Peters, R. S., Meyer, B., Borner, J., Petersen, M.,
1229 Aberer, A. J., Stamatakis, A., Walzl, M. G., Minh, B. Q., Haeseler, von A., Ebersberger, I., Pass, G.
1230 N., Misof, B. 2013. Decisive Data Sets in Phylogenomics: Lessons from Studies on the Phylogenetic
1231 Relationships of Primarily Wingless Insects. Molecular Biology and Evolution 31: 239–249.

1232 Delsuc, F., Brinkmann, H., Philippe, H. 2005. Phylogenomics and the reconstruction of the Tree of Life.
1233 Nature Reviews Genetics 6: 361–375.
1234 dos Reis, M., Donoghue, P. C. J., Yang, Z. 2016. Bayesian molecular clock dating of species divergences
1235 in the genomics era. Nature Reviews Genetics 17:71–80.
1236 Douzery, E. J. P., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., Ranwez, V. 2014.
1237 OrthoMaM v8: A Database of Orthologous Exons and Coding Sequences for Comparative Genomics
1238 in Mammals. Molecular Biology and Evolution 31:1923–1928.
1239 Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy.
1240 Systematic Botany 17: 144-163.
1241 Drew, B. T. 2013. Data deposition: Missing data mean holes in Tree of Life. Nature 493: 305.
1242 Drew, B. T., Gazis, R., Cabezas, P., Swithers, K. S., Deng, J., Rodriguez, R., Katz, L. A., Crandall, K. A.,
1243 Hibbett, D. S., Soltis, D. E. 2013. Lost Branches on the Tree of Life. PLoS Biology 11:e1001636.
1244 DOI: 10.1371/journal.pbio.1001636.
1245 Dunn, C. W., Howinson, M., Zapata, F. 2013. Agalma: an automated phylogenomics workflow. BMC
1246 Bioinformatics 14: 330.
1247 Edwards, S. V. 2009a. Natural selection and phylogenetic analysis. Proceedings of the National Academy
1248 of Sciences of the United States of America 106: 8799–8800.
1249 Edwards, S. V. 2009b. Is a new and general theory of molecular systematics emerging? Evolution 63:1–
1250 19.
1251 Edwards, S. V., Beerli, P. 2000. Perspective: gene divergence, population divergence, and the variance in
1252 coalescence time in phylogeographic studies. Evolution 54: 1839–1854.
1253 Edwards, S. V., Cloutier, A., Baker, A. J. 2017. Conserved Nonexonic Elements: A Novel Class of
1254 Marker for Phylogenomics. Systematic Biology 66: 1028–1044.
1255 Edwards, S. V., Liu, L., Pearl, D. K. 2007. High-resolution species trees without concatenation.
1256 Proceedings of the National Academy of Sciences USA 104: 5936–5941.
1257 Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G., Moritz, C. 2016a. Reticulation, divergence, and the

1258 phylogeography-phylogenetics continuum. Proceedings of the National Academy of Sciences USA
1259 113: 8025–8032.
1260 Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S.,
1261 Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., Davis, C. C. 2016b. Implementing and
1262 testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Molecular
1263 Phylogenetics and Evolution 94: 447–462.
1264 Elmer, K. R., Meyer, A. 2011. Adaptation in the age of ecological genomics: insights from parallelism
1265 and convergence. Trends in Ecology and Evolution 26: 298–306.
1266 Ellegren, N., Galtier, N. 2016. Determinants of genetic diversity. Nature Reviews Genetics 17: 422–433.
1267 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., Foll, M. 2013. Robust demographic
1268 inference from genomic and SNP data. PLoS Genet 9: e1003905.
1269 Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., Glenn, T. C. 2012.
1270 Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary
1271 Timescales. Systematic Biology 61: 717–726.
1272 Faurby, S., Svenning, J. C. 2015. A species-level phylogeny of all extant and late Quaternary extinct
1273 mammals using a novel heuristic-hierarchical Bayesian approach. Molecular Phylogenetics and
1274 Evolution 84: 14–26.
1275 Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters.
1276 American Journal of Human Genetics 25:471–492.
1277 Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125: 1–15.
1278 Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. Annual Review of
1279 Genetics 22: 521–565.
1280 Felsenstein, J. 2012. A comparative method for both discrete and continuous characters using the
1281 threshold model. American Naturalist 179: 145–156.
1282 Fernández, R, Laumer C. E., Vahtera, V., Libro, S., Kaluziak, S., Sharma, P. P., Pérez-Morro, A. R.,
1283 Edgecombe, G. D., Giribert, G. 2014. Evaluating Topological Conflict in Centipede Phylogeny

1284 Using Transcriptomic Data Sets. Molecular Biology and Evolution 31: 1500–1513.
1285 Figuet, E., Ballenghien, M., Romiguier, J., Galtier, N. 2015. Biased Gene Conversion and GC-Content
1286 Evolution in the Coding Sequences of Reptiles and Vertebrates. Genome Biology and Evolution 7:
1287 240–250.
1288 Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology 19: 99–113.
1289 Fong, J. J., Brown, J. M., Fujita, M. K., Boussau, B. 2012. A Phylogenomic Approach to Vertebrate
1290 Phylogeny Supports a Turtle-Archosaur Affinity and a Possible Paraphyletic Lissamphibia. PLoS
1291 ONE 7: e48990–14.
1292 Frantzeskakis, L., Kracher, B., Kusch, S., Yoshikawa-Maekawa, M., Bauer, S., Pedersen, C., Spanu, P.
1293 D., Maekawa, T., Schulze-Lefert, P., Panstruga, R. 2018. Signatures of host specialization and a
1294 recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery
1295 mildew pathogen. BMC Genomics 19: 381.
1296 Fredman, D., White, S. J., Potter, S., Eichler, E. E, Dunnen, J. T. D., Brookes, A. J. 2004. Complex SNP-
1297 related sequence variation in segmental genome duplications. Nature Genetics 36:861–866.
1298 Garrick, R. C., Sunnucks, P., Dyer, R. J. 2010. Nuclear gene phylogeography using PHASE: dealing with
1299 unresolved genotypes, lost alleles, and systematic bias in parameter estimation. BMC Evolutionary
1300 Biology 10: 118.
1301 Garrick, R. C., Bonatelli, I. A., Hyseni, C., Morales, A., Pelletier, T. A., Perez, M. F., Rice, E., Satler, J.
1302 D., Symula, R. E., Thomé, M. T. C. 2015. The evolution of phylogeographic data sets. Molecular
1303 Ecology 24: 1164–1171.
1304 Genome 10K Community of Scientists. 2009. A proposal to obtain whole-genome sequence for 10,000
1305 vertebrate species. Journal of Heredity 100: 659–674.
1306 Gernhard, T., 2008a. The conditioned reconstructed process. Journal of Theoretical Biology 253: 769–
1307 778.
1308 Gernhard, T. 2008b. New analytic results for speciation times in neutral models. Bulletin of Mathematical
1309 Biology 70: 1082–1097.

1310 Ghiurcuta, C. G., Moret, B. M. 2014. Evaluating synteny for improved comparative studies.
1311 Bioinformatics 30: i9–i18.
1312 Goodwin, S., McPherson, J. D., McCombie, 2016. Coming of age: ten years of next-generation
1313 sequencing technologies. Nature Reviews Genetics 17: 333–351.
1314 Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Systematic
1315 Biology 47: 9–17.
1316 Gratton, P., Marta, S., Bocksberger, G., Winter, M., Trucchi, E., Kühl, H. 2016. A world of sequences:
1317 can we use georeferenced nucleotide databases for a robust automated phylogeography? Journal of
1318 Biogeography 44: 475–486.
1319 Griffiths, R. C., Marjoram, P. 1996. An ancestral recombination graph. *in* ‘‘IMA Volume on
1320 Mathematical Population Genetics’’ (P. Donnelly and S. Tavare, Eds.), pp. 257–270, Springer-
1321 Verlag, New York.
1322 Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X.,
1323 Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., Shabalov, I. 2014 MycoCosm portal:
1324 gearing up for 1000 fungal genomes. Nucleic Acids Research 42: D699–704.
1325 Guang, A., Zapata, F., Howison, M., Lawrence, C. E., Dunn, C. W. 2016. An Integrated Perspective on
1326 Phylogenetic Workflows. Trends in Ecology and Evolution 31: 116–126.
1327 Guerrero R. F., Hahn M. W. 2018. Quantifying the risk of hemiplasy in phylogenetic inference.
1328 Proceedings of the National Academy of Sciences USA 201811268. DOI: 10.1073/pnas.1811268115.
1329 Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New Algorithms and
1330 Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.
1331 Systematic Biology 59: 307–321.
1332 Gusfield, D. 2015. Persistent phylogeny. In: New York, New York, USA: ACM Press, 443–451.
1333 Hahn, M. W. 2018. Molecular Population Genomics. Oxford Univ. Press.
1334 Hahn, M. W., Nakhleh, L. 2016. Irrational exuberance for resolved species trees. Evolution 70: 7–17.
1335 Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., Wall., J. D. 2011. Genetic evidence for

1336 archaic admixture in Africa. Proceedings of the National Academy of Sciences 108: 15123–15128.
1337 Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51: 1341–
1338 1351.
1339 Harmon, L. J., Baumes, J., Hughes, C., Soberón, J., Specht, C. D., Tumer, W., Lisle, C., Thacker, R. W.
1340 2013. Arbor: Comparative Analysis Workflows for the Tree of Life. PLoS Currents 5:
1341 ecurrents.tol.099161de5eabdee073fd3d21a44518dc.
1342 Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., Brumfield, R. T. 2016. Sequence Capture
1343 versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Systematic Biology
1344 65: 910–924.
1345 He, D., R. Sierra, Pawlowski, J., Baldauf, S. L. 2016. Reducing long-branch effects in multi-protein data
1346 uncovers a close relationship between *Alveolata* and *Rhizaria*. Molecular Phylogenetics and
1347 Evolution 101: 1–7.
1348 Heath, T. A., Huelsenbeck, J. P., Stadler, T. 2014. The fossilized birth–death process for coherent
1349 calibration of divergence-time estimates. Proceedings of the National Academy of Sciences USA
1350 111: E2957–E2966.
1351 Hedtke, S. M., Townsend, T. M., Hillis, D. M. 2006. Resolution of Phylogenetic Conflict in Large Data
1352 Sets by Increased Taxon Sampling. Systematic Biology 55: 522–529.
1353 Heled, J., Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. Molecular
1354 Biology and Evolution 27: 570–580.
1355 Hey, J., Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and
1356 divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*.
1357 Genetics 167: 747–760.
1358 Hey, J., Nielsen, R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte
1359 Carlo methods in population genetics. Proceedings of the National Academy of Sciences USA 104:
1360 2785–2790.
1361 Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46: 627–640.

1362 Hey, J. 2010. Isolation with Migration Models for More Than Two Populations. Molecular Biology and
1363 Evolution 27: 905–920.
1364 Hibbins, M. S., Hahn, M. W. 2018. The timing and direction of introgression under the multispecies
1365 network coalescent. bioRxiv doi: https://doi.org/10.1101/328575
1366 Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R., Bejerano, G. 2012. A "forward
1367 genomics" approach links genotype to phenotype using independent phenotypic losses among related
1368 species. Cell Reports 2: 817–23.
1369 Hillis, D. M. 1996. Inferring complex phylogenies. Nature 383: 130–131.
1370 Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Systematic
1371 Biology 47: 3–8.
1372 Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K.
1373 A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D. IV,
1374 McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T.,
1375 Cranston, K. A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive Tree of Life.
1376 Proceedings of the National Academy of Sciences USA 112: 12764–12769.
1377 Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed,
1378 L. K., Storfer, A., Whitlock, M. C. 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls,
1379 Practical Solutions, and Future Directions. American Naturalist 188: 379–397.
1380 Hobolth, A., Christensen, O. F., Mailund, T., Schierup, M. H. 2007. Genomic relationships and speciation
1381 times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS
1382 Genet. 3, e7.
1383 Huang, H. T., He, Q. I., Kubatko, L. S., Knowles, L. L. 2010. Sources of error inherent in species-tree
1384 estimation: impact of mutational and coalescent effects on accuracy and implications for choosing
1385 among different methods. Systematic Biology 59: 573–583.
1386 Huang, H., Sukumaran, J., Smith, S. A., Knowles, L. L. 2017. Cause of gene tree discord? Distinguishing
1387 incomplete lineage sorting and lateral gene transfer in phylogenetics. PeerJ Preprints

https://doi.org/10.7287/peerj.preprints.3489v1.

Huang, W., Zhou, G., Marchand, M., Ash, J. R., Morris, D., Van Dooren, P., Brown, J. M., Gallivan, K. A., Wilgenbusch, J. C. 2016. TreeScaper: Visualizing and Extracting Phylogenetic Signal from Sets of Trees. Molecular Biology and Evolution 33: 3314–3316.

Huber, K. T., Oxelman, B., Lott, M., Moulton, V. 2006. Reconstructing the evolutionary history of polyploids from multilabeled trees. Molecular Biology and Evolution 23: 1784–91.

Hunt, G. 2013. Testing the link between phenotypic evolution and speciation: an integrated palaeontological and phylogenetic analysis. Methods in Ecology and Evolution 4: 714–23.

Hunt, G., Slater, G. 2016. Integrating Paleontological and Phylogenetic Approaches to Macroevolution. Annual Review of Ecology, Evolution, and Systematics 47: 189–213.

Huson, D. H. 2006. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution 23: 254–267.

Hykin, S. M., Bi, K., McGuire, J. A. 2015. Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing. PLoS ONE 10(10): e0141579.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., McVean, G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics 44: 226–232.

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., Philippe, H. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nature Ecology and Evolution 1: 1370–1378.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldon, T., Capella-Gutierrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V.,

1414 Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M.
1415 V., Alfaro-Nunez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield,
1416 P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng,
1417 Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J.,
1418 Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson,
1419 W., Koepfli, K. P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R.,
1420 Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alstrom, P., Edwards, S. V., Stamatakis, A.,
1421 Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun. W, Gilbert, M. T. P., Zhang, G. 2014.
1422 Whole-genome analyses resolve early branches in the Tree of Life of modern birds. Science
1423 346:1320–1331.

1424 Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. 2006. Phylogenomics: the beginning of
1425 incongruence? Trends in Genetics 22: 225–231.

1426 Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K, Mooers, A. O. 2012. The global diversity of birds in
1427 space and time. Nature 491: 444–448.

1428 Jhwueng, D. C., O'Meara, B. 2015. Trait evolution on phylogenetic networks. BioRxiv
1429 doi: https://doi.org/10.1101/023986

1430 Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y., Tighe, S. W. 2017. Real-Time DNA Sequencing in
1431 the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer. Journal of Biomolecular
1432 Techniques 28: 2–7.

1433 Jombart, T., Kendall, M., Almagro-Garcia, J., Colijn, C. 2017. TREESPACE: Statistical exploration of
1434 landscapes of phylogenetic trees. Molecular Ecology Resources 17: 1385–1392.

1435 Jones, M. R., Good, J. M. 2016. Targeted capture in evolutionary and ecological genomics. Molecular
1436 Ecology 25: 185–202.

1437 Jones, G. R. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the
1438 multispecies coalescent. Journal of Mathematical Biology 74: 447–467.

1439 Jones, G. R. 2018. Divergence estimation in the presence of incomplete lineage sorting and migration.

1440 Systematic Biology, syy041, https://doi.org/10.1093/sysbio/syy041
1441 Jones, G. R., Sagitov, S., Oxelman, B. 2013. Statistical Inference of Allopolyploid Species Networks in
1442 the Presence of Incomplete Lineage Sorting. Systematic Biology 62: 467–478.
1443 Jones, G. R., Aydin, Z., Oxelman, B. 2015. DISSECT: an assignment-free Bayesian discovery method for
1444 species delimitation under the multispecies coalescent. Bioinformatics 31: 991–998.
1445 Kainer, D., Lanfear, R. 2015. The Effects of Partitioning on Phylogenetic Inference. Molecular Biology
1446 and Evolution 32: 1611–1627.
1447 Kaiser, V. B., van Tuinen, M., Ellegren, H. 2007. Insertion events of CR1 retrotransposable elements
1448 elucidate the phylogenetic branching order in galliform birds. Molecular Biology and Evolution 24:
1449 338–347.
1450 Khost, D. E., Eickbush, D. G., Larracuente, A. M. 2017. Single-molecule sequencing resolves the detailed
1451 structure of complex satellite DNA loci in *Drosophila melanogaster*. Genome Research 27: 709–721.
1452 Kidd, D. M. 2010. Geophylogenies and the Map of Life. Systematic Biology 59: 741–752.
1453 Kim, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and
1454 increasing numbers of taxa. Systematic Biology 46: 363–374.
1455 King, N., Rokas, A. 2017. Embracing Uncertainty in Reconstructing Early Animal Evolution. Current
1456 Biology 27: R1081–R1088.
1457 Klopfstein, S., Massingham, T., Goldman, N. 2017. More on the Best Evolutionary Rate for Phylogenetic
1458 Analysis. Systematic Biology 66: 769–785.
1459 Knowles, L. L., Smith, S. A. Huang, H., Sukumaran, J. 2018. A matter of phylogenetic scale:
1460 distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord
1461 in recent versus deep diversification histories. American Journal of Botany 105: 376–384.
1462 Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese, D. A., Cannon, J.
1463 T., Moroz, L. L., Lieb, B., Halanych, K. M. 2017. Phylogenomics of Lophotrochozoa with
1464 Consideration of Systematic Error. Systematic Biology 66: 256–282.
1465 Kowada, L. A. B., Doerr, D., Dantas, S., Stoye, J. 2016. New Genome Similarity Measures Based on

1466 Conserved Gene Adjacencies. In: Singh M. (eds) Research in Computational Molecular Biology.
1467 RECOMB 2016. Lecture Notes in Computer Science, vol 9649. Springer, Cham.
1468 Kozlov, A. M., Aberer, A. J., Stamatakis, A. 2015. ExaML version 3: a tool for phylogenomic analyses
1469 on supercomputers. Bioinformatics 31: 2577–2579.
1470 Kriegs, J. O., Zemann, A., Churakov, G., Matzke, A., Ohme, M., Zischler, H., Brosius, J., Kryger, U.,
1471 Schmitz, J. 2010. Retroposon Insertions Provide Insights into Deep Lagomorph Evolution. Molecular
1472 Biology and Evolution 27: 2678–2681.
1473 Ksepka, D. T., Benton, M. J., Carrano, M. T., Gandolfo, M. A., Head, J. J., Hermsen, E. J., Joyce, W. G.,
1474 Lamm, K. S., Patané, J. S. L., Phillips, M. J., Polly, P. D., van Tuinen, M., Ware, J. L., Warnock, R.
1475 C. M., Parham, J. F. 2011. Synthesizing and databasing fossil calibrations: divergence dating and
1476 beyond. Biology Letters 2011: 801–803.
1477 Kubatko, L. S., Degnan, J. H. 2007. Inconsistency of phylogenetic estimates from concatenated data
1478 under coalescence. Systematic Biology 56: 17–24.
1479 Kubatko, L. S., Gibbs, H. L., Bloomquist, E. W. 2011. Inferring species-level phylogenies and taxonomic
1480 distinctiveness using multilocus data in *Sistrurus* rattlesnakes. Systematic Biology 60: 393–409.
1481 Kuhn, T.S., Mooers, A. Ø., Thomas, G. H. 2011. A Simple Polytomy Resolver for Dated Phylogenies.
1482 Methods in Ecology and Evolution 2: 427–36.
1483 Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky, Pond, S. L, Tamura, K. 2012. Statistics and
1484 Truth in Phylogenomics. Molecular Biology and Evolution 29:457–472.
1485 Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. 2013. Blobology: exploring raw genome data for
1486 contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Frontiers in
1487 Genetics 4: 237.
1488 Kunin, V., Goldovsky, L., Darzentas, N. 2005. The net of life: reconstructing the microbial phylogenetic
1489 network. Genome Research 15: 954–959.
1490 Lamichhaney, S., Berglund, B., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A.,
1491 Promerova, M., Rubin, C.J., Wang, C., Zamani, N., Grant, B.R., Grant, P.R., Webster, M.T.,

1492 Andersson, L. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing.
1493 Nature 518: 371–375.
1494 Lanier, H. C., Knowles, L. L. 2012. Is Recombination a Problem for Species-Tree Analyses? Systematic
1495 Biology 61: 691–701.
1496 Lanier, H. C., Knowles, L. L. 2015. Applying species-tree analyses to deep phylogenetic histories:
1497 Challenges and potential suggested from a survey of empirical phylogenetic studies. Molecular
1498 Phylogenetics and Evolution 83: 191–199.
1499 Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A. 2014. Selecting optimal partitioning
1500 schemes for phylogenomic datasets. BMC Evolutionary Biology 14: 82.
1501 Lartillot, N., Rodrigue, N., Stubbs, D., Richer, J. 2013. PhyloBayes MPI: Phylogenetic Reconstruction
1502 with Infinite Mixtures of Profiles in a Parallel Environment. Systematic Biology 62: 611–615.
1503 Larson-Johnson, K. 2016. Phylogenetic investigation of the complex evolutionary history of dispersal
1504 mode and diversification rates across living and fossil Fagales. New Phytologist 209: 418–35.
1505 Leaché, A. D., Oaks, J. R. 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in
1506 Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 48: 69–84.
1507 Leaché, A. D., Chavez, A. S., Jones, L. N. 2015. Phylogenomics of phrynosomatid lizards: conflicting
1508 signals from sequence capture versus restriction site associated DNA sequencing. Genome Biology
1509 7: 706–719.
1510 Leaché, A. D., Harris, R. B., Rannala B., Yanz Z. 2014. The influence of gene flow on species tree
1511 estimation: A simulation study. Systematic Biology 63: 17–30.
1512 Leigh, J. W., Lapointe, F-J., Lopez, P., Bapteste, E. 2011. Evaluating Phylogenetic Congruence in the
1513 Post-Genomic Era. Genome Biology and Evolution 3: 571–587.
1514 Lelieveld, S. H., Spielman, M., Mundlos, S., Veltman, J. A., Gilissen, C. 2015. Comparison of Exome
1515 and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Human
1516 Mutation 36: 815–822.
1517 Lemmon, A. R., Emme, S. A., Lemmon, E. M. 2012. Anchored Hybrid Enrichment for Massively High-

1518 Throughput Phylogenomics. Systematic Biology 61: 727–744.
1519 Lemmon, E. M., Lemmon, A. R. 2013. High-Throughput Genomic Data in Systematics and
1520 Phylogenetics. Annual Review of Ecology Evolution and Systematics 44: 99–121.
1521 Leonardi, M., Librado, P., Der Sarkissian, C., Schubert, M., Alfarhan, A. H., Alquraishi, S. A., Al-
1522 Rasheid, K. A. S., Gamba, C., Willerslev, E., Orlando, L. 2017. Evolutionary Patterns and Processes:
1523 Lessons from Ancient DNA. Systematic Biology 66: e1–e29.
1524 Leprevost, F. V., Barbosa, V. C., Francisco, E. L., Perez-Riverol, Y., Carvalho, P. C. 2014. On best
1525 practices in the development of bioinformatics software. Frontiers in Genetics 5: 199.
1526 Liang, Y., Liao, Bo., Zhu, W. 2017. An Improved Binary Differential Evolution Algorithm to Infer
1527 Tumor Phylogenetic Trees. BioMed Research International, Article ID 5482750, 13.
1528 Li, G., Figueiró, H. V., Eizirik, E., Murphy, W. J. 2018. Recombination-aware phylogenomics unravels
1529 the complex divergence of hybridizing species. bioRxiv doi: https://doi.org/10.1101/485904
1530 Lin, Y., Hu, F., Tang, J. Moret, B. M. 2013. Maximum likelihood phylogenetic reconstruction from high-
1531 resolution whole-genome data and a tree of 68 eukaryotes. Pacific Symposium on Biocomputing
1532 285–296.
1533 Lischer, H. E., Excoffier, L., Heckel, G. 2014. Ignoring heterozygous sites biases phylogenomic estimates
1534 of divergence times: implications for the evolutionary history of Microtus voles. Molecular Biology
1535 and Evolution 31: 817–831.
1536 Liu, L., Pearl, D. K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions
1537 of a species phylogeny using estimated gene tree distributions. Systematic Biology 56: 504–514.
1538 Liu, L., Wu, S., Yu, L. 2015. Coalescent methods for estimating species trees from phylogenomic data.
1539 Journal of Systematics and Evolution 53: 380–390.
1540 Liu, L., Xi, Z., Wu, S., Davis, C. C., Edwards, S. V. 2015. Estimating phylogenetic trees from genome-
1541 scale data. Annals of the New York Academy of Sciences 1360: 36–53.
1542 Liu, L., Xi, Z., Davis, C. C. 2015. Coalescent Methods Are Robust to the Simultaneous Effects of Long
1543 Branches and Incomplete Lineage Sorting. Molecular Biology and Evolution 32: 791–805.

1544 Liu, L., Yu, L., Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species
1545 trees under the coalescent model. BMC Evolutionary Biology 10: 302.
1546 Liu, L., Yu, L., Pearl, D. K., Edwards, S. V. 2009. Estimating species phylogenies using coalescence
1547 times among sequences. Systematic Biology 58: 468–477.
1548 Liu, L., Zhang, J., Rheindt, F. E., Lei, F., Qu, Y., Wang, Y., Sullivan, C., Nie, W., Wang, J., Yang, F.,
1549 Chen, J., Edwards, S. V., Meng, J., Wu, S. 2017. Genomic evidence reveals a radiation of placental
1550 mammals uninterrupted by the KPg boundary. Proceedings of the National Academy of Science of
1551 the USA 114: E7282–7290.
1552 Long, J. C. 1991. The genetic structure of admixed populations. Genetics 127: 417–418.
1553 Lott, M., Spillner, A., Huber, K. T., Moulton, V. 2009. PADRE: A package for analyzing and displaying
1554 reticulate evolution. Bioinformatics 25: 1199–2000.
1555 Lu, H., Giordano, F., Ning, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly.
1556 Genomics, Proteomics & Bioinformatics 14: 265–279.
1557 Lynch, V. J., Bedoya-Reina, O. C., Ratan, A., Sulak, M., Drautz-Moses, D. I., Perry, G. H., Miller, W.,
1558 Schuster, S. C. 2015. Elephantid Genomes Reveal the Molecular Bases of Woolly Mammoth
1559 Adaptations to the Arctic. Cell Reports 12: 217–228.
1560 Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46: 523–536.
1561 Magee, A. F., May, M. R., Moore, B. R. 2014. The Dawn of Open Access to Phylogenetic Data. PLoS ONE
1562 9:e110268. DOI: 10.1371/journal.pone.0110268.
1563 Mai, U., Mirarab, S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of
1564 phylogenetic trees. BMC Genomics 19(Suppl. 5): 272.
1565 Mallet, J., Besansky, N., Hahn, M. W. 2015. How reticulated are species? BioEssays 38: 140–149.
1566 Manceau, M., Lambert, A., Morlon, H. 2015. Phylogenies support out-of-equilibrium models of
1567 biodiversity. Ecology Letters 18: 347–356.
1568 Manceau, M., Lambert, A., Morlon, H. 2017. A unifying comparative phylogenetic framework including
1569 traits coevolving across interacting lineages. Systematic Biology 66: 551–568.

1570 Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., Lysak, M. A. 2010. Fast Diploidization in
1571 Close Mesopolyploid Relatives of *Arabidopsis*. The Plant Cell 22: 2277–2290.
1572 Marcovitz, A., Jia, R., Bejerano, G. 2016. "Reverse Genomics" Predicts Function of Human Conserved
1573 Noncoding Elements. Molecular Biology and Evolution 33: 1358–1369.
1574 Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., The International Wheat Genome
1575 Sequencing Consortium, Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., Olsen,
1576 O. A. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:
1577 1250092.
1578 Marcussen, T., Heier, L., Brysting, A. K., Oxelman, B., Jakobsen, K. S. 2015. From gene trees to a dated
1579 allopolyploid network: Insights from the angiosperm genus *Viola* (Violaceae). Systematic Biology
1580 64: 84–101.
1581 Matzke, A., Churakov, G., Berkes, P., Arms, E. M., Kelsey, D., Brosius, J., Kriegs, J. O., Schmitz, J.
1582 2012. Retroposon Insertion Patterns of Neoavian Birds: Strong Evidence for an Extensive Incomplete
1583 Lineage Sorting Era. Molecular Biology and Evolution 29: 1497–1501.
1584 Matzke, N. J., Wright, A. 2016. Inferring node dates from tip dates in fossil Canidae: the importance of
1585 tree priors. Biology Letters 12: 20160328.
1586 McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., Brumfield, R. T. 2013. Applications of
1587 next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and
1588 Evolution 66: 526–538.
1589 McCormack, J. E., Tsai, W. L. E., Faircloth, B. C. 2016. Sequence capture of ultraconserved elements
1590 from bird museum specimens. Molecular Ecology Resources 16: 1189–1203.
1591 McCormack, J. E., Rodríguez-Gómez, F., Tsai, W. L. E., Faircloth, B. C. 2017. Transforming Museum
1592 Specimens into Genomic Resources. Pp. 143–156 *in* M. S. Webster (editor), The Extended
1593 Specimen: Emerging Frontiers in Collections-based Ornithological Research. Studies in Avian
1594 Biology (no. 50), CRC Press, Boca Raton, FL.
1595 McTavish, E. J., Drew, B. T., Redelings, B., Cranston, K. A. 2017. How and why to build a unified Tree

of Life. BioEssays 1700114.

McTavish, E. J., Hinchliff, C. E., Allman, J. F., Brown, J., Cranston, K. A., Holder, M. T., Rees, J. A., Smith, S. A. 2015. Phylesystem: A git-based data store for community-curated phylogenetic estimates. Bioinformatics 31: 2794–2800.

Mendes, F. K., Fuentes-González, J. A., Schraiber, J. G., Hahn, M. W. 2018. A multispecies coalescent model for quantitative traits. eLife 7:e36482.

Mendes, F. K., Hahn, Y., Hahn, M. W. 2016. Gene tree discordance can generate patterns of diminishing convergence over time. Molecular Biology and Evolution 33: 3299–3307.

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K. S., Montgomery, S. B., Wheeler, M., Buchan, J. G., Lambert, C. C., Eng, K. S., Hickey, L., Korlach, J., Ford, J., Ashley, E. A. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine 20: 159–163.

Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu1, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M. F., Shunkov, M. V, Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J., Pääbo, S. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222–226.

Mirarab, S., Warnow, T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31: i44–i52.

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

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspock, U., Aspock, H., Bartel, D., Blanke, A., Berger, S., Bohm, A., Buckley, T. R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y, Jermiin, L. S.,

1622 Kawahara, A. Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H.,
1623 Machida, R., Mashimo, Y., Kapli, P., McKenna, D. D., Meng, G., Nakagaki, Y., Navarrete-Heredia,
1624 J. L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B. M., Schutte, K.,
1625 Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N. U., Tan, M.,
1626 Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T.,
1627 Walzl, M. G., Wiegmann, B. M., Wilbrandt, J., Wipfler, B., Wong, T. K., Wu, Q., Wu, G., Xie, Y.,
1628 Yang, S., Yang, Q., Yeates, D. K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y.,
1629 Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J.,
1630 Wang, J., Kjer, K. M., Zhou. X. 2014. Phylogenomics resolves the timing and pattern of insect
1631 evolution. Science 346: 763–767.
1632 Mitchell, A., Mitter, C., Regier, J. C. 2000. More taxa or more characters revisited: Combining data from
1633 nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera).
1634 Systematic Biology 49: 202–224.
1635 Montague, M. J., Li, G., Gandolfi, B., Khan, R., Aken, B. L., Searle, S. M. J., Minx, P., Hillier, L. W.,
1636 Koboldt, D. C., Davis, B. W., Driscoll, C. A., Barr, C. S., Blackistone, K., Quilez, J., Lorente-
1637 Galdos, B., Marques-Bonet, T., Alkan, C., Thomas, G. W. C., Hahn, M. W., Menotti-Raymond, M.,
1638 O'Brien, S. J., Wilson, R. K., Lyons, L. A., Murphy, W. J., Warren, W. C. 2014. Comparative
1639 analysis of the domestic cat genome reveals genetic signatures underlying feline biology and
1640 domestication. Proceedings of the National Academy of Sciences USA 111: 17230–17235.
1641 Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., Worm, B. 2011. How Many Species Are There on
1642 Earth and in the Ocean? PLoS Biology 9(8):e1001127.
1643 Morlon, H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters 17: 508–525.
1644 Morlon, H., Parsons, T.L., Plotkin, J. 2011. Reconciling molecular phylogenies with the fossil record
1645 Proceedings of the National Academy of Sciences USA 108: 16327–16332.
1646 Murphy, W. J., Larkin, D. M., Everts-van der Wind, A, Bourque, G., Tesler, G., Auvil, L., Beever, J. E,
1647 Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A.,

1648 Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M.,
1649 Womack, J. E., O'Brien, S. J., Pevzner, P. A., Lewin, H. A. 2005. Dynamics of mammalian
1650 chromosome evolution inferred from multispecies comparative maps. Science 309: 613–617.
1651 Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S., Miller, W. 2007. Using genomic data to
1652 unravel the root of the placental mammal phylogeny. Genome Research 17: 413–421.
1653 Nabhan, A. R., Sarkar, I. N. 2012. The impact of taxon sampling on phylogenetic inference: a review of
1654 two decades of controversy. Briefings in Bioinformatics 13: 122–134.
1655 Nee, S., Mooers, A. O., Harvey, P. H. 1992. Tempo and mode of evolution revealed from molecular
1656 phylogenies. Proceedings of the National Academy of Sciences USA 89: 8322–8326.
1657 Nguyen, L-T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. 2015. IQ-TREE: A Fast and Effective
1658 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and
1659 Evolution 32: 268–274.
1660 Ogilvie, H. A., Bouckaert, R. R., Drummond, A. J. 2017. StarBEAST2 Brings Faster Species Tree
1661 Inference and Accurate Estimates of Substitution Rates. Molecular Biology and Evolution 34: 2101–
1662 2114.
1663 Olave, M., Sola E., Knowles L. L. 2014. Upstream analyses create problems with DNA-based species
1664 delimitation. Systematic Biology 63: 263–271.
1665 Oxelman, B., Yoshikawa, N., McConaughy, B. L., Luo, J., Denton, A. L., Hall, B. D. 2004. RPB2 Gene
1666 Phylogeny in Flowering Plants, with Particular Emphasis on Asterids. Molecular Phylogenetics and
1667 Evolution 32: 462–79.
1668 Pamilo, P., Nei, M. 1988. Relationships between gene trees and species trees. Molecular Biology and
1669 Evolution 5: 568–583.
1670 Park, S. D. E., Magee, D. A., McGettigan, P. A., Teasdale, M. D., Edwards, C. J., Lohan, A. J, Murphy,
1671 A., Braud, M., Donoghue, M. T., Liu, Y., Chamberlain, A. T, Rue-Albrecht, K., Schroeder, S.,
1672 Spillane, C., Tai, S., Bradley, D. G., Sonstegard, T. S., Loftus, B. J., McHugh, D. E. 2015. Genome
1673 sequencing of the extinct Eurasian wild aurochs, *Bos primigenius*, illuminates the phylogeography

and evolution of cattle. Genome Biology 16: 234.

Patel, S., Kimball., R.T., Braun, E. L. 2013. Error in Phylogenetic Estimation for Bushes in the Tree of Life. Journal of Phylogenetics and Evolutionary Biology 1: 110.

Payseur, B. A., Rieseberg, L. H. 2016. A genomic perspective on hybridization and speciation. Molecular Ecology 25: 2337–2360.

Pease, J. B., Haak, D. C., Hahn, M. W., Moyle, L. C. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PloS Biology 14: e1002379.

Pennell, M. W., FitzJohn, R. G., Cornwell, W. K. 2016. A simple approach for maximizing the overlap of phylogenetic and comparative data. Methods in Ecology and Evolution 7: 751–758.

Pennell, M. W., Harmon, L. J. 2013. An integrative view of phylogenetic comparative methods: Connections to population genetics, community ecology, and paleobiology. Annals of the New York Academy of Sciences 1289: 90–105.

Peterson, A. T., Moyle, R. G., Nyári, Á. S., Robbins, M. B., Brumfield, R. T., Remsen, J. V. Jr. 2007. The need for proper vouchering in phylogenetic studies of birds. Molecular Phylogenetics and Evolution 45: 1042–1044.

Pfeil, B. E., C. L. Brubaker, L. A. Craven, Crisp, M. D. 2004. Paralogy and orthology in the Malvaceae rpb2 gene family: Investigation of gene duplication in Hibiscus. Molecular Biology and Evolution 21: 1428–1437.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J, Manuel, M., Wörheide, G., Baurain, D. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biology 9: e1000602.

Philippe, H., de Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., Delsuc, F. 2017. Pitfalls in supermatrix phylogenomics. European Journal of Taxonomy 283: 1–25.

Piel, W. H., Chan, L., Dominus, M. J., Ruan, J., Vos, R. A., Tannen, V. 2009. Treebase v. 2: A Database of Phylogenetic Knowledge. e-Biosphere.

Pleijel, F., Jondelius, U., Norlinder, E., Nygren, A., Oxelman, B., Schander, C., Sundberg, P., Thollesson,

1700 M. 2008. Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic
1701 studies. Molecular Phylogenetics and Evolution 48: 369–371.
1702 Poe, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Systematic Biology 47: 18–31.
1703 Potts, A. J., Hedderson, T. A., Grimm, G. W. 2014. Constructing Phylogenies in the Presence Of Intra-
1704 Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron. Systematic
1705 Biology 63: 1–16.
1706 Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P. 2015. A comprehensive phylogeny of
1707 birds (Aves) using targeted next-generation DNA sequencing. Nature 526: 569–573.
1708 Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of
1709 Lissamphibia. Systematic Biology 60: 466–81.
1710 Pyron, R. A. 2015. Post-molecular systematics and the future of phylogenetics. Trends in Ecology &
1711 Evolution 30: 384–389.
1712 Pyron, R. A., Burbrink, F. T. 2012. Trait-dependent diversification and the impact of paleontological data
1713 on evolutionary hypothesis testing in New World ratsnakes (tribe Lampropeltini). Journal of
1714 Evolutionary Biology 25: 497–508.
1715 Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow,
1716 H. P., Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent,
1717 Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.
1718 Quintero, I., P. Keil., W. Jetz., Crawford, F. W. 2015. Historical Biogeography Using Species
1719 Geographical Ranges. Systematic Biology 64: 1059–1073.
1720 Rabosky, D. L. 2015. No substitute for real data: a cautionary note on the use of phylogenies from birth-
1721 death polytomy resolvers for downstream comparative analyses. Evolution 69: 3207–3216.
1722 Ramadugu, C., Pfeil, B. E., Manjunath, K. L., Lee, R. F., Maureira-Butler, I. J., Roose, M. L. 2013.
1723 Coalescence simulation testing of hybridization versus lineage sorting in *Citrus* (Rutaceae) using six
1724 nuclear genes. PLoS One 8: e68410.
1725 Rannala, B., Yang, Z. H. 2003. Bayes estimations of species divergence times and ancestral population

1726 sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.
1727 Rannala, B., Yang Z. H. 2008. Phylogenetic inference using whole genomes. Annual Review of
1728 Genomics and Human Genetics 9: 217–231.
1729 Ranwez, V., Delsuc, F. D. R., Ranwez, S., Belkhir, K., Tilak, M-K., Douzery, E. J. 2007. OrthoMaM: A
1730 database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary
1731 Biology 7: 241–12.
1732 Rasmussen, M. D., Hubisz, M. J., Gronau, I., Siepel, A. 2014. Genome-Wide Inference of Ancestral
1733 Recombination Graphs. PLoS Genetics 10(5): e1004342.
1734 Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K-L., Harshman,
1735 J., Huddleston, C. J., Kingston, S., Marks, B. D., Miglia, K. J., Moore, W. S., Sheldon, F. H., Witt, C.
1736 C., Yuri, T., Braun, E. L. 2017. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data
1737 Type Influences the Avian Tree of Life more than Taxon Sampling. Systematic Biology 66: 857–
1738 879.
1739 Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H.,
1740 Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M. N.,
1741 Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R.,
1742 Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A.,
1743 Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X.,
1744 Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., Hurles, M. E.
1745 2006. Global variation in copy number in the human genome. Nature 444: 444–454.
1746 Reid, N. M., Hird, S. M., Brown, J. M., Pelletier, T. A., McVay, J. D., Satler, J. D., Carstens, B. C. 2014.
1747 Poor fit to the multispecies coalescent is widely detectable in empirical data. Systematic Biology 63:
1748 322–333.
1749 Rhoads, A., Au, K. F. 2015. PacBio Sequencing and Its Applications. Genomics, Proteomics &
1750 Bioinformatics 13: 278–289.
1751 Ricketts, C., Popic, V., Toosi, H., Hajirasouliha, I. 2018. Using LICHeE and BAMSE for Reconstructing

1752 Cancer Phylogenetic Trees. Current Protocols in Bioinformatics 62: e49.
1753 Rogozin, I. B., Thomson, K., Csürös, M., Carmel, L., Koonin, E. V. 2008. Homoplasy in genome-wide
1754 analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of
1755 homologous series. Biology Direct 2008 3: 7.
1756 Rogozin, I. B., Wolf, Y. I., Babenko, B. N., Koonin, E. V. 2006. Dollo parsimony and the reconstruction
1757 of genome evolution. In: Albert VA ed. Parsimony, Phylogeny, and Genomics. Oxford University
1758 Press, p.p. 190–200.
1759 Rokas, A. 2005. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon
1760 Number to Phylogenetic Accuracy. Molecular Biology and Evolution 22: 1337–1344.
1761 Rokas, A. 2011. Phylogenetic Analysis of Protein Sequence Data Using the Randomized Axelerated
1762 Maximum Likelihood (RAXML) Program. Current Protocols in Molecular Biology 96: 19.11.1–
1763 19.11.14.
1764 Rokas, A., Holland, P. W. H. 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecology
1765 and Evolution 15: 454–459.
1766 Rokas, A., Williams, B. L., King, N., Carroll, S. B. 2003. Genome-scale approaches to resolving
1767 incongruence in molecular phylogenies. Nature 425: 798–804.
1768 Romiguier, J., Cameron, S. A., Woodard, S. H., Fischman, B. J., Keller, L., Praz, C. J. 2016.
1769 Phylogenomics Controlling for Base Compositional Bias Reveals a Single Origin of Eusociality in
1770 Corbiculate Bees. Molecular Biology and Evolution 33: 670–678.
1771 Romiguier, J., Roux, C. 2017. Analytical Biases Associated with GC-Content in Molecular Evolution.
1772 Frontiers in Genetics 8: 16.
1773 Roncal, J., Guyot, R., Hamon, P., Crouzillat, D., Rigoreau, M., Konan, O. N., Rakotomalala, J. J., Nowak,
1774 M. D., Davis, A. P., de Kochko, A. 2016. Active transposable elements recover species boundaries
1775 and geographic structure in Madagascan coffee species. Molecular Genetics and Genomics 291: 155–
1776 168.

1777 Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., Rasnitsyn, A. P. 2012. A
1778 total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera.
1779 Systematic Biology 61: 973–99.
1780 Rosenberg, N. A., Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic
1781 polymorphisms. Nature Reviews Genetics 3: 380–390.
1782 Rosindell, J., Cornell, S. J., Hubbell S. P., Etienne, R. S. 2010. Protracted speciation revitalizes the
1783 neutral theory of biodiversity. Ecology Letters 13: 716–727.
1784 Rosindell, J., Harmon, L. J. 2012. OneZoom: A Fractal Explorer for the Tree of Life. PLoS Biology
1785 10(10): e1001406.
1786 Rosindell, J., Harmon, L. J., Etienne, R. S. 2015. Unifying ecology and macroevolution with
1787 individual-based theory. Ecology Letters 18: 472–482.
1788 Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G. D., Longhorn, S. J., Peterson, K. J.,
1789 Pisani, D., Philippe, H., Telford, M. J. 2011. A congruent solution to arthropod phylogeny:
1790 phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proceedings of the
1791 Royal Society B 278: 298.
1792 Ruane, S., Austin, C. C. 2017. Phylogenomics using formalin-fixed and 100+ year-old intractable natural
1793 history specimens. Molecular Ecology Resources 17: 1003–1008.
1794 Ryan, J. F., Pang, K., Schnitzler, C. E., Nguyen, A-D., Moreland, R. T., Simmons, D. K., Koch, B. J.,
1795 Francis, W. R., Havlak, P., Comparative Sequencing Program NISC, Smith, S. A., Putnam, N. H.,
1796 Haddock, S. H. D., Dunn, C. W., Wolfsberg, T. G., Mullikin, J. C., Martindale, M. Q., Baxevanis, A.
1797 D. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type
1798 evolution, Science 342: 1242592.
1799 Sagitov, S., Bartoszek, K. 2012. Interspecies correlation for neutrally evolving traits. Journal of
1800 Theoretical Biology 309: 11–19.
1801 Sanderson, M. J. 1992. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A
1802 Penalized Likelihood Approach. Molecular Biology and Evolution 19: 101–109.

1803 Sanderson, M. J., Donoghue, M. J., Piel, W. H., Eriksson, T. 1994. TreeBASE: a prototype database of
1804 phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of
1805 Botany 81: 183.
1806 Sayyari, E., Mirarab, S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet
1807 Frequencies. Molecular Biology and Evolution 33: 1654–1668.
1808 Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A.,
1809 Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C.,
1810 Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E.
1811 V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P.,
1812 Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M.,
1813 Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S.,
1814 Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L.,
1815 Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E.,
1816 Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O.
1817 A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C.,
1818 Durbin, R. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:
1819 169–175.
1820 Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., Kosiol. C. 2016. Reversible polymorphism-
1821 aware phylogenetic models and their application to tree inference. Journal of Theoretical Biology
1822 407: 362–370.
1823 Schwartz, R., Schäffer, A. A. 2017. The evolution of tumour phylogenetics: Principles and practice.
1824 Nature Reviews Genetics 18: 213–229.
1825 Scotch, M., Sarkar, I. N., Mei, C., Leaman, R., Cheung, K-H., Ortiz, P., Singraur, A., Gonzalez, G. 2011.
1826 Enhancing phylogeography by improving geographical information from GenBank. Journal of
1827 Biomedical Informatics 44: S44–S47.
1828 Shen, X-X., Hittinger, C. T., Rokas, A. 2017. Contentious relationships in phylogenomic studies can be

1829 driven by a handful of genes. Nature Ecology and Evolution 1: 0126.
1830 Shi, C.-M., Yang, Z. 2018. Coalescent-Based Analyses of Genomic Sequence Data Provide a Robust
1831 Resolution of Phylogenetic Relationships among Major Groups of Gibbons. Molecular Biology and
1832 Evolution 35: 159–179.
1833 Siepel, A. 2009. Phylogenomics of primates and their ancestral populations. Genome Research 19: 1929–
1834 1941.
1835 Simion, P., Philippe, H., Baurain, D., Muriel, J., Richter, D. J., Di Franco, A., Roure, B., Satoh, N.,
1836 Quéinnec, E., Ereskovsky, A. 2017. A Large and Consistent Phylogenomic Dataset Supports
1837 Sponges as the Sister Group to All Other Animals. Current Biology 27: 958–967.
1838 Siu-Ting. K., Gower, D. J., Pisani, D., Kassahun, R., Gebresenbet, F., Menegon, M., Mengistu, A. A.,
1839 Saber, S. A., de Sá, R., Wilkinson, M., Loader, S. P. 2014. Evolutionary relationships of the
1840 Critically Endangered frog *Ericabatrachus baleensis* Largen, 1991 with notes on incorporating
1841 previously unsampled taxa into large-scale phylogenetic analyses. BMC Evolutionary Biology 14:44.
1842 Sjödin, P., Jakobsson, M. 2012. Population genetic nature of copy number variation. Population Genetic
1843 Nature of Copy Number Variation. In: Feuk L. (eds) Genomic Structural Variants. Methods in
1844 Molecular Biology (Methods and Protocols), vol 838. Springer, New York, NY.
1845 Slater, G. J. 2015 Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent
1846 trait evolution. Proceedings of the National Academy of Sciences USA 112: 4897– 4902.
1847 Smith, S. A., Brown, J. W. 2018. Constructing a broadly inclusive seed plant phylogeny. American
1848 Journal of Botany 105: 302–314.
1849 Smith, S. A., Moore, M. J., Brown, J. W., Yang, Y. 2015. Analysis of phylogenomic datasets reveals
1850 conflict, concordance, and gene duplications with examples from animals and plants. BMC
1851 Evolutionary Biology 15: 150.
1852 Smith, S. A., O'Meara, B. C. 2012. treePL: divergence time estimation using penalized likelihood for
1853 large phylogenies. Bioinformatics 28: 2689–2690.
1854 Solís-Lemus, C., Ané, C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under

1855 Incomplete Lineage Sorting. PLoS Genet 12: e1005896.
1856 Solís-Lemus, C., Knowles, L. L., Ané, C. 2015. Bayesian species delimitation combining multiple genes
1857 and traits in a unified framework. Evolution 69: 492–507.
1858 Solís-Lemus, C., Bastide, P., Ané, C. 2017. PhyloNetworks: A Package for Phylogenetic Networks.
1859 Molecular Biology and Evolution 34: 3292–3298.
1860 Song, S., Liu, L., Edwards, S. V., Wu, S. 2012. Resolving conflict in eutherian mammal phylogeny using
1861 phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of
1862 Sciences USA 112: E6079–E6079.
1863 Sousa, F., Bertrand, Y. J. K., Doyle, J. J., Oxelman, B., Pfeil, B. E. 2017. Using genomic location and
1864 coalescent simulation to investigate gene tree discordance in *Medicago* L. Systematic Biology 66:
1865 934–949.
1866 Springer, M. S., Gatesy, J. 2016. The gene tree delusion. Molecular Phylogenetics and Evolution 94: 1–
1867 33.
1868 Staats, M., Erkens, R. H. J., van de Vossenberg, B., Wieringa, J. J., Kraaijeveld, K., Stielow, B., Geml, J.,
1869 Richardson, J. E., Bakker, F. T. 2013. Genomic treasure troves: complete genome sequencing of
1870 herbarium and insect museum specimens. PLoS ONE 8: e69189.
1871 Stadler, T., 2009. On incomplete sampling under birth-death models and connections to the sampling-
1872 based coalescent. Journal of Theoretical Biology 261: 58–68.
1873 Stadler, T. 2013. Recovering speciation and extinction dynamics based on phylogenies. Journal of
1874 Evolutionary Biology 26: 1203–1219.
1875 Stadler, T., Steel, M., 2012. Distribution of branch lengths and phylogenetic diversity under
1876 homogeneous speciation models. Journal of Theoretical Biology 297: 33–40.
1877 Stamatakis, A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large
1878 Phylogenies. Bioinformatics 30: 1312–1313.
1879 Stephens, M., Smith, N.J., Donnelly, P. 2001. A new statistical method for haplotype reconstruction from
1880 population data. The American Journal of Human Genetics 68: 978–989.

1881 Stoltzfus, A., O'Meara, B., Whitacre, J., Mounce, R., Gillespie, E. L., Kumar, S., Rosauer, D. F., Vos R.
1882 A. 2012. Sharing and re-use of phylogenetic trees (and associated data) to facilitate synthesis. BMC
1883 Research Notes 5:574
1884 Struck, T. H. 2013. The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid
1885 Relationships. PLoS ONE 8: e62892.
1886 Suarez, A. V., Tutsui, N. D. 2004. The value of museum collections for research and society. BioScience
1887 54: 66–74.
1888 Suh, A., Paus, M., Kiefmann, M., Churakov, G., Franke, F. A., Brosius, J., Kriegs, J. O., Schmitz, J.
1889 2011. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nature
1890 Communications 2: 443.
1891 Suh, A., Smeds, L. A., Ellegren, H. 2015. The Dynamics of Incomplete Lineage Sorting across the
1892 Ancient Adaptive Radiation of Neoavian Birds. PLoS Biology 13: e1002224–18.
1893 Sukumaran, J., Knowles, L. L., 2017. Multispecies coalescent delimits structure, not species. Proceedings
1894 of the National Academy of Sciences USA 114: 1607–1612.
1895 Talavera, G., Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously
1896 aligned blocks from protein sequence alignments. Systematic Biology 56: 564–577.
1897 Tang, J., Moret, B. M. E., Cui, L., dePamphilis, C. W. 2004. Phylogenetic reconstruction from arbitrary
1898 gene-order data. in Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering
1899 592–599.
1900 Todd, E. V., Black, M. A., Gemmell, N. J. 2016. The power and promise of RNA-seq in ecology and
1901 evolution. Molecular Ecology 25: 1224–1241.
1902 Tonini, J. F. R., Beard, K. H., Ferreira, R. B., Jetz, W., Pyron, R. A. 2016. Fully-sampled phylogenies of
1903 squamates reveal evolutionary patterns in threat status. Biological Conservation 204: 23–31.
1904 Toprak, Z., Pfeil, B.E., Jones, G., Marcussen, T., Ertekin, A.S., Oxelman, B. 2016. Species Delimitation
1905 Without Prior knowledge: DISSECT Reveals Extensive Cryptic Speciation in the *Silene aegyptiaca*
1906 complex (Caryophyllaceae). Molecular Phylogenetics and Evolution 102: 1–8.

1907 Troudet, J., Vigness-Lebbe, R., Grandcolas, P., Legendre, F. 2018. The Increasing Disconnection of
1908 Primary Biodiversity Data from Specimens: How Does It Happen and How to Handle It? Systematic
1909 Biology 67: 1110–1119.
1910 Turney, S., Cameron, E. R., Cloutier, C. A, Buddle, C. M. 2015. Non-repeatable science: assessing the
1911 frequency of voucher specimen deposition reveals that most arthropod research cannot be verified.
1912 PeerJ 3: e1168–16.
1913 Ulitsky, I. 2016. Evolution to the rescue: using comparative genomics to understand long non-coding
1914 RNAs. Nature Reviews Genetics 17: 601–614.
1915 Urrutia, E., Chen, H., Zhou, Z., Zhang, N. R., Juang, Y. 2018. Integrative pipeline for profiling DNA
1916 copy number and inferring tumor phylogeny. Bioinformatics 34: 2126–2128.
1917 Villar, D., Berthelot, C., Alridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R.,
1918 Erichsen, J. T., Jasinska, A. J., Turner, J. M., Bertelsen, M. F., Murchison, E. P., Flicek, P., Odom, D.
1919 T. 2015. Enhancer evolution across 20 mammalian species. Cell 160: 554–566.
1920 Vision, T. 2010. Open data and the social contract of scientific publishing. *BioScience* 60: 330.
1921 Warnow, T. 2018. Supertree Construction: Opportunities and Challenges. arXiv:1805.03530v1.
1922 Wen, D., Nakhleh L. 2018. Co-estimating Reticulate Phylogenies and Gene Trees from Multi-locus
1923 Sequence Data. Systematic Biology 67: 439–457.
1924 Wen, D., Yu Y., Hahn M. W., Nakhleh L. 2016. Reticulate evolutionary history and extensive
1925 introgression in mosquito species revealed by phylogenetic network analysis. Molecular Ecology 25:
1926 2361–2372.
1927 Wesche, P. L., Gaffney, D. J., Keightley, P. D. 2004. DNA Sequence Error Rates in Genbank Records
1928 Estimated Using the Mouse Genome as a Reference. DNA Sequence 15(5–6): 362–64.
1929 Wiedenhoeft, J., Brugel, E., Schliep, A. 2016. Fast Bayesian Inference of Copy Number Variants using
1930 Hidden Markov Models with Wavelet Compression. PLoS Computational Biology 12(5): e1004871.
1931 Will, K. P., Mishler, B. D., Wheeler, Q. D. 2005. The perils of DNA Barcoding and the need for
1932 integrative taxonomy. Systematic Biology 54: 844–851.

1933 Wilson, G., Aruliah, D. A., Brown, C. T., Chue-Hong N. P., Davis, M., Guy, R. T., Haddock, S. H. D,
1934 Huff, K. D., Mitchell, I. M., Plumbley, M. D, Waugh, B., White, E. P., Wilson, P. 2014. Best
1935 Practices for Scientific Computing. PLoS Biology 12(1): e1001745.
1936 Xi, Z., Ruhfel, B. R., Schaefer, H., Amorim, A. M., Sugumaran, M., Wurdack, K. J., Endress, P. K.,
1937 Matthews, M. L., Stevens, P. F., Mathews, S., Davis, C. C. 2012. Phylogenomics and a posteriori
1938 data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proceedings of the
1939 National Academy of Sciences USA 109: 17519–17524.
1940 Xi, Z., Liu, L., Davis, C. C. 2015. Genes with minimal phylogenetic information are problematic for
1941 coalescent analyses when gene tree estimation is biased. Molecular Phylogenetics and Evolution 92:
1942 63–71.
1943 Xu, B., Yang, Z. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model.
1944 Genetics 204: 1353–1368.
1945 Yang Z, Rannala B, 2014. Unguided species delimitation using DNA sequence data from multiple loci.
1946 Molecular Biology and Evolution 31: 3125–3135.
1947 Yu, Y., Dong J., Liu K. J., Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary
1948 histories. Proceedings of the National Academy of Sciences USA 111: 16448–16453.
1949 Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A, FitzJohn, R. G., McGlinn, D. J.,
1950 O'Meara, B. C., Moles, A. T., Reich, P. B., Royer, D. L., Soltis, D. E., Stevens, P. F., Westoby, M.,
1951 Wright, I. J., Aarssen, L., Bertin, R. I. Calaminus, A., Govaerts, R., Hemmings, F., Leishman, M. R.,
1952 Oleksyn, J., Soltis, P. S., Swenson, N. G., Warman, L., Beaulieu, J. M. 2014. Three keys to the
1953 radiation of angiosperms into freezing environments. Nature 506: 89–92.
1954 Zhang, C., Ogilvie, H. A., Drummond, A. J., Stadler, T. 2018. Bayesian Inference of Species Networks
1955 from Multilocus Sequence Data. Molecular Biology and Evolution 35: 504–517.
1956
1957

**Figures**

Figure 1. *A posteriori* marker selection from whole genome alignments for phylogenomics and phylogeography. Whole genome analysis (A) permits researchers to choose different markers for specific purposes (B – D). By contrast, subsampling methods such as Rad-seq or hybrid capture, which dominate phylogenomics today, usually yield a specific set of markers that the researcher has chosen *a priori*. The generation of WGA thus greatly increases the use of genomic data in biological research, beyond the initial goals of the researcher producing those data. Here, we show how a hypothetical WGA that includes seven different loci (different colors) for four individuals allows extracting sequence data to generate gene trees (B), identifying SNPs to genotype individuals (C), and measuring copy depth to infer CNVs across genomic regions (D). Ultimately, these different kinds of data can be translated into species tree inferences (B – D). In the case of CNVs, only locus number 3 (orange) shows significant CNV. Because CNVs are measured as continuous characters (i.e., copy depth), the orange shading represents a hypothetical evolutionary scenario of copy number variation of genomic region number 3 within the inferred species tree, which is incongruent with those based on sequence and SNP data from other loci in the genome.

Figure 2. Trends in phylogenomic data sets since the emergence of HTS. Based on a sample of 164 phylogenomic papers published since 2004 (see Supplementary Table S1), we observed no increase in the number of species per data set over time (A). On the other hand, there is a significant increase in the number of loci (B), total alignment length (C), and total data set size, as measured by the product of species times locus number (Data set size 1, E) and species times total alignment length (Data set size 2, F). Moreover, the advent of HTS does not support the notion of a tradeoff between the number of species and the number of loci in phylogenomic studies.

Figure 3. Some examples of violations of the multispecies coalescent. In event A, there is gene flow; in event B there is homoploid hybridization; in event C, there is a gene duplication; and in event D, incomplete lineage sorting. All of these processes contribute to gene tree heterogeneity but fall outside the standard multispecies coalescent model. Importantly, all of these processes also yield strictly dichotomous gene trees, whereas recombination (not illustrated here) does not.

Figure 4. Gene duplication and loss creates patterns that can mimic incomplete lineage sorting and can lead to spurious inferences of the species history. Genes and genomes of three species A, B, and C. Multi-colored bars show (parts of) their genomes with a number of loci indicated in different colors. The orange gene is duplicated in species A and it was lost in species B. The blue gene was duplicated before the divergence between species A and the ancestor of species B and C. However, one of these copies was lost in species A, whereas both copies were maintained in species B and C. Reconstruction of the orange gene tree based on extant diversity will yield a wrong inference of its history due to the absence of data for species B. On the other hand, a phylogenetic reconstruction of the blue gene is difficult to predict. Depending on which of the duplicates are sampled for species B and C, different outcomes can be expected regarding the relationship among the three species. The duplication and loss history of these two genes may cause serious issues for phylogenetic reconstruction because no specific pattern can be expected between them.

Figure 5. Complex patterns of gene lineages with polyploidization and interspecific gene flow. Genes and genomes of four species A, B, C and D. Multi-colored bars show (parts of) genomes with a number of loci indicated in different colors. Two gene trees, one orange and one blue, evolve within the species network. Species B is an allopolyploid containing two genomes.

Figure 6. Gradual speciation or isolation-with migration. After starting to split, gene flow between species decreases gradually. Such a gradual decrease in the extent of gene flow between species might present an

especially useful extension of the standard multispecies coalescent model. Colors depict different gene pools and their gradual change along branches describes how species gradually differentiate despite the existence of migration over time. Thickness and color intensity of arrows show that gene flow becomes weaker as species gradually isolate.

Figure 7. Two possible species phylogenies producing similar observations at present time. On the left (A), there is a species tree with gene flow. On the right (B), there is a species network with homoploid hybridization. Distinguishing two such scenarios usually requires simulations and comparison of observed and expected summary statistics.

## Figure 1(on next page)

*A posteriori* marker selection from whole genome alignments for phylogenomics and phylogeography.

Whole genome analysis (A) permits researchers to choose different markers for specific purposes (B – D). By contrast, subsampling methods such as Rad-seq or hybrid capture, which dominate phylogenomics today, usually yield a specific set of markers that the researcher has chosen *a priori*. The generation of WGA thus greatly increases the use of genomic data in biological research, beyond the initial goals of the researcher producing those data. Here, we show how a hypothetical WGA that includes seven different loci (different colors) for four individuals allows extracting sequence data to generate gene trees (B), identifying SNPs to genotype individuals (C), and measuring copy depth to infer CNVs across genomic regions (D). Ultimately, these different kinds of data can be translated into species tree inferences (B – D). In the case of CNVs, only locus number 3 (orange) shows significant CNV. Because CNVs are measured as continuous characters (i.e., copy depth), the orange shading represents a hypothetical evolutionary scenario of copy number variation of genomic region number 3 within the inferred species tree, which is incongruent with those based on sequence and SNP data from other loci in the genome.

A

Whole-genome alignments

Ind. 1
Ind. 2
Ind. 3
Ind. 4

COPY 1
COPY 2
COPY 3

Selected markers

B

Sequence data

Ind. 1
Ind. 2
Ind. 3
Ind. 4

C

SNPs

Ind. 1
Ind. 2
Ind. 3
Ind. 4

A / G
T / C
A / T
G / T
A / C

D

Genome-level characters (e.g., CNVs)

Copy depth

Ind. 4
Ind. 3
Ind. 2
Ind. 1

Genome position

Sp. 1 Sp. 2 Sp. 3 Sp. 4

Sp. 1 Sp. 2 Sp. 3 Sp. 4

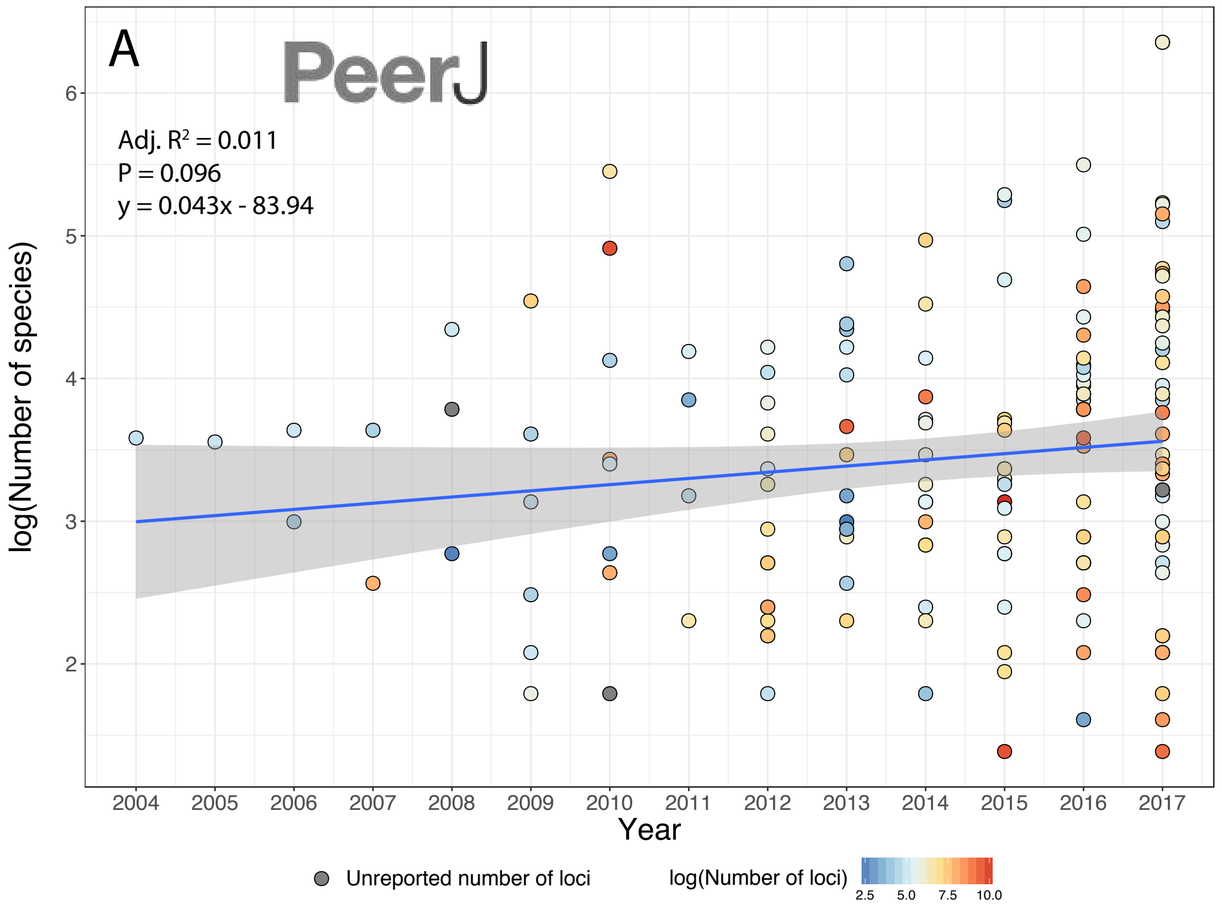
Sp. 1 Sp. 2 Sp. 3 Sp. 4

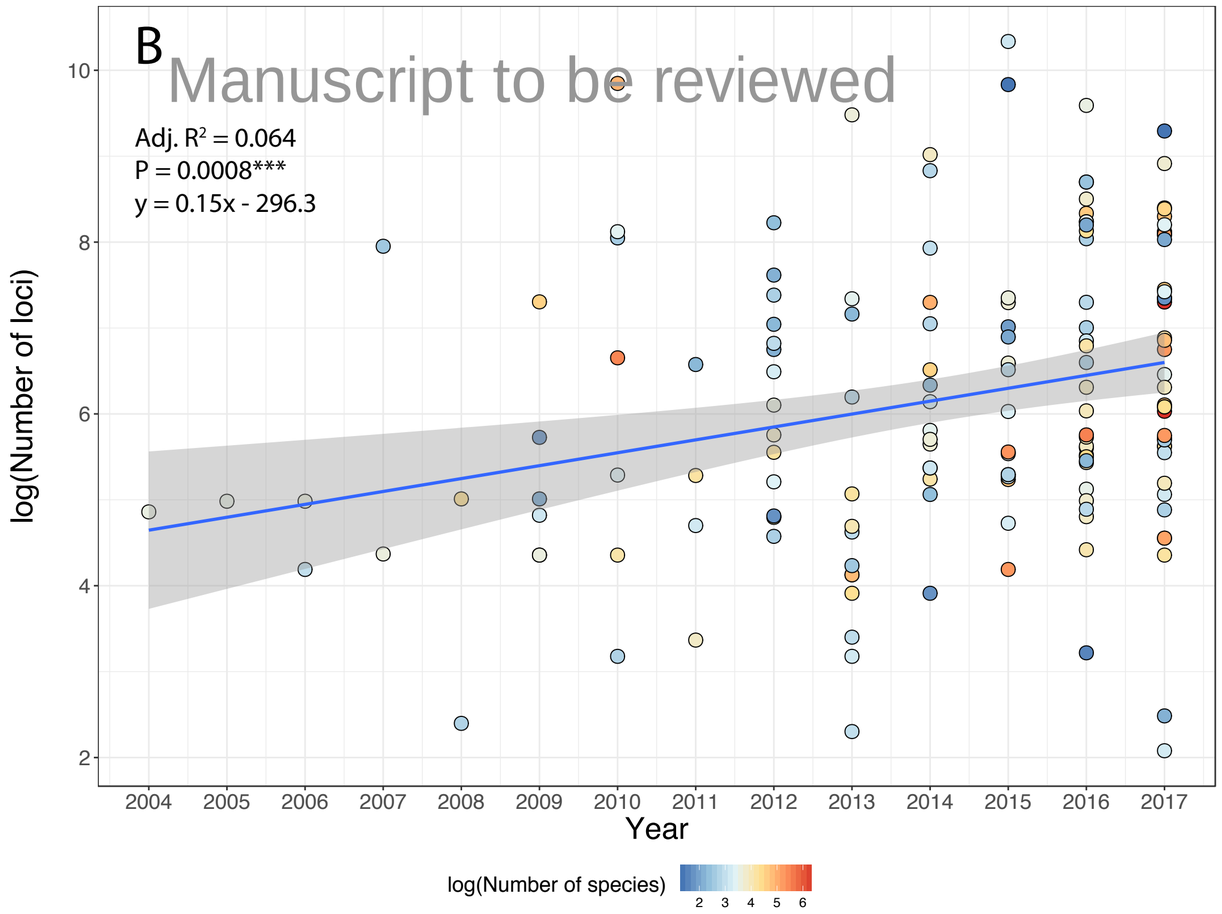
## Figure 2(on next page)

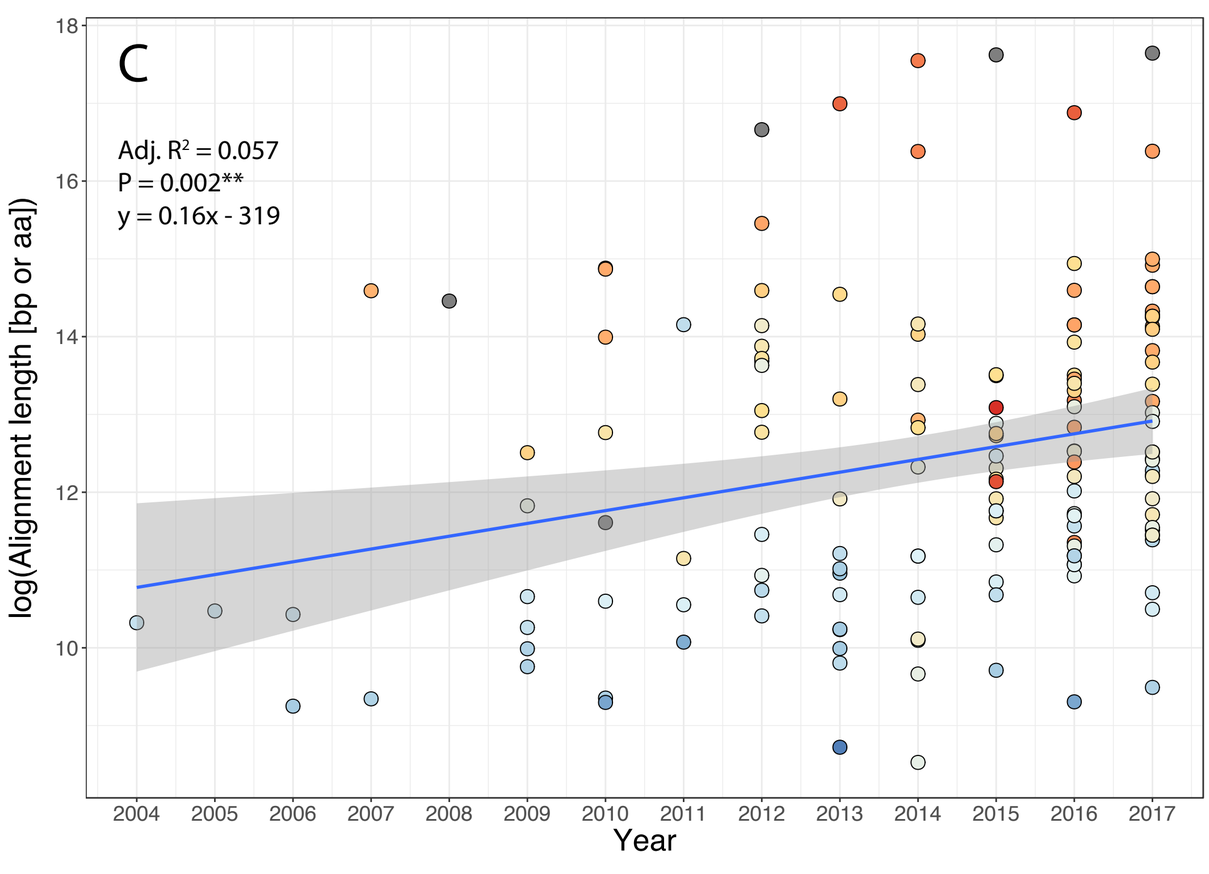
Trends in phylogenomic data sets since the emergence of HTS.

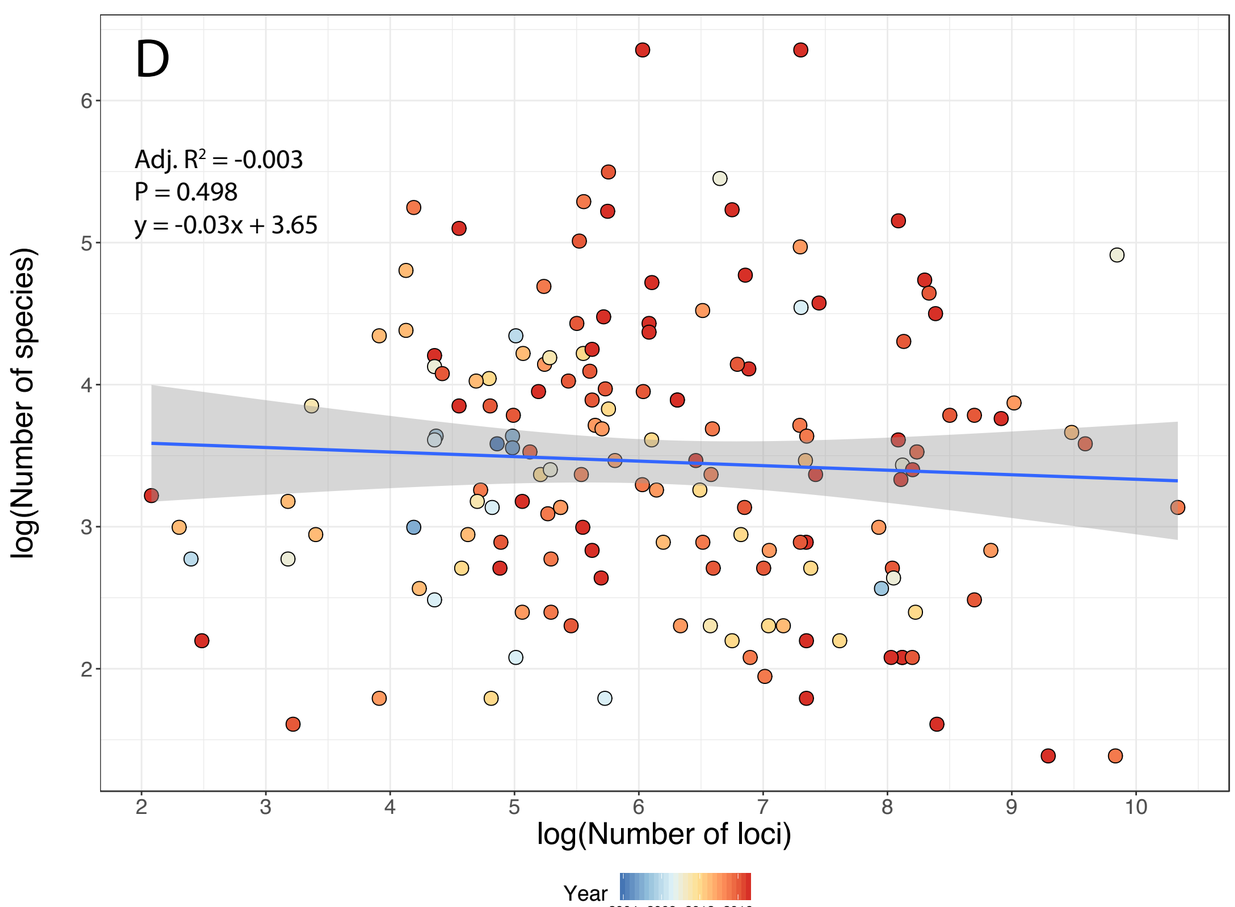
Based on a sample of 164 phylogenomic papers published since 2004 (see Supplementary Table S1), we observed no increase in the number of species per data set over time (A). On the other hand, there is a significant increase in the number of loci (B), total alignment length (C), and total data set size, as measured by the product of species times locus number (Data set size 1, E) and species times total alignment length (Data set size 2, F). Moreover, the advent of HTS does not support the notion of a tradeoff between the number of species and the number of loci in phylogenomic studies.

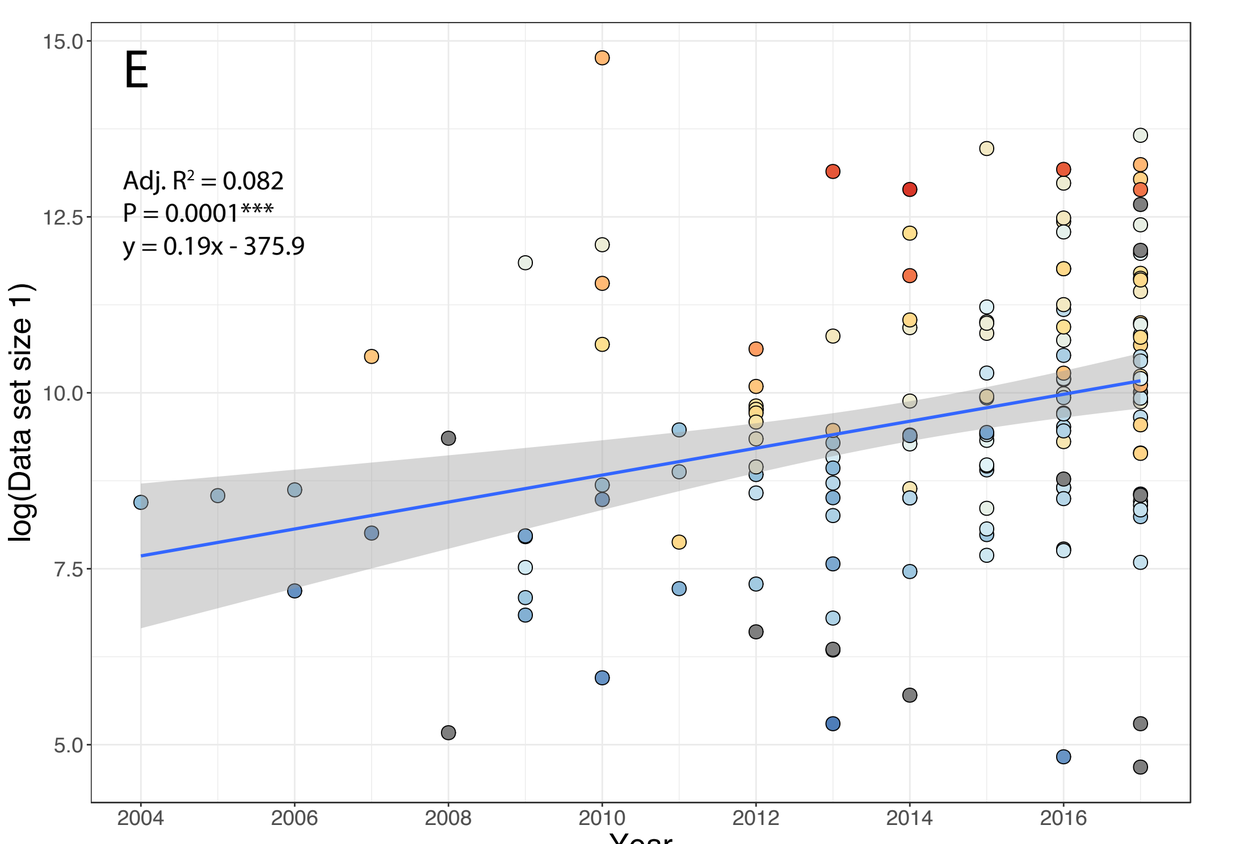
PeerJ
Manuscript to be reviewed

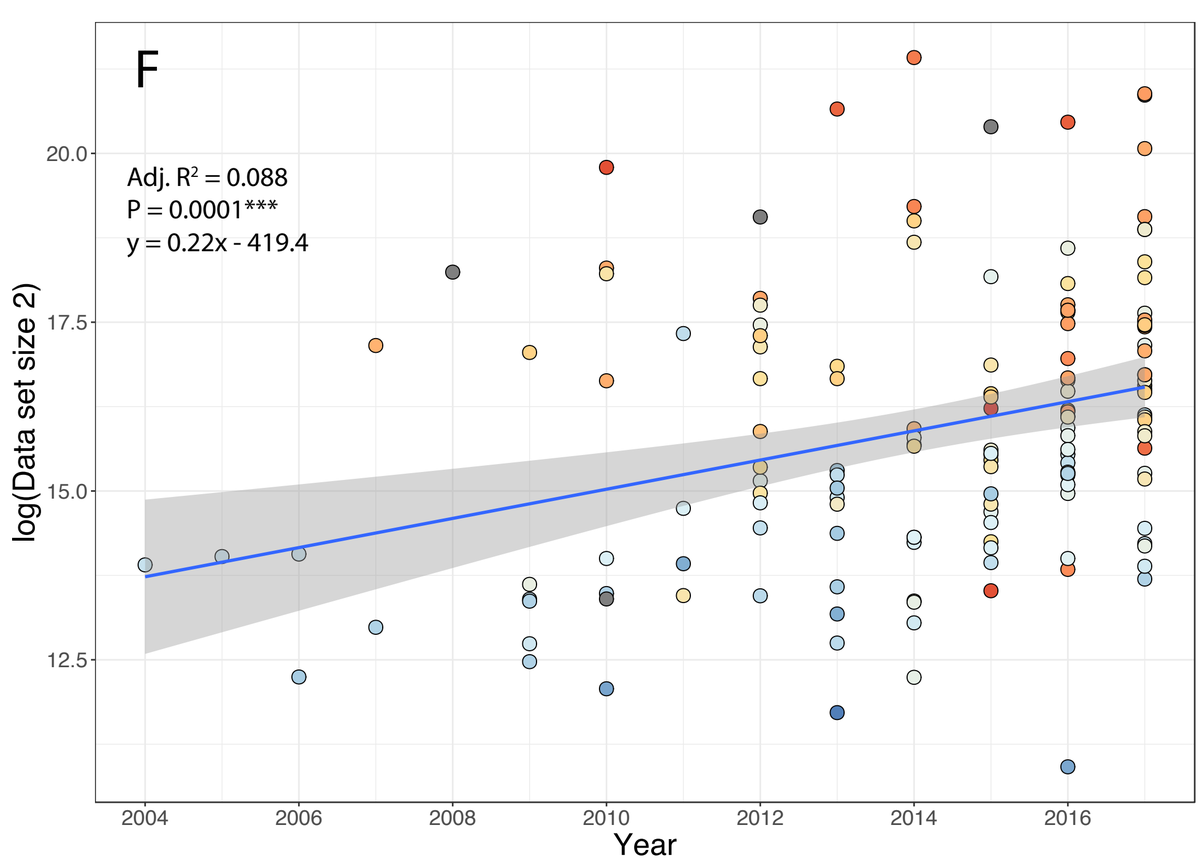


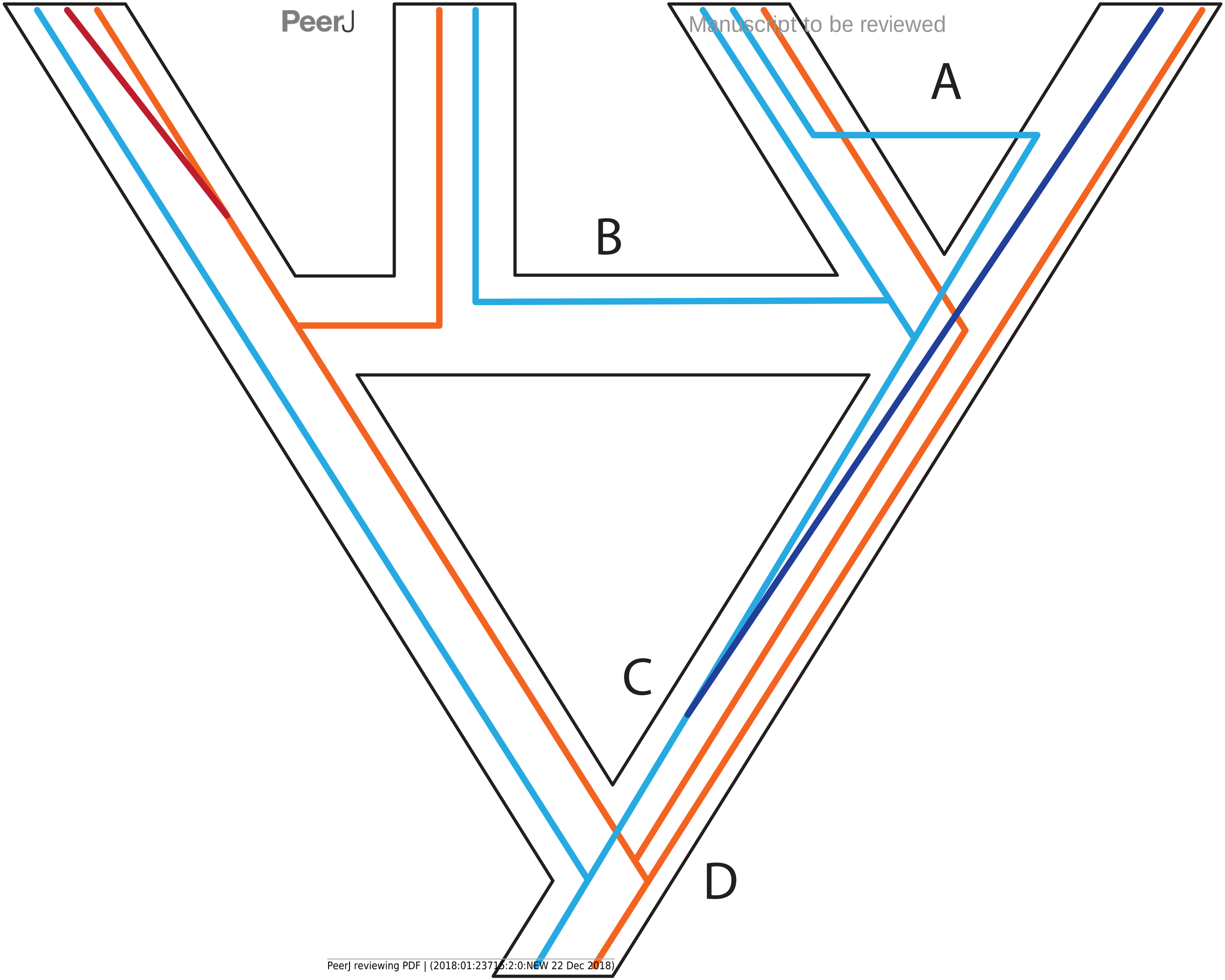



## Figure 3(on next page)

Some examples of violations of the multispecies coalescent.

In event A, there is gene flow; in event B there is homoploid hybridization; in event C, there is a gene duplication; and in event D, incomplete lineage sorting. All of these processes contribute to gene tree heterogeneity but fall outside the standard multispecies coalescent model. Importantly, all of these processes also yield strictly dichotomous gene trees, whereas recombination (not illustrated here) does not.

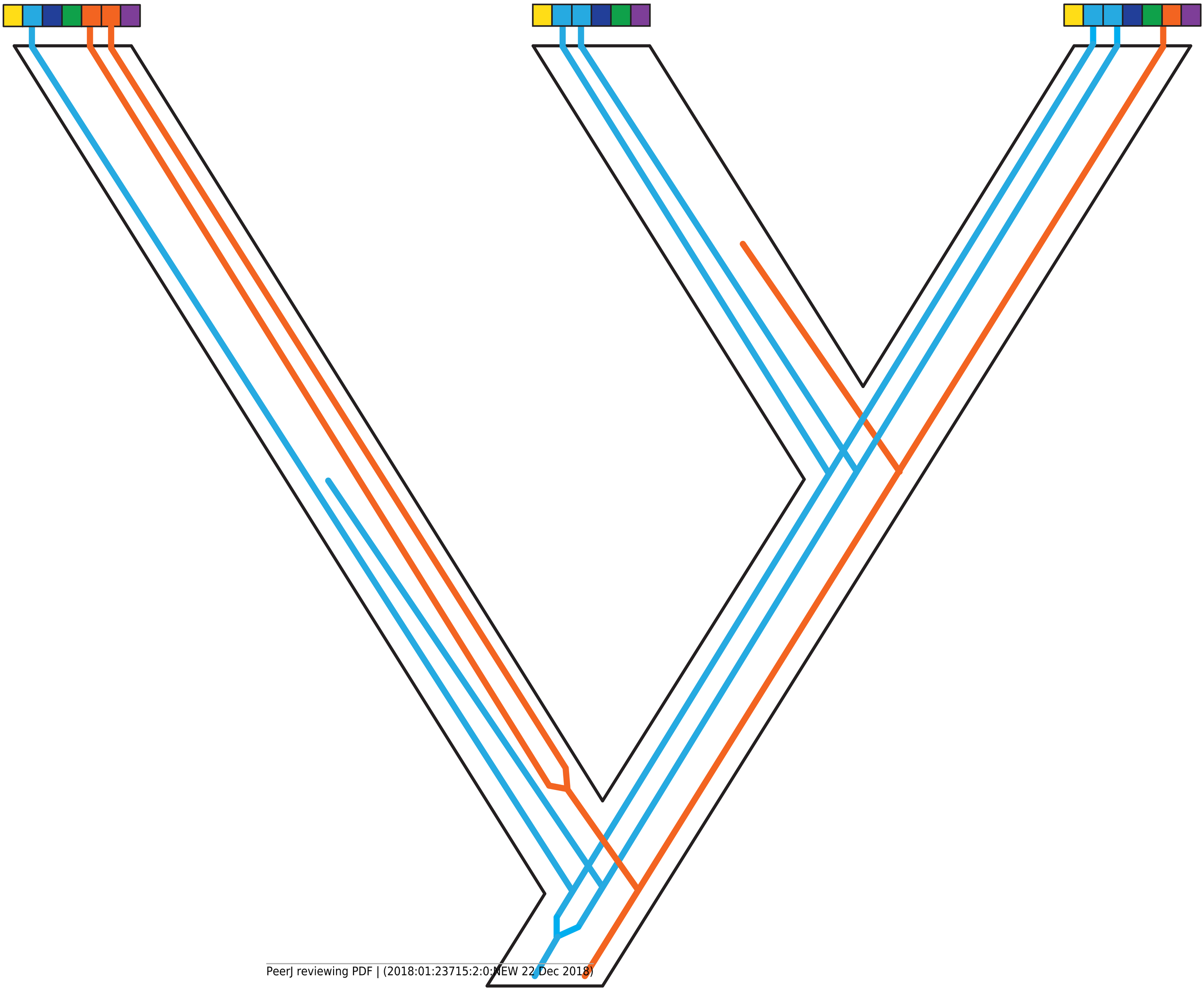
Manuscript to be reviewed



## Figure 4(on next page)

Gene duplication and loss creates patterns that can mimic incomplete lineage sorting and other processes. can lead to spurious inferences of the species history.

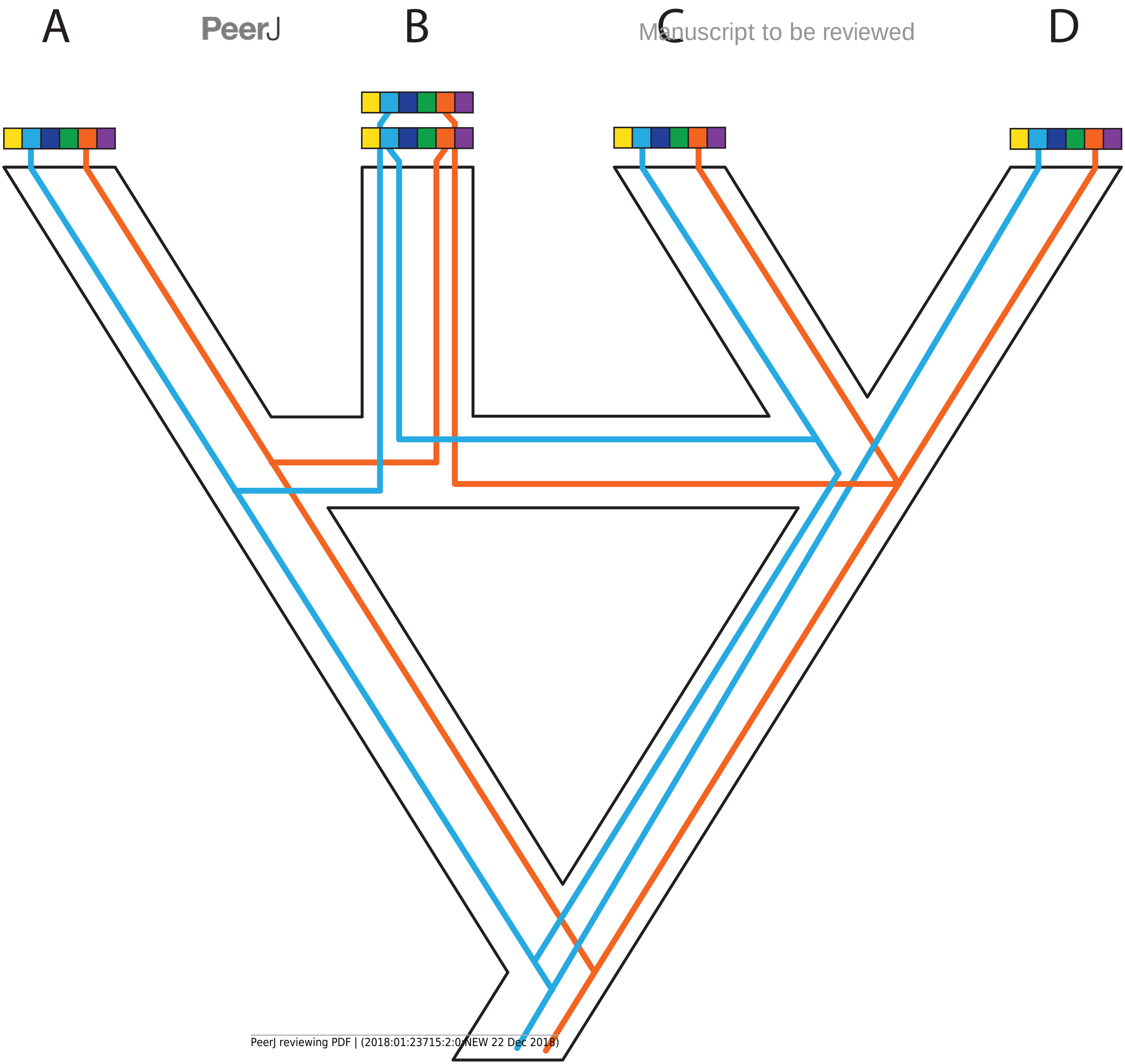
Genes and genomes of three species A, B, and C. Multi-colored bars show (parts of) their genomes with a number of loci indicated in different colors. The orange gene is duplicated in species A and it was lost in species B. The blue gene was duplicated before the divergence between B and C. However, both copies are maintained in species B and only one copy persists in species C. species A and the ancestor of species B and C. However, one of these copies was lost in species A, whereas both copies were maintained in species B and C. Reconstruction of the orange gene tree based on extant diversity will yield a wrong inference of its history due to the absence of data for species B. On the other hand, a phylogenetic reconstruction of the blue gene is difficult to predict. Depending on which of the duplicates are sampled for species B and C, different outcomes can be expected regarding the relationship among the three species. The duplication and loss history of these two genes may cause serious issues for phylogenetic reconstruction because no specific pattern can be expected between them.



## Figure 5(on next page)

Complex patterns of gene lineages with polyploidization and interspecific gene flow.

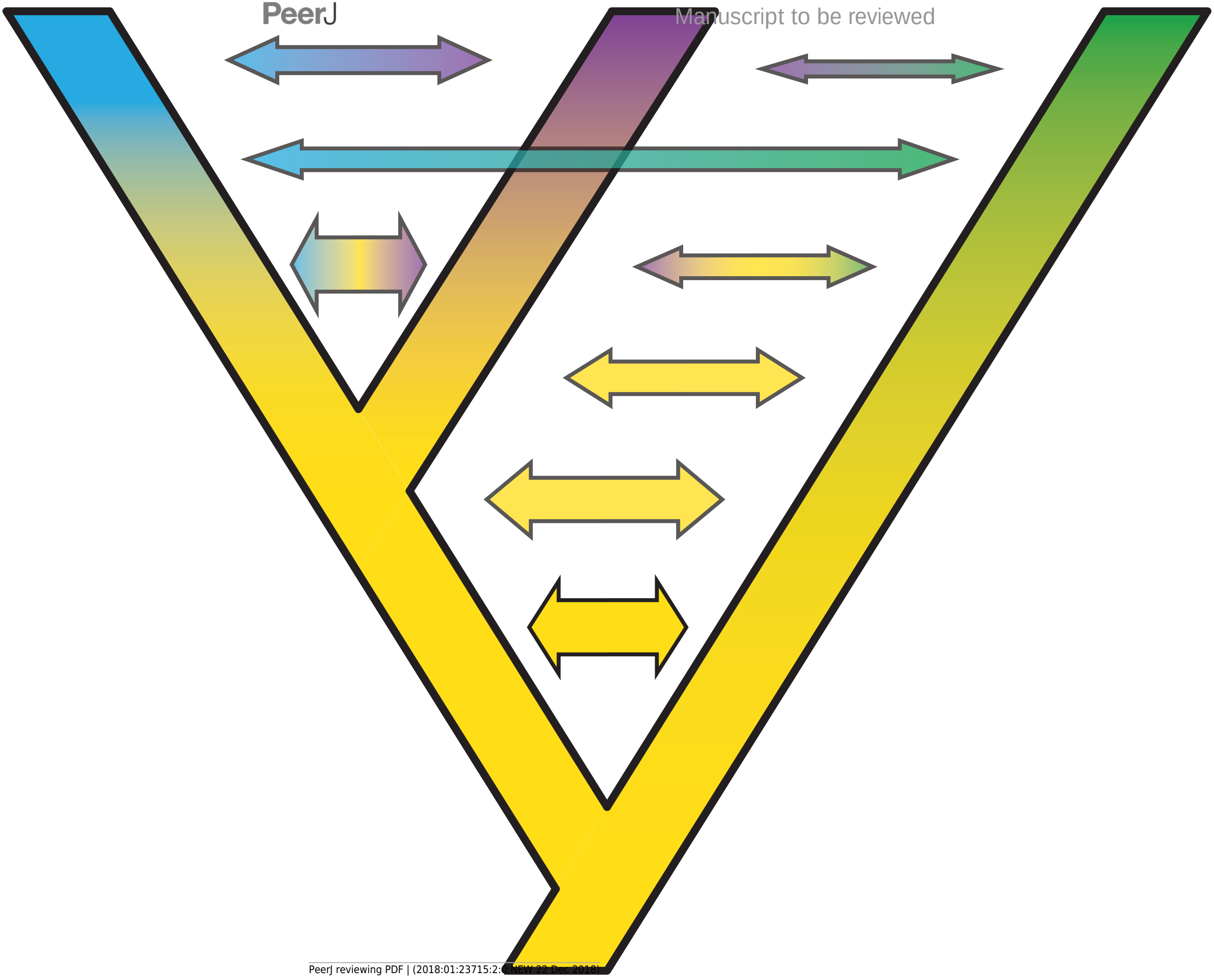
Genes and genomes of four species A, B, C and D. Multi-colored bars show (parts of) genomes with a number of loci indicated in different colors. Two gene trees, one orange and one blue, evolve within the species network. Species B is an allopolyploid containing two genomes.


A
B
C
D

## Figure 6(on next page)

Gradual speciation, or isolation-with migration.

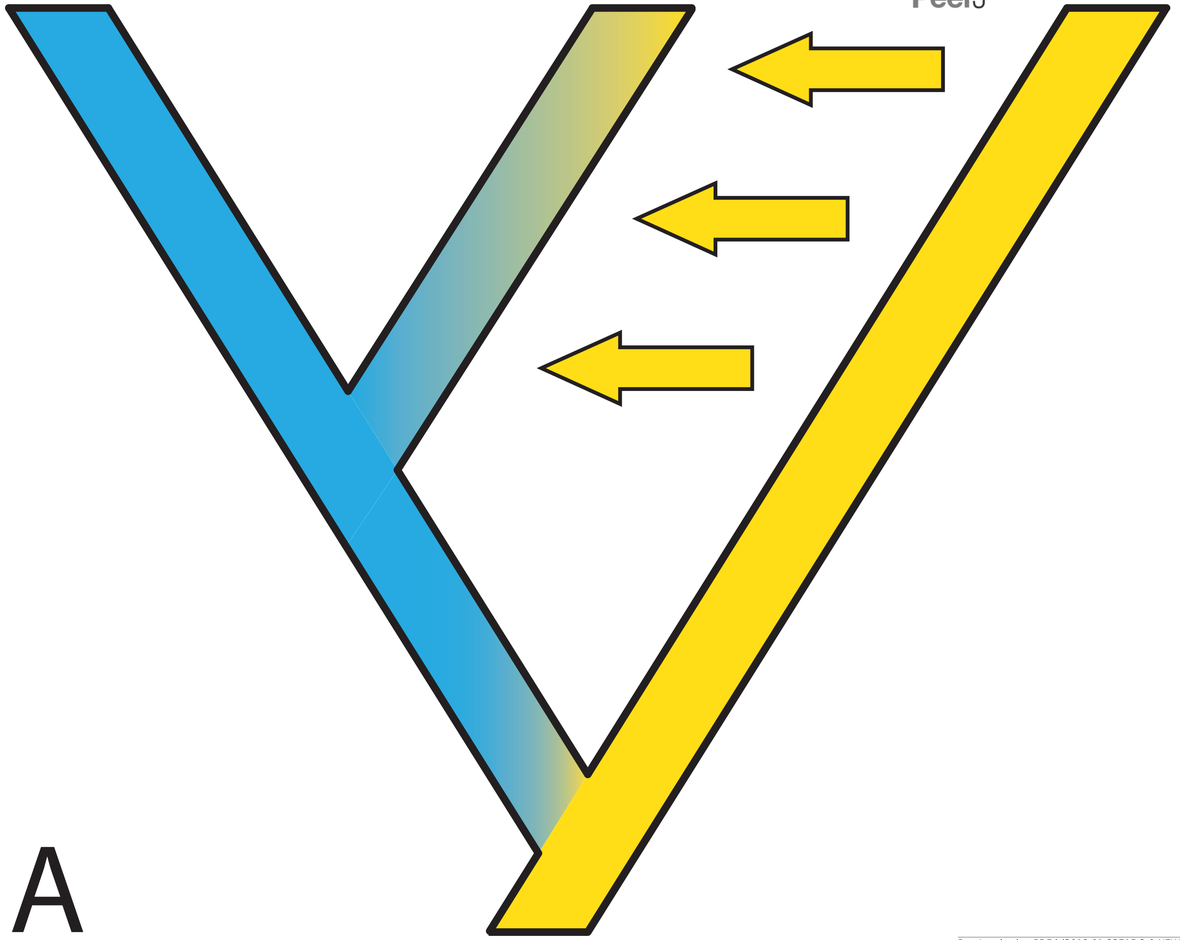
After starting to split, gene flow between species decreases gradually. Such a gradual decrease in the extent of gene flow between species might present an especially useful extension of the standard multispecies coalescent model. Colors depict different gene pools and their gradual change along branches describes how species gradually differentiate despite the existence of migration over time. Thickness and color intensity of arrows show that gene flow becomes weaker as species gradually isolate.



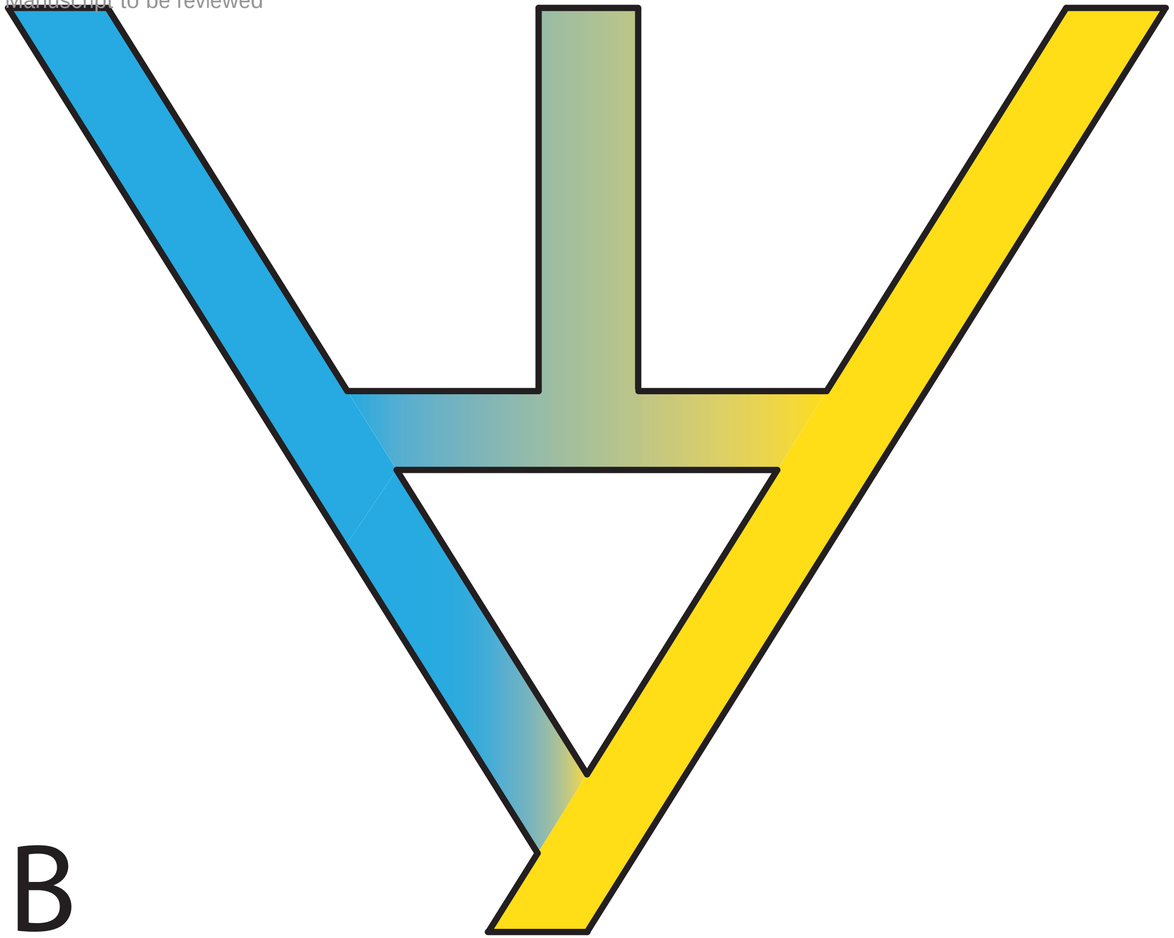
## Figure 7 (on next page)

Two possible species phylogenies producing similar observations at present time.

On the left (A), there is a species tree with gene flow. On the right (B), there is a species network with homoploid hybridization. Distinguishing two such scenarios usually requires simulations and comparison of observed and expected summary statistics.



A



B

## Table 1(on next page)

Definitions of core concepts used in this article

1 **Table 1. Definitions of core concepts used in this article**
2

| Concept | Definition |
|---|---|
| **ToL**<br>The Tree of Life | This idea, originally articulated by Darwin and others, refers to the grand vision of understanding the branching pattern of all life on earth. Today the idea conveys the use of morphological and molecular data to reconstruct the phylogenetic relationships of all life forms. In some usages, the idea also includes reconstructing reticulate evolutionary events, such as introgression and hybridization, which are now thought to be common in many lineages. |
| **HTS**<br>High-throughput sequencing | Also referred to as "next generation sequencing", this term refers to the plethora of new DNA and RNA sequencing technologies that in the last fifteen years have allowed biologists to dramatically increase the number of bases sequenced for a given species or clade. HTS technologies can be applied to sequencing whole genomes or transcriptomes and have been embraced by phylogeneticists interested in increasing the size of comparative molecular data sets. See Goodwin, McPherson & McCombie (2016) for a review on the progress of HTS. |
| **MSC**<br>The multispecies coalescent model | A generalization of the standard, single population coalescent model to multiple species related in a phylogeny. The MSC applies the single-population coalescent model to each branch of a phylogenetic tree, including both terminal and internal branches. In the MSC, alleles sampled in terminal species will coalesce to a smaller number of ancestral alleles at a rate depending on the effective population size within the branch. The gene tree lineages in a branch of the species tree do not necessarily coalesce within that branch as one goes backwards in time; multiple alleles may persist into ancestral branches. This phenomenon is called incomplete lineage sorting (see next definition). The decrease in the number of alleles and the time to coalescence to a single allele in a lineage follows the standard neutral coalescent model, until all alleles coalesce from all species. See Rannala & Yang (2003) and Degnan & Rosenberg (2009) for a full discussion. |
| **ILS**<br>Incomplete lineage sorting | This phenomenon, originally described by John Avise (see Avise et al. 1987) refers to the tendency of alleles in an ancestral species to persist across multiple speciation events, resulting in a situation in which the gene tree ("allele tree") differs from the species tree. In ILS, alleles fail to "sort" by genetic drift as species diverge from one another, resulting in different species retaining the same alleles, or their descendants, causing discordance with the overarching species or population tree. The language of this phrase is looking forward in time, as opposed to the language of coalescence, which looks backwards in time. See Degnan & Rosenberg 2009 for a full discussion. |

| | |
|---|---|
| **GDL**<br>Gene duplication and loss | This concept describes the process by which a gene in an ancestral species can duplicate, forming paralogs and one or more of the paralogs can subsequently be deleted from the genome, resulting in complex patterns of relationships among paralogs and orthologs. Gene duplication is another mechanism, in addition to ILS, that can render the gene tree different from the species tree. As a result of gene (paralog) loss, inferring the correct ortholog/paralog relationships and history of branching events in a multigene family can be challenging. Phylogenetic models incorporating GDL try to use patterns in multigene families to deduce the branching history of the constituent species. See Degnan & Rosenberg (2009) and Souza et al., (2017) for a full discussion. |
| **ARG**<br>Ancestral recombination graph | This is a complete record of the coalescent and recombination events in the history of a set of DNA sequences. As a consequence of incorporating recombination events, ARGs do not necessarily depict trees, but often have a network structure. The accurate estimation of ARGs remains challenging but they enhance our ability to estimate recombination rates, ancestral effective population sizes, population divergence times, rates of gene flow between populations, and detect selective sweeps. See Griffiths & Marjoram (1996), Siepel (2009), and Rasmussen et al., (2014) for a full discussion. |
| **LGT**<br>Lateral gene transfer | This process occurs when genes jump taxonomic and phylogenetic boundaries, moving between unrelated species and therefore causing discordances between genetic and lineage history. LGT, along with ILS and GDL was among the three primary causes of discordance between gene and species trees identified by Maddison (1997). LGT has been documented to occur between bacterial lineages, between bacteria and viruses, and between these two and eukaryotes, including plants and vertebrates. If not identified prior to phylogenetic analysis, LGT can cause many algorithms for phylogenetic inference to fail. Without prior identification, LGT essentially amounts to errors in data sets and sequence alignments. At the same time, LGT can be a source of adaptation and evolutionary novelty for recipient genomes and has had a major impact on the history of life. See Gogarten & Townsend (2005) for a full review. |

## Table 2(on next page)

Challenges in the fields of coalescent-based phylogenomics and implications for unraveling character evolution and the Tree of Life.

**Table 2. Challenges in the fields of coalescent-based phylogenomics and implications for unraveling character evolution and the Tree of Life.**

| Category | Challenge | Proposed strategy |
|---|---|---|
| Data | Integration and assessment of large amounts of data with heterogeneous phylogenetic signal. | Protocols for marker selection should assess markers' biological relevance and adequacy for the study organism, given the temporal and spatial scales in question, and not only logistical convenience. *A posteriori* (after data generation) marker selection from whole-genome alignments can be useful to inform these aspects as well as minimize the effects of missing data and varying data quality. Until then, researchers should attempt a higher standardization of markers to facilitate combinatory analyses.<br><br>To discern true phylogenomic heterogeneity from noise and error as well as to identify violations of the MSC, adequate filtering of large phylogenomic datasets should be conducted based on biological and statistical properties of markers (e.g., analyses of gene-tree outliers and rogue taxa).<br><br>Further research on filtering methods as well as on their impact on phylogenomic estimation is still required. |
| | Inclusion of additional character types into phylogenomic analyses. | Research efforts focused on the adequate identification and utilization of rare genomic changes other than nucleotide substitutions, such as indels, transpositions, inversions, CNVs, and chromosomal rearrangements. Development of new methods not only to infer phylogenetic hypotheses based on these characters but also to integrate them with more traditional sequence data. |
| Phylogenetic inference models and methods | Analyses of genome-scale data for large numbers of species within a coalescent framework. | Continue the development of models and methods that allow simultaneous gene tree and species tree estimation within a Bayesian framework (e.g., Ogilvie, Bouckaert & Drummond, 2017) for increasingly large and complex datasets.<br><br>For the time being, two-step methods, particularly those based on biological models and permitting statistical tests |

| | | |
|---|---|---|
| | | of topologies in a likelihood framework (e.g., Liu et al., 2010), are useful tools to incorporate coalescent information into species tree inference. |
| | Detection and incorporation of violations of the MSC into phylogenomic inferences. | Extensions of the MSC should seek the inference of reticulate evolutionary histories (i.e., multispecies network coalescent; Yu et al., 2014; Wen, Yu & Nakhleh, 2016) by simultaneously incorporating violations of the MSC (e.g., Wen & Nakhleh, 2018 [reticulation and ILS]; Jones, Sagitov & Oxelman, 2013 [allopolyploidy]). Inference methods dealing with GDL and recombination are of high priority.<br><br>Further development of conceptual approaches aimed at detecting and quantifying different underlying biological processes of phylogenetic history (e.g., Jones 2018 [ILS and migration], Blischak et al., 2018 [hybridization], Hibbins & Hahn, 2018 [direction and timing of introgression], Sousa et al., 2017 [ILS and GDL], Li et al., 2018 [recombination rates]). As proposed by Degnan (2018), using the MSC as a null model within a model selection approach can be a powerful tool to identify violations of the MSC and to deepen our understanding of the biological consequences of these processes. |
| Models that integrate phylogenomics and comparative analyses | Integrating different phylogenetic signals into comparative analyses. | Methods and models should attempt to incorporate gene tree incongruence into macroevolutionary models of character evolution. Similarly, integrative studies aiming at unraveling character evolution at the molecular, phenotypic, and clade levels. |
| | Understanding the genomic bases of character evolution in species trees vs. gene trees. | Methods that estimate phenotype-genotype associations incorporating heterogeneity across gene trees or that at least take into account differential state probabilities stemming from gene tree discordance (e.g., Guerrero & Hahn, 2018). Similarly, extensions of the MSC for quantitative traits that take into account genealogical heterogeneity represent a promising avenue for research and implementation (e.g., Mendes et al., 2018). |

| | | |
|---|---|---|
| Best practices for building the ToL | Increasing the number of species represented in the ToL while ensuring reproducibility and encouraging community participation | Natural history museums must be central players for providing and analyzing genome-scale data. Genetic resources and specimen collections are fundamental for allowing the acquisition of data for extinct and poorly accessible species. Open access community initiatives must continue to be relevant repositories of the ToL. Adequate methods for curation of data and analytical tools must continue to be a high priority. |